Statistical Learning I
Quang Tran
Assignment 1: Chapter 2 (2,7,8), Chapter 3 (3,4,5,8,9,10,13)

Chapter 2:

2)

a) regression, inference: quantitative output of CEO salary based on CEO firm's features
n: the top 500 firms in the US
p: profit, number of employees, industry

b) classification, prediction: predicting a new product whether is a success or failure
n: 20 similar products previously launched
p: price charge, marketing budget, competition price, 10 other variables

c) regression, prediction: predicting % change in US dollars
n: all weekly data of 2012 (52 weeks)
p: % change in US market, % change in German market, % change in British market

7)

a)

Obs	X_1	X_2	X_3	Y	Distance to (0,0,0)
1	0	3	0	Red	3
2	2	0	0	Red	2
3	0	1	3	Red	~3.2
4	0	1	2	Green	~2.2
5	-1	0	1	Green	~1.4
6	1	1	1	Red	~1.7

b) Prediction: Green, the obs #5 is the closet neighbor for K=1

c) Prediction: Red, the obs #2, #5, #6 are the closet neighbor for K=3 (Red, Green, Red)

d) The best value for K would be small, if the Bayes decision boundary for this problem highly non-linear. Since a small K would be flexible for non-linear boundary, whereas a large K would try to fit a more linear boundary – it takes more points into consideration.

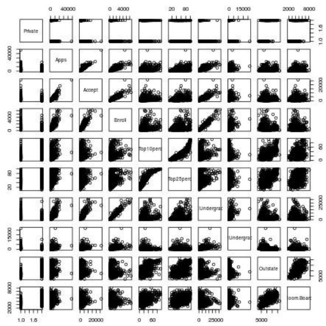
8)

a) >college = read.csv(file="College.csv")

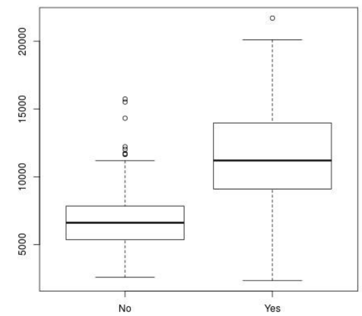
b)> rownames(college) = college[,1]
> fix(college)
> college = college[,-1]
> fix(college)

c)
i)> summary(college)
```
 Private      Apps         Accept        Enroll      Top10perc
No :212   Min.   :   81   Min.   :   72   Min.   :  35   Min.   : 1.00
Yes:565   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242   1st Qu.:15.00
...
```

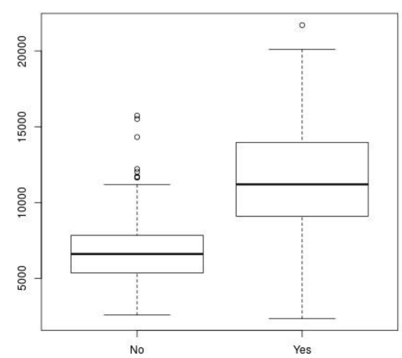
```
ii)> jpeg("pairs.jpg")
> pairs(college[,1:10])
> dev.off()
```



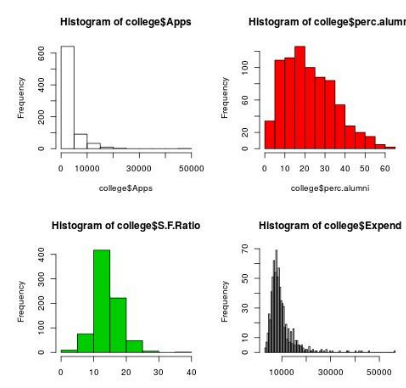
```
iii)> jpeg("boxplots.jpg")
> plot(college$Private, college$Outstate)
> dev.off()
```



