Used Car Market Around Clemson, SC: Prices and Predictors.

Alex Matandos

Introduction

Data from the St. Louis Federal Reserve points out that prices that urban consumers face for new and used cars have risen from June, 2020 to October, 2021 by the staggering amount of approximately 23.4%. Reasons behind such could be more than tenfold, due to the effects of COVID-related restrictions: disruption of global supply chains (labor and inventory bottlenecks), rising energy prices, geopolitical instability, hoarding by sellers, and so forth.

Regardless of the reason, the scenario presented represents an interesting opportunity to investigate, once again, the used car market and gauge the significance of certain predictors for prices. Thus, the author have scraped the craigslist webpage for the used car market around the Clemson, SC area, being sold by the owners, and obtained the price and the following covariates: car age, mileage, title status, condition, type of fuel, type of transmission, drivetrain, engine, color, size, type, and VIN (Vehicle Identification Number) availability.

For the regression, the main independent variables were the odometer readings and the age of the car reported on the listings as the predictors of asking prices. Controls were added to gauge the effects of the aforementioned regarding the condition of the car and whether the car was from a Japanese manufacturer.

Literature Review

Due to the considerable information asymmetry and uncertainty in the decision-making process, the car market represents an interesting place to research economic behavior.

Akerlof (1970) discuss the used car market through a theoretical framework, as an example of market interaction under uncertainty. While crudely representing reality, by the author's own statements, important conclusions are derived: (1) that information asymmetry arises as sellers can properly gauge the quality of the car being sold, having driven it for a given length of time, during its ownership; (2) as buyers have less knowledge about the quality of the car, "lemons" can be sold by the same price as good cars; (3) but sellers are sufficiently aware that used cars cannot have the same valuation as the brand new (otherwise, one could sell its own "lemon" by the price of the showroom model, and buy the latter with the money from the trade), thus the uncertainty due to the lack of differentiation, from the buyer's perspectives, results in no good cars being offered in the market at all , as prices shall not meet neither the true value of his car nor the expected value of a new car.

Jae-Cheol Kim's paper from 1985 expands on Akerlof (1970) by adding the following: (1) agents are free to shift between being buyers and sellers, as transaction costs for such are null; (2) quality of used cars are endogenous, as a function of the maintenance level; (3) no matter the maintenance level of the used car, the only value of maintenance that matters is the one done in the period in which the car was new; (4) the model now features two periods; (5) agents, faced with a von Neumann-Morgenstern utility function, and starting with no car in the first period, can choose to (a) buy a new car, maintain it, and sell it in the next period; (b) buy a new car, maintain it, and keep it; (c) buy a used car, and sell it in the second period; or (d) buy no car at all. Thus, an equilibrium different from the one feature in Akerlof (1970) may arise: the actual better quality used cars are the ones being traded if there are more agents that are indifferent between scenarios (a) and (b).

Berry, Levinsohn, and Pakes (1995) developed techniques for analyzing demand and supply with data for US automobile industry, and derive equilibrium conditions. Their data on product characteristics come from annual issues of the *Automotive News Market Data Book*, from 1971 to 1990. Their OLS model for prices (in log scale), derived from marginal cost pricing when setting the markup term equal to zero, used as covariates "cost-shifters", similar to hedonic price regressions used on real estate. The authors find that all their predictors (horsepower-weight ratio, air conditioning, fuel economy (MPG), size, and trend (related to government regulation in the market)) all were positively correlated to price increase, except for fuel economy.

Data Description

Data was obtained through web scrapping craigslist, the American classified advertisements websites, for used cars around Clemson, South Carolina region, on November 12, 2021. The "uncleaned" data features 519 observations and the following variables were obtained: manufacturer and model, year, mileage, fuel type, title status, transmission, engine cylinders, drivetrain, paint color, size, type, condition, Vehicle Identification Number, and nationality (we separated according to the three most important producers: USA, Germany, and Japan), and so forth, totaling 19 variables.

