CHAPTER 12: MISSING DATA

A serious problem in data analysis is the existence of missing data. We concentrate on missing responses in a dependent data situation. Missing and unbalanced data are very different, the latter might be by design and so will not lead to bias.

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. If data on air quality are being collected, there will be missing observations on days on which a pollution monitor was faulty.
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.

In 1, the missingness will not depend on either the measurement on the pollution either on the missing day, or any previous days. In 2, the missingness will be a function of the responses on previous occasions, while in 3 it depends on the actual measurement that would have been recorded.

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 $\boldsymbol{Y}_i^M$.

We distinguish three situations:

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.

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.

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.

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.

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 as

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

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

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

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

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

$$p(\boldsymbol{Y}^O, \boldsymbol{R} \mid \boldsymbol{\theta}, \boldsymbol{\phi}) = \int p(\boldsymbol{Y}^O, \boldsymbol{Y}^M \mid \boldsymbol{\theta}) \, d\boldsymbol{Y}^M \Pr(\boldsymbol{R} \mid \boldsymbol{Y}^O, \boldsymbol{\phi}) = p(\boldsymbol{Y}^0 \mid \boldsymbol{\theta}) \Pr(\boldsymbol{R} \mid \boldsymbol{Y}^O, \boldsymbol{\phi})$$

Hence we have the log-likelihood $\log p(\boldsymbol{Y}^0 \boldsymbol{\theta}) + \log \Pr(\boldsymbol{R} \mid \boldsymbol{Y}^O, \boldsymbol{\phi})$ and can ignore the second term and don't have to model the missingness mechanism.

Bayesian Inference via Data Augmentation

Data augmentation is a auxiliary variable method that treats the missing observations as unknown parameters – this can lead to simple MCMC schemes.

General formulation: we have posterior

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

MCMC scheme:

1. Auxiliary variables:
$$\boldsymbol{Y}^M \sim p(\boldsymbol{Y}^M \mid \boldsymbol{Y}^O, \boldsymbol{\theta}).$$
2. Model parameters:
$$\boldsymbol{\theta} \sim p(\boldsymbol{\theta} \mid \boldsymbol{Y}^O, \boldsymbol{Y}^M).$$

The auxiliary variable scheme may be modified to $p(\boldsymbol{Y}^M \mid \boldsymbol{Y}^O, \boldsymbol{\theta}) \sim p(\boldsymbol{Y}^M \mid \boldsymbol{\theta})$, depending on the missing data model, as we now illustrate.

Example: Censoring Model

Suppose we have data Y_i measured at times t_i, $j = 1, ..., n$, but measurements *below the lower limit of detection*, D (assumed known) are not recorded. Also suppose that the data generating model (likelihood) is:

$$Y \mid \boldsymbol{\beta}, \sigma \sim_{ind} \mathrm{N}(\eta(\boldsymbol{\beta}, t), \sigma^2).$$

Clearly setting such measurements to zero or ignoring the measurements will lead to bias in estimation.

Figure 47 illustrates for a set of simulated data in which the true slope was -0.01; the slope estimates are -0.0099, -0.0095 and -0.0087 for the full data, set equal to D and ignored schemes, respectively.

