UNIVERSITY OF EDINBURGH
SCHOOL OF MATHEMATICS
INCOMPLETE DATA ANALYSIS

Workshop 2

1. (Adapted from Woodward, M. (2014) *Epidemiology: Study Design and Data Analysis*, 3rd Ed., Chapman & Hall/CRC Series in Statistical Science.)
   The following table shows a small artificial dataset with three missing values

Subject number	Age (years)	Sex	Vitamin D (ng/ml)
1	47	Male	1.8
2	76	Male	23.6
3	52	Male	2.2
4	59	Male	NA
5	43	Female	3.1
6	39	Female	4.7
7	45	Male	NA
8	60	Male	12.4
9	57	Female	44.2
10	38	Female	NA
11	63	Female	3.0
12	61	Male	8.5
13	36	Male	57.6

   The file `dataex1.Rdata` is available on Learn. Save the data on your preferred directory and then read it in `R` by typing `load("dataex1.Rdata")`.

   (a) Carry out a complete case analysis to find the mean value of vitamin D overall, and by sex. Also compute the associated standard error of the mean.

   (b) Impute the missing values of vitamin D by mean imputation. Use these filled in values to estimate the mean vitamin D with corresponding standard error.

   (c) Impute the missing values for vitamin D by regression imputation, regressing on sex and age. Write down the regression equation used and the three imputed values obtained. Use these to estimate the mean vitamin D with associated standard error.

   (d) The same as in (c) but now using stochastic regression imputation.

   (e) Suppose that hot deck imputation is to be used with strata defined by sex and age (≤ 50 years and > 50 years). Estimate the mean vitamin D with associated standard error.

2. (Adapted from `http://www4.stat.ncsu.edu/ davidian/st790/`)
In this exercise you will study missingness in regression analysis. Here the full data are $(Y, \mathbf{X})$ where Y is a scalar response and $\mathbf{X}$ is a vector of covariates and we are interested in estimation of $\boldsymbol{\beta}$ in a linear regression model of the form $\mathbf{X}^T\boldsymbol{\beta}$ for $E(Y \mid \mathbf{X})$. Suppose that it is possible for Y to be missing and that either the covariates are all observed or all missing. We write $R = (R_Y, R_X)$ where $R_Y = 1$ if Y is observed and $R_Y = 0$ if Y is missing, and similarly for R_X and $\mathbf{X}$. You will write a program to carry out a simulation study to investigate the following result: the complete case ordinary least squares estimator (i.e., the ordinary least squares estimator applied to complete cases) is consistent under any missing data mechanism, provided that the probability of observing a complete case (an individual for whom both Y and $\mathbf{X}$ are observed) is a function of the covariates and not of the response variable. Note that by a complete case we mean an individual for whom both Y and $\mathbf{X}$ is observed. Algebraically, our assumption is the following $\Pr\{(R_Y, R_X) = (1,1) \mid Y, \mathbf{X}\} = \Pr\{(R_Y, R_X) = (1,1) \mid \mathbf{X}\}$.

The objective of a simulation study is to approximate the properties of an estimator by generating some large number, say S, of independent data sets from a known situation and computing the estimates for each data set. The sample mean of the estimates over all S data sets is an estimate of the mean of the sampling distribution of the estimator; similarly, the standard deviation of the estimates over the S data sets is an estimate of the standard deviation of the sampling distribution (how good these quantities are at capturing the true features of the sampling distribution obviously depends on the size of S).

If an estimator is consistent, we expect, for reasonably large sample sizes where we might expect large sample theory to be a good approximation, the sample mean of the S estimates to be very close to the true value of the parameter being estimated. For a regression parameter with p elements $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ and estimator $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_p)'$, we can assess this by computing for each element β_k, $k = 1, \ldots, p$, the *Monte Carlo mean* $S^{-1}\sum_{s=1}^{S}\widehat{\beta}_{s,k}$ and *Monte Carlo bias*

$$S^{-1}\sum_{s=1}^{S}\widehat{\beta}_{s,k} - \beta_{0,k},$$

where $\widehat{\boldsymbol{\beta}}_s = (\widehat{\beta}_{s,1}, \ldots, \widehat{\beta}_{s,p})'$ is the estimate calculated for the sth data set and $\boldsymbol{\beta}_0 = (\beta_{0,1}, \ldots, \beta_{0,p,})'$ is the true value of $\boldsymbol{\beta}$ used to generate the data sets. We would expect the Monte Carlo mean/bias to be close to the true value/zero for a consistent estimator. Likewise, the Monte Carlo standard deviation, the usual standard deviation of the S estimates, reflects the variation in the sampling distribution of $\widehat{\boldsymbol{\beta}}$.

Write an `R` program to carry out a simulation of the performance of the complete case least squares estimator, denoted by $\widehat{\boldsymbol{\beta}}^{cc}$, under a known data generation scenario. Your program should have the following features:

- Each of the S data sets should contain n observations, where $S = 1000$ and $n = 200$. For each data set, and for each $i = 1, \ldots, n$, generate independently $X_{1i} \sim \mathrm{N}(10, 3^2)$ (standard deviation 3) and X_{2i} as Bernoulli with $\Pr(X_2 = 1) = 0.4$. Then generate

$$Y_i = \beta_1 + \beta_2 X_{1i} + \beta_3 X_{2i} + \varepsilon_i,$$

  where ε_i are independent $\mathrm{N}(0, 8^2)$ (standard deviation 8), and $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)' = (20, 5, -5)'$. Thus, the true value $\boldsymbol{\beta}_0 = (20, 5, -5)'$.

- You will run your program three times, once under each of below scenarios (i)–(iii). In each case, for each of the S data sets and for each $i = 1, \ldots, n$, generate an indicator of observing a complete case, by calculating

$$\pi_i = \frac{e^{U_i}}{1 + e^{U_i}}, \quad U_i = \psi_1 + \psi_2 X_{1i} + \psi_3 X_{2i} + \psi_4 Y_i + \psi_5 X_{1i} Y_i + \psi_6 X_{2i} Y_i,$$

  and then generating C_i, the indicator of whether or not the ith observation is a complete case ($R_i = (1, 1)$), as Bernoulli with $\Pr(C_i = 1 \mid X_{1i}, X_{2i}, Y_i) = \pi_i$. For each of the following cases, take $\boldsymbol{\psi} = (\psi_1, \psi_2, \psi_3, \psi_4, \psi_5, \psi_6)'$ as

  (i) $\boldsymbol{\psi} = c(2, -0.025, 0.5, 0, 0, 0)'$

  (ii) $\boldsymbol{\psi} = c(6, 0, 0, -0.075, -0.003, 0.05)'$

  (iii) $\boldsymbol{\psi} = c(0.5, 0, 0, 0, 0, 0)'$

  Thus, you should write your program so that it can be run for each case by simply changing the value of $\boldsymbol{\psi}$.

- For each case, for each of the S data sets of size n, calculate two least squares estimates of $\boldsymbol{\beta}$ in the above linear model: (1) the ideal complete data estimate that could be calculated if the complete data were available, which will serve as a 'gold standard' and (2) the complete case estimate based only on the observations with $C_i = 1$. For each data set $s = 1, \ldots, S$, save the values of the regression coefficients.

- For each case, calculate the Monte Carlo mean, bias, and standard deviation of the S. estimates. Discuss whether or not the results are in line with the claimed results.