# Statistical Learning HW 1

Christian Conroy

February 10, 2018

## 1. (3 pts) Textbook #3.7.3

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 ??^0 = 50, ??^1 = 20, ??^2 = 0.07, ??^3 = 35, ??^4 = 0.01, ??^5 = ???10.

(a) Which answer is correct, and why?

i. For a fixed value of IQ and GPA, males earn more on average than females.

ii. For a fixed value of IQ and GPA, females earn more on average than males.

iii. For a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough.

iv. For a fixed value of IQ and GPA, females earn more on average than males provided that the GPA is high enough.

iii is the correct answer. For a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough. When IQ and GPA are fixed, then the difference between males and females is demonstrated by B3 + B5. While B3 indicates that females earn more on average than males at a theoretical GPA of 0, the negative coefficient on B5 indicates that females on average will only earn more on average than males at lower GPAs but that males will earn more on average than females as the GPA surpassesapproximately 3.5.

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

```
50 + (20*4) + (.07*110) + (35*1) + (.01*4*110) + (-10*1*4)

## [1] 137
```

The predicted salary of a female with IQ of 110 and a GPA of 4.0 is $137,000.

(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. While the coefficient may indicate that the magnitude of the GPA/IQ interaction term is small, we would need to examine the hypothesis test with a null that B4 = 0 to determine whether there is evidence of an interaction effect. Specifically, we can examine whether the p-value is lower than 0.05 (95% confidence level) or potentially .10 (90% confidence level) depending on what threshold we apply. This is the same as observing whether the t statistic falls outside of the critical value in absolute value terms. This is all to say that we can only determine if there is evidence of an interaction effect by assessing whether the coefficient is statistically significant.

## 2. (3 pts) Textbook #3.7.8

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.

```
data("Auto")
head(Auto)

##   mpg cylinders displacement horsepower weight acceleration year origin
## 1  18         8          307        130   3504         12.0   70      1
## 2  15         8          350        165   3693         11.5   70      1
## 3  18         8          318        150   3436         11.0   70      1
## 4  16         8          304        150   3433         12.0   70      1
## 5  17         8          302        140   3449         10.5   70      1
## 6  15         8          429        198   4341         10.0   70      1
##                        name
## 1 chevrolet chevelle malibu
## 2         buick skylark 320
## 3        plymouth satellite
## 4             amc rebel sst
## 5               ford torino
## 6          ford galaxie 500

reg1 <- lm(mpg ~ horsepower, data = Auto)
summary(reg1)

## 
## Call:
## lm(formula = mpg ~ horsepower, data = Auto)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -13.571  -3.259  -0.344   2.763  16.924 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 39.93586    0.71750    55.7   <2e-16 ***
## horsepower  -0.15784    0.00645   -24.5   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 4.91 on 390 degrees of freedom
## Multiple R-squared:  0.606,  Adjusted R-squared:  0.605 
## F-statistic:  600 on 1 and 390 DF,  p-value: <2e-16
```

Comment on the output. For example: 122 3. Linear Regression i. Is there a relationship between the predictor and the response?

According to the results of our regression analysis, there appears to be a relationship between the predictor and the response. The coefficient of -0.15785 indicates a negative relationship between horsepower and mpg and the low p-value indicates that the coefficient is statistically significant at the 99% confidence level. However, it is important to note that this is only a univariate regression model and it therefore may suffer from endogeneity. We can conclude a correlation but we cannot conclude causation.

ii. How strong is the relationship between the predictor and the response?

The magnitude is fairly small, indicating that a unit increase in horsepower only slightly impacts the mpg of the vehicle. It would take much larger changes in the horsepower to have any larger magnitude effect. Again, according to the model, this is a statistically significant effect.

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

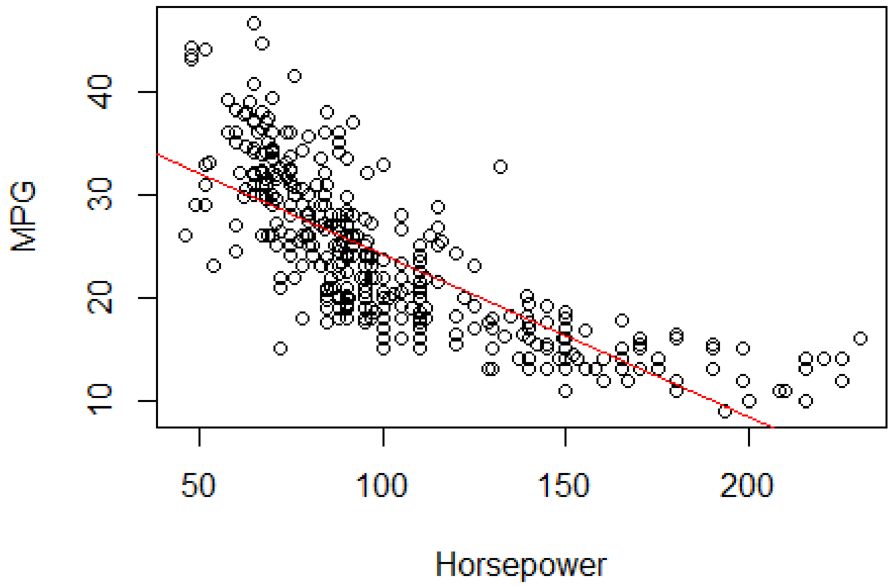
The relationship between the predictor and the response variable is negative.

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

```
confint(reg1)

##               2.5 % 97.5 %
## (Intercept) 38.525 41.347
## horsepower  -0.171 -0.145

predict(reg1, data.frame(horsepower=98), interval="predict")

##    fit  lwr  upr
## 1 24.5 14.8 34.1
```

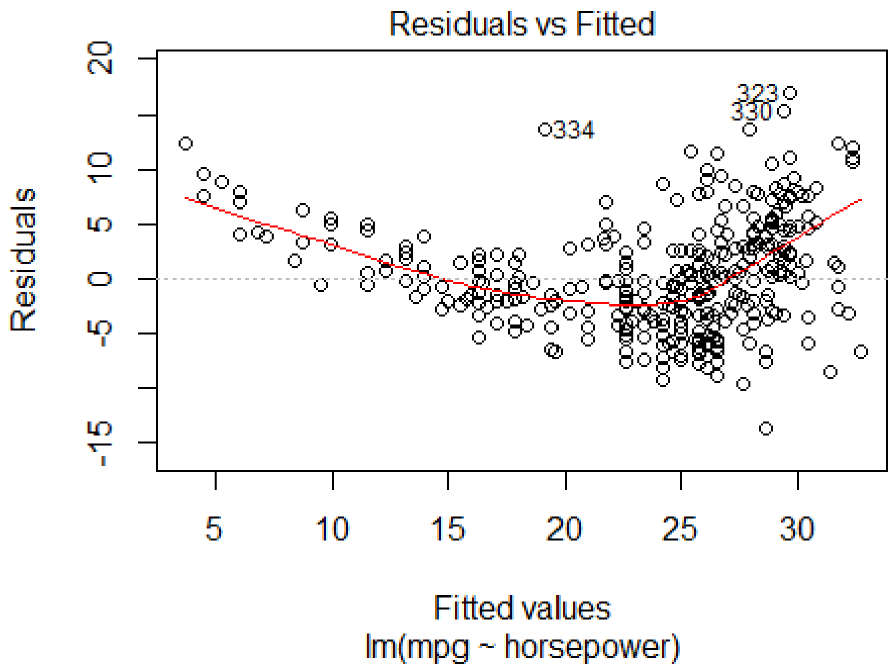
The predicted mpg associated with a horsepower of 98 is 24.5 with a prediction interval of 14.8 to 34.1. The 95% confidence interval for the horsepower coefificent is -0.171 to -0.145.

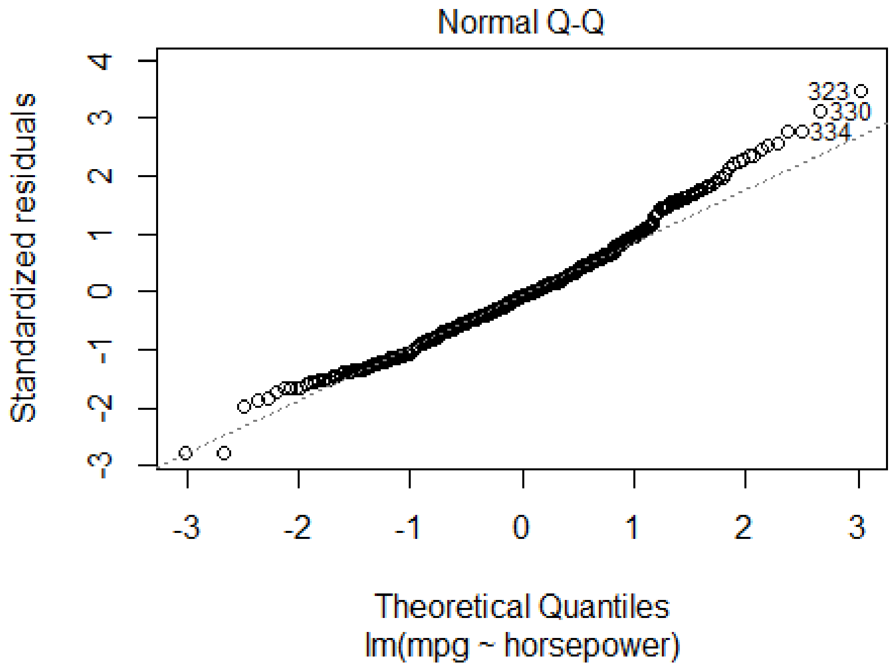
(b) Plot the response and the predictor. Use the abline() function to display the least squares regression line.

