ISL Chapter 3 Excercises

Jonathan Bryan

July 26, 2017

1. Describe the null hypotheses to which the p-values given in Table 3.4 correspond. Explain what conclusions you can draw based on these p-values. Your explanation should be phrased in terms of sales, TV, radio, and newspaper, rather than in terms of the coefficients of the linear model.

The null hypothesis tested for the coefficients in the linear regression model asserts that there is no relationship between the predictor and the response variable, which mathematically means the coefficient is zero. At an alpha level of 0.05 we can say that both TV and radio advertising spending are significant predictors for the sales response variable. More specifically, the model says that every dollar towards TV advertising is associated with an additional $46 in sales and every dollar towards radio advertising is associated with $189 dollars in sales.

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

Both KNN classifiers and regressions are non-parametric models. Both methods use a distance metric to calculate a pre-scpecified K number of data point's nearest neighbors. KNN classifiers then use a voting method to predict a class based upon the labeled classes of the neighbors. KNN regressions average the numerical values of the neighbors to generate an estimate.

3. Suppose we have a data set with five predictors, X1 = GPA, X2 = IQ, X3 = Gender (1 for Female and 0 for Male), X4 = Interaction between GPA and IQ, and X5 = Interaction between GPA and Gender. The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get $\hat{\beta}_0$ = 50, $\hat{\beta}_1$ = 20, $\hat{\beta}_2$ = 0.07, $\hat{\beta}_3$ = 35, $\hat{\beta}_4$ = 0.01, $\hat{\beta}_5$ = -10.

(a) Which answer is correct, and why?

For a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough. This is because while the model estimates $35,000 increase in salary for females, if GPA is higher than 3.5 then the GPA-Gender interaction term for females overwhelms the $35,000 increase. (iii) is correct.

(b) Predict the salary of a female with IQ of 110 and a GPA of 4.0.

$50 + 20(4.0) + 0.07(110) + 35(1) + 0.01(4.0 * 110) - 10(4.0 * 1) = \$137,100$

(c) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer

False. "Evidence" for an interaction effect would be evaluated for example by using hypothesis testing or analysis of vairance of the coeffieicent not its numerical scale. The effect may in fact be large because IQ is on a numerically larger scale compared to GPA and the binary 0 or 1 for gender. In addition, IQ is multiplied by GPA, increasing the relative scale of the interaction term.

4. I collect a set of data (n = 100 observations) containing a single predictor and a quantitative response. I then fit a linear regression model to the data, as well as a separate cubic regression, i.e. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$**.**

(a)Suppose that the true relationship between X and Y is linear, i.e. $Y = \beta_0 + \beta_1 X + \epsilon$**. Consider the training residual sum of squares (RSS) for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.**

This would depend on the training sample and therefore there is not enough information. This is because, while the true relationship between Y and X maybe be linear, the cubic regression may fit the training sample "closer" than the linear regression, therefore giving us a lower training residual sum of squares. Alternatively, the nonlinear curves of the cubic model may overshoot or undershoot the training data.

(b) Answer (a) using test rather than training RSS.

The testing resdiual sum of squares would be expected to be lower for the linear model rather than than the cubic regression model. This is because the flexbility of the cubic regression model will to closely track the training data and lead to larger residuals when fit to the testing data.

(c) Suppose that the true relationship between X and Y is not linear, but we don't know how far it is from linear. Consider the training RSS for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

There is not enough information to tell. The size of the residuals for each model will be highly dependent on how nonlinear the data is. Small deviations in linearity may not disadvantage the linear regression in comparison to the cubic regression model, however, if the relationship between the predictor and response was in fact quadratic, cubic, or higher, we would expect the cubic model to perform better.

(d) Answer (c) using test rather than training RSS.

This answer is similar to part (c). The outcome is highly dependent on the nonlinearity of the population data. One difference is that we would expect the performance of the cubic regression model to perform much better in comparison to the training data (lower test residual sum of squares) as nonlinearity increases.

5. Consider the fitted values that result from performing linear regression without an intercept. In this setting, the ith fitted value takes the form

$$\hat{y_i} = x_i \hat{\beta}$$

,

where

$$\hat{\beta} = \frac{\sum_{j=1}^{n} x_j y_j}{\sum_{k=1}^{n} x_k^2}$$

.

Show that we can write

$$\hat{y_i} = \sum_{j=1}^{n} a_j y_j$$

.

What is a_j?

$$\hat{y_i} = x_i(\frac{\sum_{j=1}^{n} x_j y_j}{\sum_{k=1}^{n} x_k^2})$$
$$\hat{y_i} = \sum_{j=1}^{n}(\frac{x_j x_i}{\sum_{k=1}^{n} x_k^2})y_j$$
$$a_j = (\frac{\sum_{j=1}^{n} x_j x_i}{\sum_{k=1}^{n} x_k^2})$$

6. Using (3.4), argue that in the case of simple linear regression, the least squares line always passes through the point $(\bar{x}, \bar{y})$.

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$
$$y = \beta_0 + x_i \beta_1$$
$$y = \bar{y} - \bar{x}\beta_1 + x_i \beta_1$$
$$x_i = \bar{x},\ y = \bar{y} - \bar{x}\beta_1 + \bar{x}\beta_1 = \bar{y}$$

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 $\bar{x} = \bar{y} = 0$.

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$
$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$
$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i)^2}$$
$$\hat{y_i} = \hat{\beta}_0 + \hat{\beta}_1 x_1$$
$$\hat{y_i} = \bar{y} + \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}\bar{x} + \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} x_i$$
$$\hat{y_i} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} x_i$$
$$\hat{y_i} = \frac{\sum_{i=1}^n (x_i)(y_i)}{\sum_{i=1}^n (x_i)^2} x_i$$
$$R^2 = 1 - \frac{\sum(y_i - \frac{\sum_{i=1}^n (x_i)(y_i)}{\sum_{i=1}^n (x_i^2)} x_i)^2}{\sum(y_i)^2}$$
$$R^2 = 1 - \frac{\sum(y_i^2 - 2 * \frac{\sum_{i=1}^n (x_i)(y_i)}{\sum_{i=1}^n (x_i^2)} x_i y_i + (\frac{\sum_{i=1}^n (x_i)(y_i)}{\sum_{i=1}^n (x_i^2)})^2 x_i^2)}{\sum(y_i^2)}$$
$$R^2 = \frac{\sum(y_i^2) - \sum y_i^2 + \sum 2 * \frac{\sum_{i=1}^n (x_i)(y_i)}{\sum_{i=1}^n (x_i^2)} x_i y_i - \sum (\frac{\sum_{i=1}^n (x_i)(y_i)}{\sum_{i=1}^n (x_i^2)})^2 x_i^2)}{\sum(y_i^2)}$$
$$R^2 = \frac{2\sum \frac{\sum_{i=1}^n (x_i)(y_i)}{\sum_{i=1}^n (x_i^2)} x_i y_i - \sum (\frac{\sum_{i=1}^n (x_i^2)(y_i^2) x_i^2}{\sum_{i=1}^n (x_i^4)})}{\sum(y_i^2)}$$
$$R^2 = \frac{2\frac{\sum_{i=1}^n (x_i^2)(y_i^2)}{\sum_{i=1}^n (x_i^2)} - (\frac{\sum_{i=1}^n (x_i^2)(y_i^2)}{\sum_{i=1}^n (x_i^2)})}{\sum(y_i^2)}$$
$$R^2 = \frac{\frac{\sum_{i=1}^n (x_i^2)(y_i^2)}{\sum_{i=1}^n (x_i^2)}}{\sum(y_i^2)}$$
$$R^2 = \frac{\sum(x_i^2)(y_i^2)}{\sum(y_i^2)\sum(x_i^2)}$$
$$Cor(X,Y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$
$$Cor(X,Y)^2 = (\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}})^2$$
$$Cor(X,Y)^2 = \frac{(\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}))^2}{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}$$
$$Cor(X,Y)^2 = \frac{(\sum_{i=1}^n (x_i)(y_i))^2}{\sum_{i=1}^n (x_i)^2 \sum_{i=1}^n (y_i)^2}$$
$$R^2 = \frac{\sum(x_i^2)(y_i^2)}{\sum(y_i^2)\sum(x_i^2)} = Cor(X,Y)^2 = \frac{(\sum_{i=1}^n (x_i)(y_i))^2}{\sum_{i=1}^n (x_i)^2 \sum_{i=1}^n (y_i)^2}$$

8. This question involves the use of simple linear regression on the Auto data set.

(a) Use the lm() function to perform a simple linear regression with mpg as the response and horsepower as the predictor. Use the summary() function to print the results. Comment on the output.

```
#Load data
library(ISLR)
auto = Auto
```

```
#Fit simple linear regression
lm.fit = lm(mpg ~ horsepower, data = auto)
summary(lm.fit)
```

```
##
## 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
```

(i) Is there a relationship between the predictor and the response?

Given the high F-statistic of the model we can say the model itself and thus the sole predictor is significant. In addition, the p-value of the horsepower coefficient is lower than 0.01 so we can say the horsepower is a significant predictor and therefore there is a relationship between the predictor and the response.

(ii) How strong is the relationship between the predictor and the response?

For every 1 unit increase in horsepower, there is a correspondiing 0.15 decrease in mpg. The model explains 60 percent of the variation in mpg.

(iii) Is the relationship between the predictor and the response positive or negative?

The horsepower coefficient is negative (-0.157) and a 99% confidence interval does not include a change of sign, therefore we can say the relationship between horsepower and mpg is negative.

(iv) What is the predicted mpg associated with a horsepower of 98? What are the associated 95% confidence and prediction intervals?

