HW2

賴冠維

2020/10/12

## 清理資料、資料型態轉換

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(magrittr)  
library(ggfortify)

## Loading required package: ggplot2

library(ISLR)

Auto = read.table("C:/Users/Lai/Desktop/統計學習/Auto.data",header = T)

for (i in 1:7) {  
 Auto[,i] = as.numeric(Auto[,i])  
}

## Warning: 強制變更過程中產生了 NA

for (i in 8:9) {  
 Auto[,i] = as.factor(Auto[,i])  
}  
Auto = Auto %>% na.omit()

## 資料介紹

* mpg:miles per gallon
* cylinders:Number of cylinders between 4 and 8
* displacement:Engine displacement (cu. inches)
* horsepower:Engine horsepower
* weight:Vehicle weight (lbs.)
* acceleration:Time to accelerate from 0 to 60 mph (sec.)
* year:Model year (modulo 100)
* origin:Origin of car (1. American, 2. European, 3. Japanese)
* name:Vehicle name

str(Auto)

## 'data.frame': 392 obs. of 9 variables:  
## $ mpg : num 18 15 18 16 17 15 14 14 14 15 ...  
## $ cylinders : num 8 8 8 8 8 8 8 8 8 8 ...  
## $ displacement: num 307 350 318 304 302 429 454 440 455 390 ...  
## $ horsepower : num 130 165 150 150 140 198 220 215 225 190 ...  
## $ weight : num 3504 3693 3436 3433 3449 ...  
## $ acceleration: num 12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...  
## $ year : num 70 70 70 70 70 70 70 70 70 70 ...  
## $ origin : Factor w/ 3 levels "1","2","3": 1 1 1 1 1 1 1 1 1 1 ...  
## $ name : Factor w/ 304 levels "amc ambassador brougham",..: 49 36 231 14 161 141 54 223 241 2 ...  
## - attr(\*, "na.action")= 'omit' Named int [1:5] 33 127 331 337 355  
## ..- attr(\*, "names")= chr [1:5] "33" "127" "331" "337" ...

## (8)

### (a)

* 1. 可以從個別t檢定看出，截距項及mpg皆以極趨近0的p-value(\*\*\*) 拒絕虛無假設，代表此變數(horsepower)對mpg之間有關係。
  2. 關係的強度我們可以從 Multicple R-squared : 0.6059 , Adjusted R-squared : 0.6049 這兩個值看出此線性回歸模型對mpg解釋的程度,此模型高達0.6代表解釋mpg程度尚佳。
  3. 我們可以從horsepower項的 Estimate的值為-0.157845看出，mpg與horsepower兩者為負相關，符合我們的想像，馬力大的車通常較耗油。

rg = lm(mpg~horsepower,data = Auto)  
summary(rg)

##   
## Call:  
## lm(formula = mpg ~ horsepower, data = Auto)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -13.5710 -3.2592 -0.3435 2.7630 16.9240   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 39.935861 0.717499 55.66 <2e-16 \*\*\*  
## horsepower -0.157845 0.006446 -24.49 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.906 on 390 degrees of freedom  
## Multiple R-squared: 0.6059, Adjusted R-squared: 0.6049   
## F-statistic: 599.7 on 1 and 390 DF, p-value: < 2.2e-16

* 1. 使用predict()預設之信賴區間即為95%的信賴區間，其預測值為24.46708，而預測區間與信賴區間相比多了一個標準差，因此Intervals的區間更寬 因為預測區間為估計一個“個別值”，而信賴區間為估計一個“平均值”，因此有此結果。

predict(rg,data.frame(horsepower = 98),interval = "confidence")

## fit lwr upr  
## 1 24.46708 23.97308 24.96108

predict(rg,data.frame(horsepower = 98),interval = "prediction")

## fit lwr upr  
## 1 24.46708 14.8094 34.12476

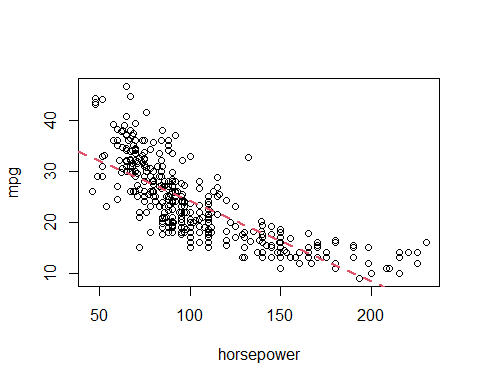
### (b)

