CMDA 4654: Intermed Data Analytics & ML

Homework 1

Christopher J. Mobley

Due: Friday, 26FEB16

1. For each of parts (a) through (d), indicate whether we would generally expect the performance of a flexible statistical learning method to be better or worse than an inflexible method. Justify your answer.

   a) The sample size n is extremely large, and the number of predictors p is small.

      i. A flexible model will fit the data better than an inflexible model. Since, the sample size, n, is extremely large and the number of predictors, p, small overfitting is unlikely. So, a flexible model would be the better statistical learning method in this case.

   b) The number of predictors p is extremely large, and the number of observations n is small.

      i. As previously stated, a flexible model will fit the data better than an inflexible model. Since, the sample size, n, is small and the number of predictors, p, is extremely large overfitting is likely. So, an inflexible model would be the better statistical learning method in this case.

   c) The relationship between the predictors and response is highly non-linear.

      i. With a highly non-linear relationship between the predictors and response, a more flexible model will be needed in order to fit the data adequately. So, a flexible model would be the better statistical learning method in this case.

   d) The variance of the error terms, i.e. σ^2 = Var(), is extremely high.

      i. As previously stated, a flexible model will fit the data better than an inflexible model. Since, the variance of the error term, σ^2, is extremely high, overfitting is likely. So, an inflexible model would be the better statistical learning method in this case.

2. Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, provide n and p.

a) We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.

   i. Due to the fact the CEO's salary is a quantitative value and we are attempting to understand which firms factors it, this would be a regression problem in which we are more interested in inference. N = 500 and p = 3 (profit, # of employees, industry)

b) We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.

   i. Due to the fact that whether a product will succeed or not is a qualitative value and we are attempting to determine whether or not a specific product will succeed or fail based on similar product specifications, this would be a classification problem in which we are more interested in prediction. N = 20 and p = 13 (product price, marketing budget, competition price and ten other variables)

c) We are interesting in predicting the % change in the US dollar in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week we record the % change in the dollar, the % change in the US market, the % change in the British market, and the % change in the German market.

   i. Due to the fact that the % change in the US dollar is a quantitative value and we are attempting to predict it based on the weekly % changes in world stock markets, this is a regression problem in which we are more interested in prediction. N = 52 (# of weeks in a year) and p = 3 (weekly % change in US, British, and German market)

3. What are the advantages and disadvantages of a very flexible (versus a less flexible) approach for regression or classification? Under what circumstances might a more flexible approach be preferred to a less flexible approach? When might a less flexible approach be preferred?

a) The advantages of a flexible approach vs an inflexible one for regression or classification is that a flexible approach is able to fit non-linear models better than an inflexible approach. However, with a small sample size, n, or a large number of predictors, p, a flexible model is more susceptible to variation in the data making it more likely to create a model that is overfitted. Thus a more flexible model would be the preferred approach when the goal is making predication. However, when the goal is making an inference an inflexible model would be the better choice.

4. This exercise relates to the College data set, which can be found in the file College.csv. It contains a number of variables for 777 different universities and colleges in the US. The variables are

- Private : Public/private indicator
- Apps : Number of applications received
- Accept : Number of applicants accepted
- Enroll : Number of new students enrolled
- Top10perc : New students from top 10 % of high school class
- Top25perc : New students from top 25 % of high school class
- F.Undergrad : Number of full-time undergraduates
- P.Undergrad : Number of part-time undergraduates
- Outstate : Out-of-state tuition
- Room.Board : Room and board costs
- Books : Estimated book costs
- Personal : Estimated personal spending
- PhD : Percent of faculty with Ph.D.'s
- Terminal : Percent of faculty with terminal degree
- S.F.Ratio : Student/faculty ratio
- perc.alumni : Percent of alumni who donate
- Expend : Instructional expenditure per student
- Grad.Rate : Graduation rate

Before reading the data into R, it can be viewed in Excel or a text editor.

a) Use the read.csv() function to read the data into R. Call the loaded data college. Make sure that you have the directory set to the correct location for the data.

b) Look at the data using the fix() function. You should notice that the first column is just the name of each university. We don’t really want R to treat this as data. However, it may be handy to have these names for later. Try the following commands:

```
> rownames (college )=college [,1]
```

```
> fix(college)
```

You should see that there is now a row.names column with the

name of each university recorded. This means that R has given each

row a name corresponding to the appropriate university. R will not try to

perform calculations on the row names. However, we still need to

eliminate the first column in the data where the names are stored. Try

```
> college =college [,-1]
```

```
> fix(college)
```

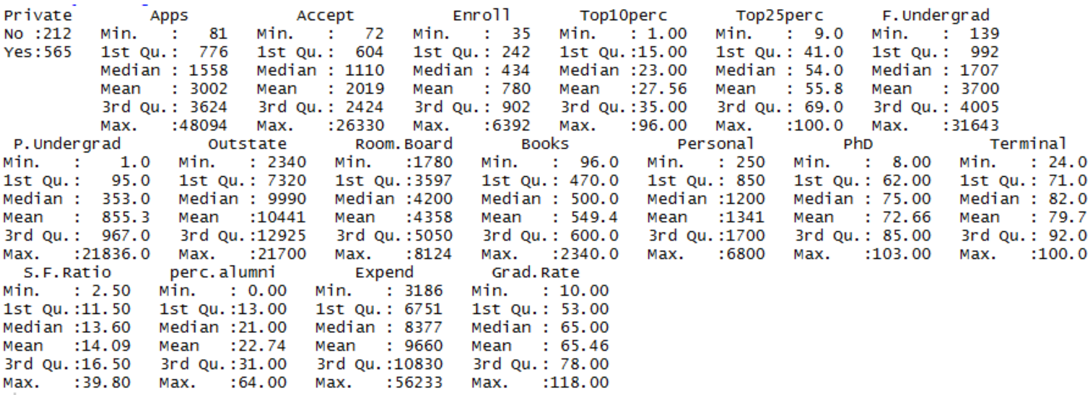
Now you should see that the first data column is Private. Note that

another column labeled row.names now appears before the Private

column. However, this is not a data column but rather the name that R

is giving to each row.

c) i. Use the summary() function to produce a numerical summary of the variables in the data set.



```
 Private        Apps           Accept          Enroll        Top10perc       Top25perc      F.Undergrad
 No :212   Min.   :   81   Min.   :   72   Min.   :  35   Min.   : 1.00   Min.   :  9.0   Min.   :  139
 Yes:565   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242   1st Qu.:15.00   1st Qu.: 41.0   1st Qu.:  992
           Median : 1558   Median : 1110   Median : 434   Median :23.00   Median : 54.0   Median : 1707
           Mean   : 3002   Mean   : 2019   Mean   : 780   Mean   :27.56   Mean   : 55.8   Mean   : 3700
           3rd Qu.: 3624   3rd Qu.: 2424   3rd Qu.: 902   3rd Qu.:35.00   3rd Qu.: 69.0   3rd Qu.: 4005
           Max.   :48094   Max.   :26330   Max.   :6392   Max.   :96.00   Max.   :100.0   Max.   :31643
  P.Undergrad         Outstate       Room.Board        Books           Personal         PhD            Terminal
 Min.   :    1.0   Min.   : 2340   Min.   :1780   Min.   :  96.0   Min.   : 250   Min.   :  8.00   Min.   : 24.0
 1st Qu.:   95.0   1st Qu.: 7320   1st Qu.:3597   1st Qu.: 470.0   1st Qu.: 850   1st Qu.: 62.00   1st Qu.: 71.0
 Median :  353.0   Median : 9990   Median :4200   Median : 500.0   Median :1200   Median : 75.00   Median : 82.0
 Mean   :  855.3   Mean   :10441   Mean   :4358   Mean   : 549.4   Mean   :1341   Mean   : 72.66   Mean   : 79.7
 3rd Qu.:  967.0   3rd Qu.:12925   3rd Qu.:5050   3rd Qu.: 600.0   3rd Qu.:1700   3rd Qu.: 85.00   3rd Qu.: 92.0
 Max.   :21836.0   Max.   :21700   Max.   :8124   Max.   :2340.0   Max.   :6800   Max.   :103.00   Max.   :100.0
   S.F.Ratio      perc.alumni        Expend        Grad.Rate
 Min.   : 2.50   Min.   : 0.00   Min.   : 3186   Min.   : 10.00
 1st Qu.:11.50   1st Qu.:13.00   1st Qu.: 6751   1st Qu.: 53.00
 Median :13.60   Median :21.00   Median : 8377   Median : 65.00
 Mean   :14.09   Mean   :22.74   Mean   : 9660   Mean   : 65.46
 3rd Qu.:16.50   3rd Qu.:31.00   3rd Qu.:10830   3rd Qu.: 78.00
 Max.   :39.80   Max.   :64.00   Max.   :56233   Max.   :118.00
```

Figure 1. Summary of college dataset

ii. Use the pairs() function to produce a scatterplot matrix of the first ten columns or variables of the data. Recall that you can reference the first ten columns of a matrix A using A[,1:10].

