

# SOLUTIONS MANUAL

FOR

Korosteleva, O. (2018). *Advanced Regression Models with SAS and R*, CRC Press

By

OLGA KOROSTELEVA

Department of Mathematics and Statistics

California State University, Long Beach

## TABLE OF CONTENTS

CHAPTER 1 .....	3
CHAPTER 2 .....	24
CHAPTER 3 .....	58
CHAPTER 4 .....	92
CHAPTER 5 .....	131
CHAPTER 6 .....	163
CHAPTER 7 .....	187
CHAPTER 8 .....	218
CHAPTER 9 .....	284
CHAPTER 10 .....	315

# CHAPTER 1

**EXERCISE 1.1.** Show that the normal distribution belongs to the exponential family of distributions.

$$f(y, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\} = \exp\left\{-\frac{1}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y^2 - 2y\mu + \mu^2)\right\}. \text{ Let } \theta = \mu$$

$$\text{and } \phi = \sigma^2. \text{ Then, we can write } f(y, \theta, \phi) = \exp\left\{-\frac{1}{2}\ln(2\pi\phi) - \frac{1}{2\phi}(y^2 - 2y\theta + \theta^2)\right\}$$

$$= \exp\left\{\frac{y\theta - \frac{\theta^2}{2}}{\phi} - \frac{1}{2}\ln(2\pi\phi) - \frac{y^2}{2\phi}\right\} = \exp\left\{\frac{y\theta - c(\theta)}{\phi} + h(y, \phi)\right\} \text{ where } c(\theta) = \frac{\theta^2}{2}, \text{ and}$$

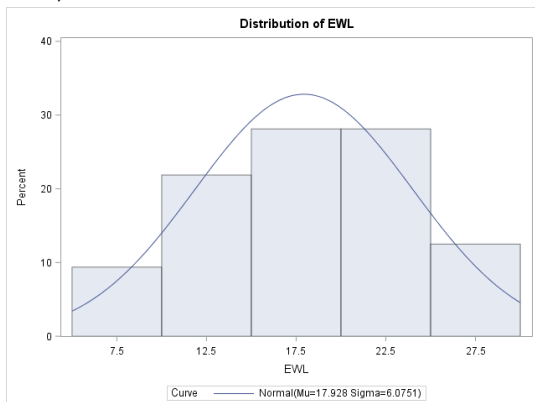
$$h(y, \phi) = -\frac{1}{2}\ln(2\pi\phi) - \frac{y^2}{2\phi}.$$

**EXERCISE 1.2.** (a) Verify the normality of the response variable, then fit the linear regression model to the data. State the fitted model. Give estimates for all parameters.

In SAS:

```
data weightloss;
input drug$ age gender$ EWL @@;
cards;
A 49 F 14.2  A 54 M 25.4  A 37 F 14.1  A 43 F 20.0  A 57 M 11.7 A 48 M 16.6
A 34 F 15.9  A 51 F 17.4  A 54 F 22.8  A 45 F 16.7  A 36 M 12.7 A 57 M 15.0
A 44 M 8.4   A 56 M 11.2  A 44 M 17.3  A 47 M 20.5  A 44 F 6.7  B 52 F 29.4
B 51 M 21.9  B 44 F 23.6  B 53 F 23.8  B 55 M 7.4   B 30 F 23.1 B 47 M 16.8
B 26 M 14.1  B 56 F 24.6  B 28 F 17.8  B 34 M 27.8  B 43 M 10.6 B 55 M 26.8
B 52 F 15.7  B 54 F 23.7
;

/*running normality check*/
proc univariate;
var EWL;
histogram/normal;
run;
```



### Goodness-of-Fit Tests for Normal Distribution

Test	Statistic	p Value
Kolmogorov-Smirnov D	0.10216310	Pr > D >0.150
Cramer-von Mises	W-Sq 0.05103595	Pr > W-Sq >0.250
Anderson-Darling	A-Sq 0.28788730	Pr > A-Sq >0.250

Based on the large p-values of the normality tests and the histogram, we can conclude that the response variable follows a normal distribution.

```
/*fitting general linear model*/
proc genmod;
  class drug(ref="A") gender;
  model EWL = drug age gender / dist=normal link=identity;
run;
```

Log Likelihood -98.4395

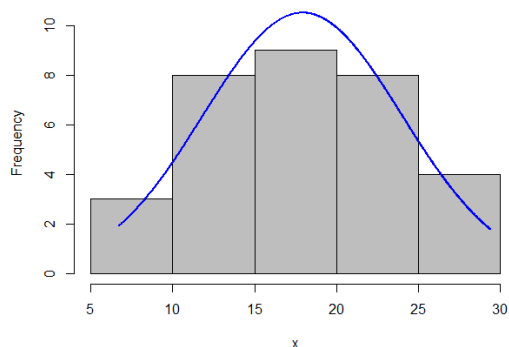
Analysis of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	9.2146	5.3301	-1.2322	19.6614	2.99	0.0838
drug	B 1	4.8103	1.8697	1.1456	8.4749	6.62	0.0101
drug	A 0	0.0000	0.0000	0.0000	0.0000	.	.
age	1	0.1102	0.1067	-0.0988	0.3192	1.07	0.3015
gender	F 1	2.7235	1.8664	-0.9346	6.3815	2.13	0.1445
gender	M 0	0.0000	0.0000	0.0000	0.0000	.	.
Scale	1	5.2451	0.6556	4.1054	6.7012		

The fitted model is  $\hat{E}(EWL) = 9.2146 + 4.8103 \cdot \text{drug}B + 0.1102 \cdot \text{age} + 2.7235 \cdot \text{female}$ , and  $\hat{\sigma} = 5.2451$ .

In R:

```
weightloss.data<- read.csv(file="C:/./Exercisel.2Data.csv", header = TRUE, sep =
",")

#running normality check
library(rcompanion)
plotNormalHistogram(weightloss.data$EWL)
```



```
shapiro.test(weightloss.data$EWL)
```

### Shapiro-wilk normality test

**w = 0.97424, p-value = 0.6234**

```
#specifying reference levels
drug.rel<- relevel(weightloss.data$drug, ref="A")
gender.rel<- relevel(weightloss.data$gender, ref="M")

#fitting general linear model
summary(fitted.model<- glm(EWL ~ drug.rel + age + gender.rel, data =
weightloss.data, family=gaussian(link=identity)))
```

### Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	9.2146	5.6981	1.617	0.1171
drug.relB	4.8103	1.9988	2.407	0.0229
age	0.1102	0.1140	0.967	0.3420
gender.relF	2.7235	1.9952	1.365	0.1831

```
#outputting estimated sigma
sigma(fitted.model)
```

**5.607257**

(b) Which regression coefficients turn out to be significant at 5%? Discuss the goodness of fit of the model.

Drug B is the only significant predictor in the model at the 5% significance level since the corresponding p-value is the only one under 0.05.

In SAS:

```
/*checking model fit*/
proc genmod;
  model EWL = / dist=normal link=identity;
run;
```

**Log Likelihood -102.6326**

```
data deviance_test;
  deviance = -2*(-102.6326 - (-98.4395));
  pvalue = 1 - probchi(deviance,3);
run;
```

```
proc print noobs;
run;
```

deviance	pvalue
8.3862	0.038669

The p-value for the deviance test is less than 0.05, indicating a good fit for the model. The R code and output are:

```
#checking model fit
null.model<- glm(EWL ~ 1, data=weightloss.data, family=gaussian(link=identity))
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))
```

8.386158

```
print(p.value<- pchisq(deviance, df=3, lower.tail=FALSE))
```

0.03867005

(c) Is one of the drugs more efficient for weight loss than the other? Interpret all estimated significant coefficients.

The estimated average EWL for subjects taking drug B is 4.8103 percent higher than that for subjects taking drug A, keeping all the other predictors fixed. It means that drug B is more efficient than drug A.

(d) According to the model, what is the predicted percent decrease in excess body weight for a 35-year-old male who is taking drug A?

The predicted percent decrease in excess body weight for a 35-year-old male who is taking drug A is computed by hand as:  $EWL^0 = 9.2146 + 0.1102 \cdot 35 = 13.0716$ .

In SAS:

```
/*using fitted model for prediction*/
data predict;
  input drug$ age gender$;
cards;
A 35 M
;

data weightloss;
  set weightloss predict;
run;

proc genmod;
  class drug gender;
  model EWL = drug age gender / dist=normal link=identity;
  output out=outdata p=pEWL;
run;

proc print data=outdata (firstobs=33) noobs;
var pEWL;
run;
```

```
      pEWL
13.0718
```

In R:

```
#using fitted model for prediction
print(predict(fitted.model, data.frame(drug.rel="A", age=35, gender.rel="M")))
```

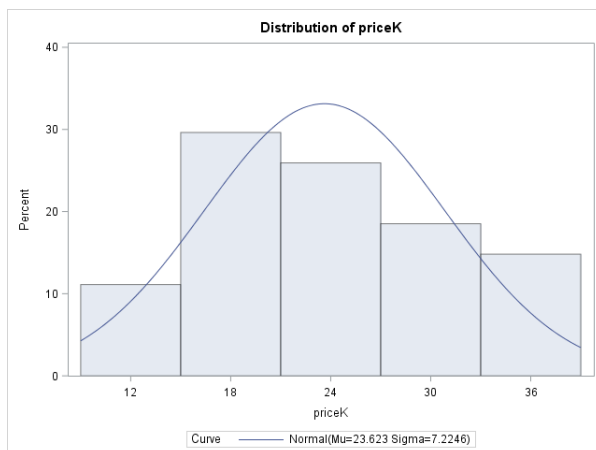
13.7178

**EXERCISE 1.3.** (a) Reduce the car price by a factor of 1000. Check that the distribution of the price is normal. Fit a general linear regression model to predict the price of a car. Write down the fitted model, specifying all estimated parameters.

In SAS:

```
data carsales;
input bodystyle$ 1-9 country$ hwy doors leather$ price @@;
priceK=price/1000;
cards;
coupe      USA      26 4 no 17445 coupe      USA      40 4 no 23500
coupe      USA      35 2 no 19600 coupe      Germany 37 4 no 23400
coupe      Germany 25 4 no 24100 coupe      Germany 24 2 no 12400
coupe      Japan   26 2 no 13300 coupe      Japan   27 4 no 15550
coupe      Japan   20 4 yes 29345 hatchback USA      30 2 no 12540
hatchback USA      39 4 no 17595 hatchback USA      38 2 no 17300
hatchback Germany 38 4 no 17800 hatchback Germany 32 4 no 22500
hatchback Germany 34 4 no 20300 hatchback Japan   38 4 yes 27300
hatchback Japan   38 2 yes 23300 hatchback Japan   38 2 yes 29300
sedan      USA      29 4 no 32000 sedan      USA      25 2 yes 34200
sedan      USA      33 4 yes 33395 sedan      Germany 40 4 no 22850
sedan      Germany 23 2 yes 36000 sedan      Germany 25 4 no 19900
sedan      Japan   40 4 yes 36700 sedan      Japan   35 4 yes 31600
sedan      Japan   37 4 no 24600
run;

/*running normality check*/
proc univariate;
var priceK;
histogram/normal;
run;
```



#### Goodness-of-Fit Tests for Normal Distribution

Test	Statistic	p value
Kolmogorov-Smirnov D	0.11287889	Pr > D > 0.150
Cramer-von Mises	W-Sq 0.05867848	Pr > W-Sq > 0.250
Anderson-Darling	A-Sq 0.37263698	Pr > A-Sq > 0.250

P-values for the normality tests are all in excess of 0.05, indicating that normality holds. The histogram also displays a distribution close to bell-shaped.

```
/*fitting general linear model*/
proc genmod;
  class bodystyle(ref="hatchback") country(ref="Japan") leather(ref="no");
  model priceK=bodystyle country hwy doors leather/dist=normal link=identity;
run;
```

Log Likelihood -67.2613

Analysis of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	wald	95% Confidence Limits	Chi-Square	Pr > ChiSq
Intercept	1	5.1353	4.6900	-4.0570	14.3276	1.20	0.2735
bodystyle coupe	1	2.2698	1.6836	-1.0301	5.5696	1.82	0.1776
bodystyle sedan	1	6.4107	1.5477	3.3772	9.4441	17.16	<.0001
bodystyle hatchback	0	0.0000	0.0000	0.0000	0.0000	.	.
country Germany	1	3.1959	1.6859	-0.1085	6.5002	3.59	0.0580
country USA	1	3.2128	1.5780	0.1199	6.3058	4.15	0.0418
country Japan	0	0.0000	0.0000	0.0000	0.0000	.	.
hwy	1	0.1305	0.1117	-0.0884	0.3494	1.36	0.2427
doors	1	1.5554	0.6630	0.2560	2.8549	5.50	0.0190
leather yes	1	12.1757	1.6217	8.9972	15.3541	56.37	<.0001
leather no	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale	1	2.9219	0.3976	2.2378	3.8150		

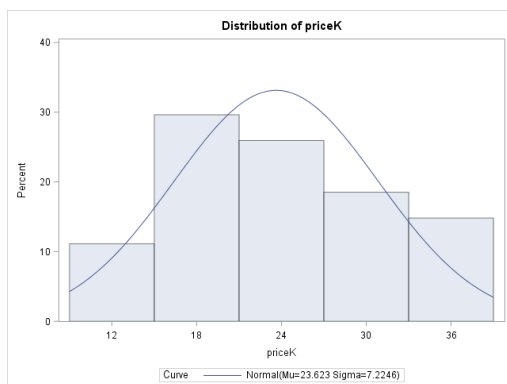
The fitted model is  $\hat{E}(\text{priceK}) = 5.1353 + 2.2698 \cdot \text{coupe} + 6.4107 \cdot \text{sedan} + 3.1959 \cdot \text{Germany} + 3.2128 \cdot \text{USA} + 0.1305 \cdot \text{hwy} + 1.5554 \cdot \text{doors} + 12.1757 \cdot \text{leather}$ , and  $\hat{\sigma} = 2.9219$ .

In R:

```
carsales.data<- read.csv(file="C:/./Exercisel.3Data.csv",header=TRUE, sep=",")

#rescaling price
priceK<- carsales.data$price/1000

#running normality check
library(rcompanion)
plotNormalHistogram(priceK)
```



```
shapiro.test(priceK)
Shapiro-wilk normality test

W = 0.95482, p-value = 0.28
```

```
#specifying reference levels
bodystyle.rel<- relevel(carsales.data$bodystyle, ref="hatchback")
country.rel<- relevel(carsales.data$country, ref="Japan")
leather.rel<- relevel(carsales.data$leather, ref="no")

#fitting general linear model
summary(fitted.model<- glm(priceK ~ bodystyle.rel + country.rel + hwy + doors +
leather.rel, data=carsales.data, family=gaussian(link=identity)))
```

**Coefficients:**

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	5.1353	5.5909	0.919	0.36986
bodystyle.relcoupe	2.2698	2.0070	1.131	0.27216
bodystyle.relsedan	6.4107	1.8450	3.475	0.00254
country.relGermany	3.1959	2.0098	1.590	0.12829
country.relUSA	3.2128	1.8812	1.708	0.10394
hwy	0.1305	0.1332	0.980	0.33937
doors	1.5554	0.7904	1.968	0.06384
leather.relyes	12.1757	1.9332	6.298	4.79e-06

```
#outputting estimated sigma
sigma(fitted.model)
```

**3.483088**

(b) How good is the model fit? Discuss the significance of the regression coefficients.

The p-value in the deviance test is way below 0.05, indicating a good model fit. Significant variables are sedan body style and leather interior.

**In SAS:**

```
/*checking model fit*/
proc genmod;
  model priceK = / dist=normal link=identity;
run;
```

**Log Likelihood -91.1942**

```
data deviance_test;
  deviance = -2*(-91.1942 - (-67.2613));
  pvalue = 1 - probchi(deviance,7);
run;
```

```
proc print noobs;
run;
```

deviance	pvalue
47.8658	3.7823E-8

**In R:**

```
#checking model fit
null.model<- glm(priceK ~ 1, data=carsales.data, family=gaussian(link=identity))
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))
47.86586

print(p.value<- pchisq(deviance, df=7, lower.tail = FALSE))
```

3.78218e-08

(c) Interpret the estimates of those regression coefficients that differ significantly from zero.

As estimated, a sedan costs on average \$6,410.70 more than a hatchback, under all other equal conditions. The estimated average price of a car with a leather interior is \$12,175.70 larger compared to a car without a leather interior.

(d) What is the predicted price of a sedan made in the USA that has 4 doors, leather seats, and runs 30 mpg on the highway?

The predicted price of a sedan that is made in the USA, has 4 doors, leather seats, and runs 30 mpg on the highway is calculated as:  $price^0 = \$1,000(5.1353 + 6.4107 + 3.2128 + 0.1305 \cdot 30 + 1.5554 \cdot 4 + 12.1757) = \$37,071.10$ .

In SAS:

```
/*using fitted model for prediction*/
data predict;
input bodystyle$ country$ hwy doors leather$;
cards;
sedan USA 30 4 yes
;

data carsales;
  set carsales predict;
run;

proc genmod;
  class bodystyle country leather;
  model priceK = bodystyle country hwy doors leather / dist=normal link=identity;
  output out=outdata p=ppriceK;
run;

data final_prediction;
set outdata;
pprice=ppriceK*1000;
run;

proc print data=final_prediction (firstobs=28) noobs;
  var pprice;
run;

  pprice
37071.14
```

In R:

```
#using fitted model for prediction
prediction<- (predict(fitted.model, data.frame(bodystyle.rel="sedan", country.rel
="USA", hwy=30, doors=4, leather.rel="yes")))
print(prediction*1000)

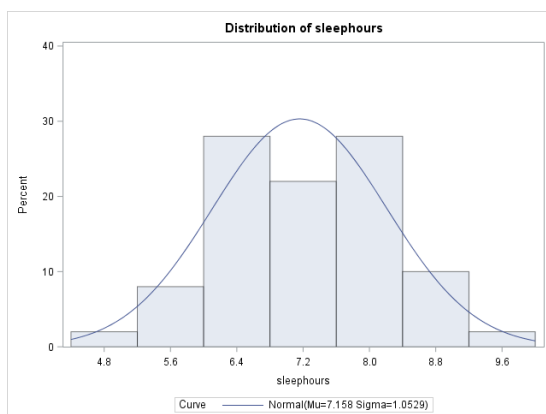
37071.14
```

**EXERCISE 1.4.** (a) Show normality of the distribution of the number of hours of sleep per night. Regress the number of hours of sleep on all the given factors. Write explicitly what the fitted model is.

In SAS:

```
data sleep;
input age gender$ quiettime nchildren stresslevel jobstatus$ nactivities pastvac
sleephours @@;
cards;
62 F 60 1 5 unempl 1 15 7.7 28 F 15 1 6 unempl 5 11 5.3
50 M 15 0 5 unempl 1 19 6.4 36 M 60 1 6 full 1 21 7.7
56 F 50 0 3 part 4 5 7.6 48 M 180 0 5 full 0 6 6.4
55 M 40 0 8 full 8 23 7.0 26 F 80 0 7 student 9 8 8.3
44 M 180 1 3 part 6 20 9.6 49 F 5 0 7 unempl 5 15 5.5
29 M 60 2 5 student 5 7 7.7 56 M 10 1 4 unempl 4 17 5.7
46 F 40 1 7 part 3 3 7.4 41 F 5 2 6 full 9 10 6.2
22 M 15 0 8 full 4 3 6.3 36 F 45 2 5 part 8 14 7.5
54 F 120 1 8 part 7 10 8.5 42 F 60 3 1 full 9 11 6.3
58 F 5 1 7 full 1 17 5.3 33 M 100 2 1 full 9 5 8.3
50 F 2 2 6 full 3 12 5.1 59 M 30 2 5 full 2 6 6.9
32 M 30 1 8 full 5 9 6.9 50 M 60 2 8 part 8 13 8.0
56 F 10 0 3 unempl 7 7 6.1 42 F 240 0 1 part 8 21 8.8
58 F 10 2 7 full 9 4 6.2 57 F 15 1 6 full 2 16 6.3
30 F 30 0 2 full 8 9 8.3 54 M 20 2 8 full 6 7 6.5
57 M 45 2 4 full 7 18 7.5 45 F 120 0 9 part 2 13 6.6
33 F 40 1 6 unempl 9 24 7.0 56 F 120 0 5 part 2 20 8.7
59 F 60 2 9 part 4 19 8.1 41 M 60 2 3 student 2 3 7.5
62 M 40 0 1 unempl 0 2 8.6 29 M 15 1 7 unempl 3 20 6.3
34 F 30 0 7 unempl 9 0 6.6 32 F 20 3 7 unempl 2 8 7.8
46 F 20 2 3 unempl 9 18 7.9 45 M 60 0 2 unempl 0 22 9.0
23 M 45 0 6 part 4 12 7.6 38 M 60 4 5 full 3 5 7.8
45 M 30 0 5 unempl 9 7 6.8 63 F 40 0 6 unempl 5 5 7.3
27 F 120 0 4 student 1 16 7.3 30 F 45 0 7 part 8 10 7.7
34 F 5 3 6 full 0 4 6.0 62 M 10 0 10 part 8 11 6.0
;

/*running normality check*/
proc univariate;
var sleephours;
histogram/normal;
run;
```



# Goodness-of-Fit Tests for Normal Distribution

Test	Statistic	p Value
Kolmogorov-Smirnov D	0.08733974	Pr > D >0.150
Cramer-von Mises	W-Sq 0.06145088	Pr > W-Sq >0.250
Anderson-Darling	A-Sq 0.32815950	Pr > A-Sq >0.250

The normality tests (p-values > 0.05) as well as the bell-shaped histogram indicate the normality of the response variable.

```
/*fitting general linear model*/
proc genmod;
  class gender(ref="F") jobstatus(ref="full");
  model sleepphours = age gender quiettime nchildren stresslevel jobstatus
    nactivities pastvac / dist=normal link=identity;
run;
```

Log Likelihood -54.6201

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	95% Confidence Limits	Pr > ChiSq	ChiSq
Intercept	1	6.8260	0.7051	5.4440	8.2080	93.72	<.0001
age	1	-0.0037	0.0093	-0.0218	0.0145	0.16	0.6932
gender M	1	0.3568	0.2132	-0.0610	0.7747	2.80	0.0942
gender F	0	0.0000	0.0000	0.0000	0.0000	.	.
quiettime	1	0.0074	0.0029	0.0018	0.0130	6.74	0.0095
nchildren	1	0.1204	0.1086	-0.0925	0.3334	1.23	0.2677
stresslevel	1	-0.1398	0.0536	-0.2450	-0.0347	6.80	0.0091
jobstatus part	1	1.0484	0.3188	0.4235	1.6732	10.81	0.0010
jobstatus student	1	0.6286	0.4358	-0.2255	1.4828	2.08	0.1492
jobstatus unempl	1	0.3818	0.2857	-0.1781	0.9418	1.79	0.1814
jobstatus full	0	0.0000	0.0000	0.0000	0.0000	.	.
nactivities	1	0.0204	0.0345	-0.0472	0.0879	0.35	0.5545
pastvac	1	0.0050	0.0170	-0.0282	0.0383	0.09	0.7663
scale	1	0.7214	0.0721	0.5930	0.8776		

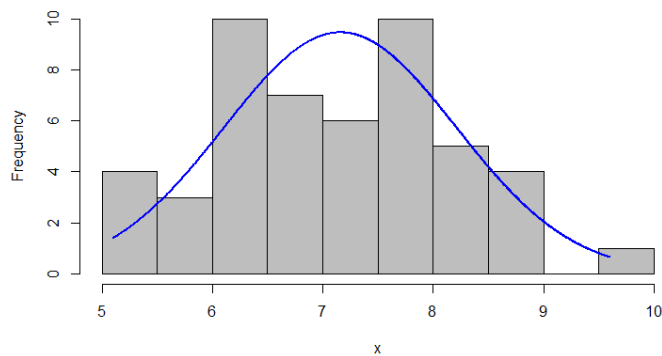
The fitted model is

$$\hat{E}(\text{sleepphours}) = 6.8260 - 0.0037 \cdot \text{age} + 0.3568 \cdot \text{male} + 0.0074 \cdot \text{quiettime} + 0.1204 \cdot \text{nchildren} - 0.1398 \cdot \text{stresslevel} + 1.0484 \cdot \text{parttime} + 0.6286 \cdot \text{student} + 0.3818 \cdot \text{unempl} + 0.0204 \cdot \text{nactivities} + 0.0050 \cdot \text{pastvac}, \text{ and } \hat{\sigma} = 0.7214.$$

In R:

```
sleep.data<- read.csv(file="C:/./Exercisel.4Data.csv", header=TRUE, sep=",")

#running normality check
library(rcompanion)
plotNormalHistogram(sleep.data$sleepphours)
```



```
shapiro.test(sleep.data$sleephours)
```

### Shapiro-wilk normality test

w = 0.98284, p-value = 0.6762

```
#specifying reference levels
gender.rel<- relevel(sleep.data$gender, ref="F")
jobstatus.rel<- relevel(sleep.data$jobstatus, ref="full")

#fitting general linear model
summary(fitted.model<- glm(sleephours ~ age + gender.rel + quiettime + nchildren
+ stresslevel + jobstatus.rel + nactivities + pastvac, data=sleep.data,
family=gaussian(link=identity)))
```

### Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	6.826002	0.798388	8.550	1.78e-10
age	-0.003656	0.010494	-0.348	0.72943
gender.relM	0.356815	0.241401	1.478	0.14741
quiettime	0.007421	0.003238	2.292	0.02738
nchildren	0.120419	0.123020	0.979	0.33368
stresslevel	-0.139828	0.060734	-2.302	0.02674
jobstatus.relpart	1.048386	0.360976	2.904	0.00603
jobstatus.relstudent	0.628623	0.493437	1.274	0.21021
jobstatus.relnempl	0.381840	0.323501	1.180	0.24501
nactivities	0.020373	0.039031	0.522	0.60465
pastvac	0.005046	0.019222	0.263	0.79430

```
#outputting estimated sigma
sigma(fitted.model)
```

0.8168443

(b) How good is the model fit? What beta coefficients are significantly different from zero at the 5% level of significance?

The p-value in the deviance test is smaller than 0.05, which indicates a good fit for the model. Significant variables at the 5% level are quiet time, stress level, and part-time employment status.

### In SAS:

```
/*checking model fit*/
proc genmod;
  model sleephours = / dist=normal link=identity;
run;
```

Log Likelihood -73.0195

```
data deviance_test;
  deviance = -2*(-73.0195 - (-54.6201));
  pvalue = 1 - probchi(deviance,10);
run;

proc print noobs;
run;
```

deviance	pvalue
36.7988	.000061312

In R:

```
#checking model fit
null.model<- glm(sleephours ~ 1, data=sleep.data, family=gaussian(link=identity))
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))
```

36.79887

```
print(p.value<- pchisq(deviance, df=10, lower.tail = FALSE))
```

6.131066e-05

(c) Interpret the estimated significant regression coefficients.

It is estimated that for each extra minute of quiet time, a person would get on average 0.0074 hours more sleep per night. For a unit increase in stress level, the estimated average number of hours of night sleep decrease by 0.1398. It is estimated that, on average, someone working part-time would get 1.0484 more hours of sleep compared to someone who is working full-time.

(d) Find the estimated number of hours of night's sleep that a 30-year-old full-time mom of three children under the age of five has, if she gets 10 minutes a day for herself, walks to the park with her kids every day of the week, estimates her stress level as 7, and who hasn't gotten any vacation for one year.

Below we calculate the predicted number of hours of night's sleep that a 30-year-old full-time mom of three children under the age of five has, if she gets 10 minutes a day for herself, walks to the park with her kids every day of the week, estimates her stress level as 7, and who hasn't gotten any vacation for one year.

$$\text{sleephours}^0 = 6.8260 - 0.0037 \cdot 30 + 0.0074 \cdot 10 + 0.1204 \cdot 3 - 0.1398 \cdot 7 + 0.0204 \cdot 7 + 0.0050 \cdot 12 = 6.3744.$$

In SAS:

```
/*using fitted model for prediction*/
data predict;
input age gender$ quiettime nchildren stresslevel jobstatus$ nactivities pastvac;
cards;
30 F 10 3 7 full 7 12
;

data sleep;
```

```

set sleep predict;
run;

proc genmod;
  class gender jobstatus;
  model sleepphours = age gender quiettime nchildren stresslevel jobstatus
    nactivities pastvac / dist=normal link=identity;
  output out=outdata p=psleepphours;
run;

proc print data=outdata (firstobs=51) noobs;
  var psleepphours;
run;

psleepphours
  6.37616

```

In R:

```

#using fitted model for prediction
print(predict(fitted.model, data.frame(age=30, gender.rel="F", quiettime=10,
nchildren=3, stresslevel=7, jobstatus.rel="full", nactivities=7, pastvac=12)))

6.376164

```

**EXERCISE 1.5.** (a) Compute the total time spent on both transitions. Verify the normality of the distribution of this variable, and fit a general linear regression model. Specify the fitted model.

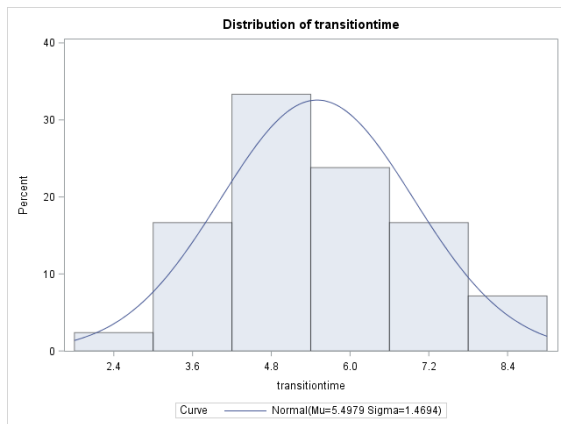
In SAS:

```

data time;
input age gender$ run t1 bike t2 swim @@;
  transitiontime=t1+t2;
cards;
55 M 24.17 2.60 37.95 2.50 5.70 59 F 34.88 2.83 52.15 3.05 5.20
24 M 32.97 2.55 59.20 3.47 5.37 53 F 22.2 1.83 46.70 2.15 5.50
51 M 27.35 1.75 42.05 2.32 3.75 38 F 32.13 2.38 50.92 2.95 6.00
66 M 25.39 1.95 41.57 2.80 3.93 30 F 24.67 1.58 48.28 2.77 5.68
43 F 42.33 2.78 63.60 4.08 7.18 47 F 28.73 2.35 45.57 3.90 6.62
26 F 29.62 2.92 51.23 3.85 4.92 45 M 22.23 2.07 38.95 2.35 4.28
29 F 26.93 2.10 44.33 2.45 7.47 34 M 17.75 0.75 33.27 1.23 3.65
39 M 37.47 2.52 55.67 4.47 8.60 54 M 36.63 3.27 43.92 3.08 7.15
26 M 34.42 2.73 52.62 2.67 9.23 36 M 27.38 2.22 39.03 2.92 7.43
42 M 21.37 2.12 35.95 1.93 3.95 49 M 29.03 4.50 38.53 3.95 8.80
42 F 28.53 3.27 49.85 3.67 8.13 42 F 25.12 1.72 39.52 2.50 4.55
42 F 26.33 1.70 48.98 2.30 5.02 41 F 36.75 3.95 62.85 3.13 6.93
15 M 25.12 1.70 44.75 3.20 7.48 48 M 26.52 4.43 40.98 3.82 6.58
37 M 28.3 2.85 41.78 3.47 6.02 55 M 31.25 2.70 43.43 3.25 5.25
42 M 24.38 1.45 37.13 1.83 3.70 25 M 33.45 2.25 51.38 4.03 7.45
12 F 27.62 2.23 55.47 2.97 4.37 23 F 28.55 2.17 54.57 2.55 7.90
49 M 33.88 2.77 54.82 3.87 6.90 53 F 26.97 1.77 42.33 3.40 6.58
45 F 26.58 1.65 44.30 2.52 5.40 33 F 32.32 2.10 54.87 2.32 6.25
63 M 40.53 3.78 69.75 3.83 12.17 50 M 33.68 3.07 43.57 3.13 5.77
43 F 34.93 2.58 62.35 2.95 7.92 24 M 22.88 1.82 39.55 2.12 4.03
44 M 29.25 2.47 45.60 2.75 9.18 51 F 36.98 3.70 46.58 5.18 7.60
;

```

```
/*running normality check*/
proc univariate;
  var transitiontime;
  histogram/normal;
run;
```



### Goodness-of-Fit Tests for Normal Distribution

Test	Statistic	p Value
Kolmogorov-Smirnov D	0.07499320	Pr > D >0.150
Cramer-von Mises	W-Sq 0.03895414	Pr > W-Sq >0.250
Anderson-Darling	A-Sq 0.26390584	Pr > A-Sq >0.250

The p-values in the normality tests are above 0.05, which means that the response variable has a normal distribution. The histogram displays a bell-shaped curve, supporting the normality conclusion.

```
/*fitting general linear model*/
proc genmod;
  class gender;
  model transitiontime = age gender run bike swim / dist=normal link=identity;
run;
```

Log Likelihood -56.4150

Analysis of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	0.5293	1.0253	-1.4803	2.5388	0.27	0.6057
age	1	0.0067	0.0128	-0.0184	0.0318	0.27	0.6032
gender	F 1	0.0961	0.3256	-0.5421	0.7343	0.09	0.7679
gender	M 0	0.0000	0.0000	0.0000	0.0000	.	.
run	1	0.1964	0.0500	0.0985	0.2943	15.46	<.0001
bike	1	-0.0565	0.0328	-0.1207	0.0078	2.97	0.0849
swim	1	0.2475	0.1024	0.0468	0.4483	5.84	0.0156
Scale	1	0.9271	0.1012	0.7486	1.1481		

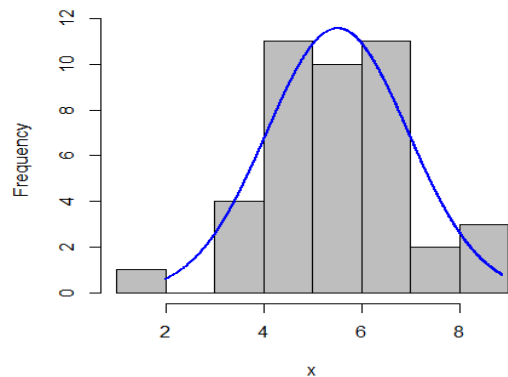
The fitted model is  $\hat{E}(\text{transitiontime}) = 0.5293 + 0.0067 \cdot \text{age} + 0.0961 \cdot \text{female} + 0.1964 \cdot \text{run} - 0.0565 \cdot \text{bike} + 0.2475 \cdot \text{swim}$ , and  $\hat{\sigma} = 0.9271$ .

In R:

```
time.data<- read.csv(file="C:/./Exercisel.5Data.csv", header=TRUE, sep=",")

#computing total transition time
transition.time<- time.data$t1 + time.data$t2

#running normality check
library(rcompanion)
plotNormalHistogram(transition.time)
```



```
shapiro.test(transition.time)
```

Shapiro-wilk normality test

w = 0.97896, p-value = 0.6216

```
#specifying reference levels
gender.rel<- relevel(time.data$gender, ref="M")
```

```
#fitting general linear model
summary(fitted.model<- glm(transition.time ~ age + gender.rel + run + bike +
swim, data=time.data, family=gaussian(link=identity)))
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	0.529266	1.107464	0.478	0.635605
age	0.006659	0.013837	0.481	0.633232
gender.relF	0.096094	0.351716	0.273	0.786250
run	0.196405	0.053953	3.640	0.000849
bike	-0.056487	0.035412	-1.595	0.119427
swim	0.247544	0.110615	2.238	0.031507

```
#outputting estimated sigma
sigma(fitted.model)
```

1.001351

(b) Discuss the model fit. Are all the predictors in that model significant at the 5% significance level?

In SAS:

```
/*checking model fit*/
proc genmod;
  model transitiontime = / dist=normal link=identity;
run;
```

Log Likelihood -74.6263

```
data deviance_test;
  deviance = -2*(-74.6263 - (-56.4150));
  pvalue = 1 - probchi(deviance,5);
run;
```

```
proc print noobs;
run;
```

deviance	pvalue
36.4226	.000000782

Since the p-value in the deviance test is tiny, the model has a good fit. The only significant predictors at the 5% level are run time and swim time.

In R:

```
#checking model fit
null.model<- glm(transition.time ~ 1, data=time.data,
family=gaussian(link=identity))
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))
```

36.42269

```
print(p.value<- pchisq(deviance, df=5, lower.tail = FALSE))
```

7.817128e-07

(c) Interpret only the estimated significant regression coefficients of this model.

The estimated average transition time increases by 0.1964 for a one-minute increase in run time. For a one-minute increase in swim time, the estimated average transition time increases by 0.2475.

(d) What is the predicted total time at transitions for the student, if his best result in the 5-kilometer run is 27:32, the 13-mile bike is 56:17, and the 200-meter swim is 8:46?

Below we compute the predicted time at transitions for the 25-year-old student with a 27:32 run, 56:17 bike, and 8:46 swim. First, we convert the times into minutes:  $27+32/60=27.53$ ,  $56+17/60=56.28$ , and  $8+46/60=8.77$ . The calculation is as follows:  $transitiontime^0 = 0.5293 + 0.0067 \cdot 25 + 0.1964 \cdot 27.53 - 0.0565 \cdot 56.28 + 0.2475 \cdot 8.77 = 5.09$ .

In SAS:

```
/*using fitted model for prediction*/
data predict;
input age gender$ run bike swim;
cards;
25 M 27.53 56.28 8.77
;
data time;
  set time predict;
run;

proc genmod;
```

```

class gender;
  model transitiontime = age gender run bike swim / dist=normal link=identity;
  output out=outdata p=ptransitiontime;
run;

proc print data=outdata (firstobs=43) noobs;
  var ptransitiontime;
run;

ptransitiontime
      5.09465

```

In R:

```

#using fitted model for prediction
print(predict(fitted.model, data.frame(age=25, gender.rel="M", run=27.53,
bike=56.28, swim=8.77)))

5.094653

```

**EXERCISE 1.6.** (a) Check that the measurements for the heart rate are coming from a normal distribution. Fit the regression model and specify all estimated parameters.

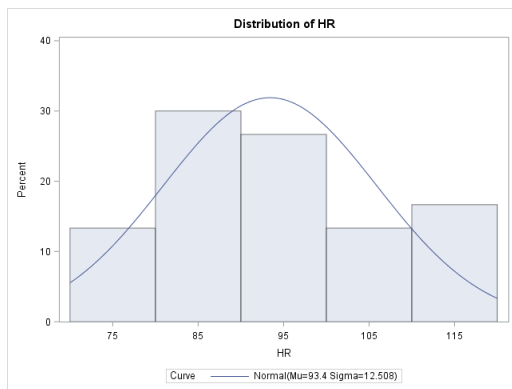
In SAS:

```

data heartrate;
length AQI $9.;
input age gender$ ethnicity$ BMI nmeds AQI$ HR @@;
cards;
48 F Black 29.9 0 good 76 56 F White 22.9 3 unhealthy 112
67 F White 23.4 1 good 94 82 M Black 29.7 0 good 92
64 F White 31.4 3 good 97 58 M White 18.9 2 moderate 79
72 F Black 25.2 0 moderate 114 70 F Black 25.9 1 moderate 115
54 M Hispanic 29.6 0 moderate 80 57 F Hispanic 20.2 2 good 81
50 F Black 23.9 1 unhealthy 97 59 F Hispanic 22.6 0 good 86
61 M Hispanic 32.8 1 good 84 69 M Hispanic 24.1 2 unhealthy 94
65 F Black 23.4 2 moderate 114 66 F Hispanic 27.8 3 good 82
74 M White 32.4 1 moderate 97 66 M Hispanic 22.9 2 good 86
53 M Hispanic 25.2 0 good 84 55 M Hispanic 24.6 0 moderate 94
73 F Hispanic 24.8 3 moderate 105 45 F Hispanic 19.0 2 unhealthy 83
71 F White 20.3 2 unhealthy 111 63 M Black 23.8 2 unhealthy 108
71 F White 21.5 2 moderate 100 62 M Hispanic 27.4 3 good 79
44 F Hispanic 17.2 0 unhealthy 86 49 M White 17.1 1 good 75
63 M Black 28.0 2 good 91 65 F Hispanic 22.2 1 moderate 106
;

/*running normality check*/
proc univariate;
  var HR;
  histogram/normal;
run;

```



### Goodness-of-Fit Tests for Normal Distribution

Test	Statistic	p value
Kolmogorov-Smirnov D	0.15627802	Pr > D 0.061
Cramer-von Mises	W-Sq 0.09496306	Pr > W-Sq 0.129
Anderson-Darling	A-Sq 0.65250988	Pr > A-Sq 0.084

Based on the histograms and the large p-values, we can conclude that the heart rate follows a normal distribution.

```
/*fitting generallinear model*/
proc genmod;
  class gender ethnicity(ref="Hispanic") AQI(ref="good");
  model HR = age gender ethnicity BMI nmeds AQI / dist=normal link=identity;
run;
```

Log Likelihood -96.2779

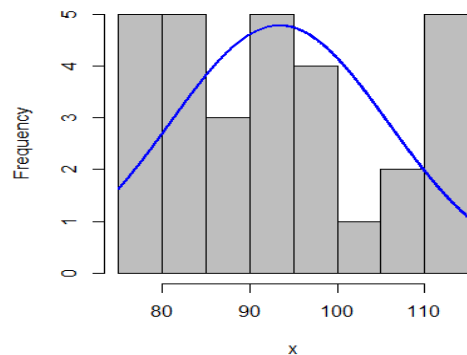
Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	wald	95% Confidence Limits	wald Chi-Square	Pr > ChiSq
Intercept		1	38.0164	10.2408	17.9449	58.0879	13.78	0.0002
age		1	0.6503	0.1472	0.3617	0.9389	19.51	<.0001
gender	F	1	7.1031	2.3608	2.4760	11.7303	9.05	0.0026
gender	M	0	0.0000	0.0000	0.0000	0.0000	.	.
ethnicity Black		1	7.5351	2.8956	1.8598	13.2104	6.77	0.0093
ethnicity white		1	2.2633	2.7895	-3.2041	7.7306	0.66	0.4172
ethnicity Hispanic		0	0.0000	0.0000	0.0000	0.0000	.	.
BMI		1	0.0431	0.3225	-0.5890	0.6751	0.02	0.8938
nmeds		1	0.4384	1.1919	-1.8976	2.7743	0.14	0.7130
AQI	moderate	1	10.8596	2.6942	5.5790	16.1402	16.25	<.0001
AQI	unhealthy	1	14.1674	3.1905	7.9142	20.4206	19.72	<.0001
AQI	good	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		1	5.9914	0.7735	4.6520	7.7165		

The fitted model is  $\hat{E}(HR) = 38.0164 + 0.6503 \cdot age + 7.1031 \cdot female + 7.5351 \cdot Black + 2.2633 \cdot White + 0.0431 \cdot BMI + 0.4384 \cdot nmeds + 10.8596 \cdot AQI_{moderate} + 14.1674 \cdot AQI_{unhealthy}$ , and  $\hat{\sigma} = 5.9914$ .

In R:

```
hr.data<- read.csv(file="C:/./Exercisel.6Data.csv", header=TRUE, sep=",")

#running normality check
library(rcompanion)
plotNormalHistogram(hr.data$HR)
```



```
shapiro.test(hr.data$HR)
```

**Shapiro-wilk normality test**

**w = 0.93047, p-value = 0.05054**

```
#specifying reference levels
gender.rel<- relevel(hr.data$gender, ref="M")
ethnicity.rel<- relevel(hr.data$ethnicity, ref="Hispanic")
AQI.rel<- relevel(hr.data$AQI, ref="good")

#fitting general linear model
summary(fitted.model<- glm(HR ~ age + gender.rel + ethnicity.rel + BMI + nmeds +
AQI.rel, data=hr.data, family=gaussian(link=identity)))
```

**Coefficients:**

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	38.01638	12.24005	3.106	0.00535
age	0.65033	0.17599	3.695	0.00134
gender.relF	7.10311	2.82173	2.517	0.02002
ethnicity.relBlack	7.53509	3.46094	2.177	0.04102
ethnicity.relWhite	2.26328	3.33411	0.679	0.50466
BMI	0.04306	0.38543	0.112	0.91210
nmeds	0.43836	1.42454	0.308	0.76133
AQI.relmoderate	10.85963	3.22023	3.372	0.00288
AQI.relunhealthy	14.16737	3.81333	3.715	0.00128

```
#outputting estimated sigma
sigma(fitted.model)
```

**7.161087**

(b) Discuss the goodness-of-fit of the model. What variables are significant predictors of heart rate at the 5% level of significance?

**In SAS:**

```
/*checking model fit*/
proc genmod;
  model HR = / dist=normal link=identity;
```

```
run;
```

Log Likelihood -117.8512

```
data deviance_test;
  deviance = -2*(-117.8512 - (-96.2779));
  pvalue = 1 - probchi(deviance,8);
run;
```

```
proc print noobs;
run;
```

deviance	pvalue
43.1466	.000000824

Since the p-value in the deviance test is tiny, the model has a good fit. The significant predictors at the 5% level are age, gender, ethnicity level Black, and both levels of AQI.

In R:

```
#checking model fit
null.model<- glm(HR ~ 1, data=hr.data, family=gaussian(link=identity))
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))
```

43.14658

```
print(p.value<- pchisq(deviance, df=8, lower.tail=FALSE))
```

8.243212e-07

(c) Give an interpretation of the estimated statistically significant regression coefficients.

As age increases by one year, the estimated average heart rate increases by 0.6503 beats per minute. The estimated average heart rate for females is 7.1031 beats per minute larger than that for males. The estimated average heart rate for Blacks is 7.5351 beats per minute larger than that for Hispanics. The estimated average heart rate for people living with moderate air quality is 10.8956 beats per minute larger than that for people living with good air quality. The estimated average heart rate for people living with moderate air quality is 14.1674 beats per minute larger than that for people living with good air quality.

(d) Compute the predicted heart rate of a 50-year-old Hispanic male who has a BMI of 20, is not taking any heart medications, and resides in an area with moderate air quality.

The predicted heart rate of a 50-year-old Hispanic male who has a BMI of 20, is not taking any heart medications, and resides in an area with moderate air quality is computed as follows:

$$HR^0 = 38.0164 + 0.6503 \cdot 50 + 0.0431 \cdot 20 + 10.8596 = 82.253.$$

In SAS:

```
/*using fitted model for prediction*/
data predict;
input age gender$ ethnicity$ BMI nmeds AQI$;
cards;
50 M Hispanic 20 0 moderate
;
```

```

data heartrate;
set heartrate predict;
run;

proc genmod;
  class gender ethnicity AQI;
  model HR = age gender ethnicity BMI nmeds AQI / dist=normal link=identity;
  output out=outdata p=pHR;
run;

proc print data=outdata (firstobs=31) noobs;
  var pHR;
run;

      pHR
82.2536

```

In R:

```

#using fitted model for prediction
print(predict(fitted.model, data.frame(age=50, gender.rel="M", ethnicity.rel="Hispanic", BMI=20, nmeds=0, AQI.rel="moderate")))

82.25361

```

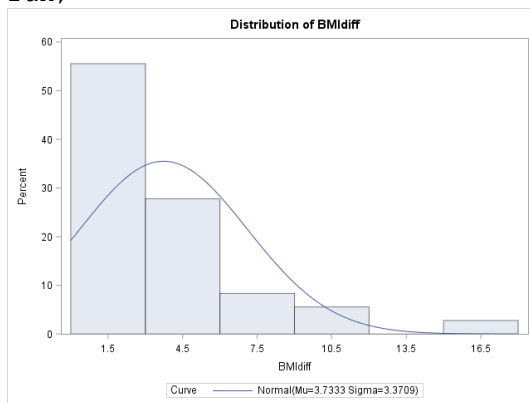
# CHAPTER 2

**EXERCISE 2.1.** (a) Is the decrease in BMI percentile (preBMI-postBMI) normally distributed? Plot a histogram and test for the normality of the distribution.

In SAS:

```
data obesity;
input gender$ age group$ preBMI postBMI @@;
    BMIdiff=preBMI-postBMI;
    female=(gender="F");
    control=(group="Cx");
cards;
F 6 Cx 85.7 83.8 F 6 Cx 93.8 92.9 F 7 Cx 93.5 92.5 F 8 Cx 90.1 89.8
F 9 Tx 92.3 90.7 F 9 Tx 90.3 88.3 F 12 Cx 87.6 85.9 F 12 Cx 87.2 84.1
F 12 Tx 96.9 94.9 F 12 Tx 85.8 81.2 F 13 Cx 96.7 94.1 F 13 Cx 93.5 92.9
F 13 Tx 92.3 87.5 F 13 Tx 85.3 83.7 F 14 Tx 95.5 78.7 F 15 Cx 91.3 89.9
F 15 Tx 95.8 87.1 F 16 Tx 90.7 87.2 M 6 Cx 92.6 88.1 M 7 Cx 95.8 94.7
M 7 Cx 90.4 89.1 M 7 Cx 91.2 88.6 M 8 Tx 94.4 87.8 M 8 Tx 93.2 87.3
M 10 Cx 93.9 91.5 M 10 Tx 96.2 91.1 M 10 Tx 89.4 87.9 M 11 Tx 86.2 77.1
M 11 Tx 95.4 84.8 M 12 Cx 97.7 95.8 M 13 Tx 85.3 80.0 M 13 Tx 86.2 82.4
M 14 Cx 85.5 83.6 M 14 Cx 97.8 93.8 M 16 Cx 95.0 93.6 M 16 Tx 93.1 86.8
;

/*running normality check of response*/
proc univariate;
var BMIdiff;
    histogram/normal;
run;
```



## Goodness-of-Fit Tests for Normal Distribution

Test	Statistic	p value
Kolmogorov-Smirnov D	0.18720025	Pr > D <0.010
Cramer-von Mises	W-Sq 0.36512474	Pr > W-Sq <0.005
Anderson-Darling	A-Sq 2.15289200	Pr > A-Sq <0.005

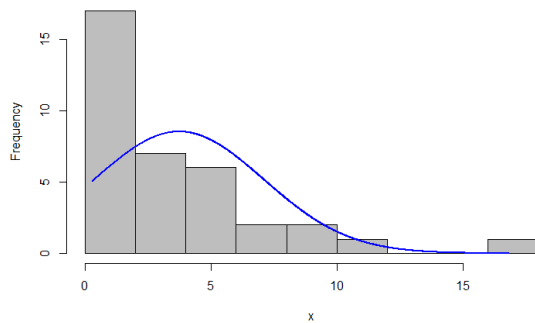
Neither the histogram nor the normality tests support the normality of the response. In fact, the distribution is right-skewed.

In R:

```
bmi.data<- read.csv(file="C:/./Exercise2.1Data.csv",header=TRUE, sep=",")

#creating the difference in BMI
BMIdiff<- bmi.data$preBMI-bmi.data$postBMI

#running normality check of response
library(rcompanion)
plotNormalHistogram(BMIdiff)
```



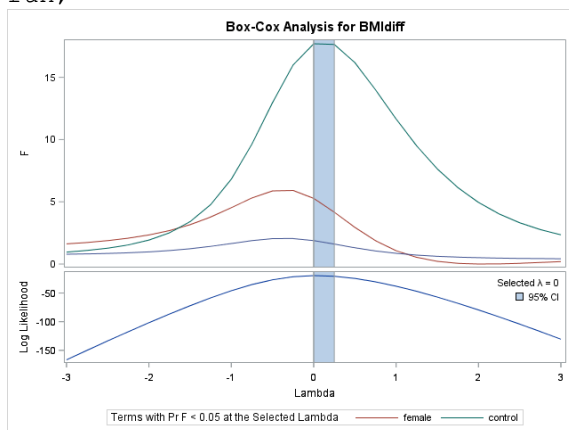
```
shapiro.test(BMIdiff)
```

Shapiro-wilk normality test  
w = 0.79159, p-value = 1.114e-05

- (b) Find the optimal lambda for Box-Cox transformation. Transform the change in BMI percentile (find the appropriate transformation in Table 2.1), and show that the transformed variable is normally distributed. Plot the histogram and do formal testing.

In SAS:

```
/*finding optimal lambda for Box-Cox transformation*/
proc transreg;
model BoxCox(BMIdiff) = identity(age female control);
run;
```



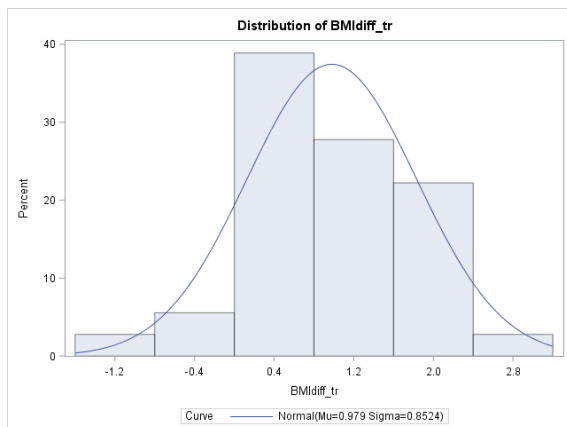
```
/*applying Box-Cox transformation with lambda=0*/
data obesity;
set obesity;
```

```

BMIdiff_tr=log(BMIdiff);
run;

/*running normality check of transformed response*/
proc univariate;
  var BMIdiff_tr;
  histogram/normal;
run;

```



#### Goodness-of-Fit Tests for Normal Distribution

Test	Statistic		p value	
Kolmogorov-Smirnov	D	0.10351850	Pr > D	>0.150
Cramer-von Mises	W-Sq	0.03884400	Pr > W-Sq	>0.250
Anderson-Darling	A-Sq	0.22534295	Pr > A-Sq	>0.250

The optimal lambda for the Box-Cox transformation is  $\lambda = 0$ , which corresponds to the log transformation. The log-transformed response variable has a normal distribution as backed up by a bell shape on the histogram and the large p-values in the normality tests.

In R:

```

#creating indicator variables
female<- relevel(bmi.data$gender, ref="M")
control<- relevel(bmi.data$group, ref="Tx")

#finding optimal lambda for Box-Cox transformation
library(MASS)

BoxCox.fit<- boxcox(BMIdiff ~ age + female + control, data=bmi.data, lambda =
seq(-3,3,1/4), interp = FALSE)
BoxCox.data<- data.frame(BoxCox.fit$x, BoxCox.fit$y)
ordered.data<- BoxCox.data[with(BoxCox.data, order(-BoxCox.fit.y)),]
ordered.data[1,]

BoxCox.fit.x BoxCox.fit.y

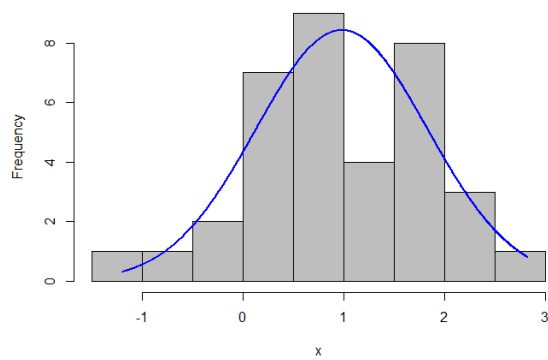
      0      -46.71658

#applying Box-Cox transformation with lambda=0
BMIdiff.tr<- log(BMIdiff)

#running normality check of transformed response
library(rcompanion)

```

```
plotNormalHistogram(BMIdiff.tr)
```



```
shapiro.test(BMIdiff.tr)
```

Shapiro-wilk normality test  
w = 0.9877, p-value = 0.9532

- (c) Fit the general regression model to the Box-Cox transformed change in BMI percentile. Does this model have a good fit?

In SAS:

```
/*fitting general linear model to transformed response*/
proc genmod;
  model BMIdiff_tr = age female control / dist=normal link=identity;
run;
```

Log Likelihood -33.2950

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	1.1438	0.4333	0.2945	1.9930	6.97	0.0083
age	1	0.0501	0.0344	-0.0174	0.1176	2.12	0.1457
female	1	-0.4986	0.2047	-0.8998	-0.0975	5.94	0.0148
control	1	-0.9384	0.2103	-1.3506	-0.5262	19.91	<.0001
Scale	1	0.6101	0.0719	0.4843	0.7687		

The fitted model is  $\hat{E}(\ln BMIdiff) = 1.1438 + 0.0501 \cdot age - 0.4986 \cdot female - 0.9384 \cdot control$ , and  $\hat{\sigma} = 0.6101$ .

```
/*checking model fit*/
proc genmod;
  model BMIdiff_tr = / dist=normal link=identity;
run;
```

Log Likelihood -44.8268

```
data deviance_test;
  deviance = -2*(-44.8268 - (-33.2950));
  pvalue = 1 - probchi(deviance, 3);
```

```
run;

proc print noobs;
run;

deviance      pvalue
23.0636      .000039169
```

Based on the small p-value in the deviance test, the model for the log-transformed response fits the data well.

In R:

```
#fitting general linear model to transformed response

summary(fitted.model<- glm(BMIdiff.tr ~ age + female + control, data=bmi.data,
family=gaussian(link=identity)))
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	1.14375	0.45959	2.489	0.018218
age	0.05008	0.03651	1.372	0.179731
femaleF	-0.49862	0.21708	-2.297	0.028317
controlCx	-0.93835	0.22307	-4.207	0.000195

```
#outputting estimated sigma
```

```
sigma(fitted.model)
```

0.6471448

```
#checking model fit
```

```
null.model<- glm(BMIdiff.tr ~ 1, family=gaussian(link=identity))
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))
```

23.06361

```
print(p.value<- pchisq(deviance, df=3, lower.tail = FALSE))
```

3.916872e-05

- (d) What predictors are significant at the 5% level? Write the interpretation of the estimated regression coefficients for the significant predictors only.

Gender and group are significant predictors. The estimated mean of log-transformed reduction in BMI percentile is 0.4986 units larger in females than in males. The estimated average log-transformed decrease in BMI percentile for the control group participants is 0.9384 units smaller than that for participants in the intervention group (conclusion: intervention works).

- (e) Predict the change in BMI percentile for a 9-year-old girl in the control group.

To predict the change in BMI percentile for a 9-year-old girl in the control group, we calculate:

$$BMIdiff^0 = \exp(1.1438 + 0.0501 \cdot 9 - 0.4986 - 0.9384) = 1.1708.$$

In SAS:

```

data predict;
input age female control;
cards;
9 1 1
;

data obesity;
set obesity predict;
run;

proc genmod;
  model BMIdiff_tr = age female control / dist=normal link=identity;
  output out=outdata p=pBMIdiff_tr;
run;

data outdata;
set outdata;
  pBMIdiff=exp(pBMIdiff_tr);
run;

proc print data=outdata(firstobs=37) noobs;
  var pBMIdiff;
run;

pBMIdiff
  1.17061

```

In R:

```

#using fitted model for prediction
pred.BMIdiff.tr<- predict(fitted.model, data.frame(female="F", age=9,
control="Cx"))
print(exp(pred.BMIdiff.tr))

1.170609

```

**EXERCISE 2.2.** (a) Construct a histogram of the score. Does the distribution look normal? Perform the test for normality. Draw conclusion.

In SAS:

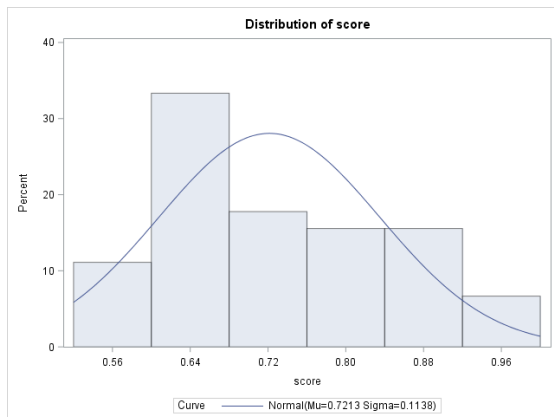
```

data QI;
length desgn $6.;
input desgn$ wrkyrs priorQI$ score @@;
  nurse=(desgn="nurse");
  doctor=(desgn="doctor");
  priorQIyes=(priorQI="yes");
  score=score/100;
cards;
nurse 16 yes 63  nurse  9  yes 93  nurse  8  yes 74
nurse  1 no 69  nurse  5  no 67  nurse  3  no 66
nurse 24 no 86  nurse  4  no 74  nurse  1  no 88
nurse 24 no 84  nurse  3  no 97  doctor 2  yes 88
doctor 5  yes 78  doctor 26 yes 82  doctor 3  no 57
doctor 3  no 88  doctor 15 no 78  doctor 4  no 65
doctor 25 no 78  staff  3  yes 62  staff 21 no 55
staff  8  no 62  staff 11 no 67  nurse  8  yes 62

```

```
nurse 22 yes 68 nurse 4 no 93 nurse 6 no 77
nurse 2 no 59 nurse 20 no 64 nurse 2 no 70
nurse 3 no 63 nurse 16 no 65 nurse 18 no 73
nurse 15 no 76 doctor 2 yes 85 doctor 7 yes 91
doctor 2 yes 69 doctor 20 no 66 doctor 13 no 55
doctor 8 no 62 doctor 14 no 61 staff 9 yes 57
staff 11 yes 69 staff 19 no 64 staff 17 no 76
;
```

```
/*running normality check of response*/
proc univariate;
  var score;
  histogram/normal;
run;
```



#### Goodness-of-Fit Tests for Normal Distribution

Test	Statistic	p Value
Kolmogorov-Smirnov D	0.14185097	Pr > D 0.023
Cramer-von Mises	W-Sq 0.14986260	Pr > W-Sq 0.023
Anderson-Darling	A-Sq 0.88141925	Pr > A-Sq 0.023

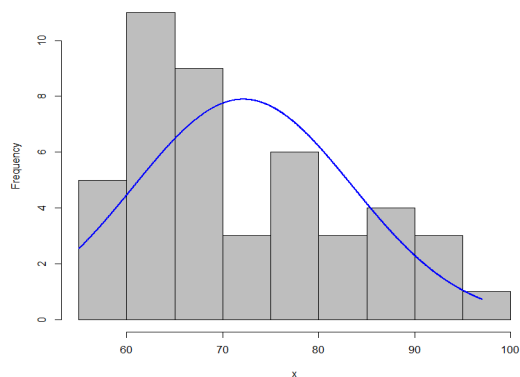
The histogram shows a right-skewed distribution. The normality tests support the conclusion that the distribution is not normal, since the p-values are below 0.05.

In R:

```
jobscore.data<- read.csv(file="C:/./Exercise2.2Data.csv", header=TRUE, sep=",")

#running normality check of response
library(rcompanion)

plotNormalHistogram(jobscore.data$score)
```



```
shapiro.test(jobscore.data$score)
```

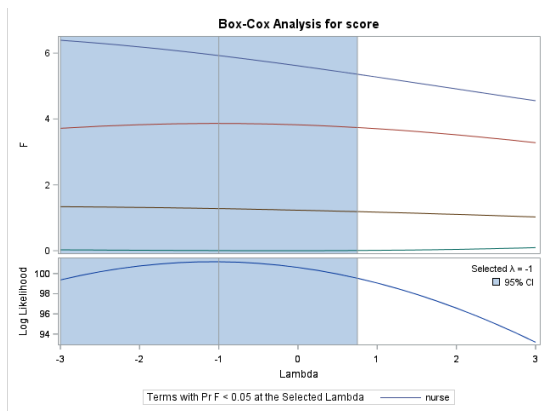
Shapiro-wilk normality test

$w = 0.94357$ ,  $p\text{-value} = 0.02913$

(b) Transform the score variable using a meaningful Box-Cox transformation and assure that it is now normally distributed by plotting the histogram and doing normality testing.

In SAS:

```
/*finding optimal lambda for Box-Cox transformation*/
proc transreg;
  model BoxCox(score) = identity(nurse doctor wrkyrs priorQIyes);
run;
```



The optimal lambda for the Box-Cox transformation is  $\lambda = -1$ , which corresponds to the inverse transformation. The inverse-transformed response variable has a normal distribution as can be seen from the histogram and the large p-values in the normality tests.

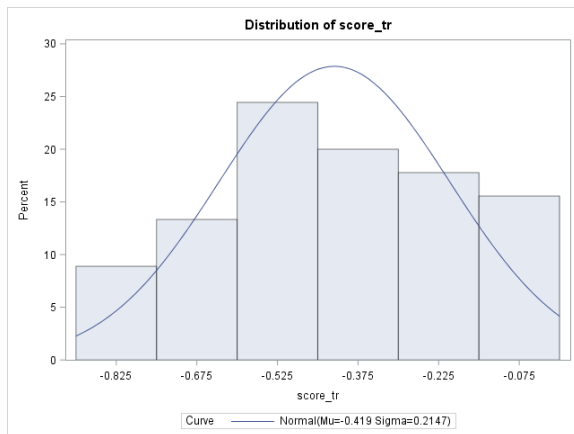
```
/*applying Box-Cox transformation with lambda=-1*/
data qi;
set qi;
  score_tr=1-(1/score);
run;

/*running normality check of transformed response*/
proc univariate;
  var score_tr;
```

```

  histogram/normal;
run;

```



### Goodness-of-Fit Tests for Normal Distribution

Test	Statistic	p Value
Kolmogorov-Smirnov D	0.08928583	Pr > D >0.150
Cramer-von Mises	W-Sq 0.07051371	Pr > W-Sq >0.250
Anderson-Darling	A-Sq 0.44867900	Pr > A-Sq >0.250

In R:

```

#creating indicator variables and rescaling score
desgn.rel<- relevel(jobscore.data$desgn, ref="staff")
priorQI.rel<- relevel(jobscore.data$priorQI, ref="no")
score<- jobscore.data$score/100

#finding optimal lambda for Box-Cox transformation
library(MASS)
BoxCox.fit<- boxcox(score ~ desgn.rel + wrkyrs + priorQI.rel,
data=jobscore.data, lambda=seq(-3,3,1/4), interp = FALSE)
BoxCox.data<- data.frame(BoxCox.fit$x, BoxCox.fit$y)
ordered.data<- BoxCox.data[with(BoxCox.data, order(-BoxCox.fit.y)),]
ordered.data[1,]

```

**BoxCox.fit.x BoxCox.fit.y**

**-1 2.940242**

```

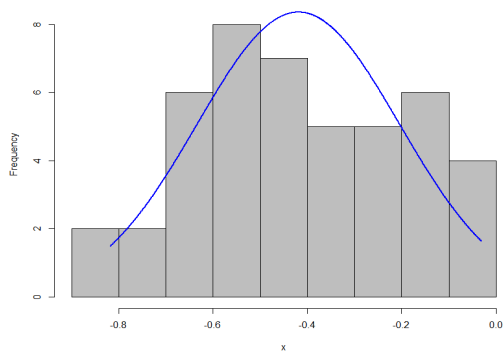
#applying Box-Cox transformation with lambda=-1
score.tr<- 1-(1/score)

```

```

#running normality check of transformed response
plotNormalHistogram(score.tr)

```



```
shapiro.test(score.tr)
```

Shapiro-wilk normality test

w = 0.96606, p-value = 0.2073

- (c) Run the general linear regression model on the transformed score. What predictors are significant at the 0.05 level?

In SAS:

```
/*fitting general linear model to transformed response*/
proc genmod;
  model score_tr = nurse doctor wrkyrs priorQIyes / dist=normal link=identity;
run;
```

Log Likelihood 9.5061

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	95% Confidence Limits	Pr > ChiSq
Intercept	1	-0.6093	0.0888	47.11	-0.7833 -0.4353	<.0001
nurse	1	0.2122	0.0822	6.67	0.0511 0.3733	0.0098
doctor	1	0.1799	0.0863	4.35	0.0108 0.3490	0.0371
wrkyrs	1	0.0002	0.0039	0.00	-0.0073 0.0078	0.9497
priorQIyes	1	0.0773	0.0644	1.44	-0.0490 0.2035	0.2304
Scale	1	0.1959	0.0206		0.1593 0.2408	

The fitted model is  $\hat{E}\left(1 - \frac{1}{score}\right) = -0.6093 + 0.2122 \cdot nurse + 0.1799 \cdot doctor + 0.0002 \cdot wrkyrs + 0.0773 \cdot priorQIyes$ , and  $\hat{\sigma} = 0.1959$ . Judging by p-values, nurse and doctor are significant predictors.

In R:

```
#fitting general linear model to transformed response
summary(fitted.model<- glm(score.tr ~ desg.n.rel + wrkyrs + priorQI.rel,
data=jobscore.data, family=gaussian(link = identity)))
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-0.609290	0.094150	-6.471	1.03e-07
desgn.reldoctor	0.179873	0.091507	1.966	0.0563
desgn.relnurse	0.212200	0.087173	2.434	0.0195

wrkysr	0.000243	0.004086	0.059	0.9529
priorQI.relyes	0.077263	0.068323	1.131	0.2649

```
#outputting estimated sigma
sigma(fitted.model)
```

0.2077762

- (d) Interpret the estimates of the significant beta coefficients. Does the model fit the data well?  
Conduct the chi-squared deviance test.

The estimated mean inverse transformed score for nurses is 0.2122 points above that for staff, and for doctors, it is 0.1799 points above that for staff.

The model doesn't really fit the data well, as seen from the large p-value of the deviance test given below.

In SAS:

```
/*checking model fit*/
proc genmod;
  model score_tr = / dist=normal link=identity;
run;
```

Log Likelihood 5.8777

```
data deviance_test;
  deviance = -2*(5.8777 - 9.5061);
  pvalue = 1 - probchi(deviance,4);
run;
```

```
proc print noobs;
run;
```

deviance	pvalue
7.2568	0.12292

In R:

```
#checking model fit
null.model<- glm(score.tr ~ 1, family = gaussian(link=identity))
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))
```

7.256887

```
print(p.value<- pchisq( deviance, df=4, lower.tail = FALSE))
```

0.1229198

- (e) Predict the score for a nurse who has worked at the center for seven years and who had previously been a co-PI on a grant that involved a quality assurance component.

We calculate  $score^0 = 100 \cdot (1 - (0.6093 + 0.2122 + 0.0002 \cdot 7 + 0.0773))^{-1} = 75.84951$ .

In SAS:

```

/*using fitted model for prediction*/
data predict;
input nurse doctor wrkyrs priorQIyes;
cards;
1 0 7 1
;

data QI;
set QI predict;
run;

proc genmod;
  model score_tr = nurse doctor wrkyrs priorQIyes / dist=normal link=identity;
  output out=outdata p=pscore_tr;
run;

data outdata;
set outdata;
  pscore=100/(1-pscore_tr);
run;

proc print data=outdata (firstobs=46) noobs;
  var pscore;
run;

```

**pscore**

**75.8653**

**In R:**

```

#using fitted model for prediction
pscore.tr<- predict(fitted.model, data.frame(design.rel="nurse", wrkyrs=7,
priorQI.rel="yes"))
print(100/(1-pscore.tr))

```

**75.86528**

**EXERCISE 2.3.** (a) Are the distances normally distributed? Plot the histogram, and do the testing. Explain.

**In SAS:**

```

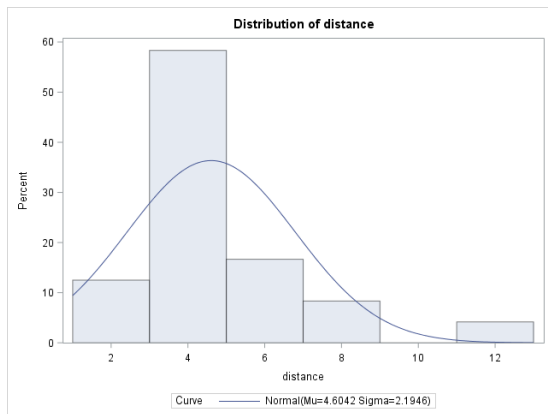
data distance;
input gender$ prior_expr$ self_eval distance @@;
male=(gender="M");
priorexpr_yes=(prior_expr="yes");
cards;
F no 2 1.9 F no 2 2.1 F yes 8 3.8 F yes 4 3.0
M no 5 4.2 F yes 10 8.2 F no 3 3.1 F no 4 2.4
F no 5 4.6 M yes 6 8.7 F no 6 4.7 M yes 7 4.2
F no 7 4.4 F yes 3 3.1 M yes 10 6.4 F yes 4 3.2
F no 6 5.1 M no 10 5.9 F no 6 5.0 M yes 3 3.6
F no 7 4.4 M yes 10 11.2 F yes 3 3.0 M yes 7 4.3
;

/*running normality check of response*/

```

```
proc univariate;
  var distance;
  histogram/normal;
run;
```

The distribution of distances is skewed to the right as depicted on the histogram. The normality tests give a small p-value indicating that the distribution is not normal.



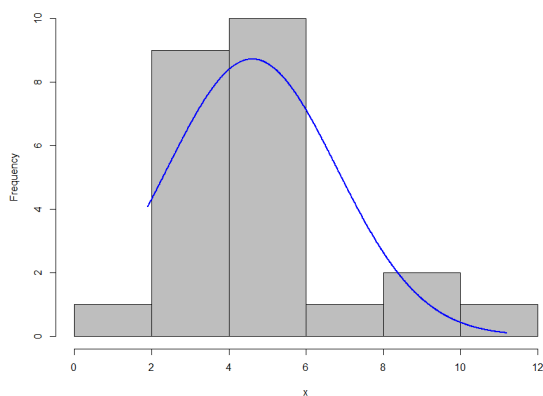
### Goodness-of-Fit Tests for Normal Distribution

Test	Statistic	p value
Kolmogorov-Smirnov D	0.20229147	Pr > D 0.012
Cramer-von Mises	W-Sq 0.19637956	Pr > W-Sq 0.005
Anderson-Darling	A-Sq 1.14009053	Pr > A-Sq <0.005

In R:

```
distance.data<- read.csv(file="C:/./Exercise2.3Data.csv", header=TRUE, sep=",")

#running normality check of response
library(rcompanion)
plotNormalHistogram(distance.data$distance)
```



```
shapiro.test(distance.data$distance)
```

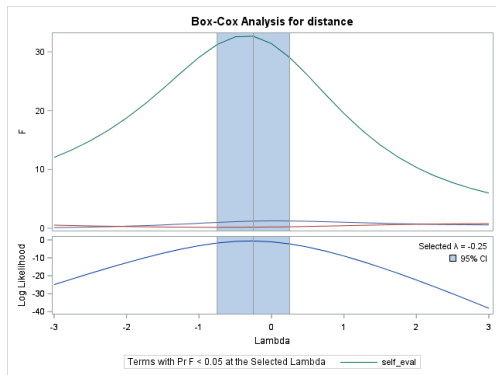
Shapiro-wilk normality test

w = 0.86065, p-value = 0.00347

- (b) Create indicator variables male and prior yes (existing prior experience), and use them to find a meaningful Box-Cox transformation that would transform the distance into a normally distributed variable. Prove its normality.

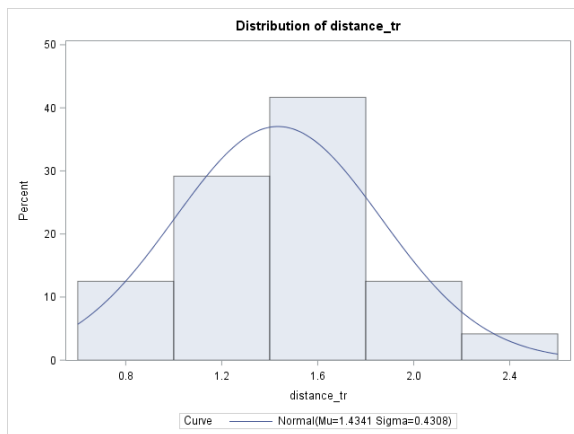
In SAS:

```
/*finding optimal lambda for Box-Cox transformation*/
proc transreg;
  model BoxCox(distance) = identity(male  priorexpr_yes  self_eval);
run;
```



```
/*applying Box-Cox transformation with lambda=0*/
data distance;
set distance;
  distance_tr=log(distance);
run;

/*running normality check of transformed response*/
proc univariate;
  var distance_tr;
  histogram/normal;
run;
```



#### Goodness-of-Fit Tests for Normal Distribution

Test	Statistic	p value
Kolmogorov-Smirnov D	0.11690632	Pr > D > 0.150
Cramer-von Mises	W-Sq 0.05030901	Pr > W-Sq > 0.250
Anderson-Darling	A-Sq 0.29858732	Pr > A-Sq > 0.250

In R:

```
#creating indicator variables
gender.rel<- relevel(distance.data$gender, ref="F")
priorexpr.rel<- relevel(distance.data$priorexpr, ref="no")

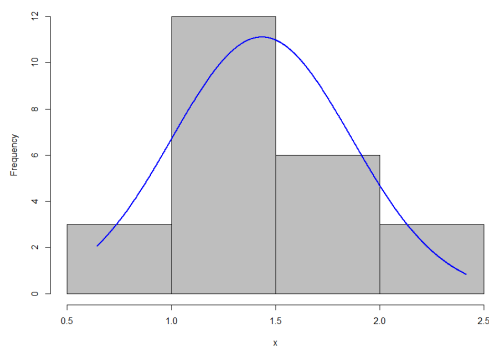
#finding optimal lambda for Box-Cox transformation
library(MASS)
BoxCox.fit<- boxcox(distance ~ gender.rel + priorexpr.rel + selfeval,
data=distance.data, lambda=seq(-3,3,1/4), interp = FALSE)
BoxCox.data<- data.frame(BoxCox.fit$x, BoxCox.fit$y)
ordered.data<- BoxCox.data[with(BoxCox.data, order(-BoxCox.fit.y)),]
ordered.data[1,]
```

**BoxCox.fit.x BoxCox.fit.y**

**-0.25      -2.144156**

```
#applying Box-Cox transformation with lambda=0
distance.tr<- log(distance.data$distance)
```

```
#running normality check of transformed response
plotNormalHistogram(distance.tr)
```



```
shapiro.test(distance.tr)
```

**Shapiro-wilk normality test**

**w = 0.97326, p-value = 0.7472**

- (c) Fit the general linear regression model to the transformed distance. Show that the model has a good fit. Discuss the significance of predictors.

In SAS:

```
/*fitting general linear model to transformed response*/
proc genmod;
  model distance_tr = male priorexpr_yes self_eval / dist=normal link=identity;
run;
```

**Log Likelihood 1.4688**

# Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald Limits	95% Confidence	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.6437	0.1206	0.4073	0.8800	28.49	<.0001
male	1	0.1402	0.1140	-0.0833	0.3637	1.51	0.2189
priorexpr_yes	1	0.0504	0.0995	-0.1446	0.2454	0.26	0.6125
self_eval	1	0.1249	0.0204	0.0851	0.1648	37.69	<.0001
Scale	1	0.2276	0.0329	0.1715	0.3020		

/\*checking model fit\*/

proc genmod;

model distance\_tr = / dist=normal link=identity;

run;

**Log Likelihood -13.3323**

data deviance\_test;

deviance = -2\*(-13.3323 - (1.4688));

pvalue = 1 - probchi(deviance,3);

run;

proc print noobs;

run;

deviance pvalue

29.6022 .000001673

Since the p-value is tiny, the model has a good fit. Self-evaluation is the only significant predictor.

In R:

#fitting general linear model for transformed response

```
summary(fitted.model<- glm(distance.tr ~ gender.rel + priorexpr.rel
+ selfeval, data=distance.data, family=gaussian(link=identity)))
```

**Coefficients:**

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	0.64370	0.13210	4.873	9.21e-05
gender.relM	0.14018	0.12491	1.122	0.275
priorexpr.relyes	0.05040	0.10900	0.462	0.649
selfeval	0.12494	0.02229	5.604	1.74e-05

#outputting estimated sigma

sigma(fitted.model)

**0.2493296**

#checking model fit

```
null.model<- glm(distance.tr ~ 1, data=distance.data,
family = gaussian(link=identity))
```

```
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))
```

**29.60224**

```
print(p.value<- pchisq(deviance,3,lower.tail=FALSE))
```

**1.673199e-06**

(d) Give interpretation for the estimates of the statistically significant regression coefficients. Use

$\alpha=0.05$ .

As the self-evaluation score increases by one unit, the estimated mean log-transformed distance increases by 0.1249.

- (e) Write down the final model that can be used for the prediction of distance. Predict the distance that a woman with no prior experience would bike if she is moderately confident about her abilities with the self-assessment value of 5.

The fitted model is  $\hat{E}(\ln(\text{distance})) = 0.6437 + 0.1402 \cdot \text{male} + 0.0504 \cdot \text{priorexperyes} + 0.1249 \cdot \text{selfeval}$ , and  $\hat{\sigma} = 0.2276$ .

The predicted value is computed as follows:  $\text{distance}^0 = \exp(0.6437 + 0.1249 \cdot 5) = 3.5544$ .

In SAS:

```
/*using fitted model for prediction*/
data predict;
input male priorexpr_yes self_eval;
cards;
0 0 5
;

data distance;
set distance predict;
run;

proc genmod;
  model distance_tr = male priorexpr_yes self_eval / dist=normal link=identity;
  output out=outdata p=pdistance_tr;
run;

data outdata;
set outdata;
pdistance=exp(pddistance_tr);
run;

proc print data=outdata (firstobs=25) noobs;
  var pdistance;
run;

pdistance
3.55522
```

In R:

```
#using fitted model for prediction
pdistance.tr<- predict(fitted.model, data.frame(gender.rel="F",
priorexpr.rel="no", selfeval=5))
print(exp(pddistance.tr))

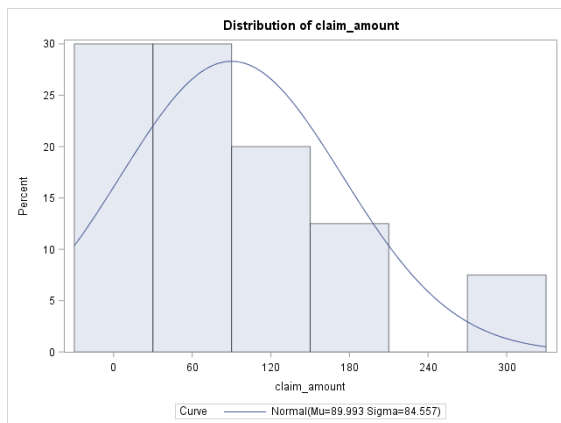
3.555221
```

**EXERCISE 2.4.** (a) Plot a histogram and carry out statistical tests for the normality of the distribution of claim amounts. Transform the variable via a Box-Cox transformation to achieve normality. Show that the transformed variable is normally distributed.

In SAS:

```
data claims;
input npolicies yrswithfirm percopenclaims claim_amount @@;
npoliciesK=npolicies/1000;
cards;
12318 4 16 19.9 29777 4 15 200.5 36980 10 12 308.5
18055 4 20 24.4 16505 20 27 48.7 19049 11 14 51.0
37112 20 26 163.2 22338 16 35 7.1 32349 16 25 1.5
26626 1 21 81.0 28547 11 17 91.0 33268 5 21 147.5
29045 13 29 63.9 18622 7 10 8.5 22784 12 11 27.0
39612 23 26 296.6 28423 7 12 129.0 17020 6 30 26.0
36930 7 24 98.6 37152 15 26 103.5 29629 9 35 107.4
32319 6 19 78.9 27103 23 25 0.3 23704 2 28 6.1
20432 21 16 58.4 30899 16 12 19.5 19052 10 23 46.9
37823 12 19 325.6 24269 14 31 5.7 23103 22 14 71.2
25556 4 32 29.3 15878 11 12 34.4 36772 17 13 50.6
19475 1 34 107.5 29241 8 29 180.2 36821 7 33 158.7
47309 11 12 124.0 15381 2 25 41.9 39857 13 11 195.0
34790 7 18 60.7
;

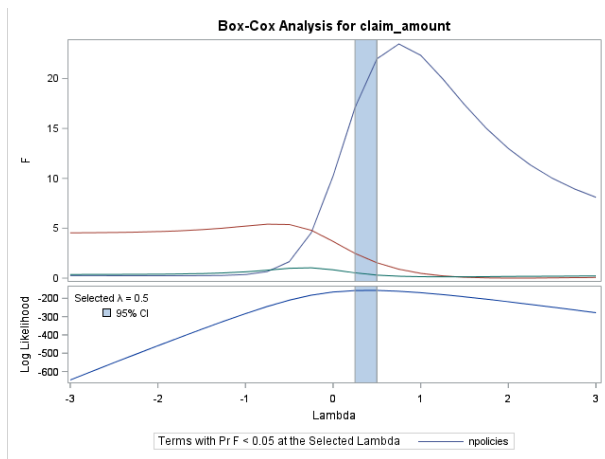
/*running normality check of response*/
proc univariate;
var claim_amount;
histogram/normal;
run;
```



#### Goodness-of-Fit Tests for Normal Distribution

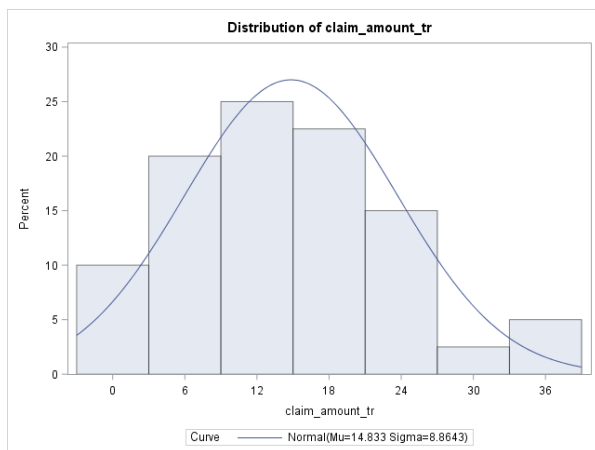
Test	Statistic	p value
Kolmogorov-Smirnov D	0.14617911	Pr > D 0.030
Cramer-von Mises	W-Sq 0.26260150	Pr > W-Sq <0.005
Anderson-Darling	A-Sq 1.70664028	Pr > A-Sq <0.005

```
/*finding optimal lambda for Box-Cox transformation*/
proc transreg;
model BoxCox(claim_amount) = identity(npoliciesK yrswithfirm percopenclaims);
run;
```



```
/*applying Box-Cox transformation with lambda=0.5*/
data claims;
set claims;
  claim_amount_tr = 2*(claim_amount**(0.5)-1);
run;

/*running normality check of transformed response*/
proc univariate;
  var claim_amount_tr;
  histogram/normal;
run;
```



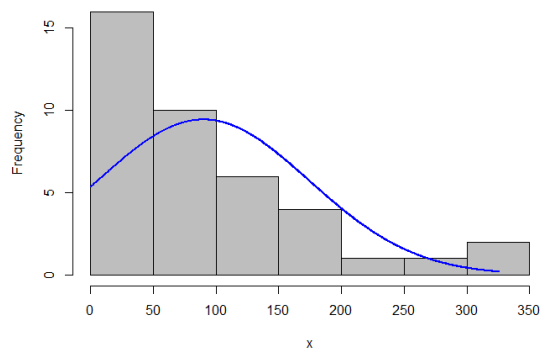
### Goodness-of-Fit Tests for Normal Distribution

Test	Statistic	p value
Kolmogorov-Smirnov D	0.06321834	Pr > D > 0.150
Cramer-von Mises	W-Sq 0.02560195	Pr > W-Sq > 0.250
Anderson-Darling	A-Sq 0.21938379	Pr > A-Sq > 0.250

In R:

```
claims.data<- read.csv(file="C:/./Exercise2.4Data.csv", header=TRUE, sep=",")

#running normality check of response
library(rcompanion)
plotNormalHistogram(claims.data$claimamount)
```



```
shapiro.test(claims.data$claimamount)
```

**Shapiro-wilk normality test**

**w = 0.85595, p-value = 0.0001259**

```
#rescaling npolicies
npoliciesK<- claims.data$npolicies/1000

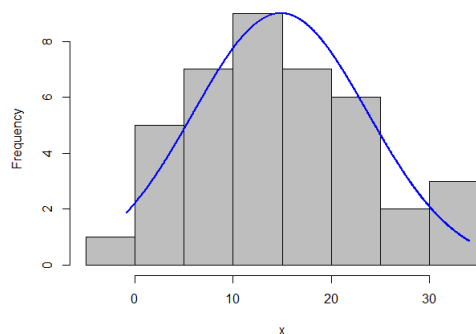
#finding optimal lambda for Box-Cox transformation
library(MASS)
BoxCox.fit<- boxcox(claimamount ~ npoliciesK + yrswithfirm + percopenclaims,
data=claims.data, interp = FALSE)
BoxCox.data<- data.frame(BoxCox.fit$x, BoxCox.fit$y)
ordered.data<- BoxCox.data[with(BoxCox.data, order(-BoxCox.fit.y)),]
ordered.data[1,]
```

**BoxCox.fit.x BoxCox.fit.y**

**0.4 -72.77094**

```
#applying Box-Cox transformation with lambda=0.5
claimamount.tr<- 2*(sqrt(claims.data$claimamount)-1)
```

```
#running normality check of transformed response
plotNormalHistogram(claimamount.tr)
```



```
shapiro.test(claimamount.tr)
```

**Shapiro-wilk normality test**

**w = 0.97601, p-value = 0.5445**

(b) Fit a linear regression model, relating the transformed claim amounts to all the other variables. Which variables are significant predictors at the 5% level?

In SAS:

```
/*fitting general linear model to transformed response*/
proc genmod;
  model claimamount_tr = npoliciesK yrswithfirm percopenclaims / dist=normal
  link=identity;
run;
```

Log Likelihood -133.7181

#### Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	95% Confidence Limits	Pr > ChiSq
Intercept	1	0.6928	5.2239	0.02	-9.5459 10.9314	0.8945
npoliciesK	1	0.6624	0.1341	24.41	0.3996 0.9251	<.0001
yrswithfirm	1	-0.2338	0.1786	1.71	-0.5838 0.1163	0.1905
percopenclaims	1	-0.0825	0.1419	0.34	-0.3607 0.1956	0.5609
Scale	1	6.8484	0.7657		5.5007 8.5262	

The fitted model is

$\hat{E} \left( 2(\sqrt{\text{claimamount}} - 1) \right) = 0.6928 + 0.6624 \cdot \text{npoliciesK} - 0.2338 \cdot \text{yrswithfirm} - 0.0825 \cdot \text{percopenclaims}$ , and  $\hat{\sigma} = \text{Scale} = 6.8484$ .

At the 5% level, only the number of policies is a significant predictor.

In R:

```
#fitting general linear model to transformed response
summary(fitted.model<- glm(claimamount.tr ~ npoliciesK + yrswithfirm +
percopenclaims, data=claims.data, family=gaussian(link=identity)))
```

#### Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	0.69275	5.50646	0.126	0.901
npoliciesK	0.66236	0.14130	4.688	3.89e-05
yrswithfirm	-0.23379	0.18826	-1.242	0.222
percopenclaims	-0.08254	0.14961	-0.552	0.585

```
#outputting estimated sigma
sigma(fitted.model)
```

7.218831

(c) Assess the model fit. Interpret estimated significant regression coefficients.

In SAS:

```
/*checking model fit*/
proc genmod;
```

```
model claimamount_tr = / dist=normal link=identity;
run;
```

**Log Likelihood -143.5324**

```
data deviance_test;
  deviance = -2*(-143.5324 - (-133.7181));
  pvalue = 1 - probchi(deviance,3);
run;
```

```
proc print noobs;
run;
```

```
deviance      pvalue
19.6286      .000202641
```

The model has a good fit since the p-value is very small.

**In R:**

```
checking model fit
null.model<- glm(claimamount.tr ~ 1, data=claims.data,
family=gaussian(link=identity))
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))
```

**19.62872**

```
print(p.value<- pchisq(deviance,3,lower.tail=FALSE))
```

**0.0002026295**

As the number of policies increases by one thousand, the estimated mean square root-transformed claim amount increases by 0.0007.

(d) Compute the predicted amount of aggregate claims for a company with 15,500 policyholders, that has been buying policies at this firm for the past three years, and that still has 15% of outstanding claims from the previous year.

The predicted value is evaluated as follows:  $claimamount^0 = \left(\frac{1}{2}(0.6928 + 0.66236 \cdot 15.5 - 0.2338 \cdot 3 - 0.0825 \cdot 15) + 1\right)^2 = 30.36$ .

**In SAS:**

```
/*using fitted model for prediction*/
data predict;
input npoliciesK yrswithfirm percopenclaims;
cards;
15.5 3 15
;
```

```
data claims;
set claims predict;
run;
```

```
proc genmod;
  model claim_amount_tr = npoliciesK yrswithfirm percopenclaims / dist=normal
  link=identity;
```

```

output out=outdata p=pclaim_amount_tr;
run;

data outdata;
set outdata;
pclaim_amount=((pclaim_amount_tr/2)+1)**2;
run;

proc print data=outdata (firstobs=41);
var pclaim_amount;
run;

pclaim_amount
30.3600

```

In R:

```

#using fitted model for prediction
pclaim.amount.tr<- predict(fitted.model, data.frame(npoliciesK=15.5,
yrswithfirm=3, percopenclaims=15))
print((pclaim.amount.tr/2+1)**2)

30.36

```

**EXERCISE 2.5.** Show that a gamma distribution with density defined by (2.3) belongs to the exponential family of distributions. Conclude that the gamma regression is a generalized linear regression. Give its link function.

$$\begin{aligned}
 f_Y(y; \alpha, \beta) &= \frac{y^{\alpha-1}}{\Gamma(\alpha)\beta^\alpha} e^{-\frac{y}{\beta}} = \exp \left\{ -\frac{y}{\beta} - \alpha \ln \beta + (\alpha - 1) \ln y - \ln \Gamma(\alpha) \right\} \\
 &= \exp \left\{ \frac{y \left( -\frac{1}{\alpha\beta} \right) - \ln \beta}{\frac{1}{\alpha}} + (\alpha - 1) \ln y - \ln \Gamma(\alpha) \right\}.
 \end{aligned}$$

If we let  $\theta = -\frac{1}{\alpha\beta}$  and  $\phi = \frac{1}{\alpha}$ , we can write  $\ln \beta = \ln \left( -\frac{\phi}{\theta} \right) = \ln \left( -\frac{1}{\theta} \right) + \ln \phi$ , and thus, we obtain

$$f_Y(y; \alpha, \beta) = \exp \left\{ \frac{y\theta - \ln \left( -\frac{1}{\theta} \right) - \ln \phi}{\phi} + \left( \frac{1}{\phi} - 1 \right) \ln y - \ln \Gamma \left( \frac{1}{\phi} \right) \right\}.$$

Letting  $c(\theta) = \ln \left( -\frac{1}{\theta} \right)$ , and  $h(y, \phi) = -\frac{\ln \phi}{\phi} + \left( \frac{1}{\phi} - 1 \right) \ln y - \ln \Gamma \left( \frac{1}{\phi} \right)$ , we obtain the form of a density that belongs to the exponential family of distributions (1.3). Thus, a gamma regression is a generalized linear regression. According to (2.4), the mean response is modeled via a log-link function.

**EXERCISE 2.6.** (a) Fit the gamma regression model with the log link function. Write down the fitted model. Check its goodness of fit.

In SAS:

```
data obesity;
input gender$ age group$ preBMI postBMI @@;
    BMIdiff=preBMI-postBMI;
cards;
F 6 Cx 85.7 83.8 F 6 Cx 93.8 92.9 F 7 Cx 93.5 92.5 F 8 Cx 90.1 89.8
F 9 Tx 92.3 90.7 F 9 Tx 90.3 88.3 F 12 Cx 87.6 85.9 F 12 Cx 87.2 84.1
F 12 Tx 96.9 94.9 F 12 Tx 85.8 81.2 F 13 Cx 96.7 94.1 F 13 Cx 93.5 92.9
F 13 Tx 92.3 87.5 F 13 Tx 85.3 83.7 F 14 Tx 95.5 78.7 F 15 Cx 91.3 89.9
F 15 Tx 95.8 87.1 F 16 Tx 90.7 87.2 M 6 Cx 92.6 88.1 M 7 Cx 95.8 94.7
M 7 Cx 90.4 89.1 M 7 Cx 91.2 88.6 M 8 Tx 94.4 87.8 M 8 Tx 93.2 87.3
M 10 Cx 93.9 91.5 M 10 Tx 96.2 91.1 M 10 Tx 89.4 87.9 M 11 Tx 86.2 77.1
M 11 Tx 95.4 84.8 M 12 Cx 97.7 95.8 M 13 Tx 85.3 80.0 M 13 Tx 86.2 82.4
M 14 Cx 85.5 83.6 M 14 Cx 97.8 93.8 M 16 Cx 95.0 93.6 M 16 Tx 93.1 86.8
;

/*fitting gamma regression model*/
proc genmod;
    class gender(ref="F") group(ref="Cx");
    model BMIdiff = gender age group / dist=gamma link=log;
run;
```

Log Likelihood -69.3482

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	95% Confidence Limits	Pr > ChiSq
Intercept	1	-0.0442	0.4154	0.01	-0.8584 0.7700	0.9153
gender	M 1	0.3862	0.2052	3.54	-0.0160 0.7884	0.0598
gender	F 0	0.0000	0.0000	.	0.0000 0.0000	.
age	1	0.0470	0.0349	1.82	-0.0214 0.1154	0.1777
group	Tx 1	0.9870	0.2104	22.01	0.5747 1.3994	<.0001
group	Cx 0	0.0000	0.0000	.	0.0000 0.0000	.
Scale	1	2.8805	0.6435		1.8591 4.4629	

The fitted model is  $\hat{E}(BMIdiff) = \exp(-0.0442 + 0.3862 \cdot male + 0.0470 \cdot age + 0.9870 \cdot Tx)$ , and  $\hat{\alpha} = \frac{1}{2.8805} = 0.3472$ .