* 兩者呈現負相關，與我們的直覺相同，馬力大的車油耗較差。

attach(Auto)

## The following object is masked from package:ggplot2:  
##   
## mpg

plot(horsepower,mpg)  
abline(rg,col = 2, lwd = 2,lty = 2)

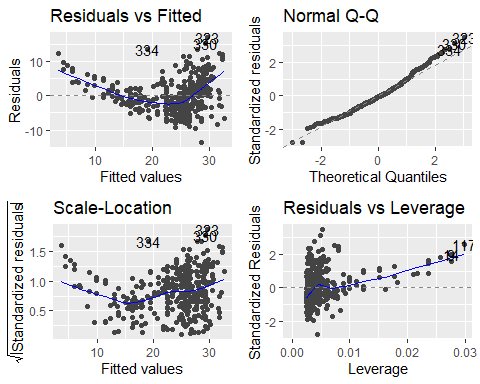


### (c)

* 由下圖可看見以下幾個結果：
  1. 殘差不隨機，有趨勢，代表解釋變數並未能對mpg有效解釋。
  2. Normal Q-Q圖可看出，殘差偏離斜直線代表殘差為不對襯分布，與我們通常對殘差的常態假設不符。
  3. 而從Leverage圖可看出，哪些觀測值偏離回歸線甚遠，可能造成回歸線預測偏離。

autoplot(rg)

## Warning: `arrange\_()` is deprecated as of dplyr 0.7.0.  
## Please use `arrange()` instead.  
## See vignette('programming') for more help  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_warnings()` to see where this warning was generated.

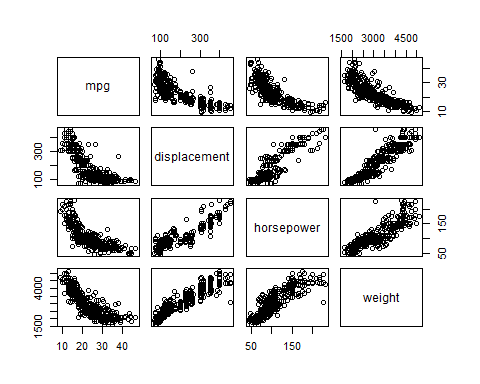


## (9)

### (a)

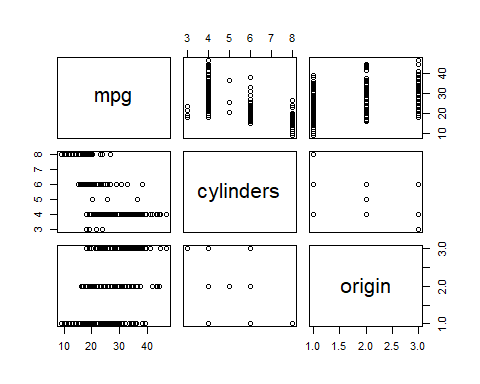
* 從散佈圖可看出，displacement、horsepower、weight對mpg呈現負相關，year對mpg為正相關，其代表：
* 引擎排氣量(displacement)越高，油耗越差。
* 馬力(horsepower)越高，油耗越差。
* 汽車重量(weight)越重，油耗越差。

x = Auto %>% select(mpg,displacement,horsepower,weight)  
pairs(x)



* 可以看出來當汽缸(cylinders)變多，mpg顯著下降，與我們的想法相符，汽缸數較多的車代表排氣量較高，因此油耗較高，而地區(origin)並未有太顯著的差別。

x = Auto %>% select(mpg,cylinders,origin)  
pairs(x)



### (b)

* 從此相關係數的表可見與上述散佈圖的結果相同。

cor(Auto[1:7])

