

How Much is Your Car Worth?

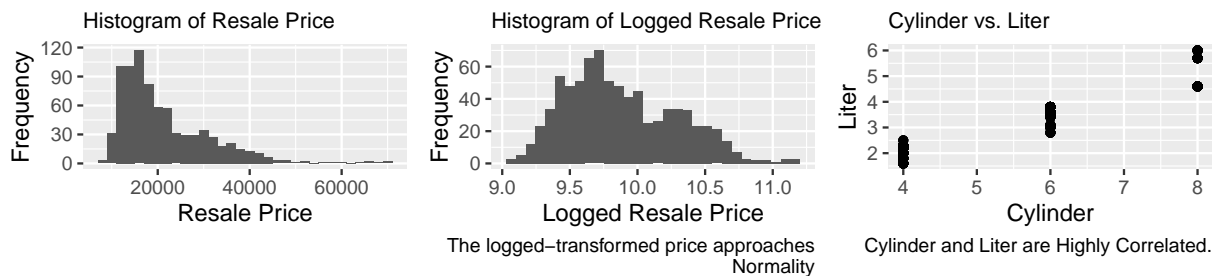
Sam Lee & Evan Miller

Abstract

The resale value of a car is an important factor to consider when buying a car. In this report, we use data from Kelley Blue Book to try and understand what factors contribute to the resale value KBB sets, and in what way. We fit multiple linear regression on the log-transformed price of a car to predict the resale value of a car given the mileage, and other key characteristics of a car. We found that the most significant factors that contribute to the predicted price of a given car are Mileage, Model, Leather, Sound, and Trim, that the model is highly predictive of the resale price of a car, and that there isn't a significant difference in the effect of mileage on the price of a car between different makes.

Introduction

The Kelley Blue Book value of a car is widely used by consumers and dealerships to determine the value of a car. In this report, we will use data from Kelley Blue Book to try and understand better what factors contribute to the resale value KBB sets, and in what way. By fitting a model, we will also be able to predict the resale value of a car given certain characteristics. Before fitting a model, there are a few issues in the dataset that need to be accounted for. As you can see in the histogram below, the resale price of the cars is right-skewed. To account for this, we will log-transform the price. We also found some correlation between some of the predictor variables in the data. For example, as you can see from the scatterplot below, the number of cylinders in the car is highly correlated with the engine size. To account for this issue, we will use variable selection techniques before fitting our model to make sure there aren't pairs of highly correlated predictors in the model.



Methodology

Method 1

The first model we proposed was a model obtained by performing AIC forward selection on the log price of a car. Modelling the log price of the car instead of the price of the car addresses well the issue of skewness of the distribution of prices as described above. Forward selection AIC on log price gave us a model with Model, Mileage, Trim, Leather, Sound, and Cruise as predictors. This process accounted for the issue of correlation between predictors also described above.

Method 2

The second model we proposed was a model obtained by performing BIC two-way selection on the log price of a car. This method accounts for the issues in the data the same way forward AIC does, but BIC penalizes the inclusion of predictors a little more than AIC and therefore results in a more parsimonious model. The BIC model included Mileage, Model, Leather, Sound, and Trim as predictors. In both of these models we assume linearity between the predictors and the log price of the car, independence between observations, normality in the residuals, and equal variance of the residuals. These assumptions will be evaluated in more detail in the model evaluation section.

Model Evaluation

For a given car i , we propose that the resale value of the car can best (and most simply) be approximated by the following linear¹ model²:

$$\begin{aligned} \ln(\text{Price}_i) = & \beta_0 + \beta_1 \text{Mileage}_i + \beta_2 \mathbb{1}(\text{Sound} = 1) + \beta_3 \mathbb{1}(\text{Leather}_i = 1) + \\ & \sum_{j=9-3}^{XLR-V8} \beta_j \mathbb{1}(\text{Model}_i = j) + \sum_{k=\text{Aero Sedan 4D}}^{\text{SVM Sedan 4D}} \beta_k \mathbb{1}(\text{Trim}_i = k) + \epsilon_i \\ & \epsilon_i \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2) \end{aligned}$$

We used a Monte Carlo cross-validation technique with a 1,000 iterations on each model, evaluating the model against 20% of out-of-sample observations each iteration. The root-mean-square error for the forward selection linear model set with AIC as a model evaluation metric was approximately 0.02624, while our selected linear model where both forward and backward variable selection were enabled with BIC as the model evaluation metric arrived at 0.02623 for the root-mean-square error. While our BIC model was marginally better, we chose this model not out of its performance (as it achieved nearly the same level of fit) but out of a better parsimonious fit.

