107-1 Statistics
LAB12: CORRELATION ANALYSIS - 2

助教：廖晧宇、吳家禎、賴冠宇

2018/12/14

CORRELATION ANALYSIS

一組樣本

$x: \{x_1 + x_2 + x_3 + \cdots\}$

$y: \{y_1 + y_2 + y_3 + \cdots\}$

跑簡單迴歸

樣本的迴歸

$\hat{y} = b_0 + b_1 x$ 估計值

$y = b_0 + b_1 x + e$ 實際值

推論統計

母體的迴歸

$\mathrm{E}(y) = \beta_0 + \beta_1 x$ 期望值

$y = \beta_0 + \beta_1 x + \varepsilon$ 實際值

CORRELATION ANALYSIS

Last week

樣本的迴歸

一組樣本
$x: \{x_1 + x_2 + x_3 + \cdots\}$
$y: \{y_1 + y_2 + y_3 + \cdots\}$

跑簡單迴歸 →

$$\hat{y} = b_0 + b_1 x$$ → 估計值

$$y = b_0 + b_1 x + e$$ → 實際值

↓ **推論統計**

This week

母體的迴歸

$$\mathrm{E}(y) = \beta_0 + \beta_1 x$$ → 期望值

$$y = \beta_0 + \beta_1 x + \varepsilon$$ → 實際值

1214實習：數量資料的相關性-Ⅱ

- 使用提供的資料(Student.csv)

1. 斜率的推論與信賴區間
2. PI與CI
   - 指令、解釋、製圖
3. 迴歸診斷與處理
   - Checking conditions
   - Corrective actions

Data: *student.csv*

0. Interpretation of the result table:

(1) The linear regression summary table

```
> # Simple linear regression
> RESULTS = lm(PartyDays ~ StudyHrs)
> summary(RESULTS)

Call:
lm(formula = PartyDays ~ StudyHrs)

Residuals:
    Min      1Q  Median      3Q     Max
-8.4688 -4.3098 -0.3893  3.7329 23.2509

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.46882    0.35326  23.973  < 2e-16 ***
StudyHrs    -0.07197    0.02177  -3.306 0.000995 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.416 on 684 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared:  0.01573,	Adjusted R-squared:  0.01429
F-statistic: 10.93 on 1 and 684 DF,  p-value: 0.0009951
```

$s.e.$ of b_1
(係數的標準誤)

$t-statistic$ of b_1
(係數標準化的t值)

$p-value$ in the hypothesis test of β_1(two.sided)
(對β_1做雙尾假說檢定的$p-value$)

significance
(檢定結果的顯著性)

slope (係數)

Intercept (截距)

x variable
(解釋變數)

Standard deviation for regression
(回歸式的殘差)

Degree of freedom
(n-k-1) 自由度

R^2

修正的R^2

0. Interpretation of the result table:

Data: student.csv

(2) Correlation coefficient

$t - statistic$ of r
(相關係數標準化的t值)

Hypothesis test of correlation coefficient (r)
相關係數的假說檢定

$$H_0: r = 0$$
$$H_a: r \neq 0$$

```
> # Correlation coefficient
> cor.test(PartyDays, StudyHrs)

        Pearson's product-moment correlation

data:  PartyDays and StudyHrs
t = -3.3062, df = 684, p-value = 0.0009951
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.19841112 -0.05104175
sample estimates:
       cor
-0.1254182
```

Correlation coefficient (r)

Degree of freedom
(n-k-1) 自由度

$p - value$ in the hypothesis test of r(two.sided)
(對r做雙尾假說檢定的$p - value$)

0. Interpretation of the result table:

Data: student.csv

(3) ANOVA table for SSE, SSR, SST

```
> # ANOVA
> anova(RESULTS)
Analysis of Variance Table

Response: PartyDays
           Df  Sum Sq Mean Sq F value    Pr(>F)    
StudyHrs    1   320.6  320.62  10.931 0.0009951 ***
Residuals 684 20062.6   29.33                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

$$SSR = \sum (\hat{y}_i - \bar{y})^2$$

回歸式子可解釋的變異

$$SSE = \sum (y_i - \hat{y}_i)^2$$

回歸式子無法解釋的變異

$$SST = \sum (y_i - \bar{y})^2$$

y的總變異

1. Inference about the slope of regression

```
> # Simple linear regression
> RESULTS = lm(PartyDays ~ StudyHrs)
> summary(RESULTS)

