106-1 生物統計學二 實習課

R : Linear Regression

周芷妤

d04849010@ntu.edu.tw

2017.10.12

大綱

- Review

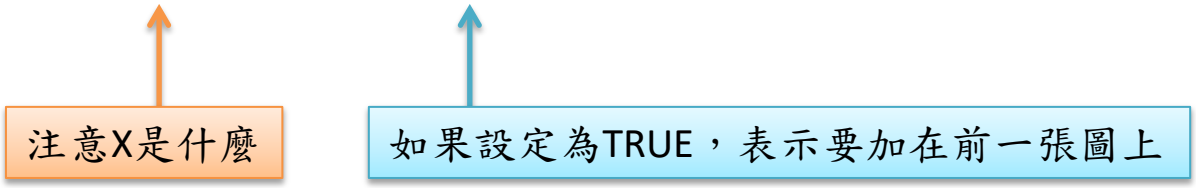
- Linear Regression
  - Diagnosis of regression model
  - Dummy variable

Review

Review

- matrix plot

matplot(X, multi.y, add = TRUE, …)



Linear Regression

Diagnosis of regression model
Dummy variable

Residual

Model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad , \quad \varepsilon_i \sim N(0,\ \sigma^2) \quad , i = 1, \ldots, n$$

誤差 ε_i

$$\varepsilon_i = Y_i - \beta_0 - \beta_1 X_i$$

利用e_i估計ε_i，進而估計σ^2

殘差 e_i

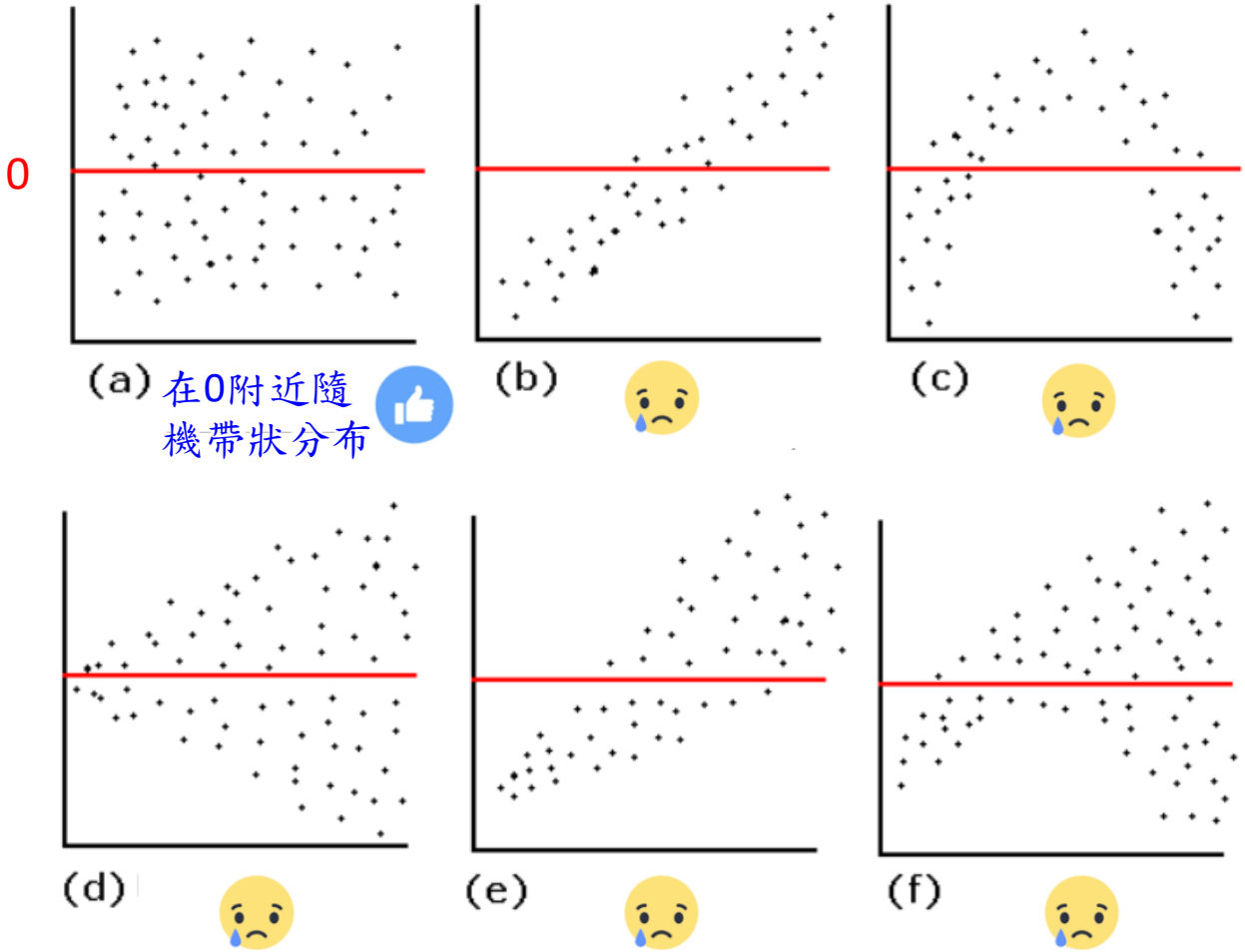
$$e_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i = Y_i - \hat{Y}_i$$