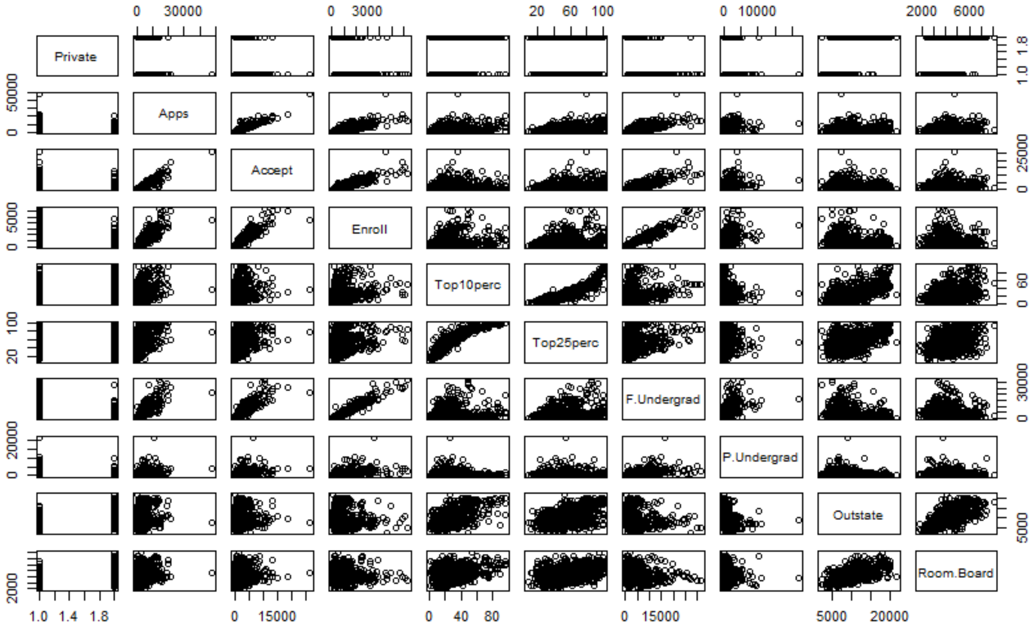


Figure 2. Scatterplot matrix of the first 10 variables of college

iii. Use the plot() function to produce side-by-side boxplots of Outstate versus Private.

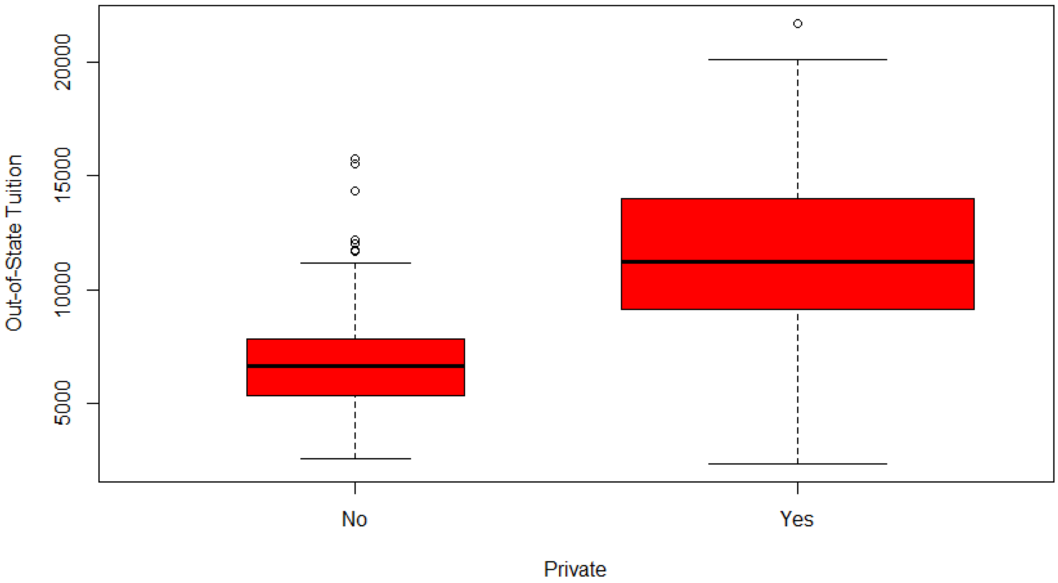
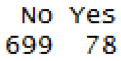


Figure 3. Side-by-side boxplots of Outstate versus Private

iv. Create a new qualitative variable, called Elite, by binning the Top10perc variable. We are going to divide universities into two groups based on whether or not the proportion of students coming from the top 10 % of their high school classes exceeds 50 %.

```
> Elite=rep("No",nrow(college ))
> Elite[college$Top10perc >50]=" Yes"
> Elite=as.factor(Elite)
> college=data.frame(college , Elite)
```

Use the summary() function to see how many elite universities there are. Now use the plot() function to produce side-by-side boxplots of Outstate versus Elite.



```
 No  Yes
699   78
```

Figure 4. Summary of elite universities

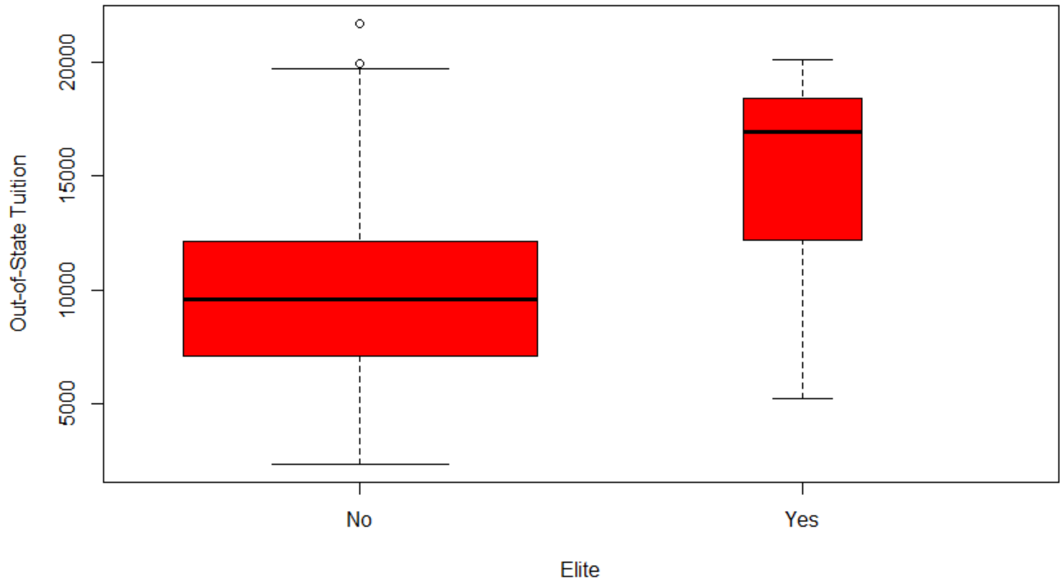


Figure 5. Side-by-side boxplots of Outstate versus Elite

v. Use the hist() function to produce some histograms with differing numbers of bins for a few of the quantitative variables. You may find the command par(mfrow=c(2,2)) useful: it will divide the print window into four regions so that four plots can be made simultaneously. Modifying the arguments to this function will divide the screen in other ways.

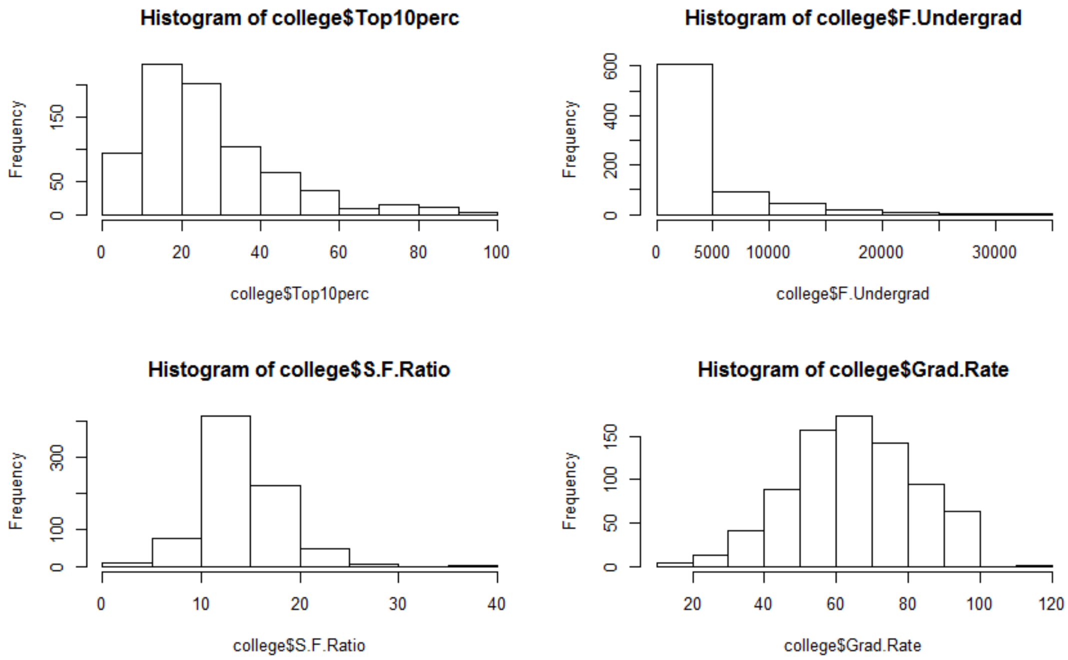


Figure 6. Histograms with differing numbers of bins

vi. Continue exploring(the data, and provide a brief summary of what you Discover.

