Kelley Blue Book Dataset
Jacob Andros

## Abstract

This analysis uses data from the Kelley Blue Book to investigate factors that influence the resale price of cars. The dataset includes around 800 cars with 11 measured factors and their resale prices. We used two multiple linear regression models to perform the analysis. One model indicates that Model, Body Type, Number of Cylinders and Mileage adequately explain and predict resale price. The other model shows that the effect of Mileage on Resale Price does not differ depending on the make of the car. We conclude that Cadillac convertibles have the highest resale prices when holding mileage constant. Although these variables adequately explain and predict resale prices, there are other factors that are worth considering, such as a car's age or accident history.

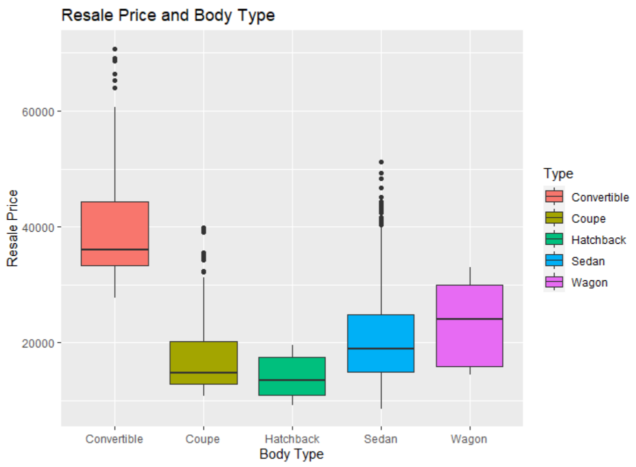
## Introduction

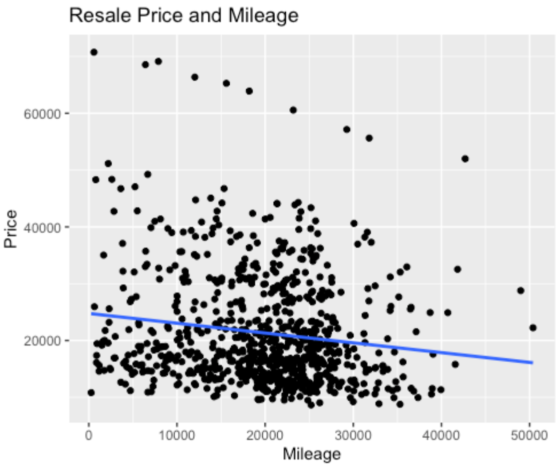
### *Background*

The Kelley Blue Book (KBB) is a trusted resource that helps car owners understand the worth of their used cars. The "blue book values" are determined by actual sales transactions and adjusted to account for seasonality and market trends. The KBB is widely known for giving car owners the most up-to-date used car pricing information. For this analysis we use a subset of KBB data to help consumers understand why their car has the KBB value it has and to give them a price range that they can expect for a resale of their car.

### *Exploratory Analysis*

The data for this analysis contains the resale price of the car and 11 explanatory variables for each car. These include the mileage, make, model, trim, body type, number of cylinders, liter (a more specific measure of engine size), number of doors, and whether or not that car has cruise control, upgraded speakers and leather seats. For exploratory purposes, we will make some graphs with a couple of these variables and the price of the car. These can be seen below.





From these graphs we might infer that resale price generally decreases as mileage increases and that convertible cars tend to have higher resale values. But to determine the exact effect that these variables have on resale price, we will fit a linear model to the data using multiple linear regression.

*Analysis Goals*

The goals of fitting the linear model are twofold:

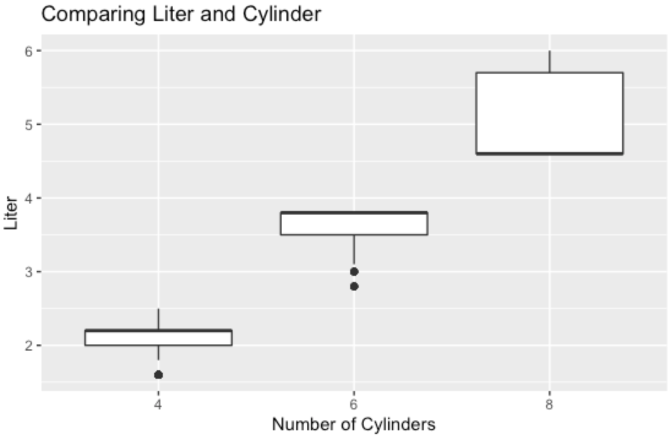
- We want to understand which characteristics of a used car contribute to a higher or lower resale value, and to what magnitude each characteristic does so. In particular, the model will be used to understand which cars retain higher resale values at 15,000 miles, and whether there is a significant interaction effect between a vehicle's make and mileage in determining its price.
- We will use the model to predict the resale value of other vehicles not in the dataset. The model will allow us to quantify uncertainty in these predictions.

**Linear Modeling**

*Variable Selection*

We chose to approach this analysis with multiple linear regression because of its interpretability and clarity. In order to use the linear model for prediction and inference, it needs to meet four assumptions. These assumptions include linearity, independence of data, normality of model residuals and equal variance, and will require a log transformation of several variables. These assumptions and transformations will be described in more detail later in the analysis.

We also need to consider collinearity between variables which can be measured by the Variance Inflation Factor (VIF). If one or more variables are highly collinear, or correlated, then it will be difficult to estimate the effects and significance of the explanatory variables accurately because it will inflate the variance of the standard errors. This can be resolved by eliminating or combining certain variables. In this data set, there are quite a few variables that are at risk of being collinear. For example, the graph below shows side by side boxplots of the number of cylinders and liters. Both are a measure of engine size, but the liter is more specific. As the graph illustrates, knowing the liter does not give us much additional information from what cylinder already gives us. If a car has 6 cylinders, we know that the liter size will be between 2.5 and 4. Because of this relationship, it will probably be best to only use one of these variables, not both.



In addition to these variables, *Make*, *Model*, and *Trim* are all highly related. *Trim* often contains the body type and number of doors, which are already given to us with other variables. Also, most trims are specific to certain makes, so that would provide redundant information. Similarly, the car model is specific to the make, so knowing the model would give us the make as well. Dealing with these types of variables can be difficult, but if all variables are included, there will be nonexistent coefficients and inflated standard errors. As stated previously, one should either eliminate some variables or combine them to avoid these problems.

For the purpose of this analysis, we have chosen two different linear models. Both are specified below.

<u>*Linear Model 1*</u>

$$log(Price)_i = \beta_0 + \beta_1 log(Mileage)_i + \beta_2 I(Type\!: Coupe)_i + \ldots$$
$$+ \beta_5 I(Type\!: Wagon)_i + \beta_6 I(Model\!: Corvette)_i + \ldots. + \beta_{36} I(Model\!: XLR - V8)_i + \beta_{37} Cylinder_i$$
$$+ \varepsilon_i$$
$$\varepsilon_i \sim N(0, \sigma^2)$$

In this equation, $\beta_0$represents the intercept, or the expected *Log-Price* of a hypothetical car with 0 miles on it, no cylinders, that is a convertible 9_3 model (the baseline level for the *Model* variable). $\beta_1$represents the average expected change in *Log-Price* as *Log-Mileage* goes up by one, holding everything else constant. $\beta_2$through $\beta_5$represent the average expected difference in *Log-Price* of a convertible car and a car of the type specified in the indicator statement, holding everything else constant. $\beta_{37}$represents the average expected change in *Log-Price* of a car as the number of cylinders increase by one, holding everything else constant. The rest of the coefficients ($\beta_6$-$\beta_{36}$) represent the expected difference in *Log-Price* between a 9_3 model and a car of the model specified in the indicator statement, holding all else constant. $\varepsilon_i$ represents the random error of each car, or the difference between a car's *Log-Price* as predicted by the model and the log of its actual price. We assume that all the random errors are distributed normally with a mean of 0.

