Statistics 3220: Final Project

Inverse Prediction

Frank Hancock, Tim Kim, Lindsey Reiter, Zulmira Yusuf

Introduction

The relationship between the mileage per gallon of a car model is of interest to a wide variety of people, including car buyers, car sellers, and manufacturers. The mileage per gallon of any given car is dependent on a lot of different factors, including cylinders, displacement, horsepower, weight (kg), acceleration, model year, origin, and car name. The purpose of the this project is to determine how is an automobile's mileage affected by variables such as vehicle weight, engine horsepower, and acceleration performance. To address this question, we used to dataset we selected and performed inverse prediction regression analysis. This should help us to predict the importance of an independent variables given the dependent variable.

We were interested in what factors might directly impact the mileage per gallon of a car, and how to be able to obtain more information when the mileage per gallon is given of a car. In the specific case, we are trying to test if inverse prediction would be the best method to gather relevant data if miles per gallon is given.

The reason why inverse prediction would work with this case is that some of the relevant variables from the dataset are hard to obtain. When looking for a car's acceleration, letting the car run repeatedly to get data is both time consuming and expensive. The same applies to gathering data for engine horsepower and weight. However, if use the inverse prediction to estimate each variable by using a given mileage per gallon value, we would be able to retrieve the data without spending too much time and money.

Data Summary

Our dataset for this project is downloaded from UCI Machine Learning Repository. This dataset was found in the StatLib Archives which is found at Carnegie Mellon University. The dataset was used in the 1983 American Statistical Association Exposition. The observations were taken over the years 1970 to 1982. There are 398 observations in the dataset. There were 6 missing values for horsepower so we removed these data points as they would not be useful in our study.

The variables used in the dataset are miles per gallon, cylinders, displacement, horsepower, weight (lbs), acceleration, model year, origin, and car name. Miles per gallon represents the amount of miles the vehicle can travel on a single gallon, or the ratio of the maximum range over the fuel tank size. Cylinders represent the number of cylinders in the vehicle's engine. Displacement represents the total volume within the cylinders in cubic inches. Horsepower is a measurement of the power output of the engine. Weight represents the amount of pounds the car weighs. Acceleration represents the time in seconds to accelerate from 0 to 60 miles per hour. Model year represents the year the vehicle was manufactured. Origin represents the geographical region the car came from. The origin was coded as 1 = USA, 2 = Europe, 3 = Japan. And finally, the car name is the brand and make of the car.

Because the independent variable in inverse prediction needed to be a quantitative value, we removed categorical variables such as car name, year and origin. From the remaining five variables, the strongest relationships to mpg were seen in the scatterplots of displacement, horsepower, and weight.

In order to choose the variables that we wanted to perform inverse prediction regression analysis on, we first created the scatterplot for each independent variable and the dependent

variable. For inverse prediction, we needed quantitative predictors that were linearly related to the dependent variable. Because it needed to be a quantitative value, car name, year and origin were not looked at because those were categorical values. From the remaining five variables, the strongest relationships to mpg were seen in the scatterplots of displacement, horsepower, and weight. The scatterplots for these variables can be found in Appendix A.

Since the scatterplots show a slight quadratic relationship, we attempted to mitigate this by transforming the x variable as a square root. As shown in Appendix B, performing this transformation does produce a scatterplot for each independent variable that looks more linear.

Besides the flattening of the scatterplots, the transformation proved useful when looking at the r-squared values. Before performing the square root function on x, the r-squared values were as followed: displacement=.6482, horsepower=.6059, weight=.6926. After the transformation, every r-squared value got closer to 1 which means that the percentage of the response variable variation that is explained by a linear model has increased, making the models stronger for prediction. This is indicated in Appendix C, in the chart we made comparing the coefficients of determination in our presentation. The r-squared values now are: displacement=.6746, horsepower=.6437, weight=.7058.

