Machine Learning: Learning from Data Using Linear Regression

Victor Wright

July 2019

Introduction

The main purpose of this paper is to inform readers of all backgrounds how linear regression can be used to model the relationship between a quantitative dependent variable and quantitative independent variable(s). What we mean by quantitative is that the variables contain real numbers. To be more precise, they are *column vectors* (a column in a data set) that contain n observations. For example, the variable $\vec{X} = \left[2.300, 4.000, 3.415, ..., 6.001\right]^T$ contains n observations that are real numbers. This variable is quantitative and has *continuous* elements. Another purpose of this paper is to present how data in quantitative variables can be transformed such that learning from the data using linear regression modeling is appropriate. The final purpose of this paper is to discuss various linear regression methods, model assessment, results of simulation, limitations of linear regression modeling, and present situations when modeling assumptions do not hold.

Explanation of the Data Source, Variables, and Modeling Objectives

The data set titled *Auto* we use to create a new data set can found at http://faculty.marshall.usc.edu/gareth-james/ISL/data.html as a CSV file. The original data set, found on the website, contains information on cars collected by David Donoho and Ernesto Ramos and used in the 1983 American Statistical Association Exposition. The Auto data set contains 392 observations (5 were deleted due to missingness) on cars manufactured between 1970 and 1982. We use the information contained in this data set to create and analyze a new data set in this paper called *ML.Auto* which is a CSV file available to download on the Wright Analytics website along with a data dictionary titled *Data.Dictionary.ML.Auto*.

As discussed in *An Overview of Linear Regression Modeling*, we consider linear regression models to be flexible in the sense that they have been used

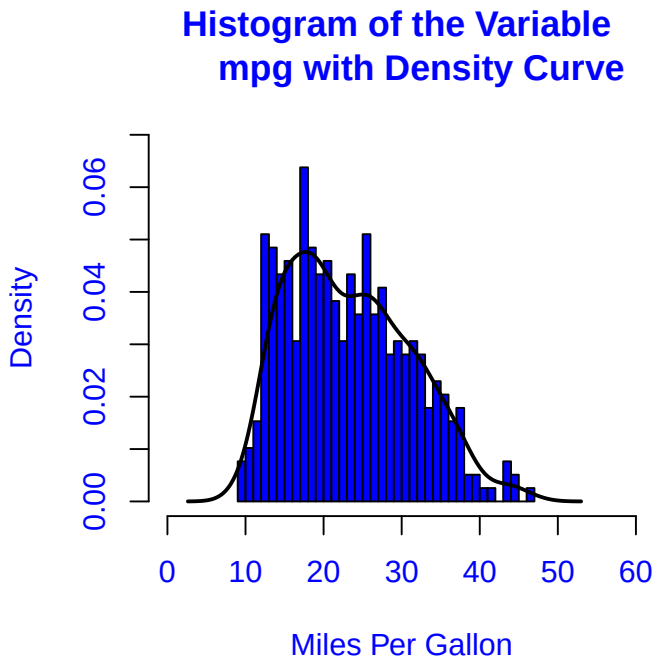
to describe phenomena in various fields. Flexibility is not declared about the functional form of the model itself. In this paper, we choose to analyze the information about these cars using linear regression. Our goal is to use linear regression models to explain how certain characteristics of these types of cars affect their miles per gallon rating. For example, the impact of a vehicle's horsepower, number of cylinders in the engine, or origin of the car on the vehicle's mpg rating. Finally, we seek out the linear regression models that adequately explain as much of a vehicle's miles per gallon as a function of a vehicle's characteristic(s). Not the one and only "best" linear regression model.

Simple Linear Regression

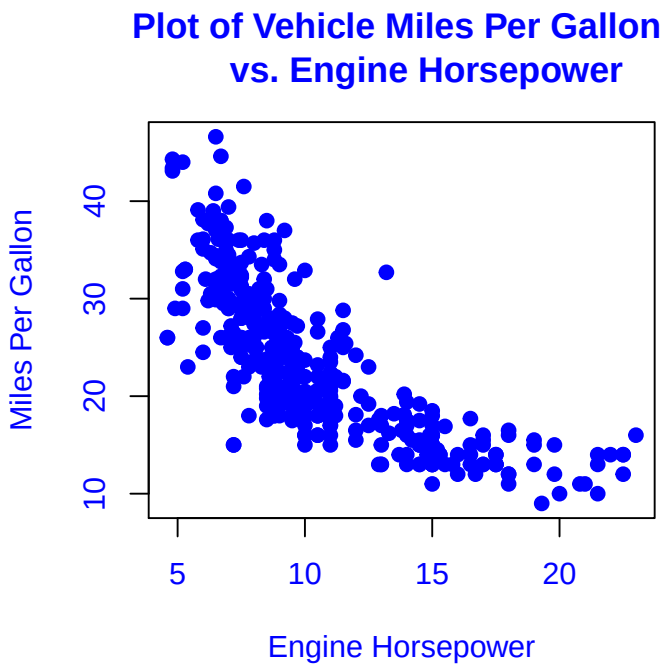
Intuition tells us that as the horsepower of a vehicle's engine increases its miles per gallon rating should decrease. One question we seek to answer is, "how does the horsepower of a vehicle's engine impact the miles per gallon that vehicle gets in general?" In other words, we want to quantify the relationship between the miles per gallon a vehicle gets given the horsepower of a vehicles engine. We may be able to quantify the relationship between these two variables if the relationship between the variables is linear using a linear regression model. The *conditional mean function*, or *simple linear regression* model we assume is

$$E[Y|X] = \beta_0 + \beta_1 X + \epsilon \tag{1}$$

for a car that has an engine with X horsepower. This equation says that for the entire population of cars that have an engine with X horsepower (manufactured in America, Europe, or Japan) the expected miles per gallon that this type of car will get is $\beta_0 + \beta_1 X + \epsilon$ miles per gallon. A histogram of the variable appears below.



