

# Missing Data in Longitudinal Studies

## Outline

- Introduction
- Missing Mechanism
- Missing Data Pattern
- Simple Solutions and Their Limitations
- Testing for MCAR
- Weighted Estimating Equations
- Data Example

## Introduction

- “*Missing data*” refers to data values that were *intended* to be collected but were not available for some reason. In contrast to “*unbalanced*” studies by design, the reason of the missingness (the “missing data mechanism”) should be considered in the analysis.
- The mechanism (“at random” or otherwise) by which data is lost is critical to understanding the impact of missing data on the conclusions and supporting modeling and inference.
- Examples of missing data:
  - A non-longitudinal example. A hormone level in blood samples were tested for 10 patients in the treatment group and another 10 from the control group. However, 2 values from the treatment group were missing. Why? The tubes were dropped or the two missing values were below the lowest detectable level?
  - In HIV clinical trials an often used outcome variable is serum HIV-RNA level. The higher this level the sicker the person tends to be and therefore less likely to continue to report in.
- Consequences of missing data:
  - Technical difficulty associated with unbalanced data.
  - Missing data usually results in loss of information. Some naive methods for dealing missing data, such as complete case analysis, result in even more severe efficiency loss.
  - Missing data can introduce potentially very serious bias.

## Missing Value Mechanisms

### Notation

- Assume that for each the  $i = 1, \dots, m$  units, outcomes  $Y_{ij}$  and covariates  $\mathbf{X}_{ij}$  are taken at times  $j = 1, \dots, n$ .
- Let  $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{in})^T$  be the complete data outcome vector, which may not be fully observed.
- Partition  $\mathbf{Y}_i = (\mathbf{Y}_i^{(o)}, \mathbf{Y}_i^{(m)})$ , where  $\mathbf{Y}_i^{(o)}$  denotes the observed data and  $\mathbf{Y}_i^{(m)}$  the missing data.
- Let  $\mathbf{R}_i = (R_{i1}, \dots, R_{in})^T$  index missingness, i.e.,  $R_{ij} = 1$  if  $Y_{ij} \in \mathbf{Y}^{(o)}$  is observed and  $R_{ij} = 0$  if  $Y_{ij} \in \mathbf{Y}^{(m)}$  (not observed).

### Missing Completely at Random

The outcomes are said to be *missing completely at random* (MCAR) if

$$\Pr(\mathbf{R}_i | \mathbf{Y}_i, \mathbf{X}_i, \psi) = \Pr(\mathbf{R}_i | \mathbf{X}_i, \psi); \quad (1)$$

i.e., the missingness is conditionally independent of the outcome given the covariates.

### Missing at Random

The outcomes are said to be *missing at random* (MAR) if

$$\Pr(\mathbf{R}_i | \mathbf{Y}_i, \mathbf{X}_i, \psi) = \Pr(\mathbf{R}_i | \mathbf{Y}_i^{(o)}, \mathbf{X}_i, \psi); \quad (2)$$

i.e., the missingness is conditionally independent of the unobserved outcomes, given the covariates and observed outcomes.

### Non-Ignorable

The missing data mechanism is *non-ignorable* (NI) or informative, if  $\mathbf{R}_i$  depends on  $\mathbf{Y}_i^{(m)}$ .

## General Comments

We will discuss missing data in more detail later. Some general comments:

- MAR is less restrictive than MCAR.
- For MCAR, both complete case analysis (using only units with complete observations) and all observation analysis using least squares are valid. They are not generally valid for MAR.
- For MAR, all observation analysis based on the likelihood is valid (no need to specify a model for the missing mechanism).
- For NI, a model for missing mechanism (and perhaps external information) is needed.

