HW1 Report

RJ Cass

Abstract - FINISH

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.

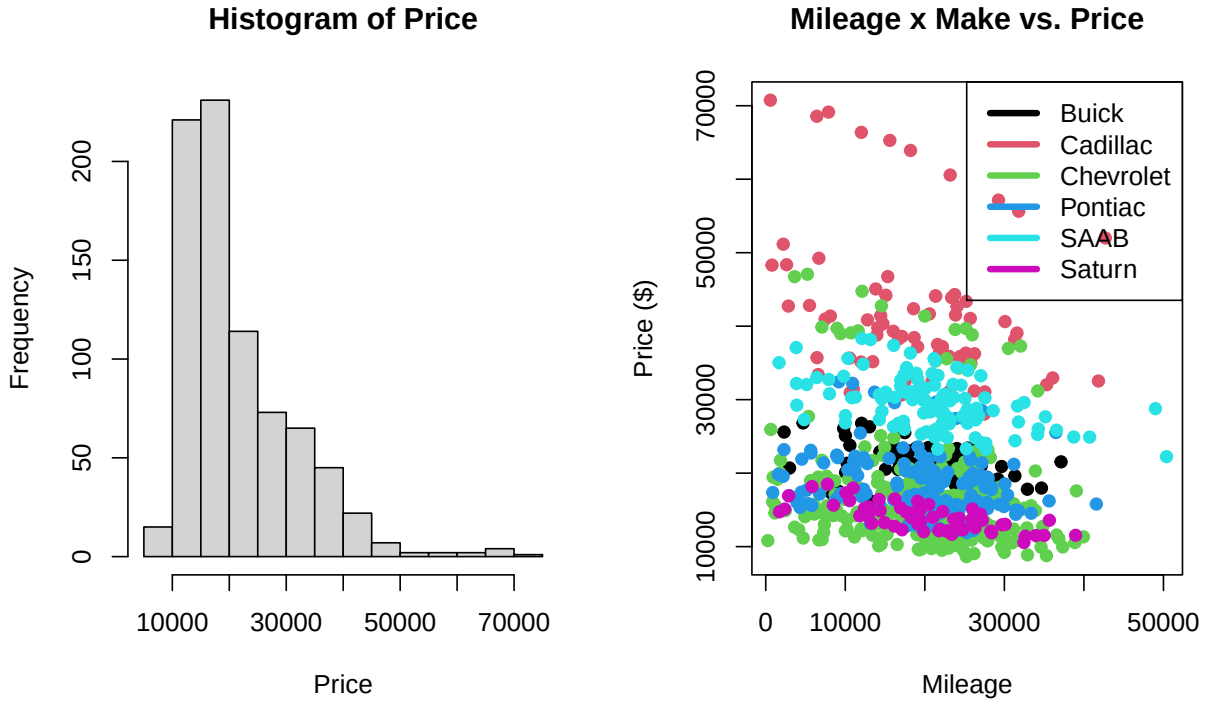


Figure 1: Left - Histogram of Price. Note that this is a right-skewed distribution, as does not meet normality assumptions. Right - Mileage versus Price with different colors for each car make. We can see that price decreases as mileage increases. Note that the rate of decrease for the red dots (Cadillac)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

2: Methodology

2.1 Models Used

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 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.

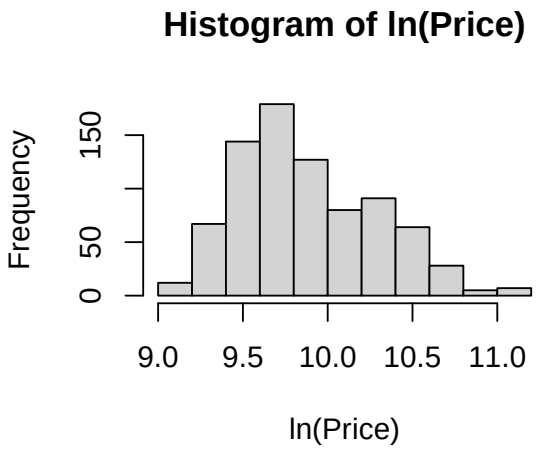
For each of these models, since they are linear models, they must follow the LINE assumptions (linear, independent, normal, and equal variance). How well the models meet these requirements are explored in section 2.2.

