Predicting Car's MPG Utilizing Multiple Linear Regression Model

Siddharth Das, Russell Chien

6/3/2022

Purpose of Linear Regression Analysis

Based on the Auto Dataset, we will create a Multiple Linear Regression Model to explain gas mileage, given other vehicle characteristics. The response variable will be in Miles Per Gallon(mpg). The regression model will include both quantitative and qualitative data.

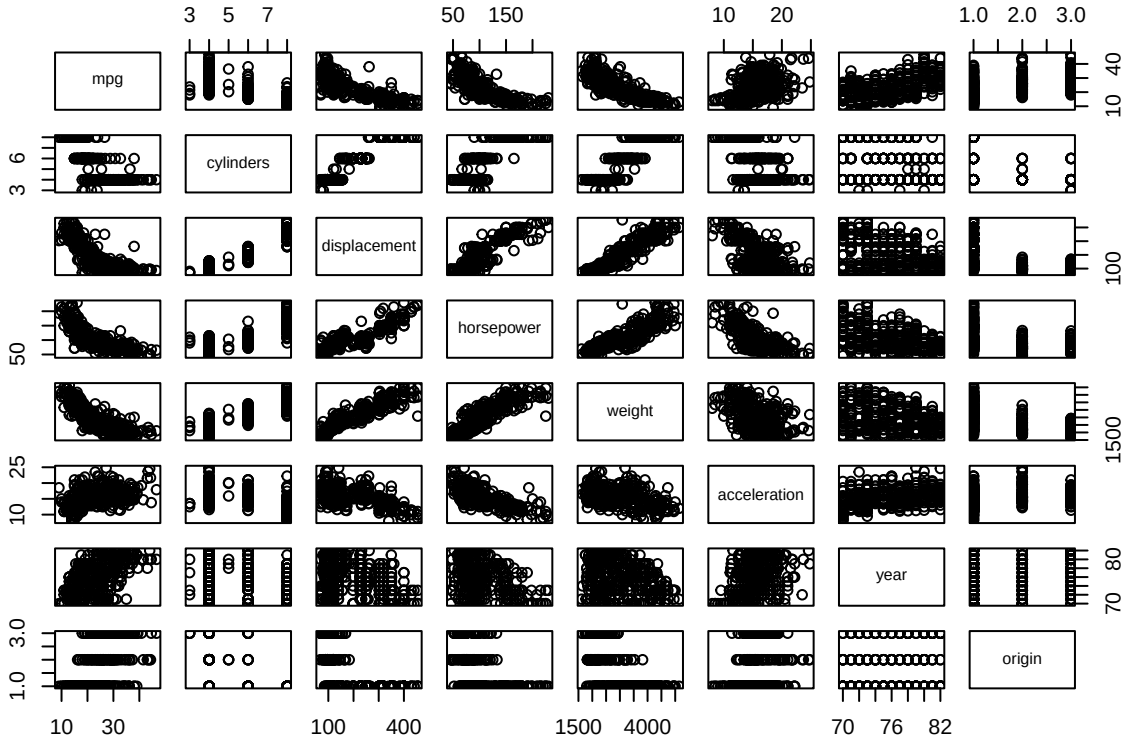
Explore the Dataset

The Auto Dataset contains 2 qualitative variables, name and origin. The other 7 variables are quantitative. There are 392 observations in this sample. Cylinders refers to the number of cylinders, between 4 and 8. Displacement refers to the engine displacement, in inches. Horsepower refers to engine horsepower. Weight refers to vehicle weight, in pounds(lbs). Acceleration is the time to accelerate from 0 to 60 mph, in seconds. Year is the model year of the car. Origin refers to the origin of the car. For the variable 'origin', 1 = American, 2 = European, and 3 = Japanese. Lastly, name refers to the vehicle name.

```
##       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                      name
##  Min.   : 8.00   Min.   :70.00   Min.   :1.000   amc matador       :  5
##  1st Qu.:13.78   1st Qu.:73.00   1st Qu.:1.000   ford pinto        :  5
##  Median :15.50   Median :76.00   Median :1.000   toyota corolla    :  5
##  Mean   :15.54   Mean   :75.98   Mean   :1.577   amc gremlin       :  4
##  3rd Qu.:17.02   3rd Qu.:79.00   3rd Qu.:2.000   amc hornet        :  4
##  Max.   :24.80   Max.   :82.00   Max.   :3.000   chevrolet chevette:  4
##                                                  (Other)           :365
```

