MISSING DATA ANALYSIS

Lecture 4

Hephaes Chuen Chau

2022-03-09

Conditional mean imputation

- Use the observed data for the variables with complete data to fit a regression model
- The fitted model is used to predict the values of the missing given the data of the other variables (provided that they exist)
- Example: Suppose Y_1 is completely observed (all cases present with the value of Y_1), and Y_2 is partially observed. A regression model is fitted so that

$$y_{2i} = \beta_0 + \beta_1 y_{1i} + \epsilon$$

where ϵ follows a normal distribution. Importantly, only the complete cases are used for fitting the model. This means $i < n$ and loop through only the cases which present with observed Y_1. - The observed data may include the outcome variable

Stochastic Regression Imputation

- When the missing data proportion is very small (say, 2%), stochastic regression imputation may be applied
- Essentially, this method adds a normal noise to the predicted value. The variance of the normal noise is given by the auxillary covariance matrix (which is just the variance-covariance matrix derived from the model we used)
  - Meaning of variance covariance matrix: If $\mathbf{Y} \sim N\left(\mathbf{X}\beta, \sigma^2\mathbf{I}\right)$, we have

$$\mathrm{E}(\mathbf{b}) = \left(\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right)\mathbf{X}\beta = \beta$$
$$\sigma^2\{\mathbf{b}\} = \mathrm{Cov}(\mathbf{b}) = \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

  - Since σ^2 is estimated by the MSE $s^2, \sigma^2\{\mathbf{b}\}$ is estimated by $s^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}$.