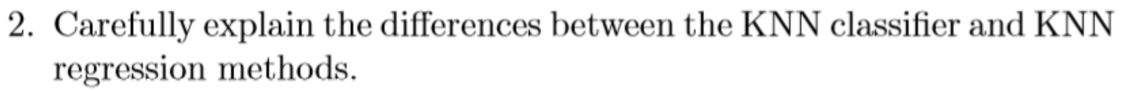
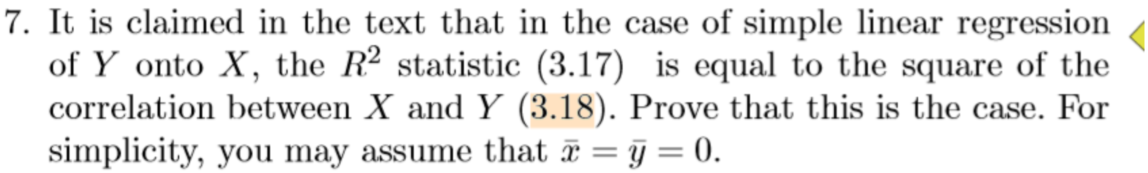
應數碩一 M072040019 梅瀚中



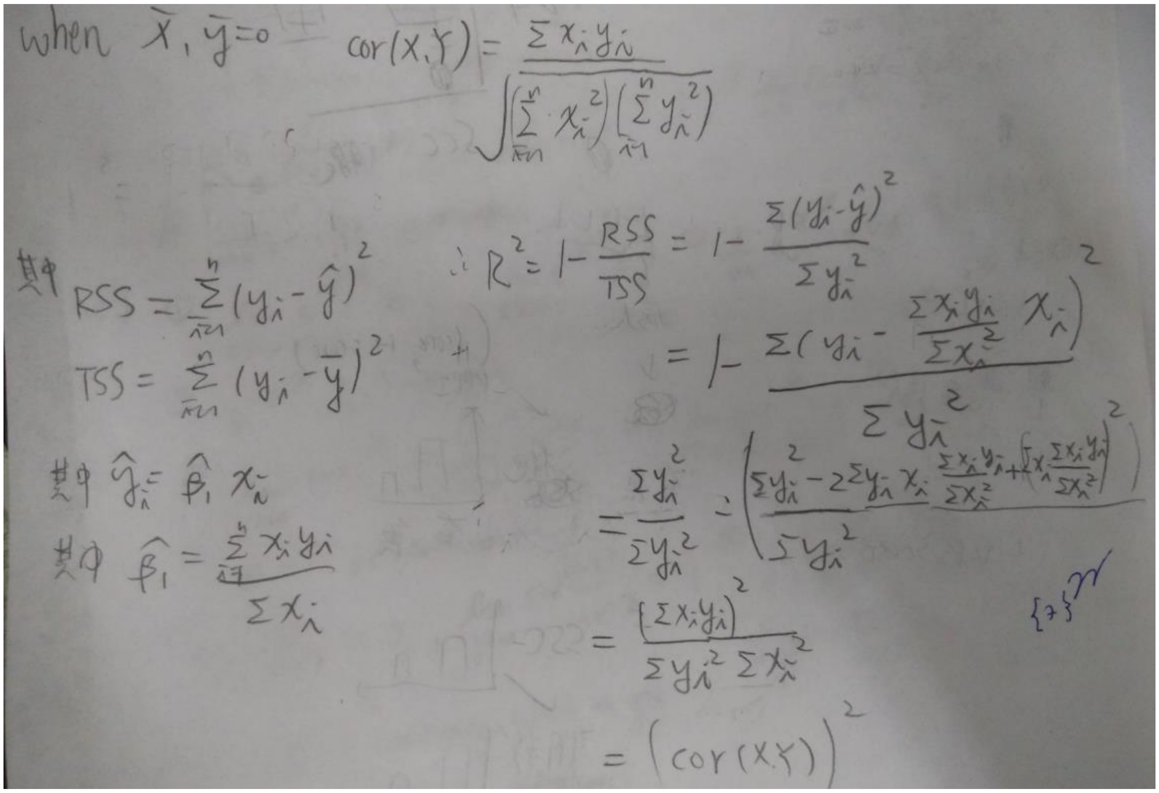
2. Carefully explain the differences between the KNN classifier and KNN regression methods.

Knn classifier 是一個分類的演算法,演算法的核心思想是如果一個樣本在空間中的 K 個最相鄰的樣本中的大多數屬於某一個類別,則該樣本也屬於這個類別

Knn regression 跟一般的 regression 一樣,應變數是屬於連續變數,想法是,一個樣本在空間中的 k 個最相臨的樣本的值(y),作平均,得到的值為此樣本的值。



7. It is claimed in the text that in the case of simple linear regression of Y onto X, the R^2 statistic (3.17) is equal to the square of the correlation between X and Y (3.18). Prove that this is the case. For simplicity, you may assume that $\bar{x} = \bar{y} = 0$.



when $\bar{x}, \bar{y} = 0$ $\quad \operatorname{cor}(X,Y) = \dfrac{\Sigma x_i y_i}{\sqrt{\left(\sum_{i=1}^{n} x_i^2\right)\left(\sum_{i=1}^{n} y_i^2\right)}}$

