Course 3 · Week 2 — Missing data, longitudinal, time series

Cheatsheet — biostats_courses

R. Heller

Missingness mechanisms

Mechanism	Definition	Complete-case analysis?
MCAR	missingness independent of data	unbiased but inefficient
MAR	missingness depends on **observed** data	biased, imputation helps
MNAR	missingness depends on **unobserved** values	imputation under model only

Test is impossible without assumptions — reason about the mechanism.

Multiple imputation with `mice`

```
library(mice)
imp <- mice(df, m = 20, seed = 42, printFlag = FALSE)
fit <- with(imp, lm(y ~ x1 + x2))
pooled <- pool(fit)
summary(pooled, conf.int = TRUE)
```

Rule of thumb: $m \geq 100 \times$ fraction of missing information, but 20 is a fine starting point.

Linear mixed models

```
library(lme4); library(lmerTest)
lmer(y ~ treatment * time + (1 | subject), data = df)
lmer(y ~ treatment * time + (time | subject), data = df)  # random slope
```

- Random intercept: each cluster has its own baseline.
- Random slope: each cluster has its own trajectory.
- ICC $\rho = \sigma_u^2/(\sigma_u^2 + \sigma_e^2)$.

REML for variance components; ML only for likelihood-ratio tests on fixed effects.

GLMMs & GEE

```
library(lme4)
glmer(y ~ x + (1 | cluster), data = df, family = binomial)

library(geepack)
geeglm(y ~ x, id = cluster, data = df,
       family = binomial, corstr = "exchangeable")
```

- GLMM: subject-specific effects, good for prediction.
- GEE: population-average effects, robust to working-correlation misspecification (when combined with robust SEs).

Time series basics

```
library(forecast)
ts_obj <- ts(x, frequency = 12)
decompose(ts_obj) |> plot()
auto.arima(ts_obj)

library(changepoint)
cpt.mean(x, method = "PELT") |> plot()
```

Decision rule for Week 2

- Missing data > 5% ⟶ at least sensitivity analysis; prefer MI.
- Repeated measurements ⟶ mixed model, not repeated-measures ANOVA.
- Binary outcome with clusters ⟶ GLMM (if subject effects matter) or GEE (if marginal effect is the estimand).
- Time series with trend / seasonality ⟶ decompose before modelling.

Common pitfalls

- Last-observation-carried-forward; usually biased.
- Imputing after transformation rather than in the right scale.
- Ignoring random effects when clusters are small in number (degrees of freedom correction).
- Quoting GEE coefficients as if they were subject-specific.

Further reading

- van Buuren, *Flexible Imputation of Missing Data*.
- Fitzmaurice, Laird, Ware, *Applied Longitudinal Analysis*.