Missing Data in Longitudinal Studies

Outline

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

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

Missing Completely at Random

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

$$\Pr(\boldsymbol{R}_i \mid \boldsymbol{Y}_i, X_i) = \Pr(\boldsymbol{R}_i \mid X_i); \tag{1}$$

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

$$\Pr(\boldsymbol{R}_i \mid \boldsymbol{Y}_i, X_i) = \Pr(\boldsymbol{R}_i \mid \boldsymbol{Y}_i^{(o)}, X_i); \tag{2}$$

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 $\boldsymbol{R}_i$ depends on $\boldsymbol{Y}_i^{(m)}$.

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 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 a likelihood-based analysis, the observed data likelihood is the joint density of $(\boldsymbol{Y}^{(o)}, \boldsymbol{R})$ which can be written as:

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

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

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

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

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

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

- Note that for likelihood-based methods, we need correctly specify the model for $\boldsymbol{Y}^{(o)}$, that is $f(\boldsymbol{y}^{(o)}; \boldsymbol{\theta})$.

- Implicitly we are assuming the marginal distribution of $\boldsymbol{Y}^{(o)}$ 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 $\boldsymbol{Y}^{(o)} \mid \boldsymbol{R} = \mathbf{1}$ is of interest. For example, the effect of a cancer treatment regime on the quality of life, given the patient survives.

- When the likelihood is the basis for inference, both MCAR and MAR are sometimes referred to without distinction as ”ignorable”.

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.
  - A trial protocol which specifies a set of circumstances under which a patient must be withdrawn from the trial on the basis of their observed measurement history defines a random dropout mechanism.

– when there is any kind of relationship between the measurement process and the dropout process, the interpretation of apparently simple trends in the mean response over time can be problematic.

A simulated data example

∗ Dropouts occurred at random, according to the following logistic regression model

$$\text{logit}(p_t) = -1 - 2y_{t-1}.$$

∗ Mean response was constant over time.

∗ Pair-wise correlation within a subject was $\rho = 0.9$ (Figure 13.1(a)) or $\rho = 0$ (Figure 13.1(b)).

