DS-6030 Homework Module 2

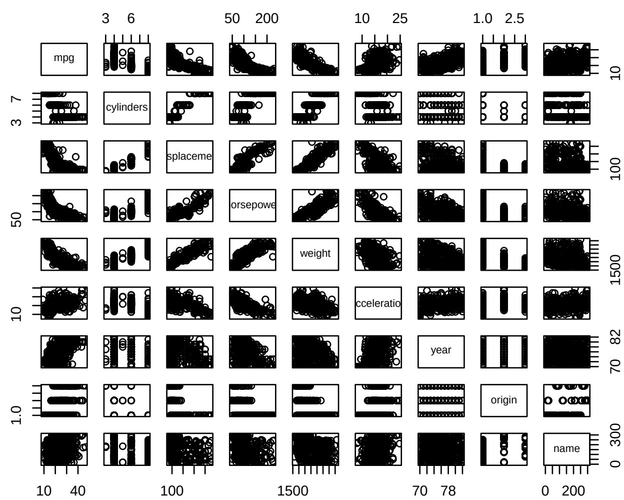
Matt Scheffel

DS 6030 | Spring 2022 | University of Virginia

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.

```
library("ISLR2")
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.

```
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
```

```r
# "name" is the last column
cor(Auto[1:8])
```

```
#>                     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.

```r
model1 = lm(mpg ~. -name, data = Auto)
summary(model1)
```

```
#> 
#> 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
```

Comment on the output. For instance:

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

Yes, multiple predictors from this model have a relationship with the response. We can tell due to their associated p-values being significant.

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

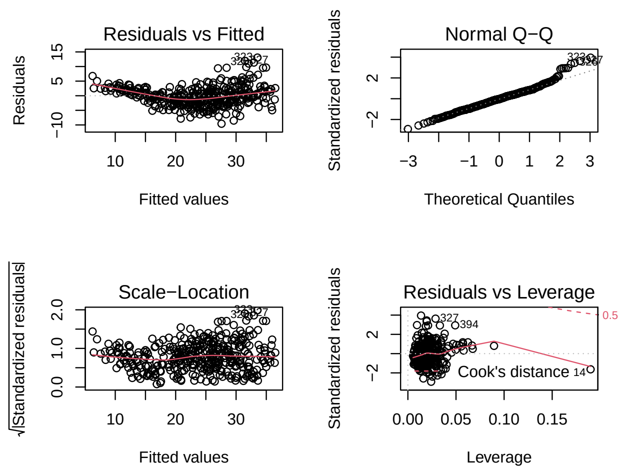
Displacement, weight, year, and origin have a statistically significant relationship to the response.

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

The coefficient for the year variable suggests that the average effect of an increase of 1 year is an increase of 0.7507727 in mpg, when all other predictors are held constant.

(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(model1)
```



The residual plot has U-shape pattern that suggests non-linear data. A few of the residuals in the upper right hand corner could be considered large outliers.However, the Residuals vs. Leverage graph shows no observations above the Cook's distance red dotted line that indicate unusually high leverage.

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

```
# two most correlated pairs
model2 <- lm(mpg ~ cylinders * displacement + displacement * weight, data = Auto[, 1:8])
summary(model2)
```

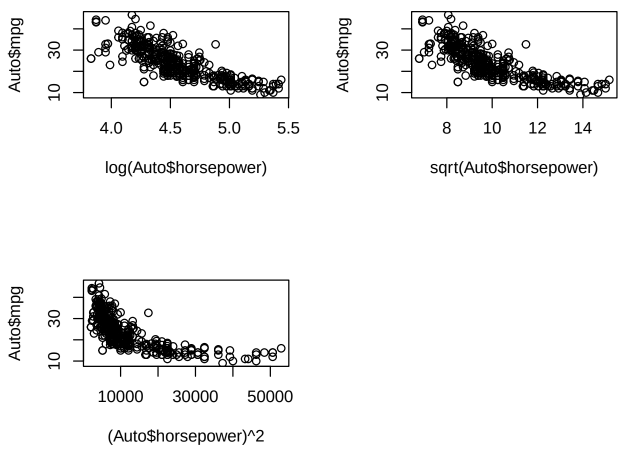
```
#>
#> Call:
#> lm(formula = mpg ~ cylinders * displacement + displacement *
#>     weight, data = Auto[, 1:8])
#>
#> Residuals:
#>      Min       1Q   Median       3Q      Max
#> -13.2934  -2.5184  -0.3476   1.8399  17.7723
#>
#> Coefficients:
#>                          Estimate Std. Error t value Pr(>|t|)
#> (Intercept)             5.262e+01  2.237e+00  23.519  < 2e-16 ***
#> cylinders               7.606e-01  7.669e-01   0.992    0.322
#> displacement           -7.351e-02  1.669e-02  -4.403 1.38e-05 ***
```

```
#> weight                  -9.888e-03  1.329e-03  -7.438 6.69e-13 ***
#> cylinders:displacement -2.986e-03  3.426e-03  -0.872    0.384
#> displacement:weight     2.128e-05  5.002e-06   4.254 2.64e-05 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 4.103 on 386 degrees of freedom
#> Multiple R-squared:  0.7272, Adjusted R-squared:  0.7237
#> F-statistic: 205.8 on 5 and 386 DF,  p-value: < 2.2e-16
```

Based on this model and the p-values, the interaction between displacement and weight appears to be statistically signifcant.

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

