數據分析報告(using R)

資管三 406257004 彭靜怡

1. 從外部匯入資料

```
getData <- function(){
  csvFile <- file.path("D:/R/insurance.csv")
  data <- read.csv(csvFile,sep=",")
}
```

Dataset (Medical Cost Personal Datasets Insurance Forecast by using Linear Regression) 此 data 在研究不同地區的健保費與年紀、性別、BMI、是否有孩子以及是否抽菸等資料是否相關。

2.敘述統計

#觀察 dataset 前五筆資料 **head(data,5)**

```
  age    sex    bmi children smoker    region   charges
1  19 female 27.900        0    yes southwest 16884.924
2  18   male 33.770        1     no southeast  1725.552
3  28   male 33.000        3     no southeast  4449.462
4  33   male 22.705        0     no northwest 21984.471
5  32   male 28.880        0     no northwest  3866.855
> 
```

#了解資料的統計特徵

查看 data summary **summary(data)**

```
> summary(data)
      age            sex                 bmi           children    
 Min.   :18.00   Length:1338        Min.   :15.96   Min.   :0.000  
 1st Qu.:27.00   Class :character   1st Qu.:26.30   1st Qu.:0.000  
 Median :39.00   Mode  :character   Median :30.40   Median :1.000  
 Mean   :39.21                      Mean   :30.66   Mean   :1.095  
 3rd Qu.:51.00                      3rd Qu.:34.69   3rd Qu.:2.000  
 Max.   :64.00                      Max.   :53.13   Max.   :5.000  
    smoker             region             charges     
 Length:1338        Length:1338        Min.   : 1122  
 Class :character   Class :character   1st Qu.: 4740  
 Mode  :character   Mode  :character   Median : 9382  
                                       Mean   :13270  
                                       3rd Qu.:16640  
                                       Max.   :63770  
```

經過此步驟我觀察出：

1. age：年齡介在 18~64 之間

2.BMI ：BMI 介於 15.96~53.13

3.children：有幾個小孩介在 0~5 個

4.charges：醫療保險費用介於 1122~6370

5.平均保險花費是 13270

6.各項欄位的極端值都蠻大的

使用 Hmisc 了解資料的統計特徵

library(Hmisc) #使用 Hmisc library

describe(data) #使用 Hmisc 的內建函數來查看 data summary

```
> describe(data)
data

 7  Variables      1338  Observations
------------------------------------------------------------------------
age 
       n  missing distinct     Info     Mean      Gmd      .05      .10 
    1338        0       47    0.999    39.21    16.21       18       19 
     .25      .50      .75      .90      .95 
      27       39       51       59       62 

lowest : 18 19 20 21 22, highest: 60 61 62 63 64
```

```
------------------------------------------------------------------------
sex
       n  missing distinct
    1338        0        2

Value      female   male
Frequency     662    676
Proportion  0.495  0.505
------------------------------------------------------------------------
bmi
       n  missing distinct     Info     Mean      Gmd      .05      .10
    1338        0      548        1    30.66    6.893    21.26    22.99
     .25      .50      .75      .90      .95
   26.30    30.40    34.69    38.62    41.11

lowest : 15.960 16.815 17.195 17.290 17.385, highest: 48.070 49.060 50.380
 52.580 53.130
------------------------------------------------------------------------


children
       n  missing distinct     Info     Mean      Gmd
    1338        0        6    0.899    1.095    1.275

lowest : 0 1 2 3 4, highest: 1 2 3 4 5

Value          0     1     2     3     4     5
Frequency    574   324   240   157    25    18
Proportion 0.429 0.242 0.179 0.117 0.019 0.013
------------------------------------------------------------------------
smoker
       n  missing distinct
    1338        0        2

Value         no   yes
Frequency   1064   274
Proportion 0.795 0.205
------------------------------------------------------------------------

region
       n  missing distinct
    1338        0        4

Value      northeast northwest southeast southwest
Frequency        324       325       364       325
Proportion     0.242     0.243     0.272     0.243
------------------------------------------------------------------------
charges
       n  missing distinct     Info     Mean      Gmd      .05      .10
    1338        0     1337        1    13270    12301     1758     2347
     .25      .50      .75      .90      .95
    4740     9382    16640    34832    41182

lowest :  1121.874  1131.507  1135.941  1136.399  1137.011
highest: 55135.402 58571.074 60021.399 62592.873 63770.428
------------------------------------------------------------------------
>
```