## mpg cylinders displacement horsepower weight  
## mpg 1.0000000 -0.7776175 -0.8051269 -0.7784268 -0.8322442  
## cylinders -0.7776175 1.0000000 0.9508233 0.8429834 0.8975273  
## displacement -0.8051269 0.9508233 1.0000000 0.8972570 0.9329944  
## horsepower -0.7784268 0.8429834 0.8972570 1.0000000 0.8645377  
## weight -0.8322442 0.8975273 0.9329944 0.8645377 1.0000000  
## acceleration 0.4233285 -0.5046834 -0.5438005 -0.6891955 -0.4168392  
## year 0.5805410 -0.3456474 -0.3698552 -0.4163615 -0.3091199  
## acceleration year  
## mpg 0.4233285 0.5805410  
## cylinders -0.5046834 -0.3456474  
## displacement -0.5438005 -0.3698552  
## horsepower -0.6891955 -0.4163615  
## weight -0.4168392 -0.3091199  
## acceleration 1.0000000 0.2903161  
## year 0.2903161 1.0000000

### (c)

* 係數中displacement、weight、year、origin為顯著通過個別t檢定，而此處可見origin為顯著，在上述的分析中並未看到此變數對mpg有顯著的解釋能力，但在此卻顯著，也代表origin可能提供別的邊際貢獻。

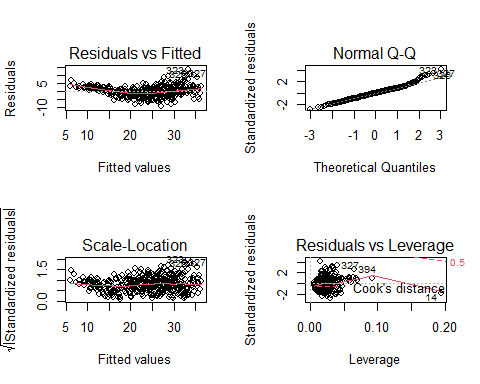
mrg = lm(mpg~.-name,data = Auto)  
summary(mrg)

##   
## Call:  
## lm(formula = mpg ~ . - name, data = Auto)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.0095 -2.0785 -0.0982 1.9856 13.3608   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.795e+01 4.677e+00 -3.839 0.000145 \*\*\*  
## cylinders -4.897e-01 3.212e-01 -1.524 0.128215   
## displacement 2.398e-02 7.653e-03 3.133 0.001863 \*\*   
## horsepower -1.818e-02 1.371e-02 -1.326 0.185488   
## weight -6.710e-03 6.551e-04 -10.243 < 2e-16 \*\*\*  
## acceleration 7.910e-02 9.822e-02 0.805 0.421101   
## year 7.770e-01 5.178e-02 15.005 < 2e-16 \*\*\*  
## origin2 2.630e+00 5.664e-01 4.643 4.72e-06 \*\*\*  
## origin3 2.853e+00 5.527e-01 5.162 3.93e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.307 on 383 degrees of freedom  
## Multiple R-squared: 0.8242, Adjusted R-squared: 0.8205   
## F-statistic: 224.5 on 8 and 383 DF, p-value: < 2.2e-16

### (d)

* 由下圖可看見以下幾個結果：
  1. 殘差不隨機，有趨勢，代表解釋變數並未能對mpg有效解釋。
  2. Normal Q-Q圖可看出，殘差偏離斜直線代表殘差為不對襯分布，與我們通常對殘差的常態假設不符。
  3. 而從Leverage圖可看出，哪些觀測值偏離回歸線甚遠，可能造成回歸線預測偏離，標記出第327,394,14筆資料可能為不正常的離群值，而明顯可見第14筆觀測值存在有高度的Leverage Effects。

par(mfrow=c(2,2))  
plot(mrg)



* 由下面結果可見，此車種的汽缸數(cylinders)、引擎排量(displacement)、馬力(horsepower)明顯高於平均，但重量(weight)卻與平均差不多，而油耗(mpg)卻明顯差很多 ，可能是因為weight的部分其他觀測值有明顯的差異。

Auto[14,"name"]

## [1] buick estate wagon (sw)  
## 304 Levels: amc ambassador brougham amc ambassador dpl ... vw rabbit custom

paste("Average mpg:",mean(Auto$mpg),"Buick Estate Wagon:",Auto[14,"mpg"])

## [1] "Average mpg: 23.4459183673469 Buick Estate Wagon: 14"

paste("Average cylinders:",mean(Auto$cylinders),"Buick Estate Wagon:",Auto[14,"cylinders"])

## [1] "Average cylinders: 5.4719387755102 Buick Estate Wagon: 8"

paste("Average displacement:",mean(Auto$displacement),"Buick Estate Wagon:",Auto[14,"displacement"])