```r
par(mfrow = c(2, 2))
plot(log(Auto$horsepower), Auto$mpg)
plot(sqrt(Auto$horsepower), Auto$mpg)
plot((Auto$horsepower)^2, Auto$mpg)
```



The log transformation helps to create the plot that appears to be the most linear.

14. This problem focuses on the collinearity problem.

(a) Perform the following commands in R.

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

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?

Form of the linear model: $Y = 2 + 2X1 + 0.3X2 + \epsilon$

Regression coefficients: 2, 2, and 0.3

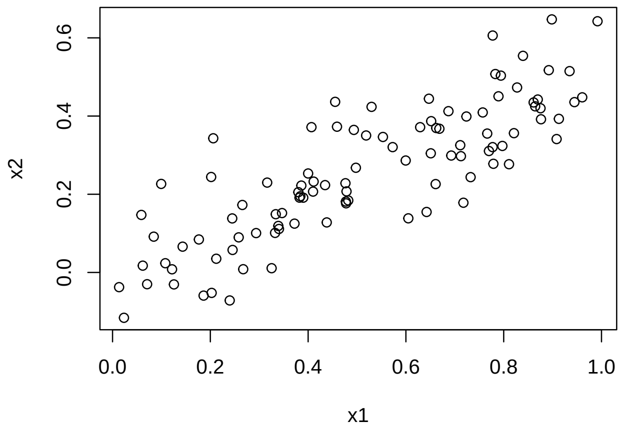
(b) What is the correlation between `x1` and `x2`? Create a scatterplot displaying the relationship between the variables.

```r
cor(x1, x2)
```

```
#> [1] 0.8351212
```

The correlation is 0.8351212.

```
plot(x1, x2)
```



(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 H0 : $\beta_1 = 0$? How about the null hypothesis H0 : $\beta_2 = 0$?

```
model3 <- lm(y ~ x1 + x2)
summary(model3)
```

```
#>
#> 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
```

Coefficient estimates: $\hat{\beta}_0 = 2.1305$, $\hat{\beta}_1 = 1.4396$, and $\hat{\beta}_2 = 1.0097$. The values are not good estimates of the true β_0, β_1, and β_2. $\hat{\beta}_0$ is the closest to its true value.

For β_1, we cannot reject the null hypothesis at a 95% level of confidence, but we can at the 99% confidence level.

For β_2, we reject the null hypothesis.

(d) Now fit a least squares regression to predict y using only x1. Comment on your results. Can you reject the null hypothesis $H0$: $\beta_1 = 0$?

```r
model4 <- lm(y ~ x1)
summary(model4)
```

```
#>
#> 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
```

In this model, the coefficient for $x1$ differs from the previous model that used $x1$ and $x2$ as predictors. In this model, $x1$ is significant with a fairly low p-value and we will reject the null hypothesis, $H0$.

(e) Now fit a least squares regression to predict `y` using only `x2`. Comment on your results. Can you reject the null hypothesis H0: $\beta_2 = 0$?

```r
model5 <- lm(y ~ x2)
summary(model5)
```

```
#>
#> 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
```

In this model, the coefficient for $x2$ differs from the previous model that used $x1$ and $x2$ as predictors. In this model, $x2$ is significant with a fairly low p-value and we will reject the null hypothesis, $H0$.

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

Yes, the results from (c)-(e) appear to contradict each other. The MLR model does not regard $x1$ and $x2$ as significant predictors, but the SLR models show that $x1$ and $x2$ are significant predictors. However, collinearity may help explain why these variables seemed insignificant in the MLR model.

(g) Now suppose we obtain one additional observation, which was unfortunately mismeasured.