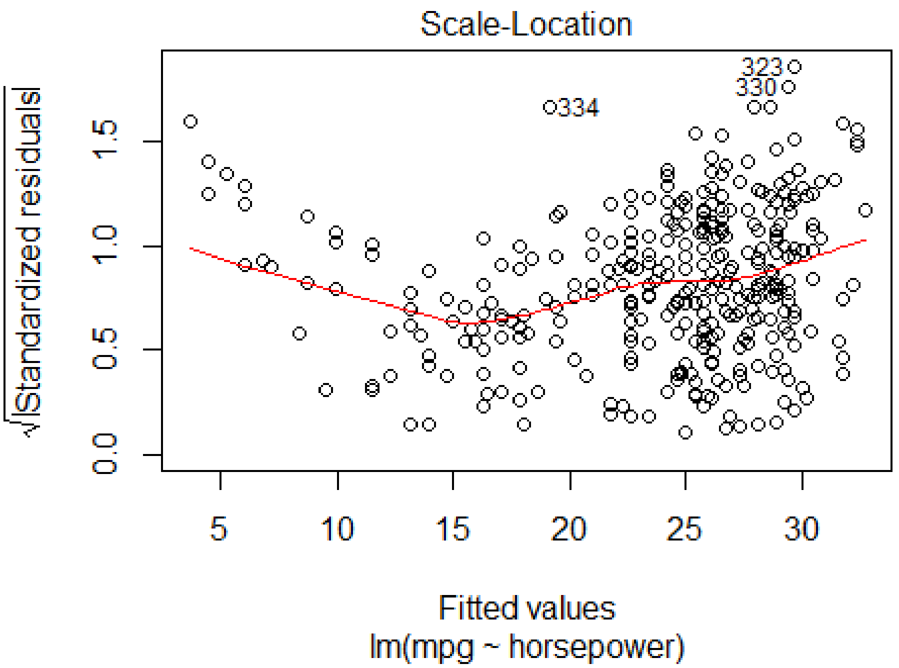


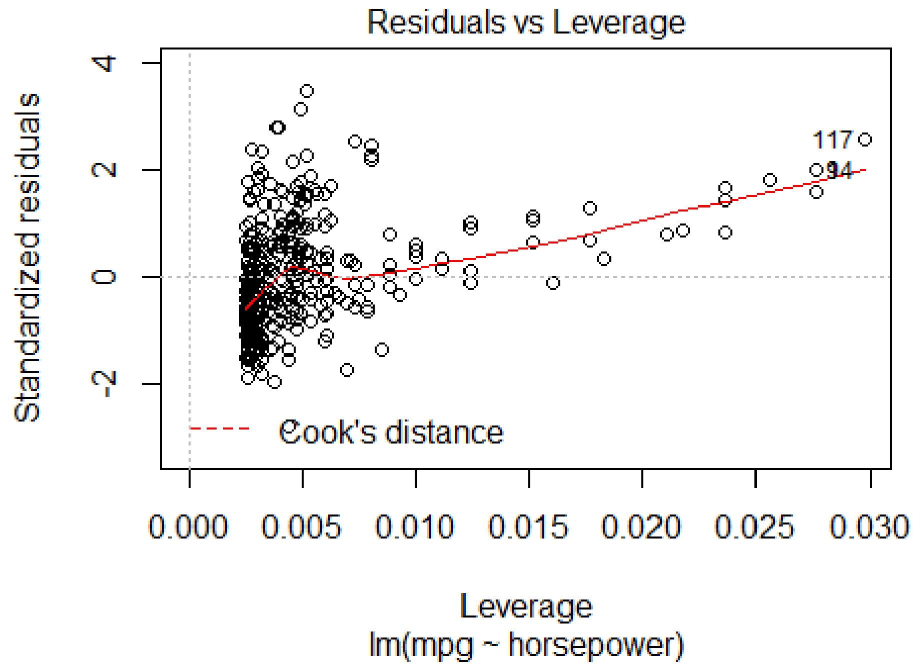
(c) Use the plot() function to produce diagnostic plots of the least squares regression fit. Comment on any problems you see with the fit.

```
plot(reg1)
```


Residuals vs Fitted
Residuals
20
10
5
0
-5
-15
5
10
15
20
25
30
Fitted values
lm(mpg ~ horsepower)
334
323
330


Normal Q-Q
Standardized residuals
4
3
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(mpg ~ horsepower)
323
330
334





The residuals vs fitted plot tells us that we might want to exercise caution in using a linear model. It appears that there might be a non-linear relationship between the predictor and outcome variable. While the qq-plot shows that the residuals slightly deviate from normal distribution, the deviation does not appear to be significant enough to merit major concern.

The scale-location plot may lead us to believe that there is a small heteroskedasticity risk, and so it might be worth it to use robust standard errors just in case (this is always a good practice anyways). You cannot see a Cook's distance line on the final residuals vs leverage plot because all cases are within the Cook's distance lines. Therefore, even if some values might appear to look like outliers, they do not seem to have a major effect on the regression output.

## 3. (3 pts) Textbook #3.7.10abc

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

```
data("Carseats")
head(Carseats)
```

```
##   Sales CompPrice Income Advertising Population Price ShelveLoc Age
## 1  9.50       138     73          11        276   120       Bad  42
## 2 11.22       111     48          16        260    83      Good  65
## 3 10.06       113     35          10        269    80    Medium  59
## 4  7.40       117    100           4        466    97    Medium  55
## 5  4.15       141     64           3        340   128       Bad  38
## 6 10.81       124    113          13        501    72       Bad  78
##   Education Urban  US
## 1        17   Yes Yes
## 2        10   Yes Yes
## 3        12   Yes Yes
## 4        14   Yes Yes
## 5        13   Yes  No
## 6        16    No Yes
```

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

```
reg2 <- lm(Sales ~ Price + Urban + US, data = Carseats)
summary(reg2)
```

```
## 
## Call:
## lm(formula = Sales ~ Price + Urban + US, data = Carseats)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -6.921 -1.622 -0.056  1.579  7.058 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 13.04347    0.65101   20.04  < 2e-16 ***
## Price       -0.05446    0.00524  -10.39  < 2e-16 ***
## UrbanYes    -0.02192    0.27165   -0.08     0.94    
## USYes        1.20057    0.25904    4.63  4.9e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
```

```
## Residual standard error: 2.47 on 396 degrees of freedom
## Multiple R-squared:  0.239,  Adjusted R-squared:  0.234
## F-statistic: 41.5 on 3 and 396 DF,  p-value: <2e-16
```

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

B0: The average sales for stores in rural locations outside of the US not taking price into account is 13.04 thousand dollars. This theoretically means assuming price is 0, but since price will not be 0, the main purpose of the intercept here is to anchor the regression line in the right place. B1: Holding all other variables in the model constant, a one unit increase in the price is associated with a decrease of .05 thousand ($50) in sales. B2. Holding all other variables in the model constant, a store located in an urban location will have .02 thousand ($20) less in sales on average compared to a store in a rural location. B3. Holding all other variables in the model constant, a store located in the US will have 1.26 thousand ($1,260) less in sales on average compared to a store in outside of the US.

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

$y_hat = \beta_0 + \beta_1 Price + \beta_2 Urban(Yes = 1) + \beta_3 US(Yes = 1)$

## 3. (3 pts) Textbook #3.7.15ab

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.

```
data("Boston")
head(Boston)

##      crim zn indus chas   nox   rm  age  dis rad tax ptratio black lstat
## 1 0.00632 18  2.31    0 0.538 6.58 65.2 4.09   1 296    15.3   397  4.98
## 2 0.02731  0  7.07    0 0.469 6.42 78.9 4.97   2 242    17.8   397  9.14
## 3 0.02729  0  7.07    0 0.469 7.18 61.1 4.97   2 242    17.8   393  4.03
## 4 0.03237  0  2.18    0 0.458 7.00 45.8 6.06   3 222    18.7   395  2.94
## 5 0.06905  0  2.18    0 0.458 7.15 54.2 6.06   3 222    18.7   397  5.33
## 6 0.02985  0  2.18    0 0.458 6.43 58.7 6.06   3 222    18.7   394  5.21
##   medv
## 1 24.0
## 2 21.6
## 3 34.7
## 4 33.4
## 5 36.2
## 6 28.7
```