```
/*checking model fit*/
proc genmod;
    model BMIdiff = / dist=gamma link=log;
run;
```

Log Likelihood -81.2031

```
data deviance_test;
    deviance = -2*(-81.2031 - (-69.3482));
    pvalue = 1 - probchi(deviance,3);
run;
```

```
proc print noobs;
run;
```

```
deviance      pvalue
23.7098      .000028719
```

The model has a good fit since the p-value is very small.

In R:

```
bmi.data<- read.csv(file="C://Exercise2.1Data.csv", header=TRUE, sep=",")

#creating the difference in BMI
BMIdiff<- bmi.data$preBMI-bmi.data$postBMI

#fitting gamma regression model
summary(fitted.model<- glm(BMIdiff ~ gender + age + group, data=bmi.data,
family=Gamma(link=log)))
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-0.04415	0.45077	-0.098	0.922581
genderM	0.38624	0.22749	1.698	0.099250
age	0.04703	0.03826	1.229	0.227963
groupTx	0.98701	0.23376	4.222	0.000187

Dispersion parameter for Gamma family taken to be 0.4599244

```
#checking model fit
null.model<- glm(BMIdiff ~ 1, data=bmi.data, family=Gamma(link=log))
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))
```

23.83593

```
print(p.value<- pchisq(deviance,3,lower.tail = FALSE))
```

2.70297e-05

(b) What variables are significant predictors in this model? Use the 5% significance level.

Only the group is a significant predictor.

(c) Interpret estimated significant regression coefficients.

The estimated average decrease in BMI percentile for patients in the intervention group is  $\exp(0.9870) \cdot 100\% = 268.32\%$  of that for patients in the control group.

(d) Predict the change in BMI percentile for a 9-year-old girl in the control group. Compare the prediction with the one obtained in Exercise 2.1.

We calculate the predicted value as follows:

$$BMIdiff^0 = \exp(-0.0442 + 0.0470 \cdot 9) = 1.4605.$$

In SAS:

```
/*using fitted model for prediction*/
data predict;
input gender$ age group$;
cards;
F 9 Cx
;

data obesity;
set obesity predict;
run;

proc genmod;
  class gender(ref="F") group(ref="Cx");
  model BMIdiff = gender age group / dist=gamma link=log;
  output out=outdata p=pBMIdiff;
run;

proc print data=outdata (firstobs=37);
  var pBMIdiff;
run;

pBMIdiff
1.46106
```

In R:

```
#using fitted model for prediction
print(predict(fitted.model, data.frame(gender="F", age=9, group="Cx"),
type="response"))

1.461058
```

**EXERCISE 2.7.** (a) Fit the gamma regression model with the log link function. Present the fitted model and discuss its goodness of fit.

In SAS:

```
data QI;
length desgn $6.;
input desgn$ wrkyrs priorQI$ score @@;
score=score/100;
cards;
nurse 16 yes 63 nurse 9 yes 93 nurse 8 yes 74
nurse 1 no 69 nurse 5 no 67 nurse 3 no 66
nurse 24 no 86 nurse 4 no 74 nurse 1 no 88
nurse 24 no 84 nurse 3 no 97 doctor 2 yes 88
doctor 5 yes 78 doctor 26 yes 82 doctor 3 no 57
doctor 3 no 88 doctor 15 no 78 doctor 4 no 65
doctor 25 no 78 staff 3 yes 62 staff 21 no 55
staff 8 no 62 staff 11 no 67 nurse 8 yes 62
nurse 22 yes 68 nurse 4 no 93 nurse 6 no 77
nurse 2 no 59 nurse 20 no 64 nurse 2 no 70
```

```

nurse 3 no 63 nurse 16 no 65 nurse 18 no 73
nurse 15 no 76 doctor 2 yes 85 doctor 7 yes 91
doctor 2 yes 69 doctor 20 no 66 doctor 13 no 55
doctor 8 no 62 doctor 14 no 61 staff 9 yes 57
staff 11 yes 69 staff 19 no 64 staff 17 no 76
;

/*fitting gamma regression model*/
proc genmod;
  class desgn(ref="staff") priorQI(ref="no");
  model score=desgn wrkyrs priorQI/dist=gamma link=log;
run;

```

Log Likelihood 39.1840

#### Analysis of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald Limits	95% Confidence	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.4624	0.0640	-0.5880	-0.3369	52.14	<.0001
desgn doctor	1	0.1340	0.0625	0.0115	0.2564	4.60	0.0320
desgn nurse	1	0.1540	0.0595	0.0375	0.2706	6.71	0.0096
desgn staff	0	0.0000	0.0000	0.0000	0.0000	.	.
wrkyrs	1	-0.0002	0.0028	-0.0056	0.0052	0.01	0.9290
priorQI yes	1	0.0532	0.0470	-0.0389	0.1454	1.28	0.2575
priorQI no	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale	1	49.8543	10.4753	33.0256	75.2584		

The fitted model is  $\hat{E}(\text{score}) = 100 \cdot \exp(-0.4624 + 0.1340 \cdot \text{doctor} + 0.1540 \cdot \text{nurse} - 0.0002 \cdot \text{wrkyrs} + 0.0532 \cdot \text{priorQIyes})$ , and  $\hat{\alpha} = \frac{1}{49.8543} = 0.020058$ .

The model doesn't really fit the data well because the p-value for the deviance test is larger than 0.05.

```

/*checking model fit*/
proc genmod;
  model score = / dist=gamma link=log;
run;

```

Log Likelihood 35.5289

```

data deviance_test;
  deviance = -2*(35.5289 - 39.1840);
  pvalue = 1 - probchi(deviance,4);
run;

```

```

proc print noobs;
run;

```

```

deviance  pvalue
7.3102    0.12038

```

In R:

```

jobscore.data<- read.csv(file="C:/./Exercise2.2Data.csv", header=TRUE, sep=",")

#rescaling values and setting references

```

```

desgn.rel<- relevel(jobscore.data$desgn, ref="staff")
priorQI.rel<- relevel(jobscore.data$priorQI, ref="no")
score<- jobscore.data$score/100

#fitting gamma regression model
summary(fitted.model<- glm(score ~ desgn.rel + wrkyrs + priorQI.rel,
data=jobscore.data, family=Gamma(link=log)))

```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	4.1427297	0.0686964	60.305	<2e-16
desgn.reldoctor	0.1339899	0.0667675	2.007	0.0516
desgn.relnurse	0.1540443	0.0636050	2.422	0.0201
wrkyrs	-0.0002455	0.0029813	-0.082	0.9348
priorQI.relyes	0.0532444	0.0498513	1.068	0.2919

Dispersion parameter for Gamma family taken to be 0.02298337

```

#checking model fit
null.model<- glm(score ~ 1, data=jobscore.data, family=Gamma(link=log))
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))

```

7.310283

```
print(p.value<- pchisq(deviance,4,lower.tail=FALSE))
```

0.1203719

(b) Discuss the significance of the beta coefficients. Interpret the estimated significant coefficients.

Indicators of a doctor and a nurse are significant at the 5% level. The estimated mean score for doctors is  $\exp(0.1340) \cdot 100\% = 114.34\%$  of that for staff. The estimated mean score for nurses is  $\exp(0.1540) \cdot 100\% = 116.65\%$  of that for staff.

(c) Predict the score for a nurse who has worked at the center for seven years and who had previously been a co-PI on a grant that involved a quality assurance component. Compare that predicted score to the one obtained in Exercise 2.2.

The predicted value is calculated as:  $score^0 = 100 \cdot \exp(-0.4624 + 0.1540 - 0.0002 \cdot 7 + 0.0532) = 77.3678$ . The predicted score in Exercise 2.2 is 75.8653 which is smaller than what we predict here.

In SAS:

```

/*using fitted model for prediction*/
data predict;
input desgn$ wrkyrs priorQI$;
cards;
nurse 7 yes
;

data QI;
set QI predict;
run;

proc genmod;

```

```

class desgn(ref="staff") priorQI(ref="no");
model score = desgn wkryrs priorQI / dist=gamma link=log;
output out=outdata p=pscore;
run;

data outdata;
set outdata;
pscore=pscore*100;
run;

proc print data=outdata (firstobs=46) noobs;
var pscore;
run;

pscore
77.3468

```

In R:

```

#using fitted model for prediction
print(predict(fitted.model, data.frame(desgn.rel="nurse", wrkys=7,
priorQI.rel="yes"), type="response"))

77.34687

```

**EXERCISE 2.8.** (a) Write out explicitly the estimated model. Check the goodness of fit of this model.

In SAS:

```

data distance;
input gender$ prior_expr$ self_eval distance @@;
cards;
F no 2 1.9 F no 2 2.1 F yes 8 3.8 F yes 4 3.0 M no 5 4.2
F yes 10 8.2 F no 3 3.1 F no 4 2.4 F no 5 4.6 M yes 6 8.7
F no 6 4.7 M yes 7 4.2 F no 7 4.4 F yes 3 3.1 M yes 10 6.4
F yes 4 3.2 F no 6 5.1 M no 10 5.9 F no 6 5.0 M yes 3 3.6
F no 7 4.4 M yes 10 11.2 F yes 3 3.0 M yes 7 4.3
;

/*fitting gamma regression model*/
proc genmod;
class gender(ref="F") prior_expr(ref="no");
model distance = gender prior_expr self_eval / dist=gamma link=log;
run;

```

Log Likelihood -33.4006

#### Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald Limits	95% Confidence	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.6494	0.1241	0.4060	0.8927	27.36	<.0001
gender M	1	0.1652	0.1156	-0.0614	0.3918	2.04	0.1531
gender F	0	0.0000	0.0000	0.0000	0.0000	.	.

# Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald Limits	95% Confidence	Wald Chi-Square	Pr > ChiSq
prior_expr yes	1	0.0571	0.1026	-0.1440	0.2583	0.31	0.5778
prior_expr no	0	0.0000	0.0000	0.0000	0.0000	.	.
self_eval	1	0.1266	0.0208	0.0859	0.1672	37.19	<.0001
Scale	1	18.9200	5.4143	10.7978	33.1517		

The fitted model is  $\hat{E}(\text{distance}) = \exp(0.6494 + 0.1652 \cdot \text{male} + 0.0571 \cdot \text{prior}_{\text{expr}_{\text{yes}}} + 0.1266 \cdot \text{self}_{\text{eval}})$ , and  $\hat{a} = \frac{1}{18.92} = 0.052854$ . This model fits well as indicated by a small p-value in the deviance test.

```
/*checking model fit*/
proc genmod;
  model distance = / dist=gamma link=log;
run;
```

Log Likelihood -48.6416

```
data deviance_test;
  deviance = -2*(-48.6416 - (-33.4006));
  pvalue = 1 - probchi(deviance,3);
run;
```

```
proc print noobs;
run;
```

```
deviance      pvalue
  30.482      .000001093
```

In R:

```
distance.data<- read.csv(file="C:/./Exercise2.3Data.csv",header=TRUE, sep=",")

#setting reference variables
gender.rel<- relevel(distance.data$gender, ref="F")
priorexpr.rel<- relevel(distance.data$priorexpr, ref="no")

#fitting gamma regression
summary(fitted.model<- glm(distance ~ gender.rel + priorexpr.rel + selfeval,
data=distance.data, family=Gamma(link=log)))
```

Coefficients:

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.64936    0.13858   4.686 0.000142
gender.relM     0.16515    0.13104   1.260 0.222054
priorexpr.relyes 0.05713    0.11435   0.500 0.622822
selfeval       0.12656    0.02339   5.411 2.69e-05
```

Dispersion parameter for Gamma family taken to be 0.06841251

```
#checking model fit
null.model<- glm(distance ~ 1, data=distance.data, family=Gamma(link=log))
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))
```

30.492

```
print(p.value<- pchisq(deviance,3,lower.tail=FALSE))
```

1.087373e-06

- (b) Which predictors would really influence the response, if changed? Give an interpretation of the estimated significant regression coefficients.

The self-evaluation score is the only significant predictor. If the self-evaluation score increases by one, the estimated mean distance increases by  $(\exp(0.1266) - 1) \cdot 100\% = 13.49\%$ .

- (c) Predict the distance that a woman with no prior experience would bike if she is moderately confident about her abilities with the self-assessment value of 5. Compare your answer to the one obtained in Exercise 2.3.

Predicted distance is  $distance^0 = \exp(0.6494 + 0.1266 \cdot 5) = 3.6053$ . In Exercise 2.3, the predicted value is 3.5544, which is smaller than the one in this exercise.

In SAS:

```
/*using fitted model for prediction*/
data predict;
input  gender$ prior_expr$ self_eval;
cards;
F no 5
;

data distance;
set distance predict;
run;

proc genmod;
  class gender(ref="F") prior_expr(ref="no");
  model distance = gender prior_expr self_eval / dist=gamma link=log;
  output out=outdata p=pdistance;
run;

proc print data=outdata (firstobs=25) noobs;
  var pdistance;
run;

pdistance
3.60448
```

In R:

```
#using fitted model for prediction
print(predict(fitted.model, data.frame(gender.rel="F", priorexpr.rel="no",
selfeval=5), type="response"))

3.604477
```

**EXERCISE 2.9.** (a) Run the gamma regression and write the predicted model. What variables are significant predictors of the claim amount? Compare this to the model in Exercise 2.4.

In SAS:

```
data claims;
input npolicies yrswithfirm percopenclaims claim_amount @@;
npoliciesK=npolicies/1000;
cards;
12318 4 16 19.9 29777 4 15 200.5 36980 10 12 308.5
18055 4 20 24.4 16505 20 27 48.7 19049 11 14 51.0
37112 20 26 163.2 22338 16 35 7.1 32349 16 25 1.5
26626 1 21 81.0 28547 11 17 91.0 33268 5 21 147.5
29045 13 29 63.9 18622 7 10 8.5 22784 12 11 27.0
39612 23 26 296.6 28423 7 12 129.0 17020 6 30 26.0
36930 7 24 98.6 37152 15 26 103.5 29629 9 35 107.4
32319 6 19 78.9 27103 23 25 0.3 23704 2 28 6.1
20432 21 16 58.4 30899 16 12 19.5 19052 10 23 46.9
37823 12 19 325.6 24269 14 31 5.7 23103 22 14 71.2
25556 4 32 29.3 15878 11 12 34.4 36772 17 13 50.6
19475 1 34 107.5 29241 8 29 180.2 36821 7 33 158.7
47309 11 12 124.0 15381 2 25 41.9 39857 13 11 195.0
34790 7 18 60.7
;

/*fitting gamma regression model*/
proc genmod;
model claim_amount = npoliciesK yrswithfirm percopenclaims / dist=gamma
link=log;
run;
```

Log Likelihood -211.7158

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	2.4656	0.6775	1.1376	3.7936	13.24	0.0003
npoliciesK	1	0.0764	0.0169	0.0432	0.1096	20.38	<.0001
yrswithfirm	1	-0.0186	0.0216	-0.0610	0.0238	0.74	0.3897
percopenclaims	1	-0.0037	0.0187	-0.0403	0.0330	0.04	0.8444
Scale	1	1.2907	0.2597	0.8701	1.9147		

The fitted model is  $\hat{E}(\text{claim amount}) = \exp\left(2.4656 + 0.0764 \cdot \frac{\text{npolicies}}{1000} - 0.0186 \cdot \text{yearwithfirm} - 0.0037 \cdot \text{percentopenclaims}\right)$  and  $\hat{\alpha} = \frac{1}{1.2907} = 0.7748$ . Only the number of policies is significant at the 5% level, which is the same as in the model in Exercise 2.4.

In R:

```
claims.data<- read.csv(file="C:/./Exercise2.4Data.csv", header=TRUE, sep=",")

#rescaling npolicies
npoliciesK<- claims.data$npolicies/1000

#fitting gamma regression
```

```
summary(fitted.model<- glm(claimamount ~ npoliciesK + yrswithfirm +
percopenclaims, data=claims.data, family=Gamma(link=log)))
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	2.465581	0.529342	4.658	4.25e-05
npoliciesK	0.076404	0.013584	5.625	2.21e-06
yrswithfirm	-0.018610	0.018098	-1.028	0.311
percopenclaims	-0.003667	0.014382	-0.255	0.800

Dispersion parameter for Gamma family taken to be 0.4815732

(b) Interpret estimates of the significant beta coefficients. How good is the model fit?

When the number of policies increases by one thousand, the estimated mean claim amount increases by  $(\exp(0.0764) - 1) \cdot 100\% = 7.94\%$ . The fit of the model is good since the p-value is tiny.

In SAS:

```
/*checking model fit*/
proc genmod;
  model claim_amount = / dist=gamma link=log;
run;
```

Log Likelihood -219.9272

```
data deviance_test;
  deviance = -2*(-219.9272 - (-211.7158));
  pvalue = 1 - probchi(deviance,3);
run;
```

```
proc print noobs;
run;
```

deviance	pvalue
16.4228	.000928679

In R:

```
#checking model fit
null.model<- glm(claimamount~ 1, data=claims.data, family=Gamma(link=log))
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))
```

16.67906

```
print(p.value<- pchisq(deviance,3,lower.tail=FALSE))
```

0.0008226888

(c) Obtain the predicted amount of aggregate claims for a company with 15,500 policyholders, that has been buying policies at this firm for the past three years, and that still has 15% of outstanding claims from the previous year. Compare the result with the one computed in Exercise 2.4.

The prediction as done by hand yields:  $claim\ amount^0 = \exp\left(2.4656 + 0.0764 \cdot \frac{15500}{1000} - 0.0186 \cdot 3 - 0.0037 \cdot 15\right) = 34.4153$ . In Exercise 2.4, the predicted claim amount is 30.36 which is much less.

In SAS:

```
/*using fitted model for prediction*/
data predict;
input npoliciesK yrswithfirm percopenclaims;
cards;
15.5 3 15
;
data claims;
set claims predict;
run;

proc genmod;
  model claim_amount = npoliciesK yrswithfirm percopenclaims / dist=gamma
    link=log;
  output out=outdata p=pclaim_amount;
run;

proc print data=outdata (firstobs=41) noobs;
  var pclaim_amount;
run;

pclaim_amount
      34.4327
```

In R:

```
#using fitted model for prediction
print(pred.claims<- predict(fitted.model, data.frame(npoliciesK=15.5,
yrswithfirm=3, percopenclaims=15), type="response"))

34.43248
```

## CHAPTER 3

**EXERCISE 3.1.** Show that the probability mass function of a Bernoulli( $\pi$ ) random variable belongs to the exponential family of distributions. Conclude that the logistic, probit, and complement log-log models are special cases of generalized linear regression. Specify the respective link functions.

$$\begin{aligned} f_Y(y; \alpha, \beta) &= \pi^y (1 - \pi)^{1-y} = \exp\{y \ln \pi + (1 - y) \ln(1 - \pi)\} \\ &= \exp\left\{y \ln \frac{\pi}{1 - \pi} + \ln(1 - \pi)\right\} = \exp\{y \cdot \theta - c(\theta)\} \end{aligned}$$

where  $\theta = \ln \frac{\pi}{1 - \pi}$  and  $c(\theta) = -\ln(1 - \pi) = -\ln\left(1 - \frac{e^\theta}{1 + e^\theta}\right) = \ln(1 + e^\theta)$ . Thus, Bernoulli( $\pi$ ) belongs to the exponential family of distributions with the location parameter  $\theta = \ln \frac{\pi}{1 - \pi}$  and dispersion parameter  $\phi = 1$ . Thus, logistic, probit, and complementary log-log models belong to the class of generalized linear models with the link functions logit, probit, and complementary log-log, respectively.

**EXERCISE 3.2.** (a) Fit a binary logistic model. Write down the fitted model. Discuss the significance of predictor variables, and the goodness of fit of the model. Use  $\alpha = 0.05$ .

In SAS:

```
/*fitting logistic model*/
data psoriasis;
input gender$ age medication$ relief @@;
cards;
M 37 A 1   F 24 A 1   F 15 A 1   M 31 B 1   F 39 B 1   M 31 B 1
M 20 A 1   M 32 A 1   M 30 A 1   F 24 B 0   M 17 B 0   F 33 B 1
M 24 A 1   M 32 A 1   F 27 A 1   M 16 A 1   F 33 A 1   F 28 A 0
M 51 B 1   F 35 B 0   M 16 B 0   F 25 A 0   M 18 A 1   F 19 A 1
M 39 B 1   M 38 B 1   M 37 B 1   F 24 B 0   F 39 B 0   F 33 B 0
;

proc genmod;
  class gender(ref="F") medication(ref="B");
  model relief(event="1") = gender age medication / dist=binomial link=logit;
run;
```

Log Likelihood -10.5042

AIC 29.0084  
AICC 30.6084  
BIC 34.6132

### Analysis of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald	95% Confidence Limits	Wald Chi-Square	Pr > ChiSq
Intercept	1	-6.7992	3.1143	-12.9031	-0.6953	4.77	0.0290
gender	M 1	3.1713	1.4710	0.2883	6.0544	4.65	0.0311
gender	F 0	0.0000	0.0000	0.0000	0.0000	.	.
age	1	0.1713	0.0869	0.0010	0.3416	3.89	0.0487
medication A	1	3.8164	1.5462	0.7860	6.8469	6.09	0.0136
medication B	0	0.0000	0.0000	0.0000	0.0000	.	.

The fitted model is  $\ln \frac{\hat{P}(\text{relief})}{1-\hat{P}(\text{relief})} = -6.7992 + 3.1713 \cdot \text{male} + 0.1713 \cdot \text{age} + 3.8164 \cdot \text{medication A}$ . Gender, age, and medication type are all significant predictors at the 5% level.

```
/*checking model fit*/
proc genmod;
  model relief = / dist=binomial link=logit;
run;
```

Log Likelihood -18.3259

```
data deviance_test;
  deviance = -2*(-18.3259 - (-10.5042));
  pvalue = 1 - probchi(deviance,3);
run;
```

```
proc print noobs;
run;
```

```
deviance      pvalue
15.6434      .001341753
```

The model has a good fit as follows from the small p-value (<0.05) in the deviance test.

In R:

```
#fitting logistic model
psoriasis.data<- read.csv(file="C:/./Exercise3.2Data.csv", header=TRUE, sep=",")

#setting reference categories
gender.rel<- relevel(psoriasis.data$gender, ref="F")
medication.rel<- relevel(psoriasis.data$medication, ref="B")

#running the model
summary(fitted.model<- glm(relief ~ gender.rel + age + medication.rel,
data=psoriasis.data,family=binomial(link=logit)))
```

Coefficients:

```
Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.79921 3.11430 -2.183 0.0290
gender.relM 3.17132 1.47097 2.156 0.0311
age 0.17131 0.08691 1.971 0.0487
medication.relA 3.81641 1.54617 2.468 0.0136
```

AIC: 29.008

```
#computing AICC
```

```

p<-4
n<-30
print(AICC<- -2*logLik(fitted.model)+2*p*n/(n-p-1))

30.60842

#outputting BIC
BIC(fitted.model)

34.6132

#checking model fit
null.model<- glm(relief ~ 1, data=psoriasis.data, family=binomial(link=logit))
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))

15.64344

print(p.value<- pchisq(deviance,3,lower.tail = FALSE))

0.001341726

```

(b) Give an interpretation of the estimated significant regression coefficients.

The estimated odds in favor of relief from psoriasis for male patients are  $\exp(3.1713) \cdot 100\% = 2,383.85\%$  of those for female patients. As age increases by one year, the estimated odds increase by  $(\exp(0.1713) - 1) \cdot 100\% = 18.68\%$ . The estimated odds for patients taking medication A are  $\exp(3.8164) \cdot 100\% = 4,544.03\%$  of those for patients taking medication B.

(c) Find the predicted probability of relief from psoriasis for a 50-year-old woman who is administered the medication A treatment.

The predicted probability is computed as:

$$P^0(\text{relief}) = \frac{\exp(-6.7992+0.1713 \cdot 50+3.8164)}{1+\exp(-6.7992+0.1713 \cdot 50+3.8164)} = 0.99625.$$

In SAS:

```

/*using fitted model for prediction*/
data predict;
input gender$ age medication$;
cards;
F 50 A
;
run;

data psoriasis;
set psoriasis predict;
run;

proc genmod;
class gender medication;
model relief(event="1") = gender age medication / dist=binomial link=logit;
output out=outdata p=presponse;
run;

proc print data=outdata (firstobs=31) noobs;
var prresponse;

```

```
run;
```

```
preponse  
0.99625
```

In R:

```
#using fitted model for prediction  
print(predict(fitted.model, data.frame(gender.rel="F", age=50,  
medication.rel="A"), type="response"))
```

```
0.9962508
```

(d) Repeat parts (a)-(c) but fit a probit model. Compare the results.

In SAS:

```
/*fitting probit model*/  
data psoriasis;  
input gender$ age medication$ relief @@;  
cards;  
M 37 A 1 F 24 A 1 F 15 A 1 M 31 B 1 F 39 B 1 M 31 B 1 M 20 A 1  
M 32 A 1 M 30 A 1 F 24 B 0 M 17 B 0 F 33 B 1 M 24 A 1 M 32 A 1  
F 27 A 1 M 16 A 1 F 33 A 1 F 28 A 0 M 51 B 1 F 35 B 0 M 16 B 0  
F 25 A 0 M 18 A 1 F 19 A 1 M 39 B 1 M 38 B 1 M 37 B 1 F 24 B 0  
F 39 B 0 F 33 B 0  
;  
  
proc genmod;  
class gender(ref="F") medication(ref="B");  
model relief(event="1") = gender age medication / dist=binomial link=probit;  
run;
```

```
Log Likelihood -10.3775
```

```
AIC 28.7549
```

```
AICC 30.3549
```

```
BIC 34.3597
```

Analysis of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	-4.0783	1.8291	-7.6633	-0.4932	4.97	0.0258
gender M	1	1.9230	0.8477	0.2615	3.5846	5.15	0.0233
gender F	0	0.0000	0.0000	0.0000	0.0000	.	.
age	1	0.1026	0.0514	0.0019	0.2033	3.99	0.0458
medication A	1	2.2335	0.8595	0.5490	3.9181	6.75	0.0094
medication B	0	0.0000	0.0000	0.0000	0.0000	.	.

The fitted model is  $\Phi^{-1}(\hat{P}(\text{relief})) = -4.0783 + 1.9230 \cdot \text{male} + 0.1026 \cdot \text{age} + 2.2335 \cdot \text{medication A}$ . All three predictors are significant at the 5% level (the same as in the logistic model).

```
/*checking model fit*/  
proc genmod;
```

```

model relief = / dist=binomial link=probit;
run;

```

**Log Likelihood -18.3259**

```

data deviance_test;
  deviance = -2*(-18.3259 - (-10.3775));
  pvalue = 1 - probchi(deviance,3);
run;

```

```

proc print noobs;
run;

```

```

deviance      pvalue
15.8968      .001190587

```

The p-value in the deviance test is less than 0.05, thus the model fits the data well (the same as the logistic model).

The estimated regression coefficients are interpreted as follows. The z-score of the estimated probability of relief from psoriasis for male patients is larger than that for female patients by 1.9230. As age increases by one year, the z-score of the estimated probability of relief increases by 0.1026. The z-score for the estimated probability of relief for a medication A patient is larger than that for a medication B patient by 2.2335.

The predicted value is obtained as:  $P^0(\text{relief}) = \Phi(-4.0783 + 0.1026 \cdot 50 + 2.2335) = \Phi(3.2852) = 0.99949$ . This predicted value is slightly larger than that in the logistic model.

In SAS:

```

/*using fitted model for prediction*/
data predict;
input gender$ age medication$;
cards;
F 50 A
;
run;

data psoriasis;
set psoriasis predict;
run;

proc genmod;
  class gender medication;
  model relief(event="1") = gender age medication / dist=binomial link=probit;
  output out=outdata p=presponse;
run;

proc print data=outdata (firstobs=31) noobs;
  var presponse;
run;

presponse
0.99949

```

In R:

```
#fitting probit model
psoriasis.data<- read.csv(file="C:/./Exercise3.2Data.csv", header=TRUE, sep=",")

#setting reference categories
gender.rel<- relevel(psoriasis.data$gender, ref="F")
medication.rel<- relevel(psoriasis.data$medication, ref="B")

#fitting the model
summary(fitted.model<- glm(relief ~ gender.rel + age + medication.rel,
data=psoriasis.data,family=binomial(link=probit)))
```

**Coefficients:**

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-4.07825	1.76524	-2.310	0.02087
gender.relM	1.92301	0.80597	2.386	0.01703
age	0.10260	0.04969	2.065	0.03892
medication.relA	2.23351	0.85603	2.609	0.00908

**AIC: 28.755**

```
#computing AICC
p<-4
n<-30
print(AICC<- -2*logLik(fitted.model)+2*p*n/(n-p-1))
```

**30.35491**

```
#outputting BIC
BIC(fitted.model)
```

**34.3597**

```
#checking model fit
null.model<- glm(relief ~ 1, data=psoriasis.data, family=binomial(link=probit))
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))
```

**15.89695**

```
print(p.value<- pchisq(deviance,3,lower.tail = FALSE))
```

**0.001190503**

```
#using fitted model for prediction
print(predict(fitted.model, data.frame(gender.rel="F", age=50,
medication.rel="A"), type="response"))
```

**0.9994907**

- (e) Repeat parts (a)-(c), fitting a complementary log-log model. Compare the results with the previous two models. Which of the three models has a better fit?

**In SAS:**

```
/*fitting complementary log-log model*/
data psoriasis;
input gender$ age medication$ relief @@;
cards;
M 37 A 1 F 24 A 1 F 15 A 1 M 31 B 1 F 39 B 1 M 31 B 1 M 20 A 1
M 32 A 1 M 30 A 1 F 24 B 0 M 17 B 0 F 33 B 1 M 24 A 1 M 32 A 1
F 27 A 1 M 16 A 1 F 33 A 1 F 28 A 0 M 51 B 1 F 35 B 0 M 16 B 0
```

```
F 25 A 0 M 18 A 1 F 19 A 1 M 39 B 1 M 38 B 1 M 37 B 1 F 24 B 0
F 39 B 0 F 33 B 0
;
```

```
proc genmod;
  class gender(ref="F") medication(ref="B");
  model relief(event="1") = gender age medication / dist=binomial link=cloglog;
run;
```

Log Likelihood -10.3062

AIC 28.6125

AICC 30.2125

BIC 34.2173

Analysis of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald	95% Confidence Limits	Wald Chi-Square	Pr > ChiSq
Intercept	1	-4.7318	2.1937	-9.0313	-0.4324	4.65	0.0310
gender	M 1	2.1558	0.9651	0.2643	4.0473	4.99	0.0255
gender	F 0	0.0000	0.0000	0.0000	0.0000	.	.
age	1	0.1069	0.0596	-0.0099	0.2236	3.22	0.0728
medication A	1	2.2361	0.9386	0.3964	4.0757	5.68	0.0172
medication B	0	0.0000	0.0000	0.0000	0.0000	.	.

The fitted model is  $1 - \hat{P}(\text{relief}) = \exp(-\exp(-4.7318 + 2.1558 \cdot \text{male} + 0.1069 \cdot \text{age} + 2.2361 \cdot \text{medication A}))$ . Gender and medication type are significant at the 5% level, whereas age is not significant. This is different from the logistic and probit models.

```
/*checking model fit*/
proc genmod;
  model relief = / dist=binomial link=logit;
run;

data deviance_test;
  deviance = -2*(-18.3259 - (-10.3062));
  pvalue = 1 - probchi(deviance,3);
run;

proc print noobs;
run;
```

```
deviance      pvalue
16.0394      .001113086
```

The complementary log-log model fits the data well since the p-value is very small.

Next, we interpret the estimated beta coefficients for the significant predictors. For male patients, the estimated probability of no relief from psoriasis is that for female patients raised to the power  $\exp(2.1558) = 8.63$ . The estimated probability of no relief for patients taking medication A is that for patients taking medication B raised to the power  $\exp(2.2361) = 9.36$ . It means that the estimated probability of no relief for males (medication A patients) is much smaller than that for females (medication B patients). This is in accordance with the logistic and probit models.

The predicted value in this model is derived as:  $P^0(\text{relief}) = 1 - \exp(-\exp(-4.7318 + 0.1069 \cdot 50 + 2.2361)) = 0.99999997$ . This predicted value is larger than those in the logistic and probit models.

In SAS:

```
/*using fitted model for prediction*/
data predict;
input gender$ age medication$;
cards;
F 50 A
;

data psoriasis;
set psoriasis predict;
run;

proc genmod;
  class gender medication;
  model relief(event="1") = gender age medication / dist=binomial link=cloglog;
  output out=outdata p=preponse;
run;

proc print data=outdata (firstobs=31) noobs;
  var preponse;
run;
```

```
preponse
  1.00000
```

In R:

```
#fitting complementary log-log model
psoriasis.data<- read.csv(file="C:/./Exercise3.2Data.csv", header=TRUE, sep=",")

#setting reference categories
gender.rel<- relevel(psoriasis.data$gender, ref="F")
medication.rel<- relevel(psoriasis.data$medication, ref="B")

#running the model
summary(fitted.model<- glm(relief ~ gender.rel + age + medication.rel,
data=psoriasis.data,family=binomial(link=cloglog)))
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-4.73174	2.12142	-2.230	0.0257
gender.relM	2.15577	0.87247	2.471	0.0135
age	0.10686	0.05708	1.872	0.0612
medication.relA	2.23606	0.96039	2.328	0.0199

AIC: 28.612

```
#computing AICC
p<-4
```

```

n<-30
print(AICC<- -2*logLik(fitted.model)+2*p*n/(n-p-1))

30.21248

#outputting BIC
BIC(fitted.model)

34.21727

#checking model fit
null.model<- glm(relief ~ 1, data=psoriasis.data, family=binomial(link=cloglog))
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))

16.03937

print(p.value<- pchisq(deviance,3,lower.tail = FALSE))

0.0011131

#using fitted model for prediction
print(predict(fitted.model, data.frame(gender.rel="F", age=50,
medication.rel="A"), type="response"))

1

```

As we can see, the complementary log-log model has the smallest values in the AIC, AICC, and BIC criteria, and thus has the best fit.

	logistic	probit	cloglog
AIC	29.0084	28.7549	28.6125
AICC	30.6084	30.3549	30.2125
BIC	34.6132	34.3597	34.2173

**EXERCISE 3.3.** (a) Fit a binary logistic model to the data. What predictors turn out to be significant at the 5% level? How good is the fit of the model?

In SAS:

```

/*fitting logistic model*/
data novel;
input success$ cover$ methods$ novels$ years @@;
cards;
yes   yes   one   many   18   no   no   one   first   7
no    yes   none  several 10   yes  yes  many  many   6
no    yes   none  several 1   no   no   one   several 1
no    no    one   first   11  yes  no   one   several 19
yes   yes   none  first   5   no   no   none  many   2
no    no    one   several 10  no   no   many  many   9
yes   no    many  several 6   yes  yes  many  many   8
no    no    one   several 12  no   no   none  many   2

```

```

yes    no    none several 17  yes yes many first 10
yes    no    none several 7   no  no one  first 12
no     yes    none several 7   no  yes none many 4
no     no     one  several 9   yes no many several 13
yes    yes    none first 6    no  no none many 2
yes    yes    one  several 7   yes yes many many 17
yes    yes    many first 18   yes yes one several 17
no     yes    none several 9   no  no one several 11
yes    yes    many first 17   no  no many many 1
no     no     many many 6     no  yes none several 1
yes    yes    many first 6     yes yes one many 4
no     yes    none many 7     no  no one first 12
no     no     one  several 7   yes yes one several 9
no     no     one  several 8   no  no one several 2
;

```

```

proc genmod;
  class cover(ref="no") methods(ref="none") novels(ref="many");
  model success(event="yes") = cover methods novels years / dist=binomial
    link=logit;
run;

```

Log Likelihood -16.0357

AIC 46.0714

AICC 49.1825

BIC 58.5607

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	-6.8762	2.3656	-11.5127	-2.2397	8.45	0.0037
cover	yes	1	3.5238	1.3120	0.9522	6.0954	7.21	0.0072
cover	no	0	0.0000	0.0000	0.0000	0.0000	.	.
methods	many	1	3.9286	1.7918	0.4168	7.4404	4.81	0.0283
methods	one	1	0.7008	1.1914	-1.6344	3.0360	0.35	0.5564
methods	none	0	0.0000	0.0000	0.0000	0.0000	.	.
novels	first	1	1.8768	1.5784	-1.2168	4.9703	1.41	0.2344
novels	several	1	1.3992	1.3254	-1.1986	3.9969	1.11	0.2911
novels	many	0	0.0000	0.0000	0.0000	0.0000	.	.
years		1	0.2907	0.1269	0.0420	0.5394	5.25	0.0220

The fitted model has the form:  $\ln \frac{P(\text{success})}{1-P(\text{success})} = -6.8762 + 3.5238 \cdot \text{catchy cover} + 3.9286 \cdot \text{many methods} + 0.7008 \cdot \text{one method} + 1.8768 \cdot \text{first novel} + 1.3992 \cdot \text{several novels} + 0.2907 \cdot \text{years}.$

The significant predictors at the 5% level are catchy cover, many promotional methods, and years the publisher was in business.

```
/*checking model fit*/
```

```

proc genmod;
  model success = / dist=binomial link=logit;
run;

```

**Log Likelihood -30.0881**

```
data deviance_test;
  deviance = -2*(-30.0881 - (-16.0357));
  pvalue = 1 - probchi(deviance,6);
run;
```

```
proc print noobs;
run;
```

```
deviance      pvalue
28.1048      .000089787
```

The fit of this model is very good as indicated by the tiny p-value in the deviance test.

**In R:**

```
#fitting logistic model
novel.data<- read.csv(file="C:/./Exercise3.3Data.csv", header=TRUE, sep=",")

#setting reference categories
cover.rel<- relevel(novel.data$cover, ref="no")
methods.rel<- relevel(novel.data$methods, ref="none")
novels.rel<- relevel(novel.data$novels, ref="many")

#fitting the model
summary(fitted.model<- glm(success~ cover.rel + methods.rel + novels.rel
+ years, data=novel.data,family=binomial(link=logit)))
```

**Coefficients:**

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-6.8762	2.3656	-2.907	0.00365
cover.relyes	3.5238	1.3120	2.686	0.00724
methods.relmany	3.9286	1.7917	2.193	0.02834
methods.relone	0.7008	1.1914	0.588	0.55642
novels.relfirst	1.8768	1.5784	1.189	0.23442
novels.relseveral	1.3992	1.3254	1.056	0.29113
years	0.2907	0.1269	2.291	0.02198

**AIC: 46.071**

```
#computing AICC
p<-7
n<-44
print(AICC<- -2*logLik(fitted.model)+2*p*n/(n-p-1))
```

**49.18253**

```
#outputting BIC
BIC(fitted.model)
```

**58.56075**

```
#checking model fit
null.model<- glm(success ~ 1, data=novel.data, family=binomial(link=logit))
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))
```

28.10479

```
print(p.value<- pchisq(deviance,6,lower.tail = FALSE))
```

8.97874e-05

(b) Give an interpretation of the estimated significant beta coefficients.

The estimated odds in favor of financial success for a novel with a catchy cover are  $\exp(3.5238) \cdot 100\% = 3,391.31\%$  of those for a novel without a catchy cover. The estimated odds for a publisher with many promotional methods are  $\exp(3.9286) \cdot 100\% = 5,083.58\%$  of those for a publisher with no promotional methods. For every additional year, a publisher was in business prior to the publication of a novel, the estimated odds in favor of the financial success of the novel increase by  $(\exp(0.2907) - 1) \cdot 100\% = 33.74\%$ .

(c) Suppose a newly established publishing house prints a novel by some previously unknown author and doesn't advertise the publication. Find the estimated probability that this novel is successful financially if it has an extremely catchy cover.

The predicted value is calculated as follows:  $P^0(\text{success}) = \frac{\exp(-6.8762+3.5238+1.8768)}{1+\exp(-6.8762+3.5238+1.8768)} = 0.18609$ .

In SAS:

```
/*using fitted model for prediction*/
data predict;
input cover$ methods$ novels$ years;
cards;
yes none first 0
;
run;

data novel;
set novel predict;
run;

proc genmod;
  class cover methods novels;
  model success(event="yes") = cover methods novels years / dist=binomial
    link=logit;
  output out=outdata p=psuccess;
run;

proc print data=outdata (firstobs=45) noobs;
  var psuccess;
run;
```

```
psuccess
0.18609
```

In R:

```
#using fitted model for prediction
print(predict(fitted.model, data.frame(cover.rel="yes", methods.rel="none",
novels.rel="first", years=0), type="response"))
```

0.186086

(f) Redo parts (a) through (c), fitting a probit model.

In SAS:

```
/*fitting probit model*/
data novel;
input success$ cover$ methods$ novels$ years @@;
cards;
yes    yes    one    many    18    no    no    one    first    7
no     yes    none   several 10    yes   yes   many   many    6
no     yes    none   several 1    no    no    one    several 1
no     no     one    first    11   yes   no    one    several 19
yes    yes    none   first    5    no    no    none   many    2
no     no     one    several 10   no    no    many   many    9
yes    no     many   several 6    yes   yes   many   many    8
no     no     one    several 12   no    no    none   many    2
yes    no     none   several 17   yes   yes   many   first    10
yes    no     none   several 7    no    no    one    first    12
no     yes    none   several 7    no    yes   none   many    4
no     no     one    several 9    yes   no    many   several 13
yes    yes    none   first    6    no    no    none   many    2
yes    yes    one    several 7    yes   yes   many   many    17
yes    yes    many   first    18   yes   yes   one    several 17
no     yes    none   several 9    no    no    one    several 11
yes    yes    many   first    17   no    no    many   many    1
no     no     many   many    6    no    yes   none   several 1
yes    yes    many   first    6    yes   yes   one    many    4
no     yes    none   many    7    no    no    one    first    12
no     no     one    several 7    yes   yes   one    several 9
no     no     one    several 8    no    no    one    several 2
;

proc genmod;
  class cover(ref="no") methods(ref="none") novels(ref="many");
  model success(event="yes") = cover methods novels years / dist=binomial
link=probit;
run;
```

Log Likelihood -16.2531

AIC 46.5062

AICC 49.6173

BIC 58.9955

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	-3.6827	1.1644	-5.9648	-1.4006	10.00	0.0016
cover	yes	1	1.8841	0.6447	0.6205	3.1477	8.54	0.0035
cover	no	0	0.0000	0.0000	0.0000	0.0000	.	.
methods	many	1	2.0626	0.9293	0.2412	3.8841	4.93	0.0265
methods	one	1	0.2483	0.6293	-0.9850	1.4816	0.16	0.6932
methods	none	0	0.0000	0.0000	0.0000	0.0000	.	.

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald	95% Confidence Limits	Wald Chi-Square	Pr > ChiSq
novels	first	1	0.9553	0.8852	-0.7796	2.6901	1.16	0.2805
novels	several	1	0.7903	0.7415	-0.6630	2.2436	1.14	0.2865
novels	many	0	0.0000	0.0000	0.0000	0.0000	.	.
years		1	0.1623	0.0679	0.0291	0.2955	5.71	0.0169

The fitted model can be written as:  $\Phi^{-1}(\hat{P}(\text{success})) = -3.6827 + 1.8841 \cdot \text{catchy cover} + 2.0626 \cdot \text{many methods} + 0.2483 \cdot \text{one method} + 0.9553 \cdot \text{first novel} + 0.7903 \cdot \text{several novels} + 0.1623 \cdot \text{years}$ .

The significant predictors at the 5% level are catchy cover, many promotional methods, and years the publisher was in business. The same as in the logistic model.

```
/*checking model fit*/
proc genmod;
  model success = / dist=binomial link=probit;
run;
```

Log Likelihood -30.0881

```
data deviance_test;
  deviance = -2*(-30.0881 - (-16.2531));
  pvalue = 1 - probchi(deviance,6);
run;

proc print noobs;
run;
```

```
deviance      pvalue
    27.67    .000108405
```

The probit model fits the data very well since the p-value is tiny. The estimated regression coefficients for significant predictors yield the following interpretation. The z-score for the estimated probability of financial success for a novel with a catchy cover is 1.8841 units larger than that for a novel without a catchy cover. The z-score for the estimated probability of success of a novel for publishing houses with many promotional methods exceeds by 2.0626 that for publishing houses with no promotional methods. As the number of years in business increases by one, the z-score increases by 0.1623.

Both, the logistic and probit models agree on the direction of influence of the significant predictors on the estimated probability of success.

The predicted probability is obtained as:  $P^0(\text{relief}) = \Phi(-3.6827 + 1.8841 + 0.9553) = \Phi(-0.8433) = 0.19953$ . This prediction exceeds that in the logistic model.

In SAS:

```
/*using fitted model for prediction*/
data predict;
input cover$ methods$ novels$ years;
cards;
yes none first 0
;
```

```
run;

data novel;
set novel predict;
run;

proc genmod;
class cover methods novels;
model success(event="yes") = cover methods novels years / dist=binomial
link=probit;
output out=outdata p=psuccess;
run;

proc print data=outdata (firstobs=45) noobs;
var psuccess;
run;

psuccess
0.19953
```

### In R:

```
#fitting probit model
novel.data<- read.csv(file="C:/./Exercise3.3Data.csv", header=TRUE, sep=",")

#setting reference categories
cover.rel<- relevel(novel.data$cover, ref="no")
methods.rel<- relevel(novel.data$methods, ref="none")
novels.rel<- relevel(novel.data$novels, ref="many")

#running the model
summary(fitted.model<- glm(success~ cover.rel + methods.rel + novels.rel + years,
data=novel.data,family=binomial(link=probit)))
```

### Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-3.68266	1.20990	-3.044	0.00234
cover.relyes	1.88408	0.67751	2.781	0.00542
methods.relmany	2.06258	0.94055	2.193	0.02831
methods.relone	0.24823	0.67040	0.370	0.71118
novels.relfirst	0.95527	0.87638	1.090	0.27570
novels.relseveral	0.79030	0.72763	1.086	0.27742
years	0.16230	0.06649	2.441	0.01465

AIC: 46.506

```
#computing AICC
p<-7
n<-44
print(AICC<- -2*logLik(fitted.model)+2*p*n/(n-p-1))
```

49.6173

```
#outputting BIC
BIC(fitted.model)
```

58.99552

```
#checking model fit
```

```
null.model<- glm(success ~ 1, data=novel.data, family=binomial(link=probit))
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))
```

27.67002

```
print(p.value<- pchisq(deviance,6,lower.tail = FALSE))
```

0.000108404

```
#using fitted model for prediction
print(predict(fitted.model, data.frame(cover.rel="yes",
methods.rel="none", novels.rel="first", years=0), type="response"))
```

0.1995286

(g) Redo parts (a) through (c) with the complementary log-log model. How good is the model fit compared to the logistic and probit models?

In SAS:

```
/*fitting complementary log-log model*/
data novel;
input success$ cover$ methods$ novels$ years @@;
cards;
yes yes one many 18 no no one first 7
no yes none several 10 yes yes many many 6
no yes none several 1 no no one several 1
no no one first 11 yes no one several 19
yes yes none first 5 no no none many 2
no no one several 10 no no many many 9
yes no many several 6 yes yes many many 8
no no one several 12 no no none many 2
yes no none several 17 yes yes many first 10
yes no none several 7 no no one first 12
no yes none several 7 no yes none many 4
no no one several 9 yes no many several 13
yes yes none first 6 no no none many 2
yes yes one several 7 yes yes many many 17
yes yes many first 18 yes yes one several 17
no yes none several 9 no no one several 11
yes yes many first 17 no no many many 1
no no many many 6 no yes none several 1
yes yes many first 6 yes yes one many 4
no yes none many 7 no no one first 12
no no one several 7 yes yes one several 9
no no one several 8 no no one several 2
;

proc genmod;
class cover(ref="no") methods(ref="none") novels(ref="many");
model success(event="yes") = cover methods novels years / dist=binomial
link=cloglog;
run;
```

Log Likelihood -15.3985

AIC 44.7971

AICC 47.9082

BIC 57.2864

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	-6.1473	2.0929	-10.2493	-2.0453	8.63	0.0033
cover	yes	1	2.8126	1.0659	0.7235	4.9017	6.96	0.0083
cover	no	0	0.0000	0.0000	0.0000	0.0000	.	.
methods	many	1	3.4196	1.5548	0.3722	6.4670	4.84	0.0279
methods	one	1	0.8348	0.9811	-1.0882	2.7578	0.72	0.3949
methods	none	0	0.0000	0.0000	0.0000	0.0000	.	.
novels	first	1	1.8391	1.4227	-0.9492	4.6275	1.67	0.1961
novels	several	1	1.2622	1.1214	-0.9358	3.4602	1.27	0.2604
novels	many	0	0.0000	0.0000	0.0000	0.0000	.	.
years		1	0.2187	0.1029	0.0171	0.4204	4.52	0.0335

The fitted model has the form  $1 - \hat{P}(\text{success}) = \exp(-\exp(-6.1473 + 2.8126 \cdot \text{catchy cover} + 3.4196 \cdot \text{many methods} + 0.8348 \cdot \text{one method} + 1.8391 \cdot \text{first novel} + 1.2622 \cdot \text{several novels} + 0.2187 \cdot \text{years}))$ . The significant at the 5% level predictors are the same as in the logistic and probit models. Namely, catchy cover, many promotional methods, and years the publisher was in business.

```
/*checking model fit*/
proc genmod;
  model success = / dist=binomial link=cloglog;
run;
```

Log Likelihood -30.0881

```
data deviance_test;
  deviance = -2*(-30.0881 - (-15.3985));
  pvalue = 1 - probchi(deviance,6);
run;
```

```
proc print noobs;
run;
```

```
deviance      pvalue
29.3792      .000051563
```

The model fits the data very well as indicated by the tiny p-value. The interpretation of the estimated significant regression coefficients goes as follows: The estimated probability of financial failure for a novel with a catchy cover is that of a novel without a catchy cover raised to the power  $\exp(2.8126) = 16.65$ . The estimated probability of financial failure of a novel for publishing houses with many promotional methods that for publishing houses with no promotional methods raised to the power  $\exp(3.4196) = 30.56$ . For every additional year the publishing house was in business, the estimated probability of failure is raised to the power  $\exp(0.2187) = 1.24$ .

This interpretation is in agreement with the ones given in the logistic and probit models (in the same direction and roughly magnitude).

The prediction in this model is calculated as:  $P^0(\text{success}) = 1 - \exp(-\exp(-6.1473 + 2.8126 + 1.8391)) = 0.200776$ . This predicted value is larger than those obtained through logistic and probit modeling.

## In SAS:

```
/*using fitted model for prediction*/
data predict;
input cover$ methods$ novels$ years;
cards;
yes none first 0
;
run;

data novel;
set novel predict;
run;

proc genmod;
  class cover methods novels;
  model success(event="yes") = cover methods novels years / dist=binomial
link=cloglog;
  output out=outdata p=psuccess;
run;

proc print data=outdata (firstobs=45) noobs;
  var psuccess;
run;
```

**psuccess**  
**0.20078**

## In R:

```
#fitting complementary log-log model
novel.data<- read.csv(file="C:/./Exercise3.3Data.csv", header=TRUE, sep=",")

#setting reference categories
cover.rel<- relevel(novel.data$cover, ref="no")
methods.rel<- relevel(novel.data$methods, ref="none")
novels.rel<- relevel(novel.data$novels, ref="many")

#running the model
summary(fitted.model<- glm(success~ cover.rel + methods.rel + novels.rel + years,
data=novel.data,family=binomial(link=cloglog)))
```

### Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-6.14718	1.98767	-3.093	0.00198
cover.relyes	2.81247	0.96108	2.926	0.00343
methods.remany	3.41940	1.37735	2.483	0.01304
methods.renone	0.83463	0.86156	0.969	0.33267
novels.refirst	1.83906	1.30348	1.411	0.15828
novels.reseveral	1.26216	1.02470	1.232	0.21805
years	0.21874	0.09438	2.318	0.02047

**AIC: 44.797**

```
#computing AICC
p<-7
n<-44
print(AICC<- -2*logLik(fitted.model)+2*p*n/(n-p-1))
```

47.9082

```
#outputting BIC
BIC(fitted.model)
```

57.28642

```
#checking model fit
null.model<- glm(success ~ 1, data=novel.data, family=binomial(link=cloglog))
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))
```

29.37912

```
print(p.value<- pchisq(deviance,6,lower.tail = FALSE))
```

5.156521e-05

```
#using fitted model for prediction
print(predict(fitted.model, data.frame(cover.rel="yes", methods.rel="none",
novels.rel="first", years=0), type="response"))
```

0.2007672

To see which of the three models fits the data the best, we compare the AIC, AICC, and BIC values. For convenience, we repeat them below.

	logistic	probit	cloglog
AIC	46.0714	46.5062	44.7971
AICC	49.1825	49.6173	47.9082
BIC	58.5607	58.9955	57.2864

The complementary log-log model has smaller values in all three criteria, and thus has a better fit.

**EXERCISE 3.4.** (a) Run the binary logistic model, regressing on all the predictors. Identify variables that are significant predictors of loan default at the 5% level of significance. Analyze the model fit.

In SAS:

```
/*fitting logistic model*/
data loan;
input LTV age income$ default$ @@;
cards;
70 41 low no 70 25 high yes 65 48 low no 65 48 high no
60 32 high yes 50 48 high no 55 53 low no 85 38 high yes
80 43 low yes 50 33 low no 60 42 low no 90 23 low yes
80 31 high no 70 37 high no 40 39 high no 80 40 low no
70 52 high no 80 29 low yes 40 44 low no 80 36 high no
90 47 high no 80 29 high no 70 24 low yes 30 42 high no
50 33 low no 80 36 low no 75 54 low no 75 29 high yes
70 38 low no 60 35 low no 95 30 low yes 80 34 low yes
```

```
75 43 low yes 75 47 high no 85 47 low yes
;
```

```
proc genmod;
  class income(ref="high");
  model default(event="yes") = LTV age income / dist=binomial link=logit;
run;
```

Log Likelihood -14.2347

**AIC 36.4693**

**AICC 37.8027**

**BIC 42.6907**

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald	95% Confidence Limits	Chi-Square	Pr > ChiSq
Intercept	1	-3.0087	4.0955	-11.0356	5.0183	0.54	0.4626
LTV	1	0.1059	0.0512	0.0054	0.2063	4.27	0.0388
age	1	-0.1616	0.0731	-0.3049	-0.0182	4.88	0.0272
income low	1	1.1162	1.0249	-0.8926	3.1250	1.19	0.2761
income high	0	0.0000	0.0000	0.0000	0.0000	.	.

The fitted model is written as:  $\ln \frac{\hat{P}(\text{default})}{1-\hat{P}(\text{default})} = -3.0087 + 0.1059 \cdot LTV - 0.1616 \cdot \text{age} + 1.1162 \cdot \text{low income}$ . LTV and age are significant at the 5% level.

```
/*checking model fit*/
proc genmod;
  model default = / dist=binomial link=logit;
run;
```

Log Likelihood -22.5019

```
data deviance_test;
  deviance = -2*(-22.5019 - (-14.2347));
  pvalue = 1 - probchi(deviance,3);
run;
```

```
proc print noobs;
run;
```

```
deviance    pvalue
16.5344    .000880948
```

The model fit is excellent as shown by the small p-value in the deviance test.

In R:

```
#fitting logistic model
rate.data<- read.csv(file="C:/./Exercise3.4Data.csv", header=TRUE, sep=",")

#running the model
summary(fitted.model<- glm(default ~ LTV + age + income, data=rate.data,
family=binomial(link=logit)))
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-3.00869	4.09545	-0.735	0.4626
LTV	0.10586	0.05124	2.066	0.0388
age	-0.16157	0.07314	-2.209	0.0272
incomelow	1.11619	1.02490	1.089	0.2761

AIC: 36.469

```
#computing AICC
p<-4
n<-35
print(AICC<- -2*logLik(fitted.model)+2*p*n/(n-p-1))
```

37.80266

```
#outputting BIC
BIC(fitted.model)
```

42.69072

```
#checking model fit
null.model<- glm(default ~ 1, data=rate.data, family=binomial(link=logit))
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))
```

16.53454

```
print(p.value<- pchisq(deviance,3,lower.tail = FALSE))
```

0.0008808876

(b) Interpret the estimated significant beta coefficients. What is your suggestion for the bank to decrease the default rate of home equity loans?

As the loan-to-value ratio increases by one, the estimated odds in favor of default increase by  $(\exp(0.1059) - 1) \cdot 100\% = 11.17\%$ . As the age of a client increases by one year, the estimated odds in favor of default change by  $(\exp(-0.1616) - 1) \cdot 100\% = -14.92\%$ , that is, decrease by 14.92%. To decrease the default rate, the bank might want to give loans with a smaller loan-to-value ratio, and/or give loans to older clients.

(c) Give a point estimate for the probability of loan default if the LTV ratio is 50%, and the borrower is a 50-year-old man with a high income.

The predicted probability is computed as:  $P^0(\text{default}) = \frac{\exp(-3.0087+0.1059 \cdot 50-0.1616 \cdot 50)}{1+\exp(-3.0087+0.1059 \cdot 50-0.1616 \cdot 50)} = 0.0030374$ .

In SAS:

```
/*using fitted model for prediction*/
data predict;
input LTV age income$;
cards;
50 50 high
;
run;

data loan;
```

```

set loan predict;
run;

proc genmod;
  class income;
  model default(event="yes") = LTV age income / dist=binomial link=logit;
  output out=outdata p=pdefault;
run;

proc print data=outdata (firstobs=36) noobs;
  var pdefault;
run;

      pdefault
    .003035760

```

In R:

```

#using fitted model for prediction
print(predict(fitted.model, data.frame(LTV=50, age=50, income="high"),
type="response"))

```

0.00303576

(d) Repeat the previous parts, fitting a probit model. How different are the results?

In SAS:

```

/*fitting probit model*/
data loan;
input LTV age income$ default$ @@;
cards;
70 41 low no 70 25 high yes 65 48 low no 65 48 high no
60 32 high yes 50 48 high no 55 53 low no 85 38 high yes
80 43 low yes 50 33 low no 60 42 low no 90 23 low yes
80 31 high no 70 37 high no 40 39 high no 80 40 low no
70 52 high no 80 29 low yes 40 44 low no 80 36 high no
90 47 high no 80 29 high no 70 24 low yes 30 42 high no
50 33 low no 80 36 low no 75 54 low no 75 29 high yes
70 38 low no 60 35 low no 95 30 low yes 80 34 low yes
75 43 low yes 75 47 high no 85 47 low yes
;

proc genmod;
  class income(ref="high");
  model default(event="yes") = LTV age income / dist=binomial link=probit;
run;

```

Log Likelihood -14.0867

**AIC 36.1733**

**AICC 37.5067**

**BIC 42.3947**

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald	95% Confidence Limits	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.6059	2.3230	-6.1589	2.9472	0.48	0.4894
LTV	1	0.0620	0.0287	0.0057	0.1183	4.67	0.0308
age	1	-0.0987	0.0431	-0.1832	-0.0141	5.24	0.0221
income low	1	0.6392	0.5932	-0.5234	1.8019	1.16	0.2812
income high	0	0.0000	0.0000	0.0000	0.0000	.	.

The fitted model is  $\Phi^{-1}(\hat{P}(\text{default})) = -1.6059 + 0.0620 \cdot \text{LTV} - 0.0987 \cdot \text{age} + 0.6392 \cdot \text{low income}$ . Significant at the 5% level are LTV and age, the same as in the logistic model.

```
/*checking model fit*/
proc genmod;
  model default = / dist=binomial link=probit;
run;
```

Log Likelihood -22.5019

```
data deviance_test;
  deviance = -2*(-22.5019 - (-14.0867));
  pvalue = 1 - probchi(deviance,3);
run;

proc print noobs;
run;
```

```
deviance      pvalue
16.8304      .000765830
```

The probit model fits the data well because the p-value is very small. The significant estimated coefficients are interpreted as follows. For a one-percent increase in the loan-to-value ratio, the z-score of the estimated probability of default increases by 0.0620 units. If the age of a client increases by one year, the z-score of the estimated probability of default decreases by 0.0987 units. The same direction is observed in the logistic model.

The predicted probability of default in this model is equal to  $P^0(\text{default}) = \Phi(-1.6059 + 0.0620 \cdot 50 - 0.0987 \cdot 50) = \Phi(-3.4409) = 0.0002899$ . This probability is a magnitude smaller than that obtained in the logistic regression.

In SAS:

```
/*using fitted model for prediction*/
data predict;
input LTV age income$;
cards;
50 50 high
;
run;

data loan;
set loan predict;
run;

proc genmod;
```

```

class income;
  model default(event="yes") = LTV age income / dist=binomial link=probit;
  output out=outdata p=pdefault;
run;

```

```

proc print data=outdata (firstobs=36) noobs;
  var pdefault;
run;

```

```

  pdefault
.000292304

```

In R:

```

#fitting probit model
rate.data<- read.csv(file="C:/./Exercise3.4Data.csv", header=TRUE, sep=",")

```

```

#running the model
summary(fitted.model<- glm(default ~ LTV + age + income, data=rate.data,
family=binomial(link=probit)))

```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-1.60587	2.34383	-0.685	0.4933
LTV	0.06200	0.02853	2.173	0.0297
age	-0.09865	0.04121	-2.394	0.0167
incomelow	0.63924	0.59365	1.077	0.2816

AIC: 36.173

```

#computing AICC
p<-4
n<-35
print(AICC<- -2*logLik(fitted.model)+2*p*n/(n-p-1))

```

37.50665

```

#outputting BIC
BIC(fitted.model)

```

42.39471

```

#checking model fit
null.model<- glm(default ~ 1, data=rate.data, family=binomial(link=probit))
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))

```

16.83055

```

print(p.value<- pchisq(deviance,3,lower.tail = FALSE))

```

0.0007657748

```

#using fitted model for prediction
print(predict(fitted.model, data.frame(LTV=50, age=50, income="high"),
type="response"))

```

0.0002923167

(e) Redo parts (a)-(c) with a complementary log-log model. Discuss the differences between the three models, if any. Which model fits the data the best?

In SAS:

```
/*fitting complementary log-log model*/
data loan;
input LTV age income$ default$ @@;
cards;
70 41 low no 70 25 high yes 65 48 low no 65 48 high no
60 32 high yes 50 48 high no 55 53 low no 85 38 high yes
80 43 low yes 50 33 low no 60 42 low no 90 23 low yes
80 31 high no 70 37 high no 40 39 high no 80 40 low no
70 52 high no 80 29 low yes 40 44 low no 80 36 high no
90 47 high no 80 29 high no 70 24 low yes 30 42 high no
50 33 low no 80 36 low no 75 54 low no 75 29 high yes
70 38 low no 60 35 low no 95 30 low yes 80 34 low yes
75 43 low yes 75 47 high no 85 47 low yes
;

proc genmod;
class income(ref="high");
model default(event="yes") = LTV age income / dist=binomial link=cloglog;
run;
```

Log Likelihood -14.2179

AIC 36.4358

AICC 37.7691

BIC 42.6572

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald	95% Confidence Limits	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.6814	3.5174	-9.5754	4.2125	0.58	0.4459
LTV	1	0.0790	0.0416	-0.0026	0.1605	3.60	0.0578
age	1	-0.1225	0.0555	-0.2314	-0.0137	4.87	0.0273
income low	1	0.8907	0.7346	-0.5491	2.3305	1.47	0.2253
income high	0	0.0000	0.0000	0.0000	0.0000	.	.

The fitted complementary log-log model can be written as  $1 - \hat{P}(\text{default}) = \exp(-\exp(-2.6814 + 0.0790 \cdot \text{LTV} - 0.1225 \cdot \text{age} + 0.8907 \cdot \text{low income}))$ . Age is a significant predictor at the 5% significance level, whereas LTV is only marginally significant at this level. This is different from what we have seen in the two previous models.

```
/*checking model fit*/
proc genmod;
model default = / dist=binomial link=cloglog;
run;
```

Log Likelihood -22.5019

```
data deviance_test;
deviance = -2*(-22.5019 - (-14.2179));
pvalue = 1 - probchi(deviance,3);
run;

proc print noobs;
run;
```

deviance	pvalue
16.568	.000867061

The complementary log-log model has a very good fit for the data since the p-value is very small.

As the age of a client increases by one year, the probability of no default is raised to the power  $\exp(-0.1225) = 0.8847$ , that is the probability of no default increases. This agrees with our findings in the logistic and probit models.

The predicted probability is calculated as:  $P^0(\text{default}) = 1 - \exp(-\exp(-2.6814 + 0.0790 \cdot 50 - 0.1225 \cdot 50)) = 0.007748$ . This probability is larger than those predicted by the logistic and probit models.

In SAS:

```
/*using fitted model for prediction*/
data predict;
input LTV age income$;
cards;
50 50 high
;
run;

data loan;
set loan predict;
run;

proc genmod;
class income;
model default (event="yes") = LTV age income / dist=binomial link=cloglog;
output out=outdata p=pdefault;
run;

proc print data=outdata (firstobs=36) noobs;
var pdefault;
run;
```

	pdefault
	.007713442

In R:

```
#fitting complementary log-log model
rate.data<- read.csv(file="C:/./Exercise3.4Data.csv", header=TRUE, sep=",")

#running the model
summary(fitted.model<- glm(default ~ LTV + age + income, data=rate.data,
family=binomial(link=cloglog)))
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-2.68156	3.27613	-0.819	0.4131
LTV	0.07896	0.04114	1.919	0.0550
age	-0.12254	0.05111	-2.398	0.0165
incomelow	0.89073	0.72016	1.237	0.2161

AIC: 36.436

```

#computing AICC
p<-4
n<-35
print(AICC<- -2*logLik(fitted.model)+2*p*n/(n-p-1))

37.7691

#outputting BIC
BIC(fitted.model)

42.65716

#checking model fit
null.model<- glm(default ~ 1, data=rate.data, family=binomial(link=cloglog))
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))

16.56811

print(p.value<- pchisq(deviance,3,lower.tail = FALSE))

0.0008670163

#using fitted model for prediction
print(predict(fitted.model, data.frame(LTV=50, age=50, income="high"),
type="response"))

0.007713725

```

The probit model has the smallest AIC, AICC, and BIC values and therefore has the best fit.

	logistic	probit	cloglog
AIC	36.4693	36.1733	36.4358
AICC	37.8027	37.5067	37.7691
BIC	42.6907	42.3947	42.6572

**EXERCISE 3.5.** (a) Model the probability of being a cardiac patient via binary logistic regression. Write the fitted model explicitly. Discuss the goodness of fit of the model and the significance of the regression coefficients. Assume  $\alpha = 0.01$  for all tests.

In SAS:

```

/*fitting logistic model*/
data cardiac;
input group A W @@;
cards;
1 8 2 1 1 2 1 2 1 1 4 0 1 2 7 1 6 3 1 2 8 1 1 9
1 3 0 1 0 2 1 3 2 1 2 7 1 2 7 1 2 8 1 6 0 1 3 5
1 1 0 1 7 1 1 4 3 1 2 4 1 5 3 1 7 1 1 8 1 1 0 6
0 0 9 0 2 1 0 0 8 0 1 3 0 3 1 0 1 4 0 0 8 0 1 6
0 0 9 0 2 2 0 4 4 0 0 6 0 3 2 0 1 2 0 2 5 0 4 0
0 8 1 0 2 7 0 0 10 0 0 5 0 0 6 0 2 1 0 0 7 0 0 6
;

```

```
proc genmod;
  model group(event="1") = A W / dist=binomial link=logit;
run;
```

**Log Likelihood -29.0551**

AIC 64.1102

AICC 64.6557

BIC 69.7238

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald	95% Confidence Limits	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.1160	0.8896	-2.8596	0.6276	1.57	0.2097
A	1	0.4378	0.1989	0.0480	0.8276	4.85	0.0277
W	1	0.0277	0.1257	-0.2188	0.2741	0.05	0.8260

The fitted logistic model is  $\ln \frac{\hat{P}(\text{cardiac})}{1-\hat{P}(\text{cardiac})} = -1.1160 + 0.4378 \cdot \# \text{ of arches} + 0.0277 \cdot \# \text{ of whorls}$ . The number of arches is a significant predictor at the 5% level.

```
/*checking model fit*/
proc genmod;
  model group = / dist=binomial link=logit;
run;
```

**Log Likelihood -33.2711**

```
data deviance_test;
  deviance = -2*(-33.2711 - (-29.0551));
  pvalue = 1 - probchi(deviance,2);
run;
```

```
proc print noobs;
run;
```

```
deviance    pvalue
      8.432  0.014758
```

The model fits the data reasonably well since the p-value is smaller than 0.05.

In R:

```
#fitting logistic model
cardiac.data<- read.csv(file="C:/./Exercise3.5Data.csv", header=TRUE, sep=",")

#running the model
summary(fitted.model<- glm(group ~ A + W, data=cardiac.data,
family=binomial(link=logit)))
```

**Coefficients:**

```
Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.11602    0.88961  -1.255  0.2097
A            0.43778    0.19888   2.201  0.0277
W            0.02765    0.12574   0.220  0.8260
```

AIC: 64.11

```

#computing AICC
p<-3
n<-48
print(AICC<- -2*logLik(fitted.model)+2*p*n/(n-p-1))

64.65568

#outputting BIC
BIC(fitted.model)

69.72383

#checking model fit
null.model<- glm(group ~ 1, data=cardiac.data, family=binomial(link=logit))
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))

8.4319

print(p.value<- pchisq(deviance,2,lower.tail = FALSE))

0.0147583

```

(b) Interpret the estimated significant regression coefficients. For which fingerprint pattern the fitted probability is the largest? For which, the lowest?

If the number of arches increased by one, the estimated odds in favor of a cardiac disease would increase by  $(\exp(0.4378) - 1) \cdot 100\% = 54.93\%$ . A person with all ten arches has the highest estimated probability  $\hat{P}(\text{cardiac}) = \frac{\exp(-1.1160+0.4378 \cdot 10)}{1+\exp(-1.1160+0.4378 \cdot 10)} = 0.9631$ . A person with all ten loops has the lowest estimated probability  $\hat{P}(\text{cardiac}) = \frac{\exp(-1.1160)}{1+\exp(-1.1160)} = 0.2468$ .

(c) Suppose the model is used to predict the probability of being a cardiac patient in a male with the dermatoglyphics reading L-L-W-W-A-W-A-L-LW. What is this predicted probability?

The predicted probability for this person is  $P^0(\text{cardiac}) = \frac{\exp(-1.1160+0.4378 \cdot 2+0.0277 \cdot 4)}{1+\exp(-1.1160+0.4378 \cdot 2+0.0277 \cdot 4)} = 0.4676$ .

In SAS:

```

/*using fitted model for prediction*/
data predict;
input A W;
cards;
2 4
;
run;

data cardiac;
set cardiac predict;
run;

proc genmod;
  model group(event="1") = A W / dist=binomial link=logit;
  output out=outdata p=pcardiac;
run;

```

```
proc print data=outdata (firstobs=49) noobs;
  var pcardiac;
run;
```

```
pcardiac
0.46758
```

In R:

```
#using fitted model for prediction
print(predict(fitted.model, data.frame(A=2, W=4), type="response"))
```

```
0.4675796
```

(h) In parts (a)-(c), fit a probit model. Compare results.

In SAS:

```
/*fitting probit model*/
data cardiac;
input group A W @@;
cards;
1 8 2 1 1 2 1 2 1 1 4 0 1 2 7 1 6 3 1 2 8 1 1 9
1 3 0 1 0 2 1 3 2 1 2 7 1 2 7 1 2 8 1 6 0 1 3 5
1 1 0 1 7 1 1 4 3 1 2 4 1 5 3 1 7 1 1 8 1 1 0 6
0 0 9 0 2 1 0 0 8 0 1 3 0 3 1 0 1 4 0 0 8 0 1 6
0 0 9 0 2 2 0 4 4 0 0 6 0 3 2 0 1 2 0 2 5 0 4 0
0 8 1 0 2 7 0 0 10 0 0 5 0 0 6 0 2 1 0 0 7 0 0 6
;

proc genmod;
  model group(event="1") = A W / dist=binomial link=probit;
run;
```

```
Log Likelihood -29.1509
```

```
AIC 64.3018
```

```
AICC 64.8473
```

```
BIC 69.9154
```

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.6300	0.5189	-1.6471	0.3871	1.47	0.2248
A	1	0.2490	0.1061	0.0409	0.4570	5.50	0.0190
W	1	0.0105	0.0760	-0.1385	0.1594	0.02	0.8904

The fitted probit model looks like  $\Phi^{-1}(\hat{P}(\text{cardiac})) = -0.6300 + 0.2490 \cdot \# \text{ of arches} + 0.0105 \cdot \# \text{ of whorls}$ . Only the number of arches is a significant predictor, which is similar to what the logistic model gives.

```
/*checking model fit*/
proc genmod;
  model group = / dist=binomial link=probit;
run;
```

Log Likelihood -33.2711

```
data deviance_test;  
  deviance = -2*(-33.2711 - (-29.1509));  
  pvalue = 1 - probchi(deviance,2);  
run;
```

```
proc print noobs;  
run;
```

deviance	pvalue
8.2404	0.016241

The probit model fits the data reasonably well, at the 5% level. If the number of arches increased by one, the z-score of the estimated probability of a cardiac disease would increase by 0.2490.

The predicted probability in this model is  $P^0(\text{cardiac}) = \Phi(-0.6300 + 0.2490 \cdot 2 + 0.0105 \cdot 4) = \Phi(-0.09) = 0.4641$ . This prediction is a tiny bit smaller than the one produced by the logistic model.

In SAS:

```
/*using fitted model for prediction*/  
data predict;  
input A W;  
cards;  
2 4  
;  
run;  
  
data cardiac;  
set cardiac predict;  
run;  
  
proc genmod;  
  model group(event="1") = A W / dist=binomial link=probit;  
  output out=outdata p=pcardiac;  
run;  
  
proc print data=outdata (firstobs=49) noobs;  
  var pcardiac;  
run;  
pcardiac  
0.46408
```

In R:

```
#fitting probit model  
cardiac.data<- read.csv(file="C:/./Exercise3.5Data.csv", header=TRUE, sep=",")  
  
#running the model  
summary(fitted.model<- glm(group ~ A + W, data=cardiac.data,  
family=binomial(link=probit)))
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-0.62995	0.53497	-1.178	0.2390
A	0.24895	0.11125	2.238	0.0252
W	0.01047	0.07694	0.136	0.8918

AIC: 64.302

```

#computing AICC
p<-3
n<-48
print(AICC<- -2*logLik(fitted.model)+2*p*n/(n-p-1))

64.84727

#outputting BIC
BIC(fitted.model)

69.91541

#checking model fit
null.model<- glm(group ~ 1, data=cardiac.data, family=binomial(link=probit))
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))

8.240317

print(p.value<- pchisq(deviance,2,lower.tail = FALSE))

0.01624194

#using fitted model for prediction
print(predict(fitted.model, data.frame(A=2, W=4), type="response"))

0.4640763

```

(e) Fit the complementary log-log model instead of the logistic model in (a) through (c). Do the models differ? Which of the three models should be preferred?

In SAS:

```

/*fitting complementary log-log model*/
data cardiac;
input group A W @@;
cards;
1 8 2 1 1 2 1 2 1 1 4 0 1 2 7 1 6 3 1 2 8 1 1 9
1 3 0 1 0 2 1 3 2 1 2 7 1 2 7 1 2 8 1 6 0 1 3 5
1 1 0 1 7 1 1 4 3 1 2 4 1 5 3 1 7 1 1 8 1 1 0 6
0 0 9 0 2 1 0 0 8 0 1 3 0 3 1 0 1 4 0 0 8 0 1 6
0 0 9 0 2 2 0 4 4 0 0 6 0 3 2 0 1 2 0 2 5 0 4 0
0 8 1 0 2 7 0 0 10 0 0 5 0 0 6 0 2 1 0 0 7 0 0 6
;

proc genmod;
  model group(event="1") = A W / dist=binomial link=cloglog;
run;

```

Log Likelihood -29.5236

AIC 65.0471  
AICC 65.5926  
BIC 70.6607

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	wald	95% Confidence Limits	wald Chi-Square	Pr > ChiSq
Intercept	1	-1.0392	0.6424	-2.2984	0.2199	2.62	0.1057

### Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald	95% Confidence Limits	Wald Chi-Square	Pr > ChiSq
A	1	0.2382	0.1039	0.0345	0.4419	5.25	0.0219
W	1	0.0167	0.0971	-0.1735	0.2070	0.03	0.8631

We write the fitted model  $1 - \hat{P}(\text{cardiac}) = \exp(-\exp(-1.0392 + 0.2382 \cdot \# \text{ of arches} + 0.0167 \cdot \# \text{ of whorls}))$ . The number of arches is the only significant predictor, as concurs with the previous two models.

```
/*checking model fit*/
proc genmod;
  model group = / dist=binomial link=cloglog;
run;
```

**Log Likelihood -33.2711**

```
data deviance_test;
  deviance = -2*(-33.2711 - (-29.5236));
  pvalue = 1 - probchi(deviance,2);
run;
```

```
proc print noobs;
run;
```

```
deviance    pvalue
      7.495  0.023577
```

The model has a good fit at the 5% level of significance. If the number of arches increased by one, the estimated probability of no cardiac disease would be raised to the power  $\exp(0.2382) = 1.27$ , that is, the probability of no cardiac disease would decrease, which is in agreement with the previous two models. The predicted probability is found as  $P^0(\text{cardiac}) = 1 - \exp(-\exp(-1.0392 + 0.2382 \cdot 2 + 0.0167 \cdot 4)) = 0.4561$ .

In SAS:

```
/*using fitted model for prediction*/
data predict;
input A W;
cards;
2 4
;
run;

data cardiac;
set cardiac predict;
run;

proc genmod;
  model group(event="1") = A W / dist=binomial link=cloglog;
  output out=outdata p=pcardiac;
run;

proc print data=outdata (firstobs=49) noobs;
  var pcardiac;
run;
```

pcardiac  
0.45612

In R:

```
#fitting complementary log-log model
cardiac.data<- read.csv(file="C:/./Exercise3.5Data.csv", header=TRUE, sep=",")

#running the model
summary(fitted.model<- glm(group ~ A + W, data=cardiac.data,
family=binomial(link=cloglog)))
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-1.03922	0.62874	-1.653	0.0984
A	0.23818	0.10916	2.182	0.0291
W	0.01674	0.09199	0.182	0.8556

AIC: 65.047

```
#computing AICC
p<-3
n<-48
print(AICC<- -2*logLik(fitted.model)+2*p*n/(n-p-1))
```

65.59256

```
#outputting BIC
BIC(fitted.model)
```

70.66071

```
#checking model fit
null.model<- glm(group ~ 1, data=cardiac.data, family=binomial(link=cloglog))
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))
```

7.495025

```
print(p.value<- pchisq(deviance,2,lower.tail = FALSE))
```

0.02357632

```
#using fitted model for prediction
print(predict(fitted.model, data.frame(A=2, W=4), type="response"))
```

0.45612

Using the AIC, AICC, and BIC criteria, we see that the logistic regression has the smallest values and thus should be preferred.

	logistic	probit	cloglog
AIC	64.1102	64.3018	65.0471
AICC	64.6557	64.8473	65.5926
BIC	69.7238	69.9154	70.6607

## CHAPTER 4

**EXERCISE 4.1.** (a) Run the cumulative logit model and specify the fitted model. Discuss the model fit. What predictors are significant at the 5% level? Interpret the estimated significant regression coefficients. Predict the probabilities of each admission status for a person whose GPA is 3.1 and GMAT score is 550.

In SAS:

```
/*fitting cumulative logit model*/
data admission;
input GPA GMAT status$ @@;
cards;
2.96 596 admit      3.14 473 admit      3.22 482 admit      3.29 527 admit
3.69 505 admit      2.46 693 admit      3.03 626 admit      3.19 663 admit
3.63 447 admit      3.59 588 admit      3.30 563 admit      3.78 591 admit
3.44 692 admit      3.48 528 admit      3.47 552 admit      3.35 520 admit
2.89 543 admit      2.28 523 admit      3.21 530 admit      3.58 564 admit
3.33 565 admit      2.80 444 border      3.13 416 border      2.89 431 border
3.01 471 border      2.91 446 border      2.75 546 border      2.73 467 border
3.12 463 border      3.08 440 notadmit      3.01 453 notadmit      3.03 414 notadmit
3.04 446 notadmit      2.89 485 notadmit      2.79 490 notadmit      2.54 446 notadmit
2.43 425 notadmit      2.20 474 notadmit      3.36 531 notadmit      2.57 542 notadmit
2.36 482 notadmit      3.66 420 notadmit
;
```

```
proc genmod;
  model status = GPA GMAT / dist=multinomial link=cumlogit;
run;
Log Likelihood -27.5443
```

```
AIC  63.0887
AICC 64.1698
BIC  70.0394
```

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept1	1	-23.3903	5.8569	-34.8697	-11.9109	15.95	<.0001
Intercept2	1	-21.8526	5.6696	-32.9647	-10.7404	14.86	0.0001
GPA	1	3.1194	1.1913	0.7845	5.4543	6.86	0.0088
GMAT	1	0.0278	0.0084	0.0113	0.0442	10.90	0.0010

The fitted model is  $\frac{\hat{P}(\text{admit})}{1-\hat{P}(\text{admit})} = \exp(-23.3903 + 3.1194 \cdot \text{GPA} + 0.0278 \cdot \text{GMAT})$ , and

$\frac{\hat{P}(\text{admit or borderline})}{\hat{P}(\text{not admit})} = \exp(-21.8526 + 3.1194 \cdot \text{GPA} + 0.0278 \cdot \text{GMAT})$ . Both GPA and GMAT scores are significant predictors. As GPA increases by one point, the estimated odds in favor of more admission increase by  $(\exp(3.1194) - 1) \cdot 100\% = 2,163.28\%$ . As the GMAT score increases by one point, the estimated odds in favor of more admission increase by  $(\exp(0.0278) - 1) \cdot 100\% = 2.819\%$ .