Call:
lm(formula = PartyDays ~ StudyHrs)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.4688 -4.3098 -0.3893  3.7329 23.2509 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  8.46882    0.35326  23.973  < 2e-16 ***
StudyHrs    -0.07197    0.02177  -3.306 0.000995 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.416 on 684 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared:  0.01573,	Adjusted R-squared:  0.01429 
F-statistic: 10.93 on 1 and 684 DF,  p-value: 0.0009951
```

- $Significance\ level:\ \alpha = 0.05$

Step 1.
- $\boldsymbol{H_0: \beta_1 = 0}$
- $\boldsymbol{H_1: \beta_1 \neq 0}$

Step 2.
- b_1 = -0.07197
- $t = \dfrac{b_1 - null\ value}{s.e.(b_1)} = \dfrac{-0.07197 - 0}{0.02177} = -3.306$

Step 3.
- $p\ value = 0.000995$

Step 4.
- $p\ value < \alpha.\ \ Reject\ H_0.$

Step 5.
- $Based\ on\ the\ hypotheses\ test\ result, \boldsymbol{\beta_1}\ is\ not\ equal\ to\ 0.$

By manual calculation:

1. Inference about the slope of regression

$$t = \frac{\text{Sample statistic} - \text{Null value}}{\text{Standard error}} = \frac{b_1 - 0}{s.e.(b_1)}$$

$$b_1 = \frac{n\left(\sum_{i=1}^{n} x_i y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\sum_{i=1}^{n} \boxed{(x_i - \bar{x})}\boxed{(y_i - \bar{y})}}{\boxed{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

or

$$b_1 = r\frac{\boxed{S_y}}{\boxed{S_x}}$$

```
sd(x)
sd(y)
```

$$s.e.(b_1) = \frac{s}{\sqrt{\boxed{\sum (x - \bar{x})^2}}} \quad \text{where } s = \sqrt{\frac{\boxed{SSE}}{n-2}}$$

```
x.mean = mean(student$x)
student$xx = (student$x - x.mean)^2
sqrt(sum(student$xx))
```

By manual calculation coding examples:

```
# (data cleaning before calculation)
student = student[,c("PartyDays","StudyHrs")]
student = na.omit(student)
```

```
> b1 = sum(student$xxyy) / sum(student$xx2); b1
[1] -0.07196698
```

```
> SSE = sum((student$PartyDays - yhat)^2); SSE
[1] 20062.55
```

```
> s = sqrt(SSE/(nrow(student)-2)); s
[1] 5.41583
```

```
#
x.mean = mean(student$StudyHrs)
y.mean = mean(student$PartyDays)

student$xx = student$StudyHrs - x.mean
student$yy = student$PartyDays - y.mean
student$xxyy = student$xx*student$yy
student$xx2 = student$xx^2

# b1
b1 = sum(student$xxyy) / sum(student$xx2); b1


# SSE
RESULTSS = lm(PartyDays ~ StudyHrs, data = student)
summary(RESULTSS)

yhat = RESULTSS$fitted.values
SSE = sum((student$PartyDays - yhat)^2); SSE


# standard error of residual
s = sqrt(SSE/(nrow(student)-2)); s
```

1. Confidence interval of the slope:

$$b_1 \pm t^* \times s.e.(b_1) = b_1 \pm t^* \times \frac{s}{\sqrt{\sum(x - \bar{x})^2}}$$

The multiplier t^*:
$t\ distribution$ with $n - 2$ degrees of freedom

```
qt(1-alpha/2, df = n-2)
```

```
> # Simple linear regression
> RESULTS = lm(PartyDays ~ StudyHrs)
> summary(RESULTS)

Call:
lm(formula = PartyDays ~ StudyHrs)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.4688 -4.3098 -0.3893  3.7329 23.2509 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  8.46882    0.35326  23.973  < 2e-16 ***
StudyHrs    -0.07197    0.02177  -3.306 0.000995 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.416 on 684 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared:  0.01573,	Adjusted R-squared:  0.01429 
F-statistic: 10.93 on 1 and 684 DF,  p-value: 0.0009951
```

2. Prediction interval and confidence interval:

```r
# 2. Prediction interval and Confidence interval ----------------------------

