Neha Maddali

Problem 1:

a.

```
> str(insurance)
'data.frame':   1338 obs. of  7 variables:
 $ age     : chr  "19" "18" "28" "33" ...
 $ gender  : Factor w/ 2 levels "female","male": 1 2 2 2 2 1 1 1 2
 1 ...
 $ bmi     : chr  "27.9" "33.77" "33" "22.705" ...
 $ children: chr  "0" "1" "3" "0" ...
 $ smoker  : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 1 1
 ...
 $ region  : Factor w/ 4 levels "northeast","northwest",..: 4 3 3
 2 2 3 3 2 1 2 ...
 $ charges : chr  "16884.924" "1725.5523" "4449.462" "21984.47061"
 ...
```

b.

```
Call:
lm(formula = charges ~ age + bmi + gender, data = insurance)

Residuals:
   Min     1Q Median     3Q    Max
-14974  -7073  -5072   6953  47348

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -6986.82    1761.04  -3.967 7.65e-05 ***
age           243.19      22.28  10.917  < 2e-16 ***
bmi           327.54      51.37   6.377 2.49e-10 ***
gendermale   1344.46     622.66   2.159    0.031 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11370 on 1334 degrees of freedom
Multiple R-squared:  0.1203,    Adjusted R-squared:  0.1183
F-statistic: 60.78 on 3 and 1334 DF,  p-value: < 2.2e-16
```

We included factors in the data set. Female was set at the baseline for the gender variable and a dummy variable was created for male. The standard error for the intercept and gendermale are very large. The F-statistic is very large and shows that at least one of the predictors are significant to the response. But the R^2 value is really small and shows that a very small portion of the variability in the response is due to the predictors.

c. Model for just males: $\hat{Y}_i = -6986.82 + 243.19X_{i1} + 327.54X_{i2} + 1344.46$
Model for just females: $\hat{Y}_i = -6986.82 + 243.19X_{i1} + 327.54X_{i2}$

d. fit_males: $\hat{Y}_i = -8012.79 + 409.87_{i1} + 238.63X_{i2}$
fit_females: $\hat{Y}_i = -4515.22 + 241.32_{i1} + 246.92X_{i2}$
where X_{i1} is bmi and Xi_2 is age

e. The models results from part d and c are different. The models for fit_males and fit_females omit gender as a predictor. The models in part c take the gender predictor into account while the other models do not.

f. Yes, this looks like a contradiction. The significant F-test statistic could allow us to see that there is a coefficient for x that is not 0 and there is a relationship between at least one of the predictors and the response. A low R^2 means that a very small proportion of the variability in the response is due to the predictors. So the two of these combined would imply that at least one predictor affects the response. However, this does not explain the variability in the response.

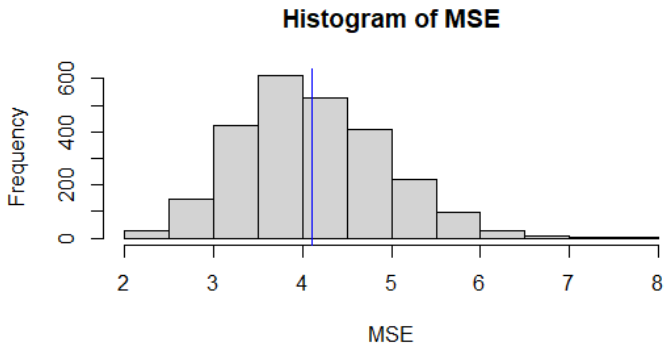
Problem 2:

a. Yes, multicollinearity is a problem for making accurate predictions. If there is multicollinearity, the standard errors for the least squares estimates could be very large. These errors could skew the predictions produced by the model and lead to less accurate predictions.