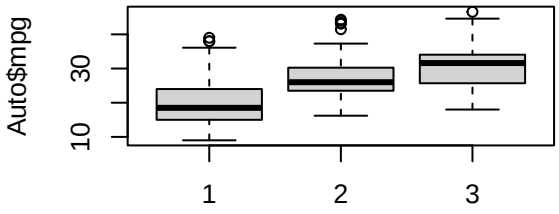
```
##                mpg cylinders displacement horsepower weight acceleration  year
## mpg           1.00     -0.78        -0.81      -0.78  -0.83         0.42  0.58
## cylinders    -0.78      1.00         0.95       0.84   0.90        -0.50 -0.35
## displacement -0.81      0.95         1.00       0.90   0.93        -0.54 -0.37
## horsepower   -0.78      0.84         0.90       1.00   0.86        -0.69 -0.42
## weight       -0.83      0.90         0.93       0.86   1.00        -0.42 -0.31
## acceleration  0.42     -0.50        -0.54      -0.69  -0.42         1.00  0.29
```

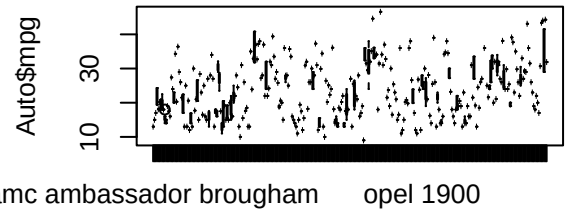
```
## year          0.58      -0.35         -0.37       -0.42  -0.31          0.29  1.00
## origin        0.57      -0.57         -0.61       -0.46  -0.59          0.21  0.18
##               origin
## mpg            0.57
## cylinders     -0.57
## displacement  -0.61
## horsepower    -0.46
## weight        -0.59
## acceleration   0.21
## year           0.18
## origin         1.00
```

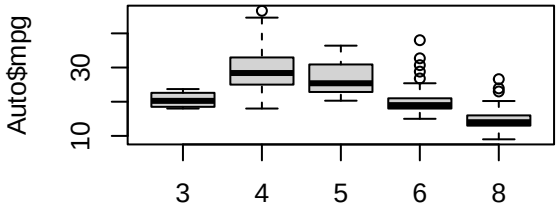


Comment on Results

The scatter plot matrix and correlation matrix indicate that each variable has a normal distribution, and there may be a few potential outliers. Year has the weakest linear correlation to the other variables. MPG has a strong negative relationship with cylinders, displacement, horsepower, and weight. Cylinders has a strong positive correlation with displacement, horsepower, and weight. Displacement has a strong positive correlation with cylinders, horsepower, and weight. Horsepower has a strong positive relationship with cylinders, displacement, and weight. Due to the multitude of correlations between cylinders, displacement, and horsepower, we suspect multicollinearity may be a problem in this regression model.







Box plots are used to correctly visualize the relationship between qualitative predictor variables and the response variable(mpg). The box plot comparing origin to mpg indicates that American cars tend to have a lower mpg, while Japanese cars tend to have a higher mpg. European cars' mpg tend to be in the middle of American and Japanese cars. The box plot comparing name to mpg is not interpretable. The box plot comparing cylinders to mpg indicates that as the number of cylinders increase, mpg initially increases, and then decreases. As the number of cylinders increases past 4, the mpg begins to decrease.

Choose 6 Predictor Variables

We chose cylinders, displacement, horsepower, weight, year, and origin as our predictor variables. Origin is the only selected qualitative predictor. The rest of the selected predictors are quantitative. We selected these 5 quantitative predictor variables because according to the scatter plot and correlation matrix, they had the strongest correlation to mpg. These relationships were either positive or negative. According to the box plots, origin had the stronger association to mpg, and the box plot was easily interpretable. The box plot with the only other qualitative variable, name, was much more difficult to interpret.

