STAT 410 PROJECT

Modeling the Probability of Stroke
(Binary Logistic + Probit + Complementary Log Log Regression Models)

Submitted to
Prof. Olga Korosteleva

Report Prepared by
Sample Student

November 30, 2022

Contents

I. Introduction

One of the few things I have been interested in since the beginning of my academic career was health science. Although I am pursuing a career path of a statistician, I believe I could blend the two fields. I found a dataset which was able to satisfy my interests, one in which it modeled the probability of a patient having a stroke based on numerous variables. With Stroke being modeled as a binary response variable and there being 11 other predictors with 5000+ rows of data, I knew this dataset was what I was looking for.

II. Background

A stroke is essentially a disease in which a blood vessel connected to the brain is blocked by a clot or bursts, preventing oxygen from reaching the brain causing brain cells to die [1]. Stroke is loosely referred to as a "brain attack" with most pain occurring in the head area as compared to a "heart attack" where pain would occur in the chest. According to the World Health Organization (WHO) stroke is the second leading cause of death globally and that one in four people are at risk of stroke in their lifetime [2].

III. Data Description

The data was found through a website called "Kaggle.com", an online community platform for data scientists, where users can find and publish datasets [3].

The data itself contained 5000+ rows of data with 11 variables including: gender (male/female/other), age, hypertension (0 /1), heart disease (0/1) , marital status(yes/no), work type (Government Job/Never Worked/Self-employed/Private/Child), residence type (Urban/Rural), average glucose level, bmi, smoking status (Formerly/Never/Smokes/Unknown), and if they had previously had a stroke (0/1) [4]. From the dataset there are three binary response variables modeled through zero for no and one for yes, which include: hypertension, heart disease, and stroke. I did have to clean the dataset in both SAS and R, removing any "N/A" values presented within the data set which transformed the data from 5000+ rows to 4700 rows. We will be modeling the prediction probability of a stroke occurring based on the predictors present within the dataset.

IV. Results

The probit model proved to be the best model of the three due to the lower AIC, AICC, and BIC values. There were five significant predictors at the .05 significance level which were: age, patient having hypertension, self-employed work type, average glucose level, and smoking status as someone who currently smokes. What was interesting was that in both models, SAS and R, the presence of heart disease was not significant at the .05 level, since in the beginning I assumed it would be. Another interesting observation was that employment status was the only significant variable in which the predictor was less than the reference, meaning that those who are self employed in this study have a lower probability of having a stroke than those who work in private company jobs. Both hypertension and smoking status had the largest unit increases for probability of having a stroke with their values almost reaching 0.3 units higher.

When running the deviance test I modeled the value for a possibility of a stroke which I had to specify using the "descending" option in the proc genmod statement since stroke was a binary variable. From the results, a log likelihood value of -864.1931 was found. Plugging the original likelihood of -679.9565 along with the new log likelihood and 16 degrees of freedom, a value of 368.473 and an extremely small *p*-value which sas displayed as "0" was found. This proved the model was a great fit with all the predictors for the data as the *p*-value at the .05 significance level was significant.

The fitted model for the data is Φ^{-1} ($\hat{p}$(stroke)) = -4.0359 + 0.228 * *Female* + -3.8547 * *Other* + 0.0342 * *age* + 0.2910 * *hypertension* + 0.1968 * *heart_disease* + 0.0704 * *ever_married_no* + -0.0875 * *govt_job* + -3.3968 * *never_worked* + -0.1961 * *self_employed* + 0.4837 * *children* + -0.0069 * *rural* + 0.0023 * *avg_glucose_level* + -0.0008 * *bmi1* + 0.1269 * *formerly smoked* + 0.0801 * *never_smoked* + 0.2695 * *smokes*. The interpretation for the predictors is as follows, for each one year increase in age the probability of having a stroke increases by .0342 units, patients who have hypertension has a .2910 increased probability of having a stroke than those without hypertension, those who identify as self-employed have a .1961 less probability of having a stroke than those who identify working for a private company, every increase in average glucose level (mg/dcl) the probability of having a stroke increases by .0023 units, and those who identify as smokers have an increased probability of .2695 units greater than those who identified as unknown smoking status.

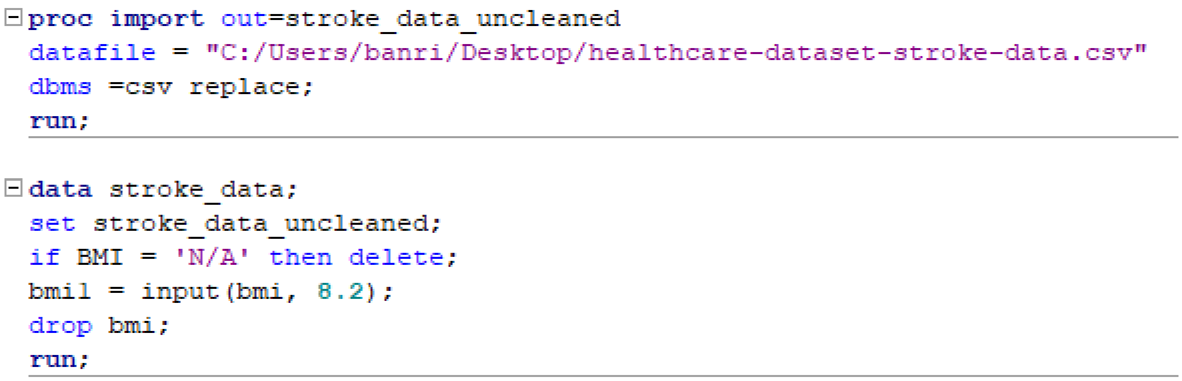
When using the model for prediction I predicted for a Female, who is 75 years old, has hypertension, no heart disease, is married, works in the private sector, lives in an urban environment, average glucose level of 100, bmi of 21.5, and has never smoked which yielded a probability of .19420 or a 19.420% chance of having a stroke. This probability seems very plausible given the statistics of the hypothesized patient. However, this differs from what was presented within the presentation as I realized I made an error when coding and was modeling for the probability of not having a stroke which was displayed as 1 - .19420 as ".80580".

V. Conclusion

Apart from realizing I made a grave error in presentation by presenting the probability of not having a stroke rather than presenting probability of having a stroke, the dataset and modeling did a satisfactory job. I was able to predict the probability of having a stroke for someone similar to my mother's statistics quite accurately. On the other hand, to improve upon the model it should be considered that the data is highly unbalanced with there being only 209 patients who had a stroke versus 4700 patients who marked not having a stroke. In retrospect, it would have been better to randomly sample from the dataset first before running any modeling. Additionally, maybe running the prediction and comparison to a patient who had already had a stroke would have been useful in order to "check" the accuracy of the model. In summary, this project was an opportunity to apply the techniques we learned in class to a topic of our own interest so it was a great learning experience.

