KBB Homework

Abe Durrant

September 20, 2021

Abstract

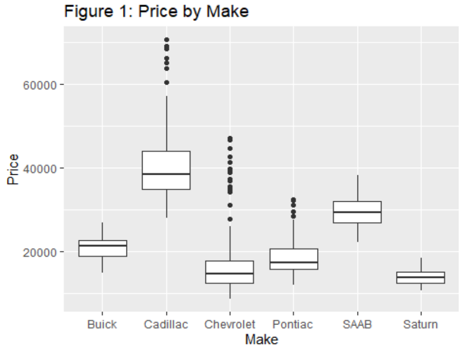
A subset of a Kelley Blue Book dataset was used to try to find insights into the pricing of used GMC cars. The model and trim of a car are what are most important in estimating a car's KBB value. The mileage, seat type and sound system all have an impact as well. Our linear regression model can use the stated variables to predict the price at which a GMC car will resale.

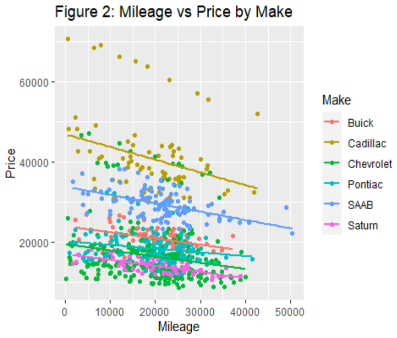
1 Exploratory Data Analysis

The following data and analysis come from a subset of data from Kelley Blue Book which is often used to estimate the value of a car at a given time. An analysis of this data can help to give consumers a better idea of why their car has the value that KBB lists as well as what a reasonable range for price might be on their specific car.

The goals of this analysis are to find what features of a car lead to a higher or lower price, to find whether the make of the car changes how much mileage affects the price, and to develop a model that can predict the price of different cars based on their features.

To begin the analysis a simple look at plots of how the price is related to other variables can be helpful. The make and model of the car both seem to have a relationship with the price of the car as seen in Figure 1 (only the make is pictured due to space). A look at mileage shows that it has a weak negative correlation with price (-.143), but it appears that the effect it has on price could be different based on the model (Figure 2). The lines have similar slopes but could be different. Later the effect between the two will be analyzed to see if it is significant.





A linear model will be used to analyze this dataset. However, there are a few potential issues. The price of the car could potentially have a different variance (measure of the spread of the data) as it gets into the more expensive range of cars. If we don't adjust this, then it will through off a lot of our error estimates and our intervals for our predictions won't be as accurate as we believe them to.

There is also a potential issue if we don't account for an interaction between some of the features. For example, if mileage is different dependent upon the make of the car and we don't account for this then our estimates will not be as accurate as they could be.

2 Models Used

In order to accomplish the goals of the analysis, two separate linear regression models were used. One model was optimized to fit the data best and for the most accurate predictions (Model 1). Another was used to examine whether or not the effect of mileage on a car's price is dependent on its make (Model 2).

Model 1 which was used to best fit the data is listed in the equation below:

$$\sqrt{price}_i = \beta_0 + \beta_{1:Model9-3_i} + \beta_{2:Model9-3HO_i} + \cdots \beta_{31:ModelXLR-V8_i} + \beta_{32:Mileage_i} + \beta_{33:TrimAeroSedan4D_i} \cdots + \beta_{70:TrimSVMHatchback4D_i} + \beta_{71:Leather_i} + \beta_{71:Sound_i} + \epsilon_i$$

The (…) represents the rest of the different model and trim types that are used as parameters but not included due to space. Epsilon is the error term for the difference between our actual value and the square root of price. B_0 represents the intercept of our model or where we expect price to be when all other parameters are 0. The other β's are representing our different parameters such as the type of model or the mileage of the car. When the type of the model is what the β has listed then that coefficient will be multiplied by 1. The square root is taken of price as price tended to have higher variance when the values were higher, and the square root helps us fit this assumption better.