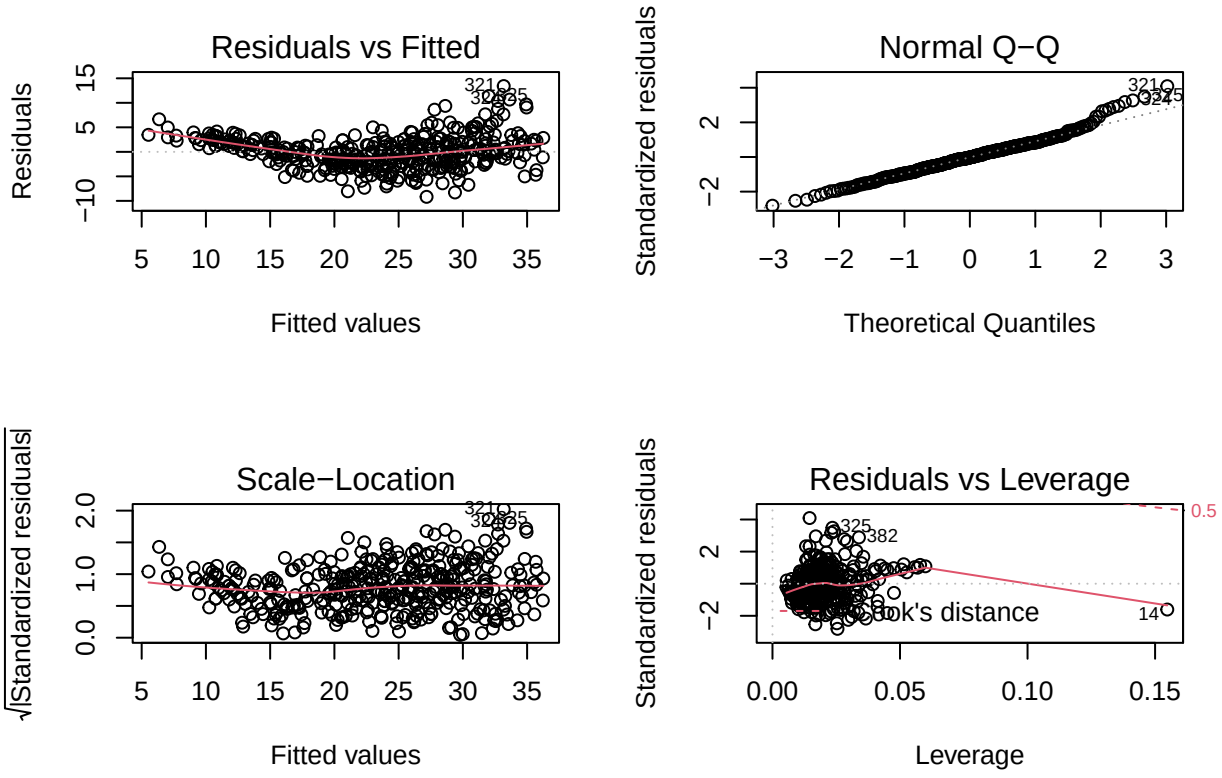
First Order Regression Model

```
##
## Call:
## lm(formula = Auto$mpg ~ Auto$cylinders + Auto$displacement +
##     Auto$horsepower + Auto$weight + Auto$year + as.factor(Auto$origin))
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.1754 -2.1139 -0.0863  1.9711 13.4207
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -1.633e+01  4.219e+00  -3.871 0.000127 ***
## Auto$cylinders           -5.028e-01  3.207e-01  -1.568 0.117742
## Auto$displacement         2.337e-02  7.613e-03   3.070 0.002292 **
## Auto$horsepower          -2.500e-02  1.078e-02  -2.320 0.020855 *
## Auto$weight              -6.460e-03  5.763e-04 -11.209  < 2e-16 ***
## Auto$year                 7.739e-01  5.161e-02  14.994  < 2e-16 ***
## as.factor(Auto$origin)2   2.635e+00  5.661e-01   4.654 4.50e-06 ***
## as.factor(Auto$origin)3   2.857e+00  5.525e-01   5.172 3.74e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.305 on 384 degrees of freedom
## Multiple R-squared:  0.8239, Adjusted R-squared:  0.8207
## F-statistic: 256.7 on 7 and 384 DF,  p-value: < 2.2e-16
```

As expected for the qualitative variable, origin, 3 levels means we have k-1 = 2 dummy variables.