<u>*Linear Model 2*</u>

$$log(Price)_i = \beta_0 + \beta_1 log(Mileage)_i + \beta_2 I(Make\!: Cadillac)_i + \ldots + \beta_6 I(Make\!: Saturn)_i +$$
$$\beta_7 I(Type\!: Coupe)_i + \ldots + \beta_{10} I(Type\!: Wagon)_i + \beta_{11} Cylinder_i +$$
$$\beta_{12} I(Make\!: Cadillac)_i * log(Mileage)_i + \ldots + \beta_{16} I(Make\!: Saturn)_i * log(Mileage)_i$$
$$\varepsilon_i \sim N(0, \sigma^2)$$

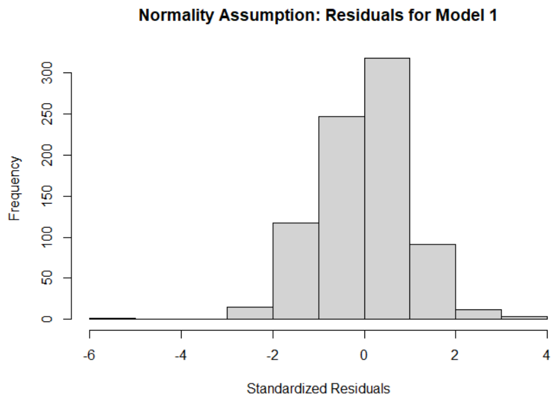
In this second model, most of the coefficients can be interpreted similarly to those in Model 1. The coefficients with an indicator statement for *Make* represent the average expected difference between a car that is a Buick and a car of the make specified in the indicator statement, holding all else constant. This model also has coefficients for the interaction which represent how the effect of mileage changes depending on the make of the car.

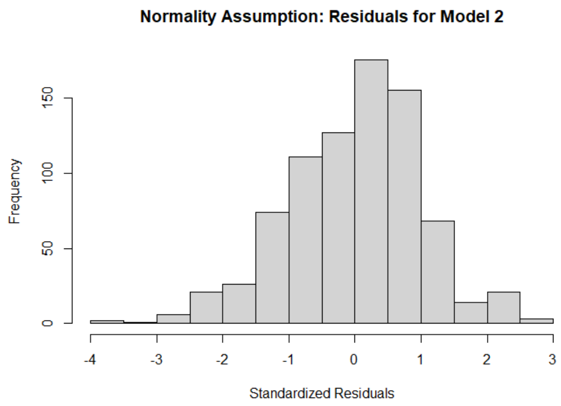
In summary, we implemented one model with *Mileage, Model, Body Type,* and *Cylinder*, and a second model with *Mileage, Make, Body Type, Cylinder*, and a *Make-Mileage* interaction. The advantage of the first model is that including *Model* gives a better overall fit, while the second model has the advantage of allowing us to study a possible *Make-Mileage* interaction. These advantages and disadvantages will be discussed further later on.

## Model Assumptions

In order to use both models for inference and/or predictions, they each need to meet certain assumptions (as described earlier) for multiple linear regression.

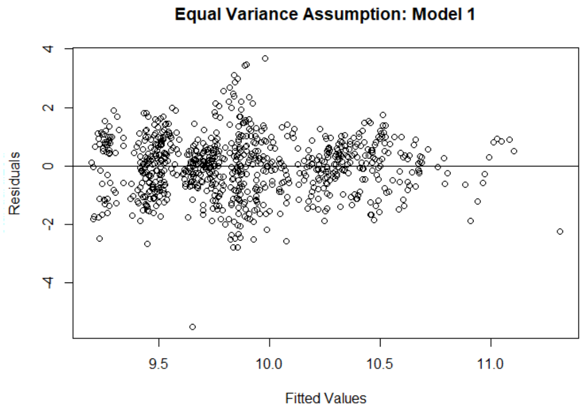
<u>Normality</u>: First, the distribution of the model residuals must be approximately normal. This assumption was the most difficult to satisfy for both models, and it was only after log-transforming both *Price* and *Mileage* that the residuals appeared somewhat normal, in both models. The histograms below demonstrate this.

