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。

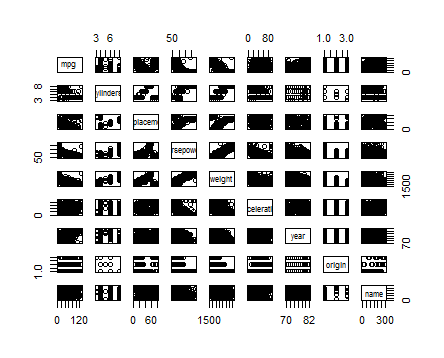
4.c. 同4.a.，train RSS不會因為真實模型而改變，越多的解釋變數會有更高的解釋力，彈性增加則training error（train RSS）下降。

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。

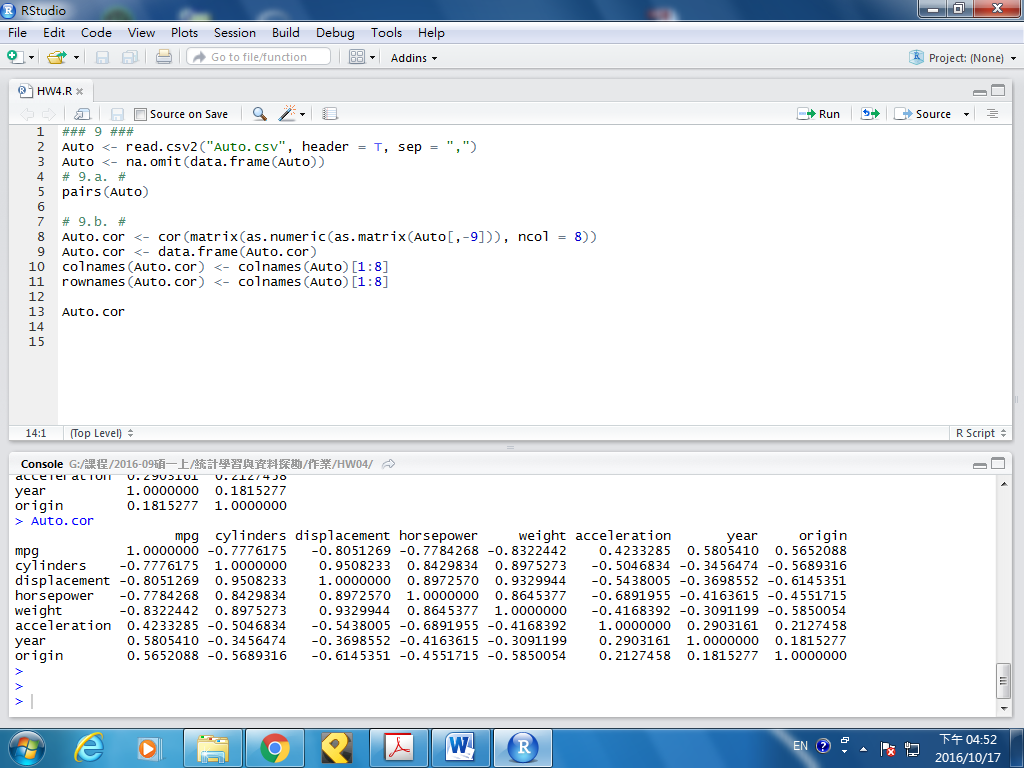
5.

6. ，take ，。

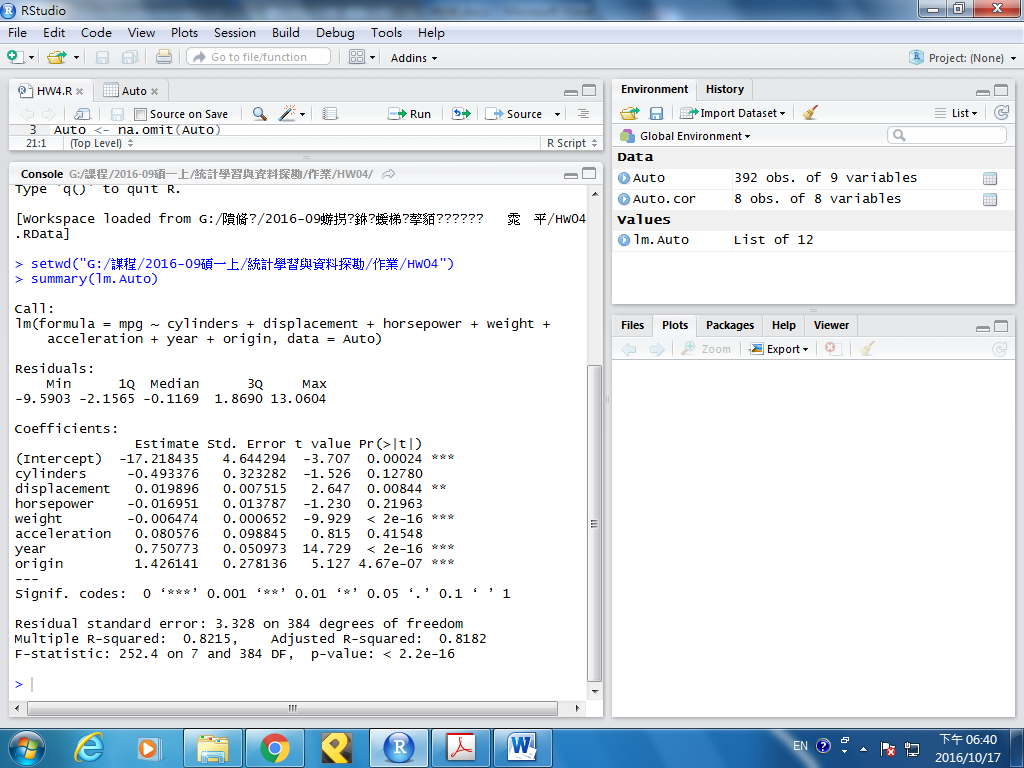
9.a.