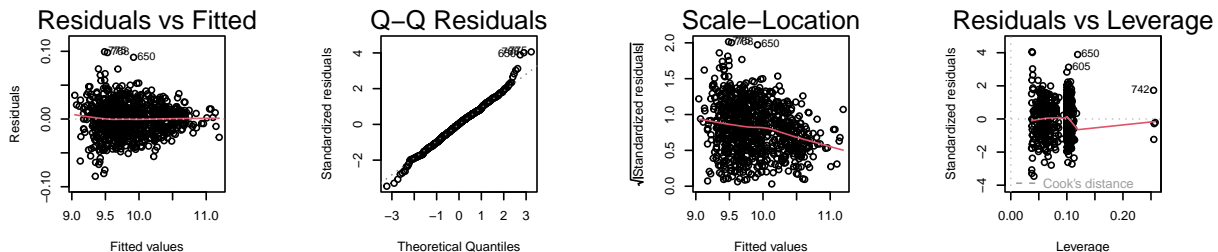
¹The set of linear coefficients of interest are vector of β coefficients. We include a vector of idiosyncratic error terms, denoted in the model as ϵ .

²Note that on the categorical variables (Model and Trim) we omitted the first alphanumerically ordered factor to preserve the rank of the linear model (e.g. the model 9-2X AWD was omitted as a covariate in the set of Model dummies and the Aero Conv 2D was omitted as a covariate in the set of Model indicator dummies.) We also note that in this regression, the Mileage covariate was standardized by subtracting the mean and dividing by its standard deviation.

We assess the in-sample fit of our model through the adjusted R^2 statistic. With the sample data, the adjusted R^2 statistic is 0.9963. Given that nearly all of the variance in the data can be explained through all model, we believe there may be some economic factors dictating the resale price. If certain models exist in perfectly competitive markets (or nearly perfectly competitive markets) this could shrink the variance of the observed prices to zero. Alternatively, if particular price controls have been imposed in the respective supply chain markets such as on certain leathers, or materials involved in sound system, then our model may be highly sensitive to the economic circumstances of from which the data was sampled—that is, if the economic circumstances change, our model may yield an improper fit. Be it as it may, the model as it stands provides as a robust and interpretable tool for predicting resale prices of cars in this market. We log-transformed price, as it typically done in financial and economic analysis; this improved our model assumptions and our confidence that we could use this in predicting resale prices.

Model Assumptions

To use this model as a valid tool for prediction, we invoke a few distributional assumptions that we pose in this section and evaluate for our proposed model.



We hold that the assumption of linearity is met as the the structure of the residuals with respect to the fitted values do not show any structure. We also assert that our distributional assumption that our error term (ϵ) is Normally distributed as shown by the QQ plot. While there are few outliers in the tails³ (see Table 1), we conclude that the heteroskedasticity is largely not an issue, although future analyses may wish model ϵ through a more robust form. Along this note, we assume each observation in our data is independent (*i.i.d.*) such that the resale price of one car does not affect the resale price of another. This may be violated if certain economic constraints are dictating the resale price of a class of cars or factors used in determining the resale price of cars.

Table 1: Problematic Observations

Price	Mileage	Model	Trim	Sound	Leather
\$17,325.27	19894	Vibe	AWD Sportwagon 4D	0	0
\$22,244.88	50387	9_3	Linear Sedan 4D	0	1

³After investigation, we do not find these outliers problematic. These observations contain unique combinations of covariates that average out the effects.

