Find your first car in San Jose

Author(s): Ren Zhi-yi, Chen Yu-kai, Xu Jie-mo
Class: 2nd year of SJSU-FCU 2+2 Bachelor's Program in Business Analytics
Student ID: D0571956, D0565634, D0571990

Instructor: Dr. Cathy W. S. Chen

Course: Introduction to Data Analytics

Department: International School of Technology and Management

Academic Year: Semester 1, 2017-2018

Abstract

This paper examines the relationship between the price of used cars in San Jose on December 23, 2017 from Carfax website. We use the multiple regression analysis to investigate 9 explanatory variables and use Stepwise Selection approach to select the best fitted model. The collected data shows the mileage which has been driven has negative correlation with second-hand cars' price. The drive wheel types such as All Wheel Drive (AWD), Front-Wheel Drive (FWD) and Rear-Wheel Drive (RWD) indicate the negative correlation as well. From the perspective of customers, they suggested that they pay more attention to the year of model and the number of images that sellers upload to the website. Second-hand cars, which are new models and have more pictures may have higher price. The quantity of cylinders that the engine has, engine displacement, the amount of miles when consuming a gallon of gasoline (MPG), whether it is for personal or business use and gearbox have strong positive correlation among our 196 randomly selected observations.

Keyword: Influential Point, Model Selection, Multicollinearity, Regression Analysis, Outlier, Residuals Analysis, Second-hand Car.

Table of Contents

1. Introduction

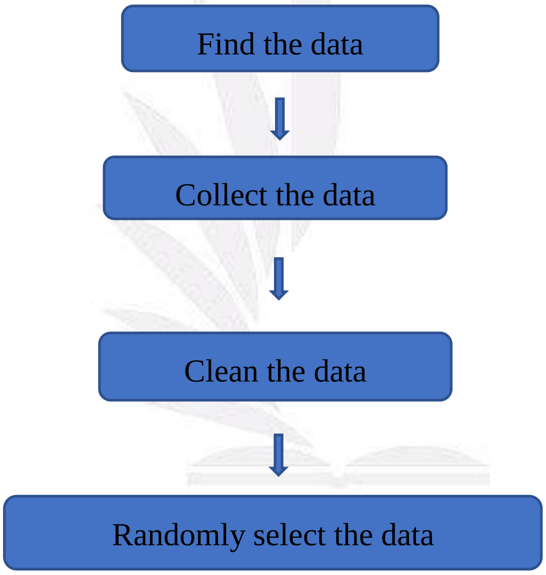
Our department cooperates with San Jose State University and we students have the opportunity of studying abroad in California to receive a dual degree. However, the area of San Jose State University lacks public transport, such as MRT (Mass Rapid Transit) system. Residents usually drive cars for their commute. Our motivation is to help our classmates find their first cars that are high quality but low price in San Jose State. In addition, we hope to provide this resource as one available reference for those marketing researchers to assess second-handed cars' price. Purchasing used cars can save money and we will not lose too much once we would like to sell it again. Moreover, used cars do not have plastic smell. Thus, purchasing second-hand cars is a wise choice. In this study, we will be considering what factors affect the price of used-cars.

The regression analysis is a very advanced statistical tool to learn more about the relationship between several independent or predictor variables and a dependent or criterion variable. For instance, personnel professionals customarily use multiple regression procedures to determine equitable compensation nowadays. With the maturity of people's concept of purchasing cars as well as more and more products that American automotive market can provide, car demanders can choose more models of cars in second-hand car markets. Purchasing used car is more suitable for young consumers' consumption ability. At the same time, according to its high hedging rate, affordable features, it is also very suitable for students who study abroad. Of course, our report has its limitations. We did not take the tax implications of buying a car into consideration.

2. Method

2.1 flow chart

Our data resource came from the website of US biggest used car chain store--Carmax. We used the “Web Crawler” by Python to collect predictor variables that we are interested in such as mileages, number of photos, etc. Then we use Excel to delete several data with missing value. After rearranging data, we got 7724 pieces of data in our dataset. Finally, we randomly picked up 196 observations by SAS as our research sample.



2.2 Explanatory variable and response variables

Table 1 presents the explanatory variable and response variables.

Explanatory Variable	Description
Price	Used car’s price in San Jose. We use the dollar as the unit.