9.b.

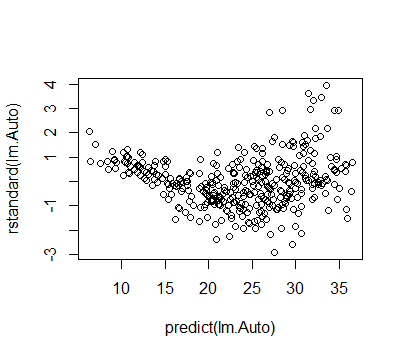
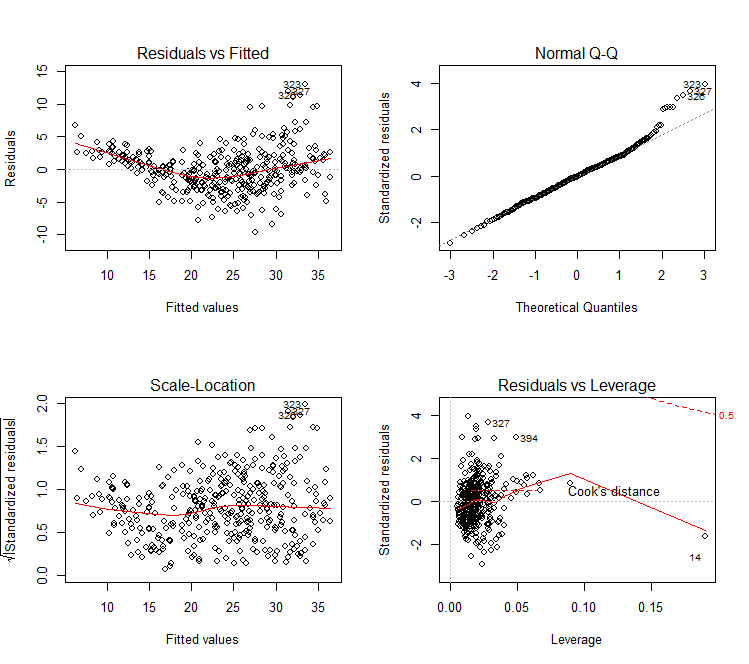


9.c.



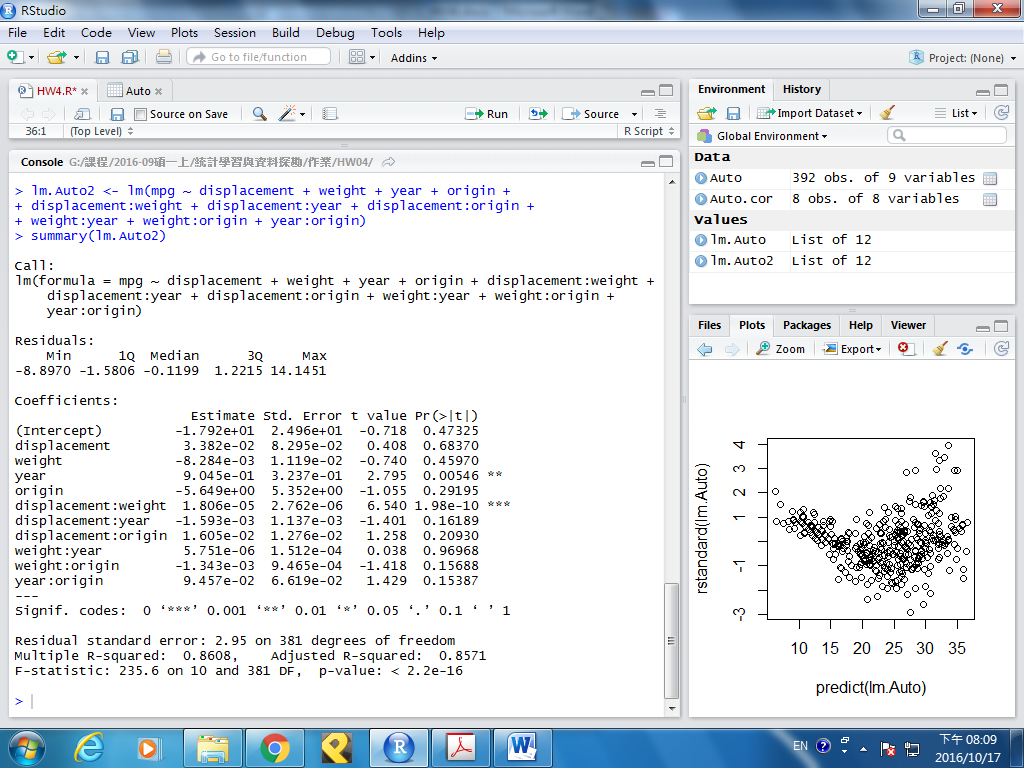
1. For the F-statistic , the p-value is small enough so that reject the null hypothesis that all beta are zero. Yes, there is a relationship between the predictors and the response.
2. The p-values of displacement, weight, year and origin are smaller than 0.05 so that they have significant relationship to the response.
3. Coefficient of the predictor, year, is 0.750773. That is, under the same condition, the mpg increases 0.750773 per year.

