Kelly Blue Book Analysis

McKay Gerratt and Carly Lundgreen

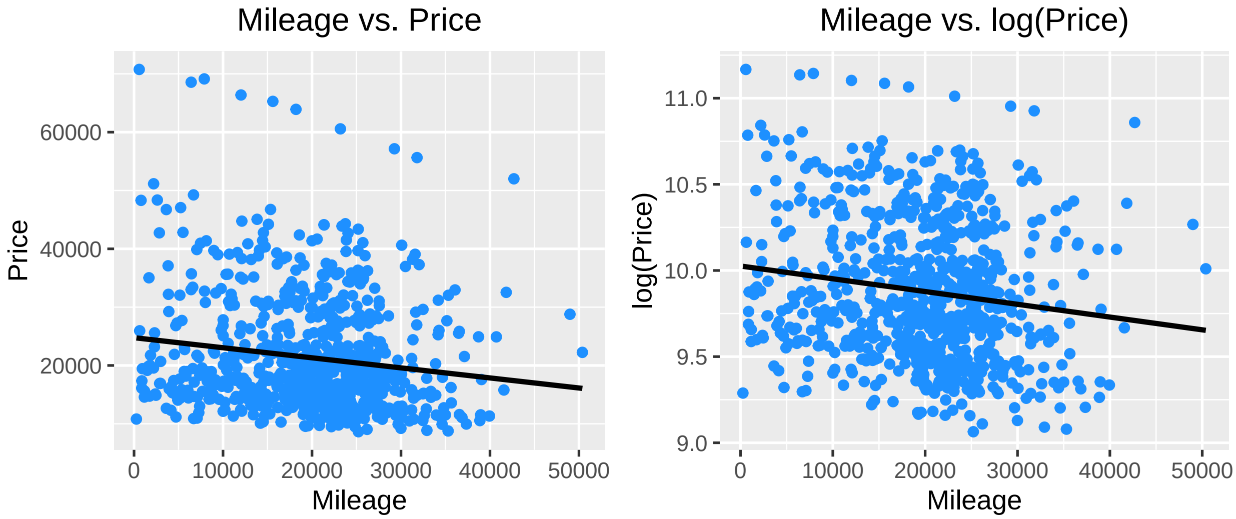
January 28, 2021

Introduction

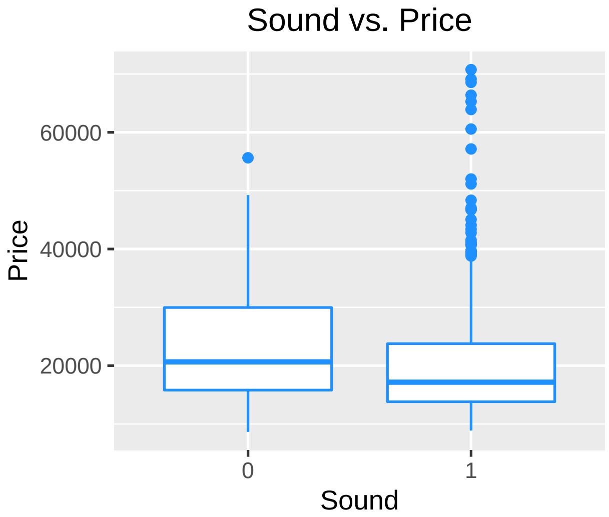
Kelley Blue Book (KBB) is a long-present resource for customers to understand resale values of used cars by helping them understand what their vehicle is worth. The values offered by KBB are based on past sale and auction prices. Geographic and market trends are also taken into account, and KBB values are updated weekly. The aim of this analysis is to understand the effects of various factors on KBB prices and to predict prices for cars with specific characteristics. Customers can then understand why their car has the resale price that KBB provides and be given a reasonable range of prices within which their car likely falls, thus allowing for more informed decisions when selling their cars (or even buying a 'new' used car).

