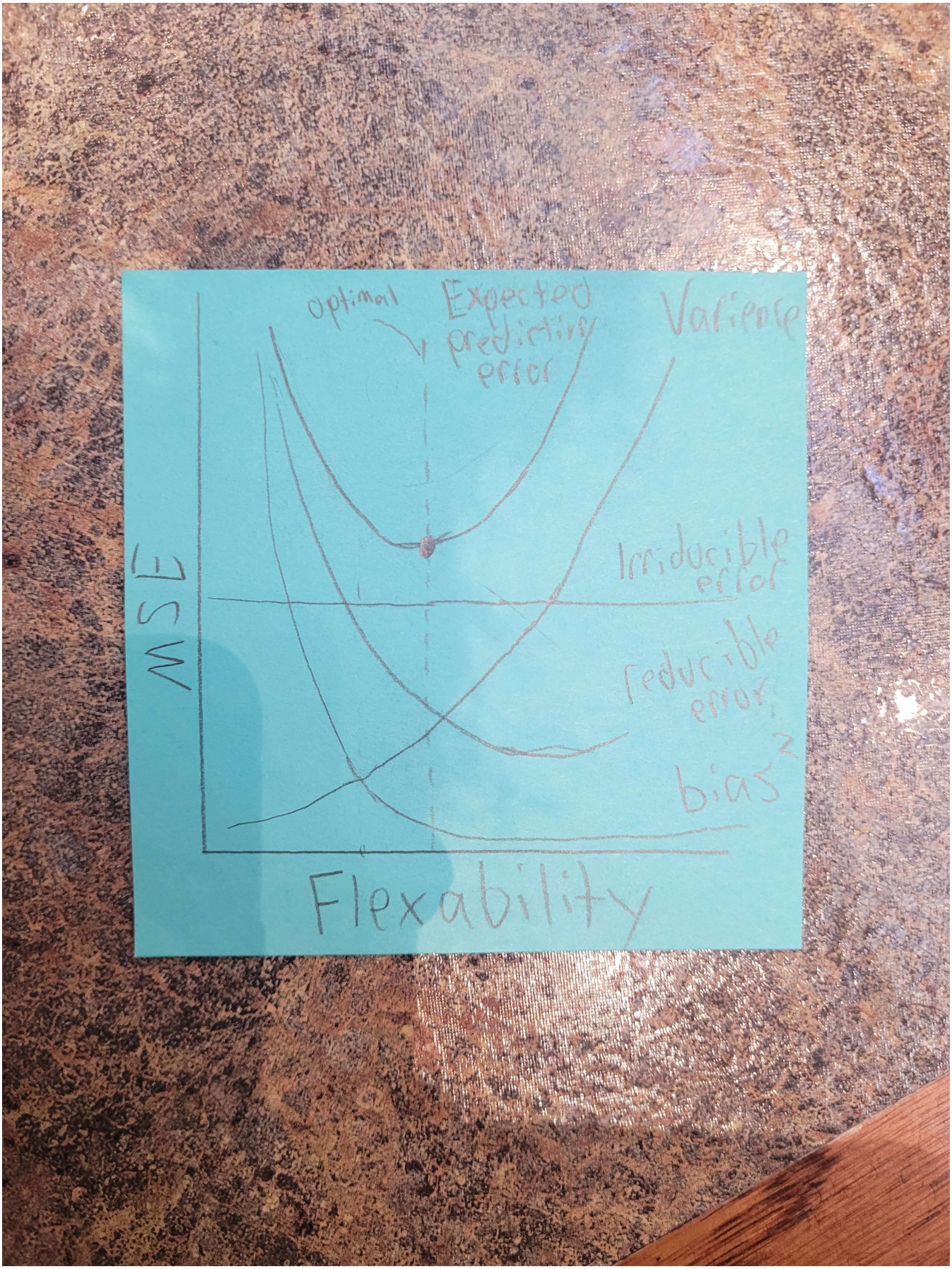
Homework 1

Tanner Huck

4/18/2017

Problem 1

a. Considering a data set of housing prices for Randomville, our goal is to develop a model that can predict house prices based on its features. The goal of this model is to generalize well to the testing data set, which need to consider and select the optimal model complexity with the bias-variance trade off in mind. In our context, bias refers to the difference between a predicted house price from our model and the true house price. High bias might signify an oversimplification of the relationship between features and the price and result in under fitting. Whereas low bias might over fit and describe the training data very well. Variance in our context is the variability of the models predictions. High variance might over fit the training data and could be sensitive to outliers and noise. Where a low variance under fits and is less sensitive to outlines. The bias-variance trade off shows the relationship between these two attributes. When thee model becomes more complicated, the bias tends to decrease and the variance tends to increase. Hence in our context, as our model’s complexity increases, we might start to fit the training data well, but we may be sensitive to outliers and over fits. Our goal is to find the optimal balances between the bias and variance that also gives us the best predictions of housing prices.

b. 

c. Consider i. a linear model with only the number of bedrooms as a feature, ii. A linear model with the number of bedrooms, square footage, and location score as features, and iii. A polynomial model of degree 10 using the number of bedrooms, square footage, and location score as features. Our goal is to rank these models

with respect to the bias, variance, training MSE, and test MSE that we would expect them to achieve.

i. This model is not very complicated and simplifies the relationship between input features and price. This leads to high bias and low variance. We can also expect a high training MSE because we under fit the training data, which also leads to a higher test MSE.

ii. This model captures more complexity compared to i because of the addition of more features. Leading to a similar result to i, but lower bias, higher variance, and lower training and testing MSE. This is because we are more flexible and has a better bias-variance trade off.

iii. This model is the most complex, with the most features added. Following the same reasoning, we expect a low bias, high variability, and low training MSE (start to fit the training data very well). However we will see a high test MSE, because we will start to over fit to the training data, which leads to poor generalization for the testing data.

Overall I would guess that model ii would have the best results. This is because i. will underfit the data and iii. will overfit the data, leaving ii. somewhere in the middle. However, based on the specifics of the data set, this may be untrue.

Problem 2

Intro) Assuming that Apara is applying for his first credit card with the bank given that his existing credit is 0 and that he currently has 1 dollar in his bank account, with an annual income of 0, we want to make some k-nearest neighbors analysis to see if he will be accepted.

a. First, we will compute the Euclidean distance between him and all other applicants. We know that the Euclidean distance between two points is defined as $\|a-b\|_2 = \sqrt{\sum_{l=1}^{m}(a_l - b_l)^2}$. Thus solving for Apara's Ecuadorian distance to all other applicants,

```
## [1] 1.414214 2.449490 5.099020 4.582576 5.385165 6.782330
```

Hence the distance to each applicant is shown above, for example, the distance to the first applicant is about 1.41. (See appendix for code).

b. Now we want to see what our prediction would be based on the first closest neighbor. Based on the distances that we found in part a, we know that Apara's closest neighbor is applicant 1. We can see that applicant 1's decision was No, hence based on $k = 1$ our prediction would be no.

c. Now finding the prediction for the nearest 3 neighbors. Again based on our distances found in part a, we know that Apara's 3 closest neighbors are applicant 1, 2, and 4. These applicants were rejected, accepted, and rejected respectfully. Since there are more rejects than accepts for the three closest neighbors, we would predict no for $k = 3$.

d. Suppose that the Bayes decision boundary is approximately linear. This means that a smaller value of k is more likely to provide better predictions and over fit. This is because a smaller k value will lead to less flexibility and make better predictions. Hence a smaller k would create a non-linear boundary which is very specific to the data, which is not representative of the true boundary that is linear, hence we would want to choose $k = 3$ when the boundary is linear.

