R 語言期中報告

經濟一 B　410510008　陳宣儒

● 匯入外部資料

氣象局的 open data -　“海平面統計-臺灣各地潮位觀測月平均海平面”

資料主題 (opendata.cwb.gov.tw)　- 進行海平面數據分析

因為這資料集是 Json 檔，所以需要先安裝 rjson 套件

R - JSON Files (tutorialspoint.com)

```
#安裝 rjson 套件
install.packages("rjson")
library(rjson)

#讀取 Json 檔案
searowdata = fromJSON(file = "C:/Users/88692/Desktop/課程/R 語言/C-B0048-001.json")
seadata = searowdata$Cwbopendata$dataset$location
```

此時資料已讀取 並擷取出數據的部分存入 seadata 中 為一個 List 型態的資料

seadata	list [3264]	List of length 3264
[[1]]	list [5]	List of length 5
SiteName	character [1]	'基隆市 基隆'
SiteId	character [1]	'1516'
ItemName	character [4]	'西元年月' '平均潮位' '最高高潮位' '最低低潮位'
ItemValue	character [4]	'200001' '-11.1' '39.5' '-102.6'
ItemUnit	character [3]	'cm' 'cm' 'cm'
[[2]]	list [5]	List of length 5
SiteName	character [1]	'基隆市 基隆'
SiteId	character [1]	'1516'
ItemName	character [4]	'西元年月' '平均潮位' '最高高潮位' '最低低潮位'
ItemValue	character [4]	'200002' '-8.3' '65.5' '-82.6'
ItemUnit	character [3]	'cm' 'cm' 'cm'
[[3]]	list [5]	List of length 5
[[4]]	list [5]	List of length 5
[[5]]	list [5]	List of length 5
[[6]]	list [5]	List of length 5
[[7]]	list [5]	List of length 5

將資料集轉成 DataFrame 的型態

```
#建立一個新的 DF
sea_dataframe=t(data.frame(seadata[[1]][[4]]))
colnames(sea_dataframe)=c(seadata[[1]][[3]])
row.names(sea_dataframe) = NULL

#選取高雄站(SiteId = 1486)的資料
i=1
for (i in c(1:3264)) {
   if (seadata[[i]]$SiteId == '1486') {
      newdata = seadata[[i]][[4]]
      sea_dataframe = rbind(sea_dataframe,newdata)
   }
   i=i+1
}

#將原本用來建立 DF 的資料刪掉
row.names(sea_dataframe) = NULL
sea_dataframe = sea_dataframe[-1,]
rm(newdata)

#取出年平均資料 (00 月的資料)
yeardata = 13* c(1:20)
sea_df_year = sea_dataframe[yeardata,]
sea_dataframe = sea_dataframe[-yeardata,]
```

	西元年月	平均潮位	最高高潮位	最低低潮位
60	200408	32.1	99.9	-23.6
61	200409	27.1	88.7	-28.4
62	200410	26.9	84.8	-23.9
63	200411	12.4	78.5	-35.2
64	200412	2.7	77.5	-62.3
65	200500	8.8	88.0	-65.7
66	200501	-4.4	73.4	-64.1
67	200502	-1.3	65.6	-62.0
68	200503	-4.4	54.4	-65.7
69	200504	4.8	57.6	-46.9
70	200505	9.2	76.4	-44.7
71	200506	15.6	84.2	-47.0
72	200507	17.9	88.0	-43.2
73	200508	17.8	83.6	-41.3
74	200509	21.4	77.2	-30.7
75	200510	15.4	67.2	-28.7
76	200511	15.6	74.4	-42.3

發現資料包含了每年的年平均資料(00 月)
但資料筆數不多 不方便做計算
所以將此資料取出 使用每月的平均做計算

➢ **資料分類整理**

1. 歷年下來的潮位變化 : sea_df

```
sea_df = data.frame(sea_dataframe[,1:2])
names(sea_df)=c('period','mean_tide_level')

#要將時間的數據轉為連續的數值
sea_df[,1] = as.numeric(sea_df[,1])
del = c(0)    #將有 NULL 值的列號存入一個 vector
j=1
p = length(sea_df[,1])
for (j in c(1:p) ) {
   sea_df[j,1] = 2000+(j-1)/12
   if(is.null(sea_df[[j,2]][1]) == TRUE) del = append(del, j);
   j=j+1
}

#刪除有 NULL 的資料列
del = del[-1]
sea_df = sea_df[-del,]

#將潮位資料轉為 numeric
sea_df[,2] = unlist(sea_df[,2])
sea_df[,2] = as.numeric(sea_df[,2]
```

