106-1 生物統計學二 實習課

R : Simple Linear Regression

周芷妤

2017.09.21

d04849010@ntu.edu.tw

大綱

- Review

- Simple Linear Regression
  - Estimation
  - Scatter plot & regression line

Review

Review

- Introduction to R
  - ✓ 簡介與安裝
  - ✓ 基本操作
  - ✓ 敘述性統計
  - ✓ 繪圖
  - ✓ 資料匯入、匯出
  - ✓ 指令查詢

資料匯入、匯出

- 資料**匯入**(read)
  - 點選方式

    read.table(file.choose(), sep=",", header=T)

  - 輸入路徑

    read.table("資料路徑.副檔名", sep=", ", header=T)

- 資料**匯出**(write)

  write.table(欲匯出的資料名,"路徑.副檔名", sep=", ", row.names=F)

- 注意斜線的方向
- 雙引號(")也可以改成單引號(')

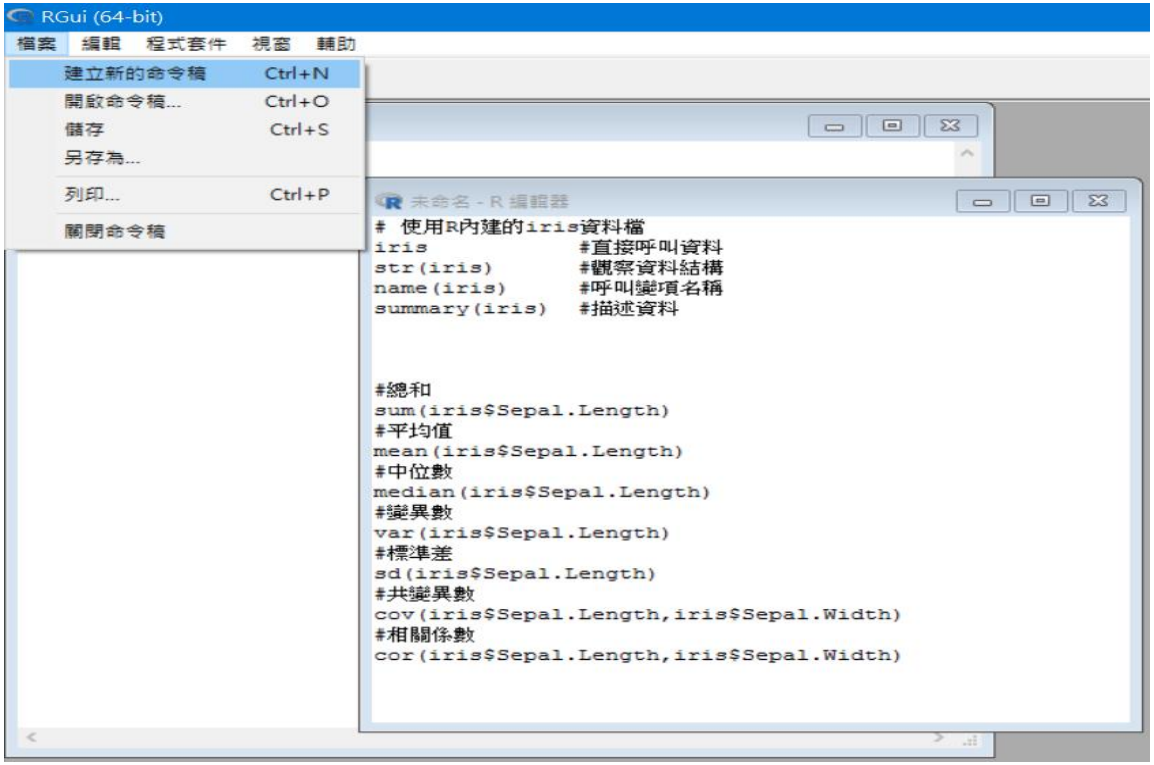
***注意**:

- 路徑："C:\\MyData1.csv " 或 "C:/MyData1.csv " 都可用
- 資料分隔方式：csv檔是以逗號(", ")分隔

  txt檔通常是以空白(" ") or tab ("\t")分隔

命令稿(Script)

- 請多使用命令稿!!
  - 將指令記在命令稿，再一起執行較方便
  - 當指令打錯時，方便除錯(debug)
  - 可將執行過的指令存檔



選擇欲執行的指令後，按F5即可執行

R內建資料

- R中內建了許多的資料庫，這些資料目前都在datasets這個package(套件)裡

- 檢視所有的資料名稱和簡介：
  - ✓ library(help=datasets)
  - ✓ data()

- 查詢資料的詳細說明：
  - ✓ help(資料名稱)

  ex: help(iris)

Simple Linear Regression

Estimation
Scatter plot & regression line

Simple linear regression

Model

$$Y = \beta_0 + \beta_1 X + \varepsilon \quad , \quad \varepsilon \sim N(0,\ \sigma^2)$$

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_i e_i^2$$

- (β_0, β_1) 的估計及意義：

迴歸係數	β_1 : 斜率 (slope)	β_0 : 截距 (intercept)
參數的意義	X每增加1個單位，Y平均增加β_1個單位	baseline effect
參數的估計	$\hat{\beta}_1 = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i (X_i - \bar{X})^2} = \frac{S_{XY}}{S_{XX}}$	$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$
性質	$\mathrm{E}[\hat{\beta}_1] = \beta_1$	$\mathrm{E}[\hat{\beta}_0] = \beta_0$

Fit linear model

- 建立迴歸模型

model <- lm(Y ~ X , data = 資料檔名稱)