## Likelihood Model with Missing Data

- Let  $\boldsymbol{\theta}$  denote the parameter for the outcome  $Y$  and  $\boldsymbol{\psi}$  the parameter for the missingness mechanism.
- For likelihood based analysis, the observed data likelihood is the joint density of  $(\mathbf{Y}^{(o)}, \mathbf{R})$  which can be written as:

$$\begin{aligned} f(\mathbf{y}^{(o)}, \mathbf{r} \mid \boldsymbol{\theta}, \boldsymbol{\psi}) &= \int f(\mathbf{y}^{(o)}, \mathbf{y}^{(m)}, \mathbf{r} \mid \boldsymbol{\theta}, \boldsymbol{\psi}) d\mathbf{y}^{(m)} \\ &= \int f(\mathbf{y}^{(o)}, \mathbf{y}^{(m)} \mid \boldsymbol{\theta}) f(\mathbf{r} \mid \mathbf{y}^{(o)}, \mathbf{y}^{(m)}, \boldsymbol{\psi}) d\mathbf{y}^{(m)} \end{aligned}$$

If  $\mathbf{R}$  does not depend on  $\mathbf{Y}^{(m)}$  (MCAR or MAR), then

$$\begin{aligned} f(\mathbf{y}^{(o)}, \mathbf{r} \mid \boldsymbol{\theta}, \boldsymbol{\psi}) &= f(\mathbf{r} \mid \mathbf{y}^{(o)}, \boldsymbol{\psi}) \int f(\mathbf{y}^{(o)}, \mathbf{y}^{(m)} \mid \boldsymbol{\theta}) d\mathbf{y}^{(m)} \\ &= f(\mathbf{r} \mid \mathbf{y}^{(o)}, \boldsymbol{\psi}) f(\mathbf{y}^{(o)} \mid \boldsymbol{\theta}). \end{aligned}$$

- If  $\boldsymbol{\psi}$  and  $\boldsymbol{\theta}$  are *separable*, i.e.,

$$\Omega(\boldsymbol{\theta}, \boldsymbol{\psi}) = \Omega(\boldsymbol{\theta}) \times \Omega(\boldsymbol{\psi}),$$

inference can be based solely on  $f(\mathbf{y}^{(o)} \mid \boldsymbol{\theta})$ .

- Note that for likelihood based methods, we need correctly specify the model for  $\mathbf{Y}^{(o)}$ , that is  $f(\mathbf{y}^{(o)} \mid \boldsymbol{\theta})$ .
- Implicitly we are assuming the marginal distribution of  $\mathbf{Y}^{(o)}$  is of interest, that is, for example, the effect of the treatment if there is no dropout.
- There can be situations where the conditional distribution of  $\mathbf{Y}^{(o)} \mid \mathbf{R} = 1$  is of interest. For example, the effect of a cancer treatment regime on the quality of life, given the patient survives.

## Missing Data Patterns

We say a **dropout** occurs if whenever  $Y_j$  is missing, so are  $Y_k$  for all  $k \geq j$ ; otherwise we say the missing values are **intermittent**.

- **Intermittent** missing can arise due to
  - a known censoring mechanism, for example, all values below a known threshold are missing.
  - a reason unrelated to the outcome, for example, a missed clinic appointment.
  - The reason for the missing is often known because the subjects with missing values are still in the study and hence the reason of missing values can be ascertained.
- **Dropouts** (attrition, lost of follow-up) are frequently lost to follow-up then we cannot be certain that the dropout is or is not related to the observed or unobserved outcome.
  - Dropout indicator: where did the dropout occur (next to last observed visit).

$$D_i = 1 + \sum_{j=1}^{n_i} R_{ij}.$$

- Dropout is an example of *monotone* missing data, meaning that once  $Y_{ik}$  is lost, all observations after time  $k$  are also lost. Some methods are designed specifically for monotone missing data.

## Simple Solutions and Their Limitations

### Last observation carried forward (LOCF)

Imputing the last observed value for the subject to the rest of their sequence:

$$Y_k = Y_{D_i-1} \text{ for all } k \geq D_i.$$

- Routinely used in the pharmaceutical industry in the analysis of randomized trials.
- If patients are expected to improve over time, LOCF should result in a conservative estimate of treatment benefits.
- In general, LOCF method is not recommended.

### Complete case analysis

Discard all incomplete sequences (i.e. use only subjects with complete sequence).

- Waste data is the dropout process is unrelated to the response process.
- When the two processes are related, this method may introduce bias. If we can make assumptions on the relationship between the two processes, then using the incomplete data can improve the efficiency of estimation.
- Only recommended when the scientific questions of interest are confined to the sub-population of completers.

