Simple Linear Regression

簡單線性回歸

助教：賴以勳、温沛得、張祐瑄

麻煩同學們填寫期末教學評鑑
Please~

簡單直線回歸

- 相關係數
- 最小平方法 →估計截距(β_0)、斜率(β_1)
- 檢定β_0, β_1
- 變方分析→決定係數
- 檢查假設(獨立、變方同質、常態性)

簡單直線回歸

(2) Sampling

Population

(1)若是Y,X有線性關係存在，線性關係為

$\mathrm{Y} = \beta_0 + \beta_1 \mathrm{X} + \epsilon$

Sample

y_i, x_i

y_i跟x_i的關係為$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
但不知道β_0和β_1

Need assumptions:
$Cov(\epsilon_i, \epsilon_j) = 0, Var(\epsilon_i) = \sigma^2$

(3) Inference

1.最小平方法估計β_0, β_1
$\rightarrow b_0, b \rightarrow \hat{y} = b_0 + b_1 x_i$
2.檢定β_0, β_1是否為0

(4) Check assumptions

Need assumptions: $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$

- 相關係數:量化變數間的相關性

$$\rho = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \rightarrow \hat{\rho} = r = \frac{S_{XY}}{S_X S_Y} = \frac{\sum_n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_n (x_i - \bar{x})^2 \sum_n (y_i - \bar{y})^2}}$$

$$(-1 \leq \rho \leq 1)$$

相關係數ρ 與回歸模型中的係數β不同

簡單直線回歸

➢ 最小平方法:估計β_0, β_1

$$\sum_n \epsilon_i^2 = \sum_n (y_i - \beta_0 - \beta_1 x_i)^2 \rightarrow \text{找到使誤差平方項最小的}\beta_0, \beta_i$$

→ 求極值(微分=0)

$$\rightarrow \hat{\beta} = b = \frac{\sum_n (x_i - \bar{x})\,(y_i - \bar{y})}{\sum_n (x_i - \bar{x})^2}$$

$$\rightarrow \widehat{\beta_0} = b_0 = \bar{y} - b\bar{x}$$

$$\hat{y}_i = \widehat{\beta_0} + \hat{\beta} x_i$$

$$y_i = \hat{y}_i + e_i = \widehat{\beta_0} + \hat{\beta} x_i + e_i$$

簡單直線回歸

➢ 檢定β_0, β_1

$$Assumption: \epsilon \sim N(0,\sigma^2) \rightarrow \frac{(n-2)\widehat{\sigma^2}}{\sigma^2} \sim \chi^2_{n-2}$$

$$H_0: \beta_0 = 0, H_a: \beta_0 \neq 0$$

$$\widehat{\beta_0} \sim \mathrm{N}\left(\beta_0, (\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})\sigma^2\right) \rightarrow \frac{\widehat{\beta_0} - \beta_0}{\sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right) MSE}} \sim t_{n-2}$$

$$\left|\frac{\widehat{\beta_0} - \beta_0}{\sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right) MSE}}\right| > t_{1-\alpha/2,n-2} \rightarrow reject\ H_0$$

$$H_0: \beta = 0, H_a: \beta \neq 0$$

$$\hat{\beta} \sim \mathrm{N}\left(\beta, \frac{\sigma^2}{S_{xx}}\right) \rightarrow \frac{\hat{\beta} - \beta}{\sqrt{\frac{MSE}{S_{xx}}}} \sim t_{n-2}$$

$$\left|\frac{\hat{\beta} - \beta}{\sqrt{\frac{MSE}{S_{xx}}}}\right| > t_{1-\alpha/2,n-2} \rightarrow reject\ H_0$$

➢ 變方分析&決定係數

SOV	df	SS	MS	F
$Regression$	1	$SSR=\sum_{n}(\hat{y}_i-\bar{y})^2$	$MSR=SSR/1$	MSR/MSE
$Error$	$n-2$	$SSE=\sum_{n}(\hat{y}_i-y_i)^2$	$MSE=SSE/(n-2)$	$H_0:\beta=0, H_a:\beta\neq 0$
$Total$	$n-1$	$SST=\sum_{n}(y_i-\bar{y}))^2$		

$$R^2=\frac{SSR}{SST}(0<R^2<1)$$

$R^2=1$表示變異可以完全由回歸模型解釋
$R^2=0$表示回歸模型一點解釋能力也沒有

簡單直線回歸:R

➢ 範例

data.txt 紀錄某地區1000多位母親與女兒的身高，想知道母親的身高與女兒的身高是否有線性關係存在，假設影響女兒身高的因子僅有母親的身高一個，請根據此假設建立一線性回歸模型，並完成變數係數檢定

