106-1 生物統計學二 實習課

R : Simple Linear Regression

周芷妤

d04849010@ntu.edu.tw

2017.10.05

大綱

- Review

- Simple Linear Regression
  - Mean response & Prediction
  - Diagnosis of regression model

Review

Review

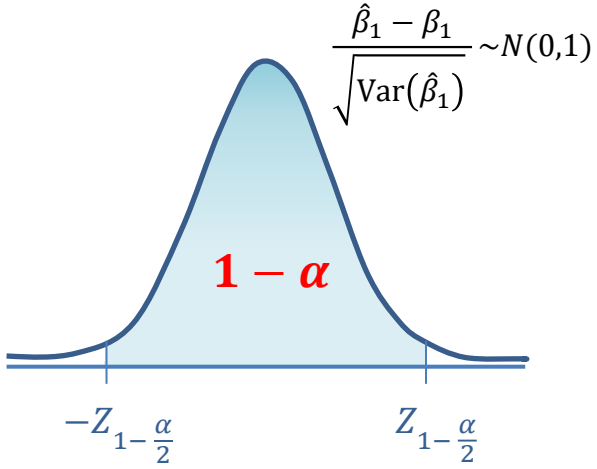
- Simple linear regression
  - ✓ Homework (extra)
  - ✓ Confidence interval

Confidence interval of β_1

- $H_0: \beta_1 = \beta_1^*$ vs. $H_1: \beta_1 \neq \beta_1^*$

假設σ^2已知

- $\hat{\beta}_1 \sim N(\beta_1, \mathrm{Var}(\hat{\beta}_1))$



$$P\left(-Z_{1-\frac{\alpha}{2}} \leq \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\mathrm{Var}(\hat{\beta}_1)}} \leq Z_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$

$$\Rightarrow \; P\left(\hat{\beta}_1 - Z_{1-\frac{\alpha}{2}} \times \sqrt{\mathrm{Var}(\hat{\beta}_1)} \leq \beta_1 \leq \hat{\beta}_1 + Z_{1-\frac{\alpha}{2}} \times \sqrt{\mathrm{Var}(\hat{\beta}_1)}\right) = 1-\alpha$$

$$\Rightarrow (1-\alpha) \times 100\% \text{ C.I. of } \beta_1 : \left[\; \hat{\beta}_1 - Z_{1-\frac{\alpha}{2}} \times \sqrt{\mathrm{Var}(\hat{\beta}_1)} \;,\; \hat{\beta}_1 + Z_{1-\frac{\alpha}{2}} \times \sqrt{\mathrm{Var}(\hat{\beta}_1)} \;\right]$$

若β_1^*不落在此區間，則拒絕H_0

- R 語法:

confint(model名稱 , level = 1- α)

Simple Linear Regression

Mean response & Prediction
Diagnosis of regression model

利用迴歸模型預測 Y

	Mean response $\mathrm{E}[Y\|X] = \beta_0 + \beta_1 X$ **at** X	**Predicted response** $Y_0 = \beta_0 + \beta_1 X + \varepsilon_0$ **at** X
Estimation	$\widehat{\mathrm{E}}[Y\|X] = \hat{\beta}_0 + \hat{\beta}_1 X$	$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 X$
$(1-\alpha)\times 100\%$ C.I.	$\widehat{\mathrm{E}}[Y\|X] \pm \mathrm{t}_{n-2,1-\frac{\alpha}{2}} \times \sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{(X-\bar{X})^2}{S_{XX}}\right)}$	$\hat{Y}_0 \pm \mathrm{t}_{n-2,1-\frac{\alpha}{2}} \times \sqrt{\hat{\sigma}^2\left(\mathbf{1} + \frac{1}{n} + \frac{(X-\bar{X})^2}{S_{XX}}\right)}$
例如	體重=50 之一群人的**平均**身高	體重=50 之**某一人**的身高

Prediction

Estimation of $\mathrm{E}[Y|X]$
Prediction of Y_0

- Y的預測及信賴區間

predict(model, newdata, interval = "confidence", level = 0.95)

語法	說明	備註
newdata	An optional data frame in which to look for variables with which to predict.	• 需指定成data.frame之形式 • 若省略，則使用原資料作預測
interval	計算的區間類型	"none" : 不計算區間(預設) "confidence" : $\mathrm{E}[Y\|X]$ 之 C.I. "prediction" : Y_0 之 C.I.
level	指定信心水準	預設為level= 0.95