Table 1. Explanatory variable description

Response Variable	**Description**
Mileage	A number of how many miles a vehicle has traveled or covered.
Year	The model year of a vehicle.
imgcount	Numbers of photos that the sellers upload to the website.
engine	How many cylinders a vehicle has.
displacement	The swept volume of all the pistons inside the cylinders of a reciprocating engine in a single movement.
mpg	The number of miles that the car can drive when consuming a gallon of gasoline. Usually, it can test the efficiency of vehicles.
Person	1 if a vehicle is for personal use; otherwise 0.
Automatic	1 if a vehicle's transmission is automatic; otherwise 0.
AWD	1 if a vehicle's transmission system is All-Wheel Drive; otherwise 0.
FWD	1 if a vehicle's transmission system is Front-Wheel Drive; otherwise 0.
RWD	1 if a vehicle's transmission system is Rear-Wheel Drive; otherwise 0.

Table 2. Response variable description

2.3 Descriptive statistics

Table 3 provides the basic summary statistics analysis including mean, standard deviation, minimum, maximum and others. We notice that the ranges of "Price" and "Mileage" are quite large. Therefore we consider to transform these two variables by taking the natural logarithm so that we can figure out the result more significant.

Variable	Mean	σ	Minimum	Q1	Median	Q3	Maximum	Range
Price	22346.84	12149.14	4899.00	13992.00	19410.00	29988.00	71984.00	67085.00
Mileage	49204.56	42114.30	827.00	23948.50	35791.00	64512.00	282584.00	281757.00
Year	2013.71	3.20	2000.00	2012.00	2015.00	2016.00	2018.00	18.00
imgcount	22.63	13.02	0.00	15.50	21.50	30.00	63.00	63.00
engine	5.15	1.42	3.00	4.00	4.00	6.00	8.00	5.00

displacement	2.87	1.12	1.40	2.00	2.50	3.50	6.40	5.00
mpg	24.61	6.18	14.00	20.00	24.00	28.00	52.00	38.00
Person	0.68	0.47	0.00	0.00	1.00	1.00	1.00	1.00
Automatic	0.96	0.20	0.00	1.00	1.00	1.00	1.00	1.00
AWD	0.17	0.38	0.00	0.00	0.00	0.00	1.00	1.00
FWD	0.46	0.50	0.00	0.00	0.00	1.00	1.00	1.00
RWD	0.30	0.46	0.00	0.00	0.00	1.00	1.00	1.00

Table 3. Summary statistics before taking natural log

Table 4 shows the results after we log the two independent variables

Variable	Mean	σ	Minimum	Q1	Median	Q3	Maximum	Range
Price	9.87	0.54	8.50	9.55	9.87	10.31	11.18	2.69
Mileage	10.46	0.89	6.72	10.08	10.49	11.07	12.55	5.83
Year	2013.71	3.20	2000.00	2012.00	2015.00	2016.00	2018.00	18.00
imgcount	22.63	13.02	0.00	15.50	21.50	30.00	63.00	63.00
engine	5.15	1.42	3.00	4.00	4.00	6.00	8.00	5.00
displacement	2.87	1.12	1.40	2.00	2.50	3.50	6.40	5.00
mpg	24.61	6.18	14.00	20.00	24.00	28.00	52.00	38.00
Person	0.68	0.47	0.00	0.00	1.00	1.00	1.00	1.00
Automatic	0.96	0.20	0.00	1.00	1.00	1.00	1.00	1.00
AWD	0.17	0.38	0.00	0.00	0.00	0.00	1.00	1.00
FWD	0.46	0.50	0.00	0.00	0.00	1.00	1.00	1.00
RWD	0.30	0.46	0.00	0.00	0.00	1.00	1.00	1.00