Methodology

Inverse prediction is used when it is more valuable to use y values to predict the x that will be needed to attain such y. Many call this a calibration problem since one uses this prediction "when inexpensive, quick, and approximate measurements (y) are related to precise, often expensive, and time-consuming measurements (X) based on n observations. The resulting regression model is then used to estimate the precise measurement x_hat(new) for a new approximate measurement y_hat(new)" (Kutner, 170). This is best understood if you wanted to increase the market share of your company which is dependent on how much you spend on advertising. It would be pretty impossible to just waste money on advertising as a way to see what market share the company will obtain. Companies will then use a y value of their desired market share they want to predict how much money will have to be spent on advertising.

To perform inverse prediction, multiple steps must be completed. The first step is to ensure that the assumptions of linear regression were met. That is, at any given value of x, the population of potential error term values has a constant variance, sigma^2, has a normal distribution, has a mean equal to zero, and has statistically independent error term values. The next step is to conduct a test of model adequacy using the null hypothesis of Beta_1=0 and the alternative hypothesis as Beta_1 is not equal to 0. If Beta_1 is not equal to 0, this means the x and y variables are linearly related. If you fail to reject the null, you should not move forward with this prediction type as if you move on with non linearly related variables, the results may be nonsensical. Once this is determined, one can fit a straight-line model to the sample to achieve a least-squares prediction equation. To make this prediction inverse, one must solve for x in terms of y. This means moving around the variables of the least squares prediction equation. This new equation can now be used to estimate x values, using y observations. Visually, this rearrangement would look like this:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$x = \frac{\hat{y} - \hat{\beta}_0}{\hat{\beta}_1}$$

After making predictions for x, one has to check whether the approximation is appropriate so one can move forward in constructing prediction intervals for the predicted x values. This is done by using the formula:

$$D = \left(\frac{t_{\alpha/2}s}{\hat{\beta}_1}\right)^2 \cdot \frac{1}{SS_{xx}}$$

The resulting value should be small; the textbook cites the maximum amount this value should be is .1. This is cited from the authors of Applied Linear Statistical Models as well as other statisticians. If it is less than .1, one can move forward to the final step of creating 95% prediction intervals. Appendix D shows an example from section 9.3 of the textbook of calculating this value for a question determining how much a company on television advertising(x) in a single month to gain a market share of 10% (y).

Inverse prediction is useful when the y variable is expensive, time-consuming, or dangerous/unethical to measure. One example of this, used in section 9.3 of the textbook, is with physicians determining the required dosage of a drug. Regression analysis showed that a linear relationship exists between decrease in blood pressure y and dosage x. Then a physician treating a new patient may want to determine the dosage to give to reduce the patients blood pressure by an amount y=y_p. The drawbacks to using this method include that using it for predictions will not be useful if there is insufficient evidence to reject the null hypothesis H_0: B1=0. Another drawback is the possible problems with extrapolation, predicting y for values of the independent variables that are outside of the experimental region. From Applied Linear Statistical Models, it is stated that there is controversy over using this prediction problem in the statistic community. It is said that predictions should be "made in direct fashion by regressing X on Y. This regression is called inverse regression" (Kutner, 170).

Analysis

From our exploratory data analysis, we moved forward in testing if the transformed models were significantly linearly related. We performed three models of adequacy tests. For all three predictor variables, we used the same null hypothesis that Beta_1 is equal to 0. If Beta_1 is equal to 0, the variable is not linearly related as there would be no slope. Our alternative that we tested was that Beta_1 was not equal to 0. At an alpha level of .05, we ran these tests through SAS and got the p-values of <.0001 for displacement, horsepower, and weight. Since all of these values are less than our alpha-level, we could confidently move forward in our inverse prediction as a linear relationship is detected.

Since all three showed linear relationships, we had to check the assumptions of linear regression. To check normality, we ran histogram plots for each model. In Appendix E, it is clear that the histogram for displacement shows the most non-normal curve, while the one for horsepower is relatively normal, and the one for weight is also relatively normal. However, we still meet our normality assumptions because regression is robust against minor deviances in normality.

