STAT 536 HW 1

Ryan Hanson

9/15/2021

Abstract

Kelly Blue Book prices are modeled based on many variables selected by their company. I will use a dataset of over 800 sales of GM cars from 2005 to create a linear regression model to predict sales prices. According to my linear regression model, mileage, make, and engine size have the largest effect on sale price.

Exploratory Data Analysis

Kelly Blue Book has always been a trusted ally to those unaware of the intricacies of the automobile resale market. It allowed those individuals to know the appropriate valuation for their car, thus allowing them to feel secure in any transaction they made. However, the actual calculations created by Kelly Blue Book are unknown to the standard user. In this article, I am going to help you understand some of the factors that Kelly Blue Book uses, the size of each factor's effect on price, and create a model which will predict the expected resale price.

The first step in creating any model is understanding the data, and what impact each variable will have on your predicted output. This dataset contains over 800 cars manufactured by General Motors in 2005, their features, and the price for which they were sold on Kelly Blue Book. These features include mileage, make, model, trim, type, cylinder count, cruise control, leather seats, engine size, number of doors, and upgraded sound system. Initially, I used boxplots to analyze the relationship between each variable and price, in order to compare means and spread. The boxplots showed a strong effect on price by car make and cylinder count.

Some of the other variable effects, such as whether a car had cruise control, an improved sound system, or leather seats had a minimal effect on the sale price. Boxplots allow me to get an initial understanding of the data, but won't show any of the interaction effect between the variables.

My exploratory data analysis also revealed that certain variables are spread out thinly over a large number of categories. For example, there are 47 types of trim (which is more specific than the model of a car) and each trim only has around ten observations. This makes the averages they supply highly unreliable and extremely susceptible to wild variations. If I use make instead of trim, I now have a variable with only six options and data counts ranging from 60-320 observations. This will allow the data to be much more reliable across a wider set of predictions.

The Kelly Blue Book dataset will be useful in creating a model, but it's far from perfect. There are many additional variables we could add to more fully encapsulate the value of the car. For example, accident history could be measured in some way. Even though these cars are all listed in excellent condition, there could be fixes that leave the car weaker than before. I would also include features such as DVD players and USB ports. These newer features are becoming more important factors in purchasing a car. There are many other features I could quantify, and we all need to recognize that no matter how much data I use to fit my model, there will always be additional factors that create error for my predictions. The purpose of variable selection in the model is to minimize that error.

If I rely on variables with a high number of discrete possibilities, I risk overfitting the data to specific cars. Overfitting is when a model is created that relies too much on the original dataset, making the model less

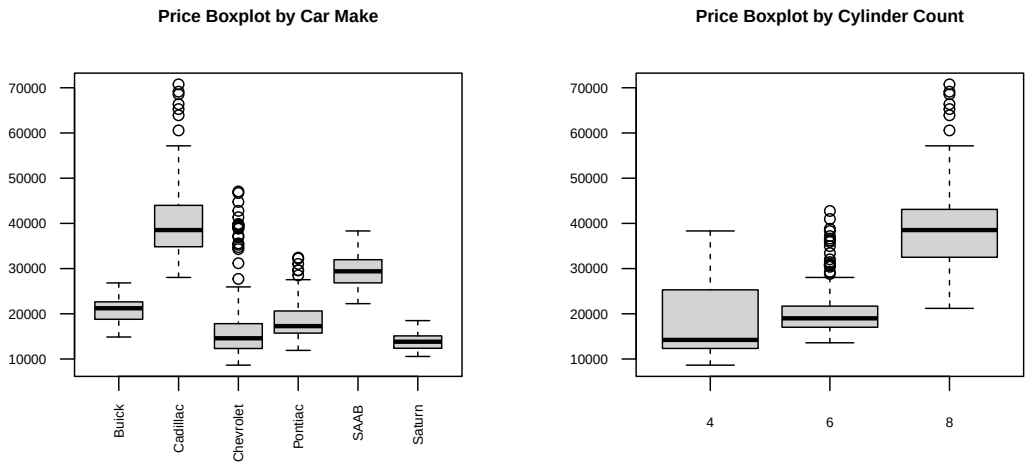

Figure 1: Boxplots showing the relationship between Price and Make/Cylinders

flexible to interpret new data. Another factor I need to recognize is interaction effect. Often, variables on their own will have no indication of price, but the interaction between two variables will have a noticeable effect. Now that I've analyzed and gained an understanding of the data, I can move onto creating a model.

Model Selection

This EDA led me to use linear regression to create a model that would predict price based off of the selected variables. My linear regression model inputs all of the data and analyzes it to create an equation to predict the output variable (Price). It analyzes the inputs and assigns them values known as βs. Regression models create a starting price, and then adjust from that price by a given amount per change in that variable using these βs. For example, the β assigned to mileage indicates how much the estimated sale price increases or decreases for every additional mile driven by the car. The equation for this specific linear regression is detailed below.