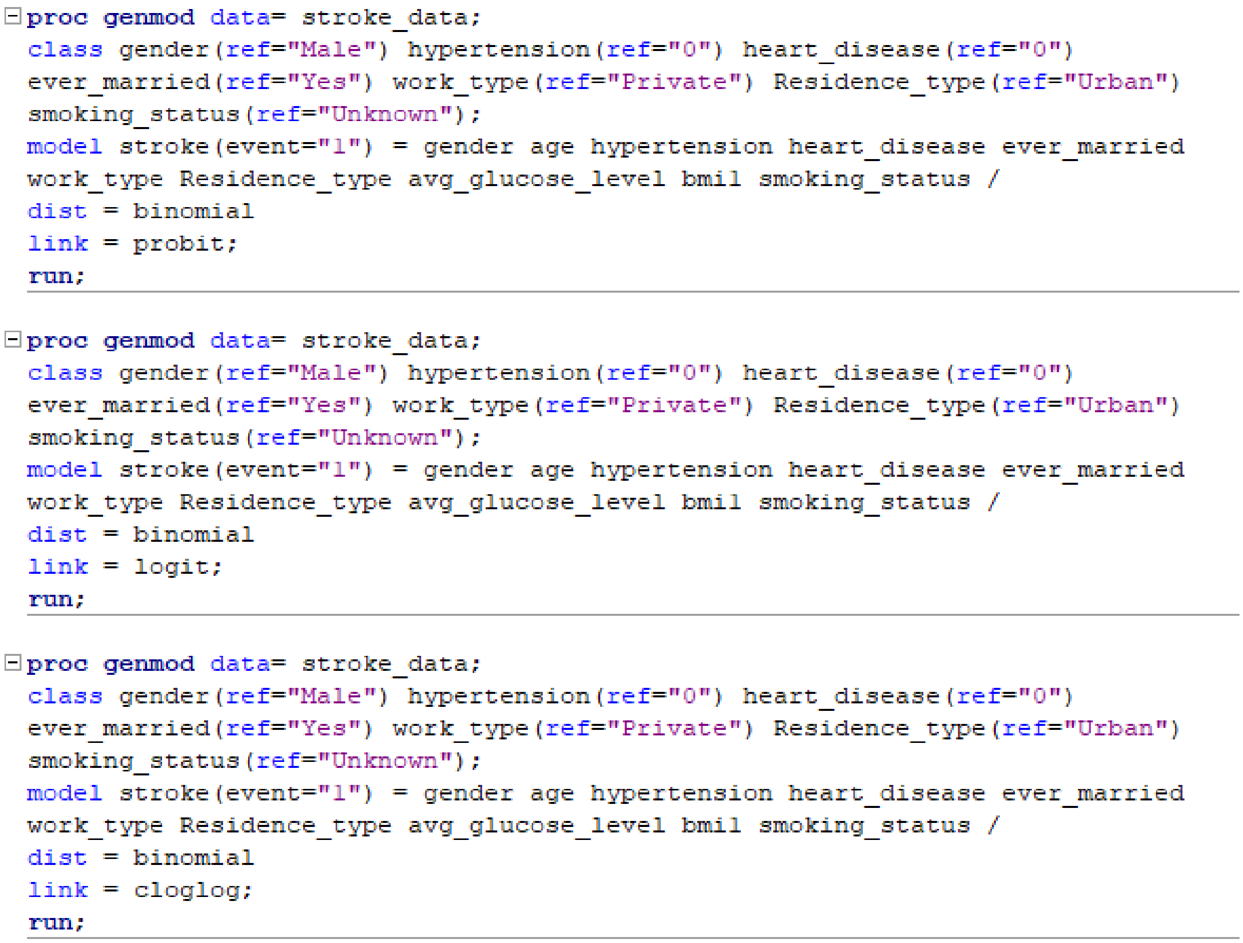
VI. Appendix

A. SAS CODE 1



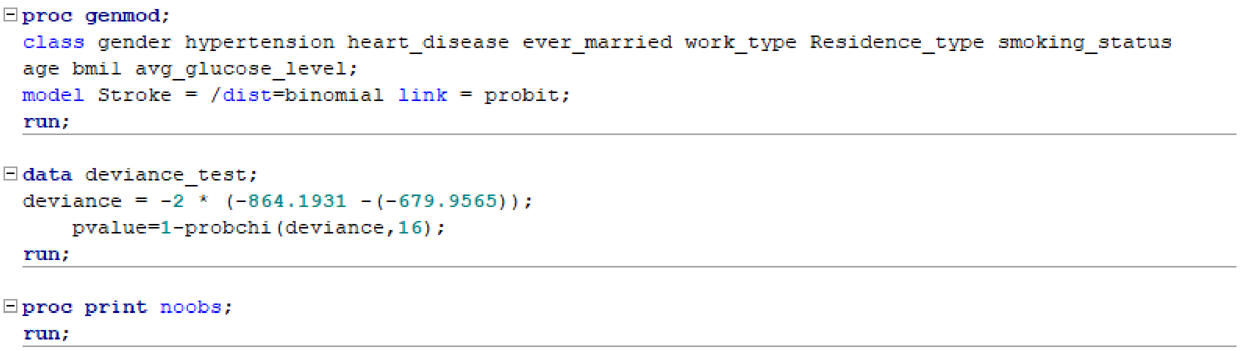
```
proc import out=stroke_data_uncleaned
datafile = "C:/Users/banri/Desktop/healthcare-dataset-stroke-data.csv"
dbms =csv replace;
run;

data stroke_data;
set stroke_data_uncleaned;
if BMI = 'N/A' then delete;
bmi1 = input(bmi, 8.2);
drop bmi;
run;
```

Picture 1: Importing the data and cleaning the dataset.



```
proc genmod data= stroke_data;
class gender(ref="Male") hypertension(ref="0") heart_disease(ref="0")
ever_married(ref="Yes") work_type(ref="Private") Residence_type(ref="Urban")
smoking_status(ref="Unknown");
model stroke(event="1") = gender age hypertension heart_disease ever_married
work_type Residence_type avg_glucose_level bmi1 smoking_status /
dist = binomial
link = probit;
run;

proc genmod data= stroke_data;
class gender(ref="Male") hypertension(ref="0") heart_disease(ref="0")
ever_married(ref="Yes") work_type(ref="Private") Residence_type(ref="Urban")
smoking_status(ref="Unknown");
model stroke(event="1") = gender age hypertension heart_disease ever_married
work_type Residence_type avg_glucose_level bmi1 smoking_status /
dist = binomial
link = logit;
run;

proc genmod data= stroke_data;
class gender(ref="Male") hypertension(ref="0") heart_disease(ref="0")
ever_married(ref="Yes") work_type(ref="Private") Residence_type(ref="Urban")
smoking_status(ref="Unknown");
model stroke(event="1") = gender age hypertension heart_disease ever_married
work_type Residence_type avg_glucose_level bmi1 smoking_status /
dist = binomial
link = cloglog;
run;
```