(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?

```r
varlist <- names(Boston)[-1]

lm.test <- vector("list", length(varlist))

for(i in seq_along(varlist)){
    lm.test[[i]] <- lm(reformulate(varlist[i], "crim"), data = Boston)
}

lapply(lm.test, summary)
```

```
## [[1]]
##
## Call:
## lm(formula = reformulate(varlist[i], "crim"), data = Boston)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -4.43  -4.22  -2.62   1.25  84.52
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.4537     0.4172   10.67  < 2e-16 ***
## zn           -0.0739     0.0161   -4.59  5.5e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.44 on 504 degrees of freedom
## Multiple R-squared:  0.0402, Adjusted R-squared:  0.0383
## F-statistic: 21.1 on 1 and 504 DF,  p-value: 5.51e-06
##
##
## [[2]]
##
## Call:
## lm(formula = reformulate(varlist[i], "crim"), data = Boston)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -11.97  -2.70  -0.74   0.71  81.81
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -2.064      0.667   -3.09   0.0021 **
## indus          0.510      0.051    9.99   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.87 on 504 degrees of freedom
## Multiple R-squared:  0.165,  Adjusted R-squared:  0.164
## F-statistic: 99.8 on 1 and 504 DF,  p-value: <2e-16
```

```
##
##
## [[3]]
##
## Call:
## lm(formula = reformulate(varlist[i], "crim"), data = Boston)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -3.74  -3.66  -3.44   0.02  85.23
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.744      0.396    9.45   <2e-16 ***
## chas          -1.893      1.506   -1.26     0.21
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.6 on 504 degrees of freedom
## Multiple R-squared:  0.00312,    Adjusted R-squared:  0.00115
## F-statistic: 1.58 on 1 and 504 DF,  p-value: 0.209
##
##
## [[4]]
##
## Call:
## lm(formula = reformulate(varlist[i], "crim"), data = Boston)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -12.37  -2.74  -0.97   0.56  81.73
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -13.7        1.7   -8.07  5.1e-15 ***
## nox             31.2        3.0   10.42  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.81 on 504 degrees of freedom
## Multiple R-squared:  0.177,  Adjusted R-squared:  0.176
## F-statistic:  109 on 1 and 504 DF,  p-value: <2e-16
##
##
## [[5]]
##
## Call:
## lm(formula = reformulate(varlist[i], "crim"), data = Boston)
##
## Residuals:
```

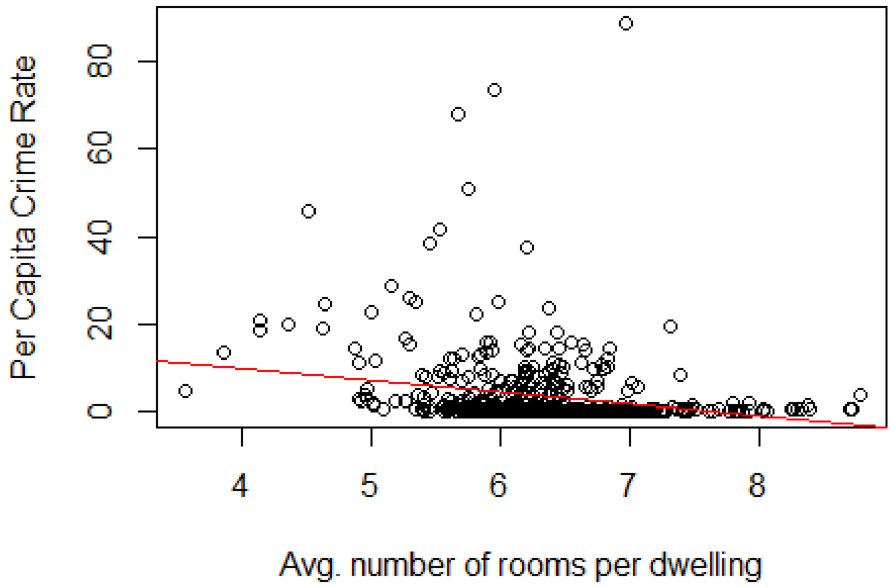
```
##    Min     1Q Median     3Q    Max 
##  -6.60  -3.95  -2.65   0.99  87.20 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   20.482      3.364    6.09  2.3e-09 ***
## rm            -2.684      0.532   -5.04  6.3e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 8.4 on 504 degrees of freedom
## Multiple R-squared:  0.0481, Adjusted R-squared:  0.0462 
## F-statistic: 25.5 on 1 and 504 DF,  p-value: 6.35e-07
## 
## 
## [[6]]
## 
## Call:
## lm(formula = reformulate(varlist[i], "crim"), data = Boston)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
##  -6.79  -4.26  -1.23   1.53  82.85 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  -3.7779     0.9440   -4.00  7.2e-05 ***
## age           0.1078     0.0127    8.46  2.9e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 8.06 on 504 degrees of freedom
## Multiple R-squared:  0.124,  Adjusted R-squared:  0.123 
## F-statistic: 71.6 on 1 and 504 DF,  p-value: 2.85e-16
## 
## 
## [[7]]
## 
## Call:
## lm(formula = reformulate(varlist[i], "crim"), data = Boston)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
##  -6.71  -4.13  -1.53   1.52  81.67 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)    9.499      0.730   13.01   <2e-16 ***
## dis           -1.551      0.168   -9.21   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.97 on 504 degrees of freedom
## Multiple R-squared:  0.144,  Adjusted R-squared:  0.142
## F-statistic: 84.9 on 1 and 504 DF,  p-value: <2e-16
##
##
## [[8]]
##
## Call:
## lm(formula = reformulate(varlist[i], "crim"), data = Boston)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -10.16  -1.38  -0.14   0.66  76.43
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.2872     0.4435   -5.16  3.6e-07 ***
## rad           0.6179     0.0343   18.00  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.72 on 504 degrees of freedom
## Multiple R-squared:  0.391,  Adjusted R-squared:  0.39
## F-statistic:  324 on 1 and 504 DF,  p-value: <2e-16
##
##
## [[9]]
##
## Call:
## lm(formula = reformulate(varlist[i], "crim"), data = Boston)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -12.51  -2.74  -0.19   1.07  77.70
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -8.52837    0.81581   -10.4   <2e-16 ***
## tax          0.02974    0.00185    16.1   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7 on 504 degrees of freedom
## Multiple R-squared:  0.34,   Adjusted R-squared:  0.338
## F-statistic:  259 on 1 and 504 DF,  p-value: <2e-16
##
##
## [[10]]
```

```
##
## Call:
## lm(formula = reformulate(varlist[i], "crim"), data = Boston)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -7.65  -3.99  -1.91   1.82  83.35
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -17.647      3.147   -5.61  3.4e-08 ***
## ptratio        1.152      0.169    6.80  2.9e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.24 on 504 degrees of freedom
## Multiple R-squared:  0.0841, Adjusted R-squared:  0.0823
## F-statistic: 46.3 on 1 and 504 DF,  p-value: 2.94e-11
##
##
## [[11]]
##
## Call:
## lm(formula = reformulate(varlist[i], "crim"), data = Boston)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -13.76  -2.30  -2.09  -1.30  86.82
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16.55353    1.42590   11.61   <2e-16 ***
## black       -0.03628    0.00387   -9.37   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.95 on 504 degrees of freedom
## Multiple R-squared:  0.148,  Adjusted R-squared:  0.147
## F-statistic: 87.7 on 1 and 504 DF,  p-value: <2e-16
##
##
## [[12]]
##
## Call:
## lm(formula = reformulate(varlist[i], "crim"), data = Boston)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -13.93  -2.82  -0.66   1.08  82.86
##
```

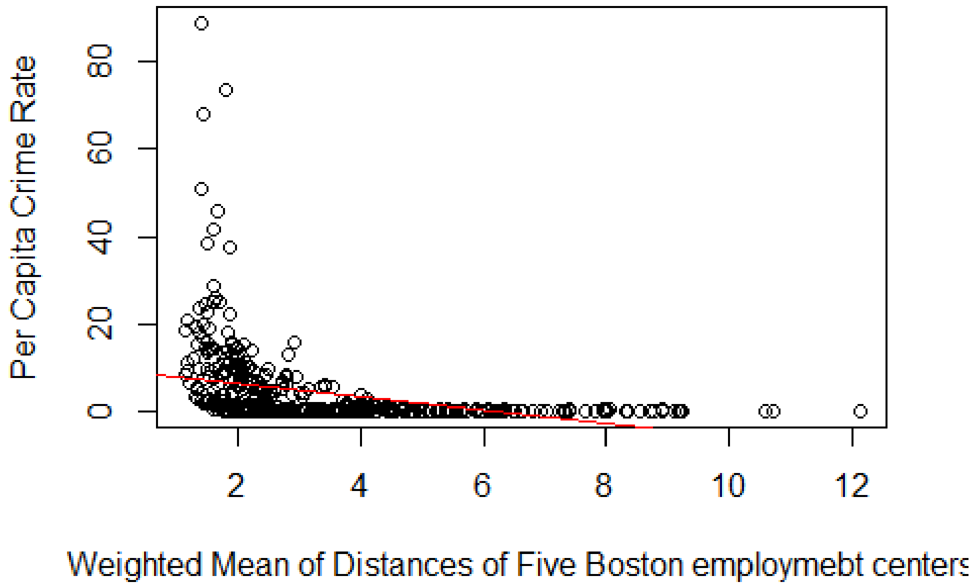
```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  -3.3305     0.6938    -4.8  2.1e-06 ***
## lstat         0.5488     0.0478    11.5  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 7.66 on 504 degrees of freedom
## Multiple R-squared:  0.208,  Adjusted R-squared:  0.206 
## F-statistic:  132 on 1 and 504 DF,  p-value: <2e-16
## 
## 
## [[13]]
## 
## Call:
## lm(formula = reformulate(varlist[i], "crim"), data = Boston)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
##  -9.07  -4.02  -2.34   1.30  80.96 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  11.7965     0.9342   12.63   <2e-16 ***
## medv         -0.3632     0.0384   -9.46   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 7.93 on 504 degrees of freedom
## Multiple R-squared:  0.151,  Adjusted R-squared:  0.149 
## F-statistic: 89.5 on 1 and 504 DF,  p-value: <2e-16
```