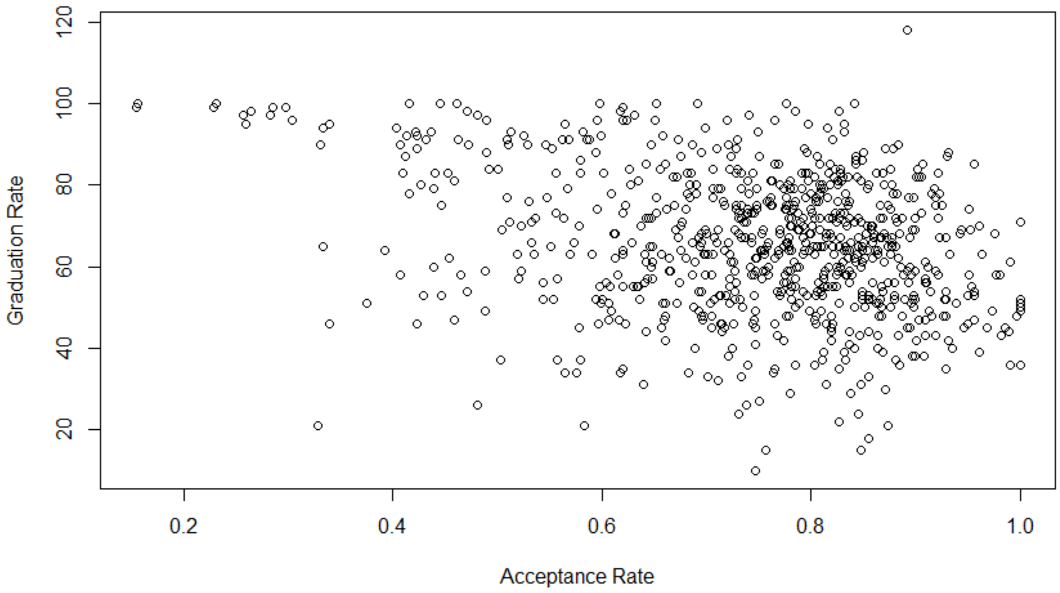


Figure 7. Plot of acceptance rate vs graduation rate

Figure 7 appears to show that colleges with low acceptance rates tend to have slightly higher graduation rates.

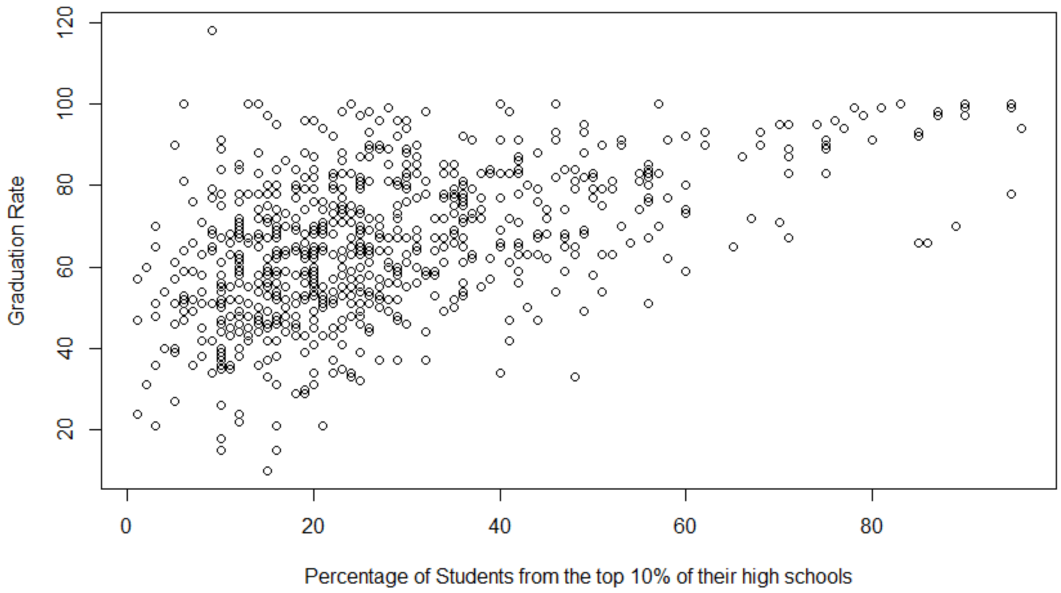


Figure 8. Plot of percentage of student from the top 10% of their high school class vs graduation rate

Figure 8 appears to show that colleges with higher percentage of student who performed at the top 10% of their high school class tend to have slightly higher graduation rates.

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

a) In Table 3.4, the null hypothesis to which the p-values correspond are H_0: β_i = 0 in the presence of the other predictors. In other words, what is the likelihood that one of the three advertising methods has no effect on sales in the presence of the other two advertising methods. A p-values less than our alpha (typically .05 or .025) means we can reject the null hypothesis. Whereas a p-value above our alpha means we fail to reject the null hypothesis. In this case, due to newspaper ads high p-value we can make inference that in the presence of the other two predictors, newspaper ads have no effect on sales. However, TV and radio ads p-value is less than our alpha (close to 0), therefore we can reject the null hypothesis and make inference that these two ad sources have effect on sales in the presence of the other two advertising methods.

6. Suppose we have a data set with five predictors, X1 = GPA, X2 = IQ, X3 = Gender (1 for Female and 0 for Male), X4 = Interaction between GPA and IQ, and X5 = Interaction between GPA and Gender. The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get $\widehat{\beta_0}$ = 50, $\widehat{\beta_1}$ = 20, $\widehat{\beta_2}$ = 0.07, $\widehat{\beta_3}$ = 35, $\widehat{\beta_4}$ = 0.01, $\widehat{\beta_5}$ = −10.

$$Y = 50 + 20GPA + 0.07IQ + 35Gender + 0.01(GPA * IQ) - 10(GPA * Gender)$$

a) Which answer is correct, and why?

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

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

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

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

1. Given the choices above, iii is the correct choice as explained below.

Male: $Y = 50 + 20GPA + 0.07IQ + 0.01(GPA * IQ)$

Female:$Y = 50 + 20GPA + 0.07IQ + 35 + 0.01(GPA * IQ) - 10GPA$

Difference: $35 - 10GPA$

Thus once GPAs are higher than 3.5, males will earn more on average than females with similar GPAs. However, when GPAs are less than 3.5 females will earn more on average than males with similar GPAs.

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

ii. Female:$Y = 50 + 20GPA + 0.07IQ + 35 + 0.01(GPA * IQ) - 10GPA$

Y = 50 + 20(4.0) + 0.07(110) + 35 + 0.01(4.0*110) – 10(4.0)

Y = $137,100.00

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

i. False, the size of the interaction terms has no bearing on whether it is statistically significant or not. This is determined by the p-value, as all the terms do not share the same scale.

7. I collect a set of data (n = 100 observations) containing a single predictor and a quantitative response. I then fit a linear regression model to the data, as well as a separate cubic regression, i.e. $Y = \beta0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + .$

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

   i. We would expect the training RSS for the cubic regression to be lower than that of the linear regression due to the fact it has a larger number of predictors, p. Thus allowing it to more tightly fit the data.

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

   i. We would expect the testing RSS for the linear regression to be lower than that of the cubic regression due to the fact that the true relationship between X and Y is linear. Thus, the cubic regression would have overfit the data resulting in a less accurate model and thus poorer performance.

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

   i. We would again expect the training RSS for the cubic relationship to be lower than that of the linear regression due to the fact that it has a larger number of predictors, p. Thus allowing it to more tightly fit the data.

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

   i. There is not enough information to tell whether the linear or cubic training RSS would be smaller since we don’t know how far from linear the true relationship is. If the model is closer to linear than cubic then we would expect the RSS of the linear model to be smaller and vice versa.

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

$$\hat{y}_i = x_i \hat{\beta},$$

where

$$\hat{\beta} = \left(\sum_{i=1}^{n} x_i y_i \right) / \left(\sum_{i'=1}^{n} x_{i'}^2 \right). \tag{3.38}$$

Show that we can write

$$\hat{y}_i = \sum_{i'=1}^{n} a_{i'} y_{i'}.$$

$$\hat{y}_i = x_i * \hat{\beta} \tag{1}$$

$$\hat{y}_i = x_i * \frac{\sum_{i'=1}^{n} x_{i'} y_{i'}}{\sum_{j=1}^{n} x_j^2} \tag{2}$$

$$\hat{y}_i = \sum_{i'=1}^{n} \left(\frac{x_i * x_{i'}}{\sum_{j=1}^{n} x_j^2} \right) y_{i'} \tag{3}$$

$$\hat{y}_i = \sum_{i'=1}^{n} a_{i'} y_{i'} \tag{4}$$

What is a_{i}'?

$$a_{i'} = \left(\frac{x_i * x_{i'}}{\sum_{j=1}^{n} x_j^2} \right) \tag{5}$$

Note: We interpret this result by saying that the fitted values from linear regression are linear combinations of the response values.