9.d.



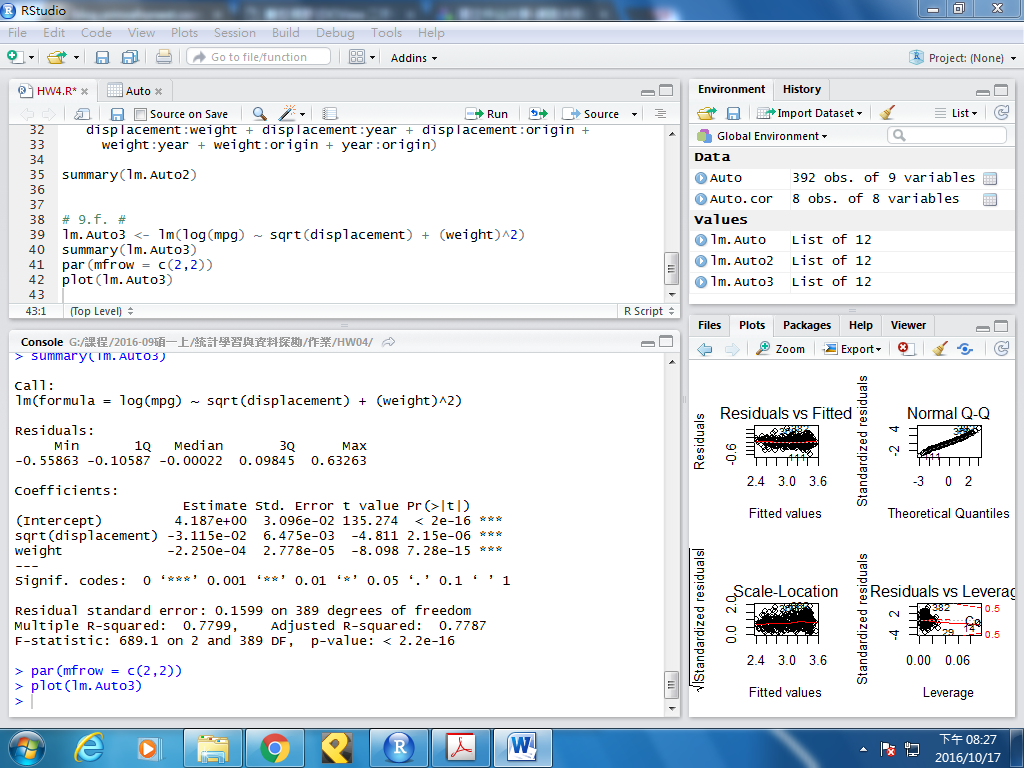
由左圖左上殘差的分布有一定程度的趨勢，而非常態分佈，故此模型估計得不好。由右圖大於3的點為離群值；由左圖右下可發現14為較高的槓桿作用。

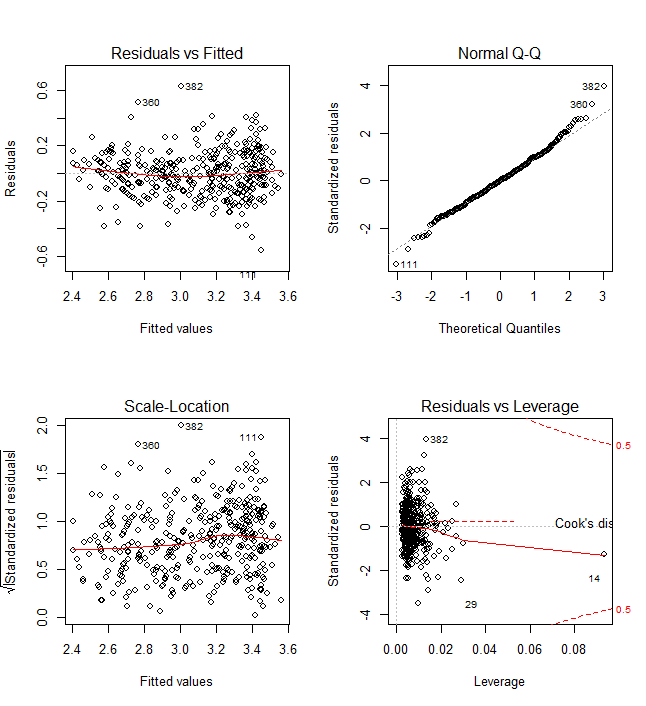
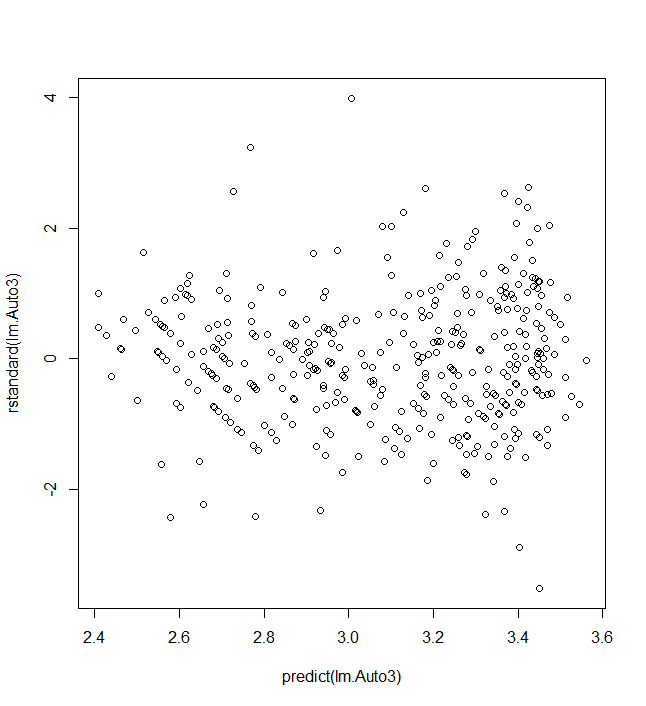
9.e.