```
/*checking model fit*/
proc genmod;
  model status = / dist=multinomial link=cumlogit;
```

```
run;
```

Log Likelihood -43.0673

```
data deviance_test;
  deviance = -2*(-43.0673 - (-27.5443));
  pvalue = 1 - probchi(deviance,2);
run;
```

```
proc print noobs;
run;
```

deviance	pvalue
31.046	.000000181

Due to the small p-value of the deviance test, the model has a good fit.

To predict probabilities of admit, borderline, and not admit for a person whose GPA is 3.1 and GMAT score is 550 we do the following calculations:

$$P^0(\text{admit}) = \frac{\exp(-23.3903 + (3.1194)(3.1) + (0.0278)(550))}{1 + \exp(-23.3903 + (3.1194)(3.1) + (0.0278)(550))} = 0.827761,$$

$$P^0(\text{admit or borderline}) = \frac{\exp(-21.8526 + (3.1194)(3.1) + (0.0278)(550))}{1 + \exp(-21.8526 + (3.1194)(3.1) + (0.0278)(550))} \\ = 0.957203,$$

from where  $P^0(\text{not admit}) = 1 - 0.957203 = 0.042797$ , and  $P^0(\text{borderline}) = 0.957203 - 0.827761 = 0.129442$ .

In SAS:

```
/*using fitted model for prediction*/
data predict;
input GPA GMAT;
cards;
3.1 550
;

data admission;
set admission predict;
run;

proc genmod;
  model status = GPA GMAT / dist=multinomial link=cumlogit;
  output out=outdata p=pstatus;
run;

proc print data=outdata (firstobs=85) noobs;
  var _level_ pstatus;
run;

_LEVEL_ pstatus
admit 0.82419
border 0.95618
```

In R:

```
#fitting cumulative logit model
admission.data<- read.csv(file="C:/./Exercise4.1Data.csv", header= TRUE, sep=",")

#rescaling predictor
GMAT.res<- admission.data$GMAT/100

#running the model
library(ordinal)
summary(fitted.model<- clm(status ~ GPA + GMAT.res, data=admission.data,
link="logit"))
```

AIC  
63.09

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
GPA	-3.1194	1.1913	-2.618	0.008833
GMAT.res	-2.7755	0.8406	-3.302	0.000961

Threshold coefficients:

	Estimate	Std. Error	z value
admit border	-23.390	5.857	-3.994
border notadmit	-21.853	5.670	-3.854

```
#computing AICC
p<- 4
n<- 42
print(AICC<- -2*logLik(fitted.model)+2*p*n/(n-p-1))
```

64.16976

```
#outputting BIC
BIC(fitted.model)
```

70.03935

```
#checking model fit
null.model<- clm(status ~ 1, data=admission.data, link="logit")
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))
```

31.04588

```
print(p.value<- pchisq(deviance, df=2, lower.tail = FALSE))
```

1.813312e-07

```
#using fitted model for prediction
print(predict(fitted.model, data.frame(GPA=3.1, GMAT.res=5.50), type="prob"))
```

admit	border	notadmit
0.8241952	0.1319818	0.04382299

(b) Redo part (a), fitting the cumulative probit model.

In SAS:

```
/*fitting cumulative probit model*/
```

```

data admission;
input GPA GMAT status$ @@;
cards;
2.96 596 admit      3.14 473 admit      3.22 482 admit      3.29 527 admit
3.69 505 admit      2.46 693 admit      3.03 626 admit      3.19 663 admit
3.63 447 admit      3.59 588 admit      3.30 563 admit      3.78 591 admit
3.44 692 admit      3.48 528 admit      3.47 552 admit      3.35 520 admit
2.89 543 admit      2.28 523 admit      3.21 530 admit      3.58 564 admit
3.33 565 admit      2.80 444 border      3.13 416 border      2.89 431 border
3.01 471 border      2.91 446 border      2.75 546 border      2.73 467 border
3.12 463 border      3.08 440 notadmit      3.01 453 notadmit      3.03 414 notadmit
3.04 446 notadmit      2.89 485 notadmit      2.79 490 notadmit      2.54 446 notadmit
2.43 425 notadmit      2.20 474 notadmit      3.36 531 notadmit      2.57 542 notadmit
2.36 482 notadmit      3.66 420 notadmit
;

```

```

proc genmod;
  model status = GPA GMAT / dist=multinomial link=cumprobit;
run;

```

Log Likelihood -27.5930

AIC 63.1860

AICC 64.2671

BIC 70.1367

Analysis of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept1	1	-13.6033	3.0779	-19.6359	-7.5707	19.53	<.0001
Intercept2	1	-12.7243	3.0012	-18.6065	-6.8421	17.98	<.0001
GPA	1	1.7356	0.6222	0.5162	2.9550	7.78	0.0053
GMAT	1	0.0165	0.0046	0.0076	0.0255	13.09	0.0003

The fitted model is of the form:  $\hat{P}(\text{admit}) = \Phi(-13.6033 + 1.7356 \cdot \text{GPA} + 0.0165 \cdot \text{GMAT})$ , and  $\hat{P}(\text{admit or borderline}) = \Phi(-12.7243 + 1.7356 \cdot \text{GPA} + 0.0165 \cdot \text{GMAT})$ .

Both GPA and GMAT are significant predictors. As GPA increases by one point, the z-scores of the estimated probabilities increase by 1.7356. For a unit increase in GMAT, the z-score increases by 0.0165.

```

/*checking model fit*/
proc genmod;
  model status = / dist=multinomial link=cumprobit;
run;

```

Log Likelihood -43.0673

```

data deviance_test;
  deviance = -2*(-43.0673 - (-27.5930));
  pvalue = 1 - probchi(deviance,2);
run;

```

```

proc print noobs;
run;

```

```

deviance      pvalue
30.9486      .000000190

```

The model has a very good fit which is evidenced by a tiny p-value in the deviance test. The predicted probabilities are found in the following manner:  $P^0(\text{admit}) = \Phi(-13.6033 + 1.7356 \cdot 3.1 + 0.0165 \cdot 550) = 0.80291$ , and  $P^0(\text{admit or borderline}) = \Phi(-12.7243 + 1.7356 \cdot 3.1 + 0.0165 \cdot 550) = 0.958279$ . Thus,  $P^0(\text{not admit}) = 1 - 0.958279 = 0.041721$ , and  $P^0(\text{borderline}) = 0.958279 - 0.80291 = 0.15537$ .

In SAS:

```
/*using fitted model for prediction*/
data predict;
input GPA GMAT;
cards;
3.1 550
;

data admission;
set admission predict;
run;

proc genmod;
  model status = GPA GMAT / dist=multinomial link=cumprobit;
  output out=outdata p=pstatus;
run;

proc print data=outdata (firstobs=85) noobs;
  var _level_ pstatus;
run;

_LEVEL_ pstatus
admit    0.80985
border   0.96048
```

In R:

```
#fitting cumulative probit model
admission.data<- read.csv(file="C:/./Exercise4.1Data.csv", header= TRUE, sep=",")

#rescaling predictor
GMAT.res<- admission.data$GMAT/100

#running the model
library(ordinal)
summary(fitted.model<- clm(status ~ GPA + GMAT.res, data=admission.data,
link="probit"))
```

AIC  
63.19

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
GPA	-1.7356	0.6222	-2.790	0.005276
GMAT.res	-1.6546	0.4573	-3.618	0.000297

Threshold coefficients:

	Estimate	Std. Error	z value
admit border	-13.603	3.078	-4.42

border|notadmit   -12.724        3.001       -4.24

```
#computing AICC
p<- 4
n<- 42
print(AICC<- -2*logLik(fitted.model)+2*p*n/(n-p-1))
```

64.26707

```
#outputting BIC
BIC(fitted.model)
```

70.13666

```
#checking model fit
null.model<- clm(status ~ 1, data=admission.data, link="probit")
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))
```

30.94857

```
print(p.value<- pchisq(deviance, df=2, lower.tail = FALSE))
```

1.903718e-07

```
#using fitted model for prediction
print(predict(fitted.model, data.frame(GPA=3.1, GMAT.res=5.50), type="prob"))
```

	admit	border	notadmit
	0.8098528	0.1506302	0.03951702

(c) Redo part (a), fitting the cumulative complementary log-log model.

In SAS:

```
/*fitting cumulative complementary log-log model*/
data admission;
input GPA GMAT status$ @@;
cards;
2.96 596 admit        3.14 473 admit        3.22 482 admit        3.29 527 admit
3.69 505 admit        2.46 693 admit        3.03 626 admit        3.19 663 admit
3.63 447 admit        3.59 588 admit        3.30 563 admit        3.78 591 admit
3.44 692 admit        3.48 528 admit        3.47 552 admit        3.35 520 admit
2.89 543 admit        2.28 523 admit        3.21 530 admit        3.58 564 admit
3.33 565 admit        2.80 444 border        3.13 416 border        2.89 431 border
3.01 471 border        2.91 446 border        2.75 546 border        2.73 467 border
3.12 463 border        3.08 440 notadmit        3.01 453 notadmit        3.03 414 notadmit
3.04 446 notadmit        2.89 485 notadmit        2.79 490 notadmit        2.54 446 notadmit
2.43 425 notadmit        2.20 474 notadmit        3.36 531 notadmit        2.57 542 notadmit
2.36 482 notadmit        3.66 420 notadmit
;

proc genmod;
  model status = GPA GMAT / dist=multinomial link=cumcll;
run;
```

Log Likelihood -29.4951

AIC        66.9901

AICC       68.0712

BIC        73.9408

### Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald	95% Confidence Limits	Chi-Square	Pr > ChiSq
Intercept1	1	-13.1371	3.0594	-19.1335	-7.1407	18.44	<.0001
Intercept2	1	-12.2653	2.9885	-18.1228	-6.4079	16.84	<.0001
GPA	1	1.6479	0.6316	0.4099	2.8858	6.81	0.0091
GMAT	1	0.0152	0.0043	0.0068	0.0236	12.46	0.0004

The fitted model is  $\hat{P}(admit) = 1 - \exp(-\exp(-13.1371 + 1.6479 \cdot GPA + 0.0152 \cdot GMAT))$  and  $\hat{P}(admit \text{ or } borderline) = 1 - \exp(-\exp(-12.2653 + 1.6479 \cdot GPA + 0.0152 \cdot GMAT))$ . GPA and GMAT are both significant predictors. As GPA increases by one point, the estimated complementary probabilities are raised to the power  $\exp(1.6479) = 5.196$ . As the GMAT score increases by one point, the estimated complementary probabilities are raised to the power  $\exp(0.0152) = 1.015$ .

```
/*checking model fit*/
proc genmod;
  model status = / dist=multinomial link=cumcll;
run;
```

**Log Likelihood -43.0673**

```
data deviance_test;
  deviance = -2*(-43.0673 - (-29.4951));
  pvalue = 1 - probchi(deviance,2);
run;
```

```
proc print noobs;
run;
```

```
deviance      pvalue
 27.1444      .000001275
```

This model has a very good fit because of the small p-value in the deviance test. To calculate the predicted probabilities, we write  $P^0(admit) = 1 - \exp(-\exp(-13.1371 + 1.6479 \cdot 3.1 + 0.0152 \cdot 550)) = 0.751647$ , and  $P^0(admit \text{ or } borderline) = 1 - \exp(-\exp(-12.2653 + 1.6479 \cdot 3.1 + 0.0152 \cdot 550)) = 0.964233$ . Hence,  $P^0(not \text{ admit}) = 1 - 0.964233 = 0.035767$ , and  $P^0(borderline) = 0.964233 - 0.751647 = 0.212586$ .

**In SAS:**

```
/*using fitted model for prediction*/
data predict;
input GPA GMAT;
cards;
3.1 550
;

data admission;
set admission predict;
run;

proc genmod;
  model status = GPA GMAT / dist=multinomial link=cumcll;
```

```

    output out=outdata p=pstatus;
run;

proc print data=outdata (firstobs=85) noobs;
  var _level_ pstatus;
run;

```

```

_LEVEL_ pstatus
admit    0.74976
border   0.96358

```

In R:

```

#fitting cumulative complementary log-log model
admission.data<- read.csv(file="C:/./Exercise4.1Data.csv", header= TRUE, sep=",")

#rescaling predictor
GMAT.res<- admission.data$GMAT/100

#running the model
library(ordinal)
summary(fitted.model<- clm(status ~ GPA + GMAT.res, data=admission.data,
link="cloglog"))

AIC
66.99
Coefficients:
      Estimate Std. Error z value Pr(>|z|)
GPA      -1.6479    0.6316  -2.609 0.009081
GMAT.res  -1.5190    0.4303  -3.530 0.000415

Threshold coefficients:
      Estimate Std. Error z value
admit|border   -13.137     3.059  -4.294
border|notadmit -12.265     2.989  -4.104

#computing AICC
p<- 4
n<- 42
print(AICC<- -2*logLik(fitted.model)+2*p*n/(n-p-1))

68.07119

#outputting BIC
BIC(fitted.model)

73.94079

#checking model fit
null.model<- clm(status ~ 1, data=admission.data, link="cloglog")
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))

27.14445

print(p.value<- pchisq(deviance, df=2, lower.tail = FALSE))

1.275435e-06

#using fitted model for prediction
print(predict(fitted.model, data.frame(GPA=3.1, GMAT.res=5.50), type="prob"))

```

admit border notadmit  
0.7497578 0.2138205 0.03642168

(d) Which of the models obtained in parts (a)-(c) has the best fit?

By the AIC, AICC, and BIC criteria, we see that the cumulative logit regression has the smallest values and thus has the best fit.

	cumulative logit	cumulative probit	Cumulative cloglog
AIC	63.0887	63.1860	66.9901
AICC	64.1698	64.2671	68.0712
BIC	70.0394	70.1367	73.9408

**EXERCISE 4.2.** (a) Regress the satisfaction score on the other variables via the cumulative logit model. How good is the model fit? Which regression coefficients are significant at  $\alpha = 0.05$ ? State the fitted model explicitly and interpret the estimated significant beta coefficients. Predict probabilities of each of the five levels of the satisfaction score for a caller who had been subscribed for 3 months, doesn't receive the magazine, and whose issue was resolved over the phone.

In SAS:

```
/*fitting cumulative logit model*/
data service;
input subscribed magazine$ resolved$ satisf @@;
cards;
5      yes   no    5  49 yes   no    5  56 no    no    3
13     yes   yes   5  27 no    yes   4  41 yes   yes   5
2      yes   yes   5  64 yes   yes   4  88 yes   yes   4
43     yes   yes   4  94 yes   no    4  8  no    no    1
9      yes   no    2  68 yes   no    4  5  no    yes   2
108    no    yes   3  21 yes   yes   4  25 yes   no    3
2      no    yes   4  11 no    no    2  98 yes   yes   5
11     no    yes   5  46 no    no    4  7  no    no    3
7      no    yes   5  9  yes   yes   5  17 no    no    2
8      no    yes   2  9  no    yes   1  95 no    no    4
60     no    yes   3  80 no    yes   4  2  yes   no    3
33     yes   yes   4  5  yes   no    3  7  no    no    1
;

proc genmod;
  class magazine(ref="yes") resolved(ref="yes");
  model satisf = subscribed magazine resolved / dist=multinomial link=cumlogit;
run;
```

Log Likelihood -47.0487

AIC 108.0974

AICC 112.0974  
BIC 119.1820

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald	95% Confidence Limits	Chi-Square	Pr > ChiSq
Intercept1	1	-4.2359	0.9678	-6.1328	-2.3390	19.16	<.0001
Intercept2	1	-2.8298	0.8126	-4.4224	-1.2373	12.13	0.0005
Intercept3	1	-1.5740	0.7442	-3.0327	-0.1153	4.47	0.0344
Intercept4	1	0.3350	0.6708	-0.9797	1.6498	0.25	0.6175
subscribed	1	-0.0105	0.0097	-0.0295	0.0085	1.18	0.2776
magazine no	1	1.9175	0.6771	0.5903	3.2447	8.02	0.0046
magazine yes	0	0.0000	0.0000	0.0000	0.0000	.	.
resolved no	1	1.4288	0.6559	0.1434	2.7143	4.75	0.0294
resolved yes	0	0.0000	0.0000	0.0000	0.0000	.	.

The fitted model is of the form  $\frac{\hat{P}(\text{satisf}=1)}{1-\hat{P}(\text{satisf}=1)} = \frac{\hat{P}(\text{very dissatisfied})}{1-\hat{P}(\text{very dissatisfied})} = \exp(-4.2359 - 0.0105 \cdot$

$\# \text{ of months subscribed} + 1.9175 \cdot \text{no magazine} + 1.4288 \cdot \text{issue not resolved}),$

$\frac{\hat{P}(\text{satisf}=1 \text{ or } 2)}{1-\hat{P}(\text{satisf}=1 \text{ or } 2)} = \frac{\hat{P}(\text{very dissatisfied or dissatisfied})}{1-\hat{P}(\text{very dissatisfied or dissatisfied})} = \exp(-2.8298 - 0.0105 \cdot$

$\# \text{ of months subscribed} + 1.9175 \cdot \text{no magazine} + 1.4288 \cdot \text{issue not resolved}),$

$\frac{\hat{P}(\text{satisf}=1,2,\text{or } 3)}{1-\hat{P}(\text{satisf}=1,2,\text{or } 3)} = \frac{\hat{P}(\text{very dissatisfied, dissatisfied, or neutral})}{\hat{P}(\text{satisfied or very satisfied})} = \exp(-1.5740 - 0.0105 \cdot$

$\# \text{ of months subscribed} + 1.9175 \cdot \text{no magazine} + 1.4288 \cdot \text{issue not resolved}),$

and  $\frac{\hat{P}(\text{satisf}=1,2,3,\text{or } 4)}{1-\hat{P}(\text{satisf}=1,2,3,\text{or } 4)} = \frac{\hat{P}(\text{very dissatisfied, dissatisfied, neutral, or satisfied})}{\hat{P}(\text{very satisfied})} = \exp(0.3350 -$   
 $0.0105 \cdot \# \text{ of months subscribed} + 1.9175 \cdot \text{no magazine} + 1.4288 \cdot \text{issue not resolved}).$

Magazine subscription and whether the issue was resolved are statistically significant at the 5% level. The number of months subscribed is not a significant predictor.

For the customers not receiving the magazine, the estimated odds are  $\exp(1.9175) \cdot 100\% = 680.39\%$  of those for a customer who receives the magazine. If the issue was not resolved, the estimated odds are  $\exp(1.4288) \cdot 100\% = 417.37\%$  of those when the issue was resolved.

```
/*checking model fit*/
proc genmod;
  model satisf = / dist=multinomial link=cumlogit;
run;
```

Log Likelihood -54.4484

```
data deviance_test;
  deviance = -2*(-54.4484 - (-47.0487));
  pvalue = 1 - probchi(deviance,3);
run;
```

```
proc print noobs;
run;
```

deviance	pvalue
14.7994	.001996353

The model has a very good fit as indicated by a small p-value. The predicted probabilities are obtained as follows.  $P^0(\text{satisf} = 1) = P^0(\text{very dissatisfied}) = \frac{\exp(-4.2359 - 0.0105 \cdot 3 + 1.9175)}{1 + \exp(-4.2359 - 0.0105 \cdot 3 + 1.9175)} = 0.087074$ ,  $P^0(\text{satisf} = 1 \text{ or } 2) = P^0(\text{very dissatisfied or dissatisfied}) = \frac{\exp(-2.8298 - 0.0105 \cdot 3 + 1.9175)}{1 + \exp(-2.8298 - 0.0105 \cdot 3 + 1.9175)} = 0.280133$ ,  $P^0(\text{satisf} = 1, 2, \text{ or } 3) = P^0(\text{very dissatisfied, dissatisfied, or neutral}) = \frac{\exp(-1.5740 - 0.0105 \cdot 3 + 1.9175)}{1 + \exp(-1.5740 - 0.0105 \cdot 3 + 1.9175)} = 0.577373$ , and  $P^0(\text{satisf} = 1, 2, 3, \text{ or } 4) = P^0(\text{very dissatisfied, dissatisfied, neutral, or satisfied}) = \frac{\exp(0.3350 - 0.0105 \cdot 3 + 1.9175)}{1 + \exp(0.3350 - 0.0105 \cdot 3 + 1.9175)} = 0.90212$ . From here, the predicted probabilities of each of the five levels of the satisfaction score are  $P^0(\text{very dissatisfied}) = 0.087074$ ,  $P^0(\text{dissatisfied}) = 0.280133 - 0.087074 = 0.19306$ ,  $P^0(\text{neutral}) = 0.577373 - 0.280133 = 0.29724$ ,  $P^0(\text{satisfied}) = 0.90212 - 0.577373 = 0.324746$ , and  $P^0(\text{very satisfied}) = 1 - 0.90212 = 0.09788$ .

In SAS:

```
/*using fitted model for prediction*/
data predict;
input subscribed magazine$ resolved$;
cards;
3 no yes
;

data service;
set service predict;
run;

proc genmod;
class magazine resolved;
model satisf = subscribed magazine resolved / dist=multinomial link=cumlogit;
output out=outdata p=psatisf;
run;

proc print data=outdata (firstobs=145) noobs;
var _level_ psatisf;
run;
```

_LEVEL_	psatisf
1	0.08707
2	0.28012
3	0.57736
4	0.90212

In R:

```
#fitting cumulative logit model
service.data<- read.csv(file="C:/./Exercise4.2Data.csv", header= TRUE, sep=",")

#making response a categorical variable
satisf.cat<- as.factor(service.data$satisf)
```

```
#specifying reference categories
magazine.rel<- relevel(service.data$magazine, ref="yes")
resolved.rel<- relevel(service.data$resolved, ref="yes")

#running the model
library(ordinal)
summary(fitted.model<- clm(satisf.cat ~ subscribed + magazine.rel
+ resolved.rel, data=service.data, link="logit"))

AIC
108.10

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
subscribed      0.010516   0.009686    1.086  0.27760
magazine.relno -1.917509   0.677141   -2.832  0.00463
resolved.relno -1.428832   0.655864   -2.179  0.02937

Threshold coefficients:
      Estimate Std. Error z value
1|2  -4.2359    0.9678   -4.377
2|3  -2.8298    0.8126   -3.483
3|4  -1.5740    0.7442   -2.115
4|5   0.3350    0.6708    0.499

#computing AICC
p<-7
n<-36
print(AICC<- -2*logLik(fitted.model)+2*p*n/(n-p-1))

112.0974

#outputting BIC
BIC(fitted.model)

119.182

#checking model fit
null.model<- clm(satisf.cat ~ 1, data=service.data, link="logit")
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))

14.7994

print(p.value<- pchisq(deviance, df=3, lower.tail = FALSE))

0.001996357

#using fitted model for prediction
print(predict(fitted.model, data.frame(subscribed=3, magazine.rel="no",
resolved.rel="yes"), type="prob"))

      1      2      3      4      5
0.08706729 0.1930496 0.2972414 0.3247592 0.09788253
```

(b) Redo part (a), running the cumulative probit model.

In SAS:

```
/*fitting cumulative probit model*/
```

```

data service;
input subscribed magazine$ resolved$ satisf @@;
cards;
5      yes    no    5  49 yes    no    5  56 no    no    3
13     yes    yes   5  27 no     yes   4  41 yes   yes   5
2      yes    yes   5  64 yes   yes   4  88 yes   yes   4
43     yes    yes   4  94 yes   no    4  8  no    no    1
9      yes    no    2  68 yes   no    4  5  no    yes   2
108    no     yes   3  21 yes   yes   4  25 yes  no    3
2      no     yes   4  11 no    no    2  98 yes  yes   5
11     no     yes   5  46 no    no    4  7  no    no    3
7      no     yes   5  9  yes   yes   5  17 no   no    2
8      no     yes   2  9  no    yes   1  95 no   no    4
60     no     yes   3  80 no    yes   4  2  yes  no    3
33     yes    yes   4  5  yes   no    3  7  no    no    1
;

proc genmod;
  class magazine(ref="yes") resolved(ref="yes");
  model satisf = subscribed magazine resolved / dist=multinomial link=cumprobit;
run;

```

Log Likelihood -46.6927

AIC 107.3854

AICC 111.3854

BIC 118.4700

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept1	1	-2.6005	0.5384	-3.6558	-1.5453	23.33	<.0001
Intercept2	1	-1.7878	0.4709	-2.7108	-0.8648	14.41	0.0001
Intercept3	1	-1.0295	0.4411	-1.8940	-0.1650	5.45	0.0196
Intercept4	1	0.1114	0.3933	-0.6596	0.8823	0.08	0.7771
subscribed	1	-0.0061	0.0057	-0.0173	0.0051	1.12	0.2892
magazine no	1	1.2027	0.3885	0.4413	1.9641	9.58	0.0020
magazine yes	0	0.0000	0.0000	0.0000	0.0000	.	.
resolved no	1	0.8809	0.3783	0.1394	1.6224	5.42	0.0199
resolved yes	0	0.0000	0.0000	0.0000	0.0000	.	.

The fitted model is written as  $\hat{P}(\text{very dissatisfied}) = \Phi(-2.6005 - 0.0061 \cdot \# \text{ of months subscribed} + 1.2027 \cdot \text{no magazine} + 0.8809 \cdot \text{issue not resolved})$ ,  
 $\hat{P}(\text{very dissatisfied or dissatisfied}) = \Phi(-1.7878 - 0.0061 \cdot \# \text{ of months subscribed} + 1.2027 \cdot \text{no magazine} + 0.8809 \cdot \text{issue not resolved})$ ,  
 $\hat{P}(\text{very dissatisfied, dissatisfied, or neutral}) = \Phi(-1.0295 - 0.0061 \cdot \# \text{ of months subscribed} + 1.2027 \cdot \text{no magazine} + 0.8809 \cdot \text{issue not resolved})$ , and  
 $\hat{P}(\text{very dissatisfied, dissatisfied, neutral, or satisfied}) = \Phi(0.1114 - 0.0061 \cdot \# \text{ of months subscribed} + 1.2027 \cdot \text{no magazine} + 0.8809 \cdot \text{issue not resolved})$ .

Subscription to the magazine and whether the issue was resolved are statistically significant at the 5% level. The number of months subscribed is not a significant predictor.

For the customers not receiving the magazine, the z-score of the estimated probability of lower satisfaction is 1.2027 points larger than that for customers who receive the magazine. The z-score of the estimated probabilities of lower satisfaction for customers whose issue was not resolved is 0.8809 points larger than that for customers whose issue was resolved.

```
/*checking model fit*/
proc genmod;
  model satisf = / dist=multinomial link=cumprobit;
run;
```

**Log Likelihood -54.4484**

```
data deviance_test;
  deviance = -2*(-54.4484 - (-46.6927));
  pvalue = 1 - probchi(deviance,3);
run;
```

```
proc print noobs;
run;
```

deviance	pvalue
15.5114	.001427894

The p-value is small, indicating a good model fit. The prediction is carried out as follows:

$P^0(\text{very dissatisfied}) = \Phi(-2.6005 - 0.0061 \cdot 3 + 1.2027) = 0.078373$ ,  
 $P^0(\text{very dissatisfied or dissatisfied}) = \Phi(-1.7878 - 0.0061 \cdot 3 + 1.2027) = 0.273121$ ,  
 $P^0(\text{very dissatisfied, dissatisfied, or neutral}) = \Phi(-1.0295 - 0.0061 \cdot 3 + 1.2027) = 0.56155$ , and  $P^0(\text{very dissatisfied, dissatisfied, neutral, or satisfied}) = \Phi(0.1114 - 0.0061 \cdot 3 + 1.2027) = 0.902478$ . Therefore, for each level of the satisfaction score, the predicted probabilities are  $P^0(\text{very dissatisfied}) = 0.078373$ ,  $P^0(\text{dissatisfied}) = 0.273121 - 0.078373 = 0.194748$ ,  $P^0(\text{neutral}) = 0.56155 - 0.273121 = 0.288429$ ,  $P^0(\text{satisfied}) = 0.902478 - 0.56155 = 0.340928$ , and  $P^0(\text{very satisfied}) = 1 - 0.902478 = 0.097522$ .

```
/*using fitted model for prediction*/
data predict;
input subscribed magazine$ resolved$;
cards;
3 no yes
;

data service;
set service predict;
run;

proc genmod;
  class magazine resolved;
  model satisf = subscribed magazine resolved / dist=multinomial link=cumprobit;
  output out=outdata p=psatisf;
run;

proc print data=outdata (firstobs=145) noobs;
  var _level_ psatisf;
run;
```

_LEVEL_	psatisf
1	0.07838
2	0.27315
3	0.56159
4	0.90249

In R:

```
#fitting cumulative probit model
service.data<- read.csv(file="C:/./Exercise4.2Data.csv", header= TRUE, sep=",")

#making response a categorical variable
satisf.cat<- as.factor(service.data$satisf)

#specifying reference categories
magazine.rel<- relevel(service.data$magazine, ref="yes")
resolved.rel<- relevel(service.data$resolved, ref="yes")

#running the model
library(ordinal)
summary(fitted.model<- clm(satisf.cat ~ subscribed + magazine.rel + resolved.rel,
data=service.data, link="probit"))
```

AIC  
107.39

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
subscribed	0.006061	0.005718	1.060	0.28919
magazine.relno	-1.202678	0.388476	-3.096	0.00196
resolved.relno	-0.880893	0.378313	-2.328	0.01989

Threshold coefficients:

	Estimate	Std. Error	z value
1 2	-2.6005	0.5384	-4.830
2 3	-1.7878	0.4709	-3.796
3 4	-1.0295	0.4411	-2.334
4 5	0.1114	0.3933	0.283

```
#computing AICC
p<-7
n<-36
print(AICC<- -2*logLik(fitted.model)+2*p*n/(n-p-1))
```

111.3854

```
#outputting BIC
BIC(fitted.model)
```

118.47

```
#checking model fit
null.model<- clm(satisf.cat ~ 1, data=service.data, link="probit")
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))
```

15.51138

```
print(p.value<- pchisq(deviance, df=3, lower.tail = FALSE))
0.001427906
```

```
#using fitted model for prediction
```

```
print(predict(fitted.model, data.frame(subscribed=3, magazine.rel="no",
resolved.rel="yes"), type="prob"))
```

```

      1      2      3      4      5
0.0783845 0.1947673 0.2884392 0.3408959 0.09751309
```

(c) Redo part (a), running the cumulative complementary log-log model.

In SAS:

```
/*fitting cumulative complementary log-log model*/
data service;
input subscribed magazine$ resolved$ satisf @@;
cards;
5      yes      no      5  49 yes      no      5  56 no      no      3
13     yes      yes     5  27 no      yes     4  41 yes     yes     5
2      yes      yes     5  64 yes     yes     4  88 yes     yes     4
43     yes      yes     4  94 yes     no      4  8  no      no      1
9      yes      no      2  68 yes     no      4  5  no      yes     2
108    no      yes     3  21 yes     yes     4  25 yes     no      3
2      no      yes     4  11 no      no      2  98 yes     yes     5
11     no      yes     5  46 no      no      4  7  no      no      3
7      no      yes     5  9  yes     yes     5  17 no      no      2
8      no      yes     2  9  no      yes     1  95 no      no      4
60     no      yes     3  80 no      yes     4  2  yes     no      3
33     yes      yes     4  5  yes     no      3  7  no      no      1
;

proc genmod;
  class magazine(ref="yes") resolved(ref="yes");
  model satisf = subscribed magazine resolved / dist=multinomial link=cumcll;
run;
```

Log Likelihood -48.3651

AIC 110.7302

AICC 114.7302

BIC 121.8149

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald	95% Confidence Limits	Chi-Square	Pr > ChiSq
Intercept1	1	-3.6044	0.7514	-5.0770	-2.1317	23.01	<.0001
Intercept2	1	-2.4709	0.5899	-3.6271	-1.3147	17.54	<.0001
Intercept3	1	-1.5993	0.5275	-2.6331	-0.5655	9.19	0.0024
Intercept4	1	-0.4294	0.4201	-1.2527	0.3939	1.05	0.3066
subscribed	1	-0.0030	0.0063	-0.0154	0.0094	0.22	0.6364
magazine no	1	1.1422	0.4226	0.3140	1.9704	7.31	0.0069
magazine yes	0	0.0000	0.0000	0.0000	0.0000	.	.
resolved no	1	1.0099	0.4301	0.1670	1.8528	5.51	0.0189
resolved yes	0	0.0000	0.0000	0.0000	0.0000	.	.

The fitted model is  $\hat{P}(\text{satisf} = 1) = \hat{P}(\text{very dissatisfied}) = 1 - \exp(-\exp(-3.6044 - 0.0030 \cdot \text{\#of months subscribed} + 1.1422 \cdot \text{no magazine} + 1.0099 \cdot \text{issue not resolved}))$ ,

$\hat{P}(\text{very dissatisfied or dissatisfied}) = 1 - \exp(-\exp(-2.4709 - 0.0030 \cdot$   
 $\# \text{of months subscribed} + 1.1422 \cdot \text{no magazine} + 1.0099 \cdot \text{issue not resolved})),$   
 $\hat{P}(\text{very dissatisfied, dissatisfied, or neutral}) = 1 - \exp(-\exp(-1.5993 - 0.0030 \cdot$   
 $\# \text{of months subscribed} + 1.1422 \cdot \text{no magazine} + 1.0099 \cdot \text{issue not resolved})),$  and  
 $\hat{P}(\text{very dissatisfied, dissatisfied, neutral, or satisfied}) = 1 - \exp(-\exp(-0.4294 - 0.0030 \cdot$   
 $\# \text{of months subscribed} + 1.1422 \cdot \text{no magazine} + 1.0099 \cdot \text{issue not resolved})).$

Significant predictors are a subscription to the magazine and whether the issue was resolved or not.

The estimated complementary probabilities for customers who don't receive the magazine are those for customers who receive the magazine raised to the power  $\exp(1.1422) = 3.133655$ . The estimated complementary probabilities for customers whose issues were not resolved are those for customers whose issues were resolved raised to the power  $\exp(1.0099) = 2.745326$ .

```
/*checking model fit*/
proc genmod;
  model satisf = / dist=multinomial link=cumcll;
run;
```

**Log Likelihood -54.4484**

```
data deviance_test;
  deviance = -2*(-54.4484 - (-48.3651));
  pvalue = 1 - probchi(deviance,3);
run;
```

```
proc print noobs;
run;
```

```
deviance      pvalue
12.1666      .006833713
```

The model has a good fit, judging by the small p-value in the goodness-of-fit test. The predicted probabilities are:  $P^0(\text{very dissatisfied}) = 1 - \exp(-\exp(-3.6044 - 0.0030 \cdot 3 + 1.1422)) = 0.081013$ ,  $P^0(\text{very dissatisfied or dissatisfied}) = 1 - \exp(-\exp(-2.4709 - 0.0030 \cdot 3 + 1.1422)) = 0.230834$ ,  $P^0(\text{very dissatisfied, dissatisfied, or neutral}) = 1 - \exp(-\exp(-1.5993 - 0.0030 \cdot 3 + 1.1422)) = 0.466045$ , and  $P^0(\text{very dissatisfied, dissatisfied, neutral, or satisfied}) = 1 - \exp(-\exp(-0.4294 - 0.0030 \cdot 3 + 1.1422)) = 0.867533$ .

The predicted probabilities for the individual levels of the satisfaction score are:

$P^0(\text{very dissatisfied}) = 0.081013$ ,  $P^0(\text{dissatisfied}) = 0.230834 - 0.081013 = 0.149821$ ,  
 $P^0(\text{neutral}) = 0.466045 - 0.230834 = 0.235211$ ,  $P^0(\text{satisfied}) = 0.867533 - 0.466045 = 0.401487$ , and  $P^0(\text{very satisfied}) = 1 - 0.867533 = 0.132467$ .

In SAS:

```
/*using fitted model for prediction*/
data predict;
input subscribed magazine$ resolved$;
cards;
```

```

3 no yes
;

data service;
set service predict;
run;

proc genmod;
  class magazine resolved;
  model satisf = subscribed magazine resolved / dist=multinomial link=cumcll;
  output out=outdata p=psatisf;
run;

proc print data=outdata (firstobs=145) noobs;
  var _level_ psatisf;
run;

```

_LEVEL_	psatisf
1	0.08101
2	0.23084
3	0.46604
4	0.86752

In R:

```

fitting cumulative complementary log-log model
service.data<- read.csv(file="C:/./Exercise4.2Data.csv", header= TRUE, sep=",")

#making response a categorical variable
satisf.cat<- as.factor(service.data$satisf)

#running the model
library(ordinal)
summary(fitted.model<- clm(satisf.cat ~ subscribed + magazine + resolved,
data=service.data, link="cloglog"))

```

AIC  
110.73

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
subscribed	0.002997	0.006339	0.473	0.63641
magazine.relno	-1.142173	0.422566	-2.703	0.00687
resolved.relno	-1.009893	0.430080	-2.348	0.01887

Threshold coefficients:

	Estimate	Std. Error	z value
1 2	-3.6044	0.7514	-4.797
2 3	-2.4709	0.5899	-4.189
3 4	-1.5993	0.5275	-3.032
4 5	-0.4294	0.4201	-1.022

```

#computing AICC
p<-7
n<-36
print(AICC<- -2*logLik(fitted.model)+2*p*n/(n-p-1))

```

114.7302  
#outputting BIC

```
BIC(fitted.model)
```

121.8149

```
#checking model fit
null.model<- clm(satisf.cat ~ 1, data=service.data, link="cloglog")
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))
```

12.16653

```
print(p.value<- pchisq(deviance, df=3, lower.tail = FALSE))
```

0.006833923

```
#using fitted model for prediction
print(predict(fitted.model, data.frame(subscribed=3, magazine="no",
resolved="yes"), type="prob"))
```

```
      1      2      3      4      5
0.0810142 0.1498225 0.2352053 0.4014798 0.1324782
```

(d) Discuss the relative fit of the models obtained in parts (a)-(c).

By the AIC, AICC, and BIC criteria, we see that the cumulative probit regression model has the smallest values and hence has the best fit. The cumulative complementary log-log model has the worse fit.

	cumulative logit	cumulative probit	Cumulative cloglog
AIC	108.0974	107.3854	110.7302
AICC	112.0974	111.3854	114.7302
BIC	119.1820	118.4700	121.8149

**EXERCISE 4.3.** (a) Categorize the amount spent into three categories: "<\$10,000," "\$10,000-<\$30,000," and "\$30,000+". Fit a cumulative logit model. Write down the fitted model, discuss its fit, and interpret estimated significant coefficients. Predict probabilities of each expenditure bracket for a company that has been in business for 4 years, and buys electronics from the supply corporation on a regular basis.

In SAS:

```
/*fitting cumulative logit model*/
data expense;
input inbusiness$ 1-9 firsttime$ type$ 15-25 amount;
cards;
< 1 year   yes stationery   5690
1-5 years  yes stationery   14454
5+ years   yes electronics  20489
5+ years   no  stationery   13115
< 1 year   no  electronics  44885
```

```

< 1 year no electronics 28182
< 1 year no furniture 40982
< 1 year no stationery 10160
1-5 years no furniture 51363
5+ years yes electronics 29448
5+ years no stationery 2093
< 1 year no furniture 127133
1-5 years yes furniture 21593
< 1 year no furniture 220909
1-5 years no electronics 17000
1-5 years yes electronics 22812
1-5 years yes electronics 13090
1-5 years no electronics 24336
5+ years yes stationery 452
< 1 year yes stationery 3600
5+ years yes furniture 2450
< 1 year no electronics 12230
5+ years yes stationery 2451
1-5 years no stationery 1110
< 1 year yes electronics 69280
< 1 year yes furniture 119613
< 1 year no electronics 21770
< 1 year yes electronics 64160
< 1 year no furniture 78900
< 1 year no electronics 75095
5+ years no furniture 7450
5+ years no furniture 5200
< 1 year no furniture 32099
5+ years no electronics 1997
;

```

```
/*categorizing spending amount*/
```

```
data expense;
```

```
set expense;
```

```
length amount_cat $13;
```

```
if amount <10000 then amount_cat="1. <$10K";
```

```
if amount ge 10000 and amount < 30000 then amount_cat="2. $10K-<$30K";
```

```
if amount ge 30000 then amount_cat="3. $30K+";
```

```
run;
```

```
proc genmod;
```

```
class inbusiness(ref="< 1 year") firsttime(ref="yes") type(ref="furniture");
```

```
model amount_cat = inbusiness firsttime type / dist=multinomial link=cumlogit;
run;
```

Log Likelihood -22.7726

AIC 59.5452

AICC 63.8529

BIC 70.2297

#### Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	wald 95% Confidence Limits	wald Pr > ChiSq
					Chi-Square
Intercept1	1	-4.8694	1.4721	-7.7546 -1.9842	10.94 0.0009
Intercept2	1	-1.2584	1.0212	-3.2599 0.7431	1.52 0.2178
inbusiness 1-5 years	1	1.6606	0.9266	-0.1554 3.4766	3.21 0.0731
inbusiness 5+ years	1	4.5406	1.3818	1.8323 7.2490	10.80 0.0010

### Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald	95% Confidence Limits	Wald Chi-Square	Pr > ChiSq
inbusiness < 1 year	0	0.0000	0.0000	0.0000	0.0000	.	.
firsttime no	1	0.1112	0.8040	-1.4647	1.6870	0.02	0.8900
firsttime yes	0	0.0000	0.0000	0.0000	0.0000	.	.
type electronics	1	0.7418	0.8609	-0.9454	2.4291	0.74	0.3888
type stationery	1	3.8781	1.4425	1.0509	6.7053	7.23	0.0072
type furniture	0	0.0000	0.0000	0.0000	0.0000	.	.

The fitted model is  $\frac{P(\text{amount} < \$10,000)}{1 - P(\text{amount} < \$10,000)} = \exp(-4.8694 + 1.6606 \cdot \text{in business 1 to 5 years} + 4.5406 \cdot \text{in business} \geq 5 \text{ years} + 0.1112 \cdot \text{not first time} + 0.7418 \cdot \text{electronics} + 3.8781 \cdot \text{stationery})$ , and  $\frac{P(\text{amount} < \$30,000)}{1 - P(\text{amount} < \$30,000)} = \exp(-1.2584 + 1.6606 \cdot \text{in business 1 to 5 years} + 4.5406 \cdot \text{in business} \geq 5 \text{ years} + 0.1112 \cdot \text{not first time} + 0.7418 \cdot \text{electronics} + 3.8781 \cdot \text{stationery})$ .

Being in business for 5 or more years and purchasing stationery are significant predictors.

The estimated odds for companies that are in business for 5 or more years are  $\exp(4.5406) \cdot 100\% = 9,374.703\%$  of those who are less than 1 year. The estimated odds for companies that purchase stationery are  $\exp(3.8781) \cdot 100\% = 4,833.23\%$  of those for companies that purchase furniture.

```
/*checking model fit*/
proc genmod;
  model amount_cat = / dist=multinomial link=cumlogit;
run;
```

Log Likelihood -37.1492

```
data deviance_test;
  deviance = -2*(-37.1492 - (-22.7726));
  pvalue = 1 - probchi(deviance, 5);
run;
```

```
proc print noobs;
run;
```

```
deviance    pvalue
28.7532 .000025921
```

The model fits the data well as supported by the small magnitude of the p-value in the deviance test. As for predicted probabilities, they are derived as follows. The company we want to predict the probabilities for has been in business between 1 and 5 years, is not a first-time buyer, and purchases electronics. Therefore,  $P^0(\text{amount} < \$10,000) = \frac{\exp(-4.8694 + 1.6606 + 0.1112 + 0.7418)}{1 + \exp(-4.8694 + 1.6606 + 0.1112 + 0.7418)} = 0.086606$ , and  $P^0(\text{amount} < \$30,000) = \frac{\exp(-1.2584 + 1.6606 + 0.1112 + 0.7418)}{1 + \exp(-1.2584 + 1.6606 + 0.1112 + 0.7418)} = 0.778199$ .

Thus, the predicted probabilities for each expenditure bracket are  $P^0(\text{amount} < \$10,000) = 0.086606$ ,  $P^0(\$10,000 \leq \text{amount} < \$30,000) = 0.778199 - 0.086606 = 0.691593$ , and  $P^0(\text{amount} \geq \$30,000) = 1 - 0.778199 = 0.221801$ .

## In SAS:

```
/*using fitted model for prediction*/
data predict;
input inbusiness $ 1-9 firsttime$ type$ 14-24;
cards;
1-5 years no electronics
;

data expense;
set expense predict;
run;

proc genmod;
  class inbusiness firsttime type;
  model amount_cat = inbusiness firsttime type / dist=multinomial link=cumlogit;
  output out=outdata p=ptype;
run;

proc print data=outdata (firstobs=69) noobs;
var _level_ ptype;
run;
```

_LEVEL_	ptype
1. <\$10K	0.08660
2. \$10K-<\$30K	0.77820

## In R:

```
#fitting cumulative logit model
expense.data<- read.csv(file="C:/./Exercise4.3Data.csv", header= TRUE, sep=",")

#specifying reference categories
inbusiness.rel<- relevel(expense.data$inbusiness, ref="< 1 year")
firsttime.rel<- relevel(expense.data$firsttime, ref="yes")
type.rel<- relevel(expense.data$type, ref="furniture")

#categorizing response variable
amount.cat<- as.factor(ifelse(expense.data$amount < 10000, "1.<$10,000",
ifelse(expense.data$amount >= 30000, "3.$30,000+", "2.$10,000-<$30,000")))

#running the model
library(ordinal)
summary(fitted.model<- clm(amount.cat ~ inbusiness.rel + firsttime.rel
+ type.rel, data=expense.data, link="logit"))
```

AIC  
59.55

### Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
inbusiness.rel1-5 years	-1.6606	0.9266	-1.792	0.07309
inbusiness.rel5+ years	-4.5406	1.3818	-3.286	0.00102
firsttime.relno	-0.1112	0.8040	-0.138	0.89003
type.relelectronics	-0.7418	0.8609	-0.862	0.38882
type.relstationery	-3.8781	1.4425	-2.689	0.00718

Threshold coefficients:

	Estimate	Std. Error	z value
1.<\$10,000 2.\$10,000-<\$30,000	-4.869	1.472	-3.308
2.\$10,000-<\$30,000 3.\$30,000+	-1.258	1.021	-1.232

```
#computing AICC
p<- 7
n<- 34
print(AICC<- -2*logLik(fitted.model)+2*p*n/(n-p-1))
```

63.85288

```
#outputting BIC
BIC(fitted.model)
```

70.22971

```
#checking model fit
null.model<- clm(amount.cat ~ 1, data=expense.data, link="logit")
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))
```

28.75325

```
print(p.value<- pchisq(deviance, df=5, lower.tail = FALSE))
```

2.592039e-05

```
#using fitted model for prediction
print(predict(fitted.model, data.frame(inbusiness.rel="1-5 years",
firsttime.rel="no", type="electronics"), type="prob"))
```

1.<\$10,000	2.\$10,000-<\$30,000	3.\$30,000+
0.08660481	0.6915989	0.2217963

(b) Fit a cumulative probit model to the data, and answer the questions in part (a).

In SAS:

```
/*fitting cumulative probit model*/
data expense;
input inbusiness$ 1-9 firsttime$ type$ 15-25 amount;
cards;
< 1 year yes stationery 5690
1-5 years yes stationery 14454
5+ years yes electronics 20489
5+ years no stationery 13115
< 1 year no electronics 44885
< 1 year no electronics 28182
< 1 year no furniture 40982
< 1 year no stationery 10160
1-5 years no furniture 51363
5+ years yes electronics 29448
5+ years no stationery 2093
< 1 year no furniture 127133
1-5 years yes furniture 21593
< 1 year no furniture 220909
1-5 years no electronics 17000
1-5 years yes electronics 22812
```

```

1-5 years yes electronics 13090
1-5 years no electronics 24336
5+ years yes stationery 452
< 1 year yes stationery 3600
5+ years yes furniture 2450
< 1 year no electronics 12230
5+ years yes stationery 2451
1-5 years no stationery 1110
< 1 year yes electronics 69280
< 1 year yes furniture 119613
< 1 year no electronics 21770
< 1 year yes electronics 64160
< 1 year no furniture 78900
< 1 year no electronics 75095
5+ years no furniture 7450
5+ years no furniture 5200
< 1 year no furniture 32099
5+ years no electronics 1997
;

```

```
/*categorizing spending amount*/
```

```
data expense;
```

```
set expense;
```

```
length amount_cat $13;
```

```
if amount <10000 then amount_cat="1. <$10K";
```

```
if amount ge 10000 and amount < 30000 then amount_cat="2. $10K-<$30K";
```

```
if amount ge 30000 then amount_cat="3. $30K+";
```

```
run;
```

```
proc genmod;
```

```
class inbusiness(ref="< 1 year") firsttime(ref="no") type(ref="furniture");
```

```
model amount_cat = inbusiness firsttime type / dist=multinomial link=cumprobit;
run;
```

Log Likelihood -22.8168

AIC 59.6336

AICC 63.9413

BIC 70.3181

#### Analysis of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	wald Confidence	95% Limits	wald Chi-Square	Pr > ChiSq
Intercept1	1	-2.7301	0.7096	-4.1208	-1.3393	14.80	0.0001
Intercept2	1	-0.7096	0.4827	-1.6557	0.2365	2.16	0.1416
inbusiness 1-5 years	1	0.9682	0.5508	-0.1113	2.0477	3.09	0.0788
inbusiness 5+ years	1	2.4398	0.6659	1.1347	3.7449	13.43	0.0002
inbusiness < 1 year	0	0.0000	0.0000	0.0000	0.0000	.	.
firsttime yes	1	0.0503	0.4621	-0.8553	0.9560	0.01	0.9133
firsttime no	0	0.0000	0.0000	0.0000	0.0000	.	.
type electronics	1	0.4350	0.5216	-0.5874	1.4573	0.70	0.4043
type stationery	1	2.0446	0.6890	0.6943	3.3950	8.81	0.0030
type furniture	0	0.0000	0.0000	0.0000	0.0000	.	.

The fitted model can be written as:

as  $\hat{P}(\text{amount} < \$10,000) = \Phi(-2.7301 + 0.9682 \cdot \text{in business 1 to 5 years} + 2.4398 \cdot \text{in business} \geq 5 \text{ years} + 0.0503 \cdot \text{not first time} + 0.4350 \cdot \text{electronics} + 2.0446 \cdot \text{stationery})$ , and  $\hat{P}(\text{amount} < \$30,000) = \Phi(-0.7096 + 0.9682 \cdot \text{in business 1 to 5 years} + 2.4398 \cdot \text{in business} \geq 5 \text{ years} + 0.0503 \cdot \text{not first time} + 0.4350 \cdot \text{electronics} + 2.0446 \cdot \text{stationery})$ . Being in business for 5 or more years and purchasing stationery are significant predictors. The z-score of the estimated probability of lower amount spent for companies that have been in business for 5 or more years is larger by 2.4398 than that for companies that have been in business less than one year. The z-score of the estimated probability for companies that purchase stationery is 2.0446 points larger than that for companies that purchase furniture.

```
/*checking model fit*/
proc genmod;
  model amount_cat = / dist=multinomial link=cumprobit;
run;
```

**Log Likelihood -37.1492**

```
data deviance_test;
  deviance = -2*(-37.1492 - (-22.8168));
  pvalue = 1 - probchi(deviance,5);
run;

proc print noobs;
run;
```

deviance	pvalue
28.6648	.000026976

The model has an excellent fit as indicated by the small p-value in the deviance test. The predicted probabilities can be obtained as follows:

$P^0(\text{amount} < \$10,000) = \Phi(-2.7301 + 0.9682 + 0.0503 + 0.4350) = 0.100872$ , and  
 $P^0(\text{amount} < \$30,000) = \Phi(-0.7096 + 0.9682 + 0.0503 + 0.4350) = 0.771532$ .  
 The predicted probabilities for individual expenditure brackets are  $P^0(\text{amount} < \$10,000) = 0.100872$ ,  $P^0(\$10,000 \leq \text{amount} < \$30,000) = 0.771532 - 0.100872 = 0.67066$ , and  
 $P^0(\text{amount} \geq \$30,000) = 1 - 0.771532 = 0.228468$ .

**In SAS:**

```
/*using fitted model for prediction*/
data predict;
input inbusiness $ 1-9 firsttime$ type$ 14-24;
cards;
1-5 years no electronics
;
run;

data expense;
set expense predict;
run;

proc genmod;
  class inbusiness(ref="< 1 year") firsttime(ref="no") type(ref="furniture");
  model amount_cat = inbusiness firsttime type / dist=multinomial link=cumprobit;
  output out=outdata p=pvalue;
run;
```

```
proc print data=outdata (firstobs=69) noobs;
var _level_ ptype;
run;
```

```
_LEVEL_      ptype
1. <$10K      0.09227
2. $10K-<$30K 0.75603
```

In R:

```
#fitting cumulative probit model
expense.data<- read.csv(file="C:/./Exercise4.3Data.csv", header= TRUE, sep=",")

#specifying reference categories
inbusiness.rel<- relevel(expense.data$inbusiness, ref="< 1 year")
firsttime.rel<- relevel(expense.data$firsttime, ref="yes")
type.rel<- relevel(expense.data$type, ref="furniture")

#categorizing response variable
amount.cat<- as.factor(ifelse(expense.data$amount < 10000, "1.<$10,000",
ifelse(expense.data$amount >= 30000, "3.$30,000+", "2.$10,000-<$30,000")))

#running the model
library(ordinal)
summary(fitted.model<- clm(amount.cat ~ inbusiness.rel + firsttime.rel
+ type.rel, data=expense.data, link="probit"))
```

AIC  
59.63

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
inbusiness.rel1-5 years	-0.96820	0.55078	-1.758	0.078772
inbusiness.rel5+ years	-2.43983	0.66588	-3.664	0.000248
firsttime.relyes	-0.05032	0.46208	-0.109	0.913289
type.relelectronics	-0.43497	0.52163	-0.834	0.404347
type.relstationery	-2.04464	0.68897	-2.968	0.003001

Threshold coefficients:

	Estimate	Std. Error	z value
1.<\$10,000 2.\$10,000-<\$30,000	-2.7301	0.7096	-3.847
2.\$10,000-<\$30,000 3.\$30,000+	-0.7096	0.4827	-1.470

```
#computing AICC
p<- 7
n<- 34
print(AICC<- -2*logLik(fitted.model)+2*p*n/(n-p-1))
```

63.94131

```
#outputting BIC
BIC(fitted.model)
```

70.31814

```
#checking model fit
null.model<- clm(amount.cat ~ 1, data=expense.data, link="probit")
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))
```

28.66482

```
print(p.value<- pchisq(deviance, df=5, lower.tail = FALSE))
```

2.697567e-05

```
#using fitted model for prediction
print(predict(fitted.model, data.frame(inbusiness.rel="1-5 years",
firsttime.rel="no", type.rel="electronics"), type="prob"))
```

1.<\$10,000	2.\$10,000-<\$30,000	3.\$30,000+
0.09227224	0.663757	0.2439708

(c) Repeat part (a) with a cumulative complementary log-log model.

In SAS:

```
/*fitting cumulative complementary log-log model*/
data expense;
input inbusiness$ 1-9 firsttime$ type$ 15-25 amount;
cards;
< 1 year   yes stationery   5690
1-5 years  yes stationery   14454
5+ years   yes electronics  20489
5+ years   no  stationery   13115
< 1 year   no  electronics  44885
< 1 year   no  electronics  28182
< 1 year   no  furniture    40982
< 1 year   no  stationery   10160
1-5 years  no  furniture    51363
5+ years   yes electronics  29448
5+ years   no  stationery   2093
< 1 year   no  furniture    127133
1-5 years  yes furniture     21593
< 1 year   no  furniture    220909
1-5 years  no  electronics  17000
1-5 years  yes electronics  22812
1-5 years  yes electronics  13090
1-5 years  no  electronics  24336
5+ years   yes stationery   452
< 1 year   yes stationery   3600
5+ years   yes furniture     2450
< 1 year   no  electronics  12230
5+ years   yes stationery   2451
1-5 years  no  stationery   1110
< 1 year   yes electronics  69280
< 1 year   yes furniture    119613
< 1 year   no  electronics  21770
< 1 year   yes electronics  64160
< 1 year   no  furniture    78900
< 1 year   no  electronics  75095
5+ years   no  furniture     7450
5+ years   no  furniture     5200
< 1 year   no  furniture    32099
5+ years   no  electronics  1997
;
/*categorizing spending amount*/
```

```

data expense;
set expense;
length amount_cat $13;
if amount <10000 then amount_cat="1. <$10K";
if amount ge 10000 and amount < 30000 then amount_cat="2. $10K-<$30K";
if amount ge 30000 then amount_cat="3. $30K+";
run;

proc genmod;
  class inbusiness(ref="< 1 year") firsttime(ref="no") type(ref="furniture");
  model amount_cat = inbusiness firsttime type / dist=multinomial link=cumcll;
run;

```

Log Likelihood -23.6537

AIC 61.3074

AICC 65.6151

BIC 71.9920

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald	95% Confidence Limits	Wald Pr > Chi-Square	ChiSq
Intercept1	1	-3.5603	0.9430	-5.4084	-1.7121	14.25	0.0002
Intercept2	1	-1.1928	0.5920	-2.3532	-0.0324	4.06	0.0439
inbusiness 1-5 years	1	1.2523	0.6415	-0.0049	2.5096	3.81	0.0509
inbusiness 5+ years	1	2.5159	0.8228	0.9033	4.1286	9.35	0.0022
inbusiness < 1 year	0	0.0000	0.0000	0.0000	0.0000	.	.
firsttime yes	1	0.1249	0.5370	-0.9276	1.1774	0.05	0.8161
firsttime no	0	0.0000	0.0000	0.0000	0.0000	.	.
type electronics	1	0.5535	0.6549	-0.7302	1.8371	0.71	0.3981
type stationery	1	1.9368	0.7914	0.3857	3.4879	5.99	0.0144
type furniture	0	0.0000	0.0000	0.0000	0.0000	.	.

The fitted model is  $\hat{P}(\text{amount} < \$10,000) = 1 - \exp(-\exp(-3.5603 + 1.2523 \cdot \text{in business 1 to 5 years} + 2.5159 \cdot \text{in business} \geq 5 \text{ years} + 0.1249 \cdot \text{first time} + 0.5535 \cdot \text{electronics} + 1.9368 \cdot \text{stationery}))$ , and  $\hat{P}(\text{amount} < \$30,000) = 1 - \exp(-\exp(-1.1928 + 1.2523 \cdot \text{in business 1 to 5 years} + 2.5159 \cdot \text{in business} \geq 5 \text{ years} + 0.1249 \cdot \text{first time} + 0.5535 \cdot \text{electronics} + 1.9368 \cdot \text{stationery}))$ .

Being in business for 5 or more years and purchasing stationery are significant predictors.

The estimated complementary probabilities for companies that are in business for 5 or more years are those for companies that are in business for less than one year raised to the power  $\exp(2.5159) = 12.37774$ . The estimated complementary probabilities for companies that purchase stationery are those for companies that purchase furniture raised to the power  $\exp(1.9368) = 6.936519$ .

```

/*checking model fit*/
proc genmod;
  model amount_cat = / dist=multinomial link=cumcll;
run;

```

Log Likelihood -37.1492

```
data deviance_test;
  deviance = -2*(-37.1492 - (-23.6537));
  pvalue = 1 - probchi(deviance,5);
run;

proc print noobs;
run;
```

deviance	pvalue
26.991	.000057273

The fit of the model is very good since the p-value is very small. The predicted probabilities are computed as  $P^0(\text{amount} < \$10,000) = 1 - \exp(-\exp(-3.5603 + 1.2523 + 0.5535)) = 0.158857$ , and  $P^0(\text{amount} < \$30,000) = 1 - \exp(-\exp(-1.1928 + 1.2523 + 0.5535)) = 0.842126$ . Thus, the predicted probabilities for each expenditure bracket are  $P^0(\text{amount} < \$10,000) = 0.158857$ ,  $P^0(\$10,000 \leq \text{amount} < \$30,000) = 0.842126 - 0.158857 = 0.683269$ , and  $P^0(\text{amount} \geq \$30,000) = 1 - 0.842126 = 0.157874$ .

In SAS:

```
/*using fitted model for prediction*/
data predict;
input inbusiness $ 1-9 firsttime$ type$ 14-24;
cards;
1-5 years no electronics
;
run;

data expense;
set expense predict;
run;

proc genmod;
  class inbusiness firsttime type;
  model amount_cat = inbusiness firsttime type / dist=multinomial link=cumcll;
  output out=outdata p=ptype;
run;

proc print data=outdata (firstobs=69) noobs;
var _level_ ptype;
run;
```

_LEVEL_	ptype
1. <\$10K	0.15886
2. \$10K-<\$30K	0.84213

In R:

```
#fitting cumulative complementary log-log model
expense.data<- read.csv(file="C:/./Exercise4.3Data.csv", header= TRUE, sep=",")

#specifying reference categories
inbusiness.rel<- relevel(expense.data$inbusiness, ref="< 1 year")
firsttime.rel<- relevel(expense.data$firsttime, ref="no")
type.rel<- relevel(expense.data$type, ref="furniture")
```

```
#categorizing response variable
amount.cat<- as.factor(ifelse(expense.data$amount < 10000, "1.<$10,000",
ifelse(expense.data$amount >= 30000, "3.$30,000+", "2.$10,000-<$30,000")))

#running the model
library(ordinal)
summary(fitted.model<- clm(amount.cat ~ inbusiness.rel + firsttime.rel
+ type.rel, data=expense.data, link="cloglog"))
```

AIC  
61.31

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
inbusiness.rel1-5 years	-1.2523	0.6415	-1.952	0.05091
inbusiness.rel5+ years	-2.5159	0.8228	-3.058	0.00223
firsttime.relyes	-0.1249	0.5370	-0.233	0.81610
type.relelectronics	-0.5535	0.6549	-0.845	0.39808
type.relstationery	-1.9368	0.7914	-2.447	0.01439

Threshold coefficients:

	Estimate	Std. Error	z value
1.<\$10,000 2.\$10,000-<\$30,000	-3.560	0.943	-3.776
2.\$10,000-<\$30,000 3.\$30,000+	-1.193	0.592	-2.015

```
#computing AICC
p<- 7
n<- 34
print(AICC<- -2*logLik(fitted.model)+2*p*n/(n-p-1))
```

65.61512

```
#outputting BIC
BIC(fitted.model)
```

71.99195

```
#checking model fit
null.model<- clm(amount.cat ~ 1, data=expense.data, link="cloglog")
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))
```

26.99101

```
print(p.value<- pchisq(deviance, df=5, lower.tail = FALSE))
```

5.727298e-05

```
#using fitted model for prediction
print(predict(fitted.model, data.frame(inbusiness.rel="1-5 years",
firsttime.rel="no", type.rel="electronics"), type="prob"))
```

1.<\$10,000	2.\$10,000-<\$30,000	3.\$30,000+
0.1588638	0.683266	0.1578701

(d) Which of the three fitted models has the best fit?

Comparing the values for AIC, AICC, and BIC criteria, we determine that the cumulative logit model has the smallest values and thus has the best fit.

	cumulative logit	cumulative probit	Cumulative cloglog
AIC	59.5452	59.6336	61.3074
AICC	63.8529	63.9413	65.6151
BIC	70.2297	70.3181	71.9920

**EXERCISE 4.4.** (a) Assuming that the outcome is measured on the nominal scale, run the generalized logit model. Use the correct prediction as the reference category. Write down the fitted model explicitly.

In SAS:

```
/*fitting generalized logit model*/
data forecast;
input elevation water$ winddir windspeed outcome$ @@;
cards;
146 yes 270 2 FA 841 no 360 13 FA 672 yes 360 4 FA
312 no 250 5 FA 126 yes 170 8 FA 607 no 360 8 FA
748 no 270 15 FA 620 yes 290 5 FA 5431 no 200 2 FD
2181 yes 310 8 FD 645 yes 170 7 FD 433 no 270 6 FD
360 no 140 15 FD 4227 yes 200 2 FD 14 yes 150 7 C
1026 no 290 1 C 17 yes 180 2 C 20 yes 270 6 C
15 yes 0 3 C 1135 no 20 13 C 21 yes 30 8 C
98 no 140 8 C 36 yes 10 3 C 8 yes 270 10 C
26 yes 0 3 C 13 yes 170 9 C 9 yes 270 6 C
18 yes 200 12 C 96 no 200 8 C 60 yes 240 9 C
;

proc logistic;
class water(ref="yes") / param=ref;
model outcome(ref="C") = elevation water winddir windspeed / link=glogit;
run;
```

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	25.8630	8	0.0011

Analysis of Maximum Likelihood Estimates

Parameter	outcome	DF	Estimate	Standard Error	wald Chi-Square	Pr > ChiSq
Intercept	FA	1	-7.7634	3.5821	4.6971	0.0302
Intercept	FD	1	-5.2150	3.0482	2.9270	0.0871
elevation	FA	1	0.00109	0.00195	0.3105	0.5774
elevation	FD	1	0.00299	0.00181	2.7306	0.0984
water	no FA	1	0.2216	1.3405	0.0273	0.8687
water	no FD	1	0.0499	1.5847	0.0010	0.9749
winddir	FA	1	0.0250	0.0120	4.3335	0.0374
winddir	FD	1	0.00756	0.00864	0.7668	0.3812

### Analysis of Maximum Likelihood Estimates

Parameter	outcome	DF	Estimate	Standard Error	wald Chi-Square	Pr > ChiSq
windspeed	FA	1	0.1116	0.1756	0.4038	0.5251
windspeed	FD	1	0.1420	0.1962	0.5237	0.4693

The fitted model has the form

$$\frac{\hat{P}(\text{false alarm})}{\hat{P}(\text{correct prediction})} = \exp(-7.7634 + 0.00109 \cdot \text{elevation} + 0.2216 \cdot \text{not near water} + 0.0250 \cdot \text{wind direction} + 0.1116 \cdot \text{wind speed}),$$

and

$$\frac{\hat{P}(\text{failure to detect})}{\hat{P}(\text{correct prediction})} = \exp(-5.2150 + 0.00299 \cdot \text{elevation} + 0.0499 \cdot \text{not near water} + 0.00756 \cdot \text{wind direction} + 0.1420 \cdot \text{wind speed}).$$

In R:

```
#fitting generalized logit model
forecast.data<- read.csv(file="C:/./Exercise4.4Data.csv", header= TRUE, sep=",")
```

```
#specifying reference categories
outcome.rel<- relevel(forecast.data$outcome, ref="C")
water.rel<- relevel(forecast.data$water, ref="yes")
```

```
#running the model
library(nnet)
summary(fitted.model<- multinom(outcome.rel ~ elevation + winddir + windspeed +
water.rel, data=forecast.data))
```

Coefficients:

	(Intercept)	elevation	winddir	windspeed	water.relno
FA	-7.764088	0.001089404	0.025033087	0.1116808	0.22092498
FD	-5.218338	0.002986672	0.007571138	0.1421470	0.04887474

```
#checking model fit
null.model<- multinom(outcome.rel ~ 1, data=forecast.data)
print(deviance<- deviance(null.model)-deviance(fitted.model))
```

25.86298

```
print(p.value<- pchisq(deviance, df=8, lower.tail = FALSE))
```

0.001108523

(b) How good is the model fit? Which variables are significant predictors at the 10% level of significance?

The model fits the data well since in the deviance test, the p-value = 0.0011 < 0.05. Wind direction is significant at 5% (p-value=0.0374) as a predictor for odds in favor of a false alarm against a correct prediction. The elevation is a significant predictor at the 10% level (p-value=0.0984) for odds in favor of failure to detect against a correct prediction.

(c) Give an interpretation of the estimated significant coefficients.

As wind direction increases by one degree clockwise, the estimated odds in favor of a false alarm against a correct prediction increase by  $(\exp(0.0250) - 1) \cdot 100\% = 2.53\%$ . As elevation increases by one foot, the estimated odds in favor of failure to detect against a correct prediction increase by  $(\exp(0.00299) - 1) \cdot 100\% = 0.299\%$ .

(d) Find predicted probabilities of each outcome of the weather forecast for an airport that is located 2,000 feet above sea level, away from a large body of water, in the presence of wind at 5 knots blowing from the east.

The calculations below show how the predicted probabilities are computed.

$$P^0(\text{correct prediction}) = (1 + \exp(-7.7634 + 0.00109 \cdot 2000 + 0.2216 + 0.0250 \cdot 90 + 0.1116 \cdot 5) + \exp(-5.2150 + 0.00299 \cdot 2000 + 0.0499 + 0.00756 \cdot 90 + 0.1420 \cdot 5))^{-1} = (1 + 0.077786 + 9.072973)^{-1} = 0.0985,$$

$$P^0(\text{false alarm}) = P^0(\text{correct prediction}) \cdot \exp(-7.7634 + 0.00109 \cdot 2000 + 0.2216 + 0.0250 \cdot 90 + 0.1116 \cdot 5) = 0.0985 \cdot 0.077786 = 0.0077, \text{ and}$$

$$P^0(\text{failure to detect}) = P^0(\text{correct prediction}) \cdot \exp(-5.2150 + 0.00299 \cdot 2000 + 0.0499 + 0.00756 \cdot 90 + 0.1420 \cdot 5) = 0.0985 \cdot 9.072973 = 0.8938.$$

In SAS:

```
/*using fitted model for prediction*/
data predict;
input elevation water$ winddir windspeed;
cards;
2000 no 90 5
;

data forecast;
set forecast predict;
run;

proc logistic;
class water;
model outcome(ref="C") = elevation water winddir windspeed / link=glogit;
output out=outdata p=poutcome;
run;

proc print data=outdata (firstobs=91) noobs;
var _level_ poutcome;
run;
```

_LEVEL_	poutcome
C	0.09931
FA	0.00772
FD	0.89297

In R:

```
#using fitted model for prediction
print(predict(fitted.model, data.frame(elevation=2000, winddir=90, windspeed=5,
water.rel="no"), type="prob"))
```

C                      FA                      FD  
0.09933447 0.00773318 0.89293235

**EXERCISE 4.5.** (a) Regress the ankle condition on age and gender by running the generalized logit regression model for the nominal response. Use “sprained” as the reference category.

In SAS:

```
/*fitting generalized logit model*/
data ankle;
input age gender$ condition$ @@;
cards;
7 female sprained 9 male torn 11 male broken
12 male broken 8 male torn 8 female torn
9 female broken 13 male broken 13 male torn
15 female sprained 16 female sprained 11 male torn
12 male broken 10 female sprained 9 female torn
8 male sprained 8 female sprained 7 female torn
15 male broken 17 male broken 18 male broken
18 female sprained 18 female torn 16 female torn
12 male broken
;

proc logistic;
class gender(ref="female") / param=ref;
model condition(ref="sprained") = age gender / link=glogit;
run;
```

Testing Global Null Hypothesis: BETA=0  
Test                      Chi-Square DF Pr > ChiSq  
Likelihood Ratio 12.4676    4   0.0142

Analysis of Maximum Likelihood Estimates						
Parameter	condition	DF	Estimate	Standard Error	wald Chi-Square	Pr > ChiSq
Intercept	broken	1	-4.6675	3.1632	2.1773	0.1401
Intercept	torn	1	0.3343	1.6797	0.0396	0.8422
age	broken	1	0.2109	0.2024	1.0861	0.2973
age	torn	1	-0.0454	0.1379	0.1082	0.7422
gender	male broken	1	4.1963	1.6559	6.4216	0.0113
gender	male torn	1	1.5574	1.2735	1.4956	0.2213

In R:

```
#fitting generalized logit model
ankle.data<- read.csv(file="C:/./Exercise4.5Data.csv", header= TRUE, sep=",")

#specifying reference categories
condition.rel<- relevel(ankle.data$condition, ref="sprained")
gender.rel<- relevel(ankle.data$gender, ref="female")

#running the model
library(nnet)
```

```
summary(fitted.model<- multinom(condition.rel ~ age + gender.rel,
data=ankle.data))
```

Coefficients:

	(Intercept)	age	gender.relmale
broken	-4.6658881	0.21077974	4.195560
torn	0.3352406	-0.04544743	1.556776

```
#checking model fit
summary(null.model<- multinom(condition.rel ~ 1, data=ankle.data))
print(deviance<- deviance(null.model)-deviance(fitted.model))
```

12.46759

```
print(p.value<- pchisq(deviance, df=4, lower.tail = FALSE))
```

0.01419266

(b) Write down the estimated model. Discuss its goodness of fit.

The fitted model is  $\frac{\hat{P}(\text{broken})}{\hat{P}(\text{sprained})} = \exp(-4.6675 + 0.2109 \cdot \text{age} + 4.1963 \cdot \text{male})$ , and  
 $\frac{\hat{P}(\text{torn})}{\hat{P}(\text{sprained})} = \exp(0.3343 - 0.0454 \cdot \text{age} + 1.5574 \cdot \text{male})$ .

The p-value for the deviance test is 0.0142 which is less than 0.05, indicating a good fit of the model.

(c) Interpret the estimates of the regression coefficients that significantly differ from zero.

The only statistically significant predictor is gender in the model for the odds of a broken ankle vs. sprained one. The estimated odds for males are  $\exp(4.1963) \cdot 100\% = 6,644.01\%$  of those for females.

(d) What are the predicted probabilities of each type of ankle injury for a 9-year-old girl?

The prediction is carried out as follows:  $P^0(\text{sprained}) = (1 + \exp(-4.6675 + 0.2109 \cdot 9) + \exp(0.3343 - 0.0454 \cdot 9))^{-1} = (1 + 0.0627 + 0.9284)^{-1} = 0.5022$ ,  
 $P^0(\text{broken}) = P^0(\text{sprained}) \cdot \exp(-4.6675 + 0.2109 \cdot 9) = 0.5022 \cdot 0.0627 = 0.0315$ , and  
 $P^0(\text{torn}) = P^0(\text{sprained}) \cdot \exp(0.3343 - 0.0454 \cdot 9) = 0.5022 \cdot 0.9284 = 0.4663$ .

In SAS:

```
/*using fitted model for prediction*/
data predict;
input age gender$;
cards;
9 female
;

data ankle;
set ankle predict;
run;

proc logistic;
class gender;
model condition = age gender / link=glogit;
output out=outdata p=pcondition;
```

```
run;

proc print data=outdata (firstobs=76) noobs;
  var _level_ pcondition;
run;
```

```
_LEVEL_  pcondition
broken    0.03148
sprained  0.50216
torn      0.46636
```

In R:

```
#using fitted model for prediction
print(predict(fitted.model, data.frame(age=9, gender.rel="female"), type="prob")
```

```
      sprained   broken      torn
0.50210809 0.03149866 0.46639325
```

**EXERCISE 4.6.** (a) Regress the communication status on the other variables. Treat it as a nominal variable. Use the zero level as a reference. Write down the fitted model.

In SAS:

```
/*fitting generalized logit model*/
```

```
data datingsite;
input status agediff heightdiff drinking$ @@;
cards;
3 -3 -1 0 3 3 -2 1 3 2 -3 1 3 0 1 1 3 -5 0 1 3 -6 -6 1
3 2 -5 1 3 0 -4 1 3 4 -7 1 3 -1 -8 1 3 -5 1 1 3 -2 2 1
3 -6 -4 1 3 -7 -6 0 2 -5 -1 0 2 -18 0 1 2 -8 3 0 2 4 0 1
2 -4 2 1 2 1 -8 1 2 0 -7 1 2 4 -3 0 1 8 -7 1 1 1 0 1
1 11 0 0 1 -4 -7 0 1 7 -6 1 1 14 -6 1 1 -1 -8 0 1 -5 -4 0
1 -1 -7 0 1 -3 -8 1 1 8 -4 1 1 4 -5 1 0 3 -8 1 0 4 3 0
0 -6 3 0 0 2 -2 0 0 6 3 1 0 6 3 0
;
```

```
proc logistic;
  class drinking (ref="0") / param=ref;
  model status(ref="0") = agediff heightdiff drinking / link=glogit;
run;

Testing Global Null Hypothesis: BETA=0
Test          Chi-Square DF Pr > ChiSq
Likelihood Ratio    29.3421 9    0.0006
```

Analysis of Maximum Likelihood Estimates						
Parameter	status	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1	-0.7248	0.9372	0.5982	0.4393
Intercept	2	1	-0.9584	1.0239	0.8761	0.3493
Intercept	3	1	-1.4013	1.0981	1.6285	0.2019
agediff	1	1	0.0652	0.1258	0.2689	0.6041

Analysis of Maximum Likelihood Estimates						
Parameter	status	DF	Estimate	Standard Error	Wald Pr > ChiSq	Chi-Square
agediff	2	1	-0.2946	0.1429	4.2524	0.0392
agediff	3	1	-0.2717	0.1373	3.9140	0.0479
heightdiff	1	1	-0.5116	0.2083	6.0299	0.0141
heightdiff	2	1	-0.1767	0.1913	0.8535	0.3556
heightdiff	3	1	-0.2532	0.1843	1.8868	0.1696
drinking	1 1	1	-0.1375	1.3391	0.0105	0.9182
drinking	1 2	1	1.8911	1.4309	1.7466	0.1863
drinking	1 3	1	3.0551	1.4460	4.4638	0.0346

The fitted model is written as

$$\frac{\hat{P}(\text{user sent message})}{\hat{P}(\text{neither sent messages})} = \exp(-0.7248 + 0.0652 \cdot \text{agediff} - 0.5116 \cdot \text{heightdiff} - 0.1375 \cdot \text{same drinking preference}),$$

$$\frac{\hat{P}(\text{candidate sent message})}{\hat{P}(\text{neither sent messages})} = \exp(-0.9584 - 0.2946 \cdot \text{agediff} - 0.1767 \cdot \text{heightdiff} + 1.8911 \cdot \text{same drinking preference}),$$

and

$$\frac{\hat{P}(\text{exchanged messages})}{\hat{P}(\text{neither sent messages})} = \exp(-1.4013 - 0.2717 \cdot \text{agediff} - 0.2532 \cdot \text{heightdiff} + 3.0551 \cdot \text{same drinking preference}).$$

In R:

```
#fitting generalized logit model
datingsite.data<- read.csv(file="C:/./Exercise4.6Data.csv", header= TRUE,
sep=",")

#specifying reference categories
status.rel<- relevel(as.factor(datingsite.data$status), ref="0")

#running the model
library(nnet)
summary(fitted.model<- multinom(status.rel ~ agediff + heightdiff + drinking,
data=datingsite.data))

Coefficients:
(Intercept)    agediff heightdiff  drinking
1  -0.7248097  0.06523868 -0.5116114 -0.137546
2  -0.9583625 -0.29461474 -0.1767453  1.891054
3  -1.4012291 -0.27171217 -0.2531624  3.055023

#checking model fit
summary(null.model<- multinom(status.rel ~ 1, data=datingsite.data))
print(deviance<- deviance(null.model)-deviance(fitted.model))
```

29.34211

```
print(p.value<- pchisq(deviance, df=9, lower.tail = FALSE))
```

0.0005672988

(b) Evaluate the goodness-of-fit of the model. What predictors are significant at the 5% level of significance?

The model fits the data well which follows from a small p-value in the deviance test.

The height difference is a significant predictor of odds in favor of the user sending a message as opposed to neither sending a message. The age difference is significant in predicting the odds of a candidate sending a message vs. neither sending message. Drinking preference is significant in predicting odds in favor of exchanged messages vs. neither sending messages.

(c) Give an interpretation of the estimated significant beta coefficients.

As the height difference between a user and a candidate increases by one inch, the estimated odds in favor of the user sending a message vs. neither side sending a message change by  $(\exp(-0.5116) - 1) \cdot 100\% = -40.05\%$ , that is, decrease by 40.05%. As the age difference between a user and candidate increases by one year, the estimated odds in favor of the candidate sending a message vs. neither sending a message change by  $(\exp(-0.2946) - 1) \cdot 100\% = -25.52\%$ , that is, decrease by 25.52%. If the user and candidate have the same drinking preferences, the estimated odds in favor of exchanged messages vs. neither sending messages are  $\exp(3.0551) \cdot 100\% = 2,122.33\%$  of estimated odds for user and candidate with different drinking preferences.

(d) Find the predicted probabilities for each communication status for the case of the same age, height, and drinking preference.

The predicted probabilities are  $P^0(0) = P^0(\text{neither sent messages}) = (1 + \exp(-0.7248 - 0.1375) + \exp(-0.9584 + 1.8911) + \exp(-1.4013 + 3.0551))^{-1} = (1 + 0.42219 + 2.541362 + 5.226804)^{-1} = 0.10881$ ,  $P^0(1) = P^0(\text{user sent message}) = (0.10881) (0.42219) = 0.045938$ ,  $P^0(2) = P^0(\text{candidate sent message}) = (0.10881) (2.541362) = 0.276525$ ,  $P^0(3) = P^0(\text{exchanged messages}) = (0.10881) (5.226804) = 0.568727$ .

In SAS:

```
/*using fitted model for prediction*/
data predict;
input agediff heightdiff drinking$;
cards;
0 0 1
;

data datingsite;
set datingsite predict;
run;

proc logistic;
class drinking;
model status = agediff heightdiff drinking/ link=glogit;
```

```

    output out=outdata p=pstatus;
run;

proc print data=outdata (firstobs=161) noobs;
  var _level_ pstatus;
run;

```

```

_LEVEL_ pstatus
  0 0.10881
  1 0.04594
  2 0.27653
  3 0.56873

```

In R:

```

#using fitted model for prediction
print(predict(fitted.model, data.frame(agediff=0, heightdiff=0, drinking=1),
  type="prob"))

```

```

      0      1      2      3
0.10881062 0.04593619 0.27652468 0.56872852

```

# CHAPTER 5

**EXERCISE 5.1.** (a) Run the Poisson regression model. Discuss the significance of predictors at the 5% level of significance.

In SAS:

```
data defectives;
input ndefectives experience shift$ @@;
cards;
2 3.1 morning 5 2.1 morning 3 8.0 morning 3 7.6 morning 2 5.9 morning
2 4.0 morning 1 1.7 morning 0 1.8 morning 0 8.2 morning 1 8.1 morning
3 3.0 day 3 7.7 day 2 6.3 day 2 8.1 day 2 7.7 day
1 2.4 day 1 3.0 day 1 4.6 day 0 2.1 day 2 3.0 day
5 8.2 evening 4 4.0 evening 4 6.2 evening 3 2.9 evening 2 2.1 evening
2 1.9 evening 1 6.7 evening 1 3.4 evening 1 7.6 evening 6 5.1 night
4 3.2 night 4 7.6 night 4 2.5 night 3 6.2 night 3 2.0 night
5 4.0 night
;

/*fitting Poisson regression model*/
proc genmod;
class shift(ref="day");
model ndefectives = experience shift / dist=poisson link=log;
run;
```

Log Likelihood -3.6998

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter	DF	Estimate	Standard Error	Wald	95% Confidence Limits	Wald Chi-Square	Pr > ChiSq	ChiSq
Intercept	1	0.3571	0.3373	-0.3040	1.0183	1.12	0.2897	
experience	1	0.0355	0.0471	-0.0568	0.1278	0.57	0.4507	
shift evening	1	0.4081	0.3198	-0.2188	1.0350	1.63	0.2019	
shift morning	1	0.1009	0.3342	-0.5541	0.7559	0.09	0.7627	
shift night	1	0.9067	0.3063	0.3064	1.5070	8.76	0.0031	
shift day	0	0.0000	0.0000	0.0000	0.0000	.	.	

Only the night shift is a significant predictor.

In R:

```
defectives.data<- read.csv(file="C:/./Exercise5.1Data.csv", header= TRUE,
sep=",")

shift.rel<- relevel(defectives.data$shift, ref="day")

#fitting Poisson Regression model
summary(fitted.model<- glm(ndefectives ~ experience + shift.rel,
data=defectives.data, family=poisson(link=log)))
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	0.35714	0.33734	1.059	0.28973
experience	0.03552	0.04709	0.754	0.45066

shift.relevening	0.40813	0.31985	1.276	0.20195
shift.relmorning	0.10090	0.33419	0.302	0.76270
shift.relnight	0.90671	0.30630	2.960	0.00307

(b) Write down the estimated model. How good is the fit of the model?

The fitted model has the rate  $\hat{\lambda} = \exp(0.3571 + 0.0355 \cdot \text{months of experience} + 0.4081 \cdot \text{evening shift} + 0.1009 \cdot \text{morning shift} + 0.9067 \cdot \text{night shift})$ . This model has a good fit since the deviance test p-value is less than 0.05.

In SAS:

```
/*checking model fit*/
proc genmod;
  model ndefectives = / dist=poisson link=log;
run;
```

**Log Likelihood -9.3440**

```
data deviance;
  deviance = -2*(-9.3440 - (-3.6998));
  pvalue = 1 - probchi(deviance,4);
run;
```

```
proc print noobs;
run;
```

deviance	pvalue
11.2884	0.023507

In R:

```
#checking model fit
null.model<- glm(ndefectives ~ 1, data=defectives.data,family=poisson(link=log))
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))
```

**11.28837**

```
print(p.value<- pchisq(deviance, df=4, lower.tail = FALSE))
```

**0.02350729**

(c) Give an interpretation of the estimated significant coefficients.

During the night shift, the estimated average number of defective items is  $\exp(0.9067) \cdot 100\% = 247.61\%$  of that during the day shift.

(d) Predict the number of defective items produced during a night shift by an operator with six months of experience.

The predicted number of defective items is  $ndefectives^0 = \exp(0.3571 + 0.0355 \cdot 0.5 + 0.9067) = 3.602219$ .

In SAS:

```
/*using fitted model for prediction*/
data predict;
input experience shift$;
cards;
0.5 night
;

data defectives;
set defectives predict;
run;

proc genmod;
class shift(ref="day");
model ndefectives = experience shift / dist=poisson link=log;
output out=outdata p=pndefectives;
run;
proc print data=outdata (firstobs=37) noobs;
var pndefectives;
run;
```

pndefectives

3.60245

In R:

```
#using fitted model for prediciton
print(predict(fitted.model, data.frame(experience=0.5, shift.rel="night"),
type="response"))
```

3.602453

**EXERCISE 5.2.** (a) Fit the Poisson model to the data and specify estimated parameters. What variables are statistically significant predictors of the number of car accidents? Use  $\alpha = 0.05$ .

In SAS;

```
data autoinsurance;
input naccidents gender$ age miles @@;
cards;
1 M 27 90    1 M 60 70    1 M 36 160    2 M 32 80    2 M 27 150    2 M 58 150
2 M 38 105   3 M 42 75    3 M 55 170   3 M 42 70    3 M 30 110   3 M 54 170
4 M 36 120   4 M 47 145   5 M 20 25    5 M 67 160   5 M 33 140   5 M 41 50
5 M 43 150   6 M 59 130   7 M 65 90    9 M 68 180   0 F 33 110   0 F 40 190
0 F 36 190   0 F 57 140   1 F 47 160   1 F 59 70    1 F 55 180   2 F 44 170
2 F 36 100   2 F 40 170   2 F 58 60    3 F 53 200   3 F 29 180   3 F 51 150
3 F 49 150   4 F 32 180   4 F 51 90    4 F 43 90    4 F 43 20    4 F 31 120
4 F 50 130   4 F 36 50    5 F 40 100   6 F 48 170   6 F 57 180   8 F 66 130
;

/*fitting Poisson regression model*/
proc genmod;
class gender(ref="F");
```

```
model naccidents = gender age miles / dist=poisson link=log;
run;
```

Log Likelihood 33.3456

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald	95% Confidence Limits	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.4492	0.3708	-0.2776	1.1759	1.47	0.2258
gender	M 1	0.2189	0.1609	-0.0965	0.5343	1.85	0.1738
gender	F 0	0.0000	0.0000	0.0000	0.0000	.	.
age	1	0.0171	0.0067	0.0039	0.0303	6.47	0.0110
miles	1	-0.0013	0.0018	-0.0048	0.0022	0.52	0.4712

The fitted model has the mean  $\hat{\lambda} = \exp(0.4492 + 0.2189 \cdot \text{male} + 0.0171 \cdot \text{age} - 0.0013 \cdot \text{miles})$ . Only age is a statistically significant predictor of the number of car accidents.

In R:

```
autoinsurance.data<- read.csv(file="C:/./Exercise5.2Data.csv", header=TRUE,
sep=",")
```

```
gender.rel<- relevel(autoinsurance.data$gender, ref="F")
```

```
#fitting Poisson regression model
summary(fitted.model<- glm(naccidents ~ gender.rel + age + miles,
data=autoinsurance.data, family=poisson(link=log)))
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	0.449155	0.370804	1.211	0.226
gender.relM	0.218899	0.160938	1.360	0.174
age	0.017103	0.006726	2.543	0.011
miles	-0.001283	0.001781	-0.721	0.471

(b) Check the goodness-of-fit of the model.

The model fits the data well as indicated by a p-value below 0.05.

In SAS:

```
/*checking model fit*/
proc genmod;
model naccidents = / dist=poisson link=log;
run;
```

Log Likelihood 29.0520

```
data deviance;
deviance = -2*(29.0520 - 33.3456);
pvalue = 1 - probchi(deviance,3);
run;

proc print noobs;
run;
```

deviance	pvalue
8.5872	0.035314

In R:

```
#checking model fit
null.model<- glm(naccidents ~ 1, data=autoinsurance.data,
family=poisson(link=log))
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))

8.587217

print(p.value<- pchisq(deviance, df=3, lower.tail = FALSE))

0.03531361
```

(c) Interpret the estimated significant regression coefficients.

As the policyholder's age increases by one year, the estimated average number of car accidents caused by the policyholder increases by  $(\exp(0.0171) - 1) \cdot 100\% = 1.7247\%$ .

(d) Give a predicted value of the total number of auto accidents caused by a 35-year-old woman who has driven a total of one hundred thousand miles.

The predicted number of auto accidents is derived as  $naccidents^0 = \exp(0.4492 + 0.0171 \cdot 35 - 0.0013 \cdot 100) = 2.5035$ .

In SAS:

```
/*using fitted model for prediction*/
data predict;
input gender$ age miles;
cards;
F 35 100
;

data autoinsurance;
set autoinsurance predict;
run;

proc genmod;
class gender(ref="F");
model naccidents = gender age miles / dist=poisson link=log;
output out=outdata p=pnaccidents;
run;

proc print data=outdata (firstobs=49) noobs;
var pnaccidents;
run;

pnaccidents
2.50791
```

In R:

```
#using fitted model for prediction
print(predict(fitted.model, data.frame(gender.rel="F", age=35, miles=100),
type="response"))
```

2.507913

**EXERCISE 5.3.** (a) Fit a Poisson regression model for the number of calls. Discuss the model fit.

In SAS:

```
data howlingsurvey;
input ncalls time$ windspeed water$ @@;
cards;
2 dusk 0 yes 2 dusk 1 yes 3 dusk 0 no
2 night 6 no 3 dusk 2 no 4 night 3 yes
5 dusk 1 yes 3 night 5 yes 4 night 5 yes
7 night 0 yes 1 dusk 6 yes 2 night 1 no
4 dusk 2 yes 6 night 2 yes 5 dusk 3 yes
2 night 3 yes 3 dusk 0 yes 0 dusk 3 no
1 dusk 3 yes 2 dusk 3 yes 7 night 2 yes
5 dusk 0 yes 2 night 0 yes 4 night 2 no
6 night 1 yes 3 night 3 yes 0 dusk 1 no
1 dusk 3 no 4 night 3 yes 1 dusk 0 yes
4 dusk 2 yes 1 dusk 2 yes
;

/*fitting Poisson regression model*/
proc genmod;
  class time(ref="dusk") water(ref="no");
  model ncalls = time windspeed water / dist=poisson link=log;
run;
```

Log Likelihood 19.7922

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter	DF	Estimate	Standard Error	Wald	95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	0.6052	0.2980	0.0212	1.1892		4.13	0.0423
time night	1	0.5577	0.2083	0.1494	0.9659		7.17	0.0074
time dusk	0	0.0000	0.0000	0.0000	0.0000		.	.
wind	1	-0.0991	0.0634	-0.2233	0.0251		2.45	0.1179
water yes	1	0.5557	0.2814	0.0041	1.1073		3.90	0.0483
water no	0	0.0000	0.0000	0.0000	0.0000		.	.

