MATH 4322 Homework 2

Instructor: Dr. Cathy Poliak

Spring 2023

## Instructions

* Due February 9, 2023 at 11:59 pm
* Answer all questions fully
* Submit the answers in one file, preferably PDF, then upload in BlackBoard.
* These questions are from **Introduction to Statistical Learning, 2nd edition**, chapters 3 and 6.

## Problem 1

The following output is based on predicting sales based on three media budgets, TV, radio, and newspaper.

1. Give the estimated model to predict sales.  
   **y = 2.939 + .046(TV) + .189(radio) - .001(newspaper)**
2. Describe the null hypothesis to which the p-values given in the Coefficients table correspond. Explain this in terms of the sales, TV, radio, and newspaper, rather than in terms of the coefficients of the linear modelS.  
   **H0:βi=0, Ha:βi≠0**   
   **Using the information provided above, I will evaluate each variable.**   
   **Intercept = p<.0001; null hypothesis is rejected**  
   **TV = p<.0001; null hypothesis is rejected**  
   **radio = p<.0001; null hypothesis is rejected**  
   **newspaper = p=.86; null hypothesis is accepted**
3. Are there any variables that may not be significant in predicting sales?  
   **Newspaper**

## Problem 2

Based on the previous problem,the following is the output from the full model:

Below is based on the model

Below is based on the model

1. Determine the AIC for all three models.  
   **212.78, 210.81, 474.51**
2. Determine the for all three models.  
   **6.86, 4.9, 554.9**
3. Determine the adjusted for all three models.  
   **.8956, .8962, .6099**
4. Determine the RSE for all three models.  
   **1.686, 1.681, 3.259**
5. Which model best fits to predict sales based on these statistics?  
   **Model 2 with TV & Radio variables**

## Problem 3

Suppose we have a data set with five predictors, =GPA, = IQ, = Gender (1 for Female and 0 for Male), = Interaction between GPA and IQ, and = 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 , , , ,, .

1. Which answer is correct, and why?
   1. For a fixed value of IQ and GPA, males earn more on average than females.
   2. 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.**

**LSR: y^=50+20GPA+0.07IQ+35Gender+0.01GPA×IQ−10GPA×Gender; Males: y^=50+20GPA+0.07IQ+0.01GPA×IQ; Females: y^=85+10GPA+0.07IQ+0.01GPA×IQ. So the starting salary for males is higher than for females on average if 50+20GPA≥85+10GPA which is equivalent to GPA ≥ 3.5**

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

1. Predict the salary of a female with IQ of 110 and a GPA of 4.0.  
   **y^=85+40+7.7+4.4=137.1 which means the starting salary is equal to 137100$**
2. 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.  
   **To verify that the interaction of GPA/IQ term is very small we must test the hypothesis H0:β4^=0 and look at the p-value associated with given statistic to determine out answer. So our answer is False for the statement above.**

## Problem 4

We perform best subset, forward stepwise, and backward stepwise selection on a single data set. For each approach, we obtain models, containing predictors. Answer true or false to the following statements.

1. The predictors in the -variable model identified by forward stepwise are a subset of the predictors in the -variable model identified by forward stepwise selection.  
   **True**
2. The predictors in the k-variable model identified by backward stepwise are a subset of the predictors in the -variable model identified by backward stepwise selection.  
   **True**
3. The predictors in the -variable model identified by backward stepwise are a subset of the predictors in the -variable model identified by forward stepwise selection.  
   **False**
4. The predictors in the -variable model identified by forward stepwise are a subset of the predictors in the -variable model identified by backward stepwise selection.  
   **False**
5. The predictors in the -variable model identified by best subset are a subset of the predictors in the - variable model identified by best subset selection.  
   **False**

## Problem 5

This question involves the use of simple linear regression on the *Auto* data set. This can be found in the ISLR2 package in R.

