

Missing Data

So far, we conducted all our analyses on the basis of complete data. This is a blissful, yet highly unusual setting.

We use the following definition for missing data, borrowed from (Roderick JA Little 2019):

“Missing data are unobserved values that would be meaningful for analysis if observed; in other words, a missing value hides a meaningful value.”

We distinguish the following patterns of missingness:

- **Monotonic missingness/ dropout:** All values by a subject after a certain time are missing. More specifically, if responses are missing at visit $n \in \mathbb{N}$, then responses are also missing for every subsequent visit $n + m$, for all $m \in \mathbb{N}$. *Example:* Subject drop-out from the clinical study.
- **Intermittent missingness:** Subjects miss one or several visits, but return for later visits. *Example:* A subject with data collected at baseline and Time 1 (Week 2), a missing value at Time 2 (Week 4) and a non-missing value at Time 3 (Week 8).

Note that, following the nomenclature introduced by (Roderick JA Little 2019), we use the term missing data *pattern*, to describe which data are missing in the data matrix of subject responses, and the term missing data *mechanism*, which describes the relationship between missing and observed values in the subject responses.

Our dataset contains a second variable `chgdrops`, which is subject to missingness. Let's rerun our initial MMRM with `chgdrops` as dependent variable, baseline value, visit, baseline by visit interaction and treatment by visit interaction as fixed effects and an unstructured covariance matrix for visits within each subject.

This formulation is very similar to the one at the beginning of the former chapter. How do the results differ in terms of LS Means of change from baseline by treatment arm over time?

```
fit_cat_time <- mmrm::mmrm(  
  formula = chgdrops ~ basval*avisit + trt*avisit + us(avisit | subject),  
  data = all2,  
  control = mmrm_control(method = "Kenward-Roger")  
)
```

```
# summary(fit_cat_time)

model_lsmeans <- emmeans::emmeans(fit_cat_time, ~trt*avisit, weights = "proportional")
model_lsmeans
```

trt	avisit	emmean	SE	df	lower.CL	upper.CL
1	Week 2	-4.10	0.900	47.0	-5.91	-2.29
2	Week 2	-5.29	0.899	47.0	-7.10	-3.48
1	Week 4	-6.42	0.974	46.5	-8.38	-4.46
2	Week 4	-8.52	0.951	44.8	-10.43	-6.60
1	Week 8	-9.73	1.142	40.4	-12.03	-7.42
2	Week 8	-12.62	1.114	40.1	-14.88	-10.37

Confidence level used: 0.95

```
emmeans::emmeans(fit_cat_time, ~trt*avisit, weights = "proportional") %>%
  contrast(
    list(
      "Difference in LS Means at Week 8" = c(0, 0, 0, 0, -1, 1),
      "Difference in longitudinal LS Means to Week 8" = c(-1, 1, -1, 1, -1, 1)/3
    )
  )
```

contrast	estimate	SE	df	t.ratio
Difference in LS Means at Week 8	-2.90	1.60	40.3	-1.814
Difference in longitudinal LS Means to Week 8	-2.06	1.23	46.8	-1.671

p.value

0.0772

0.1014

To understand the nature of the differences between the model using **change** as a response variable and the one with **chgdrops**, we need to look closer into the extent of missing data and understand its nature.

Missing Data Mechanisms

To understand the nature of missing data in our clinical trial, we consider the following taxonomy, introduced by (Roderick JA Little 2019). We differentiate between the following three types of missing data:

- **Missing Completely at Random (MCAR):** Conditional on all covariates in our analysis, the probability of missingness does not depend on either observed or unobserved values of the response variable.
- **Missing at Random (MAR):** Conditional on all covariates and observed response values in our analysis, the probability of missingness does not depend on the unobserved values of the response variable.
- **Missing not at Random (MNAR):** Conditional on all covariates and observed response values in our analysis, the probability of missingness does depend on the unobserved values of the response variable.

(Mallinckrodt and Lipkovich 2016) give the following interpretation around the three types of missingness:

“With MCAR, the outcome variable is not related to the probability of dropout (after taking into account covariates). In MAR, the observed values of the outcome variable are related to the probability of dropout, but the unobserved outcomes are not (after taking into account covariates and observed outcomes). In MNAR the unobserved outcomes are related to the probability of dropout even after the observed outcomes and covariates have been taken into account.”

The following two sections outline handling strategies for missing data. However, the best approach to handle missing data is to minimise its extent. While the occurrence of missing data can rarely be avoided at all (think about the collection of questionnaire data in oncology studies and the missing data after subjects die), it is important to pursue an “as complete as can be” data collection.