displacement與weight的交叉項對模型顯著影響。

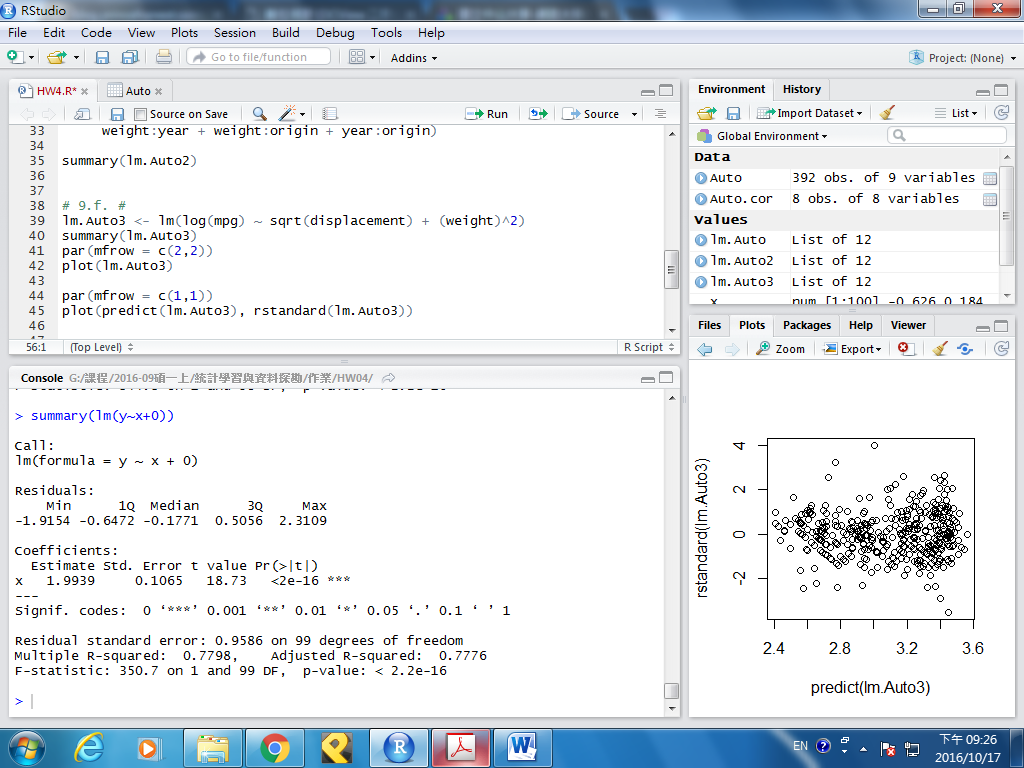
9.f.



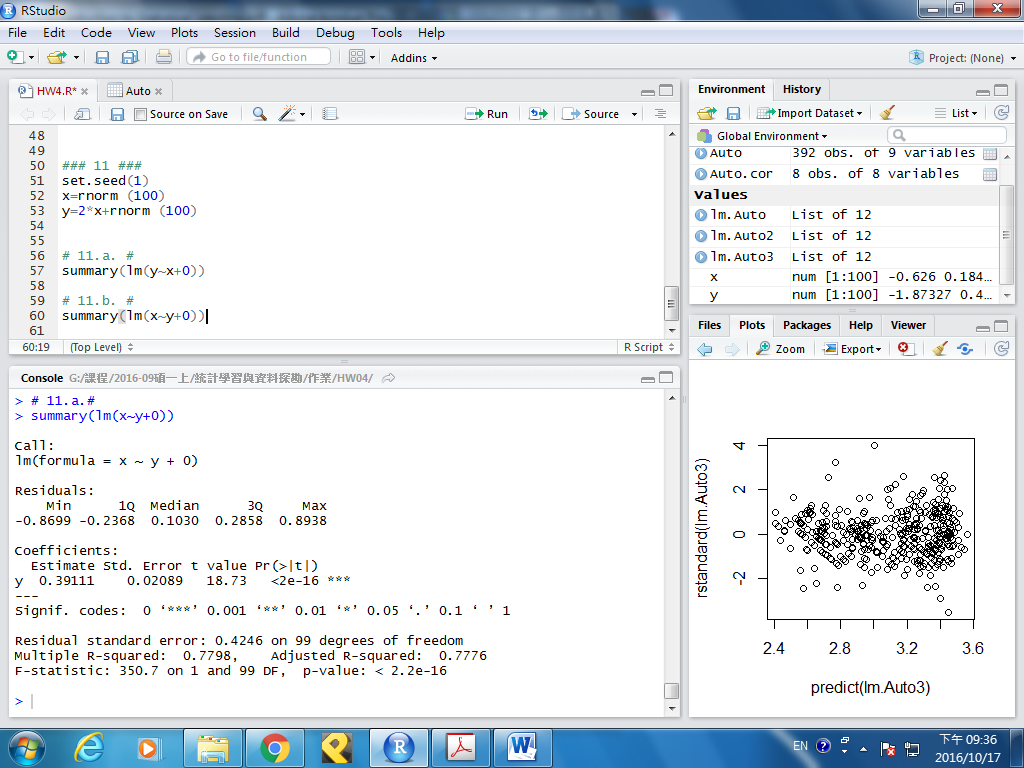
對mpg取log、對displacement開根號、對weight取平方得到以上結果。每個變數對模型的影響皆為顯著。residual v.s. fitted 圖中，比9.d.顯得分三均勻，故模型也較9.d.好；leverage圖中，各點分布更加集中靠左，惟14依然有較強的槓桿作用。Outlier的部分則可以看到有少部分的點大於3，屬於離群值。