sea_df

Filter

	period	mean_tide_level
1	2000.000	15.2
2	2000.083	18.4
3	2000.167	24.6
4	2000.250	21.8
5	2000.333	28.4
6	2000.417	30.1
7	2000.500	38.9
8	2000.583	38.9
9	2000.667	39.4
10	2000.750	34.2
11	2000.833	23.3
12	2000.917	18.3
13	2001.000	21.5
14	2001.083	23.9
15	2001.167	20.6
16	2001.250	28.3
17	2001.333	29.3
18	2001.417	35.0
19	2001.500	40.8

2. 一年內 1~12 月的潮位變化 : sea_df2

```
sea_df2 = data.frame(sea_dataframe[,c(1,1,2)])
names(sea_df2)=c('year','month','tide_level')

#將時間的數據轉為連續的數值
sea_df2[,1] = as.numeric(sea_df2[,1])
sea_df2[,2] = as.numeric(sea_df2[,2])

k=1
for (k in c(1:p) ) {
   sea_df2[k,1] = sea_df2[k,1]%/%100
   k=k+1
}

k=1
for (k in c(1:p) ) {
   sea_df2[k,2] = k%%12
   if(k%%12==0)    sea_df2[k,2] = 12;
   k=k+1
}

#刪除有 NULL 的資料列
sea_df2 = sea_df2[-del,]

#將潮位資料轉為 numeric
sea_df2[,3] = unlist(sea_df2[,3])
sea_df2[,3] = as.numeric(sea_df2[,3])
```

sea_df2

Filter

	year	month	tide_level
1	2000	1	15.2
2	2000	2	18.4
3	2000	3	24.6
4	2000	4	21.8
5	2000	5	28.4
6	2000	6	30.1
7	2000	7	38.9
8	2000	8	38.9
9	2000	9	39.4
10	2000	10	34.2
11	2000	11	23.3
12	2000	12	18.3
13	2001	1	21.5
14	2001	2	23.9
15	2001	3	20.6
16	2001	4	28.3
17	2001	5	29.3
18	2001	6	35.0
19	2001	7	40.8

- **基本敘述統計**

#先建立與 sea_df 資料集的連結

attach(sea_df)

➢ **個別資料統計**

#平均值

mean(mean_tide_level)

#中位數

median(mean_tide_level)

#標準差

sd(mean_tide_level)

#變異數

var(mean_tide_level)

#全距

range(mean_tide_level)

#四分位距

quantile(mean_tide_level)

```
> mean(mean_tide_level)
[1] 21.6755
> median(mean_tide_level)
[1] 21.7
> sd(mean_tide_level)
[1] 10.57423
> var(mean_tide_level)
[1] 111.8144
> range(mean_tide_level)
[1] -4.4 46.1
> quantile(mean_tide_level)
  0%  25%  50%  75% 100%
-4.4 14.4 21.7 29.5 46.1
```

➢ **整體統計特徵**

summary(mean_tide_level)

library('Hmisc')

describe(mean_tide_level)