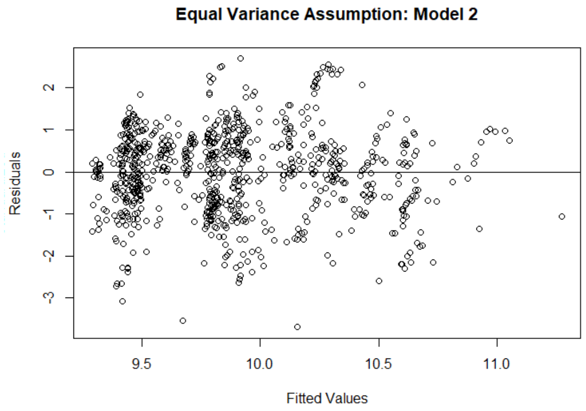




Both sets of model residuals appear approximately normal (symmetric and single-peaked) in shape. A KS-test for normality yields p-values of .131 (model 1) and .054 (model 2). Both of these p-values are low, but using α = .05, we would not be able to firmly conclude non-normality in either model.

Equal Variance: The spread of the model residuals must be approximately constant across its fitted values, or predictions (also known as homoscedasticity). The plots of fitted values vs. residuals below demonstrate this for each model.

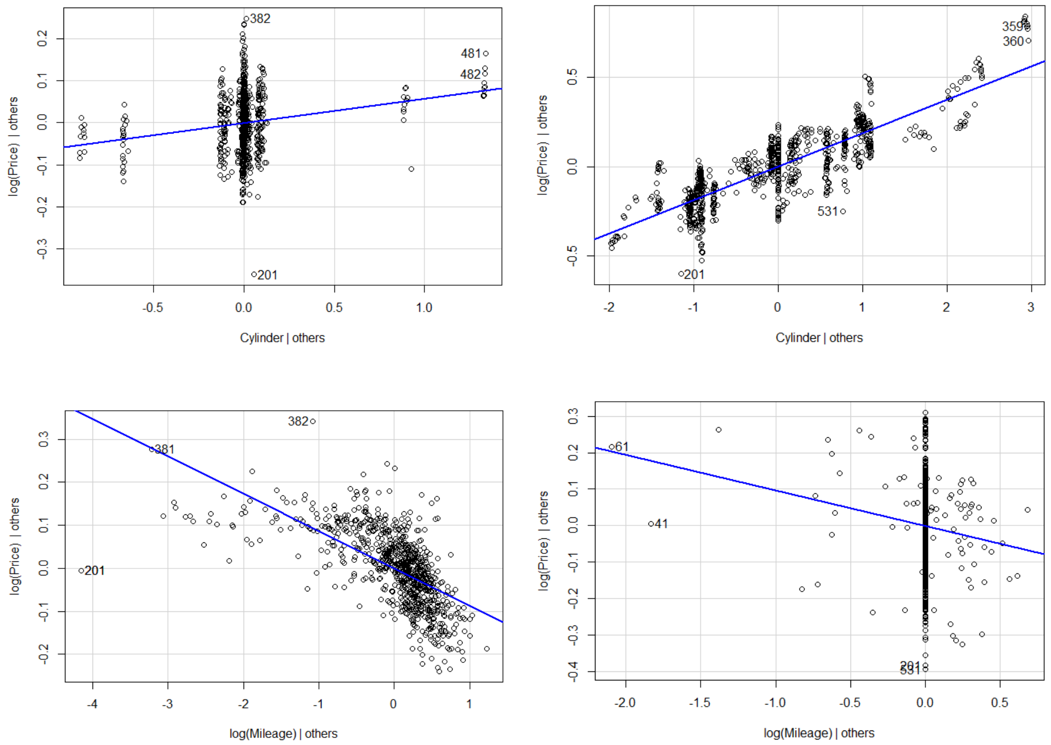




By simply looking at the plots, it doesn't look like there are any large fluctuations in residual variance for either model. Since the fitted values plots still don't show any egregious trends, we will proceed with the analysis, under the assumption of equal variance. It may be worth looking into fitting a heteroscedastic model, because the p-values for a BP-test under each model are extremely low.

Linearity: The linearity condition mainly applies to the quantitative explanatory variables in the model and their relation to the response variable. The only quantitative covariates in either model are *Log-Mileage* and *Cylinder*, so we will need to look at the added-variable plots for each of these variables under each model.

Model 1 Model 2



The linearity between *Cylinder* and *Log-Price* looks at least moderately strong for each model. The trend is more questionable for *Log-Mileage*, especially for model 2 where it is difficult to decipher any trend at all. However, there is no major curvature or clear non-linear trends present, so we will assume that linearity is sufficiently met.

