Estimating the Value of Used Cars

Felicia Liu and Naman Patel

December 6, 2022

Introduction

In the United States, the market for used cars has been steadily growing for the last few years. Between 2021 and 2022, the market grew 4.7% from $153.1 billion to $160.4 billion.[1] Growth rates are projected to be even higher in the coming years, and for many used car sellers, an important area of interest is knowing how the age of their car might affect the car's market price.

While sellers can often find the predicted value of their car through sources like Kelley Blue Book, we think it is important to use data to further examine the effect of the car's age on its price. For example, when there are market fluctuations in used car prices, sellers can be informed about the fairness of the prices by looking at the predicted relationship between a car's age and its price.

This study estimates the listing price of a used car based on, primarily, its age and additional features. We utilize car listings on the automotive research and shopping website CarGurus to see how much sellers can expect to earn, based on other sellers' listings in the area for the same vehicle. We will apply a set of regression models to investigate this question and provide advice to sellers.

Data and Methodology

The data in this analysis comes from used car listings on the CarGurus website in September of 2020 and each row represents a distinct listing. It was compiled and made open-source on Kaggle by Ananay Mital.[2] From a total of approximately 3 million rows, we utilized 500,000 rows due to hardware limitations in importing such a massive dataset into R. After exploring and cleaning this reduced dataset, 15,000 rows were used to fit the models and generate the statistics presented in this report.

In the data, there are two main variables that we operationalize for our intended purpose. The first is the year manufactured variable that represents when the listed car was made. We subtract the year manufactured from 2021 to get the variable of interest of vehicle "age".[3]. Our target variable is the price variable which represents the listing price of the car, which operationalizes the effective "value" of the car.

From our starting point of 500,000 rows, we first filtered out any listings that had a listing price greater than $100,000, leaving us with approximately 497,000 rows. This was done to prevent high-leverage ultra-luxurious and vintage car outliers found during data exploration from disproportionately influencing our statistics. Furthermore, this makes our analysis and results more useful for the average individual. We then removed any columns that did not have any predictive power such as Vehicle Identification (VIN) number and listing ID. Next, we filtered out any rows with missing values for an attribute used in any of the three presented models, leaving us with around 118,000 rows. We then took a random sample of 15,000 rows and created our models of interest.

Before getting into model specification, we would like to acknowledge an important caveat of our model and results interpretation. Since all of our data is from 2020, we effectively compare the price of a car bought in

[1] https://carsurance.net/insights/used-car-sales-statistics/

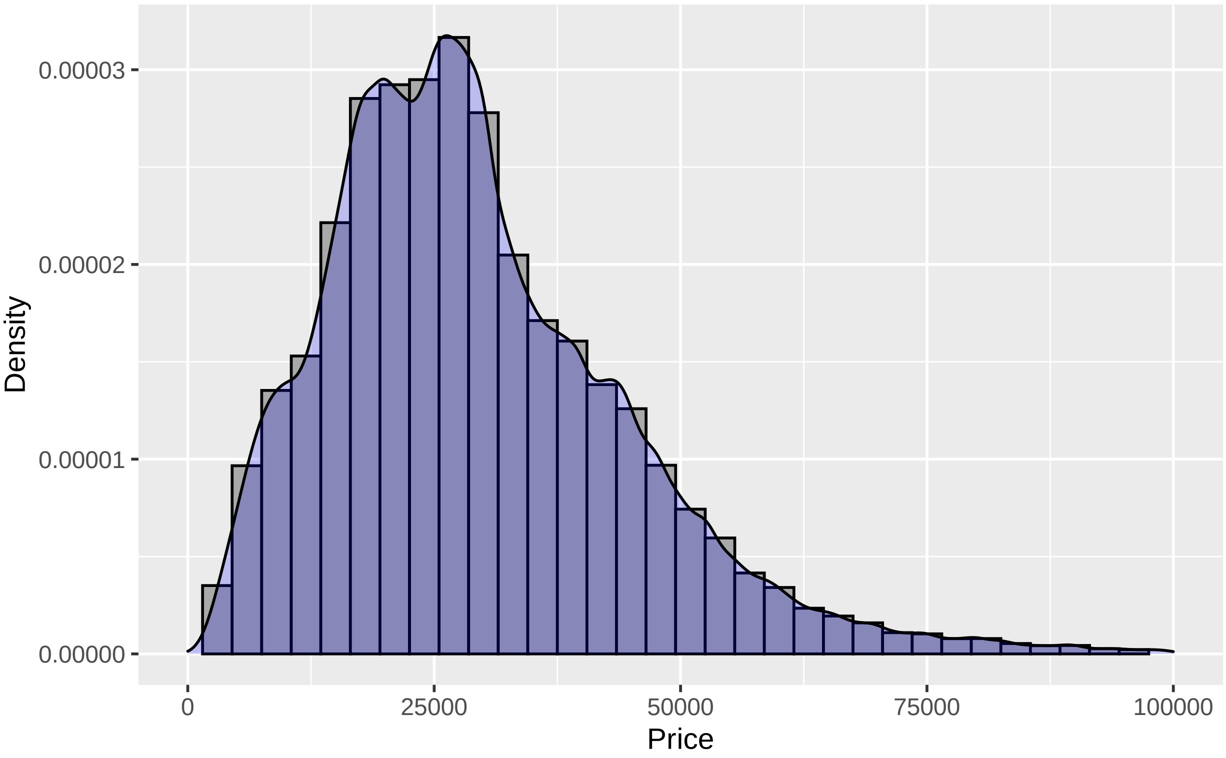
[2] https://www.kaggle.com/datasets/ananaymital/us-used-cars-dataset

