CHAPTER 14: MISSING DATA

A serious problem in data analysis is the existence of missing data. We concentrate on missing responses in a dependent data situation.

Implications of missing data:

1. Data are unbalanced – not a problem given modern regression techniques.
2. Information loss.
3. Depending on the mechanism of missingness, bias in estimation may result.

Missing data can arise in numerous ways, and understanding the mechanism is crucial to appropriate modeling assumptions.

In a longitudinal study, if *drop-out* occurs at a certain time then no additional data are observed after that point.

Examples:

1. In a health-air pollution study an individual may be unavailable for measurement because he/she took a job in another area.
2. In a clinical trial, patients may be removed from the study if their longitudinal measurements are below/above some limit.
3. Censoring – measurement instruments may be inaccurate below a lower limit of detection, this limit is then reported.
4. The value of the outcome may itself determine the missingness, but the outcome is unobserved.

In 1, the missingness will not be a problem unless the person moved area because of health problems. In 2, the missingness will be a function of the responses on previous occasions, while in 3 and 4 it depends on the actual measurement that would have been recorded.

Example: Simulated Data

Data were simulated in which the data ($m = 200, n_i = 10, i = 1, ..., m$) were generated from a linear mixed model in which intercepts and slopes are random (and independent), with measurement error and $\beta_0 = 100, \beta_1 = -5$.

Figure 40 shows the resultant data.

We then simulated drop-out by a mechanism in which if the outcome falls below 65, the subsequent observations are lost (but we retain the initial one below 65).

Figure 41 shows the data that we actually observe (509 data points were lost).

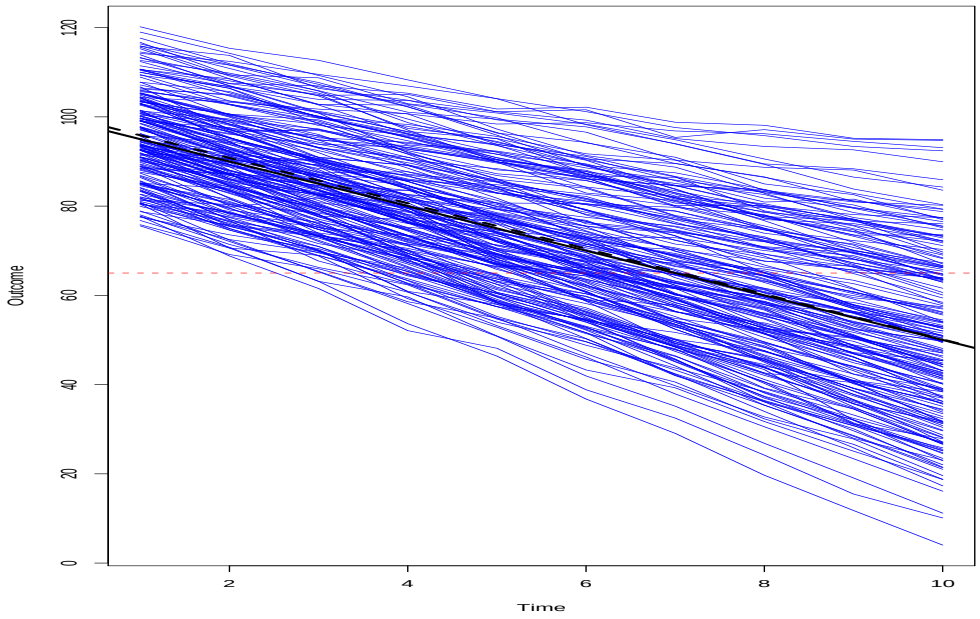

Figure 40: Full simulated data set: solid line is truth and dashed the LS line.