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.

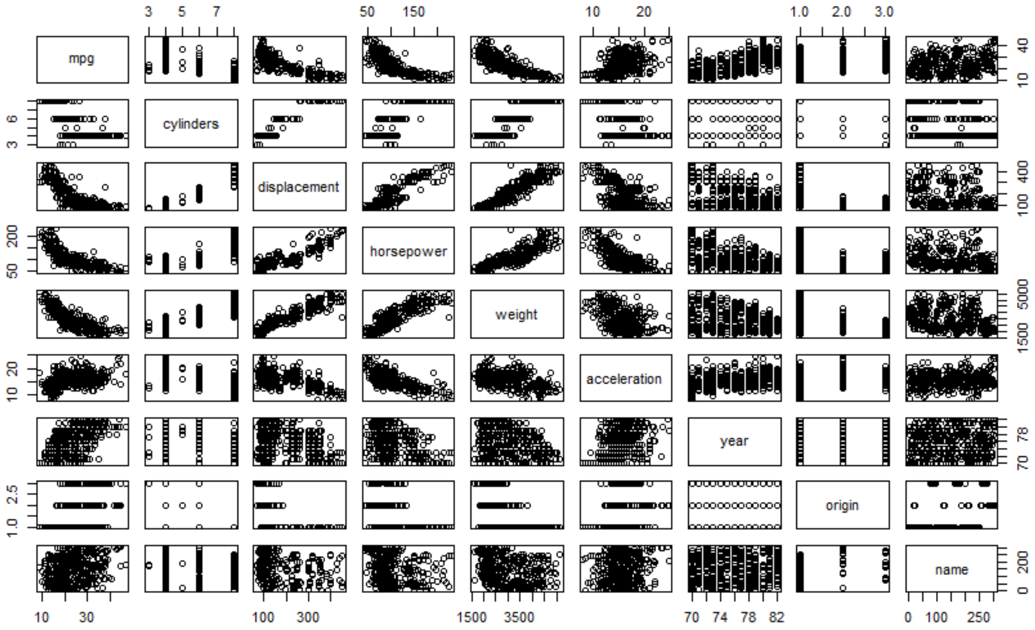
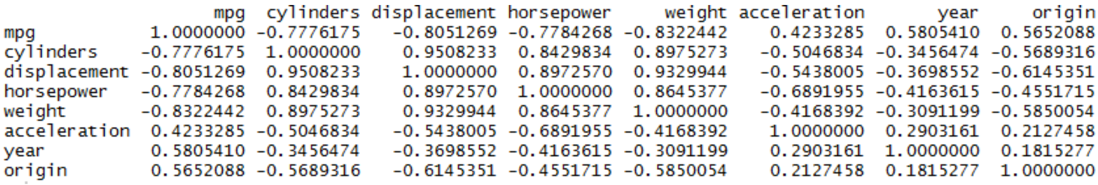


Figure 9. Scatterplot matrix of the Auto dataset variables

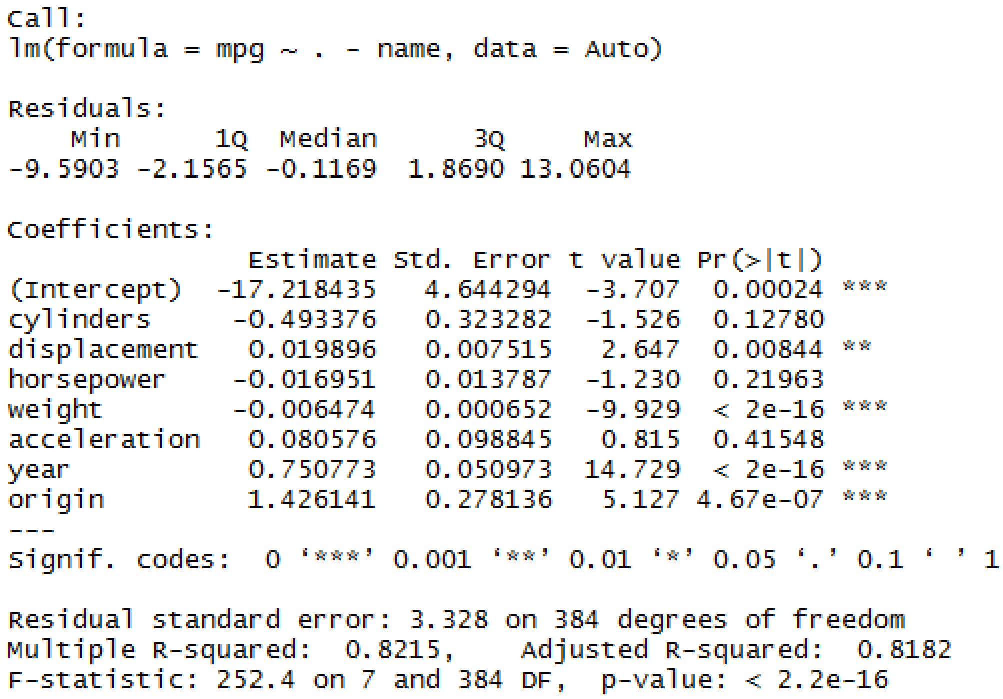
b) Compute the matrix of correlations between the variables using the function cor(). You will need to exclude the name variable, cor() which is qualitative.



```
                    mpg  cylinders displacement horsepower     weight acceleration       year     origin
mpg           1.0000000 -0.7776175   -0.8051269 -0.7784268 -0.8322442    0.4233285  0.5805410  0.5652088
cylinders    -0.7776175  1.0000000    0.9508233  0.8429834  0.8975273   -0.5046834 -0.3456474 -0.5689316
displacement -0.8051269  0.9508233    1.0000000  0.8972570  0.9329944   -0.5438005 -0.3698552 -0.6145351
horsepower   -0.7784268  0.8429834    0.8972570  1.0000000  0.8645377   -0.6891955 -0.4163615 -0.4551715
weight       -0.8322442  0.8975273    0.9329944  0.8645377  1.0000000   -0.4168392 -0.3091199 -0.5850054
acceleration  0.4233285 -0.5046834   -0.5438005 -0.6891955 -0.4168392    1.0000000  0.2903161  0.2127458
year          0.5805410 -0.3456474   -0.3698552 -0.4163615 -0.3091199    0.2903161  1.0000000  0.1815277
origin        0.5652088 -0.5689316   -0.6145351 -0.4551715 -0.5850054    0.2127458  0.1815277  1.0000000
```

Figure 10. Correlation matrix of the Auto dataset variables

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



```
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
```

Figure 11. Summary of the linear model with the response mpg

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

   1. Yes, there is a relationship between the predictors and the response, mpg, as seen by the low p-value of the predictors, p, in Figure 11, or probability that each predictor has no effect on the response in the presence of the other predictors.

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

   1. The displacement, weight, year and origin predictors appear to have a statistically significant relationship to the response as determined by their low p-value seen in Figure 11.

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

   1. The coefficient for the year variable is 0.750773. This is indicated that the year predictor has a positive relationship on mpg. In other words, vehicle appear to becoming more fuel efficient over time.

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?

