HW1

Norman Hong

February 8, 2019

3.3

Suppose we have a data set with 5 predictors, $X_1 = GPA$, $X_2 = IQ$, $X_3 = Gender$ (1 = Female and 0 = male), X_4 = Interaction between GPA and IQ, and X_5 = interaction between GPA and Gender. The response is starting salary after graduation (in t housands of dollars). Suppose we use least squares to fit the model, and get $\hat{\beta}_0 = 50$, $\hat{\beta}_1 = 20$, $\hat{\beta}_2 = .07$, $\hat{\beta}_3 = 35$, $\hat{\beta}_4 = .01$, and $\hat{\beta}_5 = -10$.

(a)

For a fixed value of IQ and GPA, males earn more on average than females provided that GPA is high enough. The interaction term between GPA and Gender causes females to earn less than males once gpa is high enough.
The regression model is: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \hat{\beta}_4 X_1 X_2 + \hat{\beta}_5 X_1 X_3$. When $X_3 = 1$, this corresponds to the equation for females, the regression equation is: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 + \hat{\beta}_4 X_1 X_2 + \hat{\beta}_5 X_1$. When $X_3 = 0$, the regression equation for males is: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_4 X_1 X_2$. The difference between the two equations is $\hat{y_f} - \hat{y_m} = \hat{\beta}_3 + \hat{\beta}_5 X_1$. Since $\hat{\beta}_5 = -10$, if gpa is large enough, the difference in income between females and males turn from positive to negative.

(b)

$Y = 50 + 20X_1 + .07X_2 + 35X_3 + .01X_4 - 10X_5 = 50 + 20X_1 + .07X_2 + 35X_3 + .01X_1X_2 - 10X_1X_3$ $= 50 + 20(4) + 110(.07) + 35X_3 + .01(110)(4) + (-10)(4)X_3 = 130 + 7.7 + 4.4 - 5 = 137$

(c)

False. The size of the coefficient is relative to the scale used to measure the variables. In other words, the magnitude of the coefficients is relative to the units used. The only way to determine if there is statistical evidence for an interaction effect is to look at the t-tets.

3.8

(a)

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

There is a relationship between mpg and horsepower because the coefficient for hosepower is very statistically significant. For every 1 unit increase in horsepower, mpg decreases by .15 units. The standard deviation of the coefficient is very small, which could be the reason for the very low p value. The 95% confidence interval does not include 0. The associated 95% confidence interval for the response variable when horsepower is 98 is between 23.97 and 24.96. The 95% prediction interval for the response variable when horsepower is 98 is from 14.80 to 34.12. The R^2 coefficient is .605, which implies that about 60.5% of the variation in Y is explained by the regression or explained by the variable X.

```r
reg <- lm(mpg ~ horsepower, data=Auto)
summary(reg)
```

```
##
## 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
```

```
confint(reg)
```

```
##                  2.5 %     97.5 %
## (Intercept) 38.525212 41.3465103
## horsepower  -0.170517 -0.1451725
```

```
predict(reg, data.frame(horsepower=98), interval='confidence')
```

```
##        fit      lwr      upr
## 1 24.46708 23.97308 24.96108
```

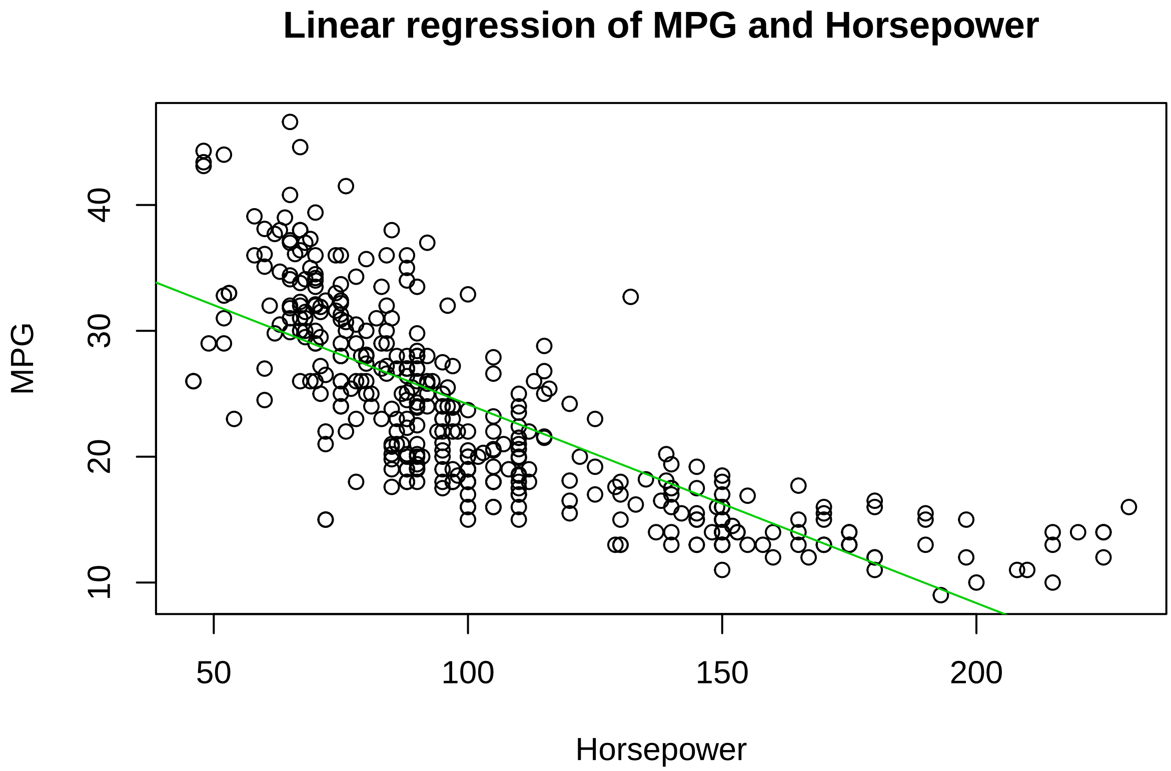
```
predict(reg, data.frame(horsepower=98), interval='prediction')
```

```
##        fit     lwr      upr
## 1 24.46708 14.8094 34.12476
```

(b)

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

```
plot(x=Auto$horsepower, y=Auto$mpg, xlab='Horsepower', ylab='MPG',
     main="Linear regression of MPG and Horsepower")
abline(reg, col=3)
```

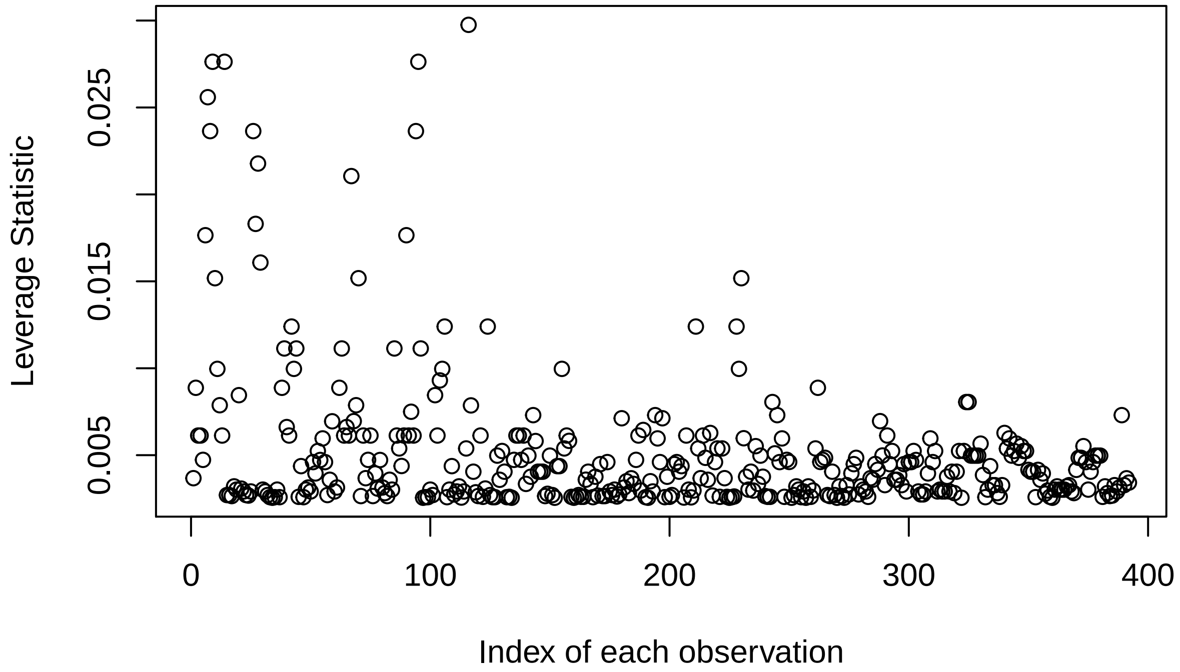


(c)