$$\rightarrow \hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} e_i^2 = \text{MSE}$$

Residual analysis

Assumption
- $E[Y|X] \perp \varepsilon$
- $X \perp \varepsilon$
- $i \perp \varepsilon$
- Normality of ε

Y軸: e_i　　X軸: i or X_i or $\hat{Y}_i$



Residual plot

```
fit <- lm(Petal.Width ~ Petal.Length, data = iris)
# 定義變項
e <- fit$residuals
y_hat <- fit$fitted.values
x <- iris$Petal.Length
n <- length(x)
```

為了方便，也可不重新定義，直接使用原變項名稱

```
par(mfrow = c(2,2))
```

par(mfrow=c(*nrows*, *ncols*))
將圖以*nrows*列×*ncols*行合併成一張，以列的方式排滿後再換至下一列

$\hat{Y}_i$ vs e_i

```
plot(y_hat, e, main="Fitted values vs Residuals",
     xlab = expression(hat(Y[i])), ylab = expression(e[i]))
```

X_i vs e_i

```
plot(x, e, main="X vs Residuals",
     xlab = expression(X[i]), ylab = expression(e[i]))
```

i vs e_i

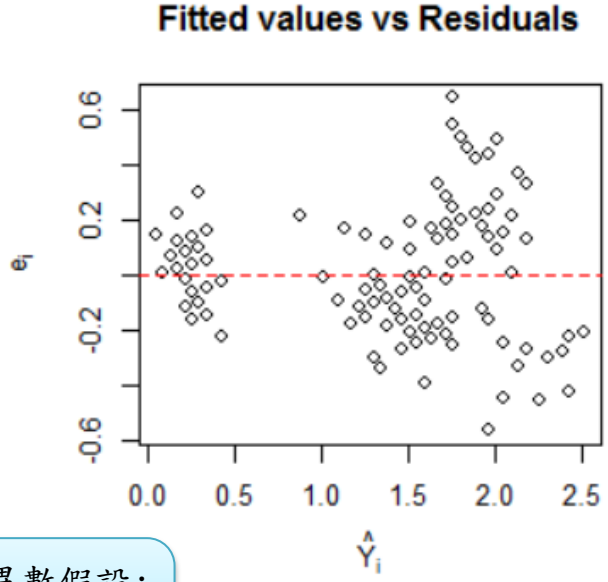
```
plot(1:n, e, main="i vs Residuals",
     xlab = "i", ylab = expression(e[i]))
```

Q – Q plot

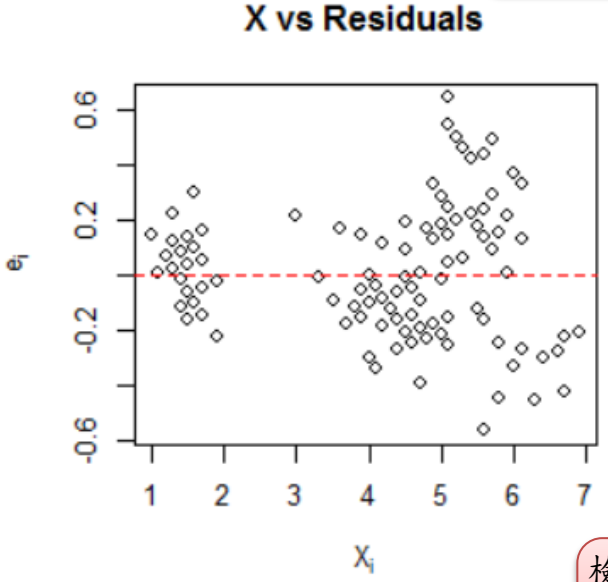
```
qqnorm(e)
qqline(e)
```

若要加上 $e = 0$ 的輔助線
abline(0, 0, lty = 2, col = "red")