## Testing for MCAR

### Test Setup

- Let  $t_j$ ,  $j = 1, \dots, n$  be intended times of measurement on each case. We observed  $\mathbf{y}_i = (y_{i1}, \dots, y_{in_i})^T$ ,  $n_i \leq n$ , where  $y_{ij}$  is observed at time  $t_{ij}$ .
- The null hypothesis is that the probability of a subject drops out at time  $t_{ij}$  is independent of  $y_{i1}, \dots, y_{i,j-1}$ .
- Note that the covariates (e.g. treatment group) can make the dropout process appear to be informative (confounders), even if it were completely random within each treatment group. We should test the above hypothesis using homogeneous subgroups (test within each treatment group).
- Method (Diggle, 1989):
  1. Apply separate tests at each time point within each treatment group.
  2. Analyze the resulting sample of p-values for departure from the uniform distribution on  $(0,1)$ .

## Test Statistic

- For each  $k = 1, \dots, n - 1$  (where a dropout occurs), choose  $h_k(y_1, \dots, y_k)$  to be the “score” of the responses up to that time.
- Let  $R_k$  be the number of subjects with  $n_i \geq k$ , that is, the number of subjects still under observation at time  $t_k$ ,
- Let  $r_k$  be the number of subjects with  $n_i = k$ , that is the number of subjects that are about to drop out at time  $t_k$ .
- Under  $H_0$ , the scores  $h_{ik} = h_k(y_{i1}, \dots, y_{ik})$  for the  $r_k$  dropouts should be a random sample from the complete set of  $R_k$  scores at time  $t_k$ .
- Use the test statistic,

$$\bar{h}_k = \frac{1}{r_k} \sum_{i:n_i=k} h_{ik},$$

which is the mean of the  $r_k$  scores. Under the null hypothesis, the large sample distribution of  $\bar{h}_{ik}$  is normal, with

$$\text{mean: } \bar{H}_k = \frac{1}{R_k} \sum_{i=1}^{R_k} h_{ik},$$

variance:  $S_k^2(R_k - r_k)/(r_k R_k)$ , where

$$S_k^2 = \frac{1}{R_k - 1} \sum_{i=1}^{R_k} (h_{ik} - \bar{H}_k)^2.$$

- When some of the  $r_k$  and  $R_k$  are small, the large sample distribution may be a poor approximation, we may use the complete randomization distribution to do the **exact test**, or **Monte Carlo test** (permuting the missingness indicator and the observed values).
  
- Repeat this for all  $k = 1, \dots, (n - 1)$  to get a set of p-values. The  $p$ -values for different  $k$ 's are independent and have a uniform distribution under the null hypothesis.
  - Informal graphical analyses: plotting the empirical distribution for each treatment group.
  - Formal test of departure from uniformity: Kolmogorov-Smirnov statistics (Diggle, 1989).
  
- A parametric variation is to use a logistic regression of the missingness indicator (i.e. missing or not) on  $h_{ik}$  among subjects who have not dropped out.

**Choice of  $h_k(\cdot)$** 

The aim is to choose the function so that extreme values of the scores constitute evidence against MCAR dropouts. An example is a weighted averages

$$h_k(i) = h_k(y_{i1}, \dots, y_{ik}) = \sum_{j=1}^k w_j y_{ij}, \text{ where } \sum_{j=1}^k w_j = 1$$

Choice of weights,  $w_j$ , reflect analysts' knowledge or judgment about the extent to which the past influences dropouts.