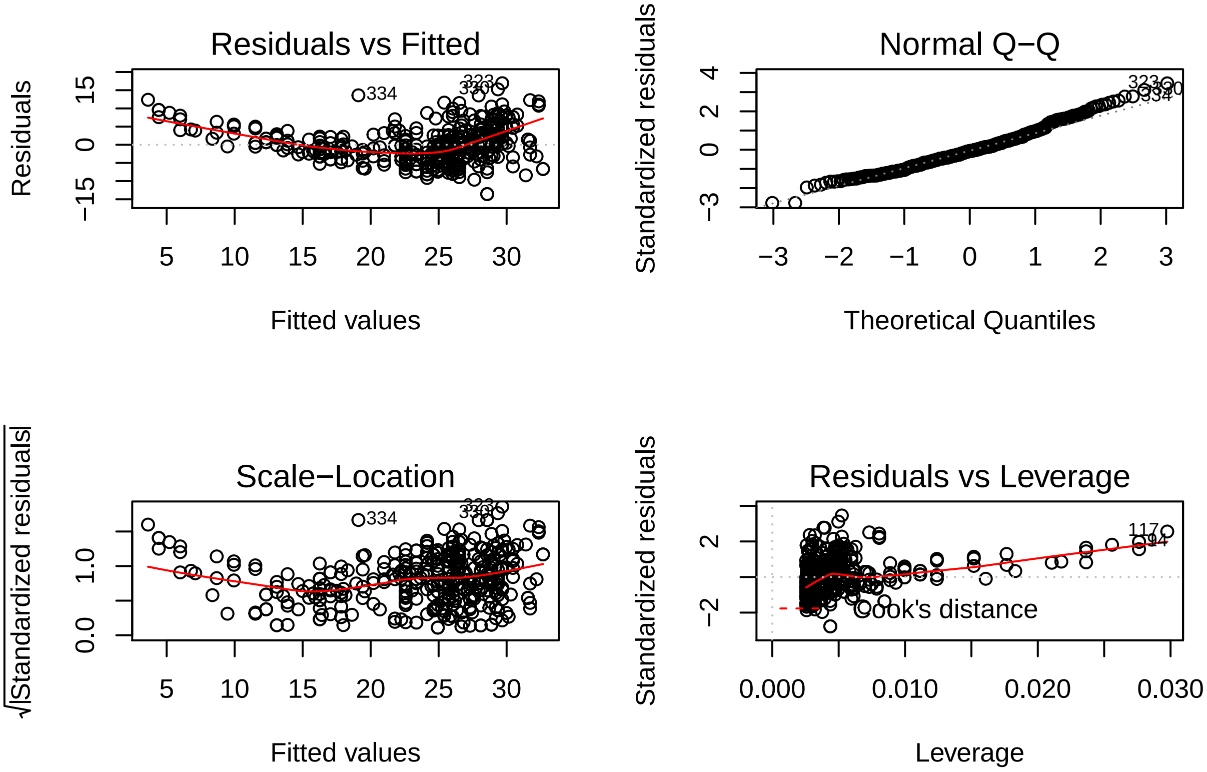
Use the plot() function to produce diagnostic plots of the least squares regression fit.

The plot of fitted values vs residuals show a non-linear relationship. The residuals don't have consistent values. This indicates heteroscedasticity. Smaller fitted values tend to have error terms with high magnitude, fitted values around 15 to 20 tend to have error terms with low magnitude, and large fitted values tend to have residuals that range from -10 to 15. The fitted values vs residuals also indicate a point with residual of 14 and fitted value of 19. This point is an outlier because it does not follow the overall pattern. The observation index vs leverage plot does not indicate that any high-leverage data points. The qqplot shows that the standardized residuals are normally distributed.

```r
# residuals plot
# plot(predict(reg), residuals(reg), xlab='fitted values', ylab='residuals')
# plot(predict(reg), rstudent(reg), xlab='fitted values', ylab='studentized residuals')
# plot of leverage vs index of observation.
plot(hatvalues(reg), xlab='Index of each observation', ylab='Leverage Statistic')
```



```
# plot(hatvalues(reg), rstudent(reg), xlab='Leverage', ylab='Studentized residuals')
par(mfrow=c(2,2))
plot(reg)
```



3.10

This question should be answered using the Carseats data set.

(a)

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

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

```
##
## Call:
## lm(formula = Sales ~ Price + Urban + US, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9206 -1.6220 -0.0564  1.5786  7.0581
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.043469   0.651012  20.036  < 2e-16 ***
## Price       -0.054459   0.005242 -10.389  < 2e-16 ***
## UrbanYes    -0.021916   0.271650  -0.081    0.936
## USYes        1.200573   0.259042   4.635 4.86e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.472 on 396 degrees of freedom
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2335
## F-statistic: 41.52 on 3 and 396 DF,  p-value: < 2.2e-16
```

(b)

Provide an interpretation of each coefficient in the model. Some of the variables in the model are qualitative.

If price, urban, and US variables are all 0, then the unit sale of carseats is predicted to be 13.04. Since the sales variable is in thousands, the unit sale of carseats is 13,000. For a 1 unit increase in price, the sale of carseats decreases by .054 units. Stores in urban locations experience a decrease of .02 units of carseats sold when compared to a rural location. Stores in the US experience an increase in the sale of carseats by 1.2 units when compared to locations outside the United States. The intercept, Price and US coefficients are statistically significant at the , whereas the Urban coefficient is not statistically significant.

(c)

Write out the model in equation form, being careful to handle the qualitative variables properly. $Sales = \beta_0 + \beta_1 Price + \beta_2 Urban + \beta_3 US$
$Sales = \beta_0 + \beta_1 Price + \beta_2 + \beta_3 US$ for urban stores
$Sales = \beta_0 + \beta_1 Price + \beta_3 US$ for rural stores
$Sales = \beta_0 + \beta_1 Price + \beta_2 Urban + \beta_3$ for stores in United states
$Sales = \beta_0 + \beta_1 Price + \beta_2 Urban$ for stores outside the United States
$Sales = \beta_0 + \beta_1 Price + \beta_2 + \beta_3$ for urban stores inside United States.
$Sales = \beta_0 + \beta_1 Price$ for rural stores outside of United States.

3.15.

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 statistically significant association between the predictor and the response? Create some plots to back up your assertions.

The coefficient for zn, indus, nox, rm, age, dis, rad, tax, ptratio, black, lstat, and medv variables are all statistically significant at alpha level of .001. The chas variable is not statistically significant.

```r
summary(lm(Boston$crim ~ Boston$zn))
```

```
##
## Call:
## lm(formula = Boston$crim ~ Boston$zn)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.429 -4.222 -2.620  1.250 84.523
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.45369    0.41722  10.675  < 2e-16 ***
## Boston$zn   -0.07393    0.01609  -4.594 5.51e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.435 on 504 degrees of freedom
## Multiple R-squared:  0.04019,    Adjusted R-squared:  0.03828
## F-statistic:  21.1 on 1 and 504 DF,  p-value: 5.506e-06
```

```r
summary(lm(Boston$crim ~ Boston$indus))
```

```
##
## Call:
## lm(formula = Boston$crim ~ Boston$indus)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.972  -2.698  -0.736   0.712  81.813
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -2.06374    0.66723  -3.093  0.00209 **
## Boston$indus   0.50978    0.05102   9.991  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.866 on 504 degrees of freedom
## Multiple R-squared:  0.1653, Adjusted R-squared:  0.1637
## F-statistic: 99.82 on 1 and 504 DF,  p-value: < 2.2e-16
```

```r
summary(lm(Boston$crim ~ Boston$chas))
```

```
##
## Call:
## lm(formula = Boston$crim ~ Boston$chas)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.738 -3.661 -3.435  0.018 85.232
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.7444     0.3961   9.453   <2e-16 ***
## Boston$chas   -1.8928     1.5061  -1.257    0.209
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.597 on 504 degrees of freedom
## Multiple R-squared:  0.003124,    Adjusted R-squared:  0.001146
## F-statistic: 1.579 on 1 and 504 DF,  p-value: 0.2094
```

```r
summary(lm(Boston$crim ~ Boston$nox))
```

```
##
## Call:
## lm(formula = Boston$crim ~ Boston$nox)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.371  -2.738  -0.974   0.559  81.728
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -13.720      1.699  -8.073 5.08e-15 ***
## Boston$nox    31.249      2.999  10.419  < 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:  0.1772, Adjusted R-squared:  0.1756
## F-statistic: 108.6 on 1 and 504 DF,  p-value: < 2.2e-16
```

```r
summary(lm(Boston$crim ~ Boston$rm))
```

```
##
## Call:
## lm(formula = Boston$crim ~ Boston$rm)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.604 -3.952 -2.654  0.989 87.197
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   20.482      3.365   6.088 2.27e-09 ***
## Boston$rm     -2.684      0.532  -5.045 6.35e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.401 on 504 degrees of freedom
## Multiple R-squared:  0.04807,    Adjusted R-squared:  0.04618
## F-statistic: 25.45 on 1 and 504 DF,  p-value: 6.347e-07
```

```r
summary(lm(Boston$crim ~ Boston$age))
```

```
##
## Call:
## lm(formula = Boston$crim ~ Boston$age)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.789 -4.257 -1.230  1.527 82.849
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.77791    0.94398  -4.002 7.22e-05 ***
## Boston$age   0.10779    0.01274   8.463 2.85e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.057 on 504 degrees of freedom
## Multiple R-squared:  0.1244, Adjusted R-squared:  0.1227
## F-statistic: 71.62 on 1 and 504 DF,  p-value: 2.855e-16
```

```r
summary(lm(Boston$crim ~ Boston$dis))
```

```
## 
## Call:
## lm(formula = Boston$crim ~ Boston$dis)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -6.708 -4.134 -1.527  1.516 81.674 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   9.4993     0.7304  13.006   <2e-16 ***
## Boston$dis   -1.5509     0.1683  -9.213   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 7.965 on 504 degrees of freedom
## Multiple R-squared:  0.1441, Adjusted R-squared:  0.1425 
## F-statistic: 84.89 on 1 and 504 DF,  p-value: < 2.2e-16
```

```r
summary(lm(Boston$crim ~ Boston$rad))
```

```
## 
## Call:
## lm(formula = Boston$crim ~ Boston$rad)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -10.164  -1.381  -0.141   0.660  76.433 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -2.28716    0.44348  -5.157 3.61e-07 ***
## Boston$rad   0.61791    0.03433  17.998  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 6.718 on 504 degrees of freedom
## Multiple R-squared:  0.3913, Adjusted R-squared:   0.39 
## F-statistic: 323.9 on 1 and 504 DF,  p-value: < 2.2e-16
```

```r
summary(lm(Boston$crim ~ Boston$tax))
```

```
## 
## Call:
## lm(formula = Boston$crim ~ Boston$tax)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -12.513  -2.738  -0.194   1.065  77.696 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -8.528369   0.815809  -10.45   <2e-16 ***
## Boston$tax   0.029742   0.001847   16.10   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.997 on 504 degrees of freedom
## Multiple R-squared:  0.3396, Adjusted R-squared:  0.3383
## F-statistic: 259.2 on 1 and 504 DF,  p-value: < 2.2e-16
```

```r
summary(lm(Boston$crim ~ Boston$ptratio))
```

```
##
## Call:
## lm(formula = Boston$crim ~ Boston$ptratio)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.654 -3.985 -1.912  1.825 83.353
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -17.6469     3.1473  -5.607 3.40e-08 ***
## Boston$ptratio   1.1520     0.1694   6.801 2.94e-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:  0.08407,    Adjusted R-squared:  0.08225
## F-statistic: 46.26 on 1 and 504 DF,  p-value: 2.943e-11
```

```r
summary(lm(Boston$crim ~ Boston$black))
```

```
##
## Call:
## lm(formula = Boston$crim ~ Boston$black)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.756  -2.299  -2.095  -1.296  86.822
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  16.553529   1.425903  11.609   <2e-16 ***
## Boston$black -0.036280   0.003873  -9.367   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.946 on 504 degrees of freedom
## Multiple R-squared:  0.1483, Adjusted R-squared:  0.1466
## F-statistic: 87.74 on 1 and 504 DF,  p-value: < 2.2e-16
```

```r
summary(lm(Boston$crim ~ Boston$lstat))
```