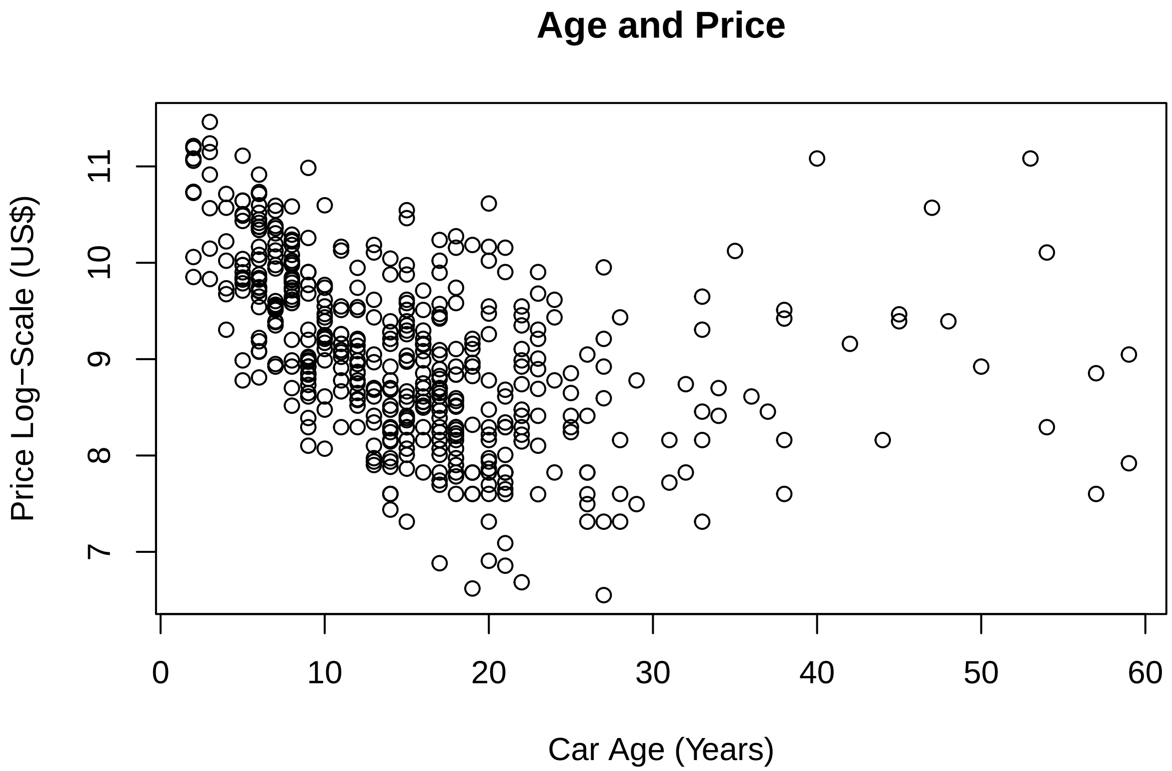
Raw data, such as the one used, is bound to feature discrepancies that the web scrapping program itself cannot overcome. Thus, some corrections are needed to present the data more clearly. Among them: (1) cars with blank values for the mileage represent either an broken or rolled back odometers, and they've been dropped; (2) strange values for odometer readings (i.e. single digits or other absurdly high/low values) have been fixed by searching for true mileage in the description for the given advertisement, otherwise dropped; (3) removed observations regarding advertisements that were not about cars on sale (e.g. truck trailer, house, etc.); (4) similarly, prices were corrected when possible, otherwise observations were dropped; (5) removed double entries by checking inside metadata which represented the most recent (represented by a later expiring date of the advertisement). Thus, we reach the number of 471 unique observations.