To make sure there was no correlation of the residuals, we ran residual analysis tests for each model. The error term is normally distributed which indicates that the residuals are also

normally distributed. To examine if the residuals for each model had a mean of zero, we analyzed the residual plots. The residual plot showed a fairly random pattern which indicates that the a linear model provides a decent fit to data for each variables. To check if the residuals for each model had equal variance, we looked at the residuals versus fitted values plot. The vertical width of the scatter plot doesn't increase or decrease across the values, so we can assume that the variance in the error term is constant. From above tests, we can conclude that all variables fit the assumption, so we can proceed to perform inverse prediction using linear regression.

From the individual model outputs in SAS, we were able to create the prediction equations that we will use for inverse prediction. After algebraically rearranging the original equations, we got the three functions for inverse prediction: Displacement (x_hat)= (y_p - 47.11839) / -1.75878, Horsepower (x_hat) = (y_p- 58.70517) / -3.50352, Weight(x_hat) = (y_p - 69.67218) / -0.85560. We then decided to predict displacement, horsepower, and weight if our y_p is 20 miles per gallon. Our predictions for displacement, horsepower, and weight were 15.4189, 11.0475, and 58.0554, respectively. We then had to square these values since we performed a square root function on the data earlier to get a displacement equal to 237.741, horsepower equal to 122.047, and a weight equal to 3370.43 pounds.

Before making the 95% prediction intervals, we had to check if each of these prediction values was an appropriate approximation. The calculations for determining if the predictions were appropriate is shown in Appendix F. Each prediction had a calculated D-value of less than .1, satisfying the need for the value to be small. This allowed us to create prediction intervals at 95% confidence for each prediction value.

We decided to calculate our prediction intervals by hand which is shown below;

$$Displacement\ interval: 237.741 \pm (2.228)\left(\frac{4.458081}{-1.75878}\right)\sqrt{1+\left(\frac{1}{393}\right)+\frac{(237.741-193.42588)^2}{4394100.83}} = 237.741 \pm 5.65588$$

$$Horsepower\ interval: 122.047 \pm (2.228)\left(\frac{4.664822612}{-3.50352}\right)\sqrt{1+\left(\frac{1}{393}\right)+\frac{(122.047-104.4694)^2}{568378.77}} = 122.047 \pm 2.97109$$

$$Weight\ interval: 3370.43 \pm (2.228)\left(\frac{4.239166192}{-0.85560}\right)\sqrt{1+\left(\frac{1}{393}\right)+\frac{(3370.43-2977.584)^2}{273234943.52}} = 3370.43 \pm 11.0560$$

These all seemed like reasonable prediction intervals, especially since we had a widespread range of observation values. After conducting the prediction interval, we then pulled out all the observation with 20 miles per gallon so that we could take the average of displacement, horsepower, and weight of all these observations to be our one observation point. The average for displacement for selected observations was 192.438. It didn't fall in the prediction interval (232.086, 243.398) that was conducted from the inverse prediction equation.The average for horsepower was 99.875, also didn't fall in the prediction interval of (119.076, 125.018). The average of weight was 3009.94 also didn't fall in the prediction interval of (3359.37, 3381.49).

Conclusion

Each predictor showed a slightly strong negative correlation with our dependent variable, miles per gallon. It made sense that each predictor has related to miles per gallon in a negative way because the heavier the car and the more energy needed for horsepower or pushing the cylinders leads to less available fuel for miles per gallon. Also, the r-squared values are all positive in our findings because the negative correlation coefficient, r, results in a positive coefficient of determination. All three predictors passed the model of adequacy test, in which we were able to reject the null hypothesis that Beta_1=0. This means that all three showed a linear