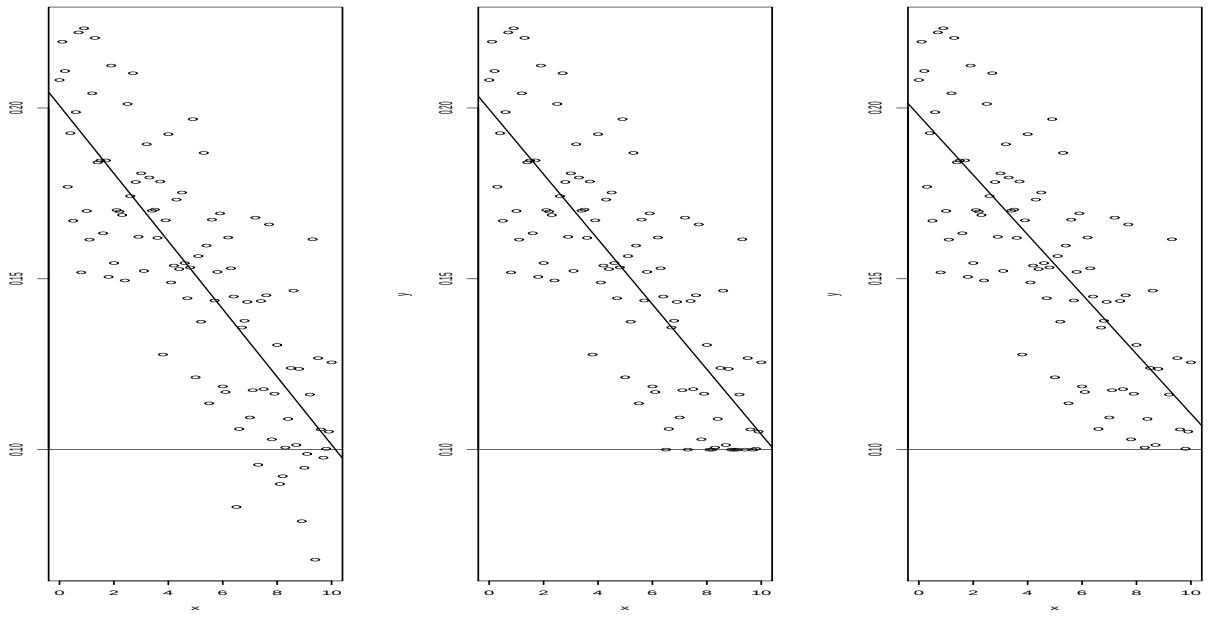

Figure 47: All data (left), assigned to lower limit (middle), ignored (right). Horizontal line is the lower limit of detection.

Suppose that the last c measurements are censored, the remaining $n-c$ being uncensored. Then

$$\begin{aligned} p(\boldsymbol{y} \mid \theta) &= \prod_{i=1}^{n-c} p(y_i \mid \boldsymbol{\beta}, \sigma^2) \prod_{i=c+1}^{n} \Pr(Y_i < D \mid \boldsymbol{\beta}, \sigma^2) \\ &= \prod_{i=1}^{n-c} \phi\left(\frac{y_i - \eta(\boldsymbol{\beta}, t_i)}{\sigma}\right) \prod_{i=c+1}^{n} \Phi\left(\frac{D - \eta(\boldsymbol{\beta}, t_i)}{\sigma}\right) \end{aligned}$$

where

$$\phi(z) = (2\pi)^{-1/2} \exp(-z^2/2)$$

and

$$\Phi(z_0) = \Pr(Z < z_0) = \int_{-\infty}^{z_0} \phi(z)\, dz$$

where Z is an $\mathrm{N}(0,1)$ random variable.

To perform likelihood or Bayesian inference we need to numerically evaluate the distribution function of a normal distribution for each likelihood calculation.

Data Augmentation Scheme

Letting $\boldsymbol{Y}^O = \{Y_i, i = 1, ..., n-c\}$ and $\boldsymbol{Y}^M = \{Y_i, i = n-c+1, ..., n\}$, we iterate between

1. $y_i \mid \boldsymbol{\beta}, \sigma \sim \text{TruncNorm}(\eta(\boldsymbol{\beta}, t_i), \sigma^2)$, on $(-\infty, D)$, $i = n-c+1, ..., n$.
2. $\boldsymbol{\beta} \mid y_1, ..., y_n, \sigma^2 \propto \prod_{i=1}^n p(y_i \mid \boldsymbol{\beta}, \sigma^2)\pi(\boldsymbol{\beta})$. Usual (uncensored) posterior.
3. $\sigma^2 \mid y_1, ..., y_n, \boldsymbol{\beta} \propto \prod_{i=1}^n p(y_i \mid \boldsymbol{\beta}, \sigma^2)\pi(\sigma^2)$. Usual (uncensored) posterior.

Related Example: Survival Analysis with Censored Data

From the `WinBUGS` manual: *Mice: Weibull regression*

Dellaportas and Smith (1993) analyse data from Grieve (1987) on photocarcinogenicity in four groups, each containing 20 mice, who have recorded a survival time and whether they died or were censored at that time.

A portion of the data, giving survival times in weeks, are shown below.

A * indicates censoring.

