# Index of Lecture 13–L



— in addition to webpage:

`http://www.utexas.edu/its/rc/answers/general/gen25.html`

# Introduction to Missing Data

First implications of missing values:

- many methods based on explicit formulae no longer accessible (e.g., multivariate analysis) $\Rightarrow$ shift to methods that are not based on explicit formulae,
- may lead to substantial loss of data if entire record with one or several components missing is discarded (e.g., multiple regression, longitudinal data analysis with only complete subjects) $\Rightarrow$ try to work around this (e.g., drop predictors with many missing values),
- when data contains missing values, analysis of non-missing values valid only under strict (and often unrealistic assumptions), otherwise $\Rightarrow$ bias!
  — serious problem but usually ignored...

Today's lecture — only an introduction to missing data:

- review of types of missingness and their implications,
- example of test for missingness of simplest type,
- handout: overview + pointers to references and software,
- first step towards development of a module for VHM 831 on missing data (Summer 2006).

Missing values may occur in either or both of:

- predictors ($x$),
- outcomes ($Y$).

# Classification and Terminology

Notation:

- observations $U_i$ (outcome or predictor),
- missing value indicator $R_i = \begin{cases} 1 & \text{if } U_i \text{ observed} \\ 0 & \text{if } U_i \text{ missing} \end{cases}$

for $i = 1, \ldots, n$.

Three types of missing values (mv):

- missing completely at random (MCAR):
  $P(R|U) = P(R)$,
  i.e., prob. of mv does not depend on the data $U$,
- missing at random (MAR):
  $P(R|U) = P(R|U_{\text{obs}})$,
  i.e., prob. of mv only depends on observed values $U_{\text{obs}}$,
- non-ignorable, informative (NINR):
  $P(R|U) = P(R|U_{\text{obs}}, U_{\text{mis}})$,
  i.e., prob. of mv depends also on unobserved values $U_{\text{mis}}$ (the values we would have got if they were not missing).

Additional terminology for longitudinal data:

- drop-out: one obs. missing $\Rightarrow$ all subsequent obs. missing as well,
- intermittent mv: otherwise,

# Examples of Types of Missingness

Missing completely at random (MCAR):
$\sim$ missingness occurs as if tossing a coin,

- external factors (e.g., weather, loss of samples),
- planned missingness in experimental design (e.g., some subjects measured less frequently than others).

Missing at random (MAR):

- longitudinal data: mv depends on history but not on actual value,
- based on treatment protocol: treatment protocol may specify conditions for withdrawal of a subject from the trial,
- mv of outcome dependent on predictors: considered as MAR by some researchers, but as MCAR by others if analysis is conditional on covariates.

Non-ignorable non-response (NINR):

- censoring (e.g., by detection limit),
- prob. of mv depends on actual value (that would be obtained).

# Classical "Rules" for Missingness

MCAR:

- analysis of non-missing data valid for all statistical procedures, but not necessarily most efficient,
- *only* situation where complete case analysis (including only complete records) generally valid,
- assumption rarely realistic in practice.

MAR:

- analysis of non-missing data valid for all likelihood-based procedures (includes ML estimation and Bayesian approach, excludes GEE[1] and quasi-likelihood),
  - approach may still be inefficient,
  - inference about non-missing data may be undesired focus of analysis (e.g., in survival studies),
- assumption more realistic than MCAR.

NINR:

- analysis requires model for missingness, but such models difficult/impossible to check from the data,
- difficult field: many different methods, no consensus.

Dropouts generally easier to deal with analytically than intermittent mv (although latter may be more informative).

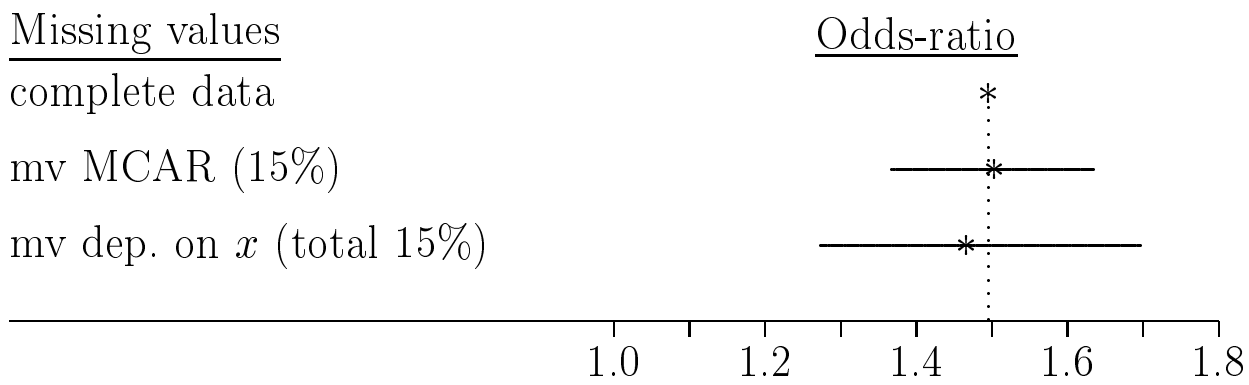
[1] Modifications to GEE for non-MCAR exist but are difficult to use.