b. Correlation between x1 and x2 = 0.9404249

c. MSE_test = 4.173413

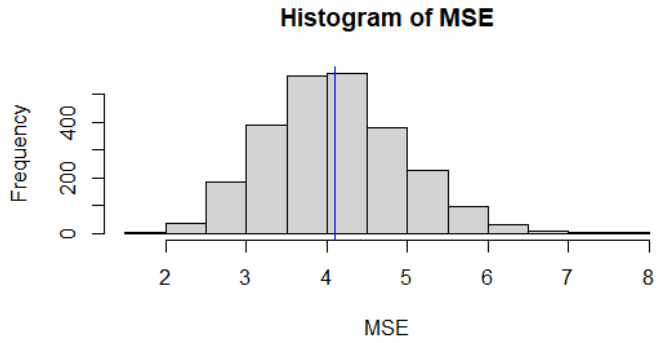
d. Mean = 4.108063



The histogram is pretty bell shaped around the average test MSE. There are some high outliers.

e. Correlation between x1 and x2 = 0.03316596

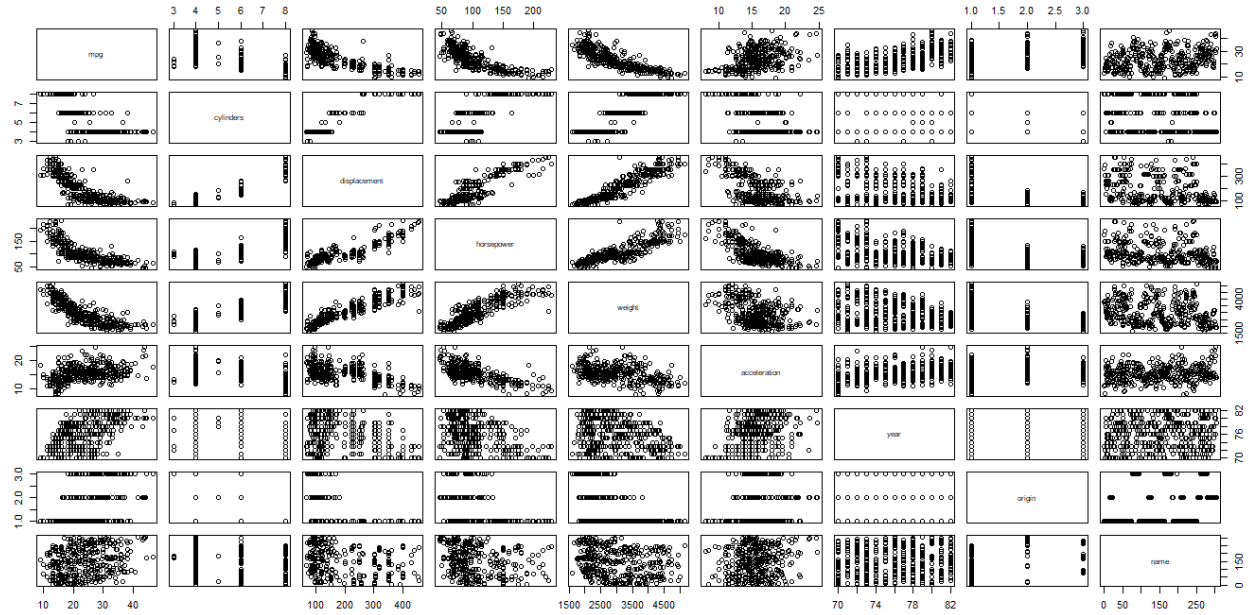
f. Mean = 4.099373



The histogram is also pretty bell shaped around the mean test MSE. This histogram and the one from part d look pretty similar with some high outliers.

g. Multicollinearity is a problem based on our simulation studies. The different models (multicollinearity vs no multicollinearity) gives us different MSEs and average MSEs after multiple iterations.

Problem 3:

a. 

