商業分析：SAS / R HW2

108208004 經濟三 白植允

1.[20pts] a. 生成一筆資料：

Xi=a+ ε Xi=a+ ε ,i=1,...,20i=1,...,20

- aa為 0~10 任意數字。 ε ε ~N(0,2)~N(0,2)
- 注意：XX 必須在 0~11 內。

```
#1 a.
set.seed(1)
a <- sample(1:10,20,replace = T)
x <- c()
for (i in 1:20) {
  x[i] = a[i] + rnorm(1,0,2)
}
```

```
> x
 [1]  4.5706002  6.2498618  6.9101328  0.9676195  3.8876724  8.6424424  3.1878026  4.8379547
 [9]  2.5642726  5.1491300  1.0212966 11.2396515  5.8877425  9.6884090  4.0584952  8.0436999
[17]  5.8358831  7.7173591  8.7944245  9.7753432
```

[20pts] b. Cauchy(θ , 1) 的密度函數，取 log 後一次微分如下，請寫出此 function

f(θ)=−2∑i=1n θ −xi{1+(θ −xi)2}f(θ)=−2∑i=1n θ −xi{1+(θ −xi)2}

```
#1 b.
f <- function(theta,x){
  result <- 0
  for( i in x){
    result <- (theta-i)/(1+(theta-i)^2) + result
    }
  return(-2*result)
}
```

[10pts] c.代入 a 生成的資料至 b 的 function，並令 θ =0.3 θ =0.3。

```
#1 c.
f(theta = 0.3,x)
```

```
> f(theta = 0.3,x)
[1] 8.411675
```

2.

[10pts] a. 根據 Build_year，建立一個新類別變數 year_type，1899 年以前的房子為”centennial”，1900~1959 年為”old”，1960 年以上為”new”。

```
#2 a.
houseprice <- read.csv("houseprice.csv",sep = ',')
library(tidyverse)
houseprice <- houseprice %>%
  mutate( year_type = ifelse(Build_year<1900,"centennial",
                             ifelse(Build_year<1960,"old","new")))
```

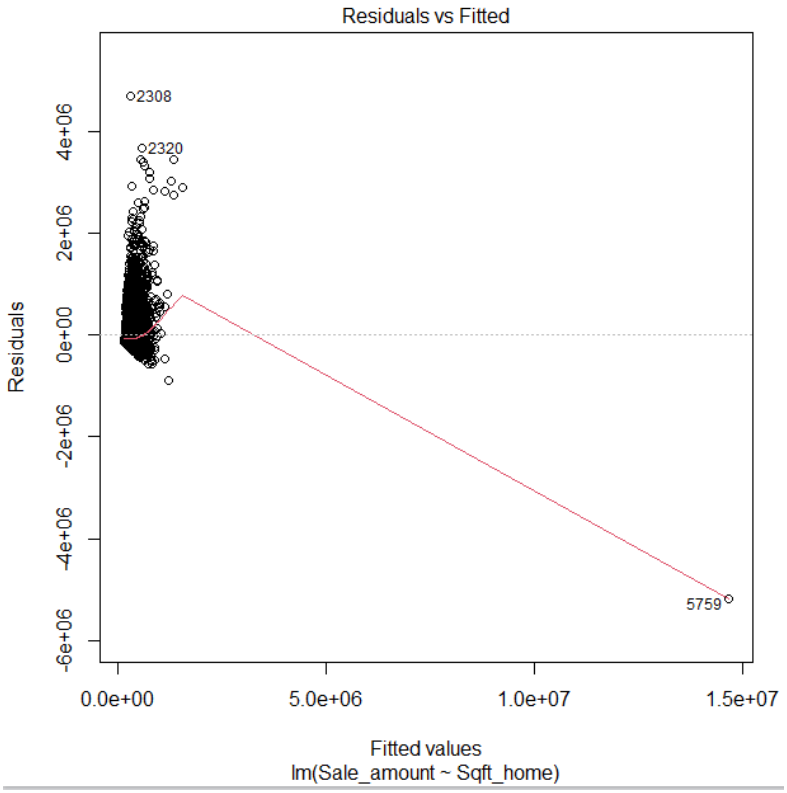
Record	Sale_amount	Sale_date	Beds	Baths	Sqft_home	Sqft_lot	Type	Build_year	Town	University	year_type
1	295000	2016/5/31	5	3.00	2020	38332.8	Single Family	1976	Ames, IA	Iowa State University	new
2	240000	2016/6/20	4	2.00	1498	54014.4	Single Family	2002	Ames, IA	Iowa State University	new
3	385000	2016/5/31	5	4.00	4000	85813.2	Single Family	2001	Ames, IA	Iowa State University	new
4	268000	2016/4/12	3	2.50	2283	118918.8	Single Family	1972	Ames, IA	Iowa State University	new
5	186000	2016/4/5	3	1.25	1527	15681.6	Single Family	1975	Ames, IA	Iowa State University	new
6	302500	2016/3/2	4	3.00	3117	33105.6	Single Family	1976	Ames, IA	Iowa State University	new
7	223000	2016/6/2	3	2.00	1218	25264.8	Single Family	1975	Ames, IA	Iowa State University	new