經過此步驟我觀察出：

1.此資料在收集四個不同的地區的醫療保險費

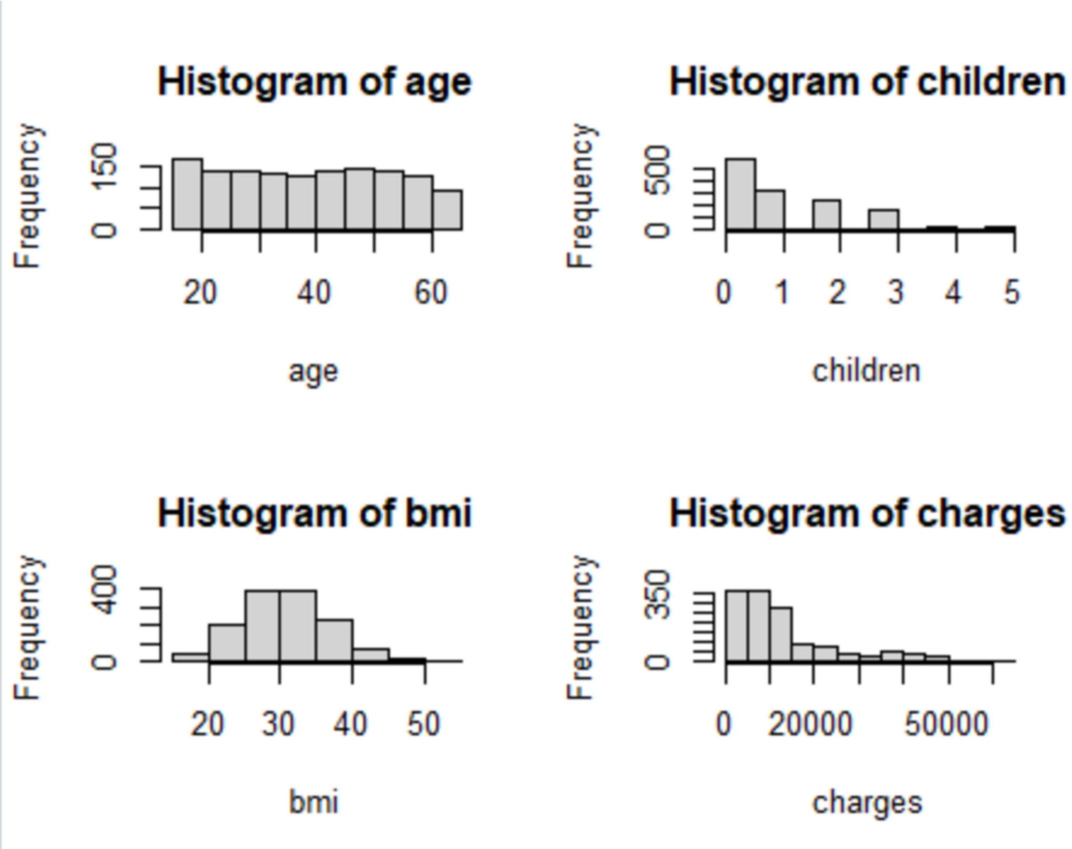
2.男女比接近 1:1

3.年齡、BMI 差距頗大
4.有接近八成的受試者不抽菸

#用視覺化的方式列出欄位資訊

(只針對 BMI,charges,children 這三欄做分析，因為只有這三欄是數字資料)

```
oldpar <- par(mfcol=c(2,2))    #讓圖表顯示成兩列兩欄
titles <- names(data)     #取得欄位名稱
for(i in c(1,3,4,7)){     #1,3,4,7 的欄位是我們想要取得的三個欄位
hist(x=data[,i],main=paste("Histogram of",titles[i]),xlab=titles[i])
}
par(oldpar)
```



