111 學年度第二學期科學計算軟體作業八

系級:測量系 114　　姓名:黃薇庭

學號:F64101032

1. 讀取 HW_Database.csv 資料集，內含 2020/04/01 – 2020/05/09 共 39 日之臺南空氣品質監測站 $PM_{2.5}$、NO_2、Temperature、RH 數值，試以線性迴歸說明 $PM_{2.5}$、Temperature 及 RH 對 NO_2 之影響效應，內容需含下列項目(**20%× 5**，答題提醒: **注意標註 p 值**並說明是否達統計之顯著水準，若未達到或錯誤皆會斟酌扣分)：

   *小提醒：資料集中 $PM_{2.5}$ 變數名稱為 PM25

   (1). 各項因子之 **Pearson 相關檢定**(說明各項因子與 NO_2 之關係以及其是否具統計上之意義)。

   *選定資料集中特定欄位的方式: Dataset[5:8] (選擇 Dataset 中 5-8 欄進行作業)。

   *也可以刪除不需要的欄位之後再做分析。

   (2). 以 **ANOVA** 方式探討該線性模型之**整體配適度**如何，**是否達統計上之顯著水準，模式中有無影響 NO_2 之因子**(**注意若使用 anova_alt 需要先執行定義函數的語法，於下兩頁**，source: https://community.rstudio.com/t/anova-table-for-full-linear-model/42074/10 created by RussS)。

   (3). 根據模式輸出之結果報表，說明**哪幾項因子(不含常數項/截距)對於 NO_2**

之影響具統計上之顯著意義，整體模式之解釋能力(決定係數)為何？

(4). 列出完整模型公式(均四捨五入至小數點第二位)。

(5). 模型校正後決定係數值，及其計算公式為何(需詳列公式，原始 R2 四捨五入至小數點第二位後才帶入，結果也四捨五入至小數點第二位)？

ANS:

由下圖可以發現，

NO2 與 PM25 為高度正相關(r=0.75)，然而 p-value=0.0000<0.05，達到統計上之顯著水準

NO2 與 Temperature 為負相關 (r=-0.57)，然而 p-value=0.0002<0.05，達到統計上之顯著水準

NO2 與 RH 為負相關(r=-0.24)，然而 p-value=0.1455>0.05，未達到統計上之顯著水準

```
> library("Hmisc")
> Ex1 <- data.frame (PM25=c(data1$PM25),NO2=c(data1$NO2),Temperature=c(data1$Te
mperature),RH=c(data1$RH))
> rcorr(as.matrix(Ex1),type=c("pearson"))
             PM25   NO2 Temperature    RH
PM25         1.00  0.75       -0.58 -0.45
NO2          0.75  1.00       -0.57 -0.24
Temperature -0.58 -0.57        1.00  0.10
RH          -0.45 -0.24        0.10  1.00

n= 39


P
            PM25   NO2    Temperature RH
PM25               0.0000 0.0001      0.0041
NO2         0.0000        0.0002      0.1455
Temperature 0.0001 0.0002              0.5342
RH          0.0041 0.1455 0.5342
```

```
+ }
> anova_alt(Ex1_lm)
Analysis of Variance Table

       Df     SS     MS      F          P
Source  3 193.07 64.357 17.346 4.5234e-07
Error  35 129.86  3.710
Total  38 322.93  8.498
```

p-value = 4.5234e-7 < 0.05 達顯著水準，表示整體模型至少有一項因子對於 NO2 有影響。

```
> anova(lm(NO2~Temperature,data=data1))
Analysis of Variance Table

Response: NO2
            Df Sum Sq Mean Sq F value    Pr(>F)
Temperature  1 103.39 103.392  17.425 0.000174 ***
Residuals   37 219.54   5.934
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Temperature 則為 p = 0.000174 < 0.05 達顯著水準，表示 Temperature 對 NO2 有影響。

```
> anova(lm(NO2~PM25,data=data1))
Analysis of Variance Table

Response: NO2
          Df Sum Sq Mean Sq F value    Pr(>F)
PM25       1 182.53 182.530  48.102 3.455e-08 ***
Residuals 37 140.40   3.795
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

PM25 則為 p = 3.455e-8 < 0.05 達顯著水準，表示 PM25 對 NO2 有影響。

```
> anova(lm(NO2~RH,data=data1))
Analysis of Variance Table

Response: NO2
           Df  Sum Sq Mean Sq F value Pr(>F)
RH          1  18.209 18.2093   2.211 0.1455
Residuals 37 304.723  8.2358
```

RH 則為 p = 0.1455 > 0.05 未達顯著水準,表示 RH 對 NO2 沒有影響。

```
> Ex1_lm<-lm(NO2 ~ PM25+Temperature+RH, data=Ex1)
> summary(Ex1_lm)

Call:
lm(formula = NO2 ~ PM25 + Temperature + RH, data = Ex1)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5295 -1.0611  0.4441  1.3051  4.0397

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.27225    6.01909   1.208    0.235
PM25         0.28689    0.06210   4.620 5.03e-05 ***
Temperature -0.17109    0.12995  -1.317    0.197
RH           0.03474    0.04661   0.745    0.461
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.926 on 35 degrees of freedom
Multiple R-squared:  0.5979,    Adjusted R-squared:  0.5634
F-statistic: 17.35 on 3 and 35 DF,  p-value: 4.523e-07
```