We can see that the data is somewhat skewed to the right (the right-hand tail of the density curve is longer than the left-hand tail) of the mean of the variable *mpg*. So the distribution of mpg does not appear to be exactly normal. The scatterplot of mpg vs. horsepower can be found below.



It is easy to see that the linearity assumption between miles per gallon and engine horsepower is violated. However, we train the learning method on all of the data to see if other linear regression modeling assumptions hold. We do this to illustrate issues in regression diagnostic plots. We use the statistical computing R program to estimate the model parameters and include the R regression output of mpg regressed on horsepower of the engine below.

```
Call:
lm(formula = mpg ~ horsepower.10s)

Residuals:
     Min       1Q   Median       3Q      Max
-13.5710  -3.2592  -0.3435   2.7630  16.9240

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    39.93586    0.71750   55.66   <2e-16 ***
horsepower.10s -1.57845    0.06446  -24.49   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.906 on 390 degrees of freedom
```

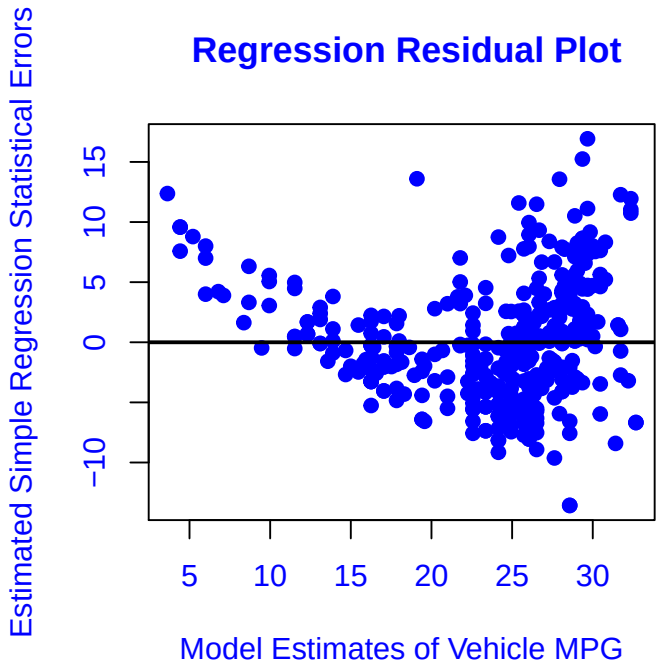
```
Multiple R-squared:  0.6059,        Adjusted R-squared:  0.6049
F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

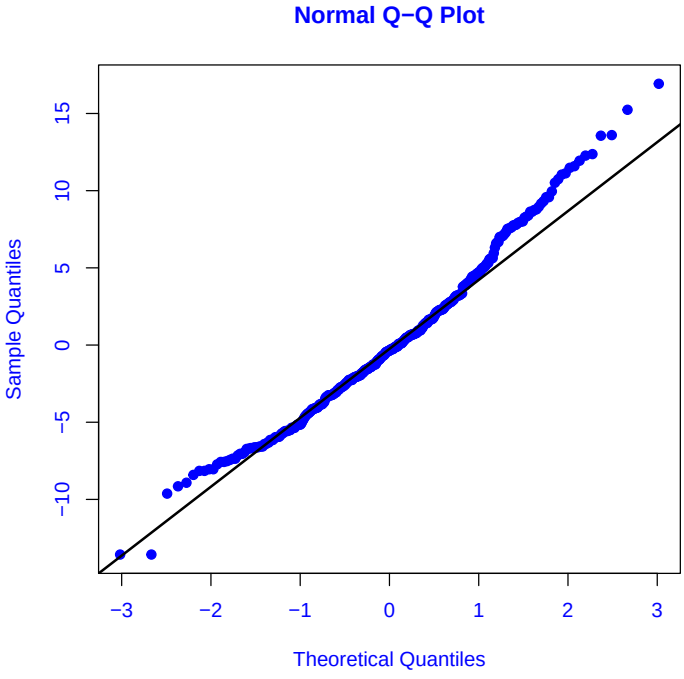
R returns the estimated model

$$E[Y|X] = 39.94 - 1.578X \tag{2}$$

. We can see that a car's miles per gallon rating may depend linearly on the horsepower of the vehicles engine. This is because of the small p-values obtained from the *t-test for slope* as well as the small p-value from the *F-test for model significance.*

Next, we interpret the *dependency* between the miles per gallon of a car that has an engine with X horsepower. The sign of the simple linear regression model slope is negative. Therefore, this means that we have evidence of an *inverse relationship* between the miles per gallon of the vehicle and the horsepower of its engine. This is because we have strong evidence that the slope of the *true regression line* is significantly different from zero. So for every unit increase in X (an increase of 10 units of horsepower), we can expect that the miles per gallon of the vehicle will decrease by 1.580 mpg. Or, if there is a unit decrease in X (a decrease of 10 units of horsepower), the mpg of the vehicle will increase by about 1.580 mpg. Note, we also have strong evidence that the intercept of the true regression line is significantly different from zero. However, in this case, the intercept doesn't have a practical interpretation because it is the mpg of a car that has a zero horsepower engine. Does this inference about the slope make sense? We need a way to see if the relationship between mpg and horsepower is truly linear and if the residuals have a constant variance. To do this we produce a *residual plot* and *normal Q-Q plot* which are often referred to as *regression model diagnostic plots* below.


Regression Residual Plot
Estimated Simple Regression Statistical Errors
−10
0
5
10
15
5
10
15
20
25
30
Model Estimates of Vehicle MPG


Normal Q−Q Plot
Sample Quantiles
−10
−5
0
5
10
15
−3
−2
−1
0
1
2
3
Theoretical Quantiles

The residual plot indicates that the residuals do not display constant variance. There is a strong functional form in this estimated regression model's residuals which suggests a need to transform data in some of the variables. Further, there is a strong departure from normality in the errors which also suggests the need to transform some of the data. Since this is the case, we should be careful with any inferential or predictive conclusions made using this model.

Transformations of a Single Variable

We do know that some of the linear regression modeling assumptions were not met when we analyzed diagnostic plots for the model in the previous section. We saw in the previous section that there is a strong functional form in the data when we analyzed the scatterplot of vehicle miles per gallon as a function of engine horsepower. Does this mean that the relationship between mpg and horsepower is nonlinear? Does it mean that modeling the dependency of mpg on horsepower with a simple linear regression model is incorrect? Not necessarily. We say not necessarily due to the idea of *sampling uncertainty*. In other words, if we were to draw another random sample from the population of these types of cars, construct the same variables which contain measurements of horsepower and mpg ratings, and draw a scatter plot of mpg as a function of horsepower, the data could display linearity between the *untransformed* variables. In other words, we are not sure if the sample used to train the simple linear regression model was representative of all vehicles manufactured in America, Europe, and Japan (the population of these types of cars).

It is important to note that we are not out to seek the one and only "best" model, or *f*, that can be used to explain mpg for these types of cars. We are seeking a robust model, explains a significant proportion of mpg for these types of vehicles, and is not overly complex and easy to interpret. Since we have somewhat encountered a barrier in the previous section, we use some alternative modeling approaches in this section in hopes to obtain such a model. Since we noted a functional form in the scatterplot of mpg as a function of horsepower we first consider modeling mpg with a *polynomial regression model* which is a special case of multiple linear regression. The form of such a model of degree k is

$$E[Y|X, X^2, ..., X^k] = \beta_0 + \beta_1 X + \beta_2 X^2 + ... + \beta_k X^k + \epsilon \tag{3}$$

. We will not just pick the degree of the polynomial out of a hat and estimate the polynomial regression model. We will compare results of $LOOCV$ (Leave-One-Out Cross-Validation) and *K-Fold Cross-Validation* to estimate *test mean square errors*. Then estimate the polynomial using the data whose degree minimizes the estimated test mean square error.