Table 4. Summary statistics before taking natural log

2.4 Scatter plot

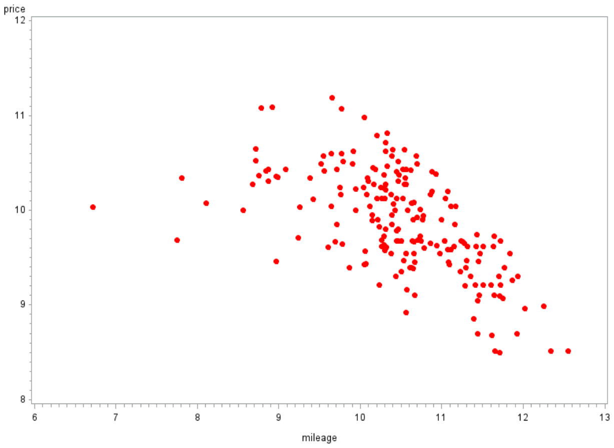


Figure 1.1 mileage versus price

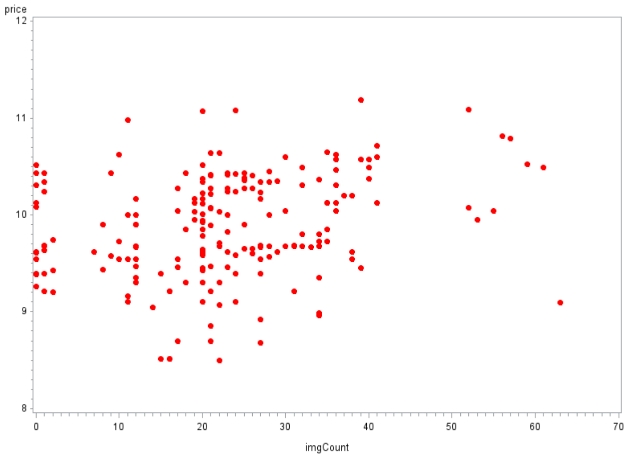


Figure 1.2 imgCount versus

Figure 1.1 illustrates a negative correlation between mileage and price , which indicates more miles the second-hand cars have been driven may cause lower price.

As the plots in figure 1.2 show, there is a positive correlation between used cars' price and the number of photos uploaded in description pages .

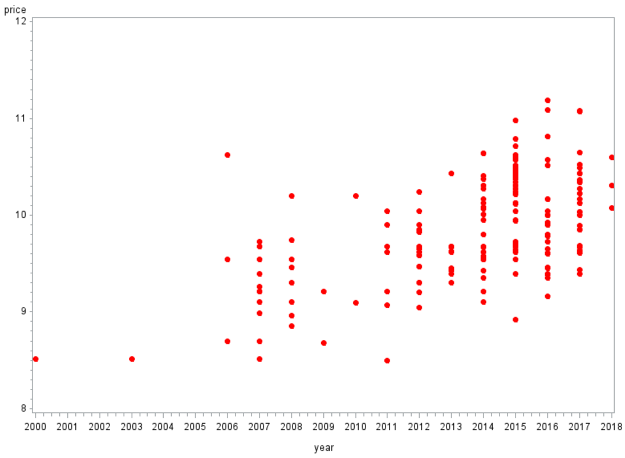


Figure 1.3 year versus price

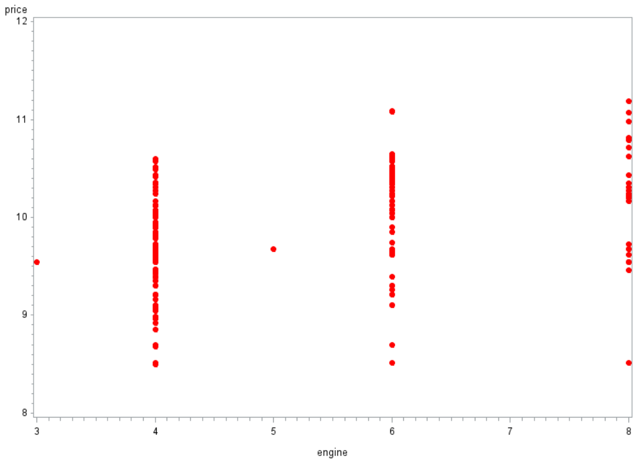


Figure 1.4 engine versus price

Figure 1.3 shows a positive correlation between the model year of cars and sales price. The newer types of cars they are, the higher price they will cost. A positive correlation can be observed in Figure 1.4, which means more cylinders may lead to higher price.