Figure 13.1 (a): $\rho = 0.9$

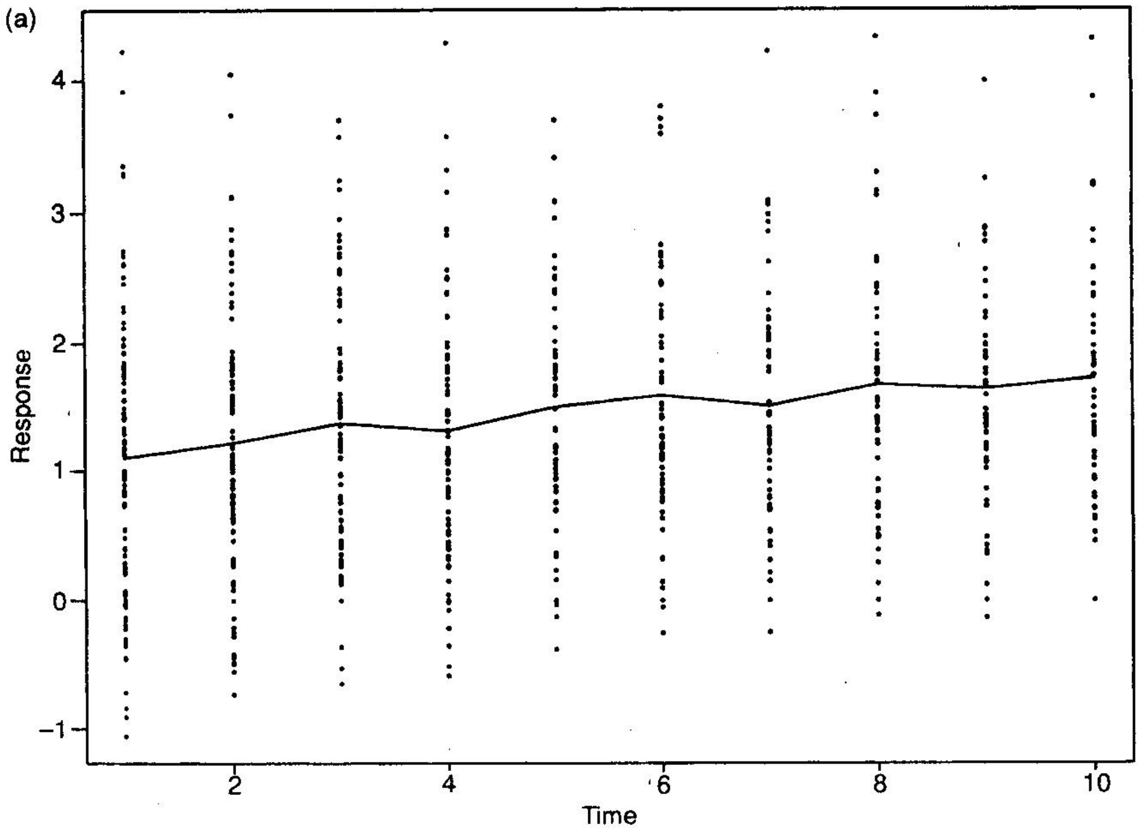
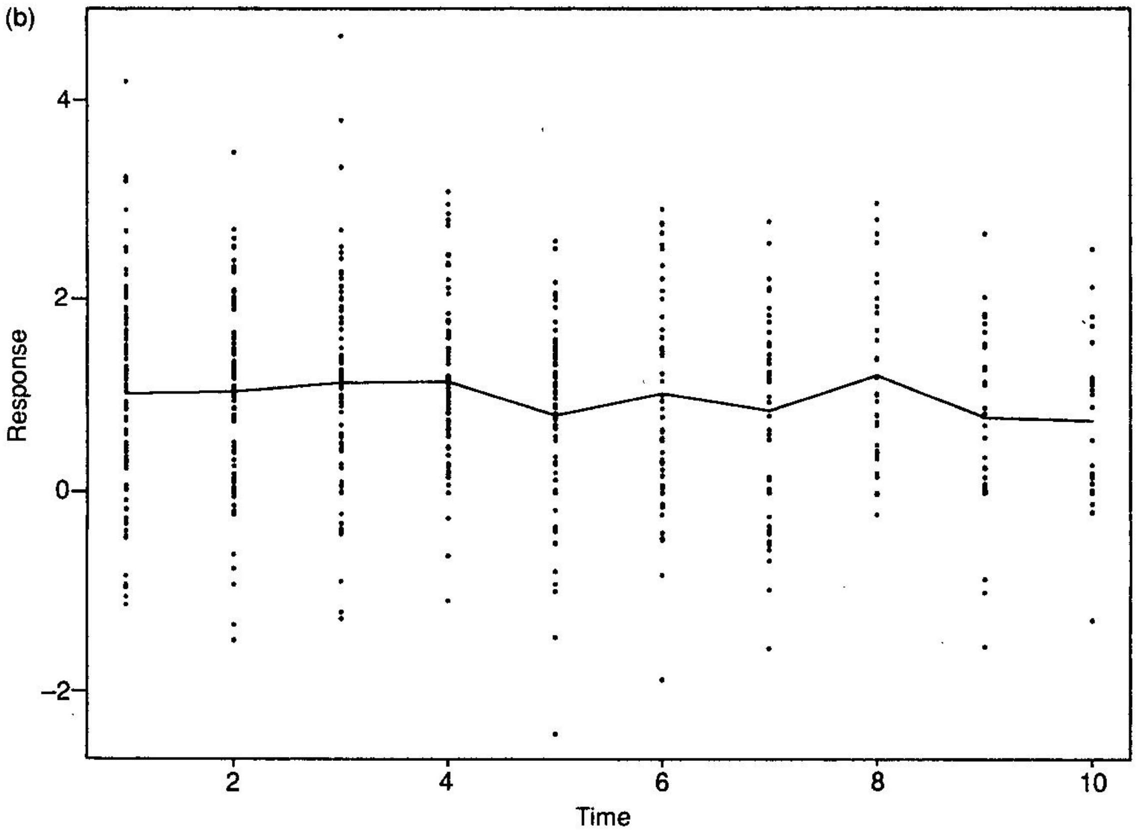


Figure 13.1 (b): $\rho = 0$



Simple Solutions and Their Limitations

Complete case analysis

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

- Waste data if 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.

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.

Testing for MCAR

Test Setup

- Let t_j, $j = 1, \ldots, n$ be intended times of measurement on each case. We observed $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i})^T$, $n_i \leq n$, 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 $y_{i1}, \ldots, y_{i,j-1}$.

- 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 $k = 1, \ldots, n-1$ (where a dropout could potentially occur), choose $h_k(y_1, \ldots, y_k)$ 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 $h_{ik} = h_k(y_{i1}, \ldots, y_{ik})$ for the r_k dropouts should be a random sample from the complete set of R_k scores at time t_k.

