HW1 Report

RJ Cass

Abstract

Kelley Blue Book (KBB) has an extensive dataset covering car conditions and their sale price. We created a linear model of the Square Root of Price to predict a resonable resale value for a car. We found that the factors that impact Price are Mileage, Make, Model, Trim, Sound, and Leather. We also found that the Make impacts how quickly a car loses value as mileage increases, with Chevrolets and Saturns best retaining their value. We also showed that a Cadillac XLR-V8 with upgraded sound and leather interior has the best value at 15k miles. Finally, we showed that the car parameters given in the research problem, a resonable resale range is $34522.6 to $35516.23.

1: Introduction

The Kelley Blue Book (KBB) dataset is intended to help consumers know what is a reasonable sale price for a car given its current condition. We want to use this data to 1) Understand which variables are most important in determining resale value 2) Consider what factors might not be included in this dataset which could contribute 3) Identify if their is any interaction between car make and mileage on determining the sale price 4) Create a model to identify which factors give the highest resale value for a car at 15k miles 5) Predict price range for a car with given values

As shown in Figure 1 - Left, price does not appear to be normally distributed. As such we will need to perform a transformation on the Price data so we can meet the assumptions for our model. If we do not perform this transformation, the resulting model will not properly explain the relationship between the covariates and the output.

Looking at how price trends based on mileage (Figure 1 - Right), price does trend downwards as mileage increases. Furthermore, examining the general decrease of price by car make, it appears there is an interaction between car make and mileage (it seems that the decrease of price due to mileage for Cadillac cars is greater than other makes). If we do not account for this interaction effect, our prediction model will not provide accurate values for price range.

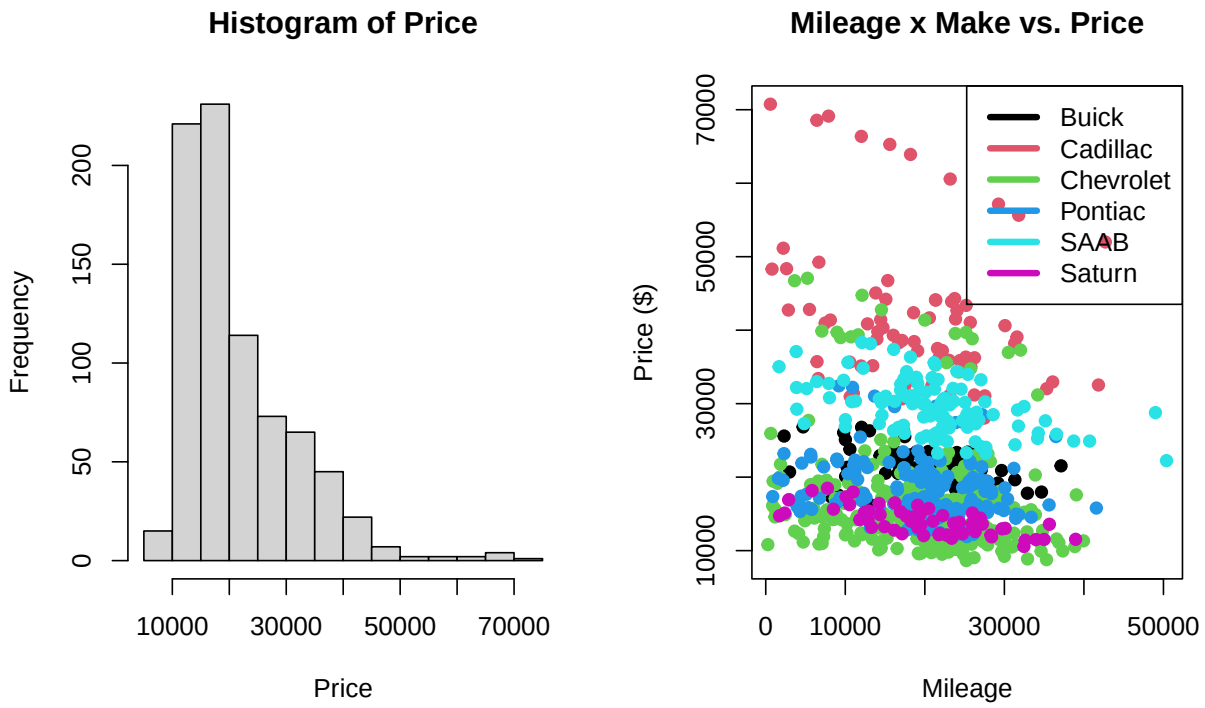


Figure 1: Left - Histogram of Price. Note that this is a right-skewed distribution, as does not meet normality assumptions. Right - Mileage versus Price with different colors for each car make. We can see that price decreases as mileage increases. Note that the rate of decrease for the red dots (Cadillac) appears to be larger.