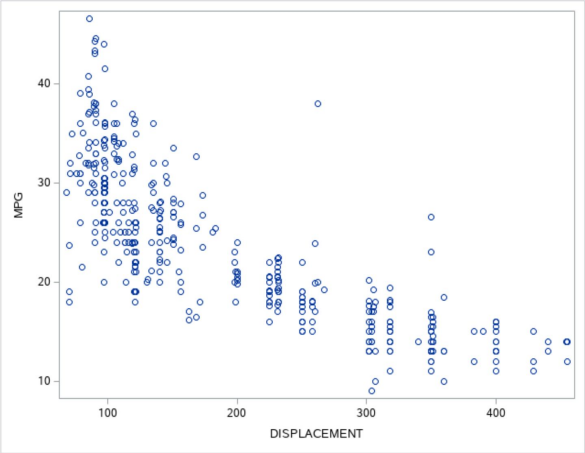
relationship with miles per gallon. Furthermore, all prediction values were appropriate approximations.

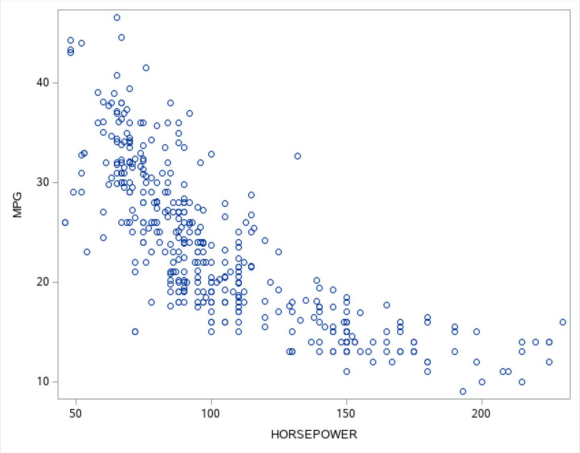
We concluded that weight was the best independent variable to use in inverse prediction. If we were to plug in a miles per gallon value, this would give us the best chance in predicting the weight needed to achieve such a miles per gallon. Weight had the largest r-squared value which means it had the highest percentage of variation in y explained by x. This means that this simple model was better than the other two to use when predicting. Although weight had the strongest coefficient of determination, it had the widest prediction interval out of all predictor variables. We decided that this was still okay to use since the weight varies a lot between different brands of cars, the models within the brands, and over time. As the cars got more modern, it would make sense for the cars at that weight to be able to perform at a better miles per gallon rate with new efficient technologies. Those at heavier weights may have been able to perform at better miles per gallons as well. This variability is shown in the large prediction intervals, for not just weight but for displacement and horsepower.

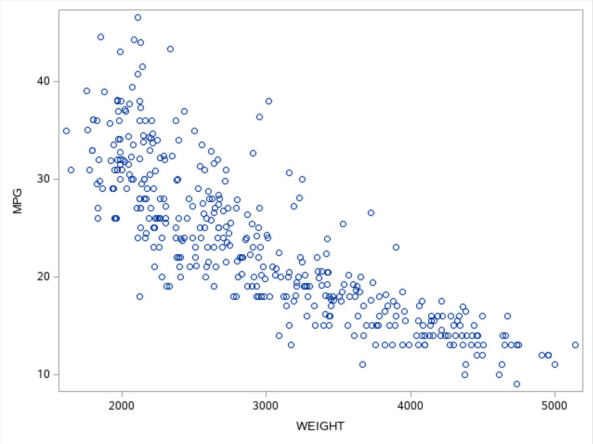
On the same lines, it was hard to compare our predictions obtained with the observed data. This is because many cars had 20 miles per gallon as a y-value, but did not have the same values for each variable. To try and account for this, we took the average of the observations of displacement, horsepower, and weight for all cars that had 20 miles per gallon. This most likely affected our comparisons of the prediction intervals with the actual data, but it did not seem like we would be accurate in just picking one car's observed displacement, horsepower, and weight to compare.