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

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

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

1. Dropouts influenced immediately by an abnormally high/low measurement:

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

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

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

- 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

$$\begin{aligned} \text{mean: } \bar{H}_k &= \frac{1}{R_k} \sum_{i=1}^{R_k} h_{ik}, \\ \text{variance: } S_k^2 &(R_k - r_k)/(r_k R_k), \text{ where} \\ S_k^2 &= \frac{1}{R_k - 1} \sum_{i=1}^{R_k} (h_{ik} - \bar{H}_k)^2. \end{aligned}$$

- 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 $k = 1, \cdots, (n-1)$ 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.
  - The implicit assumption that the separate p-values are mutually independent is valid because once a subject drops out he never returns (that is the definition of dropout!).
  - Informal graphical analyses: plotting the empirical distribution of the p-values for each treatment group.
  - Formal test of departure from uniformity: Kolmogorov-Smirnov statistics (Diggle, 1989).
    - One-sided Kolmogorov-Smirnov statistic, $D_+ = \sup\{\hat{F}(p) - p\}$, where $\hat{F}()$ is the empirical cumulative distribution function of the p-values.
    - Then use Barnard's Monte Carlo testing idea to rank the observed value of D_+ among simulated values based on discrete uniform distributions under the null hypothesis.

- 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.

  If use p_k to denote the probability that a subject will drop out at t_k, we assume

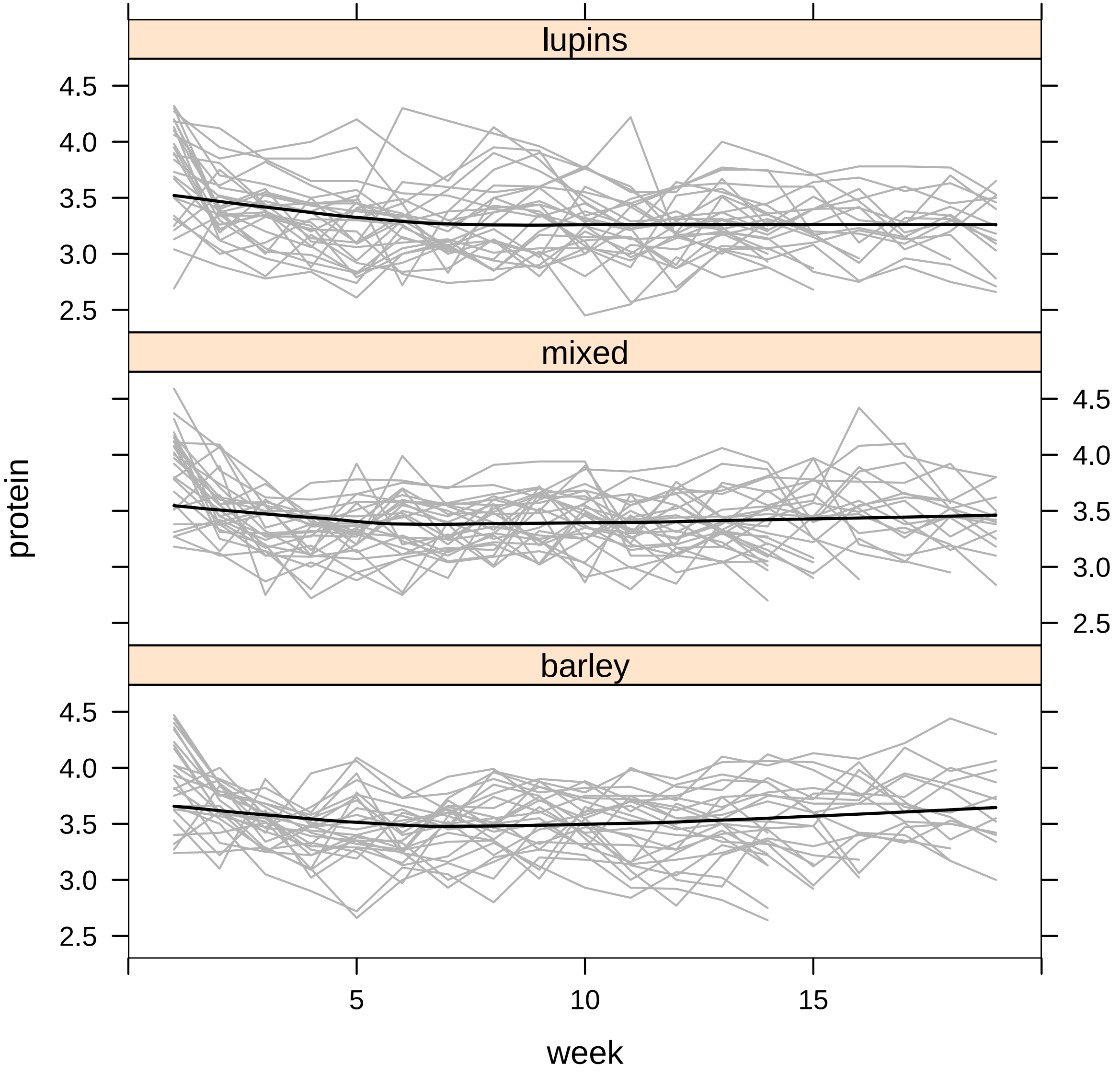
  $$\log\{p_k/(1-p_k) = \alpha + \beta h_k\}.$$