Example : IRIS data

```
attach(iris)
fit <- lm(Petal.Width ~ Petal.Length)      # 建立迴歸模型
```

- 計算Mean response 的信賴區間

```
> pred.clim <- predict(fit, interval = "confidence")
> pred.clim
           fit           lwr       upr
1   0.21898206  0.163271179 0.2746929
2   0.21898206  0.163271179 0.2746929
3   0.17740652  0.120166743 0.2346463
4   0.26055760  0.206352581 0.3147626
```

- 計算Predicted response 的信賴區間

```
> pred.plim <- predict(fit, interval = "prediction")
Warning message:
In predict.lm(fit, interval = "prediction") :
  predictions on current data refer to _future_ responses

> pred.plim
           fit          lwr       upr
1   0.21898206 -0.19284194 0.6308061
2   0.21898206 -0.19284194 0.6308061
3   0.17740652 -0.23462710 0.5894401
4   0.26055760 -0.15106540 0.6721806
```

Example : IRIS data

- 指定 covariate (X) 值

```
X_new <-  seq(min(Petal.Length), max(Petal.Length), by = 0.1)
```

- Mean response 的信賴區間

將 X_new 代入Petal.Length 計算

```
> pred.clim2 <- predict(fit, newdata = data.frame(Petal.Length = X_new), interval = "confidence")
> pred.clim2
         fit          lwr       upr
1  0.05267990 -0.009267579 0.1146274
2  0.09425544  0.033895801 0.1546151
3  0.13583098  0.077041072 0.1946209
4  0.17740652  0.120166743 0.2346463
```

- Predicted response 的信賴區間

```
> pred.plim2 <- predict(fit, newdata = data.frame(Petal.Length = X_new), interval = "prediction")
> pred.plim2
         fit         lwr       upr
1  0.05267990 -0.36003405 0.4653938
2  0.09425544 -0.31822316 0.5067340
3  0.13583098 -0.27642084 0.5480828
4  0.17740652 -0.23462710 0.5894401
```

Add C.I. to scatter plot

- 繪製迴歸線 + 兩種信賴區間的圖

(1)先將五條線Y軸的值合併

- 用cbind將矩陣以行的方式合併
- 用[, -1]將矩陣第一行拿掉

multi.y <- **cbind**(pred.clim, pred.plim[,-1])

```
> pred.clim
        fit       lwr       upr
1 0.21898206 0.163271179 0.2746929
2 0.21898206 0.163271179 0.2746929
3 0.17740652 0.120166743 0.2346463
4 0.26055760 0.206352581 0.3147626
```

```
> pred.plim
        fit       lwr       upr
1 0.21898206 -0.19284194 0.6308061
2 0.21898206 -0.19284194 0.6308061
3 0.17740652 -0.23462710 0.5894401
4 0.26055760 -0.15106540 0.6721806
```

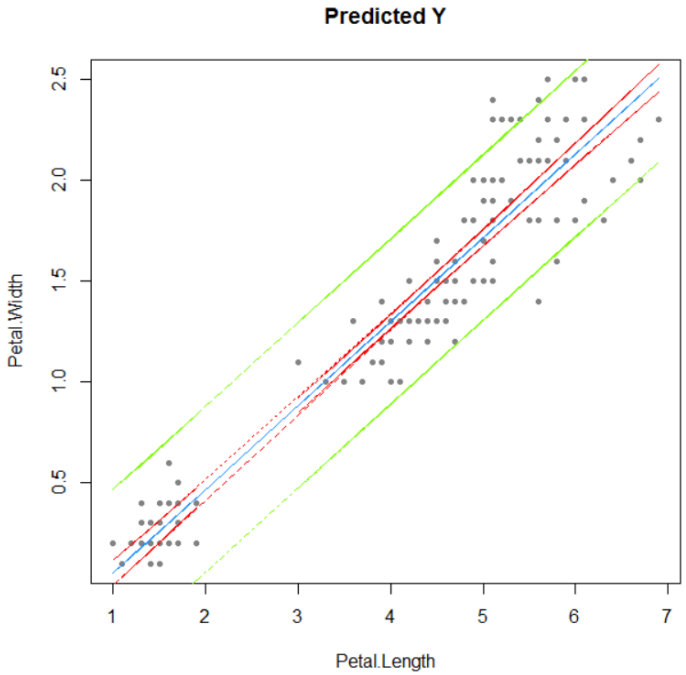
```
> multi.y
        fit       lwr       upr       lwr       upr
1 0.21898206 0.163271179 0.2746929 -0.19284194 0.6308061
2 0.21898206 0.163271179 0.2746929 -0.19284194 0.6308061
3 0.17740652 0.120166743 0.2346463 -0.23462710 0.5894401
4 0.26055760 0.206352581 0.3147626 -0.15106540 0.6721806
```