2: Methodology

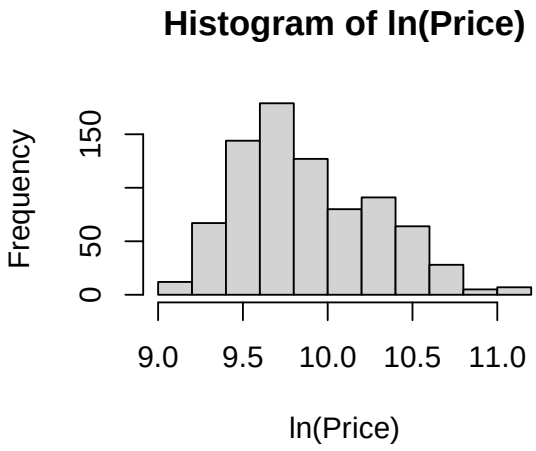
2.1 Proposed Models

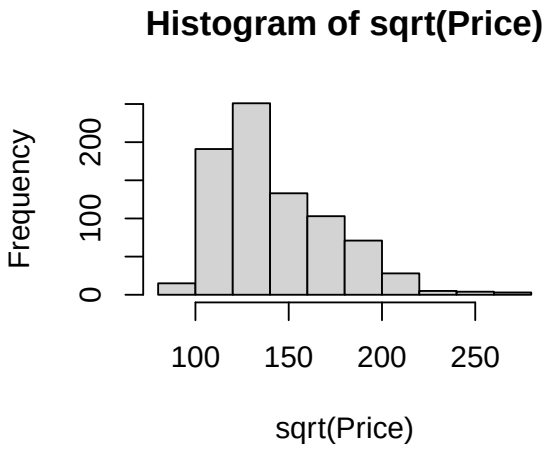
In order to account for the non-normality of Price, we are considering 2 models. The first model is a linear model using $ln(Price)$ (Model 1). The second model is a linear model using $\sqrt{Price}$ (Model 2). For each model we performed variable selection using the hybrid AIC method. In each case we found the interaciton between Make and Mileage to be significant and have included it in the model.

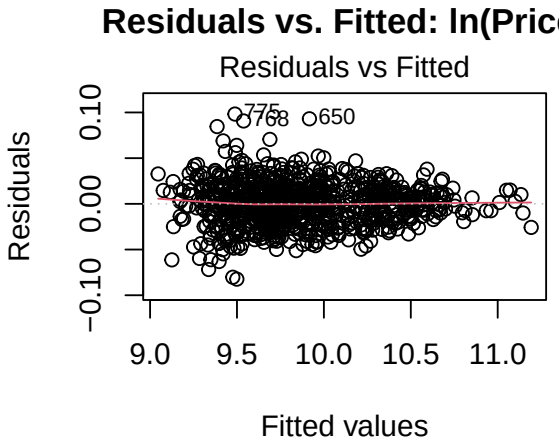
$$ln(Price) = \beta_0 + \beta_1 I_{Mileage} + \beta_2 I_{Sound} + \beta_3 I_{Leather} + \beta_4 I_{Cruise} + \sum_{i=1}^{n} \beta_{5_i} I_{Make_i * Mileage} + \sum_{i=1}^{n} \beta_{6_i} I_{Trim_i} + \sum_{i=1}^{n} \beta_{7_i} I_{Model_i} + \sum_{i=1}^{n} \beta_{8_i} I_{Make_i} \quad (1)$$

$$\sqrt{Price} = \beta_0 + \beta_1 I_{Mileage} + \beta_2 I_{Sound} + \beta_3 I_{Leather} + \sum_{i=1}^{n} \beta_{4_i} I_{Make_i * Mileage} + \sum_{i=1}^{n} \beta_{5_i} I_{Trim_i} + \sum_{i=1}^{n} \beta_{6_i} I_{Model_i} + \sum_{i=1}^{n} \beta_{7_i} I_{Make_i} \tag{2}$$