Residual Plots



The Residuals vs. Fitted plot is used to check for non linearity, unequal variance, and outliers. A horizontal line without distinct patterns indicates a linear relationship, which is desired. Since the line of the plot is not horizontal and appears to curve down and then up, the relationship is not perfectly linear, which is a problem. No changes in the distance between the residuals and their mean(0) indicates constant variance, which we want. Since the distance between the residuals and their mean often fluctuates, there is an unequal variance, which is a problem. In addition, the residuals vary more for the larger fitted values, indicating heteroscedasticity. Lastly, there appear to be a few potential outliers around the larger fitted values, which is a problem. Those potential outliers include observations 321 and 325.

The Normal Q-Q plot is used to examine if the residuals are not normally distributed. A normal distribution is indicated by the residual points closely following the straight dashed line. Since the residual points closely follow the dashed line until the positive end of the Theoretical Quantiles, where it follows less closely, we can assume the residuals are normally distributed. We assume that the residuals are not perfectly normally distributed, but the distribution is close enough to normal to be acceptable.

The Scale-Location plot is used to check for non linearity, unequal variance, and outliers. Since the line is not horizontal and has a distinct curve, the relationship is not linear, which is a problem. Given that the points are not equally spread and the line is not horizontal, there is an unequal variance, which is a problem. Observations 321 and 325 appear to be potential outliers.

The Residuals vs. Leverage plot is used to identify outliers and extreme values that may influence the regression results when they are included or excluded from the analysis. This plot confirms that observations 325 and 382 are outliers, since they are > 3 or < -3 standard deviations from the mean. High leverage points are observations with a leverage statistic that greatly exceeds p/n. Since $p/n = 7/392 = 0.019$, and observation 14 has a leverage statistic > 0.15, we suspect observation 14 may have high leverage.

Overall, non linearity, unequal variance, and outliers are problems that are present in this model. In addition, the correlation matrix indicated that multicollinearity may be a problem as well.

Multicollinearity

```
## Loading required package: carData
```

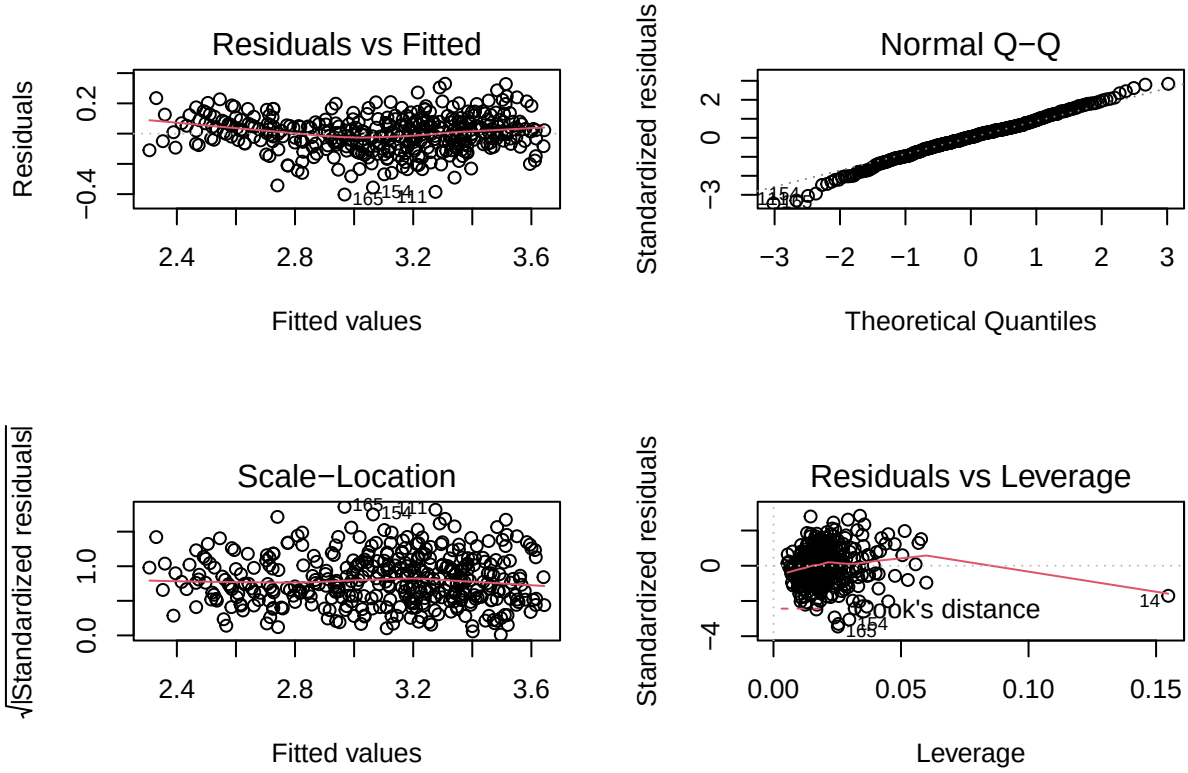
```
##                             GVIF Df GVIF^(1/(2*Df))
## Auto$cylinders         10.710418  1        3.272678
## Auto$displacement      22.715812  1        4.766111
## Auto$horsepower         6.158672  1        2.481667
## Auto$weight             8.576838  1        2.928624
## Auto$year               1.293979  1        1.137532
## as.factor(Auto$origin)  2.095801  2        1.203199
```