Future research on what miles per gallon could be used to predict would have to come from a more refined dataset. It would have been helpful if we were just looking at one brand of cars or one model. New research could look at more modern cars to get a better look at what can be predicted from miles per gallon. Furthermore, our lack of a true best model may have just been because miles per gallon may not be the best to use to "calibrate" measurements. There may be better variables that are easy to obtain that can then be used to predict what independent variable is needed to produce such a result. Multiple variables are needed to work together to generate mileage per gallon. Because of this, the simplification of inverse prediction of just choosing one variable did not seem useful for such a topic.

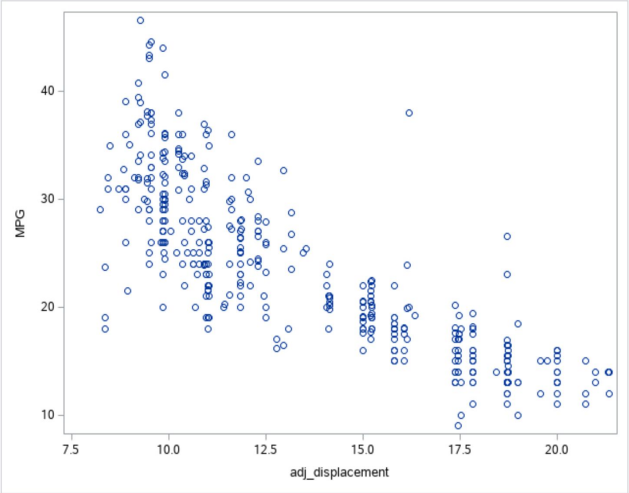
Appendix A - Scatterplots for Displacement vs MPG, Horsepower vs MPG, and Weight vs MPG

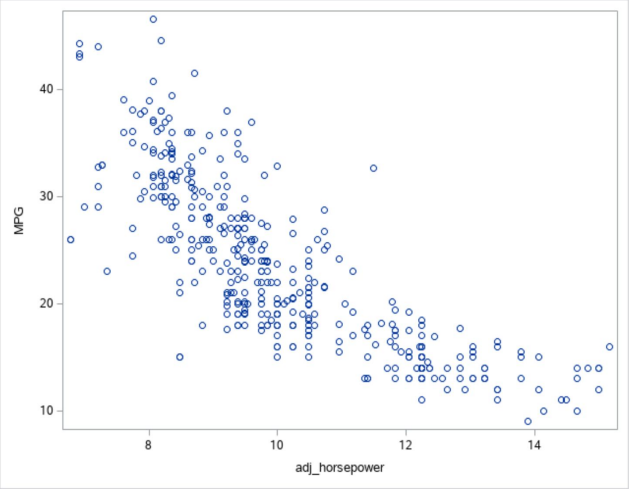


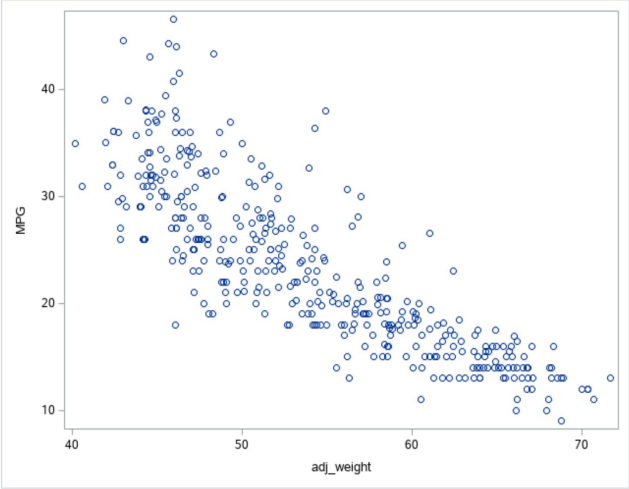




Appendix B- Scatterplots After Applying y=sqrt(x) Transformation







Appendix C - Table of R^2 Values Before and After Transformation

	Displacement	Horsepower	Weight
R^2 Before Transformation	0.6482	0.6059	0.6926
R^2 After Transformation	0.6746	0.6437	0.7058