In considering the assumptions necessary for these models, the first we considered is independence. In this case, due to our own experience with cars, we feel confident assuming independence in these factors (primarily in mileage: some factors may be related such as make/model, as well as special features like leather/cruise with trim). Linearity was shown in Figure 1. As shown in Figure 1 - Top, Model 1 more closelt matches the normality assumption, but both are an improvement over the untransformed data. To check the equal veriance assumption, we examine the fitted residuals of each model (Figure 2 - Bottom). We see that Model 2 has a more constant distribution of residuals, indicating it more cloesly meets the assumptions necessary for the model.







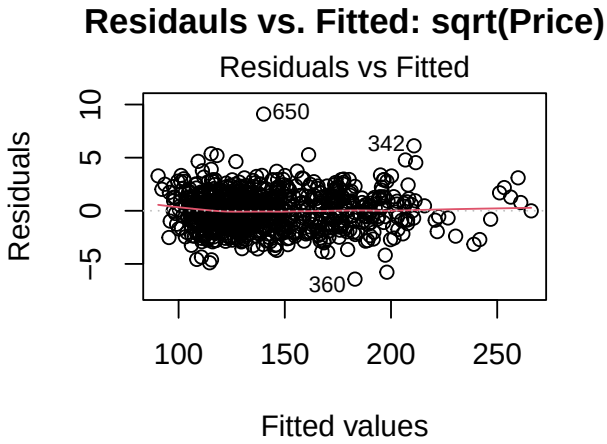


Figure 2: Top - Histograms showing the distribution of the transformed Price values. Note they are less skewed than the histogram in Figure 1, but still show show right-skewedness. Bottom - Plots of residuals versus fitted values. The residuals for Model 2 are more consistent across the fitted values.

2.2 Evaluation of Models

Comparing the adjusted R^2 values for each model, for Model 1 we get a value of 0.996. For Model 2 we get a value of 0.997. These models perform almost identically in explaining the variance in the provided data set. Furthermore, when evaluating prediction capabilites, the models perform almost identially, with Model 1 having a prediction R^2 of 0.997 and Model 2 0.997.

Comparing the models directly, we see that Model 1 identified $Cruise$ as being significant, whereas Model 2 excludes it from the model. It's important to note that even though Model 2 does not include $Cruise$ as a predictor, it did not lose any predictive power. Thus, the simplicity (one variable less) of Model 2, combined with nearly identical performance in our desired measures, leads us to select Model 2 to answer our research question.

3: Results

3.1 Estimation of Model Parameters

The estimates for the parameters are given in Table 2 in the appendix. However, we want to highlight some particularly impactful parameters. The coefficient for Mileage is -0.0005828: this means that as mileage increases by 1 mile, the expected value of the square root of the cost of the car decreases by \$0.0005828. The coefficient for the interaction of Mileage and Cadillac is -.0002621 meaning that for a Cadillac car, as the mileage increases by 1 mile, the expected value of the square root of the cost of the car decreases by \$0.0002621 when compared to a Buick. Following with the example of Cadillacs, the coefficient for the $Make = Cadillac$ parameter is 104.895 meaning that the expected value of the square root of the price of a Cadillac increases by \$104.895 when compared to a Buick. Similarly, the other coefficients relating to multi-level factors are an indication of how much the square root of price increases (or decreases) when that value is present vs. the baseline value.

3.2 Addressing Research Questions

1. What variables are important in predicting the price of car?

We found that the following variables are important in explaining the price of a car: Mileage, Make, Model, Trim, Sound, Leather, and the interaction between Mileage and Make.

2. What other factorsr might be important in predicting car price?

The model we selected describes the provided data extremely well. However, based on personal experience, we believe that the state of a car's title is also an important factor to consider (Clean, Rebuilt, etc.). It may be that all the cars in our dataset had 1 type of title, thus it did not play an important role in our analysis. However, if we were to generalize this model we believe Title Status would need to be included.

3. Does the make of a car impact the rate at which mileage impacts price?

Yes, as shown in Figure 1 - Right, we identified that the make of a car does interact with mileage, resulting in some makes maintaining value as a function of mileage better than others. In particular, Chevrolet and Saturn hold their value better than the other makes.