##   
## Call:  
## lm(formula = mpg ~ horsepower)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -13.5710 -3.2592 -0.3435 2.7630 16.9240   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 39.935861 0.717499 55.66 <2e-16 \*\*\*  
## horsepower -0.157845 0.006446 -24.49 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.906 on 390 degrees of freedom  
## Multiple R-squared: 0.6059, Adjusted R-squared: 0.6049   
## F-statistic: 599.7 on 1 and 390 DF, p-value: < 2.2e-16

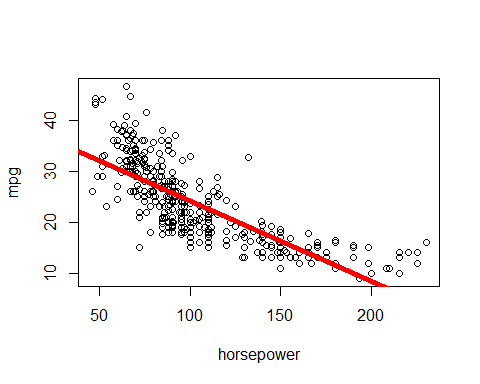
## 2.5 % 97.5 %  
## (Intercept) 38.525212 41.3465103  
## horsepower -0.170517 -0.1451725

## fit lwr upr  
## 1 24.46708 23.97308 24.96108

## fit lwr upr  
## 1 24.46708 14.8094 34.12476

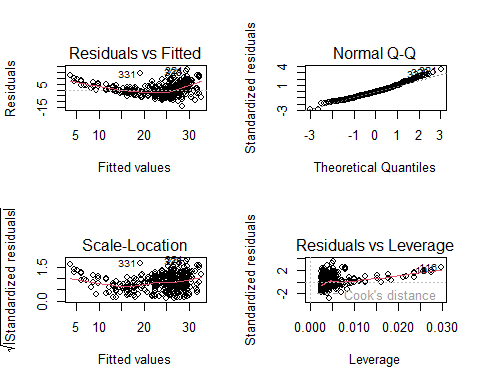
1. Use the lm() function to perform a simple linear regression with *mpg* as the response and *horsepower* (*hp*) as the predictor. Use the summary() function to print the results. Comment on the output. For example:
   1. Is there a relationship between the predictor and the response?  
      **Yes there is strong evidence that the horsepower variable has a relationship with mpg**
   2. How strong is the relationship between the predictor and the response?  
      **There is a very strong relationship between the given predictor and response**
   3. Is the relationship between the predictor and the response positive or negative?  
      **It is a negative relationship with a slope of -.157845**
   4. What is the predicted mpg associated with a horsepower of 98? What are the associated 95% confidence and prediction intervals? Give an interpretation of these intervals.  
      **The predicted mpg associated with a horsepower of 98 given that the model is y=39.935861 -0.157845(mpg) is == 24.467**  
      **The 95% confidence interval that the average mpg of a car with horsepower of 98 is between: 23.97308 and 24.96108**  
      **The 95% prediction interval that the mpg of a car with horsepower of 98 is between: 14.8094 and 34.12476**
2. Plot the response and the predictor. Use the abline() function to display the least squares regression line.  
   **The abline with the least squares regresssion line is shown below.**

## Warning: In lm.fit(x, y, offset = offset, singular.ok = singular.ok, ...) :  
## extra argument 'col' will be disregarded



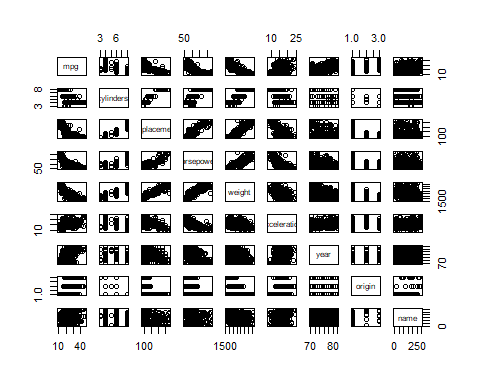
## integer(0)

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



## Problem 6

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

1. Produce a scatterplot matrix which includes all of the variables in the data set.  
   
2. Compute the matrix of correlations between the variables using the function cor(). You will need to exclude the name variable, cor() which is qualitative.