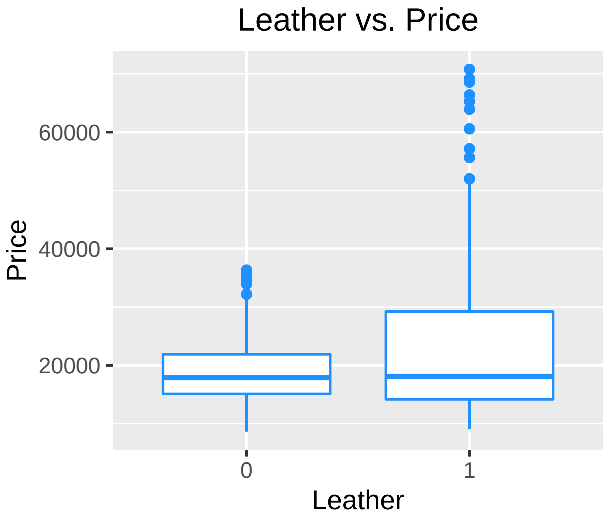
The dataset used as part of this analysis is a subset of data from Kelley Blue Book. There are 804 GM cars included in the sample. The response variable considered in this analysis is `Price`, which is the suggested retail price of a used 2005 GM car in excellent condition. Various factors such as `Mileage` (total number of driven miles), `Make`, `Model`, and `Trim` (specific type of model) of the vehicle, the `Type` of vehicle (i.e. sedan, coupe), number of cylinders (`Cylinder`), number of liters in the car's engine (`Liter`), the number of doors on the car (`Doors`), and whether the car has cruise control (`Cruise`), leather seats (`Leather`), or upgraded speakers (`Sound`).

The scatterplots below illustrate the relationship between mileage and price. On the left is a plot of price against mileage, and it is clear that there is a level of right-skewness (i.e. large spread among higher prices) that may lead to problems when fitting a linear model. In particular, the plot on the left looks like it might be more difficult to justify a linear relationship between mileage and price. Other multiple linear regression (the details of which are discussed in following sections) assumptions may be invalidated with the difference in variance (spread) between low and high resale prices. Invalidation of these modeling assumptions would invalidate the results of inference procedures such as confidence intervals and hypothesis tests that we will conduct as part of this analysis. The log-transformation of price leads to the reduction of that difference in spread between low and high prices. Modeling assumptions will be discussed in the following section, but the log-transformation of price does significantly improve the validity of these assumptions.

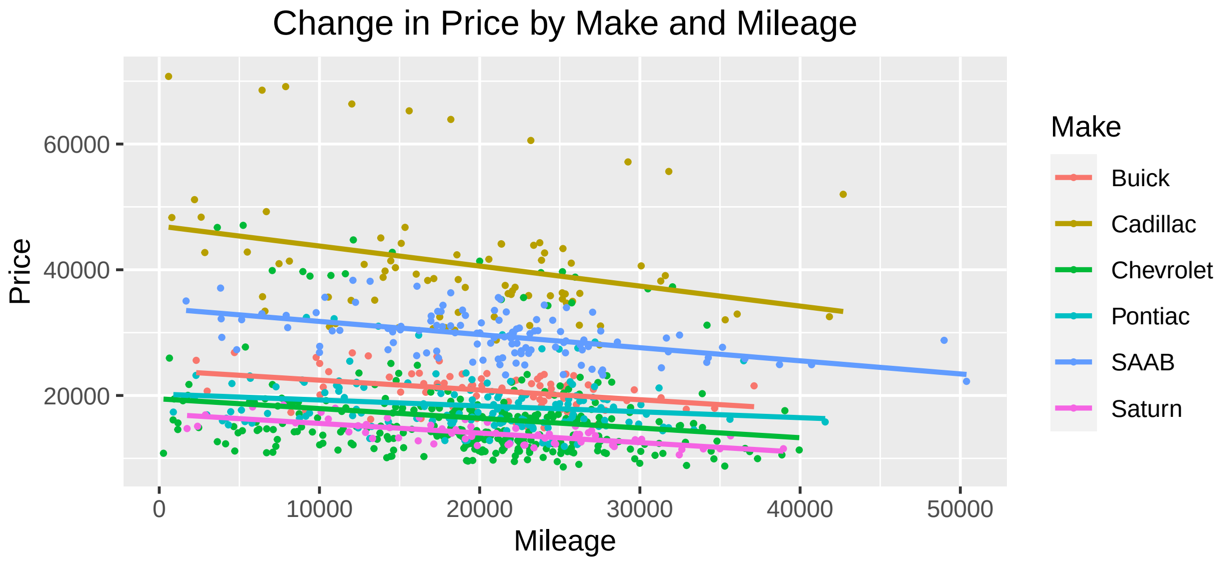


A couple of other plots of interest comparing prices for cars with and without upgraded speakers, as well as with and without leather seats, are shown below. Note that a '0' indicates a car without updated speakers or without leather seats, and a ‘1’ indicates the opposite. We can see that cars with upgraded speakers appear to have more of a right-skewness (larger spread among higher prices) compared to cars without upgraded speakers, though cars with updated speakers have a lower median price. Similarly, prices of cars without leather seats have smaller variance than cars with leather seats, with median prices that are approximately the same.