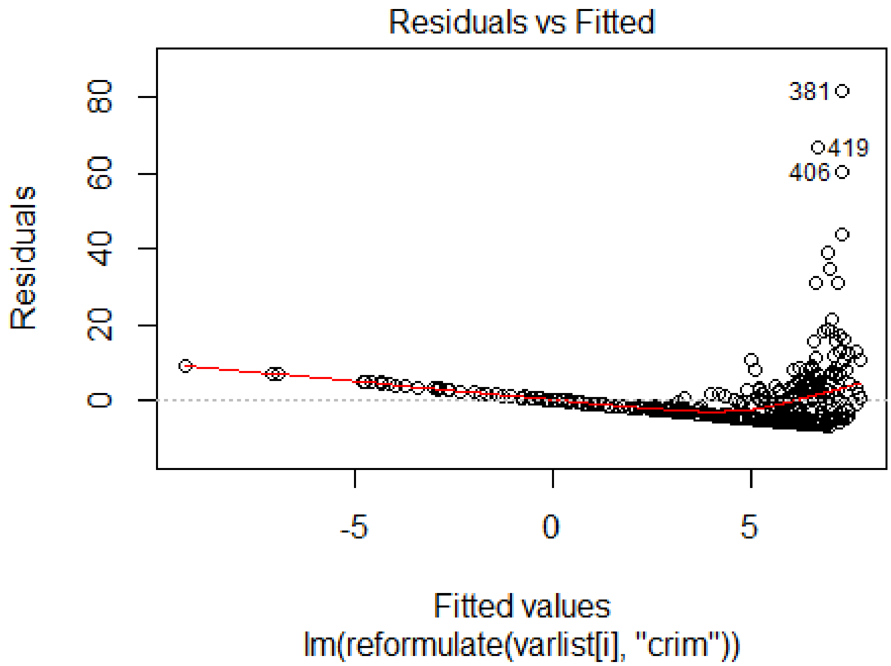
If we run a simple linear regression of crim on each respective variable in the model as covariates, we find a statistically significant effect for every covariate except for chas, the Charles River dummy variable signifying if the tract bounds the river.

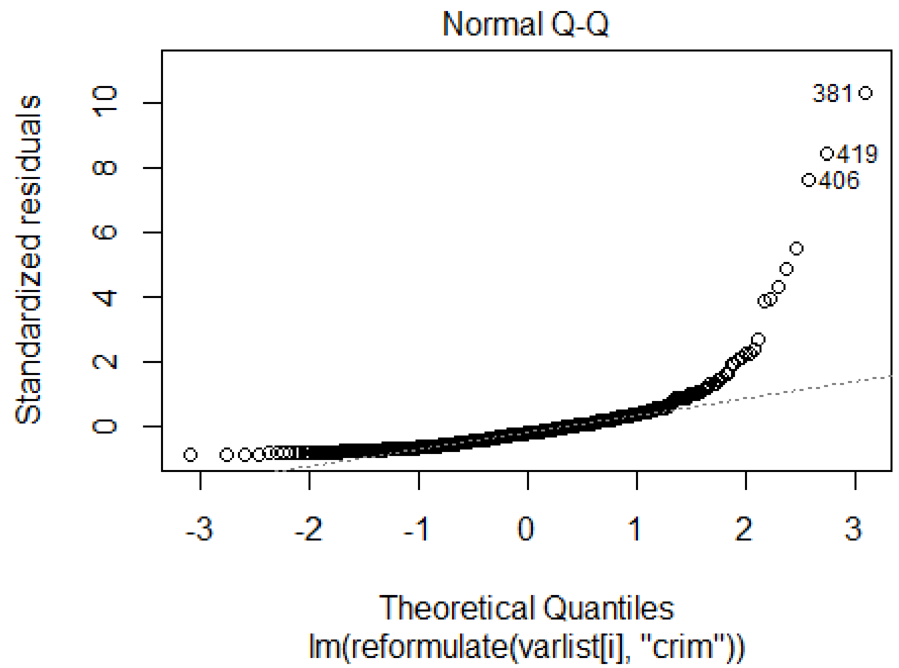
Create some plots to back up your assertions.

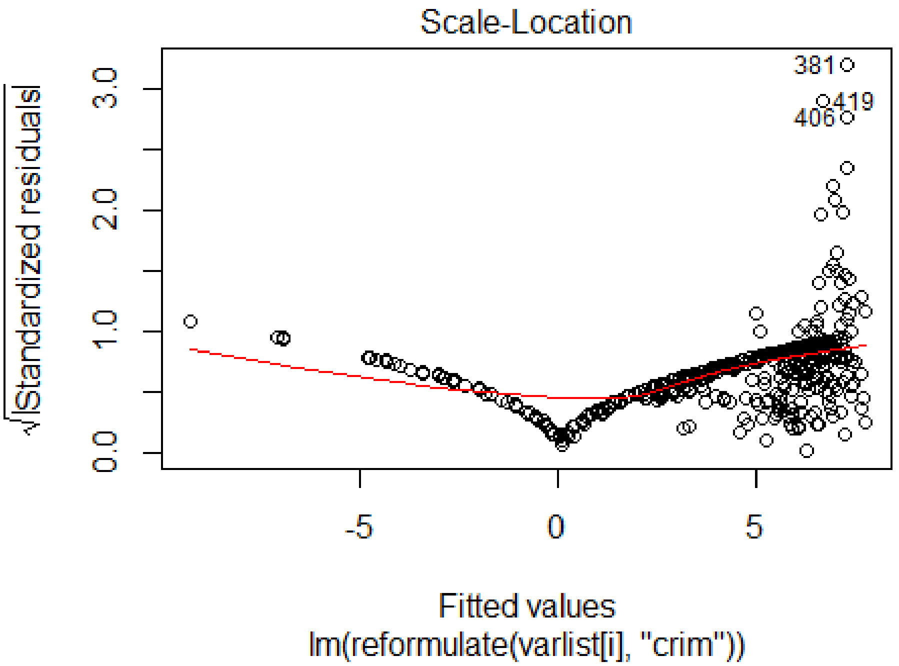


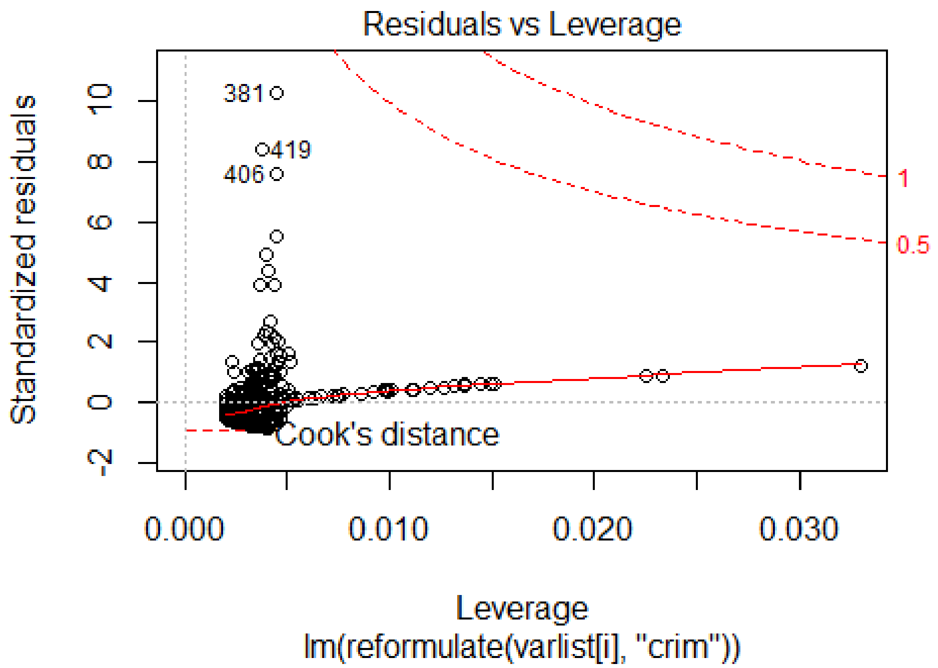
As the above plot for the relationship between, most of the univariate regression models show statistically significant relationships that are spurious and misleading. The conclusions above therefore do little to provide evidence related to any relationship to per capita crime rates.


Per Capita Crime Rate
0
20
40
60
80
2
4
6
8
10
12
Weighted Mean of Distances of Five Boston employmebt centers


Residuals vs Fitted
Residuals
0
20
40
60
80
381
419
406
-5
0
5
Fitted values
lm(reformulate(varlist[i], "crim"))


Normal Q-Q
Standardized residuals
0
2
4
6
8
10
381
419
406
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(reformulate(varlist[i], "crim"))


Scale-Location
√|Standardized residuals|
0.0
1.0
2.0
3.0
381
419
406
-5
0
5
Fitted values
lm(reformulate(varlist[i], "crim"))



As the above plot shows, not only is it insufficient to conclude that a statistically significant relationship between the dis variable and per capita crime rates is enough, but it is also wrong to conclude that a linear model is necessarily the best fit. Other plots like the residuals vs fitted also indicate as much.