The variables selected were mileage, make, door count, cylinder, engine liters, and the interaction effect between doors and make. I selected these variables by starting with only mileage, then adding one variable at a time, determining whether it improved the model, then moving to the next variable until I had a model that fit the data effectively without overfitting. Interestingly, the door count had a minimal effect on the model, but the interaction between make and door had a statistically significant effect. Finally, ϵ represents the error term to account for individual variation.

$$\begin{aligned}\text{Price}_i = {} & \beta_0 + \beta_1 Mileage + \beta_2 I_{\text{Make}_{Cadillac}} + \beta_3 t I_{\text{Make}_{Chevrolet}} + \beta_4 I_{\text{Make}_{Pontiac}} + \beta_5 I_{\text{Make}_{SAAB}} + \\ & \beta_6 I_{\text{Make}_{Saturn}} + \beta_7 I_{\text{Doors}_4} + \beta_8 I_{\text{Liter}} + \beta_9 I_{\text{Cylinder}_6} + \beta_{10} I_{\text{Cylinder}_8} + \beta_{11} I_{\text{Make}_{Cadillac}} I_{\text{Doors}_4} + \\ & \beta_{12} I_{\text{Make}_{Chevrolet}} I_{\text{Doors}_4} + \beta_{13} I_{\text{Make}_{Pontiac}} I_{\text{Doors}_4} + \beta_{14} I_{\text{Make}_{SAAB}} I_{\text{Doors}_4} + \epsilon_i\end{aligned}$$

In order to use linear regression, my model needs to meet certain assumptions. The relationship between my predictions and outcomes need to be linear. The residuals, the difference between the predicted value and our observed value, need to be normally distributed and have a comparable variance.

First, I will analyze the relationship between my model predictions and the observed values. Figure 2 shows that the relationship is approximately linear as the red line representing it is horizontal and centered near zero.

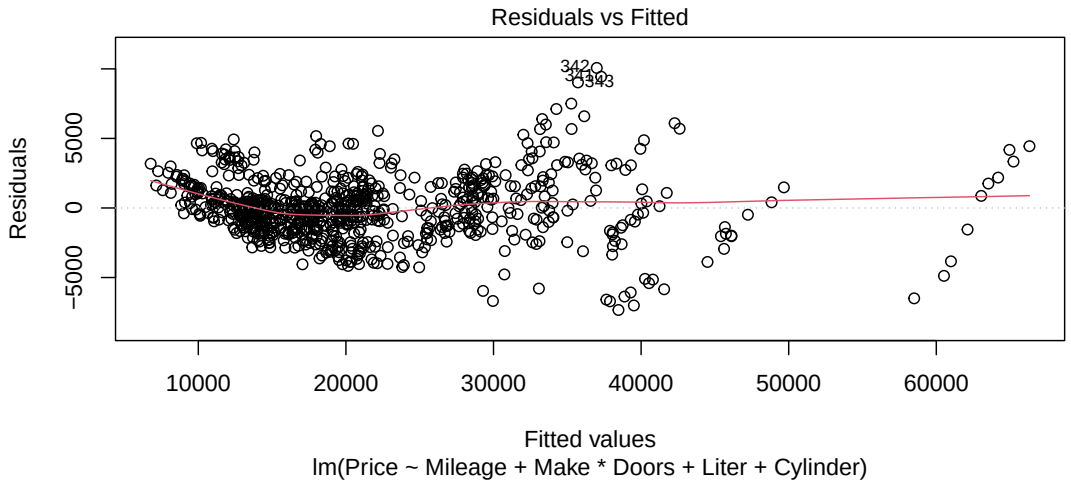

Figure 2: Our horizontal red line centered around zero shows the data is approximately linear