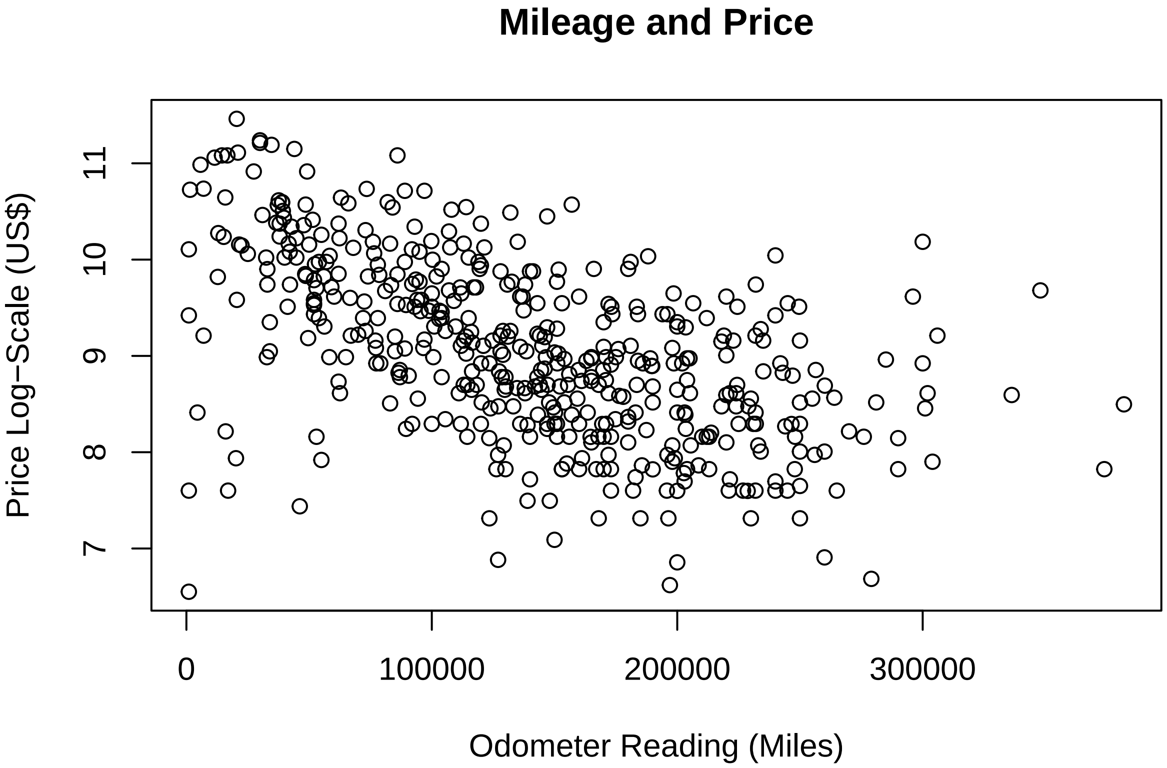
Summarizing the data, we have the following: average asking price in the sample is $13,006; average odometer reading is 139,614 miles; average car is 15.69 years old; 'like new' and 'excellent' conditions represent 38.85% of the sample; American cars represent 21.02% of the observations, Japanese cars, 28.66%, and Germans, 9.554%.

Table 1: Variables Summary

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
PRICE	471	13, 006.140	13, 518.210	700	4, 350	16, 525	95, 000
MODEL.YEAR	471	2, 006.314	9.671	1, 963	2, 002	2, 013	2, 020
CAR.AGE	471	15.686	9.671	2	9	20	59
JAPAN	471	0.287	0.453	0	0	1	1
GERMANY	471	0.096	0.294	0	0	0	1
USA	471	0.210	0.408	0	0	0	1
ODOMETER	471	139, 614.200	73, 518.930	1, 000	86, 000	190, 000	382, 000
QUALITY.CAR	471	0.389	0.488	0	0	1	1
CYLINDERS	337	6.012	1.545	4.000	4.000	8.000	8.000
LOG.PRICE	471	9.051	0.925	6.551	8.378	9.713	11.462

Plotting prices in logarithmic scale on car ages and mileage reveals the expected behavior of such predictors: older cars and/or with more usage represents a depreciated asset, thus value decreases, indicating an negative correlation between these covariates (increasing age and/or mileage) and prices.


Age and Price
Price Log−Scale (US$)
7
8
9
10
11
0
10
20
30
40
50
60
Car Age (Years)



Methods

Generally speaking, modelling car prices that sellers offer to potential buyers revolve around the following: what is common knowledge to both, what is information that's only available to the seller, and buyer preferences.

For purposes of simplicity, let's consider that the common knowledge is represented by the covariates scraped in craigslist; nonetheless, such variables only represent approximations regarding the true value of the car (even when the overall condition is reported) as only the owner has the correct measure of it (e.g. recklessness of the driver, receipts of oil changes and maintenance, damage not reported, and so forth) even though services like CARFAX may greatly reduce transaction costs (given that the potential buyer has access to the VIN).

With that said, let's start with the following base model (1):

$$log(PRICE)_i = \beta_0 + \beta_1 LOG.ODOMETER_i + \beta_2 CAR.AGE_i + u_i$$

Where the dependent variable is the asking price on log scale, $ODOMETER$ is the mileage, and $CAR.AGE$ is how old the car is (in terms of years). Makes sense to use both these variables as both are an indicative of the depreciation of the automobile, either through use and/or time.