On the other hand, if the boundary is highly non-linear, than we would want to choose the small value for k. This is because a larger k will have more flexibility and will make better predictions when there is noise/outliers in the data, creating a more linear boundary. Meaning that for a highly non-linear boundary, we would choose $k = 1$ over

$k = 3$. However, this may not hold for all data sets.

e. Note that in our data set, the bank balance/debt (x1) and annual income (x2) values, are of the same order of magnitude as the number of credit lines (x3). This is important because if the features were not all approximately on the same scale, the features with larger values would have a greater impact when making predictions. This is because the features with larger values will lead to a higher Euclidean distance, which is the distance we used in the k-NN classifier. Meaning that higher values will have greater impact on the prediction, even if it is not more important in reality.

f. Assume now that the bank also makes decisions based on the categorical values of ethnicity and gender to make decisions. This is problematic both ethically and mathematically. Ethically, making decisions based on someones ethnicity or gender can violate privacy. Collecting and storing someones gender and ethnicity in data might not have the best intentions. This can also lead to discrimination, even with good intentions; making decisions based on someones ethnicity or gender can lead to bias and make people feel bad about themselves. Mathematically, assume that we assign a number for each gender and ethnicity. This is problematic because when we solve for the Euclidean distance, someone of a specific ethnicity or gender may always have a higher or lower distance, which leads to bias. We would also have to assign each gender and ethnicity a numerical value, but how would you choose which groups is which value?

Problem 3

Intro) Consider a data set of n independent observations of (X, Y). We want to predict Y with X using the linear model $Y = \beta_0 + \beta_1 x_i + \epsilon_i$ where $\epsilon_i \sim N(0, \sigma^2)$ for each $i = 1, \ldots, n$. We estimate the regression coefficients using the least squares approach. We want to show that the coefficient of determination, R^2 is equal to the $(Cov(X, Y))^2$, the sure of the correlation between X and Y.

Proof) Starting with what we know about R^2, we can write it as

$$R^2 = 1 - \frac{RSS}{TSS}$$

where RSS is the Residual Sum of Squares and TSS is the Total Sum of Squares

then substituting in the formulas for RSS and TSS

$$= 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$= 1 - \frac{\sum_{i=1}^{n}(y_i^2 - 2y_i\hat{y}_i + (\hat{y}_i)^2)}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$= 1 - \frac{\sum_{i=1}^{n}y_i^2 - \sum_{i=1}^{n}2y_i\hat{y}_i + \sum_{i=1}^{n}(\hat{y}_i)^2}{\sum_{i=1}^{n}y_i^2 - \sum_{i=1}^{n}2\bar{y}y_i + \sum_{i=1}^{n}\bar{y}^2}$$

$$= \frac{\sum_{i=1}^{n}y_i^2 - \sum_{i=1}^{n}2\bar{y}y_i + \sum_{i=1}^{n}\bar{y}^2 - \sum_{i=1}^{n}y_i^2 + \sum_{i=1}^{n}2y_i\hat{y}_i - \sum_{i=1}^{n}(\hat{y}_i)^2}{\sum_{i=1}^{n}y_i^2 - \sum_{i=1}^{n}2\bar{y}y_i + \sum_{i=1}^{n}\bar{y}^2}$$

$$= \frac{\sum_{i=1}^{n}(-\hat{y}_i + \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$= \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}i - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}$$

$$= \left(\frac{\mathrm{Cov}(X, Y)}{\sigma_X\sigma_Y}\right)^2$$

$$= (\mathrm{Cor}(X, Y))^2$$

Hence R^2 is equal to $(Cov(X, Y))^2$.

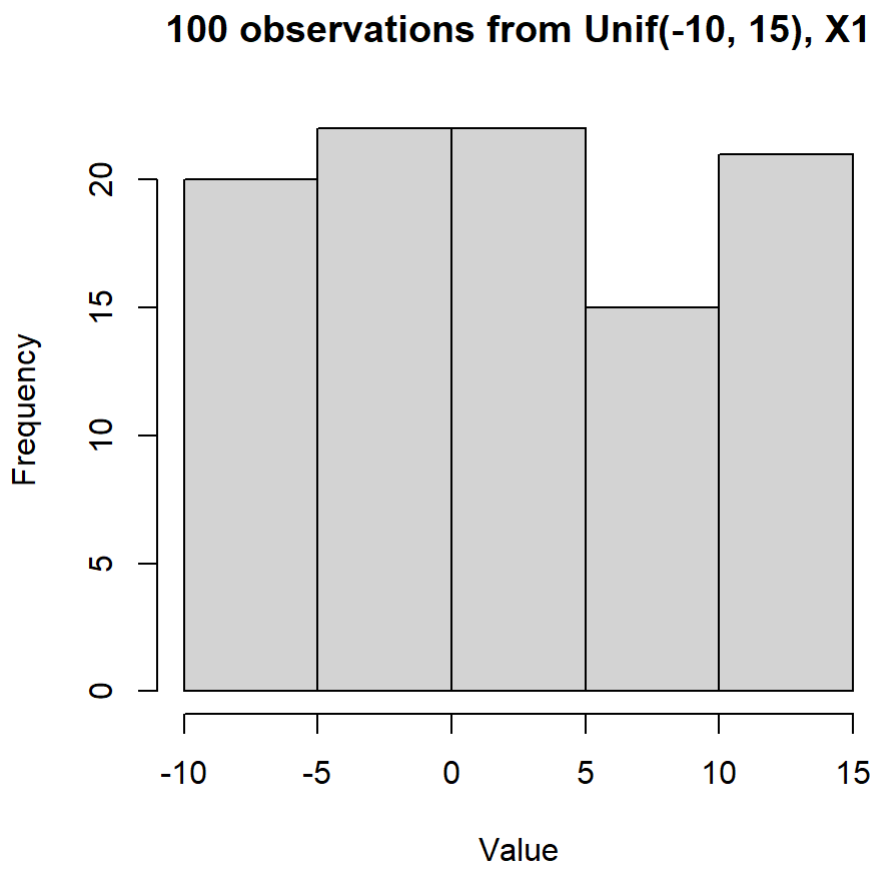
Problem 4

a. First, we will generate 100 values from a $Unif(-10, 15)$ distribution and 100 values from a $Exp(0.5)$ distribution. Then creating a histogram of our values.