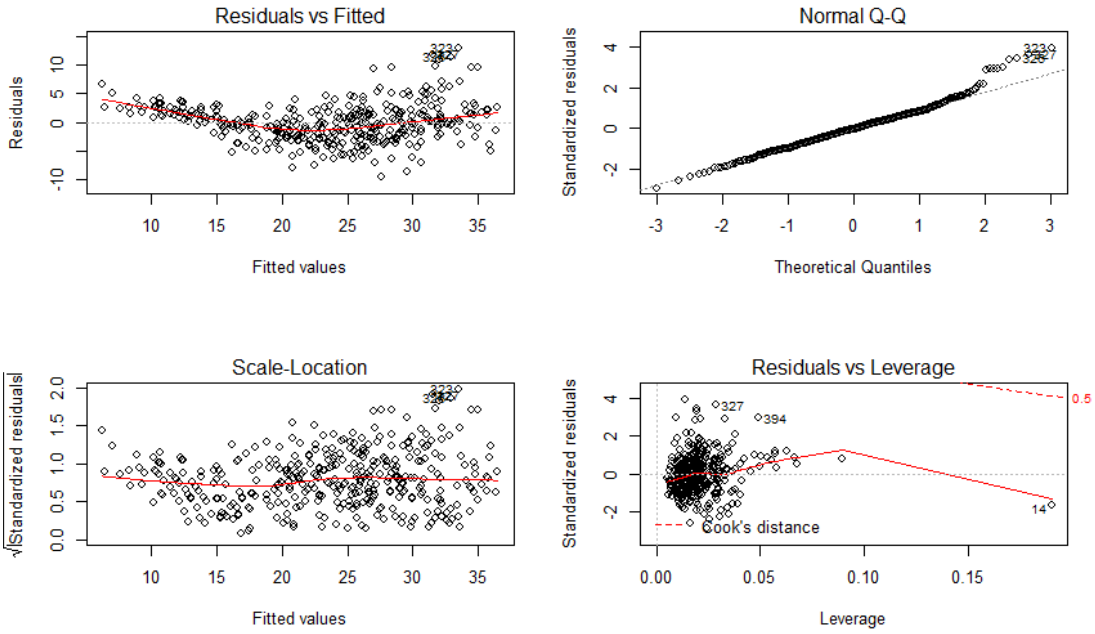
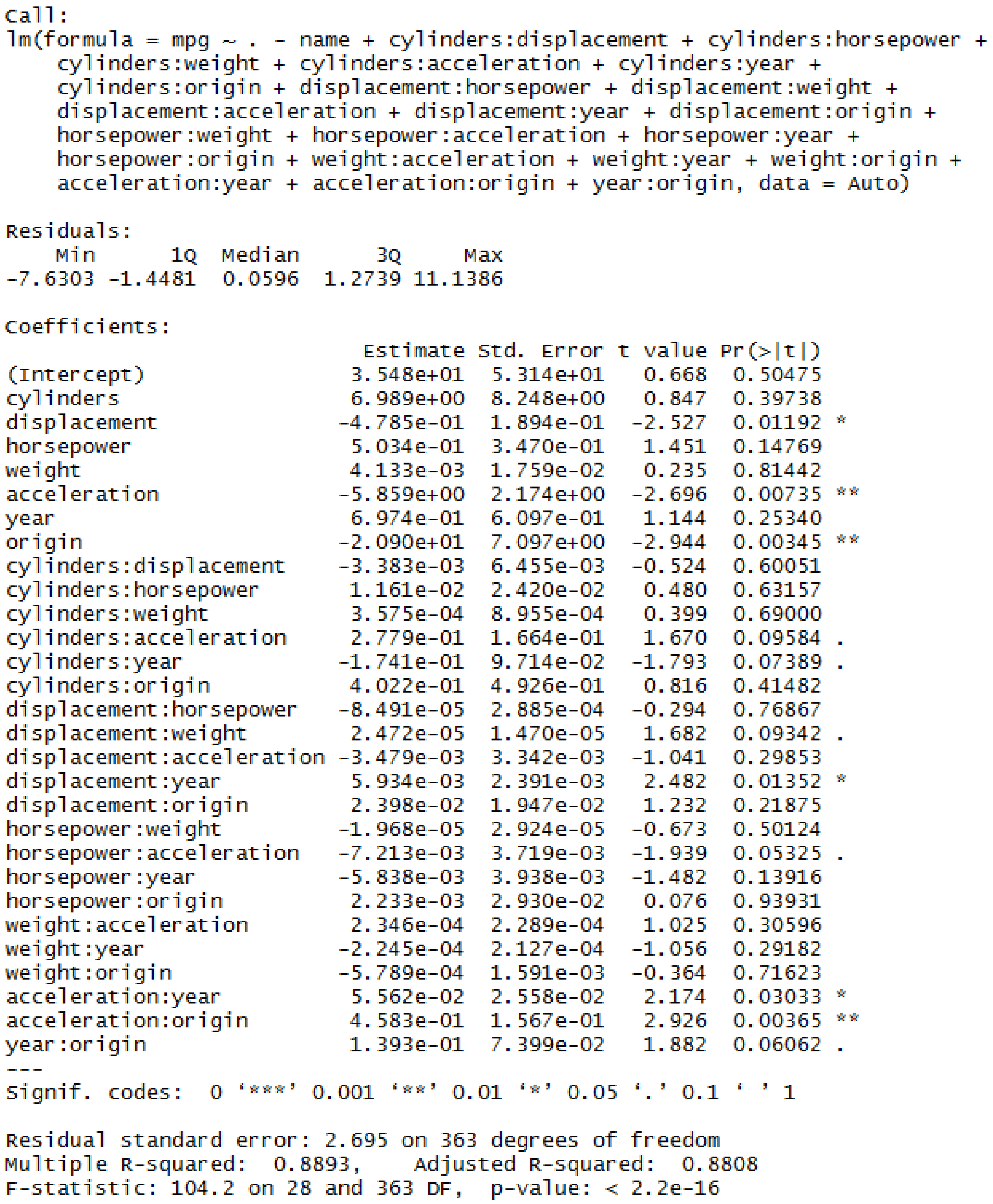


Figure 12. Diagnostic plots of the linear regression fit

The residual plot at the top left of Figure 12 has a discernable pattern which indicates that there is some non-linearity in the data and thus a non-linear model would fit the data better. In regards to outliers, the plots do not show any significant outliers. In regards to points with unusually high leverage, point 14 appears to have a high leverage thought not a very high standardized residual.

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

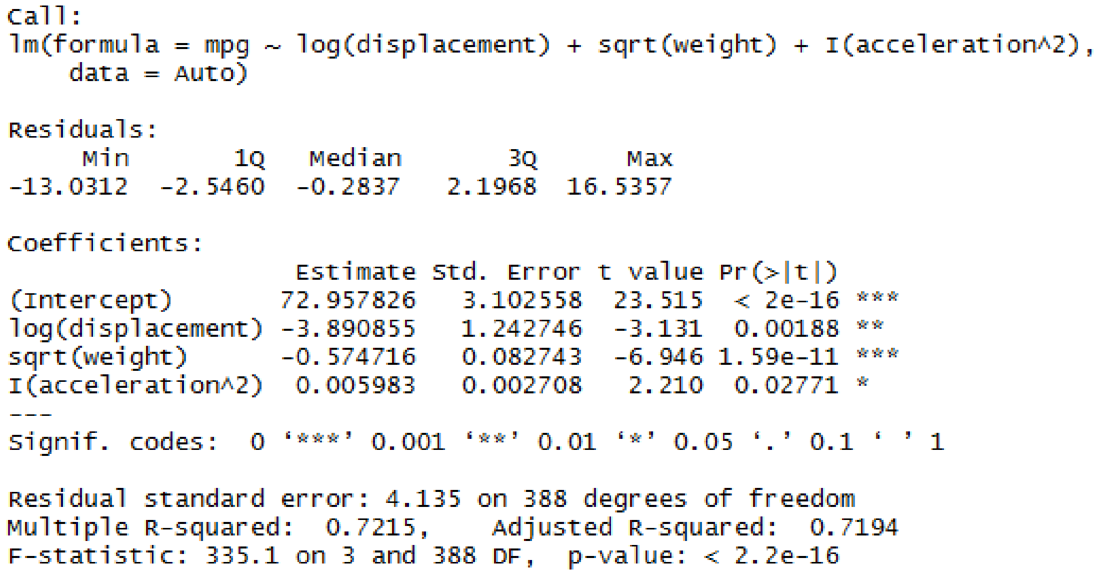


```
Call:
lm(formula = mpg ~ . - name + cylinders:displacement + cylinders:horsepower +
    cylinders:weight + cylinders:acceleration + cylinders:year +
    cylinders:origin + displacement:horsepower + displacement:weight +
    displacement:acceleration + displacement:year + displacement:origin +
    horsepower:weight + horsepower:acceleration + horsepower:year +
    horsepower:origin + weight:acceleration + weight:year + weight:origin +
    acceleration:year + acceleration:origin + year:origin, data = Auto)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.6303 -1.4481  0.0596  1.2739 11.1386 

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)   
(Intercept)                3.548e+01  5.314e+01   0.668  0.50475   
cylinders                  6.989e+00  8.248e+00   0.847  0.39738   
displacement              -4.785e-01  1.894e-01  -2.527  0.01192 * 
horsepower                 5.034e-01  3.470e-01   1.451  0.14769   
weight                     4.133e-03  1.759e-02   0.235  0.81442   
acceleration              -5.859e+00  2.174e+00  -2.696  0.00735 **
year                       6.974e-01  6.097e-01   1.144  0.25340   
origin                    -2.090e+01  7.097e+00  -2.944  0.00345 **
cylinders:displacement    -3.383e-03  6.455e-03  -0.524  0.60051   
cylinders:horsepower       1.161e-02  2.420e-02   0.480  0.63157   
cylinders:weight           3.575e-04  8.955e-04   0.399  0.69000   
cylinders:acceleration     2.779e-01  1.664e-01   1.670  0.09584 . 
cylinders:year            -1.741e-01  9.714e-02  -1.793  0.07389 . 
cylinders:origin           4.022e-01  4.926e-01   0.816  0.41482   
displacement:horsepower   -8.491e-05  2.885e-04  -0.294  0.76867   
displacement:weight        2.472e-05  1.470e-05   1.682  0.09342 . 
displacement:acceleration -3.479e-03  3.342e-03  -1.041  0.29853   
displacement:year          5.934e-03  2.391e-03   2.482  0.01352 * 
displacement:origin        2.398e-02  1.947e-02   1.232  0.21875   
horsepower:weight         -1.968e-05  2.924e-05  -0.673  0.50124   
horsepower:acceleration   -7.213e-03  3.719e-03  -1.939  0.05325 . 
horsepower:year           -5.838e-03  3.938e-03  -1.482  0.13916   
horsepower:origin          2.233e-03  2.930e-02   0.076  0.93931   
weight:acceleration        2.346e-04  2.289e-04   1.025  0.30596   
weight:year               -2.245e-04  2.127e-04  -1.056  0.29182   
weight:origin             -5.789e-04  1.591e-03  -0.364  0.71623   
acceleration:year          5.562e-02  2.558e-02   2.174  0.03033 * 
acceleration:origin        4.583e-01  1.567e-01   2.926  0.00365 **
year:origin                1.393e-01  7.399e-02   1.882  0.06062 . 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.695 on 363 degrees of freedom
Multiple R-squared:  0.8893,	Adjusted R-squared:  0.8808 
F-statistic: 104.2 on 28 and 363 DF,  p-value: < 2.2e-16
```

Figure 13. Summary of linear model with all interactions terms

Figure 13 shows that several of the interactions terms appear to be statistically significant in the presence of the other predictors as evident by their low p-values.