# PI
predict(RESULTS, data.frame(StudyHrs = c(10,20,30)), interval = "prediction", level = 0.95)

# CI
predict(RESULTS, data.frame(StudyHrs = c(10,20,30)), interval = "confidence", level = 0.95)
```

```
> # PI
> predict(RESULTS, data.frame(StudyHrs = c(10,20,30)), interval = "prediction", level = 0.95)
       fit       lwr      upr
1 7.749149 -2.893102 18.39140
2 7.029480 -3.615933 17.67489
3 6.309810 -4.355902 16.97552
> # CI
> predict(RESULTS, data.frame(StudyHrs = c(10,20,30)), interval = "confidence", level = 0.95)
       fit      lwr      upr
1 7.749149 7.321306 8.176993
2 7.029480 6.529145 7.529814
3 6.309810 5.483410 7.136209
```

Prediction interval (預測區間，預測y的可能範圍)

主角：one observation (一個觀察值 e.g. 一個人)

Confidence interval (信賴區間，預測$E(y)$的可能範圍)

主角：parameter (母體參數，這裡是期望值)

$\hat{y}_i$ (估計值)

By manual calculation:

2. Prediction interval and confidence interval

- Prediction interval (預測區間，預測y的可能範圍)

$$\hat{y} \pm t^{*}\sqrt{s^{2}+\left[s.e.(fit)\right]^{2}}$$

where $s.e.(fit) = s\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^{2}}{\sum(x_i-\bar{x})^{2}}}$

t^* found from Table A.2 with df = n – 2.

- Confidence interval (信賴區間，預測$E(y)$的可能範圍)

$$\hat{y} \pm t^{*} \times s.e.(fit)$$

where $s.e.(fit) = s\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^{2}}{\sum(x_i-\bar{x})^{2}}}$

t^* found from Table A.2 with df = n – 2.

2. Plotting

Reviewed:

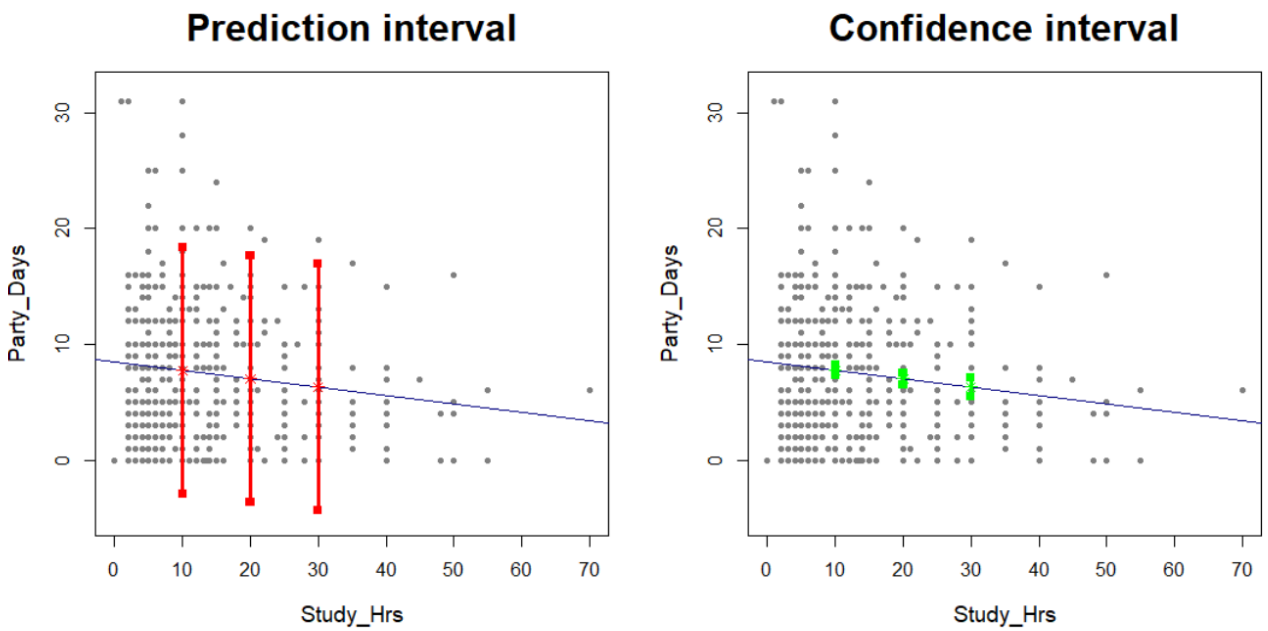
- Draw multiple lines by using a for-loop

```
for (i in 1:length()) {

  plot()

  points(x = x1, y = y1)
  lines(x = c(x1,x2, ...), y = c(y1,y2, ...))