## mpg cylinders displacement horsepower weight   
## Min. : 9.00 Min. :3.000 Min. : 68.0 Min. : 46.0 Min. :1613   
## 1st Qu.:17.00 1st Qu.:4.000 1st Qu.:105.0 1st Qu.: 75.0 1st Qu.:2225   
## Median :22.75 Median :4.000 Median :151.0 Median : 93.5 Median :2804   
## Mean :23.45 Mean :5.472 Mean :194.4 Mean :104.5 Mean :2978   
## 3rd Qu.:29.00 3rd Qu.:8.000 3rd Qu.:275.8 3rd Qu.:126.0 3rd Qu.:3615   
## Max. :46.60 Max. :8.000 Max. :455.0 Max. :230.0 Max. :5140   
## acceleration year origin   
## Min. : 8.00 Min. :70.00 Min. :1.000   
## 1st Qu.:13.78 1st Qu.:73.00 1st Qu.:1.000   
## Median :15.50 Median :76.00 Median :1.000   
## Mean :15.54 Mean :75.98 Mean :1.577   
## 3rd Qu.:17.02 3rd Qu.:79.00 3rd Qu.:2.000   
## Max. :24.80 Max. :82.00 Max. :3.000

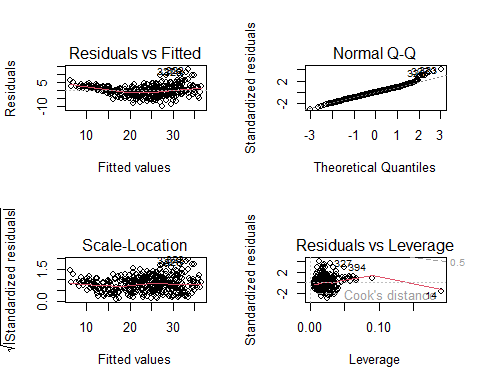
## mpg cylinders displacement horsepower weight  
## mpg 1.0000000 -0.7776175 -0.8051269 -0.7784268 -0.8322442  
## cylinders -0.7776175 1.0000000 0.9508233 0.8429834 0.8975273  
## displacement -0.8051269 0.9508233 1.0000000 0.8972570 0.9329944  
## horsepower -0.7784268 0.8429834 0.8972570 1.0000000 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

1. 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:

##   
## Call:  
## lm(formula = mpg ~ . - name, data = auto)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.5903 -2.1565 -0.1169 1.8690 13.0604   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -17.218435 4.644294 -3.707 0.00024 \*\*\*  
## cylinders -0.493376 0.323282 -1.526 0.12780   
## displacement 0.019896 0.007515 2.647 0.00844 \*\*   
## horsepower -0.016951 0.013787 -1.230 0.21963   
## weight -0.006474 0.000652 -9.929 < 2e-16 \*\*\*  
## acceleration 0.080576 0.098845 0.815 0.41548   
## year 0.750773 0.050973 14.729 < 2e-16 \*\*\*  
## origin 1.426141 0.278136 5.127 4.67e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.328 on 384 degrees of freedom  
## Multiple R-squared: 0.8215, Adjusted R-squared: 0.8182   
## F-statistic: 252.4 on 7 and 384 DF, p-value: < 2.2e-16

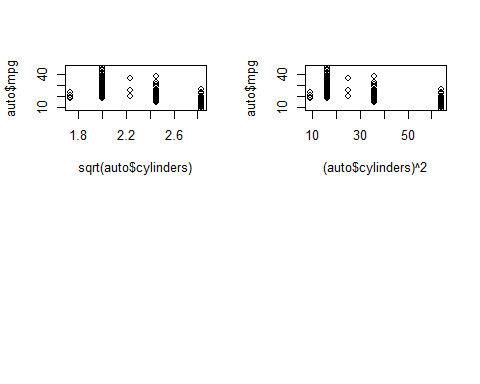
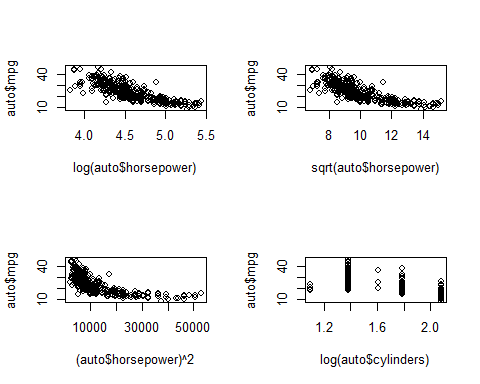
i. Is there a relationship between the predictors and the response?   
\*\*Testing the hypothesis H0:βi=0 ∀i; The p-value corresponding to the F-statistic is 2.037^-139 this indicates a relationship between mpg and the other predictors.\*\*   
  
ii. Which predictors appear to have a statistically significant relationship to the response?   
\*\*By checking the p-values given in the summary I concluded that origin, year, weight, and displacement had a significant relationship with mpg while acceleration, horsepower, and cylinders did not.\*\*   
  
iii. What does the coefficient for the year variable suggest?   
\*\*The summary suggests that for every 1 year increase in the year attribute there will be an increate of 0.750773 in the mpg attribute assuming all other predictors remain constant. This in turn means that vehichles get more fuel efficent as time goes on in yearly models at a rate of about .75 mpg.\*\*

