Outline of the problem

- Missing values in longitudinal trials are a big issue
- First aim should be to reduce proportion
- Ethics dictate that it can't be avoided
- Information lost can't be conjured up
- There is no magic method to fix it
- Magnitude of problem varies across areas
  - 8-week depression trial: 25%–50% may drop out by final visit
  - 12-week asthma trial: maybe only 5%–10%

Outline of the lecture

Part I: Missing data

Part II: Multiple imputation

1. Introduction to missing data

<table>
<thead>
<tr>
<th>Variables</th>
<th>Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>?</td>
</tr>
<tr>
<td>?</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>?</td>
<td></td>
</tr>
<tr>
<td>?</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

What is missing data?

- The missingness hides a real value that is useful for analysis purposes.

Survey questions:
1. What is your total annual income for FY 2008?
2. Who are you voting for in the 2009 election for the European parliament???
What is missing data?

Clinical trials:

Start \hspace{1cm} \text{Finish}

censored at this point in time

Missingness

- If matters why data are missing. Suppose you are modelling weight (Y) as a function of sex (X). Some respondents wouldn’t disclose their weight, so you are missing some values for Y. There are three possible mechanisms for the nondisclosure:

  1. There may be no particular reason why some respondents told you their weights and others didn’t. That is, the probability that Y is missing may have no relationship to X or Y. In this case our data is missing completely at random.
  2. One sex may be less likely to disclose its weight. That is, the probability that Y is missing depends only on the value of X. Such data are missing at random.
  3. Heavy (or light) people may be less likely to disclose their weight. That is, the probability that Y is missing depends on the unobserved value of Y itself. Such data are not missing at random.

Example: The analgesic trial

- Single-arm trial with 530 patients recruited (491 selected for analysis)
- Analgesic treatment for pain caused by chronic nonmalignant disease
- Treatment was to be administered for 12 months
- We will focus on Global Satisfaction Assessment (GSA)
- GSA scale goes from 1—very good to 5—very bad
- GSA was rated by each subject 4 times during the trial, at months 3, 6, 9, and 12

Missing data patterns & mechanisms

- **Pattern**: Which values are missing?
- **Mechanism**: Is missingness related to the response?

\[(Y_i, R_i) = \text{Data matrix, with COMPLETE DATA} \]

\[R_{ij} = \text{Missing data indicator matrix} \]

\[R_{ij} = \begin{cases} 
1, & Y_{ij} \text{ missing} \\
0, & Y_{ij} \text{ observed} 
\end{cases} \]

\(Y^{o}_i = \text{Observed part of } Y \)

\(Y^{m}_i = \text{Missing part of } Y \)

Missing data patterns & mechanisms

“**Pattern**” concerns the distribution of R

“**Mechanism**” concerns the distribution of R given Y

Rubin (Biometrika 1976) distinguishes between:

- **Missing Completely at Random (MCAR)**
  \[P(R|Y) = P(R) \text{ for all } Y \]

- **Missing at Random (MAR)**
  \[P(R|Y) = P(R|Y^*) \text{ for all } Y^* \]

- **Not Missing at Random (NMAR)**
  \[P(R|Y) \text{ depends on } Y^* \]
**Missing At Random (MAR)**

- What are the most general conditions under which a valid analysis can be done using only the observed data, and no information about the missingness value mechanism?
  \[ P(\mathbf{R}^{*}, \mathbf{Y}^{*}) \]
- The answer to this is when, given the observed data, the missingness mechanism does not depend on the unobserved data. Mathematically,
  \[ P(\mathbf{R}^{*}, \mathbf{Y}^{*}) = P(\mathbf{R}^{*}) \]
- This is termed **Missing At Random**, and is equivalent to saying that the behaviour of two units who share observed values have the same statistical behaviour on the other observations, whether observed or not.

**Example**

<table>
<thead>
<tr>
<th>Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unit 1</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Unit 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
</tbody>
</table>

- As units 1 and 2 have the same values where both are observed, given these observed values, under MAR variables 3, 5 and 6 from unit 2 have the same distribution (NB not the same value!) as variables 3, 5 and 6 from unit 1.
- Note that under MAR the probability of a value being missing will generally depend on observed values, so it does not correspond to the intuitive notion of "random". The important idea is that the missing value mechanism can expressed solely in terms of observations that are observed.
- Unfortunately, this can rarely be definitively determined from the data at hand!