}
```

New:

- Draw multiple graphs in the same plot

```
# split the plot
par(mfrow=c(1,2))

# 1st graph
plot()

# 2nd graph
plot()
```



```r
# Plotting
fit = PI[,1]
PI.low = PI[,2]
PI.high = PI[,3]

CI.low = CI[,2]
CI.high = CI[,3]
```

```r
xx.test = c(10,20,30)
```

continued...

```r
par(mfrow=c(1,2))
# plot PI
plot(PartyDays ~ StudyHrs,
     data = student, pch = 20, col="gray50",
     main="Prediction interval", xlab="Study_Hrs", ylab="Party_Days",
     ylim = c(-5,32),
     cex.main = 2, cex.lab = 1.2)

abline(RESULTS, col="navy") #regression line

for (i in 1:length(xx.test)) {
  lines(c(xx.test[i],xx.test[i]), c(PI.low[i],PI.high[i]), col = "red", lwd = 3)
  points(xx.test[i], PI.low[i], col = "red", pch = 15) #lower point
  points(xx.test[i], PI.high[i], col = "red", pch = 15) #upper point
  points(xx.test[i], fit[i], col = "red", pch = 8) # estimated value
}

# plot CI
plot(PartyDays ~ StudyHrs,
     data = student, pch = 20, col="gray50",
     main="Confidence interval", xlab="Study_Hrs", ylab="Party_Days",
     ylim = c(-5,32),
     cex.main = 2, cex.lab = 1.2)