1. Dropouts influenced immediately by an abnormally higher/lower measurement:

$$h_k(i) = h_k(y_{i1}, \dots, y_{ik}) = y_{ik}$$

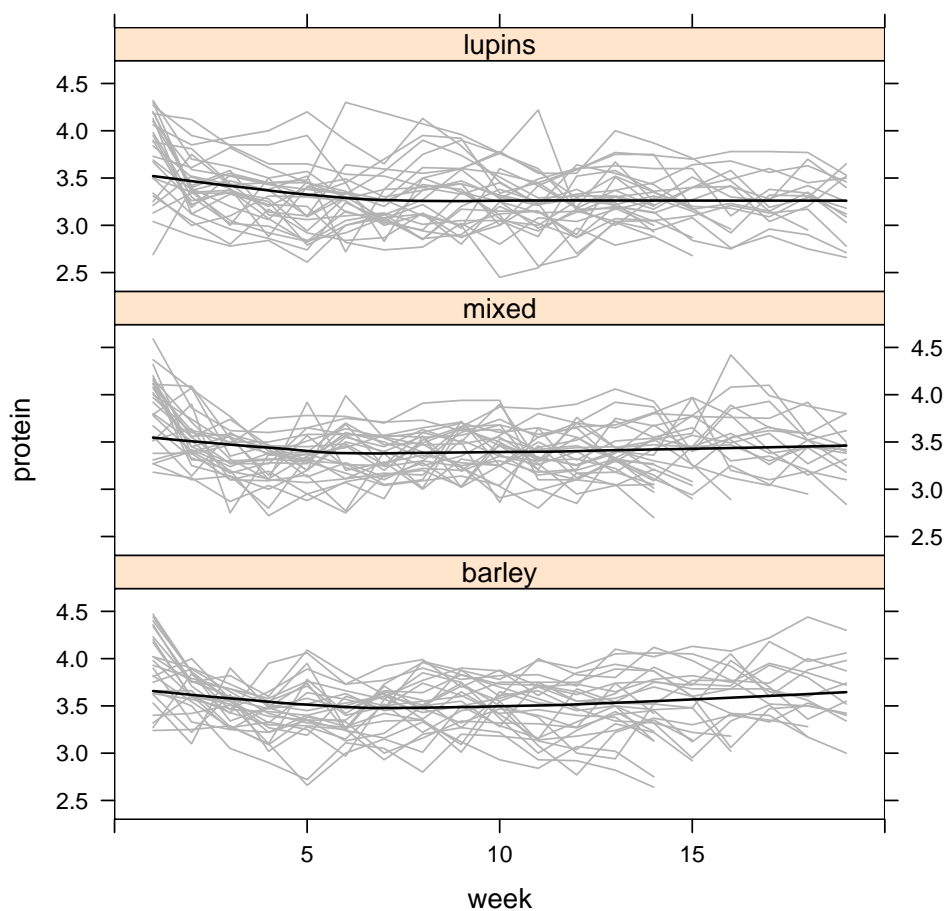
2. Dropouts influenced by a sustained sequence of higher/lower measurements

$$h_k(i) = h_k(y_{i1}, \dots, y_{ik}) = \frac{1}{k} \sum_{j=1}^k y_{ij}$$

## Example: Dropouts in Milk Protein Data

- Milk was collected weekly from 79 cows and analyzed for its protein content.
- Three diets (completely randomized): (1) barley, (2) a mixture of barley and lupins, (3) lupins.
- Goal: to determine how diet affects the protein in milk.
- Staggered entry; time is measured in weeks since calving; Study was terminated at week 19 after the earliest calving.
- 11 intermittent missing values.
- 38 “dropouts” at weeks 15, 16, 17, and 19.
- Is the observed rise in the mean response near the end of the experiment connected to the dropout process?
- Test that dropouts are completely random using Monte Carlo test with score

$$h_k(y_1, \dots, y_k) = y_k$$



Dropouts and completers by week and diet

Dropout time	Diet		
	Barley	Mixed	Lupins
Week 15	6	7	7
Week 16	2	3	4
Week 17	2	1	1
Week 19	2	2	1
Completers	13	14	14
Total	25	27	27

p-values of tests for completely random dropouts

Dropout time	Diet		
	Barley	Mixed	Lupins
Week 15	0.001	0.001	0.012
Week 16	0.016	0.001	0.011
Week 17	0.022	0.053	0.254
Week 19	0.032	0.133	0.206

- The p-values range from 0.001 to 0.254. The hypothesis of completely random dropouts can be rejected.
- Dropouts predominate in cows whose protein measurement in preceding weeks are below average

## Comments

- The greatest distinction is between the two “ignorable” missing mechanisms and the non-ignorable or informative missing. With informative missing/censoring, bias is typically unavoidable.
- To model non-ignorable missing, external information/assumptions are needed. Those assumptions are not verifiable based on the available data.
- It is difficult to verify if the missingness is MAR versus NI.
- The methods to test the null hypothesis of MCAR are not very useful in practice where the “default” should be MAR which is a much weaker assumption and there are methods that can handle MAR.
- If the missing value is MCAR, it is often easier to analyze, in particular, ordinary GLS or GEE methods will work.
- Ordinary GEE methods will result in biased estimates if MAR because

$$E(\mathbf{Y}\mathbf{R}) \neq E(\mathbf{Y})E(\mathbf{R}).$$

## Weighted Estimating Equations

- WEE or weighted GEE is an extension of GEE to deal with monotone missing data (i.e., dropouts).
- GEE based on complete data

$$\mathbf{S}_{\boldsymbol{\beta}}^*(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^m \left( \frac{\partial \boldsymbol{\mu}_i^*}{\partial \boldsymbol{\beta}} \right)^T \text{Var}(\mathbf{Y}_i^*)^{-1} (\mathbf{Y}_i^* - \boldsymbol{\mu}_i^*) = 0$$