<u>Independence:</u> We assume that the observations in the data are independent from one another. We are told that it is a “representative sample”, so we will assume that this condition is met for both models.

## Performance Evaluation

### *Model Fit*

As previously stated, Model 1 holds the advantage of having better overall fit than Model 2. This is evidenced by each model’s R-squared value, which is interpreted as the percentage of variation in *Log-Price* explained by the covariates in the model. Model 1 has an R-squared of .973, and Model 2 has an R-squared of .921. Both adequately explain the variation of *Log-Price*, but Model 1’s fit is slightly better.

### *Predictive Diagnostics*

In addition to comparing the fit for each model, we can compare their respective predictive accuracies. Given below, for each model, are four key measures for assessing predictive accuracy using a leave-one-out cross validation procedure.

| **Model 1** | **Model 2** |
|---|---|

| | | | |
|---|---|---|---|
| *Bias* | -41.6 | *Bias* | -113.95 |
| *RMSE* | 1,132.6 | *RMSE* | 2,035.8 |
| *Width* | 5,955 | *Width* | 9,902 |
| *Coverage* | .9565 | *Coverage* | .929 |

Model 1, which we already established as having better fit, also beats out model 2 in every predictive diagnostic given here.

- Both models are slightly biased to predicting lower prices (as shown by the negative biases), but model 2 has a stronger bias than model 1.
- Model 1's RMSE (root mean square error) is 1,133, meaning that on average, the predictions from model 1 are off by $1,133. Meanwhile, Model 2's RMSE is almost double that, at $2036. This sounds like a large amount to be off by, but the prices in the dataset range from $8,000 to $70,000, with a standard deviation of about $9,985. In comparison to this range and standard deviation, the predictions for Model 1 are actually quite accurate, on average.
- The average width of 95% prediction intervals in model 1 is just short of $6,000; in model 2, it is nearly $10,000. Model 1 is able to predict with more certainty than Model 2.
- Model 1's average coverage is just below 0.96, while Model 2's is about 0.93. Both these values are close to .95, so the prediction intervals are on target.

**Findings**

*Interaction between Make and Mileage*

In order to determine if there is a significant interaction between *Make* and *Mileage*, we will use Model 2. An interaction between *Make* and *Mileage* would imply that the change in *Log-Price* from additional mileage is different depending upon the make of the car. We ran an ANOVA F-test for all the interaction effects. The F-statistic from this analysis of variance was 1.056, with a p-value of 0.38. With a p-value this high, we cannot conclude that any of the interaction variables were significant, since including them didn't improve the model fit by much. Having established that the change in *Log-Price* from higher mileage does not differ based on the make of the car, Model 1 will be used to perform inference and prediction for the rest of the analysis.

*How does each factor affect the resale value?*

A summary of the effect estimates on *Log-Price* for model 1 (including *Log-Mileage, Body Type, Cylinder,* and *Model*) is given below, with the effects for *Model* abbreviated due to having over 30 levels.

| **Covariate** | **Symbol** | **Estimate** | **95% Confidence Interval** |
|---|---|---|---|
| *Intercept* | $\beta_0$ | 10.9 | (10.76, 11.04) |
| *Log-Mileage* | $\beta_1$ | -0.087 | (-0.094, -0.079) |
| *Type-Coupe* | $\beta_2$ | -0.19 | (-0.23, -0.16) |
| *Type-Hatchback* | $\beta_3$ | -0.18 | (-0.22, -0.14) |
| *Type-Sedan* | $\beta_4$ | -0.21 | (-0.24, -0.18) |
| *Type-Wagon* | $\beta_5$ | -0.18 | (-0.23, -0.13) |

|  |  |  |  |
|---|---|---|---|
| *Model (32 levels)* | $\beta_6 - \beta_{36}$ | Various | Various |
| *Cylinders* | $\beta_{37}$ | 0.058 | (0.038, 0.078) |

It is clear that as *Log-Mileage* increases, *Log-Price* decreases. Even when back-transforming, this still implies that cars with higher mileage sell at lower prices, on average. As for type, all of the *Body Type* coefficients are negative because the baseline level of this variable accounts for convertible cars. As might be expected, convertible cars have a tendency to be quite expensive, so all of the other body types have coefficients that are negative *in relation to* convertibles. In *Model*, the highest effect estimate belongs to the Cadillac XLR-V8 model at 0.5. Finally, we can see that cars with more cylinders in the engine tend to sell at higher prices.