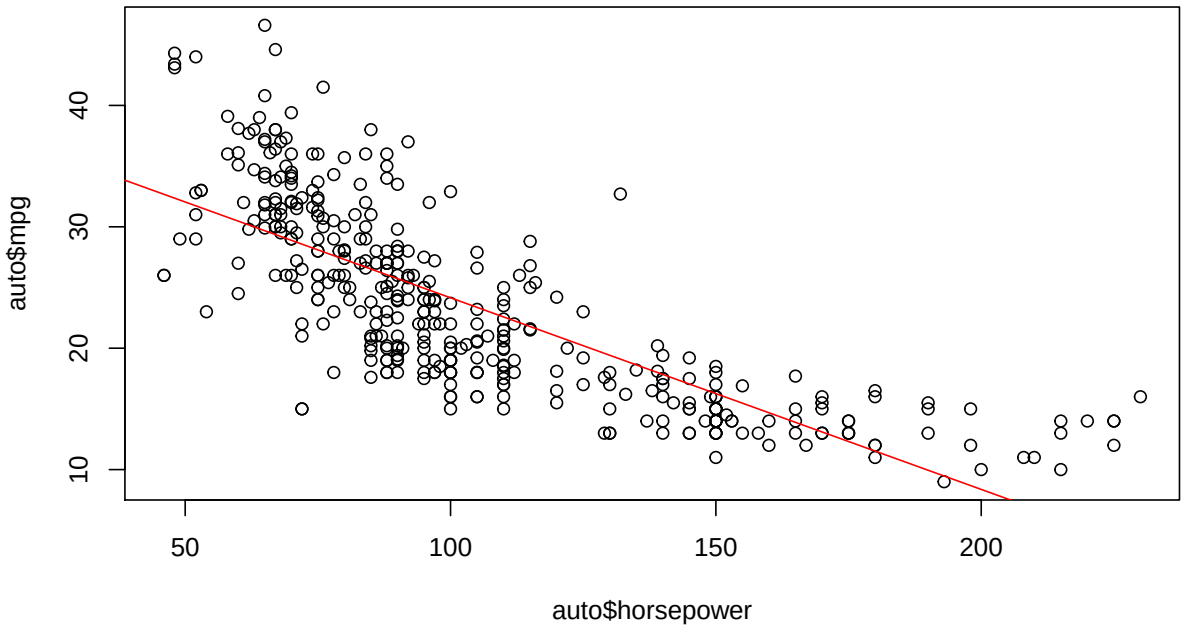
```
#mpg prediction for 98 horsepower
mpg_pred_pi = predict(lm.fit, data.frame(horsepower = c(98)), interval = "confidence")
```

```
mpg_pred_ci = predict(lm.fit, data.frame(horsepower = c(98)), interval = "prediction")
print(paste0("The point estimate for mpg at 98 horsepower is ",mpg_pred_pi[1], ", while the confidence i
```

```
## [1] "The point estimate for mpg at 98 horsepower is 24.4670771525124, while the confidence interval i
```

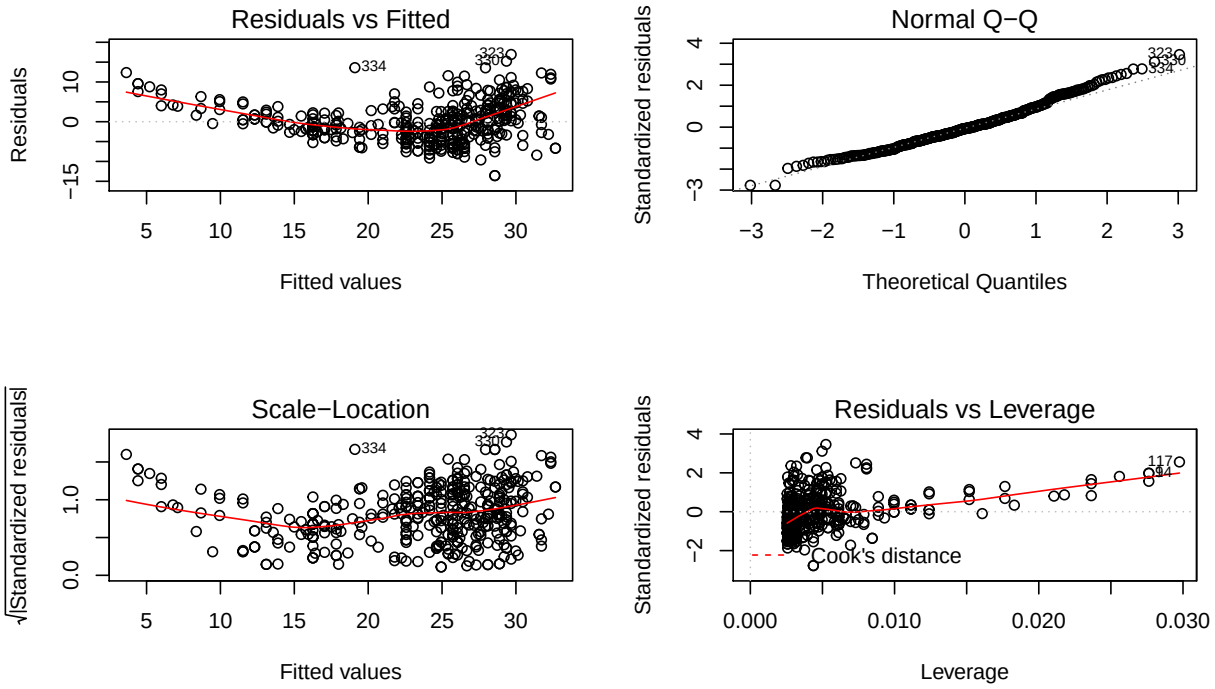
(b)

```
plot(auto$horsepower, auto$mpg)
abline(lm.fit, col="red")
```



(C)

```
par(mfrow=c(2,2))
plot(lm.fit)
```

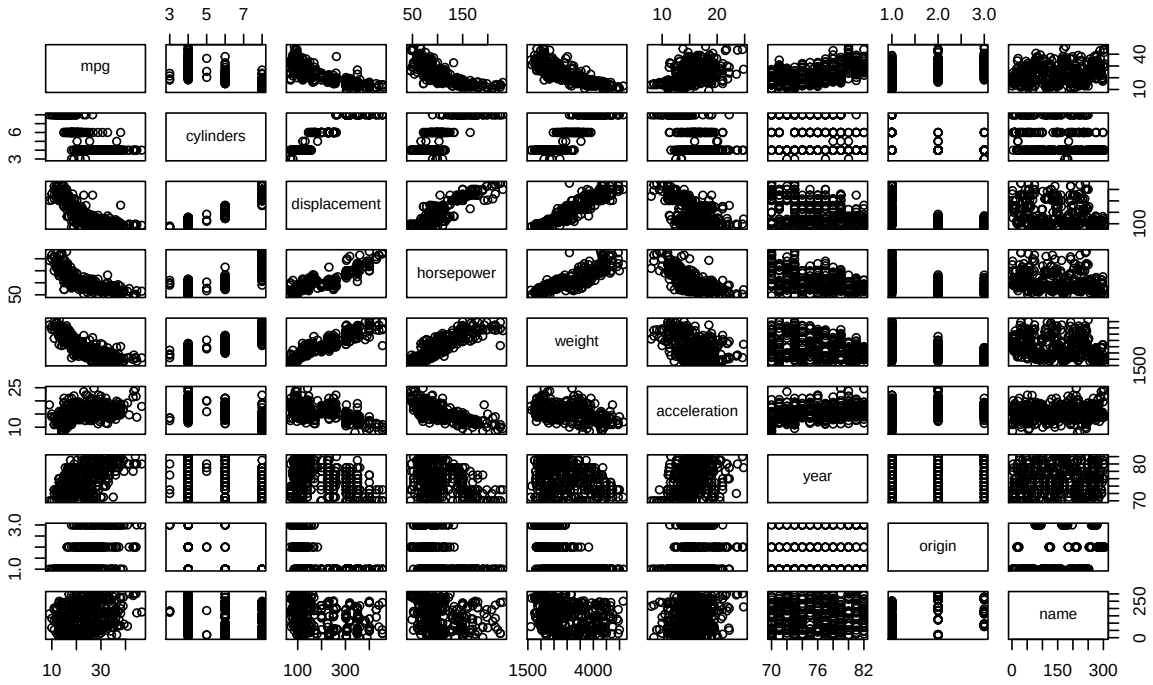


Residuals show positive bias at the lower and upper extremes of mpg, while negative near the median. This could be because the true relationship between horsepower and mpg is polynomial in nature, possibly quadratic. The Normal Q-Q plot demonstrates deviations from normality at the lower and upper extremes of the standardized residuals. Finally there are number of high leverage points but very little outliers.

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.

```
#Scatterplot Matrix
pairs(auto)
```



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

```
cor(auto[,-which(names(auto) == "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) 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.

```
lm.fit2 = lm(mpg ~ . -name, data = auto)
summary(lm.fit2)
```

```
##
## 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) Is there a relationship between the predictors and the response?

The adjusted R-squared of the model is approximately 0.81 which means the model caputres a 81% of the variation in mpg. The F-statistic is large enough as to reject the null hypothesis that there is no relation between the predictors and the response variable.

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

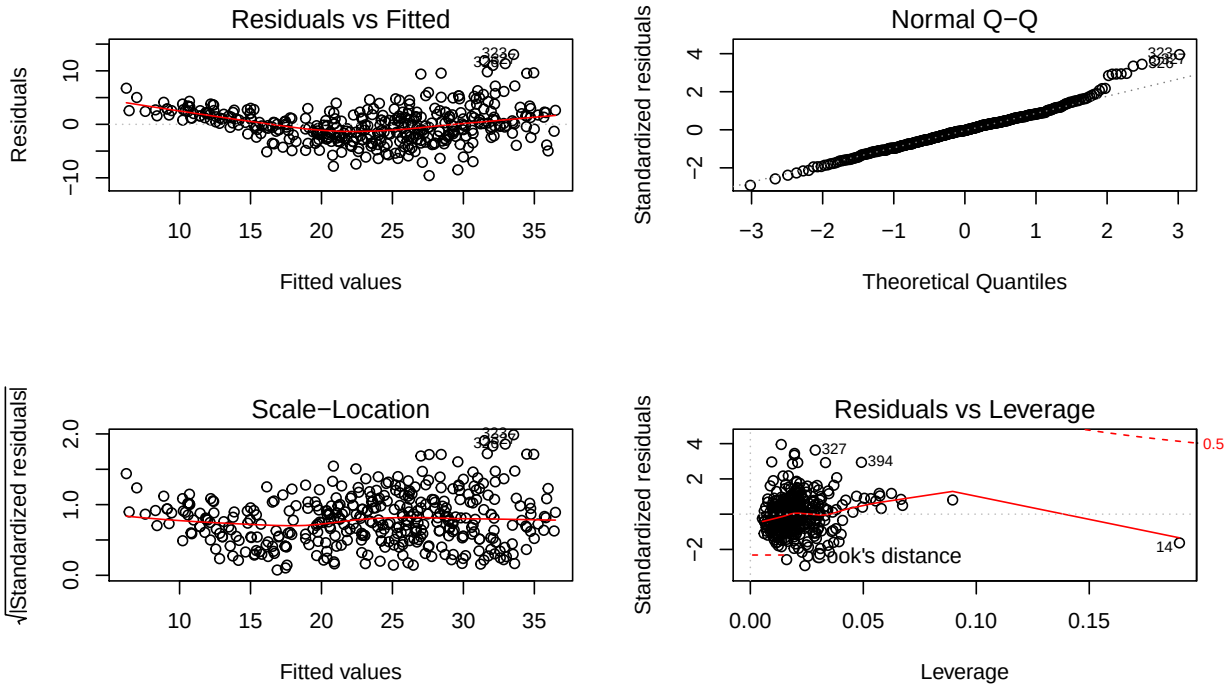
Displacement, weight, year, and origin are all statistically significant at an alpha level of 0.01.