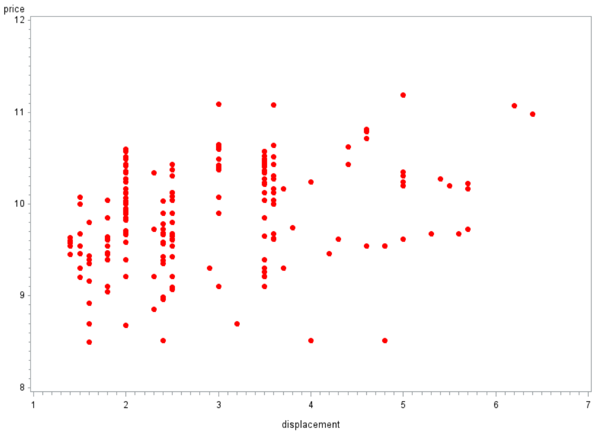


Figure 1.5 displacement versus price

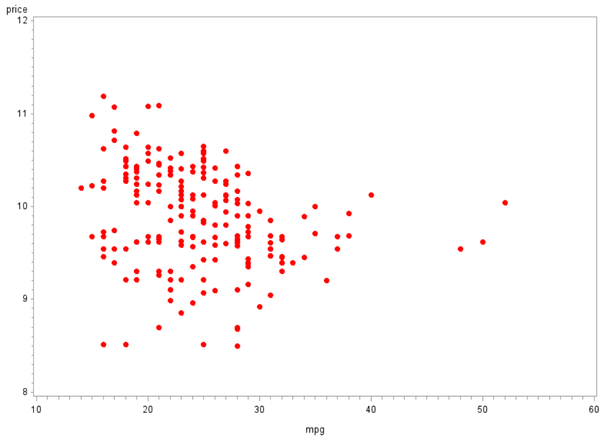
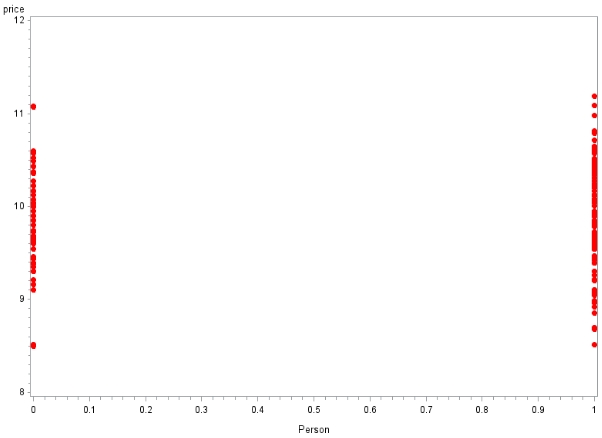


Figure 1.6 mpg versus price

The scatter plot shown in Figure 1.5 has positive correlation relationship which indicates that the larger displacement is, the higher price they have. Figure 1.6 renders a positive correlation in mileage versus engines, which demonstrates the more cylinders a vehicle has, the higher price the seller may sell it out.



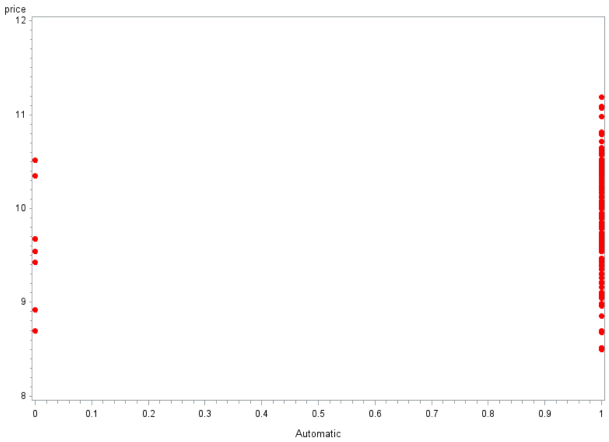


Figure 1.7 Person versus price

In the figures 1.7 and 1.8, the scatter plots of 2 dummy variables test whether personal use or automatic transmission have effect. In these patterns, the price will not be affected by the 2 dummy variables.