## [1] "Average displacement: 194.411989795918 Buick Estate Wagon: 455"

paste("Average horsepower:",mean(Auto$horsepower),"Buick Estate Wagon:",Auto[14,"horsepower"])

## [1] "Average horsepower: 104.469387755102 Buick Estate Wagon: 225"

paste("Average weight:",mean(Auto$weight),"Buick Estate Wagon:",Auto[14,"weight"])

## [1] "Average weight: 2977.58418367347 Buick Estate Wagon: 3086"

paste("Average year:",mean(Auto$year),"Buick Estate Wagon:",Auto[14,"year"])

## [1] "Average year: 75.9795918367347 Buick Estate Wagon: 70"

### (e)

* 從上述散佈圖可發現weight、cylinders及weight、displacement之間有高度相關，可能存在有共線性的問題，這時候加入交互項來解決此問題，從表1、表2皆可看到交互項通過個別t檢定，拒絕虛無假設，對mpg有顯著的解釋能力。

mrg1 = lm(mpg~weight\*cylinders,data = Auto)  
summary(mrg1)

##   
## Call:  
## lm(formula = mpg ~ weight \* cylinders, data = Auto)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -14.4916 -2.6225 -0.3927 1.7794 16.7087   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 65.3864559 3.7333137 17.514 < 2e-16 \*\*\*  
## weight -0.0128348 0.0013628 -9.418 < 2e-16 \*\*\*  
## cylinders -4.2097950 0.7238315 -5.816 1.26e-08 \*\*\*  
## weight:cylinders 0.0010979 0.0002101 5.226 2.83e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.165 on 388 degrees of freedom  
## Multiple R-squared: 0.7174, Adjusted R-squared: 0.7152   
## F-statistic: 328.3 on 3 and 388 DF, p-value: < 2.2e-16

mrg2 = lm(mpg~weight\*displacement,data = Auto)  
summary(mrg2)

##   
## Call:  
## lm(formula = mpg ~ weight \* displacement, data = Auto)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -13.8664 -2.4801 -0.3355 1.8071 17.9429   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.372e+01 1.940e+00 27.697 < 2e-16 \*\*\*  
## weight -8.931e-03 8.474e-04 -10.539 < 2e-16 \*\*\*  
## displacement -7.831e-02 1.131e-02 -6.922 1.85e-11 \*\*\*  
## weight:displacement 1.744e-05 2.789e-06 6.253 1.06e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.097 on 388 degrees of freedom  
## Multiple R-squared: 0.7265, Adjusted R-squared: 0.7244   
## F-statistic: 343.6 on 3 and 388 DF, p-value: < 2.2e-16

* ’:’代表單獨放交互項，此處放displacement與cylinders的交互項，結果如下表，顯著拒絕虛無假設，對mpg有解釋能力。

mrg3 = lm(mpg~displacement:cylinders,data = Auto)  
summary(mrg3)

##   
## Call:  
## lm(formula = mpg ~ displacement:cylinders, data = Auto)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11.705 -3.426 -0.450 2.704 17.715   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 30.9896203 0.3905111 79.36 <2e-16 \*\*\*  
## displacement:cylinders -0.0061177 0.0002462 -24.85 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.863 on 390 degrees of freedom  
## Multiple R-squared: 0.6128, Adjusted R-squared: 0.6119   
## F-statistic: 617.4 on 1 and 390 DF, p-value: < 2.2e-16

### (f)

* 上述說到殘差具有趨勢，可由加入平方項、根號項、log項等方式解決此問題， 表1為mpg對所有變數並加入平方項後的結果，再剔除掉不顯著的變數後得到表2。
* 由診斷圖可見，殘差的趨勢、Leverage的趨勢皆被消除，代表加入此平方項有顯著的效果。

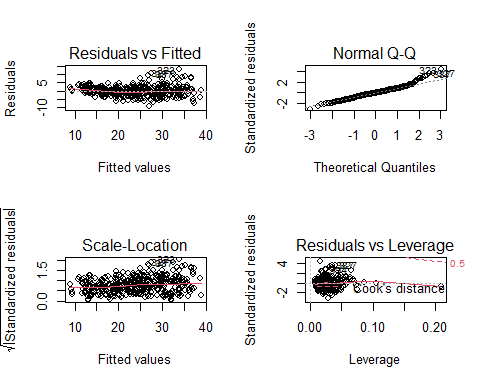
mrg4 = lm(mpg~.- name + I(weight^2),data = Auto)  
  
mrg5 = lm(mpg~.- name   
 - acceleration   
 - cylinders   
 + I(weight^2),data = Auto)  
  
summary(mrg4)