其中 $RSS = \sum_{i=1}^{n} (y_i - \hat{y})^2$

$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$

其中 $\hat{y}_i = \hat{\beta}_1 x_i$

其中 $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n} x_i y_i}{\Sigma x_i}$

$$\therefore R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\Sigma (y_i - \hat{y})^2}{\Sigma y_i^2}$$

$$= 1 - \frac{\Sigma \left(y_i - \frac{\Sigma x_i y_i}{\Sigma x_i^2} x_i \right)^2}{\Sigma y_i^2}$$

$$= \frac{\Sigma y_i^2}{\Sigma y_i^2} - \left(\frac{\Sigma y_i^2 - 2 \Sigma y_i x_i \frac{\Sigma x_i y_i}{\Sigma x_i^2} + \left(\Sigma x_i \frac{\Sigma x_i y_i}{\Sigma x_i^2} \right)^2}{\Sigma y_i^2} \right)$$

$$= \frac{(\Sigma x_i y_i)^2}{\Sigma y_i^2 \, \Sigma x_i^2}$$

$$= (\operatorname{cor}(X,Y))^2$$

{7}

9. This question involves the use of multiple linear regression on the `Auto` data set.

   (a) Produce a scatterplot matrix which includes all of the variables in the data set.

   (b) Compute the matrix of correlations between the variables using the function `cor()`. You will need to exclude the `name` variable, which is qualitative.

   (c) Use the `lm()` function to perform a multiple linear regression with `mpg` as the response and all other variables except `name` as the predictors. Use the `summary()` function to print the results. Comment on the output. For instance:

      i. Is there a relationship between the predictors and the response?

      ii. Which predictors appear to have a statistically significant relationship to the response?

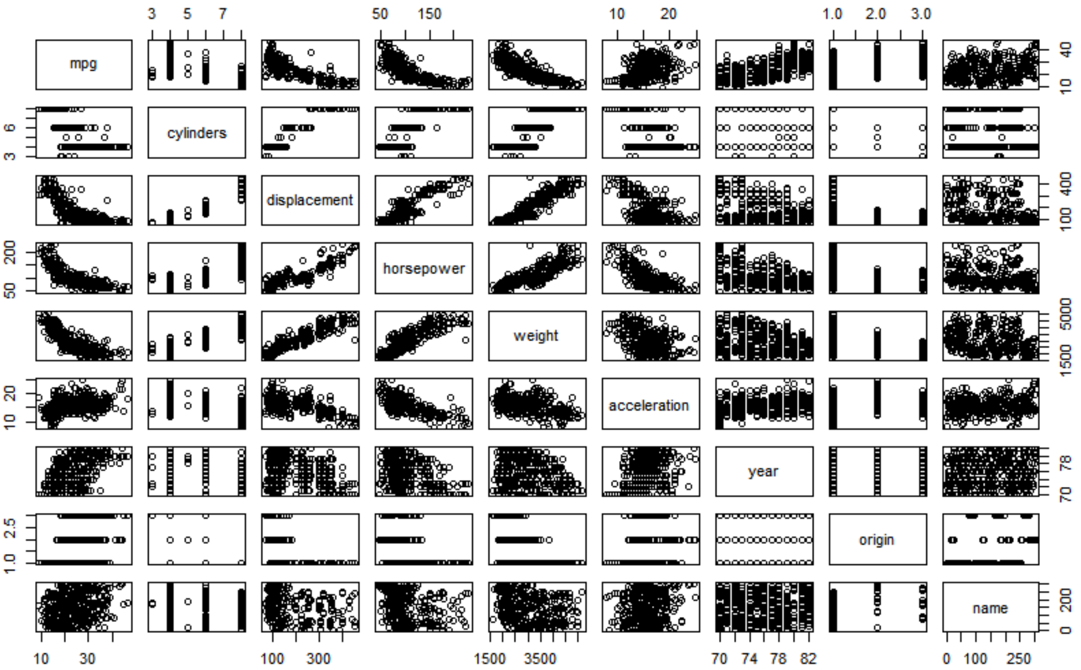
      iii. What does the coefficient for the `year` variable suggest?

   (d) Use the `plot()` function to produce diagnostic plots of the linear regression fit. Comment on any problems you see with the fit. Do the residual plots suggest any unusually large outliers? Does the leverage plot identify any observations with unusually high leverage?

   (e) Use the `*` and `:` symbols to fit linear regression models with interaction effects. Do any interactions appear to be statistically significant?

   (f) Try a few different transformations of the variables, such as $\log(X)$, $\sqrt{X}$, X^2. Comment on your findings.

(a)



(b)

```
> cor(Auto[1:8])
                    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
```

(c)