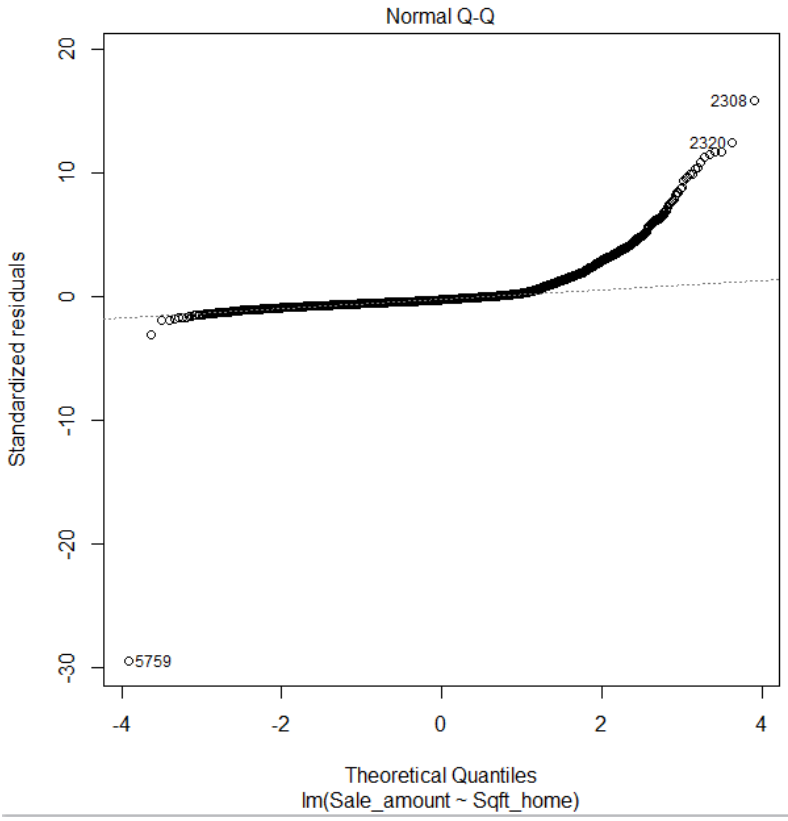
[40pts] b. 決定好你的最佳配適模型後，總結你的發現並根據解釋變數預測房屋價格。