```
x1 <- c(x1, 0.1)
x2 <- c(x2, 0.8)
y <- c(y, 6)
```

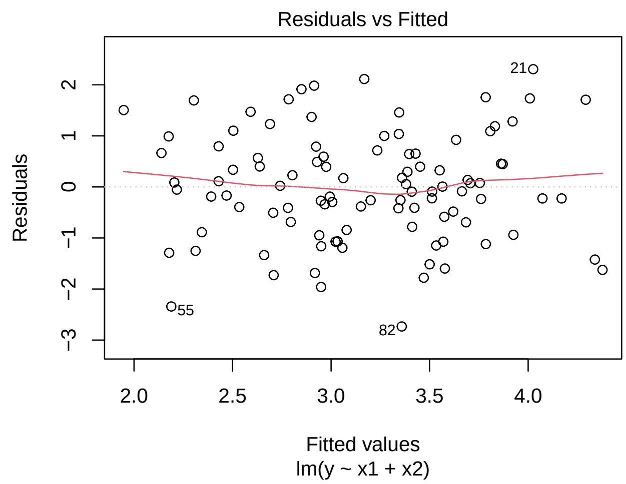
Re-fit the linear models from (c) to (e) using this new data. What effect does this new observation have on the each of the models? In each model, is this observation an outlier? A high-leverage point? Both? Explain your answers.

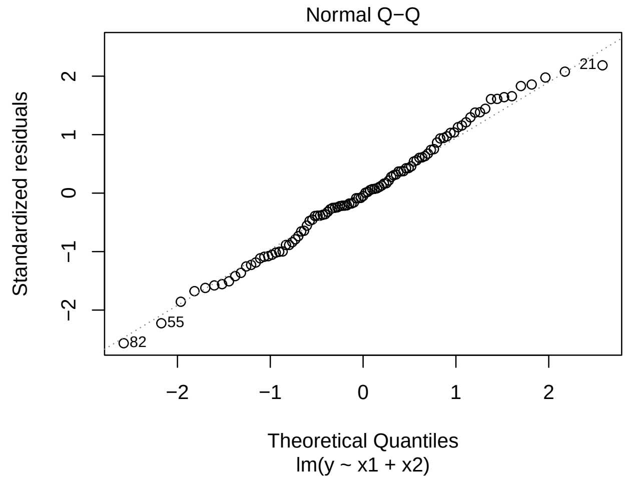
```
model6 <- lm(y ~ x1 + x2)
model7 <- lm(y ~ x1)
model8 <- lm(y ~ x2)
```

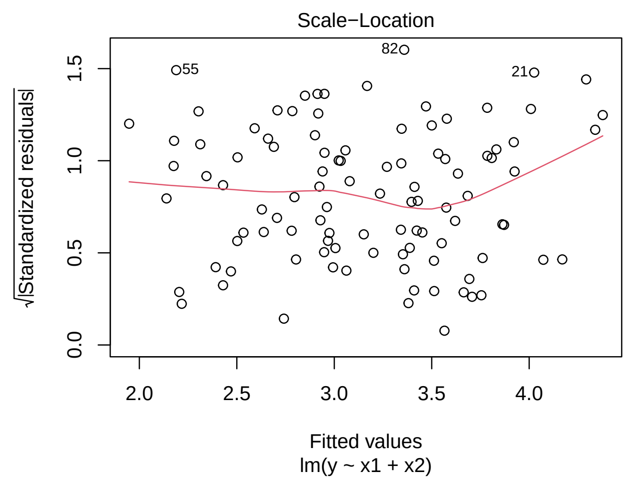
```
summary(model6)
```

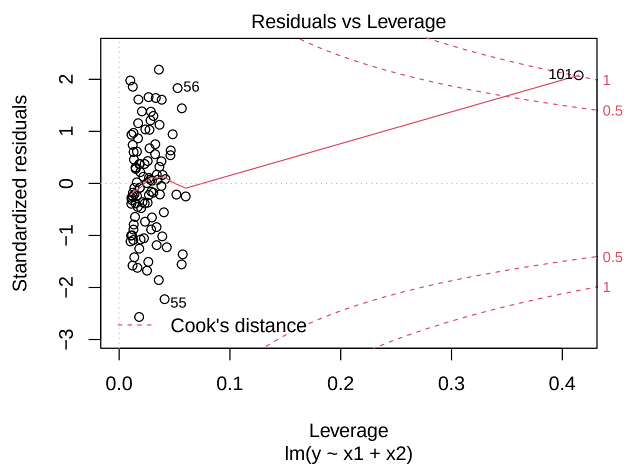
```
#>
#> Call:
#> lm(formula = y ~ x1 + x2)
#>
#> Residuals:
#>      Min       1Q   Median       3Q      Max
#> -2.73348 -0.69318 -0.05263  0.66385  2.30619
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)   2.2267     0.2314   9.624 7.91e-16 ***
#> x1            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
```

```
plot(model6)
```







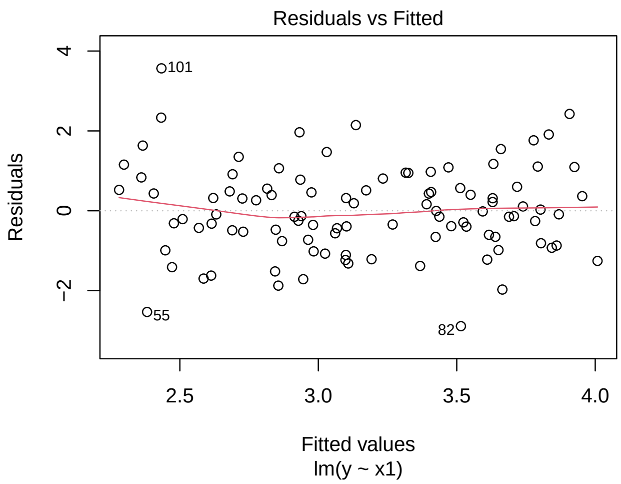


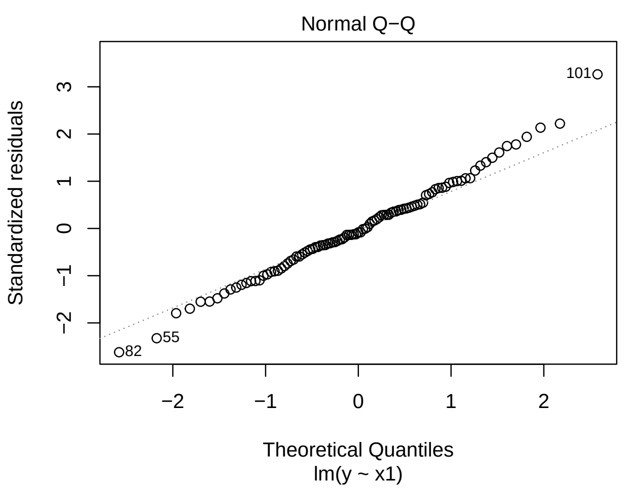