由此可觀察出：
1.BMI 大多介在 20-40
2.大多數沒有小孩
3.醫療保險費大多小於 20000

觀察在不同性別底下，Ages, BMI, children, charges 三個變數的關係

cor.all <- by(data[,c(1,3,4,7)],INDICES = data$sex,cor)

print(cor.all)

```
> print(cor.all)
data$sex: female
                age        bmi   children   charges
age      1.00000000 0.09721409 0.07849989 0.3245748
bmi      0.09721409 1.00000000 0.02215070 0.1614187
children 0.07849989 0.02215070 1.00000000 0.0584917
charges  0.32457479 0.16141865 0.05849170 1.0000000
----------------------------------------
data$sex: male
                age         bmi    children
age      1.00000000 0.123088412 0.008689940
bmi      0.12308841 1.000000000 0.002385175
children 0.00868994 0.002385175 1.000000000
charges  0.28236853 0.225847080 0.074496435
            charges
age      0.28236853
bmi      0.22584708
children 0.07449643
charges  1.00000000
> 
```

由此可觀察出：

不管男性或是女性醫療保險花費都和年齡以及 BMI 成正相關，與 children 成微弱正相關，因為相關性趨近於 0 表示花費跟有幾個小孩沒什麼關連性。

3.常態檢定

檢測 age,BMI,children,charges 是否是常態分布

#age

qqnorm(data$age,main="age")　# 常態機率圖

qqline(data$age,col = "Red")　#畫出最佳斜線

print(shapiro.test(data$age[0:5000]))　#shapiro-wilk 檢定

 # BMI

qqnorm(data$bmi,main="BMI")

qqline(data$bmi= "Blue")

print(shapiro.test(data$bmi[0:5000]))

 # children

qqnorm(data$ children,main=”children”)

qqline(data$ children = "Green")

print(shapiro.test(data$ children [0:5000]))

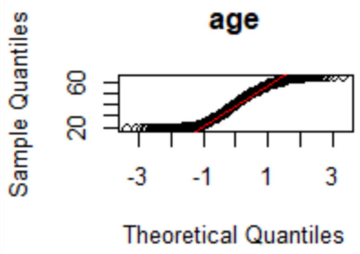
 # charges

qqnorm(data$ charges,main=”charges”)

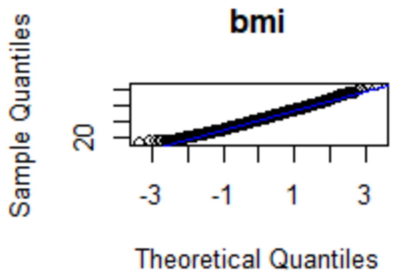
qqline(data$ charges = "Yellow")

print(shapiro.test(data$ charges [0:5000]))

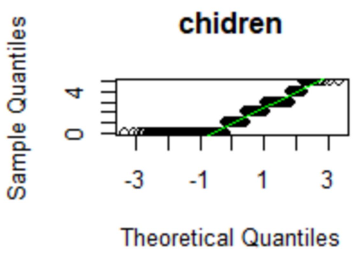
```
> qqnorm(data$age,main="age")
> qqline(data$age,col = "Red")
> print(shapiro.test(data$age[0:5000]))

        Shapiro-Wilk normality test

data:  data$age[0:5000]
W = 0.9447, p-value < 2.2e-16
```



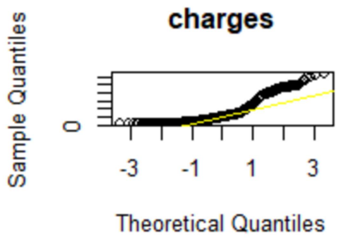

```
>  qqnorm(data$bmi,main="bmi")
>  qqline(data$bmi,col = "Blue")
>  print(shapiro.test(data$bmi[0:5000]))

        Shapiro-Wilk normality test

data:  data$bmi[0:5000]
W = 0.99389, p-value = 2.605e-05
```




```
> qqnorm(data$children,main="chidren")
> qqline(data$children,col = "Green")
> print(shapiro.test(data$children[0:5000]))

        Shapiro-Wilk normality test

data:  data$children[0:5000]
W = 0.82318, p-value < 2.2e-16
```



```
> qqnorm(data$charges,main="charges")
> qqline(data$charges,col = "Yellow")
> print(shapiro.test(data$charges[0:5000]))

        Shapiro-Wilk normality test

data:  data$charges[0:5000]
W = 0.81469, p-value < 2.2e-16
```



