Statistical Learning and Data mining

Homework 4

M052040003 鍾冠毅

4.a. 使用越多的解釋變數可以得到更強的解釋力，則 TSS – RSS 增加，在 TSS 不變的狀況下，train RSS 將會下降，故使用三階多項式迴歸模型將有更低的 train RSS。

4.b. 當原本的模型應該是一階模型，卻使用三階模型，則會造成模型 overfitted 的情況，此時使用 test data，在三階模型會造成較高的 RSS，故使用一階模型有較低的 test RSS。

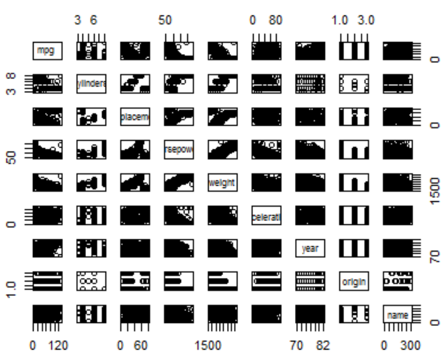
4.c. 同 4.a.，train RSS 不會因為真實模型而改變，越多的解釋變數會有更高的解釋力，彈性增加則 training error（train RSS）下降。

4.d. 資訊不足，test RSS 越低，則模型將估計得越好，若真實的迴歸模型是三階多項式，則其 test RSS 較一階多項式低。本題未告知真實模型為何，故無法回答。

5. $\hat{y}_i = x_i\hat{\beta} = x_i \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i'=1}^{n} {x_{i'}}^2} = \frac{\sum_{i'=1}^{n} x_{i'} x_i y_i}{\sum_{k=1}^{n} {x_k}^2} = \sum_{i'=1}^{n} \frac{x_{i'} x_i}{\sum_{k=1}^{n} {x_k}^2} y_{i'} = \sum_{i'=1}^{n} a_{i'} y_{i'}$，$a_{i'} = \frac{x_{i'} x_i}{\sum_{k=1}^{n} {x_k}^2}$

6. $y = \hat{\beta}_0 + \hat{\beta}_1\, x = \left(\bar{y} - \hat{\beta}_1\, \bar{x}\right) + \hat{\beta}_1\, x = \bar{y} - \hat{\beta}_1(x - \bar{x})$，take $x = \bar{x}$，$y = \bar{y}$。

9.a.



9.b.

```
> Auto.cor
                    mpg  cylinders displacement horsepower     weight acceleration       year     origin
mpg           1.0000000 -0.7776175   -0.8051269 -0.7784268 -0.8322442    0.4233285  0.5805410  0.5652088
cylinders    -0.7776175  1.0000000    0.9508233  0.8429834  0.8975273   -0.5046834 -0.3456474 -0.5689316
displacement -0.8051269  0.9508233    1.0000000  0.8972570  0.9329944   -0.5438005 -0.3698552 -0.6145351
horsepower   -0.7784268  0.8429834    0.8972570  1.0000000  0.8645377   -0.6891955 -0.4163615 -0.4551715
weight       -0.8322442  0.8975273    0.9329944  0.8645377  1.0000000   -0.4168392 -0.3091199 -0.5850054
acceleration  0.4233285 -0.5046834   -0.5438005 -0.6891955 -0.4168392    1.0000000  0.2903161  0.2127458
year          0.5805410 -0.3456474   -0.3698552 -0.4163615 -0.3091199    0.2903161  1.0000000  0.1815277
origin        0.5652088 -0.5689316   -0.6145351 -0.4551715 -0.5850054    0.2127458  0.1815277  1.0000000
```

9.c.