4. What characteristics give a car the highest value if it has 15k miles?

Using the selected model, the values that give the highest price of a car at 15k miles is a Cadillac XLR-V8 (which by default has is a Trim: Hardtop Conv 2D), with upgraded sound and leather interior.

5. What is a reasonable resale value for a Cadillac CTS 4D Sedan with 17,000 miles, 6 cylinder, 2.8 liter engine, cruise control, upgraded speakers and leather seats?

Using the selected model we predict that a reasonable price for a car meeting the stated specifications is between $34522.6 and $35516.23.

4: Conclusion

In this analysis, we addressed the research questions by creating a linear model of the Square Root of price. The model meets most of the required assumptions fairly well, but does still have some non-normality in the distirbution of price. However, it resulted in a simpler model than the other considered model, which is why we chose it.

Through this model we determined that the variables that impact price are: Mileage, Make, Model, Trim, Sound, and Leather. We also detremined that there may be other factors not included which may be important, such as title status. We showed that the make of the car does impact the rate at which the car loses value per mile, with Chevrolet and Saturn cars best maintaining their value. We calculated that given a car at 15k miles, the values that would maximize the value of that car are a Cadillac XLR-V8 with upgraded sound and a leather interior. Finally, we predict that a resonable resale price for a car with the given condition is between $34522.6 and $35516.23.

Moving forward, we believe it would be useful to perform this same analysis but including the Title Status. We believe this would have a large impact on resale price and would allow more accurate predictions of price. We also feel it would be useful to include a wider variety of makes, particularly from different regions/countries, as perceptions of quality of makes from different countries may have an impacton resale price.

5. Appendix

Table 2: The calculated coefficients of the parameters used in the selected model

Table 1: Model Variable Coefficients

	coef(aic_root)	lower_bound	upper_bound
(Intercept)	159.9128030	157.6280607	162.1975452
Mileage	-0.0005828	-0.0006441	-0.0005215
MakeCadillac	104.8950688	102.1803234	107.6098143
MakeChevrolet	-24.3910388	-26.6836113	-22.0984663
MakePontiac	-25.8128345	-28.4310887	-23.1945803
MakeSAAB	14.3650222	11.6397278	17.0903166
MakeSaturn	-22.9253805	-25.6451712	-20.2055898
Model9_3 HO	18.2266874	16.6152894	19.8380855
Model9_5	22.5659713	20.2594801	24.8724625
Model9_5 HO	24.5249413	22.2406730	26.8092097
Model9-2X AWD	6.3811918	3.2585404	9.5038432
ModelAVEO	-27.3099000	-29.0387253	-25.5810747
ModelBonneville	19.8119337	18.1983134	21.4255539
ModelCavalier	-18.7538413	-20.2448464	-17.2628361
ModelCentury	-25.8474619	-27.4573229	-24.2376008
ModelClassic	-12.5175801	-14.4715830	-10.5635771
ModelCobalt	-13.9745992	-15.4645602	-12.4846381
ModelCorvette	61.9214106	59.9266668	63.9161545
ModelCST-V	-40.1016996	-42.3804431	-37.8229560
ModelCTS	-80.2452976	-82.5151385	-77.9754566
ModelDeville	-71.1226915	-73.4016295	-68.8437536
ModelG6	12.8276594	10.6220147	15.0333041
ModelGrand Am	-1.8040073	-4.1135048	0.5054902
ModelGrand Prix	3.3190223	1.0989709	5.5390736
ModelGTO	45.2965410	42.9165598	47.6765222
ModelImpala	3.5227123	1.8162500	5.2291747
ModelIon	-15.5463875	-17.6327404	-13.4600347
ModelL Series	NA	NA	NA
ModelLacrosse	1.3526588	-0.9247820	3.6300995
ModelLesabre	-3.2182817	-5.5011290	-0.9354345
ModelMalibu	-2.8023575	-4.4705685	-1.1341465
ModelMonte Carlo	NA	NA	NA
ModelPark Avenue	NA	NA	NA
ModelSTS-V6	-59.4943896	-61.7695829	-57.2191963
ModelSTS-V8	-45.7947936	-48.0733175	-43.5162697
ModelSunfire	-13.1998045	-15.5813344	-10.8182745
ModelVibe	NA	NA	NA
ModelXLR-V8	NA	NA	NA
TrimAero Sedan 4D	-19.4443416	-21.0578138	-17.8308694
TrimAero Wagon 4D	-14.6115452	-16.8897303	-12.3333602