```
> set.seed(1)
> # LOOCV #
```

```
> library(boot)
> LOOCV.Error <- rep(x = 0, times = 10)
> for ( i in 1:10){
+   LOOCV.poly <- glm(mpg ~ poly(horsepower.10s, degree = i), data = ML.Auto )
+ LOOCV.Error[i] <- cv.glm(data = ML.Auto, glmfit = LOOCV.poly)$delta[1]
+
+ }
> LOOCV.Error

 [1] 24.23151 19.24821 19.33498 19.42443 19.03321 18.97864 18.83305 18.96115
 [9] 19.06863 19.49093

> # 10-fold cross validation #
> set.seed(18)
> Five.fold.error <- rep(x = 0, times = 10)
> for (j in 1:10){
+   K.fold.poly <- glm(mpg ~ poly(horsepower.10s, degree = j), data = ML.Auto)
+   Five.fold.error[j] <- cv.glm(data = ML.Auto, glmfit = K.fold.poly, K = 10)$delta[1]
+ }
> Five.fold.error

 [1] 24.17665 19.22218 19.42568 19.47748 19.03218 19.11929 18.70025 18.87910
 [9] 19.38766 21.15394
```

Looking at the LOOCV results we choose the fifth degree polynomial which sacrifices a little bit of accuracy but preserves simplicity as compared to the a seventh degree polynomial option which has the lowest estimated test MSE. Further, the estimated test MSE for this model is 19.03321 mpg^2 or 4.4mpg on average. The 10-fold cross-validation results suggest using a seventh degree polynomial. Its test MSE is 18.70960 mpg^2 or 4.32 mpg on average. Although, the fifth degree polynomial test MSE is not much larger. It is 18.86876 mpg^2 or 4.34 mpg on average. Since the discrepancy is small (using both validation methods) we choose to fit a fifth degree polynomial to reduce model complexity.

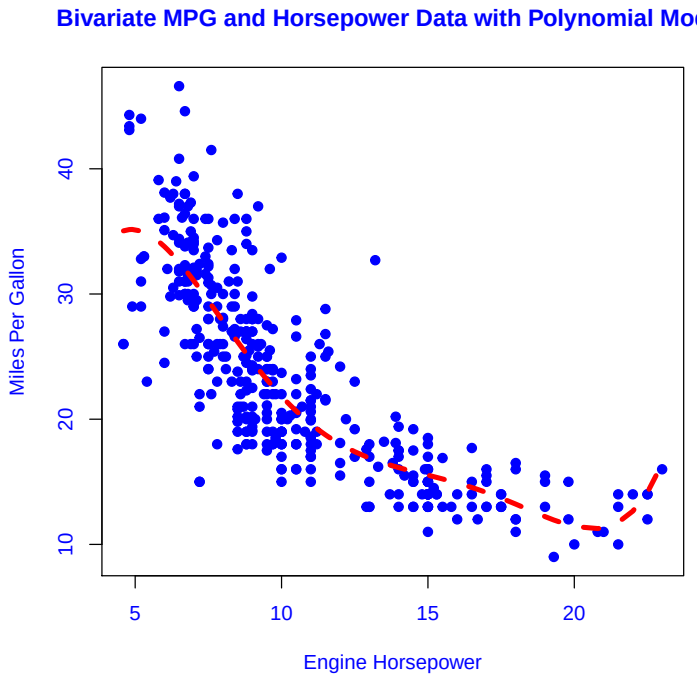
```
Call:
lm(formula = mpg ~ poly(horsepower.10s, degree = 5), data = ML.Auto)