Add C.I. to scatter plot

(2) 將五條線畫在同一張圖(scatter plot)上

plot(Petal.Length, Petal.Width, main = "Predicted Y", pch=20, col="gray51")

matplot(Petal.Length, multi.y, add = TRUE, type = "l",
col = c('dodgerblue','red','red','lawngreen','lawngreen'))



語法	說明	備註
pch	point character	詳見補充
add	是否加在原先畫的圖上	預設為 FALSE
type	以什麼形式畫圖	"p" : points "l" : lines "b" : both ,…,etc.

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. 請先繪製 Height vs FEV 之scatter plot (請注意X 和 Y分別是什麼變項)

2. 請建立迴歸模型，並分別計算 mean response 和 predicted response 之 95%信賴區間

3. 請在scatter plot 上加迴歸線、mean response 和 predicted response 之95%信賴區間

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

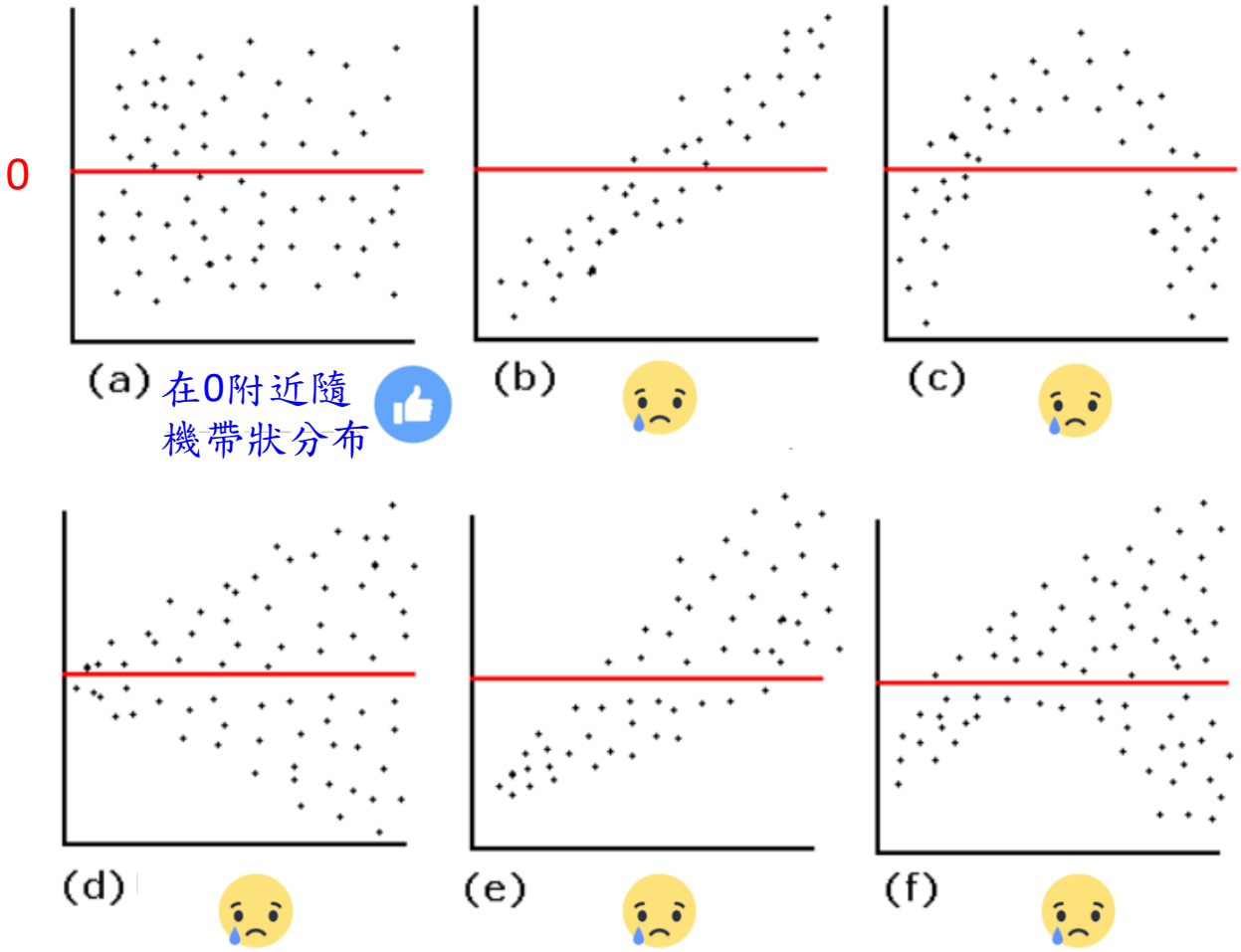
$$e_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i = Y_i - \hat{Y}_i$$