```
summary(model7)
```

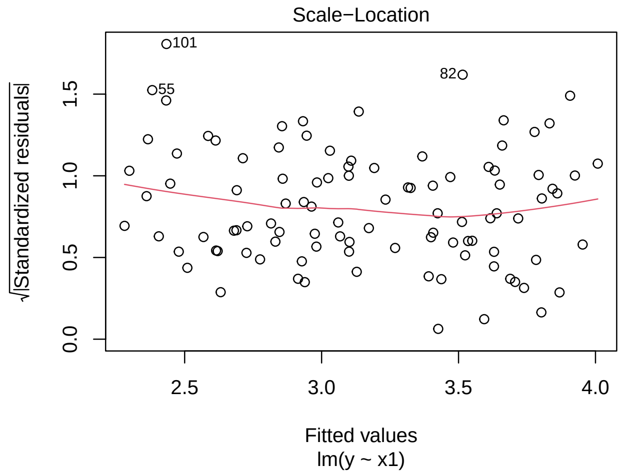
```
#>
#> Call:
#> lm(formula = y ~ x1)
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -2.8897 -0.6556 -0.0909  0.5682  3.5665
```

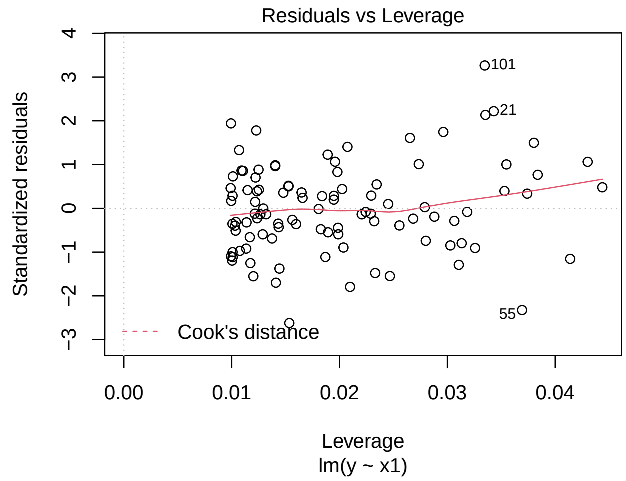
```
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)    
#> (Intercept)   2.2569     0.2390   9.445 1.78e-15 ***
#> x1            1.7657     0.4124   4.282 4.29e-05 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 1.111 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
```

```
plot(model7)
```





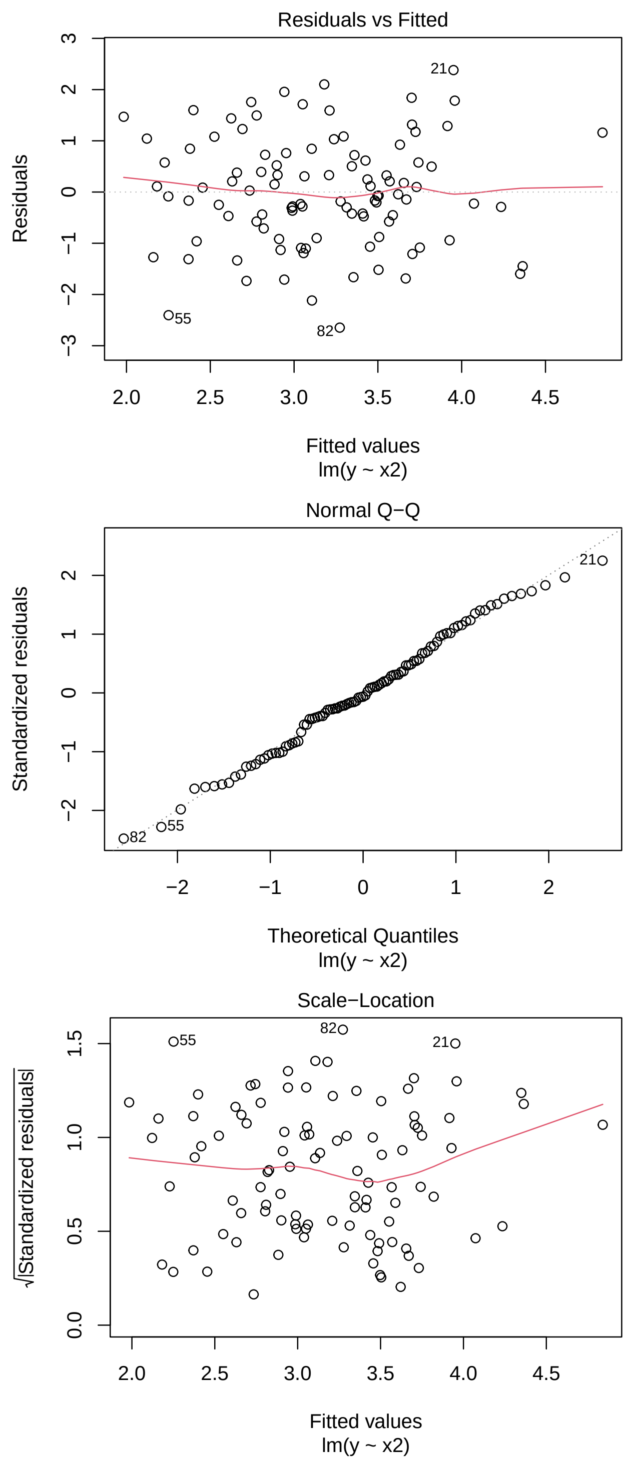


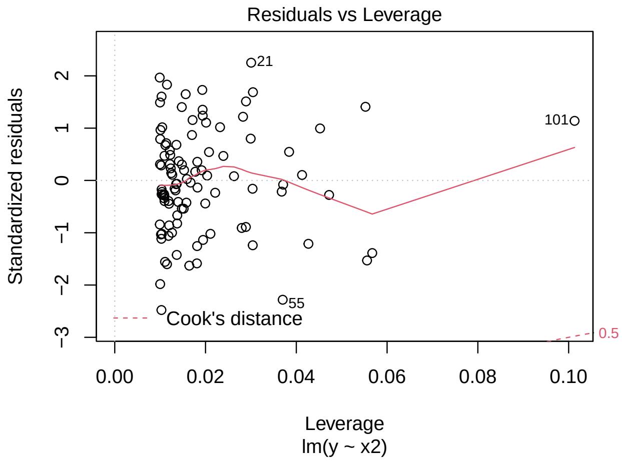