abline(RESULTS, col="navy") #regression line

for (i in 1:length(xx.test)) {
  lines(c(xx.test[i],xx.test[i]), c(CI.low[i],CI.high[i]), col = "green", lwd = 3)
  points(xx.test[i], CI.low[i], col = "green", pch = 15) #lower point
  points(xx.test[i], CI.high[i], col = "green", pch = 15) #upper point
  points(xx.test[i], fit[i], col = "green", pch = 8) # estimated value
}
```

3. Checking conditions

14.5 Checking Conditions for Regression Inference

Conditions:

1. **Form** of the equation that links the mean value of *y* to *x* **must be correct**.
2. **No extreme outliers** that influence the results unduly.
3. **Standard deviation** of values of *y* from the mean *y* is **same** regardless of value of *x*.
4. For individuals in the population with same value of *x*, the distribution of *y* is a **normal distribution**. Equivalently, the distribution of deviations from the mean value of *y* is a normal distribution. This can be relaxed if the *n* is large.
5. **Observations** in the sample are **independent** of each other.

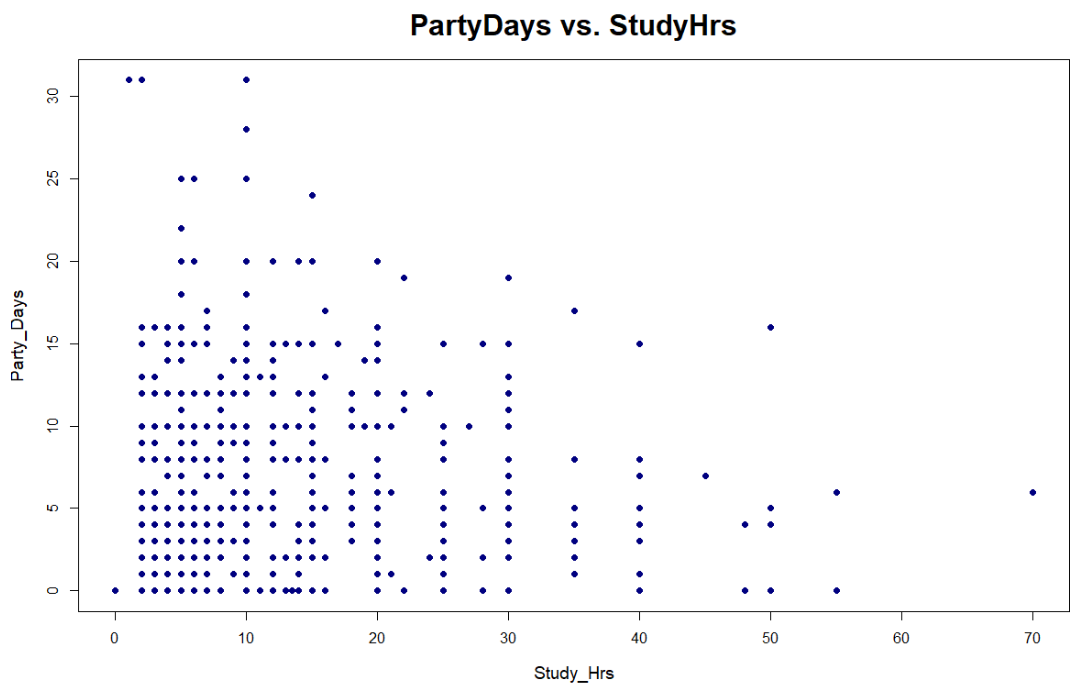
→ **Scatterplot of x vs. y**

→ **Scatterplot of $residuals$ vs. x**

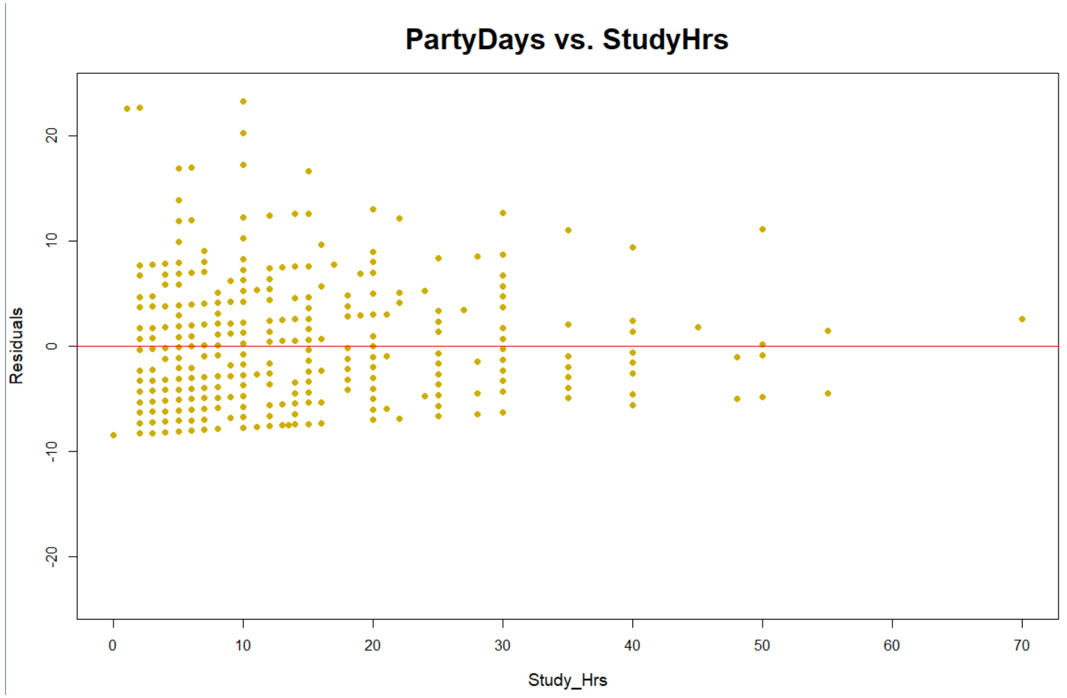
→ **Histogram of $residuals$**

Scatterplots

Scatterplot of x vs. y



Scatterplot of $residuals$ vs. x



Pre-work: Check is there any NA value and removed if needed:

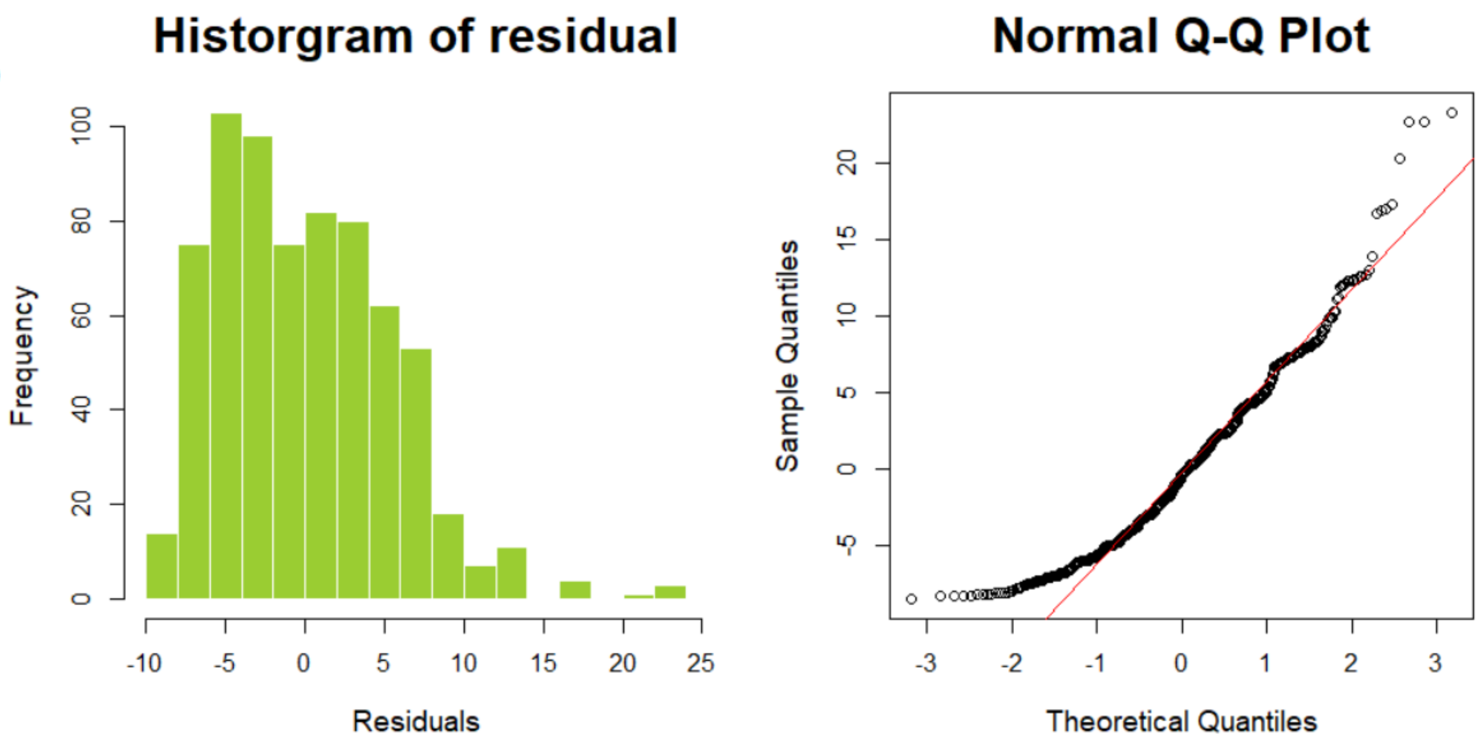
```
# Check na value and data cleaning
PartyDays = student$PartyDays
StudyHrs = student$StudyHrs
length(PartyDays[is.na(PartyDays)])  #0
length(StudyHrs[is.na(StudyHrs)]) #4