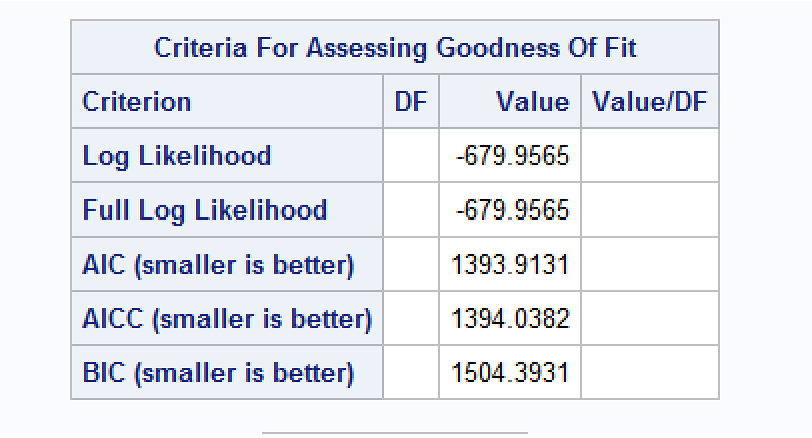
Picture 2: Running the Probit, Logit, and Complementary Log-log models.



```
proc genmod;
class gender hypertension heart_disease ever_married work_type Residence_type smoking_status
age bmil avg_glucose_level;
model Stroke = /dist=binomial link = probit;
run;

data deviance_test;
deviance = -2 * (-864.1931 -(-679.9565));
    pvalue=1-probchi(deviance,16);
run;

proc print noobs;
run;
```

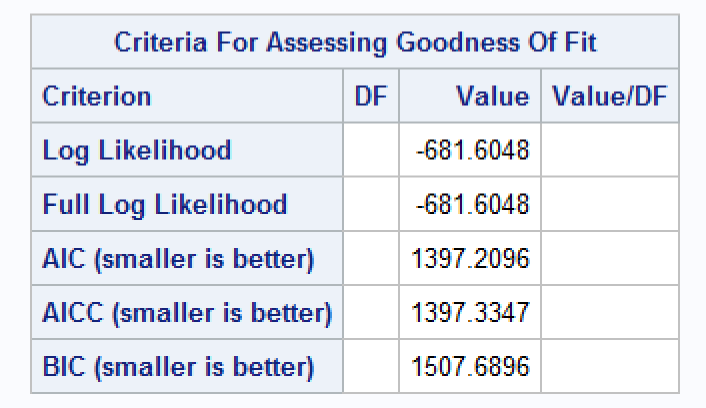
Picture 3: Verifying model fit.

B. SAS CODE 1 OUTPUT



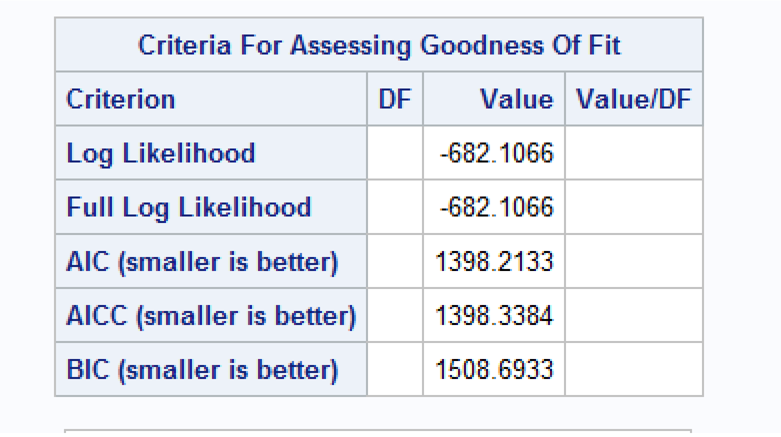
Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Log Likelihood		-679.9565	
Full Log Likelihood		-679.9565	
AIC (smaller is better)		1393.9131	
AICC (smaller is better)		1394.0382	
BIC (smaller is better)		1504.3931	

Picture 4: Probit output.



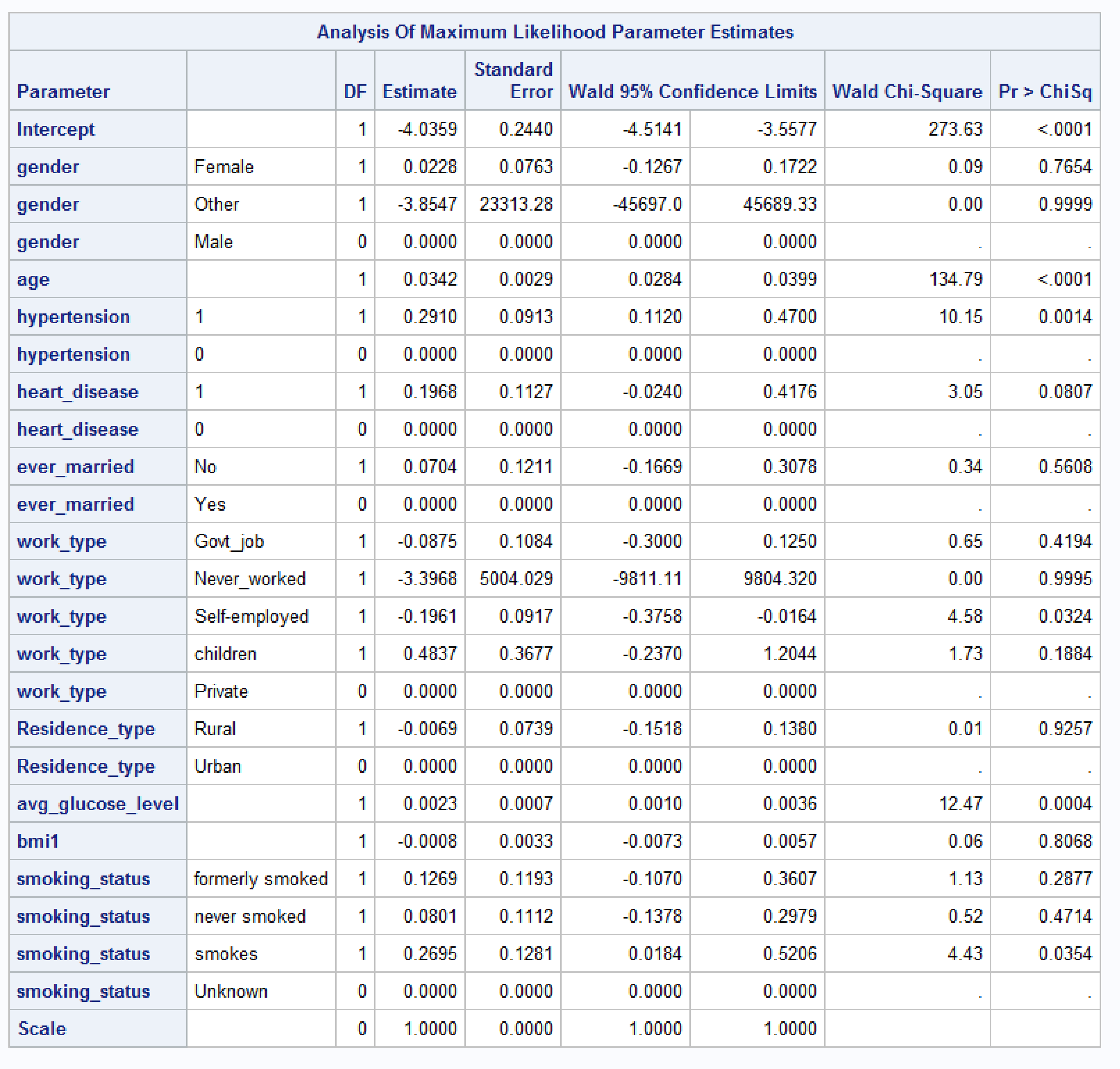
Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Log Likelihood		-681.6048	
Full Log Likelihood		-681.6048	
AIC (smaller is better)		1397.2096	
AICC (smaller is better)		1397.3347	
BIC (smaller is better)		1507.6896	