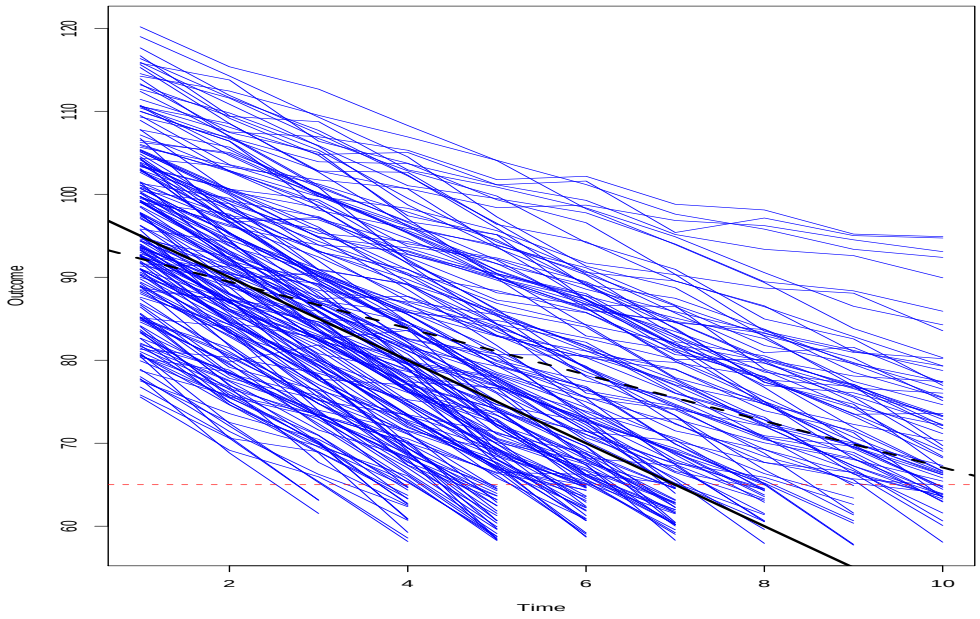

Figure 41: Simulated data set with drop-out: solid line is truth and dashed the LS line.

Mechanisms of Missingness

The impact of missing data depends crucially on the mechanism of missingness, that is the probability model for missingness.

We let $\boldsymbol{R}_i$ be a vector of response indicators for the i-th units so that

$$R_{ij} = \begin{cases} 1 & \text{if } Y_{ij} \text{ is observed} \\ 0 & \text{if } Y_{ij} \text{ is missing} \end{cases}$$

We partition the complete data vector $\boldsymbol{Y}_i = (\boldsymbol{Y}_i^O, \boldsymbol{Y}_i^M)$ into those components that are observed, $\boldsymbol{Y}_i^O$, and those that are missing $\boldsymbol{Y}_i^M$.

There are two ways of factoring the data:

$$\begin{aligned} p(\boldsymbol{Y}, \boldsymbol{R} \mid \boldsymbol{x}) &= p(\boldsymbol{Y} \mid \boldsymbol{x}) \times p(\boldsymbol{R} \mid \boldsymbol{Y}, \boldsymbol{x}) \\ p(\boldsymbol{Y}, \boldsymbol{R} \mid \boldsymbol{x}) &= p(\boldsymbol{Y} \mid \boldsymbol{R}, \boldsymbol{x}) \times p(\boldsymbol{R} \mid \boldsymbol{x}) \end{aligned}$$

The first is known as a selection model (individuals are selected according to their outcome), and the second as a pattern mixture model (we "mix" pattern specific models). We concentrate on the former.

Three situations are distinguished:

1. Missing completely at random (MCAR).
2. Missing at random (MAR).
3. Not missing at random (NMAR).

each of which we now discuss in detail.

Unfortunately the terminology is confusing!

Missing Completely at Random (MCAR)

Data are MCAR if

$$\Pr(R_{ij} = 1 \mid \boldsymbol{Y}^O, \boldsymbol{Y}^M, \boldsymbol{x}) = \Pr(R_{ij} \mid \boldsymbol{x}),$$

so that the missingness does not depend on the response data, observed or unobserved. Can depend on $\boldsymbol{x}$, e.g. design in linear regression.