$E[\mathbf{Y}_i^* - \boldsymbol{\mu}_i^*] = 0 \Rightarrow \mathbf{S}_{\boldsymbol{\beta}}^*$  is an unbiased estimating function.

- GEE based on observed data

$$\mathbf{S}_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^m \left( \frac{\partial \boldsymbol{\mu}_i^*}{\partial \boldsymbol{\beta}} \right)^T \text{Var}(\mathbf{Y}_i^*)^{-1} \Delta_i (\mathbf{Y}_i^* - \boldsymbol{\mu}_i^*)$$

where  $\Delta_i = \text{diag}(R_{i1}, \dots, R_{in})$ .

- If dropouts are completely at random, then  $\Delta_i$  is uncorrelated with  $\mathbf{Y}_i^*$  and

$$E[\Delta_i (\mathbf{Y}_i^* - \boldsymbol{\mu}_i^*)] = 0$$

$\mathbf{S}_{\boldsymbol{\beta}}$  is an unbiased estimating function.

- If dropouts are MAR instead, then

$$E[\Delta_i (\mathbf{Y}_i^* - \boldsymbol{\mu}_i^*)] \neq 0$$

$\mathbf{S}_{\boldsymbol{\beta}}$  is not unbiased which leads to inconsistent estimates of  $\boldsymbol{\beta}$ .

- Define

$$p_{ij} \equiv \Pr(R_{ij} = 1 \mid \mathbf{Y}_i^*),$$

the probability that subject  $i$  has not dropped out by time  $t_j$ , given the subject's response vector.

- To restore the unbiasedness of  $\mathbf{S}_\beta$  for the complete population we need to weight the contribution of  $y_{ij}$  by the inverse of  $p_{ij}$ . This leads to the weighted estimation equation (WEE)

$$\mathbf{S}_\beta(\beta, \alpha) = \sum_{i=1}^m \left( \frac{\partial \boldsymbol{\mu}_i^*}{\partial \beta} \right)^T \text{Var}(\mathbf{Y}_i^*)^{-1} P_i^{-1} \Delta_i (\mathbf{Y}_i^* - \boldsymbol{\mu}_i^*)$$

where  $P_i = \text{diag}(p_{i1}, \dots, p_{in})$ . Now we have

$$\mathbb{E}[p_{ij}^{-1} R_{ij} (y_{ij}^* - \mu_{ij}^*)] = 0.$$

- The extended GEE requires that the dropout probability  $p_{ij}$  can be *consistently* estimated.

## Therapy for Schizophrenia Trial

- A randomized trial of 523 patients in six treatment groups: placebo, haloperidol 20 mg and risperidone at doses levels 2, 6, 10 and 16 mg.
- The primary outcome is Positive and Negative Symptom Rating Scale (PANSS), a measure of psychiatric disorder. (the smaller the better).
- The design requires the score be taken at weeks -1, 0, 1, 2, 4, 6, and 8.
- Only 253 patients have complete observations. The reasons for dropouts are