Price	Mileage	Model	Trim	Sound	Leather
\$25,959.12	17431	9-2X AWD	Linear Wagon 4D	0	1

Results

The most significant factors that contribute to the predicted resale price of a given car are summarized below in Table 2. Due to the nature of our saturated regression—by virtue of including every possible dummy variable for every categorical variable—it is not surprising that the most significant factors are from these sets of indicator variables.

Significant Factor	Estimate	95% Confidence Interval
Lowest Coefficients		
Model = AVEO	-1.0497	(-1.0928, -1.0066)
Model = Cavalier	-0.8975	(-0.9420, -0.8530)
Model = Sunfire	-0.8486	(-0.8917, -0.8055)
Model = Ion	-0.8361	(-0.8786, -0.7936)
Model = Cobalt	-0.8119	(-0.8554, -0.7684)
Highest Coefficients		
Model = XLR-V8	0.7936	(0.7521, 0.8351)
Model = CST-V	0.3994	(0.3615, 0.4373)
Trim = SS Sedan 4D	0.3792	(0.3419, 0.4165)
Trim = SS Coupe 2D	0.3557	(0.3187, 0.3927)
Model = STS-V8	0.3446	(0.3083, 0.3809)
Intercept		
Intercept	10.3813	(10.0028, 10.7598)

Table 2: Coefficients and 95% Confidence Intervals for the most significant factors

As noted earlier, the adjusted R^2 statistic is remarkably high (and perhaps conspicuously so due to idiosyncratic economic reasons). Nonetheless we would attribute any of the noise in our model and perhaps issue a word of caution that this model stands precariously on the verge of robustness and predictability: We do not have any measure of maintenance use, history of previous ownership, nor accident history. These unobserved covariates would bias our coefficients if they are, in fact, correlated with the current covariates we have included in the model. Hence we strongly caution against interpreting our results as causal.

Model Predictions

Our model concludes (and significantly so) that as mileage increases, price decreases⁴. However, both stepwise models failed to include an interaction term between mileage and the categorical variable and make. Upon investigation, we found that none of the interaction terms between mileage and make in the full linear model were significant. Hence, we conclude that the amount of decrease in value does not significantly depend on the make of the car.

⁴Our estimated coefficient on (standardized) mileage was -0.008150159 ($-0.0083695, -0.007931$) for every 1,000 miles.

Using our model we predict that the following car (limited to our model’s specifications⁵) will have the highest resale value at 15,000 miles:

Table 3: Highest value car with mileage=15,000 miles

Model	Trim	Leather	Sound
XLR-V8	SS Sedan 4D	1	1

For a Cadillac CTS 4d Sedan with 17,000 miles, 6 cylinder, 2.8 liter engine, cruise control, upgraded speakers, and leather seats, we performed a 95% prediction interval using our model. Our model predicts that this car will resale at \$30,346.55 (\$28,826.75, \$31,946.47).

Conclusion

This study examines the key factors that influence the resale value of a car using data from Kelley Blue Book. We determined the most significant factors that were of interest—notably, we concluded that while mileage contributed to the depreciation of a car, there was no inner dependency between the mileage of a car and its make. We also aimed to predict specific resale values for particular cars according to the this report’s specifications. For example, a Cadillac CTS 4D Sedan with 17,000 miles was predicted to have a resale value of approximately \$30,346.55, given its specific characteristics such as leather seats and upgraded sound system. We analyzed and compared two linear regression models using forward and backward selection methods. Both models used a logged transformed price in its evaluation.

We also acknowledge several limitations that may prove insightful for future research and model adaptations. Notably we caution against a causal application of our model. Our model account for unobserved factors such as maintenance history or accident records. We also admit the potential need for additional structure in the error term (ϵ) in our model to accomodate for potential dependencies. We would also encourage exploring non-linear models as well to investigate potential non-linearities found in our model.

⁵It should be noted that due to the nature of how we selected our model, beyond these specifications, the resale model should be fairly robust to any another modifications (i.e. the make of the car).