Residuals:
     Min       1Q   Median       3Q      Max
-15.4326  -2.5285  -0.2925   2.1750  15.9730

Coefficients:
                                   Estimate Std. Error t value Pr(>|t|)
(Intercept)                         23.4459     0.2185 107.308  < 2e-16 ***
poly(horsepower.10s, degree = 5)1 -120.1377     4.3259 -27.772  < 2e-16 ***
poly(horsepower.10s, degree = 5)2   44.0895     4.3259  10.192  < 2e-16 ***
poly(horsepower.10s, degree = 5)3   -3.9488     4.3259  -0.913  0.36190
```

```
poly(horsepower.10s, degree = 5)4   -5.1878     4.3259  -1.199  0.23117
poly(horsepower.10s, degree = 5)5   13.2722     4.3259   3.068  0.00231 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.326 on 386 degrees of freedom
Multiple R-squared:  0.6967,        Adjusted R-squared:  0.6928
F-statistic: 177.4 on 5 and 386 DF,  p-value: < 2.2e-16
```



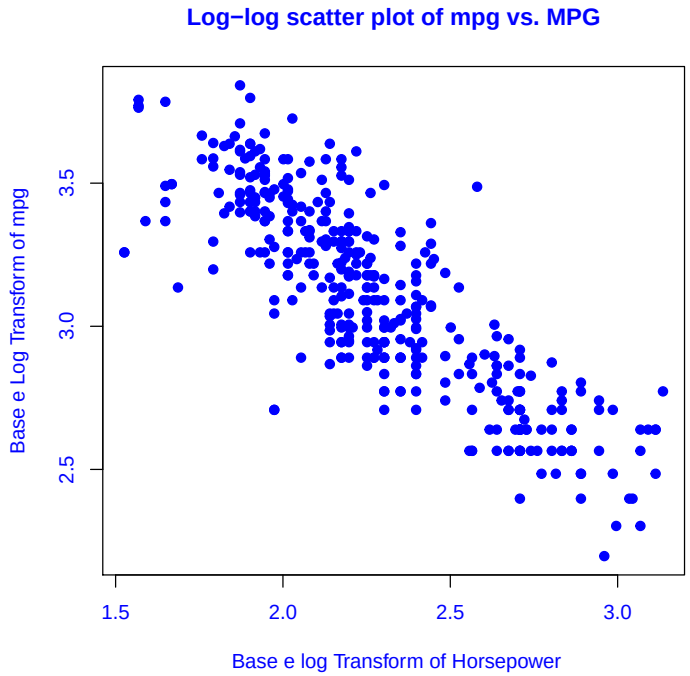
Even though some of the powers appear to be insignificant, we keep them in the model to obey a concept called *model hierarchy* that applies to fitted multiple linear regression models. The estimated model is

$$E[Y|X, X^2, ..., X^5] = 23.4 - 120.1X + 44.1X^2 - 3.9X^3 - 5.2X^4 + 13.3X^5 \quad (4)$$

. If one plots and analyzes a normal QQ-plot and residual plot you will see that the errors do not have a constant variance and they present a strong departure from normality. Our search continues!

Another approach is to transform the variables in a data set to obtain linearity between a dependent variable and one or more independent variables rather than trying to capture curvature with a polynomial. More formally this is called *transforming for linearity*. Transformation methods do have formal names and

criteria that have been constructed by statisticians. To discuss them is beyond the scope of this paper. However, we will present the idea of *log transforms* on variables which are common transforms for linearity. The scatter plot of the *log-log transform (log base e)* is shown below.



We see that the relationship is linear. Therefore, it is acceptable to conclude that the population line has the form

$$\ln(E[Y|X]) = \beta_0 + \beta_1 \ln(X) + \epsilon \tag{5}$$

in the transformed scale of the varibles. Where β_1 in (1) is not the same as β_1 in (5) due to the transform. In other words, the transformed data produces different estimates for both β_0 and β_1.

```
Call:
lm(formula = log(mpg) ~ log(horsepower.10s))

Residuals:
     Min       1Q   Median       3Q      Max
-0.65230 -0.12176  0.00788  0.11631  0.63730

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         5.02222    0.06102   82.30   <2e-16 ***
```