This implies that

$$\mathrm{E}[Y_{ij}|R_{ij} = 1, \boldsymbol{x}_i] = \mathrm{E}[Y_{ij}|\boldsymbol{x}_i]$$

No selection bias.

Missing at Random (MAR)

Data are MCAR if

$$\Pr(R_{ij} = 1 \mid \boldsymbol{Y}^O, \boldsymbol{Y}^M, \boldsymbol{x}) = \Pr(R_{ij} \mid \boldsymbol{Y}^O, \boldsymbol{x}),$$

so that the missingness may depend on observed values.

This implies that

$$\mathrm{E}[Y_{ij}|R_{ij} = 1, \boldsymbol{x}_i] \neq \mathrm{E}[Y_{ij}|\boldsymbol{x}_i]$$

which suggests that the GEE approach might be in trouble in terms of biased parameter estimates.

Not Missing at Random (NMAR)

If the missingness depends on $\boldsymbol{Y}^M$, i.e.

$$\Pr(R_{ij} = 1 \mid \boldsymbol{Y}^O, \boldsymbol{Y}^M, \boldsymbol{x}) = \Pr(R_{ij} \mid \boldsymbol{Y}^O, \boldsymbol{Y}^M, \boldsymbol{x}).$$

In this case the mechanism is also sometimes referred to as non-ignorable.

This selection bias is not fixable, since we don't know the outcomes that caused the problems. Models can be postulated, but are not checkable from the observed data alone.

In general it is obviously best if we know why the data are missing.

Approaches

Complete-case analysis

A simple approach is to exclude units that did not provide data at all intended occasions. Clearly there is a loss of information in this process, and bias will result unless the data are MCAR. Not to be recommended.

Available-case analysis

This approach uses the largest set of available data for estimating parameters. Will provide biased estimates unless the data are MCAR.

Last observation carried forward

In a longitudinal setting we could simply "fill-in" the missing values, extrapolating from the last observed value. As a general method not to be recommended.

Imputation

An appealing approach is to "fill-in", or impute, the missing values and then carry out a conventional analysis. Complex models for the missingness can be incorporated (closely related to *data augmentation* which we describe later).

Likelihood-based approach

Let $\boldsymbol{\theta}$ be the parameters of the model for $\boldsymbol{Y}$, and $\boldsymbol{\phi}$ the parameters for $\boldsymbol{R}$.

In general, a natural way to decompose the data is

$$\begin{aligned} p(\boldsymbol{Y}^O, \boldsymbol{Y}^M, \boldsymbol{R} \mid \boldsymbol{\theta}, \boldsymbol{\phi}) &= p(\boldsymbol{Y}^O, \boldsymbol{Y}^M \mid \boldsymbol{\theta}, \boldsymbol{\phi}) \times \Pr(\boldsymbol{R} \mid \boldsymbol{Y}^O, \boldsymbol{Y}^M, \boldsymbol{\theta}, \boldsymbol{\phi}) \\ &= p(\boldsymbol{Y}^O, \boldsymbol{Y}^M \mid \boldsymbol{\theta}) \times \Pr(\boldsymbol{R} \mid \boldsymbol{Y}^O, \boldsymbol{Y}^M, \boldsymbol{\phi}) \end{aligned}$$

where we have also assumed that the data and missingness models have distinct parameters.

We require a distribution for the observed data, $\boldsymbol{Y}^O, \boldsymbol{R}$:

$$p(\boldsymbol{Y}^O, \boldsymbol{R} \mid \boldsymbol{\theta}, \boldsymbol{\phi}) = \int p(\boldsymbol{Y}^O, \boldsymbol{Y}^M \mid \boldsymbol{\theta}) \times \Pr(\boldsymbol{R} \mid \boldsymbol{Y}^O, \boldsymbol{Y}^M, \boldsymbol{\phi}) \; d\boldsymbol{Y}^M.$$

This is an example of a selection model.