**If data are MCAR or MAR**, you can ignore the missing data mechanism and use **multiple imputation** and **maximum likelihood**.

**If data are NMAR**, you can’t ignore the missing data mechanism; two approaches to NMAR data are **selection models** and **pattern mixture**.

- Suppose \( Y \) is weight in pounds; if someone has a heavy weight, they may be less inclined to report it. So the value of \( Y \) affects whether \( Y \) is missing; the data are NMAR. Two possible approaches for such data are selection models and pattern mixture.
- **Selection models**. In a selection model, you simultaneously model \( Y \) and the probability that \( Y \) is missing. Unfortunately, a number of practical difficulties are often encountered in estimating selection models.
- **Pattern mixture** (Rubin 1987). When data is NMAR, an alternative to selection models is multiple imputation with pattern mixture. In this approach, you perform multiple imputations under a variety of assumptions about the missing data mechanism. In ordinary multiple imputation, you assume that those people who report their weights are similar to those who don’t. In a pattern-mixture model, you may assume that people who don’t report their weights are an average of 20 pounds heavier. This is of course an arbitrary assumption; the idea of pattern mixture is to try out a variety of plausible assumptions and see how much they affect your results. Pattern mixture is a more natural, flexible, and interpretable approach.

**Simple analysis strategies**

1. **Complete Case (CC) analysis**
   - **Advantages**:
     - Easy
     - Does not invent data
   - **Disadvantages**:
     - Inefficient
     - Discarding data is bad
     - CC are often biased samples

2. **Analyze as incomplete (summary measures, GEE, …)**
   - **Advantages**:
     - Does not invent data
   - **Disadvantages**:
     - Restricted in what you can infer
     - Maximum likelihood methods may be computationally intensive or not feasible for certain types of models.
Analysis strategies

(3) Analysis after single imputation

Advantages:
- Rectangular file
- Good for multiple users

Disadvantages:
- Naïve imputations not good
- Invents data - inference is distorted by treating imputations as the truth
**Ignorability**

- Let us decide to use likelihood based estimation.
- The full data likelihood contribution for subject $i$ is:

  $$L(\theta, \phi | Y_i, D_i) \propto L(Y_i | Y_i, D_i, \theta, \phi)$$

- Base inference on the observed data:

  $$L(\theta, \phi | Y_i, D_i) \propto L(Y_i | \theta, \phi)$$

  with

  $$L(Y_i | \theta, \phi) = \prod_{i=1}^{n} f(Y_i | \theta, \phi)$$

  $$= \prod_{i=1}^{n} f(Y_i | \theta, \phi) \psi(Y_i)$$

  In a likelihood setting the term ignorable is often used to refer to MAR mechanism. It is the mechanism which is ignorable - not the missing data.

**Direct likelihood maximisation**

- Mechanism is MAR

  \[ \theta \text{ and } \psi \text{ distinct} \]

  \[ \text{Interest in } \theta \]

  Use observed information matrix

<table>
<thead>
<tr>
<th>Outcome type</th>
<th>Modeling strategy</th>
<th>Software</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaussian</td>
<td>Linear mixed model</td>
<td>SAS proc MIXED</td>
</tr>
<tr>
<td>Non-Gaussian</td>
<td>Generalized linear mixed model</td>
<td>SAS proc GLIMMIX, NLMIXED</td>
</tr>
</tbody>
</table>

**Example 1: Growth data**

- Taken from Pothof and Roy, Biomechanics (1964)
- Research question:

  Is dental growth related to gender?

- The distance from the center of the pituitary to the maxillary fissure was recorded at ages 8, 10, 12, and 14, for 13 girls and 16 boys.

**Growth data**

- Data
  - Complete cases
  - LOCF imputed data
  - All available data
- Model
  - Unstructured group by time mean
  - Unstructured covariance matrix
- Analysis methods
  - Direct likelihood
    - ML
    - REML
  - MANOVA
  - ANOVA per time point
### Example: The depression trial

- **Clinical trial**: experimental drug versus standard drug
- **L0F patients**
- **Response**: change versus baseline in HAM-D17 score
- **5 post-baseline visits**: 4-8

Patients are evaluated both pretreatment and posttreatment with the 17-item Hamilton Rating Scale for Depression (Ham-D-17).