Figure 1.8 Automatic versus price

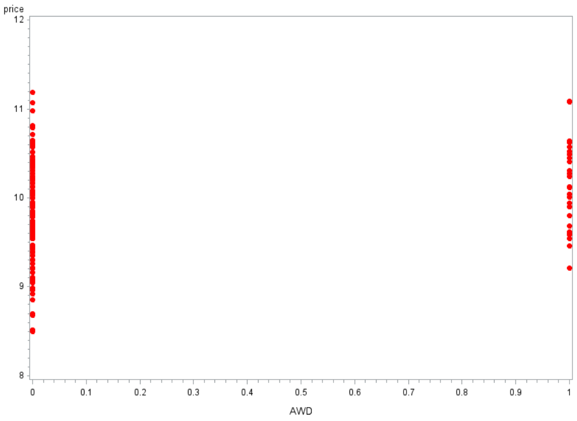


Figure 1.9 AWD versus price

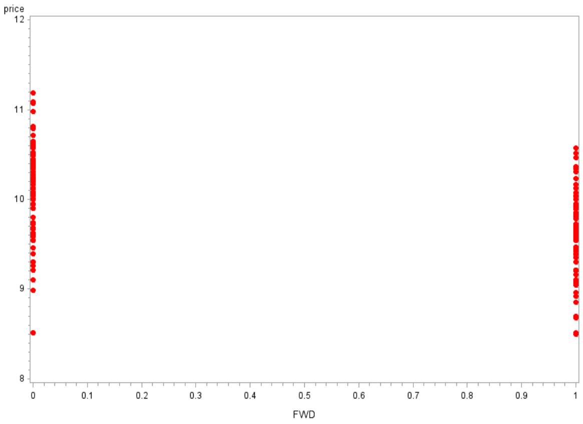


Figure 1.10 FWD versus price

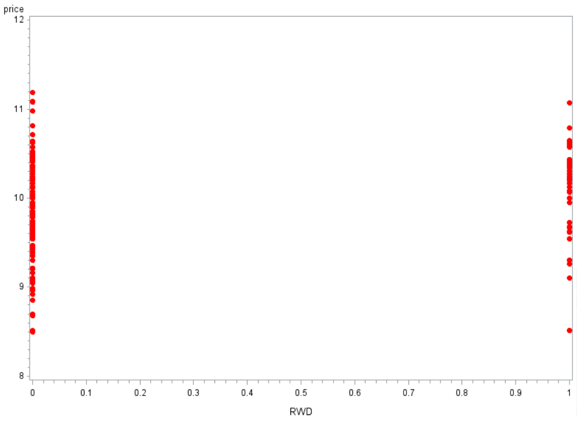


Figure 1.11 RWD versus price

In Figures 1.9, 1.10 and 1.11, the scatter plots of 3 dummy variables show the price is not affected by the wheel drive type, whether it is all-wheel-drive or front-drive-type or rear-wheel-drive.

2.5 Correlation plot

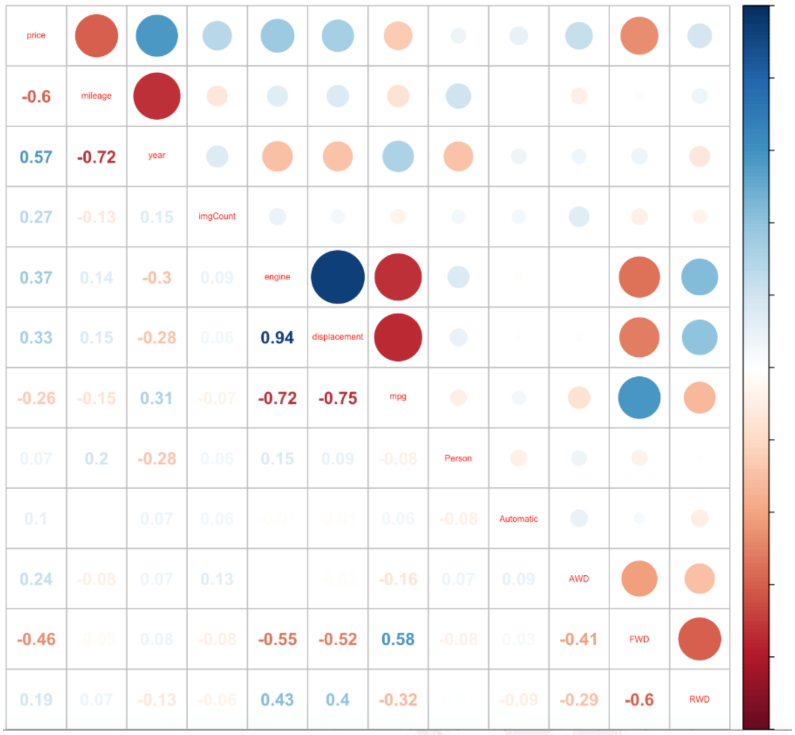


Figure 2 correlation matrix of variables

The correlation matrix table reflects price and mileage; mpg; FWD have positive correlations. That the price has significant positive correlation with mileage shows the less miles they drive, the higher price they may sell.