It is important to understand what these “effect estimates” actually mean in context. Interpretations of these estimates (for a categorical example and a quantitative example) are as follows:

- The effect estimate for *Type-Coupe* is -0.19. This means that the *Log-Price* of coupe vehicles is, on average, 0.19 lower than convertible vehicles (convertible is the baseline level for *Type*) when holding all other variables constant. To account for uncertainty, we say we are 95% confident that this difference is somewhere between -0.23 and -0.16.
- The effect estimate for *Cylinders* is 0.058. This means that, if other variables are held constant, a vehicle’s *Log-Price* is .058 higher for each additional cylinder in its engine, on average. We are also 95% confident that this effect is between .038 and .078.

*Vehicles with the highest resale values*

It was also of interest to know if there were any particular vehicles that retained a higher resale value at 15,000 miles (and what characteristics they held). We took all the cars in the dataset and removed their *Log-Mileage* observation, meaning that we had all combinations of car models, body types, and cylinder counts in the dataset. We then ran predictions on all of these observations with *Log-Mileage* fixed at log(15000), using Model 1. The 5 most expensive vehicles are given in the table below, with their estimated resale values at 15,000 miles and corresponding 95% confidence intervals.

| **Model** | **Body Type** | **Cylinders** | **Est. Price** | **95% Conf. Interval** |
|---|---|---|---|---|
| Cadillac XLR-V8 | Convertible | 8 | $61,657 | ($59070, $64359) |
| Cadillac CST-V | Sedan | 8 | $45,158 | ($43264, $47134) |
| Chevrolet Corvette | Convertible | 8 | $43,043 | ($41575, $44563) |
| Cadillac STS-V8 | Sedan | 8 | $42,212 | ($40441, $44060) |
| Cadillac STS-V6 | Sedan | 6 | $37,429 | ($35859, $39067) |

These results are consistent with the effect estimates that we saw previously. The Cadillac XLR-V8 model had the highest effect estimate of anything in the model. It also has a convertible body (recall that all the other body types had negative estimates in relation to convertibles) and 8 cylinders (which is the highest cylinder count in the dataset). Overall, we see that Cadillacs, convertibles, and 8-cylinder vehicles tend to retain the highest resale values at 15,000 miles.

*Prediction Example*

In addition to all the inference conducted in this analysis, Model 1 can be used to predict the resale value of a car not already in the dataset. We want to predict the resale price of a Cadillac CTS Sedan with 17,000 miles and a 6-cylinder engine. A 95% prediction interval for that estimate is ($26180, $34783). We would expect this Cadillac CTS to sell for about $30,176, and we are 95% confident that it would sell for somewhere between $26,180 and $34,783.

**Conclusion**

The two linear models met both of the goals stated at the beginning of the analysis. Using Model 2, we established that the effect of a car's mileage on price does not depend on its make. Then, using Model 1, we found out that a car's model and body type play significant roles in determining the resale price. Model 1 also showed that as a car's *Log-Mileage* increased or *Cylinders* decreased, the *Log-Price* was likely to decrease as well.

Even though Model 1 has both excellent fit and strong predictive accuracy, it is not perfect. The equal variance and residual normality assumptions that were assessed earlier in the analysis were questionable at best, and failing to properly meet these assumptions could mean that our ANOVA F-test and/or confidence intervals for the effect estimates are inaccurate. For that reason, we suggest building a heteroscedastic linear model to account for the changing variance. Also, obtaining more variables from Kelley Blue Book's data to include in the models could help even out some of the heteroscedasticity and residual skewness. Other variables (not found in this dataset) that would likely help predict a car's resale price include:

- Age (in years)
- Fuel efficiency (in miles per gallon)
- The state in which it sold (states on the east coast often have better deals on used vehicles than states in the mountain west).
- Whether or not it had been in an accident (if there are two cars that share all the same characteristics but one has been in an accident and one has a clean title, the clean-title car is likely to sell for more).