  Test the hypothesis that $\beta = 0$.

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

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


lupins
mixed
barley
protein
week
4.5
4.0
3.5
3.0
2.5
5
10
15

Dropouts and completers by week and diet

	Diet		
Dropout time	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 by using $h_k(y_1, \ldots, y_k) = y_k$

	Diet		
Dropout time	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 firmly by even the "eye-ball" test.
- A Kolmogorov-Smirnov test gives $p = 0.001$.

- We can also use the parametric way.

Let p_{gk} denote the probability that a subject in the gth treatment group will drop out right after the kth time point. We assume that

$$\text{logit}(p_{gk}) = \alpha_{gk} + \beta_{gk} h_k,$$

where $h_k(y_1, \ldots, y_k) = y_k$.

Table 13.2

Model for logit(p_{gk})	Residual deviance	df
1. $\alpha_{gk} + \beta_{gk} h_k$	111.97	210
2. $\alpha_{gk} + \beta_g h_k$	116.33	219
3. $\alpha_{gk} + \beta_k h_k$	118.63	218
4. $\alpha_{gk} + \beta h_k$	119.32	221
5. α_{gk}	197.66	222
6. $\alpha_g + \alpha_k + \beta h_k$	124.16	227
7. $\alpha_g + \beta h_k$	131.28	230
8. $\alpha + \beta h_k$	139.04	232

- Dropouts predominate in cows whose protein measurement in preceding weeks are below average.
- The parametric analysis allows a more detailed description of departures from the MCAR.
- Limitation! The alternative hypothesis (for both the non-parametric and parametric approaches) includes MAR. Hence, rejection of MCAR does not determine the subsequent analysis strategy for the data.

Comments

- The methods to test the null hypothesis of MCAR are not very useful in practice where the alternative is MAR.
- The greatest distinction is between the "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.
- If the missing mechanism is MCAR, it is often easier to analyze, in particular, any likelihood-based methods or the GEE method will work.
- The ordinary form of the GEE method assumes that the dropouts are MCAR, otherwise the consistency is lost.

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

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

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

- GEE based on observed data

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

  where $\Delta_i = \mathrm{diag}(R_{i1}, \ldots, R_{in})$.

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

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

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

- If dropouts are MAR instead, then

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

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

- Define

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

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

- To restore the unbiasedness of $\boldsymbol{S}_{\boldsymbol{\beta}}$ 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)

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

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

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

- The extended GEE requires that the dropout probability $(1 - p_{ij})$ 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 does 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