### The depression trial

- **Complete case analysis**
  ```r
dc(depression, imputation, time visita, response(change), out=x(c));
  ``
  - performs analysis on CC data set
- **LOCF analysis**
  ```r
dc(depression, imputation, time visita, response(change), out=x(c));
  ``
  - performs analysis on LOCF data
- **Direct likelihood analysis**
  ```r
  fit linear mixed model to incomplete data
  ``

### 5. Part II: Multiple imputation

<table>
<thead>
<tr>
<th>Method</th>
<th>Estimate (s.e.)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC</td>
<td>-1.94 (1.77)</td>
<td>0.0995</td>
</tr>
<tr>
<td>LOCF</td>
<td>-1.63 (1.08)</td>
<td>0.132</td>
</tr>
<tr>
<td>MAR</td>
<td>-2.38 (1.16)</td>
<td>0.0419</td>
</tr>
</tbody>
</table>

- Treatment effect at visit 8 (last follow-up measurement):

Observe the slightly significant p-value under the MAR model.

### Single imputation

- **Substitute a value for each missing value**
  - Some of the values to choose this value:
    - **Mean fill-in case**
      - drop rows with missing values
    - **Complete case analysis**
      - drop rows with missing values
    - **Horizon imputation**
      - drop rows with missing values
      - drop rows with missing values
      - drop rows with missing values
      - drop rows with missing values
    - **Complete case analysis**
      - drop rows with missing values

**Disadvantage**

In general, single imputation results in the sample size being over-estimated with the variance and correlation of the estimates underestimated.
General principles

- Valid under MAR
- An alternative to direct likelihood
- Three steps:
  1. The missing values are filled in $M$ times $\Rightarrow M$ complete data sets
  2. The $M$ complete data sets are analyzed by using standard procedures
  3. The results from the $M$ analyses are combined into a single inference

Informal justification

- We need to estimate $\theta$ from the data (e.g., from the complete cases)
- Plug in the estimated $\hat{\theta}$ and use $f(y_i^*|y_i, \hat{\theta})$ to impute the missing data
- We need to acknowledge that $\theta$ is a random variable; its uncertainty needs to be included in the imputation process
- Given this distribution use:
  - draw a random $\theta^*$ from the distribution of $\theta$
  - put this $\theta^*$ is to draw a random $y_{ij}^*$ from $f(y_{ij}^*|y_i, \theta^*)$

The algorithm

1. Draw $\mathbf{\theta}^*$ from its posterior distribution
2. Draw $Y_{ij}^*$ from $f(y_{ij}^*|y_i, \mathbf{\theta}^*)$
3. To estimate $\beta$, first calculate the estimate of the parameter of interest, and it estimated variance, using the completed data, $(\mathbf{Y}, \mathbf{Y}^*)$:
   $$\hat{\beta} = \mathbf{b}^* = \mathbf{b}^* (Y, Y^*),$$
   The within imputation variance is
   $$U = \frac{1}{M} \sum_{m=1}^{M} U_m,$$
4. Repeat steps 2 and 3 a number of $M$ times
   $$\Rightarrow \hat{\beta} = \mathbf{b}^* (n = 1, \ldots, M)$$

Pooling information

- With $M$ imputations, the estimate of $\beta$ is
  $$\hat{\beta} = \frac{1}{M} \sum_{m=1}^{M} \hat{\beta}_m$$
- Further, one can make normally based inferences for $\beta$ with $$(\hat{\beta} - \beta) \sim N(0, V),$$
  where
  $$V = W + \left( \frac{1}{M} + \frac{1}{M^2} \right) \text{Var}\left(\beta\right).$$

Hypothesis testing

- Two “sample sizes”:
  - $n$: The sample size of the data set
  - $M$: The number of imputations
- Both play a role in the asymptotic distribution (Li, Raghunathan, and Rubin 1991)
  $$H_0: \theta = \theta_0;$$
  $$1 - p = P(F_{n,M} > F)$$
MI in practice

A simulation-based approach to missing data

1. Generate $M > 1$ plausible versions of $Y^*$. Complete Cases

2. Analyze each of the $M$ datasets by standard complete-data methods.

3. Combine the results across the $M$ datasets ($M = 3-5$ is usually OK).

How many datasets to create?

The efficiency of an estimator based on $M$ imputations is $(1 + \gamma/M)^{-1}$, where $\gamma$ is the fraction of missing information.