```
summary(model8)
```

```
#>
#> 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
```

```
plot(model8)
```


Residuals vs Fitted
Residuals
Fitted values
lm(y ~ x2)
21
55
82
Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(y ~ x2)
21
55
82
Scale−Location
√|Standardized residuals|
Fitted values
lm(y ~ x2)
55
82
21



In the first new model using $x1$ and $x2$ as predictors, the last point is a high-leverage point. R squared is slightly higher in this model and $x2$ is significantly significant.

In the second new model with $x1$ as the predictor, the last point can be considered an outlier. R squared decreases in this model and $x1$ is significant.

In the third new model with $x2$ as the predictor, there does not appear to be a significant leverage point or outlier. R squared increases in this model and $x1$ is significant.

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.

```
#library(ISLR2)
Boston <- ISLR2::Boston
head(Boston)
```

```
#>      crim zn indus chas   nox    rm  age    dis rad tax ptratio lstat medv
#> 1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3  4.98 24.0
#> 2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8  9.14 21.6
#> 3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8  4.03 34.7
#> 4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7  2.94 33.4
#> 5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7  5.33 36.2
#> 6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7  5.21 28.7
```

```
attach(Boston)
```

```
model9 <- lm(crim ~zn)
summary(model9)
```

```
#>
#> Call:
#> lm(formula = crim ~ 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 ***
#> 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
```

```
model10 <- lm(crim ~ indus)
summary(model10)
```

```
#>
#> Call:
#> lm(formula = crim ~ 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 **
#> 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
```

```
chas <- as.factor(chas)
model11 <- lm(crim ~ chas)
summary(model11)
```

```
#>
#> Call:
#> lm(formula = crim ~ 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 ***
#> chas1        -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
```

```
model12 <- lm(crim ~ nox)
summary(model12)
```

```
#>
#> Call:
#> lm(formula = crim ~ 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 ***
#> 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
```

```
fit.rm <- lm(crim ~ rm)
summary(fit.rm)
```

```
#>
#> Call:
#> lm(formula = crim ~ 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 ***
#> 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
```

```
model13 <- lm(crim ~ age)
summary(model13)
```

```
#>
#> Call:
#> lm(formula = crim ~ 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 ***
#> 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
```

```
model14 <- lm(crim ~ dis)
summary(model14)
```

```
#>
#> Call:
#> lm(formula = crim ~ 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 ***
#> 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
```

```
model15 <- lm(crim ~ rad)
summary(model15)
```

```
#>
#> Call:
#> lm(formula = crim ~ 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 ***
#> 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
model16 <- lm(crim ~ tax)
summary(model16)
```

```
#> 
#> Call:
#> lm(formula = crim ~ 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 ***
#> 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
model17 <- lm(crim ~ ptratio)
summary(model17)
```

```
#> 
#> Call:
#> lm(formula = crim ~ 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 ***
#> 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
model18 <- lm(crim ~ lstat)
summary(model18)
```

```
#> 
#> Call:
#> lm(formula = crim ~ lstat)
#> 
#> Residuals:
#>     Min      1Q  Median      3Q     Max 
```

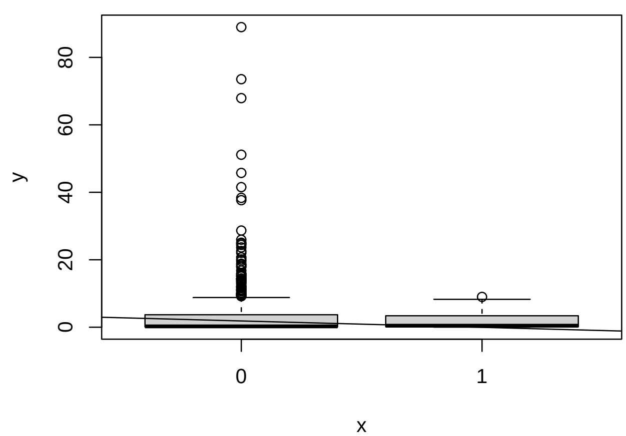
```
#> -13.925  -2.822  -0.664   1.079  82.862
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept) -3.33054    0.69376  -4.801 2.09e-06 ***
#> lstat        0.54880    0.04776  11.491  < 2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 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
```