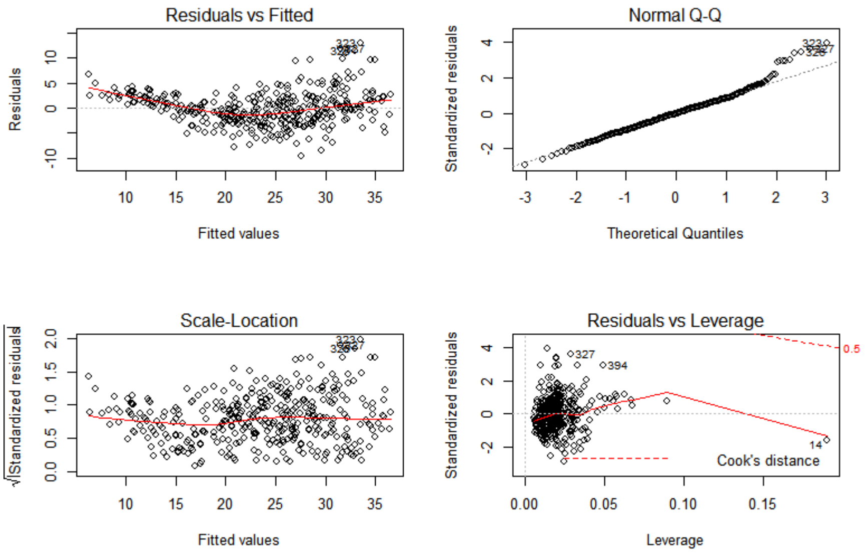
```
Call:
lm(formula = mpg ~ . - name, data = Auto)

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)p-value<2.2e-16,所以整體來說 y 跟 x 有關係。

(ii)displacement,weight,year 和 origin 變數都和 mpg 有顯著關係。

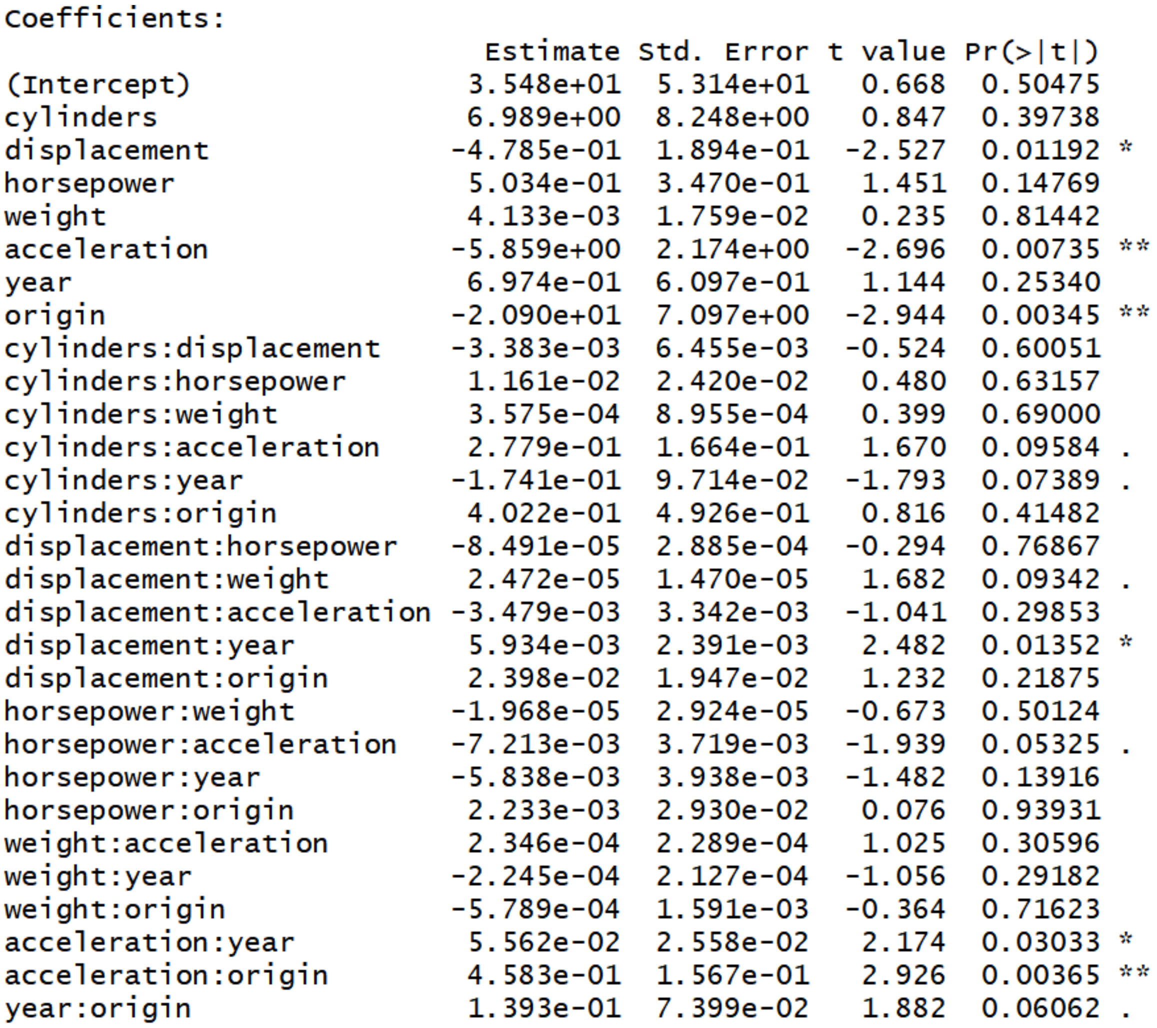
(III)平均來說,year 增加一單位,mpg 會增加 0.750773 單位

(d)



我們看左下方的圖,圖中左上方的幾個點的 standardized residuals 有一點大,但是還沒有明確的超過範圍。右下方的圖中我們可以看到點 14 是一個高的槓桿值。

(e)



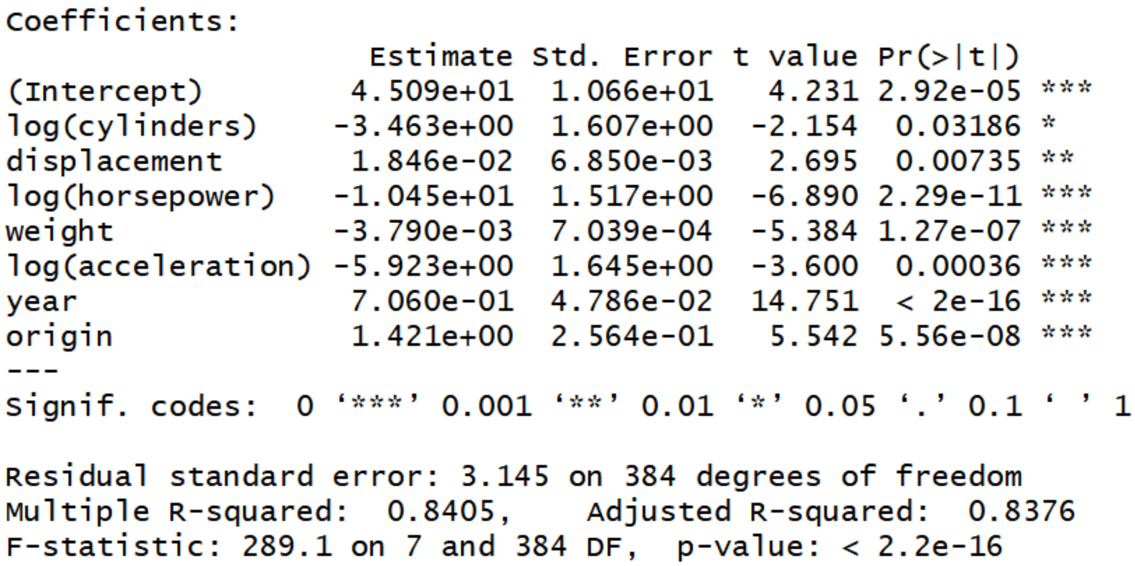
```
Coefficients:
                           Estimate Std. Error t value Pr(>|t|)   
(Intercept)               3.548e+01  5.314e+01   0.668  0.50475   
cylinders                 6.989e+00  8.248e+00   0.847  0.39738   
displacement             -4.785e-01  1.894e-01  -2.527  0.01192 * 
horsepower                5.034e-01  3.470e-01   1.451  0.14769   
weight                    4.133e-03  1.759e-02   0.235  0.81442   
acceleration             -5.859e+00  2.174e+00  -2.696  0.00735 **
year                      6.974e-01  6.097e-01   1.144  0.25340   
origin                   -2.090e+01  7.097e+00  -2.944  0.00345 **
cylinders:displacement   -3.383e-03  6.455e-03  -0.524  0.60051   
cylinders:horsepower      1.161e-02  2.420e-02   0.480  0.63157   
cylinders:weight          3.575e-04  8.955e-04   0.399  0.69000   
cylinders:acceleration    2.779e-01  1.664e-01   1.670  0.09584 . 