透過此步驟我觀察到

1.在 Shapiro-Wilk 檢定中得出的 Age,bmi,children,charges 的 p-value 都小於 0.05

2. charges 的常態機率圖中斜線與真實分布圖存在較大的差異;而 age,bmi,chidren 的常態機率圖中斜線與真實分布圖存在較小的差異

3.由前面兩個敘述可以得知前三欄資料為“常態性分布”，而最後一欄資料則為“非常態性分布”

---

4. 簡單線性回歸分析

#使用 ggplot2 package

install.packages(“ggplot2”)

library(ggplot2)

#建立模型

bmi <- data$bmi

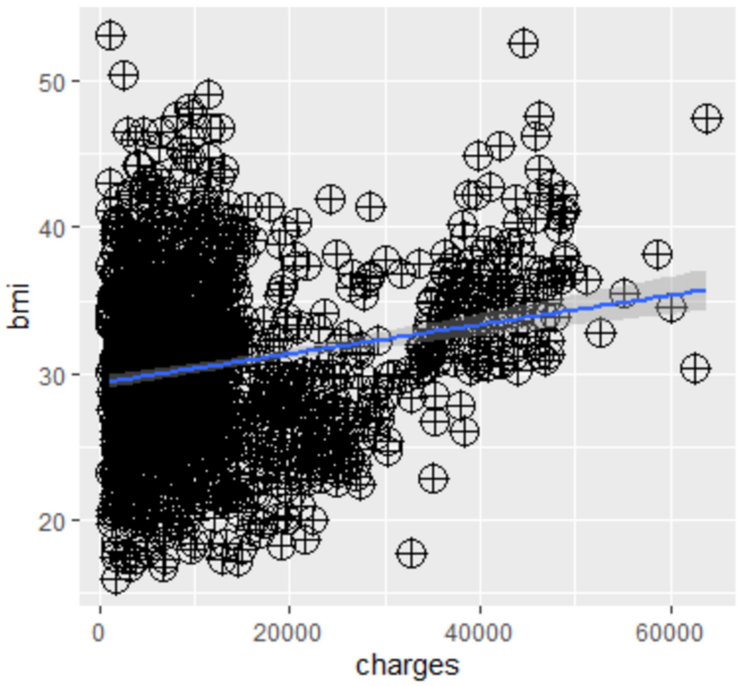
charges<- data$charges

LM<-lm(bmi ~charges,data=data)

dev.off() #避免錯誤發生

畫出分布加預測圖

ggplot(data, aes(x=data$charges, y=data$bmi)) + geom_point(shape = 10, size = 5) + geom_smooth(method = lm) + labs(x = "charges", y = "bmi")



#取得方程式參數

summary(LM)

```
> summary(LM)

Call:
lm(formula = bmi ~ charges, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-14.8424  -4.1030  -0.2401   3.8467  23.6758 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2.934e+01  2.426e-01 120.956  < 2e-16 ***
charges     9.988e-05  1.350e-05   7.397 2.46e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.979 on 1336 degrees of freedom
Multiple R-squared:  0.03934,	Adjusted R-squared:  0.03862 
F-statistic: 54.71 on 1 and 1336 DF,  p-value: 2.459e-13
```

由此觀察出：

1.此回歸模型公式可寫成 bmi = (2.934e+01)+(9.988e-05)*charges+e

2.Adjusted R-squared 偏小，表示此模型的預測能力偏低

#殘差性常態性檢定

shapiro.test(LM$residual[0:5000])

```
> shapiro.test(LM$residual[0:5000])

        Shapiro-Wilk normality test

data:  LM$residual[0:5000]
W = 0.99249, p-value = 2.544e-06
```

由此觀察出：p-value 極低，故殘差值的常態性假設是不成立的

5. 利用預測函數取得結果

```
new <- data.frame(charges=30 )    #給予一個新的值
result <- predict(LM, newdata = new)    #進行預測
print(result)    # 列印結果
```

```
>
> new <- data.frame(charges=40000)#給予一個新的值
> result <- predict(LM, newdata = new)# 進行預測
> print(result) # 列印結果
       1
33.33309
>
> |
```

由此可知，醫療保險花費在 40000 時，模型預測出的 BMI 為 33.33