11.a.



在的假設下顯著，意即拒絕此假設。

11.b.



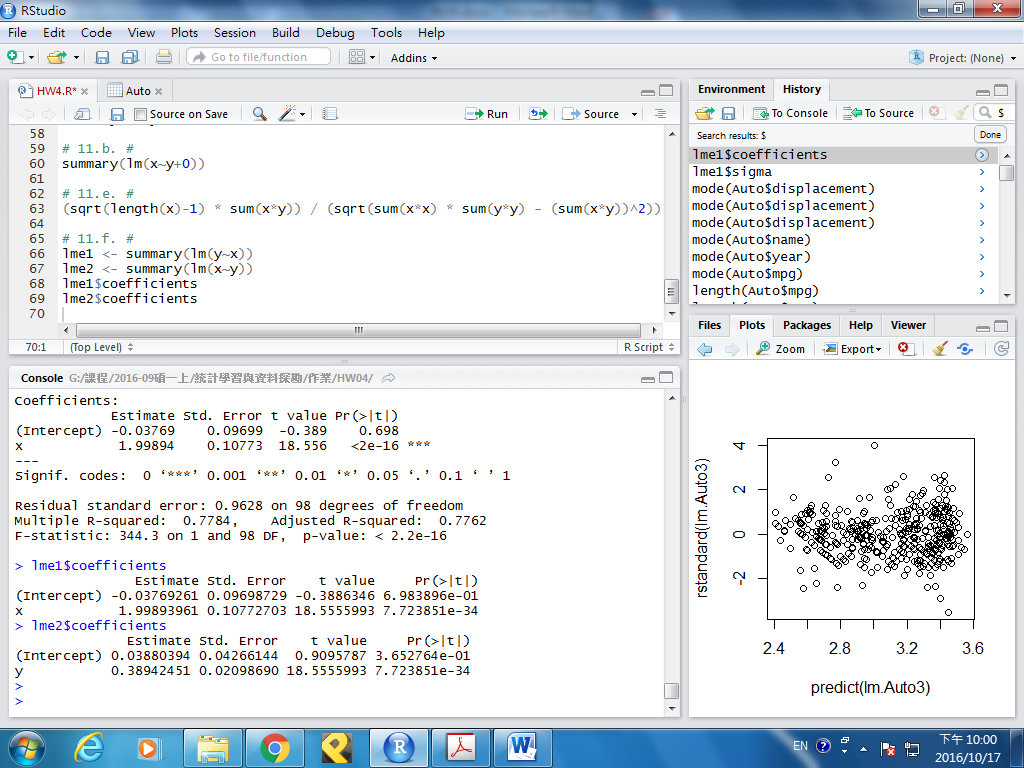
在的假設下顯著，意即拒絕此假設。

11.c. 在11.a. 可將方程式表為，也可以在11.b.中表為。

11.d. 18.73 與上述相同

11.e. 11.a.和11.b.所得之t統計量一樣。

11.f.