```
model19 <- lm(crim ~ medv)
summary(model19)
```

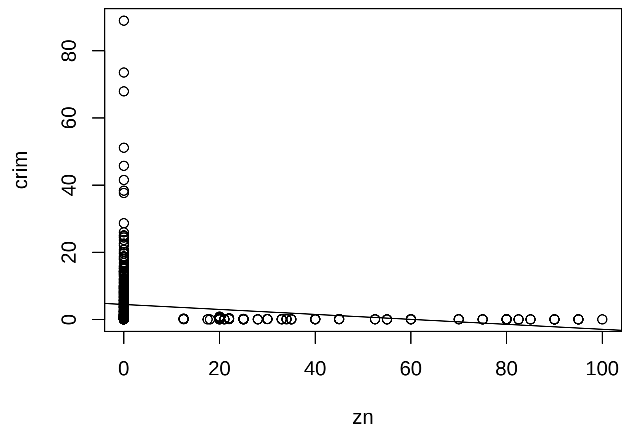
```
#>
#> Call:
#> lm(formula = crim ~ medv)
#>
#> Residuals:
#>    Min     1Q Median     3Q    Max
#> -9.071 -4.022 -2.343  1.298 80.957
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept) 11.79654    0.93419   12.63   <2e-16 ***
#> medv        -0.36316    0.03839   -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
```

Each predictors besides "chas" has a p-value of less than 0.05, indicating that there is a statistically significant association between those predictors and the response.

```
plot(chas,crim)
abline(model11)
```



```r
plot(zn,crim)
abline(model9)
```



(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 $H_0 : \beta_j = 0$?

```r
model.all.variables <- lm(crim ~ ., data = Boston)
summary(model.all.variables)
```

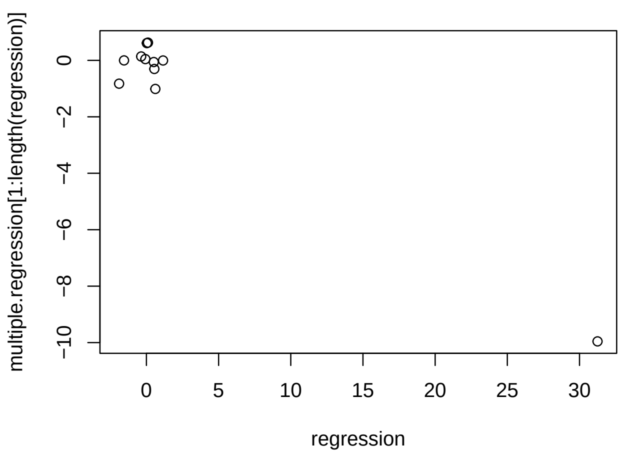
```
#>
#> Call:
#> lm(formula = crim ~ ., data = Boston)
#>
#> Residuals:
#>    Min     1Q Median     3Q    Max
#> -8.534 -2.248 -0.348  1.087 73.923
#>
#> Coefficients:
#>               Estimate Std. Error t value Pr(>|t|)
#> (Intercept) 13.7783938  7.0818258   1.946 0.052271 .
#> zn           0.0457100  0.0187903   2.433 0.015344 *
#> indus       -0.0583501  0.0836351  -0.698 0.485709
#> chas        -0.8253776  1.1833963  -0.697 0.485841
#> nox         -9.9575865  5.2898242  -1.882 0.060370 .
#> rm           0.6289107  0.6070924   1.036 0.300738
#> age         -0.0008483  0.0179482  -0.047 0.962323
#> dis         -1.0122467  0.2824676  -3.584 0.000373 ***
#> rad          0.6124653  0.0875358   6.997 8.59e-12 ***
#> tax         -0.0037756  0.0051723  -0.730 0.465757
#> ptratio     -0.3040728  0.1863598  -1.632 0.103393
#> lstat        0.1388006  0.0757213   1.833 0.067398 .
#> medv        -0.2200564  0.0598240  -3.678 0.000261 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 6.46 on 493 degrees of freedom
#> Multiple R-squared:  0.4493, Adjusted R-squared:  0.4359
#> F-statistic: 33.52 on 12 and 493 DF,  p-value: < 2.2e-16
```

A relatively low R squared value suggests that this MLR model does not fit the data well. In this fitted multiple regression model "zn", "dis", "rad", and "medv" are found to be statistically significant. The other variables have high p-values and we do not reject the null hypothesis for them. Thus, we reject the null hypothesis for "zn", "dis", "rad", and "medv".

(c) How do your results from (a) compare to your results from (b)? Create a plot displaying the univariate regression coefficients from (a) on the x-axis, and the multiple regression coefficients from (b) on the y-axis. That is, each predictor is displayed as a single point in the plot. Its coefficient in a simple linear regression model is shown on the x-axis, and its coefficient estimate in the multiple linear regression model is shown on the y-axis.

```
regression <- vector("numeric",0)
regression <- c(regression, model9$coefficient[2])
regression <- c(regression, model10$coefficient[2])
regression <- c(regression, model11$coefficient[2])
regression <- c(regression, model12$coefficient[2])
regression <- c(regression, model13$coefficient[2])
regression <- c(regression, model14$coefficient[2])
regression <- c(regression, model15$coefficient[2])
regression <- c(regression, model16$coefficient[2])
regression <- c(regression, model17$coefficient[2])
regression <- c(regression, model18$coefficient[2])
regression <- c(regression, model19$coefficient[2])
multiple.regression <- vector("numeric", 0)
multiple.regression <- c(multiple.regression, model.all.variables$coefficients)
multiple.regression <- multiple.regression[-1]

#plot(regression, multiple.regression)
plot(regression, multiple.regression[1:length(regression)])
```




```
#unsure why original plot will not work - error says x and y lengths differ
```

The results differ because univariate regression and multiple regression have significantly different coefficients. The slope of the univariate regression model shows the average effect of an increase in the predictor while ignoring all the other predictors from the dat However, the multiple regression holds other predictors fixed, and the slope represents the average effect of an increase in the predictor.

(d) Is there evidence of non-linear association between any of the predictors and the response? To answer this question, for each predictor X, fit a model of the form

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon.$$