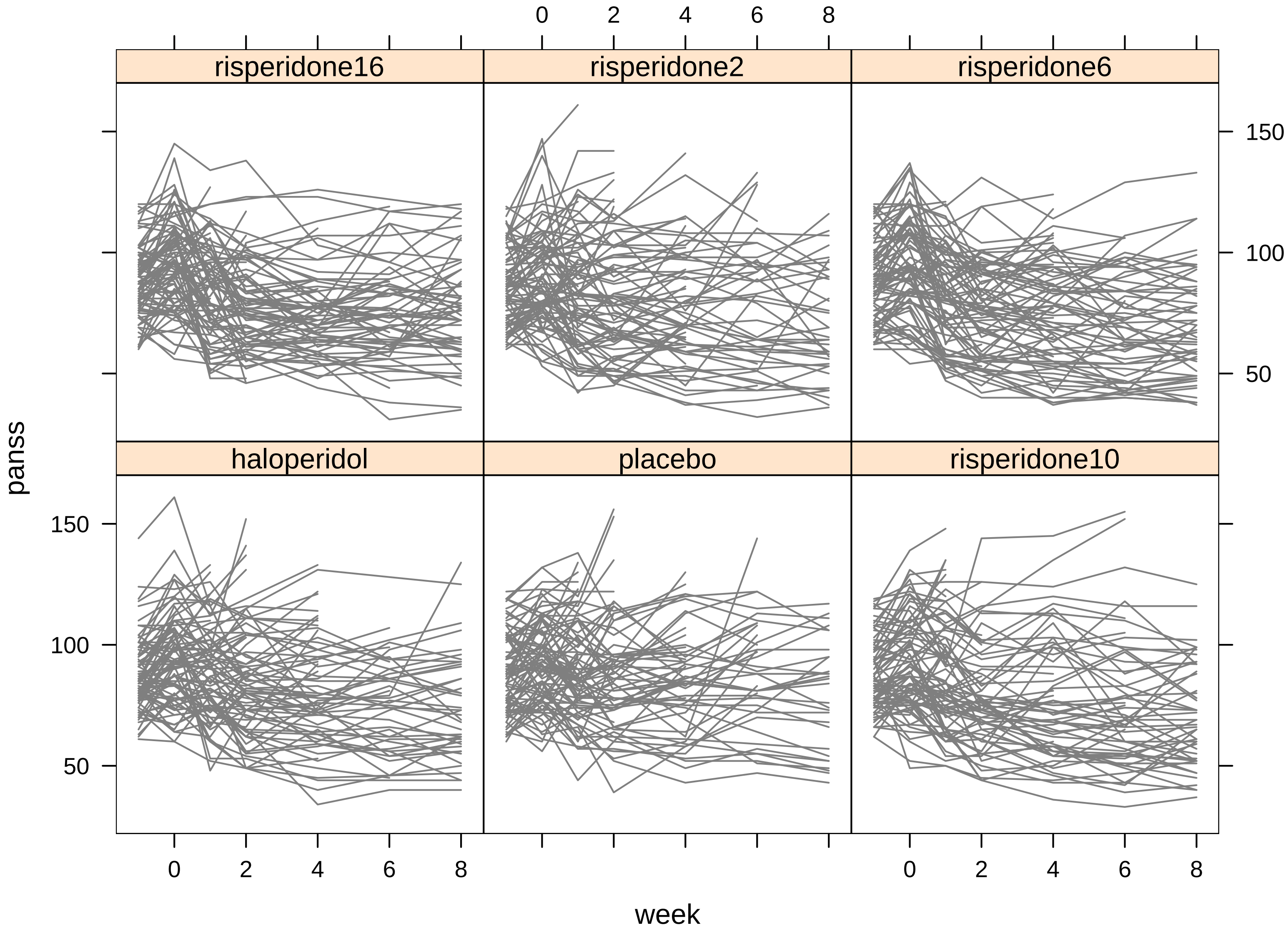
- Numbers of dropouts and completes by treatment group

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.

