105-1: EE4052

計算機程式設計

Computer Programming

Unit 15: 線性關係

連 豊 力

臺大電機系

Sep 2016 - Jan 2017

大綱

什麼是 abline

什麼是 線性回歸模型

U10：加入圖形元件 – 點 線 框

- points( )　　# 打點
- lines( )　　# 畫線
- abline( )　　# 畫 y = b x + a 的直線
- segments( )　　# 畫線段
- arrows( )　　# 畫箭頭
- box( )　　# 在原圖形最外圍加上框框
- lty　　# 直線的樣式
- lwd　　# 直線的寬度

U12：遺漏值處理

- 回歸模型預測值補差法：
- nhanes2[, 4]　　# 針對第4組數據
- sub <- which(is.na(nhanes2[, 4]) == TRUE)
- dataTR <- nhanes2[-sub,]
- dataTE <- nhanes2[sub,]
- dataTE
- lmout <- lm(chl ~ age, data = dataTR)
  - # 利用 dataTR 中 age 為引數，chl 為因變數，建構線性回歸模型
- dataTE[, 4] <- round(predict(lmout, dataTE))
  - # 用回歸模型預測值取代之
- dataTE

大綱

三個資料庫

nhanes2, cars, iris

資料庫

- install.packages("mice") # 安裝 mice 軟體套件
- library(mice) # 載入 mice 軟體套件
- data(nhanes2)
- nrow(nhanes2) # nhanes2 資料集的橫列數
- ncol(nhanes2) # nhanes2 資料集的直行數
- summary(nhanes2) # nhanes2 資料集的概括資訊
- head(nhanes2)

```
> head( nhanes2 )
    age  bmi  hyp chl
1 20-39   NA <NA>  NA
2 40-59 22.7   no 187
3 20-39   NA   no 187
4 60-99   NA <NA>  NA
5 20-39 20.4   no 113
6 60-99   NA <NA> 184
```

```
> summary( nhanes2 )
    age          bmi          hyp          chl
 20-39:12   Min.   :20.40   no  :13   Min.   :113.0
 40-59: 7   1st Qu.:22.65   yes : 4   1st Qu.:185.0
 60-99: 6   Median :26.75   NA's: 8   Median :187.0
            Mean   :26.56             Mean   :191.4
            3rd Qu.:28.93             3rd Qu.:212.0
            Max.   :35.30             Max.   :284.0
            NA's   :9                 NA's   :10
```

遺漏值處理：回歸模型預測值補差法

- 回歸模型預測值補差法：
- data0 <- nhanes2 # 針對第2, 4組數據
- subNA <- which(is.na(nhanes2[, 4]) == TRUE | is.na(nhanes2[, 2]) == TRUE)
- dataOK <- nhanes2[-subNA,]
- dataNA <- nhanes2[subNA,]
- dataOK
- dataNA
- lmchlbmi <- lm(chl ~ bmi, data = dataOK)
  - # 利用 dataOK 中 bmi 為引數，chl 為因變數，建構線性回歸模型

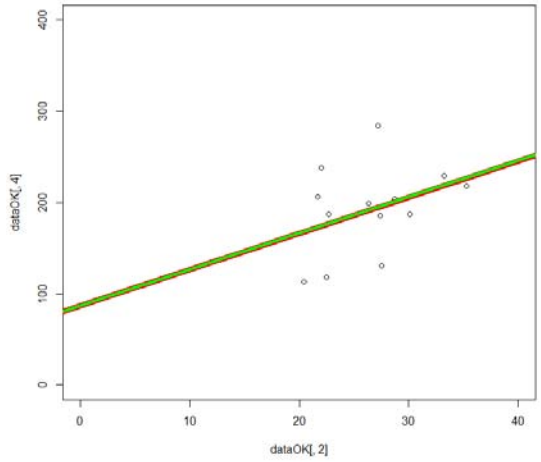
```
> lmout

Call:
lm(formula = chl ~ bmi, data = dataOK)

Coefficients:
(Intercept)          bmi
     87.130        3.963
```

chl = 3.963 * bmi + 87.130

遺漏值處理：回歸模型預測值補差法

計算機程式設計 – 2016F
Chap 15: 線性關係
Feng-Li Lian @ NTU-EE

- abline( ) # 畫 y = b x + a 的直線
- plot(dataOK[, 2], dataOK[, 4], xlim = c(0, 40), ylim = c(0, 400))
- abline(a = 87.130, b = 3.963, col = "red", lwd = 8)
- abline(lmchlbmi, col = "green", lwd = 4)