For further detail, it is added the variable $QUALITY.CAR$, that takes the value of 1 if the observation is either described as 'like new' or 'excellent' in terms of condition, and 0 otherwise (2):

$$\begin{aligned} log(PRICE)_i = \beta_0 + \beta_1 LOG.ODOMETER_i + \beta_2 CAR.AGE_i + \beta_3 QUALITY.CAR_i \\ + \beta_4 LOG.ODOMETER_i QUALITY.CAR_i + \beta_5 CAR.AGE_i QUALITY.CAR_i + u_i \end{aligned}$$

Japanese car manufacturers are known for their reliability. To gauge if there's an impact on prices, we add the dummy variable $JAPAN$ to indicate if the vehicle is from a Japanese brand or not (3):

$$\begin{aligned}
log(PRICE)_i = \beta_0 + \beta_1 LOG.ODOMETER_i + \beta_2 CAR.AGE_i + \\
\beta_3 QUALITY.CAR_i + \beta_4 JAPAN_i + \beta_5 LOG.ODOMETER_i QUALITY.CAR_i \\
+\beta_6 CAR.AGE_i QUALITY.CAR_i + \beta_7 LOG.ODOMETER_i JAPAN_i + \\
+\beta_8 CAR.AGE_i JAPAN_i + \beta_9 LOG.ODOMETER_i QUALITY.CAR_i JAPAN_i + \\
\beta_{10} CAR.AGE_i QUALITY.CAR_i JAPAN_i + u_i
\end{aligned}$$

As mentioned previously, important variables are missing that clearly would impact the asking price, were it common knowledge: maintenance records, how many previous owners, damage and accident reports, and so on. Using the VIN, further more, only partially helps, as access to the car record sheet is behind a paywall (40 dollars per report, as of November 30, 2021, on CARFAX). Thus, it's safe to say that the hypothesis of unbiased parameters will not be met as $E(u_i|x_i) \neq 0$, and it is not possible to gauge causal relationships. Nonetheless, the other OLS assumptions does hold, as the sample is random for the region considered, the model is linear in its parameters, and no perfect linear relationships between the covariates.

Results

For model 1, both "LOG.ODOMETER" and "CAR.AGE" have their expected *ceteris paribus* effects on "LOG.PRICE: a 1% increase in odometer reading is associated, on average, with 0.45% decrease in prices; a increase in the age of the car by one year, on the other hand, is correlated with a decrease in prices by 3.5%. Their economic impact is quite relevant and both parameters are statistically significant for 99% confidence level.

Table 2: Regression Results

	Dependent variable:		
	LOG.PRICE		
	(1)	(2)	(3)
LOG.ODOMETER	-0.4566923000^{***} (0.0409922600)	-0.2718172000^{***} (0.0522920200)	-0.3044782000^{***} (0.0610514900)
CAR.AGE	-0.0351241800^{***} (0.0036202670)	-0.0233152800^{***} (0.0042930880)	-0.0196464900^{***} (0.0046315280)
QUALITY		3.5118290000^{***} (0.9601104000)	4.2218140000^{***} (1.1273680000)
JPN			-2.5966550000^{*} (1.3831530000)
LOG.ODOMETER*QUALITY		-0.2544280000^{***} (0.0831151500)	-0.3200410000^{***} (0.0976906600)
CAR.AGE*QUALITY		0.0008013752 (0.0083809770)	0.0015885150 (0.0088567020)
LOG.ODOMETER*JPN			0.2519552000^{**} (0.1158203000)
CAR.AGE*JPN			-0.0357407000^{***} (0.0116785400)
QUALITY*JPN			−1.6941280000 (2.0765530000)
LOG.ODOMETER*QUALITY*JPN			0.1521020000 (0.1823470000)
CAR.AGE*QUALITY*JPN			−0.0092466430 (0.0245884200)
Constant	14.9082100000^{***} (0.4772800000)	12.3267200000^{***} (0.6372302000)	12.7080800000^{***} (0.7453517000)
Observations	471	471	471
R^2	0.3325941000	0.4193214000	0.4556918000
Adjusted R^2	0.3297420000	0.4130776000	0.4426474000
Residual Std. Error	0.7574175000 (df = 468)	0.7087689000 (df = 465)	0.6906839000 (df = 459)
F Statistic	116.6112000000^{***} (df = 2; 468)	67.1574500000^{***} (df = 5; 465)	34.9338400000^{***} (df = 11; 459)
Note:			*p<0.1; **p<0.05; ***p<0.01