根據 T 檢定結果之 p-value，可知截距 $p = 0.235 > 0.05$ 未達顯著水準，截距為 7.27225。

PM25 則為 $p = 5.03e\text{-}5 < 0.05$ 達顯著水準,對於 NO2 的影響具有統計上之意義，PM25 之迴歸係數 = 0.28689

Temperature 則為 $p = 0.197 > 0.05$ 未達顯著水準，對於 NO2 之影響沒有統計上之意義，其迴歸係數 = -0.17109

RH 則為 $p = 0.461 > 0.05$ 未達顯著水準，對於 NO2 之影響沒有統計上之意義，RH 之迴歸係數 =0.03474

整體模式對於 NO2 之決定係數 = 0.5979，能夠解釋 NO2 變化的 59.79%，p-value = 4.523e-7 < 0.05，達統計上之顯著水準,是為統計上可信之結果。

模型公式:

NO2(y) = 7.27225 + 0.28689 ×PM25 -0.17109 × Temperature +0.03474 ×RH

校正後決定係數:

$$R^2{}_{adj} = 1 - (\frac{39-1}{39-1-4}) \times (1 - 0.5979) \approx 0.5506$$

#code:

```
####################
# Anova for full model fitting #
####################
# Definition function
anova_alt = function (object, reg_collapse=TRUE,...)
{
  if (length(list(object, ...)) > 1L)
    return(anova.lmlist(object, ...))
  if (!inherits(object, "lm"))
    warning("calling anova.lm(<fake-lm-object>) ...")
  w <- object$weights
  ssr <- sum(if (is.null(w)) object$residuals^2 else w * object$residuals^2)
  mss <- sum(if (is.null(w)) object$fitted.values^2 else w *
               object$fitted.values^2)
  if (ssr < 1e-10 * mss)
    warning("ANOVA F-tests on an essentially perfect fit are unreliable")
  dfr <- df.residual(object)
  p <- object$rank
  if (p > 0L) {
    p1 <- 1L:p
    comp <- object$effects[p1]
    asgn <- object$assign[stats:::qr.lm(object)$pivot][p1]
    nmeffects <- c("(Intercept)", attr(object$terms, "term.labels"))
    tlabels <- nmeffects[1 + unique(asgn)]
    ss <- c(vapply(split(comp^2, asgn), sum, 1), ssr)
    df <- c(lengths(split(asgn, asgn)), dfr)
    if(reg_collapse){
      if(attr(object$terms, "intercept")){
        collapse_p<-2:(length(ss)-1)
        ss<-c(ss[1],sum(ss[collapse_p]),ss[length(ss)])
        df<-c(df[1],sum(df[collapse_p]),df[length(df)])
        tlabels<-c(tlabels[1],"Source")
      } else{
        collapse_p<-1:(length(ss)-1)
        ss<-c(sum(ss[collapse_p]),ss[length(ss)])
        df<-c(df[1],sum(df[collapse_p]),df[length(df)])
        tlabels<-c("Regression")
      }
    }
  }else {
    ss <- ssr
    df <- dfr
    tlabels <- character()
    if(reg_collapse){
      collapse_p<-1:(length(ss)-1)
      ss<-c(sum(ss[collapse_p]),ss[length(ss)])
      df<-c(df[1],sum(df[collapse_p]),df[length(df)])
    }
  }

  ms <- ss/df
  f <- ms/(ssr/dfr)
  P <- pf(f, df, dfr, lower.tail = FALSE)
  table <- data.frame(df, ss, ms, f, P)
  table <- rbind(table,
                 colSums(table))
  if (attr(object$terms, "intercept")){
    table$ss[nrow(table)]<- table$ss[nrow(table)] - table$ss[1]
  }
  table$ms[nrow(table)]<-table$ss[nrow(table)]/table$df[nrow(table)]
  table[length(P):(length(P)+1), 4:5] <- NA
  dimnames(table) <- list(c(tlabels, "Error","Total"),
                          c("Df","SS", "MS", "F",
                            "P"))
  if (attr(object$terms, "intercept")){
    table <- table[-1, ]
    table$MS[nrow(table)]<-table$MS[nrow(table)]*
(table$Df[nrow(table)])/(table$Df[nrow(table)]-1)
```

```
    table$Df[nrow(table)]<-table$Df[nrow(table)]-1
  }
  structure(table, heading = c("Analysis of Variance Table\n"),
            class = c("anova", "data.frame"))
}

install.packages("Hmisc")
library("Hmisc")
setwd("/Users/huangweiting/coding/INTRODUCTION TO SCIENTIFIC COMPUTING SOFTWARE
/W10_ClassData")
getwd()
data1<-read.csv("HW_Database.csv")
Ex1 <- data.frame
(PM25=c(data1$PM25),NO2=c(data1$NO2),Temperature=c(data1$Temperature),RH=c(data1$RH))
rcorr(as.matrix(Ex1),type=c("pearson"))

Ex1_lm<-lm(NO2 ~ PM25+Temperature+RH, data=Ex1)
summary(Ex1_lm)

anova_alt(Ex1_lm)
anova(lm(NO2~Temperature,data=data1))
anova(lm(NO2~PM25,data=data1))
anova(lm(NO2~RH,data=data1))
```