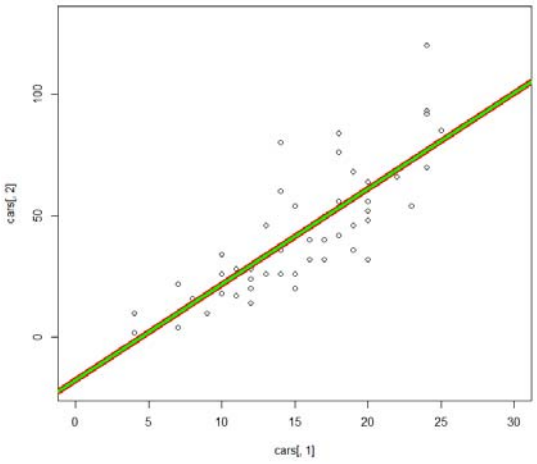


chl = 3.963 * bmi + 87.130

```
> lmchlbmi

Call:
lm(formula = chl ~ bmi, data = dataOK)

Coefficients:
(Intercept)          bmi
     87.130        3.963
```

另一個資料：cars

- cars
- plot(cars[, 1], cars[, 2], xlim = c(0, 30), ylim = c(-20, 130))
- lmcars <- lm(dist ~ speed, data = cars)
- abline(a = -17.579, b = 3.932, col = "red", lwd = 8)
- abline(lmcars, col = "green", lwd = 4)

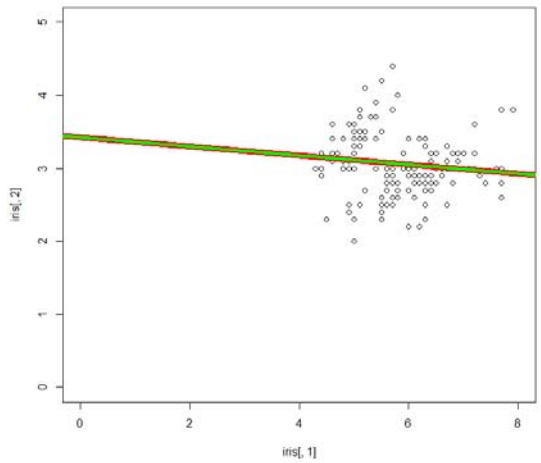


chl = 3.932 * speed – 17.579

```
> lmcars

Call:
lm(formula = dist ~ speed, data = cars)

Coefficients:
(Intercept)        speed
    -17.579        3.932
```


另一個資料：iris

- iris
- plot(iris[, 1], iris[, 2], xlim = c(0, 8), ylim = c(0, 5))
- lmiris1 <- lm(Sepal.Width ~ Sepal.Length, data = iris)
- abline(a = 3.41895, b = -0.06188 , col = "red", lwd = 8)
- abline(lmiris1, col = "green", lwd = 4)

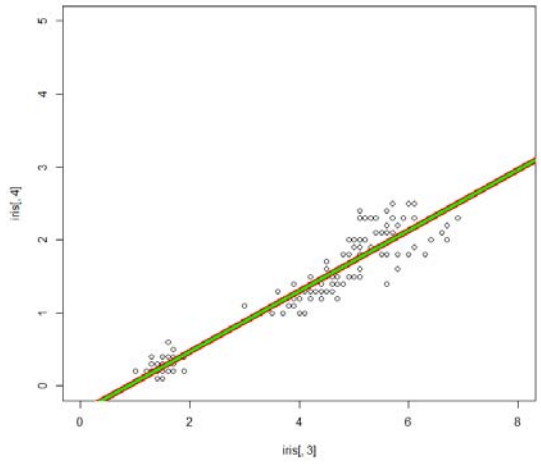


Sepal.Width = -0.06188 * Sepal.Length + 3.41895

```
> lmiris1

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

Coefficients:
 (Intercept)  Sepal.Length
     3.41895      -0.06188
```

另一個資料：iris

- iris
- plot(iris[, 3], iris[, 4], xlim = c(0, 8), ylim = c(0, 5))
- lmiris2 <- lm(Petal.Width ~ Petal.Length, data = iris)
- abline(a = -0.3631, b = 0.4158, col = "red", lwd = 8)
- abline(lmiris2, col = "green", lwd = 4)




```
Petal.Width = 0.4158 * Petal.Length - 0.3631
```