cylinders:year           -1.741e-01  9.714e-02  -1.793  0.07389 . 
cylinders:origin          4.022e-01  4.926e-01   0.816  0.41482   
displacement:horsepower  -8.491e-05  2.885e-04  -0.294  0.76867   
displacement:weight       2.472e-05  1.470e-05   1.682  0.09342 . 
displacement:acceleration -3.479e-03  3.342e-03  -1.041  0.29853   
displacement:year         5.934e-03  2.391e-03   2.482  0.01352 * 
displacement:origin       2.398e-02  1.947e-02   1.232  0.21875   
horsepower:weight        -1.968e-05  2.924e-05  -0.673  0.50124   
horsepower:acceleration  -7.213e-03  3.719e-03  -1.939  0.05325 . 
horsepower:year          -5.838e-03  3.938e-03  -1.482  0.13916   
horsepower:origin         2.233e-03  2.930e-02   0.076  0.93931   
weight:acceleration       2.346e-04  2.289e-04   1.025  0.30596   
weight:year              -2.245e-04  2.127e-04  -1.056  0.29182   
weight:origin            -5.789e-04  1.591e-03  -0.364  0.71623   
acceleration:year         5.562e-02  2.558e-02   2.174  0.03033 * 
acceleration:origin       4.583e-01  1.567e-01   2.926  0.00365 **
year:origin               1.393e-01  7.399e-02   1.882  0.06062 . 
```

Displacement 和 year 的交互作用項有在 alpha 設定為 0.05 之下有顯著效應,

Acceleration 和 year 的交互作用項有在 alpha 設定為 0.05 之下有顯著效應,

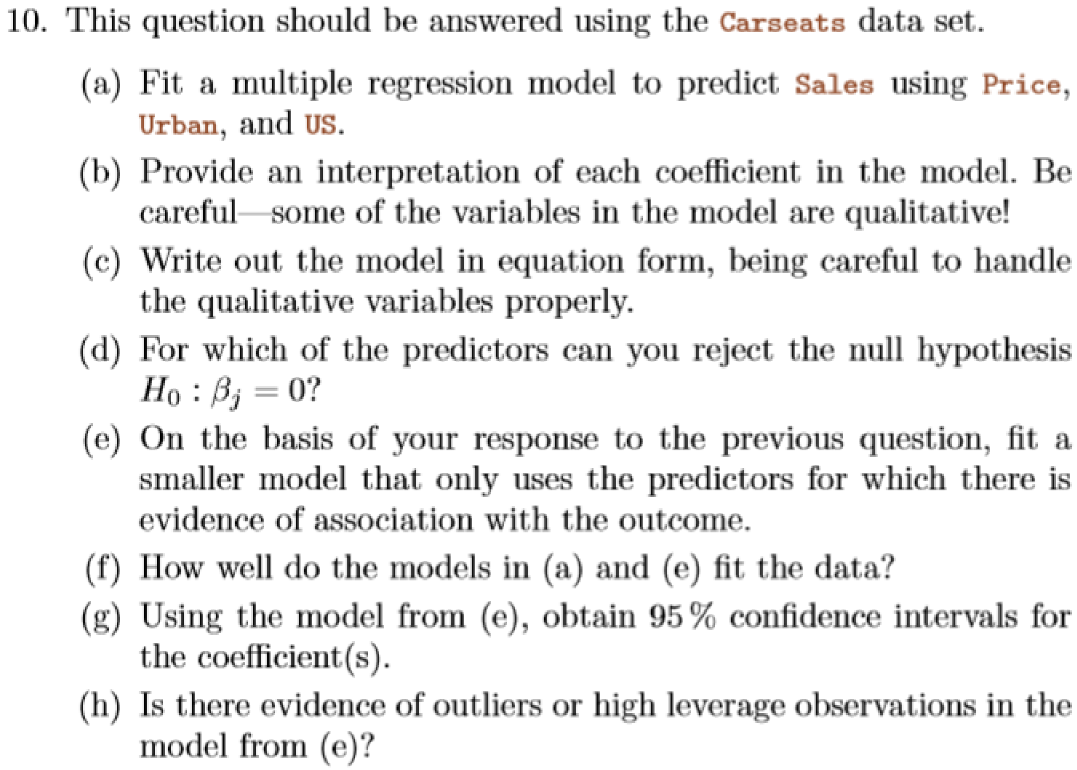
Acceleration 和 origin 的交互作用項有顯著效應。

(f)