Variance Inflation Factor(VIF) is a useful indicator to detect multicollinearity. The rule is that if the Variance Inflation Factor is greater than 10 or 5, then multicollinearity is high. Considering that cylinders, displacement, horsepower, and weight all have a VIF close to or greater than 10, we can conclude that multicollinearity is high in this model.

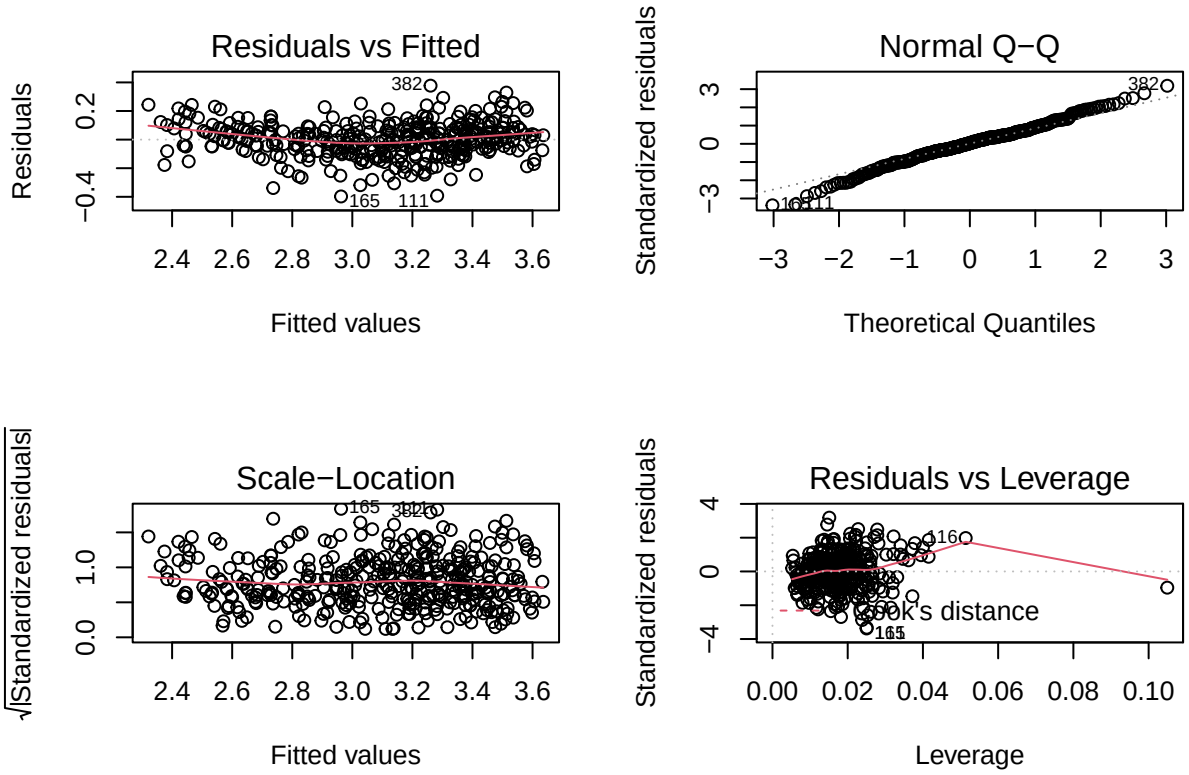
Remedial Measures

To address the lack of linearity, we can try variable transformations, or polynomial regression. To address the unequal variance, we can try response variable transformations, or weighted least squares. An alternate option is to apply a concave function to the response variable 'mpg', as they are useful with alleviating unequal variance. To understand the reason for the outliers, we can refit the model without the outliers to

gain insight into the data. We can analyze how the model changes to understand the affects of the outliers. To solve the issue of multicollinearity, we can remove the highly correlated independent variable(s), or we can linearly combine the independent variables.



