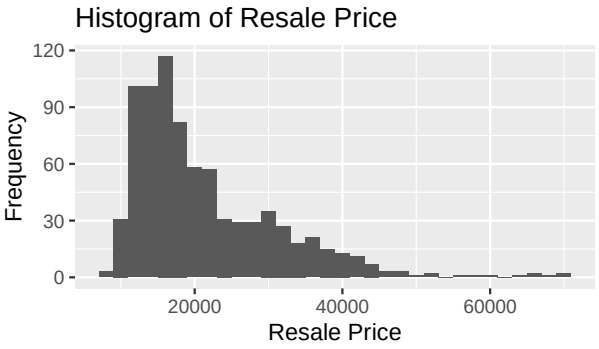
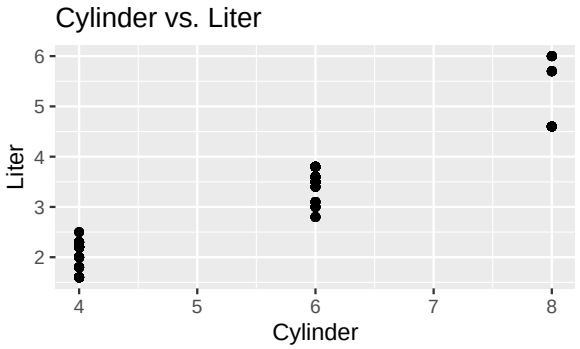
How Much is Your Car Worth?

Sam Lee & Evan Miller

Introduction

Kelley Blue Book is one of the most well known and trusted car evaluation resources. 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

Proposed Methods

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

In this section, we will propose the statistical model we arrived at and describe the methods we used to reach our conclusion.

For a given car i, we propose that the resale value of the car can best (and most simply) be approximated by the following linear model[1]:

$$ln(\text{Price}_i) = \beta_0 + \beta_1 \text{Mileage}_i + \beta_2 \mathbb{1}(\text{Sound} = 1) + \beta_3 \mathbb{1}(\text{Leather}_i = 1) + \sum_{j=\text{9-3}}^{XLR-V8} \beta_j \mathbb{1}(\text{Model}_i = j) + \sum_{k=\text{Aero Sedan 4D}}^{\text{SVM Sedan 4D}} \beta_k \mathbb{1}(\text{Trim}_i = k) + \epsilon_i$$

$$\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

We used a Monte Carlo cross-validation techinque with a 1,000 interations 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 evalutation 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

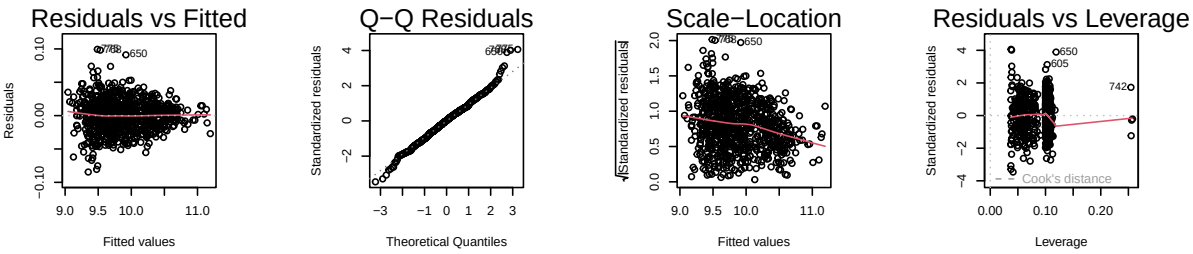
[1]Note that on the categorical variables (Model and Trim) we omitted the first alphanumerically ordered factor to preserve the rank of the linear model (e.g. the model 9-2X AWD was omitted as a covariate in the set of Model dummies and the Aero Conv 2D was omitted as a covariate in the set of Model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

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.

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 tails[2] (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.

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

Results

Using our model that we established in the preceding section, we will proceed to answer our research questions.

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.

Table 1: Problematic Observations

Price	Mileage	Model	Trim	Sound	Leather
17325.27	19894	Vibe	AWD Sportwagon 4D	0	0
22244.88	50387	9_3	Linear Sedan 4D	0	1
25959.12	17431	9-2X AWD	Linear Wagon 4D	0	1

Model Predictions

Our model concludes (and significantly so) that as mileage increases, price decreases[3]. 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.

Using our model we predict that the following car (limited to our model's specifications[4]) 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 Caillac 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).

[3]Our estimated coefficient on (standardized) mileage was −0.008150159 (−0.0083695,−0.007931) for every 1,000 miles.

[4]It 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).