106-1 生物統計學二實習課

R : Linear Regression

周芷妤

d04849010@ntu.edu.tw

2017.10.19

大綱

- Review

- Linear Regression
  - Interaction
  - Partial F-test

Review

Review

- Linear regression
  - ✓ Homework 2
  - ✓ Homework 2 (Lab)

Linear Regression

Interaction
Partial F-test

Multiple Linear Regression

Linear Model	$Y = \mathrm{E}[\,Y \mid X_1, \ldots, X_p\,] + \varepsilon$ $= \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon \quad , \quad \varepsilon \sim N(0,\ \sigma^2)$
Estimation	• LSE $\rightarrow \ \hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)$ • MSE $\rightarrow \ \hat{\sigma}^2$
Hypothesis Testing	• 單一係數 $\rightarrow \ \mathrm{H}_0 : \beta_j = 0 \quad (j = 0, \ldots, p)$ • 整體係數 $\rightarrow \ \mathrm{H}_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$
Test statistic	• $\mathrm{t} = \dfrac{\hat{\beta}_j - 0}{\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_j)}} \overset{\mathrm{H}_0}{\sim} \mathrm{t}_{n-(p+1)}$ • $\mathrm{F} = \dfrac{\mathrm{SSR}/p}{\mathrm{SSE}/(n-(p+1))} = \dfrac{\mathrm{MSR}}{\mathrm{MSE}} \overset{\mathrm{H}_0}{\sim} \mathrm{F}_{p,\, n-(p+1)}$

$\mathrm{F} = \mathrm{t}^2$ (only when $p = 1$)

Fit linear model 多個解釋變項

- 建立迴歸模型

  model <- **lm**(Y ~ X1 + X2 + … + Xp , data = 資料檔名稱)

- 產生model配適結果的總結

  summary(model)

- Analysis of Variance :

  anova(model)

變異來源	df	Sum of Square	Mean Square	F	p-value
Regression	p	SSR	MSR= $\frac{SSR}{p}$	$\frac{MSR}{MSE}$	p-value
Error	$n-(p+1)$	SSE	MSE= $\frac{SSE}{n-(p+1)}$		
Total	$n-1$	SST			

Example **Simple** Linear regression

Model

$Y = \beta_0 + \beta_1 X_1 + \varepsilon$,
$\varepsilon \sim N(0, \sigma^2)$

```
> attach(iris)
> fit_m0 <- lm(Petal.Width ~ Petal.Length)
> summary(fit_m0)

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

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

> anova(fit_m0)
Analysis of Variance Table

Response: Petal.Width
              Df Sum Sq Mean Sq F value    Pr(>F)
Petal.Length   1  80.26  80.260  1882.5 < 2.2e-16 ***
Residuals    148   6.31   0.043
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$F = t^2$

Example **Multiple** Linear regression

Model

$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$,

$\varepsilon \sim N(0, \sigma^2)$

```
> attach(iris)
> fit_m1 <- lm(Petal.Width ~ Petal.Length + Sepal.Length)
> summary(fit_m1)

Call:
lm(formula = Petal.Width ~ Petal.Length + Sepal.Length)

Residuals:
     Min       1Q   Median       3Q      Max
-0.60598 -0.12560 -0.02049  0.11616  0.59404

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.008996   0.182097  -0.049   0.9607
Petal.Length  0.449376   0.019365  23.205   <2e-16 ***
Sepal.Length -0.082218   0.041283  -1.992   0.0483 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2044 on 147 degrees of freedom
Multiple R-squared:  0.929,     Adjusted R-squared:  0.9281
F-statistic: 962.1 on 2 and 147 DF,  p-value: < 2.2e-16
```

在相同Sepal.Length下 (調整Sepal.Length後)，
Petal.Length 每增加一單位，
Petal.Width平均增加 0.45個單位

Interaction

以一項交互作用為例，其餘以此類推

- 多個解釋變數 X + 交互作用: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$

lm(Y ~ X1*X2, data = 資料檔名稱)

- R會自動將X_1和X_2放入model
- 也可先將交互作用項定義成新變數，但需自行將其他變數加入model中 (eg: X3 <- X1*X2)
- 如果只要放交互作用項用 X1:X2