```
log(horsepower.10s) -0.84185    0.02641  -31.88   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1793 on 390 degrees of freedom
Multiple R-squared:  0.7227,        Adjusted R-squared:  0.722
F-statistic:  1016 on 1 and 390 DF,  p-value: < 2.2e-16
```

The model diagnostic plots above show that the errors exhibit constant variance, are independent, and follow a normal distribution. Is this the one and only "best" model? No. Probably not. Also, when model error diagnostic plots show patterns such as these, any statistical inferences made about the impact of the independent variable(s) on a dependent variable are valid only if there is a significant relationship between the dependent variable and *significant independent variable(s)*. In other words, such models are *robust* and predictions and inference made using these models are often accurate. However, if any of the independent variables are not significant, it does not make much sense to interpret their slopes. This is because we cannot rule out the fact that the population slope may be zero. We next analyze and interpret the log-log regression model.

The estimated log-log regression model is

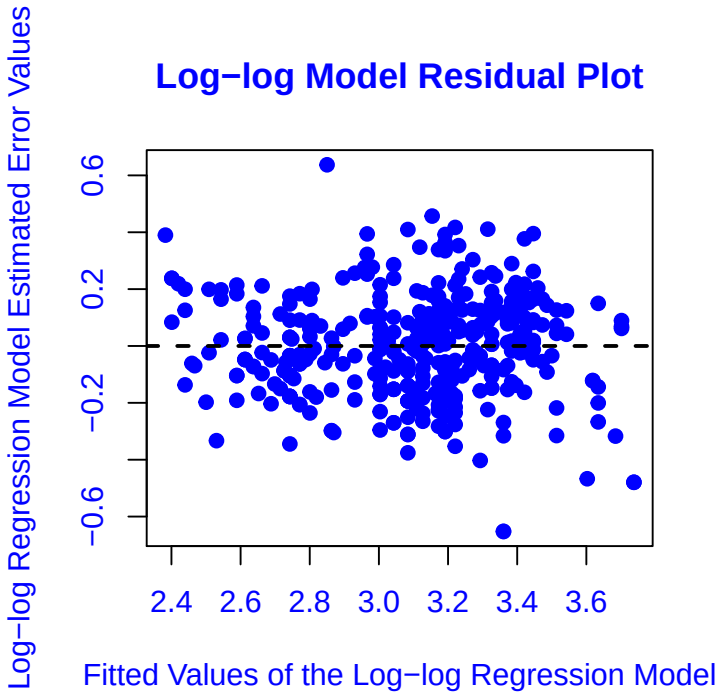
$$\ln(E[Y|X]) = 5.02 - 0.84\ln(X) \tag{6}$$

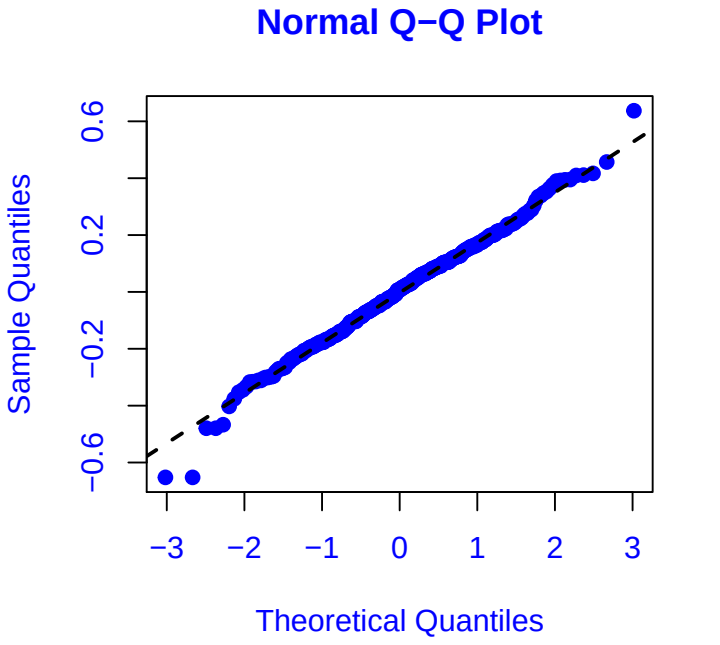
The LOOCV estimated test MSE is found to be 1.03 mpg as well as the ten-fold cross-validation estimated test MSE. So the average error that this model will make when predicting the gas mileage of a car whose mpg rating was not in the training set is this value. The typical error the model makes on the training data is $\exp(0.1793) = 1.20$ mpg. The *R-Squared* of the model is 72%. So overall, this model is fairly effective. The intercept has no practical interpretation. This is because when $X = 0$ we have $\ln(E[Y \mid X]) = \infty$. Before we interpret the estimated slope of the log-log regression model, we first *bootstrap* the coefficient estimate and its standard error. This will help any doubts about the value of $\hat{\beta}_1$ obtained from the ordinary least square estimate of β_1 in the log-log model.

Below we show R code how we boot strap $\hat{\beta}_1$. We have

```
> # Define a function that estimates the log-log slope estimate #
> log.frame <- data.frame(mpg = log(mpg), horsepower.10s = log(horsepower.10s))
> beta1.hat <- function(data, index){
+   X <- data$horsepower.10s[index]
+   Y <- data$mpg[index]
+   return ( sum( (X - mean(X)) * (Y - mean(Y)) )/ sum((X - mean(X)) ^2) )
+ }
> beta1.hat(data = log.frame, index = 1:392)

[1] -0.841847
```


Log−log Model Residual Plot
Log−log Regression Model Estimated Error Values
0.6
0.2
−0.2
−0.6
2.4
2.6
2.8
3.0
3.2
3.4
3.6
Fitted Values of the Log−log Regression Model


Normal Q−Q Plot
Sample Quantiles
0.6
0.2
−0.2
−0.6
−3
−2
−1
0
1
2
3
Theoretical Quantiles

```
> # Create a bootstrap sample and recalculate the estimate #
>
> set.seed(12)
> beta1.hat( data = log.frame , index = sample(x = 392, size = 392, replace = T))

[1] -0.884816

> # Estimate the parameter B = 1000 times using B = 1000 bootstrap samples and #
> # Average the these values in order to reduce the variance of the estimate #
>
> boot(data = log.frame, statistic = beta1.hat, R = 1000)

ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = log.frame, statistic = beta1.hat, R = 1000)


Bootstrap Statistics :
     original      bias    std. error
t1* -0.841847 0.001052799  0.02755081

>
```

We next do the same for the model intercept. Although it has no practical interpretation, we simply include it to fit the model to the data.