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

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

# 定義變項
e <- fit$residuals
y_hat <- fit$fitted.values
x <- iris$Petal.Length
n <- length(x)
```

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

為了方便，也可不重新定義，直接使用原變項名稱

$\hat{Y}_i$ 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

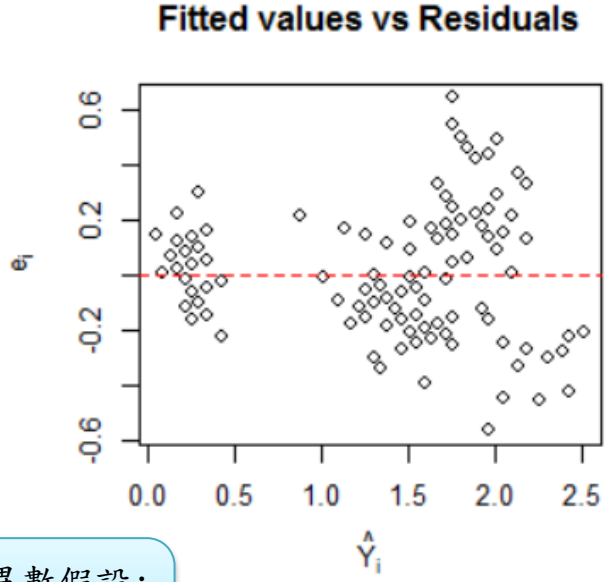
```
plot(1:n, e, main="i vs Residuals",
     xlab = "i", ylab = expression(e[i]))
```

Q – Q plot

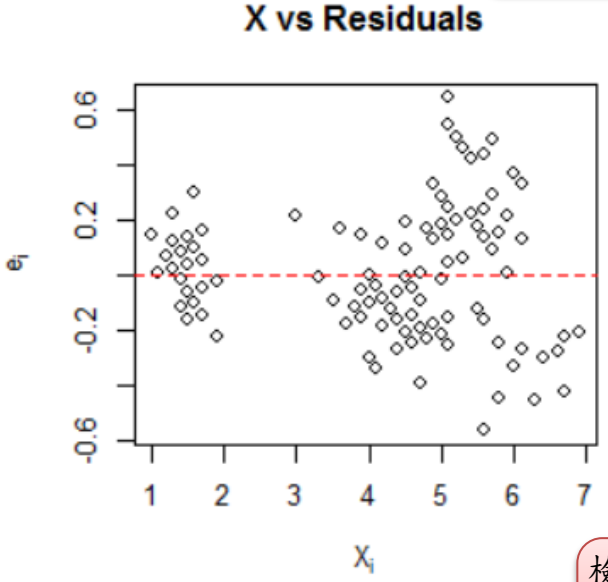
```
qqnorm(e)
qqline(e)
```

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

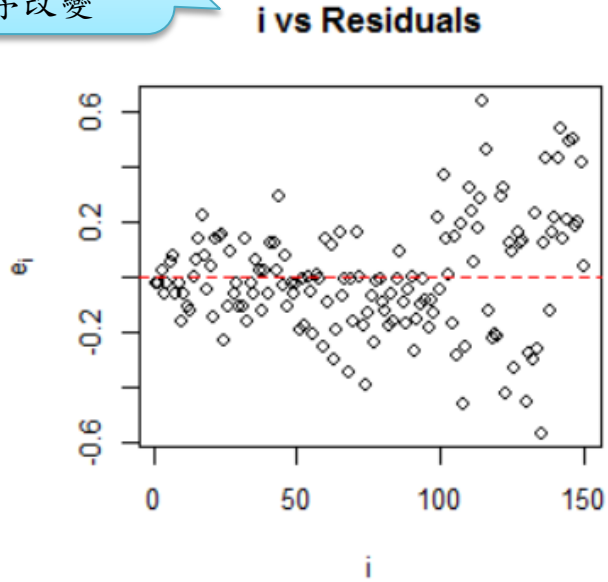
檢查同質變異數假設:
殘差的變異是否會隨$\hat{Y}$改變



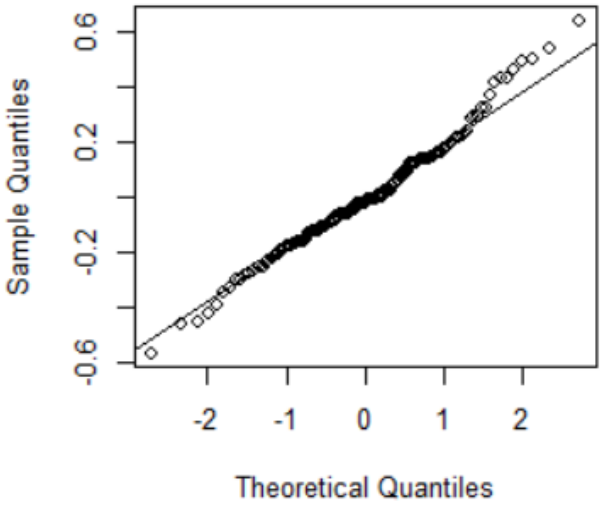
檢查同質變異數假設:
殘差的變異是否會隨X改變



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

4. 請依據第2題所建立的模型，進行殘差分析，檢查誤差

(1) 是否符合同質變異數假設？

(2) 是否服從常態分布?

(試著以一張圖來呈現所有殘差圖)

Homework

- **FEV (forced expiratory volume)**：兒童肺功能是否受到抽菸影響?

＊請先匯入資料檔(需附上code)： fev.csv (逗號分隔)