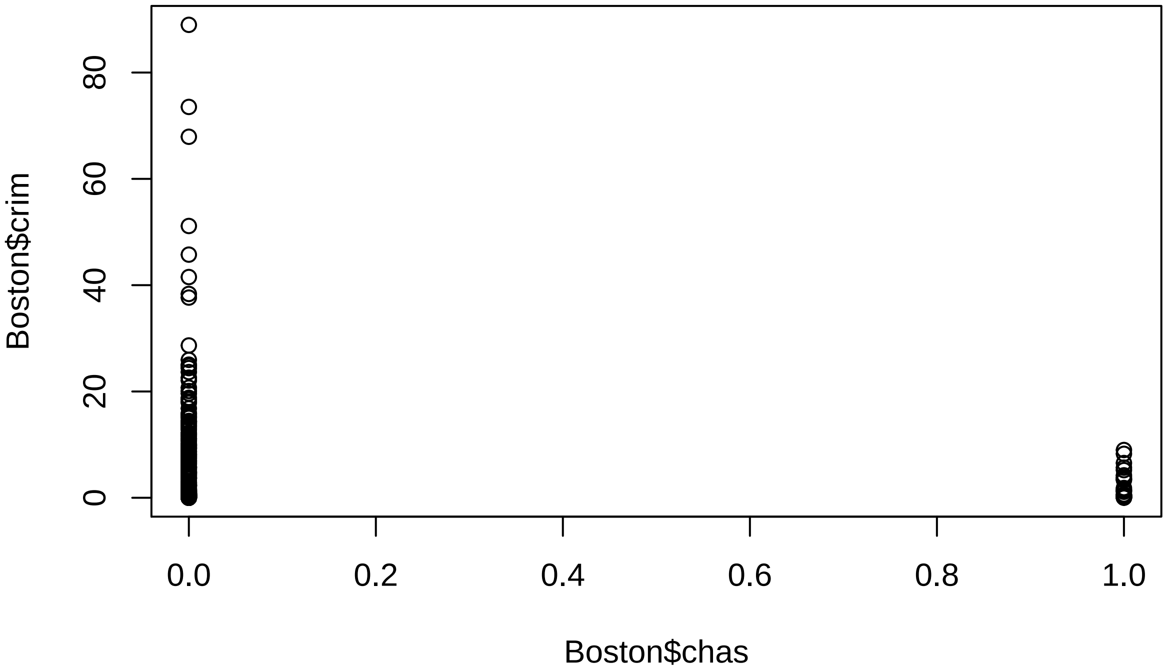
```
##
## Call:
## lm(formula = Boston$crim ~ Boston$lstat)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.925  -2.822  -0.664   1.079  82.862
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -3.33054    0.69376  -4.801 2.09e-06 ***
## Boston$lstat  0.54880    0.04776  11.491  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.664 on 504 degrees of freedom
## Multiple R-squared:  0.2076, Adjusted R-squared:  0.206
## F-statistic:   132 on 1 and 504 DF,  p-value: < 2.2e-16
```

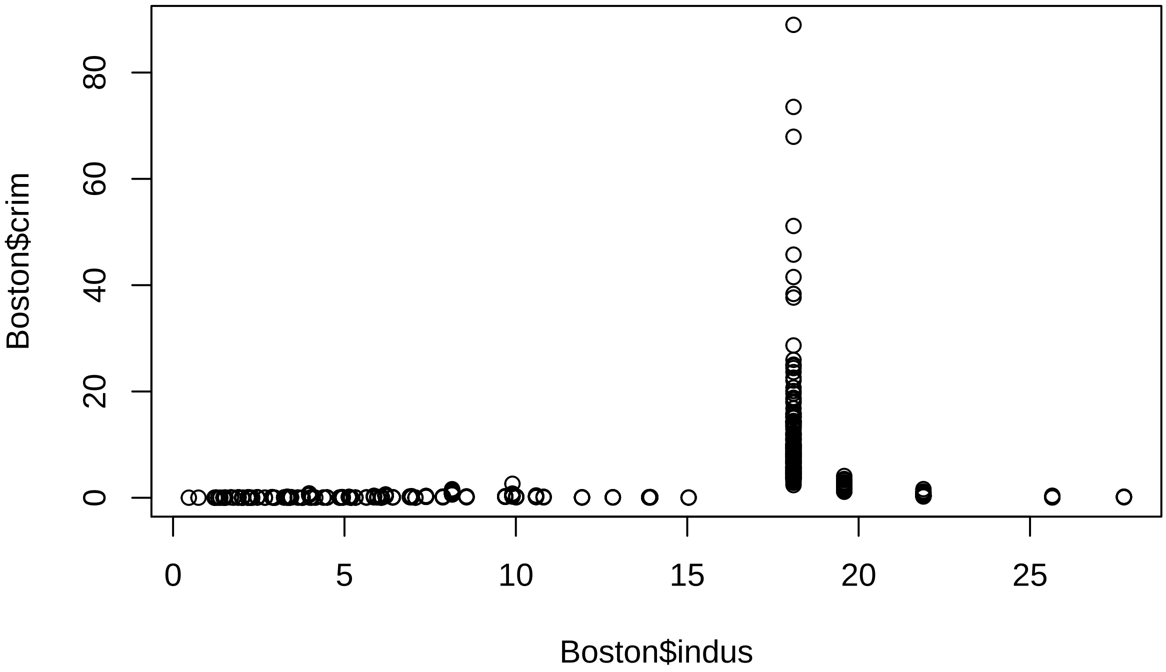
```r
summary(lm(Boston$crim ~ Boston$medv))
```

```
##
## Call:
## lm(formula = Boston$crim ~ Boston$medv)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.071 -4.022 -2.343  1.298 80.957
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.79654    0.93419   12.63   <2e-16 ***
## Boston$medv -0.36316    0.03839   -9.46   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.934 on 504 degrees of freedom
## Multiple R-squared:  0.1508, Adjusted R-squared:  0.1491
## F-statistic: 89.49 on 1 and 504 DF,  p-value: < 2.2e-16
```