Our linear regression model is a good choice for this problem as it helps us in discovering what parameters have the biggest effect on the KBB value of a car as well as what those values are. It is highly interpretable and can also provide us with strong predictions when it fits the data well. However, it is limited if the assumptions aren't met and can be susceptible to overfit when not done correctly.

Model 2 is similar to that of Model 1, however rather than use a coefficient for each car model, there will be one for each make of car as well as a term for the interaction between mileage and each make of car.

Both of these models operate under the same assumptions. We are assuming that the data follows a linear relationship, that the observations are independent of one another, that the distribution of price around our estimates is normally distributed, and that the variance is equal across the dataset. These assumptions will be justified in the next section.

3 Model Justification and Performance Evaluation

The variables in the Model 1 were selected by using a best subsets selection method using the R package OLS. This method examines all the combinations of the variables and returns the combination that best meets the criteria set. The model chosen for Model 1 met the criteria best as far as a fit of the data and for use in inference (AIC and adjusted R-squared values were used).

Model 2 was chosen using the same algorithm but inserting Make into the model instead of the car's model. An interaction effect was also put into the model between Make and Mileage. This was done to see if there was an effect between the two.

Both models met the assumption of linearity. When looking at the added variable plots for the models it appeared that all relationships were linear. The assumption of independence was met as we have no information that would lead us to believe that the observations weren't independent of one another. Our assumption of normality is met by looking at the histogram of standard residuals (Figure 3). In this figure we can see that they are approximately normally distributed. There was one observation that looked like a potential outlier but when tested using a metric called Cook's Distance it didn't qualify as an outlier.

Figure 3: Histogram of the Standardized Residual

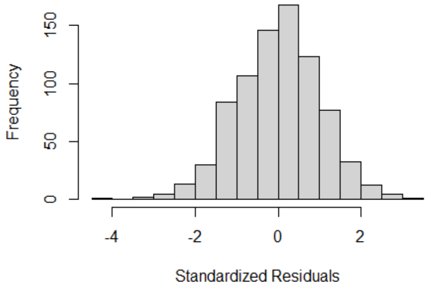
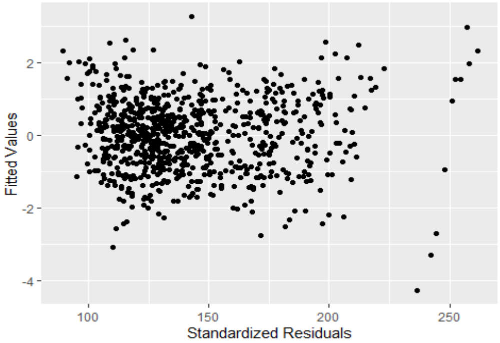