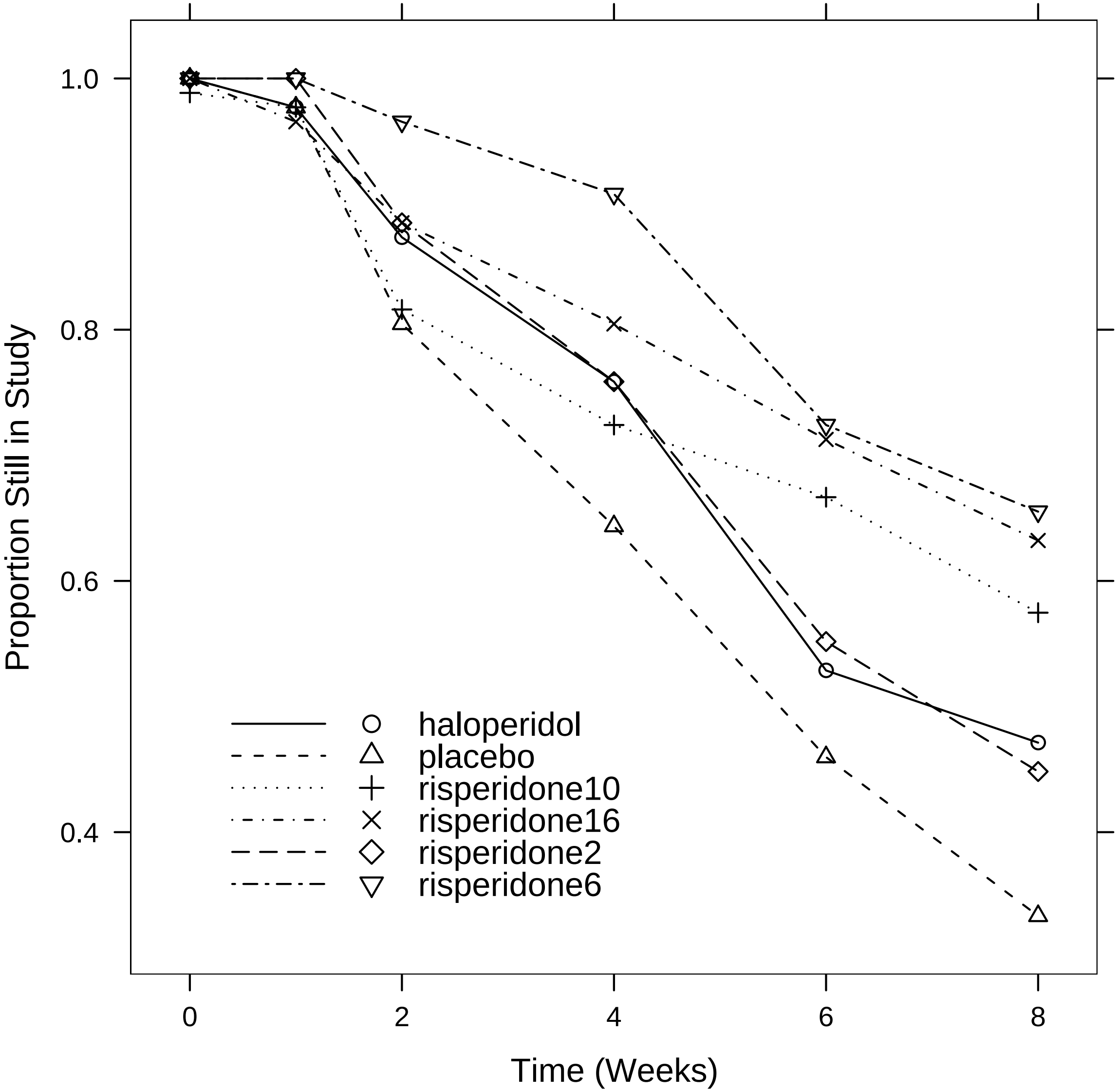
Raw data – by treatment groups

```
> panssL <- reshape (panss[,1:8], direction = "long",
+                    idvar = "id", timevar = "week", times = otime,
+                    varying = names (panss)[2:8])
> panssL$week <- rep (otime, each = nrow (panss))
> panssL <- panssL[order (panssL$group, panssL$id, panssL$week),]
> xyplot (panss ~ week | group, group = id, type = "l",
+         data = panssL,
+         col = "gray50")
```


risperidone16
risperidone2
risperidone6
haloperidol
placebo
risperidone10
panss
week
0
2
4
6
8
50
100
150