```
> # Create a function that calculates the estimate of the log-log model slope #
> beta0.hat <- function(data, index){
+   X_0 <- data$horsepower.10s[index]
+   Y_0 <- data$mpg[index]
+   return(mean(Y_0) - (-0.841847)*mean(X_0))
+ }
> beta0.hat(data = log.frame, index = 1:392)

[1] 5.022225

> # Recalculate the estimate of the log-log model slope with a bootstrap sample #
>
> set.seed(10)
> beta0.hat(data = log.frame, index = sample(x = 392, size = 392, replace = T))

[1] 5.006397

> # Estimate the slope B = 1000 times with B = 1000 bootstrap samples #
> # and average them reducing the variability of the estimate #
>
> boot(data = log.frame, statistic = beta0.hat, R = 1000)
```

```
ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = log.frame, statistic = beta0.hat, R = 1000)


Bootstrap Statistics :
    original        bias    std. error
t1* 5.022225 -0.0002294325 0.009121353
```

We have obtained the desired bootstrapped model parameter estimates. Since this is a log-log regression model, the interpretation of $\hat{\beta}_1$ is *for every percent increase in horsepower of a vehicle engine we expect that the vehicle's mpg will decrease by 0.8418* % that has that engine. This parameter is called the *elastic impact* of engine horsepower on mpg ratings of vehicles manufactured in America, Europe, or Japan in this particular application. The code was included to show the reader how to bootstrap two different parameter estimates which are also statistics. In fact, bootstrap simulations can be applied to any statistic. To wrap up this section, we use our model to make a prediction, produce a confidence interval about that prediction, an estimation interval, and produce plots of our model fitted to the data.

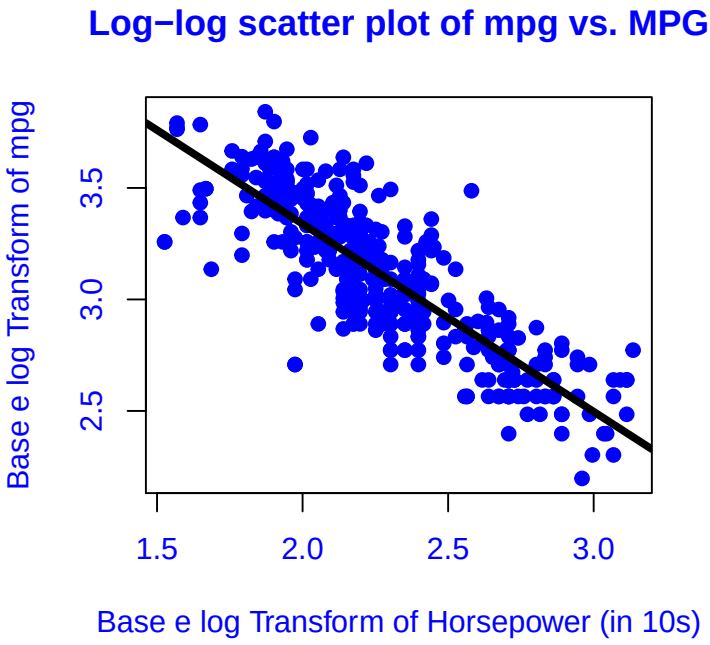
The minimum horsepower in the data set is 46 and the maximum horsepower is 230. Let's estimate the mpg of one of these types of cars that has a 175 horsepower engine. The fitted value and 95% confidence and prediction intervals are