Picture 5: Logit Output.



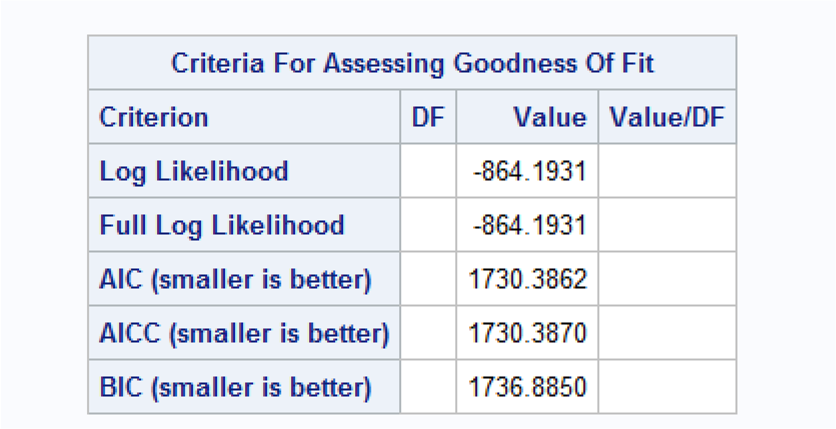
Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Log Likelihood		-682.1066	
Full Log Likelihood		-682.1066	
AIC (smaller is better)		1398.2133	
AICC (smaller is better)		1398.3384	
BIC (smaller is better)		1508.6933	

Picture 6: Complementary Log-log output.



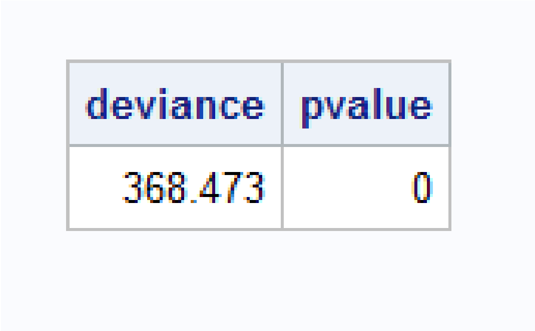
Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	-4.0359	0.2440	-4.5141	-3.5577	273.63	<.0001
gender	Female	1	0.0228	0.0763	-0.1267	0.1722	0.09	0.7654
gender	Other	1	-3.8547	23313.28	-45697.0	45689.33	0.00	0.9999
gender	Male	0	0.0000	0.0000	0.0000	0.0000	.	.
age		1	0.0342	0.0029	0.0284	0.0399	134.79	<.0001
hypertension	1	1	0.2910	0.0913	0.1120	0.4700	10.15	0.0014
hypertension	0	0	0.0000	0.0000	0.0000	0.0000	.	.
heart_disease	1	1	0.1968	0.1127	-0.0240	0.4176	3.05	0.0807
heart_disease	0	0	0.0000	0.0000	0.0000	0.0000	.	.
ever_married	No	1	0.0704	0.1211	-0.1669	0.3078	0.34	0.5608
ever_married	Yes	0	0.0000	0.0000	0.0000	0.0000	.	.
work_type	Govt_job	1	-0.0875	0.1084	-0.3000	0.1250	0.65	0.4194
work_type	Never_worked	1	-3.3968	5004.029	-9811.11	9804.320	0.00	0.9995
work_type	Self-employed	1	-0.1961	0.0917	-0.3758	-0.0164	4.58	0.0324
work_type	children	1	0.4837	0.3677	-0.2370	1.2044	1.73	0.1884
work_type	Private	0	0.0000	0.0000	0.0000	0.0000	.	.
Residence_type	Rural	1	-0.0069	0.0739	-0.1518	0.1380	0.01	0.9257
Residence_type	Urban	0	0.0000	0.0000	0.0000	0.0000	.	.
avg_glucose_level		1	0.0023	0.0007	0.0010	0.0036	12.47	0.0004
bmi1		1	-0.0008	0.0033	-0.0073	0.0057	0.06	0.8068
smoking_status	formerly smoked	1	0.1269	0.1193	-0.1070	0.3607	1.13	0.2877
smoking_status	never smoked	1	0.0801	0.1112	-0.1378	0.2979	0.52	0.4714
smoking_status	smokes	1	0.2695	0.1281	0.0184	0.5206	4.43	0.0354
smoking_status	Unknown	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		0	1.0000	0.0000	1.0000	1.0000		

Picture 7: Estimated Probit coefficients and *p*-values. Significant variables here include: Age, Hypertension answered 1, work type as "self-employed", average glucose level, and smoking status as "smokes".



Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Log Likelihood		-864.1931	
Full Log Likelihood		-864.1931	
AIC (smaller is better)		1730.3862	
AICC (smaller is better)		1730.3870	
BIC (smaller is better)		1736.8850	

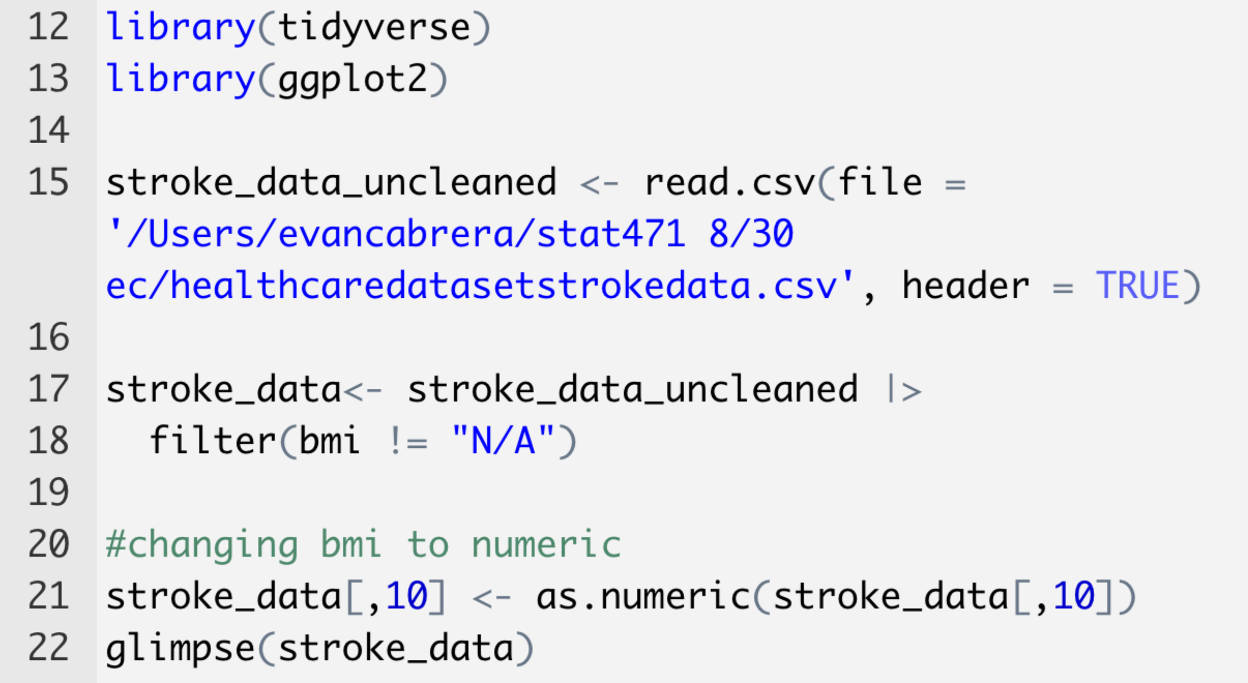
Picture 8: Acquiring the 2nd log likelihood value.



deviance	pvalue
368.473	0

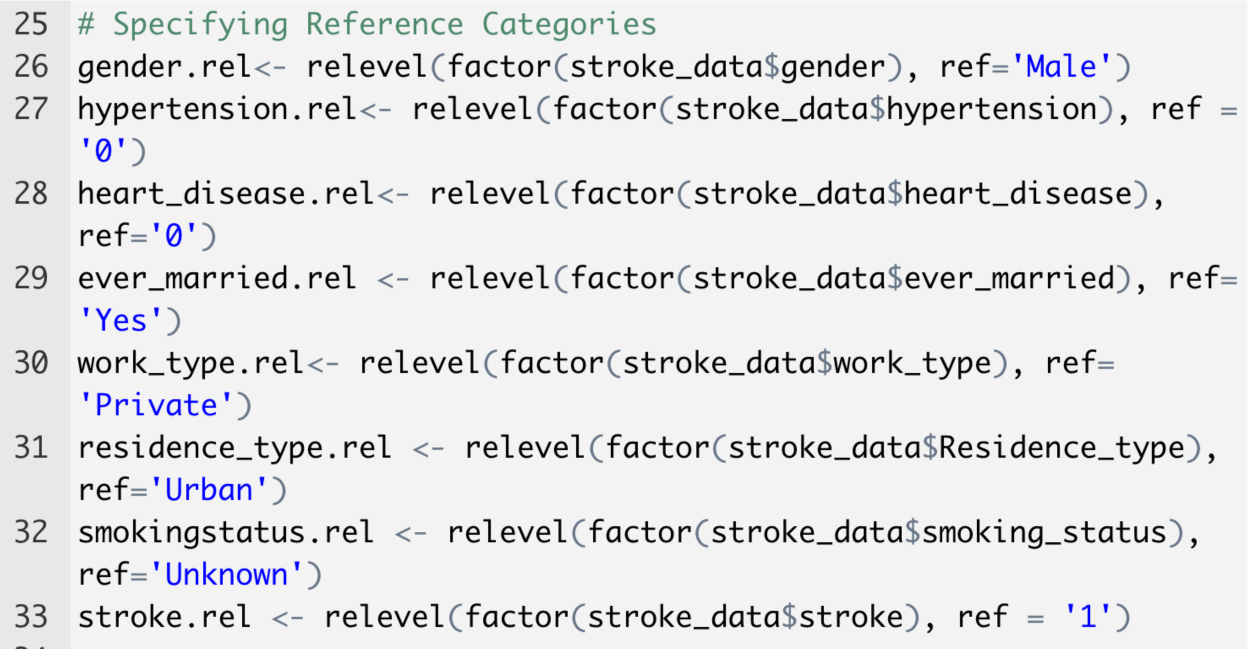
Picture 9: Verifying model fit. Model is a good fit due to the small *p*-value.

