1. Experimental design

Controlled experimental designs

1. Balanced.
2. Straightforward computation of estimates, testing contrasts and ANOVA table.

With missing responses

1. Unbalanced.
2. Identifiability problem.
3. Computation may be onerous.

When the responses are MAR, there exist some methods that fill in missing values and yield correct estimates of estimable (identifiable) parameters.

They only apply for models with a fixed-effect linear model plus one error term.
2. Linear regression analysis with complete data

Data

1. Design matrix: $X_{n \times p}$, where the $i$th row is $x_i = (x_{i1}, \ldots, x_{ip})$.
2. Response: $Y_{n \times 1}$.

Linear regression model

$$ Y = X \beta + e. $$

Important assumptions:

1. Correct mean structure.
2. The error terms $e_i$’s are independently distributed.
3. The error terms $e_i$’s have the same variance, i.e., $Var(e_i) = \sigma^2$.
4. The error terms $e_i$’s are normally distributed.
3. **Least square**

Least square estimates for linear regression

1. $\hat{\beta} = (X^T X)^{-1} X^T Y$.

2. $\text{Var}(\hat{\beta}) = (X^T X)^{-1} \sigma^2$.

3. $\hat{\sigma}^2 = \sum_{i=1}^{n} \frac{(y_i - x_i \hat{\beta})^2}{n-p}$.

4. $\frac{(n-p)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}$.

5. Let $V = (v_{ij})_{p \times p} = (X^T X)^{-1} \hat{\sigma}^2$, then $\frac{\hat{\beta}_j - \beta_j}{\sqrt{v_{jj}}} \sim t_{n-p}, j = 1, 2, \ldots, p$.

**Hypothesis testing**

Suppose $C_{p \times q}$ is a matrix to specify $q$ linear combinations of $\beta$ that to be tested: $C^T \beta = 0$, then

1. The sum of squares attributable to these restrictions is

   $$S = (C^T \hat{\beta})^T \{C^T (X^T X)^{-1} C\}^{-1} (C^T \hat{\beta}).$$

2. Test statistics: $F = \frac{S/q}{\hat{\sigma}^2} \sim F_q, n-p$. 

Biostat 2065: Analysis of Incomplete Data, 2005.
4. If the outcome $Y$ is MAR

Intuitively, the incomplete cases do not supply any information regarding the conditional distribution of $Y$ given $X$

\[
\text{MAR} \quad \implies p(M \mid X, Y) = p(M \mid X) \\
\implies M \perp Y, \text{ given } X. \\
\implies p(Y \mid X, M = 1) = p(Y \mid X, M = 0) = p(Y \mid X).
\]

The correct analysis is to carry out the least square analysis on the complete cases

1. Residual sum squares. $SS(\beta) = \sum_{i=m+1}^{n} (y_i - x_i\beta)^2$, assuming that the first $m$ cases are incomplete cases, i.e., with $y$ missing.

2. LS estimates. $\hat{\beta}_* = (X_*^T X_*)^{-1} X_*^T Y_*$, where $(X_*, Y_*)$ are the last $(n - m)$ rows of $(X, Y)$, respectively.

3. Estimate of $\sigma^2$. $\hat{\sigma}^2_* = SS(\hat{\beta}_*)/(n - m - p)$.

4. Correct sum of squares for testing $\lambda = C^T \beta = 0$, where $C_{p \times 1}$ is a vector. $SS_* = (C^T \hat{\beta}_*)^2 / \{C^T (X_*^T X_*)^{-1} C\}$.

5. Cannot use simple programs for ANOVA where data are balanced.

Biostat 2065: Analysis of Incomplete Data, 2005.
5. **Yates’s method**

1. Run least square analysis on the remaining \( n - m \) complete cases and obtain estimates \( \widehat{\beta}_* \).

2. Impute missing values by \( \tilde{y}_i = x_i \widehat{\beta}_* , i = 1, \ldots, m \). Regard the filled-in data as a complete data set.