1. Use the plot() function to produce diagnostic plots of the linear regression fit based on the predictors that appear to have a statistically signifianct relationship to the response. 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?  
   **There is evidence that we can assume that there is normal distribution. We can also see on the plots that there is evidence that there is some non linearity in the data. The Residuals vs Leverage plot also shows us a few outliers and one high Leverge point (14).**



1. Use the \* and/or : symbols to fit linear regression models with interaction effects. Do any interactions appear to be statistically significant?  
   **I used the \* symbol and found that the summary showed me the following: The interactions between ‘horsepower’ and ‘weight’ is significant, the interactions between ‘acceleration’ and ‘year’ is significant, and the interactions between ‘cylinders’ and ‘displacement’ are not significant.**

##   
## Call:  
## lm(formula = mpg ~ cylinders \* displacement + horsepower \* weight +   
## acceleration \* year, data = auto[, 1:8])  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.3265 -1.5779 0.0389 1.3483 11.6961   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.162e+02 1.853e+01 6.274 9.53e-10 \*\*\*  
## cylinders -1.803e-01 4.776e-01 -0.377 0.7061   
## displacement -2.867e-02 1.425e-02 -2.013 0.0449 \*   
## horsepower -2.261e-01 2.609e-02 -8.664 < 2e-16 \*\*\*  
## weight -1.019e-02 9.020e-04 -11.296 < 2e-16 \*\*\*  
## acceleration -7.081e+00 1.158e+00 -6.113 2.41e-09 \*\*\*  
## year -6.719e-01 2.417e-01 -2.780 0.0057 \*\*   
## cylinders:displacement 2.790e-03 2.067e-03 1.350 0.1779   
## horsepower:weight 5.154e-05 6.727e-06 7.661 1.53e-13 \*\*\*  
## acceleration:year 9.113e-02 1.502e-02 6.069 3.10e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.819 on 382 degrees of freedom  
## Multiple R-squared: 0.8726, Adjusted R-squared: 0.8696   
## F-statistic: 290.6 on 9 and 382 DF, p-value: < 2.2e-16



1. Try a few different transformations of the variables, such as *log(X)*, , . Comment on your findings.  
   **When trying a few different transformations such as log(X), sqrt(x), and X^2 I found that no real linear relationships were seen with ‘mpg’, ‘horsepower’, ‘cylinders’, ‘acceleration’ as I found them to be potentially quadratic. I found that I was able to make the relationships for ‘horsepower’ somewhat linear from using the log(X) transformation.**

## Problem 7

This problem involves the Boston data set, from the ISLR2 pakcage. 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.

##   
## 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

##   
## 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

##   
## 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

##   
## 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

##   
## 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

##   
## Call:  
## lm(formula = crim ~ 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 \*\*\*  
## 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

##   
## 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

##   
## 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

##   
## 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

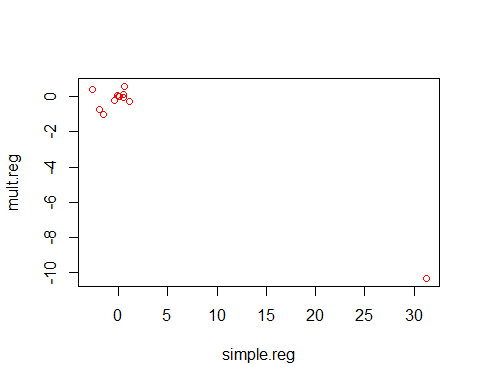
##   
## Call:  
## lm(formula = crim ~ age)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.789 -4.257 -1.230 1.527 82.849   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -3.77791 0.94398 -4.002 7.22e-05 \*\*\*  
## age 0.10779 0.01274 8.463 2.85e-16 \*\*\*  
## ---  
## 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