```
/*checking model fit*/
proc genmod;
  model ncalls = / dist=poisson link=log;
run;
```

Log Likelihood 12.8090

```
data deviance;
```

```

deviance = -2*(12.8090 - 19.7922);
pvalue = 1 - probchi(deviance,3);
run;

proc print noobs;
run;

```

```

deviance      pvalue
13.9664      .002951247

```

The model has a very good fit due to a small p-value in the deviance test.

In R:

```

howlingsurvey.data<- read.csv(file="C:/./Exercise5.3Data.csv", header=TRUE,
sep=",")

```

```

#reference levels
time.rel<- relevel(howlingsurvey.data$time, ref="dusk")
water.rel<- relevel(howlingsurvey.data$water, ref="no")

```

```

#fitting poisson regression model
summary(fitted.model<- glm(ncalls ~ time.rel + windspeed + water.rel,
data=howlingsurvey.data, family=poisson(link=log)))

```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	0.60517	0.29796	2.031	0.04225
time.relnight	0.55765	0.20827	2.678	0.00742
wind	-0.09909	0.06336	-1.564	0.11786
water.relyes	0.55568	0.28145	1.974	0.04834

```

#checking model fit
null.model<- glm(ncalls ~ 1, data=howlingsurvey.data,family = poisson(link=log))
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))

```

13.96632

```

print(p.value<- pchisq(deviance, df=3, lower.tail = FALSE))

```

0.002951363

(b) Specify the fitted model. Give estimates of all parameters. Which variables are significant at the 5%?

The fitted model has the rate  $\hat{\lambda} = \exp(0.6052 + 0.5577 \cdot \text{night time} - 0.0991 \cdot \text{wind speed} + 0.5557 \cdot \text{water source})$ . The nighttime and the presence of a water source are significant predictors.

(c) Give an interpretation of estimated significant regression coefficients.

When a howling session is conducted at night time, the estimated mean number of calls is  $\exp(0.5577) \cdot 100\% = 174.67\%$  of that for a howling session conducted at dusk time. If there is a water source in the wilderness, the estimated mean number of calls is  $\exp(0.5557) \cdot 100\% = 174.32\%$  of that when there is no water source.

(d) What is the predicted number of wolves that would call back during a howling session conducted at dusk, in a wilderness with no water source, if the wind's speed is 5 mph?

The predicted value is evaluated as  $ncalls^0 = \exp(0.6052 - 0.0991 \cdot 5) = 1.11594$ .

In SAS:

```
/*using fitted model for prediction*/
data predict;
input time$ windspeed water$;
cards;
dusk 5 no
;

data howlingsurvey;
set howlingsurvey predict;
run;

proc genmod;
  class time(ref="dusk") water(ref="no");
  model ncalls = time windspeed water / dist=poisson link=log;
  output out=outdata p=pncalls;
run;

proc print data=outdata (firstobs=33) noobs;
  var pncalls;
run;
```

```
pncalls
1.11596
```

In R:

```
#using fitted model for prediction
print(predict(fitted.model, data.frame(time.rel="dusk", windspeed=5,
water.rel="no"), type="response"))
```

```
1.115965
```

**EXERCISE 5.4.** (a) Model the number of defective items via the zero-truncated Poisson regression model. Display the fitted model. List the significant predictors.

In SAS:

```
data defectives;
input ndefectives experience shift$ @@;
cards;
2 3.1 morning 5 2.1 morning 3 8.0 morning 3 7.6 morning 2 5.9 morning
2 4.0 morning 1 1.7 morning 0 1.8 morning 0 8.2 morning 1 8.1 morning
3 3.0 day 3 7.7 day 2 6.3 day 2 8.1 day 2 7.7 day
1 2.4 day 1 3.0 day 1 4.6 day 0 2.1 day 2 3.0 day
5 8.2 evening 4 4.0 evening 4 6.2 evening 3 2.9 evening 2 2.1 evening
2 1.9 evening 1 6.7 evening 1 3.4 evening 1 7.6 evening 6 5.1 night
4 3.2 night 4 7.6 night 4 2.5 night 3 6.2 night 3 2.0 night
5 4.0 night
;
```

```

data defectives;
set defectives;
if ndefectives>0;
run;

proc format;
value $shiftfmt "morning"="morning" "day"="ref" "evening"="evening"
               "night"="night";
run;

/*fitting zero-truncated Poisson model*/
proc fmm;
  class shift;
  model ndefectives = experience shift / dist=truncpoisson;
format shift $shiftfmt.;
run;

```

-2 Log Likelihood 99.3494

Parameter Estimates for Truncated Poisson Model					
Effect	shift	Estimate	Standard Error	z Value	Pr >  z
Intercept		0.1920	0.4411	0.44	0.6633
experience		0.03355	0.05607	0.60	0.5495
shift	evening	0.4764	0.4062	1.17	0.2409
shift	morning	0.3648	0.4256	0.86	0.3914
shift	night	1.0631	0.3812	2.79	0.0053
shift	ref	0	.	.	.

In the fitted model, the estimated parameter  $\hat{\lambda} = \exp(0.1920 + 0.03355 \cdot \text{months of experience} + 0.4764 \cdot \text{evening shift} + 0.3648 \cdot \text{morning shift} + 1.0631 \cdot \text{night shift})$ . Here only the indicator of the night shift is a significant predictor.

In R:

```

defectives.data<- read.csv(file="C:/./Exercise5.1Data.csv", header= TRUE,
sep=",")

#eliminating zeros
defectives.data<- defectives.data[which(defectives.data$ndefectives != 0),]

#fitting zero-truncated Poisson model
library(VGAM)
summary(fitted.model<- vglm(ndefectives ~ experience + shift,
data=defectives.data, family = pospoisson()))

```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	0.19203	0.44108	0.435	0.66330
experience	0.03355	0.05606	0.598	0.54951
shiftevening	0.47641	0.40615	1.173	0.24080
shiftmorning	0.36479	0.42558	0.857	0.39135
shiftnight	1.06314	0.38115	2.789	0.00528

(b) Discuss the model fit.

The p-value in the deviance test is below 0.05, which supports a good fit of the model.

In SAS:

```
/*checking model fit*/
proc fmm;
  model ndefectives = / dist=truncpoisson;
run;
```

**-2 Log Likelihood 109.3**

```
data deviance;
  deviance = 109.3 - 99.3494;
  pvalue = 1 - probchi(deviance,4);
run;
```

```
proc print noobs;
run;
```

deviance	pvalue
9.9506	0.041268

In R:

```
#checking model fit
null.model<- vglm(ndefectives ~ 1, data=defectives.data, family = pospoisson())
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))
```

**9.904659**

```
print((p.value<- pchisq(deviance, df=4, lower.tail = FALSE)))
```

**0.04206469**

(c) Interpret estimated significant coefficients.

During the night shift, the estimated average number of defective items is  $\exp(1.0631) \cdot 100\% = 289.53\%$  of that during the day shift.

(d) Predict the number of defective items produced during a night shift by an operator with six months of experience.

The predicted number of defective items is

$$ndefectives^0 = \frac{\exp(0.1920+0.03355 \cdot 0.5+1.0631)}{1-\exp(-\exp(0.1920+0.03355 \cdot 0.5+1.0631))} = 3.671155.$$

In SAS:

```
/*using fitted model for prediction*/
data predict;
input experience shift$;
cards;
6 night
;
```

```

data defectives;
set defectives predict;
run;

proc fmm;
class shift;
model ndefectives = experience shift / dist=truncpoisson;
output out=outdata pred=pndefectives;
run;

proc print data=outdata (firstobs=34) noobs;
var pndefectives;
run;

```

```

pndefectives
    3.67138

```

In R:

```

#using fitted model for prediction
print(predict(fitted.model, data.frame(experience=0.5, shift="night"),
type="response"))

```

```

3.671376

```

**EXERCISE 5.5.** In the setting of Exercise 5.2, remove those policyholders who caused no accidents. Run the zero-truncated Poisson regression model on the remaining data.

(a) Write down the fitted model. Are there any significant predictors at the 5% level?

In SAS:

```

data autoinsurance;
input naccidents gender$ age miles @@;
cards;
1 M 27 90    1 M 60 70    1 M 36 160    2 M 32 80    2 M 27 150    2 M 58 150
2 M 38 105   3 M 42 75    3 M 55 170   3 M 42 70    3 M 30 110   3 M 54 170
4 M 36 120   4 M 47 145   5 M 20 25    5 M 67 160   5 M 33 140   5 M 41 50
5 M 43 150   6 M 59 130   7 M 65 90    9 M 68 180   0 F 33 110   0 F 40 190
0 F 36 190   0 F 57 140   1 F 47 160   1 F 59 70    1 F 55 180   2 F 44 170
2 F 36 100   2 F 40 170   2 F 58 60    3 F 53 200   3 F 29 180   3 F 51 150
3 F 49 150   4 F 32 180   4 F 51 90    4 F 43 90    4 F 43 20    4 F 31 120
4 F 50 130   4 F 36 50    5 F 40 100   6 F 48 170   6 F 57 180   8 F 66 130
;

```

```

data autoinsurance;
set autoinsurance;
if naccidents>0;
run;

```

```

proc format;
value $genderfmt "F"="ref" "M"="M";
run;

```

```
/*fitting zero-truncated Poisson model*/
proc fmm;
  class gender;
  model naccidents = gender age miles / dist=truncpoisson;
  format gender $genderfmt.;
run;
```

-2 Log Likelihood 168.2

Parameter Estimates for Truncated Poisson Model					
Effect	gender	Estimate	Standard Error	z Value	Pr >  z
Intercept		0.4869	0.3982	1.22	0.2213
gender	M	0.08270	0.1705	0.48	0.6277
gender	ref	0	.	.	.
age		0.01574	0.007256	2.17	0.0301
miles		-0.00020	0.001906	-0.11	0.9159

The fitted rate  $\hat{\lambda} = \exp(0.4869 + 0.0827 \cdot \text{male} + 0.01574 \cdot \text{age} - 0.0002 \cdot \text{miles})$ . Only age is a significant predictor.

In R:

```
autoinsurance.data<- read.csv(file="C:/./Exercise5.2Data.csv", header=TRUE,
sep=",")

#eliminating zeros
autoinsurance.data<- autoinsurance.data[which(autoinsurance.data$naccidents !=
0),]

#fitting zero-truncated Poisson model
library(VGAM)
summary(fitted.model<- vglm(naccidents ~ gender + age + miles,
data=autoinsurance.data, family = pospoisson()))
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	0.4869342	0.3981630	1.223	0.2213
genderM	0.0826963	0.1705286	0.485	0.6277
age	0.0157404	0.0072558	2.169	0.0301
miles	-0.0002013	0.0019055	-0.106	0.9159

(b) Discuss the fit of the model.

The model doesn't fit the data well because the p-value is larger than 0.05.

In SAS:

```
/*checking model fit*/
proc fmm;
  model naccidents = / dist=truncpoisson;
run;
```

-2 Log Likelihood 173.2

```
data deviance;
  deviance = 173.2 - 168.2;
  pvalue = 1 - probchi(deviance,3);
run;
```

```
proc print noobs;
run;
```

```
deviance    pvalue
      5    0.17180
```

In R:

```
#checking model fit
null.model<- vglm(naccidents ~ 1, data=autoinsurance.data, family = pospoisson())
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))
```

```
5.087076
```

```
print(p.value<- pchisq(deviance, df=3, lower.tail = FALSE))
```

```
0.1655309
```

(c) Interpret the estimated significant beta coefficients.

If the age of a policyholder increases by one year, the estimated mean number of car accidents caused by the policyholder increases by  $(\exp(0.01574) - 1) \cdot 100\% = 1.5865\%$ .

(d) Give a predicted value of the total number of auto accidents caused by a 35-year-old woman who has driven a total of one hundred thousand miles.

The predicted number of car accidents is

$$naccidents^0 = \frac{\exp(0.4869 + 0.01574 \cdot 35 - 0.0002 \cdot 100)}{1 - \exp(-\exp(0.4869 + 0.01574 \cdot 35 - 0.0002 \cdot 100))} = 2.962657.$$

In SAS:

```
/*using fitted model for prediction*/
data predict;
input gender$ age miles;
cards;
F 35 100
;

data autoinsurance;
set autoinsurance predict;
run;

proc fmm;
  class gender;
  model naccidents = gender age miles / dist=truncpoisson;
  output out=outdata pred=pnaccidents;
```

```
run;

proc print data=outdata (firstobs=45) noobs;
  var pnaccidents;
run;
```

```
pnaccidents
      2.95245
```

In R:

```
#using fitted model for prediction
print(predict(fitted.model, data.frame(gender="F", age=35, miles=100),
type="response"))
```

```
2.952453
```

**EXERCISE 5.6.** (a) Model the number of wolves through a zero-truncated Poisson regression model. Estimate all parameters. Are there any significant predictors at the 0.05 level?

In SAS:

```
data howlingsurvey;
input ncalls time$ windspeed water$ @@;
cards;
2 dusk 0 yes 2 dusk 1 yes 3 dusk 0 no 2 night 6 no 3 dusk 2 no
4 night 3 yes 5 dusk 1 yes 3 night 5 yes 4 night 5 yes 7 night 0 yes
1 dusk 6 yes 2 night 1 no 4 dusk 2 yes 6 night 2 yes 5 dusk 3 yes
2 night 3 yes 3 dusk 0 yes 0 dusk 3 no 1 dusk 3 yes 2 dusk 3 yes
7 night 2 yes 5 dusk 0 yes 2 night 0 yes 4 night 2 no 6 night 1 yes
3 night 3 yes 0 dusk 1 no 1 dusk 3 no 4 night 3 yes 1 dusk 0 yes
4 dusk 2 yes 1 dusk 2 yes
;
```

```
data howlingsurvey;
set howlingsurvey;
if (ncalls>0);
run;
```

```
proc format;
value $timefmt "dusk"="ref" "night"="night";
value $waterfmt "no"="zref" "yes"="yes";
run;
```

```
/*fitting zero-truncated Poisson model*/
proc fmm;
  class time water;
  model ncalls = time windspeed water / dist=truncpoisson;
format time $timefmt. water $waterfmt.;
run;
```

```
-2 Log Likelihood 104.6
```

Parameter Estimates for Truncated Poisson Model					
Effect	time	water	Estimate	Standard Error	z Value Pr >  z
Intercept			0.7345	0.3513	2.09 0.0366
time	night		0.5652	0.2291	2.47 0.0136
time	ref		0	.	. .
windspeed			-0.1154	0.06949	-1.66 0.0969
water		yes	0.4085	0.3244	1.26 0.2080
water		zref	0	.	. .

The fitted model has the rate  $\hat{\lambda} = \exp(0.7345 + 0.5652 \cdot \text{night time} - 0.1154 \cdot \text{wind speed} + 0.4085 \cdot \text{water source})$ . The nighttime is the only significant predictor at the 5% level.

In R:

```
howlingsurvey.data<- read.csv(file="C:/./Exercise5.3Data.csv", header=TRUE,
sep=",")

#eliminating zeros
howlingsurvey.data<- howlingsurvey.data[which(howlingsurvey.data$ncalls != 0),]

#fitting zero=truncated Poisson model
library(VGAM)
summary(fitted.model<- vglm(ncalls ~ time + windspeed + water,
data=howlingsurvey.data, family=pospoisson()))
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	0.73449	0.35133	2.091	0.0366
timenight	0.56518	0.22907	2.467	0.0136
windspeed	-0.11538	0.06949	-1.660	0.0969
wateryes	0.40847	0.32443	1.259	0.2080

(b) Test the goodness-of-fit of the model.

The fit of the model is good because the p-value in the deviance test is below 0.05.

In SAS:

```
/*checking model fit*/
proc fmm;
  model ncalls = / dist=truncpoisson;
run;
```

-2 Log Likelihood 114.1

```
data deviance;
  deviance = 114.1 - 104.6;
  pvalue = 1 - probchi(deviance,3);
run;
```

```
proc print noobs;
run;
```

```
deviance    pvalue
      9.5    0.023331
```

In R:

```
#checking model fit
null.model<- vglm(ncalls ~ 1, data=howlingsurvey.data, family=pospoisson())
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))
```

9.512972

```
print(p.value<- pchisq(deviance, df=3, lower.tail = FALSE))
```

0.02319376

(c) Give an interpretation of the estimated significant regression coefficients.

When a howling session is conducted at night time, the estimated mean number of calls is  $\exp(0.5652) \cdot 100\% = 175.98\%$  of that for a howling session conducted at dusk time.

(d) Find the predicted number of wolves that would call back during a howling session conducted at dusk, in a wilderness with no water source, if the wind's speed is 5 mph.

The predicted value is  $ncalls^0 = \frac{\exp(0.7345-0.1154 \cdot 5)}{1-\exp(-\exp(0.7345-0.1154 \cdot 5))} = 1.69695$ .

In SAS:

```
/*using fitted model for prediction*/
data predict;
input time$ windspeed water$;
cards;
dusk 5 no
;

data howlingsurvey;
set howlingsurvey predict;
run;

proc fmm;
  class time water;
  model ncalls = time windspeed water / dist=truncpoisson;
  output out=outdata pred=pncalls;
run;

proc print data=outdata (firstobs=31) noobs;
  var pncalls;
run;

pncalls
1.69704
```

In R:

```
#using fitted model for prediction
print(predict(fitted.model, data.frame(time="dusk", windspeed=5, water="no"),
  type="response"))
```

1.697039

**EXERCISE 5.7.** Consider the zero-inflated Poisson regression model defined by (5.6) - (5.8).

(a) Show that the expected value of  $y$  is  $E(y) = (1 - \pi)\lambda$ .

$$E(Y) = (1 - \pi) \sum_{y=1}^{\infty} y \cdot \frac{\lambda^y e^{-\lambda}}{y!} = (1 - \pi) \sum_{y=0}^{\infty} y \cdot \frac{\lambda^y e^{-\lambda}}{y!} = (1 - \pi)\lambda.$$

- (b) Prove that the estimated gamma coefficients in the expression for  $\hat{\lambda}$  yield the same interpretation as in the Poisson regression model.

Since  $\pi$  and  $\lambda$  are modeled through non-overlapping sets of predictors, when we interpret estimated gamma coefficients, we can assume that  $\pi$  has a fixed value. Therefore, if  $x_{m+1}$  is continuous, then  $\exp(\hat{\gamma}_1)$  represents the ratio of the estimated expected values of  $y$  for  $x_{m+1} + 1$  and  $x_{m+1}$ :

$$\frac{\hat{E}(y)|_{x_{m+1}+1}}{\hat{E}(y)|_{x_{m+1}}} = \frac{(1 - \pi)\hat{\lambda}|_{x_{m+1}+1}}{(1 - \pi)\hat{\lambda}|_{x_{m+1}}} = \frac{\exp(\hat{\gamma}_0 + \hat{\gamma}_1(x_{m+1} + 1) + \hat{\gamma}_2 x_{m+2} + \cdots + \hat{\gamma}_{k-m} x_k)}{\exp(\hat{\gamma}_0 + \hat{\gamma}_1 x_{m+1} + \hat{\gamma}_2 x_{m+2} + \cdots + \hat{\gamma}_{k-m} x_k)} = \exp(\hat{\gamma}_1).$$

If  $x_{m+1}$  is a 0-1 variable, then  $\exp(\hat{\gamma}_1)$  represents the ratio of the estimated expected values of  $y$  for  $x_{m+1} = 1$  and  $x_{m+1} = 0$ . Indeed,

$$\frac{\hat{E}(y)|_{x_{m+1}=1}}{\hat{E}(y)|_{x_{m+1}=0}} = \frac{(1 - \pi)\hat{\lambda}|_{x_{m+1}=1}}{(1 - \pi)\hat{\lambda}|_{x_{m+1}=0}} = \frac{\exp(\hat{\gamma}_0 + \hat{\gamma}_1 \cdot 1 + \hat{\gamma}_2 x_{m+2} + \cdots + \hat{\gamma}_{k-m} x_k)}{\exp(\hat{\gamma}_0 + \hat{\gamma}_1 \cdot 0 + \hat{\gamma}_2 x_{m+2} + \cdots + \hat{\gamma}_{k-m} x_k)} = \exp(\hat{\gamma}_1).$$

**EXERCISE 5.8.** (a) Fit the zero-inflated Poisson regression to model the number of runs in the previous two months. Check if the pace is significantly associated with the inflation of zeros. Write down the fitted model.

In SAS:

```
data races;
input nraces gender$ age run$ pace @@;
cards;
0 F 33 10K 10.04 5 M 26 Full 7.17 0 M 32 10K 11.14
3 F 27 5K 9.18 0 M 48 5K 7.52 4 F 47 10K 11.59
1 M 51 5K 9.44 2 F 49 5K 9.53 0 M 54 10K 8.48
3 F 27 5K 11.71 2 M 24 10K 7.56 0 F 14 5K 13.78
3 M 35 Full 7.34 0 M 50 5K 7.51 0 M 44 5K 8.92
6 F 37 5K 10.71 0 M 54 5K 8.72 2 F 51 10K 7.41
1 F 51 5K 12.28 4 F 35 10K 6.98 2 M 25 10K 12.01
3 M 34 5K 6.78 0 M 28 5K 11.66 0 F 39 10K 12.31
2 M 32 Full 6.58 5 F 44 Full 7.46 0 F 49 10K 11.11
2 M 52 Full 9.2 1 M 30 5K 6.41 1 M 43 10K 7.7
1 M 30 10K 10.01 0 M 53 5K 7.56 2 F 46 Full 8.34
0 F 28 5K 9.67 2 F 50 Full 10.07 2 F 54 5K 7.58
;
/*fitting zero-inflated Poisson model*/
proc genmod;
class gender run(ref="5K");
model nraces = gender age run/ dist=zip;
zeromodel pace;
run;
```

Log Likelihood -15.1014

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter	DF	Estimate	Standard Error	Wald	95% Confidence Limits	Wald Chi-Square	Pr > ChiSq	
Intercept	1	1.6313	0.5632	0.5274	2.7352	8.39	0.0038	
gender F	1	1.0230	0.3036	0.4280	1.6180	11.35	0.0008	
gender M	0	0.0000	0.0000	0.0000	0.0000	.	.	
age	1	-0.0443	0.0148	-0.0733	-0.0153	8.98	0.0027	
run 10K	1	0.1854	0.3608	-0.5218	0.8925	0.26	0.6074	
run Full	1	0.7547	0.3237	0.1203	1.3891	5.44	0.0197	
run 5K	0	0.0000	0.0000	0.0000	0.0000	.	.	

Analysis Of Maximum Likelihood Zero Inflation Parameter Estimates								
Parameter	DF	Estimate	Standard Error	Wald	95% Confidence Limits	Wald Chi-Square	Pr > ChiSq	
Intercept	1	-8.7064	3.9941	-16.5346	-0.8782	4.75	0.0293	
pace	1	0.7255	0.3628	0.0144	1.4366	4.00	0.0455	

In the fitted regression model, the estimated parameters are  $\hat{\pi} = \frac{\exp(-8.7064 + 0.7255 \cdot \text{pace})}{1 + \exp(-8.7064 + 0.7255 \cdot \text{pace})}$ , and  $\hat{\lambda} = \exp(1.6313 + 1.0230 \cdot \text{female} - 0.0443 \cdot \text{age} + 0.1854 \cdot 10K + 0.7547 \cdot \text{full marathon})$ . The pace is a significant predictor for the probability of a structural zero in the number of races in the past four months (that is, the probability of the first race ever), and gender and age significantly predict the average number of races.

In R:

```

races.data<-read.csv(file="C:/./Exercise5.8Data.csv", header = TRUE, sep=",")

#specifying reference levels
run.rel<- relevel(races.data$run, ref="5K")
gender.rel<- relevel(races.data$gender, ref="M")

#fitting zero-inflated Poisson model
library(pscl)
summary(fitted.model<- zeroinfl(nraces ~ gender.rel + age + run.rel | pace,
data=races.data))

```

```

Count model coefficients (poisson with log link):
      Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.63134    0.56313   2.897 0.003769
gender.relF   1.02296    0.30360   3.369 0.000753
age          -0.04431    0.01478  -2.998 0.002719
run.rel10K    0.18539    0.36080   0.514 0.607359
run.relFull   0.75468    0.32369   2.331 0.019727

```

```

Zero-inflation model coefficients (binomial with logit link):
      Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.7064    3.9941  -2.18  0.0293
pace         0.7255    0.3628   2.00  0.0455

```

(b) Discuss the model fit.

The model has a good fit because the p-value is small.

In SAS:

```
/*checking model fit*/
proc genmod;
  model nraces = / dist=zip;
  zeromodel;
run;
```

**Log Likelihood -24.8739**

```
data deviance;
  deviance = -2*(-24.8739 - (-15.1014));
  pvalue = 1 - probchi(deviance, 5);
run;

proc print noobs;
run;
```

deviance	pvalue
19.545	.001520755

In R:

```
#checking model fit
null.model<- zeroinfl(nraces ~ 1, data=races.data)
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))
```

**19.54518**

```
print(p.value<- pchisq(deviance, df=5, lower.tail = FALSE))
```

**0.001520639**

(c) Interpret the estimated significant coefficients.

As the pace increases by one minute per mile, the estimated odds in favor of running the first race ever increase by  $(\exp(0.7255) - 1) \cdot 100\% = 106.58\%$ . The average number of races run in the past four months for females is  $\exp(1.0230) \cdot 100\% = 278.15\%$  of that for males. As age increases by one year, the average number of races run in the past four months changes by  $(\exp(-0.0443) - 1) \cdot 100\% = -4.33\%$ , or decreases by 4.33%.

(d) Calculate the predicted number of races in the past four months for a female runner, aged 45, who ran at an average pace of 10 minutes per mile, if she ran 10K.

The predicted value is  $nraces^0 = \frac{\exp(1.6313+1.0230-0.0443 \cdot 45+0.1854)}{1+\exp(-8.7064+0.7255 \cdot 10)} = 1.8884$ .

In SAS:

```
/*using fitted model for prediction*/
data predict;
input gender$ age run$ pace;
cards;
F 45 10K 10
;

data races;
set races predict;
run;
```

```
proc genmod;
  class gender run;
  model nraces = gender age run / dist=zip;
  zeromodel pace;
  output out=outdata p=pnraces;
run;

proc print data=outdata (firstobs=37) noobs;
  var pnraces;
run;
```

pnraces  
1.88771

In R:

```
#using fitted model for prediction
print(predict(fitted.model, data.frame(gender.rel="F", run.rel="10K", age=45,
pace=10)))
```

1.88771

**EXERCISE 5.9.** (a) Model these data using a zero-inflated Poisson regression with the variable grade responsible for structural zeros, and homework and gender predicting the counting portion. Write the model explicitly, estimating all parameters. Which predictors are significant at the 5% significance level?

In SAS:

```
data readingclub;
input grade hw$ gender$ nbooks @@;
cards;
3 no M 3 3 yes M 3 2 no F 4 2 yes M 3 3 no F 2 1 yes F 0 1 yes F 4
2 no F 0 1 no M 0 3 no M 1 3 yes F 3 2 no F 4 3 no M 0 2 no M 0
1 yes F 5 3 yes M 2 1 no F 1 3 no F 4 1 no F 0 2 yes F 2 3 no F 4
1 no M 2 2 no M 0 2 no F 4 3 no F 5 2 yes F 0 2 yes M 3 2 no M 3
3 yes F 4 3 yes M 3 3 yes F 1 1 no M 0 2 no M 0 1 yes M 0 2 yes F 6
2 yes F 2 2 no F 3 2 no F 0 3 no F 5 3 yes M 2 1 no M 0 3 no F 2
2 yes F 0 2 no M 2 2 no M 0 3 no F 3 1 yes F 1 2 no F 0 1 yes M 1
2 yes M 2
;
/*fitting zero-inflated Poisson model*/
proc genmod;
  class hw(ref="no") gender;
  model nbooks = hw gender / dist=zip;
  zeromodel grade;
run;
```

Log Likelihood -16.9395

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	0.6308	0.2564	0.1283	1.1334	6.05	0.0139
hw	yes	1	0.0715	0.2200	-0.3598	0.11	0.7453
hw	no	0	0.0000	0.0000	0.0000	.	.
gender	F	1	0.4976	0.2455	0.0165	4.11	0.0426
gender	M	0	0.0000	0.0000	0.0000	.	.

Analysis Of Maximum Likelihood Zero Inflation Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	1.4649	1.0651	-0.6227	3.5525	1.89	0.1690
grade	1	-1.2981	0.5616	-2.3988	-0.1975	5.34	0.0208

The fitted parameters are  $\hat{\pi} = \frac{\exp(1.4649 - 1.2981 \cdot \text{grade})}{1 + \exp(1.4649 - 1.2981 \cdot \text{grade})}$ , and  $\hat{\lambda} = \exp(0.6308 + 0.0715 \cdot \text{part of } hw + 0.4976 \cdot \text{female})$ . Grade level is a significant predictor of odds in favor of not turning in the list of books read, and gender is a significant predictor of the average number of books read.

In R:

```
readingclub.data<- read.csv(file="C:/./Exercise5.9Data.csv", header = TRUE,
sep=",")
```

```
#setting reference levels
```

```
hw.rel<- relevel(readingclub.data$hw, ref="no")
```

```
gender.rel<- relevel(readingclub.data$gender, ref="M")
```

```
#fitting zero-inflated Poisson model
```

```
library(pscl)
```

```
summary(fitted.model<- zeroinfl(nbooks ~ hw.rel + gender.rel | grade,
data=readingclub.data))
```

Count model coefficients (poisson with log link):

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	0.63085	0.25643	2.460	0.0139
hw.relyes	0.07149	0.22003	0.325	0.7453
gender.relF	0.49765	0.24547	2.027	0.0426

Zero-inflation model coefficients (binomial with logit link):

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	1.4649	1.0651	1.375	0.1690
grade	-1.2981	0.5616	-2.312	0.0208

(b) Is it a reliable model? Present the quantitative argument for the goodness-of-fit of the model.

The model is reliable because it has a good fit as indicated by a small p-value in the deviance test.

In SAS:

```
/*checking model fit*/
proc genmod;
  model nbooks = / dist=zip;
```

```

    zeromodel;
run;

```

**Log Likelihood -22.4440**

```

data deviance;
    deviance = -2*(-22.4440 - (-16.9395));
    pvalue = 1 - probchi(deviance, 3);
run;

```

```

proc print noobs;
run;

```

```

deviance    pvalue
    11.009   0.011677

```

**In R:**

```

#checking model fit
null.model<- zeroinfl(nbooks ~ 1, data=readingclub.data)
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))

```

**11.00896**

```

print(p.value<- pchisq(deviance, df=3, lower.tail = FALSE))

```

**0.01167754**

(c) How are the estimated significant coefficients interpreted?

As the grade level increases by one, the estimated odds in favor of not turning in the list of books read change by  $(\exp(-1.2981) - 1) \cdot 100\% = -72.70\%$ , that is, decrease by 72.7%. The estimated average number of books read by females is  $\exp(0.4976) \cdot 100\% = 164.48\%$  of that read by males.

(d) What is the predicted number of books read by a second-grade girl for whom the reading is part of the homework?

The predicted number of books is computed as  $nbooks^0 = \frac{\exp(0.6308+0.0715+0.4976)}{1+\exp(1.4649-1.2981 \cdot 2)} = 2.510019$ .

**In SAS:**

```

/*using fitted model for prediction*/
data predict;
input grade hw$ gender$;
cards;
2 yes F
;

```

```

data readingclub;
set readingclub predict;
run;

```

```

proc genmod;
    class hw(ref="no") gender;
    model nbooks = hw gender / dist=zip;

```

```

    zeromodel grade;
    output out=outdata p=pnbooks;
run;

proc print data=outdata (firstobs=51) noobs;
  var pnbooks;
run;

```

pnbooks  
2.51027

In R:

```

#using fitted model for prediction
print(predict(fitted.model, data.frame(grade=2, gender.rel="F", hw.rel="yes")))

```

2.510273

**EXERCISE 5.10.** (a) Run a ZIP model with smoking predicting the probability of excess zeros. Fit the model, and estimate the parameters. Discuss the significance of predictors.

In SAS:

```

data healthsurvey;
input BMI age gender$ smoking$ nattacks @@;
cards;
25.1 61 F no 2 27.1 33 F yes 0 26.8 61 F no 1 23.9 53 F yes 1
26.9 59 M yes 2 18.8 45 F no 0 25.2 54 M yes 2 23.5 75 M yes 5
29.7 64 F no 3 24.5 55 F no 1 21.5 63 M yes 2 37.9 52 M no 0
22.6 43 M no 0 23.0 56 F no 1 28.1 50 F no 0 24.8 86 M yes 6
30.6 74 M yes 4 33.7 71 F yes 3 26.4 66 F yes 1 27.4 25 F no 0
28.6 50 F no 0 20.0 65 F yes 3 31.5 58 F no 2 25.8 64 M yes 5
38.3 56 F no 1 41.4 45 F no 0 31.2 26 F no 0 18.5 42 M yes 0
32.2 26 F no 0 23.9 65 F no 3 31.3 52 M no 2 25.6 32 F no 0
33.2 31 M no 0 23.8 60 M no 2 31.4 55 M no 1 34.0 53 F no 1
27.4 42 M no 3 20.6 61 F yes 2 28.3 64 M no 3 30.1 52 M yes 3
;

/*fitting zero-inflated Poisson model*/
proc genmod;
  class gender(ref="F") smoking(ref="yes");
  model nattacks = BMI gender / dist=zip;
  zeromodel age smoking;
run;

```

Log Likelihood -12.1919

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	wald	95% Confidence Limits	wald Chi-Square	Pr > ChiSq
Intercept	1	1.3685	0.9123	-0.4194	3.1565	2.25	0.1336
BMI	1	-0.0309	0.0338	-0.0972	0.0355	0.83	0.3617
gender M	1	0.5667	0.2613	0.0545	1.0789	4.70	0.0301

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald	95% Confidence Limits	Wald Chi-Square	Pr > ChiSq
gender	F	0	0.0000	0.0000	0.0000	.	.

Analysis Of Maximum Likelihood Zero Inflation Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald	95% Confidence Limits	Wald Chi-Square	Pr > ChiSq
Intercept	1	16.6073	7.0103	2.8673	30.3472	5.61	0.0178
age	1	-0.3640	0.1420	-0.6423	-0.0858	6.58	0.0103
smoking	no	1	1.0721	1.9415	-2.7333	4.8774	0.30
smoking	yes	0	0.0000	0.0000	0.0000	.	.

The fitted ZIP model has parameters  $\hat{\pi} = \frac{\exp(16.6073 - 0.3640 \cdot \text{age} + 1.0721 \cdot \text{doesn't smoke})}{1 + \exp(16.6073 - 0.3640 \cdot \text{age} + 1.0721 \cdot \text{doesn't smoke})}$ , and  $\hat{\lambda} = \exp(1.3685 - 0.0309 \cdot \text{BMI} + 0.5667 \cdot \text{male})$ . Age is a significant predictor of odds in favor of excess zeros in the number of asthma attacks (never had an asthma attack), and gender is a significant predictor of the average number of asthma attacks.

In R:

```
healthsurvey.data<- read.csv(file="C:/./Exercise5.10Data.csv", header = TRUE,
sep=",")

#setting reference levels
gender.rel<- relevel(healthsurvey.data$gender, ref="F")
smoking.rel<- relevel(healthsurvey.data$smoking, ref="yes")

#fitting zero-inflated Poisson model
library(pscl)
summary(fitted.model<- zeroinfl(nattacks ~ BMI + gender.rel | age + smoking.rel,
data=healthsurvey.data))
```

Count model coefficients (poisson with log link):

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	1.36855	0.91207	1.500	0.1335
BMI	-0.03087	0.03383	-0.912	0.3615
gender.relM	0.56670	0.26132	2.169	0.0301

Zero-inflation model coefficients (binomial with logit link):

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	16.6068	7.0060	2.370	0.0178
age	-0.3640	0.1419	-2.566	0.0103
smoking.relno	1.0721	1.9414	0.552	0.5808

(b) How good is the model fit?

The p-value is very small, thus the model has a good fit.

In SAS:

```
/*checking model fit*/
proc genmod;
  model nattacks = / dist=zip;
  zeromodel;
run;
```

Log Likelihood -30.1687

```
data deviance;
  deviance = -2*(-30.1687 - (-12.1919));
  pvalue = 1 - probchi(deviance,4);
run;

proc print noobs;
run;
```

deviance	pvalue
35.9536	.000000296

In R:

```
#checking model fit
null.model<- zeroinfl(nattacks ~ 1, data=healthsurvey.data)
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))
```

35.95365

```
print(p.value<- pchisq(deviance, df=4, lower.tail = FALSE))
```

2.957927e-07

(c) Interpret the estimates of the significant regression coefficients.

As age increases by one year, the estimated odds in favor of no asthma attacks ever change by  $(\exp(-0.364) - 1) \cdot 100\% = -30.51\%$ , that is, decrease by 30.51%. The estimated mean number of asthma attacks for males is  $\exp(0.5667) \cdot 100\% = 176.24\%$  of that for females.

(d) Calculate the predicted value for the number of severe asthma attacks for a male patient, aged 60, whose BMI is 21.2, and who is currently a smoker.

The predicted value is  $nattacks^0 = \frac{\exp(1.3685 - 0.0309 \cdot 21.2 + 0.5667)}{1 + \exp(16.6073 - 0.3640 \cdot 60)} = 3.577968$ .

In SAS:

```
/*using fitted model for prediction*/
data predict;
input BMI age gender$ smoking$;
cards;
21.2 60 M yes
;

data healthsurvey;
set healthsurvey predict;
run;

proc genmod;
class gender(ref="F") smoking(ref="yes");
model nattacks = BMI gender / dist=zip;
zeromodel age smoking;
output out=outdata p=pnattacks;
```

```
run;

proc print data=outdata (firstobs=41) noobs;
  var pnattacks;
run;

pnattacks
  3.58072
```

In R:

```
#using fitted model for prediction
print(predict(fitted.model, data.frame(BMI=21.2, age=60, gender.rel="M",
smoking.rel="yes")))
```

3.5807

**EXERCISE 5.11.** (a) Show that the expected value of the response variable in the hurdle Poisson model has the form  $E(y|x_1, \dots, x_k) = \frac{(1-\pi)\lambda}{1-\exp(-\lambda)}$ .

We write  $E(y|x_1, \dots, x_k) = (1 - \pi) \sum_{y=1}^{\infty} y \cdot \frac{\lambda^y e^{-\lambda}}{y!(1-e^{-\lambda})} = \frac{(1-\pi)}{(1-e^{-\lambda})} \sum_{y=0}^{\infty} y \cdot \frac{\lambda^y e^{-\lambda}}{y!} = \frac{(1-\pi)\lambda}{(1-e^{-\lambda})}$ .

(b) Argue that the estimated regression coefficients in  $\pi$  and  $\lambda$  have the same interpretation as in binary logistic and Poisson regression models, respectively.

The probability of  $y = 0$  is modeled through the logistic function

$$\pi = \frac{\exp(\beta_0 + \beta_1 x_1 + \dots + \beta_m x_m)}{1 + \exp(\beta_0 + \beta_1 x_1 + \dots + \beta_m x_m)}$$

and thus, the estimated regression coefficients  $\hat{\beta}_1, \dots, \hat{\beta}_m$  are interpreted the same way as in the binary logistic regression.

Now,  $\lambda$  is modeled as  $\lambda = \exp(\gamma_0 + \gamma_1 x_{m+1} + \dots + \gamma_{k-m} x_k)$  and  $E(y|x_1, \dots, x_k) = \frac{(1-\pi)\lambda}{(1-e^{-\lambda})}$ .

Since the sets of predictors in  $\pi$  and  $\lambda$  are non-overlapping, when we interpret the estimated gamma coefficients, we can assume that  $\pi$  is constant. We also can assume that the denominator  $1 - \exp(-\lambda)$  is negligibly small due to  $\lambda$  being an exponential function itself. Therefore, the estimated gamma coefficients can be interpreted the same way as in the Poisson regression model.

**EXERCISE 5.12.** (a) Run the hurdle Poisson regression to model the number of computers. Assume that if observations are positive, the number of computers is related to the number of books and periodicals, whereas the zero values are governed by expenditure per student. Write down the fitted model.

In SAS:

```

data libraries;
input ncomps nbooks njrnls budget @@;
cards;
0 8.2 0 0.00 19 11.7 10 16.45 0 2.0 0 5.29
13 8.2 8 23.5 5 30.0 2 6.33 16 14.1 15 7.20
12 9.5 0 3.07 6 21.8 0 4.00 12 9.0 11 4.39
22 5.0 20 17.07 0 15.7 4 1.82 7 19.3 66 9.09
6 20.8 2 10.49 28 11.0 30 0.47 0 9.3 0 0.06
11 12.7 14 0.00 17 15.6 14 22.22 22 9.0 16 0.00
32 18.3 23 22.22 0 12.0 5 0.17 6 8.8 12 7.14
1 14.0 60 1.83 5 12.5 32 24.66 7 3.0 5 7.07
3 16.3 40 12.0 1 6.5 40 13.85 3 8.5 4 18.22
4 10.0 20 30.49 7 18.0 100 0.81 0 11.5 2 0.61
3 9.1 0 9.19 13 10.4 36 25.67 36 7.5 55 7.89
0 19.7 8 1.00
;

/*fitting hurdle poisson model*/
proc fmm;
  model ncomps = nbooks njrnls / dist=truncpoisson;
  model+ / dist=constant;
  probmodel budget;
run;

```

-2 Log Likelihood 311.1

Parameter Estimates for Truncated Poisson Model

Component Effect	Estimate	Standard Error	z Value	Pr >  z
1 Intercept	2.7338	0.1407	19.43	<.0001
1 nbooks	-0.02574	0.01059	-2.43	0.0151
1 njrnls	0.001832	0.002426	0.76	0.4502

Parameter Estimates for Mixing Probabilities

Component Effect	Estimate	Standard Error	z Value	Pr >  z
1 Intercept	-0.4086	0.6625	-0.62	0.5374
1 budget	0.4490	0.2174	2.07	0.0389

In the fitted hurdle Poisson model, the estimates of the parameters are:

$$\hat{\pi} = \frac{\exp(0.4086 - 0.4490 \cdot \text{budget})}{1 + \exp(0.4086 - 0.4490 \cdot \text{budget})} \text{ and } \hat{\lambda} = \exp(2.7338 - 0.02574 \cdot \text{nbooks} + 0.001832 \cdot$$

*njournals*). Note that the estimated regression coefficients for  $\hat{\pi}$  have to be taken with the opposite sign.

Budget is a significant predictor of odds in favor of no computers, and the number of books is a significant predictor of the number of computers when they are present.

In R:

```

libraries.data<- read.csv(file="C:/../Exercise5.12Data.csv", header=TRUE, sep=",")

#fitting hurdle poisson model
library(pscl)

```

```
summary(fitted.model<- hurdle(ncomps ~ nbooks + njrnls | budget,
data=libraries.data, dist="poisson", zero.dist="binomial", link="logit"))
```

Count model coefficients (truncated poisson with log link):

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	2.733815	0.140691	19.431	<2e-16
nbooks	-0.025741	0.010592	-2.430	0.0151
njrnls	0.001832	0.002425	0.755	0.4500

Zero hurdle model coefficients (binomial with logit link):

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-0.4086	0.6625	-0.617	0.5374
budget	0.4490	0.2174	2.065	0.0389

(b) Discuss the model fit.

In the deviance test, the p-value is very small, hence the fit of the model is good.

In SAS:

```
/*checking model fit*/
proc fmm;
  model ncomps = / dist=truncpoisson;
  model+ / dist=constant;
  probmodel;
run;
```

**-2 Log Likelihood 330.3**

```
data deviance;
  deviance = 330.3 - 311.1;
  pvalue = 1 - probchi(deviance,3);
```

```
proc print noobs;
run;
```

deviance	pvalue
19.2	.000248561

In R:

```
#checking model fit
null.model<- hurdle(ncomps ~ 1, data=libraries.data, dist="poisson",
zero.dist="binomial", link="logit")
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))
```

**19.21102**

```
print(p.value<- pchisq(deviance, df=3, lower.tail = FALSE))
```

**0.0002472598**

(c) Interpret estimated significant parameters. State the practical conclusion.

As the budget increases by one dollar per student, the estimated odds in favor of no computers present change by  $(\exp(-0.449) - 1) \cdot 100\% = -36.17\%$ , that is, decrease by 36.17%. As the number of books increases by one thousand, the estimated mean of a positive number of computers changes by  $(\exp(-0.02574) - 1) \cdot 100\% = -2.54\%$ , or decreases by 2.54%.

(d) What is the predicted number of computers in a library with 10,000 books, 25 periodicals, and an annual budget of \$15 per student?

The predicted value is

$$ncomputers^0 = \left(1 - \frac{\exp(0.4086 - 0.4490 \cdot 15)}{1 + \exp(0.4086 - 0.4490 \cdot 15)}\right) \frac{\exp(2.7338 - 0.02574 \cdot 10 + 0.001832 \cdot 25)}{1 - \exp(-\exp(2.7338 - 0.02574 \cdot 10 + 0.001832 \cdot 25))}$$

= 12.43378.

In SAS:

```
/*using fitted model for prediction*/
data predict;
input nbooks njrnls budget;
cards;
10 25 15
;

data libraries;
set libraries predict;
run;

proc fmm;
  model ncomps = nbooks njrnls / dist=truncpoisson;
  model+ / dist=constant;
  probmodel budget;
  output out=outdata pred=pncomps;
run;

proc print data=outdata (firstobs=35) noobs;
  var pncomps;
run;

pncomps
12.4339
```

In R:

```
#using fitted model for prediction
print(predict(fitted.model, data.frame(nbooks=10, njrnls=25, budget=15)))

12.43384
```

**EXERCISE 5.13.** (a) Fit the hurdle Poisson model to verify the hypotheses. Identify all parameters in the predicted model. Is the conclusion supportive of the research hypotheses?

In SAS:

```

data adherence;
input ndaysnoms gender$ age nothermeds @@;
cards;
0 F 87 12 2 M 65 3 0 M 85 3 1 F 68 3 5 F 76 18 1 F 72 9 4 F 73 5
1 M 64 0 2 M 71 1 7 F 81 5 0 M 89 7 4 F 87 8 2 M 78 9 0 F 87 9
1 M 77 4 1 F 71 2 2 M 65 1 5 F 68 7 4 M 73 4 4 F 72 3 0 M 86 13
3 F 66 4 5 F 70 5 1 M 70 5 3 M 62 3 M 93 15 5 F 70 1 3 F 68 11
3 M 75 2 2 M 88 11
;

/*fitting hurdle poisson model*/
proc fmm;
class gender;
model ndaysnoms = gender age / dist=truncpoisson;
model+ / dist=constant;
probmodel nothermeds;
run;

```

-2 Log Likelihood 104.5

Parameter Estimates for Truncated Poisson Model					
Component	Effect	gender	Estimate	Standard Error	z Value Pr >  z
1	Intercept		-1.0213	1.4841	-0.69 0.4913
1	gender	F	0.7288	0.3049	2.39 0.0168
1	gender	M	0	.	. .
1	age		0.02155	0.01995	1.08 0.2800

Parameter Estimates for Mixing Probabilities					
Component	Effect		Estimate	Standard Error	z Value Pr >  z
1	Intercept		3.0554	1.0511	2.91 0.0037
1	nothermeds		-0.2288	0.1122	-2.04 0.0414

The fitted hurdle Poisson model has the estimated parameters:

$\hat{\pi} = \frac{\exp(-3.0554 + 0.2288 \cdot \text{# of other meds})}{1 + \exp(-3.0554 + 0.2288 \cdot \text{# of other meds})}$  and  $\hat{\lambda} = \exp(-1.0213 + 0.7288 \cdot \text{female} + 0.02155 \cdot \text{age})$ . Note that the estimated regression coefficients for  $\hat{\pi}$  have to be taken with the opposite sign since SAS (as well as R) estimates the regression coefficients in  $1 - \pi$ .

The number of other medications is significant as a predictor of  $\pi$ , and gender is significant as a predictor of  $\lambda$ . The research hypotheses are supposed to the extent that the regression coefficients have positive estimates thus women and older patients have a higher mean positive response (even though age is not a significant factor), and the number of other medications is positively associated with the odds in favor of zero response.

In R:

```

adherence.data<- read.csv(file="C:/./Exercise5.13Data.csv", header=TRUE, sep=",")

gender.rel<- relevel(adherence.data$gender, ref="M")

#fitting hurdle Poisson model

```

```
library(pscl)
summary(fitted.model<- hurdle(ndaysnomeds ~ gender.rel + age | nothermeds,
data=adherence.data, dist="poisson", zero.dist = "binomial", link="logit"))
```

Count model coefficients (truncated poisson with log link):

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-1.02120	1.48329	-0.688	0.4912
gender.relF	0.72882	0.30489	2.390	0.0168
age	0.02155	0.01992	1.082	0.2794

Zero hurdle model coefficients (binomial with logit link):

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	3.0554	1.0511	2.907	0.00365
nothermeds	-0.2288	0.1122	-2.039	0.04144

(b) How good is the model fit?

The fit is good as the small p-value indicates.

In SAS:

```
/*checking model fit*/
proc fmm;
  model ndaysnomeds = / dist=truncpoisson;
  model+ / dist=constant;
  probmodel;
run;
```

**-2 Log Likelihood 117.2**

```
data deviance;
  deviance = 117.2 - 104.5;
  pvalue = 1 - probchi(deviance,3);
run;
```

```
proc print noobs;
run;
```

deviance	pvalue
12.7	.005332402

In R:

```
#checking model fit
null.model<- hurdle(ndaysnomeds ~ 1, data=adherence.data, dist="poisson",
zero.dist="binomial", link="logit")
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))
```

**12.7902**

```
print(p.value<- pchisq(deviance, df=3, lower.tail = FALSE))
```

**0.005113002**

(c) Give the interpretation of estimated significant regression coefficients.

$$\hat{\pi} = \frac{\exp(-3.0554 + 0.2288 \cdot \# \text{ of other meds})}{1 + \exp(-3.0554 + 0.2288 \cdot \# \text{ of other meds})} \text{ and } \hat{\lambda} = \exp(-1.0213 + 0.7288 \cdot \text{female} + 0.02155 \cdot \text{age}).$$

As the number of other medications increases by one, the estimated odds in favor of 100% adherence increase by  $(\exp(0.2288) - 1) \cdot 100\% = 25.71\%$ . The estimated mean positive number of days without heart medication for women is  $\exp(0.7288) \cdot 100\% = 207.26\%$  of that for men.

(d) Predict the number of days with missed heart medication for a 78-year-old male patient who is prescribed to take only that one medication.

The predicted value is  $ndaysnoms^0 = \left(1 - \frac{\exp(-3.0554)}{1 + \exp(-3.0554)}\right) \frac{\exp(-1.0213 + 0.02155 \cdot 78)}{1 - \exp(-\exp(-1.0213 + 0.02155 \cdot 78))} = 2.159158$ .

In SAS:

```
/*using fitted model for prediction*/
data predict;
input gender$ age nothermeds;
cards;
M 78 0
;

data adherence;
set adherence predict;
run;

proc fmm;
class gender;
model ndaysnoms = gender age / dist=truncpoisson;
model+ / dist=constant;
probmodel nothermeds;
output out=outdata pred=pndaysnoms;
run;

proc print data=outdata (firstobs=31) noobs;
var pndaysnoms;
run;

pndaysnoms
2.15893
```

In R:

```
#using fitted model for prediction
print(predict(fitted.model, data.frame(gender.rel="M", age=78, nothermeds=0)))

2.158935
```

## CHAPTER 6

**EXERCISE 6.1.** Consider a random experiment consisting of a sequence of independent trials each with outcomes of success or failure. And let  $p$  denote the probability of success.

(a) Let  $X$  be the number of successes observed until the  $r$ th failure. The probability of success is  $p$ . The distribution of  $X$  is a negative binomial random variable with the probability mass function

$$P(X = x) = \binom{x+r-1}{x} p^x (1-p)^r, x = 0, 1, 2, \dots$$

We can write this function as  $P(X = x) = \exp(x \cdot \ln p + r \cdot \ln(1-p) + \ln \binom{x+r-1}{x})$ . If we let  $\theta = \ln p$ , we can rewrite this expression as  $P(X = x) = \exp(x \cdot \theta + r \cdot \ln(1 - e^\theta) + \ln \binom{x+r-1}{x})$ , which has the form as in (1.3) with  $\phi = 1$ ,  $c(\theta) = -r \cdot \ln(1 - e^\theta)$ , and  $h(x, \phi) = \ln \binom{x+r-1}{x}$  and so this distribution belongs to the exponential family of distributions.

(b) Substituting  $p = \frac{\lambda}{r+\lambda}$ , we get  $P(X = x) = \binom{x+r-1}{x} \left(\frac{\lambda}{r+\lambda}\right)^x \left(1 - \frac{\lambda}{r+\lambda}\right)^r = \left(\frac{r}{r+\lambda}\right)^r \binom{x+r-1}{x} \left(\frac{\lambda}{r+\lambda}\right)^x$

$= \left(\frac{r}{r+\lambda}\right)^r \frac{\Gamma(x+r)}{x! \Gamma(r)} \left(\frac{\lambda}{r+\lambda}\right)^x$ ,  $x = 0, 1, 2, \dots$ . The mean is  $E(X) = \frac{pr}{1-p} = \frac{\lambda r}{(r+\lambda)(1-\frac{\lambda}{r+\lambda})} = \frac{\lambda r}{r} = \lambda$ .

The variance is  $Var(X) = \frac{pr}{(1-p)^2} = \frac{\lambda r}{(r+\lambda)(1-\frac{\lambda}{r+\lambda})^2} = \frac{\lambda r(r+\lambda)}{r^2} = \lambda + \frac{\lambda^2}{r}$ .

(c) We use Stirling's formula  $r! \simeq \sqrt{2\pi r} r^r e^{-r}$  to write  $\lim_{r \rightarrow \infty} \left(\frac{r}{r+\lambda}\right)^r \frac{\Gamma(x+r)}{x! \Gamma(r)} \left(\frac{\lambda}{r+\lambda}\right)^x =$   
 $\frac{\lambda^x}{x!} \lim_{r \rightarrow \infty} \frac{r^r}{(r+\lambda)^{r+x}} \frac{\sqrt{2\pi(x+r)}(x+r)^{x+r} e^{-(x+r)}}{\sqrt{2\pi r} r^r e^{-r}} = \frac{\lambda^x}{x!} e^{-x} \lim_{r \rightarrow \infty} \frac{(x+r)^{x+r}}{(r+\lambda)^{r+x}} = \frac{\lambda^x}{x!} e^{-x} \lim_{r \rightarrow \infty} \left(1 + \frac{x-\lambda}{r+\lambda}\right)^{r+\lambda+(x-\lambda)}$   
 $= \frac{\lambda^x}{x!} e^{-x} e^{x-\lambda} = \frac{\lambda^x}{x!} e^{-\lambda}.$

**EXERCISE 6.2.** (a) Model mussel mortality via the negative binomial regression. Present the fitted model. What predictors turn out to be significant at the 5% level?

In SAS:

```
data mussels;
  input max_temp min_temp feeding_level$ ndead_mussels @@;
  cards;
77 60 high 0 88 59 high 1 78 62 high 1 85 60 high 2 78 61 high 0
89 63 high 0 92 62 high 2 75 58 high 0 80 59 med 1 90 61 med 2
74 63 med 4 92 62 med 6 83 62 med 8 75 63 med 3 76 61 med 2
86 62 med 1 92 62 low 2 89 64 low 3 96 68 low 19 86 62 low 7
74 61 low 3 88 62 low 12 97 63 low 9 91 61 low 7
;

/*fitting negative binomial model*/
proc genmod;
  class feeding_level(ref="high");
  model ndead_mussels = max_temp min_temp feeding_level / dist=negbin;
run;
```

Log Likelihood 72.2132

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald	95% Confidence Limits	Chi-Square	Pr > ChiSq
Intercept	1	-11.4507	3.9298	-19.1529	-3.7485	8.49	0.0036
max_temp	1	0.0303	0.0224	-0.0136	0.0741	1.83	0.1764
min_temp	1	0.1418	0.0687	0.0072	0.2765	4.27	0.0389
feeding_level low	1	1.7786	0.4871	0.8239	2.7332	13.33	0.0003
feeding_level med	1	1.4125	0.4844	0.4632	2.3619	8.50	0.0035
feeding_level high	0	0.0000	0.0000	0.0000	0.0000	.	.
Dispersion	1	0.0940	0.1063	0.0103	0.8614		

The fitted model has the estimated parameters  $\hat{\lambda} = \exp(-11.4507 + 0.0303 \cdot \text{maxtemp} + 0.1418 \cdot \text{mintemp} + 1.7786 \cdot \text{low feeding} + 1.4125 \cdot \text{high feeding})$  and  $\hat{\tau} = \frac{1}{0.0940} = 10.6383$ .

At the 5% level, minimum temperature and both feeding levels (low and medium) are significant predictors.

In R:

```
mussels.data<- read.csv(file="C:/./Exercise6.2Data.csv", header=TRUE, sep=",")

#specifying reference level
feeding.level.rel<- relevel(mussels.data$feeding.level, ref="high")

#fitting negative binomial model
library(MASS)

summary(fitted.model<- glm.nb(ndeath.mussels ~ max.temp + min.temp
+ feeding.level.rel, data=mussels.data))
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-11.45070	3.98601	-2.873	0.00407
max.temp	0.03026	0.02222	1.361	0.17340
min.temp	0.14185	0.06925	2.048	0.04052
feeding.level.relow	1.77858	0.48825	3.643	0.00027
feeding.level.relmed	1.41255	0.48347	2.922	0.00348

Theta: 10.6

(b) How good is the model fit?

The p-values are very small, so it can be concluded that the model fits the data well.

In SAS:

```
/*fitting negative binomial model*/
proc genmod;
  class feeding_level;
  model ndeath_mussels = max_temp min_temp feeding_level / dist=negbin;
run;
```

Log Likelihood 58.5757

```
data deviance;
  deviance = -2*(58.5757 - 72.2132);
  pvalue = 1 - probchi(deviance,4);
run;

proc print noobs;
run;
```

```
deviance      pvalue
    27.275 .000017489
```

In R:

```
#checking model fit
null.model<- glm.nb(ndead.mussels ~ 1, data = mussels.data)
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))

27.27491

print(p.value<- pchisq(deviance, df=4, lower.tail = FALSE))

1.749015e-05
```

(c) How would you interpret the estimated significant coefficients?

If the minimum temperature increases by one degree, the estimated mean number of dead mussels increases by  $(\exp(0.1418) - 1) \cdot 100\% = 15.23\%$ . The estimated mean number of dead mussels for low-fed specimens is  $\exp(1.7786) \cdot 100\% = 592.16\%$  of that for high-fed specimens, and for medium-fed ones, it is  $\exp(1.4125) \cdot 100\% = 410.62\%$  of those with a high feeding regimen.

(d) Predict the number of dead mussels that were fed a high level of food, and were located in an area with a maximum temperature of 75 degrees and minimum temperature of 60 degrees.

The predicted value is computed as  $n_{deadmussels}^0 = \exp(-11.4507 + 0.0303 \cdot 75 + 0.1418 \cdot 60) = 0.5116$ .

In SAS:

```
/*using fitted model for prediction*/
data predict;
input max_temp min_temp feeding_level$;
cards;
75 60 high
;

data mussels;
set mussels predict;
run;

proc genmod;
  class feeding_level;
  model ndead_mussels = max_temp min_temp feeding_level / dist=negbin;
  output out=outdata p=pndead_mussels;
run;
```

```
proc print data=outdata (firstobs=25) noobs;
  var pndead_mussels;
run;
```

```
pndead_mussels
      0.51132
```

In R:

```
#using fitted model for prediction
print(predict(fitted.model, data.frame(max.temp=75, min.temp=60,
feeding.level.rel="high"), type="response"))
```

```
0.5113223
```

**EXERCISE 6.3.** (a) Is the negative binomial regression appropriate in modeling the amount of weekly allowance? Fit the model and discuss the significance of the predictor variables.

In SAS:

```
data daily_allowance;
input age gender$ job$ allowance @@;
cards;
15 M yes 0 18 F yes 3 18 M yes 3 14 F no 6
16 F yes 2 17 F yes 1 18 F yes 1 15 F no 4
16 M yes 1 16 F no 9 16 M no 3 16 M no 10
16 F yes 0 14 M no 9 17 M yes 1 15 M no 0
15 M no 12 18 M no 3 15 M no 4 18 M yes 0
15 F no 8 15 M no 5 15 M no 5 14 M no 4
16 F yes 3 17 M no 2 18 M yes 2 17 F yes 11
15 M no 6 16 M no 12
;

/*fitting negative binomial model*/
proc genmod;
  class gender job;
  model allowance = age gender job / dist=negbin;
run;
```

Log Likelihood 81.7552

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	-0.1297	2.6095	-5.2442	4.9849	0.00	0.9604
age		1	0.0359	0.1506	-0.2593	0.3311	0.06	0.8116
gender	F	1	0.4523	0.3192	-0.1733	1.0779	2.01	0.1565
gender	M	0	0.0000	0.0000	0.0000	0.0000	.	.
job	no	1	1.2630	0.4257	0.4286	2.0974	8.80	0.0030
job	yes	0	0.0000	0.0000	0.0000	0.0000	.	.
Dispersion		1	0.3037	0.1550	0.1117	0.8258		

In the fitted model, the estimated parameters  $\hat{\lambda} = \exp(-0.1297 + 0.0359 \cdot \text{age} + 0.4523 \cdot \text{female} + 1.2630 \cdot \text{no job})$  and  $\hat{r} = \frac{1}{0.3037} = 3.2927$ . Only the indicator of no job is significant at the 5% level.

In R:

```
allowance.data<- read.csv(file="C:/./Exercise6.3Data.csv", header=TRUE, sep=",")

#specifying reference levels
gender.rel<- relevel(allowance.data$gender, ref="M")
job.rel<- relevel(allowance.data$job, ref="yes")

#fitting negative binomial model
library(MASS)
summary(fitted.model<- glm.nb(allowance ~ age + gender.rel + job.rel,
data=allowance.data))
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-0.1297	2.3982	-0.054	0.95688
age	0.0359	0.1387	0.259	0.79570
gender.relF	0.4523	0.3067	1.475	0.14033
job.relno	1.2630	0.3937	3.208	0.00134

Theta: 3.29

b) How good is the model fit?

The p-value in the deviance test is below 0.05, indicating a good fit.

In SAS:

```
/*checking model fit*/
proc genmod;
  model allowance = / dist=negbin;
run;

Log Likelihood 75.8189

data deviance;
  deviance = -2*(75.8189 - 81.7552);
  pvalue = 1 - probchi(deviance,3);
run;

proc print noobs;
run;
```

deviance	pvalue
11.8726	.007832568

In R:

```
#checking model fit
null.model<- glm.nb(allowance ~ 1, data=allowance.data)
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))
```

11.87264

```
print(p.value<- pchisq(deviance, df=3, lower.tail = FALSE))
```

0.00783244

(c) Interpret the estimated significant regression coefficients.

The estimated average weekly allowance for students who haven't held a summer job is  $\exp(1.2630) \cdot 100\% = 353.6\%$  than that for students who held a summer job.

(d) Predict the amount of weekly allowance for a male student, age 16, who hasn't held a summer job.

The predicted amount of weekly allowance is  $allowance^0 = \exp(-0.1297 + 0.0359 \cdot 16 + 1.2630) = 5.5163$  or \$27.58.

In SAS:

```
/*using fitted model for prediction*/
data predict;
input age gender$ job$;
cards;
16 M no
;

data daily_allowance;
set daily_allowance predict;
run;

proc genmod;
  class gender job;
  model allowance = age gender job / dist=negbin;
  output out=outdata p=pallowance;
run;

proc print data=outdata (firstobs=31) noobs;
  var pallowance;
run;

pallowance
  5.51684
```

In R:

```
#using fitted model for prediction
print(predict(fitted.model, data.frame(age=16, gender.rel="M", job.rel="no"),
  type="response"))

5.516843
```

**EXERCISE 6.4.** (a) Argue that a zero-truncated negative binomial regression would be appropriate to model the number of rented kayaks. Fit the model. Discuss the significance of predictors.

In SAS:

```
data statepark;
  input nkayaks partysize routelength camped$ @@;
  cards;
6 12 1 yes 2 4 3 yes 3 7 12 no 2 6 3 no 1 3 2 no
2 7 6 yes 2 4 2 no 1 3 6 yes 3 9 12 yes 5 10 4 no
1 2 1 no 2 6 12 yes 4 9 4 no 1 3 2 no 3 7 2 no
7 14 3 no 2 7 12 no 2 6 12 no 3 18 6 yes 2 4 1 yes
2 4 4 yes 4 9 12 yes 3 10 2 no 1 2 3 no 12 12 4 no
10 12 12 yes 2 7 6 yes 3 8 12 no 7 14 3 no 1 3 6 yes
;

/*fitting zero-truncated negative binomial model*/
proc fmm;
  class camped;
  model nkayaks = partysize routelength camped / dist=truncnegbin;
run;
```

-2 Log Likelihood 96.4664

Parameter Estimates for Truncated Negative Binomial Model					
Effect		camped Estimate	Standard Error	z Value	Pr >  z
Intercept		-0.8576	0.5795	-1.48	0.1389
partysize		0.1833	0.04275	4.29	<.0001
routelength		0.02644	0.03529	0.75	0.4537
camped	no	0.2731	0.3045	0.90	0.3698
camped	yes	0	.	.	.
Scale Parameter		0.1116	0.1073		

There are no zeros and variability is larger than for the Poisson distribution, therefore a zero-truncated negative binomial model should be appropriate. The fitted model has the estimated parameters  $\hat{\lambda} = \exp(-0.8576 + 0.1833 \cdot \text{party size} + 0.02644 \cdot \text{route length} + 0.2731 \cdot \text{no camping})$  and  $\hat{\phi} = \frac{1}{0.1116} = 8.9606$ . Only party size is a significant predictor.

In R:

```
statepark.data<- read.csv(file="C:/./Exercise6.4Data.csv", header = TRUE,
sep=",")

#specifying reference level
camped.rel<- relevel(statepark.data$camped, ref="yes")

#fitting zero-truncated negative binomial model
library(VGAM)
summary(fitted.model<- vglm(nkayaks ~ partysize + routelength + camped.rel,
data=statepark.data, family=posnegbinomial()))
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept):1	-0.85770	0.55260	-1.552	0.121
(Intercept):2	2.19308	1.15141	NA	NA
partysize	0.18331	0.03440	5.329	9.89e-08
routelength	0.02644	0.03582	0.738	0.460
camped.relno	0.27304	0.29855	0.915	0.360

$$\hat{r} = \exp(2.19308) = 8.962776.$$

(b) Discuss model fit.

In the deviance test, the p-value is very small, suggesting that the model has a good fit.

In SAS:

```
/*checking model fit*/
proc fmm;
  model nkayaks = / dist=truncnegbin;
run;
```

2 Log Likelihood 121.1

```
data deviance;
  deviance = 121.1 - 96.4664;
  pvalue= 1 - probchi(deviance,3);
run;
```

```
proc print noobs;
run;
```

deviance	pvalue
24.6336	.000018418

In R:

```
#checking model fit
null.model<- vglm(nkayaks ~ 1, data=statepark.data, family=posnegbinomial())
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))
```

24.63644

```
print(p.value<- pchisq(deviance, df=3, lower.tail = FALSE))
```

1.839315e-05

(c) Interpret the estimated significant regression coefficients, whatever are possible to interpret.

Interpretation is traditionally omitted due to the complexity of the expression for the expected value.

(d) Predict the number of rented kayaks for a party of 5 people who plan to take a 6-hour route and camp overnight.

$$\text{The predicted value is } nkayaks^0 = \frac{\exp(-0.8576+0.1833 \cdot 5+0.02644 \cdot 6)}{1-(1+\exp(-0.8576+0.1833 \cdot 5+0.02644 \cdot 6)/8.9606)^{-8.9606}} = 1.8073.$$

In SAS:

```
/*using fitted model for prediction*/
data predict;
input partysize routelength camped$;
cards;
5 6 yes
;

data statepark;
set statepark predict;
run;

proc fmm;
class camped;
model nkayaks = partysize routelength camped / dist=truncnegbin;
output out=outdata pred=pnkayaks;
run;

proc print data=outdata (firstobs=31) noobs;
var pnkayaks;
run;

pnkayaks
1.80725
```

In R:

```
#using fitted model for prediction
print(predict(fitted.model, data.frame(partysize=5,routelength=6,
camped.rel="yes"), type="response"))

1.807214
```

**EXERCISE 6.5.** (a) Run the zero-truncated negative binomial model to regress the number of new videos on the other variables. Write the predicted model. What predictors turn out to be significant at the 5% level?

In SAS:

```
data vlogs;
length type $10;
input nnewvideos nvideos nsubscr nvIEWS type$ @@;
cards;
3 81 3.9 205.8 lifeadvice 4 188 27 213.6 fashion
1 55 10.1 176.8 products 4 123 14.4 59.7 science
1 65 5 508.7 lifeadvice 2 118 3.5 280.6 comedy
3 119 4.7 25.7 fashion 1 47 4.4 135.8 products
2 405 58 423.6 comedy 4 160 10.9 212.8 science
4 123 1.3 204.1 fashion 1 96 1.1 449 comedy
2 44 2.7 217.7 fashion 1 71 8 12.3 lifeadvice
4 190 6.7 433.3 lifeadvice 1 59 9.5 90.4 science
1 36 9.2 423.9 products 3 511 92.5 158.4 products
2 112 4.2 225.7 products 4 156 32.4 140.8 comedy
```

```

3 212 1.3 121.1 products 4 86 4.2 160.2 fashion
12 517 85.4 163.7 lifeadvice 7 100 8 91.5 news
9 130 2.8 38.9 lifeadvice 2 34 2.4 151.9 fashion
30 396 7.6 118.4 comedy 12 52 0.9 617.2 lifeadvice
9 43 7.7 542.6 comedy 22 304 2.6 150.5 news
10 430 1.4 242.1 comedy 2 76 15.2 106.7 fashion
2 53 3.6 121.1 fashion 9 98 1 160.2 news
19 56 4.7 163.7 news 4 102 0.9 91.5 fashion
2 43 0.5 38.9 fashion 14 81 3.2 151.9 products
4 86 3.2 118.4 products 10 90 2.6 617.2 products
;

proc format;
value $typefmt "fashion"="zref_fashion";
run;

/*fitting zero-truncated poisson model*/
proc fmm;
class type;
model nnewvideos = nvideos nsubscr nviews type / dist=truncnegbin;
format type $typefmt.;
run;

```

-2 Log Likelihood 191.9

Parameter Estimates for Truncated Negative Binomial Model					
Effect	type	Estimate	Standard Error	z Value	Pr >  z
Intercept		0.3626	0.3995	0.91	0.3640
nvideos		0.004948	0.001676	2.95	0.0032
nsubscr		-0.02072	0.009700	-2.14	0.0326
nviews		0.000955	0.000947	1.01	0.3132
type	comedy	0.3338	0.5390	0.62	0.5357
type	lifeadvice	0.5773	0.4984	1.16	0.2467
type	news	1.5310	0.5000	3.06	0.0022
type	products	0.3360	0.4561	0.74	0.4613
type	science	0.01012	0.6614	0.02	0.9878
type	Zref_fashion	0	.	.	.
Scale Parameter		0.4648	0.2165		

The fitted model has the estimated parameters  $\hat{\lambda} = \exp(0.3626 + 0.004948 \cdot nvideos - 0.02072 \cdot nsubscr + 0.000955 \cdot nviews + 0.3338 \cdot comedy + 0.5773 \cdot life\ advice + 1.5310 \cdot news + 0.3360 \cdot products + 0.01012 \cdot science)$  and  $\hat{\phi} = \frac{1}{0.4648} = 2.1515$ . At the 5% significance level, the significant predictors are the total number of videos, the number of subscribers, and the type of video “news”.

In R:

```

vlogs.data<- read.csv(file="C:/./Exercise6.5Data.csv", header=TRUE, sep=",")

#specifying reference level
type.rel<- releval(vlogs.data$type, ref="fashion")

#fitting zero-truncated negative binomial model
library(VGAM)

```

```
summary(fitted.model<- vglm(nnewvideos ~ nvideos + nsubscr + nviews + type.rel,
data=vlogs.data, family = posnegbinomial()))
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept):1	0.3625437	0.3899071	0.930	0.35246
(Intercept):2	0.7660984	0.4623516	1.657	0.09753
nvideos	0.0049484	0.0016068	3.080	0.00207
nsubscr	-0.0207244	0.0094039	-2.204	0.02754
nviews	0.0009554	0.0010078	0.948	0.34314
type.relcomedy	0.3336973	0.5483930	0.609	0.54286
type.rellifeadvice	0.5772579	0.4939778	1.169	0.24257
type.relnews	1.5310712	0.5050049	3.032	0.00243
type.relproducts	0.3359591	0.4596850	0.731	0.46487
type.relscience	0.0101134	0.6663842	0.015	0.98789

$\hat{r} = \exp(0.76660984) = 2.152457.$

(b) Does the model have a good fit?

The fit of the model is good, as supported by a small p-value in the deviance test.

In SAS:

```
/*checking model fit*/
proc fmm;
  model nnewvideos = / dist=truncnegbin;
run;
```

**-2 Log Likelihood 212.8**

```
data deviance;
  deviance = 212.8 - 191.9;
  pvalue = 1 - probchi(deviance,8);
run;
```

```
proc print noobs;
run;
```

deviance	pvalue
20.9	.007417868

In R:

```
#checking model fit
null.model<- vglm(nnewvideos ~ 1, data = vlogs.data, family=posnegbinomial())
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))
```

**20.89688**

```
print(p.value<- pchisq(deviance, df=8, lower.tail = FALSE))
```

**0.007426463**

(c) Interpret estimated significant regression coefficients, if possible.

Interpretation is not possible for this regression.

(d) Find the predicted number of new videos for a vlogger who posted a total of 87 videos on popular science, has 50,000 subscribers, and has 254,000 views.

The predicted number of new videos is computed as:

$$nnewvideos^0 = \frac{\exp(0.3626+0.004948 \cdot 87 - 0.02072 \cdot 50 + 0.000955 \cdot 254 + 0.01012)}{1 - (1 + \exp(0.3626+0.004948 \cdot 87 - 0.02072 \cdot 50 + 0.000955 \cdot 254 + 0.01012)/2.1515)^{-2.1515}}$$

= 1.793459.

In SAS:

```
/*using fitted model for prediction*/
data predict;
input nvideos nsubscr nviews type$;
cards;
87 50 254 science
;

data vlogs;
set vlogs predict;
run;

proc fmm;
  class type;
  model nnewvideos = nvideos nsubscr nviews type / dist=truncnegbin;
  output out=outdata pred=pnnewvideos;
run;

proc print data=outdata (firstobs=41) noobs;
  var pnnewvideos;
run;

pnnewvideos
    1.79337
```

In R:

```
#using fitted model for prediction
print(predict(fitted.model, data.frame(nvideos=87, nsubscr=50, nviews=254,
type.rel="science"), type="response"))

1.793339
```

**EXERCISE 6.6.** (a) Fit a zero-inflated negative binomial regression to model the number of claims made in the past five years. Model the probability of structural absence of claims as a function of the number of claims made in the previous five years. Model the positive responses as related to age and gender. What predictors are significant at the 5% level?

In SAS:

```
data insurance;
input nclaimspast5ys nclaimsprev5ys age gender$ @@;
```

```

cards;
1 1 39 M 1 2 66 M 7 0 56 M 3 4 43 F 4 1 42 F 4 2 52 M 0 0 39 F
4 6 68 M 6 1 41 F 0 1 54 F 4 2 50 F 6 4 57 M 5 4 47 F 1 2 43 M
1 1 36 M 1 2 55 F 5 5 57 F 8 5 53 M 0 1 72 M 0 1 67 F 8 3 69 F
0 2 70 M 7 2 70 M 3 1 54 F 2 1 38 M 3 1 50 F 0 1 62 M 8 2 54 M
0 0 59 M 0 1 61 F 0 0 69 F 8 3 57 F 0 0 57 M 12 5 72 F 0 2 42 M
6 2 42 F 7 2 66 M 7 4 53 M 6 0 52 M 3 3 57 F
;

/*fitting zero-inflated poisson model*/
proc genmod;
  class gender;
  model nclaimspast5ys = age gender / dist=zinb;
  zeromodel nclaimsprev5ys;
run;

```

Log Likelihood -82.0185

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.1080	0.5354	-1.1574	0.9414	0.04	0.8401
age	1	0.0297	0.0092	0.0116	0.0477	10.40	0.0013
gender F	1	0.1279	0.1842	-0.2332	0.4889	0.48	0.4876
gender M	0	0.0000	0.0000	0.0000	0.0000	.	.
Dispersion	1	0.0309	0.0725	0.0003	3.0534		

Analysis Of Maximum Likelihood Zero Inflation Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	0.8826	0.7204	-0.5294	2.2947	1.50	0.2205
nclaimsprev5ys	1	-1.3297	0.5397	-2.3876	-0.2718	6.07	0.0138

The estimated parameters of the fitted model are  $\hat{\pi} = \frac{\exp(0.8826 - 1.3297 \cdot \text{nclaimsprev5ys})}{1 + \exp(0.8826 - 1.3297 \cdot \text{nclaimsprev5ys})}$ ,

$\hat{\lambda} = \exp(-0.1080 + 0.0297 \cdot \text{age} + 0.1279 \cdot \text{female})$ , and  $\hat{\pi} = \frac{1}{0.0309} = 32.36246$ .

The number of claims in the previous five years is a significant predictor of  $\pi$ , while age is a significant predictor of  $\lambda$ .

In R:

```

insurance.data<- read.csv(file="C:/./Exercise6.6Data.csv", header=TRUE, sep=",")

#specifying reference level
gender.rel<- relevel(insurance.data$gender, ref="M")

#fitting zero-inflated negative binomial model
library(pscl)
summary(fitted.model<- zeroinfl(nclaimspast5ys ~ age + gender.rel |
nclaimsprev5ys, data=insurance.data, dist="negbin"))

```

Count model coefficients (negbin with log link):

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-0.108006	0.536028	-0.201	0.84031

age	0.029678	0.009213	3.221	0.00128
gender.relf	0.127864	0.184205	0.694	0.48760
Log(theta)	3.475381	2.344614	1.482	0.13827

Zero-inflation model coefficients (binomial with logit link):

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	0.8826	0.7204	1.225	0.2205
nclainsprev5ys	-1.3297	0.5397	-2.464	0.0138

Theta = 32.3101

(b) Interpret the estimated significant coefficients.

As the number of claims in the previous five years increases by one, the estimated odds in favor of no claim in the past five years change by  $(\exp(-1.3297) - 1) \cdot 100\% = -73.5443\%$ , that is, decrease by 73.5443%. As the age of a policyholder increases by one year, the estimated average number of claims in the past five years increases by  $(\exp(0.0297) - 1) \cdot 100\% = 3.014544\%$ .

(c) How good is the model fit? Give a quantitative answer.

In the deviance test, the p-value is very small, thus, the model has a good fit.

In SAS:

```
/*checking model fit*/
proc genmod;
  model nclainspast5ys = / dist=zinb;
  zeromodel;
run;

Log Likelihood -92.1407

/*checking model fit*/
proc genmod;
  model nclainspast5ys = / dist=zinb;
  zeromodel;
run;

data deviance;
  deviance = -2*(-92.1407 - (-82.0185));
  pvalue = 1 - probchi(deviance,3);
run;

proc print noobs;
run;

deviance      pvalue
20.2444      .000151052
```

In R:

```
#checking model fit
null.model<- zeroinfl(nclainspast5ys ~ 1, data=insurance.data, dist="negbin")
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))

20.2445
```

```
print(p.value<- pchisq(deviance, df=3, lower.tail = FALSE))
```

0.0001510455

(d) What is the predicted number of claims made in the past five years by a 55-year-old female policyholder who has made no claims in the previous five years?

The predicted number of claims is  $claims^0 = \frac{\exp(-0.1080+0.0297 \cdot 55+0.1279)}{1+\exp(0.8826)} = 1.528956$ .

In SAS:

```
/*using fitted model for prediction*/
data predict;
input nclaimsprev5ys age gender$;
cards;
0 55 F
;

data insurance;
set insurance predict;
run;

proc genmod;
  class gender;
  model nclaimspast5ys = age gender / dist=zinb;
  zeromodel nclaimsprev5ys;
  output out=outdata p=pnclaimspast5ys;
run;

proc print data=outdata (firstobs=41) noobs;
  var pnclaimspast5ys;
run;
```

```
pnclaimspast5ys
      1.52697
```

In R:

```
#using fitted model for prediction
print(predict(fitted.model, data.frame(nclaimsprev5ys=0, age=55,
gender.rel="F")))
```

1.526969

**EXERCISE 6.7.** (a) Fit a zero-inflated negative binomial model, regressing the probability of structural zeros of the DMFT index on age. Regress positive observations of DMFT index on gender and levels of oral hygiene. Write down the predicted model. Discuss the significance of predictors at the 5% significance level.

In SAS:

```
data dental;
```

```

input DMFTindex age gender $ oralhygiene $ @@;
cards;
0 28 F high 2 30 F med 0 26 F high 15 55 M high 8 40 F med
2 19 M med 0 24 F med 8 77 F low 5 48 F high 3 21 F med
11 59 M med 9 50 M high 1 24 F med 0 26 M med 1 23 F high
2 24 F med 1 21 M low 2 40 M med 0 31 F med 11 29 M low
0 20 F high 0 25 F high 1 22 F high 7 37 M med 2 56 F med
15 63 M high 0 21 M med 5 55 F high 0 25 F high 2 68 M low
4 25 M med 6 59 F low 9 58 F med 0 37 M med 0 18 M high
16 73 M med 3 23 M med 8 65 M med
;

/*fitting zero-inflated negative binomial model*/
proc genmod;
  class gender(ref="F") oralhygiene(ref="med");
  model DMFTindex = oralhygiene gender / dist=zinb;
  zeromodel age;
run;

```

Log Likelihood -88.8603

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald	95% Confidence Limits	Wald Chi-Square	Pr > ChiSq
Intercept	1	1.2352	0.2762	0.6938	1.7766	20.00	<.0001
oralhygiene high	1	0.2468	0.3623	-0.4633	0.9569	0.46	0.4957
oralhygiene low	1	0.1577	0.4125	-0.6508	0.9662	0.15	0.7022
oralhygiene med	0	0.0000	0.0000	0.0000	0.0000	.	.
gender M	1	0.6655	0.3130	0.0521	1.2789	4.52	0.0335
gender F	0	0.0000	0.0000	0.0000	0.0000	.	.
Dispersion	1	0.4268	0.2007	0.1698	1.0729		

Analysis Of Maximum Likelihood Zero Inflation Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald	95% Confidence Limits	Wald Chi-Square	Pr > ChiSq
Intercept	1	2.1073	1.4259	-0.6874	4.9019	2.18	0.1394
age	1	-0.1013	0.0507	-0.2007	-0.0018	3.98	0.0460

The estimated parameters of the fitted model are  $\hat{\pi} = \frac{\exp(2.1073 - 0.1013 \cdot \text{age})}{1 + \exp(2.1073 - 0.1013 \cdot \text{age})}$ ,  $\hat{\lambda} = \exp(1.2352 + 0.2468 \cdot \text{high oral hygiene} + 0.1577 \cdot \text{low oral hygiene} + 0.6655 \cdot \text{male})$ , and  $\hat{r} = \frac{1}{0.4268} = 2.343018$ .

Age is a significant predictor of  $\pi$ , and gender is a significant predictor of  $\lambda$ .

In R:

```

dental.data<- read.csv(file="C:/./Exercise6.7Data.csv", header = TRUE, sep=",")

#specifying reference levels
gender.rel<- relevel(dental.data$gender, ref="F")
oralhygiene.rel<- relevel(dental.data$oralhygiene, ref="med")

#fitting zero-inflated negative binomial model

```

```
library(pscl)
summary(fitted.model<- zeroinfl(DMFTindex ~ gender.rel + oralhygiene.rel | age,
data=dental.data, dist="negbin"))
```

```
Count model coefficients (negbin with log link):
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    1.2352    0.2762   4.472 7.75e-06
gender.relM     0.6655    0.3130   2.126  0.0335
oralhygiene.relhigh 0.2468    0.3623   0.681  0.4957
oralhygiene.rellow 0.1577    0.4125   0.382  0.7022
```

```
Zero-inflation model coefficients (binomial with logit link):
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    2.10729    1.42547   1.478  0.1393
age           -0.10126    0.05073  -1.996  0.0459 *
```

Theta = 2.3429

(b) Analyze the fit of the model.

The p-value in the deviance test is less than 0.05, indicating a good fit of the model.

In SAS:

```
/*checking model fit*/
proc genmod;
  model DMFTindex = / dist=zinb;
  zeromodel;
run;
```

Log Likelihood -95.2421

```
data deviance;
  deviance = -2*(-95.2421 - (-88.8603));
  pvalue = 1 - probchi(deviance,4);
run;
```

```
proc print noobs;
run;
```

```
deviance    pvalue
 12.7636    0.012491
```

In R:

```
#checking model fit
null.model<- zeroinfl(DMFTindex ~ 1, data=dental.data, dist="negbin")
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))
```

12.76369

```
print(p.value<- pchisq(deviance, df=4, lower.tail = FALSE))
```

0.01249009

(c) Give an interpretation of the estimated significant coefficients.

As age increases by one year, the estimated odds in favor of the zero value of the DMFT index change by  $(\exp(-0.1013) - 1) \cdot 100\% = -9.63381\%$ , or decrease by 9.63%. The estimated mean value of the DMFT index for males is  $\exp(0.6655) \cdot 100\% = 194.5463\%$  of that for females.

(d) Find the predicted value of the DMFT index for a man, aged 28, with a high level of oral hygiene.

The predicted value is computed as follows:

$$DMFT\ index^0 = \frac{\exp(1.2352 + 0.2468 + 0.6655)}{1 + \exp(2.1073 - 0.1013 \cdot 28)} = 5.776951.$$

In SAS:

```
/*using fitted model for prediction*/
data predict;
input age gender$ oralhygiene$;
cards;
28 M high
;

data dental;
set dental predict;
run;

proc genmod;
  class gender(ref="F") oralhygiene(ref="med");
  model DMFTindex = oralhygiene gender / dist=zinb;
  zeromodel age;
  output out=outdata p=pDMFTindex;
run;

proc print data=outdata (firstobs=39) noobs;
  var pDMFTindex;
run;

pDMFTindex
  5.77492
```

In R:

```
#using fitted model for prediction
print(predict(fitted.model, data.frame(age=28, gender.rel="M",
oralhygiene.rel="high")))
```

5.774924

**EXERCISE 6.8.** (a) Fit a hurdle negative binomial regression to model the number of claims made in the past five years. Model the probability of zero claims as a function of the number of claims made in the previous five years. Model the positive responses as related to age and gender. Write down the fitted model explicitly. What predictors are significant at the 5% level?

In SAS:

```

data insurance;
input nclaimspast5ys nclaimsprev5ys age gender$ @@;
cards;
1 1 39 M 1 2 66 M 7 0 56 M 3 4 43 F 4 1 42 F 4 2 52 M 0 0 39 F
4 6 68 M 6 1 41 F 0 1 54 F 4 2 50 F 6 4 57 M 5 4 47 F 1 2 43 M
1 1 36 M 1 2 55 F 5 5 57 F 8 5 53 M 0 1 72 M 0 1 67 F 8 3 69 F
0 2 70 M 7 2 70 M 3 1 54 F 2 1 38 M 3 1 50 F 0 1 62 M 8 2 54 M
0 0 59 M 0 1 61 F 0 0 69 F 8 3 57 F 0 0 57 M 12 5 72 F 0 2 42 M
6 2 42 F 7 2 66 M 7 4 53 M 6 0 52 M 3 3 57 F
;

/*fitting hurdle negative binomial model*/
proc fmm;
class gender;
model nclaimspast5ys = age gender / dist=truncnegbin;
model+ / dist=constant;
probmodel nclaimsprev5ys;
run;

```

-2 Log Likelihood 163.4

Parameter Estimates for Truncated Negative Binomial Model					
Effect	gender	Estimate	Standard Error	z Value	Pr >  z
Intercept		-0.1879	0.5726	-0.33	0.7428
age		0.03093	0.009793	3.16	0.0016
gender	F	0.1253	0.1936	0.65	0.5173
gender	M	0	.	.	.
Scale Parameter		0.04455	0.08328		

Parameter Estimates for Mixing Probabilities				
Effect	Estimate	Standard Error	z Value	Pr >  z
Intercept	-0.8814	0.7050	-1.25	0.2112
nclaimsprev5ys	1.2816	0.5071	2.53	0.0115

The estimated parameters of the fitted model are  $\hat{\pi} = \frac{\exp(0.8814 - 1.2816 \cdot \text{nclaimsprev5ys})}{1 + \exp(0.8814 - 1.2816 \cdot \text{nclaimsprev5ys})}$ ,

$\hat{\lambda} = \exp(-0.1879 + 0.03093 \cdot \text{age} + 0.1253 \cdot \text{female})$ , and  $\hat{r} = \frac{1}{0.04455} = 22.44669$ .

The number of claims in the previous five years is a significant predictor of  $\pi$ , while age is a significant predictor of  $\lambda$ .

In R:

```

insurance.data<- read.csv(file="C:/../Exercise6.6Data.csv", header=TRUE, sep=",")

#specifying reference level
gender.rel<- relevel(insurance.data$gender, ref="M")