(iii) What does the coefficient for the year variable suggest?

The coefficient for year is 0.75. This suggests that, holding all other predictors fixed, the mpg inceases by 0.75 for each additional year.

(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?

```
par(mfrow=c(2,2))
plot(lm.fit2)
```



Positive residual outliers exist at the upper end of mpg and at the very lower end. There is a preponderance of negative residual outliers in the middle of the mpd data between 15 and 30. Two obervations have leverage substantially above 0.02 (derived using p+1/n which is a dianostic test for high leverage observations), including observation 14 which has very high leverage.

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

```
lm.fit3 = lm(mpg ~ . +cylinders:horsepower+ year:origin+weight:acceleration -name, data = auto)
summary(lm.fit3)
```

```
##
## Call:
## lm(formula = mpg ~ . + cylinders:horsepower + year:origin + weight:acceleration -
##     name, data = auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.4518 -1.7308 -0.0691  1.2564 12.3733
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         15.4372225 10.3005141   1.499   0.1348
## cylinders           -3.7644629  0.5616873  -6.702 7.42e-11 ***
## displacement        -0.0046291  0.0070553  -0.656   0.5121
## horsepower          -0.2816329  0.0356290  -7.905 2.93e-14 ***
## weight              -0.0018075  0.0017026  -1.062   0.2891
## acceleration         0.2142857  0.2934048   0.730   0.4656
```

```
## year                  0.5757343  0.1010375   5.698 2.43e-08 ***
## origin               -7.1395988  4.2340565  -1.686   0.0926 .
## cylinders:horsepower  0.0350908  0.0048037   7.305 1.64e-12 ***
## year:origin           0.1031158  0.0544358   1.894   0.0589 .
## weight:acceleration  -0.0001305  0.0000982  -1.329   0.1847
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.917 on 381 degrees of freedom
## Multiple R-squared:  0.8639, Adjusted R-squared:  0.8604
## F-statistic: 241.9 on 10 and 381 DF,  p-value: < 2.2e-16
```

The cylinders and horsepower interaction effect is significant at an alpha level of .01. Neither the year-origin or weight-accelration is significant at an alpha level of 0.05.

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

```
lm.fit4 = lm(mpg ~ (horsepower^2) + log(displacement) +sqrt(year), data = auto )
summary(lm.fit4)
```

```
##
## Call:
## lm(formula = mpg ~ (horsepower^2) + log(displacement) + sqrt(year),
##     data = auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.887  -2.066  -0.319   1.785  14.154
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -32.205112   8.655246  -3.721 0.000228 ***
## horsepower         -0.019541   0.009373  -2.085 0.037738 *
## log(displacement)  -9.347652   0.651990 -14.337  < 2e-16 ***
## sqrt(year)         12.121339   0.938692  12.913  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.561 on 388 degrees of freedom
## Multiple R-squared:  0.7935, Adjusted R-squared:  0.7919
## F-statistic: 496.9 on 3 and 388 DF,  p-value: < 2.2e-16
```

Horsepower was raised to a power of two, the log was taken of displacement, and the square root was taken of year. Each of these transformed variables was significant and the model achieve an adjusted R-squared of 0.79 which is close to the full model excluding name.

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

```
library(car) #For diagnostics
carseats = Carseats
```

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

```
lm.car = lm(Sales ~ Price + Urban + US, data=carseats)
```

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

```
summary(lm.car)
```

```
##
## 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
```

The model estimates each dollar increase in price results in 0.05 less unit sales, holding all other covariates constant. Urban sales locations produce 0.02 less sales while US locations being in an estimated 1.2 more in unit sales, again holding all other variables constant respectively.

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

$Sales = 13.043 + (-0.054) * Price + (-0.022) UrbanYes + (1.2) USYes + \epsilon$

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

At an alpha level of 0.01 we can reject the null hypothesis for Price and US.

(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.

```
lm.car2 = lm(Sales ~ Price + US, data = carseats)
summary(lm.car2)
```

```
##
## Call:
## lm(formula = Sales ~ Price + US, data = carseats)
##
```

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

Both models have F-statistics high enough to reject the null hypothesis that the response has no relationship with the predictors. However, both models have low, albeit nearly equal, R-squared values (approximately 0.23). Therefore both models fit the data somewhat poorly.

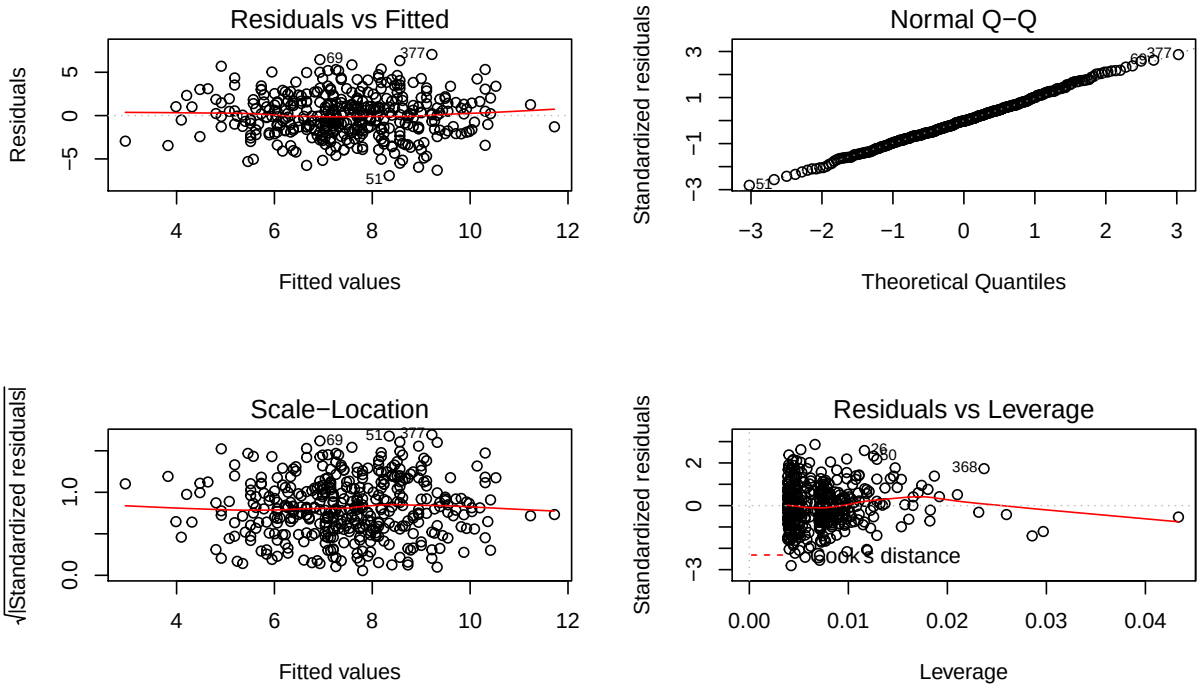
(g) Using the model from (e), obtain 95 % confidence intervals for the coefficient(s).

```
confint(lm.car2)
```

```
##                   2.5 %      97.5 %
## (Intercept) 11.79032020 14.27126531
## Price       -0.06475984 -0.04419543
## USYes        0.69151957  1.70776632
```

