Going Once! Going Twice! Prius!

Determining the influence of mileage on the price of the popular Toyota hybrid

Arjuna Keshavan | Darren Benemelis | Hari Nandakumar | Varun Naidu

2023-08-11

Introduction

The Toyota Prius and California are inextricably linked. Whether in the form of memes, catalytic converter theft, or otherwise, there is no doubt that the Toyota Prius represents more than just a small hatchback in the eyes of Californians.

To contextualize the California driving landscape, it is necessary to understand the realities that drivers face in the nation's most populous state. California gas prices have been among the highest in the country for the past two decades (source), and consumers have increasingly been looking for ways to minimize the bill at the pump. One of these ways is through the purchasing of hybrid vehicles - the Toyota Prius being the most common one for much of the last decade. In fact, the Prius was the highest-selling new car in all of California at the beginning of the 2010s decade (source).

As one of the premier car auction companies in California, and thus the whole United States, the used Toyota Prius represents a significant opportunity. Car share programs, dealers, and more are excited to get their hands on Priuses to sell their respective products to customers. While we are still interested in auctioning other vehicles, we are particularly interested in the pricing structure of the Prius given its immense popularity in our state, and the upward trend of gas prices in California.

Our goal is not only to sell Priuses for the highest price point at auction but also to maximize our margins from purchase to auction. To better inform our decision-making process, we study how the mileage (and other traits) of a vehicle impacts our metric of success - the selling price. Utilizing car auction sale data from the past two years (2014 and 2015), we employ several regression models to help us understand this relationship and ultimately help us increase profit margins at auctions.

Data and Methodology

For this explanatory analysis, we used the Used Car Auction Prices dataset (source) obtained from auctions in the United States and Canada. The dataset contains records of used car sales in 2014 and 2015 collected via private scraping methods. Each observation represents one used car sold at auction. The data contains sixteen columns with several categorical, ordinal, and metric features of a car. The units of the odometer, a key column in our analysis, are in miles, whereas the selling price is in US dollars for observations in the US.

As mentioned previously, our primary focus is the sale of the Toyota Prius and its variants in California. Hence, this was our first level filter. We used the state column to filter sales specifically in California, narrowing our dataset to 73148 observations. The make column provides information on the car's brand and was used to filter out all Toyotas in California, and the model column was used to filter all cars that match the Prius name, which gave us the list of Priuses and all its variants sold in California, a total of 418 observations. The resulting dataset contained ten observations without valid data, which were filtered out. The VIN column gives us the Vehicle Identification Number, which is a unique identifier for every car in the United States (source). To maintain uniqueness and independence, we only considered the first sale of each car, i.e. data on the resale of cars with the same VIN were filtered out. We decided to leave out other variants of the Prius (Plug-In, V and C). The Plug-In hybrid was released in 2015 and there is limited data available, and the other variants also have lower observational frequency. Furthermore, the traditional Prius represents the most common business use case for us, given its popularity in California. Our final dataset consists of 355 observations.

To operationalize our data, we return to the concept of the selling price, which is the price at which the car was sold at auction, as well as mileage. Our data includes the sale price at auction in the column "*sellingprice*". This column, which includes the sale price in US dollars, can directly be used as a variable to operationalize the sale price concept, as it tells us the price at which the car was sold at auction. Furthermore, we operationalize the concept of mileage through the "odometer" column/variable, which tells us the cumulative mileage (in miles) of a vehicle. Finally, we can further operationalize concepts relating to car wear-and-tear with the "*condition*" column and the "*saledate*" column. The "condition" column provides a ordinal scale from 1 to 5 of car condition, with 1 being the lowest value and 5 being the highest value. The "*saledate*" column along with the make year of the vehicle can be used to estimate the age of a particular vehicle.