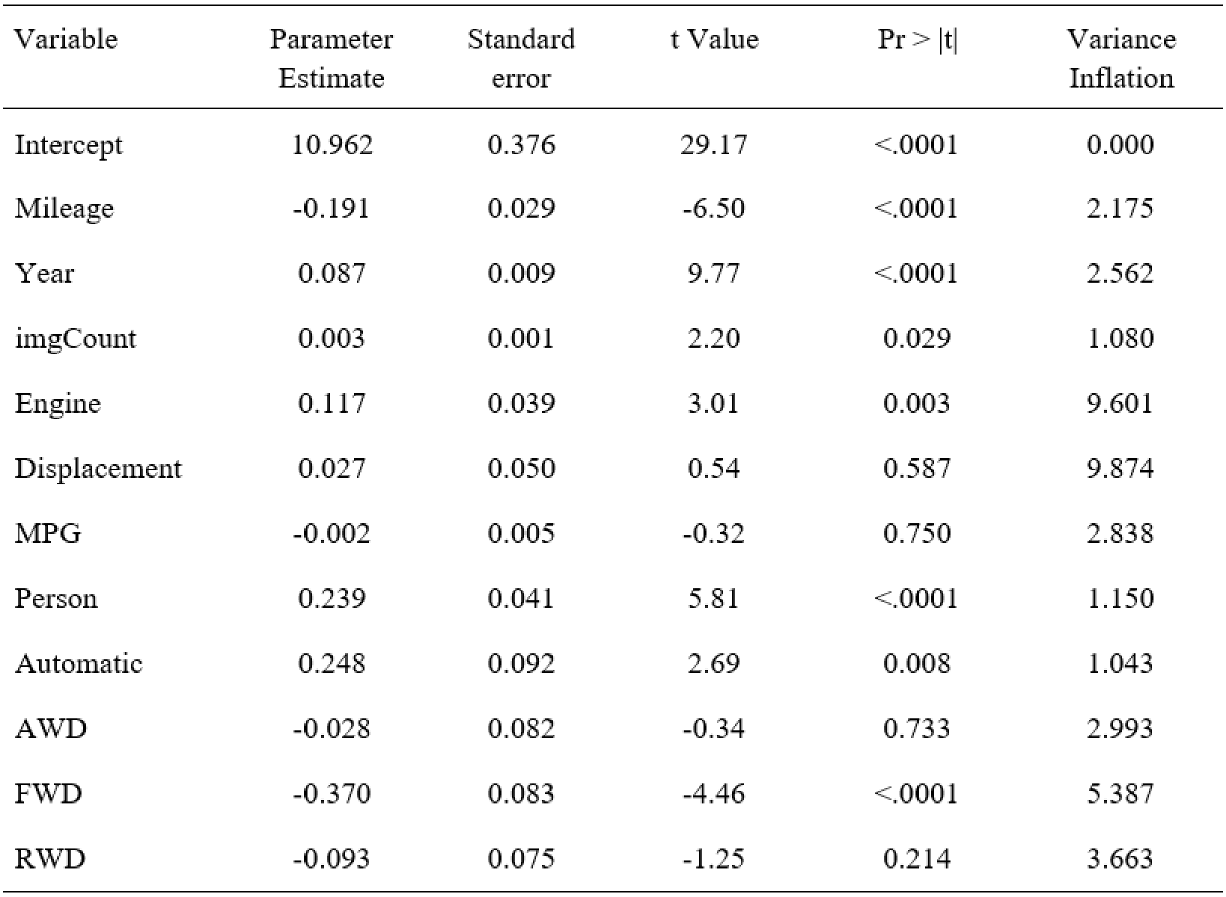
First, between the engine and displacement, there is a significant positive relationship. In normal situation, when the engine has more cylinders, the displacement will be larger.

Secondly, the relationship between displacement and mile per gallon shows a negative relationship. The larger the displacement is, the more gasoline is needed to mix with air.

Thirdly, since engine and displacement have highly positive correlations, that the more engines the car has means the displacement is larger as well, hence, the engine and MPG also show a negative relationship.

In the end, the meaning of "year" in our variables is the model year of the car, but not the year that most of the people consider as how many years one vehicle has been produced. So, they have negative correlation between year and mileage.