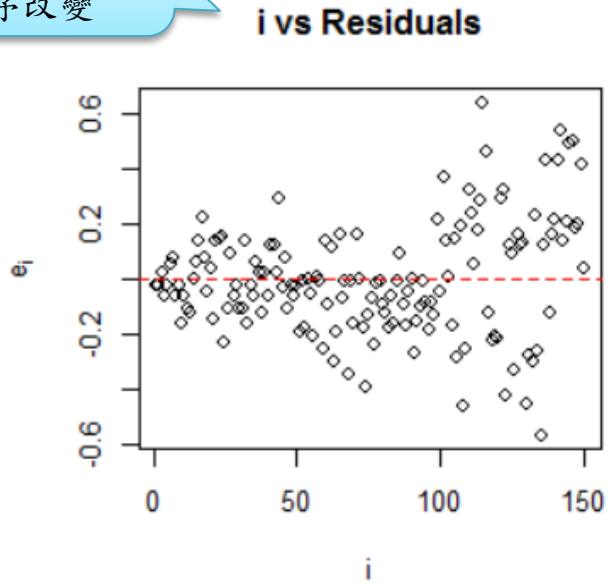
檢查同質變異數假設:
殘差的變異是否會隨$\hat{Y}$改變



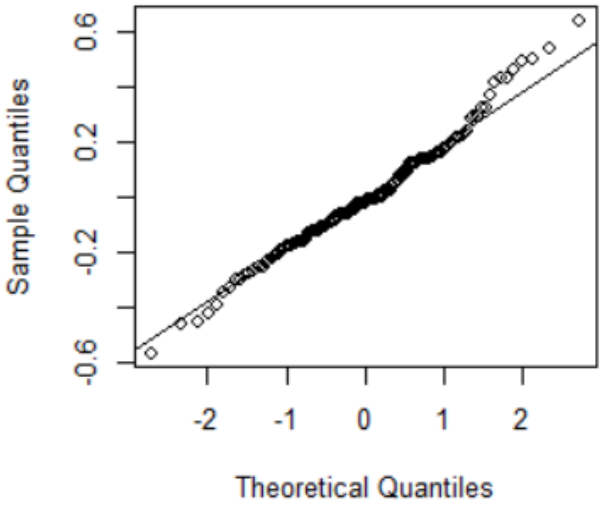
檢查同質變異數假設:
殘差的變異是否會隨X改變



檢查同質變異數假設:
殘差的變異是否會隨
資料收集順序改變



檢查Normality of ε：
若資料點越接近45°線
表示越服從Normal



課堂練習

- **FEV (forced expiratory volume)**：兒童肺功能是否受到身高影響?

  ＊資料檔：FEV.csv (逗號分隔)

1. 請利用lm先進行迴歸分析

2. 請依據第1題所建立的模型及結果，進行殘差分析，檢查誤差
   - 是否符合同質變異數假設？
   - 是否服從常態分布?
   (可以試著用一張圖來呈現所有殘差圖)

Dummy variable

- 當解釋變項(X)包含類別變項時，其代碼可能沒有數值上的意義或並非數值，因此利用dummy variables來建立迴歸模型

- 針對一個具有q (> 2) 種類別的變項，挑選特定一個類別作為對照組(reference)，產生$q-1$個dummy variables

- 以**lm**進行迴歸分析時，只要變項屬於factor，R會自動轉成dummy variables
  - ✓ 先確認變項是否為factor → **is.factor(**變項名稱**)**
  - ✓ 若不是，則改變變項屬性 → **as.factor(**變項名稱**)**

- 可自行定義dummy variables

- 將連續變項轉類別：對於數值不感興趣，想看特定族群相對於對照組的影響

Coding book

地區	d1	d2
北	1	0
中	0	1
南	0	0

Reference →

Fit linear model 類別變項

(1) 直接用類別變項進行迴歸

```
> attach(iris)
> levels(Species)          # 查看有哪些類別
[1] "setosa"     "versicolor" "virginica"
> is.factor(Species)
[1] TRUE
>
> fit_1 <- lm(Petal.Width ~ Species)
> summary(fit_1)

Call:
lm(formula = Petal.Width ~ Species)

Residuals:
    Min      1Q  Median      3Q     Max
-0.626 -0.126 -0.026   0.154   0.474

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        0.24600    0.02894    8.50 1.96e-14 ***
Speciesversicolor  1.08000    0.04093   26.39  < 2e-16 ***
Speciesvirginica   1.78000    0.04093   43.49  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2047 on 147 degrees of freedom
Multiple R-squared:  0.9289,    Adjusted R-squared:  0.9279
F-statistic:   960 on 2 and 147 DF,  p-value: < 2.2e-16
```