##   
## Call:  
## lm(formula = mpg ~ . - name + I(weight^2), data = Auto)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8.9387 -1.6686 -0.1062 1.7273 12.8215   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.291e-01 4.613e+00 0.028 0.97768   
## cylinders -2.819e-01 2.898e-01 -0.973 0.33118   
## displacement 1.750e-02 6.917e-03 2.529 0.01183 \*   
## horsepower -2.543e-02 1.235e-02 -2.059 0.04019 \*   
## weight -2.062e-02 1.570e-03 -13.134 < 2e-16 \*\*\*  
## acceleration 6.445e-02 8.836e-02 0.729 0.46623   
## year 8.236e-01 4.683e-02 17.586 < 2e-16 \*\*\*  
## origin2 1.850e+00 5.160e-01 3.585 0.00038 \*\*\*  
## origin3 1.493e+00 5.172e-01 2.886 0.00412 \*\*   
## I(weight^2) 2.224e-06 2.326e-07 9.559 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.974 on 382 degrees of freedom  
## Multiple R-squared: 0.8581, Adjusted R-squared: 0.8548   
## F-statistic: 256.7 on 9 and 382 DF, p-value: < 2.2e-16

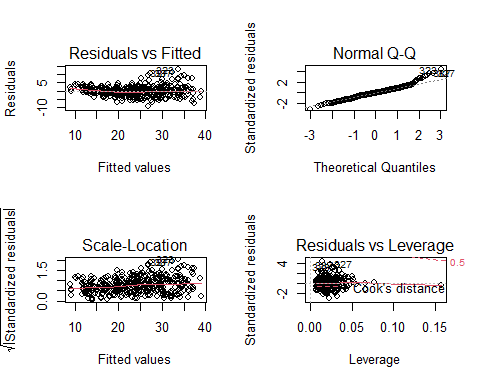
summary(mrg5)

##   
## Call:  
## lm(formula = mpg ~ . - name - acceleration - cylinders + I(weight^2),   
## data = Auto)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8.9157 -1.6289 -0.0723 1.6161 12.8276   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.401e-01 4.192e+00 0.224 0.822685   
## displacement 1.242e-02 5.181e-03 2.397 0.016987 \*   
## horsepower -2.991e-02 9.653e-03 -3.099 0.002084 \*\*   
## weight -2.061e-02 1.539e-03 -13.392 < 2e-16 \*\*\*  
## year 8.221e-01 4.667e-02 17.614 < 2e-16 \*\*\*  
## origin2 1.824e+00 5.149e-01 3.542 0.000445 \*\*\*  
## origin3 1.434e+00 5.134e-01 2.793 0.005478 \*\*   
## I(weight^2) 2.244e-06 2.318e-07 9.683 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.972 on 384 degrees of freedom  
## Multiple R-squared: 0.8576, Adjusted R-squared: 0.855   
## F-statistic: 330.3 on 7 and 384 DF, p-value: < 2.2e-16

par(mfrow=c(2,2))  
plot(mrg4)



plot(mrg5)



## 10

### (a)

Carseats為關於兒童車用座椅的資料，400筆觀測值代表不同店家，共有11個變數，變數敘述如下：

* Sales:Unit sales (in thousands) at each location
* CompPrice:Price charged by competitor at each location
* Income:Community income level (in thousands of dollars)
* Advertising:Local advertising budget for company at each location (in thousands of dollars)
* Population:Population size in region (in thousands)
* Price:Price company charges for car seats at each site
* ShelveLoc:A factor with levels Bad, Good and Medium indicating the quality of the shelving location for the car seats at each site
* Age:Average age of the local population
* Education:Education level at each location
* Urban:A factor with levels No and Yes to indicate whether the store is in an urban or rural location
* US:A factor with levels No and Yes to indicate whether the store is in the US or not
* Urban 為binary的變數，代表店家是否在都會區，未通過個別t檢定。
* 整體模型的R squared僅0.2393，代表此模型表現欠佳，尚有許多變異未解釋，有改進的空間。

library(ISLR)  
data("Carseats")  
  
fit = lm(Sales~Price+Urban+US,data = Carseats )  
summary(fit)