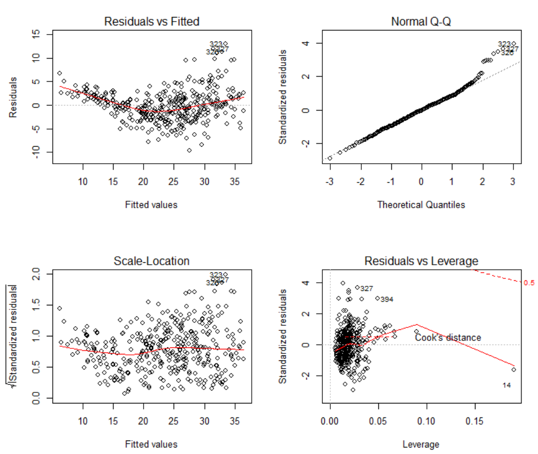
```
Call:
lm(formula = mpg ~ cylinders + displacement + horsepower + weight +
    acceleration + year + origin, data = Auto)

Residuals:
    Min      1Q  Median      3Q     Max
-9.5903 -2.1565 -0.1169  1.8690 13.0604

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -17.218435   4.644294  -3.707  0.00024 ***
cylinders     -0.493376   0.323282  -1.526  0.12780
displacement   0.019896   0.007515   2.647  0.00844 **
horsepower    -0.016951   0.013787  -1.230  0.21963
weight        -0.006474   0.000652  -9.929  < 2e-16 ***
acceleration   0.080576   0.098845   0.815  0.41548
year           0.750773   0.050973  14.729  < 2e-16 ***
origin         1.426141   0.278136   5.127 4.67e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.328 on 384 degrees of freedom
Multiple R-squared:  0.8215,    Adjusted R-squared:  0.8182
F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

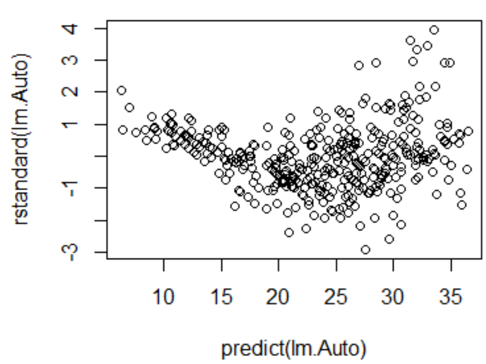
i. For the F-statistic , the p-value is small enough so that reject the null hypothesis that all beta are zero. Yes, there is a relationship between the predictors and the response.

ii. The p-values of displacement, weight, year and origin are smaller than 0.05 so that they have significant relationship to the response.

iii. Coefficient of the predictor, year, is 0.750773. That is, under the same condition, the mpg increases 0.750773 per year.

9.d.





由左圖左上殘差的分布有一定程度的趨勢，而非常態分佈，故此模型估計得不好。由右圖大於 3 的點為離群值；由左圖右下可發現 14 為較高的槓桿作用。

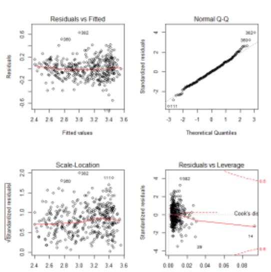
9.e.

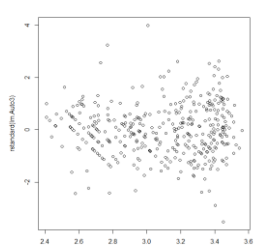
```
Call:
lm(formula = mpg ~ displacement + weight + year + origin + displacement:weight +
    displacement:year + displacement:origin + weight:year + weight:origin +
    year:origin)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8970 -1.5806 -0.1199  1.2215 14.1451

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         -1.792e+01  2.496e+01  -0.718  0.47325
displacement         3.382e-02  8.295e-02   0.408  0.68370
weight              -8.284e-03  1.119e-02  -0.740  0.45970
year                 9.045e-01  3.237e-01   2.795  0.00546 **
origin              -5.649e+00  5.352e+00  -1.055  0.29195
displacement:weight  1.806e-05  2.762e-06   6.540 1.98e-10 ***
displacement:year   -1.593e-03  1.137e-03  -1.401  0.16189
displacement:origin  1.605e-02  1.276e-02   1.258  0.20930
weight:year          5.751e-06  1.512e-04   0.038  0.96968
weight:origin       -1.343e-03  9.465e-04  -1.418  0.15688
year:origin          9.457e-02  6.619e-02   1.429  0.15387
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.95 on 381 degrees of freedom
Multiple R-squared:  0.8608,    Adjusted R-squared:  0.8571
F-statistic: 235.6 on 10 and 381 DF,  p-value: < 2.2e-16
```

displacement 與 weight 的交叉項對模型顯著影響。

9.f.