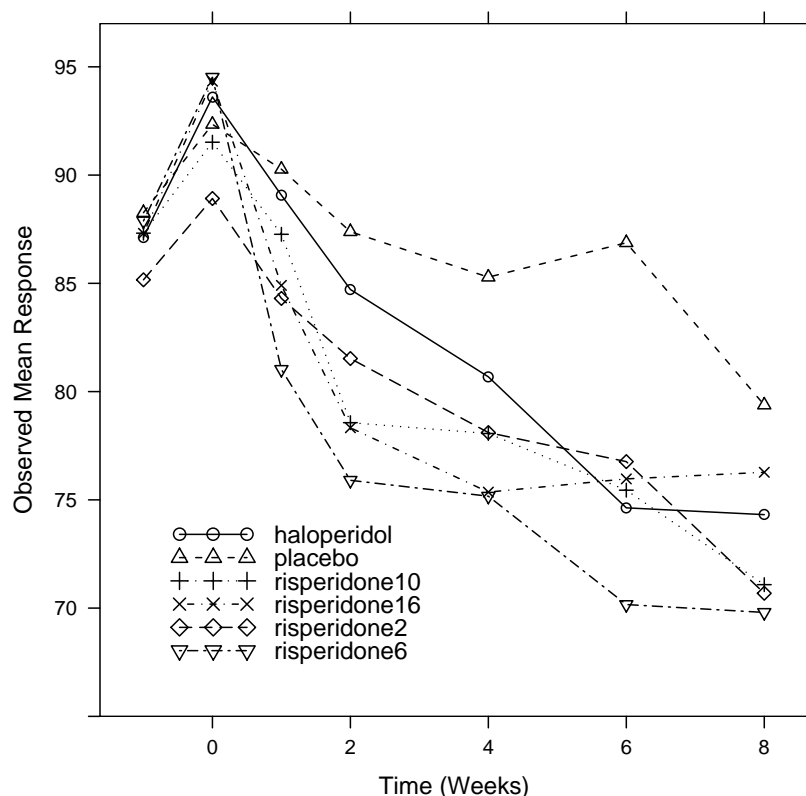
Inadequate response	183
Adverse experience	26
Uncooperative	25
Withdrew consent	19
Other reason	7
Abnormal lab result	4
Inter-current illness	3
Lost to follow-up	3

- Numbers of dropouts and completers by treatment group

Treatment	p	h	r2	r6	r10	r16	Total
Dropouts	61	51	51	34	39	34	270
Completers	27	36	36	52	48	54	253
Total	88	87	87	86	87	88	523

- “Inadequate response” is the most common reason for dropout and there is a higher proportion of dropouts in the placebo group.

## Observed mean response profile – by treatment groups



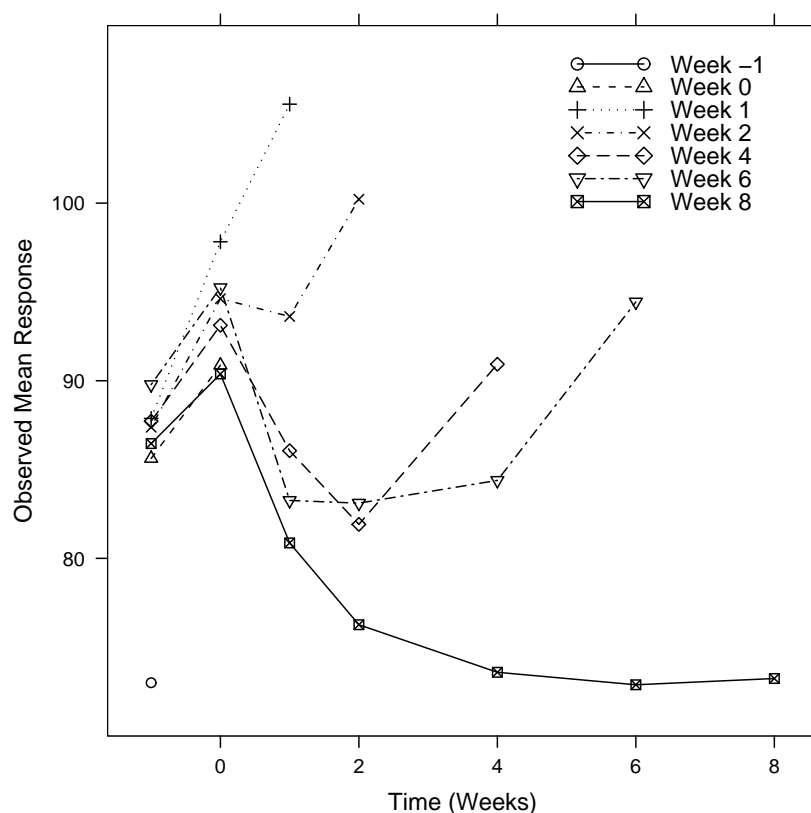
```

> panss <- read.table("../data/panss.data")
> names(panss) <- c("group", paste("panss", 0:6, sep = "."))
> gnames <- c("haloperidol", "placebo",
+           "risperidone10", "risperidone16", "risperidone2",
+           "risperidone6")
> panss$group <- factor(panss$group, labels = gnames)
> otime <- c(-1, 0, 1, 2, 4, 6, 8)

> temp <- by(panss[,2:8], panss$group, mean, na.rm = TRUE)
> group.m <- data.frame(group = rep(gnames, each = length(otime)),
+                       time = rep(otime, length(gnames)),
+                       mean = as.vector(do.call("cbind", temp)))
> xyplot(mean ~ time, group = group, data = group.m,
+        type = "o", col = 1, lty = 1:6, pch = 1:6,
+        ylim = c(65, 97), ylab = "Observed Mean Response",
+        xlab = "Time (Weeks)",
+        key = list(coner = c(0, 0), x = 0.2, y = 0.35,
+        lines = list(lty = 1:6, pch = 1:6, type = "o"),
+        text = list(gnames)))

