106-1 生物統計學二 實習課

R : Simple Linear Regression

周芷妤

d04849010@ntu.edu.tw

2017.09.28

大綱

- Review

- Simple Linear Regression
  - Coefficient of determination
  - Hypothesis testing：T-test、F-test
  - Confidence interval

Review

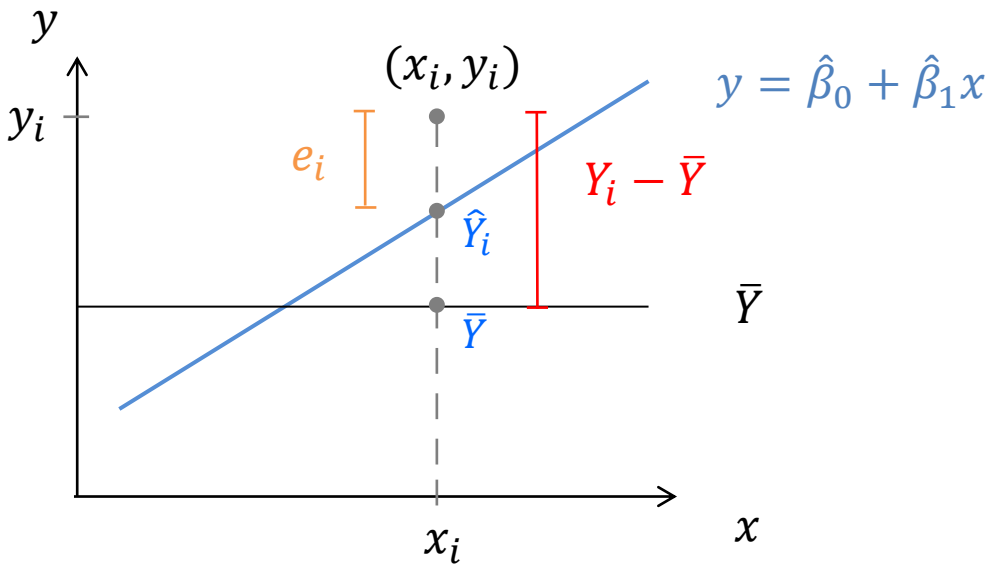
Review

- Simple linear regression
  - ✓ Homework1

Simple Linear Regression

Coefficient of determination
Hypothesis testing：T-test、F-test
Confidence interval

SST = SSE + SSR 變異數分解



$$\sum_i (Y_i - \bar{Y})^2 = \sum_i (Y_i - \hat{Y}_i)^2 + \sum_i (\hat{Y}_i - \bar{Y})^2$$

$$\mathbf{SST} = \mathbf{SSE} + \mathbf{SSR}$$

Total sum of square | Error sum of square | Regression sum of square

Coefficient of determination

Coefficient of determination	$R^2 = 1 - \frac{\sum_i (Y_i - \hat{Y}_i)^2}{\sum_i (Y_i - \bar{Y})^2} = 1 - \frac{SSE}{SST}$
Meaning	代表Model解釋了多少比例的變異 (亦即，迴歸可以解釋的部分)
Properties	• $0 \leq R^2 \leq 1$ (希望越大越好) → 但是可能會受到共變數(X)之個數的影響， (隨共變數個數增加，R^2越大) • $R^2 = [\,\rho(X, Y)\,]^2$ (only for $p = 1$)
Adjusted R^2	$R^2_{adj} = 1 - \frac{SSE/(n-(p+1))}{SST/(n-1)}$

Hypothesis testing

Simple Linear Model	$Y = \mathrm{E}[\,Y \mid X\,] + \varepsilon$ $= \beta_0 + \beta_1 X + \varepsilon \quad , \quad \varepsilon \sim N(0,\ \sigma^2)$
Hypothesis Testing	$\mathrm{H}_0 : \beta_1 = 0 \quad$ vs. $\quad \mathrm{H}_1 : \beta_1 \neq 0$
Test statistic	$\mathrm{t} = \dfrac{\hat{\beta}_1 - 0}{\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_1)}} = \dfrac{\hat{\beta}_1 - 0}{\sqrt{\frac{\hat{\sigma}^2}{S_{XX}}}} \overset{\mathrm{H}_0}{\sim} \mathrm{t}_{n-2}$
Confidence interval	$(1-\alpha) \times 100\%$ C.I. of β_1 : $\hat{\beta}_1 \pm \mathrm{t}_{n-2,1-\frac{\alpha}{2}} \times \sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_1)}$
$\mathrm{F} = \mathrm{t}^2$ (when $p = 1$)	$\mathrm{F} = \dfrac{\mathrm{SSR}/1}{\mathrm{SSE}/(n-2)} = \dfrac{\mathrm{MSR}}{\mathrm{MSE}} = \mathrm{t}^2 \overset{\mathrm{H}_0}{\sim} \mathrm{F}_{1,n-2}$

Fit linear model

- 建立迴歸模型
  - ✓ model <- lm(response ~ covariate , data = 資料檔名稱)
  - ✓ attach(資料檔名稱)
    model <- lm(response ~ covariate)
- 產生model配適結果的總結
  - ✓ summary(model)

Example : IRIS data

```
> fit <- lm(Petal.Width ~ Petal.Length, data = iris)
> summary(fit)

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

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
```

Petal.Width 作為response (Y)
Petal.Length 作為covariate (X)

$\hat{\sigma} = \sqrt{\text{MSE}}$

$$R^2 = 1 - \frac{SSE}{SST} \quad , \quad R^2_{adj} = 1 - \frac{SSE/(n-(p+1))}{SST/(n-1)}$$

Example : IRIS data

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

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
```

$$t = \frac{\hat{\beta}_1 - 0}{\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_1)}}$$

p - value

$F = t^2$

Confidence interval of $(\beta_0,\ \beta_1)$

- 計算迴歸參數$(\beta_0,\ \beta_1)$之$(1-\alpha)\times 100\%$信賴區間

confint(model名稱 , level = 1- α)

語法	說明	備註
level	指定信心水準	ex : $\alpha = 0.05 \rightarrow$ level= 0.95

- Example : IRIS data

```
> # 計算迴歸參數的信賴區間
> confint(fit, level=0.95)
                   2.5 %     97.5 %
(Intercept)   -0.4416501 -0.2845010
Petal.Length   0.3968193  0.4346915
```

95% C.I. of β_0 : $\hat{\beta}_0 \pm t_{n-2,1-\frac{\alpha}{2}} \times \sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_0)}$

95% C.I. of β_1 : $\hat{\beta}_1 \pm t_{n-2,1-\frac{\alpha}{2}} \times \sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_1)}$

課堂練習 (加分作業)

- **Inheritance of Height** ：女兒身高(Dheight)會不會受到母親身高(Mheight)的影響?

＊請先匯入資料檔： heights.txt (Tab分隔，單位: inch)

1.以Mheight為X軸、Dheight為Y軸：
- 請畫出scatter plot
- 並根據scatter plot 說明Mheight及Dheight的關係

2.請建立迴歸模型來描述此問題：
- 寫出迴歸模型_____、反應變數Y為_____、共變數X為_____
- 此模型**假設**X和Y呈現_____相關
- 計算X之迴歸係數的數值，截距為_____，斜率為_____，當X增加 1 inch， Y_____
- 判定係數 R^2 =_____

3.評估模型的好壞：
- 請在scatter plot 加上迴歸線
- 你認為此模型足以用來描述這筆資料嗎? (試以R^2、假說檢定或信賴區間來說明)

4. 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2017.9.30(六)中午12點