	coef(aic_root)	lower_bound	upper_bound
TrimArc Conv 2D	9.7822892	8.1726678	11.3919105
TrimArc Sedan 4D	-10.9636134	-12.5975272	-9.3296997
TrimArc Wagon 4D	-8.9227882	-11.2073515	-6.6382250
TrimAWD Sportwagon 4D	4.3805313	2.7383737	6.0226888
TrimConv 2D	14.3452422	11.9782631	16.7122213
TrimCoupe 2D	1.7799573	0.0499860	3.5099286
TrimCustom Sedan 4D	-9.9963959	-11.6173669	-8.3754250
TrimCX Sedan 4D	-9.4468647	-11.0576500	-7.8360793
TrimCXL Sedan 4D	-4.5310854	-6.1667506	-2.8954202
TrimCXS Sedan 4D	NA	NA	NA
TrimDHS Sedan 4D	15.7359152	13.4681553	18.0036751
TrimDTS Sedan 4D	17.6699267	15.3983433	19.9415101
TrimGT Coupe 2D	1.5670487	-0.7260742	3.8601716
TrimGT Sedan 4D	4.9603032	2.9925086	6.9280978
TrimGT Sportwagon	3.0517882	1.4162330	4.6873433
TrimGTP Sedan 4D	15.7714048	13.5653750	17.9774346
TrimGXP Sedan 4D	7.6685084	6.0416263	9.2953904
TrimHardtop Conv 2D	NA	NA	NA
TrimL300 Sedan 4D	NA	NA	NA
TrimLimited Sedan 4D	NA	NA	NA
TrimLinear Conv 2D	20.5971738	18.9832992	22.2110483
TrimLinear Sedan 4D	NA	NA	NA
TrimLinear Wagon 4D	-11.4952102	-13.7843869	-9.2060336
TrimLS Coupe 2D	5.0041578	3.2727701	6.7355456
TrimLS Hatchback 4D	6.9767495	5.3629559	8.5905430
TrimLS MAXX Hback 4D	6.6351539	4.6968351	8.5734727
TrimLS Sedan 4D	6.4792167	5.0092995	7.9491338
TrimLS Sport Coupe 2D	5.8678257	3.8827072	7.8529442
TrimLS Sport Sedan 4D	7.9307648	5.9505777	9.9109518
TrimLT Coupe 2D	16.9386075	14.5685278	19.3086872
TrimLT Hatchback 4D	7.7121594	6.0672185	9.3571003
TrimLT MAXX Hback 4D	8.5337323	6.5934895	10.4739750
TrimLT Sedan 4D	7.0178790	5.5498694	8.4858887
TrimMAXX Hback 4D	6.1593875	4.2131294	8.1056456
TrimQuad Coupe 2D	9.3362862	7.4317955	11.2407769
TrimSE Sedan 4D	-5.1414272	-6.7622658	-3.5205886
TrimSedan 4D	2.3132768	0.7151129	3.9114407
TrimSLE Sedan 4D	NA	NA	NA
TrimSpecial Ed Ultra 4D	7.9138179	5.6475331	10.1801027
TrimSportwagon 4D	NA	NA	NA
TrimSS Coupe 2D	21.5094353	19.1313156	23.8875550
TrimSS Sedan 4D	24.2783015	22.2432198	26.3133833
TrimSVM Hatchback 4D	-0.6113548	-2.2376473	1.0149377
TrimSVM Sedan 4D	NA	NA	NA
Sound	0.5956842	0.2773191	0.9140494
Leather	1.0949100	0.7221452	1.4676748

	coef(aic_root)	lower_bound	upper_bound
Mileage:MakeCadillac	-0.0002621	-0.0003387	-0.0001856
Mileage:MakeChevrolet	0.0000771	0.0000109	0.0001434
Mileage:MakePontiac	0.0000053	-0.0000677	0.0000784
Mileage:MakeSAAB	-0.0001183	-0.0001922	-0.0000445
Mileage:MakeSaturn	0.0000365	-0.0000472	0.0001202