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



```
Call:
lm(formula = mpg ~ log(displacement) + sqrt(weight) + I(acceleration^2),
    data = Auto)

Residuals:
     Min       1Q   Median       3Q      Max 
-13.0312  -2.5460  -0.2837   2.1968  16.5357 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)       72.957826   3.102558  23.515  < 2e-16 ***
log(displacement) -3.890855   1.242746  -3.131  0.00188 ** 
sqrt(weight)      -0.574716   0.082743  -6.946 1.59e-11 ***
I(acceleration^2)  0.005983   0.002708   2.210  0.02771 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.135 on 388 degrees of freedom
Multiple R-squared:  0.7215,    Adjusted R-squared:  0.7194 
F-statistic: 335.1 on 3 and 388 DF,  p-value: < 2.2e-16
```

Figure 14. Summary of linear model with transformed variables

Figure 14 shows that the log of the displacement, the square root of the weight, and the square of the acceleration all appear to be statistically relevant which make since in light of the diagnostic plots of the linear regression fit shown in Figure 12.

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

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

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

c) Using x and eps, generate a vector y according to the model

$$Y = -1 + 0.5X + \epsilon. \qquad (3.39)$$

What is the length of the vector y? What are the values of β0 and β1 in this linear model?

i. The vector y has a length of 100. β0 has a value of -1. Whereas, β1 has a value of 0.5.

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

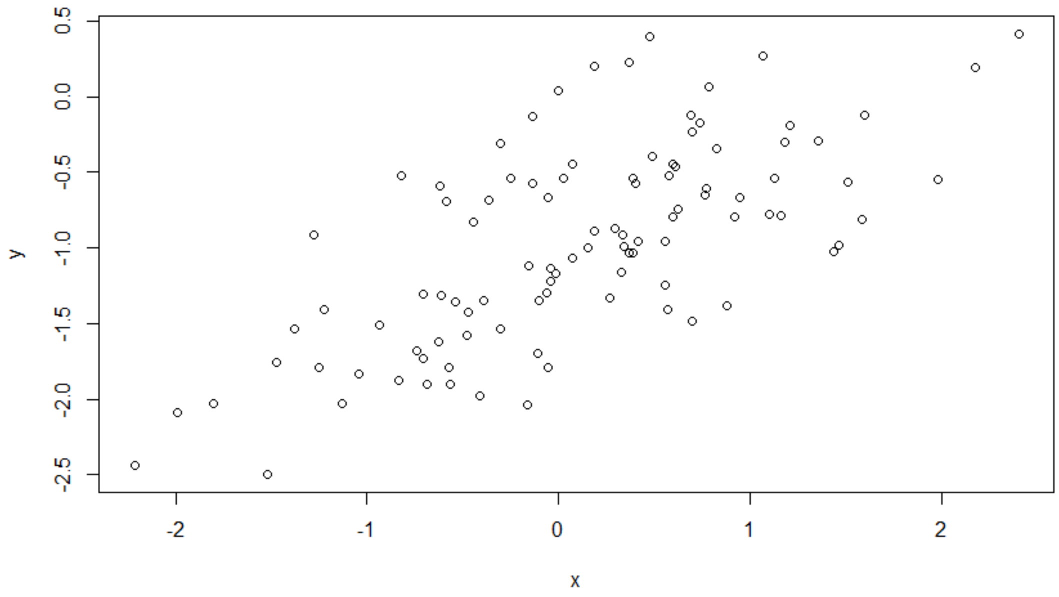
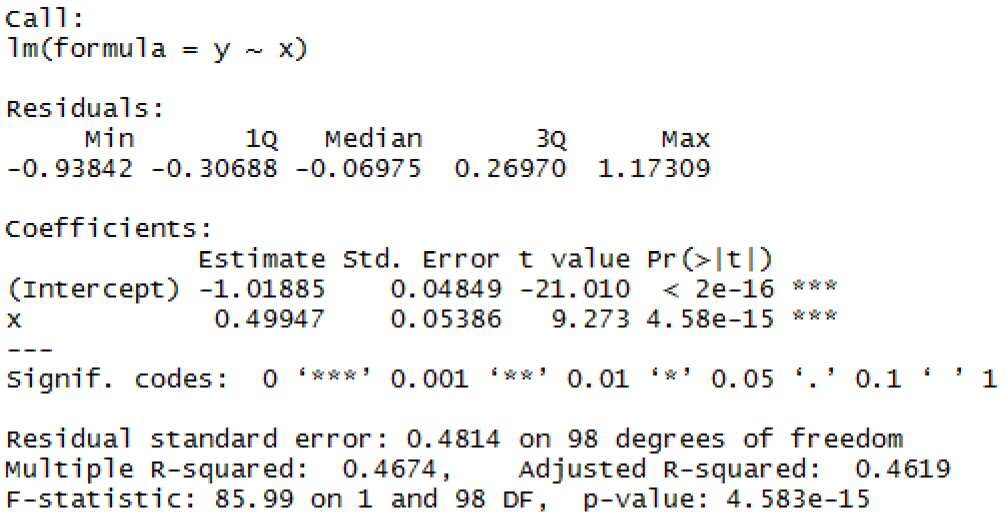


Figure 15. Scatterplot of y vs x

Figure 15 shows that the relationship between x and y appears to be linear with a positive slope, which is expected since the true fit equation 3.39, which is a linear line with a positive slope.

e) Fit a least squares linear model to predict y using x. Comment on the model obtained. How do β^0 and β^1 compare to β0 and β1?



```
Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.93842 -0.30688 -0.06975  0.26970  1.17309 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.01885    0.04849 -21.010  < 2e-16 ***
x            0.49947    0.05386   9.273 4.58e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4814 on 98 degrees of freedom
Multiple R-squared:  0.4674,	Adjusted R-squared:  0.4619 
F-statistic: 85.99 on 1 and 98 DF,  p-value: 4.583e-15
```

Figure 16. Summary of linear model of y given x

Figure 16 shows that the model of equation 3.39 is very close to the original equation. β^0 and β^1 are within .02 of the original β0 and β1.

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

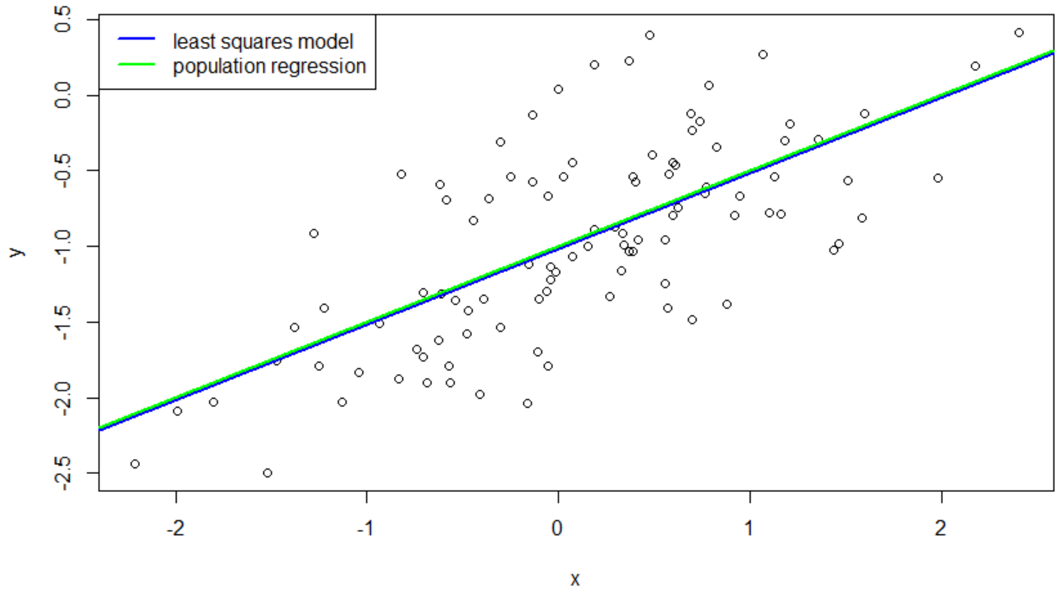
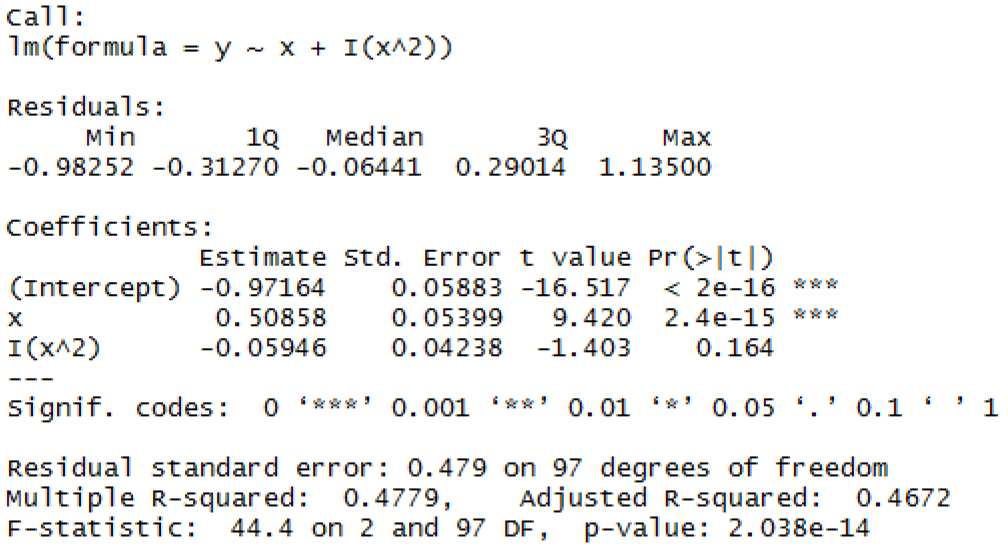