語法	說明	備註
Y (反應變數的名稱)	response variable	
X (共變數的名稱)	covariate	
data	指定欲分析的資料檔	Ex : data = mydata1

Example : IRIS data

- 使用iris資料
  - Petal.Width 作為response (Y)
  - Petal.Length 作為covariate (X)

有用attach時，可不加 data = 資料名稱

```
> attach(iris)
> # 建立迴歸模型
> fit <- lm(Petal.Width ~ Petal.Length, data = iris)
>
> # 查看內容包含變項
> names(fit)
 [1] "coefficients"  "residuals"     "effects"       "rank"
 [5] "fitted.values" "assign"        "qr"            "df.residual"
 [9] "xlevels"       "call"          "terms"         "model"
>
> # 查看迴歸係數
> fit$coefficients
 (Intercept) Petal.Length
  -0.3630755    0.4157554
```

Example : IRIS data

```
> # 利用公式計算迴歸係數
> SXY <- sum( (Petal.Width - mean(Petal.Width))*(Petal.Length - mean(Petal.Length)) )
> SXX <- sum( (Petal.Length - mean(Petal.Length))^2 )
> SXY/SXX
[1] 0.4157554
>
> # 計算ei總和及ei*xi總和
> sum(fit$residual)
[1] 6.31873e-16
> sum(fit$residual*Petal.Length)
[1] -3.191891e-16
```

Example : IRIS data

```
> fit <- lm(Petal.Width ~ Petal.Length, data = iris)
> summary(fit)

Call:
lm(formula = Petal.Width ~ Petal.Length, data = iris)

Residuals:
     Min       1Q   Median       3Q      Max
-0.56515 -0.12358 -0.01898  0.13288  0.64272

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.363076   0.039762  -9.131  4.7e-16 ***
Petal.Length  0.415755   0.009582  43.387  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2065 on 148 degrees of freedom
Multiple R-squared:  0.9271,    Adjusted R-squared:  0.9266
F-statistic:  1882 on 1 and 148 DF,  p-value: < 2.2e-16
```

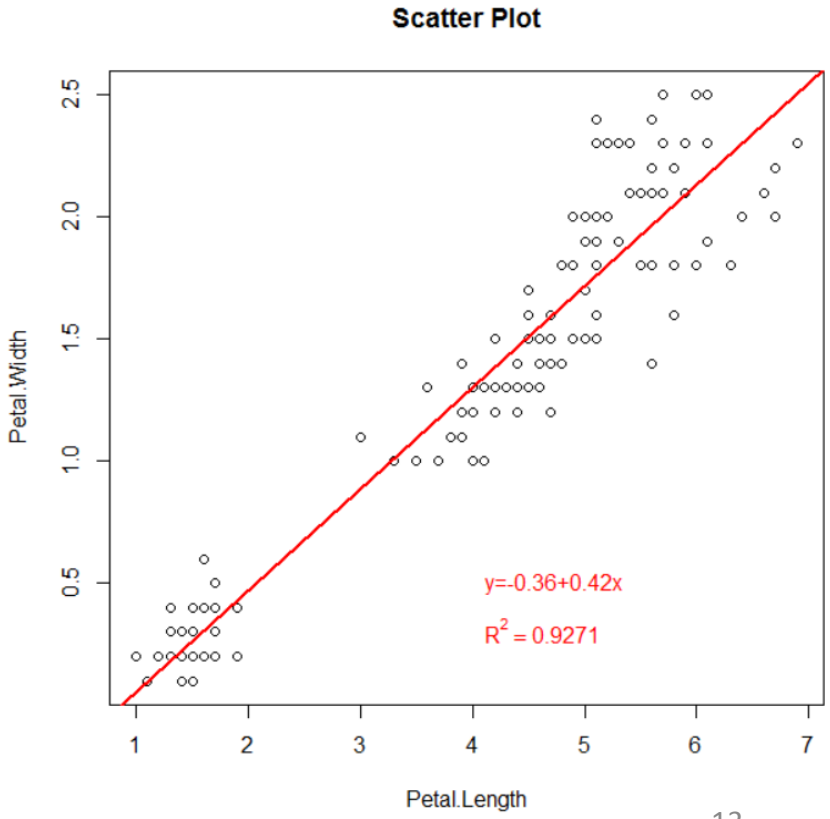
Petal.Length 每增加一單位，Petal.Width平均增加 0.42個單位

Scatter plot & regression line

注意X和Y的位置

```
> plot(Petal.Length, Petal.Width , main="Scatter Plot")
> abline(fit, col="red", lwd=2)                                  #畫迴歸線
> text(4, 0.5, "y=-0.36+0.42x", col="red", pos=4, cex=1 )        #加文字敘述
> text(4, 0.3, expression(R^2==0.9271), col="red", pos=4, cex=1 )  #加文字敘述
```

語法	說明	備註
lwd	線條粗細	數字越大越粗
pos	文字位置 (以指定的座標來看)	1: below 2: left 3: above 4: right
cex	圖標or文字大小	可用於plot及text指令中



Homework

1.請先匯入example-1資料，並顯示前幾筆資料

2.請以身高為反應變數(Y)，體重為共變數(X)建立迴歸模型，

(a)計算出迴歸係數的數值

(b)嘗試說明迴歸係數在資料上的意義

3. 請畫出身高、體重的scatter plot

(圖上須包含主標題、迴歸線、迴歸線係數)

4. 請將code及output貼上word檔，上傳至ceiba作業區

*最晚上傳期限為2017.9.23(六)中午12點