Figure 4: Fitted Values vs Standardized Residuals



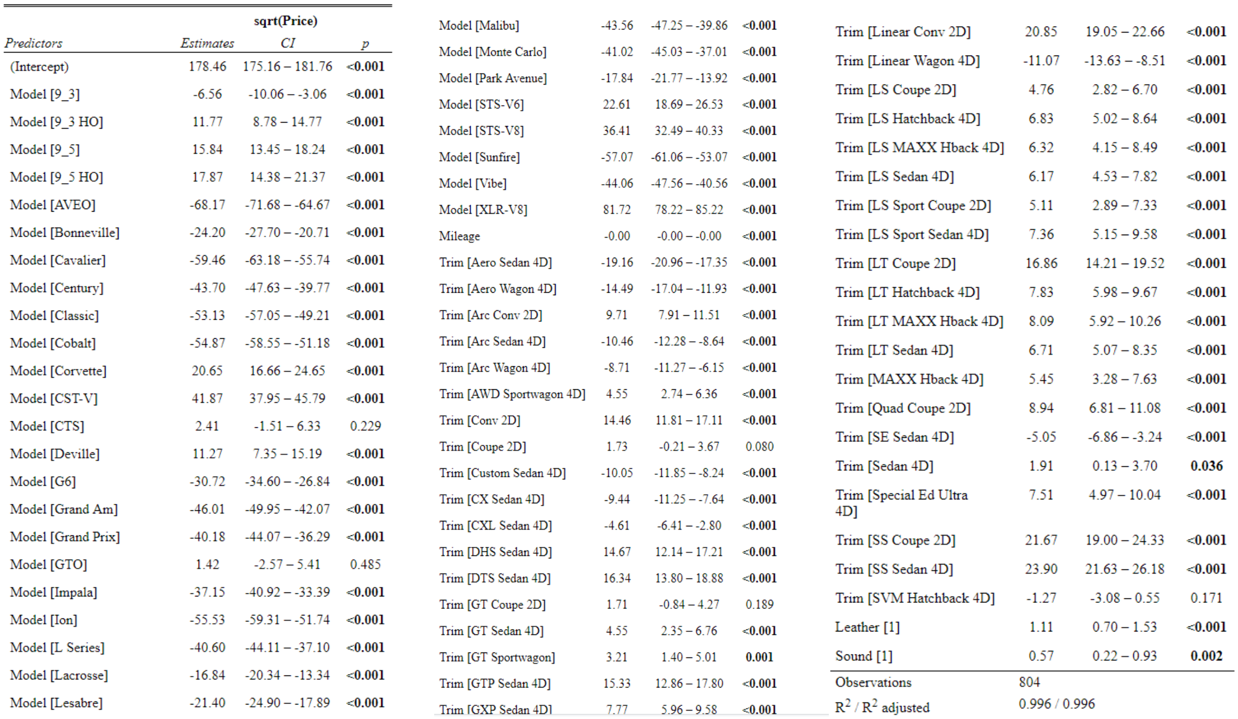
Finally, our last assumption was of equal variance. To meet this assumption the fitted values vs. residuals plot was examined to see if there was any difference in the variance across the fitted values (Figure 4). The square root of price was taken as the original plot had higher variance among the larger fitted values.

Both models met all of the assumptions (only Model 1 plots are included). Model 1 will be used to answer most of the questions we had originally so we will examine its fit here. The R-Squared value for Model one came out to be .996. This is a very high R-squared value and means that we are fitting the data very well. In fact, about 99.6% of the variation across our data is explained by our model.

To see how well Model 1 predicts, 100 cross validation studies were run where the data was split into a training set (to train the model) and a test set. After running these studies, the model had an average bias of 4.52 and an average RPMSE of 547.43. These are both very good numbers, the bias measures how much we were off on our predictions on average while the RPMSE measures the average distance each prediction was off. This means that overall, we weren't only slightly too high on average on predictions, and we were generally within $547.43 on our price prediction. Compared to the data's standard deviation of $9,884.853 this is very good.

4 Results

Estimates of the model parameters are given in Table 1 below. Here we can see what factors lead to higher or lower resale values based on the estimate for their coefficient:



	sqrt(Price)		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	178.46	175.16 – 181.76	<**0.001**
Model [9_3]	-6.56	-10.06 – -3.06	<**0.001**
Model [9_3 HO]	11.77	8.78 – 14.77	<**0.001**
Model [9_5]	15.84	13.45 – 18.24	<**0.001**
Model [9_5 HO]	17.87	14.38 – 21.37	<**0.001**
Model [AVEO]	-68.17	-71.68 – -64.67	<**0.001**
Model [Bonneville]	-24.20	-27.70 – -20.71	<**0.001**
Model [Cavalier]	-59.46	-63.18 – -55.74	<**0.001**
Model [Century]	-43.70	-47.63 – -39.77	<**0.001**
Model [Classic]	-53.13	-57.05 – -49.21	<**0.001**
Model [Cobalt]	-54.87	-58.55 – -51.18	<**0.001**
Model [Corvette]	20.65	16.66 – 24.65	<**0.001**
Model [CST-V]	41.87	37.95 – 45.79	<**0.001**
Model [CTS]	2.41	-1.51 – 6.33	0.229
Model [Deville]	11.27	7.35 – 15.19	<**0.001**
Model [G6]	-30.72	-34.60 – -26.84	<**0.001**
Model [Grand Am]	-46.01	-49.95 – -42.07	<**0.001**
Model [Grand Prix]	-40.18	-44.07 – -36.29	<**0.001**
Model [GTO]	1.42	-2.57 – 5.41	0.485
Model [Impala]	-37.15	-40.92 – -33.39	<**0.001**
Model [Ion]	-55.53	-59.31 – -51.74	<**0.001**
Model [L Series]	-40.60	-44.11 – -37.10	<**0.001**
Model [Lacrosse]	-16.84	-20.34 – -13.34	<**0.001**
Model [Lesabre]	-21.40	-24.90 – -17.89	<**0.001**
Model [Malibu]	-43.56	-47.25 – -39.86	<**0.001**
Model [Monte Carlo]	-41.02	-45.03 – -37.01	<**0.001**
Model [Park Avenue]	-17.84	-21.77 – -13.92	<**0.001**
Model [STS-V6]	22.61	18.69 – 26.53	<**0.001**
Model [STS-V8]	36.41	32.49 – 40.33	<**0.001**
Model [Sunfire]	-57.07	-61.06 – -53.07	<**0.001**
Model [Vibe]	-44.06	-47.56 – -40.56	<**0.001**
Model [XLR-V8]	81.72	78.22 – 85.22	<**0.001**
Mileage	-0.00	-0.00 – -0.00	<**0.001**
Trim [Aero Sedan 4D]	-19.16	-20.96 – -17.35	<**0.001**
Trim [Aero Wagon 4D]	-14.49	-17.04 – -11.93	<**0.001**
Trim [Arc Conv 2D]	9.71	7.91 – 11.51	<**0.001**
Trim [Arc Sedan 4D]	-10.46	-12.28 – -8.64	<**0.001**
Trim [Arc Wagon 4D]	-8.71	-11.27 – -6.15	<**0.001**
Trim [AWD Sportwagon 4D]	4.55	2.74 – 6.36	<**0.001**
Trim [Conv 2D]	14.46	11.81 – 17.11	<**0.001**
Trim [Coupe 2D]	1.73	-0.21 – 3.67	0.080
Trim [Custom Sedan 4D]	-10.05	-11.85 – -8.24	<**0.001**
Trim [CX Sedan 4D]	-9.44	-11.25 – -7.64	<**0.001**
Trim [CXL Sedan 4D]	-4.61	-6.41 – -2.80	<**0.001**
Trim [DHS Sedan 4D]	14.67	12.14 – 17.21	<**0.001**
Trim [DTS Sedan 4D]	16.34	13.80 – 18.88	<**0.001**
Trim [GT Coupe 2D]	1.71	-0.84 – 4.27	0.189
Trim [GT Sedan 4D]	4.55	2.35 – 6.76	<**0.001**
Trim [GT Sportwagon]	3.21	1.40 – 5.01	**0.001**
Trim [GTP Sedan 4D]	15.33	12.86 – 17.80	<**0.001**
Trim [GXP Sedan 4D]	7.77	5.96 – 9.58	<**0.001**
Trim [Linear Conv 2D]	20.85	19.05 – 22.66	<**0.001**
Trim [Linear Wagon 4D]	-11.07	-13.63 – -8.51	<**0.001**
Trim [LS Coupe 2D]	4.76	2.82 – 6.70	<**0.001**
Trim [LS Hatchback 4D]	6.83	5.02 – 8.64	<**0.001**
Trim [LS MAXX Hback 4D]	6.32	4.15 – 8.49	<**0.001**
Trim [LS Sedan 4D]	6.17	4.53 – 7.82	<**0.001**
Trim [LS Sport Coupe 2D]	5.11	2.89 – 7.33	<**0.001**
Trim [LS Sport Sedan 4D]	7.36	5.15 – 9.58	<**0.001**
Trim [LT Coupe 2D]	16.86	14.21 – 19.52	<**0.001**
Trim [LT Hatchback 4D]	7.83	5.98 – 9.67	<**0.001**
Trim [LT MAXX Hback 4D]	8.09	5.92 – 10.26	<**0.001**
Trim [LT Sedan 4D]	6.71	5.07 – 8.35	<**0.001**
Trim [MAXX Hback 4D]	5.45	3.28 – 7.63	<**0.001**
Trim [Quad Coupe 2D]	8.94	6.81 – 11.08	<**0.001**
Trim [SE Sedan 4D]	-5.05	-6.86 – -3.24	<**0.001**
Trim [Sedan 4D]	1.91	0.13 – 3.70	**0.036**
Trim [Special Ed Ultra 4D]	7.51	4.97 – 10.04	<**0.001**
Trim [SS Coupe 2D]	21.67	19.00 – 24.33	<**0.001**
Trim [SS Sedan 4D]	23.90	21.63 – 26.18	<**0.001**
Trim [SVM Hatchback 4D]	-1.27	-3.08 – 0.55	0.171
Leather [1]	1.11	0.70 – 1.53	<**0.001**
Sound [1]	0.57	0.22 – 0.93	**0.002**
Observations	804		
R^2 / R^2 adjusted	0.996 / 0.996		