```
model.1 <- lm(crim ~ poly(zn, 3))
summary(model.1)
```

```
#>
#> Call:
#> lm(formula = crim ~ poly(zn, 3))
#>
#> Residuals:
#>    Min     1Q Median     3Q    Max
#> -4.821 -4.614 -1.294  0.473 84.130
#>
#> Coefficients:
#>               Estimate Std. Error t value Pr(>|t|)
#> (Intercept)    3.6135     0.3722   9.709  < 2e-16 ***
#> poly(zn, 3)1 -38.7498     8.3722  -4.628  4.7e-06 ***
#> poly(zn, 3)2  23.9398     8.3722   2.859  0.00442 **
#> poly(zn, 3)3 -10.0719     8.3722  -1.203  0.22954
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 8.372 on 502 degrees of freedom
#> Multiple R-squared:  0.05824,    Adjusted R-squared:  0.05261
#> F-statistic: 10.35 on 3 and 502 DF,  p-value: 1.281e-06
```

```r
model.2 <- lm(crim ~ poly(indus, 3))
summary(model.2)
```

```
#>
#> Call:
#> lm(formula = crim ~ poly(indus, 3))
#>
#> Residuals:
#>    Min     1Q Median     3Q    Max
#> -8.278 -2.514  0.054  0.764 79.713
#>
#> Coefficients:
#>                  Estimate Std. Error t value Pr(>|t|)
#> (Intercept)        3.614      0.330  10.950  < 2e-16 ***
#> poly(indus, 3)1   78.591      7.423  10.587  < 2e-16 ***
#> poly(indus, 3)2  -24.395      7.423  -3.286  0.00109 **
#> poly(indus, 3)3  -54.130      7.423  -7.292  1.2e-12 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 7.423 on 502 degrees of freedom
#> Multiple R-squared:  0.2597, Adjusted R-squared:  0.2552
#> F-statistic: 58.69 on 3 and 502 DF,  p-value: < 2.2e-16
```

```r
model.3 <- lm(crim ~ poly(nox, 3))
summary(model.3)
```

```
#>
#> Call:
#> lm(formula = crim ~ poly(nox, 3))
#>
#> Residuals:
#>    Min     1Q Median     3Q    Max
#> -9.110 -2.068 -0.255  0.739 78.302
```

```
#>
#> Coefficients:
#>                Estimate Std. Error t value Pr(>|t|)
#> (Intercept)      3.6135     0.3216  11.237  < 2e-16 ***
#> poly(nox, 3)1  81.3720     7.2336  11.249  < 2e-16 ***
#> poly(nox, 3)2 -28.8286     7.2336  -3.985 7.74e-05 ***
#> poly(nox, 3)3 -60.3619     7.2336  -8.345 6.96e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 7.234 on 502 degrees of freedom
#> Multiple R-squared:  0.297,  Adjusted R-squared:  0.2928
#> F-statistic: 70.69 on 3 and 502 DF,  p-value: < 2.2e-16
```

```
model.4 <- lm(crim ~ poly(rm, 3))
summary(model.4)
```

```
#>
#> Call:
#> lm(formula = crim ~ poly(rm, 3))
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -18.485  -3.468  -2.221  -0.015  87.219
#>
#> Coefficients:
#>               Estimate Std. Error t value Pr(>|t|)
#> (Intercept)    3.6135     0.3703   9.758  < 2e-16 ***
#> poly(rm, 3)1 -42.3794     8.3297  -5.088 5.13e-07 ***
#> poly(rm, 3)2  26.5768     8.3297   3.191  0.00151 **
#> poly(rm, 3)3  -5.5103     8.3297  -0.662  0.50858
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 8.33 on 502 degrees of freedom
#> Multiple R-squared:  0.06779,    Adjusted R-squared:  0.06222
#> F-statistic: 12.17 on 3 and 502 DF,  p-value: 1.067e-07
```

```
model.5 <- lm(crim ~ poly(age, 3))
summary(model.5)
```

```
#>
#> Call:
#> lm(formula = crim ~ poly(age, 3))
#>
#> Residuals:
#>    Min     1Q Median     3Q    Max
#> -9.762 -2.673 -0.516  0.019 82.842
#>
#> Coefficients:
#>                Estimate Std. Error t value Pr(>|t|)
#> (Intercept)      3.6135     0.3485  10.368  < 2e-16 ***
#> poly(age, 3)1  68.1820     7.8397   8.697  < 2e-16 ***
#> poly(age, 3)2  37.4845     7.8397   4.781 2.29e-06 ***
#> poly(age, 3)3  21.3532     7.8397   2.724  0.00668 **
```

```
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 7.84 on 502 degrees of freedom
#> Multiple R-squared:  0.1742, Adjusted R-squared:  0.1693
#> F-statistic: 35.31 on 3 and 502 DF,  p-value: < 2.2e-16
```