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

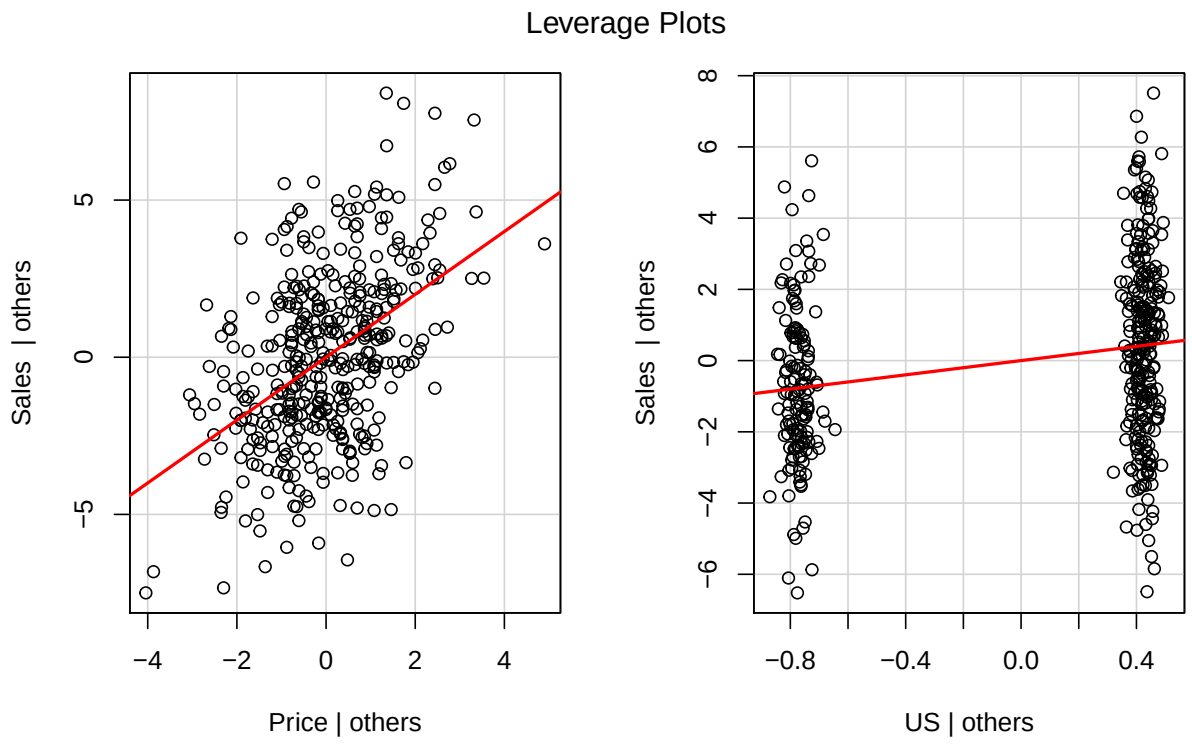
```
par(mfrow=c(2,2))
plot(lm.car2)
```



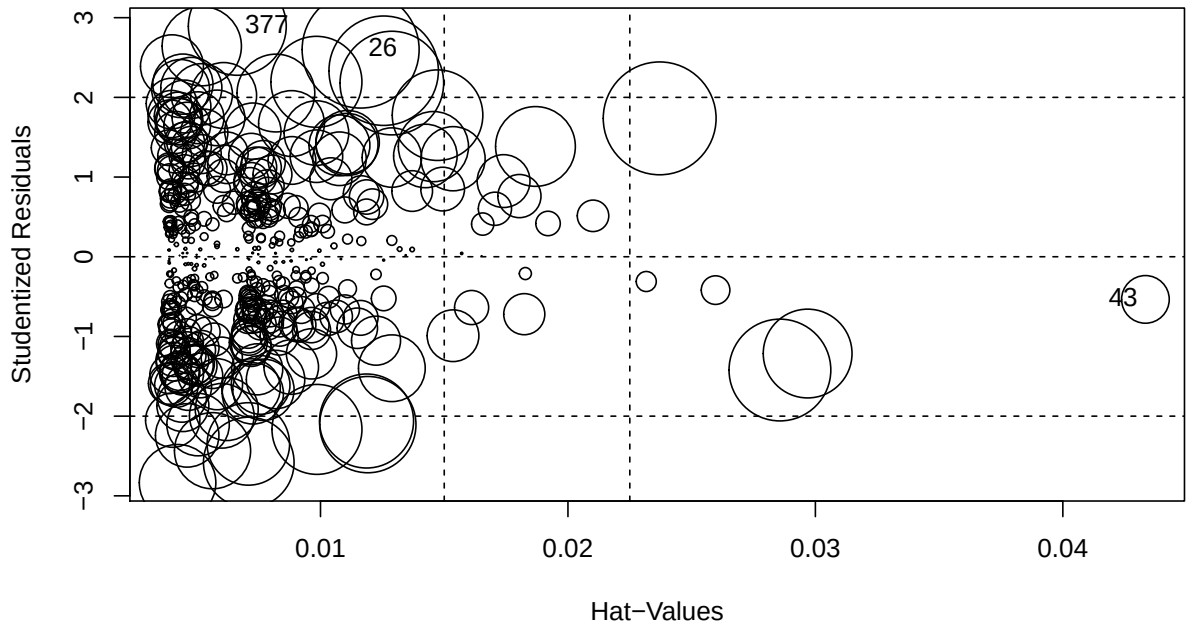

```
outlierTest(lm.car2)
```

```
##
## No Studentized residuals with Bonferonni p < 0.05
## Largest |rstudent|:
##     rstudent unadjusted p-value Bonferonni p
## 377 2.891521          0.0040452           NA
```

```
leveragePlots(lm.car2)
```




```
influencePlot(lm.car2)
```




```
##         StudRes         Hat       CookD
## 26   2.5996518 0.011621599 0.026109457
```

```
## 43  -0.5349931 0.043337657 0.004329756
## 377  2.8915213 0.006637175 0.018282191
```

The residual plots show approximately normally distributed residuals. Analysis of the residuals vs. leverage plot show no values with extraordinary Cook's distances. The Bonferroni-adjusted outlier test indicates no outliers, however, it shows the observation with the largest studentized residual is observation 377.

11. In this problem we will investigate the t-statistic for the null hypothesis $H_0 : \beta = 0$ in simple linear regression without an intercept. To begin, we generate a predictor x and a response y as follows:

```
set.seed(1)
x = rnorm(100)
y = 2*x + rnorm(100)
```

(a) Perform a simple linear regression of y onto x, without an intercept. Report the coefficient estimate $\hat{\beta}$, the standard error of this coefficient estimate, and the t-statistic and p-value associated with the null hypothesis $H_0 : \beta = 0$. Comment on these results. (You can perform regression without an intercept using the command lm(y ~ x+0).)

```
fit.y_no_intercept = lm(y ~ x + 0)
summary(fit.y_no_intercept)
```

```
##
## 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
```

$\hat{\beta} = 1.99$ for the x predictor. The p-value for the F-statistic of the model is below 0.05 so we can say the model is significant and the R^2 is 0.77 so the model fits the data fairly well.

(b) Now perform a simple linear regression of x onto y without an intercept, and report the coefficient estimate, its standard error, and the corresponding t-statistic and p-values associated with the null hypothesis $H_0 : \beta = 0$. Comment on these results.

```
fit.x_no_intercept = lm(x ~ y + 0)
summary(fit.x_no_intercept)
```

```
##
## 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
```

$\hat{\beta} = 0.39$ for the y predictor. Similar to the previous model, the p-value for the F-statistic of the model is below 0.05 so we can say the model is significant and the R^2 is 0.77 so the model fits the data fairly well.

(c) What is the relationship between the results obtained in (a) and (b)?

They are both modeling the same line so they have different coefficients but the same regression diagnostics.

(d) For the regression of Y onto X without an intercept, the t-statistic for $H_0 : \beta = 0$ takes the form $\hat{\beta}/SE(\hat{\beta})$, where $\hat{\beta}$ is given by (3.38), and where

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

. (These formulas are slightly different from those given in Sections 3.1.1 and 3.1.2, since here we are performing regression without an intercept.) Show algebraically, and confirm numerically in R, that the t-statistic can be written as

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

(f) In R, show that when regression is performed with an intercept, the t-statistic for $H_0 : \beta_1 = 0$ is the same for the regression of y onto x as it is for the regression of x onto y.

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?

The coefficient will be the same When the relationship between Y and X is one-to-one and doesn't require a transformation.

(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.

```
x = rnorm(100)
y = sqrt(x+10) + rnorm(100)
lm.x_fit_random = lm(y ~ x + 0)
lm.y_fit_random = lm(x ~ y + 0)
summary(lm.x_fit_random)
```

```
##
## Call:
## lm(formula = y ~ x + 0)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## 0.3787 2.6224 3.2507 3.9229 5.2443
##
## Coefficients:
##   Estimate Std. Error t value Pr(>|t|)
## x   0.3556     0.3272   1.087     0.28
##
## Residual standard error: 3.369 on 99 degrees of freedom
## Multiple R-squared:  0.01179,    Adjusted R-squared:  0.001806
## F-statistic: 1.181 on 1 and 99 DF,  p-value: 0.2798
```

```
summary(lm.y_fit_random)
```

```
##
## Call:
## lm(formula = x ~ y + 0)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.92531 -0.57973 -0.08839  0.57499  2.53354
##
## Coefficients:
##   Estimate Std. Error t value Pr(>|t|)
## y  0.03315    0.03051   1.087     0.28
##
## Residual standard error: 1.029 on 99 degrees of freedom
## Multiple R-squared:  0.01179,    Adjusted R-squared:  0.001806
## F-statistic: 1.181 on 1 and 99 DF,  p-value: 0.2798
```

(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.

```
x = rnorm(100)
y = x + rnorm(100, mean = 0, sd = 0.5)
lm.x_fit_random = lm(y ~ x + 0)
lm.y_fit_random = lm(x ~ y + 0)
summary(lm.x_fit_random)
```

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

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.27853 -0.32730 -0.01274  0.32936  1.02101
##
## Coefficients:
##   Estimate Std. Error t value Pr(>|t|)
## x  0.96321    0.04137   23.28   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4817 on 99 degrees of freedom
## Multiple R-squared:  0.8456, Adjusted R-squared:  0.844
## F-statistic: 542.2 on 1 and 99 DF,  p-value: < 2.2e-16
```

```
summary(lm.y_fit_random)
```

```
##
## Call:
## lm(formula = x ~ y + 0)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.86028 -0.33297  0.02367  0.33125  1.39336
##
## Coefficients:
##   Estimate Std. Error t value Pr(>|t|)
## y   0.8779     0.0377   23.28   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4599 on 99 degrees of freedom
## Multiple R-squared:  0.8456, Adjusted R-squared:  0.844
## F-statistic: 542.2 on 1 and 99 DF,  p-value: < 2.2e-16
```

????

13. In this exercise you will create some simulated data and will fit simple linear regression models to it. Make sure to use set.seed(1) prior to starting part (a) to ensure consistent results.

(a) Using the rnorm() function, create a vector, x, containing 100 observations drawn from a N(0, 1) distribution. This represents a feature, X.

```
set.seed(1)
x = rnorm(100, mean = 0, sd = 1)
```

(b) Using the rnorm() function, create a vector, eps, containing 100 observations drawn from a N(0, 0.25) distribution i.e. a normal distribution with mean zero and variance 0.25.