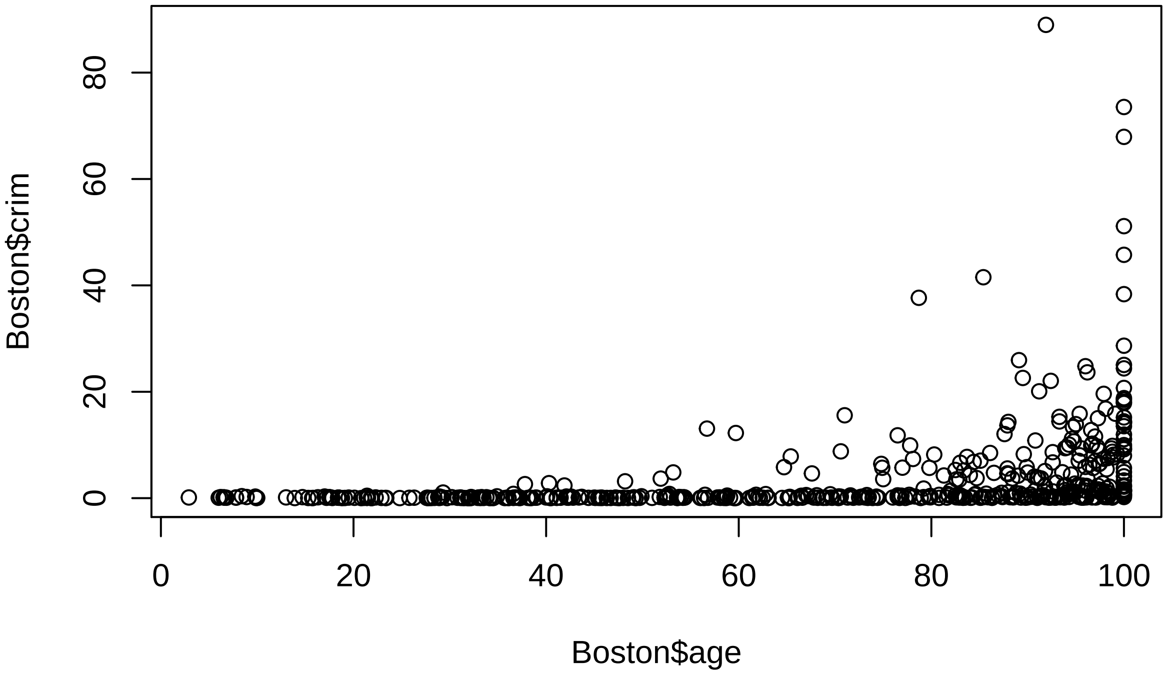
```r
plot(Boston$chas, Boston$crim)
```



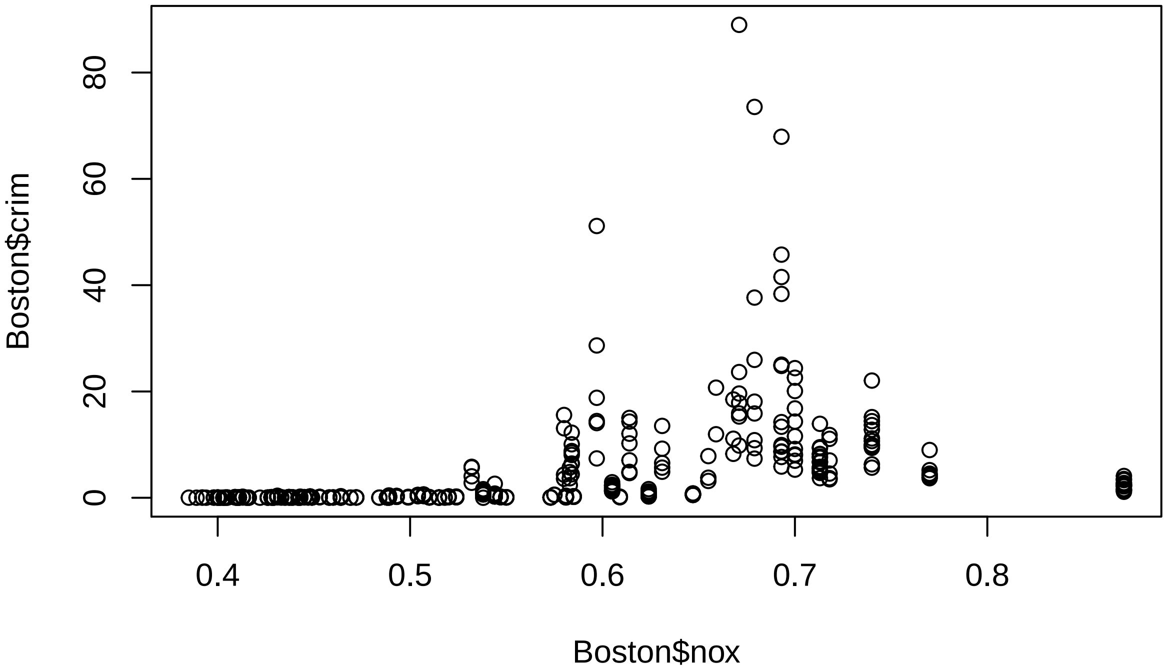
```
plot(Boston$indus, Boston$crim)
```



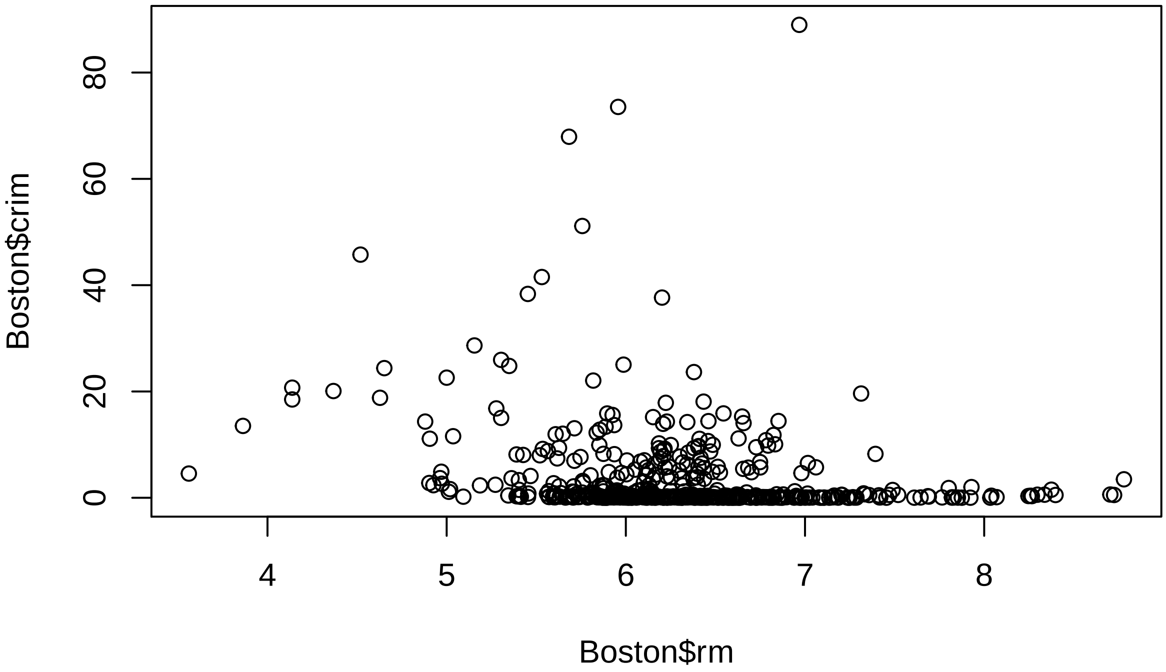
```
plot(Boston$age, Boston$crim)
```



```
plot(Boston$nox, Boston$crim)
```



```r
plot(Boston$rm, Boston$crim)
```



(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.

The F-statistic tells us that the null is rejected in favor of the alternative hypothesis where there is atleast 1 variable with a coefficient that does not equal 0. The F statistic does not tell us which one though. At alpha level of .001, the only coefficients were we can reject the null hypothesis are for the dis and rad variables. This is very differet from the previous simple bivariate linear models because when controlling for other variables, most variables are no longer statistically significant at alpha of .001. The zn, intercept, and black variables are statistically significant at alpha of .05, whereas in the bivarate models, these coefficients were all statistically significant at alpha level of .001. The medv variable is statistically significant with p value of .09. The nox and lstat variables are statistically significant at alpha of .1. The rm, age, indus, tax, and ptratio are no longer statistically significant in the multivariate model. This is a huge difference when compared to the simple linear regression model where these coefficients were statistically significant at alpha of .001. The chas variable remained statistically insignificant.

```
summary(lm(crim ~ zn+indus+chas+nox+rm+age+dis+rad+tax+ptratio+black+lstat+medv, data=Boston))
```

```
##
## 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
## -9.924 -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
```

Extra 10

Consider the built-in data set cars. Fit a linear regression model to the data. What are the 3 observations with the largest standardized residuals (in magnitude)? What are their leverages? Where are the 3 observations with the largest leverages? What are their standardized residuals? Use the R function influence.lm

The indexes of the 3 observations with the highest standardized residuals in magniude are 49, 23, and 35. The leverage of these 3 observations are .073, .021, .024, respectively. The indexes of the 3 observations with the highest leverage are 1, 2, and 50. The corresponding standardized residuals are .26, .81, and .28, respectively.

```
fit <- lm(dist ~ speed, data=cars)
sort(abs(rstudent(fit)), decreasing=TRUE)[1:3]
```

```
##       49       23       35
## 3.184993 3.022829 2.098482
```

```
lev <- hatvalues(fit)
lev[c(49, 23, 35)]
```

```
##         49         23         35
## 0.07398540 0.02143066 0.02493431
```

```
sort(lev, decreasing=T)[1:3]
```

```
##          1          2         50
## 0.11486131 0.11486131 0.08727007
```

```
rstudent(fit)[c(1,2,50)]
```

```
##         1         2        50
## 0.2634500 0.8160784 0.2877453
```

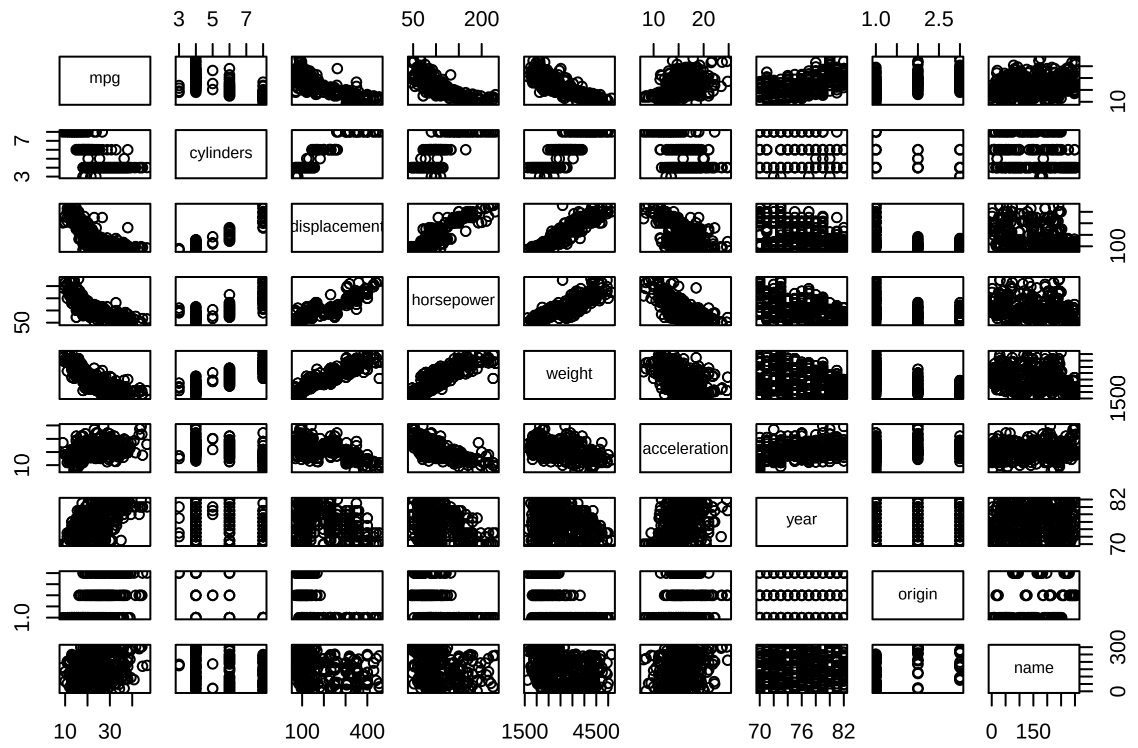
3.9

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.

```r
pairs(Auto)
```



(b)