Finally, the plot below illustrates the relationship between the mileage and make of cars. We can see that Cadillacs certainly have, on average, higher resale values than other makes included in this dataset. Similarly, it does not appear that the effect of mileage on price changes very significantly from make to make. This will be discussed further in the Results section.



Methodology

In order to answer the questions of interest, we will fit a multiple linear regression (hereafter referred to as MLR) model to the data. An MLR model’s strength lies in its interpretability; unlike several other methods, an MLR model can be used to make inference on relationships included in the analysis. However, MLR models rely on assumptions, many of which may be unrealistic.

After performing best subset selection and transforming variables in order to meet model assumptions, the following model was fit to the data:

$$log(y_i) = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{i34}\beta_{34} + x_{i35}\beta_{35} + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

y_i represents the price of a car given its odometer, interior leather, sound system, and model.

x_i is a row of data representing the i^{th} car's characteristics (for example, x_{12} holds the information about the 1st car's seat material).

β_0 represents the expected log(price) of a SAAB 9-2X AWD (the baseline category against which all $\beta's$ are compared) that has not been driven and does not include both leather seats and an upgraded sound system.

β_1 represents the expected change in log(price) of a vehicle given a mile increase in mileage holding all else constant.

β_2 represents the expected change in log(price) from a car without leather seats to the same car with leather seats holding all else constant.

β_3 represents the expected change in log(price) from a not upgraded sound system to an upgraded sound system holding all else constant.

$\beta_4 \cdots \beta_{35}$ represent the expected change in log(price) from a SAAB 9-2X AWD to another model of car, each associated with a specific coefficient, holding all else constant.

ϵ_i represents the difference from the actual value of the i^{th} car's log(price) to that predicted by the model.

σ represents how much we expect the model to differ from the actual log(price) of any particular car on average.

For interpretation purposes, the expected change in log(price) is equivalent to the expected percent change in resale price. In order to fit this model to the data, four assumptions must hold. The data must be linear, the observations must be independent from one another, the residuals must follow a normal distribution, and the variance must be equal across all observations.