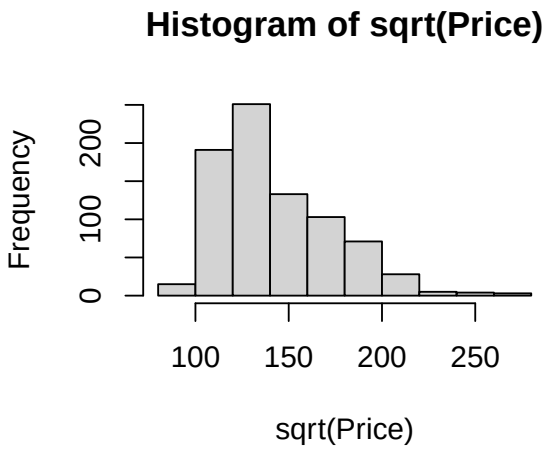
$$ln(Price) = \beta_0 + \beta_1 I_{Mileage} + \beta_2 I_{Sound} + \beta_3 I_{Leather} + \beta_4 I_{Cruise} + \sum_{i=1}^{n} \beta_{5_i} I_{Make_i * Mileage} + \sum_{i=1}^{n} \beta_{6_i} I_{Trim_i} + \sum_{i=1}^{n} \beta_{7_i} I_{Model_i} + \sum_{i=1}^{n} \beta_{8_i} I_{Make_i} \quad (1)$$

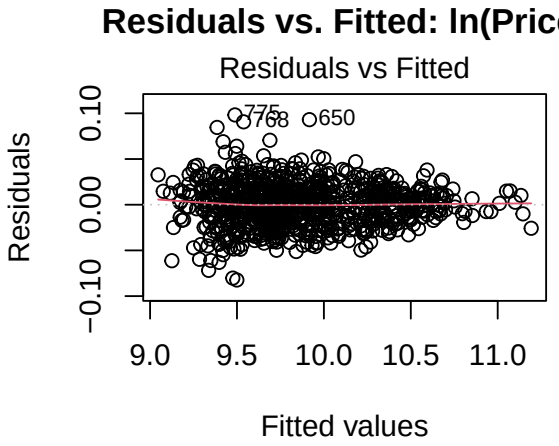
$$\sqrt{Price} = \beta_0 + \beta_1 I_{Mileage} + \beta_2 I_{Sound} + \beta_3 I_{Leather} + \sum_{i=1}^{n} \beta_{4_i} I_{Make_i * Mileage} + \sum_{i=1}^{n} \beta_{5_i} I_{Trim_i} + \sum_{i=1}^{n} \beta_{6_i} I_{Model_i} + \sum_{i=1}^{n} \beta_{7_i} I_{Make_i} \tag{2}$$

2.2 Evaluation of Models

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 2. Having transformed the data, the transformed price distibution now more closely matches the normal distribution (though still not exactly).







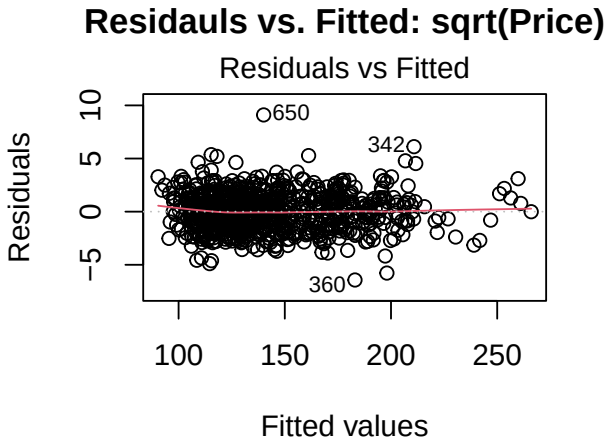


Figure 2: Top - Histograms showing the distribution of the transformed Price values. Note they are less skewed than the histogram in Figure 1, but still show show right-skewedness. Bottom - Plots of residuals versus fitted values. The residuals for Model 2 are more consistent across the fitted values.