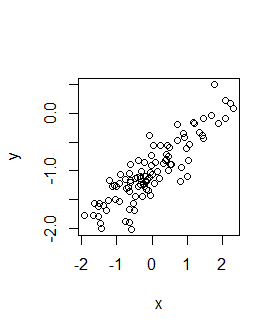
斜率的t-value一樣。

13.a. see the appendix

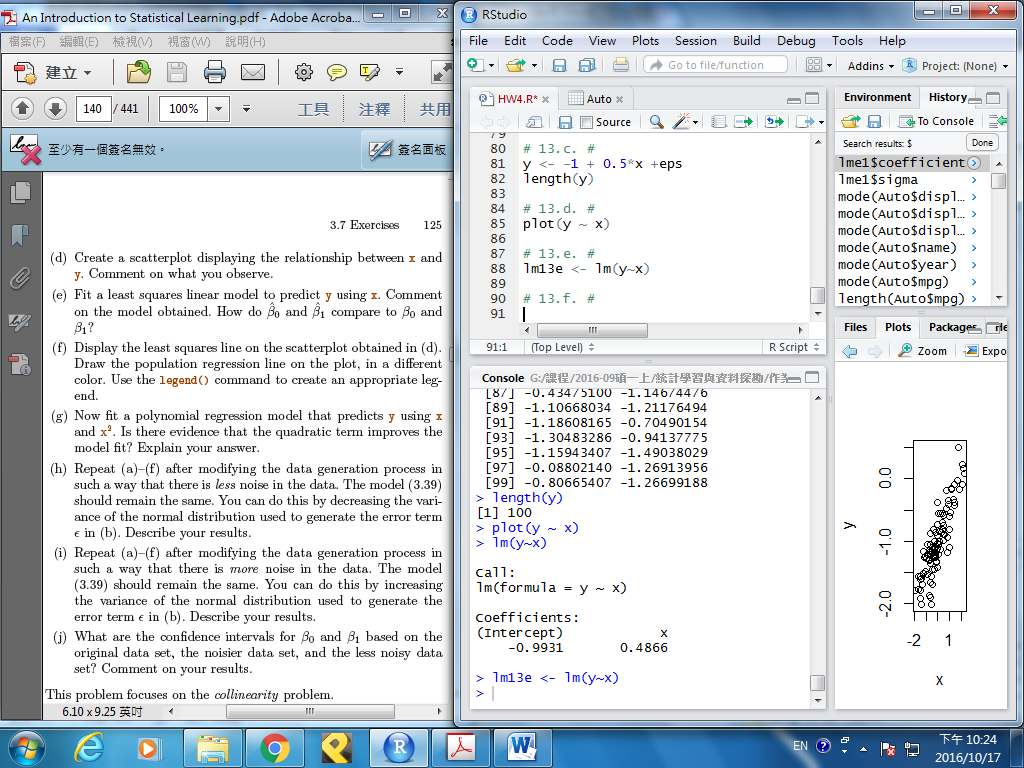
13.b. see the appendix

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

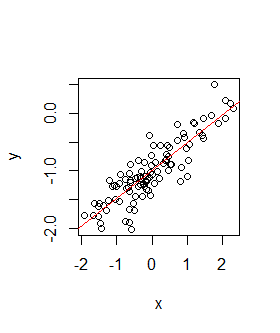
13.d.

分布接近一條右上斜直線。

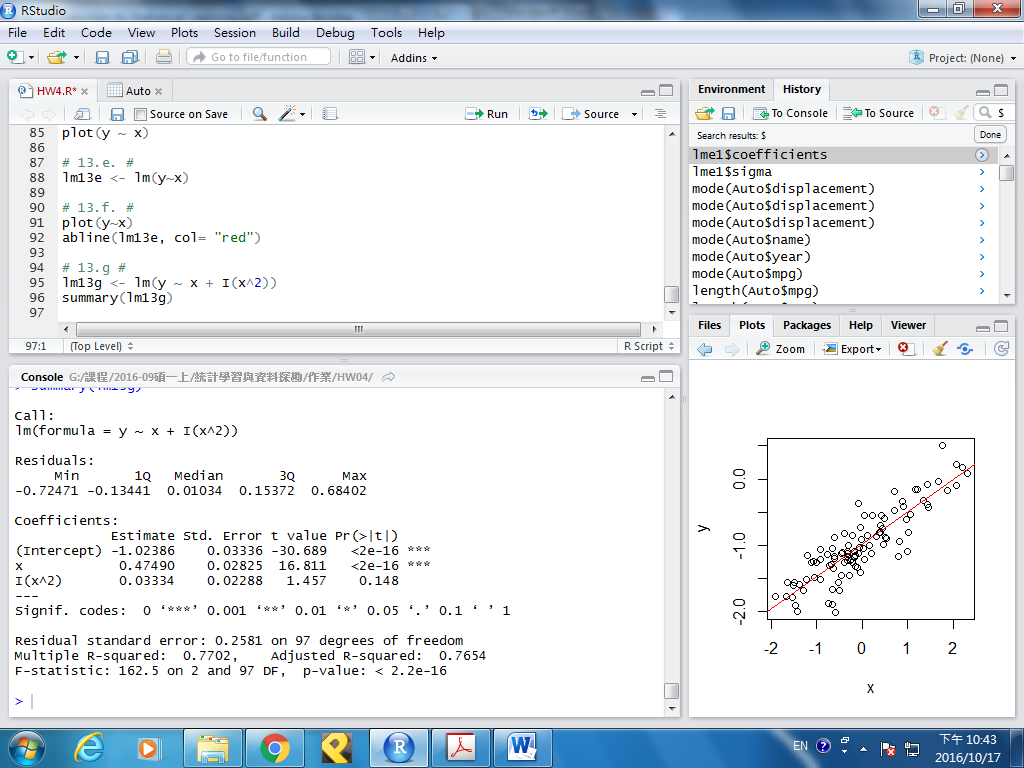
13.e.