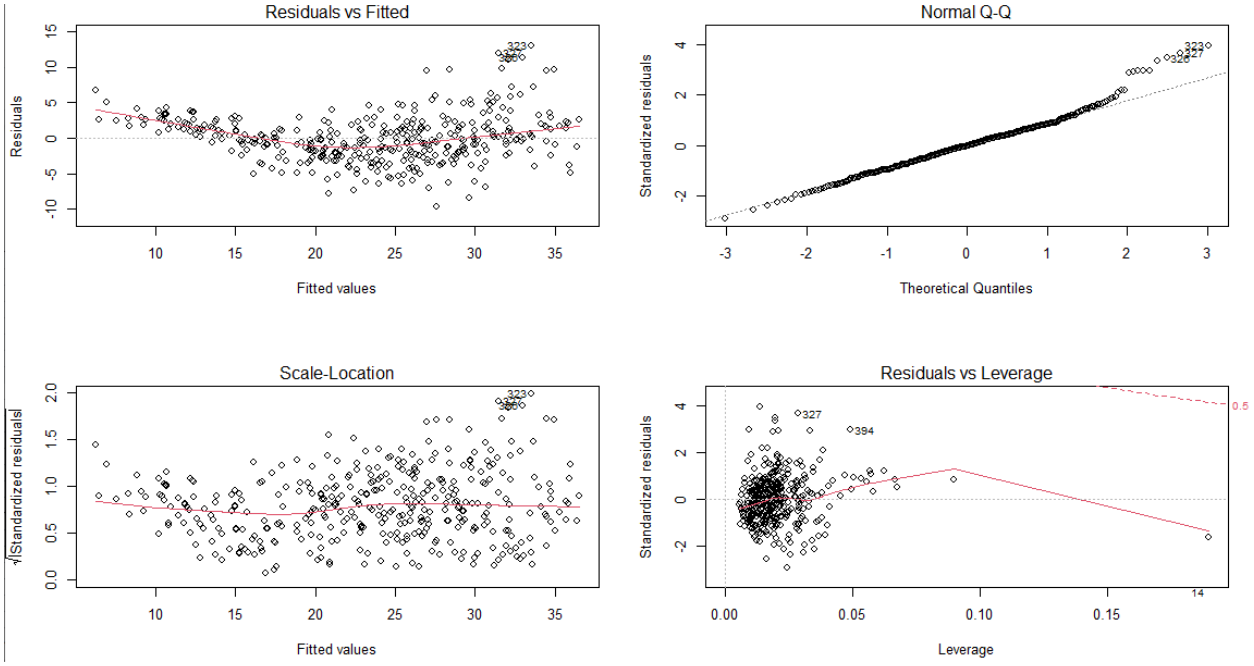
Origin, cylinders and year do not look like continuous variables.
There are linear relationships between displacement and horsepower, displacement and weight, displacement and acceleration, horsepower and weight, and horsepower and acceleration.
There are non-linear relationships with the response for variables like mpg and displacement, mpg and horsepower, and mpg and weight.

b.
```
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
```

c. H_0: $B_1 = B_2 \ldots B_8 = 0$, H_1: at least one B_j is non-zero
Test statistic (F*) = 252.4
Null distribution: $F_{8, 384}$
p-value: < 2.2e-16
Conclusion: p-value is less than 0.05 so we reject H_0. There is evidence of a relationship between the result and at least one of the predictors at a significance level 0.05.