C. R CODE 1



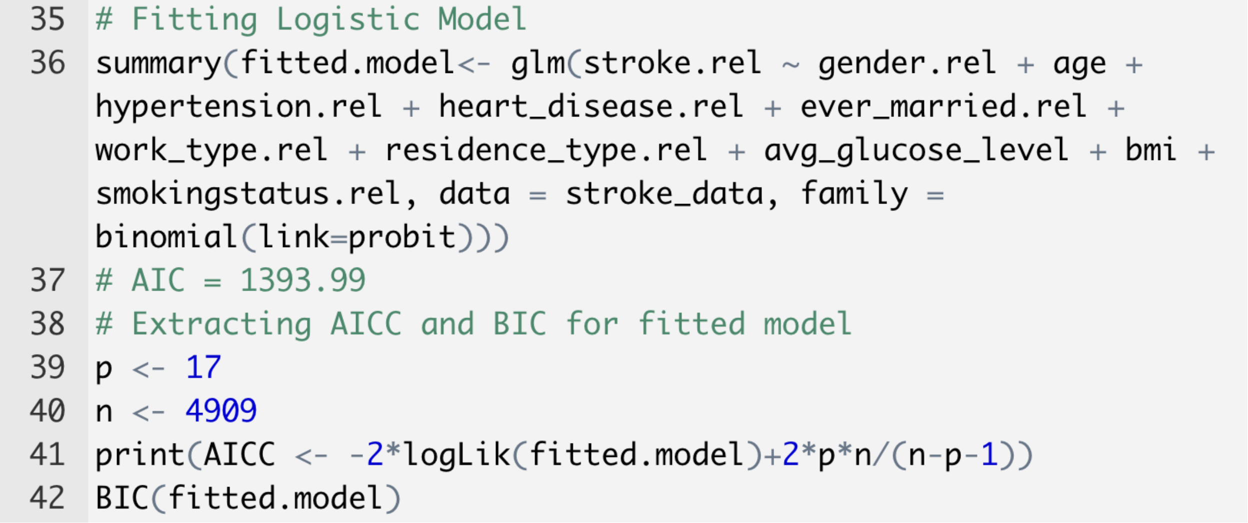
```r
library(tidyverse)
library(ggplot2)

stroke_data_uncleaned <- read.csv(file = '/Users/evancabrera/stat471 8/30 ec/healthcaredatasetstrokedata.csv', header = TRUE)

stroke_data<- stroke_data_uncleaned |>
  filter(bmi != "N/A")

#changing bmi to numeric
stroke_data[,10] <- as.numeric(stroke_data[,10])
glimpse(stroke_data)
```

Picture 10: Reading in the data and cleaning the data in R.



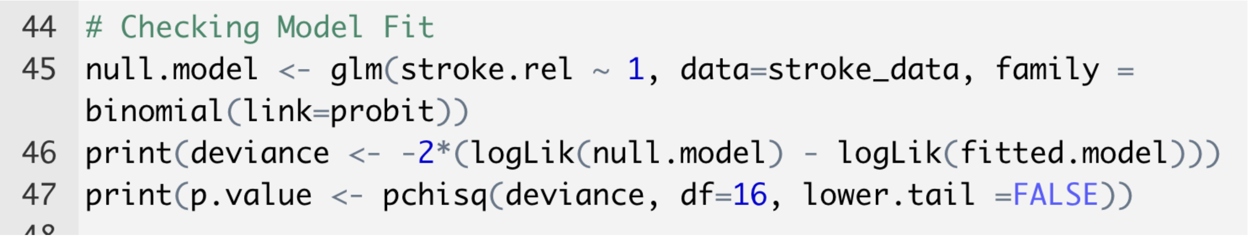
```r
# Specifying Reference Categories
gender.rel<- relevel(factor(stroke_data$gender), ref='Male')
hypertension.rel<- relevel(factor(stroke_data$hypertension), ref = '0')
heart_disease.rel<- relevel(factor(stroke_data$heart_disease), ref='0')
ever_married.rel <- relevel(factor(stroke_data$ever_married), ref= 'Yes')
work_type.rel<- relevel(factor(stroke_data$work_type), ref= 'Private')
residence_type.rel <- relevel(factor(stroke_data$Residence_type), ref='Urban')
smokingstatus.rel <- relevel(factor(stroke_data$smoking_status), ref='Unknown')
stroke.rel <- relevel(factor(stroke_data$stroke), ref = '1')
```

Picture 11: Specifying Reference categories.



```r
# Fitting Logistic Model
summary(fitted.model<- glm(stroke.rel ~ gender.rel + age +
hypertension.rel + heart_disease.rel + ever_married.rel +
work_type.rel + residence_type.rel + avg_glucose_level + bmi +
smokingstatus.rel, data = stroke_data, family =
binomial(link=probit)))
# AIC = 1393.99
# Extracting AICC and BIC for fitted model
p <- 17
n <- 4909
print(AICC <- -2*logLik(fitted.model)+2*p*n/(n-p-1))
BIC(fitted.model)
```

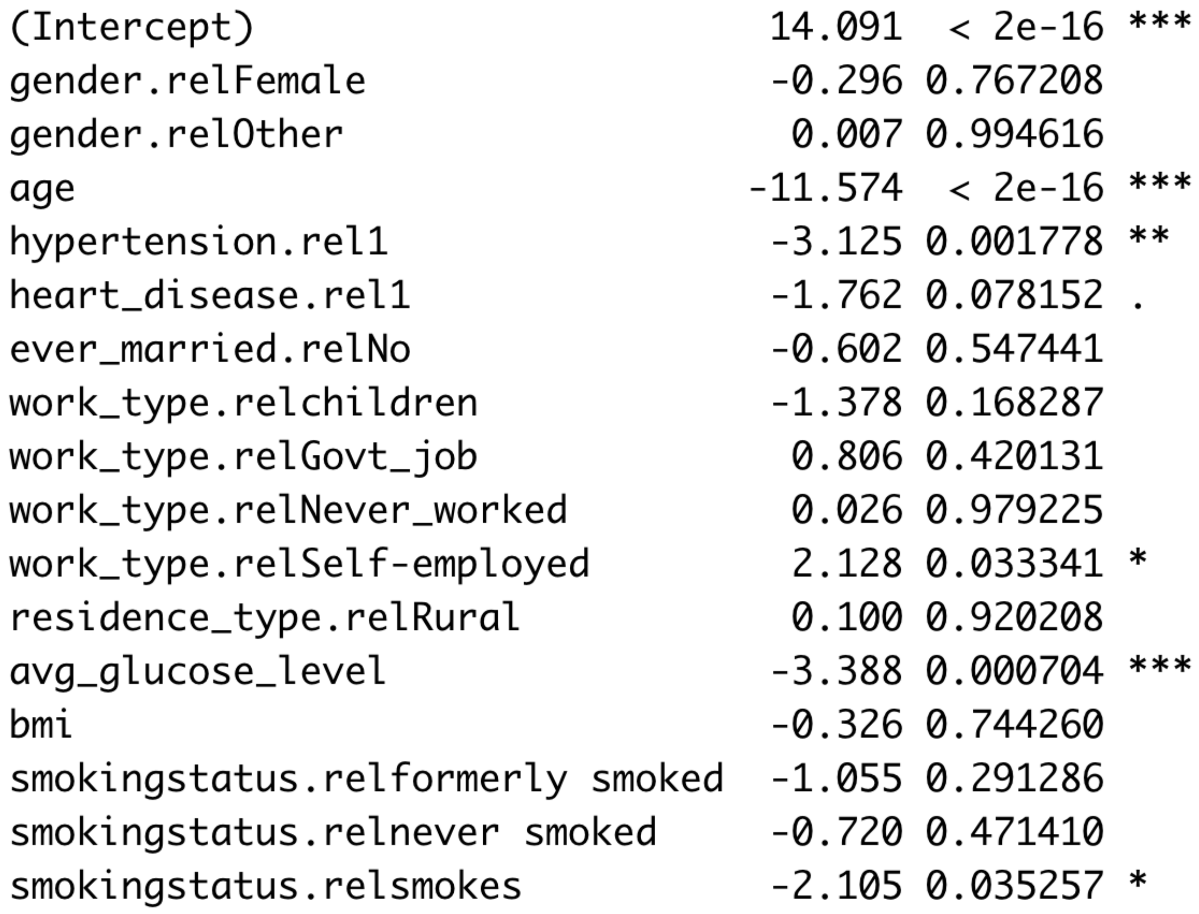
Picture 12: Fitting the model and acquiring AIC, AICC, and BIC.



```r
# Checking Model Fit
null.model <- glm(stroke.rel ~ 1, data=stroke_data, family =
binomial(link=probit))
print(deviance <- -2*(logLik(null.model) - logLik(fitted.model)))
print(p.value <- pchisq(deviance, df=16, lower.tail =FALSE))
```