To address the issues of non-linearity and unequal variance, we tried a multitude of response variable transformations. The attempted transformations included quadratic functions, and concave functions such as taking the square root or log of mpg. We found that the log transformation of mpg was the best option for a multitude of reasons. Based on the Residuals vs Fitted plot, we can see that the residuals of log(mpg) are much closer to 0 than with the original response variable, mpg. As a result, the horizontal line is more straight with less fluctuation, indicating improved linearity. In addition, it appears that points have become more equally spread around their mean(0), indicating a more equal variance. While the lack of linearity and unequal variance have improved, we cannot assume that there is perfect linearity, or a perfectly equal variance. This is because the horizontal line is still not perfectly straight, and the points are not exactly equidistant from the mean(0). To our surprise, the Normal Q-Q plot indicates that the concave function improved the normality of the distribution, as the dots follow the dashed line even closer than before.




```
##                              GVIF Df GVIF^(1/(2*Df))
## Auto$cylinders           6.108653  1        2.471569
## Auto$horsepower          4.892762  1        2.211959
## Auto$weight              7.144142  1        2.672853
## Auto$year                1.280339  1        1.131521
## as.factor(Auto$origin)   1.789885  2        1.156661
```

To address the issue of multicollinearity, we dropped displacement, which had the largest Variance Inflation Factor. Based on the residual plots, the most significant difference is found in the Normal Q-Q plot. The Normal Q-Q plot indicates that the distribution has become marginally less normal. While we prefer the most normal distribution, we accept this deficit for 2 reasons. First, the distribution is still close enough to normal to be acceptable. More importantly, by removing displacement from the model, the Variance Inflation Factor for all the predictor variables have dropped, with the highest VIF being less than 7.2. As a result, considering the totality of the implemented remedies, we have been able to improve the lack of linearity, unequal variance, normality of the distribution, and multicollinearity.

Brute Force Algorithms

```
## Subset selection object
## Call: regsubsets.formula(y_log ~ Auto$cylinders + Auto$horsepower +
##     Auto$weight + Auto$year + as.factor(Auto$origin), data = Auto,
##     nvmax = p - 1)
## 6 Variables  (and intercept)
##                          Forced in Forced out
```

```
## Auto$cylinders              FALSE      FALSE
## Auto$horsepower             FALSE      FALSE
## Auto$weight                 FALSE      FALSE
## Auto$year                   FALSE      FALSE
## as.factor(Auto$origin)2     FALSE      FALSE
## as.factor(Auto$origin)3     FALSE      FALSE
## 1 subsets of each size up to 5
## Selection Algorithm: exhaustive
##          Auto$cylinders Auto$horsepower Auto$weight Auto$year
## 1  ( 1 ) " "            " "             "*"         " "
## 2  ( 1 ) " "            " "             "*"         "*"
## 3  ( 1 ) " "            " "             "*"         "*"
## 4  ( 1 ) " "            " "             "*"         "*"
## 5  ( 1 ) " "            "*"             "*"         "*"
##          as.factor(Auto$origin)2 as.factor(Auto$origin)3
## 1  ( 1 ) " "                     " "
## 2  ( 1 ) " "                     " "
## 3  ( 1 ) "*"                     " "
## 4  ( 1 ) "*"                     "*"
## 5  ( 1 ) "*"                     "*"

##   Nvar     R2adj         CP       BIC
## 1    1 0.7661793 352.680239 -558.7159
## 2    2 0.8700418  24.616454 -783.9891
## 3    3 0.8721104  19.041709 -785.3167
## 4    4 0.8748370  11.432433 -788.8047
## 5    5 0.8770742   5.401466 -790.9180

##   R2adj CP BIC
## 1     5  5   5
```