student = student[,c("PartyDays","StudyHrs")]
student = na.omit(student)

PartyDays = student$PartyDays
StudyHrs = student$StudyHrs
```

Draw the scatterplot:

```
dev.off()
# Scatterplot of x vs. y
plot(PartyDays ~ StudyHrs,
     pch = 16, cex = 1, col = "navy",
     main="PartyDays vs. StudyHrs",
     xlab="Study_Hrs", ylab="Party_Days",
     cex.main = 2, cex.lab = 1.2)


# Scatterplot of residuals vs. x
plot(res ~ StudyHrs,
     pch = 16, cex = 1, col = "gold3",
     main="PartyDays vs. StudyHrs",
     xlab="Study_Hrs", ylab="Residuals",ylim = c(-24,24),
     cex.main = 2, cex.lab = 1.2)

abline(h = 0, col = "red")
```

Histogram of residuals and QQ plot

```r
# Histogram of residuals
par(mfrow=c(1,2))

hist(res, breaks=20, border = "white", col = "olivedrab3",
     main = "Historgram of residual", xlab = "Residuals",
     cex.main = 2, cex.lab = 1.2)

#QQ plot
qqnorm(res, cex.main = 2, cex.lab = 1.2)
qqline(res,col="red" )
```



作業11 數量資料的相關性-II

- 練習題5題(Ch. 14)
  - 14.16; 14.28; 14.36; 14.54; 14.56
- R程式練習題(繳交程式碼與執行結果)
  - 使用vehicles.csv資料檔案

1. – 估計簡單迴歸式：vehicle=b_0+b_1*GDP，檢定 H_a: $\beta_1 \neq 0$，估計β_1之信賴區間
2. – 估計GDP=c(9000,13000,17000)之PI及CI，簡要解釋，製圖
3. – 進行迴歸診斷與處理

1(1) 以五步驟進行假說檢定。
1(2) 計算斜率的信賴區間。

2(1) 計算PI及CI。
2(2) 分別解釋其代表的意義。
2(3) 繪製PI及CI。

3(1) 逐條診斷與解釋。
3(2) 處理。