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

```r
cor(Auto[-9]) # column 9 is the names varibable, which is not numeric
```

```
##                     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 variable except name as the predictors. Use the summary() function to print the results. Comment on the output. For instnace: i. Is there a relationship between the predictors and the response? ii. Which predictors appear to have a statistically significant relationship to the response? iii. What does the coefficient for the year variable suggest?

If all variables are 0, then the predicted mpg is -17. The intercept is statistically significant at alpha of .001. There is statistical significance to believe that the intercept is not 0. Weight, year and origin have coefficients that are statistically significant at alpha of .001. So it seems that the coefficients are not 0 when controlling for the variables included in the model. Displacement is statistically significant at alpha of .01. If we were to do a hypothesis test with alpha of .05, then there is statistical evidence that displacement, weight, year, and origin variables are correlated to the response. There is no statistical evidence to believe that cylinders, horsepower, and acceleration variables have a relationship with mpg. The coefficient for the year variable suggests that a 1 unit increase in year is correlated with a .750 increase in mpg.

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

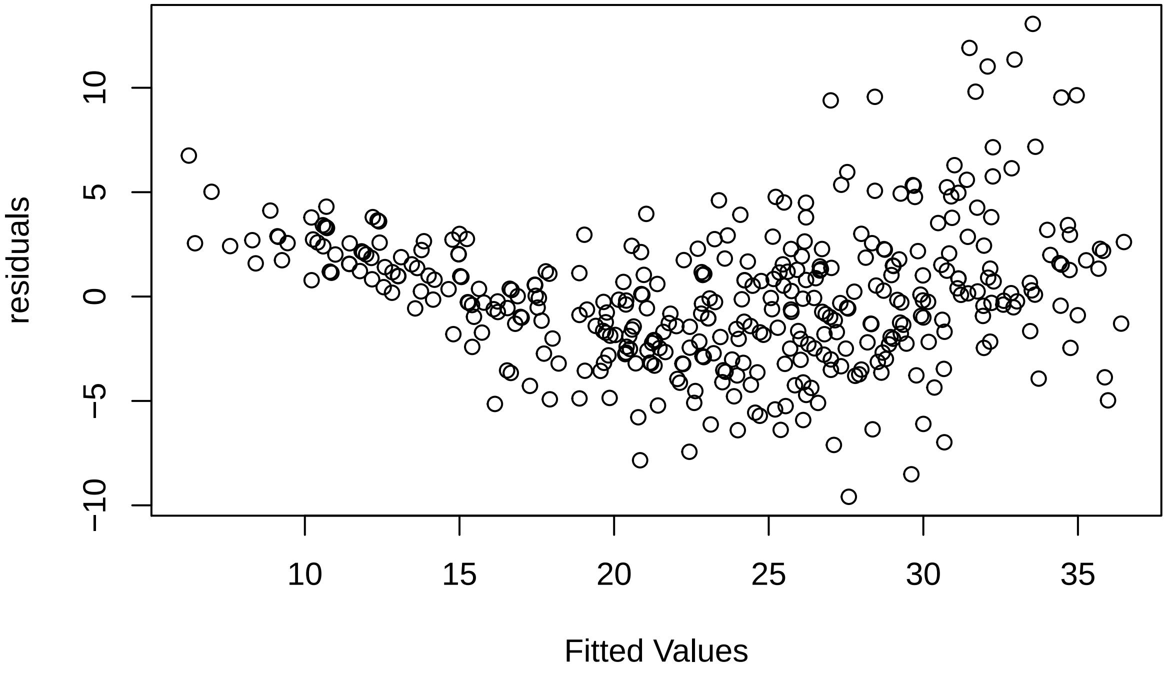
```
##
## Call:
## lm(formula = mpg ~ cylinders + displacement + horsepower + weight +
##     acceleration + year + origin, data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5903 -2.1565 -0.1169  1.8690 13.0604
##
## Coefficients:
##                Estimate Std. Error t value 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
```

(d)

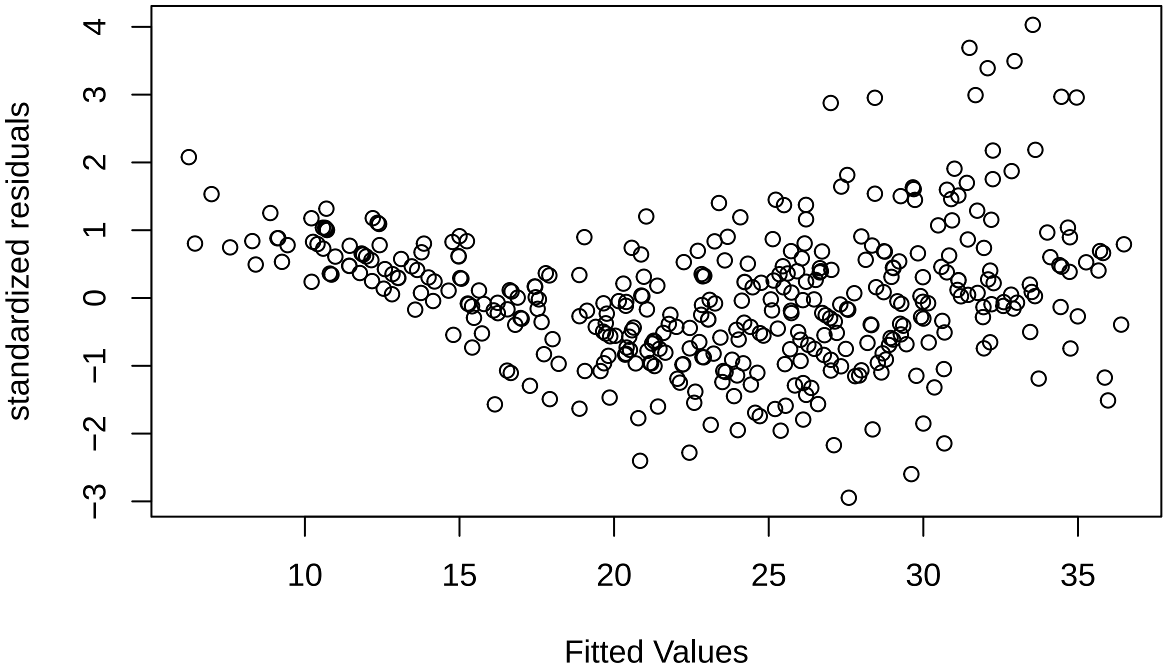
Use the plot() function to produce diagnostic plots of the linear regression fit. Comment on any problems you see with the fit. Do the residual plots suggest any unusually large outliers? Does the leverage plot identify any observations with unusually high leverage?

The residual plots does not suggest any unusually large outliers that deviate substantially from the overall pattern. The residual plot does suggest that there is a non-linear fit and the error term does not have constant or approximately constant values for each observation. The residual plot also suggets that there is a lot of observed values that differ substantially from the corresponding predicted values. The leverage plot indicates that there is one observation that is a high-leveraged point. This point has a leverage of about .18

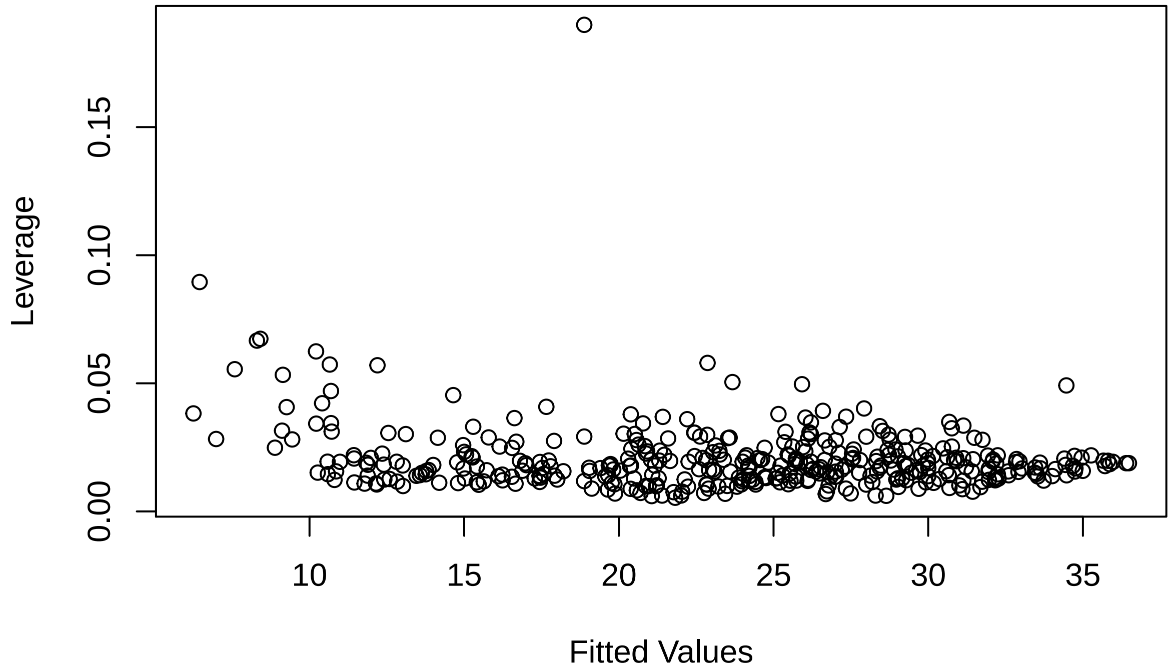
```r
plot(predict(lin.fit), residuals(lin.fit), xlab='Fitted Values', ylab='residuals')
```



```r
plot(predict(lin.fit), rstudent(lin.fit), xlab='Fitted Values', ylab='standardized residuals')
```



```
plot(predict(lin.fit), hatvalues(lin.fit), xlab='Fitted Values', ylab='Leverage')
```



(e)

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

The horsepower:weight interaction and horsepower:year interaction are both statistically significant at alpha level of .001. This means if we performed a hypothesis at alpha level of .001, this would be statistical evidence to reject the null in favor of the alternative hypothesis. In other words, there is statistical evidence for a relationship between the interaction term and the response variable.