##   
## 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

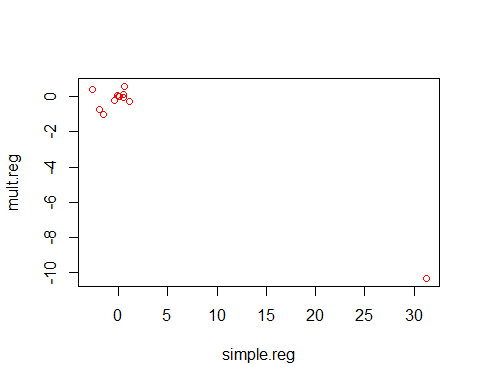
##   
## Call:  
## lm(formula = crim ~ lstat)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -13.925 -2.822 -0.664 1.079 82.862   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -3.33054 0.69376 -4.801 2.09e-06 \*\*\*  
## lstat 0.54880 0.04776 11.491 < 2e-16 \*\*\*  
## ---  
## 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

##   
## 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

##   
## Call:  
## lm(formula = crim ~ ., data = Boston)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.924 -2.120 -0.353 1.019 75.051   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 17.033228 7.234903 2.354 0.018949 \*   
## zn 0.044855 0.018734 2.394 0.017025 \*   
## indus -0.063855 0.083407 -0.766 0.444294   
## chas -0.749134 1.180147 -0.635 0.525867   
## nox -10.313535 5.275536 -1.955 0.051152 .   
## rm 0.430131 0.612830 0.702 0.483089   
## age 0.001452 0.017925 0.081 0.935488   
## dis -0.987176 0.281817 -3.503 0.000502 \*\*\*  
## rad 0.588209 0.088049 6.680 6.46e-11 \*\*\*  
## tax -0.003780 0.005156 -0.733 0.463793   
## ptratio -0.271081 0.186450 -1.454 0.146611   
## black -0.007538 0.003673 -2.052 0.040702 \*   
## lstat 0.126211 0.075725 1.667 0.096208 .   
## medv -0.198887 0.060516 -3.287 0.001087 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.439 on 492 degrees of freedom  
## Multiple R-squared: 0.454, Adjusted R-squared: 0.4396   
## F-statistic: 31.47 on 13 and 492 DF, p-value: < 2.2e-16



1. 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.  
   **We have to test H0:β1=0. When testing I found that ‘chas’ was the only predictor which the p value was more than .05. This means that all the predictors other than ‘chas’; there is a significant association between each predictor and response excluding ‘chas’.**
2. 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 ?  
   **After fitting the multiple regression model using all the predictors given in the dataset our results show us that we can reject the null hypothesis for the following predictors: zn”, “dis”, “rad”, “black” and “medv”.**
3. 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.  
   **‘nox’ is approximately -10 in single variate model and 31 in multiple regression model.**  
   **It is evident that there is a difference between the multiple and simple regression coefficients. This can be explained for the simple regression case by the slope representing the average affect of an increase of 1 in the predictor ignoring the other predictors. For the multiple regression coefficients, the slope term represents the average effect of an increase in the given predictor, the other predictors will also be fixed in this case.**



1. 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

* **For the following features, the p-value tells us that the coefficients are not significant: rad, rm, zn, tax,black, lstat.**  
  **For the following features, the p-value tells us that the coefficients are significant: indus, age, dis, pratio, medv, nox.**  
  **So we have determined which non-linear associate association is visible between which coefficients**

##   
## 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

##   
## 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

##   
## 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

##   
## 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

##   
## 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

##   
## 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

##   
## 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

##   
## 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

##   
## 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

##   
## Call:  
## lm(formula = crim ~ poly(black, 3))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -13.096 -2.343 -2.128 -1.439 86.790   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.6135 0.3536 10.218 <2e-16 \*\*\*  
## poly(black, 3)1 -74.4312 7.9546 -9.357 <2e-16 \*\*\*  
## poly(black, 3)2 5.9264 7.9546 0.745 0.457   
## poly(black, 3)3 -4.8346 7.9546 -0.608 0.544   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.955 on 502 degrees of freedom  
## Multiple R-squared: 0.1498, Adjusted R-squared: 0.1448   
## F-statistic: 29.49 on 3 and 502 DF, p-value: < 2.2e-16