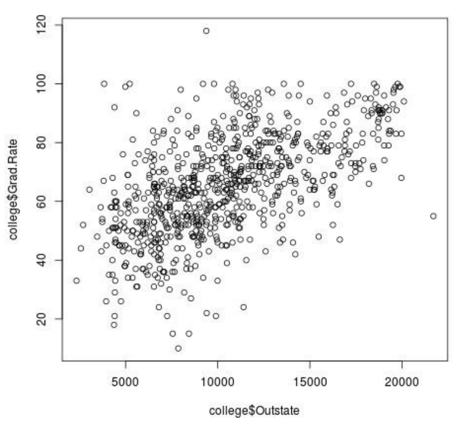

```
iv)> Elite = rep("No", nrow(college))
> Elite[college$Top10perc>50] = "Yes"
> Elite = as.factor(Elite)
> college = data.frame(college, Elite)
> summary(college$Elite)
No Yes
699  78

> jpeg("new-boxplots.jpg")
> plot(college$Private, college$Outstate)
> dev.off()
```




```
v)> jpeg("hist.jpg")
> par(mfrow=c(2,2))
> hist(college$Apps)
> hist(college$perc.alumni, col=2)
> hist(college$S.F.Ratio, col=3, breaks=10)
> hist(college$Expend, breaks=100)
> dev.off()
```




```
vi)> plot(college$Top10perc, college$Grad.Rate)
> jpeg("discovery.jpg")
> plot(college$Outstate, college$Grad.Rate)
> dev.off()
> # High tuition correlates to high graduation rate.
```



Chapter 3:

3) $Y = 50 + 20(gpa) + 0.07(iq) + 35(gender) + 0.01(gpa * iq) - 10 (gpa * gender)$

a) male (gender = 0): $50 + 20 X_1 + 0.07 X_2 + 0.01(X_1 * X_2)$
female (gender = 1): $50 + 20 X_1 + 0.07 X_2 + 35 + 0.01(X_1 * X_2) - 10 (X_1)$
Once the GPA is high enough, males earn more on average. => iii is correct.

b) Y(Gender = 1, IQ = 110, GPA = 4.0)
= 50 + 20 * 4 + 0.07 * 110 + 35 + 0.01 (4 * 110) - 10 * 4
= 137.1

c) False. We must examine the p-value of the regression coefficient to determine if the interaction term is statistically significant or not.

4) a) I would expect the polynomial regression to have a lower training RSS than the linear regression because it could make a tighter fit against data that matched with a wider irreducible error (Var(epsilon)).

b) Converse to (a), I would expect the polynomial regression to have a higher test RSS as the overfit from training would have more error than the linear regression.

c) Polynomial regression has lower train RSS than the linear fit because of higher flexibility: no matter what the underlying true relationshop is the more flexible model will closer follow points and reduce train RSS. An example is shown on Figure2.9 from Chapter 2.

d) There is not enough information to tell which test RSS would be lower for either regression given the problem statement is defined as not knowing "how far it is from linear". If it is closer to linear than cubic, the linear regression test RSS could be lower than the cubic regression test RSS. Or, if it is closer to cubic than linear, the cubic regression test RSS could be lower than the linear regression test RSS. It is dues to bias-variance tradeoff: it is not clear what level of flexibility will fit data better.

5) y_i^ = x_i*B^
B^ = sigma(x_i*y_i)/sigma(x_i^2)
y_i^ = x_i*sigma(x_i*y_i)/sigma($x_{i'}^2$) = y_i*sigma(x_i*$x_{i'}$ / $x_{i'}^2$)
$a_{i'}$ = ($x_{i'}$*x_i) / sigma($x_{i'}^2$)

8) a)

```
> Auto = read.csv("Auto.csv", header=T, na.strings="?")
> Auto = na.omit(Auto)
> summary(Auto)
      mpg          cylinders      displacement     horsepower        weight
 Min.   : 9.00   Min.   :3.000   Min.   : 68.0   Min.   : 46.0   Min.   :1613
 1st Qu.:17.00   1st Qu.:4.000   1st Qu.:105.0   1st Qu.: 75.0   1st Qu.:2225
 Median :22.75   Median :4.000   Median :151.0   Median : 93.5   Median :2804
 Mean   :23.45   Mean   :5.472   Mean   :194.4   Mean   :104.5   Mean   :2978
3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:275.8   3rd Qu.:126.0   3rd Qu.:3615
Max.   :46.60   Max.   :8.000   Max.   :455.0   Max.   :230.0   Max.   :5140
….
```

b) > attach(Auto)
> lm.fit = lm(mpg ~ horsepower)
> summary(lm.fit)

```
Call:
lm(formula = mpg ~ horsepower)

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
```

i) Yes, there is a relationship between horsepower and mpg as deterined by testing the null hypothesis of all regression coefficients equal to zero. Since the F-statistic is far larger than 1 and the p-value of the F-statistic is close to zero we can reject the null hypothesis and state there is a statistically significant relationship between horsepower and mpg.