Appendix D - Example Problem of Determining Whether a Prediction is an Appropriate Approximation

Problem is found in Section 9.3 of the textbook

$$D = \left(\frac{t_{\alpha/2}s}{\hat{\beta}_1}\right)^2 \cdot \frac{1}{SS_{xx}}$$

α=.05

$t_{\alpha/2} = t_{.025} = 2.228$

n−2=12−2=10 degrees of freedom

From the printout of the MINITAB analysis of straight-line model, s = .520 and $\hat{\beta}_1$ = .397.

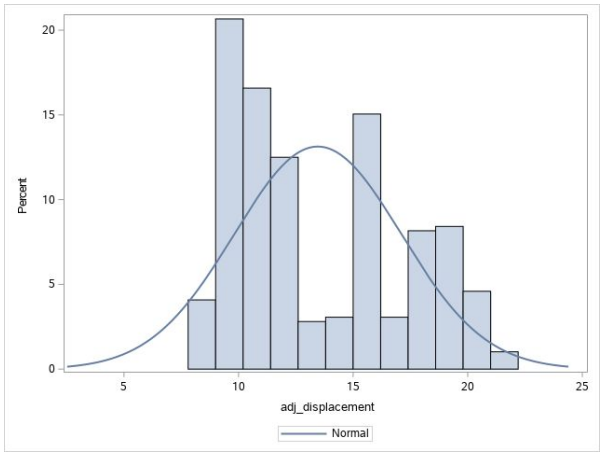
SS_{xx} is not found on the analysis printout so to calculate SS_{xx}=

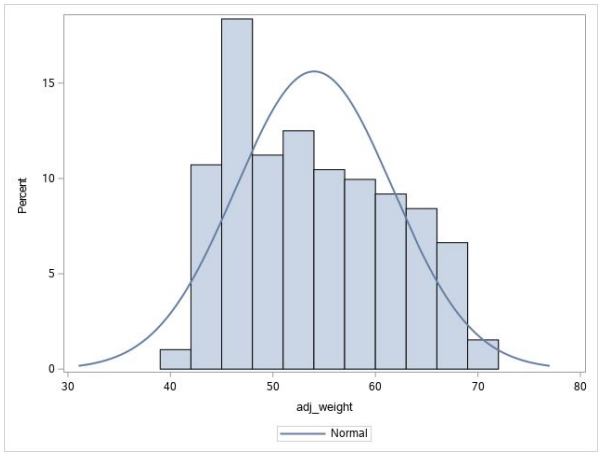
$$\begin{aligned} SS_{xx} &= \sum x^2 - n(\bar{x})^2 \\ &= 8,570 - 12\left(\frac{318}{12}\right)^2 \\ &= 8,570 - 8,427 \\ &= 143 \end{aligned}$$

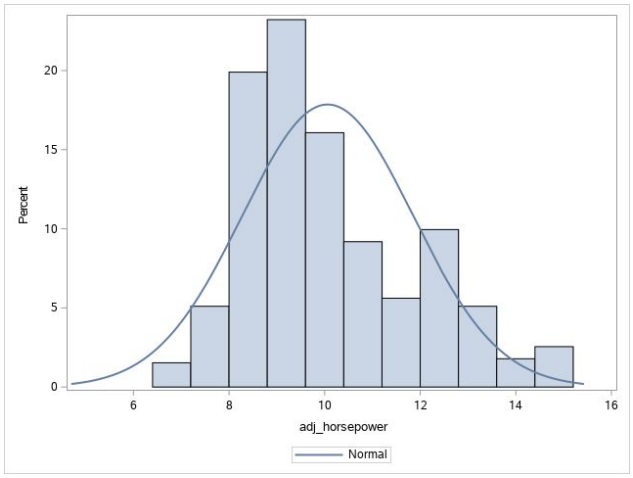
The value to use to determine if the approximation is appropriate is as followed. It is less than the 0.1 threshold which means it is small enough to move forward with constructing prediction intervals

$$\begin{aligned} D &= \left(\frac{t_{\alpha/2}s}{\hat{\beta}_1}\right)^2 \cdot \frac{1}{SS_{xx}} \\ &= \left[\frac{(2.228)(.520)}{.397}\right]^2 \cdot \frac{1}{143} = .06 \end{aligned}$$