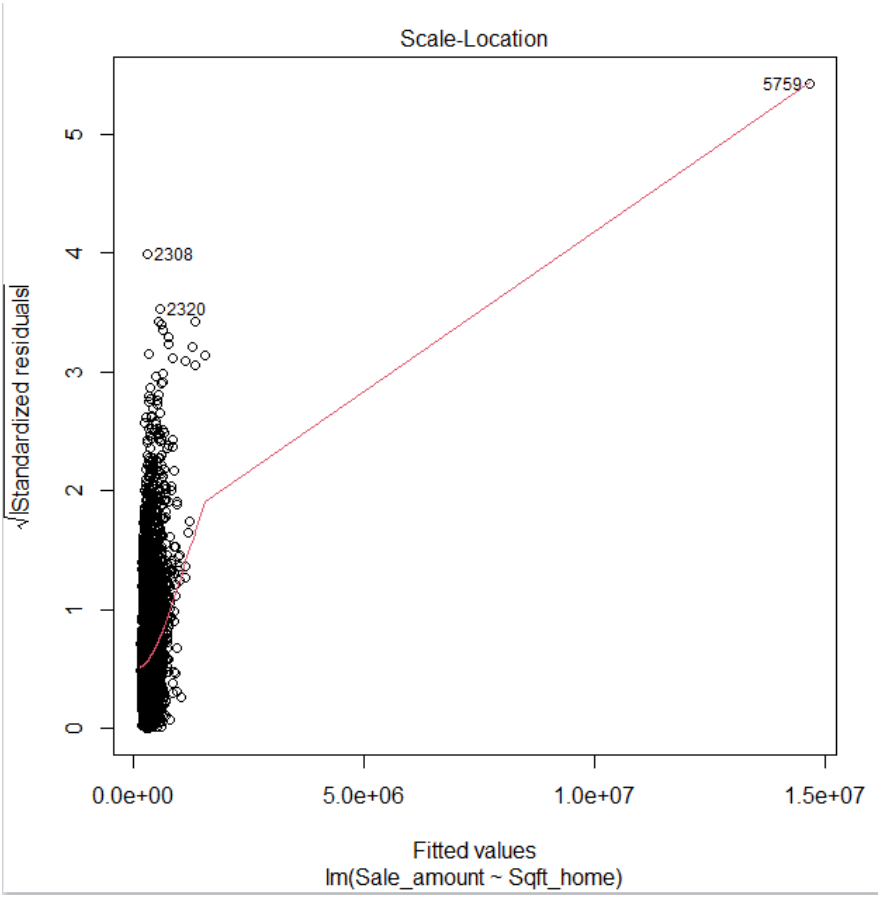
```
library(broom)
summary(houseprice)
houseprice$Type <- as.factor(houseprice$Type)
houseprice$year_type <- as.factor(houseprice$year_type)
houseprice$Town<- as.factor(houseprice$Town)
houseprice$University <- as.factor(houseprice$University)
```

=>調整數據型態

```
plot(lm(Sale_amount~Sqft_home,data = houseprice))
```


Residuals vs Fitted
Residuals
Fitted values
lm(Sale_amount ~ Sqft_home)
2308
2320
5759
4e+06
2e+06
0e+00
-2e+06
-4e+06
-6e+06
0.0e+00
5.0e+06
1.0e+07
1.5e+07


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(Sale_amount ~ Sqft_home)
2308
2320
5759
20
10
0
-10
-20
-30
-4
-2
0
2
4