ii) To calculate the residual error relative to the response we use the mean of the response and the RSE. The mean of mpg is 23.4459. The RSE of the lm.fit was 4.906 which indicates a percentage error of 20.9248%. The R^2 of the lm.fit was about 0.6059, meaning 60.5948% of the variance in mpg is explained by horsepower.

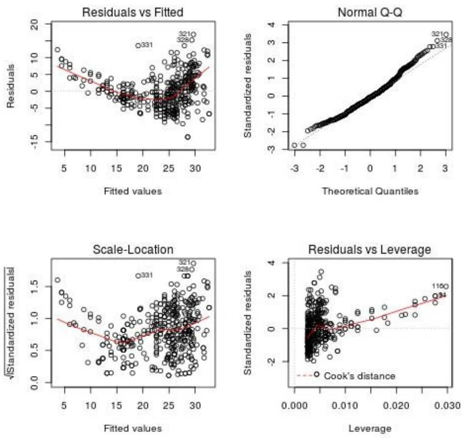
iii) The relationship between mpg and horsepower is negative. The more horsepower an automobile has the linear regression indicates the less mpg fuel efficiency the automobile will have.

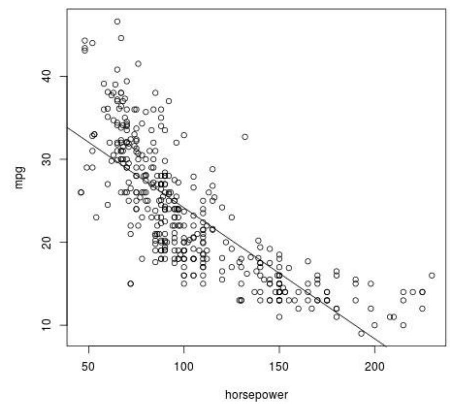
iv)
```
> predict(lm.fit, data.frame(horsepower=c(98)), interval="confidence")
       fit        lwr        upr
1 24.46708   23.97308   24.96108
> predict(lm.fit, data.frame(horsepower=c(98)), interval="prediction")
      fit       lwr        upr
1 24.46708   14.8094   34.12476
```

c)
```
> jpeg("lm-hp-mpg.jpg")
> plot(horsepower, mpg)
> abline(lm.fit)
> dev.off()
```

d)
```
> jpeg(" diagnostic-plots.jpg")
> par(mfrow=c(2,2))
> plot(lm.fit)
> dev.off()
```

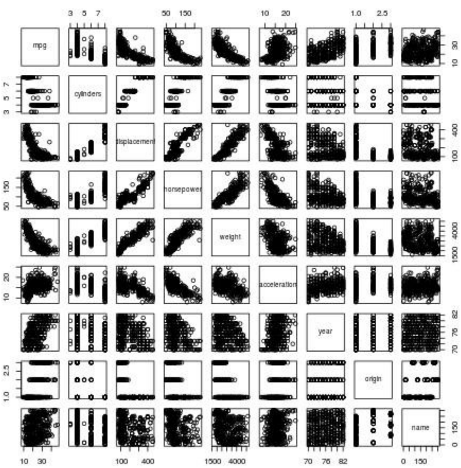
Based on the residuals plots, there is some evidence of non-linearity.