```
eps = rnorm(100, mean = 0, sd = 0.25)
```

(c) Using x and eps, generate a vector y according to the model $Y = -1 + 0.5X + \epsilon$. What is the length of the vector y? What are the values of β_0 and β_1 in this linear model?

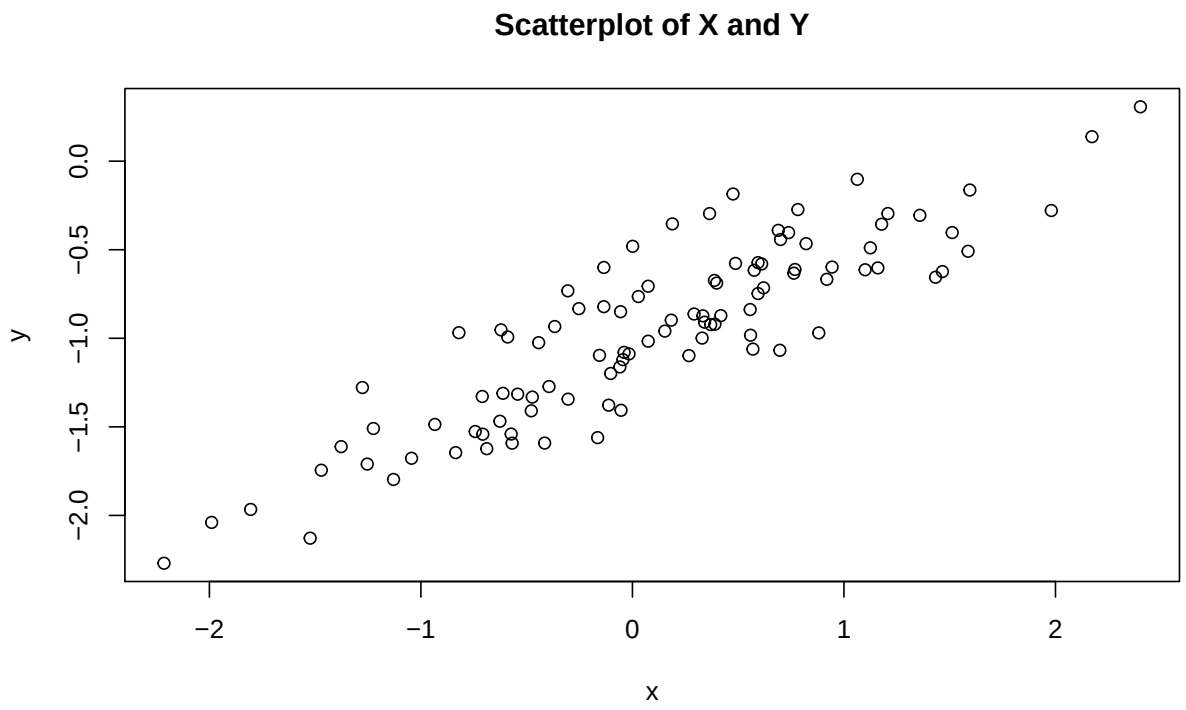
```
y = -1 + 0.5*x + eps
length(y)
```

```
## [1] 100
```

The length of the y vector is 100. β_0 = -1 and $\beta_1 = 0.5$.

(d) Create a scatterplot displaying the relationship between x and y. Comment on what you observe.

```
plot(x,y, main = "Scatterplot of X and Y")
```



We observe a positive relationship between x and y. The y-intercept is approximately -1 while one could easily draw a regression line with a slope of 0.5 through the points.

(e) Fit a least squares linear model to predict y using x. Comment on the model obtained. How do $\hat{\beta}_0$ and $\hat{\beta}_1$ compare to β_0 and β_1?

```
fit.random = lm(y ~ x)
summary(fit.random)
```

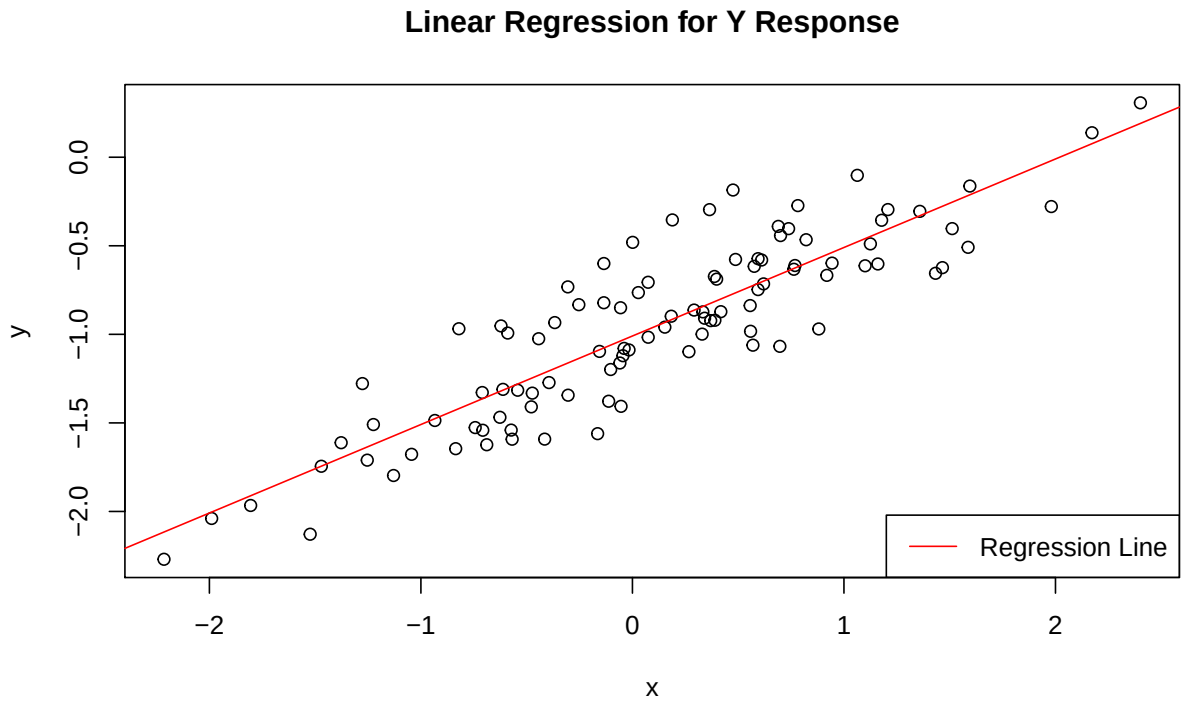
```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.46921 -0.15344 -0.03487  0.13485  0.58654
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.00942    0.02425  -41.63   <2e-16 ***
## x            0.49973    0.02693   18.56   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2407 on 98 degrees of freedom
## Multiple R-squared:  0.7784, Adjusted R-squared:  0.7762
## F-statistic: 344.3 on 1 and 98 DF,  p-value: < 2.2e-16
```

The intercept is approximately -1.01 and the coefficient of x is approximately 0.499. These are very close to the original programmed values.

(f) Display the least squares line on the scatterplot obtained in (d). Draw the population regression line on the plot, in a different color. Use the legend() command to create an appropriate legend.

```
plot(x,y, main ="Linear Regression for Y Response")
abline(fit.random, col = "red")
legend(x = "bottomright", legend = c("Regression Line"), col = "red", lty = 1)
```



(g) Now fit a polynomial regression model that predicts y using x and x^2. Is there evidence that the quadratic term improves the model fit? Explain your answer.

```
lm.fit_poly = lm(y ~ poly(x,2))
summary(lm.fit_poly)
```

```
##
## Call:
## lm(formula = y ~ poly(x, 2))
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4913 -0.1563 -0.0322  0.1451  0.5675
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.95501    0.02395 -39.874   <2e-16 ***
## poly(x, 2)1  4.46612    0.23951  18.647   <2e-16 ***
## poly(x, 2)2 -0.33602    0.23951  -1.403    0.164
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2395 on 97 degrees of freedom
## Multiple R-squared:  0.7828, Adjusted R-squared:  0.7784
## F-statistic: 174.8 on 2 and 97 DF,  p-value: < 2.2e-16
```

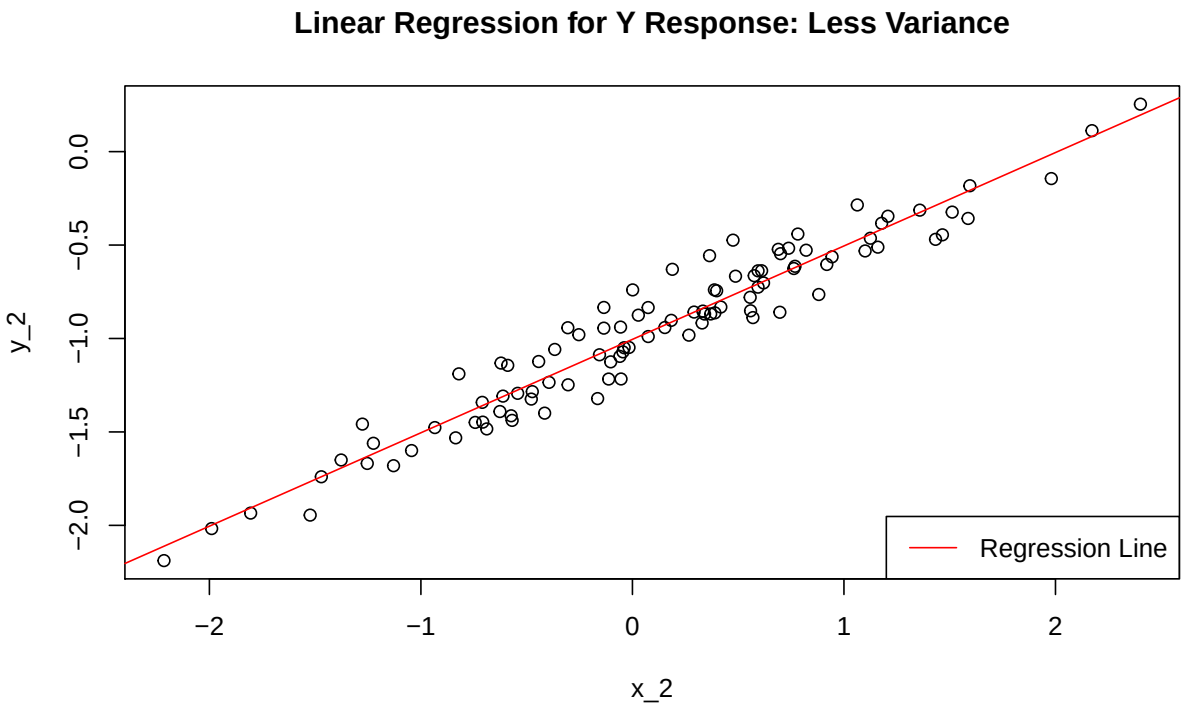
The adjusted-R squared model is nearly the same for the linear and quadratic models. The quadratic x term is not significant at an alpha level of 0.05 does little to improve the model fit.

(h) Repeat (a)-(f) after modifying the data generation process in such a way that there is less noise in the data. The model (3.39) should remain the same. You can do this by decreasing the variance of the normal distribution used to generate the error term in (b). Describe your results.