```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)       4.509e+01  1.066e+01   4.231 2.92e-05 ***
log(cylinders)   -3.463e+00  1.607e+00  -2.154  0.03186 *  
displacement      1.846e-02  6.850e-03   2.695  0.00735 ** 
log(horsepower)  -1.045e+01  1.517e+00  -6.890 2.29e-11 ***
weight           -3.790e-03  7.039e-04  -5.384 1.27e-07 ***
log(acceleration) -5.923e+00  1.645e+00  -3.600  0.00036 ***
year              7.060e-01  4.786e-02  14.751  < 2e-16 ***
origin            1.421e+00  2.564e-01   5.542 5.56e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.145 on 384 degrees of freedom
Multiple R-squared:  0.8405,	Adjusted R-squared:  0.8376 
F-statistic: 289.1 on 7 and 384 DF,  p-value: < 2.2e-16
```

把(c)原本不顯著的變數加上 log 後,變數變成顯著。



10. This question should be answered using the Carseats data set.

(a) Fit a multiple regression model to predict Sales using Price, Urban, and US.

(b) Provide an interpretation of each coefficient in the model. Be careful—some of the variables in the model are qualitative!

(c) Write out the model in equation form, being careful to handle the qualitative variables properly.

(d) For which of the predictors can you reject the null hypothesis $H_0 : \beta_j = 0$?

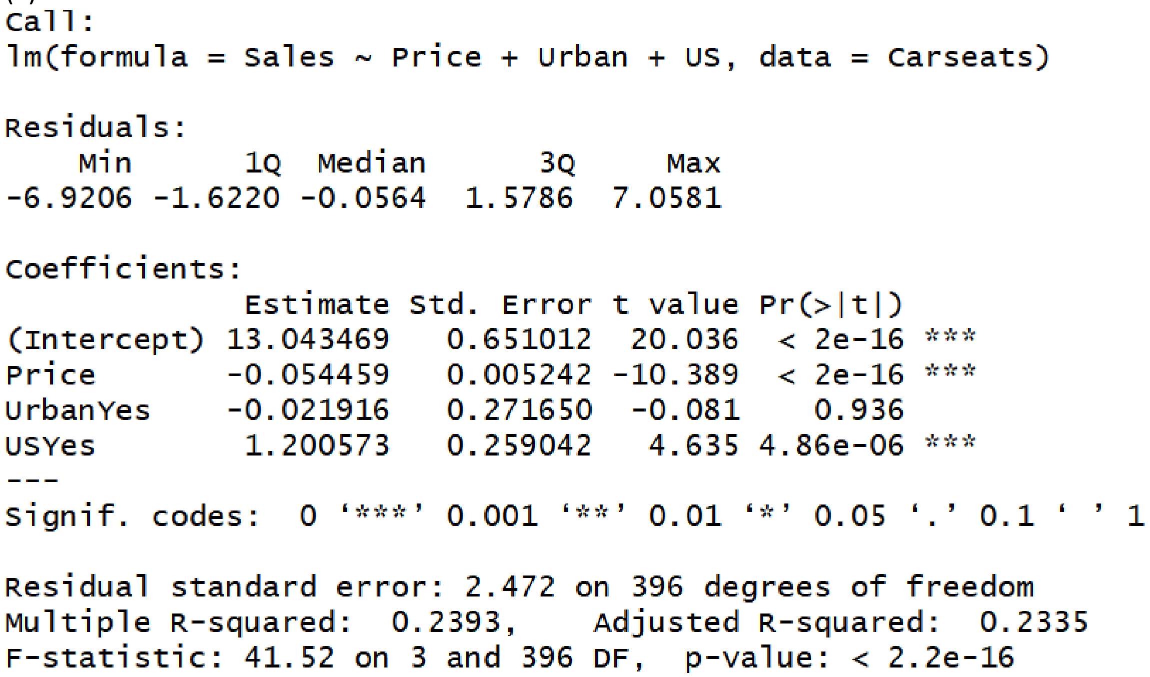
(e) On the basis of your response to the previous question, fit a smaller model that only uses the predictors for which there is evidence of association with the outcome.

(f) How well do the models in (a) and (e) fit the data?

(g) Using the model from (e), obtain 95 % confidence intervals for the coefficient(s).

(h) Is there evidence of outliers or high leverage observations in the model from (e)?

(a)