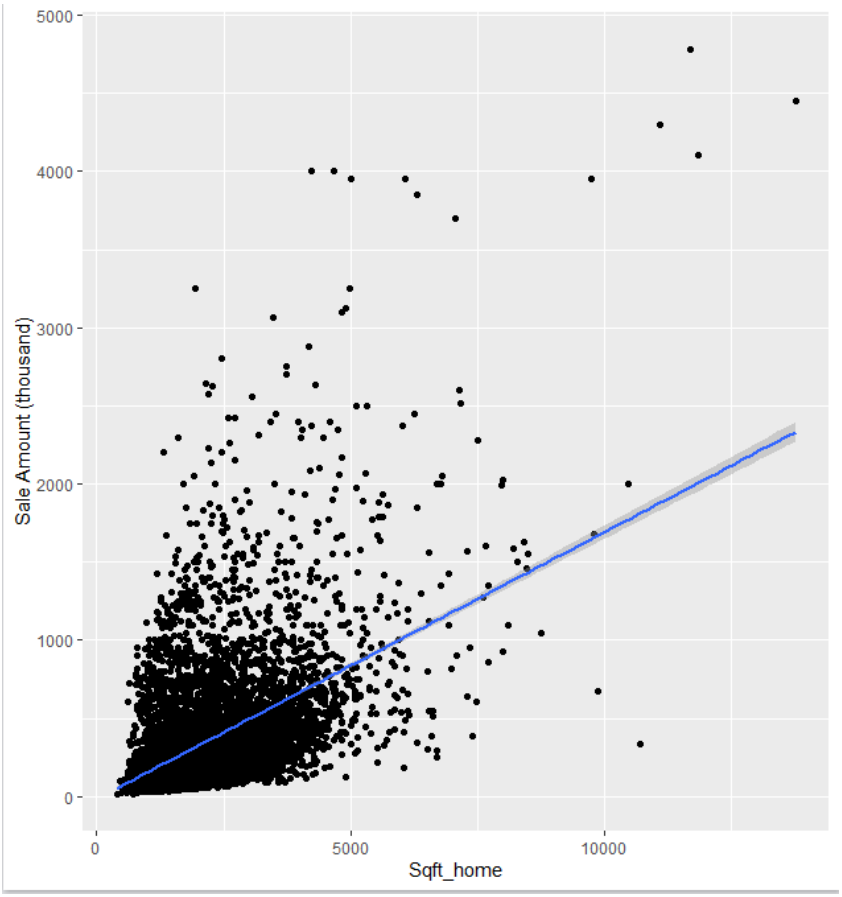


=>以 Sqft_home 為參數做簡單線性迴歸，發現 record :(2308,2320,5759)應該是 outlier，且資料有 heavy tailed 的分布

```
edited <-houseprice[-c(2308,2320,5759),]
```

=>以 edited 代替原資料作分析

```
ggplot(edited,aes(x=Sqft_home ,y=Sale_amount/1000))+geom_point()+
  geom_smooth(method = "lm")+
  labs(y='Sale Amount (thousand)')
```



=>大致上有趨勢，看起來房子平方英尺數越大房屋價格也越大。

=>隨著 Sqft_home 越來越大，變異數有越來越大的趨勢，模型有效性可能沒有到很好

```
fit <- lm(Sale_amount~Sqft_home,data = edited)
glance(fit)
anova(fit)
```

```
> glance(fit)
# A tibble: 1 x 12
  r.squared adj.r.squared   sigma statistic p.value    df  logLik    AIC    BIC deviance df.residual  nobs
      <dbl>         <dbl>   <dbl>     <dbl>   <dbl> <dbl>   <dbl>  <dbl>  <dbl>    <dbl>       <int> <int>
1     0.274         0.274 276275.     4026.       0     1 -1.49e5 2.97e5 2.97e5  8.13e14       10654 10656
> anova(fit)
Analysis of Variance Table

Response: Sale_amount
            Df     Sum Sq    Mean Sq F value    Pr(>F)
Sqft_home    1 3.0729e+14 3.0729e+14  4025.9 < 2.2e-16 ***
Residuals 10654 8.1320e+14 7.6328e+10
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

=>此模型能解釋約 27.4%的房屋價格，而 p-value 小於 0.001，有蠻高(>99.9%)的信心水準說明價格和房子多少平方英尺有線性關係

```
fit2 <- lm(Sale_amount~., data=edited)
step(fit2)
fit3 <- lm(Sale_amount ~ Beds + Baths + Sqft_home + Sqft_lot +
             Type + Build_year + Town + year_type, data = edited)
```

=>直接用電腦找出最適合的模型，其中有 Beds , Baths , Sqft_home , Sqft_lot ,

Type , Build_year , Town , year_type 做參數。