Coding book

Species		
versicolor	1	0
virginica	0	1
setosa	0	0

"versicolor組"比起"setosa組"
Petal.Width平均多1.08單位

Relevel & contrasts

- 改變reference

  relevel(變項名稱 **,"**欲當作reference的類別名稱**")**

- 查看coding的方式

  contrasts(變項名稱 **)**

```
> iris$Species_new <- relevel( Species, "versicolor" )
> levels(iris$Species_new)
[1] "versicolor" "setosa"     "virginica"
> contrasts(iris$Species_new)
           setosa virginica
versicolor      0         0
setosa          1         0
virginica       0         1
```

dummy的設定不同，會有什麼差異呢?

ifelse

ifelse(條件, a, b)
若滿足條件，則給a值，否則給b值

(2) 自行定義Dummy variables

ifelse(變項名稱 == "類別名稱", 1, 0 **)**

```
> # 建立Species的dummy variables
> # 以setosa作為reference group
> iris$d1 <- ifelse( Species=="versicolor", 1, 0)
> iris$d2 <- ifelse( Species=="virginica", 1, 0)
> View(iris)      # 檢視資料

> fit_2 <- lm(Petal.Width ~ d1 + d2, data = iris)
> summary(fit_2)

Call:
lm(formula = Petal.Width ~ d1 + d2, data = iris)

Residuals:
    Min      1Q  Median      3Q     Max
-0.626 -0.126 -0.026   0.154   0.474

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.24600    0.02894    8.50 1.96e-14 ***
d1           1.08000    0.04093   26.39  < 2e-16 ***
d2           1.78000    0.04093   43.49  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2047 on 147 degrees of freedom
Multiple R-squared:  0.9289,    Adjusted R-squared:  0.9279
F-statistic:   960 on 2 and 147 DF,  p-value: < 2.2e-16
```

Coding book

Species	d1	d2
versicolor	1	0
virginica	0	1
setosa	0	0

ifelse 延伸

(3) 連續二分

ifelse(條件1, a, **ifelse**(條件2, b, c))

ifelse(變項名稱 == "類別名稱a", 2, ifelse(變項名稱 == "類別名稱b", 1, 0))

```
> iris$D <- ifelse( Species=="versicolor", 2, ifelse( Species=="virginica", 1, 0) )
> iris$D <- as.factor(iris$D)
> levels(iris$D)
[1] "0" "1" "2"
>
> fit_3 <- lm(Petal.Width ~ D, data = iris)
> summary(fit_3)

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

Residuals:
    Min      1Q  Median      3Q     Max
-0.626 -0.126 -0.026   0.154   0.474

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.24600    0.02894    8.50 1.96e-14 ***
D1           1.78000    0.04093   43.49  < 2e-16 ***
D2           1.08000    0.04093   26.39  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2047 on 147 degrees of freedom
Multiple R-squared:  0.9289,    Adjusted R-squared:  0.9279
F-statistic:   960 on 2 and 147 DF,  p-value: < 2.2e-16
```

Species	D
versicolor	2
virginica	1
setosa	0

將連續變項轉成類別

```
> # 指定新變項Petal.Length_2 : 將Petal.Length以平均數分成兩類
> iris$Petal.Length_2 <- ifelse( Petal.Length > mean(Petal.Length), 1, 0)
> iris$Petal.Length_2 <- as.factor(iris$Petal.Length_2)
>
> fit_4 <- lm(Petal.Width ~ Petal.Length_2, data = iris)
> summary(fit_4)

Call:
lm(formula = Petal.Width ~ Petal.Length_2, data = iris)

Residuals:
    Min      1Q  Median      3Q     Max
-0.7226 -0.2226 -0.1341  0.1774  0.9544

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.34561    0.04816   7.177  3.2e-11 ***
Petal.Length_21  1.37697    0.06116  22.514  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3636 on 148 degrees of freedom
Multiple R-squared:  0.774,     Adjusted R-squared:  0.7725
F-statistic: 506.9 on 1 and 148 DF,  p-value: < 2.2e-16
```

將Petal.Length值大於平均者，設定為1，否則為0

課堂練習

- **Q：家鄉是否會影響身高?**

  ＊資料檔：MyData2.csv (逗號分隔)

1. 請以home= "S" 作為對照組進行迴歸分析
   - dummy variable的設定方式為? (coding book)
   - Mean response的估計式為? (符號請定義清楚)
   - 根據報表結果，對迴歸係數作解釋
   - 根據報表結果，回答問題"家鄉是否會影響身高"
     (根據什麼結果做了什麼結論)

2. 輸出新資料
   - 請將weight以中位數(median)做二分，
     並定義為新變項 (名稱自訂)
   - 將新資料匯出成新檔案(.csv)

Coding Book	
變項名稱	變項描述
id	ID number
sex	Male or Female
weight	Weight in kg
height	Height in cm
home	"N"：北 "M"：中 "S"：南 "other"：其他