簡單直線回歸:R

- Code

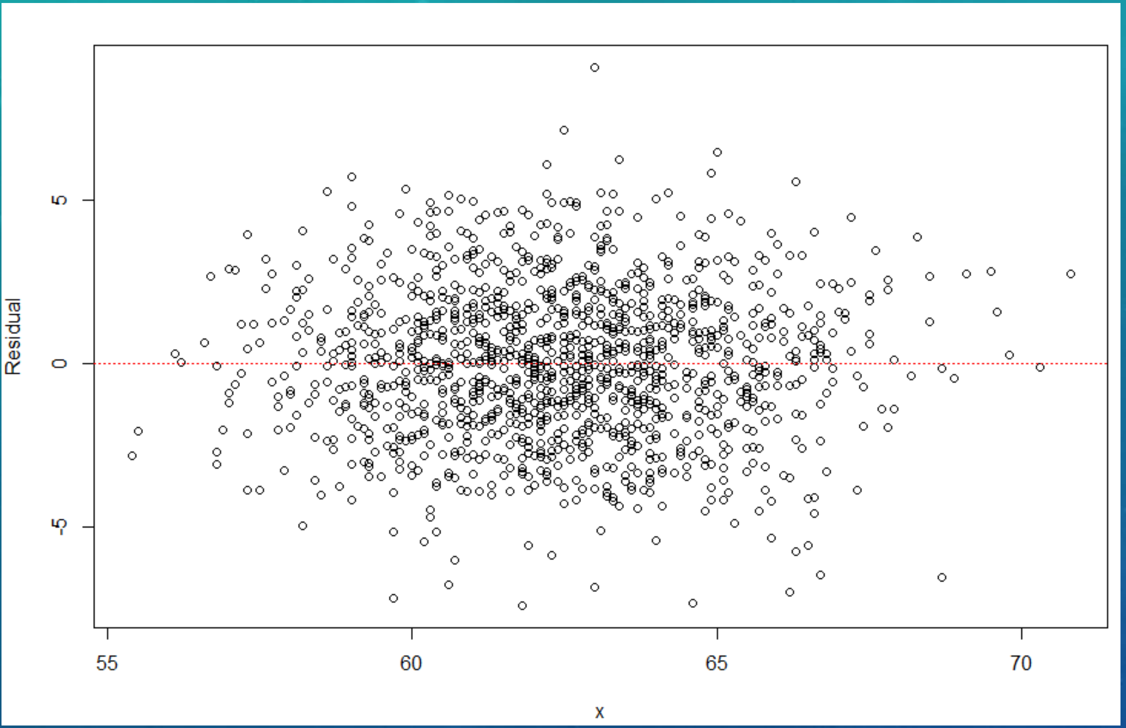
```
fit=lm(y~x)
summary(fit)
```

- Console

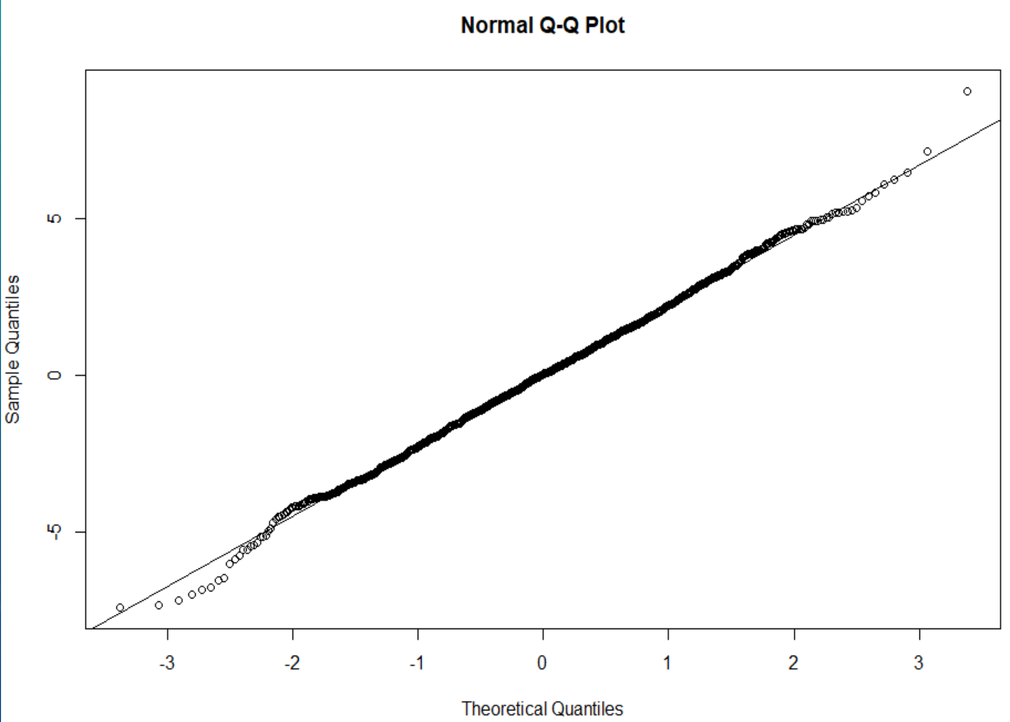
```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.397 -1.529  0.036  1.492  9.053 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 29.91744    1.62247   18.44   <2e-16 ***
x            0.54175    0.02596   20.87   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.266 on 1373 degrees of freedom
Multiple R-squared:  0.2408,	Adjusted R-squared:  0.2402 
F-statistic: 435.5 on 1 and 1373 DF,  p-value: < 2.2e-16
```

簡單直線回歸:R

➢ 檢查假設:獨立、變方同質、常態性



獨立性、變方同質



Q-Q plot

簡單直線回歸:R

- Code

```
### homogeneity of variance  ###
plot(x,fit$residuals,ylab="Residual")
abline(a=0,b=0,lty=3,col="red")

#### Q-Q plot@ ###
qqnorm(fit$residuals)
qqline(fit$residuals)
```

用殘差畫QQ plot

加線

Example 2

某化合物含有率(%)	溫度(°F)
92	182
91	185
91	186
89	188
88	190
86	193
87	194
86	195
87	192
85	197
88	192
90	185

助教信箱

張祐瑄：R09621203@ntu.edu.tw
賴以勳：R09621208@ntu.edu.tw
溫沛得：R10621204@ntu.edu.tw

祝各位同學都能all pass!!

Homework!!!!!!!