 分別與原本的斜率和截距相近

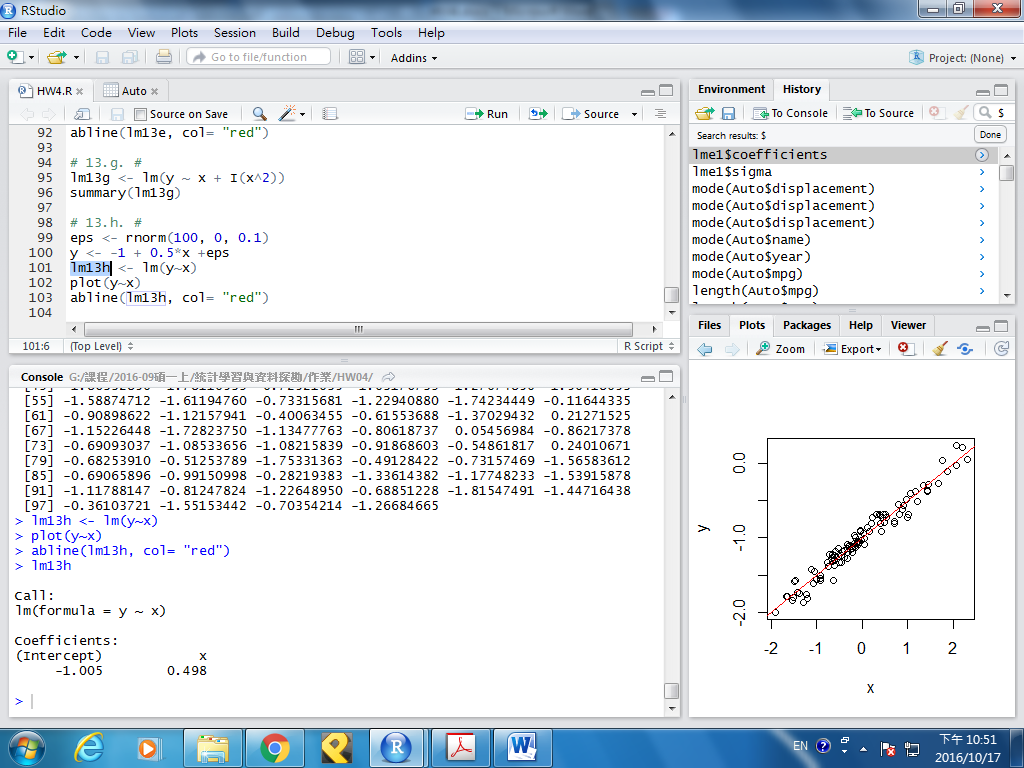
13.f.

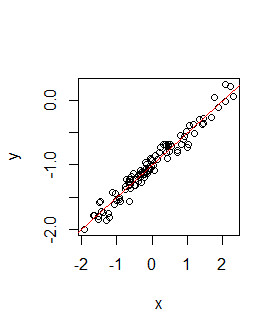
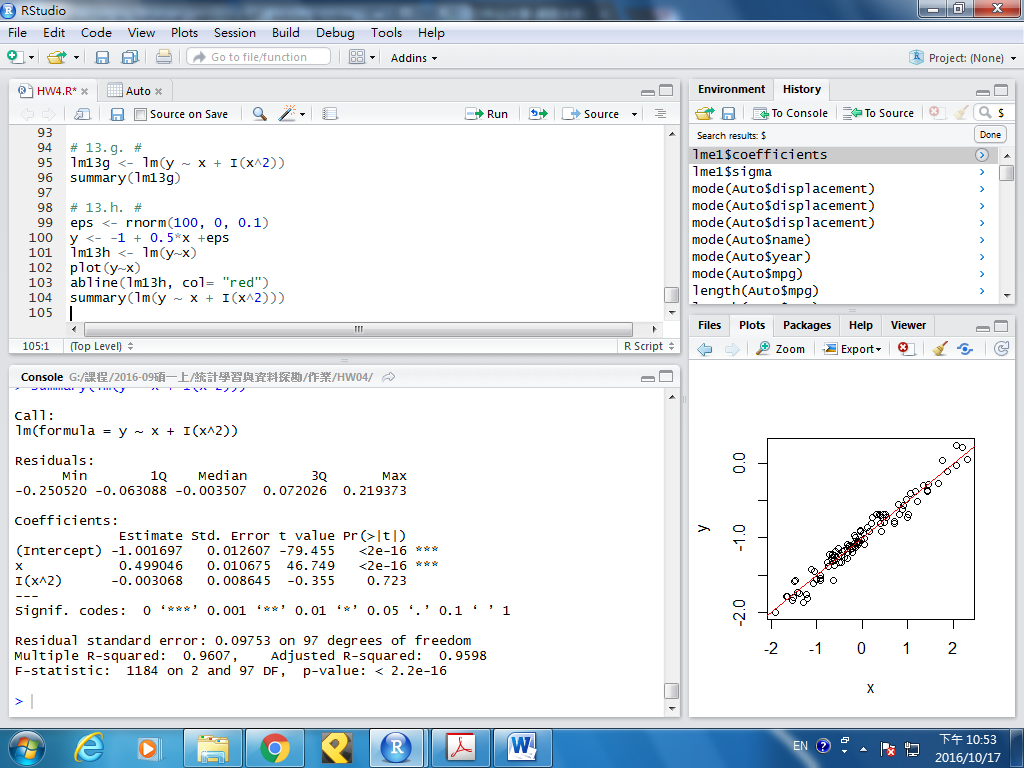


13.g.

