NONRESPONSE AND MISSINGNESS MECHANISMS

Definition. A data set with missing observations is called incomplete (or unbalanced).

There are three mechanisms for missing observations:

1. Missing Completely at Random (MCAR) – missingness doesn't depend on any characteristics of a person

2. Missing at Random (MAR) – missingness depends only on observed characteristics of a person (that is, nonrespondens are not different from respondents)

3. Missing not at Random (MNAR) or nonignorable missingness – missingness depends on nonobservable characteristics of a person (that is, nonrespondents are systematically different from respondents)

Example. A mail survey on attitudes to racial discrimination got a 45% response rate.

Scenario 1:

- Half of the letters were lost by the post-office (MCAR), but most of the others replied (who did not reply may be considered MAR, because they are not much different from those who did reply).

Scenario 2:

- No letters were lost, but a qualitative study after the survey revealed that many people in the study did not reply because they were hostile to immigrant groups (MNAR).

Example. A mail survey concerning crime victimization got a low response rate.

Scenario 1:

- Nonrespondents are not interested in the topic but otherwise look like the respondents (MAR)

Scenario 2:

- Nonrespondents are victims of a crime themselves and are afraid to respond, or moved out of town and cannot be contacted (MNAR)

HOW TO ANALYZE DATA IN THE PRESENCE OF MISSING VALUES?

1. Complete Case Method – discard all sample elements with missing values on any variables.

Definition. A data set with no missing observations is called complete (or balanced).

2. Model-Based Method – use maximum likelihood estimation procedure which works even for incomplete data. The likelihood function will depend on the assumed underlying distribution and an estimation problem at hand.

3. Reweighting Method – use sampling weights to adjust for missing values. We will come back to this topic in later lectures.

4. Imputation-Based Method – replace (impute) the missing values with some estimates. This procedure is called imputation.

Definition. A data set with imputed values is referred to as imputed data.

IMPUTATION METHODS

Remark 1. Estimates based on imputed data are usually biased, and as a rule the true variance is underestimated.

Remark 2. It makes sense to impute data only if missing values are MCAR or MAR.

Reasonable Imputation Methods:

1. Case-Mean Imputation – a missing value is substituted by an average of similar variables for the same individual, if available.

Example An instructor was grading homeworks #3 when a dog ate one homework off her table. Thus, the score for HW3 for individual #4 was missing. The missingness is MCAR.

Individual	HW1	HW2	HW3	EXAM1	EXAM2	GRADE
1	100	90	100	100	100	A
2	94	95	97	97	94	A
3	100	85	98	98	95	A
4	95	83	.	97	100	A
5	94	84	94	97	95	A
6	91	85	88	91	89	B
7	97	85	84	98	77	B
8	86	72	82	94	94	B
9	86	77	84	95	89	B
10	85	77	86	88	96	B

We can impute the missing value by the mean scores on HW1 and HW2 for this individual, hence we impute by (95+83)/2=89.

Exercise. Show that with this imputation, the overall mean for hw scores for individual #4 doesn't change, that is, show that $\frac{hw1+hw2}{2}=\frac{hw1+hw2+\frac{hw1+hw2}{2}}{3}$

2. Variable-Mean Imputation – an individual with a missing value is matched with the others with complete data based on certain variables of interest (e.g.,

demographic), and then the missing value is substituted by the average of the values of this variable for the matched individuals.

Example In the above example, it may be reasonable to impute the missing score by the average of hw3 scores for all "A" students. That is, we would substitute the missing value by (100+97+98+94)/4=97.25.

3. Case-Mode and Variable-Mode Imputations – when the variable with a missing value is categorical, the missing value is imputed by the case-mode or variable-mode.

Definition. Mode is the most frequent observation

4. Regression Imputation – impute a missing value by a fitted value when the variable with missingness is regressed on all the other variables. Suppose the data consist of variables $x_1, x_2, \ldots, x_p, x_{p+1}$. We fit the model

$$x_{p+1} = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon_p ,$$

and then predict the missing value x_{p+1}^0 from the observed values $x_1^0, x_2^0, \ldots, x_p^0$ for the individual, $x_{p+1}^0 = \hat{\beta}_0 + \hat{\beta}_1 x_1^0 + \ldots + \hat{\beta}_p x_p^0$.