(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 H0 : ??j = 0?

```
##
## Call:
## lm(formula = crim ~ zn + indus + chas + nox + rm + age + dis +
##     rad + tax + ptratio + black + lstat + medv, data = Boston)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -9.92  -2.12  -0.35   1.02  75.05
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  17.03323    7.23490    2.35   0.0189 *
## zn            0.04486    0.01873    2.39   0.0170 *
## indus        -0.06385    0.08341   -0.77   0.4443
## chas         -0.74913    1.18015   -0.63   0.5259
## nox         -10.31353    5.27554   -1.95   0.0512 .
## rm            0.43013    0.61283    0.70   0.4831
## age           0.00145    0.01793    0.08   0.9355
## dis          -0.98718    0.28182   -3.50   0.0005 ***
## rad           0.58821    0.08805    6.68  6.5e-11 ***
## tax          -0.00378    0.00516   -0.73   0.4638
## ptratio      -0.27108    0.18645   -1.45   0.1466
## black        -0.00754    0.00367   -2.05   0.0407 *
## lstat         0.12621    0.07572    1.67   0.0962 .
## medv         -0.19889    0.06052   -3.29   0.0011 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.44 on 492 degrees of freedom
## Multiple R-squared:  0.454,  Adjusted R-squared:  0.44
## F-statistic: 31.5 on 13 and 492 DF,  p-value: <2e-16
```

At the 95% level of confidence, we can reject the null hypothesis for zn, dis, rad, black, and medv. IF we extend to a 90% level of confidence, we'd also reject the null for lstat and nox.

## Extra#10

Consider the built-in data set cars (see problem 1). Fit a linear regression model to the data. What are the three observations with the largest standardized residuals (in magnitude)? What are their leverages? Where are the three observation with the largest leverage? What are their standardized residuals? Use the R function influence.lm.

```
data("cars")
head(cars)

##   speed dist
## 1     4    2
## 2     4   10
## 3     7    4
## 4     7   22
## 5     8   16
## 6     9   10

# Run the reg and get top three observations by standardized residual
reg4 <- lm(dist ~ speed, data = cars)
diagnostics <- lm.influence(reg4)
head(sort(diagnostics$wt.res, decreasing=TRUE), 3)

##   49   23   35
## 43.2 42.5 30.8

# Get leverage for three observations with highest standardized residual
diagnostics$hat[c(49, 23, 35)]

##     49     23     35
## 0.0740 0.0214 0.0249
```

```r
# Get top three observations by largest leverage
head(sort(diagnostics$hat, decreasing=TRUE), 3)

##      1      2     50
## 0.1149 0.1149 0.0873

# Get standardized residuals for three observationss with highest leverage.
diagnostics$wt.res[c(1, 2, 50)]

##     1     2    50
##  3.85 11.85  4.27
```

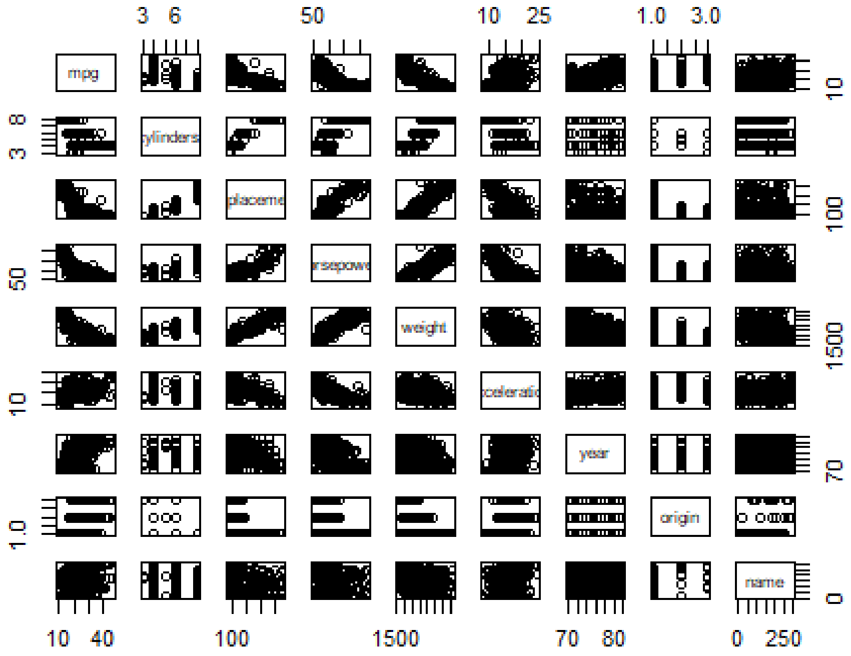
The three observations with the largest standardized residuals are observations 49, 23, and 35 with standardized residuals of 43.2, 42.5, and 30.8, respectively. Observation 49 has a leverage of 0.0740, obervation 23 a leverage of 0.0214, and observation 35 a leverage of 0.0249. The three observations with the largest leverage are observations 1, 2, and 50 with leverages of 0.1149, 0.1149, and 0.0873, respectively.The standardized residuals for observations 1, 2, and 50 are 3.85, 11.85, and 4.27, respectively.

## 6. (5 pts) Textbook #3.7.9

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

```r
data("Auto")
head(Auto)

##   mpg cylinders displacement horsepower weight acceleration year origin
## 1  18         8          307        130   3504         12.0   70      1
## 2  15         8          350        165   3693         11.5   70      1
## 3  18         8          318        150   3436         11.0   70      1
## 4  16         8          304        150   3433         12.0   70      1
## 5  17         8          302        140   3449         10.5   70      1
## 6  15         8          429        198   4341         10.0   70      1
##                        name
## 1 chevrolet chevelle malibu
## 2         buick skylark 320
## 3        plymouth satellite
## 4             amc rebel sst
## 5               ford torino
## 6          ford galaxie 500
```

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



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[,-9])

##                 mpg cylinders displacement horsepower weight acceleration
## mpg           1.000    -0.778       -0.805     -0.778 -0.832        0.423
## cylinders    -0.778     1.000        0.951      0.843  0.898       -0.505
## displacement -0.805     0.951        1.000      0.897  0.933       -0.544
## horsepower   -0.778     0.843        0.897      1.000  0.865       -0.689
## weight       -0.832     0.898        0.933      0.865  1.000       -0.417
## acceleration  0.423    -0.505       -0.544     -0.689 -0.417        1.000
## year          0.581    -0.346       -0.370     -0.416 -0.309        0.290
## origin        0.565    -0.569       -0.615     -0.455 -0.585        0.213
##                year origin
## mpg           0.581  0.565
## cylinders    -0.346 -0.569
## displacement -0.370 -0.615
## horsepower   -0.416 -0.455
## weight       -0.309 -0.585
## acceleration  0.290  0.213
## year          1.000  0.182
## origin        0.182  1.000
```

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

```
reg5 <- lm(mpg ~ cylinders + displacement + horsepower + weight +
acceleration + year + origin, data = Auto)
summary(reg5)
```

