繳交截止日期: 2020/06/05(五), 24:00
授課教師：吳漢銘 (臺北大學統計學系副教授)

請仔細閱讀每一個注意事項 (禁止討論)

1. 寫作業要點
   (a) 可參考課本、上課講義 (包含電子檔) 及其它資料。
   (b) 儘量不要與別人 (或同學) 討論，自己做，不可參考同學的答案，不可抄襲。
   (c) 程式設計題，若程式碼直接複製 (或照抄) 講義上的以不給分為原則。
   (d) 有問題者，請發 e-mail 或 FB 私訊問助教或老師。
   (e) 不按照規定作答者，酌量扣分。
   (f) 請參照下列文件第 2 ∼ 4 頁寫作規定，不按照規定作答者，會扣分。
   `http://www.hmwu.idv.tw/web/teaching/doc/R-how-homework.pdf`
2. 下載題目卷，上傳答題檔案:
   (a) 於課程網站下載題目卷。
   (b) 上傳答題檔案: 於教師網站首頁登入 [作業考試上傳區]，帳號: `dm108`。密碼: `xxx` (上課教室號碼)。
   (c) 請上傳「學號-姓名-DM-HW2.docx」。(目錄: 「20191030-exam1」。)
3. 答題檔案原則:
   (a) 請依照「R 程式作業繳交方式」，複製 Console「程式執行及結果」至答案卷。 圖形複製，請注意大小，內容數字文字需可辨識。
   (b) 程式設計題，若程式碼直接複製 (或照抄) 講義上的以不給分為原則。
   (c) 若上傳檔案格式錯誤，內容亂碼，空檔等等問題。請自行負責。
   (d) 若要重覆上傳 (第 2 次以上)，請在檔名最後加「−2」、「−3」，例如: 「學號-姓名-DM-HW2-2.docx」、「學號-姓名-DM-HW2-3.docx」等等。
   (e) 上傳兩次 (含) 以上、格式不合等等酌量扣分。
   (f) 如果上傳網站出現「You can modify the html file, but please keep the link 'www.wftpserver.com' at least.」，請將滑鼠移至「網址列」後，按「Enter」即可。若再不行，請換 (IE/Edge/Firefox/Chrome)。

我已經仔細閱讀上述各注意事項，若有違背，會自行負責。

遺失值處理、資料轉換

1. 以下為模擬具有遺失值資料 x 之 R 程式碼:

```
n <- 500
p <- 10
set.seed(123456)
library(MASS)
s <- matrix(rt(p*p, df=5), ncol = p)
sigma <- crossprod(s)
x <- mvrnorm(n, mu=rep(0, p), Sigma=sigma)
missing.percentage <- 0.1
x[sample(n*p, floor(n*p*missing.percentage))] <- NA
```

   (a) 選取完整之資料 (命名為 x.complete), 印出此資料之維度 (nc×pc)。

   (b) 模擬遺失: 將上述之資料隨機選取出比例為 missing.percentage 之觀察值 (ξ_i)，設置成 NA(命名 x.complete.na)。
   (提示: set.seed(54321); ij <- sample(1:nc*pc, floor(nc*pc*missing.percentage)))

   (c) 利用下列 5 方法各自對上述資料 (x.complete.na) 做補值: Mean Substitution K-Nearest Neighbour Imputation (K=5)、mice.impute.pmm {MICE}, mice.impute.norm {MICE}。

   (d) 計算下列指標數值，評估上述 5 種補值方法:

$$\sum_{i=1}^{m}(\hat{\xi_i} - \xi_i)^2,$$

   其中 m=floor(nc*pc*missing.percentage)、ξ_i 為模擬遺失之真實值，$\hat{\xi_i}$ 為 ξ_i 之補值。

2. 資料來源: (UCI) Concrete Compressive Strength Data,
   http://archive.ics.uci.edu/ml/datasets/Concrete+Compressive+Strength。
   說明檔見: 「Concrete_Readme.txt」

   (a) 讀取資料 Concrete_1030x9.txt，並做多重迴歸分析 (lm)，其中 y 為反應變數，{Cement, BFS, FlyA, Water, Sp, CA, FineA, Age} 為解釋變數, 印出 R^2 值。

   (b) 對資料做 (至少 5 種方法) 轉換 (部份或全部的解釋變數)，(方法其中至少包含標準化及 Box-Cox 轉換)，(可以有複合式轉換，例如標準化後，再施行另一種轉換)，並將轉換後的資料以多重迴歸方法分析，印出 R^2 值。對此資料而言，那一種轉換可以得到較高的 R^2 值?

注意: 上傳檔案之後，請刪除作答目錄 及答案卷，清空資源回收筒，關機。交回題目卷。