108 學年度第二學期
資料採礦: 期中考

日期: 2020/06/06(六), 10:10 12:00
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. 考試答題要點
   (a) 可參考課本、上課講義 (包含電子檔) 及其它資料。
   (b) 不可與別人 (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) 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。
   (e) 圖形複製，請注意大小，內容數字文字需可辨識。
   (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-Midterm.docx」。(改成自己的學號及姓名)(請注意「正確目錄」)
   (d) 若傳錯，請最終要上傳一份正確的的答題檔案。
   (e) 若上傳檔案格式錯誤，內容亂碼，空檔等等問題。請自行負責。
   (f) 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-Midterm-2.docx」、「學號-姓名-DM-Midterm-3.docx」等等。
   (g) 上傳兩次 (含) 以上、格式不合等等酌量扣分。
   (h) 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。

R: 統計分析

1. **敘述統計**

   資料來源: 政府資料開放平台。資料檔: 「癌症發生統計.csv」。https://data.gov.tw/dataset/6399

   (a) 以「癌症別: 肝及肝內膽管」為例，依「性別 (不分性別、男。女)」之類別，計算「年齡標準化發生率」之平均數、中位數及眾數。

   (b) 以「癌症別: 胃」為例，畫出各縣市別的「平均癌症發生數」的長條圖。(各縣市別的平均癌症發生數的算法: 各縣市別的歷年 (癌症診斷年) 的癌症發生數總和除以診斷年個數。)

2. **假設檢定**

   資料來源: 政府資料開放平台。資料檔: 「消費者端量測行動上網平均速率.csv」。資料網址: https://data.gov.tw/dataset/8258

   (a) 畫出資料中第一階段、第二階段的盒形圖。(提示: side-by-side boxplot)

   (b) 請問消費者端量測行動上網平均速率第一階段與第二階段是否有顯著差異? (提示: (1) 有母數及無母數方法，皆各選一用合適的檢定。(2) 有母數方法需注意資料是否符合假設)

3. **變異數分析/事後檢定**

   資料來源: 政府資料開放平台。資料名稱: 「各國證券市場成交值週轉率比較_NEW」。資料檔: 「每月 _103938_A43_t35 世界主要證券市場成交值周轉率比較 (35).csv」。資料網址: https://data.gov.tw/dataset/103938

   (a) 選取 2001/01∼2002/12 之資料 (除「上海」外)，儲存成一資料框 (data.frame)，並列印出。

   (b) (承上小題) 畫出每個地區 (臺灣, 紐約, 日本, 倫敦, 香港, 韓國, 新加坡) 之盒形圖。(提示: side-by-side boxplot)

   (c) (承上小題) 每個地區 (臺灣, 紐約, 日本, 倫敦, 香港, 韓國, 新加坡) 之成交值週轉率是否有顯著差異? 若有差異，是哪些地區有差異?

4. **迴歸分析**

資料來源: UCI Machine Learning Repository。資料名稱: 「QSAR aquatic toxicity Data Set」。資料檔: 「`qsar_aquatic_toxicity.csv`。資料網址: `https://archive.ics.uci.edu/ml/datasets/QSAR+aquatic+toxicity` 變數資訊: 此資料共 546 觀察值。變數依序為「TPSA(Tot), SAacc, H-050, MLOGP, RDCHI, GATS1p, nN, C-040, LC50」, 其中反應變數為「LC50」。

(a) 進行迴歸分析，並印出參數估計報表及 ANOVA 表格，需做簡單解釋。

(b) 進行逐步迴歸分析，選取最佳之模型。