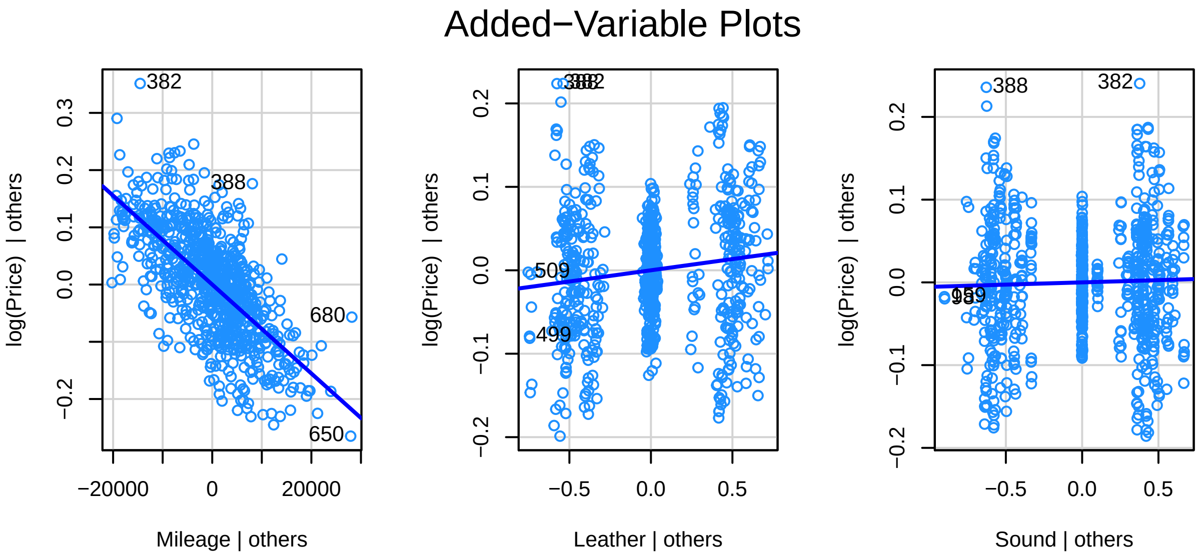
Model Assumptions and Evaluation

The variables included in the model were selected based off of a best subset selection method using the Bayesian Information Criterion, or BIC. BIC penalizes a higher number of variables in the model, thus it is preferred to make inference, which is a main goal of this analysis. Trim was omitted in order to avoid problems in fitting the model due to singularity, as well as to maintain simplicity. We still believe that the above model has high predictive power and that the variables included account for much of the variation in price.

The following paragraphs discuss the validity of assumptions made in order to fit the model.

Linearity

The following graphic shows added variable plots for mileage, leather, and sound.



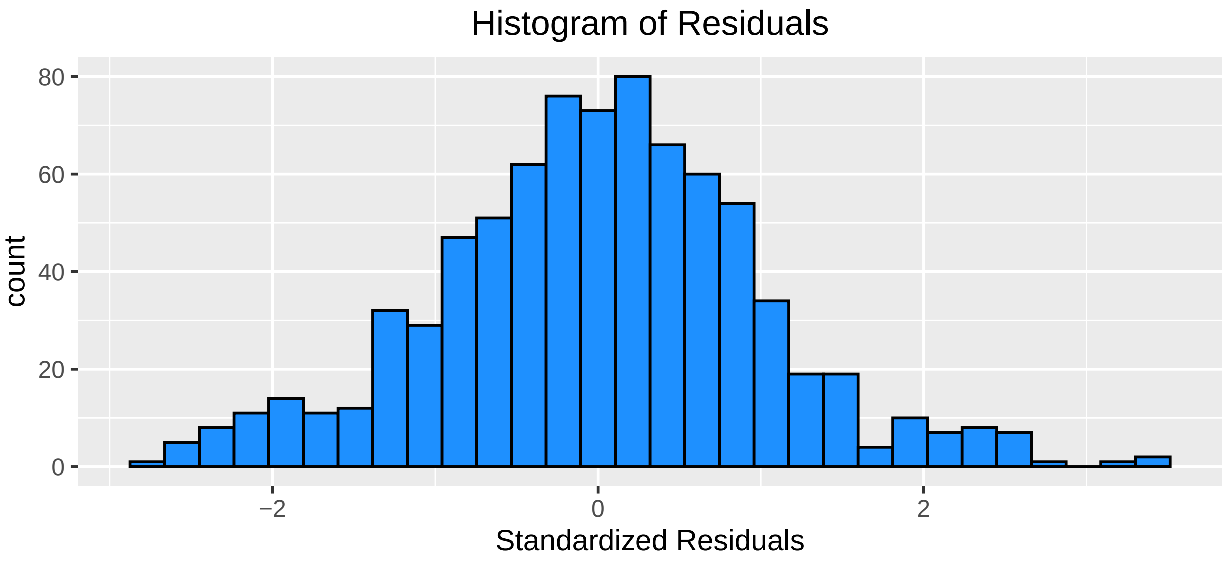
There are no patterns that would indicate that this assumption is violated.

Independence

It is reasonable to assume that the characteristics of a car in the dataset will not influence any other.