9) a) > pairs(Auto)

```
> jpeg(" pairs-auto.jpg")
> pairs(Auto)
> dev.off()
```



b) > cor(subset(Auto, select=-name))

```
                    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
origin        0.5652088 -0.5689316   -0.6145351 -0.4551715 -0.5850054
             acceleration       year     origin
mpg             0.4233285  0.5805410  0.5652088
cylinders      -0.5046834 -0.3456474 -0.5689316
displacement   -0.5438005 -0.3698552 -0.6145351
horsepower     -0.6891955 -0.4163615 -0.4551715
weight         -0.4168392 -0.3091199 -0.5850054
acceleration    1.0000000  0.2903161  0.2127458
year            0.2903161  1.0000000  0.1815277
origin          0.2127458  0.1815277  1.0000000
```

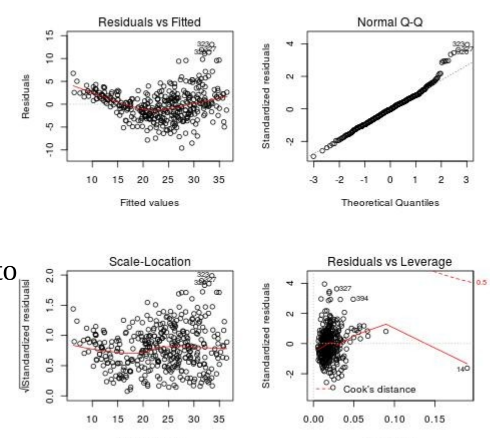
```
c) > lm.fit1 = lm(mpg~.-name, data=Auto)
> summary(lm.fit1)

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) Yes, there is a relatioship between the predictors and the response by testing the null hypothesis of whether all the regression coefficients are zero. The F -statistic is far from 1 (with a small p-value), indicating evidence against the null hypothesis.

ii) Looking at the p-values associated with each predictor's t-statistic, we see that displacement, weight, year, and origin have a statistically significant relationship, while cylinders, horsepower, and acceleration do not.

iii) The regression coefficient for year, 0.7508, suggests that for every one year, mpg increases by the coefficient. In other words, cars become more fuel efficient every year by almost 1 mpg / year.

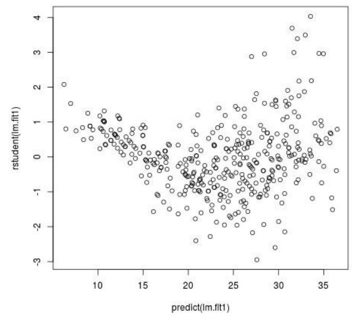
```
d) > jpeg("diagnostic-plots-auto.jpg")
> par(mfrow=c(2,2))
> plot(lm.fit1)
> dev.off()
```



The fit does not appear to be accurate because there is a discernible curve pattern to the residuals plots. From the leverage plot, point 14 appears to have high leverage, although not a high magnitude residual.

```
e) > jpeg("rstudent-auto.jpg")
> plot(predict(lm.fit1), rstudent(lm.fit1))
> dev.off()
```

There are possible outliers as seen in the plot of studentized residuals because there are data with a value greater than 3.



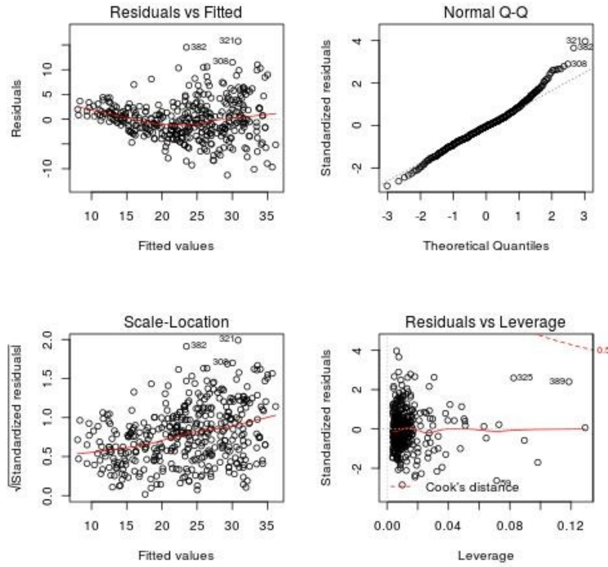

```
e) > lm.fit2 = lm(mpg~cylinders*displacement+displacement*weight)
> summary(lm.fit2)

Call:
lm(formula = mpg ~ cylinders * displacement + displacement * weight)

Residuals:
     Min       1Q   Median       3Q      Max 
-13.2934  -2.5184  -0.3476   1.8399  17.7723 