1. 請將資料分成有抽菸 (S) 及沒抽菸 (N) 兩組，以年齡(Age)為 X 軸，FEV 為 Y 軸畫出scatter plot (指定 pch = 9)

2. 請對(S)、(N) 兩組分別建立simple linear regression來描述Age如何影響 FEV，並回答下列問題
   - 年齡每增加一歲，對於這兩組FEV 分別會?
   - 兩組模型解釋變異程度為?
   - 兩組兒童年齡是否顯著影響 FEV?
   - 根據上述分析結果，是否能得到“兒童肺功能受到抽菸影響”的結論? (若是，請敘述依據什麼結果得到什麼結論；若不行，請說明理由。)

Coding Book	
變項名稱	變項描述
ID	ID number
Age	children ages in years
FEV	forced expiratory volume in liters
Height	Height in inches
Sex	Male or Female
Smoker	Non = nonsmoker Current = current smoker

Homework

3. 請分別對兩組的scatter plot加上迴歸線、mean response 和 predicted response之90%信賴區間 (指定 type = "l"，請給定新的X: 分別將兩組Age從最小到最大以0.1為間隔產生)

4. 新個案 K：男性、抽菸、15 歲、身高為 57 inches，想利用第 2 題模型來預測 FEV 。請回答下列問題
   - mean response的估計($\widehat{\mathrm{E}}[Y|X]$) 及95%信賴區間為多少?
   - predicted response的估計($\hat{Y}_K$) 及95%信賴區間為多少?

5. 依據第4題所使用的模型，請問誤差
   - 是否符合同質變異數假設 (homoscedasticity) ?
   - 是否服從常態分布?
   (畫殘差圖，並根據圖作說明)

6. 請將code及output貼上word檔(可存成pdf檔)，上傳至ceiba作業區
   *最晚上傳期限為2017.10.8(日)中午12點

補充 繪圖相關

- Color ： http://www.stat.columbia.edu/~tzheng/files/Rcolor.pdf