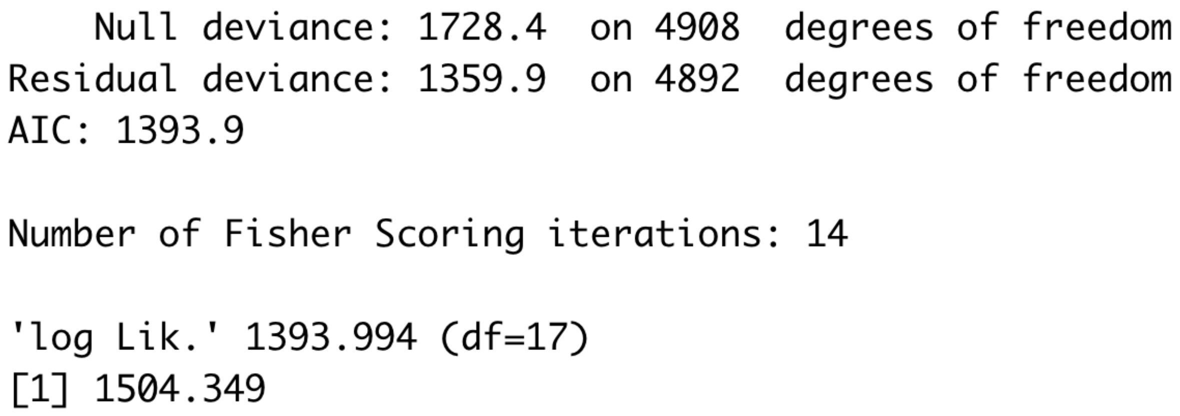
Picture 13. Verifying the model fit.

D. R CODE 1 OUTPUT



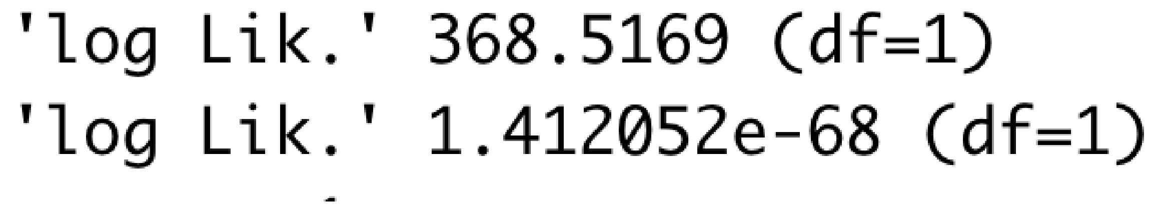
```
(Intercept)                          14.091  < 2e-16 ***
gender.relFemale                     -0.296 0.767208    
gender.relOther                       0.007 0.994616    
age                                 -11.574  < 2e-16 ***
hypertension.rel1                    -3.125 0.001778 ** 
heart_disease.rel1                   -1.762 0.078152 .  
ever_married.relNo                   -0.602 0.547441    
work_type.relchildren                -1.378 0.168287    
work_type.relGovt_job                 0.806 0.420131    
work_type.relNever_worked             0.026 0.979225    
work_type.relSelf-employed            2.128 0.033341 *  
residence_type.relRural               0.100 0.920208    
avg_glucose_level                    -3.388 0.000704 ***
bmi                                  -0.326 0.744260    
smokingstatus.relformerly smoked     -1.055 0.291286    
smokingstatus.relnever smoked        -0.720 0.471410    
smokingstatus.relsmokes              -2.105 0.035257 *  
```

Picture 14: Estimated Probit coefficients output. Significant variables here include: Age, Hypertension answered 1, self-employed work type, average glucose level, and smoking status as “smokes”.



```
    Null deviance: 1728.4  on 4908  degrees of freedom
Residual deviance: 1359.9  on 4892  degrees of freedom
AIC: 1393.9

Number of Fisher Scoring iterations: 14

'log Lik.' 1393.994 (df=17)
[1] 1504.349
```

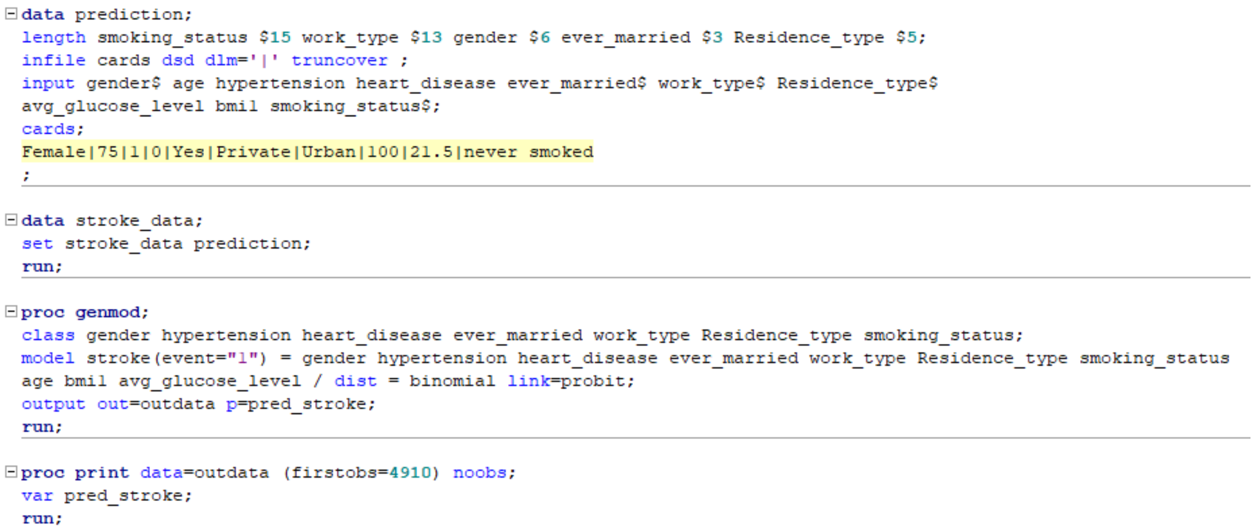
Picture 15: AICC, AIC, and BIC in R.



```
'log Lik.' 368.5169 (df=1)
'log Lik.' 1.412052e-68 (df=1)
```