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)             5.262e+01  2.237e+00  23.519  < 2e-16 ***
cylinders               7.606e-01  7.669e-01   0.992    0.322    
displacement           -7.351e-02  1.669e-02  -4.403 1.38e-05 ***
weight                 -9.888e-03  1.329e-03  -7.438 6.69e-13 ***
cylinders:displacement -2.986e-03  3.426e-03  -0.872    0.384    
displacement:weight     2.128e-05  5.002e-06   4.254 2.64e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.103 on 386 degrees of freedom
Multiple R-squared:  0.7272,	Adjusted R-squared:  0.7237 
F-statistic: 205.8 on 5 and 386 DF,  p-value: < 2.2e-16
```

From the correlation matrix, I obtained the two highest correlated pairs and used them in picking my interaction effects. From the p-values, we can see that the interaction between displacement and weight is statistically signifcant, while the interactiion between cylinders and displacement is not.

```
f) > lm.fit3 = lm(mpg~log(weight)+sqrt(horsepower)+acceleration+I(acceleration^2))
> summary(lm.fit3)

Call:
lm(formula = mpg ~ log(weight) + sqrt(horsepower) + acceleration + 
    I(acceleration^2))

Residuals:
     Min       1Q   Median       3Q      Max 
-11.2932  -2.5082  -0.2237   2.0237  15.7650 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)       178.30303   10.80451  16.503  < 2e-16 ***
log(weight)       -14.74259    1.73994  -8.473 5.06e-16 ***
sqrt(horsepower)   -1.85192    0.36005  -5.144 4.29e-07 ***
acceleration       -2.19890    0.63903  -3.441 0.000643 ***
I(acceleration^2)   0.06139    0.01857   3.305 0.001037 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 3.99 on 387 degrees of freedom
Multiple R-squared:  0.7414,     Adjusted R-squared:  0.7387
F-statistic: 277.3 on 4 and 387 DF,  p-value: < 2.2e-16
```

```
f) > jpeg("c3-9f.jpg")
> par(mfrow=c(2,2))
> plot(lm.fit3)
> dev.off()
```

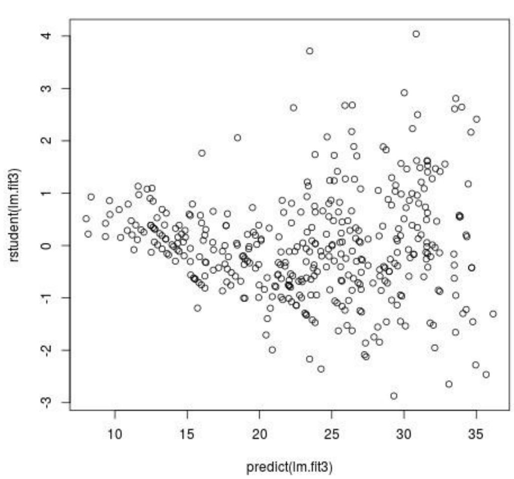



```
> jpeg("c3-9f1.jpg")
> plot(predict(lm.fit3), rstudent(lm.fit3))
> dev.off()
```

Apparently, from the p-values, the log(weight), sqrt(horsepower), and acceleration^2 all have statistical significance of some sort. The residuals plot has less of a discernible pattern than the plot of all linear regression terms. The studentized residuals displays potential outliers (>3). The leverage plot indicates more than three points with high leverage.

1) the residuals vs fitted plot indicates heteroskedasticity (unconstant variance over mean) in the model.
2) The Q-Q plot indicates somewhat unnormality of the residuals.

So, a better transformation need to be applied to our model. From the correlation matrix in 9a., displacement, horsepower and weight show a similar nonlinear pattern against our response mpg. This nonlinear pattern is very close to a log form. So in the next attempt, we use log(mpg) as our response variable. The outputs show that log transform of mpg yield better model fitting (better R^2, normality of residuals).