[3] Note: We use 2021 and not 2020 since new car models roll out before their calendar year, meaning 2021 models are available in 2020

year X and sold in 2020 with the price of car bought in year X-1 and sold in 2020 rather than comparing a car bought in year X and sold in 2020 with a car bought in year X and sold in 2021. This difference is important since waiting an extra year to sell a car has a potentially different impact on price than a car being a year older in the year it is sold since market conditions can change. That being stated, we use *(2021 - year manufactured)* as an effective proxy for age since it is still valuable to individuals who wish to glean a rough estimate for price change from, or the opportunity cost of, keeping their car an additional year.

In order to best capture the relationship between car age and price, we took a log transformation of the price target variable for multiple reasons. Statistically, from the histogram of the price distribution shown, we noticed a right skew which suggested the appropriateness of applying the log function to it.

Figure 1: Histogram of Price Distribution



We see a right skew resulting from ultra−luxurious and vintage cars.

Intuitively, by making log of price the target variable, we could prevent the unrealistic scenario of predicted price eventually becoming negative for a very old car. Logically, we also liked the resulting regression coefficient interpretation on age that provides a diminishing return percent change year over year, which is a more realistic interpretation than absolute price changes.

The regression forms fit are below:

$$\widehat{log(price)} = \beta_0 + \beta_1 \cdot (2021 - year\ manufactured) + \mathbf{Z}\gamma$$

Above, β_0 is the constant term and $(e^{\beta_1} - 1) \cdot 100$ represents the percent change from increasing the age of a car by an additional year. $\mathbf{Z}$ is a row vector of additional co-variates, and γ is a column vector of coefficients. This last term serves as an abstraction for the additional predictors used in more verbose models as seen in our results.

Results

Table 1 shows the results of our three regression models. The point estimates for the key coefficient on *(2021 - Year Manufactured)* range from approximately -0.12 to -0.10 and were highly significant across all of our models. Furthermore, the robust standard errors on this key coefficient across all three models create 95% confidence intervals that are significant since they do not include 0. To better understand our model, consider a used car being sold in 2020 that was manufactured in 2015. Applying model 3, this car would command 9.7% less in listing price compared to a car manufactured in 2016. Compared to a car manufactured in 2020, this 2015 car would command 40.0% less in listing price.