```
Call:
lm(formula = Sales ~ Price + Urban + US, data = Carseats)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.9206 -1.6220 -0.0564  1.5786  7.0581 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 13.043469   0.651012  20.036  < 2e-16 ***
Price       -0.054459   0.005242 -10.389  < 2e-16 ***
UrbanYes    -0.021916   0.271650  -0.081    0.936    
USYes        1.200573   0.259042   4.635 4.86e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.472 on 396 degrees of freedom
Multiple R-squared:  0.2393,	Adjusted R-squared:  0.2335 
F-statistic: 41.52 on 3 and 396 DF,  p-value: < 2.2e-16
```

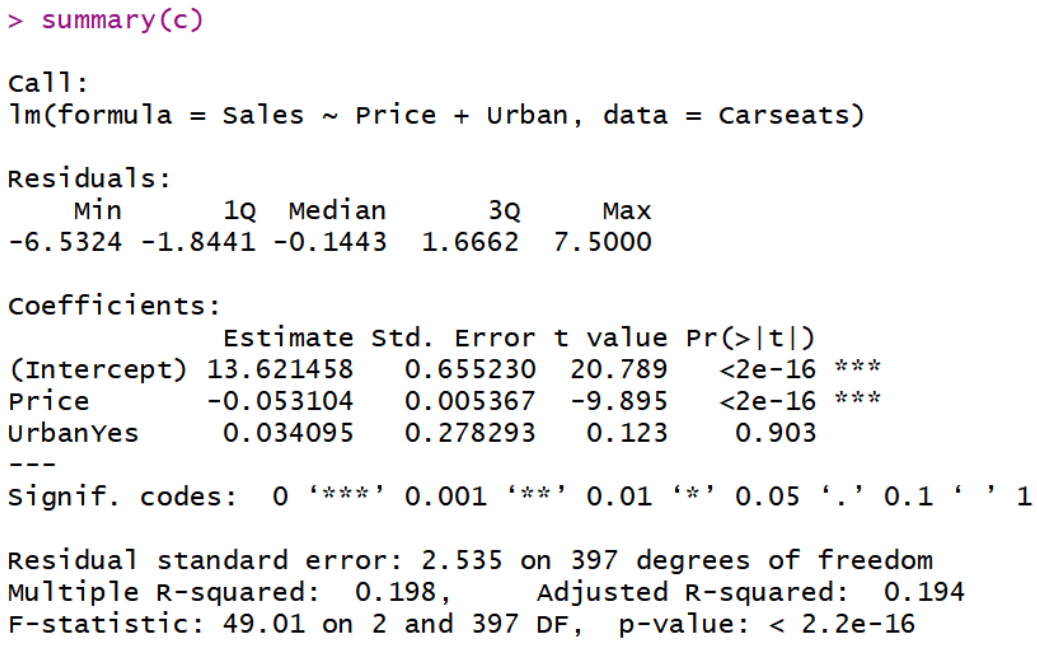
(b)

Price:公司對每個站點的汽車座椅收費
Urban: 表示商店是在城市還是鄉村
US: 表示商店是否在美國

(C)
Sales=13.043469-0.054459 price-0.021916 Urban+1.200573 US

(D)
Price 的 p-value<2e-16 ,
US 的 P-value 的 4.86e-06
這兩個變數 reject H0

(e)

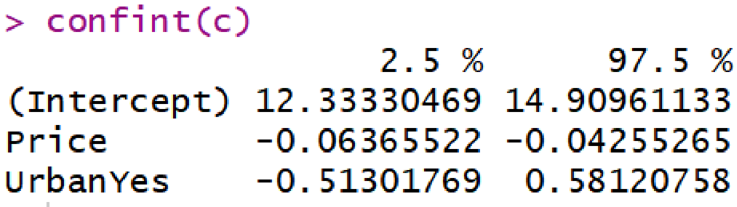