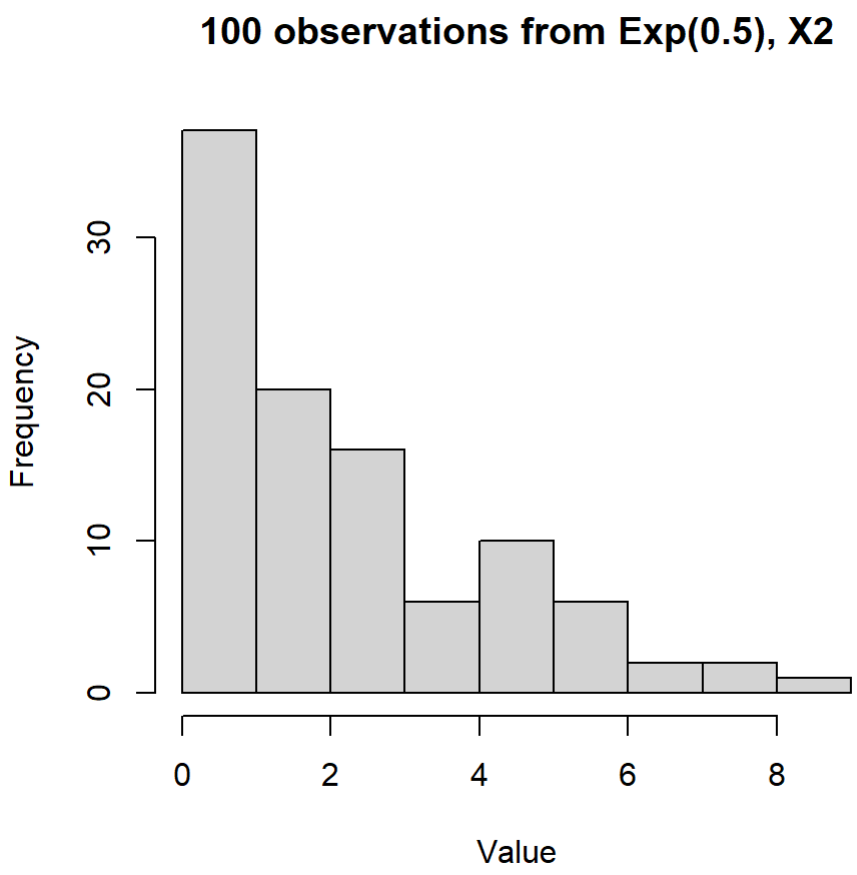
```
set.seed(333)

x1 <- runif(100, -10, 15)
x2 <- rexp(100, 0.5)

hist(x1, main = "100 observations from Unif(-10, 15), X1", xlab = "Value")
```




```
hist(x2, main = "100 observations from Exp(0.5), X2", xlab = "Value")
```

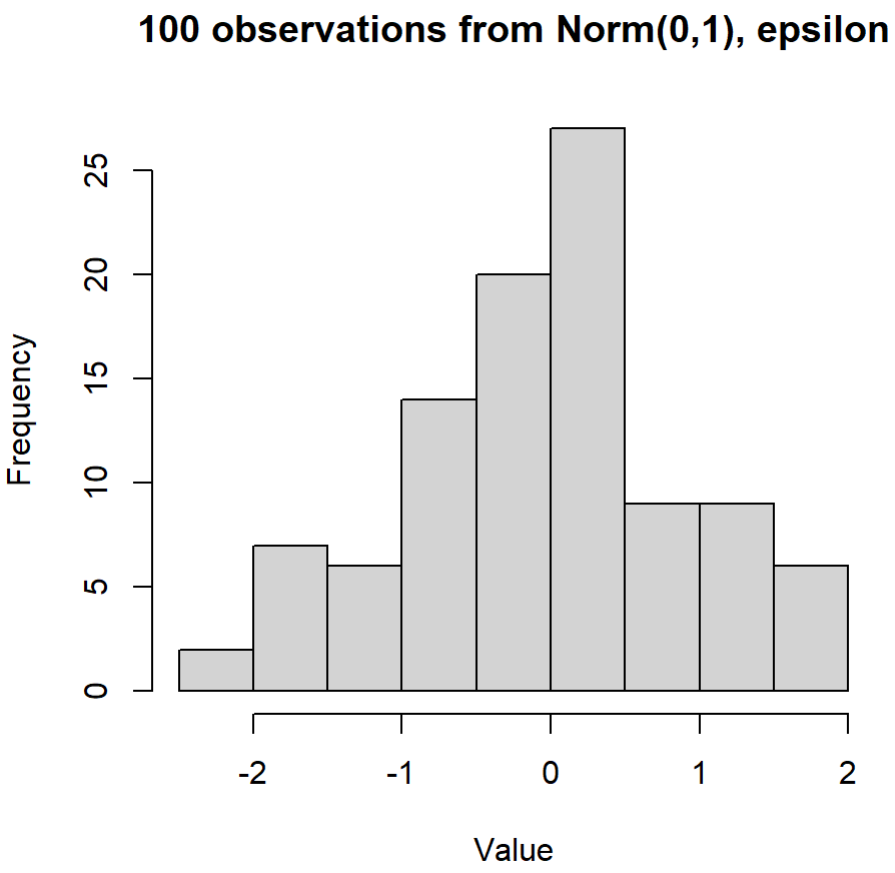


b. Second, generating 100 random values from a $N(0, \sigma_0^2)$ distribution with $\sigma_0^2 = 1$ to represent noise.

```
set.seed(333)

epsilon <- rnorm(100, 0, 1)

hist(epsilon, main = "100 observations from Norm(0,1), epsilon", xlab = "Value")
```