Remark. If the missing values are binary, then a logistic regression is fit. If the estimated probability of 1 is above 0.5, impute by 1, otherwise, by 0.

Example. In our example, we regress HW3 on HW1, HW2, EXAM 1, and EXAM2, and compute the predicted value of HW3 for HW1=95, HW2=83, EXAM1=97, EXAM2=100.

```
data imputation;
input HW1 HW2 HW3 EXAM1 EXAM2;
cards;
100   90    100   100   100
94    95    97    97    94
100   85    98    98    95
94    83    .     97    100
94    84    94    97    95
91    85    88    91    89
97    85    84    98    77
86    72    82    94    94
86    77    84    95    89
85    77    86    88    96
;

proc reg;
model HW3=HW1 HW2 EXAM1 EXAM2/cli;
run;
```

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Predict		Residual
1	100.0000	101.4488	1.1127	96.1187	106.7789	-1.4488
2	97.0000	96.9598	1.3870	91.1552	102.7645	0.0402
3	98.0000	96.4863	1.1371	91.1168	101.8559	1.5137
4	.	**95.4874**	0.9156	90.4548	100.5201	.
5	94.0000	93.0312	0.6387	88.3397	97.7227	0.9688
6	88.0000	88.3395	1.0926	83.0415	93.6375	-0.3395
7	84.0000	84.6603	1.4431	78.7510	90.5696	-0.6603
8	82.0000	83.4437	1.1391	78.0709	88.8165	-1.4437
9	84.0000	82.6498	1.1139	77.3178	87.9818	1.3502
10	86.0000	85.9806	1.2486	80.4233	91.5379	0.0194

Hence, we impute the missing value by 95.4874.

Unreasonable but Commonly-Used Imputation Methods:

1. Hot-Deck Imputation – data are ordered in some way and a missing value is substituted by an observed value of the same variable in the same dataset.

(a) Sequential Hot-Deck Imputation – the missing value is substituted by the previous observed value of the same variable.

Example. In our example, the missing value will be imputed by the value 98.

Individual	HW1	HW2	HW3	EXAM1	EXAM2	GRADE
1	100	90	100	100	100	A
2	94	95	97	97	94	A
3	100	85	**98**	98	95	A
4	95	83	.	97	100	A
5	94	84	94	97	95	A
6	91	85	88	91	89	B
7	97	85	84	98	77	B
8	86	72	82	94	94	B
9	86	77	84	95	89	B
10	85	77	86	88	96	B

(b) Random Hot-Deck Imputation – the missing value is substituted by a randomly chosen observed value of the same variable.

Example In our example, the missing value will be imputed by a randomly chosen value 82. It may be wiser to chose at random a value from among the non-missing values only for A students. Then the missing value will be imputed by, say, 97.

Individual	HW1	HW2	HW3	EXAM1	EXAM2	GRADE
1	100	90	100	100	100	A
2	94	95	97	97	94	A
3	100	85	98	98	95	A
4	95	83	.	97	100	A
5	94	84	94	97	95	A
6	91	85	88	91	89	B
7	97	85	84	98	77	B
8	86	72	**82**	94	94	B
9	86	77	84	95	89	B
10	85	77	86	88	96	B

Hot-deck imputation is widely used by the U.S. Census Bureau.

2. Cold Deck Imputation – imputed values are from a previous survey of the same or similar population.

Example. The instructor taught STAT 108 the previous semester. She finds that a person who got very similar scores on the first two homeworks received 93 for homework 3, so she imputes the missing value by 93.

Note on the name origin: The name *hot-deck* is from the days when computer programs were prepared on punched cards. The deck of cards containing the data set being analyzed was warmed by the card reader, so the term *hot deck* was used to refer to imputations made using the same data set. In *cold-deck* imputation, the imputed values are from another data set not the one running through the computer, so the deck is *cold*.

A word of caution: Both hot-deck and cold-deck imputation procedures are unreasonable in the sense that they may result in very messy data set, with pregnant men, and women with prostate cancer.