```
> summary(c)

Call:
lm(formula = Sales ~ Price + Urban, data = Carseats)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.5324 -1.8441 -0.1443  1.6662  7.5000 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 13.621458   0.655230  20.789   <2e-16 ***
Price       -0.053104   0.005367  -9.895   <2e-16 ***
UrbanYes     0.034095   0.278293   0.123    0.903    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.535 on 397 degrees of freedom
Multiple R-squared:  0.198,	Adjusted R-squared:  0.194 
F-statistic: 49.01 on 2 and 397 DF,  p-value: < 2.2e-16
```

(F)

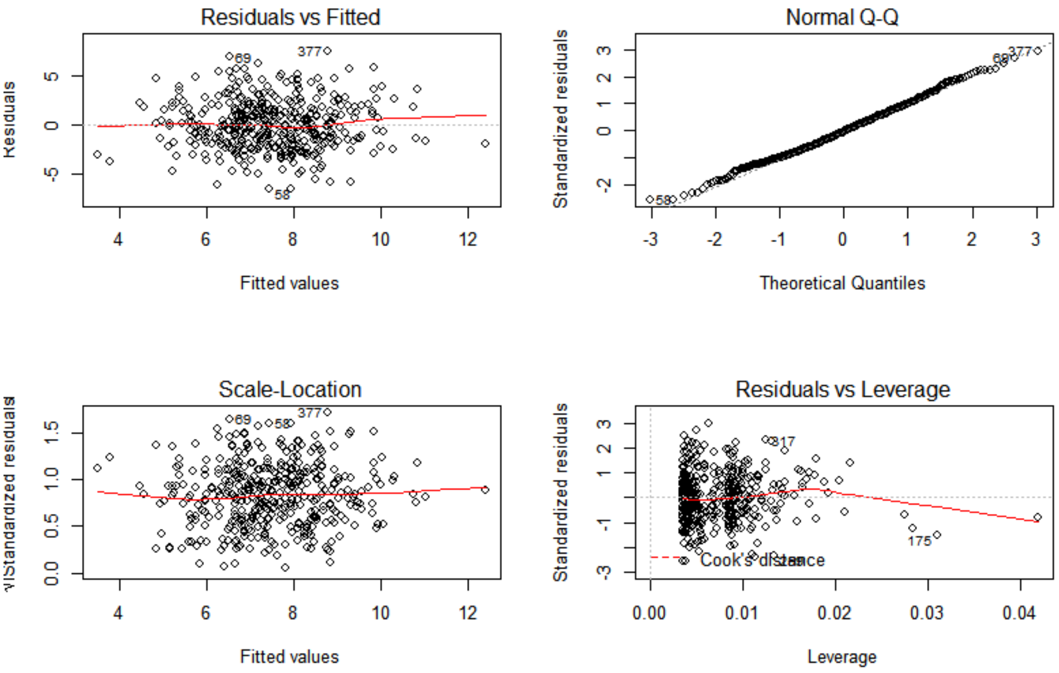
只用 price,Urban 比用 price,Urban,US 的 R^2 小,代表此模型解釋能力比較差。

(G)

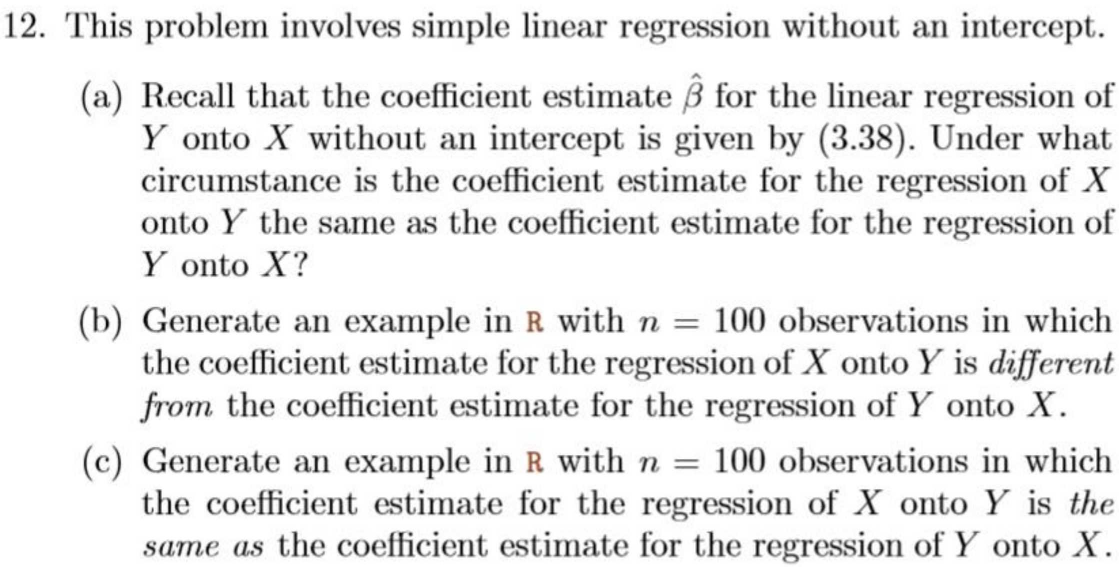