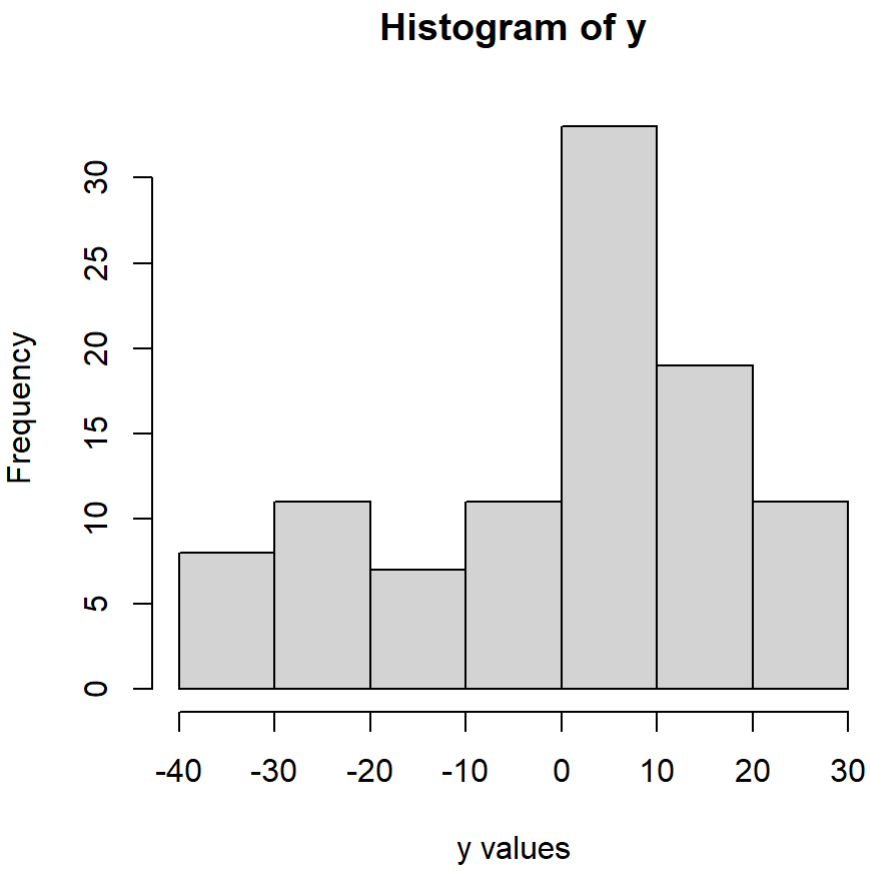


c. Third, we want to compute the response variable y using $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \epsilon$ where $\beta_0 = 1$, $\beta_1 = -3$, $\beta_2 = 2$, $\beta_3 = 0.75$, and x_1, x_2, ϵ are found in part a and b.

```
b0 <- 1
b1 <- -3
b2 <- 2
b3 <- 0.75

y <- b0 + b1*x1 + b2*x2 + b3*x1*x2 + epsilon

hist(y, main = "Histogram of y", xlab = "y values")
```



d. Fourth, creating a least squares linear model to predict y using x_1, x_2 and printing the summary to compare coefficients.

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -44.315  -5.099   0.834   3.858  36.763
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.6959     1.6593  -1.022    0.309
## x1           -1.5698     0.1465 -10.718  < 2e-16 ***
## x2            2.4184     0.5452   4.436 2.42e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.38 on 97 degrees of freedom
## Multiple R-squared:  0.6247, Adjusted R-squared:  0.617
## F-statistic: 80.75 on 2 and 97 DF,  p-value: < 2.2e-16
```

From this model we can see that the estimated coefficients for x_1, x_2 are about -1.57 and 2.42 which are somewhat close to the true parameters that we had earlier of 3 and 2 respectively. The coefficient for the intercept is -1.69.

e. Fifth, creating a least squares linear model to predict y using $x_1, x_2, x_1^2, x_2^2, x_1x_2$ and printing the summary to compare coefficients.

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x1_sqr + x2_sqr + x1_x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.18696 -0.50853  0.00502  0.52612  1.85355
##
## Coefficients:
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept)  1.047284   0.233304    4.489 2.03e-05 ***
## x1          -3.019788   0.024501 -123.250  < 2e-16 ***
## x2           1.794055   0.161701   11.095  < 2e-16 ***
## x1_sqr       0.002217   0.001996    1.110    0.270
## x2_sqr       0.034463   0.023016    1.497    0.138
## x1_x2        0.750506   0.007726   97.135  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9601 on 94 degrees of freedom
## Multiple R-squared:  0.9969, Adjusted R-squared:  0.9967
## F-statistic:  6021 on 5 and 94 DF,  p-value: < 2.2e-16
```

From this model we can see that the estimated coefficients for $x_1, x_2, x_1^2, x_2^2, x_1x_2$ are about -3.02, 1.79, 0.002, 0.03, and 0.75 respectively. The coefficient for the intercept is 1.05. Comparing these coefficients to the true coefficients and ones found in part d, these new coefficients are much closer to the true coefficients. Take x_1 for example, the new coefficient -3.02 is much closer to the true value of -3 then the value found in part d of -1.57. This implies that the additional terms are improving the model.

Then looking at the t-statistic and p-values for the x_1^2, x_2^2 variables, we can see that they are above 0.05. This tells us that we can consider dropping these variables from the model, as they do not significantly impact the output y.