```
10) > library(ISLR)
> summary(Carseats)
     Sales          CompPrice       Income        Advertising
 Min.   : 0.000   Min.   : 77   Min.   : 21.00   Min.   : 0.000
 1st Qu.: 5.390   1st Qu.:115   1st Qu.: 42.75   1st Qu.: 0.000
 Median : 7.490   Median :125   Median : 69.00   Median : 5.000
 Mean   : 7.496   Mean   :125   Mean   : 68.66   Mean   : 6.635
 3rd Qu.: 9.320   3rd Qu.:135   3rd Qu.: 91.00   3rd Qu.:12.000
 Max.   :16.270   Max.   :175   Max.   :120.00   Max.   :29.000
   Population        Price        ShelveLoc        Age         Education
 Min.   : 10.0   Min.   : 24.0   Bad   : 96   Min.   :25.00   Min.   :10.0
 1st Qu.:139.0   1st Qu.:100.0   Good  : 85   1st Qu.:39.75   1st Qu.:12.0
 Median :272.0   Median :117.0   Medium:219   Median :54.50   Median :14.0
 Mean   :264.8   Mean   :115.8                Mean   :53.32   Mean   :13.9
 3rd Qu.:398.5   3rd Qu.:131.0                3rd Qu.:66.00   3rd Qu.:16.0
 Max.   :509.0   Max.   :191.0                Max.   :80.00   Max.   :18.0
 ….
```

a) > attach(Carseats)
> lm.fit = lm(Sales~Price+Urban+US)
> summary(lm.fit)

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

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: The linear regression suggests a relationship between price and sales given the low p-value of the t-statistic. The coefficient states a negative relationship between Price and Sales: as Price increases, Sales decreases.

UrbanYes: The linear regression suggests that there isn't a relationship between the location of the store and the number of sales based on the high p-value of the t-statistic.

USYes: The linear regression suggests there is a relationship between whether the store is in the US or not and the amount of sales. The coefficient states a positive relationship between USYes and Sales: if the store is in the US, the sales will increase by approximately 1201 units.

c) Sales = 13.04 + -0.05 Price + -0.02 UrbanYes + 1.20 USYes

d) Predictors reject null hypothesis: Price and USYes, based on the p-values, F-statistic, and p-value of the F-statistic.

e) > lm.fit2 = lm(Sales ~ Price + US)
> summary(lm.fit2)