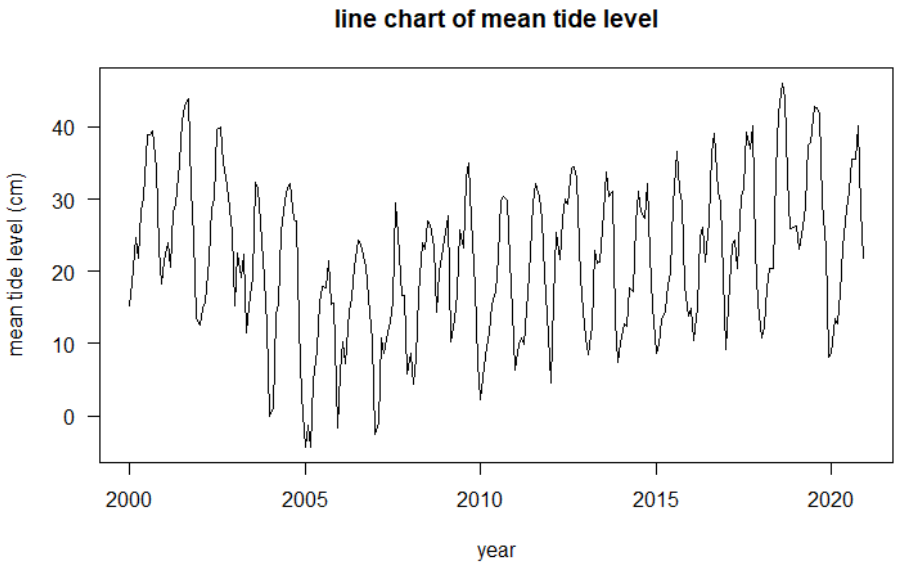
```
> summary(mean_tide_level)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  -4.40   14.40   21.70   21.68   29.50   46.10
> describe(mean_tide_level)
mean_tide_level
       n  missing distinct     Info     Mean      Gmd      .05      .10
     249        0      184        1    21.68    12.07     5.00     8.60
     .25      .50      .75      .90      .95
   14.40    21.70    29.50    35.10    39.58

lowest : -4.4 -2.6 -1.7 -1.3 -1.0, highest: 42.6 42.9 43.9 44.0 46.1
```

➢ 資料視覺化

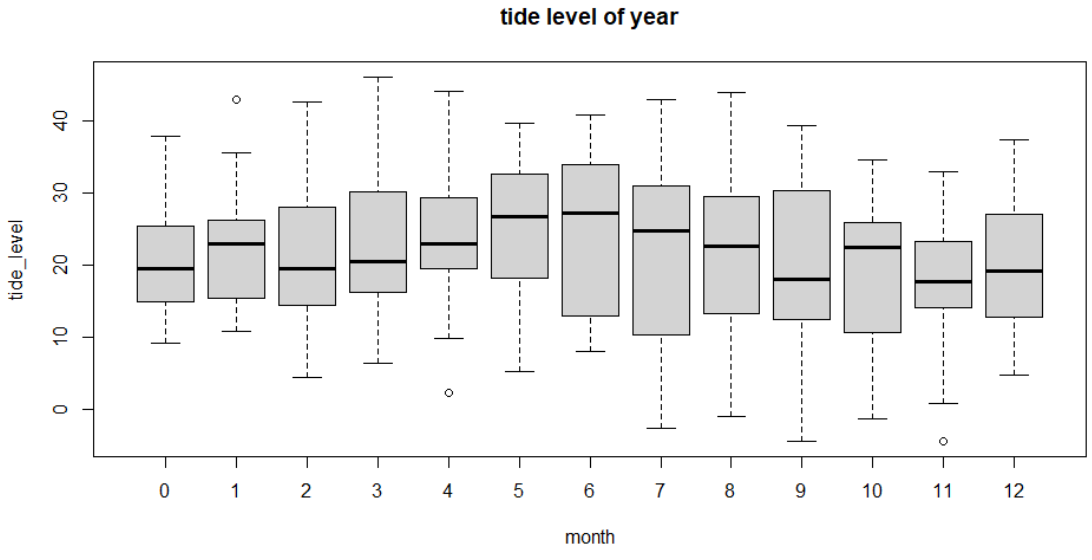
#折線圖 plot type="l" *- 歷年的海平面*

```
plot(sea_df, type="l", ann = F, xaxt = "n", yaxt = "n")
axis(1,seq(1995,2025,5),las = 1)
axis(2, las = 2)
title(xlab="year",ylab="mean tide level (cm)",
      main="line chart of mean tide level")
```