f. Lastly, we want to perform an F-test using *anova* to identify which of the models we created in part d and e fit the data best.

```
## Analysis of Variance Table
##
## Model 1: y ~ x1 + x2
## Model 2: y ~ x1 + x2 + x1_sqr + x2_sqr + x1_x2
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1     97 10446.7
## 2     94    86.7  3     10360 3746.2 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

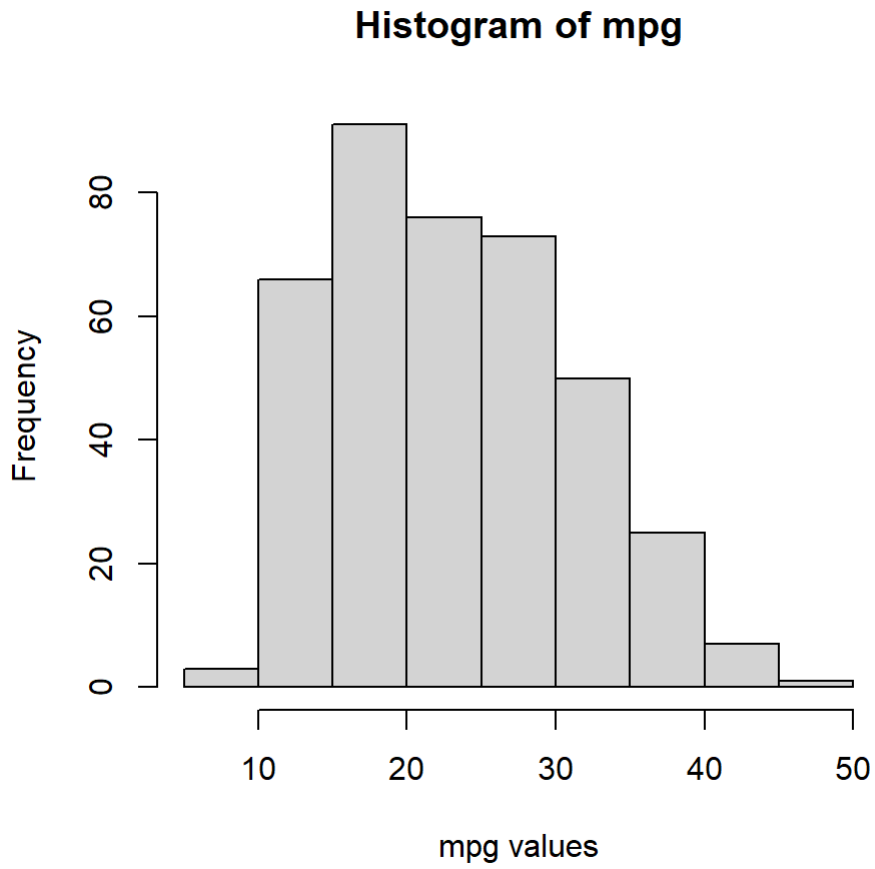
From this F-test, we can see the F-statistic is about 3746.2 and a p-value of about $2.2 * 10^{-16}$. This tells us that we have significant evidence against the null hypothesis that both models perform similarly when predicting y. We can conclude that the model created in part e with more variables is better at predicting y than the model created in

part d. Additionally, the degrees of freedom is 3, which signifies that the model in part e has 3 more variables than the model in part d.

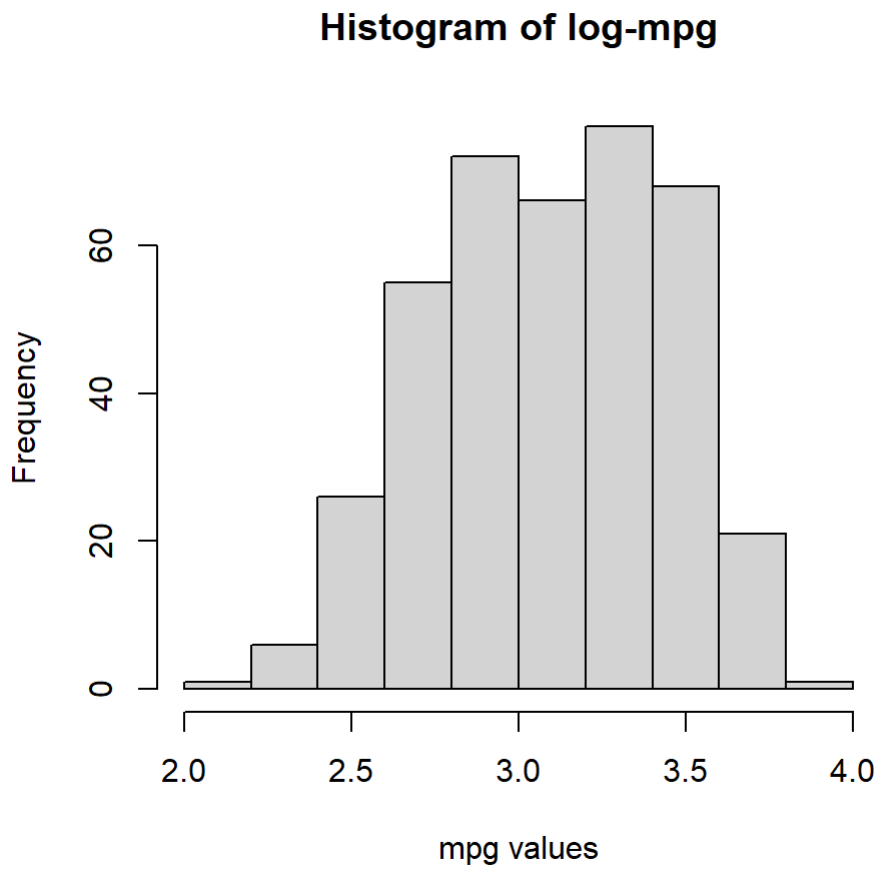
Problem 5

Intro) The *Auto* data set contains several variables like *mpg* or *cylinders*. Our goal in this problem is to predict the mileage of a car, *mpg*, using the other variables.

a. First, loading in the data and plotting a histogram of *mpg*. We want a general understanding of what the distribution looks like.

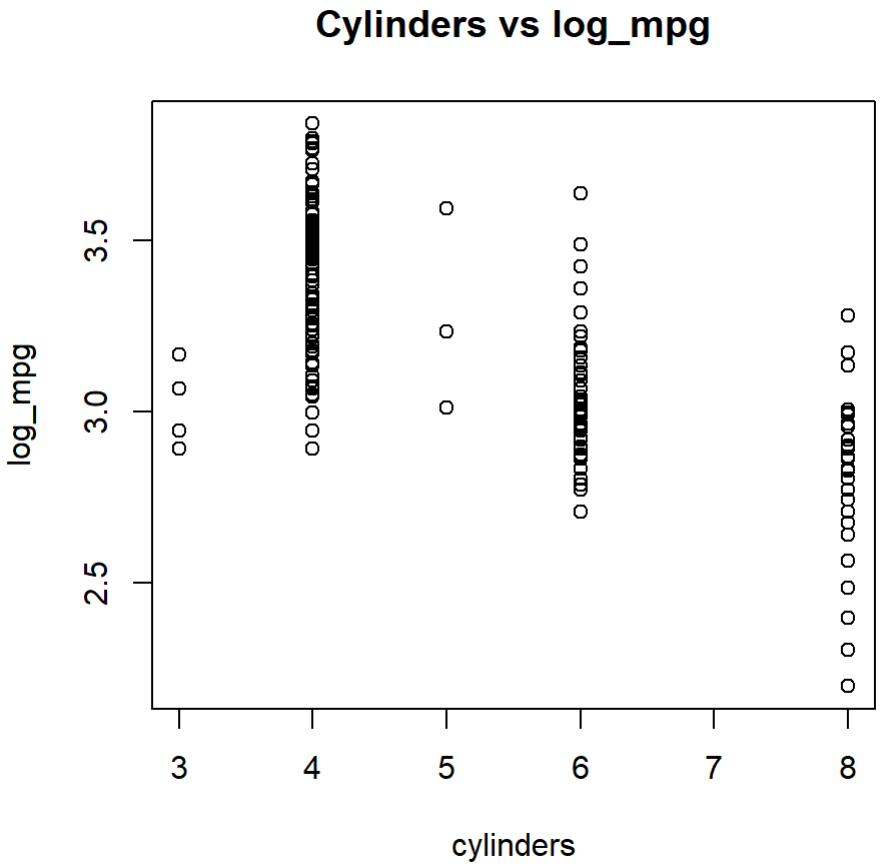


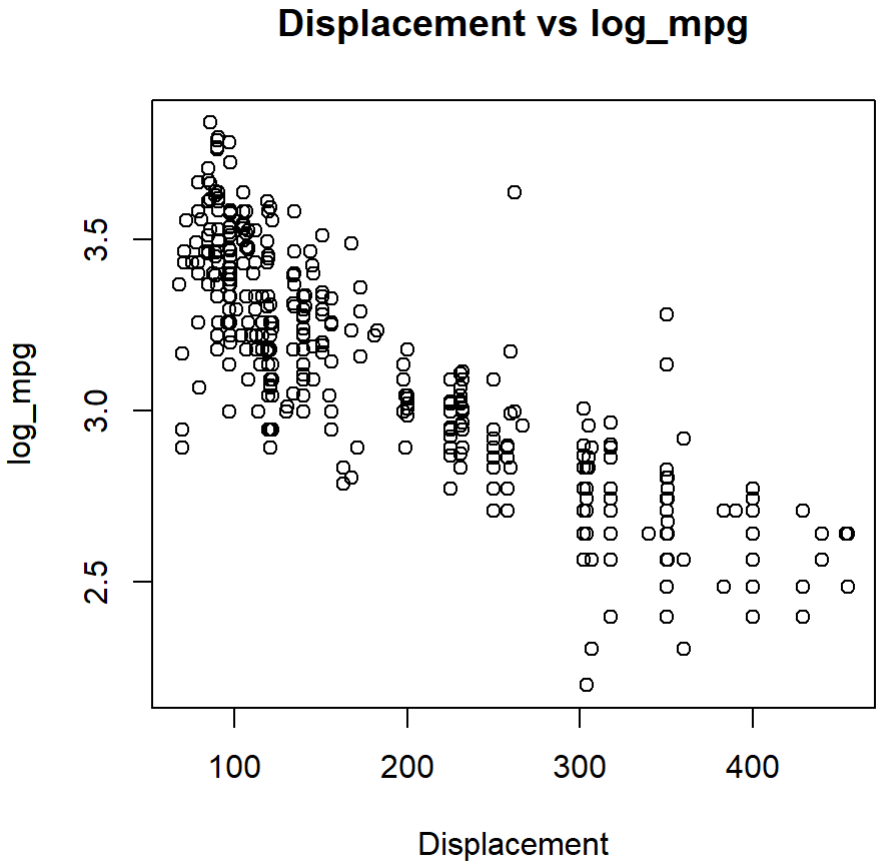
From the histogram we created we can see that *mpg* is a bit skewed to the right. Thus creating a log transform mpg we get,

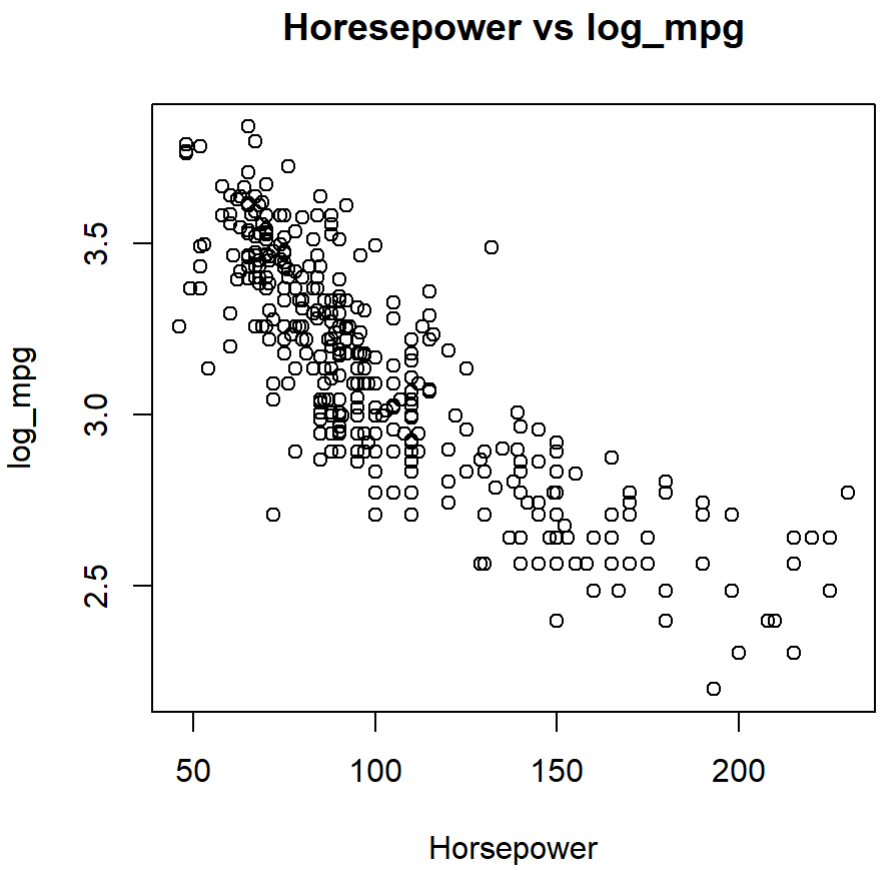


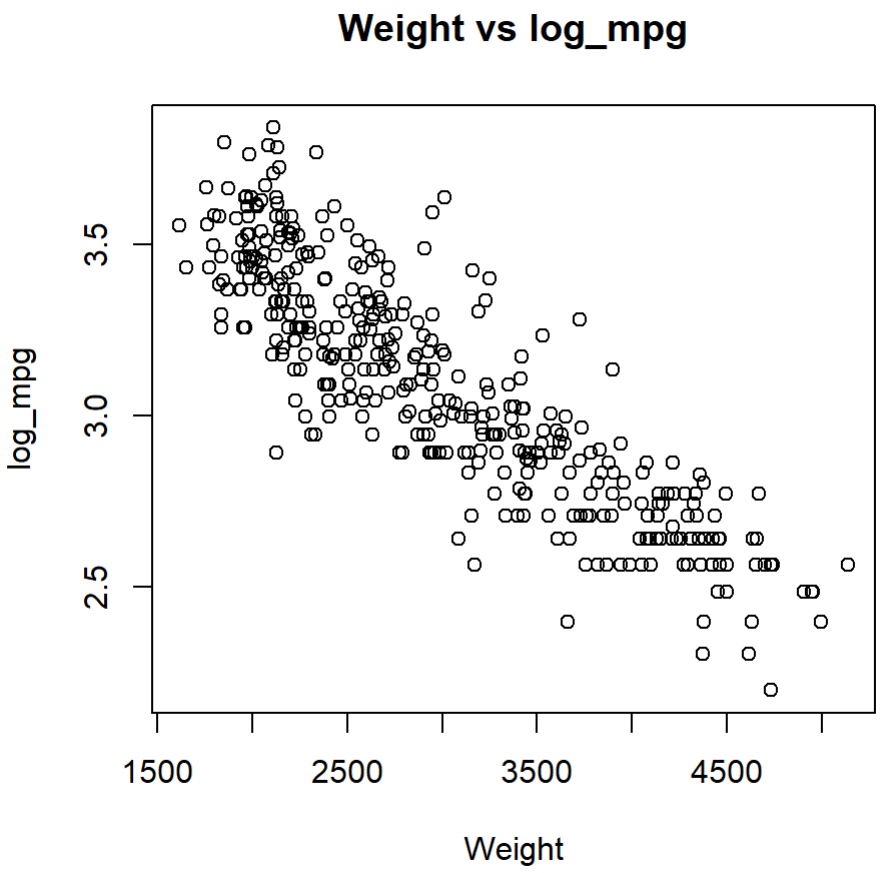
This new histogram of the logged mpg values appears to be more approximately normal.

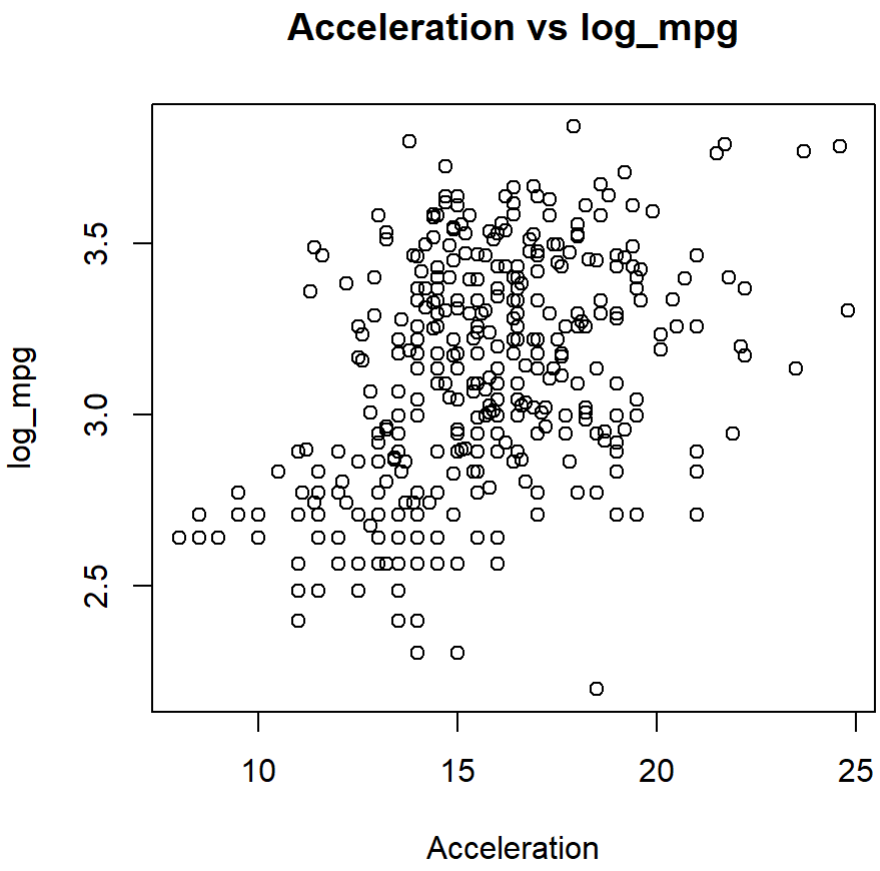
b. Secondly, creating a scatter plot between mpg and the other variables.

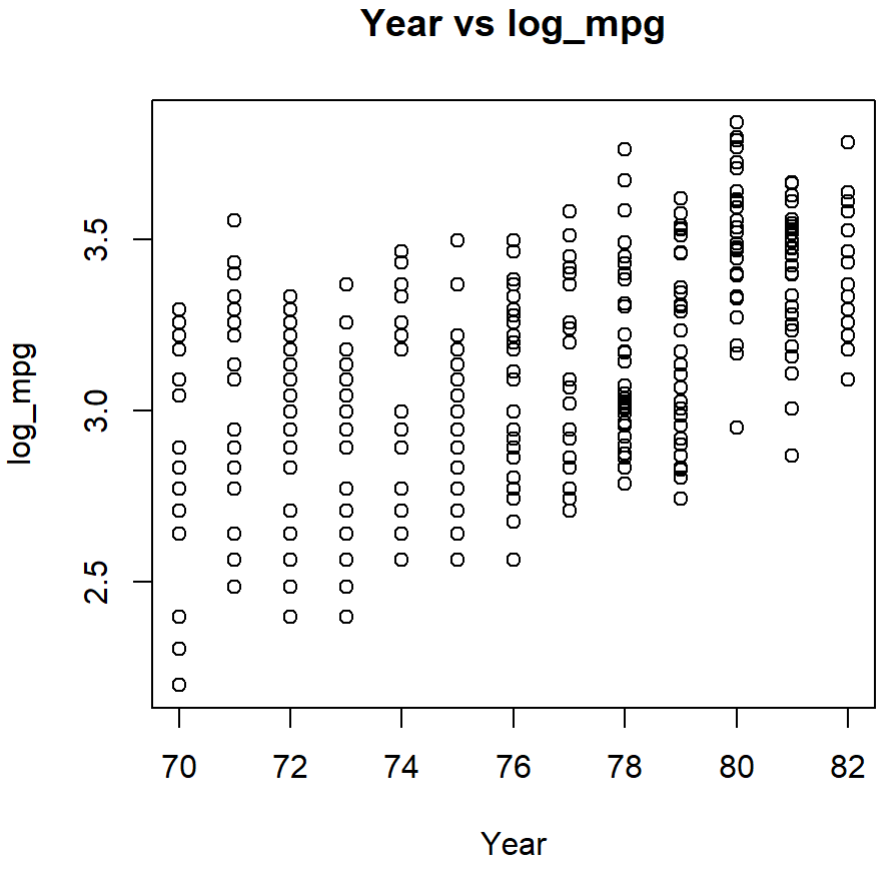

Cylinders vs log_mpg
log_mpg
2.5
3.0
3.5
3
4
5
6
7
8
cylinders


Displacement vs log_mpg
log_mpg
2.5
3.0
3.5
100
200
300
400
Displacement


Horesepower vs log_mpg
log_mpg
2.5
3.0
3.5
50
100
150
200
Horsepower


Weight vs log_mpg
log_mpg
2.5
3.0
3.5
1500
2500
3500
4500
Weight


Acceleration vs log_mpg
log_mpg
2.5
3.0
3.5
10
15
20
25
Acceleration


Year vs log_mpg
log_mpg
2.5
3.0
3.5
70
72
74
76
78
80
82
Year

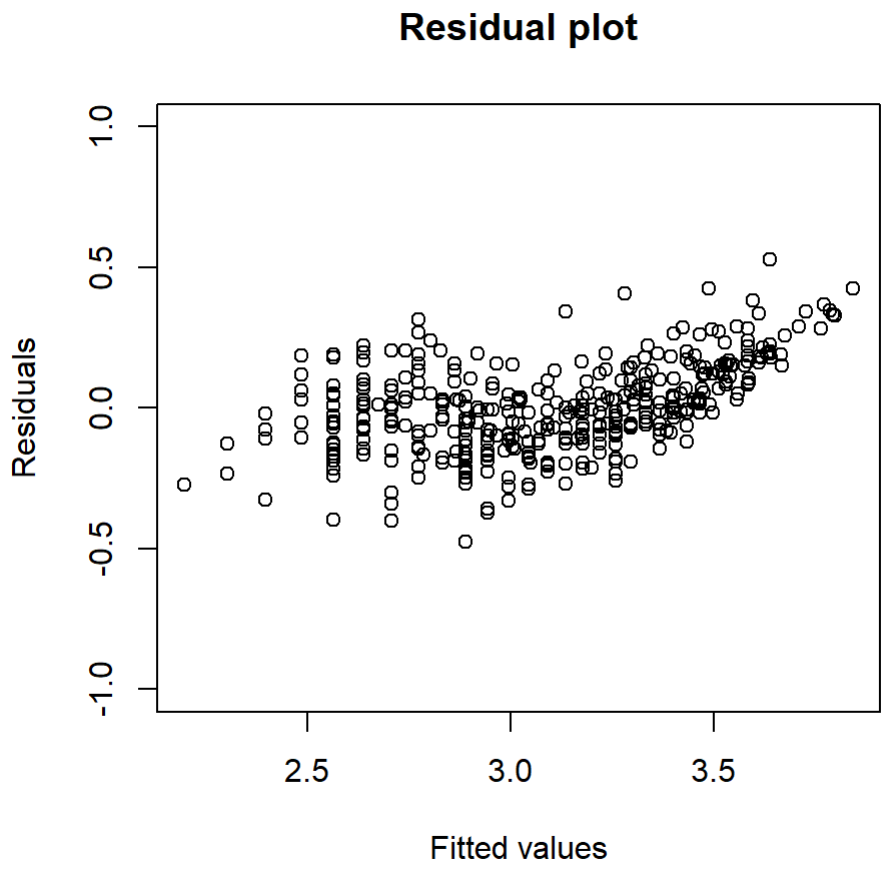
From these scatter plots, we can see that the variables that seem relevant are cylinders, displacement, horsepower, and weight. This is because these variables appear to have a negative correlation with mpg, meaning as mpg decrease the other variables increase. The other variables acceleration and year, do not appear to have as strong of a correlation.

c. Based on my results so far, I am choosing to create a linear model to predict mpg using the variables displacement, horsepower, and weight.