```
lin.fit <- lm(mpg~cylinders+displacement+origin+horsepower*weight+year*horsepower+weight*acceleration, 
summary(lin.fit)
```

```
##
## Call:
## lm(formula = mpg ~ cylinders + displacement + origin + horsepower *
##     weight + year * horsepower + weight * acceleration, data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.9131 -1.4557 -0.0789  1.3418 11.1531
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -6.502e+01  1.282e+01  -5.073 6.13e-07 ***
## cylinders              2.744e-01  2.812e-01   0.976 0.329684
## displacement          -3.826e-03  6.845e-03  -0.559 0.576504
## origin              7.607e-01  2.423e-01   3.140 0.001824 **
## horsepower          3.869e-01  1.071e-01   3.612 0.000344 ***
## weight             -5.831e-03  2.258e-03  -2.582 0.010186 *
## year                1.494e+00  1.311e-01  11.396  < 2e-16 ***
## acceleration        3.318e-01  2.824e-01   1.175 0.240757
## horsepower:weight   3.628e-05  6.890e-06   5.266 2.34e-07 ***
## horsepower:year    -7.625e-03  1.310e-03  -5.821 1.25e-08 ***
## weight:acceleration -1.642e-04  9.686e-05  -1.696 0.090755 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.812 on 381 degrees of freedom
## Multiple R-squared:  0.8735, Adjusted R-squared:  0.8702
## F-statistic: 263.1 on 10 and 381 DF,  p-value: < 2.2e-16
```

(f)

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

Displacement, horsepower, weight variables have a non-linear relationship with mpg, so I will focus on these variables. This was determined by looking at the scatterplot matrix. The model with no transformed variables show that dispalcement is statistically significant at alpha of .01, horsepower is not statistically significant, and weight is statistically significant at alpha of .001. In the model with the transformed variables, displacement is no longer significant, horsepower is now statistically significant at alpha of .01, weight is no longer significant at alpha of .001, log(displacement) is not significant, log(horsepower) is significant at alpha of .001, and sqrt(weight) is significant at alpha of .01. Because log(displacement) is not significant, it might not be necessary to include this transformed variable. The statistical evidence points to keeping log(horsepower) and sqrt(weight) in the model.

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

```
##
## Call:
## lm(formula = mpg ~ cylinders + displacement + horsepower + weight +
##     acceleration + year + origin, 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
```

```
lin.fit <- lm(mpg~cylinders+displacement+horsepower+weight+acceleration+year+origin+log(displacement)+l
summary(lin.fit)
```

```
##
## Call:
## lm(formula = mpg ~ cylinders + displacement + horsepower + weight +
##     acceleration + year + origin + log(displacement) + log(horsepower) +
##     sqrt(weight), data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3226 -1.4965 -0.1677  1.4193 12.0411
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        94.468832  11.155143   8.469 5.45e-16 ***
## cylinders          -0.121518   0.295694  -0.411  0.68133
## displacement        0.009677   0.014169   0.683  0.49503
## horsepower          0.099910   0.030893   3.234  0.00133 **
## weight              0.008354   0.004519   1.849  0.06530 .
## acceleration       -0.206725   0.100077  -2.066  0.03954 *
## year                0.771186   0.045195  17.063  < 2e-16 ***
## origin              0.633179   0.268151   2.361  0.01871 *
## log(displacement)  -2.201213   2.835058  -0.776  0.43798
## log(horsepower)   -17.127194   3.658217  -4.682 3.96e-06 ***
## sqrt(weight)       -1.370854   0.526448  -2.604  0.00958 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.909 on 381 degrees of freedom
## Multiple R-squared:  0.8646, Adjusted R-squared:  0.861
## F-statistic: 243.3 on 10 and 381 DF,  p-value: < 2.2e-16
```

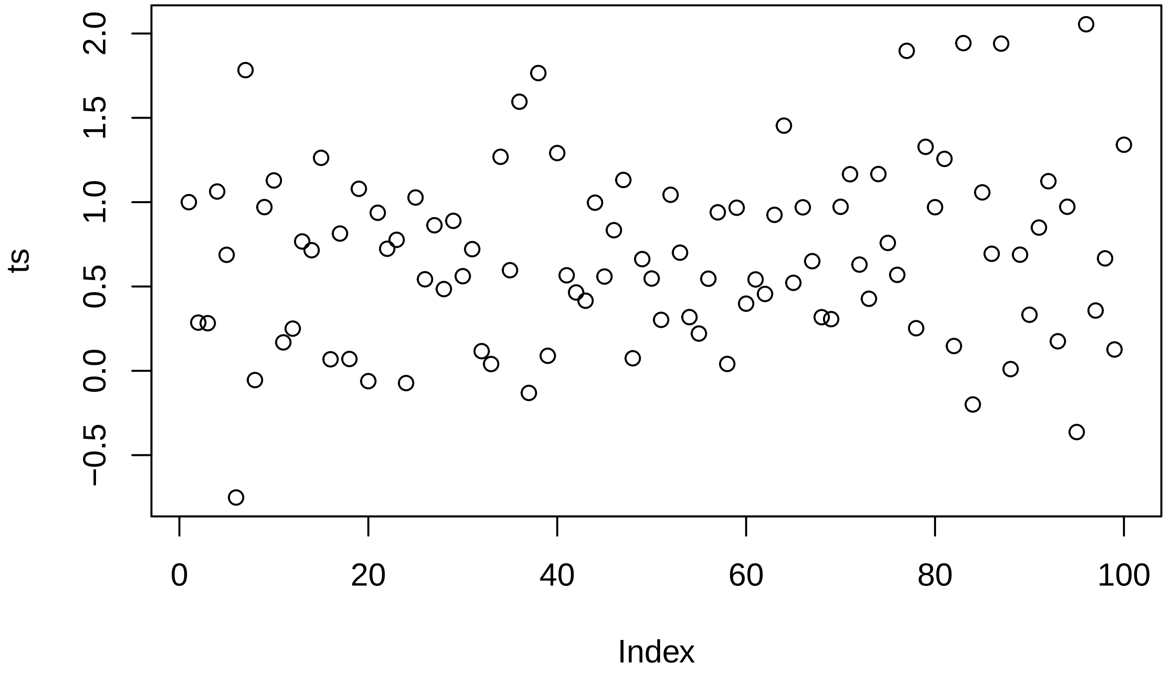
Extra 14

(a)

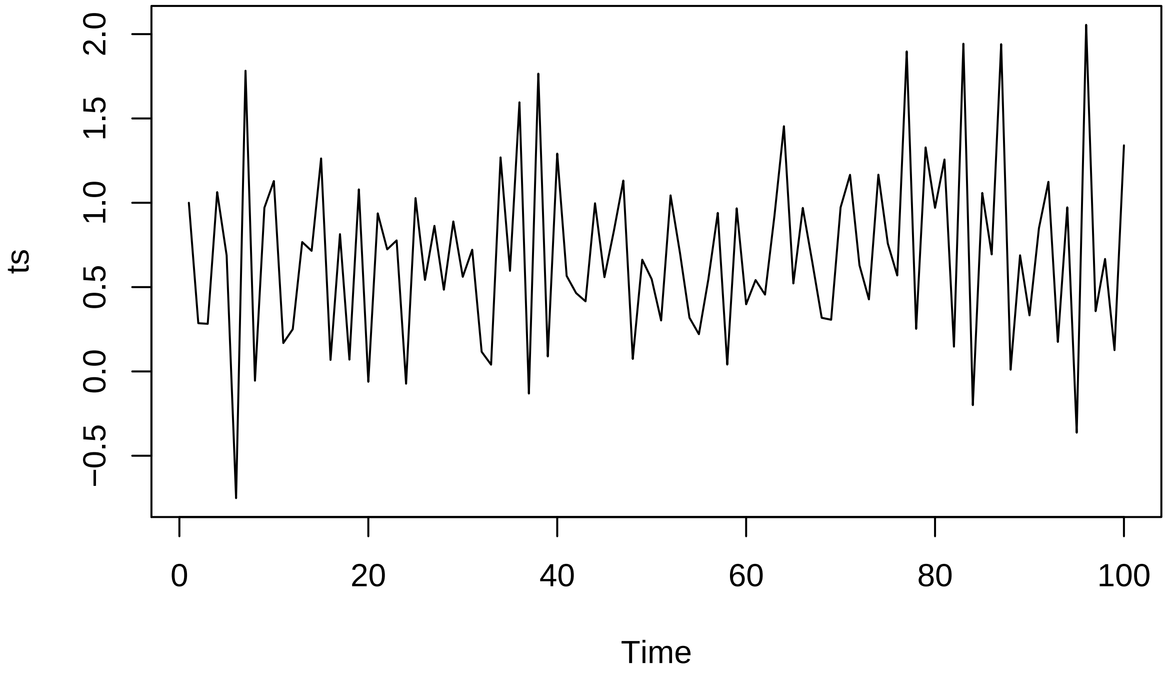
Simulate a time series X of length N=100 from the above formula, using the lag k=1, coefficients $\beta_0 = 1$ and $\beta_1 = -.5$ and error terms $\epsilon_t = N(0, 0.2^2)$. The formula tells you how to make X_t for $t \geq k+1$ from X_k. Choose X_1 arbitrarily. Plot X as a vector. Convert X into a timeseries object with function as.ts() and plot it again. Describe the plot.

The time series plot has each point connected by a line. The line connects point t to t+1 for all points.