Appendix E- Regression Assumptions







Appendix F - Calculating the D-values for Predictions

$SS_{xx} = \Sigma\, x^2 - n(\bar{x})^2$

$SS_{xx}(displacement) = 19097634.25 - 393(193.42588^2) = 4394100.83$

$SS_{xx}(horsepower) = 4857524 - 393(104.4694^2) = 568378.77$

$SS_{xx}(weight) = 3757575489 - 393(2977.584184^2) = 273234943.52$

$$D = \left(\frac{t_{\alpha/2}s}{\beta_1}\right)^2 * \frac{1}{SS_{xx}}$$

$D_{displacement} = [(2.228)(4.458081)/(-1.75878)]^2 * (1/4394100.83) = 0.000007258$

$D_{horsepower} = [(2.228)(4.664822612)/(-3.50352)]^2 * (1/568378.77) = 0.000015483$

$D_{weight} = [(2.228)(4.239166192)/(-0.85560)]^2 * (1/273234943.52) = 0.000000446$

Appendix G - Calculations of 95% Prediction Intervals

$$\hat{x} \pm t_{\alpha/2}\left(\frac{s}{\hat{\beta}_1}\right)\sqrt{1 + \frac{1}{n} + \left(\frac{(\hat{x}-\bar{x})^2}{SS_{xx}}\right)}$$

$$\textit{Displacement interval}: 237.741 \pm (2.228)\left(\frac{4.458081}{-1.75878}\right)\sqrt{1 + \left(\frac{1}{393}\right) + \frac{(237.741-193.42588)^2}{4394100.83}} = 237.741 \pm 5.65588$$

$$\textit{Horsepower interval}: 122.047 \pm (2.228)\left(\frac{4.664822612}{-3.50352}\right)\sqrt{1 + \left(\frac{1}{393}\right) + \frac{(122.047-104.4694)^2}{568378.77}} = 122.047 \pm 2.97109$$

$$\textit{Weight interval}: 3370.43 \pm (2.228)\left(\frac{4.239166192}{-0.85560}\right)\sqrt{1 + \left(\frac{1}{393}\right) + \frac{(3370.43-2977.584)^2}{273234943.52}} = 3370.43 \pm 11.0560$$