To check the equal veriance assumption, we examine the fitted residuals of each model (Figure 4). We see that the model using $\sqrt{Price}$ has a more constant distribution of residuals, indicating it may more cloesly meet the assumptions necessary for the model.

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 of (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, on 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, 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

	x
(Intercept)	159.9128030
Mileage	-0.0005828
MakeCadillac	104.8950688
MakeChevrolet	-24.3910388
MakePontiac	-25.8128345
MakeSAAB	14.3650222
MakeSaturn	-22.9253805
Model9_3 HO	18.2266874
Model9_5	22.5659713
Model9_5 HO	24.5249413
Model9-2X AWD	6.3811918
ModelAVEO	-27.3099000
ModelBonneville	19.8119337
ModelCavalier	-18.7538413
ModelCentury	-25.8474619
ModelClassic	-12.5175801
ModelCobalt	-13.9745992
ModelCorvette	61.9214106
ModelCST-V	-40.1016996
ModelCTS	-80.2452976
ModelDeville	-71.1226915
ModelG6	12.8276594
ModelGrand Am	-1.8040073
ModelGrand Prix	3.3190223
ModelGTO	45.2965410
ModelImpala	3.5227123
ModelIon	-15.5463875
ModelL Series	NA
ModelLacrosse	1.3526588
ModelLesabre	-3.2182817
ModelMalibu	-2.8023575
ModelMonte Carlo	NA
ModelPark Avenue	NA
ModelSTS-V6	-59.4943896
ModelSTS-V8	-45.7947936
ModelSunfire	-13.1998045
ModelVibe	NA
ModelXLR-V8	NA
TrimAero Sedan 4D	-19.4443416
TrimAero Wagon 4D	-14.6115452

	x
TrimArc Conv 2D	9.7822892
TrimArc Sedan 4D	-10.9636134
TrimArc Wagon 4D	-8.9227882
TrimAWD Sportwagon 4D	4.3805313
TrimConv 2D	14.3452422
TrimCoupe 2D	1.7799573
TrimCustom Sedan 4D	-9.9963959
TrimCX Sedan 4D	-9.4468647
TrimCXL Sedan 4D	-4.5310854
TrimCXS Sedan 4D	NA
TrimDHS Sedan 4D	15.7359152
TrimDTS Sedan 4D	17.6699267
TrimGT Coupe 2D	1.5670487
TrimGT Sedan 4D	4.9603032
TrimGT Sportwagon	3.0517882
TrimGTP Sedan 4D	15.7714048
TrimGXP Sedan 4D	7.6685084
TrimHardtop Conv 2D	NA
TrimL300 Sedan 4D	NA
TrimLimited Sedan 4D	NA
TrimLinear Conv 2D	20.5971738
TrimLinear Sedan 4D	NA
TrimLinear Wagon 4D	-11.4952102
TrimLS Coupe 2D	5.0041578
TrimLS Hatchback 4D	6.9767495
TrimLS MAXX Hback 4D	6.6351539
TrimLS Sedan 4D	6.4792167
TrimLS Sport Coupe 2D	5.8678257
TrimLS Sport Sedan 4D	7.9307648
TrimLT Coupe 2D	16.9386075
TrimLT Hatchback 4D	7.7121594
TrimLT MAXX Hback 4D	8.5337323
TrimLT Sedan 4D	7.0178790
TrimMAXX Hback 4D	6.1593875
TrimQuad Coupe 2D	9.3362862
TrimSE Sedan 4D	-5.1414272
TrimSedan 4D	2.3132768
TrimSLE Sedan 4D	NA
TrimSpecial Ed Ultra 4D	7.9138179
TrimSportwagon 4D	NA
TrimSS Coupe 2D	21.5094353
TrimSS Sedan 4D	24.2783015
TrimSVM Hatchback 4D	-0.6113548
TrimSVM Sedan 4D	NA
Sound	0.5956842
Leather	1.0949100

	x
Mileage:MakeCadillac	-0.0002621
Mileage:MakeChevrolet	0.0000771
Mileage:MakePontiac	0.0000053
Mileage:MakeSAAB	-0.0001183
Mileage:MakeSaturn	0.0000365