```
simTS <- function(x1,b1){
  b0 <- 1
  b1 <- b1
  xt <- rep(0, 100)
  xt[1] <- x1
  for(i in 2:100){
    e <- rnorm(1, 0, .4)
    xt[i] = b0 + b1*xt[i-1] + e
  }
  return(xt)
}
ts <- simTS(1,-.5)
plot(ts)
```



```
ts <- as.ts(ts)
plot(ts)
```

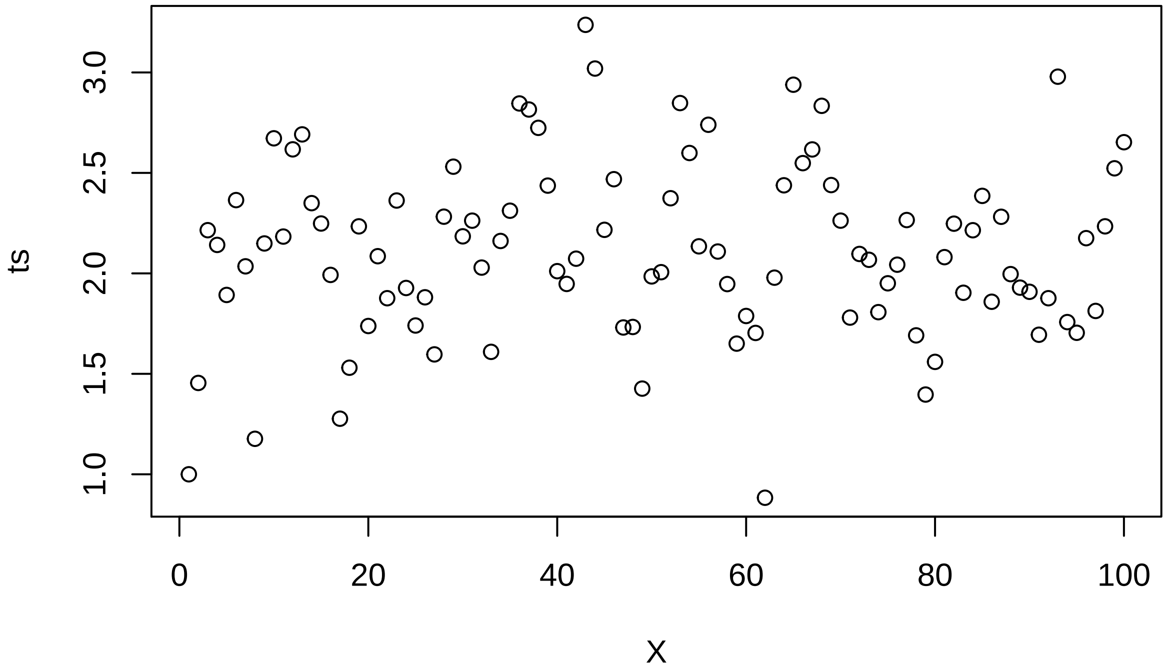


(b)

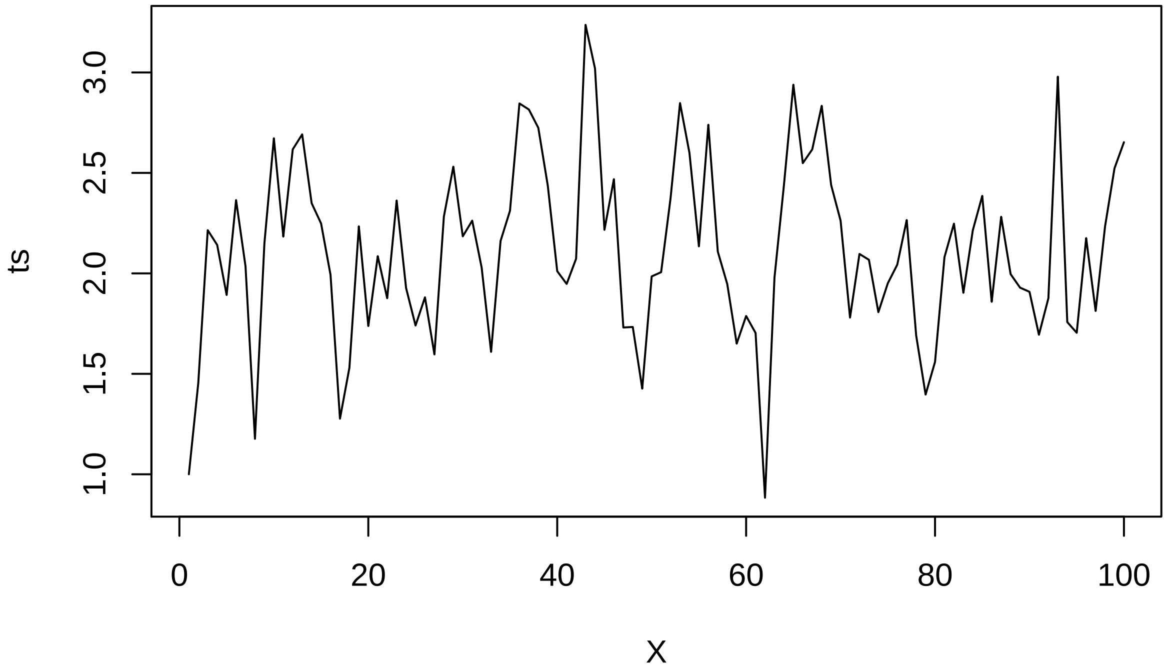
Repeat part a) with $\beta_0 = 1$, $\beta_1 = .5$. How does the plot change?

The change from point t to point t+1 is less abrupt. It looks as if the time series is a little more consistant.