# Missing Values for Predictor ($x$)

Methods for analysis of $Y$ given $x$:

- complete case analysis: (only complete records included),
  - correct if: $P(R|x,Y) = P(R|x)$, i.e. MCAR for $Y$ but possibly NINR for $x$,
  - (very) inefficient for high-dimensional $x$,
- likelihood/Bayesian estimation:
  - based on MAR and distribution for $x$, non-standard software (often EM algorithm),
- imputation = methodology to insert of replacement values for missing $x$'s,
  - based on MAR for $x$,
  - many variants available (not equally good/easily available; see handout).

250 simulations of mv on $x$, complete case analysis; medians and 95%-intervals among simulated data:



# "Simple Solutions" for Drop-outs

Review of two simple solutions (recommendations from Diggle et al., 2002):

- last observation carried forward: repeat last observed value on subject at all instances where missing,
  - apparently used routinely in clinical trials for comparison of groups,
  - if part of study protocol: randomization should safeguard against biases,
  - conservative if treatment effect is increasing over time,
  - *not recommended* in general,
- complete case analysis:
  - may be wasteful of data, may introduce bias if missingness not MCAR (selection bias of complete sample),
  - *not recommended* in general.

Simple imputations: mean substitution (using mean of existing obs.) and hot deck imputation (using value from most "similar" complete subject),

- based on MCAR (univariate; disregards predictors),
- mean substitution changes type/accuracy of obs.,
- hot deck dependent on good measure of similarity,
- *not recommended* for serious data analysis (my view).

# A Test for MCAR Drop-outs

Logistic regression version of non-parametric approach by Diggle (1989):

- idea: model $p_{ik}$, the probability of a drop-out for subject $i$ at time point $k$, as

$$\text{logit}(p_{ik}) = \alpha_k + \beta_k \, h_k(Y_{1i}, \ldots, Y_{k-1,i}),$$

  where $h_k$ is a suitably chosen function of the past observations,

- $h_k$ should be tailored to biologically interesting hypothesis, e.g.
  - $h_k(Y) = Y_{k-1}$ — dependence on last observed value,
  - $h_k(Y) = (Y_1 + \ldots + Y_{k-1}))/(k-1)$ — dependence on previous average,

- significance for $\beta$'s gives evidence of dependence on past observations $\sim$ MAR assumption,

- model should be applied separately to homogeneous groups of subjects (typically defined by values of predictors), or $\alpha$ and $\beta$ should be allowed to depend on predictors,

- data should be pooled across multiple times points; MCAR assumption $\Rightarrow$ within-subject independence.

# Example: Protein in Milk Samples

Cohort study involving 79 cows:

- examined weekly through 19 weeks after calving,
- outcome: protein percent in milk sample,
- 3 diets: barley (25), barley+lupins (27), lupins (27),
- drop-outs from week 15: 38/79 incomplete series, due to study termination (time censoring),
  - may be informative because of influence of calving date on protein content.

Logistic regression models for drop-out prob. from week 15, using previous value ($Y_{k-1}$) as predictor:

| | Residual | | $\hat{\beta}$ (SE) |
|---|---|---|---|
| Model for logit($p_{gk}$) | deviance | DF | (barley, week 15) |
| $\alpha_{gk} + \beta_{gk} h_k$ | 111.97 | 210 | -13.3 (7.0) |
| $\alpha_{gk} + \beta_g h_k + \beta_k h_k$ | 115.44 | 216 | -12.4 (4.2) |
| $\alpha_{gk} + \beta_k h_k$ | 118.63 | 218 | -8.9 (2.3) |
| $\alpha_{gk} + \beta_g h_k$ | 116.33 | 219 | -11.4 (3.6) |
| $\alpha_{gk} + \beta h_k$ | 119.32 | 221 | -8.2 (1.4) |
| $\alpha_{gk}$ | 197.66 | 222 | — |
| $\alpha_g + \alpha_k + \beta h_k$ | 124.16 | 224 | -7.6 (1.3) |
| $\alpha_k + \beta h_k$ | 132.61 | 229 | -6.4 (1.1) |
| $\alpha_g + \beta h_k$ | 131.28 | 230 | -7.1 (1.1) |
| $\alpha + \beta h_k$ | 139.04 | 232 | -6.2 (1.0) |

conclusion: strong effect of $h_k$, some treatment effect.

# Modelling Non-ignorable Drop-out

Distinction between 3 approaches:

- selection models: focus on (modelling) the probability of dropping out,

  $$P(Y, R) = P(Y)\, P(R|Y),$$

  and may postulate a specific model (e.g., logistic) for $P(R|Y)$ to obtain a full likelihood,

- pattern mixture models: focus on modelling the outcome on separate dropout patterns,

  $$P(Y, R) = P(R)\, P(Y|R),$$

  and suggest to analyse different patterns separately and compare their inference, before trying to build a combined model,

- random effects models: model dropout probabilities by random effects and assume (conditional) independence given those random effects,

  $$P(Y, R|(U_1, U_2)) = P(U_1, U_2)\, P(R|U_1)\, P(Y|U_2),$$

Comment in Diggle et al. (2002):

> In our opinion, debates about the relative merits of selection and pattern mixture models per se are unhelpful: models for dropout in longitudinal studies should be considered on their merits in the context of particular applications.