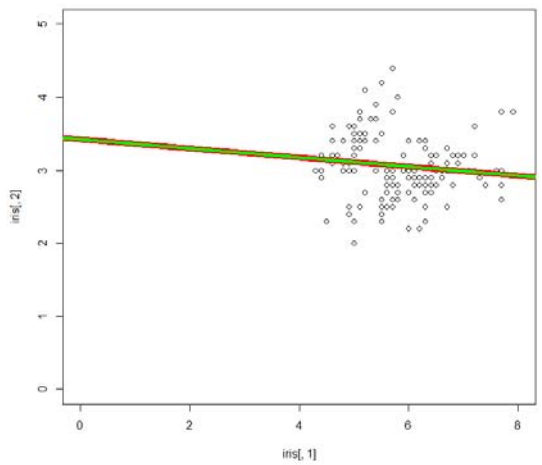
```
> lmiris2

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

Coefficients:
 (Intercept)  Petal.Length
     -0.3631        0.4158
```


另一個資料：iris

- iris

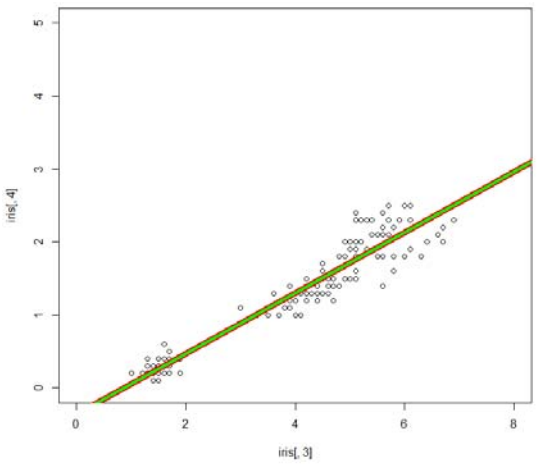
```
> lmiris1

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

Coefficients:
 (Intercept)  Sepal.Length
     3.41895      -0.06188
```

```
Sepal.Width = -0.06188 * Sepal.Length + 3.41895
```




```
> lmiris2

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

Coefficients:
 (Intercept)  Petal.Length
     -0.3631        0.4158
```

```
Petal.Width = 0.4158 * Petal.Length - 0.3631
```



U11：相關性

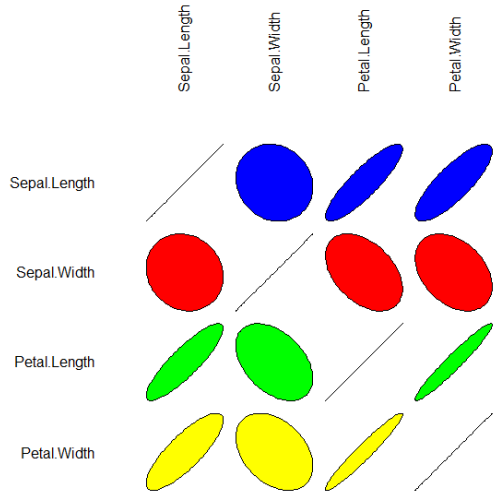
```
# cor( ), correlation 相關係數

cor( x, y )

cor_matrix <- cor( data_all, use = "pairwise" )


cor_iris <- cor( iris[, 1:4], use = "pairwise")
```

```
> cor_iris

             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    1.0000000  -0.1175698    0.8717538   0.8179411
Sepal.Width    -0.1175698   1.0000000   -0.4284401  -0.3661259
Petal.Length    0.8717538  -0.4284401    1.0000000   0.9628654
Petal.Width     0.8179411  -0.3661259    0.9628654   1.0000000
```


計算機程式設計 – 2016F
Chap 15: 線性關係
Feng-Li Lian @ NTU-EE

U11：相關性

```
# plotcoor( ), 繪製相關圖

install.packages( "ellipse" )

library( ellipse )

plotcorr( cor_iris, col = c( "blue", "red", "green", "yellow" ) )
```




```
> cor_iris

             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    1.0000000  -0.1175698    0.8717538   0.8179411
Sepal.Width    -0.1175698   1.0000000   -0.4284401  -0.3661259
Petal.Length    0.8717538  -0.4284401    1.0000000   0.9628654
Petal.Width     0.8179411  -0.3661259    0.9628654   1.0000000
```

大綱

Im: Linear Model

Least Squares Approximation

Least Squares Approximation

- 參考資料：http://www.ms.uky.edu/~ma138/Spring15/Curve_fitting.pdf

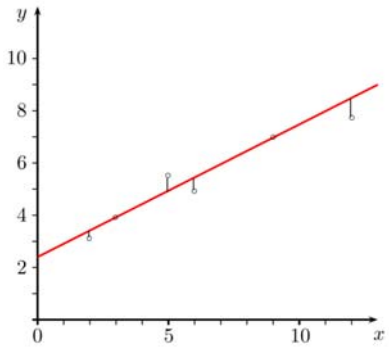


FIGURE 1: Fitting a straight line to data by the method of least squares

$$\begin{cases} ax_1+b &= y_1 \\ ax_2+b &= y_2 \\ & \vdots \\ ax_n+b &= y_n \end{cases} \quad \rightsquigarrow \quad \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$$\delta_1 = (ax_1+b)-y_1, \quad \delta_2 = (ax_2+b)-y_2, \quad \ldots, \quad \delta_n = (ax_n+b)-y_n.$$