Figure 1: Odometer distribution

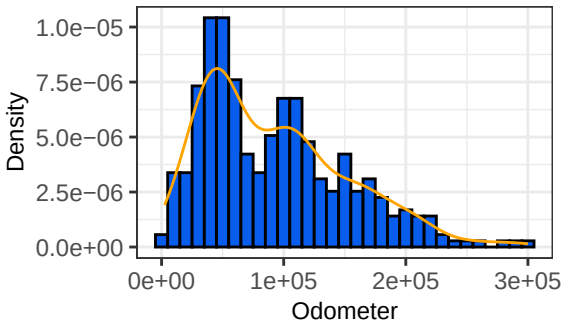


Figure 2: Selling price distribution

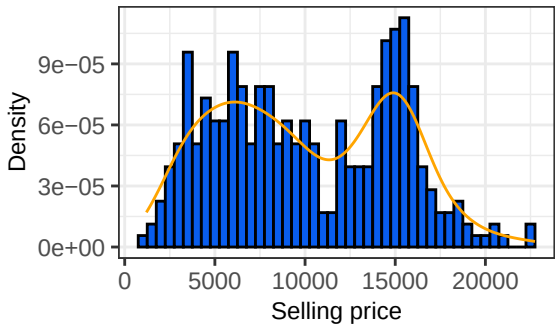


Figure 1 shows the distribution of odometer values. It appears that there is a right skew to the data, with a long tail on the right side of the histogram. There are a few high mileage values, such as one car with 299,999 miles on the odometer. While this seems like an extremely high mileage for a vehicle, anecdotal evidence suggests that there are in fact Priuses that have extremely high mileage, with a dealer reporting encountering a Prius with over 600,000 miles (source). Selling price for these Priuses appears to follow a bimodal distribution, as shown in Figure 2. While exploring the data, we discovered that age of the car also follows a bimodal distribution - when graphing age by selling price, we notice that there appears to be a relationship between the two. This suggests that there may be a local minimum in car auction data availability for vehicles around 5 years old. The maximum sale price is $22,750, which appears reasonable, given that this is still below the MSRP value of the Prius, even in 2015 (source). There was, however, one anomaly that was dropped from the dataset where a Prius with an odometer reading of 1 was sold for $900 - these were minimums for both columns. This appeared to be a significant outlier and looked like a possible mislabel, given that all cars that sold for less than $2000 had over one hundred thousand miles.