Table 1: Estimated Regressions

	Log of Listing Price		
	(1)	(2)	(3)
$(2021 - YearManufactured)$	-0.122^{***}	-0.107^{***}	-0.102^{***}
	(0.002)	(0.002)	(0.001)
Horsepower		0.002^{***}	0.001^{***}
		(0.0001)	(0.0001)
Torque		0.001^{***}	0.001^{***}
		(0.00005)	(0.0001)
Owner Count		-0.064^{***}	-0.044^{***}
		(0.005)	(0.004)
City Fuel Economy		-0.003^{***}	0.016^{***}
		(0.001)	(0.002)
Days on Market		-0.0001	-0.00003
		(0.00004)	(0.00003)
Constant	10.530^{***}	9.814^{***}	7.859^{***}
	(0.008)	(0.030)	(0.170)
Make Name			✓
Body Type			✓
Wheel System Type			✓
Engine Type			✓
Transmission Type			✓
Additional Features			✓
Observations	15,000	15,000	15,000
R^2	0.412	0.753	0.892
Residual Std. Error	0.352 (df = 14998)	0.228 (df = 14993)	0.151 (df = 14877)

Note: HC_1 robust standard errors in parentheses. Additional features are back legroom, engine displacement, damaged frame, front legroom, fuel tank volume, fuel type, accidents, highway fuel economy, is certified pre-owned, is new, latitude, longitude, maximum seating, salvage title, width, length, and height.

Given that we use a log-transformed dependent variable linear regression, our results capture the non-linear decrease in value year over year. This is demonstrated by the fact our model predicts a one-year price decrease of 9.7% but predicts a five-year price decrease of 40.0%. A more simplistic model would predict a 48.5% (9.7 * 5) decrease which would overestimate the change. Our model reflects the true phenomenon of car price depreciating more rapidly in the initial years of ownership.[4] One potentially unrealistic assumption is that depreciation follows a systematic logarithmic pattern.

Limitations

Evaluating the first assumption of independent and identically distributed observations, we believe that there is a very strong chance of geographical clustering. Cars are often priced differently across the United States based on state taxes, differing DMV fees, and registration costs. While we tried to account for this by including the general location from which the car is being sold in our third model, we are still susceptible to regional clustering. Since used car sales occur over time, we may also experience temporal autocorrelation. For one, appraisal sites such as Kelley Blue Book use historical data to price used cars. As an additional thought, periods of inflation or low interest rates may also lead to this autocorrelation or temporal clustering.

To assess the second assumption of the population distribution being represented by a unique best linear predictor, we examined the distribution of price. Since a car can only be worth so much, there is a finite variance despite a slight right skew. The second half of this assumption of no perfect co-linearity is satisfied because we dropped any linearly-dependent columns.

We recognize that our estimates may be biased by omitted variables. None of our current models take into consideration the safety features of a car. Thus, a potential omitted variable could be the presence of sophisticated airbags, which has been required in all passenger vehicles by law since 2007.[5] We can deduce that the added safety of these sophisticated airbags have a positive effect on price. Since they were not mandated until 2007, we expect a negative correlation between their presence and vehicle age (if present, age is lower). The overall OVB is then negative, meaning the direction of the OVB is “away from zero” since our measured coefficient on age is negative. Appropriate analyses can also be done for other possibly omitted variables such as electronic stability control and tire pressure monitoring systems.

Furthermore, our estimates may also be biased by an outcome variable on the RHS. We tried to prevent certain instances of this, such as not including mileage as a predictor since age is highly correlated with mileage as the older a car gets, the more it has been driven usually. However, we still have a potential violation since owner count is a predictor, and it is likely that an older car has had more owners. This means that the owner count coefficient may be absorbing some of the causal impact of age on price.

Conclusion

This study estimates the listing price of a used car based on, primarily, its age and additional features. Our models predict that the percent change from increasing the car’s age by an additional year ranges from -11.5% to -9.7%. This range is found by calculating $(e^{\beta_1} - 1) \cdot 100$, with β_1 ranging from -0.12 to -0.10. Our model accounts for the depreciation in the car’s value over time as we take the log transformation of our dependent variable, the car’s listing price. We recognize more robust estimates could have provided using a linear fixed effects model for car make and is worthwhile exploring in subsequent endeavors.

In future research, we can perhaps estimate the prices that used cars are actually being sold for, given the age and other characteristics of the car, with sales data from multiple years. Sellers may be specifically interested in knowing how much they can expect to earn based on already-sold cars, in addition to the estimated listing prices. The overall goal for this area of research is to offer informative and accessible advice to those who wish to sell their cars on online automotive shopping sites.

[4] https://www.creditkarma.com/auto/i/how-car-depreciation-affects-value

[5] https://www.iihs.org/topics/airbags