#箱線圖 boxplot *(四分位距)*

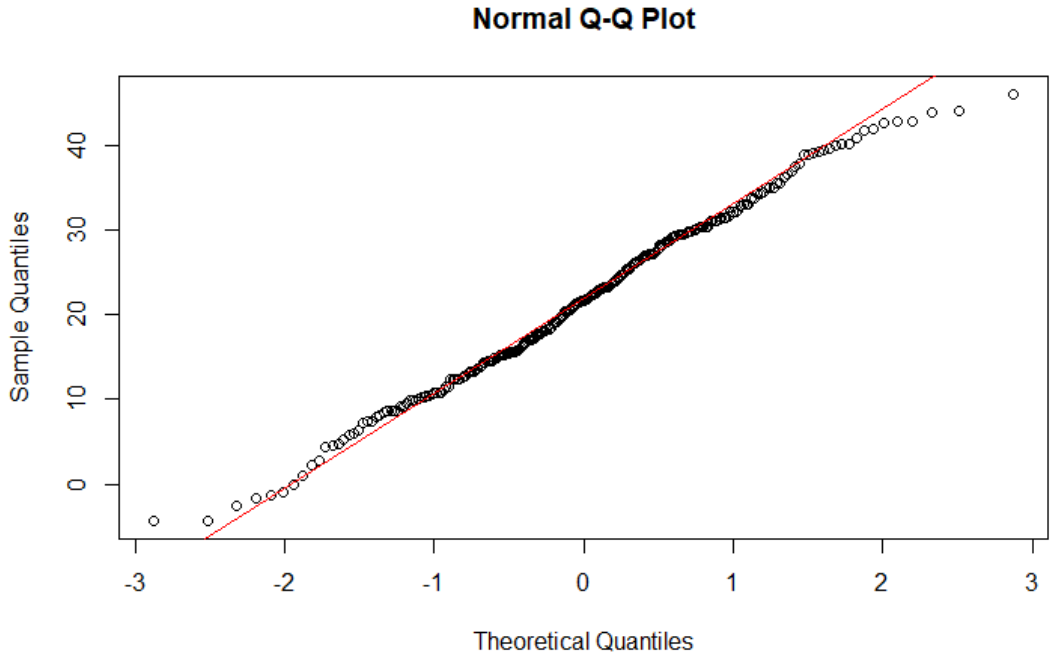
```
boxplot(tide_level~month, data = sea_df2,main="tide level of year")
```



- **常態檢定**

#常態機率圖

qqnorm(sea_df[,2]);qqline(sea_df[,2], col='Red')



***#Shapiro-Wilk** 常態性檢定*

shapiro.test(sea_df[,2])

```
        Shapiro-Wilk normality test

data:  sea_df[, 2]
W = 0.99219, p-value = 0.2116
```