```
       1
13.63568

       fit      lwr      upr
1 13.63568 13.16879 14.11912

       fit     lwr      upr
1 13.63568 9.56845 19.43175
```

. The fitted value or point estimate means that given $X = 175$ we expect that a car that has this engine to get 13.6 miles per gallon. Due to random chance (driving conditions such as city, highway, etc.), at minimum the miles per gallon the vehicle will get is 9.6 mpg and at maximum we expect it to get 19.4 mpg. This is the interpretation of the fitted value and prediction interval for an *individual car* based *only* on horsepower. That is a particular model and make. The 95% confidence interval can be interpreted as the following: we are 95% sure that all 175 horsepower engines used in these types of automobiles will produce an automobile mpg rating of 13.2 mpg at minimum, 14.1 mpg at maximum, and 13.6 mpg on average. The plot of the model with bootstrapped parameter estimates and standard errors can be found below.



Multiple Linear Regression

In the previous section, we uncovered the log-log model that was found to be effective. In this section, we seek a model where the modeling assumptions are met and has a higher R-Squared value. One way that we could find a more effective model is to consider modeling the mpg of these types of vehicles with a multiple linear regression model. Additionally, the dependent variable that we want to regress on other variables should follow a normal distribution. One can show that transforming the mpg data with a ln function forces the distribution of mpg to become approximately normal (this may not always work. Sometimes different transforms may be needed). Since this is the case, we model $\ln(mpg)$ in this section with a multiple linear regression model.

Given the application, intuition tells us that other factors than an engine's horsepower should affect a vehicle's mpg rating. For example, cars manufactured in America are notorious for having worse mpg ratings compared to vehicles manufactured by foreign companies. Additionally, things like increases in weight may require engine horsepower to increase so a vehicle can quickly obtain and maintain safe traveling speeds. Finally, one may hypothesize that the number of cylinders an engine has should also affect a vehicle's mpg rating. Suppose also we consider using the origin of the car as a factor that should affect the

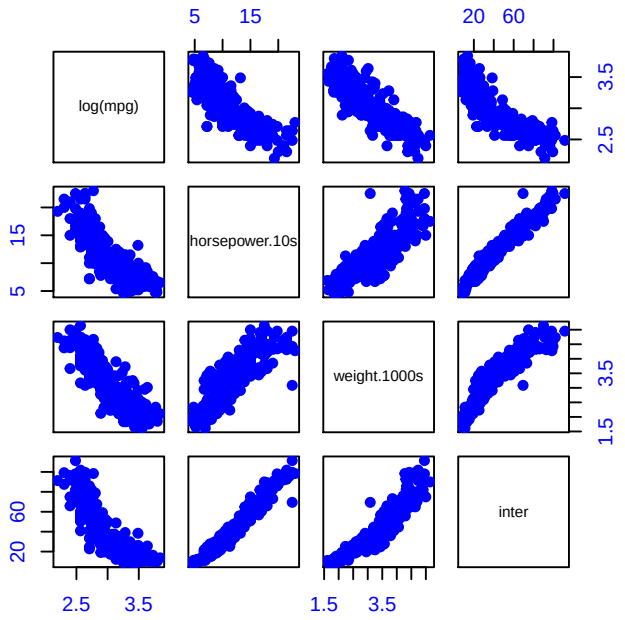
variable *mpg*. Finally, the weight of the vehicle as another variable. A possible multiple linear regression model is

$$\ln(E[Y \mid \vec{X}]) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_4 X_5 + \epsilon \quad (7)$$

where X_1, X_2 and X_3 are *binary valued* variables better know as *dummy variables*. The *coding scheme* of the dummy variables is as follows: $X_1 = 1$ if the car was manufactured in America or $X_1 = 0$ if it was not. $X_2 = 1$ if the car was manufactured in Europe or $X_2 = 0$ if it was not. $X_3 = 1$ if the cars engine has four cylinders $X_3 = 0$ if the engine has a different number of cylinders. β_4 is the slope for horsepower (X). β_5 is the slope for weight (X_5). β_6 is the slope for the *interaction* of X_4 and X_5. We can rewrite (7) as

$$\begin{cases} \beta_0 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_4 X_5 + \epsilon \\ \beta_0 + \beta_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_4 X_5 + \epsilon \\ \beta_0 + \beta_1 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_4 X_5 + \epsilon \\ \beta_0 + \beta_1 + \beta_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_4 X_5 + \epsilon \\ \beta_0 + \beta_2 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_4 X_5 + \epsilon \\ \beta_0 + \beta_2 + \beta_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_4 X_5 + \epsilon \end{cases} \quad (8)$$

But before estimating the model one should check for linearity between the dependent variable and all of the continuous predictors with scatter plots or a *scatter plot matrix* which is shown below.



The first row of the scatterplot matrix shows that the relationship between the natural log transform of mpg (the dependent variable) and horsepower, weight, and interaction of weight and horsepower (all of the continuous independent variables) is linear. So these results show that the linearity assumption is satisfied. The estimated model is

```
Call:
lm(formula = log(mpg) ~ American.Dummy + European.Dummy + X4.cylinders.dummy +
    horsepower.10s + weight.1000s + horsepower.10s:weight.1000s)

Residuals:
     Min       1Q   Median       3Q      Max
-0.37481 -0.09360 -0.01174  0.08909  0.57272