```
Call:
lm(formula = Sales ~ Price + US)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.9269 -1.6286 -0.0574  1.5766  7.0515 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 13.03079    0.63098  20.652  < 2e-16 ***
```

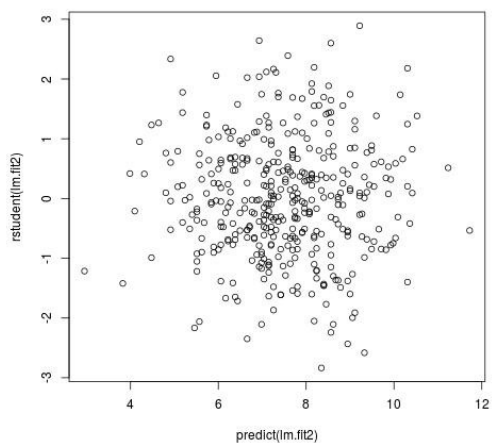
Price   -0.05448   0.00523 -10.416 < 2e-16 ***
USYes   1.19964   0.25846   4.641 4.71e-06 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.469 on 397 degrees of freedom
Multiple R-squared: 0.2393,	Adjusted R-squared: 0.2354
F-statistic: 62.43 on 2 and 397 DF, p-value: < 2.2e-16

f) Based on the RSE and R^2 of the linear regressions, they both fit the data similarly, with linear regression from (e) fitting the data slightly better.

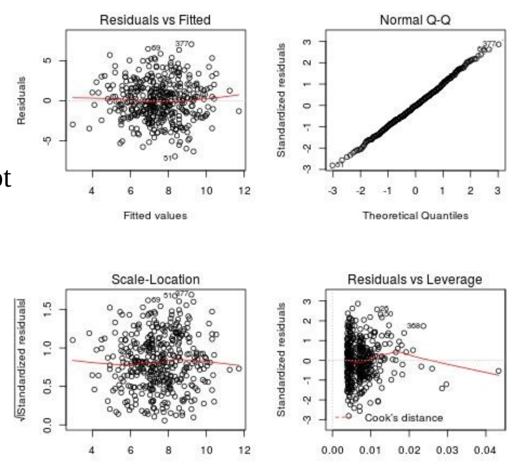
g) > confint(lm.fit2)
                  2.5 %   97.5 %
(Intercept) 11.79032020 14.27126531
Price   -0.06475984 -0.04419543
USYes   0.69151957 1.70776632



h) > jpeg("c310h.jpg")
> plot(predict(lm.fit2), rstudent(lm.fit2))
> dev.off()

All studentized residuals appear to be bounded by -3 to 3, so not potential outliers are suggested from the linear regression.

> jpeg("c310h1.jpg")
> par(mfrow=c(2,2))
> plot(lm.fit2)
> dev.off()



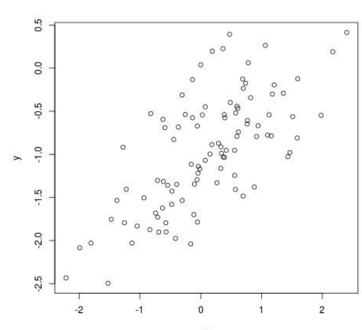
There are a few observations that greatly exceed (p+1)/n(p+1)/n (0.0076) on the leverage-statistic plot that suggest that the corresponding points have high leverage.

13)

a) > set.seed(1)
> x = rnorm(100)

b) > eps = rnorm(100, 0, sqrt(0.25))

c) > y = -1 + 0.5*x + eps
y has length 100, $b_0 = -1$, $b_1 = 0.5$

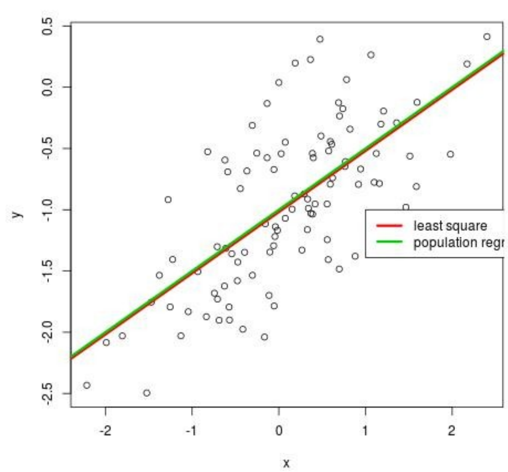


d) > jpeg("13d")
> plot(x, y)
> dev.off()
A linear relationship between x and y with a positive slope, with a variance as is to be expected.

```
e) > lm.fit = lm(y~x)
> summary(lm.fit)

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.93842 -0.30688 -0.06975  0.26970  1.17309 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.01885    0.04849 -21.010  < 2e-16 ***
x            0.49947    0.05386   9.273 4.58e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4814 on 98 degrees of freedom
Multiple R-squared:  0.4674,	Adjusted R-squared:  0.4619 
F-statistic: 85.99 on 1 and 98 DF,  p-value: 4.583e-15
```

The linear regression fits a model close to the true value of the coefficients as was constructed. The model has a large F-statistic with a near-zero p-value so the null hypothesis can be rejected.

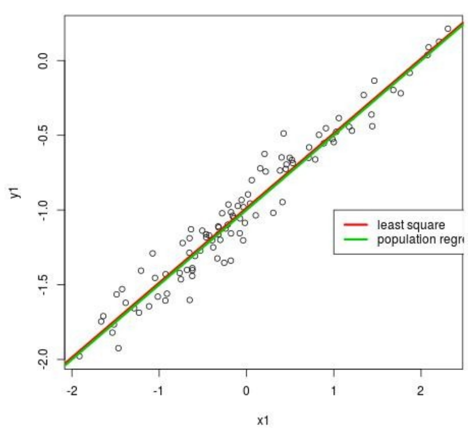
```
f) > jpeg("13f.jpg")
> plot(x, y)
> abline(lm.fit, lwd=3, col=2)
> abline(-1, 0.5, lwd=3, col=3)
> legend(-1, legend = c("least square", "population regression"), col=2:3, lwd=3)
> dev.off()
```




```
g) > lm.fit_sq = lm(y~x+I(x^2))
> summary(lm.fit_sq)