Suppose we are in the MAR situation so that

$$\Pr(\boldsymbol{R} \mid \boldsymbol{Y}^O, \boldsymbol{Y}^M, \boldsymbol{\phi}) = \Pr(\boldsymbol{R} \mid \boldsymbol{Y}^O, \boldsymbol{\phi}).$$

In this situation the likelihood is given by

$$\begin{aligned} p(\boldsymbol{Y}^O, \boldsymbol{R} \mid \boldsymbol{\theta}, \boldsymbol{\phi}) &= \int p(\boldsymbol{Y}^O, \boldsymbol{Y}^M \mid \boldsymbol{\theta}) \; d\boldsymbol{Y}^M \times \Pr(\boldsymbol{R} \mid \boldsymbol{Y}^O, \boldsymbol{\phi}) \\ &= p(\boldsymbol{Y}^0 \mid \boldsymbol{\theta}) \times \Pr(\boldsymbol{R} \mid \boldsymbol{Y}^O, \boldsymbol{\phi}) \end{aligned}$$

Hence we have the log-likelihood

$$\log p(\boldsymbol{Y}^0 \mid \boldsymbol{\theta}) + \log \Pr(\boldsymbol{R} \mid \boldsymbol{Y}^O, \boldsymbol{\phi})$$

and can ignore the second term and don't have to model the missingness mechanism.

Important Point: We need to get the model right!!!

Models for Drop-out

If the missingness is monotone, in the sense that if $R_{ij} = 0$ then $R_{ik} = 0$ for all $k > j$, then we define the drop-out time as

$$D_i = \min_k \{R_{ik} = 0\}.$$

Hence $2 \leq D_i \leq n_i + 1$, with $D_i = n_i + 1$ for an individual that does not drop out.

The reason for drop-out may be that the individual was not responding well, and their outcomes reflect this.

To examine this possibility we could fit logistic models of the form:

$$\log \left(\frac{\Pr(D_i = k | D_i \geq k, Y_{i1}, ..., Y_{ik})}{\Pr(D_i > k | D_i \geq k, Y_{i1}, ..., Y_{ik})} \right) = \phi_0 + \phi_1 Y_{ik-1}$$

and look for evidence that $\phi_1 \neq 0$.

GEE Approaches

Suppose that if the full data had been observed there would have been n_i observations on each individual, $i = 1, ..., m$.

We write the usual estimating equation as

$$\boldsymbol{G}(\boldsymbol{\beta}) = \sum_{i=1}^{m} \boldsymbol{D}_i^{\mathrm{T}} \boldsymbol{W}_i^{-1} \boldsymbol{R}_i (\boldsymbol{Y}_i - \boldsymbol{\mu}_i)$$

where $\boldsymbol{R}_i$ is the diagonal matrix with elements R_{ij}, $j = 1, ..., n_i$.

For the estimator, $\widehat{\boldsymbol{\beta}}$ to be consistent we require $\boldsymbol{G}$ to be unbiased. The random variables are now $\boldsymbol{Y}, \boldsymbol{R}$ and so we have

$$\begin{aligned}
\mathrm{E}_{Y,R}[\boldsymbol{G}(\boldsymbol{\beta})] &= \mathrm{E}_R\{\mathrm{E}_{Y|R}[\boldsymbol{G}(\boldsymbol{\beta})]\} \\
&= \sum_{i=1}^{m} \mathrm{E}_{R_i}\{\mathrm{E}_{Y_i|R_i}[\boldsymbol{D}_i^{\mathrm{T}} \boldsymbol{W}_i^{-1} \boldsymbol{R}_i (\boldsymbol{Y}_i - \boldsymbol{\mu}_i)]\} \\
&= \sum_{i=1}^{m} \mathrm{E}_{R_i}\{\boldsymbol{D}_i^{\mathrm{T}} \boldsymbol{W}_i^{-1} \boldsymbol{R}_i \mathrm{E}_{Y_i|R_i}[\boldsymbol{Y}_i - \boldsymbol{\mu}_i]\} \\
&= \sum_{i=1}^{m} \mathrm{E}_{R_i}\{\boldsymbol{D}_i^{\mathrm{T}} \boldsymbol{W}_i^{-1} \boldsymbol{R}_i \left(\mathrm{E}_{Y_i|R_i}[\boldsymbol{Y}_i] - \boldsymbol{\mu}_i\right)\}
\end{aligned}$$