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                  4.111994   0.124351  33.068  < 2e-16 ***
American.Dummy              -0.060360   0.023690  -2.548  0.01123 *
European.Dummy              -0.066574   0.025158  -2.646  0.00847 **
X4.cylinders.dummy           0.120615   0.027892   4.324 1.95e-05 ***
horsepower.10s              -0.052686   0.011032  -4.776 2.55e-06 ***
weight.1000s                -0.235002   0.038683  -6.075 2.98e-09 ***
horsepower.10s:weight.1000s  0.006617   0.002739   2.416  0.01616 *
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1485 on 385 degrees of freedom
Multiple R-squared:  0.8121,        Adjusted R-squared:  0.8092
F-statistic: 277.4 on 6 and 385 DF,  p-value: < 2.2e-16
```

. Given the small p-values for t-tests for model parameters, we have evidence that each is significantly different from zero and conclude that this model is effective since the F-test for model significance returned a small p-value as well. Secondly, if one plots the residual plot and normal Q-Q plot, you will see that the errors are approximately normal, independent, and have a constant variance. Hence, we conclude that this *semi-log* multiple linear regression model with interaction (the log transformation is applied only to the response and not the independent variables) is robust. Note, the R-Squared value is almost 10% larger than that of the log-log regression R-Squared value. The typical error this model makes on the training data is 1.16 mpg. The test MSE estimates from LOOCV and ten-fold cross-validation tell us that the model makes an average error of 1.02 mpg for unseen observations. Overall, this model is somewhat comparable to the log-log model in terms of the residual standard error and estimated test MSE metrics. Though it may be more complicated, it does explain a significant proportion of variability in the dependent variable. Also, it does allow us to make inferences about multiple quantities and factors that have an impact on the dependent variable as well. To determine if this much more comlex model is really necessary, one must conduct a *nested F-test*. We next interpret the model and use it to make a prediction and construct confidence intervals.

Obviously $\hat{\beta}_0$ does not have a practical interpretation because it is the mpg of a car that is manufactured in Japan, with a zero horsepower engine, the engine is not a four cylinder engine, and the car weighs nothing. Now since this is a semi-log multiple regression model, the interpretation of the significant variable coefficients is given in terms of *percent change*. In particular, the semi-log interpretation of the impact of any independent variable (excluding the interacting continuous variables) in this model is the following: for any single unit change in a continuous predictor, X_j, the percent change in mpg is $100 * (e^{\hat{\beta}_j} - 1)\%$. The estimated dummy variable coefficients that represent categories is the following: when dummy variable $X_s = 1$ the *automatic* percent change in the dependent variable is $100*(e^{\hat{\beta}_s}-1)\%$ holding all other independent variables constant (take on values of zero). Finally, we interpret the interaction of continuous predictors in this model. That is we want to interpret $\hat{\beta}_4 X_4 + \hat{\beta}_5 X_5 + \hat{\beta}_6 X_4 X_5$. Note that,

$$\hat{\beta}_4 X_4 + \hat{\beta}_5 X_5 + \hat{\beta}_6 X_4 X_5 = (\hat{\beta}_5 + \hat{\beta}_6 X_4) * X_5 + \hat{\beta}_4 X_4 \tag{8}$$

. So if X_5 increases by one unit of weight then the percent change in mpg is $(\hat{\beta}_5 + \hat{\beta}_6 X_4)$ since

$$(\hat{\beta}_5 + \hat{\beta}_6 X_4)(X_5 + 1) = (\hat{\beta}_5 + \hat{\beta}_6 X_4) X_5 + (\hat{\beta}_5 + \hat{\beta}_6 X_4) \tag{9}$$

for a chosen value of X_4. In layman's terms, if the car has a 175 horespower engine then the percent change in mpg for a unit increase in weight is $\Delta\% = 100*(-0.235002+0.006617*17.5)\% = -11.9\%$. So if a car has a 175 horsepower engine and its weight increases by 1000 pounds then its mpg is expected to decrease by 11.9% as compared to what it previously weighed. Is this practical? Not really since the engine horsepower remains constant for such a large increase in weight. What is practical is to examine a weight increase of 100 pounds for a car that has a 175 horsepower engine. In this case the percentage decrease is $0.10 * -11.9\% = -1.19\%$. The interpretation of the interaction term in this model was lengthy and abstract. Nevertheless, we have interpreted the ordinary least squares estimates symbolically and numerically for this model. The fitted equation below is found below.

$$\begin{cases} 4.112 - 0.053 * X_4 - 0.235 * X_5 + 0.007 * X_4X_5 \\ 4.112 + 0.121 - 0.053 * X_4 - 0.235 * X_5 + 0.007 * X_4X_5 \\ 4.112 - 0.060 - 0.053 * X_4 - 0.235 * X_5 + 0.007 * X_4X_5 \\ 4.112 - 0.060 + 0.121 - 0.053 * X_4 - 0.235 * X_5 + 0.007 * X_4X_5 \\ 4.112 - 0.067 - 0.053 * X_4 - 0.235 * X_5 + 0.007 * X_4X_5 \\ 4.112 - 0.067 + 0.121 - 0.053 * X_4 - 0.235 * X_5 + 0.007 * X_4X_5 \end{cases} \tag{10}$$

To conclude our analysis we will compute a point estimate as well as a produce 95% confidence and prediction interval. For a car that was manufactured in America, has a 155 horsepower, weighs 3500 pounds, and is not a four cylinder we list the point estimate, confidence as well as the prediction interval below and can be interpreted similarly to the point estimate as well as the confidence intervals for the log-log model. We say similary because we consider different characteristics of the vehicle in the multiple linear regression model. The interpretation of them is almost the same as the log-log fitted value, confidence, and prediction intervals. The only difference is is that they consider characteristics such as manufacturer origin, the number of cylinders in the engine, and vehicle weight.

```
       1
15.98254

       fit      lwr      upr
1 15.98254 15.42804 16.55697

       fit      lwr      upr
1 15.98254 11.90963 21.44833
```

Summary

We analyzed the miles per gallon rating of automobiles manufactured in America, Europe, and Japan between 1970 and 1982 using linear regression

modeling to learn from these data. In our journery, we saw when linear models may not be appropriate because some or all of the modeling assumptions were violated. To be precise, the simple linear regression model was not an appropriate tool to use for explaining a vehicle's miles per gallon rating using engine horsepower because the linearity assumption was violated along with the requirement of the residuals having a constant variance and follow a normal distribution. We then tried to capture the nonlinear pattern between mpg and horsepower with a polynomial regression model. Similarly, the residuals did not exhibit constant variance or normality. However, we were able to transform for linearity and sample from a normal population. The next step we took was to increase the amount of variability we could explain in the miles per gallon rating of these types of cars by fitting a multiple linear regression model (semi-log model). Since the increase in the R-Squared was large, we concluded that this was a pretty good model. These are some of the steps a modeler can take when building a regression model from sample data.

References

[1] Sanford Weisberg. *Applied linear regression.* John Wiley and Sons, 1985.

[2] William Navidi. *Principles of statistics for engineers and scientists.* McGraw-Hill, 2010.

[3] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning with applications in R.* Springer, 2017.

[4] Ann R. Cannon, George W. Cobb, Bradley A. Hartlaub, Julie M. Legler, Robin H. Lock, Thomas L. Moore, Allan J. Rossman, and Jeffrey A. Witmer. *STAT2: building models for a world of data.* W.H. Freeman, 2013.

[5] Linear regression for business statistics. URL `http://online.rice.edu/courses/linear-regression-business-statistics/`.

[6] Victor Wright. An overview of linear regression modeling, 2019.

[7] Popular applications of linear regression for businesses, Nov 2018. URL `https://analyticstraining.com/popular-applications -of-linear-regression-for-businesses/`.