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

Residuals:
     Min       1Q   Median       3Q      Max 
-0.98252 -0.31270 -0.06441  0.29014  1.13500 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.97164    0.05883 -16.517  < 2e-16 ***
x            0.50858    0.05399   9.420  2.4e-15 ***
I(x^2)      -0.05946    0.04238  -1.403    0.164    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.479 on 97 degrees of freedom
Multiple R-squared:  0.4779,	Adjusted R-squared:  0.4672 
F-statistic:  44.4 on 2 and 97 DF,  p-value: 2.038e-14
```

```
h) > set.seed(1)
> eps1 = rnorm(100, 0, 0.125)
> x1 = rnorm(100)
> y1 = -1 + 0.5*x1 + eps1
> lm.fit1 = lm(y1~x1)
> summary(lm.fit1)

Call:
lm(formula = y1 ~ x1)

Residuals:
    Min       1Q   Median       3Q      Max 
-0.29052 -0.07545  0.00067  0.07288  0.28664 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.98639    0.01129  -87.34   <2e-16 ***
x1           0.49988    0.01184   42.22   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1128 on 98 degrees of freedom
Multiple R-squared:  0.9479,	Adjusted R-squared:  0.9474 
F-statistic:  1782 on 1 and 98 DF,  p-value: < 2.2e-16

> jpeg("13h.jpg")
> plot(x1, y1)
> abline(lm.fit1, lwd=3, col=2)
> abline(-1, 0.5, lwd=3, col=3)
> legend(-1, legend = c("least square", "population regression"), col=2:3, lwd=3)
> dev.off()
```

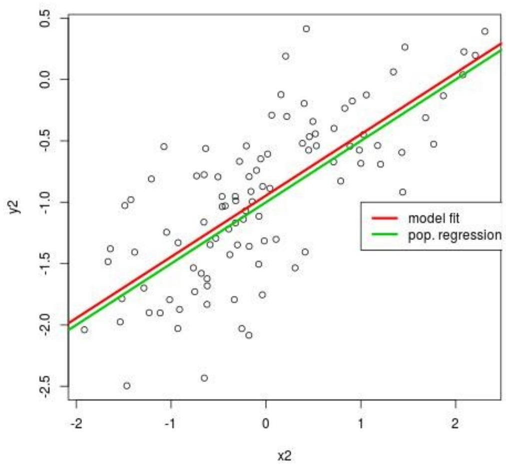
As expected, the error observed in R^2 and RSE decreases considerably.




```
i) > set.seed(1)
> eps2 = rnorm(100, 0, 0.5)
> x2 = rnorm(100)
> y2 = -1 + 0.5*x2 + eps2
> lm.fit2 = lm(y2~x2)
> summary(lm.fit2)

Call:
lm(formula = y2 ~ x2)

Residuals:
    Min       1Q   Median       3Q      Max 
-1.16208 -0.30181  0.00268  0.29152  1.14658 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.94557    0.04517  -20.93   <2e-16 ***
x2           0.49953    0.04736   10.55   <2e-16 ***
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4514 on 98 degrees of freedom
Multiple R-squared:  0.5317,	Adjusted R-squared:  0.5269
F-statistic: 111.2 on 1 and 98 DF,  p-value: < 2.2e-16

> jpeg("13i.jpg")
> plot(x2, y2)
> abline(lm.fit2, lwd=3, col=2)
> abline(-1, 0.5, lwd=3, col=3)
> legend(-1, legend = c("model fit", "pop. regression"), col=2:3, lwd=3)
> dev.off()
```

```
> confint(lm.fit)
                 2.5 %     97.5 %
(Intercept) -1.1150804 -0.9226122
x            0.3925794  0.6063602
> confint(lm.fit1)
                2.5 %     97.5 %
(Intercept) -1.008805 -0.9639819
x1           0.476387  0.5233799
> confint(lm.fit2)
                 2.5 %     97.5 %
(Intercept) -1.0352203 -0.8559276
x2           0.4055479  0.5935197
```



All intervals seem to be centered on
approximately 0.5, with the second fit’s interval being narrower than the first fit’s interval and the last fit’s interval being wider than the first fit’s interval.