```
> anova(fit,fit3)
Analysis of Variance Table

Model 1: Sale_amount ~ Sqft_home
Model 2: Sale_amount ~ Beds + Baths + Sqft_home + Sqft_lot + Type + Build_year +
    Town + year_type
  Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
1  10654 8.1320e+14
2  10597 3.0171e+14 57 5.1149e+14 315.18 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

=>根據 anova table，他的模型比較好(以 beds,baths,sqft_home,sqft_lot,type,build_year,town,year_type 做參數估計 sale_amount)，果然電腦比較厲害。

```
> glance(fit3)
# A tibble: 1 x 12
  r.squared adj.r.squared
      <dbl>         <dbl>
1     0.731         0.729
```

=>他的模型可以解釋約 73%的房屋價格，他的模型真的比較好，看來這個模型比較適合做我的最佳配飾模型。

```
> tidy(fit3)
# A tibble: 59 x 5
   term                      estimate   std.error statistic  p.value
   <chr>                        <dbl>       <dbl>     <dbl>    <dbl>
 1 (Intercept)            -592732.     199661.        -2.97 3.00e- 3
 2 Beds                     -6178.       2202.        -2.81 5.03e- 3
 3 Baths                    46617.       2889.        16.1  6.75e-58
 4 Sqft_home                  130.          2.81       46.3  0
 5 Sqft_lot                     0.161       0.0233      6.90 5.40e-12
 6 TypeMultiple Occupancy  -49764.      22220.        -2.24 2.51e- 2
 7 TypeSingle Family       118770.      12226.         9.71 3.24e-22
 8 Build_year                 237.        105.        2.25 2.46e- 2
 9 TownAmherst, MA          98803.      24437.        4.04 5.31e- 5
10 TownAnn Arbor, MI       110739.      15295.        7.24 4.79e-13
# ... with 49 more rows
```

=>附上最適模型的部分參數估計表(Beds 跟 sale_amount 的關係好像不顯著，其中一個 type 也是)。

- 程式碼

```
#1 a.
set.seed(1)
a <- sample(1:10,20,replace = T)
x <- c()
for (i in 1:20) {
   x[i] = a[i] + rnorm(1,0,2)
```

```
}

#1 b.
f <- function(theta,x){
   result <- 0
   for( i in x){
      result <- (theta-i)/(1+(theta-i)^2) + result
      }
   return(-2*result)
}

#1 c.
f(theta = 0.3,x)

#2 a.
houseprice <- read.csv("houseprice.csv",sep = ',')
library(tidyverse)
houseprice <- houseprice %>%
   mutate( year_type = ifelse(Build_year<1900,"centennial",
                              ifelse(Build_year<1960,"old","new")))

#2 b.
library(broom)
summary(houseprice)
houseprice$Type <- as.factor(houseprice$Type)
houseprice$year_type <- as.factor(houseprice$year_type)
houseprice$Town<- as.factor(houseprice$Town)
houseprice$University <- as.factor(houseprice$University)

plot(lm(Sale_amount~Sqft_home,data = houseprice))

edited <-houseprice[-c(2308,2320,5759),]

ggplot(edited,aes(x=Sqft_home ,y=Sale_amount/1000))+geom_point()+
   geom_smooth(method = "lm")+
   labs(y='Sale Amount (thousand)')

fit <- lm(Sale_amount~Sqft_home,data = edited)
```

```
glance(fit)
anova(fit)

fit2 <- lm(Sale_amount~., data=edited)
step(fit2)
fit3 <- lm(Sale_amount ~ Beds + Baths + Sqft_home + Sqft_lot +
                    Type + Build_year + Town + year_type, data = edited)
anova(fit,fit3)
glance(fit3)
tidy(fit3)
```