```
model.6 <- lm(crim ~ poly(dis, 3))
summary(model.6)
```

```
#>
#> Call:
#> lm(formula = crim ~ poly(dis, 3))
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -10.757  -2.588   0.031   1.267  76.378
#>
#> Coefficients:
#>                 Estimate Std. Error t value Pr(>|t|)
#> (Intercept)       3.6135     0.3259  11.087  < 2e-16 ***
#> poly(dis, 3)1 -73.3886     7.3315 -10.010  < 2e-16 ***
#> poly(dis, 3)2  56.3730     7.3315   7.689 7.87e-14 ***
#> poly(dis, 3)3 -42.6219     7.3315  -5.814 1.09e-08 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 7.331 on 502 degrees of freedom
#> Multiple R-squared:  0.2778, Adjusted R-squared:  0.2735
#> F-statistic: 64.37 on 3 and 502 DF,  p-value: < 2.2e-16
```

```
model.7 <- lm(crim ~ poly(rad, 3))
summary(model.7)
```

```
#>
#> Call:
#> lm(formula = crim ~ poly(rad, 3))
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -10.381  -0.412  -0.269   0.179  76.217
#>
#> Coefficients:
#>                 Estimate Std. Error t value Pr(>|t|)
#> (Intercept)       3.6135     0.2971  12.164  < 2e-16 ***
#> poly(rad, 3)1 120.9074     6.6824  18.093  < 2e-16 ***
#> poly(rad, 3)2  17.4923     6.6824   2.618  0.00912 **
#> poly(rad, 3)3   4.6985     6.6824   0.703  0.48231
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 6.682 on 502 degrees of freedom
#> Multiple R-squared:    0.4, Adjusted R-squared:  0.3965
#> F-statistic: 111.6 on 3 and 502 DF,  p-value: < 2.2e-16
```

```r
model.8 <- lm(crim ~ poly(tax, 3))
summary(model.8)
```

```
#> 
#> Call:
#> lm(formula = crim ~ poly(tax, 3))
#> 
#> Residuals:
#>     Min      1Q  Median      3Q     Max 
#> -13.273  -1.389   0.046   0.536  76.950 
#> 
#> Coefficients:
#>                Estimate Std. Error t value Pr(>|t|)    
#> (Intercept)      3.6135     0.3047  11.860  < 2e-16 ***
#> poly(tax, 3)1 112.6458     6.8537  16.436  < 2e-16 ***
#> poly(tax, 3)2  32.0873     6.8537   4.682 3.67e-06 ***
#> poly(tax, 3)3  -7.9968     6.8537  -1.167    0.244    
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#> 
#> Residual standard error: 6.854 on 502 degrees of freedom
#> Multiple R-squared:  0.3689, Adjusted R-squared:  0.3651 
#> F-statistic:  97.8 on 3 and 502 DF,  p-value: < 2.2e-16
```

```r
model.9 <- lm(crim ~ poly(ptratio, 3))
summary(model.9)
```

```
#> 
#> Call:
#> lm(formula = crim ~ poly(ptratio, 3))
#> 
#> Residuals:
#>    Min     1Q Median     3Q    Max 
#> -6.833 -4.146 -1.655  1.408 82.697 
#> 
#> Coefficients:
#>                   Estimate Std. Error t value Pr(>|t|)    
#> (Intercept)          3.614      0.361  10.008  < 2e-16 ***
#> poly(ptratio, 3)1   56.045      8.122   6.901 1.57e-11 ***
#> poly(ptratio, 3)2   24.775      8.122   3.050  0.00241 ** 
#> poly(ptratio, 3)3  -22.280      8.122  -2.743  0.00630 ** 
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#> 
#> Residual standard error: 8.122 on 502 degrees of freedom
#> Multiple R-squared:  0.1138, Adjusted R-squared:  0.1085 
#> F-statistic: 21.48 on 3 and 502 DF,  p-value: 4.171e-13
```

```r
model.10 <- lm(crim ~ poly(lstat, 3))
summary(model.10)
```

```
#> 
#> Call:
#> lm(formula = crim ~ poly(lstat, 3))
#> 
```

```
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -15.234  -2.151  -0.486   0.066  83.353
#>
#> Coefficients:
#>                 Estimate Std. Error t value Pr(>|t|)
#> (Intercept)       3.6135     0.3392  10.654   <2e-16 ***
#> poly(lstat, 3)1  88.0697     7.6294  11.543   <2e-16 ***
#> poly(lstat, 3)2  15.8882     7.6294   2.082   0.0378 *
#> poly(lstat, 3)3 -11.5740     7.6294  -1.517   0.1299
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 7.629 on 502 degrees of freedom
#> Multiple R-squared:  0.2179, Adjusted R-squared:  0.2133
#> F-statistic: 46.63 on 3 and 502 DF,  p-value: < 2.2e-16
```

```r
model.11 <- lm(crim ~ poly(medv, 3))
summary(model.11)
```

```
#>
#> Call:
#> lm(formula = crim ~ poly(medv, 3))
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -24.427  -1.976  -0.437   0.439  73.655
#>
#> Coefficients:
#>                Estimate Std. Error t value Pr(>|t|)
#> (Intercept)       3.614      0.292  12.374  < 2e-16 ***
#> poly(medv, 3)1  -75.058      6.569 -11.426  < 2e-16 ***
#> poly(medv, 3)2   88.086      6.569  13.409  < 2e-16 ***
#> poly(medv, 3)3  -48.033      6.569  -7.312 1.05e-12 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 6.569 on 502 degrees of freedom
#> Multiple R-squared:  0.4202, Adjusted R-squared:  0.4167
#> F-statistic: 121.3 on 3 and 502 DF,  p-value: < 2.2e-16
```

Based on the model, the p-values for “indus”, “nox”, “age”, “dis”, “ptratio” and “medv” suggest these predictors are statistically significant. However, I do not spot evidence of non-linearity.