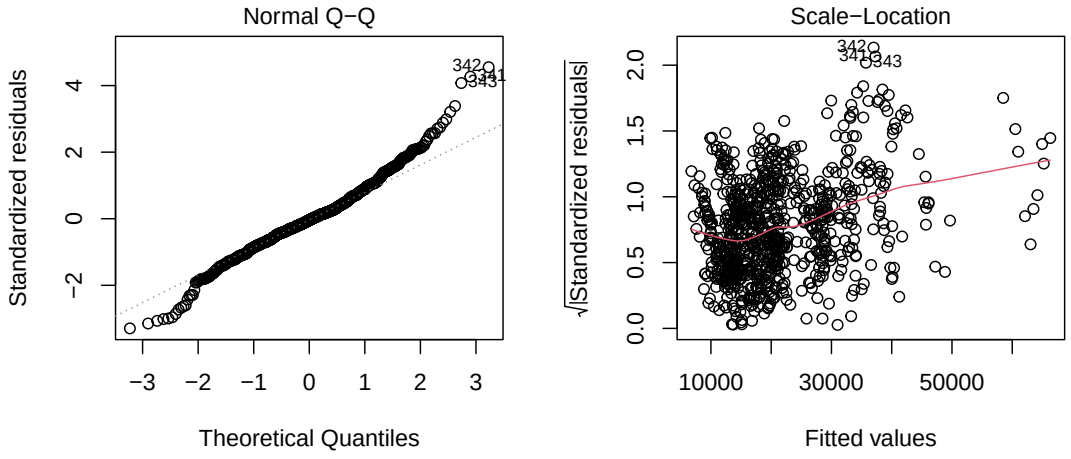

Figure 3: These plots show that the residuals are normally distributed as they follow the diagonal (left) and approimately constant variance (right)

The next assumptions I need to verify involve the residuals. Figure 3 uses a normal Q-Q plot to show that the residuals are normally distributed when they are approximately on the diagonal line. Figure 3 also contains a Scale-Location chart that proves the residuals have constant variance when the line is approximately horizontal.

One metric used to validate a model is R^2. R^2 measures how much of the dependent variable (Price) is explained by the independent variables I put into our model. In order to decrease the value of unimportant variables, I adjust the R^2 score by the number of variables in the model. This model has an adjusted R^2

score of .948, meaning that almost 95% of the price is going to be explained by the variables I have selected. This is an extremely high value that supports the reliability of the model.

The root mean square error (RMSE) can also help show the model's reliability. RMSE is the standard deviation of the difference between the observed values and my predicted values. Lower values show that the model closely follows reality. In this case, we have an RMSE of $2,225. Given the resale cost of cars, this RMSE is relatively small and shows this model is working reliably.

Model Parameters

Estimates of each of the β's are given in table 2. Unsurprisingly, the make of the car has a large effect on the sale price of the car. The size of the car engine in liters also had an extreme effect on price, raising the price by almost 7k for each additional liter. Mileage decreases the price at a rate of about $185 per every 1,000 miles driven. For the doors, make, and cylinder category, the model sets a baseline using 2 doors, Buick, and 4 cylinders, then adjusts such that 4 doors is on average sold for $913 less than a 2 door, etc.

Table 2: Estimates of the model parameters, with standard errors (SE) and brief descriptions.

Parameter	Estimate	SE	Coefficient Description
β_0	6,315	996.5	Intercept
β_1	-0.185	0.009	Mileage
β_2	35,240	1076	MakeCadillac
β_3	-2,536	708	MakeChevrolet
β_4	-4,946	791	MakePontiac
β_5	17,660	800	MakeSAAB
β_6	-2,626	463	MakeSaturn
β_7	-913	621	Doors4
β_8	6,658	311	Car Engine Liters
β_9	-5,182	553	Cylinder6
β_{10}	-5,759	1,114	Cylinder8
β_{11}	-22,750	993	MakeCadillac $\times$ Doors4 interaction
β_{12}	183	679	MakeChevrolet $\times$ Doors4 interaction
β_{13}	3,648	773	MakePontiac $\times$ Doors4 interaction
β_{14}	-5,731	791	MakeSAAB $\times$ Doors4 interaction

Prediction

Now that I have my model, I use it to give predictions and analyze the effect of certain variables. When I added an interaction effect between make and mileage, I saw that there was no significant interaction. Therefore, while mileage and make both have a significant effect on the price, the rate at which mileage decreases resale price does not change between different car makes.

Since there was minimal effect for the interaction between mileage and make, the variables that will dictate the highest resale value are independent of the mileage. Therefore, at any mileage the highest resale value will be a Cadillac with 2 doors and the largest possible engine. The largest the data shows is 6.8 liters. The model projects approximately a $70,000 car with these specifications at 15,000 miles.

In order to test the model, I created a new observation as a test case. In this instance, I used a Cadillac CTS 4D Sedan with 17,000 miles, 6 cylinders, 2.8 liter engine, cruise control, upgraded speakers, and leather seats. When I pass this through my model, it estimates the car will sell for $28,195, with a 95% confidence interval between $27,460 and $28,931.

This aligns with similar observations that were initially given for the same car type, even though I didn't use every single variable. This shows that my model works effectively in predicting new values not initially found in the dataset.

Conclusion

The model proved to be effective in answering which variables were most effective in determining predicted sale price by focusing on mileage, make, and engine size in liters. While there are certainly shortcomings with linear regression regarding outliers and necessity of linearity, it was still a useful choice to analyze the data due to minimal interaction between many of the variables, relative linearity, and a lack of outliers.

My model can become even more robust if I was able to expand the data to other makes and years in order to see the overall effect mileage, engine size, and the rest of the variables have on all different types of cars. It would also be interesting to create a variable for original sale price so I can see how the value of a car depreciates over time. For now, I can only accurately answer questions concerning price for those with 2005 GM cars.