##   
## 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

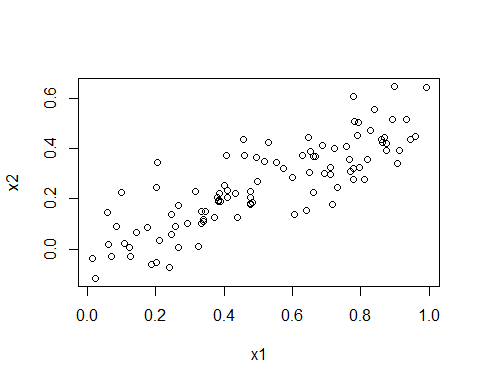
##   
## Call:  
## lm(formula = crim ~ poly(medv, 3))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -24.427 -1.976 -0.437 0.439 73.655   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.614 0.292 12.374 < 2e-16 \*\*\*  
## poly(medv, 3)1 -75.058 6.569 -11.426 < 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

## Problem 8

This problem focuses on the **collinearity** problem.

1. Perform the following commands in R:

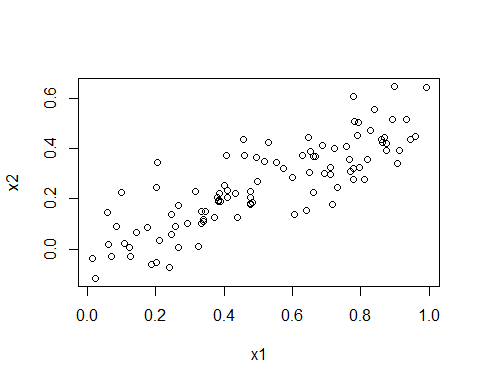
set.seed (1)  
x1=runif (100)  
x2 =0.5\* x1+rnorm (100) /10  
y=2+2\* x1 +0.3\* x2+rnorm (100)  
  
cor(x1, x2)  
plot(x1, x2)



The last line corresponds to creating a linear model in which is a function of and . Write out the form of the linear model. What are the regression coefficients?  
**The regression coefficients are B0 = 2+rnorm(100), B1 = 2, and B3 = 0.3**  
**Form is: y=2+2\* x1 +0.3\* x2+rnorm (100)**

1. What is the correlation between and ? Create a scatterplot displaying the relationship between the variables.  
   **cor = 0.8351212**

cor(x1, x2)  
plot(x1, x2)



1. Using this data, fit a least squares regression to predict y using and . Describe the results obtained. What are , , and ? How do these relate to the true , , and ? Can you reject the null hypothesis ? How about the null hypothesis ?  
   \*\*\*\*

lm.new =lm(y~x1+x2)  
summary(lm.new)

1. Now fit a least squares regression to predict using only . Comment on your results. Can you reject the null hypothesis ?  
   **Given the estimates: βo = 2.1305, β1 = 1.4396 , and β3 = 1.0097**\*  
   **Our decisions are to reject the null hypothesis for B1 only at a 99% confidence and not fail to reject it at a 95% confidence. For B2 we will reject the null hypothesis given the value above.**
2. Now fit a least squares regression to predict using only . Comment on your results. Can you reject the null hypothesis ?  
   **For the given B0=2.1124 and B1=1.9759, our decision will be to reject the null hypothesis for B1 because of the very low computated p-value. The r^2 value also shows us that x1 explains about 20% of the changes in the variable y.**