From this table we can see that many of the model types as well as trim types have the highest effect on price. Some of these include the model being an XLR-V8, STS-V8, or CST-V or the trim type being Linear Conv 2D or LT Coupe 2D. The coefficients for these are interpreted as what we expect the square root of price to go up by when they increase by 1. To get our actual prediction we would have to eventually square the result. As an example of what an estimate represents, when a car is a model of XLR-V8 we would expect the square root of its price to go up by 81.72 on average from our baseline. Our model baseline is the 9-2X AWD. There is also a 95% confidence interval included next to each coefficient which is an interval estimate where we are 95% confident that the increase in the square root of price will go up by on average when our coefficient increases by 1 in the case of mileage or is under the category in the case of the others. This provides us with a degree of uncertainty in our estimates.

While it is possible that there are other factors outside of this dataset that explain how much a car is worth, they are most likely very minor indicators. Our model explained approximately 99.6% of the variation in how much a car is worth which is a very high percent. Another variable that might explain resale value that isn't included in the dataset is whether or not the car has a clean title. It is very possible that all cars in this dataset have clean titles and thus our analysis is only indicative of cars with a clean title.

Interactions

When examining Model 2, we see that the decrease in value from additional mileage is not affected by the make of the car. When including the make in our model as well as an interaction term between make and mileage our R-squared value actually goes down. In addition to this when looking at the t-tests for our parameter estimates of our interaction term, we see that none of the interactions between any make and mileage are significant.

Of all the cars included in the data set, the one with the highest resale value at 15,000 miles is the Cadillac XLR-V8 Hardtop Convertible with 2 doors, 8-cylinder, cruise control, sound system, and leather seats. It had an estimate price of $63,954 when given 15,000 miles. This is 7,000 below its resale value at its original listed mileage of 583 miles.

Model 1 can also be used to predict the value of a car with certain characteristics. As an example, the prediction for the price of a Cadillac CTS 4D Sedan with 17,000 miles, 6-cylinder, 2.8-liter engine, cruise control, upgraded speakers and leather seats is ($29,045.91, $31,675.49). This interval is a prediction interval from Model 1 and is interpreted as us being 95% confident that a car with those characteristics would sell somewhere in that price range.

5 Conclusion

Overall, we were able to achieve the goals we had at the beginning of the analysis. We were able to discover what factors lead to a higher price in cars and how much of an effect they have. We were also able to look at how mileage didn't differ depending upon the make of the car. Finally, we were able to come up with a good prediction for a specific car's priced based upon its characteristics.

Our model could be improved in a few ways. Some of the trim types and makes and models of the car were unique and thus lead to redundancy in some parameters. We could try combining all these into one variable and this could lead to some better estimates. However, this would make interpretation more difficult. Having our price square rooted in our model also leads to harder interpretations of the coefficients. This could be changed if we are willing to have less trust in our interval estimates.

The next steps that one could take from this analysis are to look at other factors outside of this dataset that may have an effect. This is just a subset of the KBB data and only covers GMC cars. It also may be from only one particular time, so estimates may be off if the market for cars changes. An analysis involving the time and where the market is at could return better predictions for the resale value of a car.