##   
## 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)

* 要比較各變數之間的貢獻，則需要先將數值型變數進行標準化，才可擺脫單位造成的影響。
* 標準化後，下表即為回歸模型的式子以及每個變數的所估計的參數，可注意到：
* Price 參數為負，代表Price與Sales之間為負相關，代表當產品的定價越高對於銷售會產生負面的效果。
* Urban 變數為Binary變數，由於上述提到並未通過個別t檢定，其P-value顯著的不拒絕虛無假設，故在此討論其參數可能有誤。
* US 變數為Binary變數，代表商店是否位於美國，可見當為Yes時其參數為正，並且其值遠大於Price，可能代表在美國的店家的銷售明顯高過其他地區所造成。

index = sapply(1:11,function(x){  
 is.numeric(Carseats[,x])  
 }  
)  
Carseats[,index] %<>% scale()  
  
  
print(fit)

##   
## Call:  
## lm(formula = Sales ~ Price + Urban + US, data = Carseats)  
##   
## Coefficients:  
## (Intercept) Price UrbanYes USYes   
## 13.04347 -0.05446 -0.02192 1.20057

### (c)

模型中有兩個Binary變數，依這兩個變數的結果共有以下四種情況，如下：

* 當Urban、US皆為Yes
* Urban 為Yes，US為NO
* Urban為NO，US為YES
* Urban、US皆為NO

### (d)

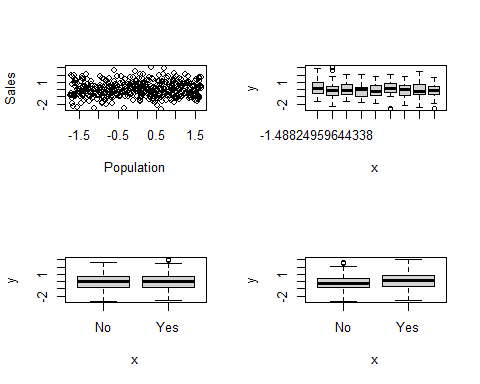
* 回歸放入全部的變數，發現Population、Education、Urban、US皆未通過個別t檢定，其餘變數皆通過個別t檢定，拒絕虛無假設。

fit1 = lm(Sales~.,data = Carseats)  
summary(fit1)

##   
## Call:  
## lm(formula = Sales ~ ., data = Carseats)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.01598 -0.24463 0.00748 0.23496 1.20797   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.73292 0.05999 -12.217 < 2e-16 \*\*\*  
## CompPrice 0.50397 0.02252 22.378 < 2e-16 \*\*\*  
## Income 0.15660 0.01828 8.565 2.58e-16 \*\*\*  
## Advertising 0.28987 0.02619 11.066 < 2e-16 \*\*\*  
## Population 0.01085 0.01933 0.561 0.575   
## Price -0.79946 0.02239 -35.700 < 2e-16 \*\*\*  
## ShelveLocGood 1.71742 0.05422 31.678 < 2e-16 \*\*\*  
## ShelveLocMedium 0.69286 0.04465 15.516 < 2e-16 \*\*\*  
## Age -0.26413 0.01825 -14.472 < 2e-16 \*\*\*  
## Education -0.01958 0.01830 -1.070 0.285   
## UrbanYes 0.04351 0.04000 1.088 0.277   
## USYes -0.06519 0.05306 -1.229 0.220   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3608 on 388 degrees of freedom  
## Multiple R-squared: 0.8734, Adjusted R-squared: 0.8698   
## F-statistic: 243.4 on 11 and 388 DF, p-value: < 2.2e-16

* 接著看Sales對這四個變數的plot，可以發現Sales在這四個變數的Outcome間皆無明顯差異，可解讀其個別對Sales並無解釋能力，故未通過個別t檢定。

attach(Carseats)  
  
par(mfrow=c(2,2))  
plot(Population,Sales)  
plot(as.factor(Education),Sales)  
plot(Urban,Sales)  
plot(US,Sales)



## (e)

* 發現去掉上述4個個別t檢定未通過的變數後，R Square並未有明顯的下降，而自由度卻有大幅的上升。
* 此舉動代表降低了估計參數的同時並未犧牲掉解釋力，故此篩選變數是一個好的選擇。

fit1 = lm(Sales~.-Population-Education-Urban-US,data = Carseats)  
summary(fit1)