```
Call:
lm(formula = log(mpg) ~ sqrt(displacement) + (weight)^2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.55863 -0.10587 -0.00022  0.09845  0.63263

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        4.187e+00  3.096e-02 135.274  < 2e-16 ***
sqrt(displacement) -3.115e-02  6.475e-03  -4.811 2.15e-06 ***
weight            -2.250e-04  2.778e-05  -8.098 7.28e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1599 on 389 degrees of freedom
Multiple R-squared:  0.7799,    Adjusted R-squared:  0.7787
F-statistic: 689.1 on 2 and 389 DF,  p-value: < 2.2e-16
```





對 mpg 取 log、對 displacement 開根號、對 weight 取平方得到以上結果。每個變數對模型的影響皆為顯著。residual v.s. fitted 圖中，比 9.d.顯得分三均勻，故模型也較 9.d.好；leverage 圖中，各點分布更加集中靠左，惟 14 依然有較強的槓桿作用。Outlier 的部分則可以看到有少部分的點大於 3，屬於離群值。

11.a.

```
Call:
lm(formula = y ~ x + 0)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.9154 -0.6472 -0.1771  0.5056  2.3109 

Coefficients:
  Estimate Std. Error t value Pr(>|t|)    
x   1.9939     0.1065   18.73   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9586 on 99 degrees of freedom
Multiple R-squared:  0.7798,	Adjusted R-squared:  0.7776 
F-statistic: 350.7 on 1 and 99 DF,  p-value: < 2.2e-16
```

在$\beta = 0$的假設下顯著，意即拒絕此假設。

11.b.

```
Call:
lm(formula = x ~ y + 0)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.8699 -0.2368  0.1030  0.2858  0.8938 

Coefficients:
  Estimate Std. Error t value Pr(>|t|)    
y  0.39111    0.02089   18.73   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4246 on 99 degrees of freedom
Multiple R-squared:  0.7798,	Adjusted R-squared:  0.7776 
F-statistic: 350.7 on 1 and 99 DF,  p-value: < 2.2e-16
```

在$\beta = 0$的假設下顯著，意即拒絕此假設。

11.c. 在 11.a. 可將方程式表為$y = 2x + \varepsilon$，也可以在 11.b.中表為$x = 0.5(y - \varepsilon)$。

11.d. 18.73 與上述相同

$$t - statistic = \frac{\hat{\beta}}{SE(\hat{\beta})} = \frac{\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i'=1}^{n} x_{i'}^2}}{\sqrt{\frac{\sum_{i=1}^{n}(y_i - x_i\hat{\beta})^2}{(n-1)\sum_{i'=1}^{n} x_{i'}^2}}} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i'=1}^{n} x_{i'}^2}\sqrt{\frac{(n-1)\sum_{i'=1}^{n} x_{i'}^2}{\sum_{i=1}^{n}(y_i - x_i\hat{\beta})^2}}$$

$$= \frac{\sum_{i=1}^{n} x_i y_i \sqrt{n-1}}{\sqrt{\sum_{i'=1}^{n} x_{i'}^2 \sum_{i=1}^{n}(y_i - x_i\hat{\beta})^2}} = \frac{\sum_{i=1}^{n} x_i y_i \sqrt{n-1}}{\sqrt{\sum_{i'=1}^{n} x_{i'}^2 \sum_{i=1}^{n}(y_i^2 - 2\hat{\beta}x_i y_i + \hat{\beta}^2 {x_i}^2)}}$$

$$= \frac{\sum_{i=1}^{n} x_i y_i \sqrt{n-1}}{\sqrt{(\sum_{i'=1}^{n} x_{i'}^2 \sum_{i=1}^{n} y_i^2) - (\sum_{i=1}^{n} x_i y_i)^2}}$$

11.e.　11.a.和 11.b.所得之 t 統計量一樣。

11.f.