```
##
## Call:
## lm(formula = log_mpg ~ displacement + horsepower + weight, data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.47436 -0.10001 -0.00689  0.09827  0.52821
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.063e+00  4.420e-02  91.922  < 2e-16 ***
## displacement -3.506e-04  2.433e-04  -1.441     0.15
## horsepower   -2.215e-03  4.736e-04  -4.677 4.02e-06 ***
## weight       -2.235e-04  2.633e-05  -8.487 4.54e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1568 on 388 degrees of freedom
## Multiple R-squared:  0.7891, Adjusted R-squared:  0.7874
## F-statistic: 483.8 on 3 and 388 DF,  p-value: < 2.2e-16
```

From my linear model the coefficients for displacement, horsepower, and weight are about $-3.506 * 10^{-4}$, $-2.215 * 10^{-3}$, and $-2.235 * 10^{-4}$ respectively. The coefficient for intercept is 4.06.

d. Now we we want to generate residual plots to check the model diagnostics.



Looking at the residual plot, it is unlikely there is any indication of heteroscedasticity because the spread of the residuals seems to decrease as the fitted values increase. There also seems to be no pattern to the data, so I will leave my model as is.

e. Next we want to expand our current model to include more predictors. I will add quadratic terms for the displacement, horsepower, and weight variables to my model. I want to add these variables to my model because they might help explain some of the non-linear relationship between mpg and th variables. I also want to add an interaction term between horsepower and weight, because these variables might be related.