Figure 3: Selling Price vs Square Root of Odometer reading of the Toyota Prius

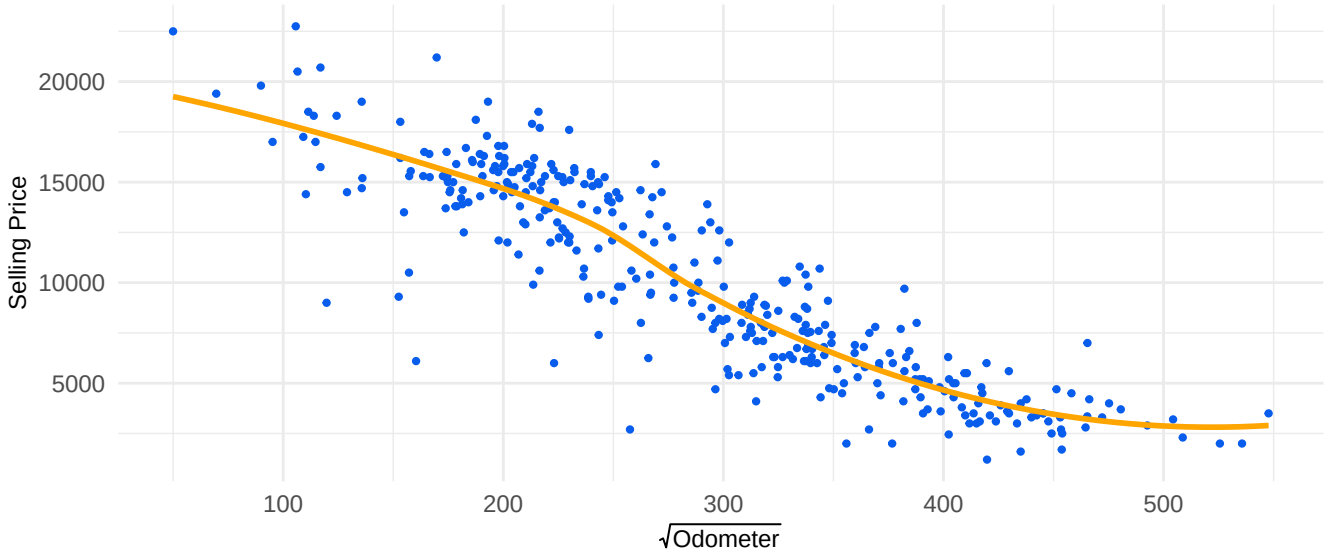


Figure 3 shows the general trend between square root of the odometer reading and selling price for the Prius. Overall, these is a downward trend and a rough linear relationship between the two. We explored a number of transformations and found that the square root of odometer reading produces a more linear relationship with the selling price. We model the interaction between selling price and odometer reading like so -

$$\widehat{sellingprice} = \beta_0 + \beta_1 \cdot \sqrt{odometer} + \mathbf{Z}\gamma$$

where β_1 represents the decrease in selling price per unit increase in the square root of odometer reading. Z represents a row vector of other covariates - condition and age, and γ is a column vector containing their respective coefficients. We also run additional regression models that cumulatively add on each covariate of interest, and one that does not transform the odometer reading. To maximize our margins from purchase to auction, we also look at the influence of car age and condition on the selling price.

Results

Table 1 shows four regression models that were developed to examine the impact of odometer readings on selling price. Each model incorporates various predictors to understand the interplay of factors affecting vehicle values. Across all four models, the coefficients associated with the odometer variable are statistically significant at the $p < 0.001$ level. This robust statistical significance indicates that the odometer is indeed a crucial factor in determining vehicle selling prices. Our main model centers on the relationship between selling prices and the square root of odometer readings. The estimated coefficient value for this predictor is -44.97. In practical terms, this suggests that, holding all other factors constant, for each one-unit increase in the square root of the odometer reading, the selling price is expected to decrease by approximately 0.45%.

Table 1: Regression Results

	Output Variable: Selling Price			
	(1)	(2)	(3)	(4)
$\sqrt{Odometer}$	-44.97^{***}		-40.22^{***}	-24.14^{***}
	(1.27)		(1.39)	(1.42)
Odometer		-0.07^{***}		
		(0.003)		
Condition			$1{,}400.11^{***}$	$1{,}223.28^{***}$
			(177.25)	(185.07)
Age				-676.36^{***}
				(38.38)
Constant	$23{,}054.43^{***}$	$16{,}763.39^{***}$	$17{,}183.77^{***}$	$16{,}896.79^{***}$
	(415.36)	(298.62)	(899.85)	(941.55)
Observations	355	355	355	355
R^2	0.80	0.75	0.84	0.91
Residual Std. Error	2,226.59 (df = 353)	2,473.39 (df = 353)	2,001.04 (df = 352)	1,473.05 (df = 351)
Note:				HC_1 Robust Standard Errors reported.

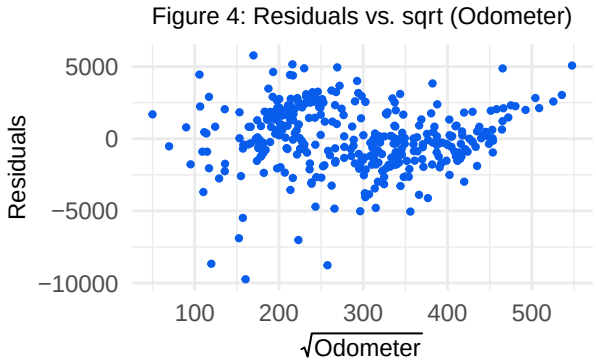
Around 80% of the variability in selling prices is accounted for by the predictors in the model. Upon extending the analysis to include age and condition predictors, the explained variability increases to 91%. This finding reveals that the impact of relatively newer cars in better condition on selling prices is substantial. Interestingly, the coefficient associated with the odometer in this model drops significantly, suggesting that even with significant odometer readings, well-maintained newer cars retain higher values. However, due to a strong correlation between vehicle age and