Appendix H - SAS Code

```
* Importing data;
data mpg1;
infile '/folders/myfolders/sasuser.v94/STAT 3220/Final Project/auto-mpg.csv' delimiter=','
firstobs=2;
input MPG CYLINDERS DISPLACEMENT HORSEPOWER WEIGHT ACCELERATION
YEAR ORIGIN NAME $;
run;

* Removing entries with missing elements;
data mpg;
set mpg1;
where HORSEPOWER ne .;
run;

* Making scatterplots and testing linear regression models for each variable;
proc sgplot data=mpg;
scatter y=mpg x=cylinders;
run;
proc reg data=mpg plots=none;
model mpg = cylinders;
run;
proc sgplot data=mpg;
scatter y=mpg x=displacement;
run;
proc reg data=mpg plots=none;
model mpg = displacement;
run;
proc sgplot data=mpg;
scatter y=mpg x=horsepower;
run;
```

```
proc reg data=mpg plots=none;
model mpg = horsepower;
run;
proc sgplot data=mpg;
scatter y=mpg x=weight;
run;proc reg data=mpg plots=none;
model mpg = weight;
run;
proc sgplot data=mpg;
scatter y=mpg x=acceleration;
run;
proc reg data=mpg plots=none;
model mpg = acceleration;
run;
proc sgplot data=mpg;
scatter y=mpg x=year;
run;
proc reg data=mpg plots=none;
model mpg = year;
run;
proc sgplot data=mpg;
scatter y=mpg x=origin;
run;
proc reg data=mpg plots=none;
model mpg = origin;
run;

* Making adjustments to data;
data mpgadj;
set mpg;
adj_displacement = displacement**(1/2);
adj_horsepower = horsepower**(1/2);
adj_weight = weight**(1/2);
run;

* Viewing data after adjustments;
proc sgplot data=mpgadj;
scatter y=mpg x=adj_displacement;
run;
proc reg data=mpgadj plots=none;
model mpg = adj_displacement / clb;
run;
proc sgplot data=mpgadj;
scatter y=mpg x=adj_horsepower;
```

```
run;
proc reg data=mpgadj plots=none;
model mpg = adj_horsepower / clb;
run;
proc sgplot data=mpgadj;
scatter y=mpg x=adj_weight;
run;
proc reg data=mpgadj plots=none;
model mpg = adj_weight / clb;
run;

* Algebraically rearranging expression to solve for x
y = Beta_0 + Beta_1 * x
x = (y_p - Beta_0) / Beta_1;

data calculations;
* Beta terms from regression;
d0 = 47.11839; d1 = -1.75878;
h0 = 58.70517; h1 = -3.50352;
w0 = 69.67218; w1 = -0.85560;

* mpg value for prediction;
y = 20;

* Computing the predicted x given y = 20;
xd = (y-d0)/d1; xh = (y-h0)/h1; xw = (y-w0)/w1;
dfinal = xd**2; hfinal = xh**2; wfinal = xw**2;

n = 393; * Number of observations;
t = 2.228; * t(alpha/2) of alpha = 0.05;
mean_disp = 193.42588;
mean_hp = 104.4694;
mean_wt = 2977.584;
s_disp = 4.458081426;
s_hp = 4.664822612;
s_wt = 4.239166192;
SS_disp = 19097634.25; * Sum of squares of x;
SS_hp = 4857524;
SS_wt = 3757575489;
* Calculating SSxx values for each variable;
SSxx_disp = SS_disp - n*(mean_disp**2);
SSxx_hp = SS_hp - n*(mean_hp**2);
SSxx_wt = SS_wt - n*(mean_wt**2);
* Calculating D statistic;
```

```
D_disp = (t*s_disp/d1)**2 * (1/SSxx_disp);
D_hp = (t*s_hp/h1)**2 * (1/SSxx_hp);
D_wt = (t*s_wt/w1)**2 * (1/SSxx_wt);
* Computing prediction intervals;
B1_disp = -1.75878;
B1_hp = -3.50352;
B1_wt = -0.85560;
x_hat_disp = 237.741;
x_hat_hp = 122.047;
x_hat_wt = 3370.43;
s_disp = sqrt(19.87449);
s_hp = sqrt(21.76057);
s_wt = sqrt(17.97053);
interval_disp = t * (s_disp/B1_disp) * sqrt(1 + 1/n + (((x_hat_disp -
mean_disp)**2)/SSxx_disp));
interval_hp = t * (s_hp/B1_hp) * sqrt(1 + 1/n + (((x_hat_hp - mean_hp)**2)/SSxx_hp));
interval_wt = t * (s_wt/B1_wt) * sqrt(1 + 1/n + (((x_hat_wt - mean_wt)**2)/SSxx_wt));
proc print data=calculations;
run;
```

Works Cited

Kutner, Michael H., et al. *Applied Linear Statistical Models*. McGraw-Hill Education, 2013.

Mendenhall, William, and Terry Sincich. *A Second Course in Statistics Regression Analysis*. Pearson Education, 2014.