```
##
## Call:
## lm(formula = mpg ~ cylinders + displacement + horsepower + weight +
##     acceleration + year + origin, data = Auto)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.590 -2.157 -0.117  1.869 13.060
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.72e+01   4.64e+00   -3.71  0.00024 ***
## cylinders    -4.93e-01   3.23e-01   -1.53  0.12780
## displacement  1.99e-02   7.51e-03    2.65  0.00844 **
## horsepower   -1.70e-02   1.38e-02   -1.23  0.21963
## weight       -6.47e-03   6.52e-04   -9.93  < 2e-16 ***
## acceleration  8.06e-02   9.88e-02    0.82  0.41548
## year          7.51e-01   5.10e-02   14.73  < 2e-16 ***
## origin        1.43e+00   2.78e-01    5.13  4.7e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.33 on 384 degrees of freedom
## Multiple R-squared:  0.821,  Adjusted R-squared:  0.818
## F-statistic:  252 on 7 and 384 DF,  p-value: <2e-16
```

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

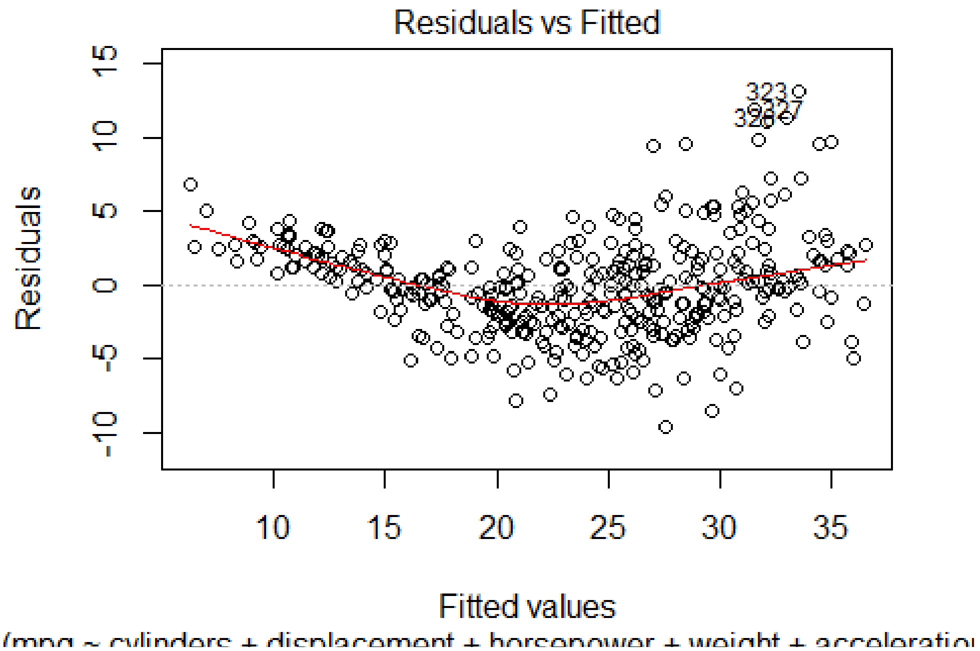
ii. Which predictors appear to have a statistically significant relationship to the response?

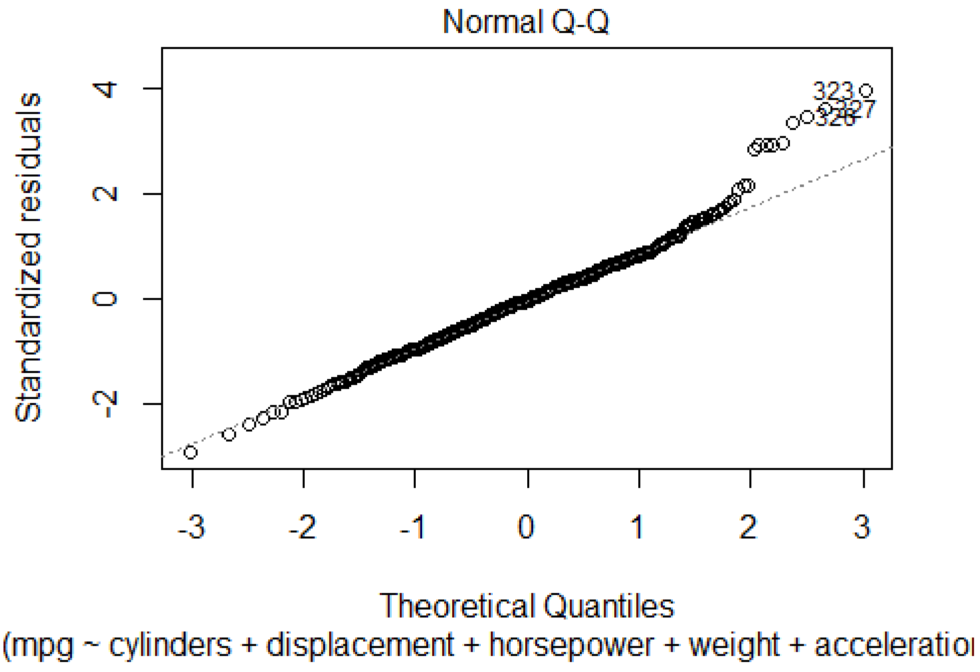
There appears to be statistically significant relationships between displacement, weight, year, and origin. The relationship for displacement and weight are negative and the relationship for year and origin are positive.

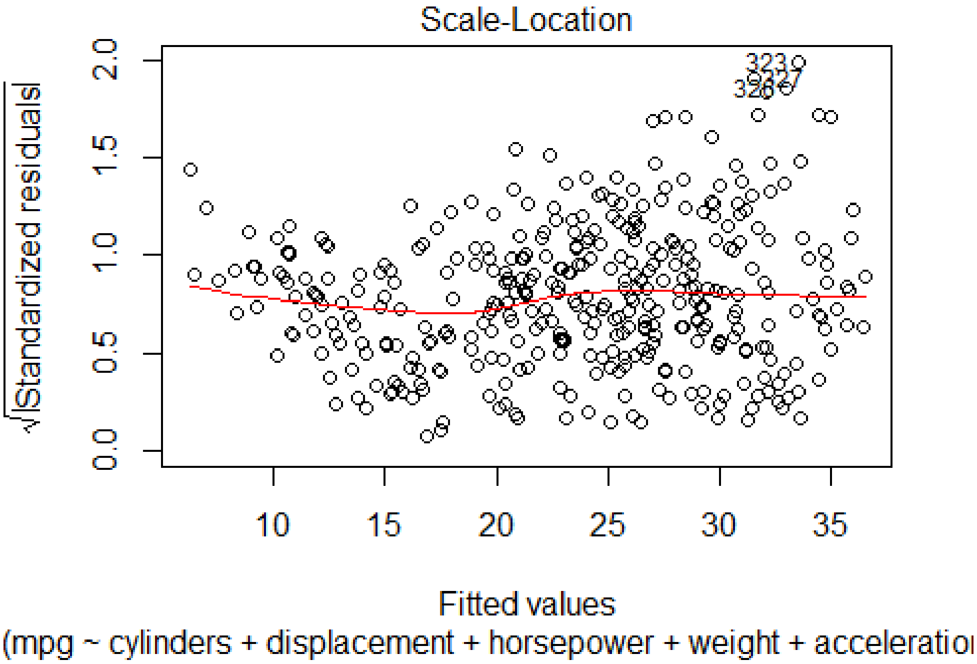
iii. What does the coefficient for the year variable suggest?

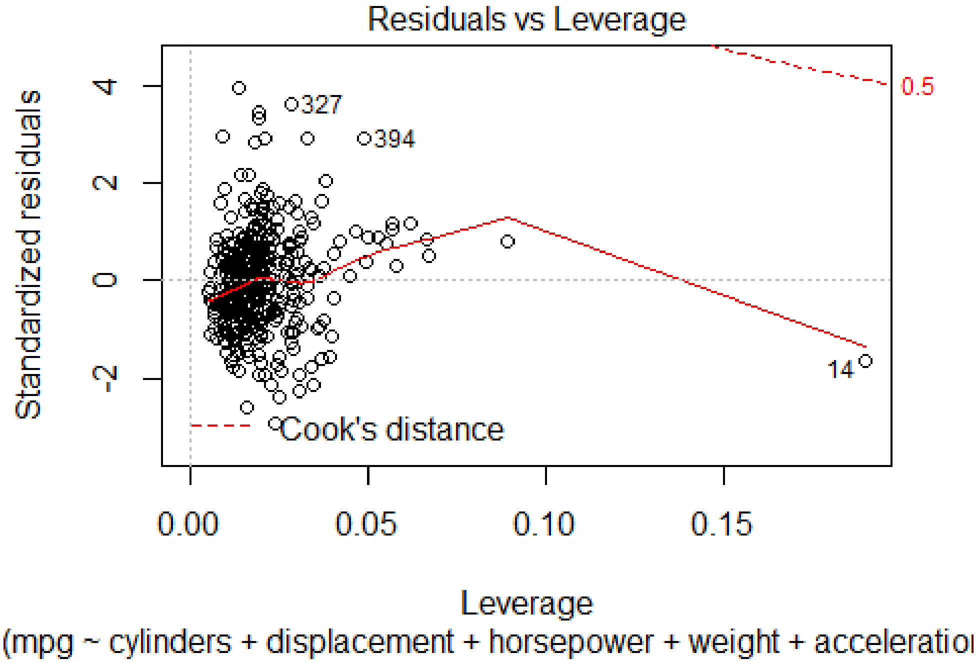
The coefficient on the year variable suggests that each additional model year for a car is associated with a .751 increase in mpg.

(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?


Residuals vs Fitted
Residuals
15
10
5
0
-5
-10
323
327
326
10
15
20
25
30
35
Fitted values
(mpg ~ cylinders + displacement + horsepower + weight + acceleratio


Normal Q-Q
Standardized residuals
4
2
0
-2
323
327
326
-3
-2
-1
0
1
2
3
Theoretical Quantiles
(mpg ~ cylinders + displacement + horsepower + weight + acceleratio





The residuals vs fitted plot shows a slight dip in the middle of the plot, but the degree is not large enough for us to confidently concluse that a non-linear fit would be best. The QQ Plot gives us a little concern in terms of whether residuals are normally distributed, especially regarding observations 323, 327, and 326. The scale-location plot seems to indicate the the assumption of homoskedasticity is met. The residuals vs leverage plot indicates that we might want to double check how coefficients change with and without observation 14.

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

```
reg6 <- lm(mpg ~ cylinders + displacement + horsepower + weight +
acceleration + year + origin + horsepower*weight + cylinders*weight +
acceleration*horsepower + acceleration*weight + origin*displacement +
year*horsepower, data = Auto)
summary(reg6)

##
## Call:
## lm(formula = mpg ~ cylinders + displacement + horsepower + weight +
##     acceleration + year + origin + horsepower * weight + cylinders *
##     weight + acceleration * horsepower + acceleration * weight +
##     origin * displacement + year * horsepower, data = Auto)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.443 -1.457 -0.047  1.357 11.699
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -4.78e+01   1.39e+01   -3.44  0.00066 ***
## cylinders               -2.64e+00   1.03e+00   -2.56  0.01081 *
## displacement            -1.90e-02   9.53e-03   -1.99  0.04719 *
## horsepower               5.00e-01   1.16e-01    4.32  2.0e-05 ***
## weight                  -1.29e-02   2.99e-03   -4.31  2.1e-05 ***
## acceleration             2.68e-01   2.87e-01    0.93  0.35196
## year                     1.44e+00   1.39e-01   10.35  < 2e-16 ***
## origin                  -8.81e-01   8.84e-01   -1.00  0.31937
## horsepower:weight        1.68e-05   1.08e-05    1.56  0.12066
## cylinders:weight         9.59e-04   3.14e-04    3.05  0.00244 **
## horsepower:acceleration -6.07e-03   2.56e-03   -2.37  0.01845 *
## weight:acceleration      6.65e-05   1.37e-04    0.49  0.62679
## displacement:origin      1.35e-02   7.73e-03    1.75  0.08145 .
## horsepower:year         -7.15e-03   1.39e-03   -5.15  4.3e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.77 on 378 degrees of freedom
## Multiple R-squared:  0.878,  Adjusted R-squared:  0.874
## F-statistic:  209 on 13 and 378 DF,  p-value: <2e-16
```

In throwing a few interaction terms into the model, it looks like we have statistical significance for the interaction term of cyliners and weight, horsepower and accleration, and horsepower and year. Displacement and origin can be considered statistically significant at the 90% confidence level as well. The year is probably an important variable because it meant that different emissions reduction standards that could affect GDP were interacting with the effect of variables like horsepower.

(f) Try a few different transformations of the variables, such as log(X), ???X, X2. Comment on your findings.

