Automobile Pricing

## Introduction

Purchasing a vehicle is often a stressful endeavor, especially when it comes to determining whether or not a vehicle you’re interested in is appropriately priced. There are myriad factors that come into play, such as fuel efficiency, weight, even style; it would thus be useful to have a means of estimating a vehicle’s worth based upon its attributes. Armed with such information, an individual could gain the upper hand in negotiation on price.

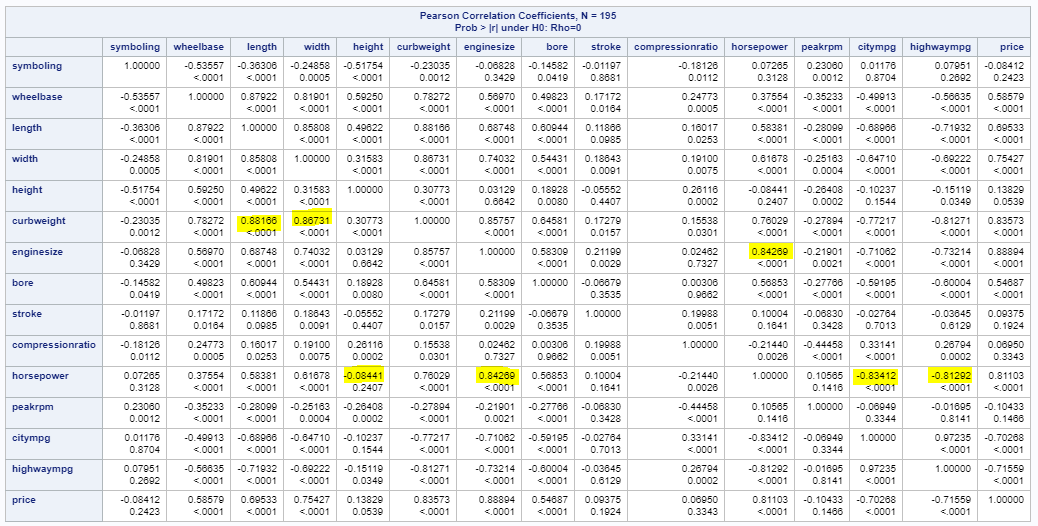
Using the data from the 1985 Ward’s Automotive Yearbook, which itself is collected from three other sources – 1985 Model Import Car and Truck Specifications (1985 Ward's Automotive Yearbook); Personal Auto Manuals (Insurance Services Office, 160 Water Street, New York, NY 10038); and Insurance Collision Report (Insurance Institute for Highway Safety, Watergate 600, Washington, DC 20037) – our goal was to investigate the data, find the significant determinants of the vehicle’s price, then use those factors to build a model that would be able to reasonably predict the price of a given make of vehicle.

## Exploratory Analysis

We began by examining the factors available to us in the data set: we found 9 physical car attributes, such as make, model, engine location, and curb-weight; 14 performance attributes, including fuel type, aspiration, engine size, horsepower, and miles-per-gallon for both city and highway; 2 insurance risk attributes, normalized losses and symboling (an indicator of insurance risk rating); and price, which we will be using as our response variable.

We then began looking through the data itself. We found that a 41 records were missing values for the Normalized Losses variable; we chose to remove the variable from the dataset, as with so many missing it, it would be insufficiently reliable to include in our analysis.

Figure 1

Once we removed the records with missing data, we compared the resulting clean data set to the original; as the means of the deleted 10 records are within range of the original data, we feel that their removal will not impact our analysis.

We then checked for correlated variables, and found several variables to have high correlation to one another (Figure 1); after looking at the specific variables involved, we determined that removal of the *Curb Weight*, *Wheelbase*, and *Highway MPG* variables would be sufficient to limit correlation between our variables.

## Regression Models (Objective 1)

In this section, we will discuss our regression model in detail.

### Restatement of Problem

Our goal is to build a regression model using our aforementioned automobile data, which would allow us to predict a vehicle’s price from some set of its characteristics. We have a wide variety of factors that we will be able to select from, some of which turned out to be correlated with other variables and were thus removed from consideration.

### Model Selection

In this section, we will review the model we produced, as well as some of the analysis of that model as we refined it over time.

#### Type of Selection

We initially used a LARS model; this resulted in a model using 9 features, with an adjusted R-square of 85%, indicating that 85% of the variation in price could be explained by these 9. However, that seemed to be an excessive number of variables, and so we decided to compare this with a LASSO model.

Using LASSO, we produced a model using 4 variables as features, with an adjusted R-square of 73%. While this is lower than that for the LARS model, the SBC is smaller (3149.56 for LARS, against 1641.95 for LASSO; see Appendix for more detailed figures), and the interpretation is simpler, with 4 variables compared to 9; as such, we decided to go with the LASSO model.