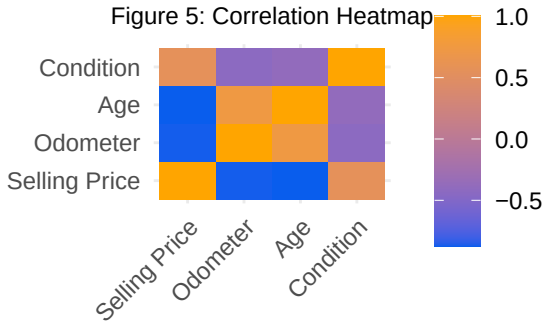
condition, isolating the precise effect of each variable becomes challenging. Consequently, this model might not be suitable for our specific use case. Notably, the model constant of $\$23,054$ holds significance as it mirrors the instant depreciation experienced by vehicles as they leave dealerships with 0 miles on the odometer. This is also portrayed in model 4 (for 0 miles on the odometer, age 0, and condition 5), where we get a base price of approximately $\$22,500$.

Limitations

With 355 observations, our models fall under the Large Sample Regressions umbrella. To achieve dependable regression estimates, our data should be independent and identically distributed (IID) and have a unique best linear predictor. We noticed some potential caveats with respect to the IID assumption. For one, we may not be able to generalize this across seasons, given that all our observations were within the winter/spring seasons. Customer behavior may be seasonally and temporally dependent - thus, our regression is limited by the same seasonal boundaries. Further, Figure 4 shows a random distribution above and below zero without showing any patterns, but the Durbin-Watson test yielded a significant p-value (0.01576). Hence, we don't have strong evidence to support IID. As mentioned previously, a high correlation exists between age and the odometer, and thus our last model (which includes age) would potentially violate the unique BLP assumption (Figure 5). In the residuals plot, we don't see many fluctuations in the size of the residuals as we traverse to the X-axis. However, a lower p-value on the Breusch-Pagan test (0.0053) rejects the null that there is no evidence of heteroskedasticity. However, this is not a prerequisite for large sample models. The histogram of residuals looked approximately normal, and the Q-Q plot comparing residuals against normal distribution presented a straight line except for tails at the end. Despite that, the Shapiro-Wilk test showed a significant p-value rejecting the evidence of normally distributed errors.

For our analysis, we focus on three predictor variables, and this could potentially lead to many omitted variables in our regression model, which could bias our estimates. We identify two such variables, service and accident histories. A negative correlation may exist between the number of accidents and the selling price, as the more accidents a car has gone through, the lower its value. As this negatively affects the selling price, there will exist a positive omitted bias shifting away from zero making the model over-confident. This applies to the number of past owners as well. The opposite could be true for service history. If regular service history is positively correlated, the bias will shift to the left, making the models under-confident. Furthermore, we may consider including generation of Prius as an additional variable, as it is possible that there are significant differences between different sets of Priuses.





Conclusion

Our study assessed the impact of mileage on the selling price of Toyota Priuses in California. We found that cars that have run more than 200,000 miles are valued at less than $\$3,000$, which could impede our profit margins. We also looked at the effect the age and condition of cars have on their value. This indicates that an assessment of the physical condition of the car may be needed to get a more accurate quote, and newer cars can confidently be sold at higher prices. Ultimately, we believe that this tool can be helpful in our company's analysis of prospective vehicles. For further investigation, it could be useful to take selling point snapshots at different periods, so that we will be able to understand the temporal effect on car prices as well. Furthermore, it would be useful to try to obtain data for other facets that affect car prices, like service history and whether or not the title is clean. We hope this analysis provides a starting point for us to continue estimating a Prius' value before it hits the lot so that we may maximize our margins in the future.