In any case, the model has “QUALITY.CAR” omitted, meaning that the parameters for both “LOG.ODOMETER” and “CAR.AGE” suffer from a negative bias, i.e., without controlling for quality, mileage and age of the car have a bigger effect on prices than their actual impact. Thus, as expected, holding all else constant, the individual effects of “LOG.ODOMETER” and “CAR.AGE” are lower in absolute terms (0.27% and 23% decrease in prices, for a 1% increase in mileage and one additional year, respectively). Interesting enough, quality cars are “penalized” in terms of asking prices, as a 1% increase in mileage is associated, all else equal, to approximately a 0.52% decrease in prices, and it is statistically significant. While algebraically such result makes sense, economically it might make for a collective head-scratching: first of all, there’s clearly some endogeneity as the seller inputs his car condition, there’s no uncommitted third party doing the actual verification of the vehicle; second, not all user made sure to report their car actual condition, thus the variable may not correctly report such.

For the final model it is controlled for Japanese cars. Now, increased odometer readings by 1% are correlated with only 0.05% increase in prices, holding all else constant, and was statistically significant for a 95% confidence level. Further controlling with quality, “like new” or “excellent” condition Japanese cars actually increases 0.10% in prices for each increase of 1% in mileage, *ceteris paribus*. While some caution is advised as the statistical significance of the latter is rather low, the economic relationship is somewhat interesting but unrealistic: the assumption of car prices rising over time due to a prolonged supply shock is fine but increased miles should, with the obvious exception of collector’s cars, be correlated with lower prices.

“CAR.AGE” did not deviate from what it was expected in models 2 and 3: older cars are, *ceteris paribus*, correlated with price decreases.

Limitations and Extensions

For further research, and with backing from the institution supporting such, the next step would be to gather even more by adding the information available on services such as CARFAX through the VIN available on some of the postings. Additional variables would reduce the bias due to omission.

Unlike Berry, Levinsohn, and Pakes (1995), the analysis done here wasn’t in the mold of the hedonic price regressions. In other words, some features present in the cars of the sample were not accounted for and they definitely exert influence over asking prices. Thus, another suggestion for enhancing the model would be to add such variables onto the dataset, through sources like the aforementioned *Automotive News Market Data Book*.

craigslist also offers geolocation data for the listings. Thus, spatial clustering and other geographical analysis could be done through softwares such as ArcGIS, and create even more interesting results.

Conclusions

If anything, the main conclusion one could have arrived through this brief paper is that scraping data from craigslist is cumbersome and somewhat unreliable. On a more serious note, the “common Joe and Jane” attitude around the website allows for gauging the asking prices at individual level, rather than at retailers and dealerships, which due to scaling or/and bargaining power can markup prices.

Thus, the OLS regression didn’t move away from the norm in terms of what is expected in the used car market: increased mileage and older cars are both correlated with lower asking. A interest result arose when controlling for quality: cars in “like new” or “excellent” condition actually were correlated with lower asking prices for percentage increases in mileage, but such result is likely due to bad data. When controlling for Japanese cars, prices actually decreased at a lower rate for each percentage increase in mileage, and was statistically significant. Perhaps indicating that Japanese autos may sell at a premium on the second hand market, and perhaps proving their claim to fame as reliable machines.

References

Akerlof, George A. *The Market for "Lemons": Quality Uncertainty and the Market Mechanism.* The Quarterly Journal of Economics 84, no. 3 (August 1970), 488-500.

Berry, Steven, James Levinsohn, and Ariel Pakes. *Automobile Prices in Equilibrium.* Econometrica 63, no. 4 (July 1995), 841-890.

FRED. *Consumer Price Index for All Urban Consumers: New and Used Motor Vehicles in U.S. City Average.* Accessed November 21, 2021.

Jae-Cheol, Kim. *The Market for "Lemons" Reconsidered: A Model of the Used Car Market with Asymmetric Information.* The American Economic Review 75, no. 4 (September 1985), 836-843.

Matandos, Alex. *alexmatandos/craigslist_cars_scrapper.* GitHub. November 30, 2021. https://github.com/alexmatandos/craigslist_cars_scrapper