平方項不顯著。

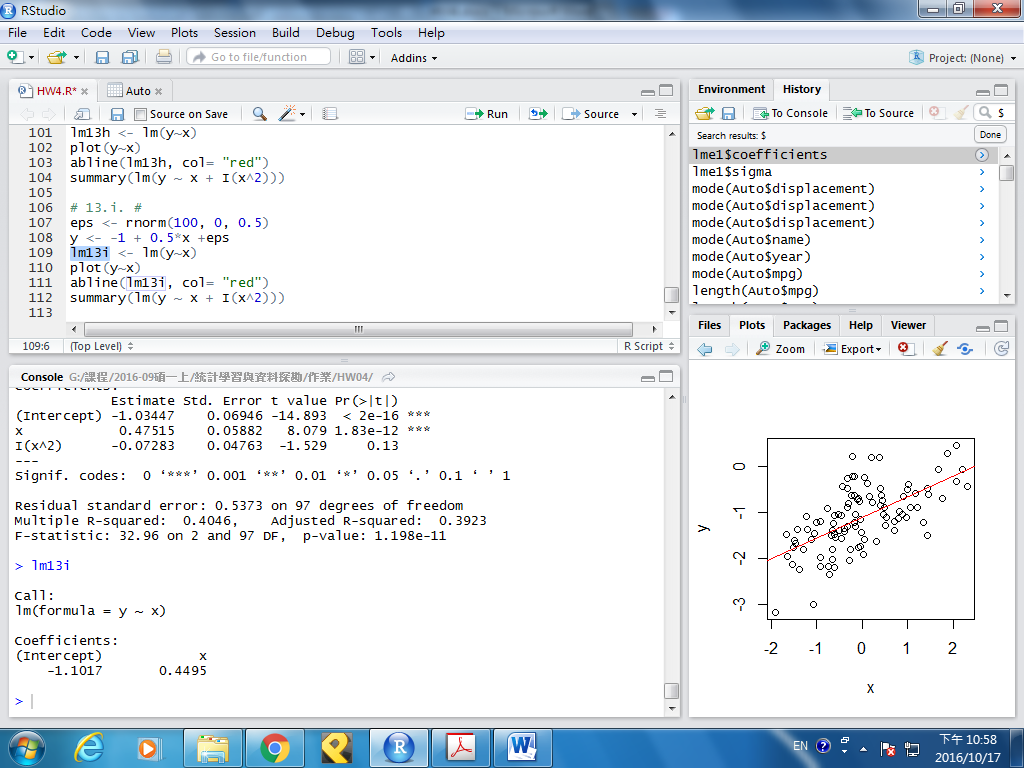
13.h.

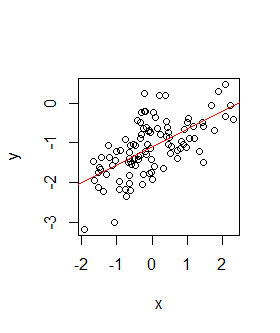
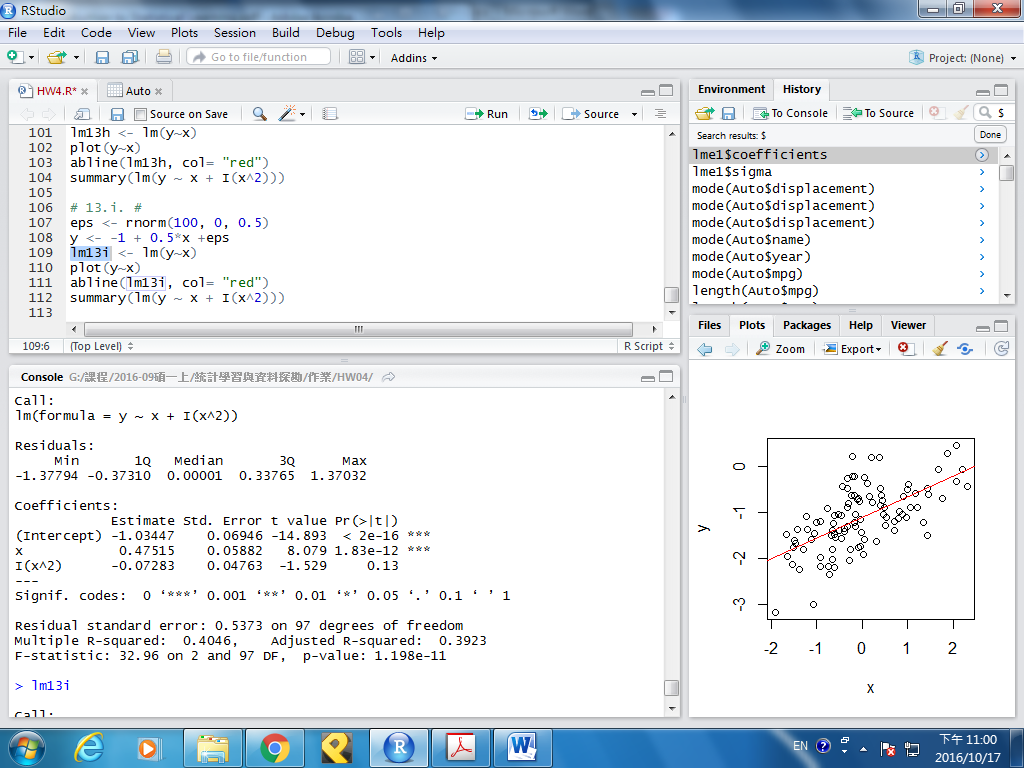
 係數估計值分別更接近-1、0.5。

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.

係數估計值分別靠近-1、0.5。

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 |
|  | -1.20914 | -0.99425 | -1.02381 | -0.98525 | -1.12130 | -0.90713 |
|  | 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)