```

## Observed mean response profile— by different dropout times



```

> last.observed <- function (x, time = NULL) {
+   if (is.null (time)) {
+     time <- 1:length (x)
+   }
+   dropout <- sapply (1:length(x),
+     function (i, y) all(y[i:length(y)]), is.na (x))
+   ifelse (any (dropout), time[which(dropout)[1]-1], time[length(x)])
+ }
> panss$last.observed <- apply (panss[,2:8], 1, last.observed, otime)
> temp <- by (panss[,2:8], panss$last.observed, mean, na.rm = TRUE)
> temp <- do.call ("cbind", temp)
> temp[is.nan(temp)] <- NA
> obs.m <- data.frame (group = rep (otime, each = 7),
+   time = rep (otime, 7),
+   mean = as.vector (temp))
> xyplot (mean ~ time, group = group, data = obs.m,
+   type = "o", col = 1, lty = 1:7, pch = 1:7,
+   ylim = c(70, 110), ylab = "Observed Mean Response",
+   xlab = "Time (Weeks)",
+   key = list (coner = c(0, 0), x = 0.65, y = 0.9,
+   lines = list (lty = 1:7, pch = 1:7, type = "o"),
+   text = list (paste ("Week", otime), divide = 2))

```

## Explore dropout mechanism

We model the probability of dropout (or not dropout) as a function of the measured response (a selection model).

- We fit a logistic regression with the most recent measurement as an explanatory variable:

$$\text{logit}(p_{ij}) = \beta_0 + \beta_1 y_{i,j-1}.$$

$\hat{\beta}_1 = 0.031$  ( $p \ll 0.05$ ) confirms that high responders are likely to drop out. Therefore, we reject completely random dropout (CRD) in favor of random dropout (RD).

- Within the RD framework, we consider two extensions:

$$\text{logit}(p_{ij}) = \beta_0 + \beta_1 y_{i,j-1} + \beta_2 y_{i,j-2},$$

and

$$\text{logit}(p_{ij}) = \beta_{0k} + \beta_1 y_{i,j-1} + \beta_2 y_{i,j-2},$$

where  $k = k(i)$  denotes the treatment for the  $i$ th subject.

Both extensions yield a significant improvement.

$\text{logit}(p_{ij})$	Log-likelihood
$\beta_0 + \beta_1 y_{i,j-1}$	-20743.85
$\beta_0 + \beta_1 y_{i,j-1} + \beta_2 y_{i,j-2}$	-20728.51
$\beta_{0k} + \beta_1 y_{i,j-1} + \beta_2 y_{i,j-2}$	-20724.73

- Can we model the relationship between  $p_{ij}$  (observed) and  $y_{ij}$  (the measurement which would have been observed had the subject not dropped out)?

- When the dropout process is informative, the analyst is presented with a dilemma:
  - doing nothing and accept biased estimates without knowing the extent of bias;
  - trying to model the dropout process (more work), which involves making untestable assumptions, to get still potentially biased estimates.
- Do sensitivity analysis for informative dropout models to assess
  - perhaps some possible causes of informative dropout are more likely than others.
  - how sensitive the results are to the different assumption of informative dropouts (and not modeling dropouts).

## Further Reading

- Chapter 13 of DHLZ.