```
reg7 <- lm(mpg ~ cylinders + log(displacement) + log(horsepower) +
log(weight) + log(acceleration) + year + origin, data = Auto)
summary(reg7)

##
## Call:
## lm(formula = mpg ~ cylinders + log(displacement) + log(horsepower) +
##     log(weight) + log(acceleration) + year + origin, data = Auto)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.800 -1.764 -0.054  1.497 12.652
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       116.7378    10.0018   11.67  < 2e-16 ***
## cylinders           0.4227     0.2833    1.49   0.1365
## log(displacement)  -1.7722     1.4163   -1.25   0.2116
## log(horsepower)    -7.1216     1.5508   -4.59  6.0e-06 ***
## log(weight)       -12.1665     2.1953   -5.54  5.6e-08 ***
## log(acceleration)  -4.9727     1.5946   -3.12   0.0020 **
## year                0.7278     0.0465   15.64  < 2e-16 ***
## origin              0.7968     0.2775    2.87   0.0043 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.05 on 384 degrees of freedom
## Multiple R-squared:  0.85,   Adjusted R-squared:  0.847
## F-statistic:  311 on 7 and 384 DF,  p-value: <2e-16
```

```
reg8 <- lm(mpg ~ cylinders + displacement + horsepower + weight +
acceleration + year + origin + sqrt(horsepower) + sqrt(acceleration), data =
Auto)
summary(reg8)

##
## Call:
## lm(formula = mpg ~ cylinders + displacement + horsepower + weight +
##     acceleration + year + origin + sqrt(horsepower) + sqrt(acceleration),
##     data = Auto)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.693 -1.675 -0.087  1.537 12.072
##
```

```
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)    
## (Intercept)         6.64e+01   1.56e+01    4.24  2.8e-05 ***
## cylinders           1.01e-01   2.93e-01    0.34  0.73050    
## displacement       -8.53e-03   7.31e-03   -1.17  0.24400    
## horsepower          3.91e-01   4.92e-02    7.95  2.2e-14 ***
## weight             -3.07e-03   6.71e-04   -4.58  6.3e-06 ***
## acceleration        1.31e+00   9.81e-01    1.34  0.18163    
## year                7.42e-01   4.53e-02   16.39  < 2e-16 ***
## origin              9.11e-01   2.52e-01    3.62  0.00034 ***
## sqrt(horsepower)   -9.87e+00   1.10e+00   -8.96  < 2e-16 ***
## sqrt(acceleration) -1.33e+01   7.91e+00   -1.69  0.09242 .  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.95 on 382 degrees of freedom
## Multiple R-squared:  0.86,   Adjusted R-squared:  0.857 
## F-statistic:  261 on 9 and 382 DF,  p-value: <2e-16
```

```
reg9 <- lm(mpg ~ cylinders + displacement + horsepower + weight +
acceleration + year + origin + I(horsepower^2) + I(acceleration^2), data =
Auto)
summary(reg9)
```

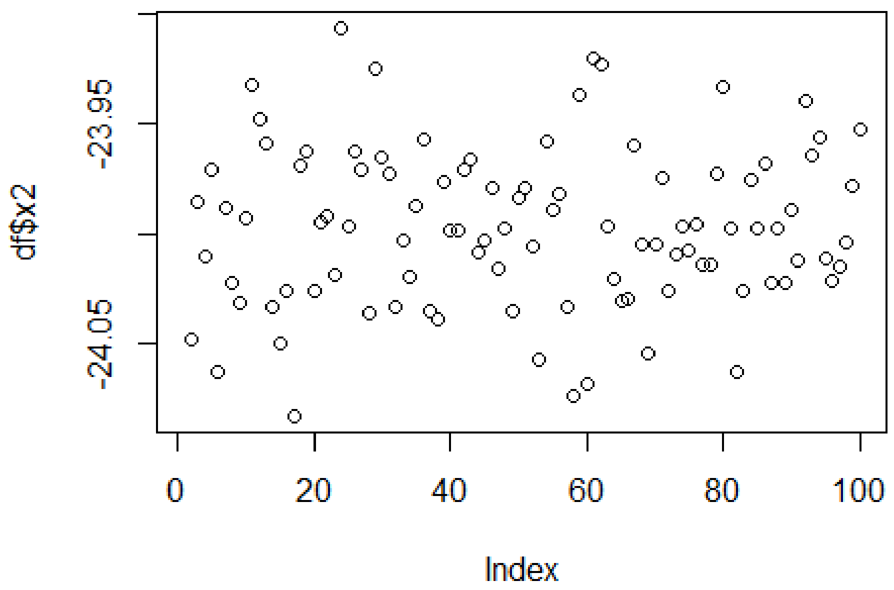
```
## 
## Call:
## lm(formula = mpg ~ cylinders + displacement + horsepower + weight + 
##     acceleration + year + origin + I(horsepower^2) + I(acceleration^2), 
##     data = Auto)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
##  -8.83  -1.69  -0.18   1.62  12.19 
## 
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)    
## (Intercept)       11.488111   6.022749    1.91   0.0572 .  
## cylinders          0.372182   0.302682    1.23   0.2196    
## displacement      -0.011009   0.007436   -1.48   0.1396    
## horsepower        -0.304181   0.034588   -8.79  < 2e-16 ***
## weight            -0.002975   0.000683   -4.36  1.7e-05 ***
## acceleration      -1.717647   0.541119   -3.17   0.0016 ** 
## year               0.738059   0.045660   16.16  < 2e-16 ***
## origin             0.995773   0.252751    3.94  9.7e-05 ***
## I(horsepower^2)    0.000924   0.000110    8.39  9.7e-16 ***
## I(acceleration^2)  0.041309   0.015846    2.61   0.0095 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.98 on 382 degrees of freedom
## Multiple R-squared:  0.858,  Adjusted R-squared:  0.854
## F-statistic:  256 on 9 and 382 DF,  p-value: <2e-16
```

For the log transformation, we can now say, for example, that a 1% change in weight is associated with a decrease of 12.1 for mpg or a 1% change in horsepower is associated with a 7.12 decline in mpg. This might be a more useful interpretation and given statistical significance, we can conclude that linear modeling may not be the best here. The results of the square root and squared models confirm that suspicion, with the suqare root and squared terms for horsepower demonstrating statistical significance, for example.

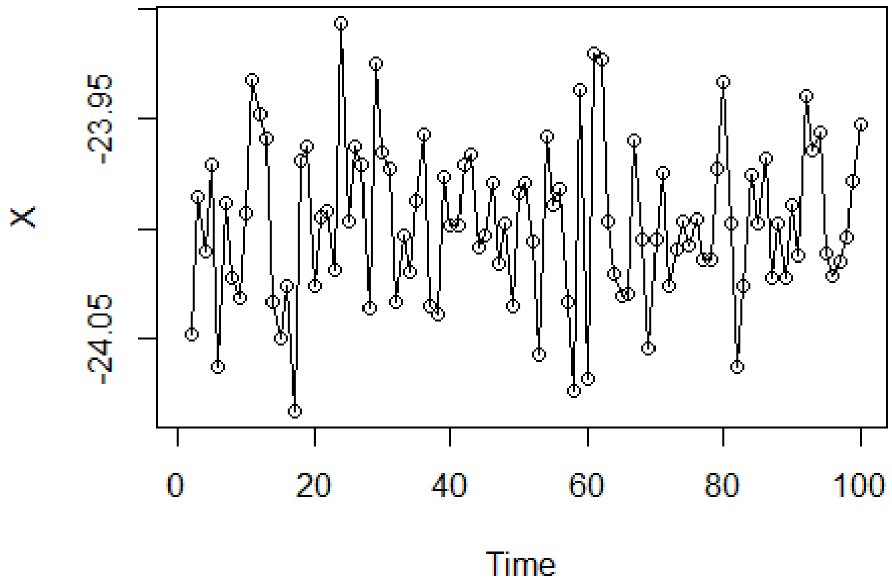
## Extra #14

Problem 14 a) Simulate a time series X of length N = 100 from the above formula, using the lag k = 1, coefficients ??0 = 1 and ??1 = ???0.5 and error terms t ??? N(0, 0.2 2 ). The formula tells you how to make Xt for t ??? k + 1. Choose X1 arbitrarily. Plot X as a vector. Convert X into a timeseries object with the function as.ts() and plot it again. Describe the plot.