```
> confint(c)
                  2.5 %      97.5 %
(Intercept) 12.33330469 14.90961133
Price       -0.06365522 -0.04255265
UrbanYes    -0.51301769  0.58120758
```

(h)



點 377,69 ,58 都是 outliers。點 175 是 high leverage



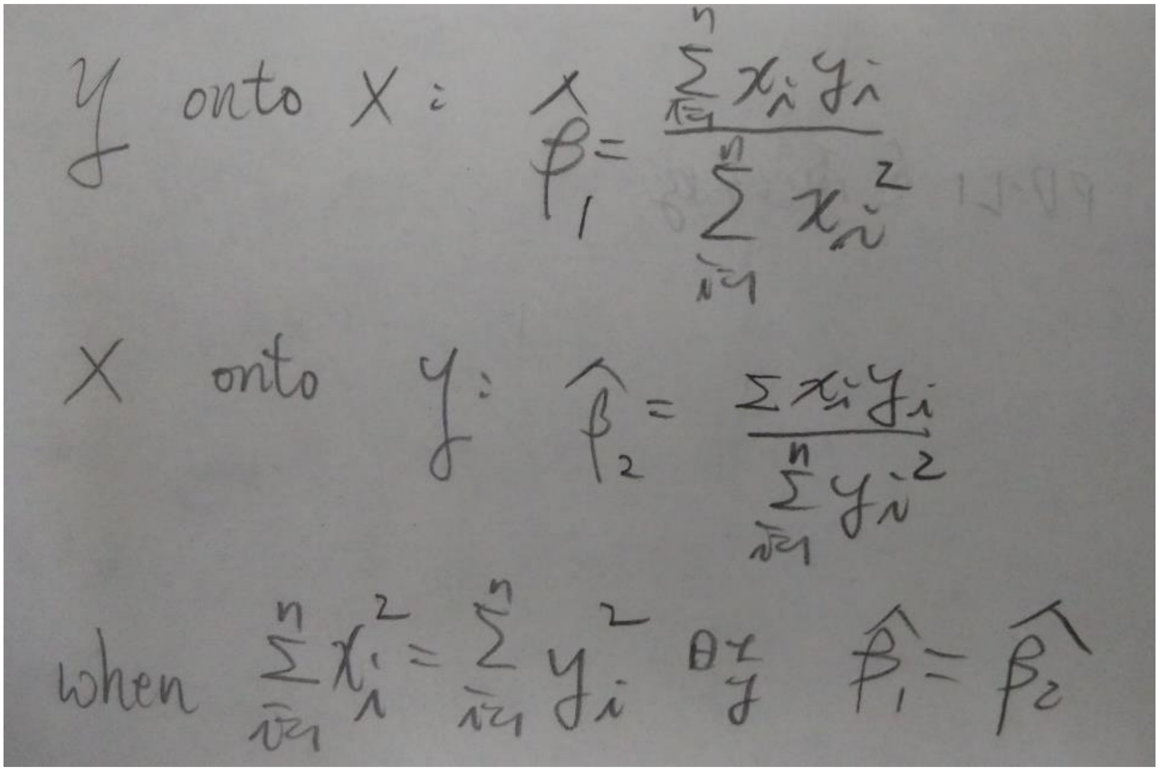
12. This problem involves simple linear regression without an intercept.

(a) Recall that the coefficient estimate $\hat{\beta}$ for the linear regression of Y onto X without an intercept is given by (3.38). Under what circumstance is the coefficient estimate for the regression of X onto Y the same as the coefficient estimate for the regression of Y onto X?

(b) Generate an example in R with $n = 100$ observations in which the coefficient estimate for the regression of X onto Y is *different from* the coefficient estimate for the regression of Y onto X.

(c) Generate an example in R with $n = 100$ observations in which the coefficient estimate for the regression of X onto Y is *the same as* the coefficient estimate for the regression of Y onto X.

(a)

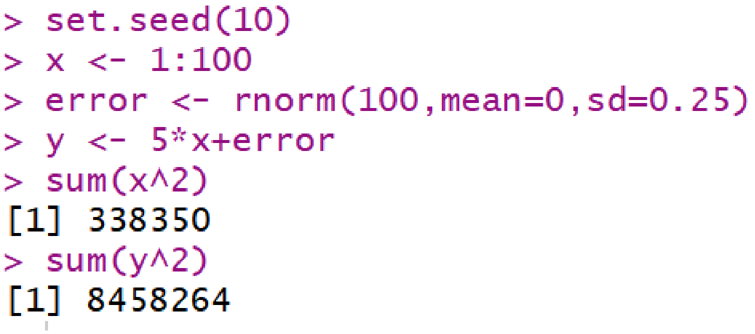


y onto X : $\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$

X onto y : $\hat{\beta}_2 = \frac{\sum x_i y_i}{\sum_{i=1}^{n} y_i^2}$

when $\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i^2$ 時 $\hat{\beta}_1 = \hat{\beta}_2$

(b) 查看無截距模型的 beta 係數時,X 和 Y 互換,只需要看分母值相不相同:

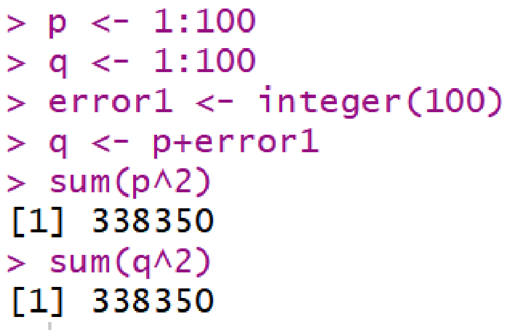


```
> set.seed(10)
> x <- 1:100
> error <- rnorm(100,mean=0,sd=0.25)
> y <- 5*x+error
> sum(x^2)
[1] 338350
> sum(y^2)
[1] 8458264
```

這裡 sum(x^2)和 sum(y^2)不同 所以估計的 beta 就不相同

(c)

把 error 設為 0



```
> p <- 1:100
> q <- 1:100
> error1 <- integer(100)
> q <- p+error1
> sum(p^2)
[1] 338350
> sum(q^2)
[1] 338350
```