```
##
## Call:
## lm(formula = log_mpg ~ displacement + horsepower + weight + displacement_sqr +
##     horsepower_sqr + weight_sqr + horse_weight, data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.49781 -0.09089 -0.00575  0.09538  0.55664
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       4.133e+00  1.383e-01  29.881  < 2e-16 ***
## displacement     -3.066e-03  7.611e-04  -4.028 6.77e-05 ***
## horsepower       -7.140e-03  2.023e-03  -3.529 0.000467 ***
## weight            5.049e-05  1.359e-04   0.371 0.710534
## displacement_sqr  5.669e-06  1.570e-06   3.612 0.000345 ***
## horsepower_sqr   -3.027e-06  1.025e-05  -0.295 0.767914
## weight_sqr       -5.387e-08  3.260e-08  -1.653 0.099226 .
## horse_weight      1.281e-06  9.737e-07   1.316 0.188974
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1503 on 384 degrees of freedom
## Multiple R-squared:  0.8081, Adjusted R-squared:  0.8046
## F-statistic: 231.1 on 7 and 384 DF,  p-value: < 2.2e-16
```

After including these additional quadratic variables, the coefficients for displacement, horsepower, and weight are about -0.00307, $-0.007.14$, and -0.0000505 and the coefficients for their squared counter parts are 0.00000567, -0.00000303, and -0.0000000539. The coefficient for intercept is 4.13. The coefficient for horsepower times weight is 0.00000128

f. Finally, doing an F-test with anova to compare the models from part d and e.