#fitting hurdle negative binomial model
library(pscl)
summary(fitted.model<- hurdle(nclaimspast5ys ~ age + gender.rel | nclaimsprev5ys,
data=insurance.data, dist="negbin", zero.dist = "binomial", link="logit"))

```

Count model coefficients (truncated negbin with log link):

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-0.187878	0.573901	-0.327	0.74339
age	0.030928	0.009815	3.151	0.00163
gender.ref	0.125341	0.193554	0.648	0.51726

Zero hurdle model coefficients (binomial with logit link):

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-0.8814	0.7050	-1.250	0.2112
nclaimsprev5ys	1.2816	0.5071	2.527	0.0115

Theta: count = 22.4472

(b) Discuss the goodness-of-fit of the model.

The model has a good fit which is supported by the small p-value in the deviance test.

In SAS:

```
/*checking model fit*/
proc fmm;
  model nclaimspast5ys = / dist=truncnegbin;
  model+ / dist=constant;
  probmodel;
run;
```

-2 Log Likelihood 184.3

```
data deviance;
  deviance = 184.3 - 163.4;
  pvalue = 1 - probchi(deviance,3);
run;
```

```
proc print noobs;
run;
```

deviance	pvalue
20.9	.000110432

In R:

```
#checking model fit
null.model<- hurdle(nclaimspast5ys ~ 1, data=insurance.data, dist="negbin",
  zero.dist="binomial", link="logit")
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))
```

20.90545

```
print(p.value<- pchisq(deviance, df=3, lower.tail = FALSE))
```

0.0001101446

(d) Interpret the estimated significant coefficients. What is the direction of the relationships?

As the number of claims made in the previous 5 years increases by 1, the estimated odds in favor of the structural absence of claims change by  $(\exp(-1.2816) - 1) \cdot 100\% = -72.24\%$ , that is, decrease by 72.24%. No interpretation of estimated regression coefficients is possible for  $\hat{\lambda}$  in hurdle models.

(e) Find the predicted number of claims made in the past five years by a 55-year-old female policyholder who has made no claims in the previous five years.

The predicted value is calculated as follows:

$$claims^0 = \frac{(1 + \exp(0.8814))^{-1} \exp(-0.1879 + 0.03093 \cdot 55 + 0.1253)}{1 - (1 + \exp(-0.1879 + 0.03093 \cdot 55 + 0.1253)/22.44669)^{-22.44669}} = 1.522482.$$

In SAS:

```
/*using fitted model for prediction*/
data predict;
input nclaimsprev5ys age gender$;
cards;
0 55 F
;

data insurance;
set insurance predict;
run;

proc fmm;
class gender;
model nclaimspast5ys = age gender / dist=truncnegbin;
model+ / dist=constant;
probmodel nclaimsprev5ys;
output out=outdata pred=pnclaimspast5ys;
run;

proc print data=outdata (firstobs=41) noobs;
var pnclaimspast5ys;
run;

pnclaimspast5ys
1.52243
```

In R:

```
#using fitted model for prediction
print(predict(fitted.model, data.frame(nclaimsprev5ys=0, age=55,
gender.rel="F")))
```

1.522423

**EXERCISE 6.9.** (a) Fit a hurdle negative binomial regression model. Specify the fitted model. Does it support the researchers' hypotheses? Discuss the significance of predictor variables at the 5% significance level.

In SAS:

```
data sportsmedicine;
input ngameinjuries gender$ nsports npracticeinjuries @@;
cards;
0 M 2 2 0 M 1 1 1 F 2 3 1 F 1 0 2 F 1 1 0 F 2 1 0 M 1 0
6 M 2 3 7 M 1 5 2 M 2 4 8 M 3 1 10 M 2 2 4 M 1 7 0 F 1 1
2 M 2 2 2 F 2 0 0 F 2 0 1 F 2 2 2 M 2 3 0 M 1 0 3 M 2 4
5 F 1 4 0 M 2 0 7 M 3 4 7 F 2 3 3 F 3 4 8 M 1 2 3 F 3 5
7 M 3 6 12 M 2 5
;

proc format;
value $genderfmt "F"="ref";
run;

/*fitting hurdle negative binomial model*/
proc fmm;
class gender;
model ngameinjuries = gender nsports / dist=truncnegbin;
model+ / dist=constant;
probmodel npracticeinjuries;
format gender $genderfmt.;
run;
```

-2 Log Likelihood 118.9

#### Parameter Estimates for Truncated Negative Binomial Model

Effect	gender	Estimate	Standard Error	z Value	Pr >  z
Intercept		0.7247	0.4761	1.52	0.1280
gender	M	0.8873	0.3251	2.73	0.0063
gender	ref	0	.	.	.
nsports		0.08047	0.1985	0.41	0.6852
Scale Parameter		0.1713	0.1530		

#### Parameter Estimates for Mixing Probabilities

Effect	Estimate	Standard Error	z Value	Pr >  z
Intercept	-1.1574	0.8030	-1.44	0.1495
npracticeinjuries	1.3498	0.5401	2.50	0.0124

The estimated parameters of the fitted hurdle model are  $\hat{\pi} = \frac{\exp(1.1574 - 1.3498 \cdot npracticeinjuries)}{1 + \exp(1.1574 - 1.3498 \cdot npracticeinjuries)}$ ,  $\hat{\lambda} = \exp(0.7247 + 0.8873 \cdot male + 0.08047 \cdot nsports)$ , and  $\hat{r} = \frac{1}{0.1713} = 5.837712$ .

The number of injuries during practice is a significant predictor of  $\pi$ , whereas gender is a significant predictor of  $\lambda$ . The slope for the number of injuries during practices is negative, indicating the relationship opposite to the hypothesized one, however, the estimated positive coefficient for the number of sports supports the second hypothesis that a larger number of sports accounts for a larger number of injuries.

In R:

```
sportsmedicine.data<- read.csv(file="C:/./Exercise6.9Data.csv", header=TRUE,
sep=",")
```

```
library(pscl)
```

```
#specifying reference level
```

```
gender.rel<- relevel(sportsmedicine.data$gender, ref="F")
```

```
#fitting hurdle negative binomial model
```

```
summary(fitted.model<- hurdle(ngameinjuries ~ gender.rel +
nsports|npracticeinjuries, data=sportsmedicine.data, dist="negbin",
zero.dist="binomial", link="logit"))
```

```
Count model coefficients (truncated negbin with log link):
```

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	0.72465	0.47606	1.522	0.12796
gender.relM	0.88730	0.32510	2.729	0.00635
nsports	0.08047	0.19854	0.405	0.68525

```
Zero hurdle model coefficients (binomial with logit link):
```

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-1.1574	0.8030	-1.441	0.1495
npracticeinjuries	1.3498	0.5401	2.499	0.0124

```
Theta: count = 5.8374
```

(b) Analyze the model fit.

The p-value in the deviance test is very small which means that the fit is very good.

In SAS:

```
/*checking model fit*/
```

```
proc fmm;
```

```
model ngameinjuries = / dist=truncnegbin;
```

```
model+ / dist=constant;
```

```
probmodel;
```

```
run;
```

**-2 Log Likelihood 140.2**

```
data deviance;
```

```
deviance = 140.2 - 118.9;
```

```
pvalue = 1 - probchi(deviance,3);
```

```
run;
```

```
proc print noobs;
```

```
run;
```

deviance	pvalue
21.3	.000091203

In R:

```
#checking model fit
```

```
null.model<- hurdle(ngameinjuries ~ 1, data=sportsmedicine.data, dist="negbin",
zero.dist="binomial", link="logit")
```

```
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))
```

21.23368

```
print(p.value<- pchisq(deviance, df=3, lower.tail = FALSE))
```

9.414351e-05

(c) Give an interpretation of the estimated significant coefficients.

As the number of injuries during practices increases by one, the estimated odds in favor of zero injuries during games change by  $(\exp(-1.3498) - 1) \cdot 100\% = -74.07\%$ , that is, decrease by 74.07%. Interpretation of estimated regression coefficients in  $\hat{\lambda}$  in the hurdle negative binomial model is traditionally omitted.

(d) Calculate the predicted number of injuries for a male athlete who throughout his college years has participated in two sports, and who has received one minor injury during practice games.

$$ninjuries^0 = \frac{(1 + \exp(1.1574 - 1.3498))^{-1} \exp(0.7247 + 0.8873 + 0.08047 \cdot 2)}{1 - (1 + \exp(0.7247 + 0.8873 + 0.08047 \cdot 2) / 5.837712)^{-5.837712}} = 3.282386.$$

In SAS:

```
/*using fitted model for prediction*/
data predict;
input gender$ nsports npracticeinjuries;
cards;
M 2 1
;

data sportsmedicine;
set sportsmedicine predict;
run;

proc fmm;
class gender;
model ngameinjuries = gender nsports / dist=truncnegbin;
model+ / dist=constant;
probmodel npracticeinjuries;
output out=outdata pred=pngameinjuries;
format gender $genderfmt.;
run;

proc print data=outdata (firstobs=31) noobs;
var pngameinjuries;
run;

pngameinjuries
      3.28221
```

In R:

```
#using fitted model for prediction
print(predict(fitted.model, data.frame(gender.rel="M", nsports=2,
npracticeinjuries=1)))

3.282217
```

# CHAPTER 7

## EXERCISE 7.1.

$$\begin{aligned}
 E(y) &= \int_0^1 \frac{y^{\mu\phi}(1-y)^{(1-\mu)\phi-1}}{B(\mu\phi, (1-\mu)\phi)} dy = \frac{B(\mu\phi+1, (1-\mu)\phi)}{B(\mu\phi, (1-\mu)\phi)} \\
 &= \frac{\Gamma(\mu\phi+1)\Gamma((1-\mu)\phi)}{\Gamma(\phi+1)} \cdot \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} = \frac{\mu\phi}{\phi} = \mu. \\
 Var(y) &= \int_0^1 \frac{y^{\mu\phi+1}(1-y)^{(1-\mu)\phi-1}}{B(\mu\phi, (1-\mu)\phi)} dy - \mu^2 = \frac{B(\mu\phi+2, (1-\mu)\phi)}{B(\mu\phi, (1-\mu)\phi)} - \mu^2 \\
 &= \frac{\Gamma(\mu\phi+2)\Gamma((1-\mu)\phi)}{\Gamma(\phi+2)} \cdot \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} - \mu^2 = \frac{(\mu\phi+1)\mu\phi}{(\phi+1)\phi} - \mu^2 = \frac{\mu(1-\mu)}{1+\phi}.
 \end{aligned}$$

**EXERCISE 7.2.** (a) Model the proportion of birds per flock that successfully reach the winter grounds. To avoid small estimates of the regression coefficients, convert mass into kilograms, wingspan into meters, and distance, into thousands of kilometers. Write out the fitted model explicitly. Which predictors are significant at the 5% level?

In SAS:

```

data birdsmigration;
input mass wingspan distance nringed nmigrated @@;
cards;
811 67 1680 70 8 261 33 2137 113 75 398 48 2159 100 51
114 56 1204 145 113 119 53 1673 72 28 151 30 543 87 71
176 70 1414 116 109 184 45 2296 90 68 250 42 1511 52 42
505 24 741 74 63 551 17 1434 114 105 716 51 2116 98 58
735 119 2171 98 35 1233 108 2442 69 13 1315 98 2061 61 38
1633 72 1955 81 24 1736 119 1297 71 70 2019 101 930 112 105
2476 100 2312 95 37
;

/*rescaling predictors and computing response*/
data birdsmigration;
set birdsmigration;
mass=mass/1000;
wingspan=wingspan/100;
distance=distance/1000;
propsuccess=nmigrated/nringed;
run;

/*fitting beta regression model*/
proc glimmix;
model propsuccess = mass wingspan distance / dist=beta link=logit solution;
run;

```

-2 Log Likelihood -12.81

Parameter Estimates					
Effect	Estimate	Standard Error	DF	t Value	Pr >  t
Intercept	2.7830	0.8642	15	3.22	0.0057
mass	-0.01502	0.4027	15	-0.04	0.9707
wingspan	0.04487	0.8796	15	0.05	0.9600
distance	-1.3186	0.4368	15	-3.02	0.0086
Scale	4.1727	1.2775	.	.	.

The fitted beta regression model has the estimated parameters

$$\hat{\mu} = \frac{\exp(2.7830 - 0.01502 \cdot \text{mass} + 0.04487 \cdot \text{wingspan} - 1.3186 \cdot \text{distance})}{1 + \exp(2.7830 - 0.01502 \cdot \text{mass} + 0.04487 \cdot \text{wingspan} - 1.3186 \cdot \text{distance})}, \text{ and } \hat{\phi} = 4.1727.$$

Only distance is a significant predictor at the 5% level.

In R:

```
birdsmigration.data<- read.csv(file="C:/../Exercise7.2Data.csv", header=TRUE,
sep=",")

#rescaling predictors and computing response
mass<- birdsmigration.data$mass/1000
wingspan<- birdsmigration.data$wingspan/100
distance<- birdsmigration.data$distance/1000
propsuccess<- birdsmigration.data$nmigrated/birdsmigration.data$nringed

#fitting beta model
library(betareg)
summary(fitted.model<- betareg(propsuccess ~ mass + wingspan + distance,
link="logit"))

Coefficients (mean model with logit link):
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.78304    0.78201   3.559 0.000373
mass         -0.01502    0.41088  -0.037 0.970848
wingspan      0.04487    0.93935   0.048 0.961901
distance     -1.31857    0.42237  -3.122 0.001797

(phi)      4.173
```

(b) Analyze the model fit.

The model has a reasonably good fit since the p-value < 0.05.

In SAS:

```
/*checking model fit*/
proc glimmix;
  model propsuccess = / dist=beta link=logit;
run;
```

2 Log Likelihood -4.25

```
data deviance;
  deviance = -4.25 - (-12.81);
  pvalue = 1 - probchi(deviance, 3);
run;
```

```
proc print noobs;
run;
```

```
deviance    pvalue
      8.56    0.035751
```

In R:

```
#checking model fit
null.model<- betareg(propsuccess ~ 1, link="logit")
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))
```

```
8.55842
```

```
print(p.value<- pchisq(deviance, df=3, lower.tail = FALSE))
```

```
0.03577627
```

(c) Give an interpretation of the estimated significant parameters.

As the distance increases by one thousand miles, the estimated ratio of the mean proportion of successfully migrated birds and the mean proportion of those that didn't migrate successfully changes by  $(\exp(-1.3186) - 1) \cdot 100\% = -73.249\%$ , or decreases by 73.249%.

(d) Predict the number of birds that successfully reach the winter grounds for a flock of 70 birds with an average mass of 600 grams, an average wingspan of 65 centimeters, and that travel a distance of 1650 kilometers.

The predicted proportion is found as:

$$nbirds^0 = 70 \cdot \frac{\exp(2.7830 - 0.01502 \cdot 0.6 + 0.04487 \cdot 0.65 - 1.3186 \cdot 1.65)}{1 + \exp(2.7830 - 0.01502 \cdot 0.6 + 0.04487 \cdot 0.65 - 1.3186 \cdot 1.65)} = (70)(0.651914) = 45.63398.$$

Only distance is a significant predictor at the 5% level.

In SAS:

```
/*using fitted model for prediction*/
data predict;
input mass wingspan distance nringed;
cards;
0.6 0.65 1.65 70
;
```

```
data birdsmigration;
set birdsmigration predict;
run;
```

```
proc glimmix;
  model propsuccess = mass wingspan distance / dist=beta link=logit solution;
```

```

    output out=outdata pred(ilink)=ppropsuccess;
run;

data outdata;
  set outdata;
  pbirds = 70*ppropsuccess;
run;

proc print data=outdata (firstobs=20) noobs;
  var pbirds;
run;

```

**pbirds**

45.6355

In R:

```

#using fitted model for prediction
print(70*predict(fitted.model, data.frame(mass=.6, wingspan=.65, distance=1.65,
nringed=70)))

```

45.63551

**EXERCISE 7.3.** (a) Model the proportion of hospitalized ER patients. Write down the fitted model. What factors are significant predictors? Use  $\alpha = 0.05$ .

In SAS:

```

data hospitals;
input perc_hospitalized location $ type $ nbeds @@;
prophospitalized=perc_hospitalized/100;
cards;
17 rural private 56    39 rural public 144    38 urban public 61
48 rural public 186    30 rural private 132    25 urban private 589
5  urban public 53    4  rural private 73    48 rural private 154
4  urban public 38    26 rural private 318    15 urban public 35
28 urban private 184    34 urban private 173    31 urban public 63
4  urban public 91    6  urban public 77    39 urban private 237
41 urban private 56    45 rural public 43    13 urban public 64
42 rural public 193    28 urban private 363    31 urban public 600
48 rural public 468    41 rural public 311    9  urban public 65
13 urban private 44    44 urban public 479    16 rural public 72
;

/*fitting beta regression model*/
proc glimmix;
  class location type(ref="private");
  model prophospitalized = location type nbeds / dist=beta link=logit solution;
run;

```

-2 Log Likelihood -41.58

Parameter Estimates						
Effect	location type	Estimate	Standard Error	DF t Value	Pr >  t	
Intercept		-1.6735	0.2951	26	-5.67	<.0001
location rural		0.5633	0.2526	26	2.23	0.0346
location urban		0	.	.	.	.
type public		0.01165	0.2572	26	0.05	0.9642
type private		0	.	.	.	.
nbeds		0.002117	0.000707	26	3.00	0.0060
scale		9.8079	2.4611	.	.	.

The estimated parameters in the fitted model are

$$\hat{\mu} = \frac{\exp(-1.6735 + 0.5633 \cdot \text{rural} + 0.01165 \cdot \text{public} + 0.002117 \cdot \text{nbeds})}{1 + \exp(-1.6735 + 0.5633 \cdot \text{rural} + 0.01165 \cdot \text{public} + 0.002117 \cdot \text{nbeds})}, \text{ and } \hat{\phi} = 9.8079.$$

Locations and number of beds are significant predictors at the 5% significance level.

In R:

```
hospitals.data<- read.csv(file="C:/./Exercise7.3Data.csv", header=TRUE, sep=",")

#computing response and specifying reference levels
prophospitalized<- hospitals.data$perhospitalized/100
location.rel<- relevel(hospitals.data$location, ref="urban")
type.rel<- relevel(hospitals.data$type, ref="private")

#fitting beta regression model
library(betareg)
summary(fitted.model<- betareg(prophospitalized ~ location.rel
+ type.rel + nbeds, data=hospitals.data, link="logit"))

Coefficients (mean model with logit link):
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.6735254  0.2787233  -6.004 1.92e-09
location.relrural  0.5632707  0.2494671   2.258 0.02395
type.relpUBLIC    0.0116530  0.2511991   0.046 0.96300
nbeds           0.0021170  0.0007219   2.933 0.00336

(phi)      9.808
```

(b) How good is the model fit?

The p-value is below 0.05, hence the fit is good.

In SAS:

```
/*checking model fit*/
proc glimmix;
  model prophospitalized = / dist=beta link=logit;
run;
```

2 Log Likelihood -30.58

```
data deviance;
```

```

deviance = -30.58 - (-41.58);
pvalue = 1 - probchi(deviance,3);
run;

proc print noobs;
run;

deviance    pvalue
      11  0.011726

```

In R:

```

#checking model fit
null.model<- betareg(prophospitalized ~ 1, data=hospitals.data, link="logit")
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))

11.00244

print(p.value<- pchisq(deviance, df=3, lower.tail = FALSE))

0.01171267

```

(c) Interpret estimated significant regression coefficients.

For rural hospitals, the estimated ratio of the mean proportion of hospitalized ER patients and the mean proportion of non-hospitalized ones is  $\exp(0.5633) \cdot 100\% = 175.6459\%$  of that for urban hospitals. As the number of beds increases by one, this estimated ratio increases by  $(\exp(0.0002117) - 1) \cdot 100\% = 0.211924\%$ .

(d) Give the predicted proportion of hospitalized ER patients for a rural public hospital with 50 beds.

The predicted proportion is

$$\text{prophospitalized}^0 = \frac{\exp(-1.6735 + 0.5633 + 0.01165 + 0.002117 \cdot 50)}{1 + \exp(-1.6735 + 0.5633 + 0.01165 + 0.002117 \cdot 50)} = 0.270379.$$

In SAS:

```

/*using fitted model for prediction*/
data predict;
input location$ type$ nbeds;
cards;
rural public 50
;

data hospitals;
set hospitals predict;
run;

proc glimmix;
class location type
model prophospitalized = location type nbeds / dist=beta link=logit solution;
output out=outdata pred(ilink)= pprophospitalized;

```

```
run;

proc print data=outdata (firstobs=31) noobs;
  var pprophospitalized;
run;
```

```
pprophospitalized
      0.27037
```

In R:

```
#using fitted model for prediction
print(predict(fitted.model, data.frame(location.rel="rural", type.rel="public",
nbeds=50)))
```

```
0.270369
```

**EXERCISE 7.4.** (a) Use the beta regression to model the proportion of by-catch. Convert depth to kilometers. Write down the fitted model.

In SAS:

```
data fishing;
input distance method $ depth percbycatch @@;
depth=depth/1000;
propbycatch= percbycatch/100;
cards;
120 trawl      250 14  115 trawl      150 6   70  trawl      300 24
130 trawl      150 6   90  seine      200 56  15  seine      350 32
15  seine      150 13  20  seine      350 23  15  longline  200 10
40  longline   150 7   115 trawl      300 8   160 trawl      200 10
160 trawl      200 10  50  trawl      150 15  10  seine      150 16
25  seine      200 22  15  seine      300 21  40  longline  100 21
60  longline   200 4   50  longline  150 17
;

/*fitting beta regression model*/
proc glimmix;
  class method;
  model propbycatch = distance method depth / dist=beta link=logit solution;
run;
```

```
-2 Log Likelihood -49.20
```

		Parameter Estimates				
Effect	method	Estimate	Standard Error	DF t Value	Pr >  t	
Intercept		-3.0476	0.8723	15	-3.49	0.0033
distance		0.006016	0.005386	15	1.12	0.2815
method	longline	0.5096	0.5864	15	0.87	0.3985
method	seine	1.4661	0.6348	15	2.31	0.0356
method	trawl	0	.	.	.	.
depth		1.5862	1.8090	15	0.88	0.3944

Parameter Estimates					
Effect	method	Estimate	Standard Error	DF	t Value Pr >  t
Scale		22.2401	7.0741	.	. .

The estimated parameters in the fitted beta regression model are

$$\hat{\mu} = \frac{\exp(-3.0476 + 0.006016 \cdot \text{distance} + 0.5096 \cdot \text{longline} + 1.4661 \cdot \text{seine} + 1.5862 \cdot \text{depth})}{1 + \exp(-3.0476 + 0.006016 \cdot \text{distance} + 0.5096 \cdot \text{longline} + 1.4661 \cdot \text{seine} + 1.5862 \cdot \text{depth})}$$

and  $\hat{\phi} = 22.2401$ .

Fishing with a seine is the only significant predictor at the 5% significance level.

In R:

```
fishing.data<- read.csv(file="C:/./Exercise7.4Data.csv", header= TRUE, sep=",")

#computing response, rescaling and specifying reference level
propbycatch<- fishing.data$percbycatch/100
depthK<- fishing.data$depth/1000
method.rel<- releval(fishing.data$method, ref="trawl")
```

```
#fitting beta regression model
library(betareg)
summary(fitted.model<- betareg(propbycatch ~ distance + method.rel
+ depthK, data=fishing.data, link="logit"))
```

Coefficients (mean model with logit link):

	Estimate	Std. Error	z	value	Pr(> z )
(Intercept)	-3.047625	0.744117	-4.096	4.21e-05	
distance	0.006016	0.004383	1.373	0.1699	
method.rellongline	0.509570	0.505675	1.008	0.3136	
method.relseine	1.466084	0.480930	3.048	0.0023	
depthK	1.586194	1.827014	0.868	0.3853	

(phi) 22.240

(b) Discuss the significance of predictor variables and model fit.

The p-value is below 0.05, indicating a good fit.

In SAS:

```
/*checking model fit*/
proc glimmix;
  model propbycatch = / dist=beta link=logit;
run;
```

-2 Log Likelihood -38.25

```
data deviance;
  deviance = -38.25 - (-49.20);
  pvalue = 1 - probchi(deviance,4);
run;
proc print noobs;
run;
```

```
deviance    pvalue
10.95 0.027132
```

In R:

```
#checking model fit
null.model<- betareg(propbycatch ~ 1, data=fishing.data, link="logit")
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))
```

```
10.95303
```

```
print(p.value<- pchisq(deviance, df=4, lower.tail = FALSE))
```

```
0.02709695
```

(c) Give an interpretation of the estimates of the regression coefficients for the significant predictors.

For the purse seining method of fishing, the estimated ratio of the mean proportion of by-catch and the mean proportion of intended fish is  $\exp(1.4661) \cdot 100\% = 433.2306\%$  of that for the trawling method of fishing.

(d) Find the predicted percent of by-catch for a trawler that fishes 80 nautical miles offshore at a depth of 250 meters.

The predicted percent of by-catch can be computed as

$$\text{prop by-catch}^0 = \frac{\exp(-3.0476 + 0.006016 \cdot 80 + 1.5862 \cdot 0.25)}{1 + \exp(-3.0476 + 0.006016 \cdot 80 + 1.5862 \cdot 0.25)} = 0.102498, \text{ or } 10.2498\%.$$

In SAS:

```
/*using fitted model for prediction*/
data predict;
input distance method$ depth;
cards;
80 trawl 0.25
;

data fishing;
set fishing predict;
run;

proc glimmix;
class method;
model propbycatch = distance method depth / dist=beta link=logit;
output out=outdata pred(ilink)=ppropbycatch;
run;

proc print data=outdata (firstobs=21) noobs;
var ppropbycatch;
run;

ppropbycatch
0.10250
```

In R:

```
#using fitted model for prediction
print(predict(fitted.model, data.frame(distance=80, method.rel="trawl",
depthK=0.25)))
```

0.1024961

**EXERCISE 7.5.** (a) Fit the zero-inflated beta regression model to the proportion of houses. Regress the probability of zero on the age of the subdivision. To achieve model convergence, normalize the average price and number of houses by a factor of 100. Discuss the significance of the predictors at the 5% level.

In SAS:

```
data realestate;
input percsold avgprice nhouses age @@;
avgprice=avgprice/100;
nhouses=nhouses/100;
propsold=percsold/100;
cards;
0      455 69  21  0      316 244 24   36.4 210 236 31   50 557 183 16
33.3 232 73  6   50      626 230 20   27.3 343 60  14   80 246 201 17
42.9 631 217 11  0      630 222 42   71.4 356 85  22   25 481 240 16
50    181 197 42  20     264 235 19   87.5 297 88  17   80 308 223 15
75    159 84  13  0      147 54  37   44.4 704 199 18  0   593 119 38
20    738 156 8   55.6 256 206 34   85.7 345 38  22   50 450 158 7
0     491 239 27  28.6 441 103 15   88.9 212 222 18  50  574 56  35
33.3 647 138 35  0      630 18  60
;

/*fitting zero-inflated beta regression model*/
proc nlmixed
parms b0=0.1 b1=0.1 g0=0.1 g1=0.1 g2=0.1 phi=0.1;
  pi0=exp(b0+b1*age)/(1+exp(b0+b1*age));
  mu=exp(g0+g1*avgprice+g2*nhouses)/(1+exp(g0+g1*avgprice+g2*nhouses));
  if(propsold=0) then loglikelihood=log(pi0);
  else loglikelihood=log(1-pi0)+lgamma(phi)-lgamma(mu*phi)-lgamma((1-mu)*phi)+
(mu*phi-1)*log(propsold)+((1-mu)*phi-1)*log(1-propsold);
  model propsold ~ general(loglikelihood);
run;
```

-2 Log Likelihood 11.1

Parameter Estimates								
Parameter	Estimate	Standard Error	DF	t Value	Pr >  t	95% Confidence Limits	Gradient	
b0	-4.5161	1.5570	30	-2.90	0.0069	-7.6959 -1.3363	4.25E-6	
b1	0.1247	0.05093	30	2.45	0.0204	0.02068 0.2287	0.000209	
g0	1.1639	0.5537	30	2.10	0.0441	0.03307 2.2948	5.63E-6	
g1	-0.2002	0.09321	30	-2.15	0.0399	-0.3906 -0.00989	9.205E-6	
g2	-0.1679	0.2442	30	-0.69	0.4970	-0.6667 0.3308	-5.58E-6	

Parameter Estimates							
Parameter	Estimate	Standard Error	DF	t Value	Pr >  t	95% Confidence Limits	Gradient
phi	5.4911	1.4991	30	3.66	0.0010	2.4294 8.5527	-2.61E-6

In the fitted model, the estimated parameters are

$$\hat{\pi}_0 = \frac{\exp(-4.5161 + 0.1247 \cdot \text{age})}{1 + \exp(-4.5161 + 0.1247 \cdot \text{age})}, \quad \hat{\mu} = \frac{\exp(1.1639 - 0.2002 \cdot \text{avgprice} - 0.1679 \cdot \text{nhouses})}{1 + \exp(1.1639 - 0.2002 \cdot \text{avgprice} - 0.1679 \cdot \text{nhouses})}, \text{ and}$$

$$\hat{\phi} = 5.4911.$$

Age is a significant predictor of  $\pi_0$ , and average house price is a significant predictor of  $\mu$ .

In R:

```
realestate.data<- read.csv(file="C:/./Exercise7.5Data.csv", header= TRUE,
sep=",")

#computing response and rescaling predictors
realestate.data$propsold<- realestate.data$percsold/100
realestate.data$avgprice.res<- realestate.data$avgprice/100
realestate.data$nhouses.res<- realestate.data$nhouses/100

#fitting zero-inflated beta regression model
library(gamlss)
summary(fitted.model<- gamlss(propsold ~ avgprice.res + nhouses.res,
mu.link="logit", nu.formula = ~ age, nu.link="logit", data=realestate.data,
family = BEZI))
```

Mu Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	1.16393	0.55373	2.102	0.0462
avgprice.res	-0.20024	0.09321	-2.148	0.0420
nhouses.res	-0.16791	0.24421	-0.688	0.4983

Sigma Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	1.703	0.273	6.238	1.9e-06

Nu Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-4.51609	1.55696	-2.901	0.00785
age	0.12468	0.05093	2.448	0.02205

(b) Present the fitted model. Does this model have a decent fit?

The model has a good fit since the p-value is small.

In SAS:

```
/*checking model fit*/
proc nlmixed;
parms b0=0.1 g0=0.1 phi=0.1;
pi0=exp(b0)/(1+exp(b0));
mu=exp(g0)/(1+exp(g0));
if(propsold=0) then loglikelihood=log(pi0);
else loglikelihood=log(1-pi0)+lgamma(phi)-lgamma(mu*phi)-lgamma((1-mu)*phi)+
```

```
(mu*phi-1)*log(propsold)+((1-mu)*phi-1)*log(1-propsold);
model propsold ~ general(loglikelihood);
run;
```

**-2 Log Likelihood 25.2**

```
data deviance;
  deviance = 25.2 - 11.1;
  pvalue = 1 - probchi(deviance,3);
run;
```

```
proc print noobs;
run;
```

```
deviance      pvalue
      14.1      .002772148
```

In R:

```
#checking model fit
null.model<- gamlss(propsold ~ 1, mu.link="logit", nu.formula= ~ 1,
nu.link="logit", data=realestate.data, family=BEZI)
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))
```

**14.12085**

```
print(p.value<- pchisq(deviance, df=3, lower.tail = FALSE))
```

**0.002745187**

(c) Interpret parameter estimates for statistically significant predictors.

As the age of subdivision increases by one year, the odds in favor of zero houses sold increase by  $(\exp(0.1247) - 1) \cdot 100\% = 13.28086\%$ . As the average price of houses in a subdivision increases by one hundred thousand dollars, the ratio  $\frac{\hat{\mu}}{1-\hat{\mu}}$  changes by  $(\exp(-0.2002) - 1) \cdot 100\% = -18.1433\%$ , that is, decreases by 18.1433%.

(d) What is the model prediction for the percent houses sold for a subdivision with 300 houses, built 50 years ago, and where houses are sold, on average, for \$450,000?

The predicted value is

$propsold^0 = (1 + \exp(-4.5161 + 0.1247 \cdot 50))^{-1} \cdot \frac{\exp(1.1639 - 0.2002 \cdot 4.5 - 0.1679 \cdot 3)}{1 + \exp(1.1639 - 0.2002 \cdot 4.5 - 0.1679 \cdot 3)} = 0.066903$ ,  
or 6.6903%.

In SAS:

```
/*using fitted model for prediction*/
data predict;
input avgprice nhouses age;
cards;
4.5 3 50
;
```

```

data realestate;
set realestate predict;
run;

proc nlmixed;
parms b0=0.1 b1=0.1 g0=0.1 g1=0.1 g2=0.1 phi=0.1;
pi0=exp(b0+b1*age)/(1+exp(b0+b1*age));
mu=exp(g0+g1*avgprice+g2*nhouses)/(1+exp(g0+g1*avgprice+g2*nhouses));
if(propsold=0) then loglikelihood=log(pi0);
else loglikelihood=log(1-pi0)+lgamma(phi)-lgamma(mu*phi)-lgamma((1-mu)*phi)+
(mu*phi-1)*log(propsold)+((1-mu)*phi-1)*log(1-propsold);
model propsold ~ general(loglikelihood);
predict (1-pi0)*mu out=outdata;
run;

proc print data=outdata (firstobs=31) noobs;
var Pred;
run;

    Pred
0.066949

```

In R:

```

#using fitted model for prediction
param.pred<- predictAll(fitted.model, newdata=data.frame(avgprice.res=4.5,
nhouses.res=3, age=50), type="response")
print((1-param.pred$nu)*param.pred$mu)

0.06694997

```

**EXERCISE 7.6.** (a) Model the proportion of first-place trophies using a zero-inflated beta regression. Use the number of pupils to predict the probability of zero. Specify the fitted model. Use an alpha of 0.05 to determine the significance of regression coefficients.

In SAS:

```

data martialarts;
input ntrophies nfirstplaces nyears nblackbelts npupils @@;
propfirst=nfirstplaces/ntrophies;
cards;
21 7 5 1 96 12 3 5 2 59 21 10 5 2 71 23 4 3 2 94 11 1 1 3 53
20 9 6 4 52 15 4 6 2 61 28 16 13 5 104 19 8 3 4 95 4 0 1 1 27
6 0 1 1 45 19 12 7 5 42 21 7 4 3 86 32 24 11 6 151 5 0 3 1 78
23 9 5 2 81 8 0 3 2 35 21 13 15 3 89 12 3 6 3 39 11 0 3 2 40
12 7 5 2 81 22 13 7 4 148 10 3 8 3 128 20 0 2 2 42 19 2 3 1 39
14 2 2 3 105
;

/*fitting zero-inflated beta regression model*/
proc nlmixed;
parms b0=0.1 b1=0.1 g0=0.1 g1=0.1 g2=0.1 phi=0.1;
pi0=exp(b0+b1*npupils)/(1+exp(b0+b1*npupils));
mu=exp(g0+g1*nyears+g2*nblackbelts)/(1+exp(g0+g1*nyears+g2*nblackbelts));

```

```

if(propfirst=0) then loglikelihood=log(pi0);
else loglikelihood=log(1-pi0)+lgamma(phi)-lgamma(mu*phi)-lgamma((1-mu)*phi)
+(mu*phi-1)*log(propfirst)+((1-mu)*phi-1)*log(1-propfirst);
model propfirst ~ general(loglikelihood);
run;

```

-2 Log Likelihood -11.6

Parameter Estimates								
Parameter	Estimate	Standard Error	DF	t Value	Pr >  t	95% Confidence Limits	Gradient	
b0	2.9956	1.7966	26	1.67	0.1074	-0.6973 6.6885	5.469E-7	
b1	-0.07133	0.03457	26	-2.06	0.0492	-0.1424 -0.00027	-0.00074	
g0	-1.9445	0.3233	26	-6.01	<.0001	-2.6091 -1.2798	0.000089	
g1	0.1274	0.03938	26	3.23	0.0033	0.04643 0.2083	0.000318	
g2	0.2268	0.1060	26	2.14	0.0420	0.008854 0.4447	0.000364	
phi	14.9642	4.6039	26	3.25	0.0032	5.5008 24.4277	-1.22E-6	

In the fitted model, the estimated parameters are

$$\hat{\pi}_0 = \frac{\exp(2.9956 - 0.07133 \cdot npupils)}{1 + \exp(2.9956 - 0.07133 \cdot npupils)}, \quad \hat{\mu} = \frac{\exp(-1.9445 + 0.1274 \cdot nyears + 0.2268 \cdot nblackbelts)}{1 + \exp(-1.9445 + 0.1274 \cdot nyears + 0.2268 \cdot nblackbelts)},$$

and  $\hat{\phi} = 14.9642$ .

All predictors are significant at the 5% level. The number of pupils is a significant predictor of  $\pi_0$ , and both the number of years and number of blackbelt instructors are significant predictors of  $\mu$ .

In R:

```

martialarts.data<- read.csv(file="C:/./Exercise7.6Data.csv", header=TRUE,
sep=",")

#computing the response variable
martialarts.data$propfirst<-
martialarts.data$nfirfirstplaces/martialarts.data$ntrophies

#fitting zero-inflated beta regression model
library(gamlss)
summary(fitted.model<- gamlss(propfirst ~ nyears + nblackbelts, mu.link="logit",
nu.formula = ~ npupils, nu.link="logit", data=martialarts.data, family=BEZI))

```

Mu Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-1.94446	0.32335	-6.014	7.03e-06
nyears	0.12740	0.03938	3.235	0.00415
nblackbelts	0.22669	0.10602	2.138	0.04502

Sigma Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	2.7057	0.3077	8.795	2.62e-08

Nu Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	2.99552	1.79441	1.669	0.111
npupils	-0.07133	0.03452	-2.066	0.052

(b) Discuss the model fit. Present the fitted model.

The p-value is tiny indicating a very good model fit.

In SAS:

```
/*checking model fit*/
proc nlmixed;
parms b0=0.1 g0=0.1 phi=0.1;
pi0=exp(b0)/(1+exp(b0));
mu=exp(g0)/(1+exp(g0));
if(propfirst=0) then loglikelihood=log(pi0);
else loglikelihood=log(1-pi0)+lgamma(phi)-lgamma(mu*phi)-lgamma((1-mu)*phi)
+(mu*phi-1)*log(propfirst)+((1-mu)*phi-1)*log(1-propfirst);
model propfirst ~ general(loglikelihood);
run;
```

**-2 Log Likelihood 15.8**

```
data deviance;
  deviance = 15.8 - (-11.6);
  pvalue = 1 - probchi(deviance,3);
run;

proc print noobs;
run;
```

deviance	pvalue
27.4	.000004853

In R:

```
#checking model fit
null.model<- gamlss(propfirst ~ 1, mu.link="logit", nu.formula = ~ 1,
nu.link="logit", data=martialarts.data, family=BEZI)
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))
```

**27.38556**

```
print(p.value<- pchisq(deviance, df=3, lower.tail = FALSE))
```

**4.887314e-06**

(c) Interpret the estimates of significant regression coefficients.

As the number of pupils increases by one, the estimated odds in favor of no first-place trophies change by  $(\exp(-0.07133) - 1) \cdot 100\% = -6.88454\%$ , that is, decrease by 6.88454%. As the studio's age increases by one year, the ratio  $\frac{\hat{\mu}}{1-\hat{\mu}}$  increases by  $(\exp(0.1274) - 1) \cdot 100\% = 13.58713\%$ . As the number of blackbelt instructors increases by one, the ratio  $\frac{\hat{\mu}}{1-\hat{\mu}}$  increases by  $(\exp(0.2268) - 1) \cdot 100\% = 25.45789\%$ .

(d) Predict the proportion of first-place trophies won by a studio that has been around for 10 years, has 85 students, and three black-belt instructors.

The predicted proportion is

$$\text{prop}^0 = (1 + \exp(2.9956 - 0.07133 \cdot 85))^{-1} \cdot \frac{\exp(-1.9445 + 0.1274 \cdot 10 + 0.2268 \cdot 3)}{1 + \exp(-1.9445 + 0.1274 \cdot 10 + 0.2268 \cdot 3)} = 0.48013.$$

In SAS:

```
/*using fitted model for prediction*/
data predict;
input nyears nblackbelts npupils;
cards;
10 3 85
;

data martialarts;
set martialarts predict;
run;

proc nlmixed;
parms b0=0.1 b1=0.1 g0=0.1 g1=0.1 g2=0.1 phi=0.1;
pi0=exp(b0+b1*npupils)/(1+exp(b0+b1*npupils));
mu=exp(g0+g1*nyears+g2*nblackbelts)/(1+exp(g0+g1*nyears+g2*nblackbelts));
if(propfirst=0) then loglikelihood=log(pi0);
else loglikelihood=log(1-pi0)+lgamma(phi)-lgamma(mu*phi)-lgamma((1-mu)*phi)
+(mu*phi-1)*log(propfirst)+((1-mu)*phi-1)*log(1-propfirst);
model propfirst ~ general(loglikelihood);
predict(1-pi0)*mu out=outdata;
run;

proc print data=outdata (firstobs=27) noobs;
var Pred;
run;
```

**Pred**  
**0.48004**

In R:

```
#using fitted model for prediction
param.pred<- predictAll(fitted.model, newdata=data.frame(nyears=10,
nblackbelts=3, npupils=85), type="response")
print((1-param.pred$nu)*param.pred$mu)
```

**0.4800583**

**EXERCISE 7.7.** (a) Fit a one-inflated beta regression to model the proportion of survived trees. Regress the probability of one on the amount of precipitation and wind speed. Discuss the significance of the predictors at 5% and 10% significance levels.

In SAS:

```
data trees;
input nplanted nsurvived pestcontrol fertilization precipitation windspeed @@;
propsurvived=nsurvived/nplanted;
cards;
125 125 3 1 18 9.6 115 68 0 0 8 13.4 250 101 1 1 17 12.8
95 85 2 2 22 10 140 48 3 1 15 15.1 75 75 3 2 27 6.3
```

```

185 163 3 3 15 12.3 20 9 3 0 18 9.4 110 83 3 1 24 13.1
80 80 0 1 18 7.8 120 117 4 1 20 9.3 90 56 5 1 15 13.9
30 30 3 0 33 8.6 90 81 4 1 23 7.7 140 119 3 1 18 11.8
70 9 3 0 32 8.4 75 71 3 3 20 13.4 150 102 5 0 16 9.7
90 73 4 1 15 9.7 160 151 6 1 18 7.8 100 46 3 1 20 12.3
85 85 4 1 22 6.8 120 85 2 1 19 6.6 180 53 3 1 29 9.4
45 12 0 1 9 13.1 35 35 1 0 7 9.4
;

/*fitting one-inflated beta regression model*/
proc nlmixed;
parms b0=0.1 b1=0.1 b2=0.1 g0=0.1 g1=0.1 g2=0.1 phi=0.1;
pi1=exp(b0+b1*precipitation+b2*windspeed)/(1+exp(b0+b1*precipitation+b2*windspeed));
mu=exp(g0+g1*pestcontrol+g2*fertilization)/(1+exp(g0+g1*pestcontrol+g2*fertilization));
if (propsurvived=1) then loglikelihood=log(pi1);
else loglikelihood=log(1-pi1)+lgamma(phi)-lgamma(mu*phi)-lgamma((1-mu)*phi)
+ (mu*phi-1)*log(propsurvived)+((1-mu)*phi-1)*log(1-propsurvived);
model propsurvived ~ general(loglikelihood);
run;

```

## 2 Log Likelihood 3.8

Parameter Estimates								
Parameter	Estimate	Standard Error	DF	t Value	Pr >  t	95% Confidence Limits	Gradient	
b0	6.9100	4.2974	26	1.61	0.1199	-1.9235 15.7435	-3.25E-7	
b1	-0.04206	0.09075	26	-0.46	0.6469	-0.2286 0.1445	-0.00003	
b2	-0.7822	0.3775	26	-2.07	0.0483	-1.5582 -0.00632	-3.3E-6	
g0	-1.1091	0.5643	26	-1.97	0.0601	-2.2690 0.05083	0.000013	
g1	0.3234	0.1364	26	2.37	0.0254	0.04312 0.6038	0.000035	
g2	0.7332	0.2822	26	2.60	0.0152	0.1532 1.3133	0.000010	
phi	4.9777	1.5005	26	3.32	0.0027	1.8934 8.0620	-7.19E-7	

In the fitted model, the estimated parameters are

$$\hat{\pi}_1 = \frac{\exp(6.91 - 0.04206 \cdot \text{precipitation} - 0.7822 \cdot \text{windspeed})}{1 + \exp(6.91 - 0.04206 \cdot \text{precipitation} - 0.7822 \cdot \text{windspeed})}$$

$$\hat{\mu} = \frac{\exp(-1.1091 + 0.3234 \cdot \text{pestcontrol} + 0.7332 \cdot \text{fertilization})}{1 + \exp(-1.1091 + 0.3234 \cdot \text{pestcontrol} + 0.7332 \cdot \text{fertilization})}, \text{ and } \hat{\phi} = 4.9777.$$

At the 5% level, wind speed is a significant predictor of  $\pi_1$ , and both pest control and fertilization are significant predictors of  $\mu$ .

In R:

```

trees.data<- read.csv(file="C:/../Exercise7.7Data.csv", header=TRUE, sep=",")

#computing response variable
trees.data$propsurvived<- trees.data$nsurvived/trees.data$nplanted

#fitting one-inflated beta model

```

```
library(gamlss)
summary(fitted.model<- gamlss(propsurvived ~ pestcontrol + fertilization,
mu.link="logit", nu.formula = ~ precipitation + windspeed, nu.link="logit",
data=trees.data, family=BEOI))
```

#### Mu Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-1.1080	0.5642	-1.964	0.0643
pestcontrol	0.3233	0.1364	2.371	0.0285
fertilization	0.7327	0.2821	2.597	0.0177

#### Sigma Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	1.6048	0.3015	5.323	3.88e-05

#### Nu Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	6.91004	4.29740	1.608	0.1243
precipitation	-0.04206	0.09075	-0.463	0.6483
windspeed	-0.78224	0.37748	-2.072	0.0521 .

(b) Present the fitted model and discuss its fit.

The deviance test produces a p-value below 0.05, which indicates a good fit for the model.

In SAS:

```
/*checking model fit*/
proc nlmixed;
parms b0=0.1 g0=0.1 phi=0.1;
p11=exp(b0)/(1+exp(b0));
mu=exp(g0)/(1+exp(g0));
if (propsurvived=1) then loglikelihood=log(p11);
else loglikelihood=log(1-p11)+lgamma(phi)-lgamma(mu*phi)-lgamma((1-mu)*phi)
+ (mu*phi-1)*log(propsurvived)+((1-mu)*phi-1)*log(1-propsurvived);
model propsurvived ~ general(loglikelihood);
run;
```

**-2 Log Likelihood 22.2**

```
data deviance;
deviance = 22.2 - 3.8;
pvalue = 1 - probchi(deviance,4);
run;
```

```
proc print noobs;
run;
```

deviance	pvalue
18.4	.001030602

In R:

```
#checking model fit
null.model<- gamlss(propsurvived ~ 1, mu.link="logit", nu.formula=~1,
nu.link="logit", data=trees.data, family=BEOI)
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))
```

18.4455

```
print(p.value<- pchisq(deviance, df=4, lower.tail = FALSE))
```

0.001009668

(c) Interpret the estimated significant parameters.

As the wind speed increases by one mph, the estimated odds in favor of 100% survival of trees change by  $(\exp(-0.7822) - 1) \cdot 100\% = -54.2601373\%$ , that is, decrease by 54.26%. As the frequency of pest control increases by one, the ratio  $\frac{\hat{\mu}}{1-\hat{\mu}}$  increases by  $(\exp(0.3234) - 1) \cdot 100\% = 38.1818\%$ .

As the frequency of fertilization increases by one, the ratio  $\frac{\hat{\mu}}{1-\hat{\mu}}$  increases by  $(\exp(0.7332) - 1) \cdot 100\% = 108.1732\%$ .

(d) Parks and Recreation Department employees are considering planting 100 trees in a hard-to-reach area where neither pest control nor soil fertilization would be feasible. They are trying to decide between an area with lower precipitation (2 inches) and stronger winds (12.5mph), and an area with higher precipitation (25 in) and lower winds (6 mph). Which of the two areas would you recommend to use based on the predicted proportion of trees that would survive for two years?

Predicted proportions of surviving trees for the two areas are

$$prop_1^0 = (1 + \exp(6.91 - 0.04206 \cdot 2 - 0.7822 \cdot 12.5))^{-1} \left( \exp(6.91 - 0.04206 \cdot 2 - 0.7822 \cdot 12.5) + \frac{\exp(-1.1091)}{1 + \exp(-1.1091)} \right) = 0.285381,$$

$$\text{and } prop_2^0 = (1 + \exp(6.91 - 0.04206 \cdot 25 - 0.7822 \cdot 6))^{-1} \left( \exp(6.91 - 0.04206 \cdot 25 - 0.7822 \cdot 6) + \frac{\exp(-1.1091)}{1 + \exp(-1.1091)} \right) = 0.821255.$$

The second area (with higher precipitation and lower winds) has a higher predicted proportion of survived trees.

In SAS:

```
/*using fitted model for prediction*/
data predict;
input pestcontrol fertilization precipitation windspeed;
cards;
0 0 2 12.5
0 0 25 6
;

data trees;
set trees predict;
run;

proc nlmixed;
parms b0=0.1 b1=0.1 b2=0.1 g0=0.1 g1=0.1 g2=0.1 phi=0.1;
pi1=exp(b0+b1*precipitation+b2*windspeed) / (1+exp(b0+b1*precipitation+b2*windspeed));
mu=exp(g0+g1*pestcontrol+g2*fertilization) / (1+exp(g0+g1*pestcontrol+g2*fertilization));
if (propsurvived=1) then loglikelihood=log(pi1);
```

```

else loglikelihood=log(1-pi1)+lgamma(phi)-lgamma(mu*phi)-lgamma((1-mu)*phi)
+ (mu*phi-1)*log(propsurvived)+((1-mu)*phi-1)*log(1-propsurvived);
model propsurvived ~ general(loglikelihood);
predict pi1+(1-pi1)*mu out=outdata;
run;

proc print data=outdata (firstobs=27) noobs;
var Pred;
run;

```

```

Pred
0.28537
0.82123

```

In R:

```

#using fitted model for prediction
param1<- predictAll(fitted.model, newdata = data.frame(pestcontrol=0,
fertilization=0, precipitation=2, windspeed=12.5), type="response")
param2<- predictAll(fitted.model, newdata = data.frame(pestcontrol=0,
fertilization=0, precipitation=25, windspeed=6), type="response")
print(param1$nu+(1-param1$nu)*param1$mu)
print(param2$nu+(1-param2$nu)*param2$mu)

0.2855535
0.8212778

```

**EXERCISE 7.8.** (a) Fit a one-inflated beta regression to model the proportion of completed sales, regressing the probability of one on the number of years of experience a salesperson has accrued. Use the significance level of 0.05. Write down the fitted model.

In SAS:

```

data sales;
input gender $ expyr bonus propsales @@;
female=(gender="F");
cards;
F 1 1.1 0.67 M 11 0.5 1 M 4 1.1 0.9 M 2 1.6 0.93
F 2 0.7 0.49 F 4 1.05 0.88 M 1 1.6 0.96 F 2 1.2 0.67
M 2 1.6 0.94 M 7 1.4 0.77 F 4 1.55 1 F 4 0.9 0.51
F 8 0.95 0.59 F 2 1.2 0.65 F 13 0.6 1 F 8 0.9 0.54
M 4 0.6 0.63 F 17 2.4 1 F 3 1.6 1 F 2 1.4 0.88
F 4 1.05 0.85 F 8 1.4 1 M 4 1.35 0.95 F 3 1 0.83
F 18 1.25 1 M 4 0.4 0.66
;

/*fitting one-inflated beta regression model*/
proc nlmixed;
parms b0=0.1 b1=0.1 g0=0.1 g1=0.1 g2=0.1 phi=0.1;
pi1=exp(b0+b1*expyr)/(1+exp(b0+b1*expyr));
mu=exp(g0+g1*female+g2*bonus)/(1+exp(g0+g1*female+g2*bonus));
if (propsales=1) then loglikelihood= log(pi1);
else loglikelihood=log(1-pi1)+lgamma(phi)-lgamma(mu*phi)-lgamma((1-mu)*phi)
+ (mu*phi-1)*log(propsales)+((1-mu)*phi-1)*log(1-propsales);

```

```
model propsales ~ general(loglikelihood);
run;
```

-2 Log Likelihood -22.6

Parameter Estimates							
Parameter	Estimate	Standard Error	DF	t value	Pr >  t	95% Confidence Limits	Gradient
b0	-3.7949	1.2965	26	-2.93	0.0070	-6.4598 -1.1300	7.533E-7
b1	0.4619	0.1933	26	2.39	0.0244	0.06453 0.8593	7.801E-6
g0	-0.3537	0.4533	26	-0.78	0.4423	-1.2854 0.5781	0.000015
g1	-0.7029	0.2599	26	-2.70	0.0119	-1.2371 -0.1686	0.000014
g2	1.8148	0.4038	26	4.49	0.0001	0.9848 2.6448	0.000021
phi	19.1898	6.1927	26	3.10	0.0046	6.4605 31.9192	3.494E-8

The fitted parameters are

$$\hat{\pi}_1 = \frac{\exp(-3.7949 + 0.4619 \cdot \text{years of experience})}{1 + \exp(-3.7949 + 0.4619 \cdot \text{years of experience})}$$

$$\hat{\mu} = \frac{\exp(-0.3537 - 0.7029 \cdot \text{female} + 1.8148 \cdot \text{bonus})}{1 + \exp(-0.3537 - 0.7029 \cdot \text{female} + 1.8148 \cdot \text{bonus})}, \text{ and } \hat{\phi} = 19.1898.$$

At the 5% level, all predictors are significant: years of experience is a significant predictor of  $\pi_1$ , and gender and bonus amount are significant predictors of  $\mu$ .

In R:

```
sales.data<- read.csv(file="C:/../Exercise7.8Data.csv", header=TRUE, sep=",")

#specifying reference level
sales.data$gender.rel<- relevel(sales.data$gender, ref="M")

#fitting one-inflated beta model
library(gamlss)
summary(fitted.model<- gamlss(propsales ~ gender.rel + bonus, mu.link="logit",
nu.formula = ~ expyr, nu.link="logit", data=sales.data, family=BEOI))
```

Mu Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-0.3532	0.4533	-0.779	0.445044
gender.relF	-0.7023	0.2599	-2.702	0.013712
bonus	1.8137	0.4038	4.492	0.000223

Sigma Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	2.9540	0.3228	9.151	1.38e-08

Nu Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-3.7949	1.2964	-2.927	0.00833
expyr	0.4619	0.1933	2.389	0.02684

(b) How good is the model fit?

The p-value is minuscule, which suggests a very good fit of the model.

In SAS:

```
/*checking model fit*/
proc nlmixed data=sales;
parms b0=0.1 g0=0.1 phi=0.1;
pi1=exp(b0)/(1+exp(b0));
mu=exp(g0)/(1+exp(g0));
if (propsales=1) then loglikelihood= log(pi1);
else loglikelihood=log(1-pi1)+lgamma(phi)-lgamma(mu*phi)-lgamma((1-mu)*phi)
  +(mu*phi-1)*log(propsales)+(1-mu)*phi-1)*log(1-propsales);
model propsales~general(loglikelihood);
run;
```

**-2 Log Likelihood 9.6**

```
data deviance;
  deviance = 9.6 - (-22.6);
  pvalue = 1 - probchi(deviance,3);
run;
```

```
proc print noobs;
run;
```

deviance	pvalue
32.2	.000000475

In R:

```
#checking model fit
null.model<- gamlss(propsales ~ 1, mu.link="logit", nu.formula = ~ 1,
nu.link="logit", data=sales.data, family=BEOI)
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))
```

**32.16519**

```
print(p.value<- pchisq(deviance, df=3, lower.tail = FALSE))
```

**4.830298e-07**

(c) Interpret the estimated significant regression coefficients.

As the number of years of experience of a salesperson increases by one, the estimated odds in favor of 100% successful sales increase by  $(\exp(0.4619) - 1) \cdot 100\% = 58.70866\%$ . For a female salesperson, the ratio  $\frac{\hat{\mu}}{1-\hat{\mu}}$  is  $\exp(-0.7029) \cdot 100\% = 49.51473\%$  of that for a male salesperson. As the amount of bonus increases by one thousand dollars, the ratio  $\frac{\hat{\mu}}{1-\hat{\mu}}$  increases by  $(\exp(1.8148) - 1) \cdot 100\% = 513.9848\%$ .

(d) Predict the proportion of completed sales for a salesman with 3 years of work experience and who received \$1,500 in bonuses the previous year.

$$\text{propsales}^0 = (1 + \exp(-3.7949 + 0.4619 \cdot 3))^{-1} \left( \exp(-3.7949 + 0.4619 \cdot 3) + \frac{\exp(-0.3537 + 1.8148 \cdot 1.5)}{1 + \exp(-0.3537 + 1.8148 \cdot 1.5)} \right) = 0.921454.$$

In SAS:

```
/*using fitted model for prediction*/
data predict;
input female expyr bonus;
cards;
0 3 1.5
;

data sales;
set sales predict;
run;

proc nlmixed;
parms b0=0.1 b1=0.1 g0=0.1 g1=0.1 g2=0.1 phi=0.1;
pi1=exp(b0+b1*expyr)/(1+exp(b0+b1*expyr));
mu=exp(g0+g1*female+g2*bonus)/(1+exp(g0+g1*female+g2*bonus));
if (propsales=1) then loglikelihood= log(pi1);
else loglikelihood=log(1-pi1)+lgamma(phi)-lgamma(mu*phi)-lgamma((1-mu)*phi)
+ (mu*phi-1)*log(propsales)+((1-mu)*phi-1)*log(1-propsales);
model propsales ~ general(loglikelihood);
predict pi1+(1-pi1)*mu out=outdata;
run;

proc print data=outdata (firstobs=27) noobs;
var Pred;
run;
```

```
Pred
0.92146
```

In R:

```
#using fitted model for prediction
param<- predictAll(fitted.model, newdata = data.frame(expyr=3, gender.rel="M",
bonus=1.5), type="response")
print(param$nu+(1-param$nu)*param$mu)
```

```
0.9213704
```

**EXERCISE 7.9.** (a) Fit the beta regression with inflated zeros and ones to model the germination rate. Regress parameter  $\mu$  on altitude normalized by a factor of 1000,  $\nu$  on EC, and  $\tau$  on soil temperature. Specify the fitted model. Determine the significance of regression coefficients at 5% and 10% levels.

In SAS:

```
data lab;
input EC soiltemp altitude germrate @@;
altitudeK=altitude/1000;
```

```

cards;
2.7 67 4368 0      1.1 67 1689 1      1.8 69 3156 0.87
1.6 67 4884 0.58  2.4 66 4926 0      1.6 63 3854 0.23
2.3 67 5146 0      1.2 64 2202 0.48  1.1 62 2759 0.82
1.9 62 2774 0.61  1.5 71 5927 0.19  1.7 61 827 0.93
2.8 62 3631 0      2.5 64 4229 0.17  1.8 69 2933 0.47
1.8 63 6110 0.32  1.5 67 461 1      2.5 67 5269 0
1.7 74 197 1      1.6 65 607 1      2.6 67 5263 0.16
1.2 69 651 1      1.7 65 863 0.8  1.5 62 4386 0.23
1.7 68 165 1      1.7 62 234 0.73
;

/*fitting zero-one-inflated beta model*/
proc nlmixed;
parms b0=0.1 b1=0.1 g0=0.1 g1=0.1 z0=0.1 z1=0.1 phi=0.1;
mu=exp(b0+b1*altitudeK)/(1+exp(b0+b1*altitudeK));
nu=exp(g0+g1*EC);
tau=exp(z0+z1*soiltemp);
pi0=nu/(1+nu+tau);
pi1=tau/(1+nu+tau);
if(germrate=0) then loglikelihood=log(pi0);
if(germrate=1) then loglikelihood=log(pi1);
if(germrate>0 and germrate<1) then loglikelihood=
log(1-pi0-pi1)+lgamma(phi)-lgamma(mu*phi)-lgamma((1-mu)*phi)+(mu*phi-
1)*log(germrate)
+((1-mu)*phi-1)*log(1-germrate);
model germrate ~ general(loglikelihood);
run;

```

-2 Log Likelihood 17.6

Parameter Estimates							
Parameter	Estimate	Standard Error	DF	t Value	Pr >  t	95% Confidence Limits	Gradient
b0	1.6634	0.4713	26	3.53	0.0016	0.6946 2.6323	-7.9E-6
b1	-0.4791	0.1262	26	-3.80	0.0008	-0.7385 -0.2197	-0.00002
g0	-13.7160	6.1607	26	-2.23	0.0349	-26.3795 -1.0525	1.974E-7
g1	5.8628	2.5790	26	2.27	0.0315	0.5616 11.1640	1.34E-6
z0	-24.8615	12.9149	26	-1.93	0.0652	-51.4086 1.6855	-8.01E-7
z1	0.3599	0.1926	26	1.87	0.0729	-0.03588 0.7558	-0.00005
phi	6.5607	2.2800	26	2.88	0.0079	1.8741 11.2474	-1.11E-7

The fitted parameters in the fitted model are

$$\hat{\mu} = \hat{E}(y | 0 < y < 1) = \frac{\exp(1.6634 - 0.4791 \cdot \text{altitudeK})}{1 + \exp(1.6634 - 0.4791 \cdot \text{altitudeK})},$$

$$\hat{\pi}_0 = \frac{\hat{v}}{1 + \hat{v} + \hat{\tau}}, \quad \text{and} \quad \hat{\pi}_1 = \frac{\hat{\tau}}{1 + \hat{v} + \hat{\tau}},$$

where  $\hat{v} = \frac{\hat{P}(y=0)}{\hat{P}(0 < y < 1)} = \exp(-13.716 + 5.8628 \cdot EC)$ , and  $\hat{\tau} = \frac{\hat{P}(y=1)}{\hat{P}(0 < y < 1)} = \exp(-24.8615 + 0.3599 \cdot \text{soiltemp})$ .

At the 5% level, altitudeK is a significant predictor of  $\mu$  and EC is a significant predictor of  $v$ . Soil temperature is a significant predictor of  $\tau$  at the 10% level.

In R:

```
lab.data<- read.csv(file="C:/./Exercise7.9Data.csv", header=TRUE, sep=",")

#rescaling variable
lab.data$altitudeK<- lab.data$altitude/1000

#fitting zero-one-inflated beta model
library(gamlss)
summary(fitted.model<- gamlss(germrate ~ altitudeK, mu.link="logit", nu.formula
= ~ EC, nu.link="log", tau.formula = ~ soiltemp, tau.link="log", data=lab.data,
family=BEINF))
```

Mu Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	1.6631	0.4713	3.529	0.00224
altitudeK	-0.4790	0.1262	-3.796	0.00122

Sigma Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-0.5594	0.2370	-2.361	0.0291

Nu Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-13.716	6.161	-2.226	0.0383
EC	5.863	2.579	2.273	0.0348

Tau Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-24.8615	12.9583	-1.919	0.0702
soiltemp	0.3599	0.1932	1.863	0.0780

(b) Discuss the model fit.

The tiny p-value indicates a very good model fit.

In SAS:

```
/*checking model fit*/
proc nlmixed;
parms b0=0.1 g0=0.1 z0=0.1 phi=0.1;
mu=exp(b0)/(1+exp(b0));
nu=exp(g0);
tau=exp(z0);
pi0=nu/(1+nu+tau);
pi1=tau/(1+nu+tau);
if(germrate=0) then loglikelihood=log(pi0);
if(germrate=1) then loglikelihood=log(pi1);
if(germrate>0 and germrate<1) then loglikelihood=
log(1-pi0-pi1)+lgamma(phi)-lgamma(mu*phi)-lgamma((1-mu)*phi)+(mu*phi-
1)*log(germrate)
+((1-mu)*phi-1)*log(1-germrate);
model germrate ~ general(loglikelihood);
run;
```

**-2 Log Likelihood 49.2**

```
data deviance;
deviance = 49.2 - 17.6;
```

```

pvalue = 1 - probchi(deviance,3);
run;

proc print noobs;
run;
deviance      pvalue
    31.6      .000000635

```

In R:

```

#checking model fit
null.model<- gamlss(germrate ~ 1, mu.link="logit", nu.formula = ~ 1,
nu.link="log", tau.formula = ~ 1, tau.link="log", data=lab.data, family=BEINF)
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))

```

31.5888

```

print(p.value<- pchisq(deviance, df=3, lower.tail = FALSE))

```

6.389023e-07

(c) Give an interpretation of the estimated significant regression coefficients.

As the plot altitude increases by one thousand feet, the ratio  $\frac{\hat{\mu}}{1-\hat{\mu}}$  changes by  $(\exp(-0.4791) - 1) \cdot 100\% = -38.0659\%$ , that is, decreases by 38.0659%. For a one-unit increase in EC, the estimated odds in favor of  $y = 0$  against  $0 < y < 1$  increase by  $(\exp(5.8628) - 1) \cdot 100\% = 35,070.75\%$ . For a one-degree increase in soil temperature, the estimated odds in favor of  $y = 1$  against  $0 < y < 1$  increase by  $(\exp(0.3599) - 1) \cdot 100\% = 43.31861\%$ .

(d) Use the fitted model to predict the germination rate for a plot with EC of 1.5 mS/cm<sup>2</sup>, soil temperature of 68°F, and altitude of 950 feet.

The predicted germination rate is computed as

$$\begin{aligned}
 \text{germrate}^0 &= \left( \exp(-24.8615 + 0.3599 \cdot 68) + \frac{\exp(1.6634 - 0.4791 \cdot 0.95)}{1 + \exp(1.6634 - 0.4791 \cdot 0.95)} \right) \times \\
 &\times (1 + \exp(-13.716 + 5.8628 \cdot 1.5) + \exp(-24.8615 + 0.3599 \cdot 68))^{-1} = 0.859213.
 \end{aligned}$$

In SAS:

```

/*using fitted model for prediction*/
data predict;
input EC soiltemp altitudeK;
cards;
1.5 68 0.95
;

data lab;
set lab predict;
run;

proc nlmixed;
parms b0=0.1 b1=0.1 g0=0.1 g1=0.1 z0=0.1 z1=0.1 phi=0.1;

```

```

mu=exp(b0+b1*altitudeK)/(1+exp(b0+b1*altitudeK));
nu=exp(g0+g1*EC);
tau=exp(z0+z1*soiltemp);
pi0=nu/(1+nu+tau);
pi1=tau/(1+nu+tau);
if(germrate=0) then loglikelihood=log(pi0);
if(germrate=1) then loglikelihood=log(pi1);
if(germrate>0 and germrate<1) then loglikelihood=
log(1-pi0-pi1)+lgamma(phi)-lgamma(mu*phi)-lgamma((1-mu)*phi)+(mu*phi-
1)*log(germrate)
+((1-mu)*phi-1)*log(1-germrate);
model germrate ~ general(loglikelihood);
predict (tau+mu)/(1+nu+tau) out=outdata;
run;

proc print data=outdata (firstobs=27) noobs;
  var Pred;
run;

```

```

    Pred
0.85938

```

In R:

```

#using fitted model for prediction
param<- predictAll(fitted.model, newdata = data.frame(EC=1.5, soiltemp=68,
altitudeK=0.95), type="response")
print((param$tau+param$mu)/(1+param$nu+param$tau))

0.8593549

```

**EXERCISE 7.10.** (a) Regress the proportion of games played, using the zero-one-inflated beta model. Regress  $\mu$  on the vertical jump and number of bench press repetitions,  $\nu$  on the broad jump, and  $\tau$  on BMI and forty-yard dash. What predictors are significant at the 5% level?

In SAS:

```

data football;
input BMI fortyyrd vertical broad bench propgames @@;
cards;
31.1 4.58 35 108 20 0 28.5 4.70 30.5 115 21 1
32.4 4.39 36 116 18 0 30.8 4.67 33 121 15 0.87
29.6 4.41 33 116 26 0.47 30.0 4.56 38 122 21 0.8
31.3 4.50 35 119 29 0.8 29.4 4.49 31.5 115 18 0.33
28.1 4.37 34.5 130 21 0.53 31.0 4.52 33.5 128 25 0.87
29.7 4.57 31.5 124 18 0.73 27.6 4.62 38.5 118 15 1
29.5 4.60 32 121 15 0.47 30.7 4.40 34.5 113 22 0.47
29.4 4.73 34 114 15 0.6 29.1 4.57 35 111 19 0.87
30.6 4.60 30.5 114 24 0.47 29.3 4.55 36 115 17 0.87
28.1 4.59 35.5 109 14 0.87 31.7 4.62 37 121 19 0.8
29.7 4.73 34 118 21 1 30.7 4.80 34 114 15 0.33
28.8 4.37 36 121 19 0.73 30.7 4.68 34 105 14 0
30.7 4.64 33 116 21 0.47 30.3 4.50 35.5 115 25 0.6

```

```

30.0 4.48 34    119 20 0.67  28.6 4.59 36    123 17 1
30.8 4.43 34    117 18 0.53  27.7 4.51 33    117 20 0.73
30.7 4.50 38    114 18 1      28.1 4.59 36    115 17 0.93
29.3 4.61 32    113 20 0      27.9 4.64 33    118 23 1
29.7 4.67 41    124 21 1      29.9 4.48 35    111 18 0
32.0 4.51 33.5  116 25 0.67  29.6 4.37 37    115 24 0.93
32.9 4.55 33    122 27 0.47  27.6 4.67 33.5  118 26 1
30.4 4.55 32    109 17 0.73  31.4 4.50 38    121 18 0.67
;

/*fitting zero-one-inflated beta model*/
proc nlmixed;
parms b0=.1 b1=.1 b2=.1 g0=.1 g1=.1 z0=.1 z1=.1 z2=.1 phi=.1;
mu=exp(b0+b1*vertical+b2*bench)/(1+exp(b0+b1*vertical+b2*bench));
nu=exp(g0+g1*broad);
tau=exp(z0+z1*BMI+z2*fortyyd);
pi0=nu/(1+nu+tau);
pi1=tau/(1+nu+tau);
if (propgames=0) then loglikelihood=log(pi0);
if(propgames=1) then loglikelihood=log(pi1);
if(propgames>0 and propgames<1) then loglikelihood=
log(1-pi0-pi1)+lgamma(phi)-lgamma(mu*phi)-lgamma((1-mu)*phi)+(mu*phi-
1)*log(propgames)
+((1-mu)*phi-1)*log(1-propgames);
model propgames ~ general(loglikelihood);
run;

```

## 2 Log Likelihood 13.6

Parameter Estimates								
Parameter	Estimate	Standard Error	DF	t Value	Pr >  t	95% Confidence Limits	Gradient	
b0	-6.3365	2.4421	42	-2.59	0.0130	-11.2649 -1.4081	-6.57E-7	
b1	0.2141	0.06943	42	3.08	0.0036	0.07398 0.3542	0.000011	
b2	-0.01399	0.03066	42	-0.46	0.6504	-0.07586 0.04787	0.000086	
g0	45.5487	19.0286	42	2.39	0.0212	7.1474 83.9500	-6.54E-6	
g1	-0.4147	0.1691	42	-2.45	0.0184	-0.7560 -0.07346	-0.00072	
z0	-27.0077	26.3559	42	-1.02	0.3114	-80.1960 26.1806	0.000024	
z1	-1.0664	0.4812	42	-2.22	0.0322	-2.0374 -0.09530	0.000706	
z2	12.4225	5.7433	42	2.16	0.0363	0.8321 24.0129	0.000110	
phi	8.6232	2.1696	42	3.97	0.0003	4.2447 13.0016	1.258E-6	

The fitted parameters in the fitted model are

$$\hat{\mu} = \hat{E}(y | 0 < y < 1) = \frac{\exp(-6.3365 + 0.2141 \cdot \text{vertical jump} - 0.01399 \cdot \text{bench press})}{1 + \exp(-6.3365 + 0.2141 \cdot \text{vertical jump} - 0.01399 \cdot \text{bench press})},$$

$$\hat{\pi}_0 = \frac{\hat{\nu}}{1 + \hat{\nu} + \hat{\tau}}, \quad \text{and} \quad \hat{\pi}_1 = \frac{\hat{\tau}}{1 + \hat{\nu} + \hat{\tau}},$$

where  $\hat{\nu} = \frac{\hat{P}(y=0)}{\hat{P}(0 < y < 1)} = \exp(45.5487 - 0.4147 \cdot \text{broad jump})$ , and  $\hat{\tau} = \frac{\hat{P}(y=1)}{\hat{P}(0 < y < 1)} = \exp(-27.0077 - 1.0664 \cdot \text{BMI} + 12.4225 \cdot \text{forty-yard dash})$ .

At the 5% level, the vertical jump is a significant predictor of  $\mu$ , the broad jump is a significant predictor of  $\nu$ , and BMI and the forty-yard dash are significant predictors of  $\tau$ .

In R:

```
football.data<- read.csv(file="C:/./Exercise7.10Data.csv", header=TRUE, sep=",")
```

```
#fitting zero-one-inflated beta model
```

```
library(gamlss)
```

```
summary(fitted.model<- gamlss(propgames ~ vertical + bench, mu.link="logit",  
nu.formula = ~ broad, nu.link="log",tau.formula = ~ BMI + fortyyd, tau.link  
="log", data=football.data, family=BEINF))
```

Mu Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-6.33600	2.44008	-2.597	0.01395
vertical	0.21407	0.06938	3.086	0.00409
bench	-0.01399	0.03066	-0.456	0.65117

Sigma Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-0.7429	0.1665	-4.463	8.9e-05

Nu Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	45.5488	18.5894	2.450	0.0197
broad	-0.4148	0.1652	-2.511	0.0171

Tau Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-27.0077	26.3530	-1.025	0.3129
BMI	-1.0664	0.4808	-2.218	0.0336
fortyyd	12.4225	5.7375	2.165	0.0377

(b) Analyze the fit of the model.

The model has a very good fit since the p-value in the deviance test is very small.

In SAS:

```
/*checking model fit*/  
proc nlmixed;  
parms b0=.1 g0=.1 z0=.1 phi=.1;  
mu=exp(b0)/(1+exp(b0));  
nu=exp(g0);  
tau=exp(z0);  
pi0=nu/(1+nu+tau);  
pi1=tau/(1+nu+tau);  
if (propgames=0) then loglikelihood=log(pi0);  
if(propgames=1) then loglikelihood=log(pi1);  
if(propgames>0 and propgames<1) then loglikelihood=  
log(1-pi0-pi1)+lgamma(phi)-lgamma(mu*phi)-lgamma((1-mu)*phi)+(mu*phi-  
1)*log(propgames)  
+((1-mu)*phi-1)*log(1-propgames);  
model propgames ~ general(loglikelihood);  
run;
```

-2 Log Likelihood 47.3

```
data deviance;
```

```

deviance = 47.3 - 13.6;
pvalue = 1 - probchi(deviance,5);
run;

proc print noobs;
run;

```

```

deviance      pvalue
    33.7      .000002732

```

In R:

```

#checking model fit
null.model<- gamlss(propgames ~ 1, mu.link="logit", nu.formula = ~ 1,
nu.link="log", tau.formula = ~ 1, tau.link="log", data=football.data,
family= BEINF)
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))

```

```

33.66176

```

```

print(p.value<- pchisq(deviance, df=5, lower.tail = FALSE))

```

```

2.780153e-06

```

(c) Give an interpretation of the estimated significant coefficients.

As the vertical jump increases by one inch, the ratio  $\frac{\hat{\mu}}{1-\hat{\mu}}$  increases by  $(\exp(0.2141) - 1) \cdot 100\% = 23.87465\%$ . For a one-inch increase in the broad jump, the estimated odds in favor of  $y = 0$  against  $0 < y < 1$  change by  $(\exp(-0.4147) - 1) \cdot 100\% = -33.9462\%$ , that is, decrease by 33.9462%. For a one-unit increase in BMI, the estimated odds in favor of  $y = 1$  against  $0 < y < 1$  change by  $(\exp(-1.0664) - 1) \cdot 100\% = -65.5754\%$ , that is, decrease by 65.5754%. For a one-second increase in the forty-dash run, the estimated odds in favor of  $y = 1$  against  $0 < y < 1$  increase by  $(\exp(12.4225) - 1) \cdot 100\% = 24,832,557.66\%$ .

(d) Predict the proportion of games that a new player will play, if his BMI is 27.8 kg/m<sup>2</sup>, the forty-dash run is 4.67 seconds, the vertical jump is 32 inches, the broad jump is 117 inches, and the bench press is 16 repetitions.

The predicted proportion of games is

$$\begin{aligned}
 \text{propgames}^0 &= \left( \exp(-27.0077 - 1.0664 \cdot 27.8 + 12.4225 \cdot 4.67) \right. \\
 &\quad \left. + \frac{\exp(-6.3365 + 0.2141 \cdot 32 - 0.01399 \cdot 16)}{1 + \exp(-6.3365 + 0.2141 \cdot 32 - 0.01399 \cdot 16)} \right) \times \\
 &\quad \times (1 + \exp(45.5487 - 0.4147 \cdot 117) + \exp(-27.0077 - 1.0664 \cdot 27.8 + 12.4225 \cdot 4.67))^{-1} \\
 &= 0.903133.
 \end{aligned}$$

In SAS:

```

/*using fitted model for prediction*/
data predict;
input BMI fortyyd vertical broad bench;
cards;
27.8 4.67 32 117 16
;

data football;
set football predict;
run;

proc nlmixed;
parms b0=.1 b1=.1 b2=.1 g0=.1 g1=.1 z0=.1 z1=.1 z2=.1 phi=.1;
mu=exp(b0+b1*vertical+b2*bench)/(1+exp(b0+b1*vertical+b2*bench));
nu=exp(g0+g1*broad);
tau=exp(z0+z1*BMI+z2*fortyyd);
pi0=nu/(1+nu+tau);
pi1=tau/(1+nu+tau);
if (propgames=0) then loglikelihood=log(pi0);
if(propgames=1) then loglikelihood=log(pi1);
if(propgames>0 and propgames<1) then loglikelihood=
log(1-pi0-pi1)+lgamma(phi)-lgamma(mu*phi)-lgamma((1-mu)*phi)+(mu*phi-
1)*log(propgames)
+((1-mu)*phi-1)*log(1-propgames);
model propgames ~ general(loglikelihood);
predict (tau+mu)/(1+nu+tau) out=outdata;
run;

proc print data=outdata (firstobs=43) noobs;
var Pred;
run;

```

**Pred**  
**0.90325**

**In R:**

```

#using fitted model for prediction
param<- predictAll(fitted.model, newdata = data.frame(BMI=27.8, fortyyd=4.67,
vertical=32, broad=117, bench=16), type="response")
print((param$tau+param$mu)/(1+param$nu+param$tau))

```

**0.9032484**

## CHAPTER 8

**EXERCISE 8.1.** (a) For any  $i \neq i'$ ,  $Cov(y_{ij}, y_{i'j'}) = Cov(\beta_0 + \beta_1 x_{1ij} + \dots + \beta_k x_{kij} + \beta_{k+1} t_j + u_{1i} + u_{2i} t_j + \varepsilon_{ij}, \beta_0 + \beta_1 x_{1i'j'} + \dots + \beta_k x_{ki'j'} + \beta_{k+1} t_{j'} + u_{1i'} + u_{2i'} t_{j'} + \varepsilon_{i'j'}) = Cov(u_{1i} + u_{2i} t_j + \varepsilon_{ij}, u_{1i'} + u_{2i'} t_{j'} + \varepsilon_{i'j'}) = Cov(u_{1i}, u_{1i'}) + Cov(u_{1i}, u_{2i'}) t_{j'} + Cov(u_{1i}, \varepsilon_{i'j'}) + Cov(u_{2i}, u_{1i'}) t_j + Cov(u_{2i}, u_{2i'}) t_j t_{j'} + Cov(u_{2i}, \varepsilon_{i'j'}) t_j + Cov(\varepsilon_{ij}, u_{1i'}) + Cov(\varepsilon_{ij}, u_{2i'}) t_{j'} + Cov(\varepsilon_{ij}, \varepsilon_{i'j'}) = 0.$

(b) For any given  $i$  and  $j \neq j'$ ,  $Cov(y_{ij}, y_{ij'}) = Cov(\beta_0 + \beta_1 x_{1ij} + \dots + \beta_k x_{kij} + \beta_{k+1} t_j + u_{1i} + u_{2i} t_j + \varepsilon_{ij}, \beta_0 + \beta_1 x_{1ij'} + \dots + \beta_k x_{kij'} + \beta_{k+1} t_{j'} + u_{1i} + u_{2i} t_{j'} + \varepsilon_{ij'}) = Cov(u_{1i} + u_{2i} t_j + \varepsilon_{ij}, u_{1i} + u_{2i} t_{j'} + \varepsilon_{ij'}) = Cov(u_{1i}, u_{1i}) + Cov(u_{1i}, u_{2i}) t_{j'} + Cov(u_{1i}, \varepsilon_{ij'}) + Cov(u_{2i}, u_{1i}) t_j + Cov(u_{2i}, u_{2i}) t_j t_{j'} + Cov(u_{2i}, \varepsilon_{ij'}) t_j + Cov(\varepsilon_{ij}, u_{1i}) + Cov(\varepsilon_{ij}, u_{2i}) t_{j'} + Cov(\varepsilon_{ij}, \varepsilon_{ij'}) = Var(u_{1i}) + Cov(u_{1i}, u_{2i})(t_j + t_{j'}) + Var(u_{2i}) t_j t_{j'} = \sigma_{u_1}^2 + \sigma_{u_1 u_2} (t_j + t_{j'}) + \sigma_{u_2}^2 t_j t_{j'}.$

(c) The response variable  $y_{ij} = \beta_0 + \beta_1 x_{1ij} + \dots + \beta_k x_{kij} + \beta_{k+1} t_j + u_{1i} + u_{2i} t_j + \varepsilon_{ij}$  has a normal distribution as the sum of normal random variables. Its mean is  $E(y_{ij}) = E(\beta_0 + \beta_1 x_{1ij} + \dots + \beta_k x_{kij} + \beta_{k+1} t_j + u_{1i} + u_{2i} t_j + \varepsilon_{ij}) = \beta_0 + \beta_1 x_{1ij} + \dots + \beta_k x_{kij} + \beta_{k+1} t_j + E(u_{1i}) + E(u_{2i}) t_j + E(\varepsilon_{ij}) = \beta_0 + \beta_1 x_{1ij} + \dots + \beta_k x_{kij} + \beta_{k+1} t_j$ , and the variance is  $Var(y_{ij}) = Var(\beta_0 + \beta_1 x_{1ij} + \dots + \beta_k x_{kij} + \beta_{k+1} t_j + u_{1i} + u_{2i} t_j + \varepsilon_{ij}) = Var(u_{1i} + u_{2i} t_j + \varepsilon_{ij}) = Var(u_{1i}) + 2Cov(u_{1i}, u_{2i}) t_j + Var(u_{2i}) t_j^2 + 2Cov(u_{1i}, \varepsilon_{ij}) + 2Cov(u_{2i}, \varepsilon_{ij}) t_j + Var(\varepsilon_{ij}) = \sigma_{u_1}^2 + 2\sigma_{u_1 u_2} t_j + \sigma_{u_2}^2 t_j^2 + \sigma^2.$

**EXERCISE 8.2.** (a) Carry out tests for the normality of the bonus and plot the histogram. Is this variable normally distributed?

In SAS:

```
data deptstore;
input id totalyears status$ bonus18 bonus19 bonus20 @@;
cards;
1 16 full 1482 1508 1543 2 7 part 673 710 895
3 11 full 933 1351 1440 4 8 part 844 958 1196
5 6 part 564 790 815 6 5 full 601 708 780
7 6 part 775 822 902 8 17 full 1209 1297 1475
9 12 full 929 1008 1255 10 9 full 983 1013 1111
11 11 full 909 1004 1084 12 6 part 387 853 999
13 4 part 476 530 627 14 6 full 780 843 925
15 10 full 717 1200 1399
;

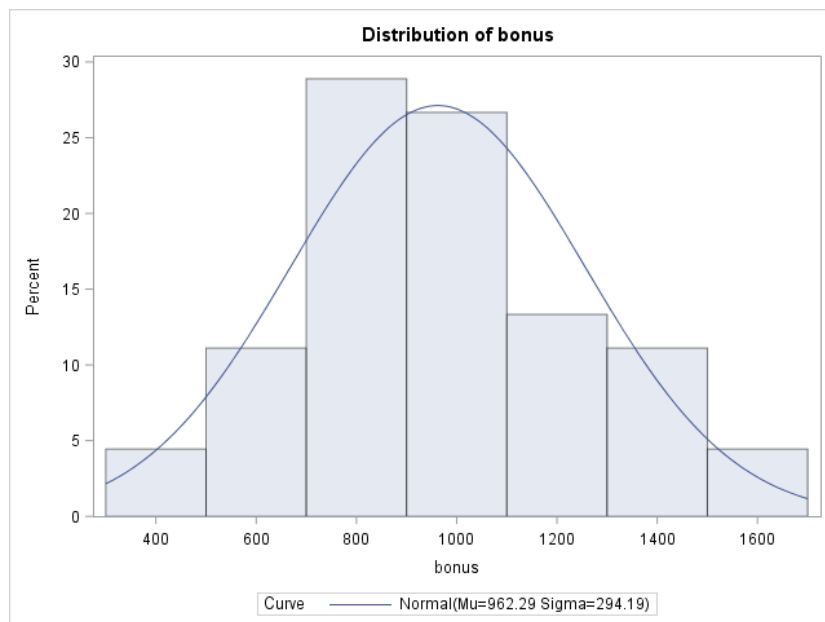
/*creating longform dataset*/
data longform;
set deptstore;
array y[3] (1.8 1.9 2.0);
array b[3] bonus18-bonus20;
```

```

do i=1 to 3;
  year=y[i];
  bonus=b[i];
  output;
end;
keep id totalyears status year bonus;
run;

/*checking normality of response*/
proc univariate;
  var bonus;
  histogram bonus/normal;
run;

```



#### Goodness-of-Fit Tests for Normal Distribution

Test	Statistic	p value
Kolmogorov-Smirnov D	0.12045943	Pr > D 0.098
Cramer-von Mises	W-Sq 0.08690779	Pr > W-Sq 0.168
Anderson-Darling	A-Sq 0.51592849	Pr > A-Sq 0.189

The histogram is roughly bell-shaped, and the normality tests all have p-values above 0.05, indicating a normal distribution of bonus.

In R:

```

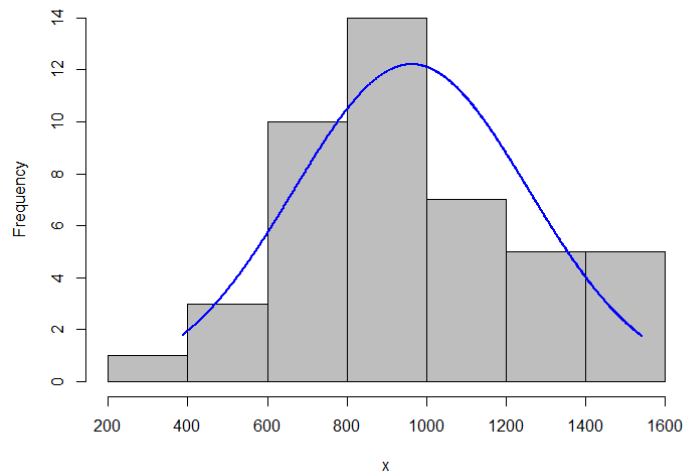
deptstore.data<- read.csv(file="C:/./Exercise8.2Data.csv", header=TRUE, sep=",")

#creating longform dataset
library(reshape2)
longform.data<- melt(deptstore.data, id.vars=c("id","totalyears","status"),
  variable.name="bonus.year", value.name="bonus")

#creating variable for time
year<- ifelse(longform.data$bonus.year=="bonus18",1.8,
  ifelse(longform.data$bonus.year=="bonus19",1.9,2.0))

```

```
#checking normality of response
library(rcompanion)
plotNormalHistogram(longform.data$bonus)
```



```
shapiro.test(longform.data$bonus)
```

Shapiro-wilk normality test

W = 0.96686, p-value = 0.222

- (a) Fit a random slope and intercept model regressing bonus on years with the company, status, and year (scaled by a factor of 10). Does the model fit the data well?

In SAS:

```
/*fitting random slope and intercept model*/
proc mixed covtest;
  class status;
  model bonus = totalyears status year / solution;
  random intercept year / subject=id type=un;
run;
```

Covariance Parameter Estimates					
Cov	Parm	Subject	Estimate	Standard Error	Z Value Pr >  Z
UN(1,1)	id		2249551	1204315	1.87 0.0309
UN(2,1)	id		-1194782	637878	-1.87 0.0611
UN(2,2)	id		636730	338378	1.88 0.0299
Residual			4608.43	1682.76	2.74 0.0031

Solution for Fixed Effects					
Effect	status	Estimate	Standard Error	DF t Value	Pr >  t
Intercept		-2246.11	457.20	12	-4.91 0.0004
totalyears		58.8982	8.2368	15	7.15 <.0001
status	full	53.9643	63.3473	15	0.85 0.4077
status	part	0	.	.	. .
year		1394.67	240.44	14	5.80 <.0001

Null Model	Likelihood	Ratio Test
DF	Chi-Square	Pr > ChiSq
3	10.93	0.0121

The deviance test has a p-value below 0.05, which confirms the model's good fit. Also, the parameters of the random-effect terms are significant at the 5% level. The covariance between the random slope and intercept is marginally significant with a p-value of 0.0611. So, using the random slope and intercept model is justified.

In R:

```
#creating reference level
status.rel<- relevel(longform.data$status, ref="part")

#fitting random slope and intercept model
library(nlme)
summary(fitted.model<- lme(bonus ~ totalyears + status.rel + year,
random = ~ 1 + year | id, data=longform.data))
```

Random effects:

	StdDev	Corr
(Intercept)	1499.85877	(Intr)
year	797.95782	-0.998
Residual	67.88526	

Fixed effects:

	Value	Std.Error	DF	t-value	p-value
(Intercept)	-2246.1138	457.2023	29	-4.912735	0.000
totalyears	58.8982	8.2368	12	7.150631	0.000
status.relfull	53.9643	63.3473	12	0.851879	0.411
year	1394.6667	240.4381	29	5.800523	0.000