Dropout rate – by treatment groups

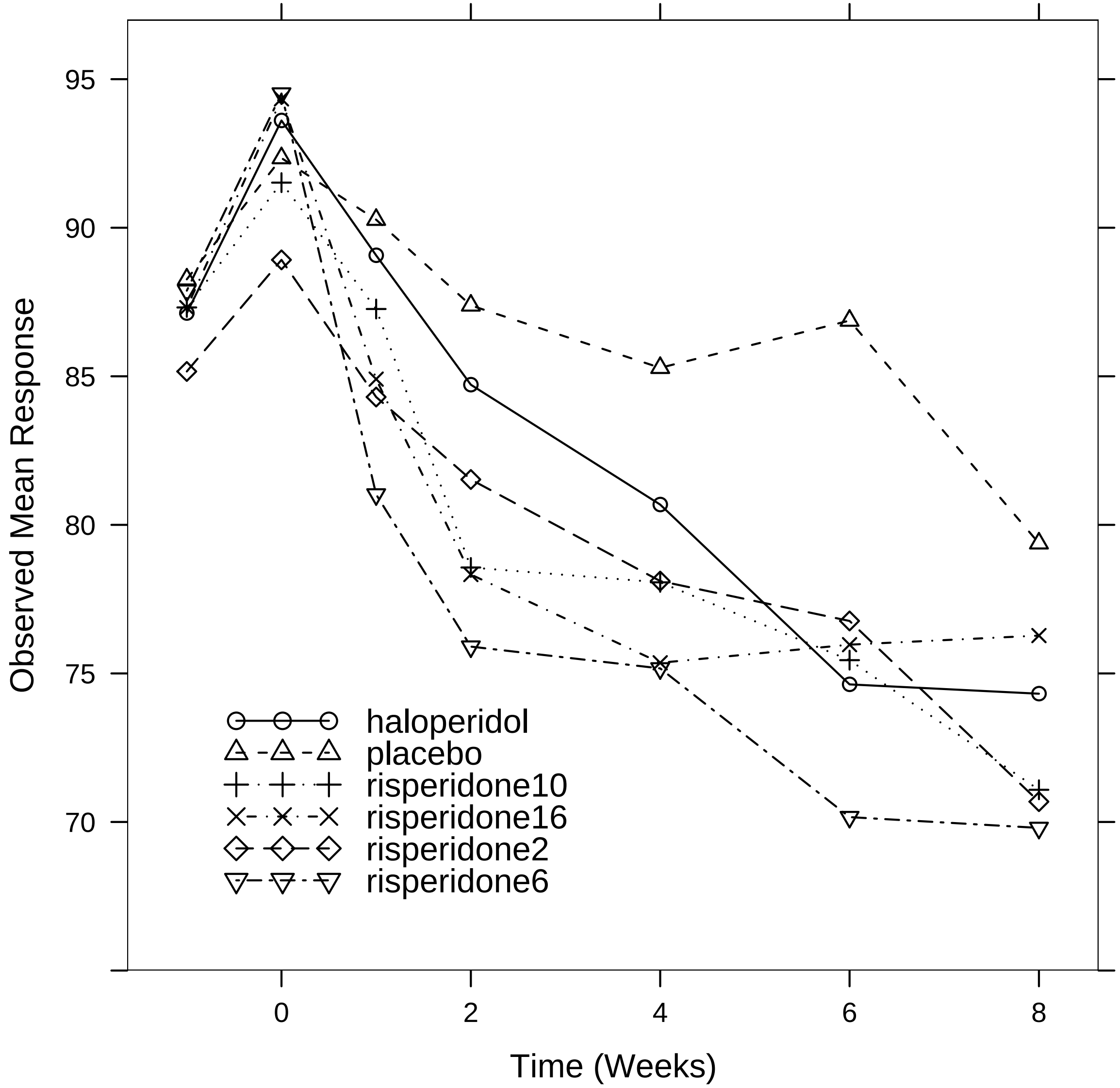
```
> last.observed <- function (x, time = NULL) {
+     if (is.null (time)) {
+         time <- 1:length (x)
+     }
+     dropout <- sapply (1:length(x),
+                        function (i, y) all(y[i:length(y)]), is.na (x))
+     ifelse (any (dropout), time[which(dropout)[1]-1], time[length(x)])
+ }
> panss$last.observed <- apply (panss[,2:8], 1, last.observed, otime)
> droptime <- table (panss$last.observed, panss$group)
> ndrop <- apply (droptime, 2, cumsum)
> ndrop <- 1 - ndrop[-7,] / ndrop[7]
>
> ndrop <- data.frame (group = rep (gnames, each = 6),
+                      time = rep (otime[-1], length (gnames)),
+                      instudy = as.vector (ndrop))
> xyplot (instudy ~ time, group = group, data = ndrop,
+         type = "o", col = 1, lty = 1:6, pch = 1:6,
+         ylab = "Proportion Still in Study",
+         xlab = "Time (Weeks)",
+         key = list (coner = c(0, 0), x = 0.2, y = 0.35,
+         lines = list (lty = 1:6), points = list (pch = 1:6),
+         text = list (gnames)))
```


Proportion Still in Study
1.0
0.8
0.6
0.4
0
2
4
6
8
Time (Weeks)
haloperidol
placebo
risperidone10
risperidone16
risperidone2
risperidone6