#### Assumption Checks

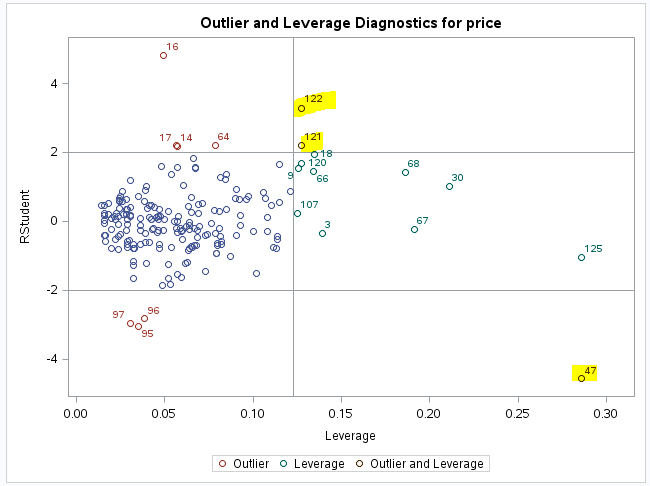
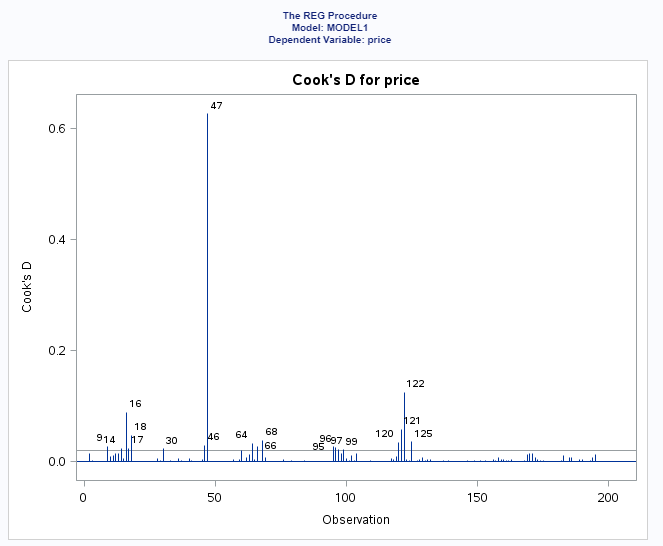
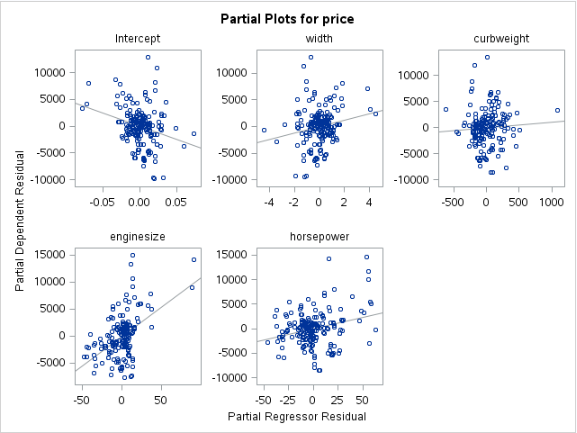


Figure 2

Observations 47, 120, and 121 are shown to have high leverage and outliers, but only 47 has a notable Cook’s D (Figure 2). In light of this, we removed observation 47 from our cleaned data set. After such removal, the only notable observation was 121, which seemed to us to be borderline and was thus retained.

Figure 3

In terms of residuals, our LASSO gave us a model using four variables (not including price, our response variable): *Width*, *Curb Weight*, *Engine Size*, and *Horsepower*. However, while the residuals were acceptable for these, we found that in the partial plots that *Curb Weight* did not have a significant slope, with a p-value of .39 (Figure 3); we decided to remove that attribute from the model, and rerun it.

After removing *Curb Weight*, our residuals still looked good, and our partial plots all had significant p-values (see Appendix, Figure 1-A: Final Residuals; and Figure 1-X: Final Partial Plots).

**Linearity:** The relationship between the dependent variable and the continue independent variables appear to be linear.

**Normality:** Residuals of the linear model are normally distributed.

**Equal Variance:** The variance of the residuals is constant for every combination of the independent variables and, thus, is also constant across all of the predicted values.

**Independence:** The car make is indicated, but not the model; the only information we have is that these are cars from the relevant 1985 line. With this in mind, we proceed with caution: in the future, we would try to obtain the model information for each observation to be certain that observations are independent.

### Parameter Interpretation

Judging from our results, it would appear that the *Price* of a vehicle increases as each of *Width*, *Engine Size*, and *Horsepower* increase; of these, *Horsepower* appears to have the least impact, while *Engine Size* has the most.

**Section on confidence intervals needs to go here**

### Conclusions

Our model appears to be sensible: specifically, the attributes that were most telling of a vehicle’s price – width, engine size, and horsepower – all seem like reasonable predictors of the price of a vehicle. Larger trucks and wider vehicles, those with increased engine size and horsepower, would presumably be more expensive, most likely due to the more intensive engineering requirements that go into larger, more powerful vehicles.

## Two Way Anova (Objective 2)

Words.

### Goal

Words.

### Main Analysis

Words.

### Conclusion

Words.

## Appendix

Words.