```
> lme1$coefficients
              Estimate Std. Error    t value     Pr(>|t|)
(Intercept) -0.03769261 0.09698729 -0.3886346 6.983896e-01
x            1.99893961 0.10772703 18.5555993 7.723851e-34
> lme2$coefficients
              Estimate Std. Error    t value     Pr(>|t|)
(Intercept) 0.03880394 0.04266144  0.9095787 3.652764e-01
y           0.38942451 0.02098690 18.5555993 7.723851e-34
```

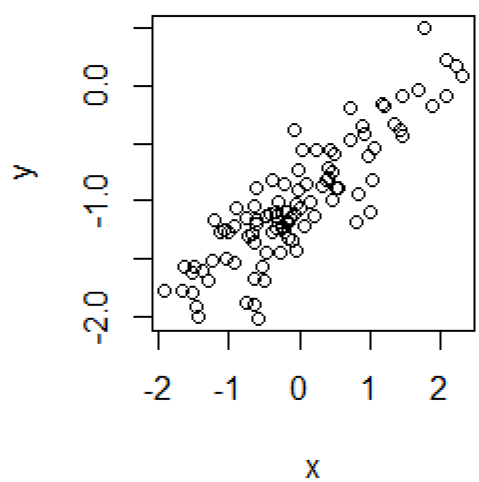
斜率的 t-value 一樣。

13.a.　see the appendix

13.b.　see the appendix

13.c.　see the appendix，長度為 100，截距為-1，斜率為 0.5

13.d.



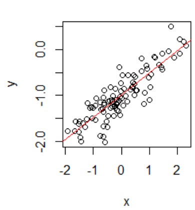
分布接近一條右上斜直線。

13.e.

```
Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
    -0.9931       0.4866
```

分別與原本的斜率和截距相近

13.f.



13.g.

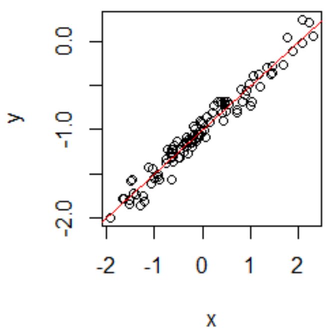
```
Call:
lm(formula = y ~ x + I(x^2))

Residuals:
     Min       1Q   Median       3Q      Max 
-0.72471 -0.13441  0.01034  0.15372  0.68402 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.02386    0.03336 -30.689   <2e-16 ***
x            0.47490    0.02825  16.811   <2e-16 ***
I(x^2)       0.03334    0.02288   1.457    0.148    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2581 on 97 degrees of freedom
Multiple R-squared:  0.7702,	Adjusted R-squared:  0.7654 
F-statistic: 162.5 on 2 and 97 DF,  p-value: < 2.2e-16
```

平方項不顯著。

13.h.

```
Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x  
     -1.005        0.498  
```

係數估計值分別更接近-1、0.5。



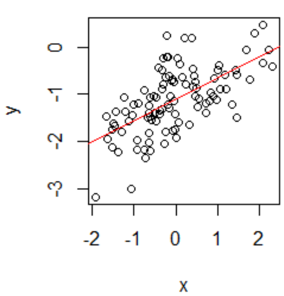

```
Call:
lm(formula = y ~ x + I(x^2))

Residuals:
      Min        1Q    Median        3Q       Max 
-0.250520 -0.063088 -0.003507  0.072026  0.219373 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.001697   0.012607 -79.455   <2e-16 ***
x            0.499046   0.010675  46.749   <2e-16 ***
I(x^2)      -0.003068   0.008645  -0.355    0.723    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09753 on 97 degrees of freedom
Multiple R-squared:  0.9607,	Adjusted R-squared:  0.9598 
F-statistic:  1184 on 2 and 97 DF,  p-value: < 2.2e-16
```

各點分布更接近一條斜直線。多項式迴歸中，平方項依然不顯著。

13.i.