Figure 17. Scatterplot of y vs x with the true equation and the linear model of that equation.

g) Now fit a polynomial regression model that predicts y using x and x2. Is there evidence that the quadratic term improves the model fit? Explain your answer.



```
Call:
lm(formula = y ~ x + I(x^2))

Residuals:
     Min       1Q   Median       3Q      Max 
-0.98252 -0.31270 -0.06441  0.29014  1.13500 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.97164    0.05883 -16.517  < 2e-16 ***
x            0.50858    0.05399   9.420  2.4e-15 ***
I(x^2)      -0.05946    0.04238  -1.403    0.164    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.479 on 97 degrees of freedom
Multiple R-squared:  0.4779,	Adjusted R-squared:  0.4672 
F-statistic:  44.4 on 2 and 97 DF,  p-value: 2.038e-14
```

Figure 18. Summary of a polynomial regression of Equation 3.39.

Figure 18 shows that the Adjusted R-squared term increased suggesting that the polynomial model fits the training data better than that of the linear model. However, the p-value of x^2 is large suggesting that there is no relationship between it and the polynomial model has been overfitted.

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

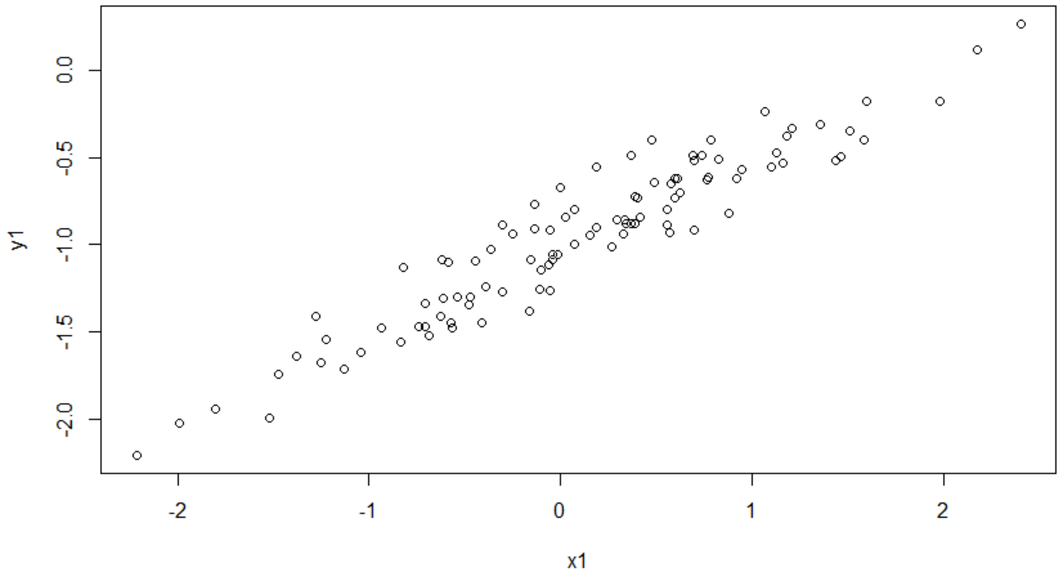
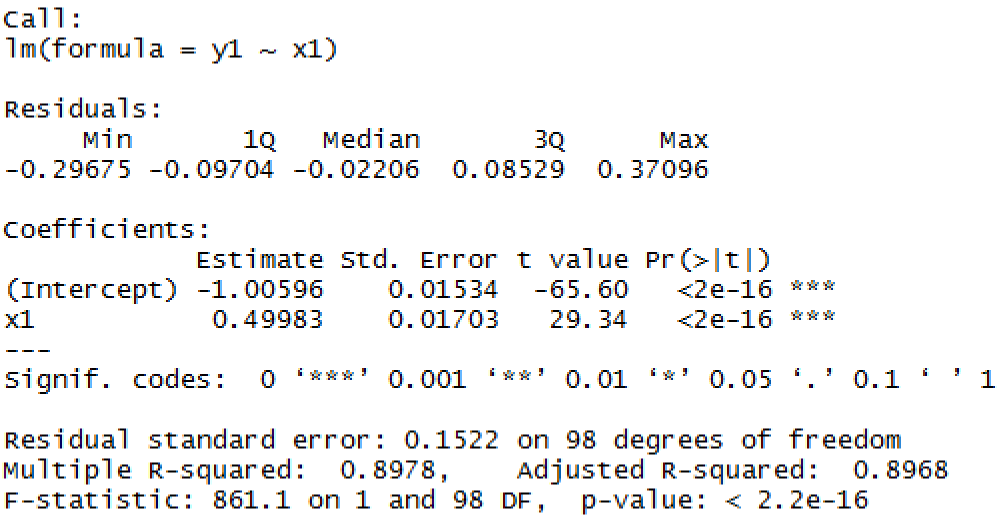


Figure 19. Scatterplot of y1 vs x1

Figure 19 shows that the relationship between x1 and y1 appears to be linear with a positive slope, which is expected since the true fit equation 3.39, which is a linear line with a positive slope. In addition, there is less variance than in Figure 15, which is expected as we reduced the variance of the error term.



```
Call:
lm(formula = y1 ~ x1)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.29675 -0.09704 -0.02206  0.08529  0.37096 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.00596    0.01534  -65.60   <2e-16 ***
x1           0.49983    0.01703   29.34   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1522 on 98 degrees of freedom
Multiple R-squared:  0.8978,	Adjusted R-squared:  0.8968 
F-statistic: 861.1 on 1 and 98 DF,  p-value: < 2.2e-16
```

Figure 20. Summary of linear model of y1 given x1

Figure 20 shows that the model of equation 3.39 is very close to the original equation. β^0 and β^1 are within .006 of the original β0 and β1 vs .02 in Figure 16.

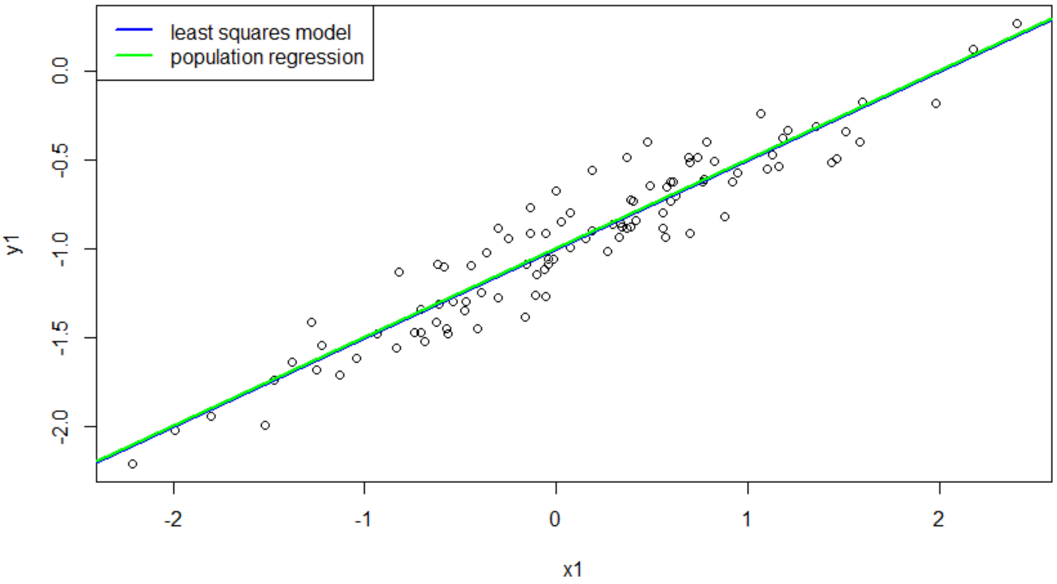


Figure 21. Scatterplot of y1 vs x1 with the true equation and the linear model of that equation.