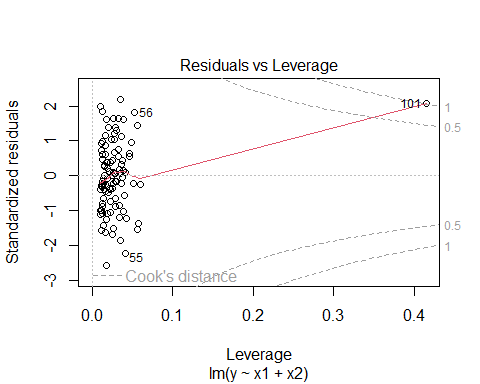
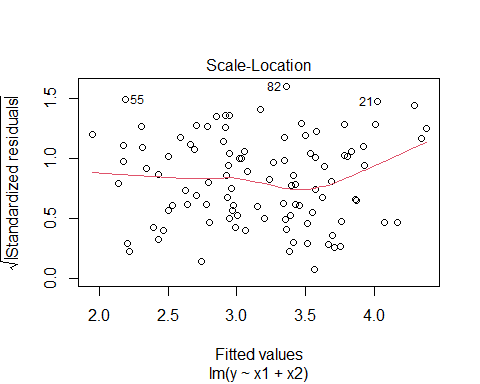
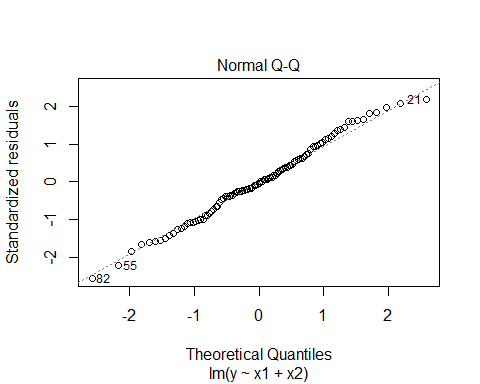
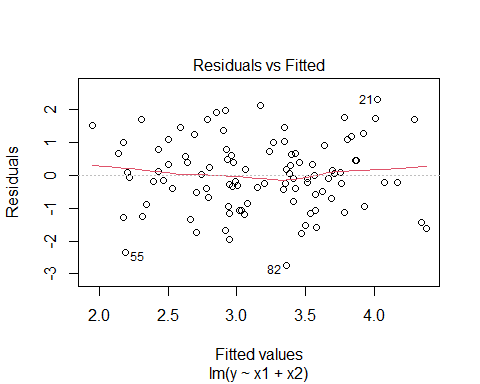
lm.new2 = lm(y~x1)  
summary(lm.new2)

1. Do the results obtained in (c)–(e) contradict each other? Explain your answer.  
   **These results do contradict eachother since for part c we found that the multi-linear regression model x1 and x2 are not significant but we see in part d that xz1 and x2 are very significant and can explain up to 20% of the changes in y.**
2. 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)  
lm.new3 =lm(y~x1+x2)  
summary(lm.new3);

##   
## 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(lm.new3);



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.  
**I found the r^2 values to be slightly greater but not enough to be considered significant. I also found the x2 variable to be the significant variable instead of x1 from our previous data model. Lastly, looking at the use of cook’s distance in the plots, the new added point is not an outlier but considered a leverage point.**

**For the second plot I found in both cases that x1 is significant however a big negative change in the r^2 value computer so this model is worse than our original. The added data point is also considered an outlier in this model.**

**For the thirds plots, x2 was considered significant and there is a big increase in the r^2 value so this means that this plot is better than our original from above. The added data point in this plot was also not an outlier or leverage point.**

## Problem 9

In this exercise, we will generate simulated data, and will then use this data to perform best subset selection.

1. Use the rnorm() function to generate a predictor of length = 100, as well as a noise vector of length = 100.  
   \*\*\*\*

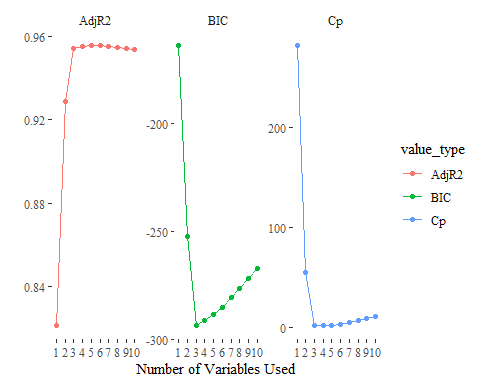
set.seed(1)  
X <- rnorm(100)  
noise <- rnorm(100)

1. Generate a response vector of length n = 100 according to the model

* where , , , and are constants of your choice.  
  \*\*\*\*

Y <- 3 + 1\*X + 4\*X^2 - 1\*X^3 + noise

1. Use the regsubsets() function to perform best subset selection in order to choose the best model containing the predictors . What is the best model obtained according to , BIC, and adjusted ? Show some plots to provide evidence for your answer, and report the coefficients of the best model obtained. Note you will need to use the data.frame() function to create a single data set containing both and .  
   **Y=16.973+3.007X+0.842X2−1.986X3 is the computed model with chosen best subsets. The model indicates that 3 variables provide the best fit.**



1. Repeat (c), using forward stepwise selection and also using backwards stepwise selection. How does your answer compare to the results in (c)?  
   **I found that after repeating this using a forward stepwise selection and also using backwards stepwise selection that the backwards stepwise model agreed with the output of the best chosen subsets in the model.**