Mouse	Irradia control	Vehicle control	Test substan	Positive control
1	12	32	22	27
.....				
18	*40	30	24	12
19	31	37	37	17
20	36	27	29	26

The survival distribution is assumed to be Weibull. That is

$$p(t_i, z_i) = r \exp(\boldsymbol{\beta} \boldsymbol{z}_i) t_i^{r-1} \exp\{-\exp(\boldsymbol{\beta} \boldsymbol{z}_i t_i r)\}$$

where t_i is the failure time of an individual with covariate vector $\boldsymbol{z}_i$ and $\boldsymbol{\beta}$ is a vector of unknown regression coefficients. The baseline hazard is given by

$$\lambda_0(t_i) = r t_i^{r-1}.$$

Setting $\mu_i = \exp(\boldsymbol{\beta} \boldsymbol{z}_i)$ gives the parameterization

$$t_i \sim \text{Weibull}(r, \mu_i)$$

For censored observations, the survival distribution is a truncated Weibull, with lower bound corresponding to the censoring time. The regression coefficients $\boldsymbol{\beta}$ are assumed a priori to follow independent Normal distributions with zero mean and "vague" precision 0.0001. The shape parameter r for the survival distribution was given a Gamma(1, 0.0001) prior, which is slowly decreasing on the positive real line.

Median survival for individuals with covariate vector $\boldsymbol{z}_i$ is given by

$$m_i = (\log 2 \exp(-\boldsymbol{\beta} \boldsymbol{z}_i))^{1/r}.$$

WinBUGS code

```
model
{
        for(i in 1 : M) {
                for(j in 1 : N) {
                        t[i, j] ~ dweib(r, mu[i])I(t.cen[i, j],)
                }
                mu[i] <- exp(beta[i])
                beta[i] ~ dnorm(0.0, 0.001)
                median[i] <- pow(log(2) * exp(-beta[i]), 1/r)
        }
        r ~ dexp(0.001)
        veh.control <- beta[2] - beta[1]
        test.sub <- beta[3] - beta[1]
        pos.control <- beta[4] - beta[1]
}
```

```
list(t = structure(.Data = c(12, 1, 21, 25, 11, 26, 27, 30, 13, 12,
21, 20, 23, 25, 23, 29, 35, NA, 31, 36, 32, 27, 23, 12, 18, NA, NA,
38, 29, 30, NA, 32, NA, NA, NA, NA, 25, 30, 37, 27, 22, 26, NA, 28,
19, 15, 12, 35, 35, 10, 22, 18, NA, 12, NA, NA, 31, 24, 37, 29, 27,
18, 22, 13, 18, 29, 28, NA, 16, 22, 26, 19, NA, NA, 17, 28, 26, 12,
17, 26), .Dim = c(4, 20)), t.cen = structure(.Data = c( 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 40, 0, 0, 0, 0, 0, 0, 0, 40, 40,
0, 0, 0, 40, 0, 40, 40, 40, 40, 0, 0, 0, 0, 0, 0, 10, 0, 0, 0, 0, 0,
0, 0, 0, 0, 24, 0, 40, 40, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 20, 0, 0,
0, 0, 29, 10, 0, 0, 0, 0, 0, 0), .Dim = c(4, 20)), M = 4, N = 20)
```

We note a number of tricks in setting up this model.

First, individuals who are censored are given a missing value in the vector of failure times t, whilst individuals who fail are given a zero in the censoring time vector t.cen.

The truncated Weibull is modelled using I(t.cen[i],) to set a lower bound. Second, we set a parameter beta[j] for each treatment group j. The contrasts beta[j] with group 1 (the irradiated control) are calculated at the end. Alternatively, we could have included a grand mean term in the relative risk model and constrained beta[1] to be zero.

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
median[1]	23.9	1.967	0.05889	20.3	23.89	28.09	1001	10000
median[2]	35.2	3.359	0.04757	29.46	34.93	42.64	1001	10000
median[3]	26.9	2.383	0.0582	22.62	26.79	31.91	1001	10000
median[4]	21.4	1.799	0.03362	18.2	21.32	25.36	1001	10000
pos.control	0.3409	.3457	0.00723	-0.327	0.3429	1.009	1001	10000
r	3.03	0.3182	0.02749	2.388	3.045	3.64	1001	10000
test.sub	-0.351	0.3459	0.004433	-1.035	-0.3541	0.3303	1001	10000
veh.control	-1.16	0.3679	0.005974	-1.893	-1.156	-0.444	1001	10000

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

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

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.

Final Comment

Dependent data are complex and difficult to analyze, but don't be afraid to apply different techniques.

Each of likelihood, Bayes and GEE have strengths and weaknesses, but can often be used in a complementary fashion.

Care is required in interpretation of parameters, however.

The End!