```
set.seed(1)
x_2 = rnorm(100, mean = 0, sd = 1)
eps_2 = rnorm(100, mean = 0, sd = 0.125)
y_2 = -1 + 0.5*x_2 + eps_2
fit.random_2 = lm(y_2 ~ x_2)
summary(fit.random_2)
```

```
##
## Call:
## lm(formula = y_2 ~ x_2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.23461 -0.07672 -0.01744  0.06742  0.29327
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.00471    0.01212  -82.87   <2e-16 ***
## x_2          0.49987    0.01347   37.12   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1203 on 98 degrees of freedom
## Multiple R-squared:  0.9336, Adjusted R-squared:  0.9329
## F-statistic:  1378 on 1 and 98 DF,  p-value: < 2.2e-16
```

```
plot(x_2,y_2, main = "Linear Regression for Y Response: Less Variance")
abline(fit.random_2, col = "red")
legend(x = "bottomright", legend = c("Regression Line"), col = "red", lty = 1)
```



When the variance is of the error term is reduced to a standard deviation of 0.125 we obeserve higher values for the adjusted R-squared (93.3), F-statistic (1378), and t values for the intercept and x term. Overall the model is a better fit when there is less noise in the data.

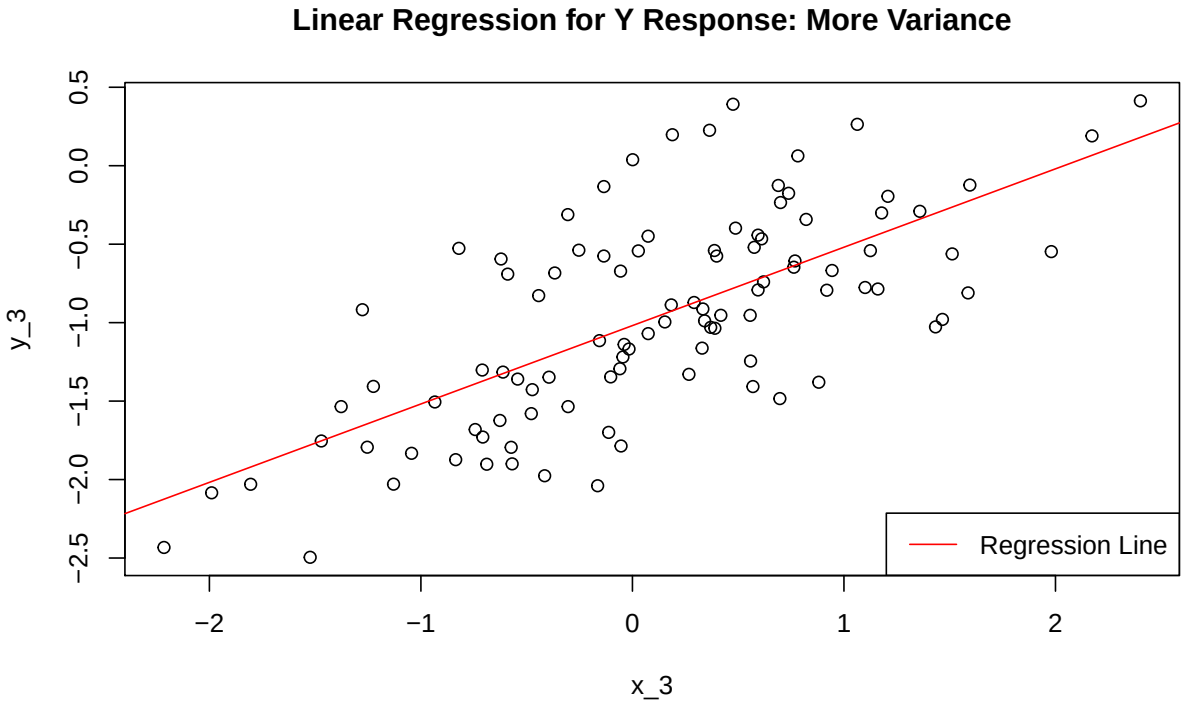
(i) Repeat (a)-(f) after modifying the data generation process in such a way that there is more noise in the data. The model (3.39) should remain the same. You can do this by increasing the variance of the normal distribution used to generate the error term in (b). Describe your results.

```
set.seed(1)
x_3 = rnorm(100, mean = 0, sd = 1)
eps_3 = rnorm(100, mean = 0, sd = 0.5)
y_3 = -1 + 0.5*x_3 + eps_3
fit.random_3 = lm(y_3 ~ x_3)
summary(fit.random_3)
```

```
##
## Call:
## lm(formula = y_3 ~ x_3)
##
## 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_3          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
```

```
plot(x_3,y_3, main = "Linear Regression for Y Response: More Variance")
abline(fit.random_3, col = "red")
legend(x = "bottomright", legend = c("Regression Line"), col = "red", lty = 1)
```



When the variance is of the error term is increased to a standard deviation of 0.5 we obeserve lower values for the adjusted R-squared (.462), F-statistic (85.99), and t values for the intercept and x term. Overall the model is a worse fit when there is more noise in the data.

(f) What are the confidence intervals for β_0 and β_1 based on the original data set, the noisier data set, and the less noisy data set? Comment on your results.

```
diff_1 = confint(fit.random)[1,2] - confint(fit.random)[1,1]
diff_12 = confint(fit.random)[2,1] - confint(fit.random)[2,2]
cbind(confint(fit.random), Difference = c(diff_1,diff_12))
```

```
##                  2.5 %     97.5 %  Difference
## (Intercept) -1.0575402 -0.9613061  0.09623406
## x            0.4462897  0.5531801 -0.10689039
```

```
diff_1 = confint(fit.random_2)[1,2] - confint(fit.random_2)[1,1]
diff_12 = confint(fit.random_2)[2,1] - confint(fit.random_2)[2,2]
cbind(confint(fit.random_2), Difference = c(diff_1,diff_12))
```

```
##                  2.5 %     97.5 %  Difference
## (Intercept) -1.0287701 -0.9806531  0.04811703
```

```
## x_2          0.4731449  0.5265901 -0.05344520
```

```
diff_1 = confint(fit.random_3)[1,2] - confint(fit.random_3)[1,1]
diff_12 = confint(fit.random_3)[2,1] - confint(fit.random_3)[2,2]
cbind(confint(fit.random_3), Difference = c(diff_1,diff_12))
```

```
##                   2.5 %     97.5 % Difference
## (Intercept) -1.1150804 -0.9226122  0.1924681
## x_3          0.3925794  0.6063602 -0.2137808
```

The confidence intervals for the data with relatively less variance has the tightest confidence intervals for the model intercept and x term. The highest variance model confidence intervals are twice as large as the middle model and nearly four times as large as the lowest variance model.

14. This problem focuses on the collinearity problem.

(a) The last line corresponds to creating a linear model in which y is a function of x1 and x2. Write out the form of the linear model. What are the regression coefficients?

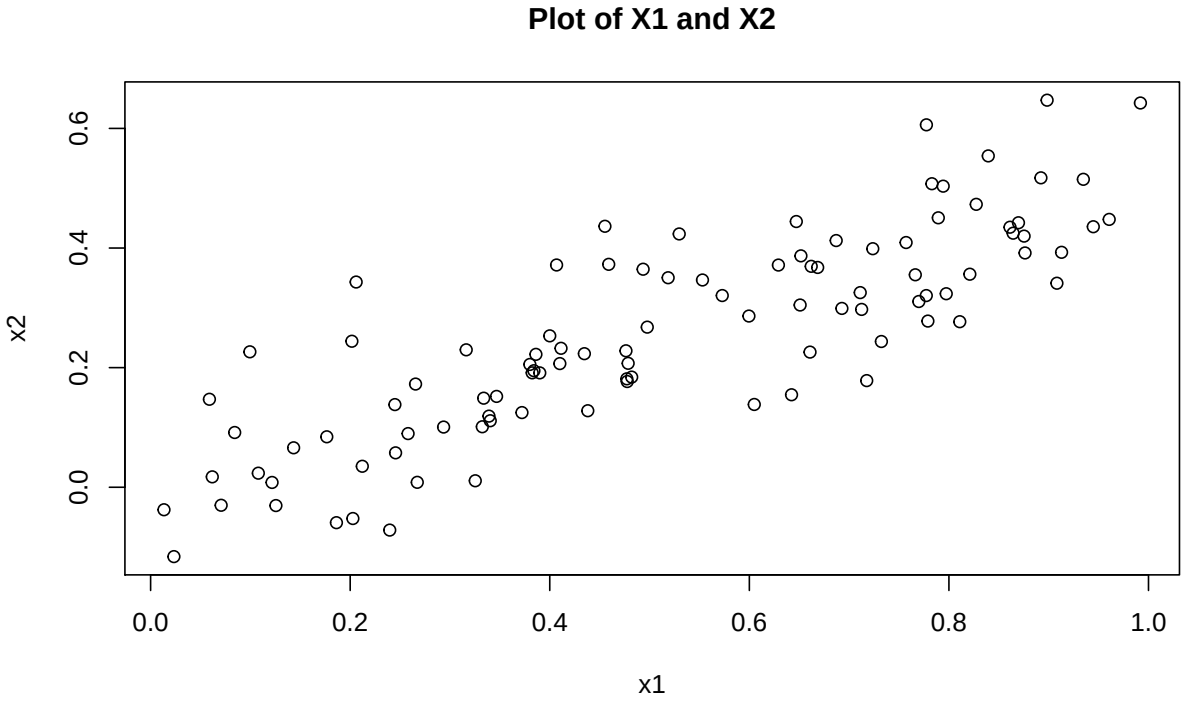
```
set.seed(1)
x1=runif (100)
x2=0.5*x1+rnorm (100)/10
y=2+2*x1+0.3*x2+rnorm (100)
```

$$y = 2 + 2 * x_1 + 0.3 * x_2 + \epsilon$$

The regression coefficients are $\beta_0 = 2,\ \beta_1 = 2,\ \beta_2 = 0.3$

(b) What is the correlation between x1 and x2? Create a scatterplot displaying the relationship between the variables.