3. Least square analysis based on the imputed dataset leads to the same estimate \( \widehat{\beta} = \widehat{\beta}_* \).

\[
SS(\beta) = \sum_{i=1}^{m} (\tilde{y}_i - x_i \beta)^2 + \sum_{i=m+1}^{n} (y_i - x_i \beta)^2
\]

4. Denote the residual variance estimate based on the imputed dataset to be \( \widehat{\sigma}^2 \), then \( \widehat{\sigma}^2 \) under-estimates the residual variance: \( \widehat{\sigma}^2 = \frac{n-m-p}{n-p} \widehat{\sigma}^2_* \).

5. Have to do the complete-case analysis at first.
6. **Healy and Westmacott**

1. An iterative algorithm.

2. Substitute some trial values for the missing values.

3. Conduct least square analysis based on the imputed data.

4. Predict the missing values with current least square estimates, and substitute these values for the corresponding missing values.

5. Repeat till the residual sum of squares stops decreasing. Then the final estimates are valid estimates for the regression parameters under the MAR assumption.

6. It is in fact an EM algorithm, to be discussed in chapter 8.

7. Though the algorithm always converges, the estimates may not make sense. i.e., when some parameters are not identifiable.

8. Convergence is slow.
7. Bartlett’s ANCOVA method

1. Fill-in trial values, say $\tilde{y}_i$, for the missing values, $i = 1, \ldots, m$.

2. Create a set of new covariates, as indicators for each individual missing value. For example, if for subject $i$, $y_i$ is missing; then a new covariate $Z_i$ is defined as $Z_i = 1$ for subject $i$ and 0 for other subjects.

3. Run ANCOVA on the imputed data with new covariates added and the model is:

   $$ Y = X\beta + Z\gamma + e. $$

4. Impute missing values by $\hat{y}_i = \tilde{y}_i - \hat{\gamma}_i z_{ii}$.

5. Residual sum squares. $SS(\beta, \gamma) = \sum_{i=1}^{m}(\tilde{y}_i - x_i\beta - \gamma_i)^2 + \sum_{i=m+1}^{n}(y_i - x_i\beta)^2$. Then $SS(\hat{\beta}, \hat{\gamma}) = SS(\hat{\beta}_*, \hat{\gamma}) = SS_*$ and $\hat{\beta} = \hat{\beta}_*$.

6. Residual variance estimate. $\hat{\sigma}^2 = SS(\hat{\beta}, \hat{\gamma})/(n - m - p) = \hat{\sigma}_*^2$.

7. The variance estimate of $\hat{\beta}$. $Var(\hat{\beta}) = \hat{\sigma}^2\{X^TX - (X^TZ)(Z^TZ)^{-1}(Z^TX)\}^{-1} = \hat{\sigma}^2_* (X_*^TX_*)^{-1}$.

Implementation of Bartlett’s method

Biostat 2065: Analysis of Incomplete Data, 2005.
8. Correct Sum of Squares Under Barlett’s ANCOVA

Barlett’s ANCOVA
Let $\lambda = C^T \beta$, where $C$ is a $p \times 1$ vector. Then based on the imputed data (complete),

1. the estimate $\widehat{\lambda} = C^T \widehat{\beta} = C^T \widehat{\beta}_*$. Correct.

2. the standard error: $SE = \hat{\sigma} \sqrt{C^T (X^T X)^{-1} C}$. Incorrect.

3. the sum of squares attributable to $C^T \lambda$ is $SS = \widehat{\lambda}^2 C^T (X^T X)^{-1} C$. Incorrect.

Correct SS for Barlett’s Method

1. Correct SE:
   
   $SE_* = \hat{\sigma}_* \sqrt{C^T (X_*^T X_*)^{-1} C}$, where $X_*$ is the design matrix for the complete cases, i.e., the $(m+1)$ to $n$ rows of $X$.

2. Correct SS:
   
   $SS_* = \widehat{\lambda}_*^2 C^T (X_*^T X_*)^{-1} C$.

Let $H^T = C^T (X^T X)^{-1} X^T Z$, then by matrix algebra

$$C^T (X_*^T X_*)^{-1} C = C^T (X^T X)^{-1} C + H^T B^{-1} H.$$ 

Then the correct $SE_*$ and $SS_*$ can be calculated as:

$$SE_* = \sqrt{\frac{n - p}{n - m - p} (SE^2 + \hat{\sigma}^2 H^T B^{-1} H)}, \quad SS_* = \frac{SS}{1 + (SS/\lambda)^2 H^T B^{-1} H}.$$