Hence, to obtain an unbiased estimating equation we require

$$\mathrm{E}[\boldsymbol{Y}_i \mid \boldsymbol{R}_i, \boldsymbol{x}_i] = \mathrm{E}[\boldsymbol{Y}_i \mid \boldsymbol{x}_i] = \boldsymbol{\mu}_i$$

so that we are fine under MCAR but not under MAR, since the distribution of $\boldsymbol{Y}_i \mid \boldsymbol{x}_i, \boldsymbol{R}_i$ is different from that of $\boldsymbol{Y}_i \mid \boldsymbol{x}_i$ under MAR.

To rectify the situation we need to modify the usual estimating equation.

Let the probability of non-dropout given history be given by

$$\pi_{ij} = \mathrm{E}[R_{ij} \mid \boldsymbol{x}_i, \boldsymbol{H}_{i,j-1}]$$

where $\boldsymbol{H}_{i,j-1} = (Y_{i1}, ..., Y_{i,j-1})$ contains the "history" of responses.

Consider the estimating equation:

$$\sum_{i=1}^{m} \boldsymbol{D}_i^{\mathrm{T}} \boldsymbol{W}_i^{-1} \boldsymbol{P}_i (\boldsymbol{Y}_i - \boldsymbol{\mu}_i)$$

where $\boldsymbol{P}_i$ is a diagonal matrix which contains terms R_{ij}/π_{ij}, for $j = 1, ..., n_i$.
We have

$$\mathrm{E}_Y \left\{ \sum_{i=1}^{m} \mathrm{E}_{R|Y} \left[\boldsymbol{D}_i^{\mathrm{T}} \boldsymbol{W}_i^{-1} \boldsymbol{P}_i (\boldsymbol{Y}_i - \boldsymbol{\mu}_i) \right] \right\} = \mathrm{E}_Y \left\{ \sum_{i=1}^{m} \boldsymbol{D}_i^{\mathrm{T}} \boldsymbol{W}_i^{-1} \mathrm{E}_{R|Y}[\boldsymbol{P}_i] (\boldsymbol{Y}_i - \boldsymbol{\mu}_i) \right\} = 0$$

since $\mathrm{E}[\boldsymbol{P}_i] = \boldsymbol{I}$ if π_{ij} is correctly specified.

In both GEE and likelihood we are basically accounting for the biased sampling scheme of MAR; likelihood does this by assuming a model, while GEE adjusts by modeling the probabilities of seeing the data.

Simulation Study: True Model is random intercepts and slopes

Model	$\widehat{\beta}_1$	(s.e.)	$\widehat{\beta}_2$	(s.e.)
Full GEE ind	101.0	0.715	−5.084	0.109
Full GEE exch	101.0	0.715	−5.084	0.109
Full LMEM1	101.0	0.981	−5.084	0.037
Full LMEM2	101.0	0.720	−5.084	0.109
MAR GEE ind	95.0	0.894	−2.796	0.134
MAR GEED exch	98.8	0.787	−4.304	0.114
MAR LMEM1	98.8	0.837	−4.282	0.041
MAR LMEM2	100.1	0.722	−5.097	0.112

Table 15: Results of GEE and LMEM analyses of full and drop-out simulated data, LMEM1 is random intercepts only, LMEM2 is random intercepts and slopes. First four rows: full data, last four rows: reduced dataset under MAR.

- Bias for GEE is bad (particularly working independence).
- Bias for LMEM if we only assume random intercepts (terrible se on $\widehat{\beta}_2$).
- LMEM with random intercepts and slopes recovers the truth.