Picture 16. Verifying model fit. Model is a good fit due to small *p*-value.

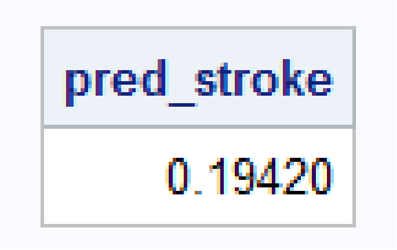
E. SAS CODE PREDICTION



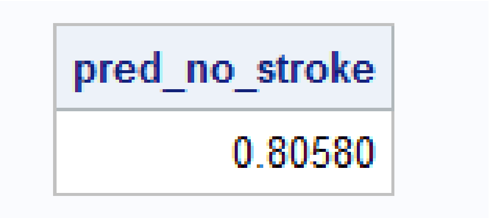
```
data prediction;
 length smoking_status $15 work_type $13 gender $6 ever_married $3 Residence_type $5;
 infile cards dsd dlm='|' truncover ;
 input gender$ age hypertension heart_disease ever_married$ work_type$ Residence_type$
 avg_glucose_level bmil smoking_status$;
 cards;
 Female|75|1|0|Yes|Private|Urban|100|21.5|never smoked
 ;

data stroke_data;
 set stroke_data prediction;
 run;

proc genmod;
 class gender hypertension heart_disease ever_married work_type Residence_type smoking_status;
 model stroke(event="1") = gender hypertension heart_disease ever_married work_type Residence_type smoking_status
 age bmil avg_glucose_level / dist = binomial link=probit;
 output out=outdata p=pred_stroke;
 run;

proc print data=outdata (firstobs=4910) noobs;
 var pred_stroke;
 run;
```

Picture 17: SAS code for predicting the probability of a patient having a stroke based on the predictors in the data set. For our example, we use: Female patient, 75 years old, has hypertension, does not have heart disease, is married, works in the private sector, lives in an urban environment, has an average glucose level of 100, has a bmi of 21.5 and has never smoked.

F. SAS CODE PREDICTION OUTPUT



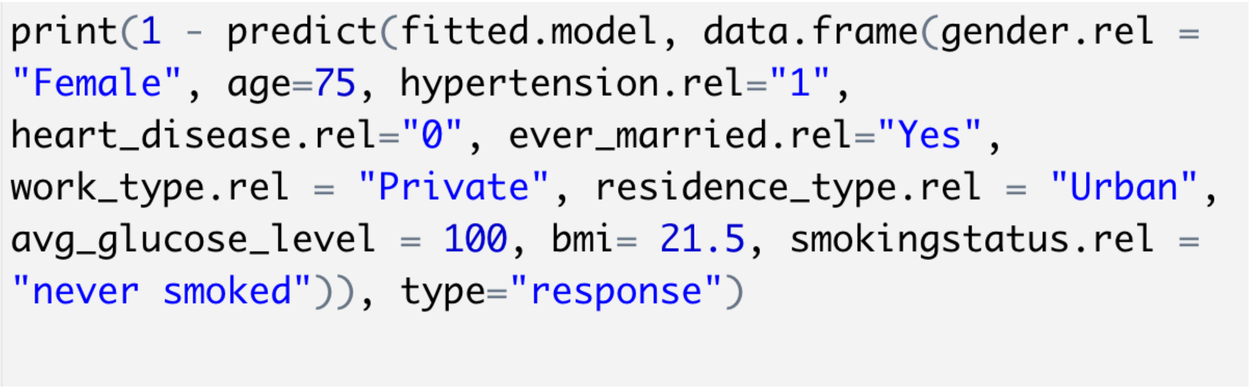
pred_stroke
0.19420



pred_no_stroke
0.80580

Picture 18: On the left, SAS output for the predicted probability of a patient having a stroke. On the right, SAS output for predicted probability of a patient not having a stroke (what was shown in the presentation).

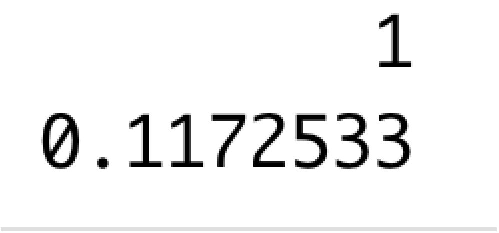
G. R CODE PREDICTION



```
print(1 - predict(fitted.model, data.frame(gender.rel =
"Female", age=75, hypertension.rel="1",
heart_disease.rel="0", ever_married.rel="Yes",
work_type.rel = "Private", residence_type.rel = "Urban",
avg_glucose_level = 100, bmi= 21.5, smokingstatus.rel =
"never smoked")), type="response")
```

Picture 19: Predicting the probability of a patient having a stroke in R.

F. R CODE PREDICTION OUTPUT



```
        1
0.1172533
```

Picture 20: Predicted probability of probit coefficient estimate, is less than SAS estimate by ".08".

VII. REFERENCES

[1] “About Stroke.” *American Stroke Association*, https://www.stroke.org/en/about-stroke. Accessed 30 November 2022.

[2] Singh, Poonam Khetrapal. “World Stroke Day.” *World Health Organization (WHO)*, 28 October 2021, https://www.who.int/southeastasia/news/detail/28-10-2021-world-stroke-day. Accessed 30 November 2022.

[3] “What is Kaggle?” *DataCamp*, https://www.datacamp.com/blog/what-is-kaggle. Accessed 30 November 2022.

[4] Fedesoriano. “Stroke Prediction Dataset.” *Kaggle.com*, 2020, https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset/discussion/232298?datasetId=1120859&sortBy=voteCount&sort=votes. Accessed 30 November 2022.