Normality

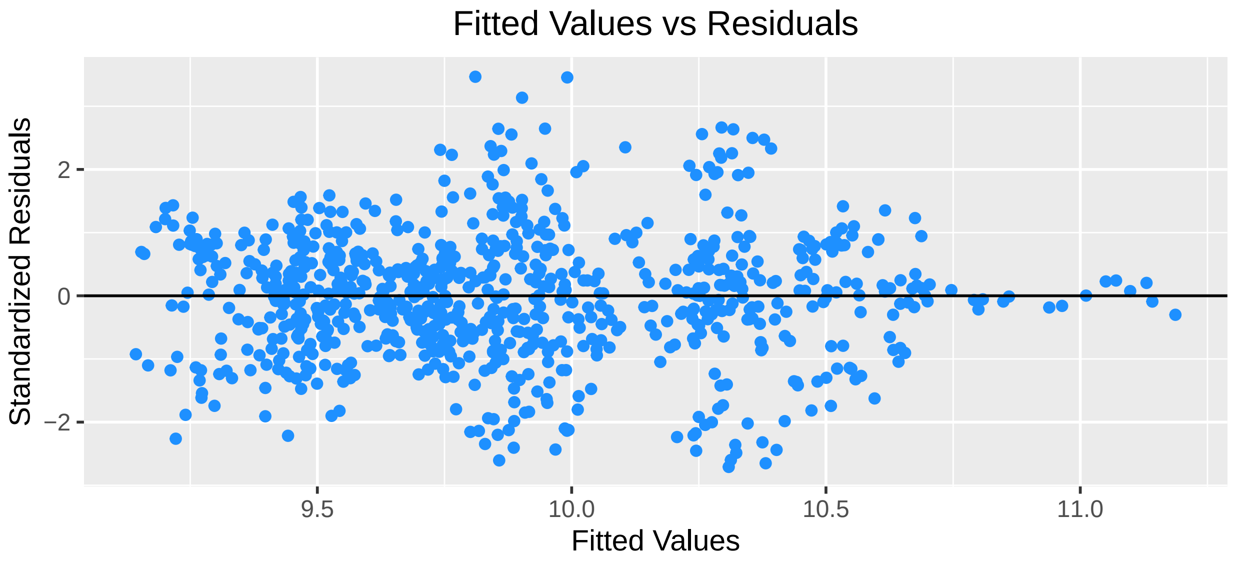
The following graphic depicts a histogram of the standardized residuals (represented by ϵ) of the model



The tails of the histogram do seem to extend a little farther than expected; however, the overall shape is similar to that of a normal. We will continue as if the assumption was met.

Equal Variance

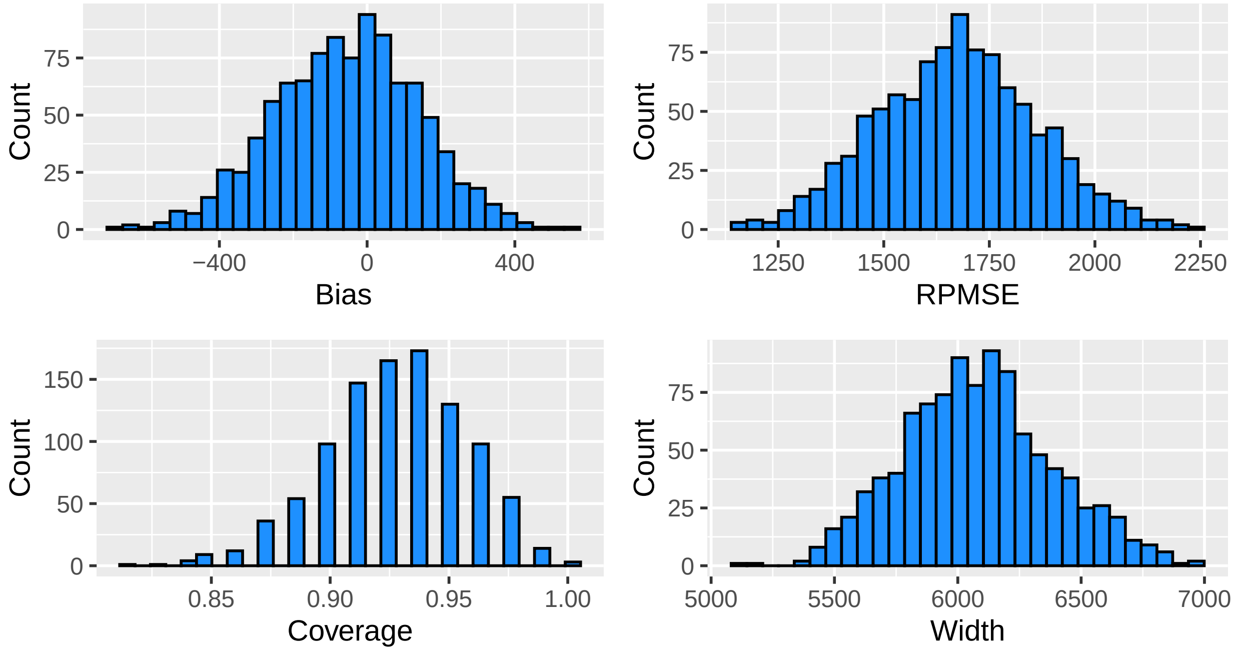
The following graphic depicts a fitted value versus residual plot, which can be used to identify trends in variance.