```
plot(x1,x2, main = "Plot of X1 and X2")
```



X1 and X2 are positively coorelated.

(c) Using this data, fit a least squares regression to predict y using x1 and x2. Describe the results obtained. What are $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$? How do these relate to the true β_0, β_1, *and* β_2? Can you reject the null hypothesis $H_0 : \beta_1 = 0$? How about the null hypothesis $H_0 : \beta_2 = 0$?

```
lm.fit_cor = lm(y ~ x1 + x2)
summary(lm.fit_cor)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8311 -0.7273 -0.0537  0.6338  2.3359
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.1305     0.2319   9.188 7.61e-15 ***
## x1            1.4396     0.7212   1.996   0.0487 *
## x2            1.0097     1.1337   0.891   0.3754
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.056 on 97 degrees of freedom
## Multiple R-squared:  0.2088, Adjusted R-squared:  0.1925
## F-statistic:  12.8 on 2 and 97 DF,  p-value: 1.164e-05
```

The model is a moderately poor fit with a very low adjusted R-squared and uncompelling albeit significant

($\alpha = 0.05$) F-statistic. $\hat{\beta}_1$ is approximately 0.6 less than the true coefficient while $\hat{\beta}_2$ is 0.7 larger.$\hat{\beta}_0$ is only 0.13 different than the true intercept. We cannot reject either null hypothesis for the coefficients given their respectively high p-values.

(d) Now fit a least squares regression to predict y using only x1. Comment on your results. Can you reject the null hypothesis $H_0 : \beta_1 = 0$?

```
lm.fit_x1 = lm(y ~ x1)
summary(lm.fit_x1)
```

```
##
## Call:
## lm(formula = y ~ x1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.89495 -0.66874 -0.07785  0.59221  2.45560
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.1124     0.2307   9.155 8.27e-15 ***
## x1            1.9759     0.3963   4.986 2.66e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.055 on 98 degrees of freedom
## Multiple R-squared:  0.2024, Adjusted R-squared:  0.1942
## F-statistic: 24.86 on 1 and 98 DF,  p-value: 2.661e-06
```

This model has moderately poor fit, much like the previous model. However, the estimated coefficient (2.11) for x1 is closer to the true value (2.0) and the p-value is below an alpha level of 0.05 so we can reject the null hypothesis.

(e) Now fit a least squares regression to predict y using only x2. Comment on your results. Can you reject the null hypothesis $H_0 : \beta_2 = 0$?

```
lm.fit_x2 = lm(y ~ x2)
summary(lm.fit_x2)
```

```
##
## Call:
## lm(formula = y ~ x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.62687 -0.75156 -0.03598  0.72383  2.44890
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.3899     0.1949   12.26  < 2e-16 ***
## x2            2.8996     0.6330    4.58 1.37e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 1.072 on 98 degrees of freedom
## Multiple R-squared:  0.1763, Adjusted R-squared:  0.1679
## F-statistic: 20.98 on 1 and 98 DF,  p-value: 1.366e-05
```

This model has worse fit compared to the previous two. However, the estimated coefficient (2.89) for x2 is not close to the true value (0.3), however, the p-value is below an alpha level of 0.05 so we can reject the null hypothesis.

(f) Do the results obtained in (c)-(e) contradict each other? Explain your answer.

Yes. Using the full model in part (c) we see that none of the coefficients are significant. However, when used as predictors in isolation, each of the predictors is deemed significant. The coefficients for each model, except for the x1 model and all estimated intercepts, are very different than the true values.

(g) Now suppose we obtain one additional observation, which was unfortunately mismeasured.Refit the linear models from (c) to (e) using this new data. What effect does this new observation have on the each of the models? In each model, is this observation an outlier? A high-leverage point? Both? Explain your answers

```
x1=c(x1, 0.1)
x2=c(x2, 0.8)
y=c(y,6)
```

```
lm.fit_cor = lm(y ~ x1 + x2)
summary(lm.fit_cor)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.73348 -0.69318 -0.05263  0.66385  2.30619
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.2267     0.2314   9.624 7.91e-16 ***
## x1            0.5394     0.5922   0.911  0.36458
## x2            2.5146     0.8977   2.801  0.00614 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.075 on 98 degrees of freedom
## Multiple R-squared:  0.2188, Adjusted R-squared:  0.2029
## F-statistic: 13.72 on 2 and 98 DF,  p-value: 5.564e-06
```

```
lm.fit_x1 = lm(y ~ x1)
summary(lm.fit_x1)
```

```
##
## Call:
## lm(formula = y ~ x1)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8897 -0.6556 -0.0909  0.5682  3.5665
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.2569     0.2390   9.445 1.78e-15 ***
## x1            1.7657     0.4124   4.282 4.29e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.111 on 99 degrees of freedom
## Multiple R-squared:  0.1562, Adjusted R-squared:  0.1477
## F-statistic: 18.33 on 1 and 99 DF,  p-value: 4.295e-05
```

```
lm.fit_x2 = lm(y ~ x2)
summary(lm.fit_x2)
```

```
##
## Call:
## lm(formula = y ~ x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.64729 -0.71021 -0.06899  0.72699  2.38074
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.3451     0.1912  12.264  < 2e-16 ***
## x2            3.1190     0.6040   5.164 1.25e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.074 on 99 degrees of freedom
## Multiple R-squared:  0.2122, Adjusted R-squared:  0.2042
## F-statistic: 26.66 on 1 and 99 DF,  p-value: 1.253e-06
```

15. This problem involves the Boston data set, which we saw in the lab for this chapter. We will now try to predict per capita crime rate using the other variables in this data set. In other words, per capita crime rate is the response, and the other variables are the predictors.

(a) For each predictor, fit a simple linear regression model to predict the response. Describe your results. In which of the models is there a statistically significant association between the predictor and the response? Create some plots to back up your assertions.

```
#Load libraries
library(MASS)
boston = Boston
```

```
#Determine which models with individual predictors have p-values less than 0.05 and store model diagnos
n = length(Boston) - 1
predictor.name = c()
predictor.pvalue = c()
predictor.fstat = c()
predictor.adj.r.squared = c()
```

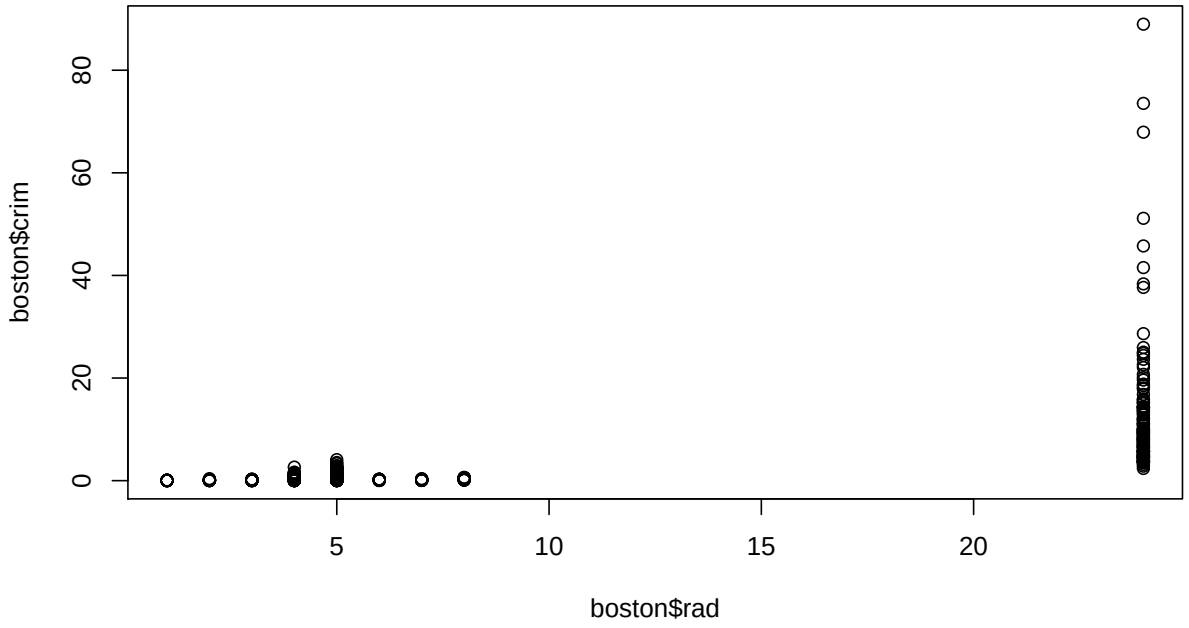
```
crim = boston$crim
for (i in 2:n){
  modified = data.frame(crim, boston[,i])
  colnames(modified) = c("crim", colnames(boston)[i])
  model = lm(crim ~ ., data = modified)
  if (summary(model)$coefficients[2,4] < 0.05){
    predictor.name = c(predictor.name, colnames(modified)[[2]])
    predictor.pvalue = c(predictor.pvalue, summary(model)$coefficients[2,4] )
    predictor.fstat = c(predictor.fstat, summary(model)$fstatistic[[1]])
    predictor.adj.r.squared = c(predictor.adj.r.squared, summary(model)$adj.r.squared)
  }
}