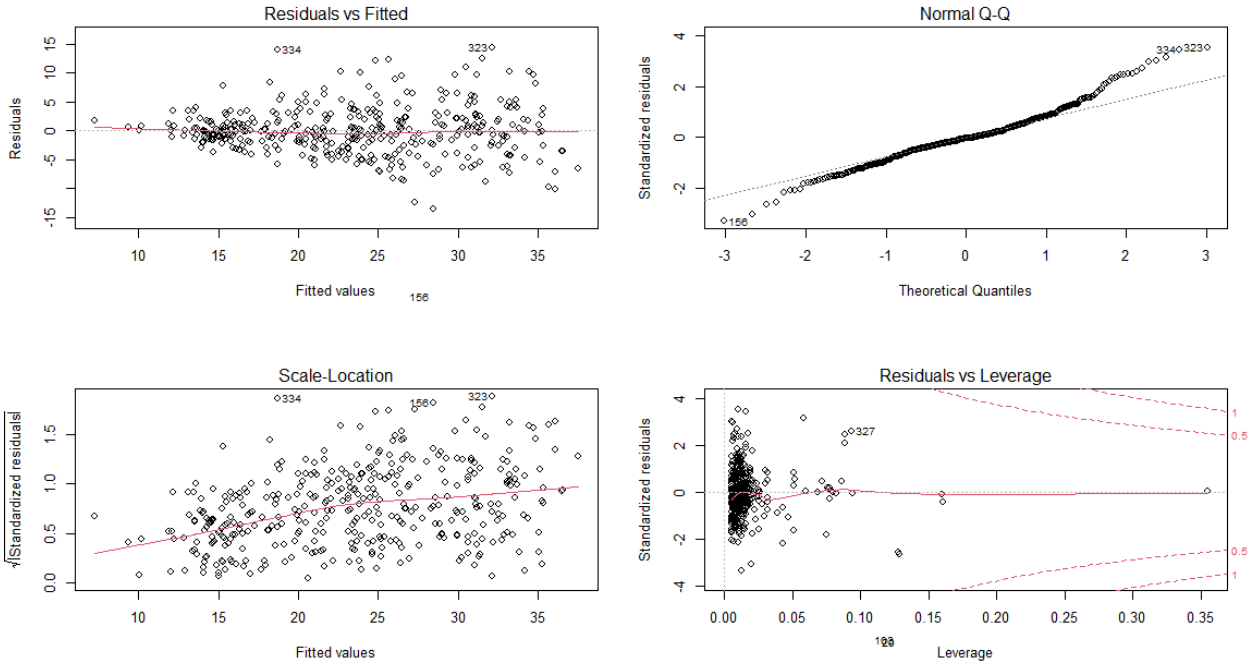
d. The coefficient for the year predictor is 0.750773 showing there is a positive relationship between year and mpg. For every one year, mpg increases by 0.750773.

e. Multicollinearity is an issue in the model. There is a pattern in the residual plot and there is no random scatter.



f. If we look at the residuals vs fitted plot we can see that there is a pattern and a funnel shape thus the constant variance assumption does not hold. The linearity assumption does not hold because there is a visible curve pattern in the residual plot.

g. I used a polynomial transformation: lm(mpg ~ poly(horsepower, 5)+acceleration, data = Auto)
My final model would be Y ~ X1 + X1^2 + X1^3 + X1^4 + X1^5 + X2
The residuals vs fitted plot shows no pattern and the values seem to be equally and randomly spaced around the horizontal axis.



Problem 4:

a. $Y_i = B_0 + B_1x_{i1} + B_2x_{i2} + \ldots + B_px_{ip} + \varepsilon_i$, i=1, …, n
b. Obtain the least squares estimates by minimizing the residual sum of squares. Choose estimators that make the sum of the squares of the difference between the predicted and actual values as small as possible.
c. I believe that the least square estimates are trustworthy. They generate a line of best fit which can be used to explain the relationship between the predictors and response. The criteria for least squares is objective and gives an unbiased estimate. Essentially a model with unbiased estimates gives an unbiased response.
d. The estimate for E(Y) for specific values of X will typically equal the true mean of Y for those specific values. We quantify any uncertainty about the estimate for E(Y) using a confidence interval which is calculated regarding the irreducible error.
e. The prediction for Y for specific values of X are those specific values of X plugged into the model. We quantify any uncertainty about the prediction for Y using a prediction interval which is calculated in regards to the reducible and irreducible error.
f. Use the test mean squared error to evaluate how good the model is at prediction. The bias-variance tradeoff says that as the model gets more flexible, bias decreases and variance increases. If the model is too flexible, there is a risk of overfitting the model
g. Statistical inference is a process where conclusions are made about a population based on a sample or subset of data. Statistical inference is useful in linear regression models because we use inference to estimate the predictors in the model.
h. 1) Non-linearity. This is a problem since it can make the predictions less accurate. Solution is to transform X.

   2) Multicollinearity. This is a problem since it increases standard errors for the estimates. Solution is to remove correlated variables/combine variables.

   3) Non-constant variance. This can affect the variances of the error terms as they might increase with the value of the response. Solution is to transform Y