<table>
<thead>
<tr>
<th>$M$</th>
<th>0.1</th>
<th>0.3</th>
<th>0.5</th>
<th>0.7</th>
<th>0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>97</td>
<td>91</td>
<td>86</td>
<td>81</td>
<td>77</td>
</tr>
<tr>
<td>5</td>
<td>98</td>
<td>94</td>
<td>91</td>
<td>88</td>
<td>85</td>
</tr>
<tr>
<td>10</td>
<td>99</td>
<td>97</td>
<td>95</td>
<td>93</td>
<td>92</td>
</tr>
<tr>
<td>20</td>
<td>100</td>
<td>99</td>
<td>98</td>
<td>97</td>
<td>96</td>
</tr>
</tbody>
</table>

MI in practice... Step 1

Generate $M > 1$ plausible versions of $Y^*$ via software, i.e. obtain $M$ different datasets.

- An assumption we make: the data are MCAR or MAR, i.e. the missing data mechanism is ignorable.
- Should use as much information as available in order to achieve the best imputation.
- If the percentage of missing data is high, we need to increase $M$.

MI in practice... Step 2

Analyze each of the $M$ datasets by standard complete-data methods.

- Let $\beta$ be the parameter of interest.
- $\hat{\beta}_j$ is the estimate of $\beta$ from the complete-data analysis of the $j$th dataset, ($j = 1…M$)
- $U_j$ is the variance of $\hat{\beta}_j$ from the analysis of the $j$th dataset.
MI in practice... Step 3

Combine the results across the $M$ datasets.

• $\hat{\beta} = M^{-1} \sum \hat{\beta}_j$ is the combined inference for $\beta$.

• Variance for $\hat{\beta}$ is $M^{-1} \sum \Sigma U_j + (1 + M^{-1}) \Theta$.

• $\Theta = 1/(M-1) \sum (\hat{\beta}_j - \beta)^2$.

• $(1 + M^{-1}) \Theta$ incorporates added uncertainty from imputation.

Software

2. SAS software (experimental)
   It is part of SAS/STAT version 8.02
   SAS institute paper on multiple imputation, gives an example and SAS code:
   SAS documentation on PROC MI
   SAS documentation on PROC MIANALYZE

Software

1. Joe Schafer’s software from his web site. ($50)
   http://www.stat.psu.edu/~jls/softwa.html#top
   Schafer has written publicly available software primarily for S-plus. There is a stand-alone
   Windows package for data that is multivariate normal.
   This web site contains much useful information regarding multiple imputation.

Software

3. SOLAS version 3.0 ($1K)
   http://www.statsol.ie/solas/solas.htm
   Windows based software that performs different types of imputation:
   • Hot-deck imputation
   • Predictive OLS/discriminant regression
   • Nonparametric based on propensity scores
   • Last value carried forward
   Will also combine parameter results across the $M$ analyses.

MI Analysis of the Orthodontic Growth Data

• The same Model 1 as before
• Focus on boys at ages 8 and 10

Results

<table>
<thead>
<tr>
<th>Boys at Age 8</th>
<th>Boys at Age 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original Data</td>
<td>22.88</td>
</tr>
<tr>
<td>Multiple Imputation</td>
<td>22.69</td>
</tr>
</tbody>
</table>

• Confidence interval for Boys at age 10: [21.96, 24.29]
Overview

- Ignore drop-out
- CC (complete-case analysis)
- Single imputation of missing values
- LOCF (last observation carried forward)
- Generate small samples from estimated distributions
- MI (multiple imputation)
- Fit model for response at all time-points
- GEE (generalized estimating equations)
- MNLM (multivariate normal linear model; also referred to as MMRM, or mixed-model repeated measures)
- Model drop-out as well as response
- SM (selection models)
- PMM (pattern-mixture models)

Properties of methods

- MCAR: drop-out independent of response
  - CC is valid, though it ignores information
  - LOCF is valid if there are no trends with time
- MAR: drop-out depends only on observations
  - CC, LOCF, GEE invalid
  - MI, MNLM, weighted GEE valid
- MNAR: drop-out depends also on unobserved
  - CC, LOCF, GEE, MI, MNLM invalid
  - SM, PMM valid if (uncheckable) assumptions true
References
- Anderson, T.W. (1956) Maximum likelihood estimates for a multivariate normal distribution when some observations are missing.

Further References