table = data.frame(predictor.name,predictor.pvalue,predictor.fstat,predictor.adj.r.squared)
table
```

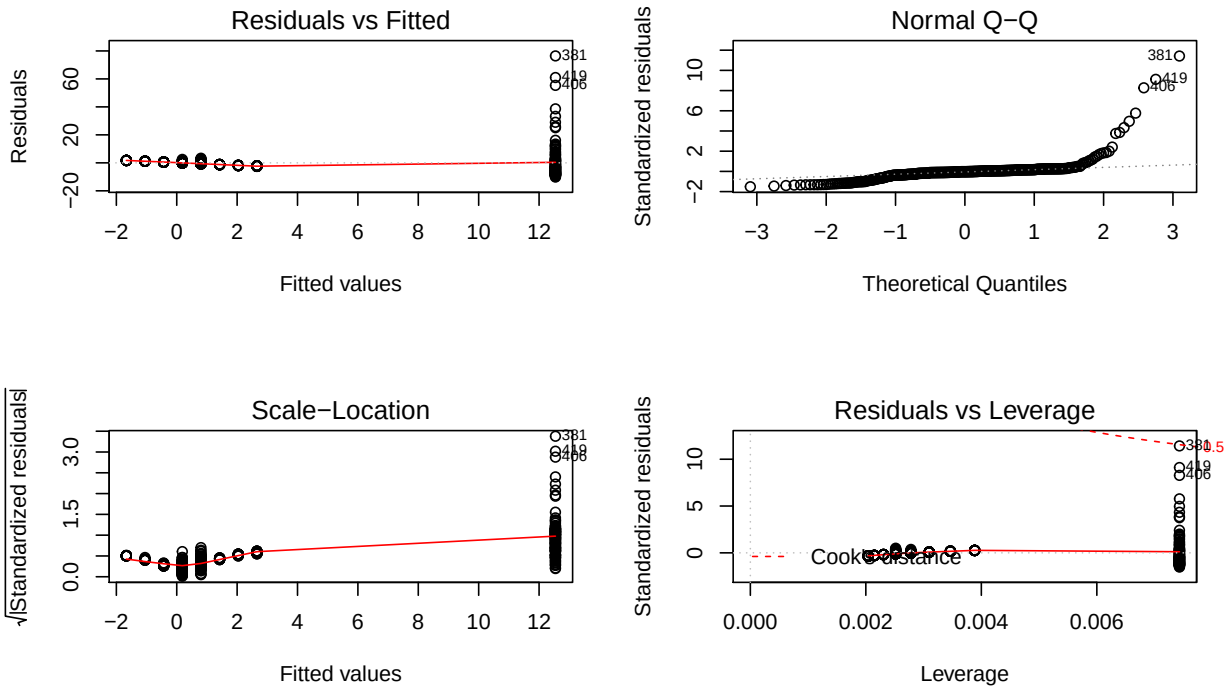
```
##    predictor.name predictor.pvalue predictor.fstat predictor.adj.r.squared
## 1              zn     5.506472e-06        21.10278              0.03828352
## 2           indus     1.450349e-21        99.81704              0.16365394
## 3             nox     3.751739e-23       108.55533              0.17558468
## 4              rm     6.346703e-07        25.45020              0.04618036
## 5             age     2.854869e-16        71.61940              0.12268419
## 6             dis     8.519949e-19        84.88781              0.14245126
## 7             rad     2.693844e-56       323.93517              0.39004886
## 8             tax     2.357127e-47       259.19029              0.33830395
## 9         ptratio     2.942922e-11        46.25945              0.08225111
## 10          black     2.487274e-19        87.73976              0.14658431
## 11          lstat     2.654277e-27       132.03512              0.20601869
```

We observe that the 11 predictors above have statistically significant relationships with per capita crime rate by town. Index of accessibility to radial highways and full-value property-tax rate per $10,000 also have moderately high adjusted R squares, providing more evidence of their significance.

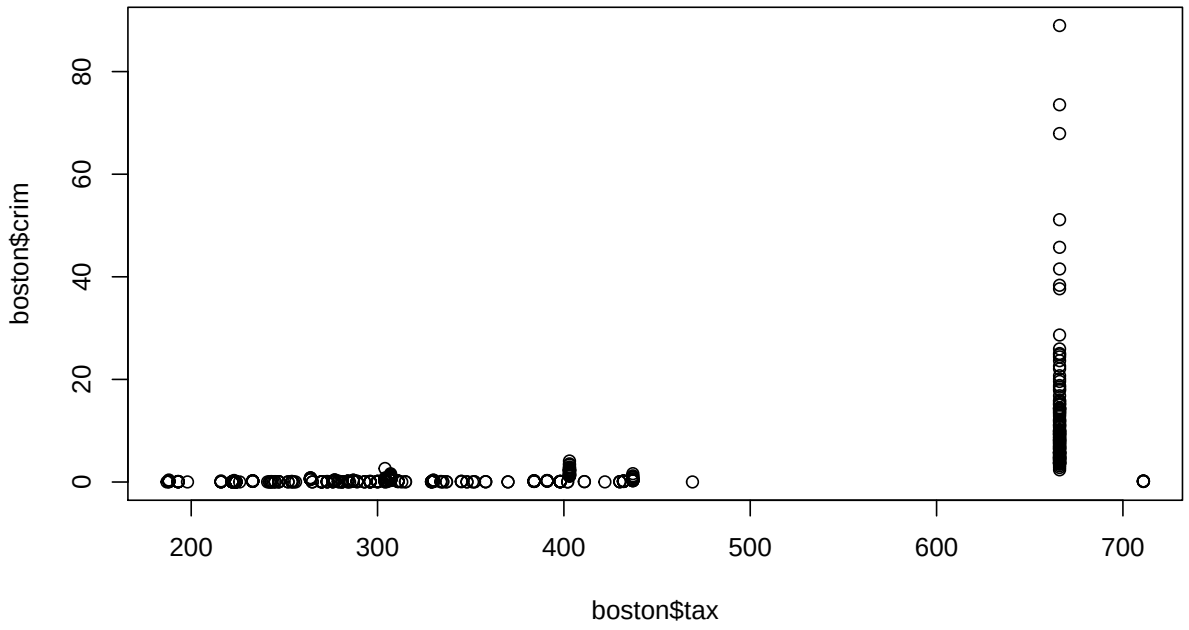
```
plot(boston$rad, boston$crim)
```



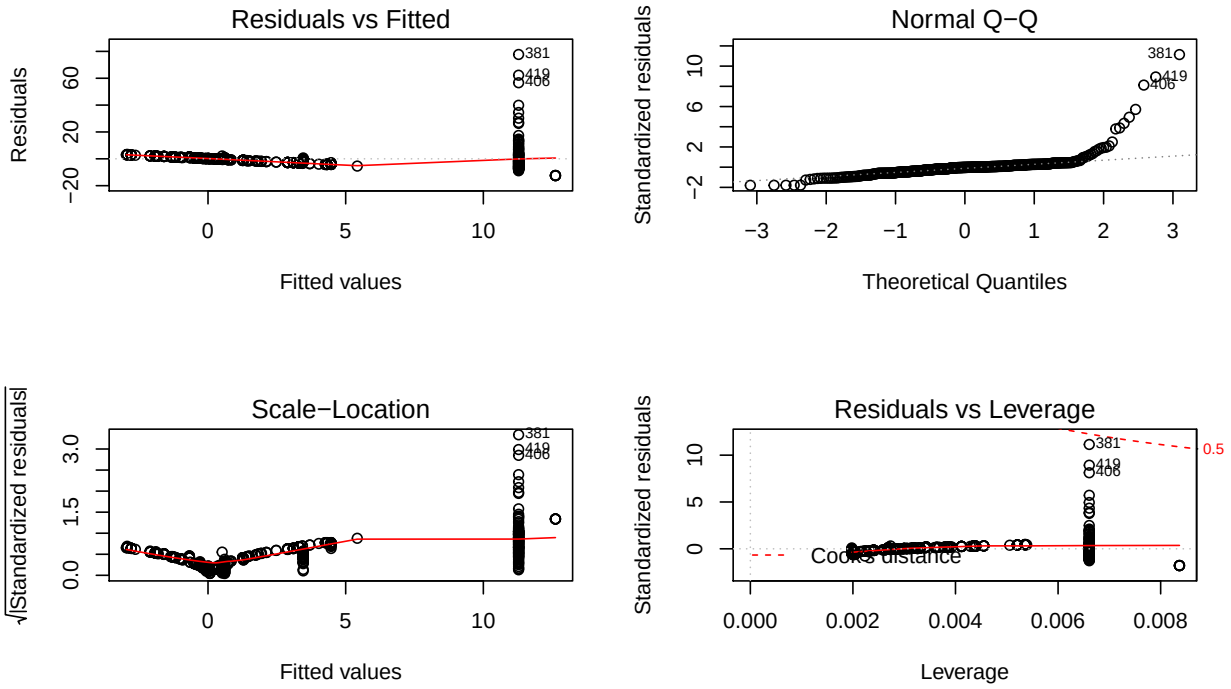

```
par(mfrow = c(2,2))
lm.fit.rad = lm(crim ~ rad, data = boston)
plot(lm.fit.rad)
```



```
plot(boston$tax, boston$crim)
```




```
par(mfrow = c(2,2))
lm.fit.tax = lm(crim ~ tax, data = boston)
plot(lm.fit.tax)
```



(b) Fit a multiple regression model to predict the response using all of the predictors. Describe your results. For which predictors can we reject the null hypothesis $H_0 : \beta_j = 0$?

```
lm.fit.boston = lm(crim ~ ., data = boston)
summary(lm.fit.boston)
```

```
##
## Call:
## lm(formula = crim ~ ., data = boston)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.924 -2.120 -0.353  1.019 75.051
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  17.033228   7.234903   2.354 0.018949 *
## zn            0.044855   0.018734   2.394 0.017025 *
## indus        -0.063855   0.083407  -0.766 0.444294
## chas         -0.749134   1.180147  -0.635 0.525867
## nox         -10.313535   5.275536  -1.955 0.051152 .
## rm            0.430131   0.612830   0.702 0.483089
## age           0.001452   0.017925   0.081 0.935488
## dis          -0.987176   0.281817  -3.503 0.000502 ***
## rad           0.588209   0.088049   6.680 6.46e-11 ***
## tax          -0.003780   0.005156  -0.733 0.463793
## ptratio      -0.271081   0.186450  -1.454 0.146611
## black        -0.007538   0.003673  -2.052 0.040702 *
```

```
## lstat         0.126211   0.075725   1.667 0.096208 .
## medv         -0.198887   0.060516  -3.287 0.001087 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.439 on 492 degrees of freedom
## Multiple R-squared:  0.454,  Adjusted R-squared:  0.4396
## F-statistic: 31.47 on 13 and 492 DF,  p-value: < 2.2e-16
```

An F-statistic of 31.47 is well above 1 so we can conlude there is some relationship between the predictors and the response. The zoning, distance to employment centers, accessibility to highways, proportion of blacks, and median home value are all significant predictors at an alpha level of 0.05.