```
true_beta_0 = 1
true_beta_1 = -0.5

df <- data.frame(time=1:100, x1=50)

for(i in 2:100){
    df$x2[i] <-  true_beta_0 + true_beta_1*df$x1[[i-1]] + rnorm(1,0,0.2^2)
}


plot(df$x2)
```



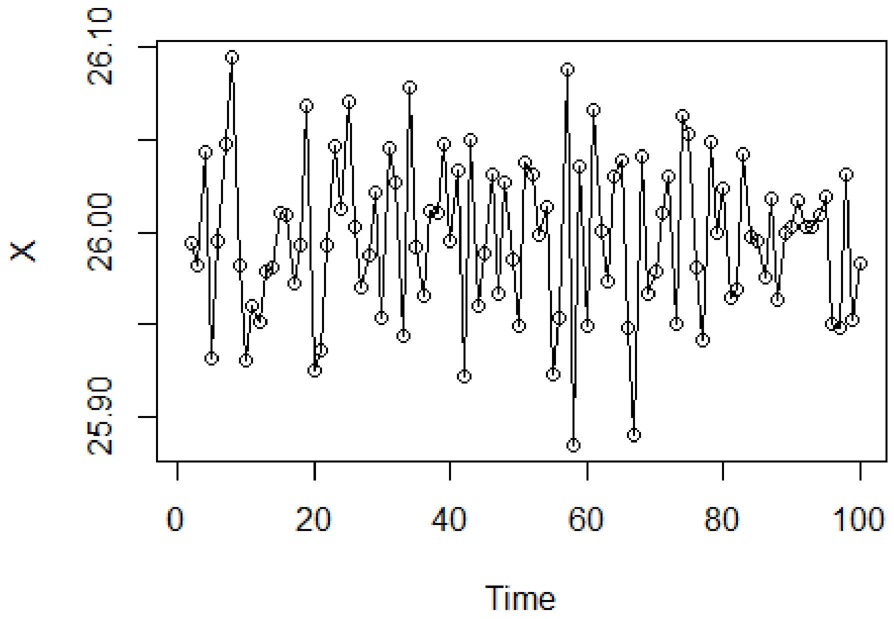
```
df$x2 <- as.ts(df$x2)

plot(df$time,df$x2, xlab = "Time", ylab = "X")
lines(df$time,df$x2)
```



There is significant time-to-time volatility demonstrated in the plot, with particularly exteeme values around t = 17, t = 57, t = 60, t = 62, and t = 84.

b) Repeat part a) with ??0 = 1, ??1 = +0.5. How does the plot change?

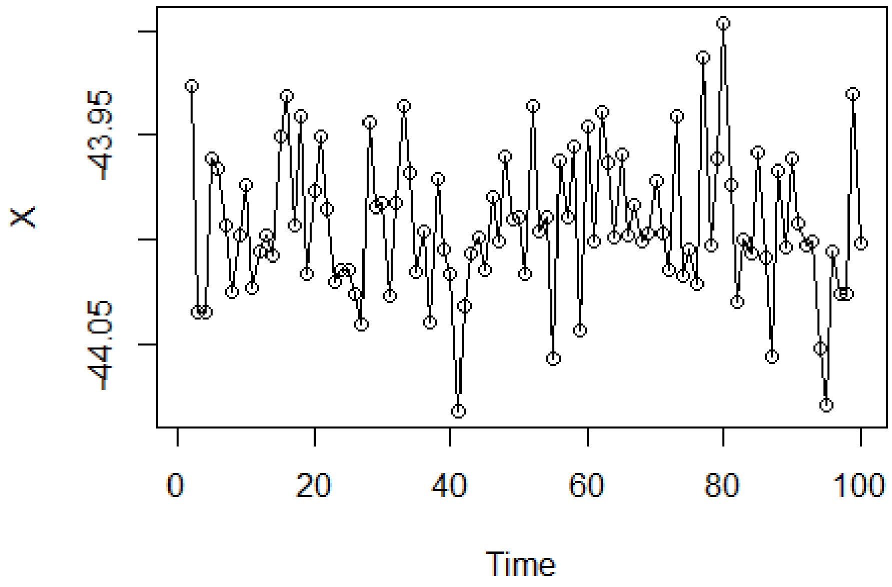
```
true_beta_0 = 1
true_beta_1 = 0.5

df <- data.frame(time=1:100, x1=50)

for(i in 2:100){
    df$x2[i] <-  true_beta_0 + true_beta_1*df$x1[[i-1]] + rnorm(1,0,0.2^2)
}

df$x2 <- as.ts(df$x2)
plot(df$time,df$x2, xlab = "Time", ylab = "X")
lines(df$time,df$x2)
```



While there are still some extreme values (t = 21 for example), there appears to be less volatility compared to the the previous plot and some values are flipped despite the original x before simulating the time series being the same.

c) Repeat part a) with ??0 = 1, ??1 = ???0.9. How does the plot change?

```
true_beta_0 = 1
true_beta_1 = -0.9

df <- data.frame(time=1:100, x1=50)

for(i in 2:100){
    df$x2[i] <-  true_beta_0 + true_beta_1*df$x1[[i-1]] + rnorm(1,0,0.2^2)
}

df$x2 <- as.ts(df$x2)
plot(df$time,df$x2, xlab = "Time", ylab = "X")
lines(df$time,df$x2)
```



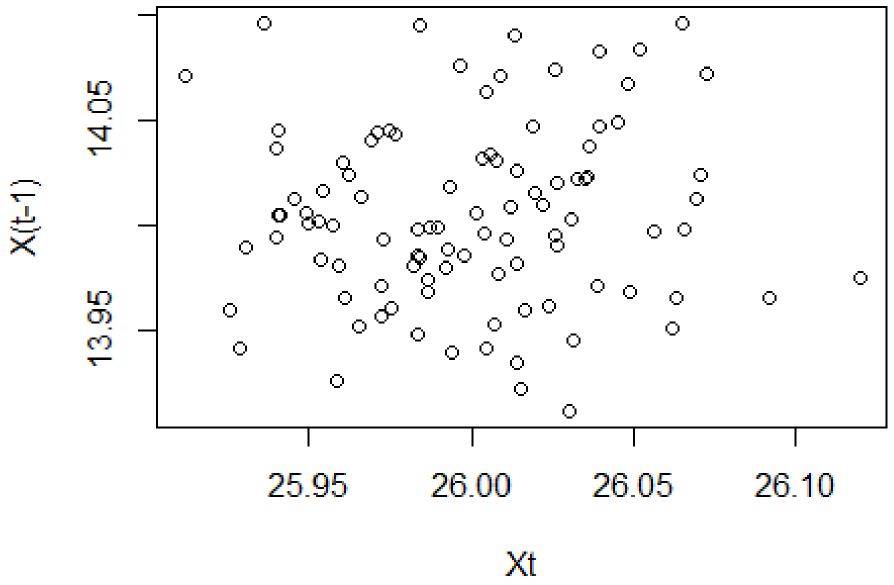
Perhaps because beta 1 is negative again in this example, the plot looks much more like the first time series plot. There are more extreme values due to the Beta 1 being higher in absolute value terms in this example compared to the original example.

## Extra #15

Simulate a time series X as in the previous problem (N = 100 observations, lag k = 1, ??0 = 1, ??1 = ???0.5, t ??? N(0, 0.2^2).

a) Make a scatterplot of Xt against xt???1 for t = 2, . . . , N and describe it.

```
true_beta_0 = 1

true_beta_1 = 0.5

df <- data.frame(time=1:100, x1=50)

for(i in 2:100){
    df$x2[i] <-  true_beta_0 + true_beta_1*df$x1[[i-1]] + rnorm(1,0,0.2^2)
}

for(i in 2:100){
    df$x3[i] <-  true_beta_0 + true_beta_1*df$x2[[i-1]] + rnorm(1,0,0.2^2)
}

plot(df$x2,df$x3, xlab = "Xt", ylab = "X(t-1)")
```



There is no clear discernible pattern in the plot. Because we are using a beta 1 of 0.5, there is a proportional halving on each new lag and thus there is a fairly consistent relationship between each value for x(t) and for x(t-1).

b) Create a data frame of N ??? 1 observations and 2 columns that contains (Xt???1, Xt) in row t. Use this to fit a linear model to predict Xt from xXt???1. Compare the estimated coefficients to the ??i. Also compare the residual standard error to the standard deviation of thet. Summarize your results and observations.

```
true_beta_0 = 1

true_beta_1 = 0.5

df <- data.frame(time=1:100, x1=50)

for(i in 2:100){
    df$x2[i] <-  true_beta_0 + true_beta_1*df$x1[[i-1]] + rnorm(1,0,0.2^2)
}

for(i in 2:100){
    df$x3[i] <-  true_beta_0 + true_beta_1*df$x2[[i-1]] + rnorm(1,0,0.2^2)
}

reg10 <- lm(x2 ~ x3, data = df)
summary(reg10)
```

```
##
## Call:
## lm(formula = x2 ~ x3, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.11280 -0.02759  0.00161  0.03248  0.08546
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  25.2169     1.2724   19.82   <2e-16 ***
## x3            0.0562     0.0909    0.62     0.54
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0426 on 96 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.00397,    Adjusted R-squared:  -0.0064
## F-statistic: 0.383 on 1 and 96 DF,  p-value: 0.538
```

A unit increase in x(t-1) is associated with a 0.097 decrease in x(t). It is interesting that this is not statistically significant, but they may be simply because of the small sample size and impact of the random error.