There is some cause for concern on the right side of the plot, as the few observations clustered there are lie closer to the center than the remainder of the observations. This would indicate that the observations do not have the same variance (as indicated in the model above by σ^2). However, we will continue with the analysis as if the assumption was met.

The model's R^2 value is 0.9718, which indicates that a car's mileage, leather seats, sound system, and model account for 97.18% of the variation in it's log(price), which is excellent.

The model was cross validated using 1,000 iterations on a test set of 10% of the data. The following graphic displays the results:



The model had an average bias of -$54.60, meaning that the model underpredicts the price of the car on average. The root predictive mean square error of $1,674.20 indicates that the model is on average off by about $1,600.00. Given that the car prices range from about $8,000 to $70,000, the model appears to be fairly accurate. The average coverage by prediction intervals was .923, which is a little lower than the expected coverage of .95, but not extremely worrisome. The average interval width of $6,080.90 is fair given the range of prices as stated previously.

Results

Of the variables that were included in our chosen model (Mileage, Model, Leather, and Sound), we found that Sound–i.e. whether the car has upgraded speakers– did not lead to significantly different resale values. There were 32 different vehicle models included in this dataset, and out of those 32 models, we found that all but the Park Avenue model led to significantly different vehicle resale amounts. The two tables below illustrate which factors lead to significantly higher and lower resale values.

Our model included an intercept term which (untransformed) represents the log(Price), or percent change in resale value, over a SAAB 9-2X AWD model with no upgraded speakers, no leather seats, and zero mileage. This particular model of car fell near the midrange of resale prices; cars with lower prices may be selling for prices lower than average, while the opposite is true for cars with higher prices.

As expected, the mileage leads to significantly lower log(resale) values on average. Based on the table of factors leading to lower resale amounts, we can see that every one-mile increase in mileage leads to between a 0.00072% and 0.00084% decrease in average resale value if all other factors are held constant. Note that the values in the tables below are still on the log scale, to allow for interpretation of these effects as a percent change in resale price. Other factors leading to lower resale values are vehicle models such as AVEO, Grand Am, G6, and Impala. In the case of the AVEO model, an AVEO vehicle has, on average, somewhere between a 75% and 89% lower resale price than a 9-2X AWD vehicle (again holding all other factors constant).

Select models such as the Corvette and Deville, as well as the presence of leather seats, lead to significantly higher resale values on average. Looking at the table of higher resale values, a vehicle with leather seats has, on average, between a 1.3% and 4.04% higher resale value than a car without leather seats, holding the other factors constant. Similarly, a Corvette has between a 36.6% and 51.8% higher resale value, on average, compared to a SAAB 9-2X AWD model vehicle, still holding other factors constant. These results are not unexpected, as the Corvette is generally a high-end sport vehicle that likely costs more than the average vehicle, so we can expect that to lead to an increase in resale value when compared with a non-sports car model.

Table 1: These factors lead to significantly lower resale values (log scale)

	Coefs	2.5%	97.5%
Mileage	-0.0000078	-0.0000084	-0.0000072
ModelAVEO	-0.823	-0.895	-0.752
ModelBonneville	-0.166	-0.24	-0.092
ModelCavalier	-0.66	-0.732	-0.588
ModelCentury	-0.433	-0.515	-0.351
ModelClassic	-0.598	-0.68	-0.515
ModelCobalt	-0.583	-0.656	-0.511
ModelG6	-0.223	-0.299	-0.147
ModelGrand Am	-0.522	-0.598	-0.446
ModelGrand Prix	-0.302	-0.376	-0.229
ModelImpala	-0.213	-0.287	-0.139
ModelIon	-0.591	-0.663	-0.519
ModelL Series	-0.412	-0.494	-0.33
ModelLacrosse	-0.138	-0.212	-0.065
ModelLesabre	-0.206	-0.282	-0.13
ModelMalibu	-0.376	-0.448	-0.304
ModelPark Avenue	-0.025	-0.101	0.05
ModelSunfire	-0.669	-0.751	-0.587
ModelVibe	-0.431	-0.505	-0.358

Table 2: These factors lead to significantly higher resale values (log scale)