Figure 21 shows that the linear model in h) is closer to the true equation than that of the previous linear model shown in Figure 17, which is as expected since we decreased the variance of the error term and thus the variance of the data.

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

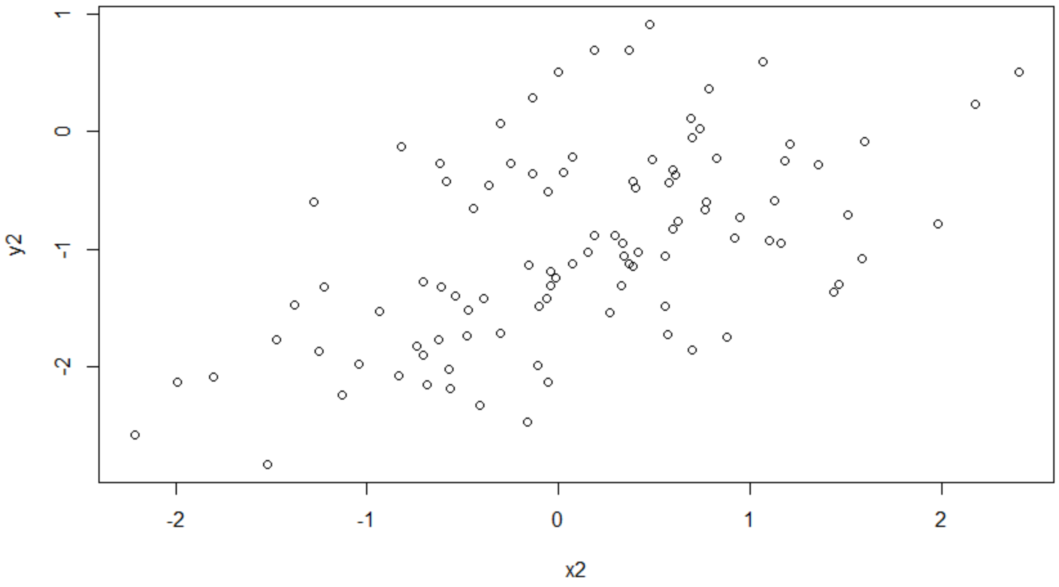
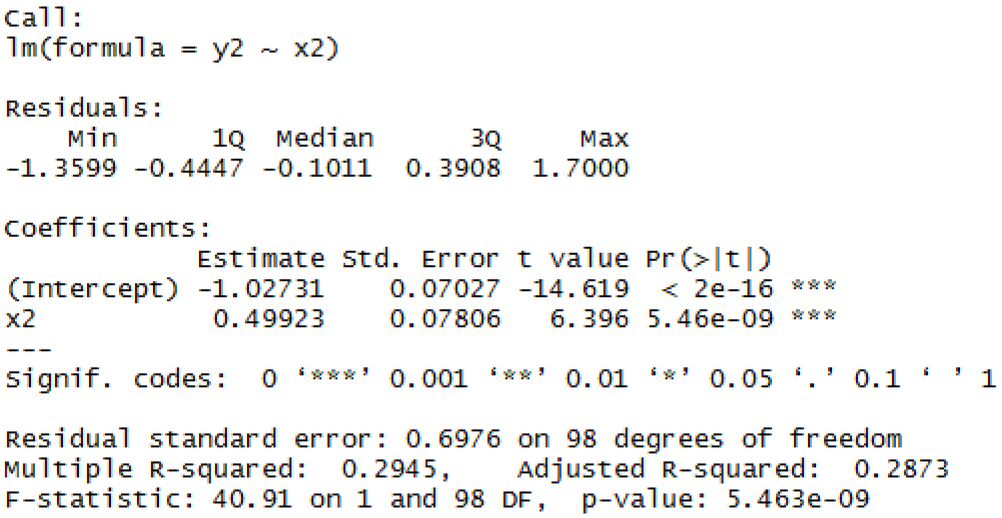


Figure 22. Scatterplot of y2 vs x2

Figure 22 shows that the relationship between x2 and y2 appears to be linear with a positive slope, which is expected since the true fit equation 3.39, which is a linear line with a positive slope. However, there is a lot of more variance than in Figure 15, which is expected as we increase the variance of the error term. This will result in a degraded linear fit.



```
Call:
lm(formula = y2 ~ x2)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.3599 -0.4447 -0.1011  0.3908  1.7000 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.02731    0.07027 -14.619  < 2e-16 ***
x2           0.49923    0.07806   6.396 5.46e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6976 on 98 degrees of freedom
Multiple R-squared:  0.2945,	Adjusted R-squared:  0.2873 
F-statistic: 40.91 on 1 and 98 DF,  p-value: 5.463e-09
```

Figure 23. Summary of linear model of y1 given x1

Figure 23 shows that the model of equation 3.39 is again very close to the original equation. βˆ0 and βˆ1 are within .03 of the original β0 and β1 vs .02 in Figure 16, which is as expected given the increased variance of the error term.

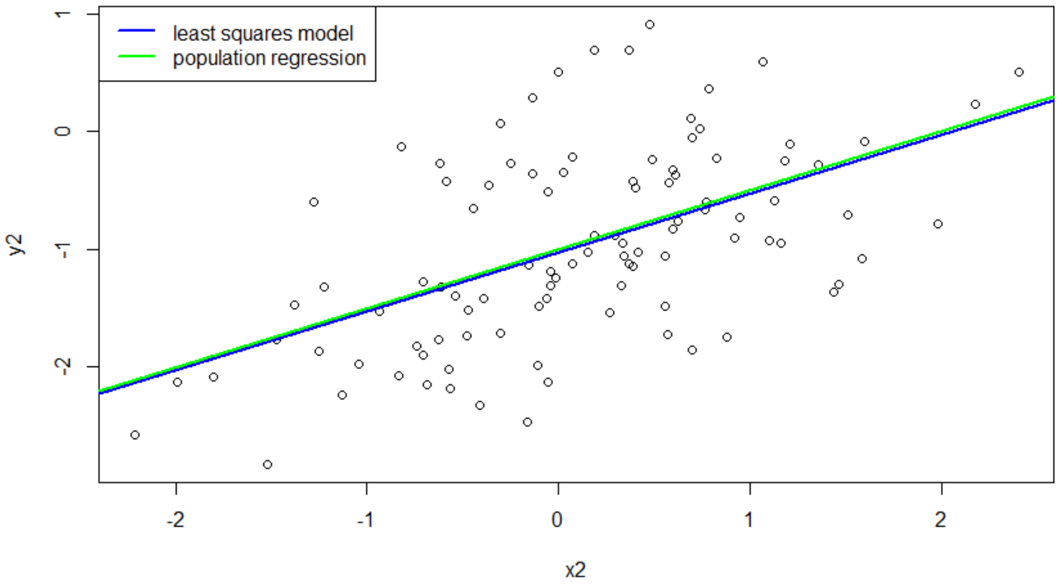
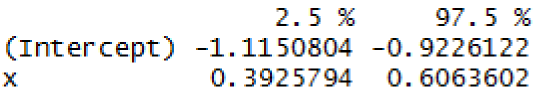


Figure 24. Scatterplot of y2 vs x2 with the true equation and the linear model of that equation.

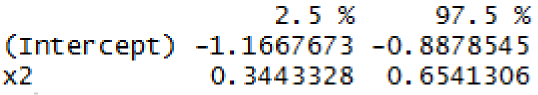
Figure 24 shows that the linear model in i) is farther away from the true equation than that of the previous linear model shown in Figure 17, which is as expected since we increased the variance of the error term and thus the variance of the data.

j) What are the confidence intervals for β0 and β1 based on the original data set, the noisier data set, and the less noisy data set? Comment on your results.



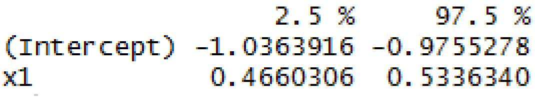
```
                 2.5 %     97.5 %
(Intercept) -1.1150804 -0.9226122
x            0.3925794  0.6063602
```

Figure 25. Confidence interval for β0 and β1 based on the original data set



```
                 2.5 %     97.5 %
(Intercept) -1.1667673 -0.8878545
x2           0.3443328  0.6541306
```

Figure 26. Confidence interval for β0 and β1 based on the noisier data set



```
                  2.5 %     97.5 %
(Intercept) -1.0363916 -0.9755278
x1           0.4660306  0.5336340
```

Figure 27. Confidence interval for β0 and β1 based on the less noisy data set

Figures 25 through 27 show that as the variance increases the confidence interval widens and vice versa.