```
## Analysis of Variance Table
##
## Model 1: log_mpg ~ displacement + horsepower + weight
## Model 2: log_mpg ~ displacement + horsepower + weight + displacement_sqr +
##     horsepower_sqr + weight_sqr + horse_weight
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    388 9.5360
## 2    384 8.6742  4   0.86184 9.5382 2.318e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From this F-test, we can see the F-statistic is about 9.53 and a p-value of about $2.3 * 10^{-7}$. This tells us that we have significant evidence against the null hypothesis that both models perform similarly when predicting the log of mpg. We can conclude that the model created in part e with the addition of quadratic variables is better at predicting log of mpg than the model created in part d. Additionally, the degrees of freedom is 3, which signifies that the model in part e has 3 more variables than the model in part d (the quadratic terms).

Problem 6

a. Considering a linear model $Y = mX + \epsilon$ where $\epsilon \sim Norm(0, \sigma^2)$. We want to find the probability density function of $Y|X = x$, given that we know σ^2.

$$f(Y|X = x, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \times e^{-\frac{(Y-mX)^2}{2\sigma^2}}$$

b. Now we want to show that the PDF found in part a takes the form $e^{\frac{y\theta - b\theta}{\alpha} + c(y,\alpha)}$.

$$\begin{aligned} f(Y|X = x, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \times e^{-\frac{(Y-mX)^2}{2\sigma^2}} \\ &= e^{-\frac{(Y-mX)^2}{2\sigma^2} - \frac{1}{2}ln(2\pi\sigma^2)} \\ &= e^{\frac{-Y^2+2YmX-(mX)^2}{2\sigma^2} - \frac{1}{2}ln(2\pi\sigma^2)} \\ &= e^{\frac{-Y^2+2Y\theta-(\theta)^2}{2\sigma^2} - \frac{1}{2}ln(2\pi\sigma^2)} \\ &= e^{\frac{2Y\theta-\frac{(\theta)^2}{2}}{2\sigma^2} - \frac{Y^2}{2\sigma^2} - \frac{1}{2}ln(2\pi\sigma^2)} \\ &= e^{\frac{y\theta-b\theta}{\alpha} + c(y,\alpha)} \end{aligned}$$

Where $\alpha = \sigma^2$, $\theta = mX$, $b(\theta) = \theta^2/2$, and $c(Y, \alpha) = -\frac{Y^2}{2\sigma^2} - \frac{1}{2}ln(2\pi\sigma^2)$

c. Now we want to show that $E[Y|X = x, \sigma^2] = b'(\theta)$ or in other words, $\theta = (b')^{-1}(E[Y|X = x, \sigma^2])$

Starting with

$$\int_{-\infty}^{\infty} \frac{d}{d\theta} f(y|x, \sigma^2)dy = \int_{-\infty}^{\infty} \frac{y - b'(\theta)}{\alpha} f(y|x, \sigma^2)dy$$

Didn't have time to finish :(

Appendix

- Problem 5 part b
- Problem 5 part c
- Problem 5 part d
- Problem 5 part e
- Problem 5 part f