We use the brute force algorithm to select the best subset of predictor variables from the previous model. The previous model we are referring to uses log(mpg) as the response variable, and drops displacement from the predictor variables. According to R2adj, Mallow's CP, and BIC, each criterion results in the same subset of 5 variables. The optimal subset of predictor variables are: weight, year, the two levels of origin, and horsepower. Cylinders appears to be the least significant variable. Thus, according to the brute force algorithm, it should be left out from the best subset. If the various criterion resulted in unique subsets, I would select the best subset according to BIC. BIC should always be preferred since it outperforms the other criterion in choosing the optimal subset. R2adj selects the subset with the highest R2adj. Differently, Mallow's CP and BIC select the subset according to the lowest corresponding value.

Stepwise Selection

Backward stepwise selection:

```
##                         Step Df     Deviance Resid. Df Resid. Dev      AIC
## 1                           NA           NA        85   438.7311 658.1489
## 2 - as.factor(Auto$origin)  0 9.094947e-13        85   438.7311 658.1489
## 3      - Auto$displacement  1 1.956203e-01        86   438.9267 656.3237
```

Forward stepwise selection:

```
##                   Step   Df    Deviance Resid. Df Resid. Dev       AIC
## 1                     NA           NA       391 23818.9935 1611.9349
## 2         + Auto$name -300 23039.24380        91   779.7497  871.5827
## 3         + Auto$year   -1   143.30415        90   636.4455  793.9777
## 4       + Auto$weight   -1   154.31560        89   482.1299  687.1251
## 5 + Auto$acceleration   -1    14.89619        88   467.2337  676.8226
## 6   + Auto$horsepower   -1    16.79556        87   450.4382  664.4719
## 7    + Auto$cylinders   -1    11.51146        86   438.9267  656.3237
```

Bi-direction stepwise selection:

```
##                   Step   Df    Deviance Resid. Df Resid. Dev       AIC
## 1                     NA           NA       391 23818.9935 1611.9349
## 2         + Auto$name -300 23039.24380        91   779.7497  871.5827
## 3         + Auto$year   -1   143.30415        90   636.4455  793.9777
## 4       + Auto$weight   -1   154.31560        89   482.1299  687.1251
## 5 + Auto$acceleration   -1    14.89619        88   467.2337  676.8226
## 6   + Auto$horsepower   -1    16.79556        87   450.4382  664.4719
## 7    + Auto$cylinders   -1    11.51146        86   438.9267  656.3237
```

Considering all the variables in the the original dataset, we can perform backward, forward, or bi-direction stepwise selections. Backward stepwise selection begins with the full model, and drops one variable at a time until the AIC is minimized. Forward stepwise selection begins with only the y intercept(empty model), and adds predictor variables until the AIC is minimized. Lastly, bi-direction stepwise selection both adds and drops variables until the AIC is minimized. However, we must keep in mind that stepwise selections often identify an incorrect model because they find the local optimal solution, rather than the global optimal solution. We chose to select the subset of variables through forward stepwise selection.

Compare Best Subsets of Predictor Variables

Interestingly, the comparison of the brute force algorithm and the forward stepwise selection yields significantly unique results. According to R2adj, Mallow's CP, and BIC, the brute force algorithm consistently selects the same 5 predictor variables: weight, year, horsepower, and the two levels of origin. However, the forward stepwise selection, which yields the same results as the other stepwise selections, selects 6 predictor variables. In addition, instead of selecting the origin as a qualitative variable, it selects name instead. This is surprising, based on the box plot interpretations and the number of levels for the variable, name. Lastly, the forward stepwise selection also adds acceleration and cylinders to the model. In conclusion, the only variables in both subset selections are weight, year, and horsepower. The brute force algorithm accounts for origin, while the stepwise selection accounts for name, acceleration, and cylinders. We would like to learn more about the reasoning for the differences in selection.