Observed mean response profile – by treatment groups

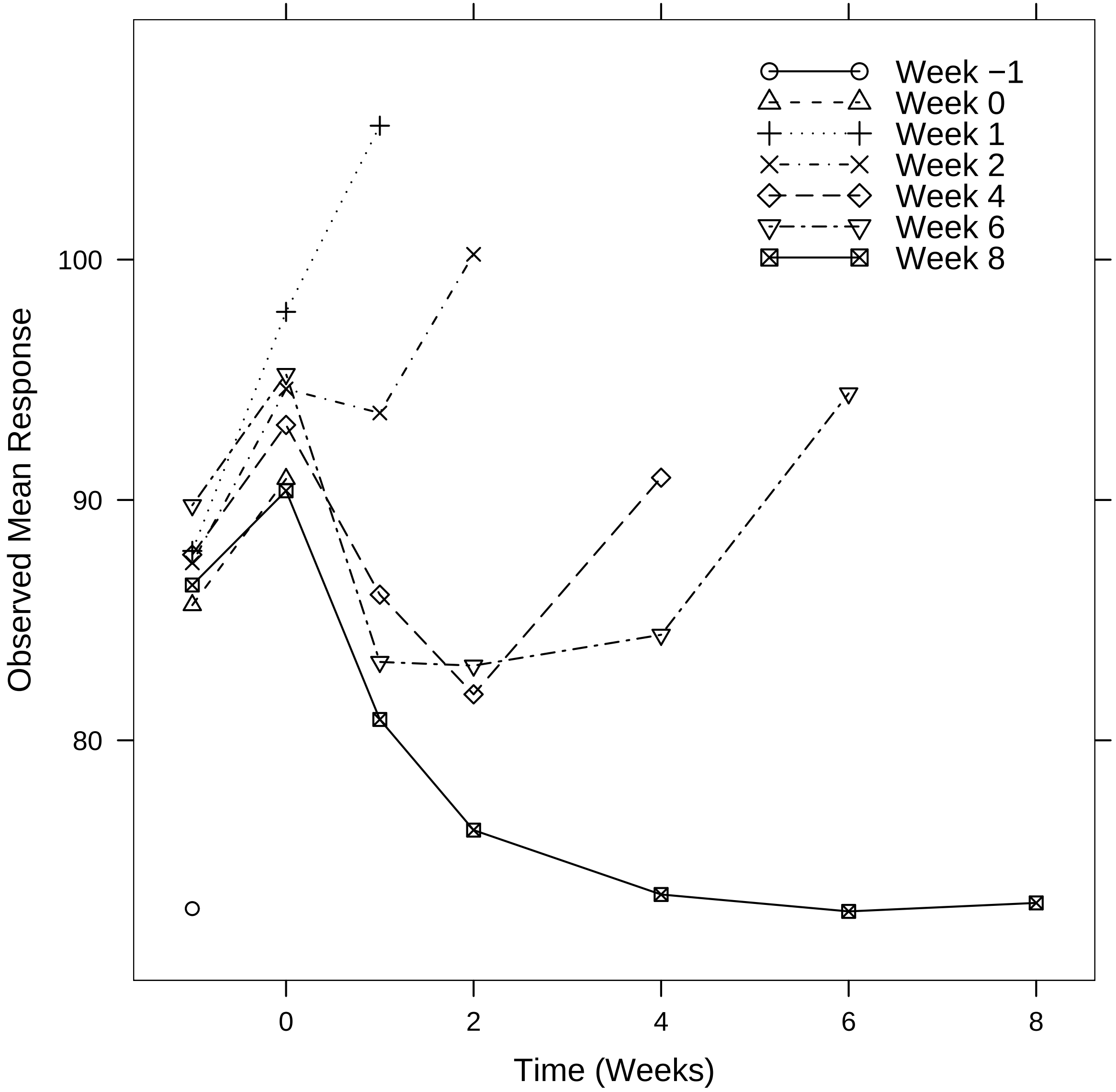
```
> panss <- read.table ("../data/panss.data")
> names (panss) <- c("group", paste ("panss", 0:6, sep = "."))
> gnames <- c("haloperidol", "placebo",
+             "risperidone10", "risperidone16", "risperidone2",
+             "risperidone6")
> panss$group <- factor (panss$group, labels = gnames)
> 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
95
90
85
80
75
70
0
2
4
6
8
Time (Weeks)
haloperidol
placebo
risperidone10
risperidone16
risperidone2
risperidone6

Observed mean response profile– by different dropout times

```
> 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))
```


Week −1
Week 0
Week 1
Week 2
Week 4
Week 6
Week 8
Observed Mean Response
100
90
80
0
2
4
6
8
Time (Weeks)

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:

  $$\text{logit}(p_{ij}) = \beta_0 + \beta_1 y_{i,j-1}. \tag{3}$$

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

- We consider two extensions of model (3):

  $$\text{logit}(p_{ij}) = \beta_0 + \beta_1 y_{i,j-1} + \beta_2 y_{i,j-2}, \tag{4}$$

  and

  $$\text{logit}(p_{ij}) = \beta_{0k} + \beta_1 y_{i,j-1} + \beta_2 y_{i,j-2}, \tag{5}$$

  where $k = k(i)$ denotes the treatment for the ith 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} (or R_{ij}, which is observed) and y_{ij} (the measurement which would have been observed had the subject not dropped out)???

  An example could be

$$\text{logit}(p_{ij}) = \beta_{0k} + \gamma y_{ij} + \beta_1 y_{i,j-1} + \beta_2 y_{i,j-2}. \tag{6}$$

- 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.
- Robins JM, Rotnitzky A, Zhao LP (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. JASA, 90, 106-21.
- Scharfstein DO, Rotnitzky A, Robins JM (1999). Adjusting for non-ignorable dropout using semiparametric non-response models (with Discussion). JASA, 94, 1096-1120.