p-value>0.05 海平面數據為常態分佈

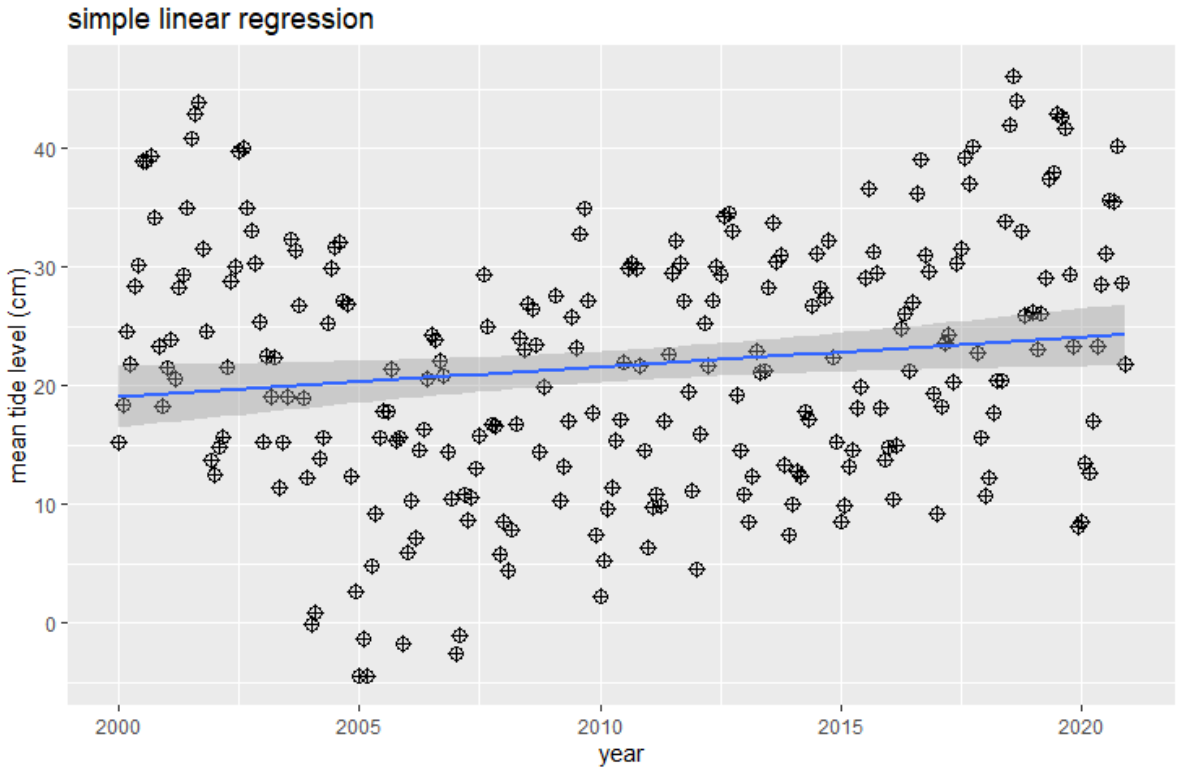
- **線性迴歸**

#建立模型

```
seaLM = lm(mean_tide_level~period, data = sea_df)
```

#畫出預測圖

```
ggplot(sea_df, aes(x = period, y = mean_tide_level))+
  geom_point(shape = 10, size = 3)+geom_smooth(method = lm)+
  labs(title = "simple linear regression",x='year',y='mean tide level (cm)')
```



臺灣海象災防環境資訊平台 (ocean.cwb.gov.tw/V2/sea_level_statistics)

#取得模型統計量

summary(seaLM)

```
Call:
lm(formula = mean_tide_level ~ period, data = sea_df)

Residuals:
     Min       1Q   Median       3Q      Max
-24.7633  -7.7931  -0.0746   7.8314  24.4082

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -478.9231   219.9787  -2.177   0.0304 *
period         0.2490     0.1094   2.276   0.0237 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.49 on 247 degrees of freedom
Multiple R-squared:  0.02054,   Adjusted R-squared:  0.01657
F-statistic: 5.179 on 1 and 247 DF,  p-value: 0.02372
```

迴歸公式 : mean tide level = -478.9231 + 0.2490 × period + e

Adjusted R-squared = 0.01657 此迴歸模型的解釋力極低

➢ **預測**

#預測 2050 年高雄的平均海平面

new = data.frame(period=2050)

result = predict(seaLM, newdata = new)

result

```
> result = predict(seaLM, newdata = new)
> result
       1
31.52683
```

在此模型的預測下 2050 年高雄測站的平均海平面會上升到 31.52683cm

#畫出預測圖

```
ggplot(sea_df, aes(x = period, y = mean_tide_level))+
  geom_point(shape = 10, size = 3)+
  geom_smooth(method = lm)+
  scale_x_continuous(breaks = c(seq(1995, new$period+5, 5)) )+
  scale_y_continuous(breaks = c(seq(-20,50,5)))+
  geom_point(x=new$period, y=result, size=5,shape=17,color="red")
```

超過圖表的預測點顯示不出來

- **複線性迴歸**

#建立模型

seaLM2 = lm(tide_level~month+year, data = sea_df2)

#取得模型統計量

summary(seaLM2)

```
Call:
lm(formula = tide_level ~ month + year, data = sea_df2)

Residuals:
     Min       1Q   Median       3Q      Max
-28.8373  -6.9602  -0.0569   6.7168  21.4152

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -434.1984   203.7118  -2.131   0.0340 *
month          1.1902     0.1798   6.618 2.26e-10 ***
year           0.2230     0.1013   2.200   0.0287 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.705 on 246 degrees of freedom
Multiple R-squared:  0.1645,    Adjusted R-squared:  0.1577
F-statistic: 24.21 on 2 and 246 DF,  p-value: 2.523e-10
```

迴歸公式 : tide level = -434.1984 + 1.1902 × month + 0.2230 × year + e
Adjusted R-squared = 0.1577 此迴歸模型的解釋力極低

- **預測**

#預測 2030 年 3 月高雄的平均海平面

new = data.frame(year=2030, month=3)

result = predict(seaLM2, newdata = new)

result

```
> result = predict(seaLM2, newdata = new)
> result
       1
22.00007
```

在此模型的預測下 2030 年 3 月高雄測站的平均海平面為 22.00007cm