```
ts <- simTS(1,.5)
plot(ts, xlab='X')
```



```
ts <- as.ts(ts)
plot(ts, xlab='X')
```

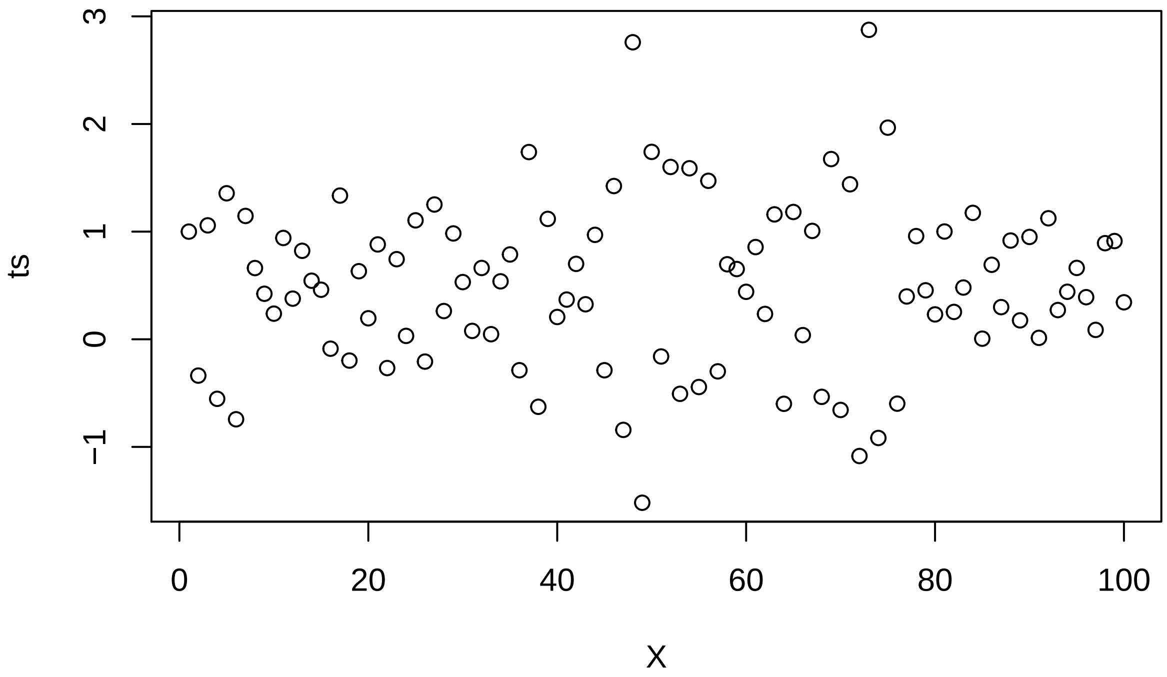


(c)

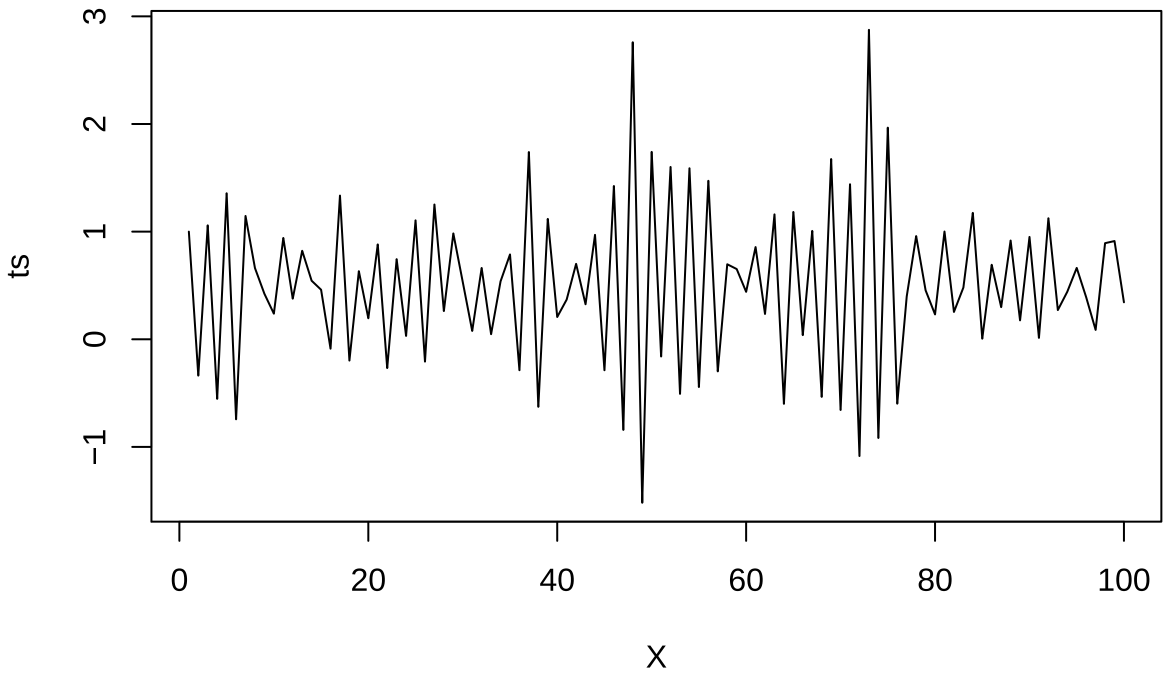
Repeat part a) with $\beta_0 = 1$, $\beta_1 = -.9$. How does the plot change?

There is more spikes in the time series plot. This corresponds to more abrupt changes from point t to t+1.

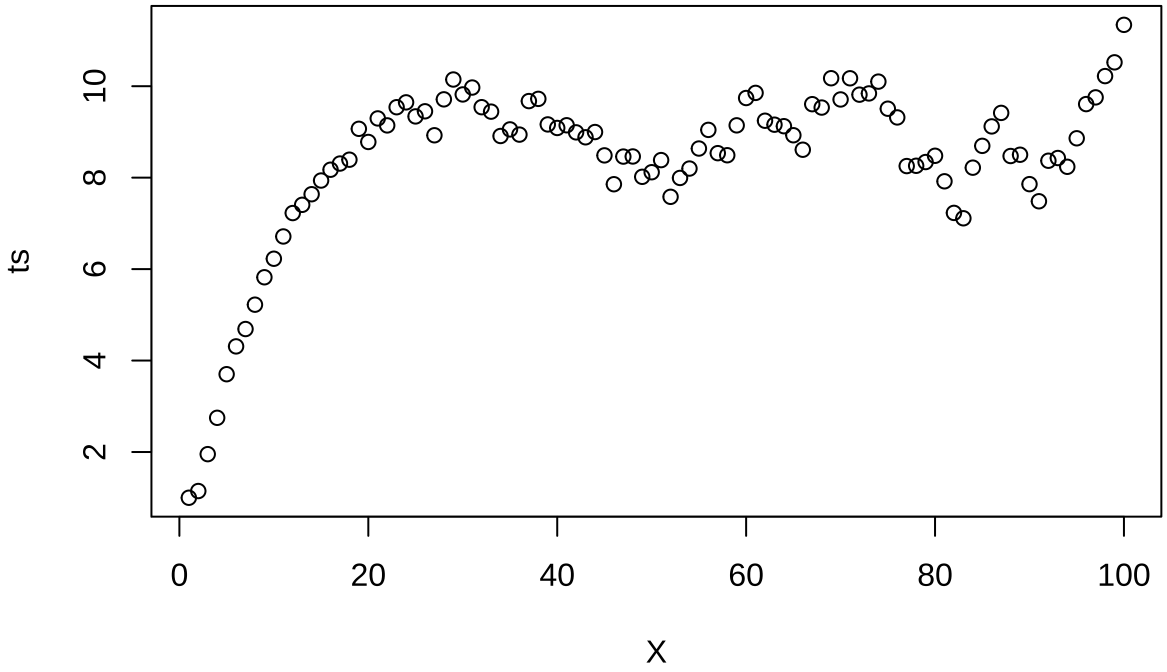
```
ts <- simTS(1,-.9)
plot(ts, xlab='X')
```



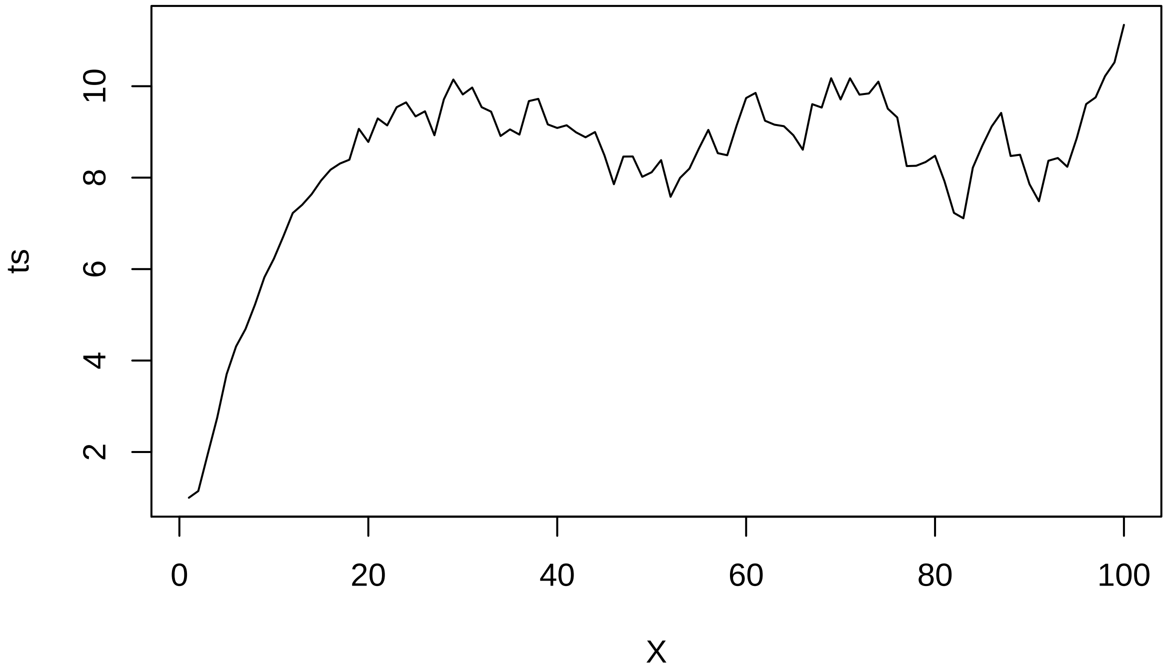
```
ts <- as.ts(ts)
plot(ts, xlab='X')
```



```r
ts <- simTS(1,.9)
plot(ts, xlab='X')
```



```
ts <- as.ts(ts)
plot(ts, xlab='X')
```



Extra 15

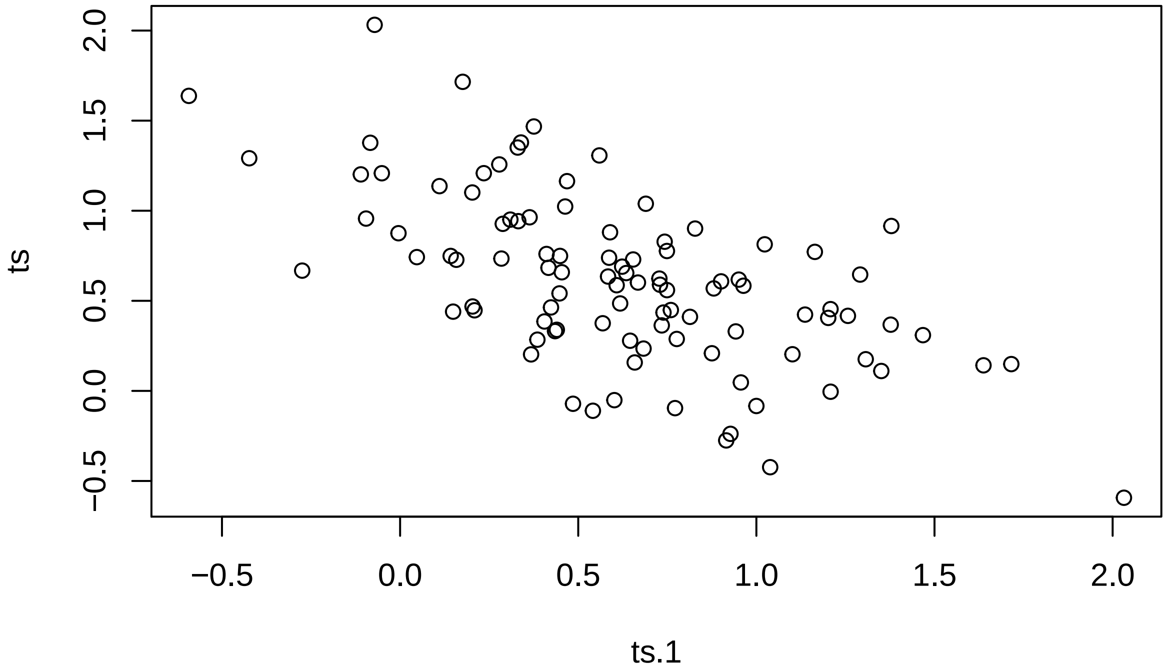
Simulate a time series X as in the previous problem $N = 100$ observations, lag $k = 1$, $\beta_0 = 1$, $\beta_1 = -.5$, and $\epsilon_t = N(0, 0.2^2)$.

(a)

Make a scatterplot of X_t against X_{t-1} for $t = 2, ...N$ and describe it.

The plot shows a clear negative relationship between X_t and X_{t-1} plus some noise.

```
ts <- simTS(1,-.5)
ts.1 <- rep(0, 100)
for (i in 2:100){
  ts.1[i] <- ts[i-1]
}
ts <- ts[-1]
ts.1 <- ts.1[-1]
plot(x=ts.1, y=ts)
```



(b)

Create a data frame of $N-1$ observations and 2 columns that contains (X_{t-1}, X_t) in row t. Use this to fit a linear model to predict X_t from X_{t-1}. Compare the estimated coefficients to the β_i. Also compare the residual standard error to the standard deviation of the ϵ_t term. Summarize your results and observations.

The estimated intercept is 1.1, which is close to 1. The estimated slope is $-.6$, which is close to $-.5$. Both coefficients are statistically significant at alpha level of .001, which means there is statistical evidence that the estimated coefficients are not 0. The residual standard error is the estimated standard deviation of the error term. In this case, they are very similar to each other.

```
data <- data.frame(X=ts.1, Y=ts)
lm.fit <- lm(Y~X, data=data)
summary(lm.fit)
```

```
##
## Call:
## lm(formula = Y ~ X, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.77466 -0.24717  0.02587  0.21609  0.98636
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.00067    0.06282  15.929  < 2e-16 ***
## X           -0.62526    0.08103  -7.716 1.07e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3716 on 97 degrees of freedom
## Multiple R-squared:  0.3804, Adjusted R-squared:  0.374
## F-statistic: 59.54 on 1 and 97 DF,  p-value: 1.069e-11
```