##   
## Call:  
## lm(formula = Sales ~ . - Population - Education - Urban - US,   
## data = Carseats)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.98184 -0.24624 0.00997 0.23839 1.17885   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.74228 0.03699 -20.07 <2e-16 \*\*\*  
## CompPrice 0.50265 0.02239 22.45 <2e-16 \*\*\*  
## Income 0.15642 0.01821 8.59 <2e-16 \*\*\*  
## Advertising 0.27294 0.01819 15.01 <2e-16 \*\*\*  
## Price -0.79913 0.02239 -35.70 <2e-16 \*\*\*  
## ShelveLocGood 1.71228 0.05400 31.71 <2e-16 \*\*\*  
## ShelveLocMedium 0.69119 0.04439 15.57 <2e-16 \*\*\*  
## Age -0.26461 0.01822 -14.52 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.361 on 392 degrees of freedom  
## Multiple R-squared: 0.872, Adjusted R-squared: 0.8697   
## F-statistic: 381.4 on 7 and 392 DF, p-value: < 2.2e-16

## (f)

* 我們可由兩模型的R Square 及 fit.values對實際Sales的plot來看，明顯可以發現(e)小題模型的解釋力比較好，從圖也可以發現預測的誤差較小。

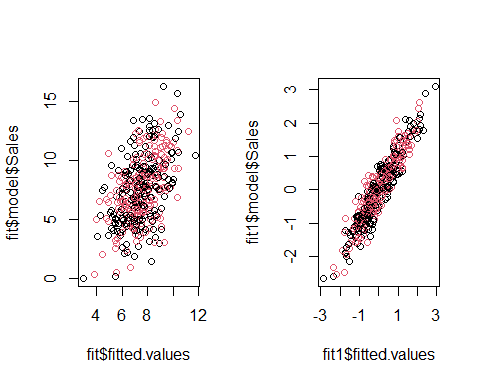
summary(fit)

##   
## 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

summary(fit1)

##   
## Call:  
## lm(formula = Sales ~ . - Population - Education - Urban - US,   
## data = Carseats)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.98184 -0.24624 0.00997 0.23839 1.17885   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.74228 0.03699 -20.07 <2e-16 \*\*\*  
## CompPrice 0.50265 0.02239 22.45 <2e-16 \*\*\*  
## Income 0.15642 0.01821 8.59 <2e-16 \*\*\*  
## Advertising 0.27294 0.01819 15.01 <2e-16 \*\*\*  
## Price -0.79913 0.02239 -35.70 <2e-16 \*\*\*  
## ShelveLocGood 1.71228 0.05400 31.71 <2e-16 \*\*\*  
## ShelveLocMedium 0.69119 0.04439 15.57 <2e-16 \*\*\*  
## Age -0.26461 0.01822 -14.52 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.361 on 392 degrees of freedom  
## Multiple R-squared: 0.872, Adjusted R-squared: 0.8697   
## F-statistic: 381.4 on 7 and 392 DF, p-value: < 2.2e-16

par(mfrow=c(1,2))  
plot(fit$fitted.values,fit$model$Sales,col=1:2)  
plot(fit1$fitted.values,fit1$model$Sales,col=1:2)



## (g)

* 可以看到(e)小題的模型的信賴區間都未包括0，皆顯著

confint(fit1)

## 2.5 % 97.5 %  
## (Intercept) -0.8150011 -0.6695679  
## CompPrice 0.4586319 0.5466658  
## Income 0.1206214 0.1922261  
## Advertising 0.2371770 0.3086930  
## Price -0.8431403 -0.7551195  
## ShelveLocGood 1.6061162 1.8184431  
## ShelveLocMedium 0.6039064 0.7784684  
## Age -0.3004333 -0.2287799

## (h)

* (e)小題的模型所畫出的Residuals vs Leverage Plot中,存在有幾個殘差較大的觀測值，但其Cook Distance並未超過0.5，皆包含在裡面，因此認為此模型中並未有存在High Leverage的觀測值。
* 而第208,298筆資料其Residuals接近+-3，而由Scale-Location Plot 圖中也可發現， 標準化後的Residuals也超過1.5，因此可能為Outliers，但仍需要再加以分析才決定是否要加以刪除。

par(mfrow=c(2,2))  
plot(fit1)