Test Whether Variables are Statistically Significant

```
## Analysis of Variance Table
##
## Model 1: y_log ~ Auto$horsepower + Auto$weight + Auto$year + as.factor(Auto$origin)
## Model 2: y_log ~ Auto$cylinders + Auto$displacement + Auto$horsepower +
##     Auto$weight + Auto$acceleration + Auto$year + as.factor(Auto$origin) +
##     Auto$name
##   Res.Df    RSS  Df Sum of Sq      F    Pr(>F)
```

```
## 1    386 5.4864
## 2     85 0.7343 301    4.7521 1.8276 0.0006088 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the comparison between the brute force algorithm model and the full model, the null hypothesis is H_o : $B_{cylinders} = B_{displacement} = B_{acceleration} = B_{name} = 0$. The alternative hypothesis is H_a : at least one of the β_i in the null hypothesis is not equal to 0.

$F* = (\frac{SSR_{brute} - SSR_{full}}{Res.Df(SSR_{brute})} - \frac{Res.Df(SSR_{full})}{SSR_{full}/Res.Df(SSR_{full})}) = 1.8276$

The test statistic follows the F distribution. The implied level of significance is alpha = 0.05.

Since the p-value $< \alpha$, we reject the null hypothesis (H_0). We conclude that the null hypothesis is false. At least one of the β_i in the null hypothesis is not equal to 0. Thus, we cannot drop all the variables from the model. We will have to drop each variable one at a time to figure out exactly which variables can be dropped, and which cannot.

Conclusions on Brute Force Algorithm

```
##
## Call:
## lm(formula = y_log ~ Auto$horsepower + Auto$weight + Auto$year +
##     as.factor(Auto$origin), data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.40778 -0.06845  0.00554  0.06835  0.37346
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                1.625e+00  1.497e-01  10.854  < 2e-16 ***
## Auto$horsepower           -9.463e-04  3.337e-04  -2.836 0.004807 **
## Auto$weight               -2.503e-04  1.574e-05 -15.901  < 2e-16 ***
## Auto$year                  3.017e-02  1.842e-03  16.384  < 2e-16 ***
## as.factor(Auto$origin)2    6.789e-02  1.852e-02   3.665 0.000282 ***
## as.factor(Auto$origin)3    6.508e-02  1.873e-02   3.476 0.000567 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1192 on 386 degrees of freedom
## Multiple R-squared:  0.8786, Adjusted R-squared:  0.8771
## F-statistic:   559 on 5 and 386 DF,  p-value: < 2.2e-16
```

The Multiple Linear Regression Model derived from the brute force algorithm appears to be a fairly good fit for the model. This conclusion is based on the p-values of each individual variable, the p-value of the whole set of predictors, and the R-squared. The definition of $b_{horsepower}$ is that for a one unit increase in horsepower, the average decrease in log(mpg) is -9.463e-04, given the other predictor variables are held constant. The definition of b_{weight} is that for a one unit increase in weight, the average decrease in log(mpg) is -2.503e-04, given the other predictor variables are held constant. The definition of b_{year} is that for a one unit increase in year, the average increase in log(mpg) is 3.017e-02, given the other predictor variables are held constant. Differently, the definition of $b_{as.factor(Auto\$origin)2}$ is the average difference in log(mpg) between cars originating from America and Europe. Lastly, the definition of $b_{as.factor(Auto\$origin)3}$ is the

average difference in log(mpg) between cars originating from America and Japan. Origin and year are the only two predictor variables with a positive relationship to log(mpg). The other predictor variables in this Multiple Linear Regression model have a negative relationship to log(mpg). Since the p-value of each individual variable is small, we know that each variable is statistically significant to the model. Since the overall p-value is small, we know that a relationship exists between the predictor variables and log(mpg). Lastly, since R-squared = 0.88, we know that 88% of the variation in Y, AKA log(mpg), is explained by the set of predictor variables. The significant predictor variables, ordered by importance, are weight, year, origin, and horsepower.

Create a New Dataset

```
newdata = data.frame(
  cylinders = c(4, 8, 8, 6, 8),
  displacement = c(100, 120, 130, 150, 180),
  horsepower = c(67, 86, 98, 143, 129),
  weight = c(2910, 2290, 2051, 2700, 2065),
  year = c(72, 73, 81, 79, 82),
  origin = factor(c(1, 2, 1, 3, 2))
)
```

Point Predictions

```
predict(mod_log2, newdata = newdata)
```

Using the model from part 9, let us predict the mpg of a car with the vehicle characteristics of 4 cylinders, 100 inches of engine displacement, 67 horsepower, a weight of 2910 pounds, a model year of 1972, and originating from America. Given these values for the predictor variables, we expect it to run $e^{2.729744} = 15.33$ miles per gallon of gas.