	Coefs	2.5%	97.5%
(Intercept)	10.241	10.170	10.311
Leather	0.027	0.013	0.040
Model9_3	0.182	0.106	0.257
Model9_3 HO	0.213	0.140	0.286
Model9_5	0.199	0.125	0.273
Model9_5 HO	0.145	0.069	0.221
ModelCorvette	0.443	0.367	0.519
ModelCST-V	0.593	0.511	0.675
ModelCTS	0.180	0.097	0.262
ModelDeville	0.378	0.304	0.452
ModelGTO	0.169	0.086	0.251
ModelSTS-V6	0.401	0.319	0.483
ModelSTS-V8	0.538	0.456	0.620
ModelXLR-V8	0.918	0.836	1.001

This dataset only included a subset of potential factors that may help explain changes in resale value. Other factors, such as the age of the vehicle, the transmission type (automatic/manual), the nationality of the vehicle, or the number of previous owners may be important to consider when modeling used car resale prices. For example, a car that is twenty years old is likely to have a lower resale price than a car manufactured last year. However, vehicle age may be highly collinear with mileage, since an older car will naturally have more miles (assuming the car is regularly driven). These would be interesting points of further analysis of this dataset, if further covariate information was accessible.

As discussed previously, an increase in mileage will generally lead to a decrease in resale price. When we performed a best subset variable selection algorithm using BIC to select a model, the make of the car was found to be unimportant in the explanation of resale price. Because of this, we inferred that the amount of decrease in value from additional mileage does not differ depending upon the make of the vehicle. To confirm this, we ran an ANOVA test comparing our fitted model (the "reduced model") to a "full model" that included the main effects from our chosen model in addition to an interaction between make and mileage. The p-value from this test was approximately 0.1, so we concluded that indeed the interaction between make and mileage was not significant.

We also found that the Cadillac XLR-V8 Hardtop Convertible (with 8 cylinders, a 4.6 liter engine, 2 doors, cruise control, upgraded speakers, and leather seats) has the highest resale value, assuming the mileage of all vehicles is 15,000. The expected resale price of this vehicle is found to be $64,536.89 with a 95% prediction interval ($55,823.76, $74,609.98).

Finally, a reasonable range of resale values for a Cadillac CTS 4D Sedan with 17,000 miles, 6 cylinders,a 2.8 liter engine, cruise control, upgraded speakers and leather seats is ($26,251.36, $35,084.40).

Conclusion

By fitting a multiple linear regression model that includes the vehicle model, mileage, the presence of leather seats, and whether the speakers have been upgraded, we were able to identify factors that led to higher and lower resale values. The mathematical model utilized log(Price) as the response variable, but all predictions were transformed out of the log-scale. Also, the adage that a car loses value the minute it is driven off the dealer's lot seems on par with our results regarding mileage–that an increase in mileage will generally lead to a decrease in a used car's resale value. In addition to these inferential insights, this model also aided in predicting the resale price of a particular vehicle–the Cadillac CTS 4D Sedan with certain specs such as leather seats and cruise control.

Our model does have reasonably good predictive power, as discussed in the performance evaluation section with metrics such as bias, RPMSE, coverage, and prediction interval width. For example, the model was found to underpredict each price by about $54.60. Given that the range of the resale prices in this dataset is about $62,000, this bias is excellent and reflects well on the model's predictive power. The predictions were found to be off by just over $1,600, on average, which again is small in comparison to the range of prices seen in this dataset.

Though the predictive power of the model was decent, however, there are still shortcomings that may require further attention. The main issue arose in the assumption of heteroskedasticity among the residuals. The log transformation fixed most of the issues with unequal variance along the regression line, but not all of them. For example, very high resale prices (around $65,000) appear to have a smaller variance than prices closer to the average. This slight departure from the 'equal variance condition' is a concern about the validity of this model, so any recommendations coming from this analysis should keep that in mind. Another shortcoming of the analysis may be the relatively few covariates. The R^2 value of our model was very high, about 0.97, so we seem to be doing fine with the data we have, but it might still be of interest to consider other covariates that may explain used car resale prices further. This may further decrease the predictive error and/or the widths/coverage of prediction intervals calculated using an MLR model.

Teamwork:

Carly wrote the introduction, conclusion, and results sections. McKay wrote the methodology and model evaluation sections, and we both collaborated on the code together in the lab.