$\sqrt{\delta_1^2+\delta_2^2+\ldots+\delta_n^2}$ is as small as possible.

$$\widehat{a} = \frac{n\Big(\sum_{i=1}^n x_i y_i\Big) - \Big(\sum_{i=1}^n x_i\Big)\Big(\sum_{i=1}^n y_i\Big)}{n\Big(\sum_{i=1}^n x_i^2\Big) - \Big(\sum_{i=1}^n x_i\Big)^2} \qquad \widehat{b} = \frac{1}{n}\Big(\sum_{i=1}^n y_i - \widehat{a}\sum_{i=1}^n x_i\Big),$$

$$y = ax+b$$

$$y = \widehat{a}\,x + \widehat{b}$$

Least Squares Approximation

- 參考資料：http://www.ms.uky.edu/~ma138/Spring15/Curve_fitting.pdf

t (sec)	0.5	1.1	1.5	2.1	2.3
T (°C)	32.0	33.0	34.2	35.1	35.7

$T = at + b,$

$$\begin{cases} 0.5\,a + b = 32.0 \\ 1.1\,a + b = 33.0 \\ 1.5\,a + b = 34.2 \\ 2.1\,a + b = 35.1 \\ 2.3\,a + b = 35.7 \end{cases} \rightsquigarrow \begin{bmatrix} 0.5 & 1 \\ 1.1 & 1 \\ 1.5 & 1 \\ 2.1 & 1 \\ 2.3 & 1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 32.0 \\ 33.0 \\ 34.2 \\ 35.1 \\ 35.7 \end{bmatrix}.$$

$$A^TA = \begin{bmatrix} 0.5 & 1.1 & 1.5 & 2.1 & 2.3 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 0.5 & 1 \\ 1.1 & 1 \\ 1.5 & 1 \\ 2.1 & 1 \\ 2.3 & 1 \end{bmatrix} = \begin{bmatrix} 13.41 & 7.5 \\ 7.5 & 5 \end{bmatrix}$$

$$A^T\mathbf{b} = \begin{bmatrix} 0.5 & 1.1 & 1.5 & 2.1 & 2.3 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 32.0 \\ 33.0 \\ 34.2 \\ 35.1 \\ 35.7 \end{bmatrix} = \begin{bmatrix} 259.42 \\ 170 \end{bmatrix}$$

$$\begin{bmatrix} 13.41 & 7.5 \\ 7.5 & 5 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 259.42 \\ 170 \end{bmatrix}$$

$$\left[\begin{array}{cc|c} 13.41 & 7.5 & 259.42 \\ 7.5 & 5 & 170 \end{array}\right] \text{ is equivalent to } \left[\begin{array}{cc|c} 1 & 0 & 2.0463 \\ 0 & 1 & 30.93 \end{array}\right]$$

$\widehat{a} = 2.0463$ and $\widehat{b} = 30.93.$ $T(t) = 2.0463\,t + 30.93$

year	1980	1985	1990	1995
population	227	237	249	262

$P(t) = at + b.$

$$\begin{bmatrix} 0 & 1 \\ 5 & 1 \\ 10 & 1 \\ 15 & 1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 227 \\ 237 \\ 249 \\ 262 \end{bmatrix}$$

$$A^TA = \begin{bmatrix} 0 & 5 & 10 & 15 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 5 & 1 \\ 10 & 1 \\ 15 & 1 \end{bmatrix} = \begin{bmatrix} 350 & 30 \\ 30 & 4 \end{bmatrix}$$

$$A^T\mathbf{b} = \begin{bmatrix} 0 & 5 & 10 & 15 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 227 \\ 237 \\ 249 \\ 262 \end{bmatrix} = \begin{bmatrix} 7605 \\ 975 \end{bmatrix}$$

$$\begin{bmatrix} 350 & 30 \\ 30 & 4 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 7605 \\ 975 \end{bmatrix}$$

$$\left[\begin{array}{cc|c} 350 & 30 & 7605 \\ 30 & 4 & 975 \end{array}\right] \text{ is equivalent to } \left[\begin{array}{cc|c} 1 & 0 & 117/50 \\ 0 & 1 & 1131/5 \end{array}\right]$$

$P(t) = 117/50 \cdot t + 1131/5.$