```
> fit_m2 <- lm(Petal.Width ~ Petal.Length*Sepal.Length)
> summary(fit_m2)

Call:
lm(formula = Petal.Width ~ Petal.Length * Sepal.Length)

Residuals:
     Min       1Q   Median       3Q      Max
-0.60607 -0.12899 -0.02035  0.11236  0.58679

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               -0.44504    0.40548  -1.098    0.274
Petal.Length               0.53603    0.07457   7.188 3.14e-11 ***
Sepal.Length               0.00257    0.08164   0.031    0.975
Petal.Length:Sepal.Length -0.01658    0.01378  -1.203    0.231
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2041 on 146 degrees of freedom
Multiple R-squared:  0.9297,    Adjusted R-squared:  0.9283
F-statistic: 643.8 on 3 and 146 DF,  p-value: < 2.2e-16
```

Partial F-test

Full Model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \beta_{k+1} X_{k+1} + \cdots + \beta_p X_p + \varepsilon$$

Reduced Model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon$$

Null hypothesis

$$\mathrm{H}_0 : \beta_{k+1} = \beta_{k+2} = \cdots = \beta_p = 0$$

Test statistic

$$\mathrm{F}^* = \frac{\frac{\mathrm{SSR}_f - \mathrm{SSR}_r}{p-k}}{\frac{\mathrm{SSE}_f}{n-(p+1)}} \sim \mathrm{F}^{\mathrm{H}_0}_{p-k,\,n-p-1}$$

Example : IRIS data

$$\text{Petal.Width} = \beta_0 + \beta_1 \times \text{Petal.Length} + \varepsilon$$

$$\text{Petal.Width} = \beta_0 + \beta_1 \times \text{Petal.Length} + \beta_2 \times \text{Sepal.Length} + \beta_3(\text{Petal.Length} \times \text{Sepal.Length}) + \varepsilon$$

★ $H_0: \beta_2 = \beta_3 = 0$ vs. H_1: at least one $\beta_j \neq 0$, $j = 2, 3$

```
> attach(iris)
> ## Reduced:
> fit_red <- lm(Petal.Width ~ Petal.Length)
>
> ## Full:
> fit_full <- lm(Petal.Width ~ Petal.Length*Sepal.Length)
>
> anova(fit_red, fit_full)
Analysis of Variance Table

Model 1: Petal.Width ~ Petal.Length
Model 2: Petal.Width ~ Petal.Length * Sepal.Length
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    148 6.3101
2    146 6.0840  2    0.2261 2.7129 0.06969 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

在顯著水準$\alpha = 0.05$之下，
我們沒有足夠的證據拒絕H_0

error sum of square

Homework

example2：性別是否影響身高?

＊請先匯入資料檔(需附上code)： example2.csv (逗號分隔)

- **M1**:以height為反應變項(Y)， sex (以Male為reference)跟age為共變項(X)建立線性迴歸模型
  - 寫下建立的迴歸模型 (定義清楚符號代表的意思)
  - 在不同的性別下，mean response 的估計式分別為?
  - 根據分析結果，請問男女身高有無差異?若有差異，請說明有什麼樣的差異?

Coding Book	
變項名稱	變項描述
age	ages in years
sex	Male or Female
height	Height in cm
weight	Weight in kg

Homework

- **M2**:增加age與sex之交互作用項
  - 寫下建立的迴歸模型 (定義清楚符號代表的意思)
  - 寫下在不同的性別下，mean response 的估計式，並解釋其各自的意義
  - 請問男女身高有無差異?
    (進行Partial F test，寫出reduced、full model及檢定流程和結論)

- **M3**:增加weight變項
  - 與M1進行Partial F test (寫出reduced、full model及檢定流程和結論)
  - 如果要做決策，你會選擇M1還是M3?請說明根據什麼結果或理由
  - 針對M3繪製Q-Q plot跟"fitted values 與residuals"的殘差圖，並根據圖說明是否符合誤差假設

- 請將code及output貼上word檔(可存成pdf檔)，上傳至ceiba作業區