3. Multiple Imputation – each missing value is imputed some fixed number of times m ($m > 1$), and then each imputed data set is analyzed separately. Typically, the same imputation method is used each time. The different results give a measure of the additional variance due to the imputation.

This method is applicable when a large number of observations are missing.

Example. In our example, let two observations for hw3 be missing.

Individual	HW1	HW2	HW3	EXAM1	EXAM2	GRADE
1	100	90	100	100	100	A
2	94	95	97	97	94	A
3	100	85	98	98	95	A
4	95	83	.	97	100	A
5	94	84	94	97	95	A
6	91	85	88	91	89	B
7	97	85	84	98	77	B
8	86	72	82	94	94	B
9	86	77	.	95	89	B
10	85	77	86	88	96	B

Suppose we use the random hot-deck method to impute both values. For subject 4, the value may be imputed by 100, 97, 98 or 94. For subject 9, the missing value may be imputed by 88, 84, 82, or 86. There is a total of (4)(4)=16 imputed datasets.

For pure illustrative purposes, suppose we would like to estimate the mean score on hw3 in the population. The 16 imputed data sets are

Imputed Data Sets															
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
97	97	97	97	97	97	97	97	97	97	97	97	97	97	97	97
98	98	98	98	98	98	98	98	98	98	98	98	98	98	98	98
100	**100**	**100**	**100**	**97**	**97**	**97**	**97**	**98**	**98**	**98**	**98**	**94**	**94**	**94**	**94**
94	94	94	94	94	94	94	94	94	94	94	94	94	94	94	94
88	88	88	88	88	88	88	88	88	88	88	88	88	88	88	88
84	84	84	84	84	84	84	84	84	84	84	84	84	84	84	84
82	82	82	82	82	82	82	82	82	82	82	82	82	82	82	82
88	**84**	**82**	**86**	**88**	**84**	**82**	**86**	**88**	**84**	**82**	**86**	**88**	**84**	**82**	**86**
86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86

Suppose we pick at random three of the 16 data sets, that is, $m=3$. Let the chosen data set be 3, 9, and 14.

	3	9	14
	100	100	100
	97	97	97
	98	98	98
	100	**98**	**94**
	94	94	94
	88	88	88
	84	84	84
	82	82	82
	82	**88**	**84**
	86	86	86
Mean	91.1	91.5	90.7
SE	7.4603	6.6207	6.6341

In every imputed data set, the mean $\bar{x}_i, i = 1,\ldots, m$, is different. The overall mean of the *m* realizations of the imputation is

$$\bar{x} = \frac{\bar{x}_1 + \ldots + \bar{x}_m}{m} = \frac{91.1 + 91.5 + 90.7}{3} = 91.1.$$

The estimated variance <u>within</u> the realizations is computed as

$$s_w^2 = \frac{s_1^2 + \ldots + s_m^2}{m} = \frac{(7.4603)^2 + (6.6207)^2 + (6.6341)^2}{3} = 47.8337.$$

The estimated variance <u>between</u> the realizations is found according to the formula

$$s_b^2 = \left(1 + \frac{1}{m}\right)\frac{\sum_{i=1}^{m}(\bar{x}_i - \bar{x})^2}{m-1} = \left(1 + \frac{1}{3}\right)\frac{(91.1-91.1)^2 + (91.5-91.1)^2 + (90.7-91.1)^2}{3-1}$$

$$= \left(\frac{4}{3}\right)(0.16) = 0.2133.$$

The overall variance and standard error of the estimated mean is given by

$$Var(\overline{x}) = s_w^2 + s_b^2 = 47.8337 + 0.2133 = 48.0470,$$

$$SE(\overline{x}) = \sqrt{Var(\overline{x})} = \sqrt{48.0470} = 6.9316.$$

Note that in this example the overall estimated standard error (6.9316) is not much different from the estimated standard errors of individual realizations (7.4603, 6.6207, and 6.6341). It means that multiple imputation is not really necessary in this case (the variability due to imputations is very small compared to the variability within the imputed datasets).