```
intervals(fitted.model)
```

Random Effects:

	lower	est.	upper
sd((Intercept))	891.8860755	1499.8587688	2522.2686935
sd(year)	476.3219559	797.9578238	1336.7779518
cor((Intercept),year)	-0.9995787	-0.9983042	-0.9931881

```
#checking model fit
null.model<- glm(bonus ~ totalyears + status.rel + year, data=longform.data)
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))
```

45.81206

```
print(p.value<- pchisq(deviance, df=3, lower.tail = FALSE))
```

6.218073e-10

(c) Write down the fitted model, specifying all estimated parameters. What predictors are significant at the 5% significance level?

The fitted model has the form  $\hat{E}(\text{bonus}) = -2246.11 + 58.8982 \cdot \text{total years} + 53.9643 \cdot \text{full-time employee} + 1394.67 \cdot \text{year}/10$ . The other parameters have estimates  $\hat{\sigma}_{u_1}^2 = 2249551$ ,  $\hat{\sigma}_{u_2}^2 = 636730$ ,  $\hat{\sigma}_{u_1 u_2} = -1194782$ , and  $\hat{\sigma}^2 = 4608.43$ . Total years with the company and year are significant predictors at the 5% level.

(d) Give an interpretation of the estimated significant regression coefficients.

As the total number of years with the company increases by one, the average bonus increases by \$58.8982. As the year increases by one, the average bonus increases by  $\$1394.67/10 = \$139.467$ .

(e) According to the fitted model, what is the predicted bonus in 2021 for a full-time employee who has been with the company for 7 years?

The predicted value is  $\text{bonus}^0 = -2246.11 + 58.8982 \cdot 7 + 53.9643 + 1394.67 \cdot \frac{21}{10} = \$1,148.949$ .

In SAS:

```
/*using fitted model for prediction*/
data predict;
input id totalyears status$ year;
cards;
21 7 full 2.1
;

data longform;
set longform predict;
run;

proc mixed covtest;
class status;
model bonus = totalyears status year / solution outpm=outdata;
random intercept year / subject=id type=un;
run;

proc print data=outdata (firstobs=46) noobs;
var Pred;
run;

Pred
1148.94
```

In R:

```
#using fitted model for prediction
print(predict(fitted.model, data.frame(status.rel="full", totalyears=7,
year=2.1), level=0))

1148.938
```

### EXERCISE 8.3. (a) Create a long-form data set.

In SAS:

```
data orthoclinic;
input id gender$ age doctor$ length1 length2 length3 score1 score2 score3 @@;
cards;
101 F 78 A 25 20 25 7.1 7.5 7.6 102 F 63 A 30 30 40 5.5 5.8 6.1
103 F 62 A 10 15 10 10.0 10.0 9.8 104 F 71 B 15 15 40 7.8 7.3 7.5
105 M 68 A 40 60 40 3.5 3.5 3.0 106 F 63 A 25 15 20 8.5 8.7 8.8
107 F 60 B 25 35 25 6.7 5.7 6.5 108 F 70 A 20 20 20 9.0 8.3 8.2
109 F 57 A 30 20 15 8.4 7.8 8.1 110 F 59 B 25 30 15 7.1 7.4 7.9
111 M 62 A 50 30 70 3.0 3.2 2.6 112 M 58 A 20 15 45 6.1 6.8 6.9
113 M 75 A 25 35 30 5.7 5.6 4.7 114 M 76 B 35 50 25 4.9 5.4 5.2
115 F 75 A 15 20 25 8.2 8.9 8.2 116 M 57 A 45 30 40 4.6 3.9 3.2
117 F 68 A 35 25 40 3.8 4.8 5.3 118 M 65 B 40 40 25 3.9 3.9 4.7
119 F 67 B 20 15 30 6.5 7.2 6.6 120 F 60 B 25 15 15 7.3 7.1 7.8
121 F 67 A 15 20 15 7.7 8.0 8.3 122 F 57 B 10 15 15 9.8 9.2 8.6
123 M 62 B 55 60 75 3.4 2.7 2.3 124 M 71 A 20 30 25 7.1 6.6 7.4
125 M 71 B 15 15 20 8.8 9.1 9.3 126 M 64 A 25 30 30 5.6 6.3 6.3
127 M 51 A 35 40 30 5.1 4.6 3.9 128 F 70 B 35 25 15 6.8 7.1 7.6
129 M 61 A 35 40 50 5.5 5.2 4.8 130 M 62 B 60 40 65 3.7 3.4 2.4
131 F 68 A 20 35 35 5.3 5.6 4.9 132 F 68 B 35 30 15 7.2 6.2 5.6
133 M 64 B 40 20 30 5.4 4.9 4.5 134 F 76 B 30 45 25 5.5 4.7 4.6
135 F 78 B 25 20 15 7.6 8.3 9.2
;

/*creating longform dataset*/
data longform;
set orthoclinic;
array l[3] length1-length3;
array s[3] score1-score3;
do visit=1 to 3;
length=l[visit];
score=s[visit];
output;
end;
keep id gender age doctor visit length score;
run;
```

In R:

```
orthoclinic.data<- read.csv(file="C:/./Exercise8.3Data.csv", header=TRUE,
sep=",")

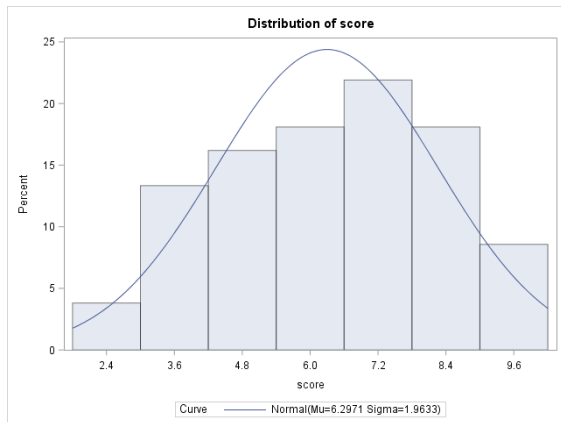
#creating longform dataset
library(reshape2)
data1<- melt(orthoclinic.data[,c("id","gender","age","doctor","length1",
"length2","length3")],id.vars=c("id","gender","age","doctor"),
variable.name="length.visit",value.name="length")
data2<- melt(orthoclinic.data[,c("score1","score2","score3")],
variable.name="score.visit", value.name="score")
longform.data<- cbind(data1,data2)

#creating variable for time
visit<- ifelse(longform.data$score.visit=="score1",1,
ifelse(longform.data$score.visit=="score2",2,3))
```

(b) Confirm that the quality of service is normally distributed by plotting a histogram and conducting normality tests.

In SAS:

```
/*checking normality of response*/  
proc univariate;  
  var score;  
  histogram score/normal;  
run;
```



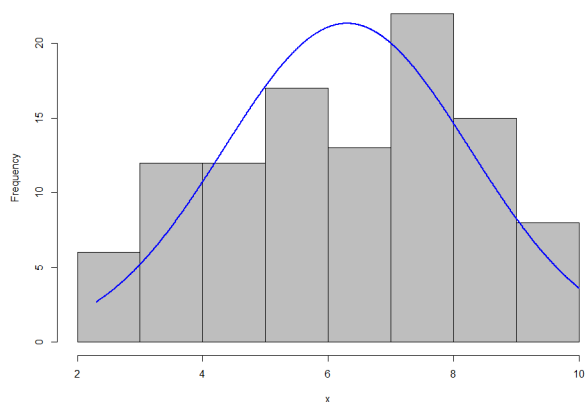
#### Goodness-of-Fit Tests for Normal Distribution

Test	Statistic	p value
Kolmogorov-Smirnov D	0.08727378	Pr > D 0.048
Cramer-von Mises	W-Sq 0.10666743	Pr > W-Sq 0.093
Anderson-Darling	A-Sq 0.64603497	Pr > A-Sq 0.092

The p-values in the normality tests are above 0.05, indicating a normal distribution. The histogram is approximately bell-shaped.

In R:

```
#plotting histogram and checking normality  
library(rcompanion)  
plotNormalHistogram(longform.data$score)
```



```
shapiro.test(longform.data$score)
```

### Shapiro-wilk normality test

W = 0.97598, p-value = 0.05332

(c) Fit a random slope and intercept model to regress the quality of service scores on all the predictor variables. Discuss the model fit.

### In SAS:

```
/*fitting random slope and intercept model*/
proc mixed covtest;
  class gender doctor;
  model score = gender age doctor length visit / solution;
  random intercept visit / subject=id type=un;
run;
```

Covariance Parameter Estimates						
Cov	Parm	Subject	Estimate	Standard Error	Z Value	Pr >  Z
UN(1,1)	id		2.2772	0.6365	3.58	0.0002
UN(2,1)	id		-0.1369	0.1277	-1.07	0.2835
UN(2,2)	id		0.1376	0.04432	3.10	0.0010
Residual			0.08137	0.01984	4.10	<.0001

Solution for Fixed Effects						
Effect	gender	doctor	Estimate	Standard Error	DF	t Value Pr >  t
Intercept			5.2216	2.5708	31	2.03 0.0509
gender	F		2.0847	0.5199	34	4.01 0.0003
gender	M		0	.	.	.
age			0.002997	0.03854	34	0.08 0.9385
doctor		A	0.1585	0.5085	34	0.31 0.7572
doctor		B	0	.	.	.
length			-0.01051	0.004784	34	-2.20 0.0349
visit			-0.04610	0.07144	34	-0.65 0.5230

Null Model Likelihood Ratio Test		
DF	Chi-Square	Pr > ChiSq
3	93.29	<.0001

The model has a good fit since the p-value of the deviance test is below 0.05.

### In R:

```
#specifying reference levels
gender.rel<- relevel(longform.data$gender, ref="M")
doctor.rel<- relevel(longform.data$doctor, ref="B")

#fitting random slope and intercept model
library(nlme)
summary(fitted.model<- lme(score ~ gender.rel + age + doctor.rel
+ length + visit, random = ~ 1 + visit | id, data=longform.data))
```

Random effects:

	StdDev	Corr
(Intercept)	1.5090310	(Intr)
visit	0.3709390	-0.245
Residual	0.2852501	

Fixed effects:

	Value	Std.Error	DF	t-value	p-value
(Intercept)	5.221580	2.5707920	68	2.031117	0.0462
gender.relF	2.084700	0.5198854	31	4.009921	0.0004
age	0.002997	0.0385434	31	0.077755	0.9385
doctor.relA	0.158469	0.5084557	31	0.311668	0.7574
length	-0.010510	0.0047844	68	-2.196820	0.0314
visit	-0.046100	0.0714364	68	-0.645336	0.5209

intervals(fitted.model)

Random Effects:

	lower	est.	upper
sd((Intercept))	1.1473569	1.5090310	1.9847134
sd(visit)	0.2705064	0.3709390	0.5086597
cor((Intercept),visit)	-0.5845323	-0.2446523	0.1682708

#checking model fit

```
null.model<- glm(score ~ gender.rel + age + doctor.rel + length + visit,
data=longform.data)
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))
```

72.63075

```
print(p.value<- pchisq(deviance, df=3, lower.tail = FALSE))
```

1.166255e-15

(d) What parameters of the random terms are significant at the 5% level? Are the scores for each patient correlated?

At the 5% level, the variances of the random slope, intercept, and random error are significant. The covariance between the intercept and slope is not significant. The scores for each patient are correlated and the use of the mixed-effects model is justified.

(e) What fixed-effects variables are significant predictors at the 5% significance level? Write down the fitted model.

Gender and length of the visit are significant predictors of the score at the 5% level. The fitted model is  $\hat{E}(\text{score}) = 5.2216 + 2.0847 \cdot \text{female} + 0.002997 \cdot \text{age} + 0.1585 \cdot \text{doctor A} - 0.01051 \cdot \text{visit length} - 0.04610 \cdot \text{visit number}$ . The other parameters have estimates  $\hat{\sigma}_{u_1}^2 = 2.2772$ ,  $\hat{\sigma}_{u_2}^2 = 0.1376$ ,  $\hat{\sigma}_{u_1 u_2} = -0.1369$ , and  $\hat{\sigma}^2 = 0.08137$ .

(f) Interpret the estimates of the significant regression coefficients.

The estimated mean score for female patients is 2.0847 points larger than that for male patients. As the length of a doctor's visit increases by one minute, the estimated mean score decreases by 0.01051 points.

(g) Predict the quality of service score that would be given by a 55-year-old male patient on his fourth visit to Dr. A with a 30-minute appointment.

Prediction is done according to the following calculation:  $score^0 = 5.2216 + 0.002997 \cdot 55 + 0.1585 - 0.01051 \cdot 30 - 0.04610 \cdot 4 = 5.045235$ .

In SAS:

```
/*using fitted model for prediction*/
data predict;
input id gender$ age doctor$ length visit;
cards;
136 M 55 A 30 4
;

data longform;
set longform predict;
run;

proc mixed covtest;
  class gender doctor;
  model score = gender age doctor length visit / outpm=outdata;
  random intercept visit / subject=id type=un;
run;

proc print data=outdata (firstobs=106) noobs;
  var Pred;
run;
```

**Pred**  
**5.04517**

In R:

```
#using fitted model for prediction
print(predict(fitted.model, data.frame(id=136,gender.rel="M",age=55,
doctor.rel="A",length=30,visit=4),level=0))
```

**5.045167**

**EXERCISE 8.4.** (a) Check that pulse has a normal distribution. Construct a histogram and conduct normality tests.

In SAS:

```
data fitness;
input id gender$ age oxygen1 runtime1 pulse1 oxygen2 runtime2 pulse2
oxygen3 runtime3 pulse3;
cards;
1 F 39 37.4 11.4 151 36.6 17.8 158 36.1 15.4 152
2 M 42 60.1 11.5 121 59.0 9.6 131 58.2 9.0 143
3 F 34 44.6 9.6 138 39.8 9.3 148 38.8 9.1 144
4 M 36 51.9 10.5 125 53.4 9.8 135 50.4 9.6 163
5 F 45 40.8 13.1 142 39.5 12.4 151 38.5 12.7 133
6 M 37 45.4 10.3 133 40.6 11.9 145 40.2 11.2 141
7 F 49 45.3 13.1 135 40.6 12.1 148 39.7 11.5 157
8 F 47 44.8 12.1 135 40.1 12.3 148 39.0 11.9 151
9 M 50 48.7 12.6 131 42.3 11.0 143 44.3 10.5 150
```

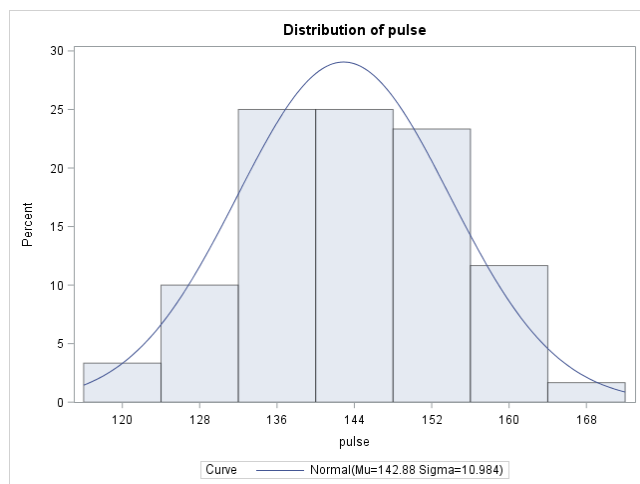
```

10 M 34 45.8 10.8 132 40.9 11.8 144 41.1 11.1 160
11 M 35 50.4 9.6 129 45.9 10.4 137 44.8 10.4 138
12 M 48 50.5 12.9 125 48.6 10.3 135 49.0 9.8 132
13 F 50 44.8 14.0 135 40.3 13.1 148 39.5 12.6 163
14 F 53 39.4 12.7 145 39.3 14.1 154 37.0 12.8 148
15 M 44 46.1 11.0 132 40.9 11.3 144 41.1 10.8 148
16 F 32 39.2 9.1 146 38.7 9.7 158 36.7 10.2 170
17 M 39 54.3 9.4 123 55.1 9.7 132 57.4 9.4 162
18 F 33 39.4 11.6 144 39.4 12.7 154 37.4 12.7 155
19 M 33 47.9 10.1 132 42.2 11.2 143 42.6 10.6 140
20 M 46 49.2 11.2 130 43.9 10.8 141 44.7 10.5 142
;

/*creating longform dataset*/
data longform;
set fitness;
array o[3] oxygen1-oxygen3;
array r[3] runtime1-runtime3;
array p[3] pulse1-pulse3;
do condition=1 to 3;
oxygen=o[condition];
runtime=r[condition];
pulse=p[condition];
output;
end;
keep id gender age oxygen runtime pulse condition;
run;

/*checking normality of response*/
proc univariate;
var pulse;
histogram pulse/normal;
run;

```



### Goodness-of-Fit Tests for Normal Distribution

Test	Statistic	p Value
Kolmogorov-Smirnov D	0.09686645	Pr > D >0.150
Cramer-von Mises	W-Sq 0.05033026	Pr > W-Sq >0.250
Anderson-Darling	A-Sq 0.32350837	Pr > A-Sq >0.250

A normal distribution is supported by the normality tests with p-values above 0.05. The histogram is roughly bell-shaped.

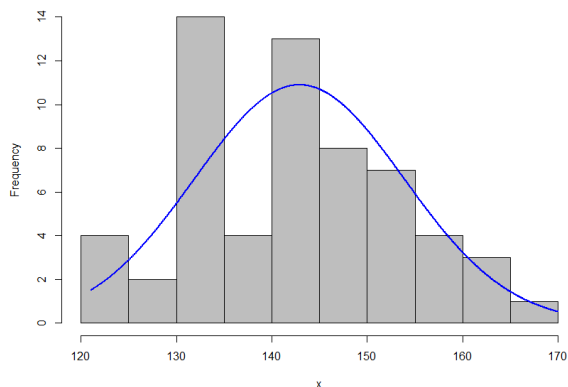
In R:

```
fitness.data<- read.csv(file="C:/./Exercise8.4Data.csv", header=TRUE, sep=",")

#creating longform dataset
library(reshape2)
data1<- melt(fitness.data[,c("id","gender","age","oxygen1","oxygen2","oxygen3")],
id.vars=c("id","gender","age"),variable.name="oxygen.cond",value.name="oxygen")
data2<- melt(fitness.data[,c("runtime1","runtime2","runtime3")],variable.name=
"runtime.cond",value.name="runtime")
data3<- melt(fitness.data[,c("pulse1","pulse2","pulse3")],variable.name=
"pulse.cond",value.name="pulse")
longform.data<- cbind(data1,data2,data3)

#creating variable for condition
condition<- ifelse(longform.data$pulse.cond=="pulse1",1,
ifelse(longform.data$pulse.cond=="pulse2",2,3))

#checking normality of response
library(rcompanion)
plotNormalHistogram(longform.data$pulse)
```



```
shapiro.test(longform.data$pulse)
```

**Shapiro-wilk normality test**

**w = 0.98398, p-value = 0.6173**

(b) Run a random slope and intercept regression model for a pulse. Discuss the model fit.

In SAS:

```
/*fitting random slope and intercept model*/
proc mixed covtest;
  class gender;
  model pulse = gender age oxygen runtime condition / solution;
  random intercept condition / subject=id type=un;
run;
```

Covariance Parameter Estimates					
Cov	Parm	Subject	Estimate	Standard Error	Z Value Pr >  Z
UN(1,1)	id		37.5141	32.7975	1.14 0.1264
UN(2,1)	id		-38.6973	20.9652	-1.85 0.0649
UN(2,2)	id		33.0452	14.6562	2.25 0.0121
Residual			21.0330	6.7194	3.13 0.0009

Solution for Fixed Effects					
Effect	gender	Estimate	Standard Error	DF t Value	Pr >  t
Intercept		172.88	8.7340	17	19.79 <.0001
gender	F	4.7249	1.4230	18	3.32 0.0038
gender	M	0	.	.	.
age		-0.1747	0.1019	18	-1.71 0.1038
oxygen		-0.9634	0.1325	18	-7.27 <.0001
runtime		0.4824	0.5270	18	0.92 0.3721
condition		6.0839	1.4984	19	4.06 0.0007

Null Model Likelihood Ratio Test		
DF	Chi-Square	Pr > ChiSq
3	28.24	<.0001

The model fits the data well as indicated by the small p-value in the deviance test.

In R:

```
#specifying reference level
gender.rel<- relevel(longform.data$gender, ref="M")

#fitting random slope and intercept model
library(nlme)
summary(fitted.model<- lme(pulse ~ gender.rel + age + oxygen + runtime
+ condition, random = ~ 1 + condition | id, control= lmeControl(opt="optim"),
data=longform.data))
```

Random effects:

	StdDev	Corr
(Intercept)	8.007742	(Intr)
condition	6.091466	-0.999
Residual	3.938579	

Fixed effects:

	Value	Std.Error	DF	t-value	p-value
(Intercept)	169.71035	10.731555	37	15.814143	0.0000
gender.relF	4.78210	1.856487	17	2.575887	0.0196
age	-0.19789	0.124495	17	-1.589534	0.1304
oxygen	-0.90924	0.167780	37	-5.419243	0.0000
runtime	0.61422	0.591748	37	1.037967	0.3060
condition	6.19390	1.531663	37	4.043907	0.0003

```
intervals(fitted.model)
```

Random Effects:

	lower	est.	upper
sd((Intercept))	4.945192	8.0077424	12.966926

```
sd(condition)          4.130192  6.0914661  8.984077
cor((Intercept),condition) -1.000000 -0.9992616  0.798489

#checking model fit
null.model<- glm(pulse ~ gender.rel + age + oxygen + runtime + condition,
data=longform.data)
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))

29.29234

print(p.value<- pchisq(deviance, df=3, lower.tail = FALSE))

1.944012e-06
```

Note that in this exercise, SAS and R give slightly different estimates of the model parameters.

(c) Specify the fitted model. What parameters of the random-effects terms are significant at the 5% level? At the 10% level? What fixed-effects terms are significant at the 5% level?

The fitted model in SAS is  $\hat{E}(\text{pulse}) = 172.88 + 4.7249 \cdot \text{female} - 0.1747 \cdot \text{age} - 0.9634 \cdot \text{oxygen} + 0.4824 \cdot \text{runtime} + 6.0839 \cdot \text{condition}$ . The estimates of the other model parameters are  $\hat{\sigma}_{u_1}^2 = 37.5141$ ,  $\hat{\sigma}_{u_2}^2 = 33.0452$ ,  $\hat{\sigma}_{u_1 u_2} = -38.6973$ , and  $\hat{\sigma}^2 = 21.033$ .

The fitted model in R is  $\hat{E}(\text{pulse}) = 169.71035 + 4.7821 \cdot \text{female} - 0.19789 \cdot \text{age} - 0.90924 \cdot \text{oxygen} + 0.61422 \cdot \text{runtime} + 6.1939 \cdot \text{condition}$ . The estimates of the other model parameters are  $\hat{\sigma}_{u_1} = 8.009942$ ,  $\hat{\sigma}_{u_2} = 6.091466$ ,  $\hat{\rho}_{u_1 u_2} = -0.999$ , and  $\hat{\sigma} = 3.938579$ .

At the 5% level, the variance of the random slope is significant. At 10%, the covariance between the random intercept and slope is significant. As for the fixed-effects terms, gender, oxygen intake, and running condition are significant predictors.

(d) Interpret the estimated regression coefficients for the significant fixed-effects terms.

For female runners, the estimated average pulse is 4.78 points larger than that for male runners. As oxygen intake increases by one unit, the estimated mean pulse decreases by 0.96 (0.91) points. As the condition number increases by one, the estimated mean pulse increases by 6.08 (6.19) points.

(e) Predict an average heart rate for a 36-year-old woman who is running on a treadmill, if her oxygen intake is 40.2 units, and her run time is 10.3 minutes per mile.

The following calculations yield the predicted value. Using the model fitted in SAS,

$$\text{pulse}^0 = 172.88 + 4.7249 - 0.1747 \cdot 36 - 0.9634 \cdot 40.2 + 0.4824 \cdot 10.3 + 6.0839 = 143.6396.$$

$$\text{Using the model fitted in R, } \text{pulse}^0 = 169.71035 + 4.7821 - 0.19789 \cdot 36 - 0.90924 \cdot 40.2 + 0.61422 \cdot 10.3 + 6.1939 = 143.3373.$$

In SAS:

```
/*using fitted model for prediction*/
data predict;
input id gender$ age oxygen runtime condition;
```

```

cards;
21 F 36 40.2 10.3 1
;

data longform;
set longform predict;
run;

proc mixed covtest;
class gender;
model pulse = gender age oxygen runtime condition / outpm=outdata;
random intercept condition / subject=id type=un;
run;

proc print data=outdata (firstobs=61) noobs;
var Pred;
run;

```

**Pred**  
**143.643**

**In R:**

```

#using fitted model for prediction
print(predict(fitted.model, data.frame(id=21, gender.rel="F", age=36,
oxygen=40.2, runtime=10.3, condition=1),level=0))

```

**143.3374**

**EXERCISE 8.5.** (a) Verify the normality of the response variable BMI by plotting the histogram and carrying out normality tests.

**In SAS:**

```

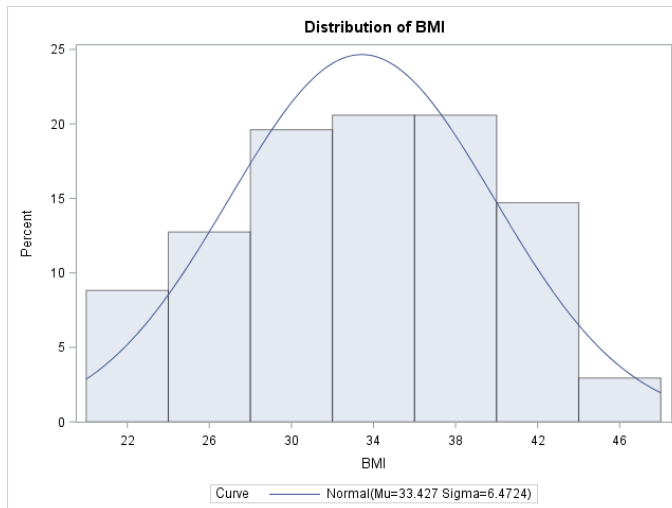
data weightloss;
input id group$ gender$ aexercise aBMI bexercise bBMI cexercise cBMI @@;
cards;
1  Int F 0 42.4 50 40.0 120 36.8 2  Int F 15 32.9 20 30.6 25 28.6
3  Int M 10 32.0 30 30.8 30 26.1 4  Int M 20 26.1 80 25.5 80 21.1
5  Int F 0 27.5 20 26.4 20 22.5 6  Int F 30 40.4 75 38.3 180 32.1
7  Int M 15 33.5 50 28.2 50 25.8 8  Int F 15 35.2 35 34.8 90 30.6
9  Int F 0 39.5 55 37.1 50 35.3 10 Int M 20 27.3 30 26.3 30 22.6
11 Int M 0 46.9 50 43.5 50 40.3 12 Int M 20 34.4 80 32.2 85 28.1
13 Int F 0 34.2 60 31.0 65 26.8 14 Int F 45 26.5 30 24.6 30 20.8
15 Int F 0 29.6 20 28.2 20 24.9 16 Int F 10 31.2 80 29.3 50 28.6
17 Cnt F 0 29.3 25 28.9 30 26.3 18 Cnt M 20 45.9 10 43.1 15 42.9
19 Cnt M 0 41.5 20 38.8 30 39.9 20 Cnt F 30 33.3 25 33.4 35 33.2
21 Cnt M 15 31.1 35 30.9 0 30.9 22 Cnt F 10 43.3 35 43.6 30 44.5
23 Cnt M 15 35.5 0 36.5 5 35.3 24 Cnt F 10 42.4 15 43.4 50 42.3
25 Cnt F 20 37.0 30 36.6 45 35.5 26 Cnt M 0 37.8 30 35.7 45 34.3
27 Cnt F 20 23.7 10 23.1 0 23.7 28 Cnt F 10 38.7 15 20.4 25 20.1
29 Cnt F 0 41.2 15 41.2 55 39.7 30 Cnt F 30 30.2 35 29.9 5 29.4
31 Cnt M 10 38.4 20 38.1 30 37.0 32 Cnt F 10 37.5 15 37.4 5 36.8
33 Cnt M 30 34.5 10 34.4 20 33.9 34 Cnt M 15 37.6 35 36.2 25 36.0

```

```
;

/*creating longform dataset*/
data longform;
set weightloss;
array m[3] (0 1 3);
array e[3] aexercise bexercise cexercise;
array b[3] aBMI bBMI cBMI;
do i=1 to 3;
  month=m[i];
  exercise=e[i];
  BMI=b[i];
  output;
end;
keep id group gender exercise BMI month;
run;

/*checking normality of response*/
proc univariate;
  var BMI;
  histogram BMI/normal;
run;
```



### Goodness-of-Fit Tests for Normal Distribution

Test	Statistic	p value
Kolmogorov-Smirnov D	0.04908222	Pr > D >0.150
Cramer-von Mises	W-Sq 0.05735715	Pr > W-Sq >0.250
Anderson-Darling	A-Sq 0.37849800	Pr > A-Sq >0.250

On the graph, we see a roughly bell-shaped histogram. In addition, the p-values in the normality tests are above 0.05, supporting the normality of BMI.

In R:

```
weightloss.data<- read.csv(file="C:/./Exercise8.5Data.csv", header=TRUE, sep=",")

#creating longform dataset
library(reshape2)
data1<- melt(weightloss.data[,c("id","group","gender","aexercise","bexercise",
```

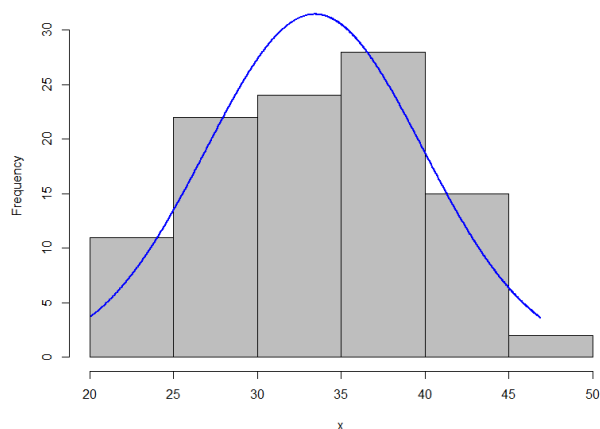
```

"cexercise")], id.vars=c("id","group","gender"),variable.name="exercise.visit",
value.name="exercise")
data2<- melt(weightloss.data[,c("aBMI","bBMI","cBMI")],variable.name=
"BMI.visit",value.name="BMI")
longform.data<- cbind(data1,data2)

#creating variable for time
month<- ifelse(longform.data$BMI.visit=="aBMI",0,
ifelse(longform.data$BMI.visit=="bBMI",1,3))

#checking normality of response
library(rcompanion)
plotNormalHistogram(longform.data$BMI)

```



```
shapiro.test(longform.data$BMI)
```

Shapiro-wilk normality test

w = 0.98317, p-value = 0.2216

(b) Fit the random slope and intercept model. How good is the model fit?

In SAS:

```

/*fitting random slope and intercept model*/
proc mixed covtest;
  class group gender(ref="F");
  model BMI = group gender exercise month / solution;
  random intercept month / subject=id type=un;
run;

```

Covariance Parameter Estimates					
Cov	Parm	Subject	Estimate	Standard Error	Z Value Pr Z
UN(1,1)	id		29.2800	8.1566	3.59 0.0002
UN(2,1)	id		2.5538	1.2409	2.06 0.0396
UN(2,2)	id		0.3306	0.3190	1.04 0.1501
Residual			3.4357	0.8496	4.04 <.0001

Solution for Fixed Effects						
Effect	group	gender	Estimate	Standard Error	DF	t Value Pr >  t
Intercept			34.5855	1.5561	31	22.23 <.0001
group	Cnt		1.1961	1.8719	33	0.64 0.5273
group	Int		0	.	.	.
gender		M	1.2370	1.8969	33	0.65 0.5189
gender		F	0	.	.	.
exercise			-0.03974	0.01121	33	-3.54 0.0012
month			-0.8445	0.2029	33	-4.16 0.0002

Null Model Likelihood Ratio Test		
DF	Chi-Square	Pr > ChiSq
3	115.71	<.0001

The model has a very good fit because of a very small p-value in the deviance test.

In R:

```
#specifying reference levels
group.rel<- relevel(longform.data$group, ref="Int")
gender.rel<- relevel(longform.data$gender, ref="F")

#fitting random slope and intercept model
library(nlme)
summary(fitted.model<- lme(BMI ~ group.rel + gender.rel + exercise + month,
random = ~ 1 + month | id, data=longform.data))
```

Random effects:

	StdDev	Corr
(Intercept)	5.4112519	(Intr)
month	0.5749658	0.821
Residual	1.8535182	

Fixed effects:

	value	Std.Error	DF	t-value	p-value
(Intercept)	34.58554	1.5561343	66	22.225291	0.0000
group.relCnt	1.19608	1.8719456	31	0.638949	0.5275
gender.relM	1.23698	1.8969761	31	0.652082	0.5192
exercise	-0.03974	0.0112112	66	-3.544454	0.0007
month	-0.84454	0.2028822	66	-4.162726	0.0001

```
intervals(fitted.model)
```

Random Effects:

	lower	est.	upper
sd((Intercept))	4.1184307	5.4112519	7.1099041
sd(month)	0.2234017	0.5749658	1.4797814
cor((Intercept),month)	-0.9626376	0.8207878	0.9996312

```
#checking model fit
null.model<- glm(BMI ~ gender.rel + group.rel + exercise + month,
data=longform.data)
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))
```

115.6489

```
print(p.value<- pchisq(deviance, df=3, lower.tail = FALSE))
```

(c) Present the fitted model and specify all estimated parameters. Discuss the significance of the parameters at the 5% significance level.

The fitted model is of the form  $\hat{E}(BMI) = 34.5855 + 1.1961 \cdot control + 1.237 \cdot male - 0.03974 \cdot exercise - 0.8445 \cdot month$ . The estimates of the other model parameters are  $\hat{\sigma}_{u_1}^2 = 29.28$ ,  $\hat{\sigma}_{u_2}^2 = 0.3306$ ,  $\hat{\sigma}_{u_1 u_2} = 2.5538$ , and  $\hat{\sigma}^2 = 3.4357$ .

At the 5% level of significance, the length of daily exercise and month are significant predictors, and the variance of the random intercept and the covariance between intercept and slope are significant, validating the need for random-effect terms in the model.

(d) Give an interpretation of the estimated significant beta coefficients. Is the intervention efficient?

As the length of daily exercise increases by one minute, the estimated average BMI decreases by 0.03974 units. It is estimated that the average BMI decreases by 0.8445 units for every additional month in the study.

(e) Compute the predicted BMI at 3 months for an intervention group female participant, if she exercises for 1 hour every day.

Predicted BMI is  $BMI^0 = 34.5855 - 0.03974 \cdot 60 - 0.8445 \cdot 3 = 29.6676$ .

In SAS:

```
/*using fitted model for prediction*/
data predict;
input id group$ gender$ exercise month;
cards;
35 Int F 60 3
;

data longform;
set longform predict;
run;

proc mixed covtest;
  class group gender;
  model BMI = group gender exercise month / outpm=outdata;
  random intercept month / subject=id type=un;
run;

proc print data=outdata (firstobs=103) noobs;
  var Pred;
run;

      Pred
29.6676
```

In R:

```
#using fitted model for prediction
```

```
print(predict(fitted.model, data.frame(id=35, gender.rel="F", group.rel="Int",
exercise=60, month=3),level=0))
```

29.66766

**EXERCISE 8.6.** Consider the data in Exercise 8.2. Answer the questions below.

(a) Fit random slope and intercept models with unstructured, Toeplitz, spatial power, autoregressive, compound symmetric, and independent covariance matrices for error terms. Present the AIC, AICC, and BIC criteria values for those models that converge. If the random slope and intercept model doesn't converge, try to fit a random intercept-only model.

In SAS:

```
data deptstore;
input id totalyears status$ bonus18 bonus19 bonus20 @@;
cards;
1 16 full 1482 1508 1543 2 7 part 673 710 895
3 11 full 933 1351 1440 4 8 part 844 958 1196
5 6 part 564 790 815 6 5 full 601 708 780
7 6 part 775 822 902 8 17 full 1209 1297 1475
9 12 full 929 1008 1255 10 9 full 983 1013 1111
11 11 full 909 1004 1084 12 6 part 387 853 999
13 4 part 476 530 627 14 6 full 780 843 925
15 10 full 717 1200 1399
;

/*creating longform dataset*/
data longform;
set deptstore;
array y[3] (1.8 1.9 2.0);
array b[3] bonus18-bonus20;
do i=1 to 3;
year=y[i];
bonus=b[i];
output;
end;
keep id totalyears status year bonus;
run;

/*fitting random slope and intercept model with
unstructured covariance matrix of error terms*/
proc mixed covtest;
class status;
model bonus = totalyears status year / solution;
random intercept year / subject=id type=un;
repeated / subject=id type=un;
run;
```

#### Fit Statistics

AIC (Smaller is Better) 566.3

AICC (Smaller is Better) 572.1

BIC (Smaller is Better) 572.7

```
/*fitting random slope and intercept model with
Toeplitz covariance matrix of error terms*/
```

```
proc mixed covtest;
  class status;
  model bonus = totalyears status year / solution;
  random intercept year / subject=id type=un;
  repeated / subject=id type=toep;
run;
```

#### Fit Statistics

AIC (Smaller is Better) 524.4  
 AICC (Smaller is Better) 526.9  
 BIC (Smaller is Better) 528.6

```
/*fitting random slope and intercept model with
spatial power covariance matrix of error terms*/
proc mixed covtest;
  class status;
  model bonus = totalyears status year / solution;
  random intercept year / subject=id type=un;
  repeated / subject=id type=sp(pow)(year) r;
run;
```

**WARNING: Did not converge.**

```
/*fitting random intercept-only model with spatial
power covariance matrix of error terms*/
proc mixed covtest;
  class status;
  model bonus = totalyears status year / solution;
  random intercept / subject=id type=un;
  repeated / subject=id type=sp(pow)(year);
run;
```

#### Fit Statistics

AIC (Smaller is Better) 518.3  
 AICC (Smaller is Better) 518.6  
 BIC (Smaller is Better) 519.7

```
/*fitting random slope and intercept model with
autoregressive covariance matrix of error terms*/
proc mixed covtest;
  class status;
  model bonus = totalyears status year / solution;
  random intercept year / subject=id type=un;
  repeated /subject=id type=ar(1);
run;
```

#### Fit Statistics

AIC (Smaller is Better) 520.4  
 AICC (Smaller is Better) 521.6  
 BIC (Smaller is Better) 523.3

```
/*fitting random slope and intercept model with
compound symmetric covariance matrix of error terms*/
proc mixed covtest;
  class status;
  model bonus = totalyears status year / solution;
  random intercept year / subject=id type=un;
```

```
repeated / subject=id type=cs;
run;
```

#### Fit Statistics

```
AIC (Smaller is Better) 523.3
AICC (Smaller is Better) 525.0
BIC (Smaller is Better) 526.9
```

```
/*fitting random slope and intercept model with
independent covariance matrix of error terms*/
proc mixed covtest;
class status;
model bonus = totalyears status year / solution;
random intercept year / subject=id type=un;
run;
```

#### Fit Statistics

```
AIC (Smaller is Better) 521.3
AICC (Smaller is Better) 522.4
BIC (Smaller is Better) 524.1
```

#### In R:

```
deptstore.data<- read.csv(file="C:/./Exercise8.2Data.csv", header=TRUE, sep=",")
```

```
#creating longform dataset
library(reshape2)
longform.data<- melt(deptstore.data, id.vars=c("id","totalyears","status"),
variable.name="bonus.year", value.name="bonus")
```

```
#creating variable for time
year<- ifelse(longform.data$bonus.year=="bonus18",1.8,
ifelse(longform.data$bonus.year=="bonus19",1.9,2.0))
```

```
#rescaling response and creating reference level
status.rel<- relevel(longform.data$status, ref="part")
```

```
#fitting random slope and intercept model with
#unstructured covariance matrix of error terms
library(nlme)
summary(un.fitted.model<- lme(bonus ~ totalyears + status.rel + year,
random = ~ 1 + year | id, data=longform.data, correlation=corSymm(),
weights=varIdent(form = ~ id | year)))
```

```
      AIC      BIC
529.393 551.6694
```

```
#computing AICC
n<- 45
p<- 14
print(AICC<- -2*logLik(un.fitted.model)+2*p*n/(n-p-1))
```

```
545.393
```

```
#fitting random slope and intercept model with
#Toeplitz covariance matrix of error terms
summary(toep.fitted.model<- lme(bonus ~ totalyears + status.rel + year,
random = ~ 1 + year| id, data=longform.data,
```

```
correlation=corARMA(form = ~ 1 | id, p=1, q=1))
```

```

      AIC      BIC
532.5445 549.6802

```

```

#computing AICC
p<-11
print(AICC<- -2*logLik(toep.fitted.model)+2*p*n/(n-p-1))

```

```
542.5445
```

```

#fitting random slope and intercept model with
#spatial power covariance matrix of error terms
summary(sppow.fitted.model<- lme(bonus ~ totalyears + status.rel + year,
random = ~ 1 + year | id, data=longform.data,
correlation=corCAR1(form = ~ 1 | id)))

```

```

      AIC      BIC
532.8405 548.2627

```

```

#computing AICC
p<- 10
print(AICC<- -2*logLik(sppow.fitted.model)+2*p*n/(n-p-1))

```

```
541.3111
```

```

#fitting random intercept-only model with
#spatial power covariance matrix of error terms
summary(sppowint.fitted.model<- lme(bonus ~ totalyears + status.rel + year,
random = ~ 1 | id, data=longform.data,
correlation=corCAR1(form = ~ 1 | id)))

```

```

      AIC      BIC
528.8405 540.8355

```

```

#computing AICC
p<- 8
print(AICC<- -2*logLik(sppowint.fitted.model)+2*p*n/(n-p-1))

```

```
534.8405
```

```

#fitting random slope and intercept model with
#autoregressive covariance matrix of error terms
summary(ar.fitted.model<- lme(bonus ~ totalyears + status.rel + year,
random = ~ 1 + year | id, data=longform.data,
correlation=corAR1(form = ~ 1 | id)))

```

```

      AIC      BIC
532.8405 548.2627

```

```

#computing AICC
p<- 10
print(AICC<- -2*logLik(ar.fitted.model)+2*p*n/(n-p-1))

```

```
541.3111
```

```

#fitting random slope and intercept model with
#compound symmetric covariance matrix of error terms
summary(cs.fitted.model<- lme(bonus ~ totalyears + status.rel + year,
random = ~ 1 + year | id, data=longform.data,
correlation=corCompSymm(form = ~ 1 | id)))

```

AIC BIC  
531.3133 546.7355

```
#computing AICC
p<- 10
print(AICC<- -2*logLik(cs.fitted.model)+2*p*n/(n-p-1))
```

539.7839

```
fitting random slope and intercept model with
#independent covariance matrix of error terms
summary(ind.fitted.model<- lme(bonus ~ totalyears + status.rel+ year,
random = ~ 1 + year | id, data=longform.data))
```

AIC BIC  
529.3133 543.0219

```
#computing AICC
p<- 8
print(AICC<- -2*logLik(ind.fitted.model)+2*p*n/(n-p-1))
```

533.3133

(b) Find the optimal model with respect to the AIC, AICC, and BIC criteria, and answer questions (c)-(e) in Exercise 8.2, using the best-fitted model.

For the models fitted in SAS, the one with the smallest AIC, AICC, and BIC is the random intercept-only model with the spatial power covariance matrix of the error terms. The values are summarized below.

	UN	Toeplitz	Sp Power Intercept	AR	CS	Ind
<b>AIC</b>	566.3	524.4	518.3	520.4	523.3	521.3
<b>AICC</b>	572.1	526.9	518.6	521.6	525	522.4
<b>BIC</b>	572.7	528.6	519.7	523.3	526.9	524.1

For the models fitted in R, according to the AICC criterion, the optimal model is the one with the independent covariance matrix of the error terms. The random intercept-only model with a spatial power covariance structure has the best fit according to the AIC and BIC criteria. The values are summed up here:

	UN	Toeplitz	Sp Power	Sp Power Intercept	AR	CS	Ind
<b>AIC</b>	529.4	532.5	532.8	528.8	532.8	531.3	529.3
<b>AICC</b>	545.4	542.5	541.3	534.8	541.3	539.8	533.3
<b>BIC</b>	551.7	549.7	548.3	540.8	548.3	546.7	543.0

Finally, we answer questions (c)-(e) in Exercise 8.2, using the best-fitted model.

In SAS:

```
/*fitting random intercept-only model with spatial
power covariance matrix of error terms*/
```

```
proc mixed covtest;
  class status;
  model bonus = totalyears status year / solution;
  random intercept / subject=id type=un;
  repeated / subject=id type=sp(pow)(year) r;
run;
```

#### Estimated R Matrix for Subject 1

Row	Col1	Col2	Col3
1	17533	10081	5184.54
2	10081	17533	10081
3	5184.54	10081	17533

#### Covariance Parameter Estimates

Cov	Parm	Subject	Estimate	Standard Error	Z Value	Pr >  Z
UN(1,1)	id		0	.	.	.
SP(POW)	id		-0.00652	0.01657	-0.39	0.6938
Residual			17533	5112.34	3.43	0.0003

#### Solution for Fixed Effects

Effect	status	Estimate	Standard Error	DF	t Value	Pr >  t
Intercept		-2255.12	391.89	12	-5.75	<.0001
totalyears		60.0189	9.0505	29	6.63	<.0001
status	full	47.2877	69.6058	29	0.68	0.5023
status	part	0	.	.	.	.
year		1394.67	202.89	29	6.87	<.0001

The fitted model is  $\hat{E}(\text{bonus}) = -2255.12 + 60.0189 \cdot \text{total years} + 47.2877 \cdot \text{full-time employee} + 1394.67 \cdot \text{year}/10$ . The random intercept has an estimated variance of zero, that is, the mixed-effect terms are non-existent. The estimated covariance matrix of the error terms is a block diagonal, with 15 blocks of the form  $\begin{pmatrix} 17533 & 10081 & 5184.54 \\ 10081 & 17533 & 10081 \\ 5184.54 & 10081 & 17533 \end{pmatrix}$ .

Total years with the company and the year the bonus was recorded are significant predictors of the bonus. As the total number of years with the company increases by one, the estimated average bonus increases by \$60.0189. Every year, the estimated average increase in bonus is \$139.467.

In R:

```
#fitting random intercept-only model with
#spatial power covariance matrix of error terms
summary(sppowint.fitted.model<- lme(bonus ~ totalyears + status.rel + year,
  random = ~ 1 | id, data=longform.data, correlation=corCAR1(form = ~ 1 | id)))
```

#### Random effects:

```
(Intercept) Residual
StdDev: 0.01464535 130.9026
```

```
Phi
0.5467992
```

#### Fixed effects:

```
value Std.Error DF t-value p-value
(Intercept) -2253.4538 386.4705 29 -5.830856 0.0000
```

totalyears	59.8379	8.9004	12	6.723065	0.0000
status.relfull	48.2277	68.4510	12	0.704558	0.4945
year	1394.6667	200.1013	29	6.969804	0.0000

```
intervals(sppowint.fitted.model)
```

```
Random Effects:
              lower      est.      upper
sd((Intercept)) 1.224958e-49 0.01464535 1.750969e+45
```

```
getVarCov(sppowint.fitted.model, type="conditional")
```

Conditional variance covariance matrix

	1	2	3
1	17135.0	9369.7	5123.3
2	9369.7	17135.0	9369.7
3	5123.3	9369.7	17135.0

To predict the amount of bonus in 2021 for a full-time employee who has been with the company for 7 years, we compute  $\text{bonus}^0 = -2255.12 + 60.0189 \cdot 7 + 47.2877 + 1394.67 \cdot 2.1 = \$1,141.107$ .

In SAS:

```
data predict;
input id totalyears status$ year;
cards;
21 7 full 2.1
;

data longform;
set longform predict;
run;

proc mixed covtest;
class status;
model bonus = totalyears status year / outpm=outdata;
random intercept / subject=id type=un;
repeated / subject=id type=sp(pow)(year);
run;

proc print data=outdata (firstobs=46) noobs;
var Pred;
run;
```

```
Pred
1141.10
```

In R:

```
print(predict(sppowint.fitted.model, data.frame(totalyears=7, status.rel="full",
year=2.1), level=0))
```

```
1142.439
```

**EXERCISE 8.7.** For the data in Exercise 8.3,

(a) Fit random slope and intercept models with unstructured, Toeplitz, spatial power, autoregressive, compound symmetric, and independent covariance matrices for error terms, whichever converge. Try to fit a random intercept-only model if convergence criteria are not met.

In SAS:

```
data orthoclinic;
input id gender$ age doctor$ length1 length2 length3 score1 score2 score3 @@;
cards;
101 F 78 A 25 20 25 7.1 7.5 7.6 102 F 63 A 30 30 40 5.5 5.8 6.1
103 F 62 A 10 15 10 10.0 10.0 9.8 104 F 71 B 15 15 40 7.8 7.3 7.5
105 M 68 A 40 60 40 3.5 3.5 3.0 106 F 63 A 25 15 20 8.5 8.7 8.8
107 F 60 B 25 35 25 6.7 5.7 6.5 108 F 70 A 20 20 20 9.0 8.3 8.2
109 F 57 A 30 20 15 8.4 7.8 8.1 110 F 59 B 25 30 15 7.1 7.4 7.9
111 M 62 A 50 30 70 3.0 3.2 2.6 112 M 58 A 20 15 45 6.1 6.8 6.9
113 M 75 A 25 35 30 5.7 5.6 4.7 114 M 76 B 35 50 25 4.9 5.4 5.2
115 F 75 A 15 20 25 8.2 8.9 8.2 116 M 57 A 45 30 40 4.6 3.9 3.2
117 F 68 A 35 25 40 3.8 4.8 5.3 118 M 65 B 40 40 25 3.9 3.9 4.7
119 F 67 B 20 15 30 6.5 7.2 6.6 120 F 60 B 25 15 15 7.3 7.1 7.8
121 F 67 A 15 20 15 7.7 8.0 8.3 122 F 57 B 10 15 15 9.8 9.2 8.6
123 M 62 B 55 60 75 3.4 2.7 2.3 124 M 71 A 20 30 25 7.1 6.6 7.4
125 M 71 B 15 15 20 8.8 9.1 9.3 126 M 64 A 25 30 30 5.6 6.3 6.3
127 M 51 A 35 40 30 5.1 4.6 3.9 128 F 70 B 35 25 15 6.8 7.1 7.6
129 M 61 A 35 40 50 5.5 5.2 4.8 130 M 62 B 60 40 65 3.7 3.4 2.4
131 F 68 A 20 35 35 5.3 5.6 4.9 132 F 68 B 35 30 15 7.2 6.2 5.6
133 M 64 B 40 20 30 5.4 4.9 4.5 134 F 76 B 30 45 25 5.5 4.7 4.6
135 F 78 B 25 20 15 7.6 8.3 9.2
;

/*creating longform dataset*/
data longform;
set orthoclinic;
array l[3] length1-length3;
array s[3] score1-score3;
do visit=1 to 3;
length=l[visit];
score=s[visit];
output;
end;
keep id gender age doctor visit length score;
run;

/*fitting random slope and intercept model with
unstructured covariance matrix of error terms*/
proc mixed covtest;
class gender doctor;
model score = gender age doctor length visit / solution;
random intercept visit / subject=id type=un;
repeated / subject=id type=un;
run;
```

#### Fit Statistics

AIC (Smaller is Better) 266.8  
AICC (Smaller is Better) 268.4  
BIC (Smaller is Better) 279.2

```

/*fitting random slope and intercept model with
Toeplitz covariance matrix of error terms*/
proc mixed covtest;
  class gender doctor;
  model score = gender age doctor length visit / solution;
  random intercept visit / subject=id type=un;
  repeated / subject=id type=toep;
run;

```

**WARNING: Did not converge.**

```

/*fitting random intercept-only model with
Toeplitz covariance matrix of error terms*/
proc mixed covtest;
  class gender doctor;
  model score = gender age doctor length visit / solution;
  random intercept / subject=id type=un;
  repeated / subject=id type=toep;
run;

```

#### Fit Statistics

AIC (Smaller is Better) 261.5  
 AICC (Smaller is Better) 261.9  
 BIC (Smaller is Better) 267.7

```

/*fitting random slope and intercept model with
spatial power covariance matrix of error terms*/
proc mixed covtest;
  class gender doctor;
  model score = gender age doctor length visit / solution;
  random intercept visit / subject=id type=un;
  repeated / subject=id type=sp(pow)(visit);
run;

```

**WARNING: Did not converge.**

```

/*fitting random intercept-only model with
spatial power covariance matrix of error terms*/
proc mixed covtest;
  class gender doctor;
  model score = gender age doctor length visit / solution;
  random intercept / subject=id type=un;
  repeated / subject=id type=sp(pow)(visit);
run;

```

#### Fit Statistics

AIC (Smaller is Better) 258.6  
 AICC (Smaller is Better) 258.7  
 BIC (Smaller is Better) 261.7

```

/*fitting random slope and intercept model with
autoregressive covariance matrix of error terms*/
proc mixed covtest;
  class gender doctor;
  model score = gender age doctor length visit / solution;
  random intercept visit / subject=id type=un;
  repeated / subject=id type=ar(1);
run;

```

**WARNING: Did not converge.**

```
/*fitting random intercept-only model with
autoregressive covariance matrix of error terms*/
proc mixed covtest;
  class gender doctor;
  model score = gender age doctor length visit / solution;
  random intercept / subject=id type=un;
  repeated / subject=id type=ar(1);
run;
```

#### **Fit Statistics**

**AIC (Smaller is Better) 258.6**

**AICC (Smaller is Better) 258.7**

**BIC (Smaller is Better) 261.7**

```
/*fitting random slope and intercept model with
compound symmetric covariance matrix of error terms*/
proc mixed covtest;
  class gender doctor;
  model score = gender age doctor length visit / solution;
  random intercept visit / subject=id type=un;
  repeated / subject=id type=cs;
run;
```

#### **Fit Statistics**

**AIC (Smaller is Better) 261.1**

**AICC (Smaller is Better) 261.7**

**BIC (Smaller is Better) 268.8**

```
/*fitting random slope and intercept model with
independent covariance matrix of error terms*/
proc mixed covtest;
  class gender doctor;
  model score = gender age doctor length visit / solution;
  random intercept visit / subject=id type=un;
run;
```

#### **Fit Statistics**

**AIC (Smaller is Better) 259.1**

**AICC (Smaller is Better) 259.5**

**BIC (Smaller is Better) 265.3**

**In R:**

```
orthoclinic.data<- read.csv(file="C:/./Exercise8.3Data.csv", header=TRUE,
sep=",")
```

```
#creating longform dataset
library(reshape2)
data1<- melt(orthoclinic.data[,c("id","gender","age","doctor", "length1",
"length2", "length3")], id.vars=c("id","gender","age","doctor"), variable.name
="length.visit",value.name="length")
data2<- melt(orthoclinic.data[,c("score1","score2","score3")], variable.name
="score.visit", value.name="score")
longform.data<- cbind(data1,data2)
```

```
#creating variable for time
visit<- ifelse(longform.data$score.visit=="score1",1,
ifelse(longform.data$score.visit=="score2",2,3))

#specifying reference levels
gender.rel<- relevel(longform.data$gender, ref="M")
doctor.rel<- relevel(longform.data$doctor, ref="B")

#fitting random slope and intercept model with
#unstructured covariance matrix of error terms
library(nlme)
summary(un.fitted.model<- lme(score ~ gender.rel + age + doctor.rel + length +
visit, random = ~ 1 + visit | id, control= lmeControl(opt="optim"),
data=longform.data, correlation=corSymm(), weights=varIdent(form=~id|length)))
```

The model doesn't converge.

```
#fitting random intercept-only model with
#unstructured covariance matrix of error terms
library(nlme)
summary(un.fitted.model<- lme(score ~ gender.rel + age + doctor.rel + length +
visit, random = ~ 1 | id, control= lmeControl(opt="optim"),
data=longform.data,correlation=corSymm(), weights=varIdent(form=~id|length)))
```

The model doesn't converge.

```
#fitting random slope and intercept model with
#Toeplitz covariance matrix of error terms
summary(toep.fitted.model<- lme(score ~ gender.rel + age + doctor.rel + length +
visit, random = ~ 1 + visit | id, data=longform.data,
correlation = corARMA(form = ~ 1 | id, p=1, q=1)))
```

AIC	BIC
274.9041	306.0456

```
#computing AICC
n<-105
p<- 12
print(AICC<- -2*logLik(toep.fitted.model)+2*p*n/(n-p-1))
```

278.2954

```
#fitting random slope and intercept model with
#spatial power covariance matrix of error terms
summary(sppow.fitted.model<- lme(score ~ gender.rel + age + doctor.rel + length +
visit, random = ~ 1 + visit| id, data=longform.data,
correlation=corCAR1(form = ~ 1 | id)))
```

AIC	BIC
273.0585	301.6048

```
#computing AICC
p<- 11
print(AICC<- -2*logLik(sppow.fitted.model)+2*p*n/(n-p-1))
```

275.8972

```
#fitting random slope and intercept model with
#autoregressive covariance matrix of error terms
summary(ar.fitted.model<- lme(score ~ gender.rel + age + doctor.rel + length +
visit, random = ~ 1 + visit| id, data=longform.data,
```

```
correlation=corAR1(form = ~ 1 | id))
```

	AIC	BIC
	272.9041	301.4504

```
#computing AICC
p<- 11
print(AICC<- -2*logLik(ar.fitted.model)+2*p*n/(n-p-1))
```

275.7428

```
#fitting random intercept-only model with
#autoregressive covariance matrix of error terms
summary(arint.fitted.model<- lme(score ~ gender.rel + age + doctor.rel + length +
visit, random = ~ 1 | id, data=longform.data, correlation=corAR1(form = ~ 1 |
id)))
getVarCov(arint.fitted.model, type="conditional")
```

	AIC	BIC
	272.612	295.9681

```
#computing AICC
p<- 9
print(AICC<- -2*logLik(arint.fitted.model)+2*p*n/(n-p-1))
```

274.5067

```
#fitting random slope and intercept model with
#compound symmetric covariance matrix of error terms
summary(cs.fitted.model<- lme(score ~ gender.rel + age + doctor.rel + length +
visit, random = ~ 1 + visit | id, data=longform.data,
correlation=corCompSymm(form = ~ 1 | id)))
```

	AIC	BIC
	273.0585	301.6048

```
#computing AICC
p<- 11
print(AICC<- -2*logLik(cs.fitted.model)+2*p*n/(n-p-1))
```

275.8972

```
#fitting random slope and intercept model with
#independent covariance matrix of error terms
summary(ind.fitted.model<- lme(score ~ gender.rel + age + doctor.rel + length +
visit, random = ~ 1 + visit | id, data=longform.data))
```

	AIC	BIC
	271.0585	297.0097

```
#computing AICC
p<- 10
print(AICC<- -2*logLik(ind.fitted.model)+2*p*n/(n-p-1))
```

273.3989

(b) Which of the fitted models has the best fit according to the AIC, AICC, and BIC criteria?

In SAS, the spatial power (the same as the autoregressive) model has the best fit. The AIC, AICC, and BIC values are given in the table below.

	UN	Toeplitz	Sp Power Intercept	AR Intercept	CS	Ind
<b>AIC</b>	266.8	261.5	258.6	258.6	261.1	259.1
<b>AICC</b>	268.4	261.9	258.7	258.7	261.7	259.5
<b>BIC</b>	279.2	267.7	261.7	261.7	268.8	265.3

According to R, the model with an independent structure has a better fit, but the next candidate is the random intercept-only autoregressive model. Note that in R, spatial power and autoregressive models are different (due possibly to dubious model convergence).

	Toeplitz	Sp Power	AR	AR Intercept	CS	Ind
<b>AIC</b>	274.9	273.1	272.9	272.6	273.1	271.1
<b>AICC</b>	278.3	275.9	275.7	274.5	275.9	273.4
<b>BIC</b>	306.0	301.6	301.5	296.0	301.6	297.0

(c) Answer parts (e)-(g) in Exercise 8.3 as applied to the best-fitted model.

In SAS:

```
/*fitting random intercept-only model with
autoregressive covariance matrix of error terms*/
proc mixed covtest;
  class gender doctor;
  model score = gender age doctor length visit / solution;
  random intercept / subject=id type=un;
  repeated / subject=id type=ar(1) r;
run;
```

Estimated R Matrix for Subject 1

Row	Col1	Col2	Col3
1	2.4954	2.3447	2.2031
2	2.3447	2.4954	2.3447
3	2.2031	2.3447	2.4954

Covariance Parameter Estimates

Cov	Parm	Subject	Estimate	Standard Error	Z Value	Pr >  Z
UN(1,1)	id		0	.	.	.
AR(1)	id		0.9396	0.01817	51.72	<.0001
Residual			2.4954	0.6142	4.06	<.0001

Solution for Fixed Effects

Effect	gender	doctor	Estimate	Standard Error	DF	t Value	Pr >  t
Intercept			4.6863	2.6628	31	1.76	0.0883
gender	F		2.2297	0.5389	68	4.14	<.0001
gender	M		0	.	.	.	.
age			0.01012	0.03984	68	0.25	0.8002
doctor		A	0.1670	0.5250	68	0.32	0.7513

Solution for Fixed Effects						
Effect	gender	doctor	Estimate	Standard Error	DF	t Value Pr >  t
doctor	B		0	.	.	.
length			-0.01130	0.005149	68	-2.19 0.0317
visit			-0.04559	0.06470	68	-0.70 0.4834

The fitted model has the form  $\hat{E}(\text{score}) = 4.6863 + 2.2297 \cdot \text{female} + 0.01012 \cdot \text{age} + 0.167 \cdot \text{doctor A} - 0.0113 \cdot \text{visit length} - 0.04559 \cdot \text{visit number}$ . The estimated variance of the random intercept is zero, and the fitted covariance matrix of the error terms is a block diagonal, with 35 blocks of the form  $\begin{pmatrix} 2.4954 & 2.3447 & 2.2031 \\ 2.3447 & 2.4954 & 2.3447 \\ 2.2031 & 2.3447 & 2.4954 \end{pmatrix}$ . Gender and visit length are significant predictors of the score at the 5% level. For female patients, the estimated average score is 2.2297 points higher than that for male patients. As visit length increases by one minute, the estimated mean score decreases by 0.0113.

In R:

```
#fitting random intercept-only model with
#autoregressive covariance matrix of error terms
summary(arint.fitted.model<- lme(score ~ gender.rel + age + doctor.rel + length
+ visit, random = ~ 1 | id, data=longform.data, correlation=corAR1(form = ~ 1 |
id)))
```

Random effects:

```
(Intercept) Residual
StdDev: 0.0005791566 1.57967
```

```
Phi
0.9396084
```

Fixed effects:

```
Value Std.Error DF t-value p-value
(Intercept) 4.686268 2.6627508 68 1.759935 0.0829
gender.relF 2.229715 0.5388793 31 4.137689 0.0002
age 0.010122 0.0398445 31 0.254043 0.8011
doctor.relA 0.167042 0.5250427 31 0.318150 0.7525
length -0.011298 0.0051494 68 -2.194026 0.0317
visit -0.045594 0.0647041 68 -0.704658 0.4834
```

```
intervals(arint.fitted.model)
```

Random Effects:

```
lower est. upper
sd((Intercept)) 7.847905e-63 0.0005791566 4.274037e+55
```

```
getVarCov(arint.fitted.model, type="conditional")
```

Conditional variance covariance matrix

```
1 2 3
1 2.4954 2.3447 2.2031
2 2.3447 2.4954 2.3447
3 2.2031 2.3447 2.4954
```

To predict the quality of service score that would be given by a 55-year-old male patient on his fourth visit to Dr. A with a 30-minute appointment, we do the following calculations:

$$\text{score}^0 = 4.6863 + 0.01012 \cdot 55 + 0.167 - 0.0113 \cdot 30 - 0.04559 \cdot 4 = 4.88854.$$

In SAS:

```
data predict;
input id gender$ age doctor$ length visit;
cards;
136 M 55 A 30 4
;

data longform;
set longform predict;
run;

proc mixed covtest;
class gender doctor;
model score = gender age doctor length visit / outpm=outdata;
random intercept / subject=id type=un;
repeated / subject=id type=ar(1);
run;

proc print data=outdata (firstobs=106) noobs;
var Pred;
run;
```

**Pred**  
**4.88872**

In R:

```
print(predict(arint.fitted.model, data.frame(id=136, gender.rel="M", age=55,
doctor.rel="A", length=30, visit=4), level=0))
```

**4.88872**

**EXERCISE 8.8.** Use the data in Exercise 8.4 to do the following analysis:

(a) Output AIC, AICC, and BIC values for random slope and intercept (or random intercept-only) models with unstructured, Toeplitz, spatial power, autoregressive, compound symmetric, and independent covariance structures for the error terms.

In SAS:

```
data fitness;
input id gender$ age oxygen1 runtime1 pulse1 oxygen2 runtime2 pulse2
oxygen3 runtime3 pulse3;
cards;
1 F 39 37.4 11.4 151 36.6 17.8 158 36.1 15.4 152
2 M 42 60.1 11.5 121 59.0 9.6 131 58.2 9.0 143
3 F 34 44.6 9.6 138 39.8 9.3 148 38.8 9.1 144
4 M 36 51.9 10.5 125 53.4 9.8 135 50.4 9.6 163
5 F 45 40.8 13.1 142 39.5 12.4 151 38.5 12.7 133
6 M 37 45.4 10.3 133 40.6 11.9 145 40.2 11.2 141
7 F 49 45.3 13.1 135 40.6 12.1 148 39.7 11.5 157
8 F 47 44.8 12.1 135 40.1 12.3 148 39.0 11.9 151
9 M 50 48.7 12.6 131 42.3 11.0 143 44.3 10.5 150
10 M 34 45.8 10.8 132 40.9 11.8 144 41.1 11.1 160
11 M 35 50.4 9.6 129 45.9 10.4 137 44.8 10.4 138
12 M 48 50.5 12.9 125 48.6 10.3 135 49.0 9.8 132
```

```

13 F 50 44.8 14.0 135 40.3 13.1 148 39.5 12.6 163
14 F 53 39.4 12.7 145 39.3 14.1 154 37.0 12.8 148
15 M 44 46.1 11.0 132 40.9 11.3 144 41.1 10.8 148
16 F 32 39.2 9.1 146 38.7 9.7 158 36.7 10.2 170
17 M 39 54.3 9.4 123 55.1 9.7 132 57.4 9.4 162
18 F 33 39.4 11.6 144 39.4 12.7 154 37.4 12.7 155
19 M 33 47.9 10.1 132 42.2 11.2 143 42.6 10.6 140
20 M 46 49.2 11.2 130 43.9 10.8 141 44.7 10.5 142
;

/*creating longform dataset*/
data longform;
set fitness;
  array o[3] oxygen1-oxygen3;
  array r[3] runtime1-runtime3;
  array p[3] pulse1-pulse3;
  do condition=1 to 3;
    oxygen=o[condition];
    runtime=r[condition];
    pulse=p[condition];
    output;
  end;
keep id gender age oxygen runtime pulse condition;
run;

/*fitting random slope and intercept model with
unstructured covariance matrix of error terms*/
proc mixed covtest;
class gender;
  model pulse = gender age oxygen runtime condition /solution;
  random intercept condition / subject=id type=un;
  repeated / subject=id type=un;
run;

Fit Statistics
AIC (Smaller is Better) 338.0
AICC (Smaller is Better) 342.1
BIC (Smaller is Better) 346.9

/*fitting random slope and intercept model with
Toeplitz covariance matrix of error terms*/
proc mixed covtest;
class gender;
  model pulse = gender age oxygen runtime condition / solution;
  random intercept condition / subject=id type=un;
  repeated / subject=id type=toep;
run;

Fit Statistics
AIC (Smaller is Better) 375.7
AICC (Smaller is Better) 376.5
BIC (Smaller is Better) 379.6

/*fitting random slope and intercept model with
spatial power covariance matrix of error terms*/
proc mixed covtest;
class gender;
  model pulse = gender age oxygen runtime condition / solution;
  random intercept condition / subject=id type=un;

```

```

    repeated / subject=id type=sp(pow) (condition);
run;

```

**WARNING: Did not converge.**

```

/*fitting random intercept-only model with
spatial power covariance matrix of error terms*/
proc mixed covtest;
  class gender;
  model pulse = gender age oxygen runtime condition / solution;
  random intercept / subject=id type=un;
  repeated / subject=id type=sp(pow) (condition);
run;

```

#### Fit Statistics

**AIC (Smaller is Better) 399.3**

**AICC (Smaller is Better) 399.5**

**BIC (Smaller is Better) 401.3**

```

/*fitting random slope and intercept model with
autoregressive covariance matrix of error terms*/
proc mixed covtest;
  class gender;
  model pulse = gender age oxygen runtime condition / solution;
  random intercept condition / subject=id type=un;
  repeated / subject=id type=ar(1);
run;

```

**WARNING: Did not converge.**

```

/*fitting random intercept-only model with
autoregressive covariance matrix of error terms*/
proc mixed covtest;
  class gender;
  model pulse = gender age oxygen runtime condition / solution;
  random intercept / subject=id type=un;
  repeated / subject=id type=ar(1);
run;

```

#### Fit Statistics

**AIC (Smaller is Better) 399.3**

**AICC (Smaller is Better) 399.5**

**BIC (Smaller is Better) 401.3**

```

/*fitting random slope and intercept model with
compound symmetric covariance matrix of error terms*/
proc mixed covtest;
  class gender;
  model pulse = gender age oxygen runtime condition / solution;
  random intercept condition / subject=id type=un;
  repeated / subject=id type=cs;
run;

```

#### Fit Statistics

**AIC (Smaller is Better) 377.8**

**AICC (Smaller is Better) 379.1**

**BIC (Smaller is Better) 382.8**

```
/*fitting random slope and intercept model with
independent covariance matrix of error terms*/
proc mixed covtest;
  class gender;
  model pulse = gender age oxygen runtime condition / solution;
  random intercept condition / subject=id type=un;
run;
```

#### Fit Statistics

```
AIC (Smaller is Better) 375.8
AICC (Smaller is Better) 376.6
BIC (Smaller is Better) 379.8
```

In R:

```
fitness.data<- read.csv(file="C:/./Exercise8.4Data.csv", header=TRUE, sep=",")

#creating longform dataset
library(reshape2)
data1<- melt(fitness.data[,c("id","gender","age","oxygen1","oxygen2","oxygen3")],
id.vars=c("id","gender","age"),variable.name="oxygen.cond",value.name="oxygen")
data2<- melt(fitness.data[,c("runtime1","runtime2","runtime3")],variable.name=
"runtime.cond",value.name="runtime")
data3<- melt(fitness.data[,c("pulse1","pulse2","pulse3")],variable.name=
"pulse.cond",value.name="pulse")
longform.data<- cbind(data1,data2,data3)

#creating variable for condition
condition<- ifelse(longform.data$pulse.cond=="pulse1",1,
ifelse(longform.data$pulse.cond=="pulse2",2,3))

#specifying reference level
gender.rel<- relevel(longform.data$gender, ref="M")

#fitting random slope and intercept model with
#unstructured covariance matrix of error terms
library(nlme)
summary(un.fitted.model<- lme(pulse ~ gender.rel + age + oxygen + runtime +
condition, random = ~ 1 + condition | id, control= lmeControl(opt="optim"),
data=longform.data, correlation=corSymm(), weights=varIdent(form = ~ id |
condition)))

      AIC      BIC
353.7616 383.5964

#computing AICC
n<- 102
p<- 15
print(AICC<- -2*logLik(un.fitted.model)+2*p*n/(n-p-1))

359.343

#fitting random slope and intercept model with
#Toeplitz covariance matrix of error terms
summary(toep.fitted.model<- lme(pulse ~ gender.rel + age + oxygen + runtime
+ condition, random = ~ 1 + condition | id, control= lmeControl(opt="optim"),
data=longform.data, correlation = corARMA(form = ~ 1 | id, p=1, q=1)))

      AIC      BIC
391.9006 415.7684
```

```
#computing AICC
p<- 12
print(AICC<- -2*logLik(toep.fitted.model)+2*p*n/(n-p-1))
```

**395.4062**

```
#fitting random slope and intercept model with
#spatial power covariance matrix of error terms
summary(sppow.fitted.model<- lme(pulse ~ gender.rel + age + oxygen + runtime
+ condition, random = ~ 1 + condition | id, control= lmeControl(opt="optim"),
data=longform.data, correlation=corCAR1(form = ~ condition | id)))
```

	AIC	BIC
	393.2069	415.0857

```
#computing AICC
p<- 11
print(AICC<- -2*logLik(sppow.fitted.model)+2*p*n/(n-p-1))
```

**396.1402**

```
#fitting random slope and intercept model with
#autoregressive covariance matrix of error terms
summary(ar.fitted.model<- lme(pulse ~ gender.rel + age + oxygen + runtime
+ condition, random = ~ 1 + condition | id, control= lmeControl(opt="optim"),
data=longform.data, correlation=corAR1(form = ~ 1 | id)))
```

	AIC	BIC
	389.9551	411.8339

```
#computing AICC
p<- 11
print(AICC<- -2*logLik(ar.fitted.model)+2*p*n/(n-p-1))
```

**392.8884**

```
#fitting random slope and intercept model with
#compound symmetric covariance matrix of error terms
summary(cs.fitted.model<- lme(pulse ~ gender.rel + age + oxygen + runtime +
condition, random = ~ 1 + condition | id, control= lmeControl(opt="optim"),
data=longform.data,
correlation=corCompSymm(form = ~ 1 | id)))
```

	AIC	BIC
	389.8923	411.7711

```
#computing AICC
p<- 11
print(AICC<- -2*logLik(cs.fitted.model)+2*p*n/(n-p-1))
```

**392.8256**

```
#fitting random slope and intercept model with
#independent covariance matrix of error terms
summary(ind.fitted.model<- lme(pulse ~ gender.rel + age + oxygen + runtime +
condition, random = ~ 1 + condition | id, control= lmeControl(opt="optim"),
data=longform.data))
```

	AIC	BIC
	391.2189	411.1088

```
#computing AICC
p<- 10
print(AICC<- -2*logLik(ind.fitted.model)+2*p*n/(n-p-1))

393.6365
```

(b) Find the best-fitted model according to the AIC, AICC, and BIC criteria.

For the models fitted in SAS, the fitting criteria values are summarized in this table:

	UN	Toeplitz	Sp Power Intercept	AR Intercept	CS	Ind
<b>AIC</b>	338.0	375.7	399.3	399.3	377.8	375.8
<b>AICC</b>	342.1	376.5	399.5	399.5	379.1	376.6
<b>BIC</b>	346.9	379.6	401.3	401.3	382.8	379.8

For the models fitted in R, the values are as follows:

	UN	Toeplitz	Sp Power	AR	CS	Ind
<b>AIC</b>	353.8	391.9	393.2	390.0	389.9	391.2
<b>AICC</b>	359.3	395.4	396.1	392.9	392.8	393.6
<b>BIC</b>	383.6	415.8	415.1	411.8	411.8	411.1

The best-fitted model is the one with the unstructured covariance matrix of the error terms.

(c) Answer questions (c)-(e) in Exercise 8.4 for the model that fits the data the best.

In SAS:

```
/*fitting random slope and intercept model with
unstructured covariance matrix of error terms*/
proc mixed covtest;
  class gender;
  model pulse = gender age oxygen runtime condition / solution;
  random intercept condition / subject=id type=un;
  repeated / subject=id type=un r;
run;
```

Estimated R Matrix for Subject 1

Row	Col1	Col2	Col3
1	9.1604	8.4295	-11.3659
2	8.4295	8.3031	-4.5490
3	-11.3659	-4.5490	136.29

Covariance Parameter Estimates

Cov	Parm	Subject	Estimate	Standard Error	Z Value	Pr >  Z
UN(1,1)	id		1.4266	219.42	0.01	0.4974
UN(2,1)	id		-1.8318	103.12	-0.02	0.9858

Covariance Parameter Estimates					
Cov	Parm	Subject	Estimate	Standard Error	Z Value Pr >  Z
UN(2,2)	id		1.8059	49.2454	0.04 0.4854
UN(1,1)	id		9.1604	61.8377	0.15 0.4411
UN(2,1)	id		8.4295	12.7894	0.66 0.5098
UN(2,2)	id		8.3031	0	. .
UN(3,1)	id		-11.3659	43.9102	-0.26 0.7958
UN(3,2)	id		-4.5490	0	. .
UN(3,3)	id		136.29	0	. .

Solution for Fixed Effects					
Effect	gender	Estimate	Standard Error	DF t Value	Pr >  t
Intercept		158.98	5.8430	17	27.21 <.0001
gender	F	6.6365	1.4390	18	4.61 0.0002
gender	M	0	. .	. .	. .
age		-0.1260	0.08944	18	-1.41 0.1759
oxygen		-0.6393	0.09872	18	-6.48 <.0001
runtime		-0.2299	0.1656	18	-1.39 0.1822
condition		9.0253	0.4501	19	20.05 <.0001

The fitted model has the estimated mean response  $\hat{E}(\text{pulse}) = 158.98 + 6.6365 \cdot \text{female} - 0.126 \cdot \text{age} - 0.6393 \cdot \text{oxygen} - 0.2299 \cdot \text{runtime} + 9.0253 \cdot \text{condition}$ . The estimated parameters for the random-effect terms are  $\hat{\sigma}_{u_1}^2 = 1.4266$ ,  $\hat{\sigma}_{u_2}^2 = 1.8059$ ,  $\hat{\sigma}_{u_1 u_2} = -1.8318$ , and the estimated covariance matrix of the error terms is a block diagonal, with 20 blocks of the form

$\begin{pmatrix} 9.1604 & 8.4295 & -11.3659 \\ 8.4295 & 8.3031 & -4.5490 \\ -11.3659 & -4.5490 & 136.29 \end{pmatrix}$ . Gender, oxygen, and condition number are significant predictors of the pulse. For female runners, the estimated average pulse is 6.6365 points higher than that for male runners. As oxygen intake increases by one unit, the estimated average pulse decreases by 0.6393 points. As the condition number increases by one, the estimated average pulse increases by 9.0253 points.

In R:

```
#fitting random intercept-only model with
#autoregressive covariance matrix of error terms
summary(arint.fitted.model<- lme(score ~ gender.rel + age + doctor.rel + length
+ visit, random = ~ 1 | id, data=longform.data, correlation=corAR1(form = ~ 1 |
id)))
```

Random effects:

	StdDev	Corr
(Intercept)	3.1670814167	(Intr)
condition	0.4693058637	-1
Residual	0.0004034296	

Fixed effects:

	Value	Std.Error	DF	t-value	p-value
(Intercept)	160.73185	6.200000	37	25.924492	0.0000
gender.relF	7.72051	1.424442	17	5.420026	0.0000
age	-0.12855	0.089524	17	-1.435918	0.1692
oxygen	-0.71493	0.098789	37	-7.236946	0.0000
runtime	-0.02773	0.176549	37	-0.157056	0.8761
condition	8.71834	0.468003	37	18.628799	0.0000

```
getVarCov(arint.fitted.model, type="conditional")
```

Conditional variance covariance matrix

	1	2	3
1	1.6276e-07	5.2205e-03	3.2633e-04
2	5.2205e-03	1.6745e+02	1.0472e+01
3	3.2633e-04	1.0472e+01	2.5288e+00

To predict an average heart rate for a 36-year-old woman who is running on a treadmill, if her oxygen intake is 40.2 units, and her run time is 10.3 minutes per mile, we do the following calculations:

$$\text{pulse}^0 = 158.98 + 6.6365 - 0.126 \cdot 36 - 0.6393 \cdot 40.2 - 0.2299 \cdot 10.3 + 9.0253 = 142.047.$$

In SAS:

```
data predict;
input id gender$ age condition oxygen runtime;
cards;
21 F 36 1 40.2 10.3
;
run;
data longform;
set longform predict;
run;

proc mixed covtest;
  class gender;
  model pulse = gender age oxygen runtime condition / outpm=outdata;
  random intercept condition / subject=id type=un;
  repeated / subject=id type=un r;
run;

proc print data=outdata (firstobs=61) noobs;
  var Pred;
run;

      Pred
142.041
```

In R:

```
print(predict(un.fitted.model, data.frame(id=21, gender.rel="F", age=36,
oxygen=40.2, runtime=10.3, condition=1), level=0))

143.5173
```

**EXERCISE 8.9.** Take the data presented in Exercise 8.5.

(a) For BMI, fit the random slope and intercept regression models (or random intercept-only models) with unstructured, Toeplitz, spatial power, autoregressive, compound symmetric, and independent covariance structures for the error terms.

In SAS:

```

data weightloss;
input id group$ gender$ aexercise aBMI bexercise bBMI cexercise cBMI @@;
cards;
1  Int F 0 42.4 50 40.0 120 36.8 2  Int F 15 32.9 20 30.6 25 28.6
3  Int M 10 32.0 30 30.8 30 26.1 4  Int M 20 26.1 80 25.5 80 21.1
5  Int F 0 27.5 20 26.4 20 22.5 6  Int F 30 40.4 75 38.3 180 32.1
7  Int M 15 33.5 50 28.2 50 25.8 8  Int F 15 35.2 35 34.8 90 30.6
9  Int F 0 39.5 55 37.1 50 35.3 10 Int M 20 27.3 30 26.3 30 22.6
11 Int M 0 46.9 50 43.5 50 40.3 12 Int M 20 34.4 80 32.2 85 28.1
13 Int F 0 34.2 60 31.0 65 26.8 14 Int F 45 26.5 30 24.6 30 20.8
15 Int F 0 29.6 20 28.2 20 24.9 16 Int F 10 31.2 80 29.3 50 28.6
17 Cnt F 0 29.3 25 28.9 30 26.3 18 Cnt M 20 45.9 10 43.1 15 42.9
19 Cnt M 0 41.5 20 38.8 30 39.9 20 Cnt F 30 33.3 25 33.4 35 33.2
21 Cnt M 15 31.1 35 30.9 0 30.9 22 Cnt F 10 43.3 35 43.6 30 44.5
23 Cnt M 15 35.5 0 36.5 5 35.3 24 Cnt F 10 42.4 15 43.4 50 42.3
25 Cnt F 20 37.0 30 36.6 45 35.5 26 Cnt M 0 37.8 30 35.7 45 34.3
27 Cnt F 20 23.7 10 23.1 0 23.7 28 Cnt F 10 38.7 15 20.4 25 20.1
29 Cnt F 0 41.2 15 41.2 55 39.7 30 Cnt F 30 30.2 35 29.9 5 29.4
31 Cnt M 10 38.4 20 38.1 30 37.0 32 Cnt F 10 37.5 15 37.4 5 36.8
33 Cnt M 30 34.5 10 34.4 20 33.9 34 Cnt M 15 37.6 35 36.2 25 36.0
;

```

```

/*creating longform dataset*/
data longform;
set weightloss;
array m[3] (0 1 3);
array e[3] aexercise bexercise cexercise;
array b[3] aBMI bBMI cBMI;
do i=1 to 3;
month=m[i];
exercise=e[i];
BMI=b[i];
output;
end;
keep id group gender exercise BMI month;
run;

/*fitting random slope and intercept model with
unstructured covariance matrix of error terms*/
proc mixed covtest;
class group(ref="Cnt") gender(ref="F");
model BMI = group gender exercise month / solution;
random intercept month / subject=id type=un;
repeated / subject=id type=un;
run;

```

#### Fit Statistics

```

AIC (Smaller is Better) 522.6
AICC (Smaller is Better) 524.2
BIC (Smaller is Better) 534.8

```

```

/*fitting random slope and intercept model with
Toeplitz covariance matrix of error terms*/
proc mixed covtest;
class group(ref="Cnt") gender(ref="F");
model BMI = group gender exercise month / solution;
random intercept month / subject=id type=un;
repeated / subject=id type=toep;
run;

```

#### Fit Statistics

AIC (Smaller is Better) 538.7  
AICC (Smaller is Better) 539.4  
BIC (Smaller is Better) 546.4

```
/*fitting random slope and intercept model with  
spatial power covariance matrix of error terms*/  
proc mixed covtest;  
  class group(ref="Cnt") gender(ref="F");  
  model BMI = group gender exercise month / solution;  
    random intercept month / subject=id type=un;  
    repeated / subject=id type=sp(pow)(month);  
run;
```

#### Fit Statistics

AIC (Smaller is Better) 516.8  
AICC (Smaller is Better) 517.5  
BIC (Smaller is Better) 524.4

```
/*fitting random slope and intercept model with  
autoregressive matrix of error terms*/  
proc mixed covtest;  
  class group(ref="Cnt") gender(ref="F");  
  model BMI = group gender exercise month / solution;  
    random intercept month / subject=id type=un;  
    repeated / subject=id type=ar(1);  
run;
```

#### Fit Statistics

AIC (Smaller is Better) 536.7  
AICC (Smaller is Better) 537.2  
BIC (Smaller is Better) 542.9

```
/*fitting random slope and intercept model with  
compound symmetric covariance matrix of error terms*/  
proc mixed data=longform covtest;  
  class group(ref="Cnt") gender(ref="F");  
  model BMI = group gender exercise month / solution;  
    random intercept month / subject=id type=un;  
    repeated / subject=id type=cs;  
run;
```

#### Fit Statistics

AIC (Smaller is Better) 544.2  
AICC (Smaller is Better) 544.9  
BIC (Smaller is Better) 551.8

```
/*fitting random slope and intercept model with  
independent covariance matrix of error terms*/  
proc mixed data=longform covtest;  
  class group(ref="Cnt") gender(ref="F");  
  model BMI = group gender exercise month / solution;  
    random intercept month / subject=id type=un;  
run;
```

### Fit Statistics

AIC (Smaller is Better) 542.2

AICC (Smaller is Better) 542.6

BIC (Smaller is Better) 548.3

In R:

```
weightloss.data<- read.csv(file="C:/./Exercise8.5Data.csv", header=TRUE, sep=",")

#creating longform dataset
library(reshape2)
data1<- melt(weightloss.data[,c("id","group","gender","aexercise","bexercise",
"cexercise")], id.vars=c("id","group","gender"),variable.name="exercise.visit",
value.name="exercise")
data2<- melt(weightloss.data[,c("aBMI","bBMI","cBMI")],variable.name=
"BMI.visit",value.name="BMI")
longform.data<- cbind(data1,data2)

#creating variable for time
month<- ifelse(longform.data$BMI.visit=="aBMI",0,
ifelse(longform.data$BMI.visit=="bBMI",1,3))

#specifying reference levels
group.rel<- relevel(longform.data$group, ref="Int")
gender.rel<- relevel(longform.data$gender, ref="M")

#fitting random slope and intercept model with
#unstructured covariance matrix of error terms
library(nlme)
ctrl <- lmeControl(opt="optim")
summary(un.fitted.model<- lme(BMI ~ group.rel + gender.rel + exercise + month,
random = ~ 1 + month | id, control=ctrl, data=longform.data, correlation =
corSymm()), weights=varIdent(form=~id|month))

      AIC      BIC
534.5973 570.6433

#computing AICC
n<- 102
p<- 14
print(AICC<- -2*logLik(un.fitted.model)+2*p*n/(n-p-1))

539.4249

#fitting random slope and intercept model with
#Toeplitz covariance matrix of error terms
summary(toep.fitted.model<- lme(BMI ~ group.rel + gender.rel + exercise
+ month, random = ~ 1 + month | id, control=ctrl, data=longform.data,
correlation = corARMA(form=~1|id, p=1, q=1)))

      AIC      BIC
553.4601 581.7819

#computing AICC
p<- 11
print(AICC<- -2*logLik(toep.fitted.model)+2*p*n/(n-p-1))

556.3935
```

```
#fitting random slope and intercept model with
#spatial power covariance matrix of error terms
summary(sppow.fitted.model<- lme(BMI ~ group.rel + gender.rel + exercise
+ month, random = ~ 1 + month | id, control=ctrl, data=longform.data,
correlation=corCAR1(form=~month|id))
```

```
      AIC      BIC
554.2215 579.9686
```

```
#computing AICC
p<- 10
print(AICC<- -2*logLik(sppow.fitted.model)+2*p*n/(n-p-1))
```

```
556.6391
```

```
#fitting random slope and intercept model with
#autoregressive covariance matrix of error terms
summary(ar.fitted.model<- lme(BMI ~ group.rel + gender.rel + exercise
+ month, random = ~ 1 + month | id, control=ctrl, data=longform.data,
correlation=corAR1(form=~1|id))
```

```
      AIC      BIC
551.595 577.3421
```

```
#computing AICC
p<-10
print(AICC<- -2*logLik(ar.fitted.model)+2*p*n/(n-p-1))
```

```
554.0126
```

```
#fitting random slope and intercept model with
#compound symmetric covariance matrix of error terms
summary(cs.fitted.model<- lme(BMI ~ group.rel + gender.rel + exercise
+ month, random = ~ 1 + month | id, control=ctrl, data=longform.data,
correlation=corCompSymm(form=~1|id))
```

```
      AIC      BIC
554.2019 579.949
```

```
#computing AICC
p<-10
print(AICC<- -2*logLik(cs.fitted.model)+2*p*n/(n-p-1))
```

```
556.6194
```

```
#fitting random slope and intercept model with
#independent covariance matrix of error terms
summary(ind.fitted.model<- lme(BMI ~ group.rel + gender.rel + exercise
+ month, random = ~ 1 + month | id, control=ctrl, data=longform.data))
```

```
      AIC      BIC
552.2045 575.3769
```

```
#computing AICC
p<-9
print(AICC<- -2*logLik(ind.fitted.model)+2*p*n/(n-p-1))
```

```
554.161
```

(b) Which of the models has the best fit according to AIC, AICC, and BIC criteria?

According to SAS, the model with spatial power covariance structure has the best fit according to the AIC, AICC, and BIC criteria. The values are summed up in this table:

	UN	Toeplitz	Sp Power	AR	CS	Ind
AIC	522.6	538.7	516.8	536.7	544.2	542.2
AICC	524.2	539.4	517.5	537.2	544.9	542.6
BIC	534.8	546.4	524.4	542.9	551.8	548.3

In R, however, the best-fitted model is the one with the unstructured covariance matrix of the error terms.

	UN	Toeplitz	Sp Power	AR	CS	Ind
AIC	534.6	553.5	554.2	551.6	554.2	552.2
AICC	539.4	556.4	556.6	554.0	556.6	554.2
BIC	570.6	581.8	580.0	577.3	579.9	575.4

(c) For the best-fitted model, do the analysis for questions (c) through (e) in Exercise 8.5.

In SAS:

```
/*fitting random slope and intercept model with
spatial power covariance matrix of error terms*/
proc mixed covtest;
  class group(ref="Cnt") gender(ref="F");
  model BMI = group gender month exercise / solution;
  random intercept month / subject=id type=un;
  repeated / subject=id type=sp(pow)(month) r;
run;
```

Estimated R Matrix for Subject 1

Row	Col1	Col2	Col3
1	2.8277	-2.4537	-1.8476
2	-2.4537	2.8277	2.1292
3	-1.8476	2.1292	2.8277

Covariance Parameter Estimates

Cov	Parm	Subject	Estimate	Standard Error	Z value	Pr >  Z
UN(1,1)	id		41.9237	13.4965	3.11	0.0009
UN(2,1)	id		3.7666	1.2958	2.91	0.0037
UN(2,2)	id		0.3456	0.1817	1.90	0.0286
SP(POW)	id		-0.8677	0.05390	-16.10	<.0001
Residual			2.8277	0.6943	4.07	<.0001

Solution for Fixed Effects

Effect	group	gender	Estimate	Standard Error	DF	t value	Pr >  t
Intercept			32.5252	1.5195	31	21.40	<.0001

Solution for Fixed Effects						
Effect	group	gender	Estimate	Standard Error	DF	t Value Pr >  t
group	Int		4.6343	1.5858	33	2.92 0.0062
group	Cnt		0	.	.	.
gender		M	2.6890	1.6155	33	1.66 0.1055
gender		F	0	.	.	.
exercise			-0.04509	0.006957	33	-6.48 <.0001
month			-0.7853	0.1395	33	-5.63 <.0001

The fitted model has the estimated mean response  $\hat{E}(BMI) = 32.5252 + 4.6343 \cdot intervention + 2.689 \cdot male - 0.04509 \cdot length\ of\ exercise - 0.7853 \cdot month$ . The estimated parameters for the random-effect terms are  $\hat{\sigma}_{u_1}^2 = 41.9237$ ,  $\hat{\sigma}_{u_2}^2 = 0.3456$ ,  $\hat{\sigma}_{u_1 u_2} = 3.7666$ , and the estimated covariance matrix of the error terms is a block diagonal, with 34 blocks of the form

$\begin{pmatrix} 2.8277 & -2.4537 & -1.8476 \\ -2.4537 & 2.8277 & 2.1292 \\ -1.8476 & 2.1292 & 2.8277 \end{pmatrix}$ . Group, exercise, and month are significant predictors of BMI.

For the intervention group, the estimated average BMI is 4.6343 points above that in the control group. As the length of daily exercise increases by one minute, the estimated average BMI decreases by 0.04509 units. It is estimated that the average BMI decreases by 0.7853 units for every additional month in the study.

Predicted BMI at 3 months for an intervention group female participant who exercises for 1 hour every day is  $BMI^0 = 32.5252 + 4.6343 - 0.04509 \cdot 60 - 0.7853 \cdot 3 = 32.0982$ .

In SAS:

```
data predict;
input id group$ gender$ exercise month;
cards;
35 Int F 60 3
;

data longform;
set longform predict;
run;

proc mixed covtest;
  class group(ref="Cnt") gender(ref="F");
  model BMI = group gender exercise month / outpm=outdata;
  random intercept month / subject=id type=un;
  repeated / subject=id type=sp(pow)(month);
run;

proc print data=outdata (firstobs=103) noobs;
  var Pred;
run;

      Pred
32.0980
```

In R:

```
#fitting random slope and intercept model with
```

```
#unstructured covariance matrix of error terms
summary(un.fitted.model<- lme(BMI ~ group.rel + gender.rel + exercise
+ month, random = ~ 1 + month | id, control=ctrl, data=longform.data,
correlation = corSymm(), weights=varIdent(form = ~ id | month)))
```

Random effects:

	StdDev	Corr
(Intercept)	6.5901276	(Intr)
month	0.7815193	0.722
Residual	2.4238856	

Fixed effects:

	Value	Std.Error	DF	t-value	p-value
(Intercept)	32.25341	1.5309623	66	21.067411	0.0000
group.relInt	5.07648	1.5803999	31	3.212151	0.0031
gender.relM	2.81462	1.6128551	31	1.745119	0.0909
exercise	-0.04517	0.0068294	66	-6.613932	0.0000
month	-0.78606	0.1397604	66	-5.624341	0.0000

```
getVarCov(un.fitted.model, type="conditional")
```

Conditional variance covariance matrix

	1	2	3
1	5.8744000	-1.7322000	-1.0604e-03
2	-1.7322000	0.5107700	3.1279e-04
3	-0.0010604	0.00031279	2.0430e-07

The fitted model in R is  $\hat{E}(BMI) = 32.25341 + 5.07648 \cdot intervention + 2.81462 \cdot male - 0.04517 \cdot length\ of\ exercise - 0.78606 \cdot month$ . The estimates of the other model parameters are  $\hat{\sigma}_{u_1} = 6.5898132$ ,  $\hat{\sigma}_{u_2} = 0.7815473$ ,  $\hat{\rho}_{u_1u_2} = 0.722$ , and the covariance matrix is

$$\begin{pmatrix} 5.8744 & -1.7322 & -0.0010604 \\ -1.7322 & 0.51077 & 0.00031279 \\ -0.0010604 & 0.00031279 & 0.0000002 \end{pmatrix}$$
. Group, length of exercise, and month are significant

predictors of the response. For the intervention group, the estimated average BMI is 5.07648 points higher than for the control group. As the length of daily exercise increases by one minute, the estimated average BMI decreases by 0.04517 units. It is estimated that the average BMI decreases by 0.78605 units for every additional month in the study.

Predicted BMI at 3 months for an intervention group female participant who exercises for 1 hour every day is  $BMI^0 = 32.25341 + 5.07648 - 0.04517 \cdot 60 - 0.78605 \cdot 3 = 32.2615$ .

In R:

```
print(predict(un.fitted.model, data.frame(id=35, group.rel="Int", gender.rel="F",
exercise=60, month=3), level=0))
```

32.2611

**EXERCISE 8.10.** Returning to the data in Exercise 8.2, answer the following questions:

(a) Fit the GEE models with unstructured, Toeplitz (in SAS only), autoregressive, compound symmetric, and independent working correlation matrices.

### In SAS:

```
data deptstore;
input id totalyears status$ bonus18 bonus19 bonus20 @@;
cards;
1 16 full 1482 1508 1543 2 7 part 673 710 895
3 11 full 933 1351 1440 4 8 part 844 958 1196
5 6 part 564 790 815 6 5 full 601 708 780
7 6 part 775 822 902 8 17 full 1209 1297 1475
9 12 full 929 1008 1255 10 9 full 983 1013 1111
11 11 full 909 1004 1084 12 6 part 387 853 999
13 4 part 476 530 627 14 6 full 780 843 925
15 10 full 717 1200 1399
;

/*creating longform dataset*/
data longform;
set deptstore;
array y[3] (1.8 1.9 2.0);
array b[3] bonus18-bonus20;
do i=1 to 3;
year=y[i];
bonus=b[i];
output;
end;
keep id totalyears status year bonus;
run;

/*fitting GEE model with unstructured working correlation matrix*/
proc genmod;
class id status;
model bonus = totalyears status year;
repeated subject = id / type=un;
run;
```

**WARNING: Iteration limit exceeded.**

```
/*fitting GEE model with Toeplitz working correlation matrix*/
proc genmod;
class id status;
model bonus = totalyears status year;
repeated subject = id / type=mdep(2);
run;
```

**QIC 51.4667**

```
/*fitting GEE model with autoregressive working correlation matrix*/
proc genmod;
class id status;
model bonus = totalyears status year;
repeated subject = id / type=ar;
run;
```

**QIC 51.6428**

```
/*fitting GEE model with compound symmetric working correlation matrix*/
proc genmod;
class id status;
model bonus = totalyears status year;
repeated subject = id / type=cs;
run;
```

## QIC 52.0199

```
/*fitting GEE model with independent working correlation matrix*/
proc genmod;
  class id status;
  model bonus = totalyears status year;
  repeated subject = id / type=ind;
run;
```

## QIC 52.0199

In R:

```
deptstore.data<- read.csv(file="C:/./Exercise8.2Data.csv", header=TRUE, sep=",")

#creating longform dataset
library(reshape2)
longform.data<- melt(deptstore.data, id.vars=c("id","totalyears", "status"),
variable.name="bonus.year", value.name="bonus")

#sorting data by id
longform.data<- longform.data[order(longform.data$id),]

#creating variable for time
year<- ifelse(longform.data$bonus.year=="bonus18",1.8,
ifelse(longform.data$bonus.year=="bonus19",1.9,2.0))

#creating reference level
status.rel<- relevel(as.factor(longform.data$status), ref="part")

library(geepack)
library(MuMIn)

#fitting GEE model with unstructured working correlation matrix
summary(un.fitted.model<- geeglm(bonus ~ totalyears + status.rel + year,
data=longform.data, id=id, family=gaussian(link="identity"),
corstr="unstructured"))
QIC(un.fitted.model)
```

**QIC**  
**54.9**

```
#fitting GEE model with autoregressive working correlation matrix
summary(ar.fitted.model<- geeglm(bonus ~ totalyears + status.rel + year,
data=longform.data, id=id, family=gaussian(link="identity"), corstr="ar1"))
QIC(ar.fitted.model)
```

**QIC**  
**56.8**

```
#fitting GEE model with compound symmetric working correlation matrix
summary(cs.fitted.model<- geeglm(bonus ~ totalyears + status.rel + year,
data=longform.data, id=id, family=gaussian(link="identity"),
corstr="exchangeable"))
QIC(cs.fitted.model)
```

**QIC**  
**57.1**

```
#fitting GEE model with independent working correlation matrix
```

```
summary(ind.fitted.model<- geeglm(bonus ~ totalyears + status.rel + year,
data=longform.data, id=id, family=gaussian(link="identity"),
corstr="independence"))
QIC(ind.fitted.model)
```

QIC  
57.1

(c) Find the best model using the QIC criterion.

In SAS, the best-fitted model is the one with the Toeplitz working correlation matrix.

	Toep	AR	CS	Ind
QIC	51.4667	51.6428	52.0199	52.0199

In R, the model with an unstructured working correlation matrix has the smallest QIC and thus has the best fit.

(c) For the model that fits the data the best, answer questions (c)-(e) in Exercise 8.2.

In SAS:

```
/*fitting GEE model with Toeplitz working correlation matrix*/
proc genmod;
  class id status;
  model bonus = totalyears status year;
  repeated subject = id / type=mdep(2) corrw;
run;
```

Working Correlation Matrix

```
      col1  col2  col3
Row1  1.0000 0.5021 -0.1012
Row2  0.5021 1.0000  0.5021
Row3 -0.1012 0.5021  1.0000
```

Analysis of GEE Parameter Estimates						
Empirical Standard Error Estimates						
Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr >  Z
Intercept	-2266.63	444.1249	-3137.10	-1396.16	-5.10	<.0001
totalyears	61.2667	7.8293	45.9214	76.6119	7.83	<.0001
status full	40.8062	52.8054	-62.6905	144.3030	0.77	0.4397
status part	0.0000	0.0000	0.0000	0.0000	.	.
year	1394.667	232.2850	939.3964	1849.937	6.00	<.0001

The fitted model is  $\hat{E}(\text{bonus}) = -2266.63 + 61.2667 \cdot \text{total years} + 40.8062 \cdot \text{full} - \text{time employee} + 1394.667 \cdot \text{year}/10$ . The working correlation matrix is

$\begin{pmatrix} 1 & 0.5021 & -0.1012 \\ 0.5021 & 1 & 0.5021 \\ -0.1012 & 0.5021 & 1 \end{pmatrix}$ . The total number of years with the company and the year are

significant predictors of the bonus. As the total number of years increases by one, the estimated average bonus increases by \$61.2667. It is estimated that, on average, the bonus amount increases by \$139.4667 every year.

The predicted bonus in 2021 for a full-time employee who has been with the company for 7 years is computed as follows:  $\text{bonus}^0 = -2266.63 + 61.2667 \cdot 7 + 40.8062 + 1394.667 \cdot 2.1 = \$1,131.844$ .

In SAS:

```
data predict;
input id totalyears status$ year;
cards;
21 7 full 2.1
;

data longform;
set longform predict;
run;

proc genmod;
class id status;
model bonus = totalyears status year;
repeated subject = id / type=mdep(2);
output out=outdata p=pbonus;
run;

proc print data=outdata (firstobs=46) noobs;
var pbonus;
run;


pbonus  
1131.84


```

In R:

```
summary(un.fitted.model<- geeglm(bonus ~ totalyears+status.rel + year,
data=longform.data, id=id, family=gaussian(link="identity"),
corstr="unstructured"))
```

Coefficients:

	Estimate	Std.err	Wald	Pr(> W )
(Intercept)	-2063.30	309.56	44.43	2.6e-11
totalyears	61.04	7.81	61.15	5.3e-15
status.relfull	41.80	51.60	0.66	0.42
year	1289.10	169.76	57.66	3.1e-14

Estimated Correlation Parameters:

	Estimate	Std.err
alpha.1:2	0.1395	0.269
alpha.1:3	-0.0891	0.269
alpha.2:3	0.8076	0.151

The fitted model is  $\hat{E}(\text{bonus}) = -2063.30 + 61.04 \cdot \text{total years} + 41.80 \cdot \text{full} - \text{time employee} + 1289.10 \cdot \text{year}/10$ . The working correlation matrix is

$\begin{pmatrix} 1 & 0.1395 & -0.0891 \\ 0.1395 & 1 & 0.8076 \\ -0.0891 & 0.8076 & 1 \end{pmatrix}$ . The total number of years with the company and the year are significant predictors of the bonus. As the total number of years increases by one, the estimated average bonus increases by \$61.04. It is estimated that, on average, the bonus amount increases by \$128.91 every year.

The predicted bonus in 2021 for a full-time employee who has been with the company for 7 years is computed as follows:  $\text{bonus}^0 = -2063.30 + 61.04 \cdot 7 + 41.80 + 1289.10 \cdot 2.1 = \$1,112.89$ .

In R:

```
print(predict(un.fitted.model, data.frame(totalyears=7, status.rel="full",
year=2.1)))
```

1113

**EXERCISE 8.11.** For the data in Exercise 8.3,

(a) Fit the generalized estimating equations models with unstructured, Toeplitz (if using SAS), autoregressive, compound symmetric, and independent working correlation matrices.

In SAS:

```
data orthoclinic;
input id gender$ age doctor$ length1 length2 length3 score1 score2 score3 @@;
cards;
101 F 78 A 25 20 25 7.1 7.5 7.6 102 F 63 A 30 30 40 5.5 5.8 6.1
103 F 62 A 10 15 10 10.0 10.0 9.8 104 F 71 B 15 15 40 7.8 7.3 7.5
105 M 68 A 40 60 40 3.5 3.5 3.0 106 F 63 A 25 15 20 8.5 8.7 8.8
107 F 60 B 25 35 25 6.7 5.7 6.5 108 F 70 A 20 20 20 9.0 8.3 8.2
109 F 57 A 30 20 15 8.4 7.8 8.1 110 F 59 B 25 30 15 7.1 7.4 7.9
111 M 62 A 50 30 70 3.0 3.2 2.6 112 M 58 A 20 15 45 6.1 6.8 6.9
113 M 75 A 25 35 30 5.7 5.6 4.7 114 M 76 B 35 50 25 4.9 5.4 5.2
115 F 75 A 15 20 25 8.2 8.9 8.2 116 M 57 A 45 30 40 4.6 3.9 3.2
117 F 68 A 35 25 40 3.8 4.8 5.3 118 M 65 B 40 40 25 3.9 3.9 4.7
119 F 67 B 20 15 30 6.5 7.2 6.6 120 F 60 B 25 15 15 7.3 7.1 7.8
121 F 67 A 15 20 15 7.7 8.0 8.3 122 F 57 B 10 15 15 9.8 9.2 8.6
123 M 62 B 55 60 75 3.4 2.7 2.3 124 M 71 A 20 30 25 7.1 6.6 7.4
125 M 71 B 15 15 20 8.8 9.1 9.3 126 M 64 A 25 30 30 5.6 6.3 6.3
127 M 51 A 35 40 30 5.1 4.6 3.9 128 F 70 B 35 25 15 6.8 7.1 7.6
129 M 61 A 35 40 50 5.5 5.2 4.8 130 M 62 B 60 40 65 3.7 3.4 2.4
131 F 68 A 20 35 35 5.3 5.6 4.9 132 F 68 B 35 30 15 7.2 6.2 5.6
133 M 64 B 40 20 30 5.4 4.9 4.5 134 F 76 B 30 45 25 5.5 4.7 4.6
135 F 78 B 25 20 15 7.6 8.3 9.2
;

/*creating longform dataset*/
data longform;
set orthoclinic;
array l[3] length1-length3;
array s[3] score1-score3;
do visit=1 to 3;
length=l[visit];
score=s[visit];
output;
end;
keep id gender age doctor visit length score;
run;

/*fitting GEE model with unstructured working correlation matrix*/
proc genmod;
class id gender doctor;
model score = gender age doctor length visit;
repeated subject = id / type=un;
```

```
run;
```

**WARNING: Iteration limit exceeded.**

```
/*fitting GEE model with Toeplitz working correlation matrix*/
proc genmod;
  class id gender doctor;
  model score = gender age doctor length visit;
  repeated subject = id / type=mdep(2);
run;
```

**QIC 121.5327**

```
/*fitting GEE model with autoregressive working correlation matrix*/
proc genmod;
  class id gender doctor;
  model score = gender age doctor length visit;
  repeated subject = id / type=ar;
run;
```

**QIC 120.1484**

```
/*fitting GEE model with compound symmetric working correlation matrix*/
proc genmod;
  class id gender doctor;
  model score = gender age doctor length visit;
  repeated subject = id / type=cs;
run;
```

**QIC 120.1274**

```
/*fitting GEE model with independent working correlation matrix*/
proc genmod;
  class id gender doctor;
  model score = gender age doctor length visit;
  repeated subject = id / type=ind;
run;
```

**QIC 117.0717**

**In R:**

```
orthoclinic.data<- read.csv(file="C:/./Exercise8.3Data.csv", header=TRUE,
sep=", ")
```

```
#creating longform dataset
library(reshape2)
data1<- melt(orthoclinic.data[,c("id","gender","age","doctor","length1",
"length2","length3")], id.vars=c("id","gender","age","doctor"),
variable.name="length.visit",value.name="length")
data2<- melt(orthoclinic.data[,c("score1","score2","score3")],
variable.name="score.visit", value.name="score")
longform.data<- cbind(data1,data2)
```

```
#sorting data by id
longform.data<- longform.data[order(longform.data$id),]
```

```
#creating variable for time
visit<- ifelse(longform.data$score.visit=="score1",1,
```

```

ifelse(longform.data$score.visit=="score2",2,3))

#specifying reference levels
gender.rel<- relevel(as.factor(longform.data$gender), ref="M")
doctor.rel<- relevel(as.factor(longform.data$doctor), ref="B")

library(geepack)
library(MuMIn)
#fitting GEE model with unstructured working correlation matrix
summary(un.fitted.model<- geeglm(score ~ gender.rel + age + doctor.rel
+ length + visit, data=longform.data, id=id, family=gaussian(link="identity"),
corstr="unstructured"))

QIC(un.fitted.model)

QIC
98.52

#fitting GEE model with autoregressive working correlation matrix
summary(ar.fitted.model<- geeglm(score ~ gender.rel + age + doctor.rel
+ length + visit, data=longform.data, id=id, family=gaussian(link="identity"),
corstr="ar1"))
QIC(ar.fitted.model)

QIC
100

#fitting GEE model with compound symmetric working correlation matrix
summary(cs.fitted.model<- geeglm(score ~ gender.rel + age + doctor.rel
+ length + visit, data=longform.data, id=id, family=gaussian(link="identity"),
corstr="exchangeable"))
QIC(cs.fitted.model)

QIC
101

#fitting GEE model with independent working correlation matrix
summary(ind.fitted.model<- geeglm(score ~ gender.rel + age + doctor.rel
+ length + visit, data=longform.data, id=id, family=gaussian(link="identity"),
corstr="independence"))
QIC(ind.fitted.model)

QIC
124

```

(b) Which of the fitted models has the best fit according to the QIC criterion?

The QIC values outputted by SAS are summarized here:

	<b>Toep</b>	<b>AR</b>	<b>CS</b>	<b>Ind</b>
<b>QIC</b>	121.533	120.148	120.127	117.072

The optimal model is the one with the independent working correlation matrix. The best model in R is the one with an unstructured working correlation matrix.

(c) Answer parts (e)-(g) in Exercise 8.3, using the best-fitted model.

In SAS:

```

/*fitting GEE model with independent working correlation matrix*/
proc genmod;
  class id gender doctor;
  model score = gender age doctor length visit;
  repeated subject = id / type=ind corrw;
run;

```

#### Working Correlation Matrix

	Col1	Col2	Col3
Row1	1.0000	0.0000	0.0000
Row2	0.0000	1.0000	0.0000
Row3	0.0000	0.0000	1.0000

#### Analysis Of GEE Parameter Estimates

##### Empirical Standard Error Estimates

Parameter		Estimate	Standard Error	95% Confidence Limits	Z	Pr >  Z
Intercept		7.9262	1.3954	5.1913 10.6612	5.68	<.0001
gender	F	1.0438	0.4448	0.1720 1.9156	2.35	0.0189
gender	M	0.0000	0.0000	0.0000 0.0000	.	.
age		0.0067	0.0198	-0.0322 0.0456	0.34	0.7358
doctor	A	0.0797	0.3108	-0.5295 0.6889	0.26	0.7977
doctor	B	0.0000	0.0000	0.0000 0.0000	.	.
length		-0.0933	0.0140	-0.1207 -0.0659	-6.68	<.0001
visit		0.0071	0.1139	-0.2160 0.2303	0.06	0.9501

The fitted model is  $\hat{E}(\text{score}) = 7.9262 + 1.0438 \cdot \text{female} + 0.0067 \cdot \text{age} + 0.0797 \cdot \text{doctor A} - 0.0933 \cdot \text{visit length} + 0.0071 \cdot \text{visit number}$ . The working correlation matrix is  $\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ .

Gender and visit length are significant predictors of the score. It is estimated that scores given by female patients are, on average, 1.0438 points larger than those given by male patients. As the length of a visit increases by one minute, the estimated average score decreases by 0.0933 points.

The predicted score that would be given by a 55-year-old male patient on his fourth visit to Dr. A with a 30-minute appointment is computed as follows:  $\text{score}^0 = 7.9262 + 0.0067 \cdot 55 + 0.0797 - 0.0933 \cdot 30 + 0.0071 \cdot 4 = 5.6038$ .

In SAS:

```

data predict;
input id gender$ age doctor$ length visit;
cards;
136 M 55 A 30 4
;

data longform;
set longform predict;
run;

proc genmod;
  class id gender doctor;
  model score = gender age doctor length visit;
  repeated subject = id / type=ind;

```

```

output out=outdata p=pscore;

run;

proc print data=outdata (firstobs=106) noobs;
  var pscore;
run;

pscore
5.60333

```

In R:

```

#fitting GEE model with independent working correlation matrix
summary(un.fitted.model<- geeglm(score ~ gender.rel + age + doctor.rel
+ length + visit, data=longform.data, id=id, family=gaussian(link="identity"),
corstr="unstructured"))

```

Coefficients:

	Estimate	Std.err	Wald	Pr(> W )
(Intercept)	4.68494	1.88506	6.18	0.013
gender.relF	2.12954	0.52062	16.73	4.3e-05
age	0.01032	0.03105	0.11	0.740
doctor.relA	0.21578	0.48814	0.20	0.658
length	-0.01046	0.00558	3.52	0.061
visit	-0.04797	0.07809	0.38	0.539

Estimated Correlation Parameters:

	Estimate	Std.err
alpha.1:2	0.858	0.0522
alpha.1:3	0.874	0.0429
alpha.2:3	0.977	0.0476

The fitted model is  $\hat{E}(\text{score}) = 4.68494 + 2.12954 \cdot \text{female} + 0.01032 \cdot \text{age} + 0.21578 \cdot \text{doctor A} - 0.01046 \cdot \text{visit length} - 0.04797 \cdot \text{visit number}$ . The working correlation matrix is  $\begin{pmatrix} 1 & 0.858 & 0.874 \\ 0.858 & 1 & 0.977 \\ 0.874 & 0.977 & 1 \end{pmatrix}$ . Gender is the only significant predictor of the score at the 5% level. It is estimated that scores given by female patients are, on average, 2.12954 points larger than that given by male patients.

The predicted score that would be given by a 55-year-old male patient on his fourth visit to Dr. A with a 30-minute appointment is computed as follows:  $\text{score}^0 = 4.68494 + 0.01032 \cdot 55 + 0.21578 - 0.01046 \cdot 30 - 0.04797 \cdot 4 = 4.96264$ .

```

print(predict(un.fitted.model, data.frame(id=136, gender.rel="M", age=55,
doctor.rel="A",length=30, visit=4)))

```

4.96

**EXERCISE 8.12.** Consider the data in Exercise 8.4.

(a) Run the generalized estimating equations models with unstructured, Toeplitz (only in SAS), autoregressive, compound symmetric, and independent working correlation matrices for the pulse.

In SAS:

```

data fitness;
input id gender$ age oxygen1 runtime1 pulse1 oxygen2 runtime2 pulse2
oxygen3 runtime3 pulse3;
cards;
1 F 39 37.4 11.4 151 36.6 17.8 158 36.1 15.4 152
2 M 42 60.1 11.5 121 59.0 9.6 131 58.2 9.0 143
3 F 34 44.6 9.6 138 39.8 9.3 148 38.8 9.1 144
4 M 36 51.9 10.5 125 53.4 9.8 135 50.4 9.6 163
5 F 45 40.8 13.1 142 39.5 12.4 151 38.5 12.7 133
6 M 37 45.4 10.3 133 40.6 11.9 145 40.2 11.2 141
7 F 49 45.3 13.1 135 40.6 12.1 148 39.7 11.5 157
8 F 47 44.8 12.1 135 40.1 12.3 148 39.0 11.9 151
9 M 50 48.7 12.6 131 42.3 11.0 143 44.3 10.5 150
10 M 34 45.8 10.8 132 40.9 11.8 144 41.1 11.1 160
11 M 35 50.4 9.6 129 45.9 10.4 137 44.8 10.4 138
12 M 48 50.5 12.9 125 48.6 10.3 135 49.0 9.8 132
13 F 50 44.8 14.0 135 40.3 13.1 148 39.5 12.6 163
14 F 53 39.4 12.7 145 39.3 14.1 154 37.0 12.8 148
15 M 44 46.1 11.0 132 40.9 11.3 144 41.1 10.8 148
16 F 32 39.2 9.1 146 38.7 9.7 158 36.7 10.2 170
17 M 39 54.3 9.4 123 55.1 9.7 132 57.4 9.4 162
18 F 33 39.4 11.6 144 39.4 12.7 154 37.4 12.7 155
19 M 33 47.9 10.1 132 42.2 11.2 143 42.6 10.6 140
20 M 46 49.2 11.2 130 43.9 10.8 141 44.7 10.5 142
;

/*creating longform dataset*/
data longform;
set fitness;
array o[3] oxygen1-oxygen3;
array r[3] runtime1-runtime3;
array p[3] pulse1-pulse3;
do condition=1 to 3;
oxygen=o[condition];
runtime=r[condition];
pulse=p[condition];
output;
end;
keep id gender age oxygen runtime pulse condition;
run;

/*fitting GEE model with unstructured working correlation matrix*/
proc genmod;
class id gender;
model pulse = gender age oxygen runtime condition;
repeated subject = id / type=un;
run;

```

#### QIC 66.0288

```

/*fitting GEE model with Toeplitz working correlation matrix*/
proc genmod;
class id gender;
model pulse = gender age oxygen runtime condition;
repeated subject = id / type=mdep(2);
run;

```

#### QIC 66.3730

```

/*fitting GEE model with autoregressive working correlation matrix*/

```

```
proc genmod;
  class id gender;
  model pulse = gender age oxygen runtime condition;
  repeated subject = id / type=ar;
run;
```

#### QIC 65.0370

```
/*fitting GEE model with compound symmetric working correlation matrix*/
proc genmod;
  class id gender;
  model pulse = gender age oxygen runtime condition;
  repeated subject = id / type=cs;
run;
```

#### QIC 64.9567

```
/*fitting GEE model with independent working correlation matrix*/
proc genmod data=longform;
  class id gender;
  model pulse = gender age oxygen runtime condition;
  repeated subject = id /type=ind;
run;
```

#### QIC 64.9492

#### In R:

```
fitness.data<- read.csv(file="C:/./Exercise8.4Data.csv", header=TRUE, sep=",")

#creating longform dataset
library(reshape2)
data1<- melt(fitness.data[,c("id","gender","age","oxygen1","oxygen2","oxygen3")],
id.vars=c("id","gender","age"),variable.name="oxygen.cond",value.name="oxygen")
data2<- melt(fitness.data[,c("runtime1","runtime2","runtime3")],variable.name=
"runtime.cond",value.name="runtime")
data3<- melt(fitness.data[,c("pulse1","pulse2","pulse3")],variable.name=
"pulse.cond",value.name="pulse")
longform.data<- cbind(data1,data2,data3)

#sorting data by id
longform.data<- longform.data[order(longform.data$id),]

#creating variable for condition
condition<- ifelse(longform.data$pulse.cond=="pulse1",1,
ifelse(longform.data$pulse.cond=="pulse2",2,3))

#specifying reference level
gender.rel<- relevel(as.factor(longform.data$gender), ref="M")

library(geepack)
library(MuMIn)
#fitting GEE model with unstructured working correlation matrix
summary(un.fitted.model<- geeglm(pulse ~ gender.rel + age + oxygen + runtime
+ condition,data=longform.data,id=id, family = gaussian(link="identity"),
corstr="unstructured"))
QIC(un.fitted.model)
```

**QIC**  
**72.9**

```
#fitting GEE model with autoregressive working correlation matrix
summary(ar.fitted.model<- geeglm(pulse ~ gender.rel + age + oxygen + runtime
+ condition,data=longform.data,id=id, family = gaussian(link="identity"),
corstr="ar1"))
QIC(ar.fitted.model)
```

**QIC**  
**72.2**

```
#fitting GEE model with compound symmetric working correlation matrix
summary(cs.fitted.model<- geeglm(pulse ~ gender.rel + age + oxygen + runtime
+ condition,data=longform.data,id=id, family = gaussian(link="identity"),
corstr="exchangeable"))
QIC(cs.fitted.model)
```

**QIC**  
**72.2**

```
#fitting GEE model with independent working correlation matrix
summary(ind.fitted.model<- geeglm(pulse ~ gender.rel + age + oxygen + runtime
+ condition,data=longform.data,id=id, family = gaussian(link="identity"),
corstr="independence"))
QIC(ind.fitted.model)
```

**QIC**  
**72.2**

(b) Compare the QIC values for the fitted models and choose the optimal one.

In SAS, the model with the independent working correlation matrix is optimal. The QIC values are summarized below:

	UN	Toep	AR	CS	Ind
QIC	66.0288	66.373	65.037	64.9567	64.9492

In R, the three models (autoregressive, compound symmetric, and independent) have the same parameter estimates. We choose the model with the independent structure as the simplest model.

(c) For the optimal model, do questions (c)-(e) in Exercise 8.4.

In SAS:

```
/*fitting GEE model with independent working correlation matrix*/
proc genmod;
  class id gender;
  model pulse = gender age oxygen runtime condition;
  repeated subject=id / type=ind corrw;
run;
```

Working Correlation Matrix

```
      col1  col2  col3
Row1 1.0000 0.0000 0.0000
Row2 0.0000 1.0000 0.0000
```

Working Correlation Matrix  
 Col1 Col2 Col3  
 Row3 0.0000 0.0000 1.0000

Analysis Of GEE Parameter Estimates							
Empirical Standard Error Estimates							
Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr >  Z	
Intercept	152.8172	15.2280	122.9708	182.6636	10.04	<.0001	
gender F	7.0418	2.4289	2.2812	11.8023	2.90	0.0037	
gender M	0.0000	0.0000	0.0000	0.0000	.	.	
age	-0.2128	0.1217	-0.4513	0.0257	-1.75	0.0803	
oxygen	-0.4401	0.2151	-0.8617	-0.0185	-2.05	0.0407	
runtime	0.1080	0.5751	-1.0192	1.2352	0.19	0.8510	
condition	6.9483	1.5668	3.8774	10.0192	4.43	<.0001	

The fitted model is  $\hat{E}(\text{pulse}) = 152.8172 + 7.0418 \cdot \text{female} - 0.2128 \cdot \text{age} - 0.4401 \cdot \text{oxygen} + 0.1080 \cdot \text{runtime} + 6.9483 \cdot \text{condition}$ . The working correlation matrix is  $\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ .

Gender, oxygen intake, and condition are significant predictors of the pulse. It is estimated that the pulse for female runners is, on average, 7.0418 points larger than that for male runners. As oxygen intake increases by one unit, the estimated mean pulse decreases by 0.4401 units. As the condition number increases by one, the estimated mean pulse increases by 9.9483.

Next, we predict an average heart rate for a 36-year-old woman who is running on a treadmill, if her oxygen intake is 40.2 units, and her run time is 10.3 minutes per mile. We compute  $\text{pulse}^0 = 152.8172 + 7.0418 - 0.2128 \cdot 36 - 0.4401 \cdot 40.2 + 0.1080 \cdot 10.3 + 6.9483 = 142.567$ .

In SAS:

```
data predict;
input id gender$ age oxygen runtime condition;
cards;
21 F 36 40.2 10.3 1
;

data longform;
set longform predict;
run;

proc genmod;
class id gender;
model pulse = gender age oxygen runtime condition;
repeated subject = id / type=ind;
output out=outdata p=ppulse;
run;

proc print data=outdata (firstobs=61) noobs;
var ppulse;
run;

ppulse
142.568
```

In R:

```
#fitting GEE model with independent working correlation matrix
summary(ind.fitted.model<- geeglm(pulse ~ gender.rel + age + oxygen + runtime
+ condition,data=longform.data,id=id,family = gaussian(link="identity"),
corstr="independence"))
```

Coefficients:

	Estimate	Std.err	wald	Pr(> w )
(Intercept)	152.817	15.895	92.43	< 2e-16
gender.relF	7.042	2.258	9.73	0.0018
age	-0.213	0.129	2.70	0.1001
oxygen	-0.440	0.266	2.74	0.0982
runtime	0.108	0.663	0.03	0.8706
condition	6.948	1.508	21.23	4.1e-06

```
print(predict(ind.fitted.model, data.frame(gender.rel="F", age=36, oxygen=40.2,
runtime=10.3, condition=1)))
```

143

**EXERCISE 8.13.** For the data given in Exercise 8.5, do the following questions:

(a) Fit the GEE models with unstructured, Toeplitz, autoregressive, compound symmetric, and independent working correlation matrices of the response variable BMI.

In SAS:

```
data weightloss;
input id group$ gender$ aexercise aBMI bexercise bBMI cexercise cBMI @@;
cards;
1 Int F 0 42.4 50 40.0 120 36.8 2 Int F 15 32.9 20 30.6 25 28.6
3 Int M 10 32.0 30 30.8 30 26.1 4 Int M 20 26.1 80 25.5 80 21.1
5 Int F 0 27.5 20 26.4 20 22.5 6 Int F 30 40.4 75 38.3 180 32.1
7 Int M 15 33.5 50 28.2 50 25.8 8 Int F 15 35.2 35 34.8 90 30.6
9 Int F 0 39.5 55 37.1 50 35.3 10 Int M 20 27.3 30 26.3 30 22.6
11 Int M 0 46.9 50 43.5 50 40.3 12 Int M 20 34.4 80 32.2 85 28.1
13 Int F 0 34.2 60 31.0 65 26.8 14 Int F 45 26.5 30 24.6 30 20.8
15 Int F 0 29.6 20 28.2 20 24.9 16 Int F 10 31.2 80 29.3 50 28.6
17 Cnt F 0 29.3 25 28.9 30 26.3 18 Cnt M 20 45.9 10 43.1 15 42.9
19 Cnt M 0 41.5 20 38.8 30 39.9 20 Cnt F 30 33.3 25 33.4 35 33.2
21 Cnt M 15 31.1 35 30.9 0 30.9 22 Cnt F 10 43.3 35 43.6 30 44.5
23 Cnt M 15 35.5 0 36.5 5 35.3 24 Cnt F 10 42.4 15 43.4 50 42.3
25 Cnt F 20 37.0 30 36.6 45 35.5 26 Cnt M 0 37.8 30 35.7 45 34.3
27 Cnt F 20 23.7 10 23.1 0 23.7 28 Cnt F 10 38.7 15 20.4 25 20.1
29 Cnt F 0 41.2 15 41.2 55 39.7 30 Cnt F 30 30.2 35 29.9 5 29.4
31 Cnt M 10 38.4 20 38.1 30 37.0 32 Cnt F 10 37.5 15 37.4 5 36.8
33 Cnt M 30 34.5 10 34.4 20 33.9 34 Cnt M 15 37.6 35 36.2 25 36.0
;

/*creating longform dataset*/
data longform;
set weightloss;
array m[3] (0 1 3);
array e[3] aexercise bexercise cexercise;
array b[3] aBMI bBMI cBMI;
do i=1 to 3;
```

```

    month=m[i];
    exercise=e[i];
    BMI=b[i];
    output;
    end;
keep id group gender exercise BMI month;
run;

/*fitting GEE model with unstructured working correlation matrix*/
proc genmod;
  class id group gender(ref="F");
  model BMI = group gender exercise month;
  repeated subject = id / type=un;
run;

```

**WARNING: Iteration limit exceeded.**

```

/*fitting GEE model with Toeplitz working correlation matrix*/
proc genmod;
  class id group gender(ref="F");
  model BMI = group gender exercise month;
  repeated subject = id / type=mdep(2);
run;

```

**QIC 113.0602**

```

/*fitting GEE model with autoregressive working correlation matrix*/
proc genmod;
  class id group gender(ref="F");
  model BMI = group gender exercise month;
  repeated subject = id / type=ar;
run;

```

**QIC 112.7227**

```

/*fitting GEE model with compound symmetric working correlation matrix*/
proc genmod;
  class id group gender(ref="F");
  model BMI = group gender exercise month;
  repeated subject = id / type=cs;
run;

```

**QIC 113.0680**

```

/*fitting GEE model with independent working correlation matrix*/
proc genmod;
  class id group gender(ref="F");
  model BMI = group gender exercise month;
  repeated subject = id / type=ind;
run;

```

**QIC 113.7968**

**In R:**

```

weightloss.data<- read.csv(file="C:/./Exercise8.5Data.csv", header=TRUE, sep=",")

#creating longform dataset
library(reshape2)

```

```

data1<- melt(weightloss.data[,c("id","group","gender","aexercise","bexercise",
"cexercise")], id.vars=c("id","group","gender"),variable.name="exercise.visit",
value.name="exercise")
data2<- melt(weightloss.data[,c("aBMI","bBMI","cBMI")],variable.name="BMI.visit",
value.name="BMI")
longform.data<- cbind(data1,data2)

#sorting data by id
longform.data<- longform.data[order(longform.data$id),]

#creating variable for time
month<- ifelse(longform.data$BMI.visit=="aBMI",0,
ifelse(longform.data$BMI.visit=="bBMI",1,3))

#specifying reference levels
group.rel<- relevel(as.factor(longform.data$group), ref="Int")
gender.rel<- relevel(as.factor(longform.data$gender), ref="F")

library(geepack)
library(MuMIn)
#fitting GEE model with unstructured working correlation matrix
summary(un.fitted.model<- geeglm(BMI ~ group.rel + gender.rel + exercise
+ month, data=longform.data, id=id, family=gaussian(link="identity"),
corstr="unstructured"))

```

#### Estimated Correlation Parameters:

	Estimate	Std.err
alpha.1:2	0.760	0.1209
alpha.1:3	0.842	0.1209
alpha.2:3	1.062	0.0556

Since one of the estimates of the correlation is above 1, the fitted model is unreliable and will not be considered further.

```

#fitting GEE model with autoregressive working correlation matrix
summary(ar.fitted.model<- geeglm(BMI ~ group.rel + gender.rel + exercise
+ month, data=longform.data, id=id, family=gaussian(link="identity"),
corstr="ar1"))
QIC(ar.fitted.model)

```

QIC  
115

```

#fitting GEE model with compound symmetric working correlation matrix
summary(cs.fitted.model<- geeglm(BMI ~ group.rel + gender.rel + exercise
+ month, data=longform.data, id=id, family=gaussian(link="identity"),
corstr="exchangeable"))
QIC(cs.fitted.model)

```

QIC  
115

```

#fitting GEE model with independent working correlation matrix
summary(ind.fitted.model<- geeglm(BMI ~ group.rel + gender.rel + exercise
+ month, data=longform.data, id=id, family=gaussian(link="identity"),
corstr="independence"))
QIC(ind.fitted.model)

```

QIC  
120

(b) Choose the best-fitted model according to the QIC criterion.

SAS selects the model with autoregressive working correlation matrix as the optimal one, according to the QIC criterion. The values are summarized in the table below.

	Toep	AR	CS	Ind
QIC	113.06	112.723	113.068	113.797

In R, the best-fitted models and with the autoregressive and compound symmetric working correlation matrices. We will choose the compound symmetric model as it is a simpler model.

(c) For the best-fitted model, do parts (c)-(e) in Exercise 8.5.

In SAS:

```
/*fitting GEE model with autoregressive working correlation matrix*/
proc genmod;
  class id group gender(ref="F");
  model BMI = group gender exercise month;
  repeated subject = id / type=ar corrw;
run;
```

```
Working Correlation Matrix
      Col1   Col2   Col3
Row1  1.0000  0.9373  0.8785
Row2  0.9373  1.0000  0.9373
Row3  0.8785  0.9373  1.0000
```

```
Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates
Parameter      Estimate Standard Error 95% Confidence Limits      Z Pr > |Z|
Intercept      32.9479    1.5936    29.8244    36.0713 20.67    <.0001
group          Cnt    4.0902    2.0527     0.0670     8.1133  1.99    0.0463
group          Int    0.0000    0.0000     0.0000     0.0000  .        .
gender         M     0.8958    1.9807    -2.9863     4.7779  0.45    0.6511
gender         F     0.0000    0.0000     0.0000     0.0000  .        .
exercise       -0.0241    0.0050    -0.0339    -0.0144 -4.87    <.0001
month          -0.9506    0.1629    -1.2699    -0.6312 -5.83    <.0001
```

The fitted model is  $\hat{E}(BMI) = 32.9479 + 4.0902 \cdot \text{control} + 0.8958 \cdot \text{male} - 0.0241 \cdot \text{exercise} - 0.9506 \cdot \text{month}$ . The working correlation matrix is  $\begin{pmatrix} 1 & 0.9373 & 0.8785 \\ 0.9373 & 1 & 0.9373 \\ 0.8785 & 0.9373 & 1 \end{pmatrix}$ .

Group, exercise, and month are significant predictors of BMI. It is estimated that the BMI for study participants in the control group is, on average, 4.0902 points larger than that for the intervention group participants. For an additional minute of daily exercise, the estimated average BMI decreases by 0.0241 points. For an additional month in the study, the estimated average BMI decreases by 0.9506 points.

Computing the predicted BMI at 3 months for an intervention group female participant, if she exercises for 1 hour every day, we get  $BMI^0 = 32.9479 - 0.0241 \cdot 60 - 0.9506 \cdot 3 = 28.6501$ .

### In SAS:

```
data predict;
input id group$ gender$ exercise month;
cards;
35 Int F 60 3
;

data longform;
set longform predict;
run;

proc genmod;
  class id group gender(ref="F");
  model BMI = group gender exercise month;
  repeated subject = id / type=ar;
  output out=outdata p=pBMI;
run;

proc print data=outdata (firstobs=103) noobs;
  var pBMI;
run;

      pBMI
28.6482
```

### In R:

```
#fitting GEE model with independent working correlation matrix
summary(cs.fitted.model<- geeglm(BMI ~ group.rel + gender.rel + exercise
+ month, data=longform.data, id=id, family=gaussian(link="identity"),
corstr="independence"))
```

#### Coefficients:

	Estimate	Std.err	wald	Pr(> w )
(Intercept)	33.21310	1.57220	446.28	< 2e-16
group.relCnt	3.57677	2.06691	2.99	0.084
gender.relM	0.99212	1.98382	0.25	0.617
exercise	-0.02777	0.00574	23.37	1.3e-06
month	-0.95011	0.20191	22.14	2.5e-06

#### Estimated Correlation Parameters:

	Estimate	Std.err
alpha	0.887	0.0651

The fitted model is  $\hat{E}(BMI) = 33.21310 + 3.57677 \cdot \text{control} + 0.99212 \cdot \text{male} - 0.02777 \cdot \text{exercise} - 0.95011 \cdot \text{month}$ . The working correlation matrix is  $\begin{pmatrix} 1 & 0.887 & 0.887 \\ 0.887 & 1 & 0.887 \\ 0.887 & 0.887 & 1 \end{pmatrix}$ . Length of exercise and month are significant predictors of BMI. For an additional minute of daily exercise, the estimated average BMI decreases by 0.02777 points. For an additional month in the study, the estimated average BMI decreases by 0.95011 points.

To predict BMI at 3 months for an intervention group female participant, if she exercises for 1 hour every day we compute  $BMI^0 = 33.21310 - 0.02777 \cdot 60 - 0.95011 \cdot 3 = 28.69657$ .

### In R:

```
print(predict(cs.fitted.model, data.frame(group.rel="Int", gender.rel="F",
exercise=60, month=3)))
```

28.7

## CHAPTER 9

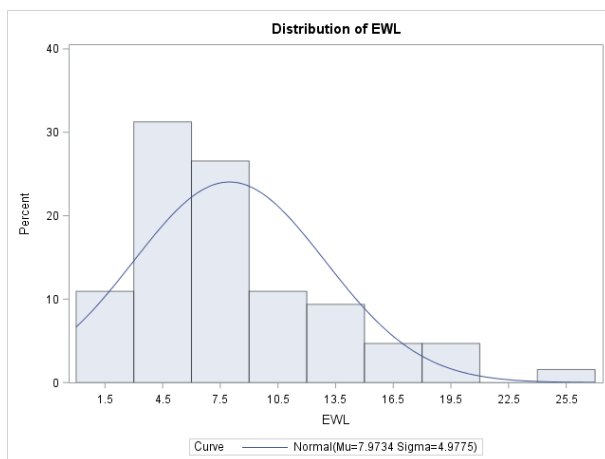
**EXERCISE 9.1.** (a) Plot the histogram of the EWL to see that this variable has a distribution with a long right tail.

In SAS:

```
data weightloss;
input patid group$ gender$ EWL1 EWL2 EWL3 EWL4 @@;
cards;
1 Tx M 11.8 16.4 7.1 4.5 2 Tx F 18.3 7.7 10.7 4.1
3 Tx F 20.1 8.2 7.2 6.3 4 Tx F 15.6 7.8 7.2 2.7
5 Tx M 12.5 8.6 9.7 5.4 6 Tx F 24.4 8.7 6.6 4.7
7 Tx F 18.8 12.3 6.7 4.5 8 Tx M 11.2 9.1 5.6 3.1
9 Cx F 13.9 14.3 4.1 5.0 10 Cx F 6.8 5.2 4.5 1.4
11 Cx M 8.1 12.7 12.3 4.9 12 Cx F 5.6 16.5 4.8 1.8
13 Cx M 9.6 9.9 3.6 3.5 14 Cx M 6.8 7.5 5.1 1.7
15 Cx F 4.7 8.3 3.2 2.4 16 Cx F 6.7 4.1 2.4 1.3
;

/*creating longform dataset*/
data longform;
set weightloss;
array e[4] EWL1 EWL2 EWL3 EWL4;
do visit=1 to 4;
    EWL=e[visit];
    output;
end;
keep patid group gender visit EWL;
run;

/*plotting histogram*/
proc univariate;
var EWL;
histogram / normal;
run;
```



The histogram shows a right-skewed distribution of excess body weight loss amount. In the normality tests, the p-values are below 0.05, confirming that the distribution is not normal.

### Goodness-of-Fit Tests for Normal Distribution

Test	Statistic	p Value
Kolmogorov-Smirnov D	0.13009496	Pr > D <0.010
Cramer-von Mises	W-Sq 0.29506648	Pr > W-Sq <0.005
Anderson-Darling	A-Sq 1.72390544	Pr > A-Sq <0.005

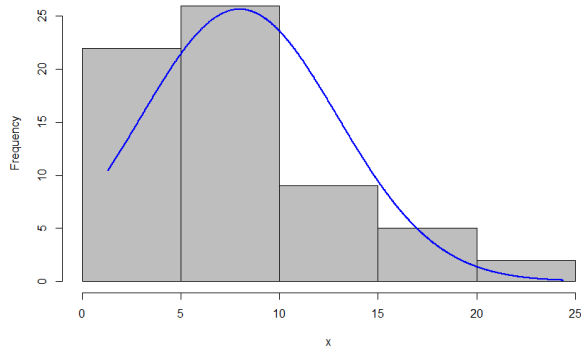
In R:

```
weightloss.data<- read.csv(file="C:/./Exercise9.1Data.csv", header=TRUE, sep=",")

#creating longform dataset
library(reshape2)
longform.data<- melt(weightloss.data, id.vars=c("patid","group","gender"),
variable.name="EWLn",value.name="EWL")

#creating variable for time
visit<- ifelse(longform.data$EWLn=="EWL1",1,ifelse(longform.data$EWLn
=="EWL2",2,ifelse(longform.data$EWLn=="EWL3",3,4)))

#plotting histogram
library(rcompanion)
plotNormalHistogram(longform.data$EWL)
```



```
shapiro.test(longform.data$EWL)
```

Shapiro-wilk normality test

w = 0.91034, p-value = 0.000201

(b) Try to run a generalized random slope and intercept model for the EWL based on a gamma distribution. If it doesn't run, fit an intercept-only model. Discuss the fit of this model. Hint: as a time variable, use visits with values 1, 2, 3, or 4.

In SAS:

```
/*fitting gamma regression model with random slope and intercept*/
proc glimmix method=Laplace;
  class group(ref="Cx") gender(ref="F");
  model EWL = group gender visit / solution dist=gamma link=log;
  random intercept visit / subject=patid type=un;
  covtest/wald;
run;
```

The model doesn't converge.

```
/*fitting gamma regression model with random intercept only*/
proc glimmix method=Laplace;
  class group(ref="Cx") gender(ref="F");
  model EWL = group gender visit / solution dist=gamma link=log;
  random intercept / subject=patid type=un;
  covtest/wald;
run;
```

-2 Log Likelihood 299.97

Covariance Parameter Estimates					
Cov	Parm	Subject	Estimate	Standard Error	Z Value Pr > Z
UN(1,1)	patid		0.04284	0.02584	1.66 0.0487
Residual			0.1131	0.02255	5.01 <.0001

Solutions for Fixed Effects						
Effect	group	gender	Estimate	Standard Error	DF t Value	Pr >  t
Intercept			2.7307	0.1452	13	18.80 <.0001
group	Tx		0.4356	0.1338	47	3.26 0.0021
group	Cx		0	.	.	.
gender		M	0.1008	0.1381	47	0.73 0.4692
gender		F	0	.	.	.
visit			-0.4192	0.03950	47	-10.61 <.0001

```
/*checking model fit*/
proc glimmix;
  class group gender;
  model EWL = group gender visit / dist=gamma link=log;
run;
```

-2 Log Likelihood 305.82

```
data deviance;
  deviance = 305.82 - 299.97;
  pvalue = 1 - probchi(deviance,1);
run;
```

```
proc print noobs;
run;
```

deviance	pvalue
5.85	0.015577

Fitted in SAS, the gamma regression model with random intercept only has a good fit as supported by the p-value below 0.05.

In R:

```
#specifying reference levels
group.rel<- relevel(longform.data$group, ref="Cx")
gender.rel<- relevel(longform.data$gender, ref="F")
```

```
#fitting gamma regression model with random slope and intercept
library(lme4)
summary(fitted.model<- glmer(EWL ~ group.rel + gender.rel
+ visit + (1 + visit | patid), data=longform.data, family=Gamma(link="log")))
```

Random effects:

Groups	Name	Variance	Std.Dev.	Corr
patid	(Intercept)	0.010594	0.10293	
	time	0.002103	0.04585	1.00
Residual		0.124364	0.35265	

Fixed effects:

	Estimate	Std. Error	t value	Pr(> z )
(Intercept)	2.75279	0.15307	17.983	<2e-16
group.relTx	0.46247	0.16694	2.770	0.0056
gender.relM	-0.01279	0.19227	-0.067	0.9470
visit	-0.42625	0.04005	-10.644	<2e-16

```
#checking model fit
null.model<- glm(EWL ~ group.rel + gender.rel + visit, data=longform.data,
family=Gamma(link="log"))
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))
```

14.65092

```
print(pvalue<- pchisq(deviance, df=3, lower.tail = FALSE))
```

0.002140621

Fitted in R, the gamma regression model with a random slope and intercept fits the data well, as suggested by a small p-value in the deviance test.

(c) What parameters are significant at the 5% level? Give the fitted model, specifying all parameter estimates.

The gamma model fitted in SAS has the form  $\hat{E}(EWL) = \exp(2.7307 + 0.4356 \cdot Tx + 0.1008 \cdot male - 0.4192 \cdot visit)$ , and  $\hat{\sigma}^2_{u_1} = 0.04284$ .

The model fitted in R has the estimated parameters  $\hat{E}(EWL) = \exp(2.75279 + 0.46247 \cdot Tx - 0.01279 \cdot male - 0.42625 \cdot visit)$ , and  $\hat{\sigma}^2_{u_1} = 0.010594$ ,  $\hat{\sigma}^2_{u_2} = 0.002103$ , and  $\hat{\rho}_{u_1 u_2} = 1$ .

The variance of the random intercept is significant at the 5% level, indicating that the fitted mixed-effects regression is appropriate. Group and time are significant predictors of EWL.

(d) Give an interpretation of the estimates of significant fixed-effects parameters. Is the new medication superior to the regularly used one?

The estimated average percent excess body weight loss for patients in the treatment group is  $\exp(0.4356) \cdot 100\% = 154.589\%$  of that for patients in the control group, suggesting that the new medication is superior to the regularly used one. With each additional visit, the estimated average percent excess body weight loss changes by  $(\exp(-0.4192) - 1) \cdot 100\% = -34.2427\%$ , that is, decreases by 34.2427%.

(e) What percent excess body weight loss can the doctors expect to see between 3 and 6 months in male patients who will be taking this new medication?

Using the model fitted in SAS, we compute the predicted percent excess body weight loss as  $EWL^0 = \exp(2.7307 + 0.4356 + 0.1008 - 0.4192 \cdot 4) = 4.9052$ .

In SAS:

```
/*using fitted model for prediction*/
data predict;
input patid group$ gender$ visit;
cards;
17 Tx M 4
;

data longform;
set longform predict;
run;

proc glimmix method=Laplace;
class group(ref="Cx") gender(ref="F");
model EWL = group gender visit / dist=gamma link=log;
random intercept / subject=patid type=un;
output out=outdata pred(ilink)=pEWL;
run;

proc print data=outdata (firstobs=65) noobs;
var pEWL;
run;

pEWL
4.90585
```

For the model fitted in R, the predicted value is  $EWL^0 = \exp(2.75279 + 0.46247 - 0.01279 - 0.42625 \cdot 4) = 4.4704$ .

In R:

```
#using fitted model for prediction
print(predict(fitted.model, data.frame(patid=17, group.rel="Tx", gender.rel="M",
visit=4), re.form=NA, type="response"))

4.470331
```

**EXERCISE 9.2.** (a) Model the logistically distributed presence of side effects via a generalized random slope and intercept model. In SAS, to obtain better estimates, scale age by a factor of 100. Discuss the model fit.

In SAS:

```
data pharma;
input patid dosage$ gender$ age week1 week3 week7 week16 @@;
cards;
```

```

1  A F 56 1 1 0 0  2  A F 53 1 1 1 0  3  A F 32 0 1 0 1
4  A F 22 0 0 0 0  5  A F 38 0 0 1 1  6  A F 42 0 1 1 1
7  A F 46 0 1 1 0  8  A M 33 1 1 1 1  9  A M 44 0 0 1 1
10 A M 34 0 1 0 0 11 A M 38 0 0 1 1 12 A M 40 0 0 1 1
13 A M 43 0 0 0 0 14 A M 44 0 0 1 0 15 A F 48 0 0 0 0
16 A F 29 0 1 0 0 17 A F 30 0 0 0 0 18 B F 30 0 0 0 0
19 B F 31 0 0 0 0 20 B F 32 1 1 0 0 21 B F 31 0 0 1 0
22 B F 50 0 0 0 1 23 B F 38 0 0 0 0 24 B M 51 0 0 1 1
25 B M 32 0 0 1 1 26 B M 25 0 0 0 0 27 B M 24 0 0 0 0
28 B M 34 0 0 0 0 29 B M 36 0 0 0 1 30 B M 44 0 0 1 1
31 B M 40 1 1 0 0 32 B M 29 0 1 0 0 33 B M 33 1 1 0 0
34 B M 38 0 0 1 0
;

/*creating longform dataset*/
data longform;
set pharma;
array w[4] (1 3 7 16);
array s[4] week1 week3 week7 week16;
do i=1 to 4;
  week=w[i];
  sideeffects=s[i];
  age=age/100;
  output;
end;
keep patid dosage gender age week sideeffects;
run;

/*fitting logistic regression model with random slope and intercept*/
proc glimmix method=Laplace;
class dosage(ref="B") gender(ref="F");
model sideeffects (event="1") = dosage gender age week / solution dist=binomial
link=logit;
random intercept week / subject=patid type=un;
covtest/wald;
run;

-2 Log Likelihood 147.31

```

Covariance Parameter Estimates						
Cov	Parm	Subject	Estimate	Standard Error	Z Value	Pr >  Z
UN(1,1)	patid		17.8005	11.6113	1.53	0.0626
UN(2,1)	patid		-3.1901	2.0170	-1.58	0.1137
UN(2,2)	patid		0.6761	0.4185	1.62	0.0531

Solutions for Fixed Effects								
Effect		dosage	gender	Estimate	Standard Error	DF	t Value	Pr >  t
Intercept				-1.5838	1.2540	31	-1.26	0.2160
dosage	A			2.1726	1.0608	67	2.05	0.0445
dosage	B			0	.	.	.	.
gender		M		0.6921	0.9485	67	0.73	0.4681
gender		F		0	.	.	.	.
age				-7.9672	2.9241	67	-2.72	0.0082
week				-0.2236	0.1787	33	-1.25	0.2194

```
/*checking model fit*/
proc glimmix;
  class dosage gender;
  model sideeffects = dosage gender age week / dist=binomial link=logit;
run;
```

**-2 Log Likelihood 162.82**

```
data deviance;
  deviance = 162.82 - 147.31;
  pvalue = 1 - probchi(deviance, 3);
run;

proc print noobs;
run;
```

```
deviance    pvalue
    15.51 .001428836
```

The model has a very good fit as evidenced by the tiny p-value in the deviance test.

In R:

```
pharma.data<- read.csv(file="C:/./Exercise9.3Data.csv", header=TRUE, sep=",")

#creating longform dataset
library(reshape2)
longform.data<- melt(pharma.data, id.vars=c("patid", "dosage","gender","age"),
  variable.name="weekn", value.name="sideeffects")

#creating variable for time
week<- ifelse(longform.data$weekn=="week1",1, ifelse(longform.data$weekn
=="week3",3,ifelse(longform.data$weekn=="week7",7,16)))

#fitting logistic model with random slope and intercept
library(lme4)
summary(fitted.model<- glmer(sideeffects ~ dosage + gender + age + week
+ (1 + week | patid), data=longform.data, family=binomial(link="logit")))
```

Random effects:

Groups	Name	Variance	Std.Dev.	Corr
patid	(Intercept)	5.5846	2.3632	
	week	0.2157	0.4644	-0.91

Fixed effects:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-4.24086	1.95170	-2.173	0.0298
dosageB	-0.97776	0.70953	-1.378	0.1682
genderM	0.59956	0.68388	0.877	0.3807
age	0.08335	0.04429	1.882	0.0598
week	-0.01254	0.10554	-0.119	0.9054

```
#checking model fit
null.model<- glm(sideeffects ~ dosage + gender + age + week,
  data=longform.data, family=binomial(link="logit"))
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))
```

**6.901126**

```
print(p.value<- pchisq(deviance, df=3, lower.tail = FALSE))
```

0.07511689

(b) Specify the fitted model, giving the estimates of all parameters. Which parameters are significant at the 5% significance level? At the 10% level?

In SAS, the estimated parameters in the fitted logistic model are:

$$\frac{\hat{P}(\text{side effects})}{1 - \hat{P}(\text{side effects})} = \exp(-1.5838 + 2.1726 \cdot \text{dosage } A + 0.6921 \cdot \text{male} - 7.9672 \cdot \text{age}/100 - 0.2236 \cdot \text{week}), \hat{\sigma}_{u_1}^2 = 17.8005, \hat{\sigma}_{u_2}^2 = 0.6761, \text{ and } \hat{\sigma}_{u_1 u_2} = -3.1901.$$

Variances of random slope and intercept are significant at the 10% level. Dosage and age are significant predictors of the probability of side effects at the 5% level.

In R, the estimated parameters in the fitted model are:

$$\frac{\hat{P}(\text{side effects})}{1 - \hat{P}(\text{side effects})} = \exp(-4.24086 - 0.97776 \cdot \text{dosage } B + 0.59956 \cdot \text{male} + 0.08335 \cdot \text{age} - 0.01254 \cdot \text{week}), \hat{\sigma}_{u_1}^2 = 5.5846, \hat{\sigma}_{u_2}^2 = 0.2157, \text{ and } \hat{\rho}_{u_1 u_2} = -0.91.$$

The variance of random intercept is significantly different from zero since  $p\text{-value} = P(Z > 5.5846/2.3632) = P(Z > 2.363152) = 0.00906 < 0.01$ . Also, age is marginally significant at the 5% level.

(c) Interpret the estimates of the significant beta coefficients. What dosage should be preferred?

For the model fitted in SAS, for the subjects taking dosage A, the estimated odds in favor of side effects are  $\exp(2.1726) \cdot 100\% = 878.1085\%$  of those for subjects taking dosage B. Thus, dosage B should be preferred. Also, as age increases by one year, the estimated odds in favor of side effects change by  $(\exp(-0.079272) - 1) \cdot 100\% = -7.62114\%$ , that is, decrease by 7.62114%.

For the model fitted in R, as age increases by one year, the estimated odds in favor of side effects change by  $(\exp(-0.08335) - 1) \cdot 100\% = -7.99709\%$ , that is, decrease by 7.99709%.

(d) Predict the probability of side effects occurring at week 7 for a 40-year-old woman taking dosage A.

Using the model fitted in SAS, we obtain

$$P^0(\text{side effects} = 1) = \frac{\exp(-1.5838 + 2.1726 - 7.9672 \cdot \frac{40}{100} - 0.2236 \cdot 7)}{1 + \exp(-1.5838 + 2.1726 - 7.9672 \cdot \frac{40}{100} - 0.2236 \cdot 7)} = 0.015318.$$

For the model fitted in R, we predict

$$P^0(\text{side effects} = 1) = \frac{\exp(-4.24086 + 0.08335 \cdot 40 - 0.01254 \cdot 7)}{1 + \exp(-4.24086 + 0.08335 \cdot 40 - 0.01254 \cdot 7)} = 0.269997.$$

In SAS:

```
/*using fitted model for prediction*/
data predict;
input patid dosage$ gender$ age week;
age=age/100;
cards;
35 A F 40 7
;

data longform;
set longform predict;
run;

proc glimmix method=Laplace;
class dosage gender;
model sideeffects (event="1") = dosage gender age week / dist=binomial
link=logit;
random intercept week / subject=patid type=un;
output out=outdata pred(ilink)=psideeffects;
run;

proc print data=outdata (firstobs=137) noobs;
var psideeffects;
run;
```

```
psideeffects
0.015315
```

In R:

```
#using model for prediction
print(predict(fitted.model, data.frame(patid=35, dosage="A", gender="F", age=40,
week=7), re.form=NA, type="response"))
```

```
0.2699819
```

**EXERCISE 9.3.** (a) Fit a random slope and intercept model (or random intercept-only model, if appropriate) for the days with occupancy below 65%. Use the Poisson distribution. Discuss the model fit.

In SAS:

```
data hotels;
input hotel region$ ADR1 OCR1 ADR2 OCR2 ADR3 OCR3 ADR4 OCR4 @@;
cards;
1 rural 88 3 76 8 74 11 78 17 2 rural 79 5 98 9 72 7 54 14
3 rural 84 2 67 4 64 9 98 13 4 rural 79 3 88 4 77 80 66 15
5 rural 68 1 75 8 58 16 80 21 6 rural 82 0 95 4 85 9 90 16
7 rural 92 4 93 8 87 13 92 20 8 rural 58 0 54 9 67 19 84 25
9 rural 84 1 87 9 94 6 92 19 10 rural 98 3 92 0 88 3 80 7
11 urban 112 1 137 11 114 5 137 23 12 urban 104 1 176 8 97 6 146 18
13 urban 195 3 171 5 175 6 137 11 14 urban 128 1 113 10 125 3 126 9
15 urban 96 2 152 10 145 5 153 10 16 urban 98 0 170 9 129 3 148 16
17 urban 119 2 121 8 128 6 147 18 18 urban 120 0 130 0 114 2 108 13
```

```
;
data longform;
set hotels;
array a[4] ADR1 ADR2 ADR3 ADR4;
array o[4] OCR1 OCR2 OCR3 OCR4;
do season=1 to 4;
  ADR=a[season];
  OCR=o[season];
  output;
end;
keep hotel region season ADR OCR;
run;

/*fitting Poisson model with random slope and intercept*/
proc glimmix method=Laplace;
class region;
model OCR = region ADR season / solution dist=poisson link=log;
random intercept season / subject=hotel type=un;
covtest/wald;
run;
```

The model doesn't converge.

```
/*fitting Poisson model with random intercept only*/
proc glimmix method=Laplace;
class region;
model OCR = region ADR season / solution dist=poisson link=log;
random intercept / subject=hotel type=un;
covtest/wald;
run;
```

-2 Log Likelihood 503.98

Covariance Parameter Estimates					
Cov	Parm	Subject	Estimate	Standard Error	Z Value Pr > Z
UN(1,1)	hotel		0.1581	0.06353	2.49 0.0064

Solutions for Fixed Effects					
Effect	region	Estimate	Standard Error	DF t Value	Pr >  t
Intercept		-1.0197	0.4744	16	-2.15 0.0473
region	rural	0.9002	0.2721	52	3.31 0.0017
region	urban	0	.	.	. .
ADR		0.009809	0.003106	52	3.16 0.0026
season		0.5615	0.03986	52	14.09 <.0001

```
/*checking model fit*/
proc glimmix;
class region;
model OCR = region ADR season / dist=poisson link=log;
run;
```

2 Log Likelihood 583.13

```
data deviance;
deviance = 583.13 - 503.98;
```

```

pvalue = 1 - probchi(deviance, 1);
run;

proc print noobs;
run;

```

```

deviance pvalue
79.15      0

```

The model fits the data very well since the value of the deviance test statistic is very large and so the p-value is very small.

In R:

```

hotels.data<- read.csv(file="C:/./Exercise9.3Data.csv", header=TRUE, sep=",")

#creating longform dataset
library(reshape2)
data1<- melt(hotels.data[,c("hotel","region","ADR1","ADR2","ADR3","ADR4")],
id.vars=c("hotel","region"), variable.name="ADRn",value.name="ADR")
data2<- melt(hotels.data[,c("OCR1","OCR2","OCR3","OCR4")],variable.name
="OCRn",value.name="OCR")
longform.data<- cbind(data1,data2)

#creating variable for time
season<- ifelse(longform.data$ADRn=="ADR1",1,ifelse(longform.data$ADRn=="ADR2",
2,ifelse(longform.data$ADRn=="ADR3",3,4)))

#specifying reference level
region.rel<- relevel(longform.data$region, ref="urban")

#fitting Poisson model with random slope and intercept
library(lme4)
summary(fitted.model<- glmer( OCR ~ region.rel + ADR + season + (1 + season|
hotel), data=longform.data, family=poisson(link="log")))

Random effects:
Groups Name          Variance Std.Dev. Corr
hotel  (Intercept)  0.215785  0.4645
      season       0.000462  0.0215   -1.00

Fixed effects:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.027042   0.476134  -2.157  0.031002
region.relrural  0.893183   0.271415   3.291  0.000999
ADR           0.009643   0.003114   3.096  0.001959
season        0.570542   0.043169  13.216 < 2e-16

#checking model fit
null.model<- glm(OCR ~ region.rel + ADR + season, data=longform.data,
family=poisson(link="log"))
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))

79.46935

print(p.value<- pchisq(deviance, df=3, lower.tail = FALSE))

3.988909e-17

```

(b) State the fitted model. Identify all significant parameters. Use  $\alpha = 0.05$ . Are responses over seasons correlated within each hotel?

The Poisson model fitted in SAS has the estimated parameters  $\hat{E}(OCR) = \exp(-1.0197 + 0.9002 \cdot rural + 0.009809 \cdot ADR + 0.5615 \cdot season)$ , and the random intercept has the estimated variance  $\hat{\sigma}_{u_1}^2 = 0.1581$ .

All predictors (region, ADR, and season) are statistically significant, and so is the random intercept.

The model fitted in R has the estimated parameters  $\hat{E}(OCR) = \exp(-1.027042 + 0.893183 \cdot rural + 0.009643 \cdot ADR + 0.570542 \cdot season)$ ,  $\hat{\sigma}_{u_1}^2 = 0.215785$ ,  $\hat{\sigma}_{u_2}^2 = 0.000462$ , and  $\hat{\rho}_{u_1 u_2} = -1$ .

(c) Interpret the estimates of all significant beta coefficients.

For the model fitted in SAS, we interpret the estimated regression coefficients as follows. For rural hotels, the estimated average number of days the hotel occupancy rate is below 65% is  $\exp(0.9002) \cdot 100\% = 246.0095\%$  of that for urban hotels. As the daily average rate increases by one dollar, the estimated average number of days the hotel occupancy rate is below 65% increases by  $(\exp(0.009809) - 1) \cdot 100\% = 0.985727\%$ . Every next season (summer to fall to winter to spring), the estimated average number of days the hotel occupancy rate is below 65% increases by  $(\exp(0.5615) - 1) \cdot 100\% = 75.33005\%$ .

For the model fitted in R, the interpretation goes as follows. For rural hotels, the estimated average number of days the hotel occupancy rate is below 65% is  $\exp(0.893183) \cdot 100\% = 244.2893\%$  of that for urban hotels. As the daily average rate increases by one dollar, the estimated average number of days the hotel occupancy rate is below 65% increases by  $(\exp(0.009643) - 1) \cdot 100\% = 0.968964\%$ . Every next season (summer to fall to winter to spring), the estimated average number of days the hotel occupancy rate is below 65% increases by  $(\exp(0.570542) - 1) \cdot 100\% = 76.92257\%$ .

(d) Predict the number of days with an occupancy rate below 65% for the winter season in a rural hotel with an average daily rate of \$75.

For the model fitted in SAS, the predicted value is  $OCR^0 = \exp(-1.0197 + 0.9002 + 0.009809 \cdot 75 + 0.5615 \cdot 3) = 9.98092$ .

In SAS:

```
/*using the fitted model for prediction*/
data predict;
input hotel region$ ADR season;
cards;
19 rural 75 3
;

data longform;
set longform predict;
run;

proc glimmix method=Laplace;
```

```

class region;
model OCR = region ADR season / dist=poisson link=log;
random intercept / subject=hotel type=un;
output out=outdata pred(ilink)=pOCR;
run;

proc print data=outdata (firstobs=73) noobs;
var pOCR;
run;

```

pOCR  
 9.97988

For the model fitted in R, the predicted value is  $OCR^0 = \exp(-1.027042 + 0.893183 + 0.009643 \cdot 75 + 0.570542 \cdot 3) = 9.98408$ .

In R:

```

#using the fitted model for prediction
print(predict(fitted.model, data.frame(hotel=19, region.rel="rural", ADR=75,
season=3), re.form=NA, type="response"))

```

9.984353

**EXERCISE 9.4.** (a) Run the random slope and intercept (possibly random intercept-only) model for the PDC, using a beta distribution. Does the model fit the data well? Use the 10% level of significance.

In SAS:

```

data adherence;
input id gender$ age edu$ pdc1 pdc2 pdc3 pdc4 @@;
cards;
1 F 38 >HS 0.05 0.25 0.62 0.87 2 M 57 >HS 0.10 0.60 0.77 0.25
3 M 46 <HS 0.05 0.10 0.15 0.15 4 M 57 <HS 0.02 0.20 0.37 0.37
5 F 39 <HS 0.23 0.90 0.93 0.95 6 F 40 <HS 0.12 0.13 0.57 0.90
7 F 66 HSgrad 0.02 0.28 0.12 0.57 8 F 50 HSgrad 0.20 0.23 0.38 0.10
9 F 43 >HS 0.08 0.72 0.87 0.97 10 F 69 HSgrad 0.18 0.50 0.75 0.63
11 F 45 <HS 0.03 0.10 0.40 0.98 12 F 41 >HS 0.13 0.82 0.93 0.65
13 F 43 HSgrad 0.05 0.75 0.98 0.38 14 M 49 HSgrad 0.17 0.37 0.20 0.58
15 F 39 >HS 0.05 0.20 0.92 0.30 16 M 47 >HS 0.03 0.57 0.67 0.60
17 M 65 >HS 0.02 0.18 0.15 0.33 18 F 59 <HS 0.03 0.07 0.18 0.37
19 F 41 >HS 0.02 0.88 0.92 0.85 20 F 49 >HS 0.05 0.13 0.05 0.03
21 F 36 HSgrad 0.08 0.20 0.32 0.13 22 M 42 HSgrad 0.13 0.33 0.22 0.37
23 M 45 HSgrad 0.03 0.15 0.33 0.70 24 M 49 <HS 0.03 0.03 0.18 0.85
25 M 56 <HS 0.03 0.67 0.48 0.50 26 M 49 HSgrad 0.07 0.15 0.20 0.12
27 M 41 HSgrad 0.12 0.23 0.53 0.32
;

/*creating longform dataset*/
data longform;
set adherence;
array p[4] pdc1 pdc2 pdc3 pdc4;
do refill=1 to 4;

```

```

    pdc=p[refill];
    output;
end;
keep id gender age edu refill pdc;
run;

/*fitting beta model with random slope and intercept*/
proc glimmix method=Laplace;
  class gender(ref="M") edu(ref="<HS");
  model pdc = gender age edu refill / solution dist=beta link=logit;
  random intercept refill / subject=id type=un;
  covtest/wald;
run;

```

The model doesn't converge.

```

/*fitting beta model with random intercept only*/
proc glimmix method=Laplace;
  class gender(ref="M") edu(ref="HSgrad");
  model pdc = gender age edu refill / solution dist=beta link=logit;
  random intercept / subject=id type=un;
  covtest/wald;
run;

```

-2 Log Likelihood -69.56

Covariance Parameter Estimates					
Cov	Parm	Subject	Estimate	Standard Error	Z Value Pr >  t
UN(1,1)	id		0.1638	0.1253	1.31 0.0955
Scale			3.4024	0.5544	6.14 <.0001

Solutions for Fixed Effects						
Effect	gender	edu	Estimate	Standard Error	DF	t Value Pr >  t
Intercept			-1.4351	0.7980	22	-1.80 0.0859
gender	F		0.4345	0.2561	80	1.70 0.0936
gender	M		0	.	.	.
age			-0.02129	0.01427	80	-1.49 0.1397
edu		<HS	0.1276	0.2999	80	0.43 0.6715
edu		>HS	0.2823	0.2982	80	0.95 0.3467
edu		HSgrad	0	.	.	.
refill			0.6161	0.08994	80	6.85 <.0001

```

/*checking model fit*/
proc glimmix;
  class gender edu;
  model pdc = gender age edu refill / dist=beta link=logit;
run;

```

-2 Log Likelihood -66.83

```

data deviance;
  deviance = -66.83 - (-69.56);
  pvalue = 1 - probchi(deviance, 1);
run;

```

```
proc print noobs;
run;
deviance    pvalue
      2.73 0.098479
```

In R:

```
adherence.data<- read.csv(file="C:/./Exercise9.4Data.csv", header=TRUE, sep=",")

#creating longform dataset and time variable
library(reshape2)
longform.data<- melt(adherence.data, id.vars=c("id","gender","age","edu"),
variable.name="pdcn",value.name="pdc")

#creating variable for time
refill<- ifelse(longform.data$pdcn=="pdc1",1,ifelse(longform.data$pdcn=="pdc2",
2,ifelse(longform.data$pdcn=="pdc3",3,4)))

#specifying reference levels
gender.rel<- relevel(longform.data$gender, ref="M")
edu.rel<- relevel(longform.data$edu, ref="HSgrad")

#fitting beta model with random slope and intercept
library(glmmTMB)
summary(glmmTMB(pdc ~ gender.rel + age + edu.rel + refill + (1 + refill | id),
data=longform.data, family=beta_family(link="logit")))
```

The model doesn't converge.