```
Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x  
    -1.1017       0.4495  
```

係數估計值分別靠近-1、0.5。




```
Call:
lm(formula = y ~ x + I(x^2))

Residuals:
     Min       1Q   Median       3Q      Max 
-1.37794 -0.37310  0.00001  0.33765  1.37032 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.03447    0.06946 -14.893  < 2e-16 ***
x            0.47515    0.05882   8.079 1.83e-12 ***
I(x^2)      -0.07283    0.04763  -1.529     0.13    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5373 on 97 degrees of freedom
Multiple R-squared:  0.4046,	Adjusted R-squared:  0.3923 
F-statistic: 32.96 on 2 and 97 DF,  p-value: 1.198e-11
```

各點分布較不像一條斜直線。多項式迴歸中，平方項依然不顯著。

13.j.

	Original		Less noisy		noisier	
	lower	upper	lower	upper	lower	upper
β_0	-1.20914	-0.99425	-1.02381	-0.98525	-1.12130	-0.90713
β_1	0.33689	0.56218	0.47775	0.51818	0.42150	0.64383

越大的 noise 造成越寬的 CI

Appendix

```
### 9 ###

Auto <- read.csv("Auto.csv", header = T, sep =
",",na.strings="?")

Auto <- na.omit(Auto)

attach(Auto)

# 9.a. #

pairs(Auto)

# 9.b. #

Auto.cor <- cor(matrix(as.numeric(as.matrix(Auto[,-9])), ncol
= 8))

Auto.cor <- data.frame(Auto.cor)

colnames(Auto.cor) <- colnames(Auto)[1:8]

rownames(Auto.cor) <- colnames(Auto)[1:8]

Auto.cor

# 9.c. #

lm.Auto <- lm(mpg ~ cylinders + displacement + horsepower
+
    weight + acceleration + year + origin, data = Auto)

summary(lm.Auto)

# 9.d. #

par(mfrow = c(2,2))

plot(lm.Auto)

par(mfrow = c(1,1))

plot(predict(lm.Auto), rstandard(lm.Auto))

# 9.e. #

lm.Auto2 <- lm(mpg ~ displacement + weight + year + origin
+
  displacement:weight + displacement:year +
displacement:origin +
    weight:year + weight:origin + year:origin)

summary(lm.Auto2)

# 9.f. #

lm.Auto3 <- lm(log(mpg) ~ sqrt(displacement) + (weight)^2)

summary(lm.Auto3)

par(mfrow = c(2,2))

plot(lm.Auto3)

par(mfrow = c(1,1))

plot(predict(lm.Auto3), rstandard(lm.Auto3))
```

```
### 11 ###

set.seed(1)

x=rnorm (100)

y=2*x+rnorm (100)


# 11.a. #

summary(lm(y~x+0))


# 11.b. #

summary(lm(x~y+0))


# 11.e. #

(sqrt(length(x)-1) * sum(x*y)) / (sqrt(sum(x*x) * sum(y*y) -
(sum(x*y))^2))


# 11.f. #

lme1 <- summary(lm(y~x))

lme2 <- summary(lm(x~y))

lme1$coefficients

lme2$coefficients


### 13 ###

set.seed(1)

# 13.a. #

x <- rnorm(100,0,1)


# 13.b. #

eps <- rnorm(100,0,0.25)


# 13.c. #

y <- -1 + 0.5*x +eps

length(y)

# 13.d. #

plot(y ~ x)


# 13.e. #

lm13e <- lm(y~x)

confint(lm13e)

# 13.f. #

plot(y~x)

abline(lm13e, col= "red")


# 13.g. #

lm13g <- lm(y ~ x + I(x^2))

summary(lm13g)


# 13.h. #

eps <- rnorm(100, 0, 0.1)

y <- -1 + 0.5*x +eps

lm13h <- lm(y~x)

plot(y~x)

abline(lm13h, col= "red")

summary(lm(y ~ x + I(x^2)))

confint(lm13h)


# 13.i. #

eps <- rnorm(100, 0, 0.5)

y <- -1 + 0.5*x +eps

lm13i <- lm(y~x)

plot(y~x)

abline(lm13i, col= "red")

summary(lm(y ~ x + I(x^2)))

confint(lm13i)
```