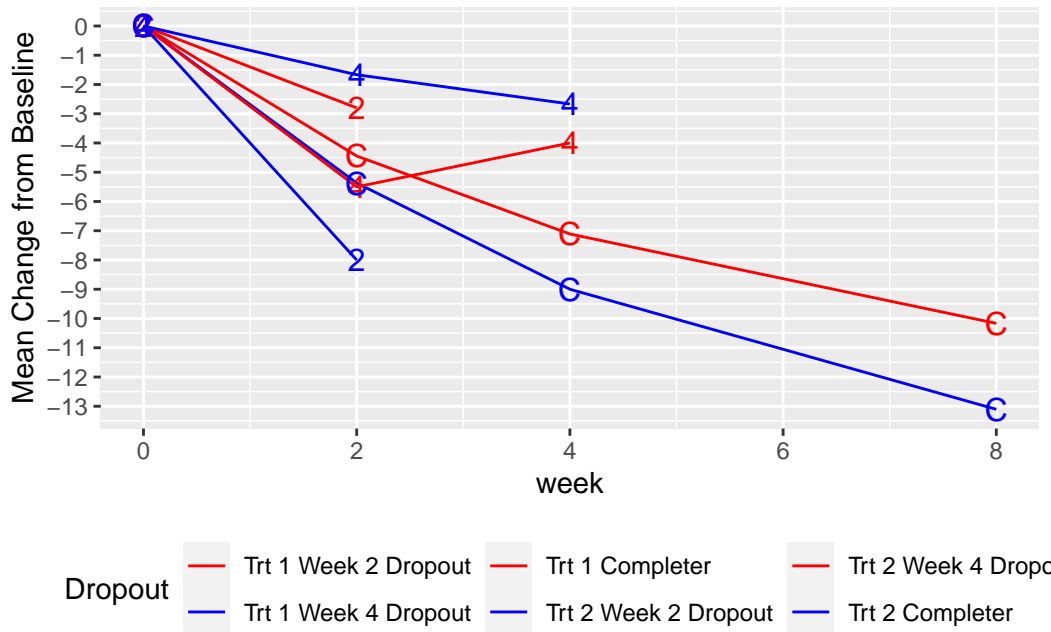
Baseline and screening data are of utmost importance in a pursuit of data completeness. If a screening value is missing, but was meant to be used as a covariate, this subjects’ whole data will be dropped from the analysis even if all responses were observed. If the baseline response variable was missing we are unable to compute a change from baseline, which also leads to the loss of this subjects’ data in the model (although LDA models are still able to provide an estimate) even if all post-baseline values were observed.

Missing data handling I (descriptive stats + visualisations)

To gain an understanding of the impact of missingness on the average response trajectories, we can plot the mean changes from baseline by visit for each drop-out group. The three drop-out groups (variable `dropout_grp`) are:

- Drop-outs at Week 2: Subjects who completed baseline and Week 2, but discontinued from the study prior to Week 4.

- Drop-outs at Week 4: Subjects who completed baseline, Week 2 and Week 4, but discontinued from the study prior to Week 8.
- Completers: Subjects who completed all visits in the study.



Exercise: Try to interpret the plot above and discuss the following topics around missingness:

- Look into the data. Which missing data pattern is present in this dataset?
- What can be seen in the plot? How does the drop-out time affect the observed mean response trajectories?
- What other aspects, apart from response, could influence a subjects' likelihood to drop-out from the study?
- Which other summaries/ visualizations can be useful to characterize and monitor the degree of missingness in clinical study data?

Solution:

```
### Work with code chunks to find the solution to the exercises
```

Missing data handling II (naive analytic approaches)

This section provides an overview of simple and most of the times overly naive methods to deal with missing data. Although we will introduce more suitable methods in the next chapter, the approaches introduced in this section have gained questionable popularity in the past, which is why we introduce them here. The following methods to compute or completely ignore missing data exist:

- Complete Case Analysis: Discard all subjects with missing observations and only conduct the analysis on subjects with complete follow-up data.
- Last observation carried forward (LOCF): Handling of monotonic missing data. The missing visits are imputed with the last non-missing value. This approach assumes a constant trend of observations after drop-out from the study, i.e. the response level remains the same as the last response under the study drug.
- Baseline observation carried forward (BOCF): Handling of monotonic missing data. The missing visits are imputed with the baseline value. This approach assumes that subjects' symptom severity or functioning (whichever was measured in the study) *bounce back* to the baseline state, prior to the initiation of the study drug.

Complete Case Analyses

Let us run a complete case analysis on the `all12` dataset.

Exercise: Fit an MMRM with response variable `chgdrops`, with baseline severity, treatment and visit as fixed effects, as well as baseline-by-visit and treatment-by-visit interaction, using an unstructured variance-covariance matrix on the `all12` completers.

- How do the results differ from the results obtained in the former chapter (response variable `change`, no missing data)?
- How do the results differ from the results obtained at the beginning of this chapter (response variable `chgdrops` with missing data)?
- Discuss the limitations of the complete case analysis. Which sources of bias can you identify?

Solution:

```
### Work with code chunks to find the solution to the exercises
```

Mallinckrodt, Craig, and Ilya Lipkovich. 2016. *Analyzing Longitudinal Clinical Trial Data*. Chapman; Hall/CRC. <https://doi.org/10.1201/9781315186634>.

Roderick JA Little, Donald B. Rubin. 2019. *Statistical Analysis with Missing Data*. Vol. 3. USA: New York, Wiley. <https://doi.org/10.1002/9781119482260>.