```
#fitting beta model with random intercept only
summary(fitted.model<- glmmTMB(pdc ~ gender.rel + age + edu.rel + refill + (1 |
id), data=longform.data, family=beta_family(link="logit")))
```

Random effects:

```
Groups Name      Variance Std.Dev.
id      (Intercept) 0.1638   0.4048
```

Overdispersion parameter for beta family (): 3.4

Conditional model:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-1.43507	0.79810	-1.798	0.0722
gender.relF	0.43451	0.25607	1.697	0.0897
age	-0.02129	0.01427	-1.491	0.1358
edu.rel<HS	0.12764	0.29991	0.426	0.6704
edu.rel>HS	0.28231	0.29822	0.947	0.3438
refill	0.61606	0.08994	6.850	7.4e-12

```
#checking model fit
library(betareg)
null.model<- betareg(pdc ~ gender.rel + age + edu.rel + refill,
data=longform.data, link="logit")
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))
```

2.728118

```
print(p.value<- pchisq(deviance, df=1, lower.tail = FALSE))
```

0.0985954

(b) Specify the fitted model and interpret all estimated significant fixed-effects parameters. Use the significance level of 0.10.

The fitted beta model has the estimated parameters

$$\hat{\mu} = \frac{\exp(-1.4351 + 0.4345 \cdot \text{female} - 0.02129 \cdot \text{age} + 0.1276 \cdot <HS + 0.2823 \cdot >HS + 0.6161 \cdot \text{refill})}{1 + \exp(-1.4351 + 0.4345 \cdot \text{female} - 0.02129 \cdot \text{age} + 0.1276 \cdot <HS + 0.2823 \cdot >HS + 0.6161 \cdot \text{refill})}, \hat{\sigma}_{u_1}^2 = 0.1638, \text{ and } \hat{\phi} = 3.4.$$

Gender and refill are significant predictors of  $\mu$ , and the variance of the random intercept is significant.

For females, the ratio of the estimated proportion of days that a patient took a diabetes medication and the estimated proportion of days that the medication wasn't taken is  $\exp(0.4345) \cdot 100\% = 154.4191\%$  of that for males. With every additional refill, the estimated ratio increases by  $(\exp(0.6161) - 1) \cdot 100\% = 85.16923\%$ .

(c) What is the predicted PDC value for the second refill of medication for a 50-year-old man with a Bachelor's degree?

$$\text{The predicted value is } PDC^0 = \frac{\exp(-1.4351 - 0.02129 \cdot 50 + 0.2823 + 0.6161 \cdot 2)}{1 + \exp(-1.4351 - 0.02129 \cdot 50 + 0.2823 + 0.6161 \cdot 2)} = 0.271881.$$

In SAS:

```
/*using fitted model for prediction*/
data predict;
input id gender$ age edu$ refill;
cards;
28 M 50 >HS 2
;

data longform;
set longform predict;
run;

proc glimmix method=Laplace;
class gender edu;
model pdc = gender age edu refill / dist=beta link=logit;
random intercept / subject=id type=un;
output out=outdata pred(ilink)=ppdc;
run;

proc print data=outdata(firstobs=109) noobs;
var ppdc;
run;

ppdc
0.27191
```

In R:

```
#using fitted model for prediction
print(predict(fitted.model, data.frame(id=28, gender.rel="M", age=50,
```

```
edu.rel=">HS", refill=2), allow.new.levels=TRUE, type="response"))
```

0.2719147

**EXERCISE 9.5.** Use the data in Exercise 9.1 to do the following:

(a) Fit the GEE models for the EWL with the gamma underlying distribution and with unstructured, Toeplitz (only in SAS), autoregressive, compound symmetric, and independent working correlation matrices.

In SAS:

```
data weightloss;
input patid group$ gender$ EWL1 EWL2 EWL3 EWL4 @@;
cards;
1 Tx M 11.8 16.4 7.1 4.5 2 Tx F 18.3 7.7 10.7 4.1
3 Tx F 20.1 8.2 7.2 6.3 4 Tx F 15.6 7.8 7.2 2.7
5 Tx M 12.5 8.6 9.7 5.4 6 Tx F 24.4 8.7 6.6 4.7
7 Tx F 18.8 12.3 6.7 4.5 8 Tx M 11.2 9.1 5.6 3.1
9 Cx F 13.9 14.3 4.1 5.0 10 Cx F 6.8 5.2 4.5 1.4
11 Cx M 8.1 12.7 12.3 4.9 12 Cx F 5.6 16.5 4.8 1.8
13 Cx M 9.6 9.9 3.6 3.5 14 Cx M 6.8 7.5 5.1 1.7
15 Cx F 4.7 8.3 3.2 2.4 16 Cx F 6.7 4.1 2.4 1.3
;

/*creating longform dataset*/
data longform;
set weightloss;
array e[4] EWL1 EWL2 EWL3 EWL4;
do visit=1 to 4;
    EWL=e[visit];
    output;
end;
keep patid group gender visit EWL;
run;

/*fitting GEE gamma model with unstructured working correlation matrix*/
proc genmod;
class patid group(ref="Cx") gender(ref="F");
model EWL = group gender visit / dist=gamma link=log;
repeated subject = patid / type=un;
run;
```

**QIC 2000.1740**

```
/*fitting GEE gamma model with Toeplitz working correlation matrix*/
proc genmod;
class patid group(ref="Cx") gender(ref="F");
model EWL = group gender visit / dist=gamma link=log;
repeated subject = patid / type=mdep(3);
run;
```

**QIC 2114.9666**

```
/*fitting GEE gamma model with autoregressive working correlation matrix*/
proc genmod;
class patid group(ref="Cx") gender(ref="F");
model EWL = group gender visit / dist=gamma link=log;
```

```

    repeated subject = patid / type=ar;
run;

```

#### QIC 1926.1829

```

/*fitting GEE gamma model with compound symmetric working correlation matrix*/
proc genmod;
  class patid group(ref="Cx") gender(ref="F");
  model EWL = group gender visit / dist=gamma link=log;
  repeated subject = patid / type=cs;
run;

```

#### QIC 1980.8110

```

/*fitting GEE gamma model with independent working correlation matrix*/
proc genmod;
  class patid group(ref="Cx") gender(ref="F");
  model EWL = group gender visit / dist=gamma link=log;
  repeated subject = patid / type=ind;
run;

```

#### QIC 1980.8110

In R:

```

weightloss.data<- read.csv(file="C:/./Exercise9.1Data.csv", header=TRUE, sep=",")

#creating longform dataset
library(reshape2)
longform.data<- melt(weightloss.data, id.vars=c("patid", "group", "gender"),
variable.name="EWLn",value.name="EWL")

#sorting data by id
longform.data<- longform.data[order(longform.data$patid),]

#creating variable for time
visit<- ifelse(longform.data$EWLn=="EWL1",1, ifelse(longform.data$EWLn=="EWL2",2,
ifelse(longform.data$EWLn=="EWL3",3,4)))

#specifying reference levels
group.rel<- relevel(as.factor(longform.data$group), ref="Cx")
gender.rel<- relevel(as.factor(longform.data$gender), ref="F")

library(geepack)
library(MuMIn)

#fitting GEE gamma model with unstructured working correlation matrix
summary(un.fitted.model<- geeglm(EWL ~ group.rel + gender.rel + visit,
data=longform.data, id=patid, family=Gamma(link="log"), corstr="unstructured"))
QIC(un.fitted.model)

```

#### QIC 2142

```

#fitting GEE gamma model with autoregressive working correlation matrix
summary(ar.fitted.model<- geeglm(EWL ~ group.rel + gender.rel + visit,
data=longform.data, id=patid, family=Gamma(link="log"), corstr="ar1"))
QIC(ar.fitted.model)

```

QIC

```
#fitting GEE gamma model with compound symmetric working correlation matrix
summary(cs.fitted.model<- geeglm(EWL ~ group.rel + gender.rel + visit,
data=longform.data, id=patid, family=Gamma(link="log"), corstr="exchangeable"))
QIC(cs.fitted.model)
```

QIC  
2113

```
#fitting GEE gamma model with independent working correlation matrix
summary(ind.fitted.model<- geeglm(EWL ~ group.rel + gender.rel + visit,
data=longform.data, id=patid, family=Gamma(link="log"), corstr="independence"))
QIC(ind.fitted.model)
```

QIC  
2113

(b) Compare model fits. Use the QIC criterion.

For the models fitted in SAS, the autoregressive has a better fit since the QIC value for this model is the smallest. The values are summarized in the table below.

	UN	Toep	AR	CS	Ind
QIC	2000.17	2114.97	1926.18	1980.81	1980.81

The best model in R is also the one with the autoregressive working correlation matrix.

(c) For the model that has the best fit, do questions (c)-(e) from Exercise 9.1.

In SAS:

```
/*fitting GEE gamma model with autoregressive working correlation matrix*/
proc genmod;
  class patid group(ref="Cx") gender(ref="F");
  model EWL = group gender visit / dist=gamma link=log;
  repeated subject = patid / type=ar corrw;
run;
```

```
Working Correlation Matrix
      col1  col2  col3  col4
Row1 1.0000 0.1504 0.0226 0.0034
Row2 0.1504 1.0000 0.1504 0.0226
Row3 0.0226 0.1504 1.0000 0.1504
Row4 0.0034 0.0226 0.1504 1.0000
```

```
Analysis of GEE Parameter Estimates
Empirical Standard Error Estimates
Parameter      Estimate Standard Error 95% Confidence Limits      Z Pr > |Z|
Intercept      2.7371    0.1329      2.4766      2.9976 20.59  <.0001
group Tx       0.4107    0.1334      0.1493      0.6721  3.08  0.0021
group Cx       0.0000    0.0000      0.0000      0.0000  .      .
gender M       0.0893    0.1396     -0.1843      0.3629  0.64  0.5225
```

Analysis Of GEE Parameter Estimates							
Empirical Standard Error Estimates							
Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr >  Z
gender	F	0.0000	0.0000	0.0000	0.0000	.	.
visit		-0.4099	0.0264	-0.4616	-0.3582	-15.54	<.0001

In the fitted GEE gamma model, the percent excess body weight loss has the estimated mean  $\hat{E}(EWL) = \exp(2.7371 + 0.4107 \cdot Tx + 0.0893 \cdot male - 0.4099 \cdot visit)$ , and the estimated

correlation matrix for each individual  $\begin{pmatrix} 1.000 & 0.1504 & 0.0226 & 0.0034 \\ 0.1504 & 1.000 & 0.1504 & 0.0226 \\ 0.0226 & 0.1504 & 1.000 & 0.1504 \\ 0.0034 & 0.0226 & 0.1504 & 1.000 \end{pmatrix}$ . Group and visit are

significant predictors of EWL. The estimated average percent excess body weight loss for patients in the treatment group is  $\exp(0.4107) \cdot 100\% = 150.7873\%$  of that for patients in the control group, suggesting that the new medication is superior to the regularly used one. With each additional visit, the estimated average percent excess body weight loss changes by  $(\exp(-0.4099) - 1) \cdot 100\% = -33.6283\%$ , that is, decreases by 33.6283%.

Next, we compute the percent excess body weight loss that the doctors can expect to see between 3 and 6 months in male patients who will be taking this new medication. We obtain

$$EWL^0 = \exp(2.7371 + 0.4107 + 0.0893 - 0.4099 \cdot 4) = 4.940665.$$

In SAS:

```
data predict;
input patid group$ gender$ visit;
cards;
17 Tx M 4
;

data longform;
set longform predict;
run;

proc genmod;
class patid group gender
model EWL = group gender visit / dist=gamma link=log;
repeated subject = patid / type=ar;
output out=outdata p=pEWL;
run;

proc print data=outdata (firstobs=65) noobs;
var pEWL;
run;

pEWL
4.93989
```

In R:

```
#fitting GEE gamma model with autoregressive working correlation matrix
summary(ar.fitted.model<- geeglm(EWL ~ group.rel + gender.rel + visit,
data=longform.data, id=patid, family=Gamma(link="log"), corstr="ar1"))
```

Coefficients:

	Estimate	Std.err	wald	Pr(> w )
(Intercept)	2.7307	0.1323	425.77	<2e-16
group.relTx	0.4159	0.1336	9.69	0.0019
gender.relM	0.0865	0.1396	0.38	0.5358
visit	-0.4091	0.0263	241.37	<2e-16

Estimated Correlation Parameters:

	Estimate	Std.err
alpha	0.194	0.172

In the fitted model, the percent excess body weight loss has a gamma distribution with the estimated mean  $\hat{E}(EWL) = \exp(2.7307 + 0.4159 \cdot Tx + 0.0865 \cdot male - 0.4091 \cdot visit)$ , and the

estimated correlation matrix for each individual 
$$\begin{pmatrix} 1 & 0.194 & (0.194)^2 & (0.194)^3 \\ 0.194 & 1 & 0.194 & (0.194)^2 \\ (0.194)^2 & 0.194 & 1 & 0.194 \\ (0.194)^3 & (0.194)^2 & 0.194 & 1 \end{pmatrix} =$$

$$\begin{pmatrix} 1 & 0.194 & 0.037636 & 0.007301 \\ 0.194 & 1 & 0.194 & 0.037636 \\ 0.037636 & 0.194 & 1 & 0.194 \\ 0.007301 & 0.037636 & 0.194 & 1 \end{pmatrix}$$
. Group and visit are significant predictors of EWL. Group and

visit are significant predictors of EWL. The estimated average percent excess body weight loss for patients in the treatment group is  $\exp(0.4159) \cdot 100\% = 151.5734\%$  of that for patients in the control group, suggesting that the new medication is superior to the regularly used one. With each additional visit, the estimated average percent excess body weight loss changes by  $(\exp(-0.4091) - 1) \cdot 100\% = -33.5752\%$ , that is, decreases by 33.5752%.

To find the percent excess body weight loss that the doctors can expect to see between 3 and 6 months in male patients who will be taking this new medication, we write

$$EWL^0 = \exp(2.7307 + 0.4159 + 0.0865 - 0.4091 \cdot 4) = 4.936714.$$

In R:

```
#using best-fitted model for prediction
print(predict(ar.fitted.model, data.frame(patid=17, group.rel="Tx",
gender.rel="M", visit=4), type="response"))
```

4.94

**EXERCISE 9.6** Use the data in Exercise 9.3 to carry out the following analysis:

(a) Fit the generalized estimating equations models for the logistically distributed presence or absence of side effects, with unstructured, Toeplitz, autoregressive, compound symmetric, and independent working correlation matrices.

In SAS:

```
data pharma;
input patid dosage$ gender$ age week1 week3 week7 week16 @@;
cards;
1 A F 56 1 1 0 0 2 A F 53 1 1 1 0 3 A F 32 0 1 0 1
```

```

4  A F 22 0 0 0 0 5  A F 38 0 0 1 1 6  A F 42 0 1 1 1
7  A F 46 0 1 1 0 8  A M 33 1 1 1 1 9  A M 44 0 0 1 1
10 A M 34 0 1 0 0 11 A M 38 0 0 1 1 12 A M 40 0 0 1 1
13 A M 43 0 0 0 0 14 A M 44 0 0 1 0 15 A F 48 0 0 0 0
16 A F 29 0 1 0 0 17 A F 30 0 0 0 0 18 B F 30 0 0 0 0
19 B F 31 0 0 0 0 20 B F 32 1 1 0 0 21 B F 31 0 0 1 0
22 B F 50 0 0 0 1 23 B F 38 0 0 0 0 24 B M 51 0 0 1 1
25 B M 32 0 0 1 1 26 B M 25 0 0 0 0 27 B M 24 0 0 0 0
28 B M 34 0 0 0 0 29 B M 36 0 0 0 1 30 B M 44 0 0 1 1
31 B M 40 1 1 0 0 32 B M 29 0 1 0 0 33 B M 33 1 1 0 0
34 B M 38 0 0 1 0
;

```

```

/*creating longform dataset*/
data longform;
set pharma;
array w[4] (1 3 7 16);
array s[4] week1 week3 week7 week16;
do i=1 to 4;
    week=w[i];
    sideeffects=s[i];
    age=age/100;
    output;
end;
keep patid dosage gender age week sideeffects;
run;

/*fitting GEE logistic model with unstructured working correlation matrix*/
proc genmod;
class patid dosage gender(ref="F");
model sideeffects (event="1") = dosage gender age week / dist=binomial
    link=logit;
repeated subject = patid / type=un;
run;

```

#### QIC 173.0832

```

/*fitting GEE logistic model with Toeplitz working correlation matrix*/
proc genmod;
class patid dosage gender(ref="F");
model sideeffects (event="1") = dosage gender age week / dist=binomial
    link=logit;
repeated subject = patid / type=mdep(3);
run;

```

#### QIC 173.5626

```

/*fitting GEE logistic model with autoregressive working correlation matrix*/
proc genmod;
class patid dosage gender(ref="F");
model sideeffects (event="1") = dosage gender age week / dist=binomial
    link=logit;
repeated subject = patid / type=ar;
run;

```

#### QIC 173.6008

```

/*fitting GEE logistic model with compound symmetric working correlation matrix*/
proc genmod;
class patid dosage gender(ref="F");

```

```

    model sideeffects (event="1") = dosage gender age week / dist=binomial
    link=logit;
    repeated subject = patid / type=cs;
run;

```

#### QIC 174.0362

```

/*fitting GEE logistic model with independent working correlation matrix*/
proc genmod;
  class patid dosage gender(ref="F");
  model sideeffects (event="1") = dosage gender age week / dist=binomial
    link=logit;
  repeated subject = patid / type=ind;
run;

```

#### QIC 174.0450

In R:

```

pharma.data<- read.csv(file="C:/./Exercise9.3Data.csv", header=TRUE, sep=",")

#creating longform dataset
library(reshape2)
longform.data<- melt(pharma.data, id.vars=c("patid", "dosage","gender","age"),
  variable.name="weekn", value.name="sideeffects")

#sorting data by id
longform.data<- longform.data[order(longform.data$patid),]

#creating variable for time
week<- ifelse(longform.data$weekn=="week1",1,
  ifelse(longform.data$weekn=="week3",3,
  ifelse(longform.data$weekn=="week7",7,16)))

#specifying reference level
dosage.rel<- releval(as.factor(longform.data$dosage), ref="B")

library(geepack)
library(MuMIn)

#fitting GEE logistic model with unstructured working
correlationsummary(un.fitted.model<- geeglm(sideeffects ~ dosage.rel + gender +
age + week, data=longform.data, id=patid, family=binomial(link="logit"),
corstr="unstructured"))

```

QIC  
171

```

#fitting GEE logistic model with autoregressive working correlation matrix
summary(ar.fitted.model<- geeglm(sideeffects ~ dosage.rel + gender + age + week,
data=longform.data, id=patid, family=binomial(link="logit"), corstr="ar1"))
QIC(ar.fitted.model)

```

QIC  
171

```

#fitting GEE logistic model with compound symmetric working correlation matrix

```

```
summary(cs.fitted.model<- geeglm(sideeffects ~ dosage.rel + gender + age + week,
data=longform.data, id=patid, family=binomial(link="logit"),
corstr="exchangeable"))
QIC(cs.fitted.model)
```

QIC  
171

```
#fitting GEE logistic model with independent working correlation matrix
summary(ind.fitted.model<- geeglm(sideeffects ~ dosage.rel + gender + age + week,
data=longform.data, id=patid, family=binomial(link="logit"),
corstr="independence"))
QIC(ind.fitted.model)
```

QIC  
171

(b) Choose the best model according to the QIC value.

Among the models fitted in SAS, the QIC-optimal model is the one with the unstructured working correlation matrix. QIC values are summarized in the table below.

	UN	Toep	AR	CS	Ind
QIC	173.0832	173.5626	173.6008	174.0362	174.0450

In R, all the models have the same value of QIC. We will choose the simplest, with the independent working correlation matrix.

(c) For the best-fitted model, answer questions (b)-(d) from Exercise 9.3.

In SAS:

```
/*fitting GEE logistic model with Toeplitz working correlation matrix*/
proc genmod;
  class patid dosage gender(ref="F");
  model sideeffects = dosage gender age week / dist=binomial link=logit;
  repeated subject = patid / type=un corrw;
run;
```

```
Working Correlation Matrix
      Col1   Col2   Col3   Col4
Row1  1.0000  0.7518 -0.1246 -0.2332
Row2  0.7518  1.0000 -0.2207 -0.2350
Row3 -0.1246 -0.2207  1.0000  0.4913
Row4 -0.2332 -0.2350  0.4913  1.0000
```

```
Analysis of GEE Parameter Estimates
Empirical Standard Error Estimates
Parameter  Estimate Standard Error 95% Confidence Limits  Z  Pr > |Z|
Intercept  -1.2221  0.4920  -2.1864  -0.2579 -2.48  0.0130
dosage     A   0.9400  0.4154   0.1258   1.7542  2.26  0.0237
dosage     B   0.0000  0.0000   0.0000   0.0000  .    .
gender     M   0.4448  0.4251  -0.3883   1.2779  1.05  0.2954
```

Analysis of GEE Parameter Estimates							
Empirical Standard Error Estimates							
Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr >  Z	
gender	F	0.0000	0.0000	0.0000	0.0000	.	.
age		-2.3683	0.9690	-4.2675	-0.4691	-2.44	0.0145
week		-0.0100	0.0349	-0.0783	0.0584	-0.29	0.7746

In the fitted GEE logistic model, the response has estimated mean  $\hat{E}(\text{side effects}) = \hat{P}(\text{side effects} = 1) =$

$$= \frac{\exp(-1.2221 + 0.94 \cdot \text{dosage A} + 0.4448 \cdot \text{male} - 2.3683 \cdot \frac{\text{age}}{100} - 0.01 \cdot \text{week})}{1 + \exp(-1.2221 + 0.94 \cdot \text{dosage A} + 0.4448 \cdot \text{male} - 2.3683 \cdot \frac{\text{age}}{100} - 0.01 \cdot \text{week})}, \text{ and the estimated}$$

$$\text{correlation matrix for each individual} \begin{pmatrix} 1.000 & 0.7518 & -0.1246 & -0.2332 \\ 0.7518 & 1.000 & -0.2207 & -0.2350 \\ -0.1246 & -0.2207 & 1.000 & 0.4913 \\ -0.2332 & -0.2350 & 0.4913 & 1.000 \end{pmatrix}.$$

Dosage and age are significant predictors at the 5% level. For the subjects taking dosage A, the estimated odds in favor of side effects are  $\exp(0.94) \cdot 100\% = 256.0\%$  of those for subjects taking dosage B. Thus, dosage B should be preferred. Also, as age increases by one year, the estimated odds in favor of side effects change by  $(\exp(-2.3683) - 1) \cdot 100\% = -90.636\%$ , that is, decrease by 90.636%.

Further, we predict the probability of side effects occurring at week 7 for a 40-year-old woman taking dosage A. The prediction is

$$P^0(\text{side effects} = 1) = \frac{\exp(-1.2221 + 0.94 - 2.3683 \cdot 40/100 - 0.01 \cdot 7)}{1 + \exp(-1.2221 + 0.94 - 2.3683 \cdot 40/100 - 0.01 \cdot 7)} = 0.214263.$$

In SAS:

```
data predict;
input patid dosage$ gender$ age week;
age=age/100;
cards;
35 A F 40 7
;

data longform;
set longform predict;
run;

proc genmod;
class patid dosage gender;
model sideeffects = dosage gender age week / dist=binomial link=logit;
repeated subject = patid / type=un;
output out=outdata p=psideeffects;
run;

proc print data=outdata (firstobs=137) noobs;
var psideeffects;
run;
```

```
psideeffects
0.21427
```

In R:

```
#fitting GEE logistic model with independent working correlation matrix
summary(ind.fitted.model<- geeglm(sideeffects ~ dosage.rel + gender + age
+ week, data=longform.data, id=patid, family=binomial(link="logit"),
corstr="independence"))
```

	Estimate	Std.err	Wald	Pr(> W )
(Intercept)	-3.5341	0.9785	13.04	0.0003
dosage.relA	0.5888	0.4019	2.15	0.1429
genderM	0.4682	0.3963	1.40	0.2375
age	0.0518	0.0222	5.44	0.0197
week	0.0374	0.0320	1.37	0.2422

Estimated Correlation Parameters:

	Estimate
alpha	0

In the fitted model, the response has estimated mean  $\hat{E}(\text{side effects}) = \hat{P}(\text{side effects} = 1) =$

$$= \frac{\exp(-3.5341 + 0.5888 \cdot \text{dosage B} + 0.4682 \cdot \text{male} + 0.0518 \cdot \text{age} + 0.0374 \cdot \text{week})}{1 + \exp(-3.5341 + 0.5888 \cdot \text{dosage B} + 0.4682 \cdot \text{male} + 0.0518 \cdot \text{age} + 0.0374 \cdot \text{week})}, \text{ and the estimated}$$

correlation matrix for each individual  $\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$ . Age is the only significant predictor at the 5%

level. In this model, the coefficient corresponding to dosage A is positive, suggesting that dosage A might be the winner. Dosage, however, is not a statistically significant predictor.

As age increases by one year, the estimated odds in favor of side effects increase by  $(\exp(0.0518) - 1) \cdot 100\% = 5.3165\%$ .

Finally, we predict the probability of side effects occurring at week 7 for a 40-year-old woman taking dosage A. The predicted value is

$$P^0(\text{side effects} = 1) = \frac{\exp(-3.5341 + 0.0518 \cdot 40 + 0.0374 \cdot 7)}{1 + \exp(-3.5341 + 0.0518 \cdot 40 + 0.0374 \cdot 7)} = 0.2314.$$

In R:

```
print(predict(ind.fitted.model, data.frame(patid=35, dosage.rel="A", gender="F",
age=40, week=7), type="response"))
```

```
0.352
```

**EXERCISE 9.7.** Consider the data given in Exercise 9.3. Answer the questions below.

(a) Fit a generalized estimating equations model for the days with occupancy below 65% based on a Poisson distribution. Try different working correlation matrices: unstructured, Toeplitz, autoregressive, compound symmetric, and independent.

In SAS:

```
data hotels;
input hotel region$ ADR1 OCR1 ADR2 OCR2 ADR3 OCR3 ADR4 OCR4 @@;
cards;
1 rural 88 3 76 8 74 11 78 17 2 rural 79 5 98 9 72 7 54 14
3 rural 84 2 67 4 64 9 98 13 4 rural 79 3 88 4 77 80 66 15
5 rural 68 1 75 8 58 16 80 21 6 rural 82 0 95 4 85 9 90 16
7 rural 92 4 93 8 87 13 92 20 8 rural 58 0 54 9 67 19 84 25
9 rural 84 1 87 9 94 6 92 19 10 rural 98 3 92 0 88 3 80 7
11 urban 112 1 137 11 114 5 137 23 12 urban 104 1 176 8 97 6 146 18
13 urban 195 3 171 5 175 6 137 11 14 urban 128 1 113 10 125 3 126 9
15 urban 96 2 152 10 145 5 153 10 16 urban 98 0 170 9 129 3 148 16
17 urban 119 2 121 8 128 6 147 18 18 urban 120 0 130 0 114 2 108 13
;

data longform;
set hotels;
array a[4] ADR1 ADR2 ADR3 ADR4;
array o[4] OCR1 OCR2 OCR3 OCR4;
do season=1 to 4;
  ADR=a[season];
  OCR=o[season];
  output;
end;
keep hotel region season ADR OCR;
run;

/*fitting GEE Poisson model with unstructured working correlation matrix*/
proc genmod;
class hotel region;
model OCR = region ADR season / dist=poisson link=log;
repeated subject = hotel / type=un;
run;
```

#### QIC -231.0202

```
/*fitting GEE Poisson model with Toeplitz working correlation matrix*/
proc genmod;
class hotel region;
model OCR = region ADR season / dist=poisson link=log;
repeated subject = hotel / type=mdep(3);
run;
```

#### QIC -227.9801

```
/*fitting GEE Poisson model with autoregressive working correlation matrix*/
proc genmod;
class hotel region;
model OCR = region ADR season / dist=poisson link=log;
repeated subject = hotel / type=ar;
run;
```

#### QIC -227.1029

```
/*fitting GEE Poisson model with compound symmetric working correlation matrix*/
proc genmod;
```

```

class hotel region;
model OCR = region ADR season / dist=poisson link=log;
repeated subject = hotel / type=cs;
run;

```

**QIC -224.0231**

```

/*fitting GEE Poisson model with independent working correlation matrix*/
proc genmod;
class hotel region;
model OCR = region ADR season / dist=poisson link=log;
repeated subject = hotel / type=ind;
run;

```

**QIC -223.9375**

**In R:**

```

#creating longform dataset
library(reshape2)
data1<-
melt(hotels.data[,c("hotel","region","ADR1","ADR2","ADR3","ADR4")],id.vars=c("hotel",
"region"), variable.name="ADrn",value.name="ADR")
data2<- melt(hotels.data[,c("OCR1","OCR2","OCR3","OCR4")], variable.name="OCrn",
value.name="OCR")
longform.data<- cbind(data1,data2)

#sorting data by id
longform.data<- longform.data[order(longform.data$hotel),]

#creating variable for time
longform.data$season<- ifelse(longform.data$ADrn=="ADR1",1,
ifelse(longform.data$ADrn=="ADR2",2,ifelse(longform.data$ADrn=="ADR3",3,4)))

#specifying reference level
longform.data$region.rel<- relevel(as.factor(longform.data$region), ref="urban")

library(geepack)
library(MuMIn)

#fitting GEE Poisson model with unstructured working correlation matrix
summary(un.fitted.model<- geeglm(OCR ~ region.rel + ADR + season,
data=longform.data, id=hotel,family=poisson(link="log"), corstr="unstructured"))
QIC(un.fitted.model)

```

**QIC  
-1811**

```

#fitting GEE Poisson model with autoregressive working correlation matrix
summary(ar.fitted.model<- geeglm(OCR ~ region.rel + ADR + season,
data=longform.data, id=hotel, family=poisson(link="log"), corstr="ar1"))
QIC(ar.fitted.model)

```

**QIC  
-1812**

```

#fitting GEE negative Poisson with compound symmetric working correlation matrix
summary(cs.fitted.model<- geeglm(OCR ~ region.rel + ADR + season,
data=longform.data, id=hotel, family=poisson(link="log"),corstr="exchangeable"))
QIC(cs.fitted.model)

```

QIC  
-1812

```
#fitting GEE negative Poisson with independent working correlation matrix
summary(ind.fitted.model<- geeglm(OCR ~ region.rel + ADR + season,
data=longform.data,id=hotel, family=poisson(link="log"), corstr="independence"))
QIC(ind.fitted.model)
```

QIC  
-1812

(b) Choose the QIC-optimal model.

For the GEE models fitted in SAS, the one with the unstructured working correlation fit has the best fit as judged by the QIC criterion.

	UN	Toeplitz	AR	CS	Ind
QIC	-231.0202	-227.9801	-227.1029	-224.0231	-223.9375

In R, the models with autoregressive, compound symmetric, and independent working correlation matrices have the smallest QIC. We choose the one with the independent structure since it is the simplest model.

(c) Answer parts (b)-(d) in Exercise 9.4 for the optimal model.

In SAS:

```
/*fitting GEE Poisson model with unstructured working correlation matrix*/
proc genmod;
  class hotel region;
  model OCR = region ADR season / dist=poisson link=log;
  repeated subject = hotel / type=un corrw;
run;
```

```
Working Correlation Matrix
      col1   col2   col3   col4
Row1  1.0000 -0.0519 -0.0137  0.0070
Row2 -0.0519  1.0000 -0.2281  0.1984
Row3 -0.0137 -0.2281  1.0000 -0.0944
Row4  0.0070  0.1984 -0.0944  1.0000
```

```
Analysis of GEE Parameter Estimates
Empirical Standard Error Estimates
Parameter      Estimate Standard 95% Confidence Limits      Z Pr > |Z|
Error
Intercept      0.1871  0.4643    -0.7229    1.0971  0.40  0.6869
region  rural   0.5636  0.2679     0.0384    1.0888  2.10  0.0354
region  urban   0.0000  0.0000     0.0000    0.0000  .    .
ADR          0.0015  0.0032    -0.0047    0.0078  0.48  0.6288
season       0.5399  0.0391     0.4632    0.6165 13.80 <.0001
```

The fitted GEE Poisson model has the estimated average number of days the hotel occupancy rate was below 65%  $\hat{\lambda} = \exp(0.1871 + 0.5636 \cdot \text{rural} + 0.0015 \cdot \text{ADR} + 0.5399 \cdot \text{season})$ . Region and season are significant predictors. The estimated working correlation matrix is

$$\begin{pmatrix} 1.0000 & -0.0519 & -0.0137 & 0.0070 \\ -0.0519 & 1.0000 & -0.2281 & 0.1984 \\ -0.0137 & -0.2281 & 1.0000 & -0.0944 \\ 0.0070 & 0.1984 & -0.0944 & 1.0000 \end{pmatrix}$$
 For rural hotels, the estimated average number of days that the hotel occupancy rate is below 65% is  $\exp(0.5636) \cdot 100\% = 175.6986\%$  of that for urban hotels. Every season (summer to fall to winter to spring), this estimated average increases by  $(\exp(0.5399) - 1) \cdot 100\% = 71.58353\%$ .

To predict the number of days with an occupancy rate below 65% for the winter season in a rural hotel with an average daily rate of \$75 we compute  $OCR^0 = \exp(0.1871 + 0.5636 + 0.0015 \cdot 75 + 0.5399 \cdot 3) = 11.9759$ .

In SAS:

```
data predict;
input hotel region$ ADR season;
cards;
19 rural 75 3
;

data longform;
set longform predict;
run;

proc genmod;
class hotel region;
model OCR = region ADR season / dist=poisson link=log;
repeated subject = hotel / type=un corrw;
output out=outdata p=pOCR;
run;

proc print data=outdata (firstobs=73) noobs;
var pOCR;
run;
```

pOCR  
12.0168

In R:

```
#fitting GEE Poisson model with independent working correlation matrix
summary(ind.fitted.model<- geeglm(OCR ~ region.rel + ADR + season,
data=longform.data, id=hotel,family=poisson(link="log"), corstr="independence"))
```

Coefficients:

	Estimate	Std.err	wald	Pr(> w )
(Intercept)	-0.07378	0.50685	0.02	0.884
region.relrural	0.58644	0.24862	5.56	0.018
ADR	0.00314	0.00351	0.80	0.371
season	0.56313	0.03776	222.40	<2e-16

The fitted model has the estimated average number of days the hotel occupancy rate was below 65%  $\hat{\lambda} = \exp(-0.07378 + 0.58644 \cdot \text{rural} + 0.00314 \cdot \text{ADR} + 0.56313 \cdot \text{season})$ . Region and season are significant predictors. The estimated working correlation matrix is

$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$ . For rural hotels, the estimated average number of days that the hotel occupancy rate

is below 65% is  $\exp(0.58644) \cdot 100\% = 179.7578\%$  of that for urban hotels. Every season (summer to fall to winter to spring), this estimated average increases by  $(\exp(0.56313) - 1) \cdot 100\% = 75.61607\%$ .

To predict the number of days with an occupancy rate below 65% for the winter season in a rural hotel with an average daily rate of \$75 we compute  $OCR^0 = \exp(-0.07378 + 0.58644 + 0.00314 \cdot 75 + 0.56313 \cdot 3) = 11.44497$ .

In R:

```
print(predict(ind.fitted.model, data.frame(hotel=19, region.rel="rural", ADR=75,
season=3), type="response"))
```

11.4

# CHAPTER 10

**EXERCISE 10.1.** For the hierarchical model with normal response defined in (10.1), show that

(a) Observations within each individual  $i$  in cluster  $m$  for different times  $j$  and  $j'$  have covariance

$$\begin{aligned} \text{Cov}(y_{ijm}, y_{ij'm}) &= \text{Cov}(\beta_0 + \beta_1 x_{1ijm} + \cdots + \beta_k x_{kijm} + \beta_{k+1} t_j + u_{1im} + u_{2im} t_j + \tau_{1m} + \tau_{2m} t_j + \varepsilon_{ijm}, \\ &\quad \beta_0 + \beta_1 x_{1ij'm} + \cdots + \beta_k x_{kij'm} + \beta_{k+1} t_{j'} + u_{1im} + u_{2im} t_{j'} + \tau_{1m} + \tau_{2m} t_{j'} + \varepsilon_{ij'm}) = \\ &= \text{Cov}(u_{1im} + u_{2im} t_j + \tau_{1m} + \tau_{2m} t_j + \varepsilon_{ijm}, u_{1im} + u_{2im} t_{j'} + \tau_{1m} + \tau_{2m} t_{j'} + \varepsilon_{ij'm}) = \\ &= \text{Cov}(u_{1im}, u_{1im}) + \text{Cov}(u_{1im}, u_{2im}) t_{j'} + \text{Cov}(u_{1im}, \tau_{1m}) + \text{Cov}(u_{1im}, \tau_{2m}) t_{j'} + \\ &\quad \text{Cov}(u_{1im}, \varepsilon_{ij'm}) + \text{Cov}(u_{2im}, u_{1im}) t_j + \text{Cov}(u_{2im}, u_{2im}) t_j t_{j'} + \text{Cov}(u_{2im}, \tau_{1m}) t_j + \\ &\quad \text{Cov}(u_{2im}, \tau_{2m}) t_j t_{j'} + \text{Cov}(u_{2im}, \varepsilon_{ij'm}) t_j + \text{Cov}(\tau_{1m}, u_{1im}) + \text{Cov}(\tau_{1m}, u_{2im}) t_{j'} + \\ &\quad \text{Cov}(\tau_{1m}, \tau_{1m}) + \text{Cov}(\tau_{1m}, \tau_{2m}) t_{j'} + \text{Cov}(\tau_{1m}, \varepsilon_{ij'm}) + \text{Cov}(\tau_{2m}, u_{1im}) t_j + \\ &\quad \text{Cov}(\tau_{2m}, u_{2im}) t_j t_{j'} + \text{Cov}(\tau_{2m}, \tau_{1m}) t_j + \text{Cov}(\tau_{2m}, \tau_{2m}) t_j t_{j'} + \text{Cov}(\tau_{2m}, \varepsilon_{ij'm}) + \\ &\quad \text{Cov}(\varepsilon_{ijm}, u_{1im}) + \text{Cov}(\varepsilon_{ijm}, u_{2im}) t_{j'} + \text{Cov}(\varepsilon_{ijm}, \tau_{1m}) + \text{Cov}(\varepsilon_{ijm}, \tau_{2m}) t_{j'} + \\ &\quad \text{Cov}(\varepsilon_{ijm}, \varepsilon_{ij'm}) = \text{Var}(u_{1im}) + \text{Cov}(u_{1im}, u_{2im}) t_{j'} + \text{Cov}(u_{2im}, u_{1im}) t_j + \text{Var}(u_{2im}) t_j t_{j'} + \\ &\quad \text{Var}(\tau_{1m}) + \text{Cov}(\tau_{1m}, \tau_{2m}) t_{j'} + \text{Cov}(\tau_{2m}, \tau_{1m}) t_j + \text{Var}(\tau_{2m}) t_j t_{j'} = \sigma_{u_1}^2 + \sigma_{\tau_1}^2 + (\sigma_{u_1 u_2} + \sigma_{\tau_1 \tau_2})(t_j + t_{j'}) + (\sigma_{u_2}^2 + \sigma_{\tau_2}^2) t_j t_{j'}. \end{aligned}$$

(b) Observations for two individuals  $i$  and  $i'$  within the same cluster  $m$  at any two times  $t_j$  and  $t_{j'}$ , equal or not, have covariance  $\text{Cov}(y_{ijm}, y_{i'jm}) = \text{Cov}(\beta_0 + \beta_1 x_{1ijm} + \cdots + \beta_k x_{kijm} + \beta_{k+1} t_j + u_{1im} + u_{2im} t_j + \tau_{1m} + \tau_{2m} t_j + \varepsilon_{ijm}, \beta_0 + \beta_1 x_{1i'jm} + \cdots + \beta_k x_{ki'jm} + \beta_{k+1} t_{j'} + u_{1im} + u_{2im} t_{j'} + \tau_{1m} + \tau_{2m} t_{j'} + \varepsilon_{i'jm}) = \text{Cov}(u_{1im} + u_{2im} t_j + \tau_{1m} + \tau_{2m} t_j + \varepsilon_{ijm}, u_{1im} + u_{2im} t_{j'} + \tau_{1m} + \tau_{2m} t_{j'} + \varepsilon_{i'jm}) = \text{Cov}(u_{1im}, u_{1im}) + \text{Cov}(u_{1im}, u_{2im}) t_{j'} + \text{Cov}(u_{1im}, \tau_{1m}) + \text{Cov}(u_{1im}, \tau_{2m}) t_{j'} + \text{Cov}(u_{1im}, \varepsilon_{i'jm}) + \text{Cov}(u_{2im}, u_{1im}) t_j + \text{Cov}(u_{2im}, u_{2im}) t_j t_{j'} + \text{Cov}(u_{2im}, \tau_{1m}) t_j + \text{Cov}(u_{2im}, \tau_{2m}) t_j t_{j'} + \text{Cov}(u_{2im}, \varepsilon_{i'jm}) t_j + \text{Cov}(\tau_{1m}, u_{1im}) + \text{Cov}(\tau_{1m}, u_{2im}) t_{j'} + \text{Cov}(\tau_{1m}, \tau_{1m}) + \text{Cov}(\tau_{1m}, \tau_{2m}) t_{j'} + \text{Cov}(\tau_{1m}, \varepsilon_{i'jm}) + \text{Cov}(\tau_{2m}, u_{1im}) t_j + \text{Cov}(\tau_{2m}, u_{2im}) t_j t_{j'} + \text{Cov}(\tau_{2m}, \tau_{1m}) t_j + \text{Cov}(\tau_{2m}, \tau_{2m}) t_j t_{j'} + \text{Cov}(\tau_{2m}, \varepsilon_{i'jm}) t_j + \text{Cov}(\varepsilon_{ijm}, u_{1im}) + \text{Cov}(\varepsilon_{ijm}, u_{2im}) t_{j'} + \text{Cov}(\varepsilon_{ijm}, \tau_{1m}) + \text{Cov}(\varepsilon_{ijm}, \tau_{2m}) t_{j'} + \text{Cov}(\varepsilon_{ijm}, \varepsilon_{i'jm}) = \text{Var}(\tau_{1m}) + \text{Cov}(\tau_{1m}, \tau_{2m}) t_{j'} + \text{Cov}(\tau_{2m}, \tau_{1m}) t_j + \text{Var}(\tau_{2m}) t_j t_{j'} = \sigma_{\tau_1}^2 + \sigma_{\tau_1 \tau_2} (t_j + t_{j'}) + \sigma_{\tau_2}^2 t_j t_{j'}$ .

(c) Observations for two individuals in different clusters are not correlated, that is, for  $i \neq i'$  and  $m \neq m'$ ,  $\text{Cov}(y_{ijm}, y_{i'jm'}) = \text{Cov}(\beta_0 + \beta_1 x_{1ijm} + \cdots + \beta_k x_{kijm} + \beta_{k+1} t_j + u_{1im} + u_{2im} t_j + \tau_{1m} + \tau_{2m} t_j + \varepsilon_{ijm}, \beta_0 + \beta_1 x_{1i'jm'} + \cdots + \beta_k x_{ki'jm'} + \beta_{k+1} t_{j'} + u_{1im'} + u_{2im'} t_{j'} + \tau_{1m'} + \tau_{2m'} t_{j'} + \varepsilon_{i'jm'}) = \text{Cov}(u_{1im} + u_{2im} t_j + \tau_{1m} + \tau_{2m} t_j + \varepsilon_{ijm}, u_{1im'} + u_{2im'} t_{j'} + \tau_{1m'} + \tau_{2m'} t_{j'} + \varepsilon_{i'jm'}) = \text{Cov}(u_{1im}, u_{1im'}) + \text{Cov}(u_{1im}, u_{2im'}) t_{j'} + \text{Cov}(u_{1im}, \tau_{1m'}) + \text{Cov}(u_{1im}, \tau_{2m'}) t_{j'} + \text{Cov}(u_{1im}, \varepsilon_{i'jm'}) + \text{Cov}(u_{2im}, u_{1im'}) t_j + \text{Cov}(u_{2im}, u_{2im'}) t_j t_{j'} + \text{Cov}(u_{2im}, \tau_{1m'}) t_j + \text{Cov}(u_{2im}, \tau_{2m'}) t_j t_{j'} + \text{Cov}(u_{2im}, \varepsilon_{i'jm'}) t_j + \text{Cov}(\tau_{1m}, u_{1im'}) + \text{Cov}(\tau_{1m}, u_{2im'}) t_{j'} + \text{Cov}(\tau_{1m}, \tau_{1m'}) + \text{Cov}(\tau_{1m}, \tau_{2m'}) t_{j'} + \text{Cov}(\tau_{1m}, \varepsilon_{i'jm'}) + \text{Cov}(\tau_{2m}, u_{1im'}) t_j + \text{Cov}(\tau_{2m}, u_{2im'}) t_j t_{j'} + \text{Cov}(\tau_{2m}, \tau_{1m'}) t_j + \text{Cov}(\tau_{2m}, \tau_{2m'}) t_j t_{j'} + \text{Cov}(\tau_{2m}, \varepsilon_{i'jm'}) t_j + \text{Cov}(\varepsilon_{ijm}, u_{1im'}) + \text{Cov}(\varepsilon_{ijm}, u_{2im'}) t_{j'} + \text{Cov}(\varepsilon_{ijm}, \tau_{1m'}) + \text{Cov}(\varepsilon_{ijm}, \tau_{2m'}) t_{j'} + \text{Cov}(\varepsilon_{ijm}, \varepsilon_{i'jm'}) = 0$ .

$$\begin{aligned} &Cov(\tau_{2m}, \varepsilon_{ijm'})t_j + Cov(\varepsilon_{ijm}, u_{1i'm'}) + Cov(\varepsilon_{ijm}, u_{2i'm'})t_{j'} + Cov(\varepsilon_{ijm}, \tau_{1m'}) + \\ &Cov(\varepsilon_{ijm}, \tau_{2m'})t_{j'} + Cov(\varepsilon_{ijm}, \varepsilon_{i'j'm'}) = 0. \end{aligned}$$

(d) The response variable  $y_{ijm} = \beta_0 + \beta_1 x_{1ijm} + \cdots + \beta_k x_{kijm} + \beta_{k+1} t_j + u_{1im} + u_{2im} t_j + \tau_{1m} + \tau_{2m} t_j + \varepsilon_{ijm}$  has a normal distribution being a linear combination of independent normally distributed random variables, has mean  $E(y_{ijm}) = E(\beta_0 + \beta_1 x_{1ijm} + \cdots + \beta_k x_{kijm} + \beta_{k+1} t_j + u_{1im} + u_{2im} t_j + \tau_{1m} + \tau_{2m} t_j + \varepsilon_{ijm}) = \beta_0 + \beta_1 x_{1ijm} + \cdots + \beta_k x_{kijm} + \beta_{k+1} t_j + E(u_{1im}) + E(u_{2im})t_j + E(\tau_{1m}) + E(\tau_{2m})t_j + E(\varepsilon_{ijm}) = \beta_0 + \beta_1 x_{1ijm} + \cdots + \beta_k x_{kijm} + \beta_{k+1} t_j$ , and variance  $Var(y_{ijm}) = Var(\beta_0 + \beta_1 x_{1ijm} + \cdots + \beta_k x_{kijm} + \beta_{k+1} t_j + u_{1im} + u_{2im} t_j + \tau_{1m} + \tau_{2m} t_j + \varepsilon_{ijm}) = Var(u_{1im} + u_{2im} t_j + \tau_{1m} + \tau_{2m} t_j + \varepsilon_{ijm}) = Cov(u_{1im} + u_{2im} t_j + \tau_{1m} + \tau_{2m} t_j + \varepsilon_{ijm}, u_{1im} + u_{2im} t_j + \tau_{1m} + \tau_{2m} t_j + \varepsilon_{ijm}) = Cov(u_{1im}, u_{1im}) + Cov(u_{1im}, u_{2im})t_j + Cov(u_{1im}, \tau_{1m}) + Cov(u_{1im}, \tau_{2m})t_j + Cov(u_{1im}, \varepsilon_{ijm}) + Cov(u_{2im}, u_{1im})t_j + Cov(u_{2im}, u_{2im})t_j^2 + Cov(u_{2im}, \tau_{1m})t_j + Cov(u_{2im}, \tau_{2m})t_j^2 + Cov(u_{2im}, \varepsilon_{ijm})t_j + Cov(\tau_{1m}, u_{1im}) + Cov(\tau_{1m}, u_{2im})t_j + Cov(\tau_{1m}, \tau_{1m}) + Cov(\tau_{1m}, \tau_{2m})t_j + Cov(\tau_{1m}, \varepsilon_{ijm}) + Cov(\tau_{2m}, u_{1im})t_j + Cov(\tau_{2m}, u_{2im})t_j^2 + Cov(\tau_{2m}, \tau_{1m})t_j + Cov(\tau_{2m}, \tau_{2m})t_j^2 + Cov(\tau_{2m}, \varepsilon_{ijm})t_j + Cov(\varepsilon_{ijm}, u_{1im}) + Cov(\varepsilon_{ijm}, u_{2im})t_j + Cov(\varepsilon_{ijm}, \tau_{1m}) + Cov(\varepsilon_{ijm}, \tau_{2m})t_j + Cov(\varepsilon_{ijm}, \varepsilon_{ijm}) = Var(u_{1im}) + Cov(u_{1im}, u_{2im})t_j + Cov(u_{2im}, u_{1im})t_j + Var(u_{2im})t_j^2 + Var(\tau_{1m}) + Cov(\tau_{1m}, \tau_{2m})t_j + Cov(\tau_{2m}, \tau_{1m})t_j + Var(\tau_{2m})t_j^2 + Var(\varepsilon_{ijm}) = \sigma_{u_1}^2 + \sigma_{\tau_1}^2 + 2(\sigma_{u_1 u_2} + \sigma_{\tau_1 \tau_2})t_j + (\sigma_{u_2}^2 + \sigma_{\tau_2}^2)t_j^2 + \sigma^2$ .

**EXERCISE 10.2.** (a) Plot a histogram for test scores and conduct normality testing. Verify that the underlying distribution may be modeled as normal.

In SAS:

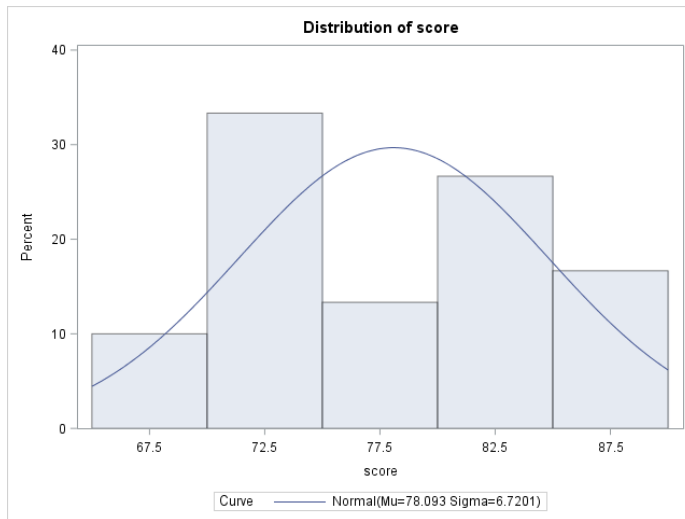
```
data schools;
input school API subject$ classsize year score @@;
cards;
1 911 ELA      20 15 78.39  1 912 ELA      22 16 79.85
1 917 ELA      23 17 81.34  1 917 ELA      22 18 82.56
1 919 ELA      24 19 83.12  1 911 Math     21 15 83.77
1 912 Math     22 16 84.90  1 917 Math     24 17 86.12
1 917 Math     23 18 88.99  1 919 Math     23 19 88.40
1 911 Science  21 15 80.19  1 912 Science  22 16 83.15
1 917 Science  24 17 84.45  1 917 Science  23 18 86.66
1 919 Science  23 19 88.43  2 732 ELA      34 15 68.03
2 745 ELA      36 16 70.67  2 751 ELA      36 17 74.17
2 753 ELA      37 18 72.78  2 753 ELA      38 19 73.18
2 732 Math     34 15 67.88  2 745 Math     34 16 68.34
2 751 Math     35 17 70.30  2 753 Math     37 18 71.22
2 753 Math     36 19 72.12  2 732 Science  34 15 72.96
2 745 Science  34 16 73.65  2 751 Science  36 17 74.58
2 753 Science  35 18 76.36  2 753 Science  35 19 76.23
;

/*plotting histogram*/
proc univariate;
var score;
```

```

  histogram/normal;
run;

```



### Goodness-of-Fit Tests for Normal Distribution

Test	Statistic	p value
Kolmogorov-Smirnov D	0.13276200	Pr > D > 0.150
Cramer-von Mises	W-Sq 0.10207261	Pr > W-Sq 0.101
Anderson-Darling	A-Sq 0.60597790	Pr > A-Sq 0.106

The underlying distribution is normal as confirmed by the histogram and the large p-values in the normality tests.

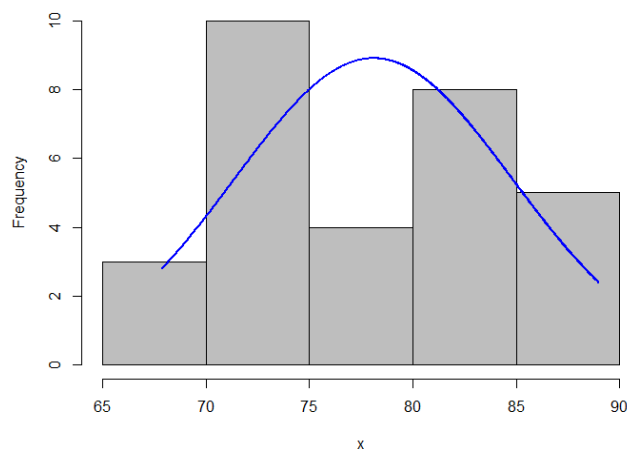
In R:

```

schools.data<- read.csv(file="C://./Exercisel0.2Data.csv", header=TRUE, sep=",")

#plotting histogram
library(rcompanion)
plotNormalHistogram(schools.data$score)

```



```
shapiro.test(schools.data$score)
```

Shapiro-wilk normality test

W = 0.93668, p-value = 0.07407

(b) Run the hierarchical model with random slopes and intercepts at the school and subject-within-school levels. If there is a problem with convergence, gradually remove the random slopes and simplify the model to random intercepts only, if necessary. Discuss the overall model fit.

In SAS:

```
/*fitting hierarchical normal model with random slopes and intercepts*/
proc mixed covtest;
  class subject school;
  model score = API classsize year / solution;
  random intercept year / subject=school type=un;
  random intercept year / subject=subject(school) type=un;
run;
```

The model doesn't converge.

```
/*fitting hierarchical normal model with random slope and intercept at
level 2, and intercept only at level 1*/
proc mixed covtest;
  class subject school;
  model score = API classsize year / solution;
  random intercept year / subject=school type=un;
  random intercept / subject=subject(school) type=un;
run;
```

The model doesn't converge.

```
/*fitting hierarchical normal model with random slope and intercept at
level 1, and intercept only at level 2*/
proc mixed covtest;
  class subject school;
  model score = API classsize year / solution;
  random intercept / subject=school type=un;
  random intercept year / subject=subject(school) type=un;
run;
```

Covariance Parameter Estimates					
Cov	Parm	Subject	Estimate	Standard Error	Z Value Pr >  Z
UN(1,1)	school		0	.	. .
UN(1,1)	subject(school)		43.5813	47.0828	0.93 0.1773
UN(2,1)	subject(school)		-2.1932	2.3457	-0.93 0.3498
UN(2,2)	subject(school)		0.1296	0.1267	1.02 0.1531
Residual			0.6425	0.2203	2.92 0.0018

In this model, the parameters of the random-effects terms are either non-estimable or non-significant.

```
/*fitting hierarchical normal model with random intercepts only*/
proc mixed covtest;
  class subject school;
  model score = API classsize year / solution;
  random intercept / subject=school type=un;
```

```
random intercept / subject=subject(school) type=un;
run;
```

Covariance Parameter Estimates				
Cov	Parm Subject	Estimate	Standard Error	Z Value Pr > Z
UN(1,1)	school	8.5624	92.3691	0.09 0.4631
UN(1,1)	subject(school)	6.4797	4.7123	1.38 0.0846
Residual		0.9194	0.2837	3.24 0.0006

Null Model Likelihood Ratio Test		
DF	Chi-Square	Pr > ChiSq
2	34.04	<.0001

Solution for Fixed Effects					
Effect	Estimate	Standard Error	DF	t Value	Pr >  t
Intercept	18.6896	22.7718	1	0.82	0.5625
API	0.05080	0.02668	21	1.90	0.0707
classsize	-0.1082	0.2235	21	-0.48	0.6334
year	1.1952	0.2322	21	5.15	<.0001

The model fits the data well as indicated by a tiny p-value in the deviance test (null model likelihood ratio test).

In R:

```
#fitting hierarchical normal model with random slopes and intercepts
library(lme4)
summary(fitted.model<- lmer(score ~ API + classsize + year + (1 + year | school)
+ (1 + year |school:subject),data=schools.data))
```

Random effects:

Groups	Name	Variance	Std.Dev.	Corr
school:subject	(Intercept)	18.14986	4.2603	
	year	0.04059	0.2015	-0.80
school	(Intercept)	122.41231	11.0640	
	year	0.19597	0.4427	-1.00
Residual		0.63146	0.7946	

Fixed effects:

	Estimate	Std. Error	t value
(Intercept)	-12.99487	21.09944	-0.616
API	0.09119	0.02231	4.087
classsize	-0.07799	0.18800	-0.415
year	1.03320	0.38184	2.706

```
#checking model fit
null.model<- glm(score ~ API + classsize + year, data=schools.data)
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))
```

29.55178

```
print(p.value<- pchisq(deviance, df=6, lower.tail = FALSE))
```

4.782027e-05

(c) Write down the fitted model. Include all estimated parameters. Use  $\alpha = 0.10$  to conclude about significant parameters of the random effects. Are the scores for each subject correlated? Are the scores for different subjects within the same school correlated?

In SAS, the fitted hierarchical normal model has the estimated parameters  $\hat{E}(\text{score}) = 18.6896 + 0.0508 \cdot \text{API} - 0.1082 \cdot \text{class size} + 1.1952 \cdot \text{year}$ ,  $\hat{\sigma}_{u_1}^2 = 6.4797$ ,  $\hat{\sigma}_{\tau_1}^2 = 8.5624$ , and  $\hat{\sigma}^2 = 0.9194$ . At  $\alpha = 0.10$ , the variance of the random intercept at the subject-within-school level is significant, thus scores within each subject are correlated (see Exercise 10.1(a)). However, since  $\sigma_{\tau_1}^2$  is not significant, and scores for different subjects within the same school are not correlated (see Exercise 10.1(b)).

In R, the fitted hierarchical normal model has the estimated parameters  $\hat{E}(\text{score}) = -12.99487 + 0.09119 \cdot \text{API} - 0.07799 \cdot \text{class size} + 1.0332 \cdot \text{year}$ ,  $\hat{\sigma}_{u_1}^2 = 18.14986$ ,  $\hat{\sigma}_{u_1 u_2} = -0.8$ ,  $\hat{\sigma}_{u_2}^2 = 0.04059$ ,  $\hat{\sigma}_{\tau_1}^2 = 122.41231$ ,  $\hat{\sigma}_{\tau_1 \tau_2} = -1$ ,  $\hat{\sigma}_{\tau_2}^2 = 0.19597$  and  $\hat{\sigma}^2 = 0.63146$ . The variances for both intercepts seem to be significantly different from zero (the ratios of estimate/stdev are large), thus in the model, scores for each subject as well as scores between different subjects within the same school are correlated.

(d) Give interpretation for all estimated significant fixed-effects coefficients. Use  $\alpha = 0.10$ .

API and year are significant predictors of the mean score. For the model fitted in SAS, the API increases by one, and the estimated average score increases by 0.0508. Each year, the estimated average score increases by 1.1952. For the model fitted in R, the API increases by one, and the estimated average score increases by 0.09119. Each year, the estimated average score increases by 1.0332.

(e) Use the fitted model to predict an average score on a math test for a class of 36 students in 2019 in a school with an API of 753.

For the model fitted in SAS, the predicted score is  $\text{score}^0 = 18.6896 + 0.0508 \cdot 753 - 0.1082 \cdot 36 + 1.1952 \cdot 19 = 75.7556$ .

In SAS:

```
/*using fitted model for prediction*/
data predict;
input school API subject$ classsize year;
cards;
2 753 Math 36 19
;

data schools;
set schools predict;
run;

proc mixed;
class subject school;
model score = API classsize year / outpm=outdata;
random intercept / subject=school type=un;
random intercept / subject=subject(school) type=un;
run;

proc print data=outdata (firstobs=31) noobs;
var Pred;
```

```
run;
  Pred
75.7565
```

For the model fitted in R, the predicted score is  $score^0 = -12.99487 + 0.09119 \cdot 753 - 0.07799 \cdot 36 + 1.0332 \cdot 19 = 72.49436$ .

In R:

```
#using fitted model for prediction
print(predict(fitted.model, data.frame(school=2, API=753, subject="Math",
classsize=36, year=19), allow.new.levels=TRUE, re.form=NA))
```

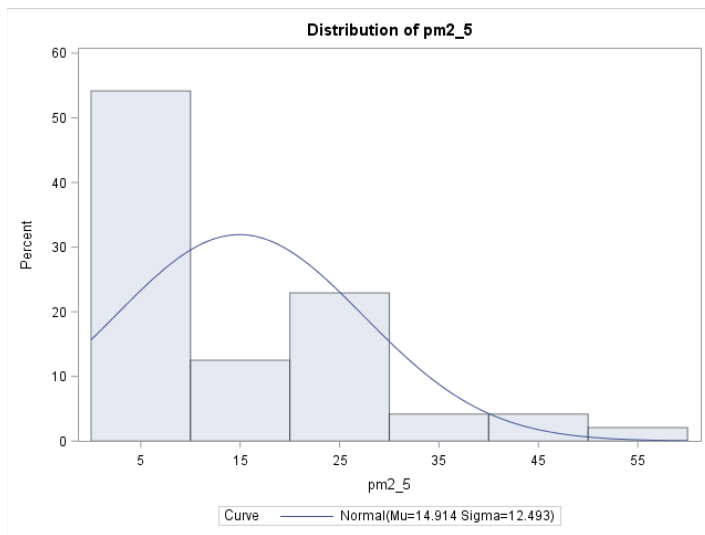
```
72.49533
```

**EXERCISE 10.3.** (a) Plot a histogram of the particulate matter (PM2.5). Describe its shape. Argue that a gamma distribution is appropriate.

In SAS:

```
data pollution;
input state$ county$ township popl pest$ pm2_5 @@;
cards;
S1 A 1 4.1 no 22.97 S1 A 2 22.0 no 23.05 S1 A 3 6.3 no 24.97
S1 A 4 3.2 no 23.77 S1 A 5 13.4 no 23.09 S1 A 6 3.9 yes 24.75
S1 A 7 3.8 yes 36.93 S1 A 8 25.6 yes 45.83 S1 B 1 12.7 no 13.19
S1 B 2 17.8 no 22.9 S1 B 3 23.7 no 31.45 S1 B 4 11.8 yes 25.40
S1 B 5 12.9 yes 44.15 S1 B 6 13.0 yes 25.16 S1 B 7 12.0 yes 54.36
S1 B 8 13.0 no 24.38 S2 C 1 9.9 no 7.25 S2 C 2 5.6 yes 28.46
S2 C 3 3.9 no 7.06 S2 C 4 7.3 no 9.33 S2 C 5 4.7 no 5.59
S2 C 6 8.9 yes 9.94 S2 C 7 6.7 yes 8.49 S2 C 8 6.5 yes 6.97
S2 D 1 6.6 no 9.13 S2 D 2 7.2 no 11.04 S2 D 3 8.3 no 8.98
S2 D 4 5.2 yes 5.75 S2 D 5 9.1 yes 11.28 S2 D 6 4.3 no 6.88
S2 D 7 6.9 yes 9.21 S2 D 8 8.5 yes 11.23 S3 E 1 6.1 no 5.44
S3 E 2 3.9 no 4.33 S3 E 3 3.5 no 5.04 S3 E 4 2.4 no 3.31
S3 E 5 4.3 no 5.24 S3 E 6 2.8 yes 14.34 S3 E 7 3.4 no 4.90
S3 E 8 3.6 no 3.59 S3 F 1 5.3 no 5.01 S3 F 2 4.5 no 5.73
S3 F 3 2.5 no 4.28 S3 F 4 3.1 yes 15.42 S3 F 5 3.5 no 3.59
S3 F 6 5.7 no 4.69 S3 F 7 7.1 no 4.06 S3 F 8 4.6 no 3.98
;

/*plotting histogram*/
proc univariate;
  var pm2_5;
  histogram/normal;
run;
```



### Goodness-of-Fit Tests for Normal Distribution

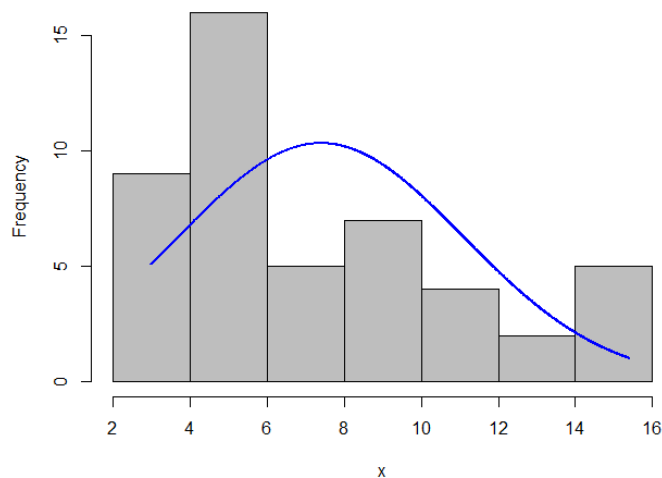
Test	Statistic	p value
Kolmogorov-Smirnov D	0.21861080	Pr > D <0.010
Cramer-von Mises	W-Sq 0.52542458	Pr > W-Sq <0.005
Anderson-Darling	A-Sq 2.98507256	Pr > A-Sq <0.005

The histogram shows a right-skewed distribution and the normality tests all conclude that the distribution is not normal. Therefore, a gamma distribution would be a better choice.

In R:

```
pollution.data<- read.csv(file="C:/./Exercise10.3Data.csv", header=TRUE, sep=",")

#plotting histogram
library(rcompanion)
plotNormalHistogram(pollution.data$pm2_5)
```



```
shapiro.test(pollution.data$pm2_5)
```

shapiro-wilk normality test

w = 0.89215, p-value = 0.0003543

(b) Run the multilevel regression model for PM2.5, based on the gamma distribution. How well does the model fit the data? Hint: Townships variable indexes repeated measures within each county.

In SAS:

```
/*fitting hierarchical gamma model with random slopes and intercepts*/
proc glimmix method=Laplace;
  class state county pest;
  model pm2_5 = pop1 pest township / solution dist=gamma link=log;
  random intercept township / subject=state type=un;
  random intercept township / subject=county(state) type=un;
  covtest/wald;
run;
```

The model converges, but the estimates of the parameters of the random-effects terms are not statistically distinguishable from zero. The model that has significant parameters is given below.

```
/*fitting hierarchical gamma model with random intercept only for county level*/
proc glimmix method=Laplace;
  class state county pest(ref="no");
  model pm2_5 = pop1 pest township / solution dist=gamma link=log;
  random intercept / subject=county(state) type=un;
  covtest/wald;
run;
```

-2 Log Likelihood 273.79

Covariance Parameter Estimates					
Cov	Parm Subject	Estimate	Standard Error	Z Value	Pr >  Z
UN(1,1)	county(state)	0.3178	0.1955	1.63	0.0521
Residual		0.09517	0.02048	4.65	<.0001

Solutions for Fixed Effects						
Effect	pest	Estimate	Standard Error	DF t Value	Pr >  t	
Intercept		2.2944	0.2657	5	8.64	0.0003
popl		0.01628	0.01134	39	1.44	0.1590
pest	yes	0.6889	0.1059	39	6.51	<.0001
pest	no	0	.	.	.	.
township		-0.05147	0.02173	39	-2.37	0.0229

```
/*checking model fit*/
proc glimmix method=Laplace;
  class state county pest;
  model pm2_5 = pop1 pest township / dist=gamma link=log;
run;
```

2 Log Likelihood 315.24

```
data deviance;
  deviance = 315.24 - 273.79;
  pvalue = 1 - probchi(deviance, 1);
run;
```

```
proc print noobs;
run;
```

```
deviance      pvalue
    41.45 1.2092E-10
```

The model has an excellent fit as judged by a negligibly small p-value in the deviance test.

In R:

```
#fitting hierarchical gamma model with random slopes and intercepts
library(lme4)
summary(glmer(pm2_5 ~ popl + pest + township + (1+township | state)
+ (1+township| state:county), data=pollution.data, family=Gamma("log")))
```

The model doesn't converge.

```
#fitting hierarchical gamma model with random intercept only for county level
summary(fitted.model<- glmer(pm2_5 ~ popl + pest + township
+ (1 | state:county), data=pollution.data, family=Gamma("log")))
```

Random effects:

Groups	Name	Variance	Std.Dev.
state:county	(Intercept)	0.1222	0.3496
Residual		0.1243	0.3526

Fixed effects:

	Estimate	Std. Error	t value	Pr(> z )
(Intercept)	2.28777	0.31448	7.275	3.47e-13
popl	0.01448	0.01078	1.344	0.1790
pestyes	0.68518	0.10150	6.751	1.47e-11
township	-0.05139	0.02083	-2.468	0.0136

```
#checking model fit
null.model<- glm(pm2_5 ~ popl + pest + township, data=pollution.data,
family=Gamma("log"))
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))
```

```
48.87942
```

```
print(p.value<- pchisq(deviance, df=1, lower.tail = FALSE))
```

```
2.72192e-12
```

(c) Write down the fitted model. Specify all estimates. Are PM2.5 readings correlated within each county? Between different counties within each state? Use  $\alpha = 0.10$ .

The fitted hierarchical gamma regression fitted in SAS has the estimated parameters  $\hat{E}(PM2.5) = \exp(2.2944 + 0.01628 \cdot \text{population size}/1000 + 0.6889 \cdot \text{pesticides used} - 0.05147 \cdot \text{township})$ ,  $\hat{\alpha} = 0.09517$ , and  $\hat{\sigma}_{county}^2 = 0.3178$ . Since this variance is statistically significant at the 10% level, it means that the responses are correlated for townships within the same county. There is no correlation between readings for different counties within the same state because the corresponding random effects are not present in the fitted model.

For the model fitted in R, the parameters have estimates  $\hat{E}(PM2.5) = \exp(2.28777 + 0.01448 \cdot \text{population size}/1000 + 0.68518 \cdot \text{pesticides used} - 0.05139 \cdot \text{township})$ ,  $\hat{\alpha} = 0.1243$ , and  $\hat{\sigma}_{\text{county}}^2 = 0.1222$ .

(d) What fixed-effects predictors are significant at the 5% level? Interpret them.

The use of pesticides and the township's number are significant predictors. In the townships where pesticides are used, the estimated average reading of PM2.5 is  $\exp(0.6889) \cdot 100\% = 199.1524\%$  (for the model fitted in R,  $\exp(0.68518) \cdot 100\% = 197.6999\%$ ) of that in townships where pesticides are not used. As the township's number increases by one, the estimated average PM2.5 reading changes by  $(\exp(-0.05147) - 1) \cdot 100\% = -5.01679\%$ , that is, decreases by 5.01679% (for the model fitted in R,  $(\exp(-0.0519) - 1) \cdot 100\% = -5.00919\%$ ).

(e) Use the fitted model to predict the level of particulate matter in a town with a population of 2,500 people if it is known that no pesticides are used in the fields that surround this town.

Using the model fitted in SAS, we compute the predicted value as follows (use township=1):

$$PM2.5^0 = \exp(2.2944 + 0.01628 \cdot 2.5 - 0.05147) = 9.812234.$$

In SAS:

```
/*using fitted model for prediction*/
data predict;
input state$ county$ township popl pest$;
cards;
S4 G 1 2.5 no
;

data pollution;
set pollution predict;
run;

proc glimmix method=Laplace;
class state county pest;
model pm2_5 = popl pest township / dist=gamma link=log;
random intercept /subject=county(state) type=un;
output out=outdata pred(ilink)=ppm2_5;
run;

proc print data=outdata (firstobs=49) noobs;
var ppm2_5;
run;

ppm2_5
9.81257
```

Using the model fitted in R, the predicted value is  $PM2.5^0 = \exp(2.28777 + 0.01448 \cdot 2.5 - 0.05139) = 9.704406$ .

In R:

```
#using fitted model for prediction
print(predict(fitted.model, data.frame(state="S4", county="G", township=1,
```

```
popl=2.5, pest="no"), allow.new.levels=TRUE, re.form=NA, type="response"))
```

9.704524

**EXERCISE 10.4.** (a) Run a three-level hierarchical model for the binary response variable. Write down the fitted model. Are the measurements correlated within each asset over time? Are the measurements for different assets within the same portfolio correlated? Use the 10% significance level. How good is the model fit?

In SAS:

```
data portfolios;
input portfolio asset type$ day1 day2 day3 day4 day5 @@;
cards;
1 1 stock 0 0 0 0 1 1 2 stock 0 1 1 0 0 1 3 bond 0 0 0 0 0
1 4 bond 1 0 0 0 0 1 5 stock 1 1 1 0 1 1 6 stock 1 0 1 1 1
1 7 stock 1 1 1 1 1 2 8 currency 0 1 1 1 1 2 9 stock 0 1 1 1 1
2 10 bond 0 1 0 0 1 2 11 stock 1 0 1 1 1 3 12 currency 1 0 1 0 1
3 13 stock 0 0 1 0 1 3 14 stock 0 0 1 1 1 4 15 stock 1 0 1 0 0
4 16 bond 1 1 1 1 0 4 17 currency 0 0 0 0 1 4 18 stock 1 1 1 1 1
4 19 currency 0 0 0 0 1 5 20 stock 0 0 1 1 1 5 21 currency 1 0 0 1 1
5 22 stock 0 0 1 1 1 5 23 bond 1 1 0 0 0 5 24 stock 1 1 1 1 1
6 25 bond 1 0 0 1 1 6 26 stock 1 1 1 1 1 6 27 stock 1 1 1 1 1
6 28 stock 1 1 1 1 1 7 29 currency 0 1 1 1 1 7 30 currency 0 0 1 1 1
7 31 bond 1 1 1 1 0 7 32 currency 1 0 1 1 1 7 33 bond 0 0 0 1 1
7 34 bond 1 0 1 0 1 7 35 stock 1 1 1 1 1 7 36 stock 1 1 1 1 1
;

/*creating longform dataset*/
data longform;
set portfolios;
array i[5] day1-day5;
do day=1 to 5;
increase=i[day];
output;
end;
keep portfolio asset type day increase;
run;

/*fitting hierarchical logistic model with random slopes and intercepts*/
proc glimmix method=Laplace;
class portfolio asset type;
model increase = type day / solution dist=binomial link=logit;
random intercept day / subject=portfolio type=un;
random intercept day / subject=asset(portfolio) type=un;
covtest/wald;
run;
```

The model doesn't converge.

The model that does converge and has a non-degenerate estimate of the parameters of the random-effects terms is as follows.

```
/*fitting hierarchical logistic model with random intercept only at asset level*/
proc glimmix method=Laplace;
class portfolio asset type(ref="bond");
```

```

model increase = type day / solution dist=binomial link=logit;
  random intercept / subject=asset(portfolio) type=un;
  covtest/wald;
run;

```

-2 Log Likelihood 211.02

Covariance Parameter Estimates					
Cov	Parm	Subject	Estimate	Standard Error	Z Value Pr >  Z
UN(1,1)	asset(portfolio)		0.7504	0.5360	1.40 0.0808

Solutions for Fixed Effects						
Effect	type	Estimate	Standard Error	DF	t Value	Pr >  t
Intercept		-1.1794	0.5912	33	-1.99	0.0544
type	currency	0.5227	0.6403	143	0.82	0.4157
type	stock	1.5515	0.5690	143	2.73	0.0072
type	bond	0	.	.	.	.
day		0.3384	0.1299	143	2.60	0.0102

The fitted hierarchical logistic model has the estimated odds

$\frac{\hat{P}(\text{increase}=1)}{1-\hat{P}(\text{increase}=1)} = \exp(-1.1794 + 0.5227 \cdot \text{currency} + 1.5515 \cdot \text{stock} + 0.3384 \cdot \text{day})$  and  $\hat{\sigma}_{\text{asset}}^2 = 0.7504$ . This variance is significant at the 10% level, therefore we can conclude that the measurements are correlated within each asset over time. However, the measurements for different assets are not correlated within each portfolio since the portfolio-level random-effects terms are not present in the model.

```

/*checking model fit*/
proc glimmix method=Laplace;
  class portfolio asset type;
  model increase = day type / dist=binomial link=logit;
run;

```

-2 Log Likelihood 215.51

```

data deviance;
  deviance = 215.51 - 211.02;
  pvalue = 1 - probchi(deviance, 1);
run;

```

```

proc print noobs;
run;

```

deviance	pvalue
4.49	0.034094

The model has a decent fit, significant at the 5% level since the p-value is less than 0.05.

In R:

```
portfolios.data<- read.csv(file="C:/./Exercise10.4Data.csv", header=TRUE,
sep=",")

#creating longform dataset
library(reshape2)
longform.data<- melt(portfolios.data, id.vars=c("portfolio","asset","type"),
variable.name="dayn",value.name="increase")

#creating variable for time
day<- ifelse(longform.data$dayn=="day1",1,ifelse(longform.data$dayn=="day2",
2,ifelse(longform.data$dayn=="day3",3, ifelse(longform.data$dayn=="day4",4,5))))

#fitting hierarchical logistic model with random slopes and intercepts
library(lme4)
summary(glmer(increase ~ type + day + (1 + day | portfolio)
+ (1 + day | portfolio:asset), data=longform.data, family=binomial("logit")))
```

The model converges but the only significant parameter of the random-effects terms is the intercept at the asset level, so a simpler model is run (the same as in SAS).

```
#fitting hierarchical logistic model with random intercept only at asset level
summary(fitted.model<- glmer(increase ~ type + day + (1 | portfolio:asset),
data=longform.data, family=binomial("logit")))
```

Random effects:

Groups	Name	Variance	Std.Dev.
portfolio:asset	(Intercept)	0.7504	0.8663

Fixed effects:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-1.1794	0.5912	-1.995	0.04606
typecurrency	0.5227	0.6403	0.816	0.41431
typestock	1.5514	0.5690	2.727	0.00640
day	0.3383	0.1299	2.605	0.00919

```
#checking model fit
null.model<- glm(increase ~ type + day, data=longform.data,
family=binomial("logit"))
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))
```

4.49183

```
print(p.value<- pchisq(deviance, df=1, lower.tail = FALSE))
```

0.03405719

(b) What predictors are significant at the 5% level? Interpret the estimated significant regression coefficients.

Stock and day are significant predictors. The estimated odds in favor of a stock going up at the closure of the stock exchange at the end of a day are  $\exp(1.5515) \cdot 100\% = 471.8543\%$  of those for a bond. Each day, the estimated odds increase by  $(\exp(0.3384) - 1) \cdot 100\% = 40.27015\%$ .

(c) According to the fitted model, what is the predicted probability of an increase in the value of a currency on the third day?

The predicted probability is  $P^0(\text{increase} = 1) = \frac{\exp(-1.1794+0.5227+0.3384 \cdot 3)}{1+\exp(-1.1794+0.5227+0.3384 \cdot 3)} = 0.588677$ .

In SAS:

```

/*using fitted model for prediction*/
data predict;
input portfolio asset type$ day;
cards;
8 37 currency 3
;

data longform;
set longform predict;
run;

proc glimmix method=Laplace;
class portfolio asset type;
model increase = type day/ dist=binomial link=logit;
random intercept / subject=asset(portfolio) type=un;
output out=outdata pred(ilink)=pincrease;
run;

proc print data=outdata (firstobs=181) noobs;
var pincrease;
run;

pincrease
0.58865

```

In R:

```

#using fitted model for prediction
print(predict(fitted.model, data.frame(portfolio=8, asset=37, type="currency",
day=3), allow.new.levels=TRUE, re.form=NA, type="response"))

0.5886426

```

**EXERCISE 10.5.** (a) Fit a four-level hierarchical regression to model the number of additional attempts: level 1 is tasks, level 2 is students, level 3 is classrooms, and level 4 is schools. Assume that the underlying distribution is Poisson. What is the fit of this model?

In SAS:

```

data test;
input school class student gender$ task1 task2 task3 task4 @@;
cards;
1 1 1 boy 1 3 4 5 1 1 2 boy 0 0 3 4 1 1 3 boy 1 2 4 5
1 1 4 girl 3 3 5 5 1 1 5 boy 1 1 4 13 1 1 6 girl 2 4 3 4
1 1 7 girl 1 2 3 7 1 2 1 boy 1 2 3 5 1 2 2 boy 10 11 7 6
1 2 3 girl 3 14 8 7 1 2 4 boy 2 2 5 6 1 2 5 girl 3 3 5 8
2 1 1 boy 11 3 4 8 2 1 2 boy 0 5 5 3 2 1 3 girl 2 6 7 9
2 1 4 boy 0 2 4 6 2 1 5 boy 2 3 3 5 2 1 6 girl 3 5 4 9
2 1 7 boy 1 3 7 3 2 2 1 girl 0 2 6 5 2 2 2 boy 0 10 4 13
2 2 3 girl 3 4 7 6 2 2 4 boy 1 2 5 3 2 2 5 boy 3 4 2 12
2 2 6 girl 1 10 6 8 2 2 7 girl 4 3 8 7 2 2 8 girl 12 5 4 5
2 3 1 girl 1 0 12 1 2 3 2 girl 0 1 2 4 2 3 3 boy 0 1 1 3
2 3 4 boy 0 1 0 2 2 3 5 girl 1 1 1 2 2 3 6 boy 0 0 0 1

```

```
;
/*creating longform dataset*/
data longform;
set test;
array a[4] task1-task4;
do task=1 to 4;
  nattempts=a[task];
  output;
end;
keep school class student gender task nattempts;
run;

/*fitting hierarchical Poisson model with random slopes and intercepts*/
proc glimmix method=Laplace;
class school class student gender;
model nattempts = gender task / solution dist=poisson link=log;
random intercept task / subject=school type=un;
random intercept task / subject=class(school) type=un;
random intercept task / subject=student(class) type=un;
covtest/wald;
run;
```

The model doesn't converge.

The model that does converge and gives nontrivial estimates of the parameters for the random-effects terms is the following model.

```
/*fitting hierarchical Poisson model with random intercepts only for class
and student levels*/
proc glimmix method=Laplace;
class school class student gender(ref="boy");
model nattempts = gender task / solution dist=poisson link=log;
random intercept / subject=class(school) type=un;
random intercept / subject=student(class) type=un;
covtest/wald;
run;
```

-2 Log Likelihood 592.37

Covariance Parameter Estimates						
Cov	Parm	Subject	Estimate	Standard Error	Z Value	Pr >  Z
UN(1,1)	class(school)		0.2072	0.1523	1.36	0.0868
UN(1,1)	student(class)		0.06298	0.03682	1.71	0.0436

Solutions for Fixed Effects						
Effect	gender	Estimate	Standard Error	DF	t Value	Pr >  t
Intercept		0.3361	0.2542	4	1.32	0.2566
gender	girl	0.2879	0.1204	107	2.39	0.0185
gender	boy	0	.	.	.	.
task		0.2947	0.04040	107	7.29	<.0001

```
/*checking model fit*/
proc glimmix method=Laplace;
class gender;
model nattempts = gender task / dist=poisson link=log;
```

```
run;
```

## 2 Log Likelihood 661.07

```
data deviance;  
  deviance = 661.07 - 592.37;  
  pvalue = 1 - probchi(deviance, 2);  
run;
```

```
proc print noobs;  
run;
```

deviance	pvalue
68.7	1.2212E-15

The model has an excellent fit due to a very small p-value in the deviance test.

In R:

```
test.data<- read.csv(file="C:/./Exercise10.5Data.csv", header=TRUE, sep=",")  
  
#creating longform dataset and numeric time variable  
library(reshape2)  
longform.data<- melt(test.data, id.vars=c("school","class","student", "gender"),  
  variable.name="taskn",value.name="nattempts")  
  
#creating variable for condition  
task<- ifelse(longform.data$taskn=="task1",1,ifelse(longform.data$taskn=="task2",  
  2,ifelse(longform.data$taskn=="task3",3,4)))  
  
#fitting hierarchical Poisson model with random slopes and intercepts  
library(lme4)  
summary(glmer(nattempts ~ gender + task + (1 + task | school) + (1 + task |  
  school:class) + (1 + task | class:student), data=longform.data,  
  family=poisson("log")))
```

The model doesn't converge.

```
#fitting hierarchical Poisson model with random intercepts only at class and  
#student levels  
summary(fitted.model<- glmer(nattempts ~ gender + task + (1 | school:class)  
  + (1 | class:student), data=longform.data, family=poisson("log")))
```

Random effects:

Groups	Name	Variance	Std.Dev.
class:student	(Intercept)	0.06297	0.2509
school:class	(Intercept)	0.20720	0.4552

Fixed effects:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	0.33609	0.25361	1.325	0.1851
gendergirl	0.28790	0.11972	2.405	0.0162
task	0.29469	0.04013	7.342	2.1e-13

```
#checking model fit  
null.model<- glm(nattempts ~ gender + task, data=longform.data, family =  
  poisson("log"))  
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))
```

68.70889

```
print(p.value<- pchisq(deviance, df=2, lower.tail = FALSE))
```

1.202414e-15

(b) Present the fitted model. What can you say about the correlation of the repeated measures for each student? Among the students in each classroom? Among the students in each school? Interpret estimated significant fixed-effects coefficients. Use the 10% level of significance.

The fitted hierarchical Poisson model has the estimated parameters  $\hat{E}(nattempts) = \exp(0.3361 + 0.2879 \cdot girl + 0.2947 \cdot task)$ ,  $\hat{\sigma}_{student}^2 = 0.06298$ , and  $\hat{\sigma}_{class}^2 = 0.2072$ . Since both variances are significantly larger than zero, the repeated measurements for each student are correlated, and the responses for students within each classroom are correlated as well. Responses for students within the same school are not correlated.

Both gender and task are significant fixed-effects predictors. The estimated average number of extra attempts for girls is  $\exp(0.2879) \cdot 100\% = 133.3624\%$  of that for boys. As the task number increases by one, the estimated average number of extra attempts increases by  $\exp(0.2947) \cdot 100\% = 134.2723\%$ .

(c) Use the fitted model to predict the number of extra attempts it would take a girl to complete the fourth task.

The predicted value is  $nattempts^0 = \exp(0.3361 + 0.2879 + 0.2947 \cdot 4) = 6.06661$ .

In SAS:

```
/*using fitted model for prediction*/
data predict;
input school class student gender$ task;
cards;
4 4 9 girl 4
;

data longform;
set longform predict;
run;

proc glimmix method=Laplace;
  class school class student gender;
  model nattempts = gender task / dist=poisson link=log;
  random intercept / subject=class(school) type=un;
  random intercept / subject=student(class) type=un;
  output out=outdata pred(ilink)=pnattempts;
run;
proc print data=outdata (firstobs=133) noobs;
  var pnattempts;
run;

pnattempts
  6.06606
```

In R:

```
#using fitted model for prediction
print(predict(fitted.model, data.frame(school=4, class=4, student=9,
gender="girl", task=4), allow.new.levels=TRUE, re.form=NA, type="response"))
```

6.066223

**EXERCISE 10.6.** (a) Argue that the data may be modeled as having a negative binomial distribution. What quantities support your argument?

The number of students who stay for Masters's degrees is an overly dispersed count data and thus may be modeled as a negative binomial random variable.

(b) Run a multilevel model, using department and year as predictors. Does the model fit the data well?

In SAS:

```
data masters;
input univ dept$ year1 year2 year3 @@;
cards;
1 bio 6 13 17 1 chem 8 7 12 1 math 10 14 13 2 bio 0 8 8
2 chem 0 9 9 2 math 0 5 8 3 bio 2 8 5 3 chem 3 3 5
3 math 18 19 26 4 bio 1 11 12 4 chem 1 5 4 4 math 5 16 17
5 bio 7 16 15 5 chem 4 4 4 5 math 8 1 6 6 bio 5 3 3
6 chem 5 3 4 6 math 23 32 45 7 bio 7 2 8 7 chem 9 12 9
7 math 7 15 16 8 bio 3 6 8 8 chem 32 11 20 8 math 8 4 13
;

/*creating longform dataset*/
data longform;
set masters;
array s[3] year1-year3;
do year=1 to 3;
nstay=s[year];
output;
end;
keep univ dept year nstay;
run;

/*fitting hierarchical negative binomial model with random slopes and
intercepts*/
proc glimmix method=Laplace;
class univ dept(ref="chem");
model nstay = dept year / solution dist=negbin link=log;
random intercept year / subject=univ type=un;
random intercept year / subject=dept(univ) type=un;
covtest/wald;
run;
```

The model doesn't converge.

The model that converges and gives positive estimates for the parameters of the random-effects terms is the following:

```
/*fitting hierarchical negative binomial model with random intercept
```

```

only at the department level*/
proc glimmix method=Laplace;
  class univ dept(ref="chem");
  model nstay = dept year / solution dist=negbin link=log;
  random intercept / subject=dept(univ) type=un;
  covtest/wald;
run;

```

-2 Log Likelihood 427.95

Covariance Parameter Estimates					
Cov	Parm	Subject	Estimate	Standard Error	Z Value Pr >  Z
UN(1,1)	dept(univ)		0.2658	0.09659	2.75 0.0030
Scale			0.07351	0.04450	1.65 0.0493

Solutions for Fixed Effects					
Effect	dept	Estimate	Standard Error	DF	t Value Pr >  t
Intercept		1.3133	0.2489	21	5.28 <.0001
dept	bio	0.02657	0.2942	21	0.09 0.9289
dept	math	0.5766	0.2899	21	1.99 0.0599
dept	chem	0	.	.	.
year		0.2680	0.06508	47	4.12 0.0002

```

/*checking model fit*/
proc glimmix;
  class dept;
  model nstay = dept year / dist=negbin link=log;
run;

```

-2 Log Likelihood 450.08

```

data deviance;
  deviance = 450.08 - 427.95;
  pvalue = 1 - probchi(deviance, 1);
run;

```

```

proc print noobs;
run;

```

deviance	pvalue
22.13	.000002548

The model has a very good fit as indicated by the tiny p-value in the deviance test.

In R:

```

masters.data<- read.csv(file="C:/./Exercisel0.6Data.csv", header=TRUE, sep=",")

#creating longform dataset
library(reshape2)
longform.data<- melt(masters.data, id.vars=c("univ","dept"),
  variable.name="yearn", value.name="nstay")

```

```
#creating variable for time
year<- ifelse(longform.data$year=="year1",1,
ifelse(longform.data$year=="year2",2,3))

#specifying reference level
dept.rel<- relevel(longform.data$dept, ref="chem")

#fitting hierarchical negative binomial model with random slopes and intercepts
library(lme4)
summary(glmer.nb(nstay ~ dept.rel + year + (1 + year | univ) + (1 + year |
univ:dept.rel), data=longform.data, family=negative.binomial("log")))
```

The model doesn't converge.

```
#fitting hierarchical negative binomial model with random
#intercept only at department level
summary(fitted.model<- glmer.nb(nstay ~ dept.rel + year + (1 | univ:dept.rel),
data=longform.data, family=negative.binomial("log")))
```

Family: Negative Binomial(13.5539) ( log )

Random effects:

Groups	Name	Variance	Std.Dev.
univ:dept.rel	(Intercept)	0.2651	0.5149

Fixed effects:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	1.30677	0.24837	5.261	1.43e-07
dept.relbio	0.02647	0.29368	0.090	0.9282
dept.relmath	0.57593	0.28943	1.990	0.0466
year	0.26652	0.06459	4.127	3.68e-05

```
#checking model fit
library(MASS)
null.model<- glm.nb(nstay ~ year, data=longform.data)
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))
```

22.18095

```
print(p.value<- pchisq(deviance, df=1, lower.tail = FALSE))
```

2.481229e-06

(c) Are the observations correlated for each department over time? For the departments within the same university? State the fitted model, specifying all parameter estimates.

In SAS, the fitted hierarchical negative binomial model has the estimated parameters  $\hat{E}(nstay) = \exp(1.3133 + 0.02657 \cdot \text{biology dept} + 0.5766 \cdot \text{math dept} + 0.268 \cdot \text{year})$ ,  $\hat{r} = 0.07351$ , and  $\hat{\sigma}_{u_1}^2 = 0.2658$ .

Since the variance of  $u_1$  is statistically significant, we can conclude that observations over time within each department are correlated. Observations for each department within the same university are not correlated since  $\tau_1$  and  $\tau_2$  are not present in the fitted model.

In R, the fitted model has parameters  $\hat{E}(nstay) = \exp(1.30677 + 0.02647 \cdot \text{biology dept} + 0.57593 \cdot \text{math dept} + 0.26652 \cdot \text{year})$ ,  $\hat{r} = \ln(13.5539) = 2.606674$ , and  $\hat{\sigma}_{u_1}^2 = 0.2651$ .

(d) Does the response change significantly over the years? Is there a difference in responses between departments? Give an interpretation of the significant regression coefficients. Use  $\alpha = 0.10$ .

Math department and year are significant predictors, thus there is a difference in the response between departments, and the response changes significantly over the years.

In a math department, the estimated average number of students who stay on for a Master's degree is  $\exp(0.5766) \cdot 100\% = 177.9976\%$  of that in a chemistry department (for the model fitted in R,  $\exp(0.57593) \cdot 100\% = 177.8784\%$ ). As the year increases by one, the estimated average number of students who stay on for a Master's degree increases by  $(\exp(0.268) - 1) \cdot 100\% = 30.73471\%$  (for the model fitted in R,  $(\exp(0.26652) - 1) \cdot 100\% = 30.54137\%$ ).

(e) What is the predicted number of students who would stay on for a Master's program in a math department in year 4?

For the model fitted in SAS, the predicted value is  $nstay^0 = \exp(1.3133 + 0.5766 + 0.268 \cdot 4) = 19.33467$ .

In SAS:

```
/*using model for prediction*/
data predict;
input univ dept$ year;
cards;
9 math 4
;

data longform;
set longform predict;
run;

proc glimmix method=Laplace;
class univ dept;
model nstay = dept year / dist=negbin link=log;
random intercept / subject=dept(univ) type=un;
output out=outdata pred(ilink)=pnstay;
run;

proc print data=outdata (firstobs=73) noobs;
var pnstay;
run;

pnstay
19.3329
```

For the model fitted in R, the predicted value is  $nstay^0 = \exp(1.30677 + 0.57593 + 0.26652 \cdot 4) = 19.08266$ .

In R:

```
#using fitted model for prediction
print(predict(fitted.model, data.frame(univ=9, dept.rel="math", year=4),
allow.new.levels=TRUE, re.form=NA, type="response"))
```

**EXERCISE 10.7.** (a) Run the multilevel regression to model the response to medication, assuming that it follows a beta distribution.

In SAS:

```
data trial;
input center subject gender$ medA medB medC medD @@;
cards;
1 101 M 0.32 0.27 0.23 0.90 1 102 M 0.17 0.16 0.35 0.40
1 103 F 0.39 0.44 0.45 0.64 1 104 M 0.14 0.47 0.63 0.76
1 105 F 0.08 0.36 0.40 0.72 1 106 F 0.61 0.53 0.64 0.79
1 107 F 0.55 0.73 0.63 0.61 1 108 M 0.40 0.47 0.46 0.99
1 109 F 0.25 0.40 0.31 0.62 1 110 M 0.34 0.48 0.29 0.63
1 111 M 0.33 0.42 0.43 0.75 1 112 F 0.21 0.39 0.74 0.98
1 113 F 0.39 0.22 0.50 0.88 1 114 M 0.33 0.30 0.26 0.19
1 115 F 0.03 0.49 0.36 0.73 2 201 M 0.31 0.46 0.53 0.81
2 202 F 0.27 0.57 0.28 0.84 2 203 M 0.26 0.42 0.38 0.90
2 204 M 0.33 0.34 0.56 0.75 2 205 F 0.29 0.45 0.57 0.81
2 206 F 0.30 0.42 0.64 0.95 2 207 F 0.34 0.42 0.55 0.77
2 208 M 0.09 0.35 0.42 0.67 2 209 M 0.25 0.44 0.62 0.73
2 210 F 0.21 0.41 0.58 0.75 3 301 F 0.23 0.41 0.50 0.86
3 302 F 0.21 0.35 0.52 0.84 3 303 M 0.21 0.43 0.68 0.72
3 304 M 0.07 0.23 0.47 0.59 3 305 M 0.11 0.28 0.50 0.78
3 306 F 0.19 0.24 0.55 0.73 3 307 M 0.15 0.23 0.39 0.82
3 308 F 0.18 0.19 0.53 0.92
;

/*creating longform dataset*/
data longform;
set trial;
array r[4] medA medB medC medD;
do med=1 to 4;
    response=r[med];
    output;
end;
keep center subject gender med response;
run;

/*fitting hierarchical beta model with random slopes and intercepts*/
proc glimmix method=Laplace;
class center subject gender;
model response = gender med / solution dist=beta link=logit;
random intercept med / subject=center type=un;
random intercept med / subject=subject(center) type=un;
covtest/wald;
run;
```

The model doesn't converge.

The model that converges and gives positive estimates for the parameters of the random-effects terms is the following:

```
/*fitting hierarchical beta model with random intercept only at subject level*/
proc glimmix method=Laplace;
```

```

class center subject gender;
model response = gender med / solution dist=beta link=logit;
random intercept / subject=subject(center) type=un;
covtest/wald;
run;

```

-2 Log Likelihood -153.06

Covariance Parameter Estimates					
Cov	Parm	Subject	Estimate	Standard Error	Pr >  Z
UN(1,1)		subject(center)	0.06449	0.04224	1.53 0.0634
Scale			11.2305	1.5764	7.12 <.0001

Solutions for Fixed Effects					
Effect	gender	Estimate	Standard Error	DF	t Value Pr >  t
Intercept		-2.0131	0.1656	31	-12.16 <.0001
gender	F	0.2497	0.1386	98	1.80 0.0746
gender	M	0	.	.	. .
med		0.7078	0.04973	98	14.23 <.0001

In R:

```

trial.data<- read.csv(file="C:/./Exercise10.7Data.csv", header=TRUE, sep=",")

#creating longform dataset
library(reshape2)
longform.data<- melt(trial.data, id.vars=c("center","subject","gender"),
variable.name="medn", value.name="response")

#creating variable for condition
med<- ifelse(longform.data$medn=="medA",1,ifelse(longform.data$medn=="medB",
2,ifelse(longform.data$medn=="medC",3,4)))

#specifying reference level
gender.rel<- relevel(longform.data$gender, ref="M")

#fitting hierarchical beta model with random slopes and intercepts
library(glmmTMB)
summary(glmmTMB(response ~ gender.rel + med + (1 + med | center) + (1 + med |
center:subject), data=longform.data, family=beta_family(link="logit")))

```

The model converges but all estimates of random-effects terms are degenerate.

```

#fitting hierarchical beta model with random intercept only at subject level
summary(fitted.model<- glmmTMB(response ~ gender.rel + med
+ (1 | center:subject), data=longform.data, family=beta_family(link="logit")))

```

Random effects:

Groups	Name	Variance	Std.Dev.
center:subject	(Intercept)	0.06449	0.2539

Overdispersion parameter for beta family (): 11.2

Conditional model:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-2.01310	0.16555	-12.160	<2e-16

gender.relF	0.24967	0.13855	1.802	0.0715
med	0.70779	0.04973	14.233	<2e-16

(c) State the model and estimate the parameters. What random-effects terms are present? Discuss the model fit. For all significance use the 10% level.

The fitted hierarchical beta model has the estimated parameters

$$\hat{E}(\text{response}) = \frac{\exp(-2.0131 + 0.2497 \cdot \text{female} + 0.7078 \cdot \text{medication})}{1 + \exp(-2.0131 + 0.2497 \cdot \text{female} + 0.7078 \cdot \text{medication})} \text{ and } \hat{\alpha} = 11.2305.$$

At the 10% significance level, the variance of the random intercept at the subject level is positive, and both gender and medication are significant predictors of the mean response.

The model fits the data well at the 10% significance level since the p-value of the deviance test is below 0.10.

In SAS:

```
/*checking model fit*/
proc glimmix;
  class gender;
  model response = gender med / dist=beta link=logit;
run;
```

**-2 Log Likelihood -149.36**

```
data deviance;
  deviance = -149.36 - (-153.06);
  pvalue = 1 - probchi(deviance, 1);
run;
```

```
proc print noobs;
run;
```

deviance	pvalue
3.7	0.054412

In R:

```
#checking model fit
library(betareg)
summary(null.model<- betareg(response ~ gender.rel + med,
data=longform.data,link="logit"))
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))
```

**3.69739**

```
print(p.value<- pchisq(deviance, df=1, lower.tail = FALSE))
```

**0.05449765**

(c) Interpret the results. Are responses correlated for each subject? For each center? Interpret estimated significant fixed-effects terms.

Since the variance of  $u_1$  is positive, and the other random-effects terms are not present in the model, we conclude that responses to medications are correlated within each subject but uncorrelated between the subjects.

For female subjects, the ratio  $\frac{\hat{\mu}}{1-\hat{\mu}}$  is  $\exp(0.2497) \cdot 100\% = 128.364\%$  of that for male subjects. As the number of medications increases by one (A=1, B=2, C=3, D=4), the ratio  $\frac{\hat{\mu}}{1-\hat{\mu}}$  increases by  $(\exp(0.7078) - 1) \cdot 100\% = 102.9521\%$ .

(d) Use the fitted model to predict the response to medication A in a female subject.

The predicted value is  $response^0 = \frac{\exp(-2.0131+0.2497+0.7078)}{1+\exp(-2.0131+0.2497+0.7078)} = 0.258151$ .

In SAS:

```
/*using fitted model for prediction*/
data predict;
input center subject gender$ med;
cards;
4 309 F 1
;

data longform;
set longform predict;
run;

proc glimmix method=Laplace;
class center subject gender;
model response = gender med / dist=beta link=logit;
random intercept / subject=subject(center) type=un;
output out=outdata pred(ilink)=presponse;
run;

proc print data=outdata (firstobs=133) noobs;
var presponse;
run;

presponse
0.25815
```

In R:

```
#using fitted model for prediction
print(predict(fitted.model, data.frame(center=4, subject=309, gender.rel="F",
med=1), allow.new.levels=TRUE, re.form=NA, type="response"))

0.2581452
```