2.6 Full Model and diagnose multicollinearity



Variable	Parameter Estimate	Standard error	t Value	Pr > |t|	Variance Inflation
Intercept	10.962	0.376	29.17	<.0001	0.000
Mileage	-0.191	0.029	-6.50	<.0001	2.175
Year	0.087	0.009	9.77	<.0001	2.562
imgCount	0.003	0.001	2.20	0.029	1.080
Engine	0.117	0.039	3.01	0.003	9.601
Displacement	0.027	0.050	0.54	0.587	9.874
MPG	-0.002	0.005	-0.32	0.750	2.838
Person	0.239	0.041	5.81	<.0001	1.150
Automatic	0.248	0.092	2.69	0.008	1.043
AWD	-0.028	0.082	-0.34	0.733	2.993
FWD	-0.370	0.083	-4.46	<.0001	5.387
RWD	-0.093	0.075	-1.25	0.214	3.663

Table 5. Parameter Estimate of the full model

From table 5, the full model is:

$$ln\left(\widehat{price}\right) = 10.962 - 0.191ln(mileage) + 0.087year + 0.003imgCount + 0.117engine + 0.027displacement - 0.002MPG + 0.239person + 0.248automatic - 0.028AWD - 0.370FWD - 0.093RWD$$

Adjusted R² for this model is 0.7888, and $\hat{\sigma}^2$ is 0.06221.

Testing whether it is multicollinearity depends on variance inflation factor(VIF).

Multicollinearity is a problem that we can run into when we're fitting a regression model. As usual, we use 10 as our standard.

In our dataset, the outputs above show that the VIF don't go over 10.

Hence, it refers to predictor variables that are not correlated with other predictors in the model.

If VIF>10, we can assume that the regression coefficients are poorly estimated due to multicollinearity, which means it has high multicollinearity.

And we can solve Multicollinearity by redefining variables, principal components, use biased estimation – ridge regression, standardized coefficients.

2.7 Discuss the outliers and influential points

Output Statistics

Obs	Residual	Std Error Residual	Student Residual	Cook's D
159	0.7607	0.235	3.242	0.127
187	0.7566	0.244	3.100	0.035

Table 6. detect the outliers and influential points

We consider Residual and Student Residual test to detect outliers and use Cook's D to identify any influential points. There are 2 observations(No.159 and No.187) whose Student Residual are greater than 3 and Residual is greater than $3\hat{\sigma}$=0.748.Therefore, it has 2 observations which can be considered as outliers.For influential point, every observation's Cook's D is less than 0.5, there is no influential point in the dataset. Since these 2 pieces of data do not affect the estimated model too much in general, we do not remove the outliers.

2.8 Model Selection

According to Freund, Wilson, and Sa (2006), we can use 5 approaches to fit the regression model. Table 7 provides the selected factors in different selection procedures. And the variables selected are marked as “√”. From table 7, one weakness of Adjusted R-Square Selection and Mallow’s C_p we can notice is that they can not recognize the group of dummy variables. Finally, we choose the result of Stepwise Regression selection as our best fitted model.

Selection procedures	Slentry/Slstay	Mileage	Year	ImgCount	Engine	Displacement	mpg	Person	Automatic	AWD	FWD	RWD
Backward Elimination	Slstay=0.15	√	√	√	√			√	√	√	√	√
Forward Selection	Slentry=0.15	√	√	√	√			√	√	√	√	√
Stepwise Regression	Slstay=0.15	√	√	√	√			√	√	√	√	√
	Slentry=0.15											
Adjusted R-Square Selection		√	√	√	√			√	√		√	√
Mallow’s C_p		√	√	√	√			√	√		√	√

Table 7. Model Selection

Backward Elimination

Backward elimination starts with all variables, tests the deletion of each variable if p-value is greater than 0.15 and deletes the variable whose loss gives the most statistically insignificant deterioration of the model fit, then repeats this process until no variables can be deleted without a statistically significant loss of fit.

Forward selection

Forward selection starts with no variables in the model, tests the addition of each variable if p-value is smaller than 0.15 and adds the variable whose inclusion gives the most statistically significant improvement of the fit, then repeats this process until no variables can improve the model.

Stepwise Regression

Stepwise Regression is one approach to fit the model with all potential variables, which combines “Backward Elimination” and “Forward Selection”. The first selection procedure is putting only one variable, and select the smallest p-value which is also smaller than 0.15, however, it would be delete if p-value is greater than 0.15 after adding other variables.

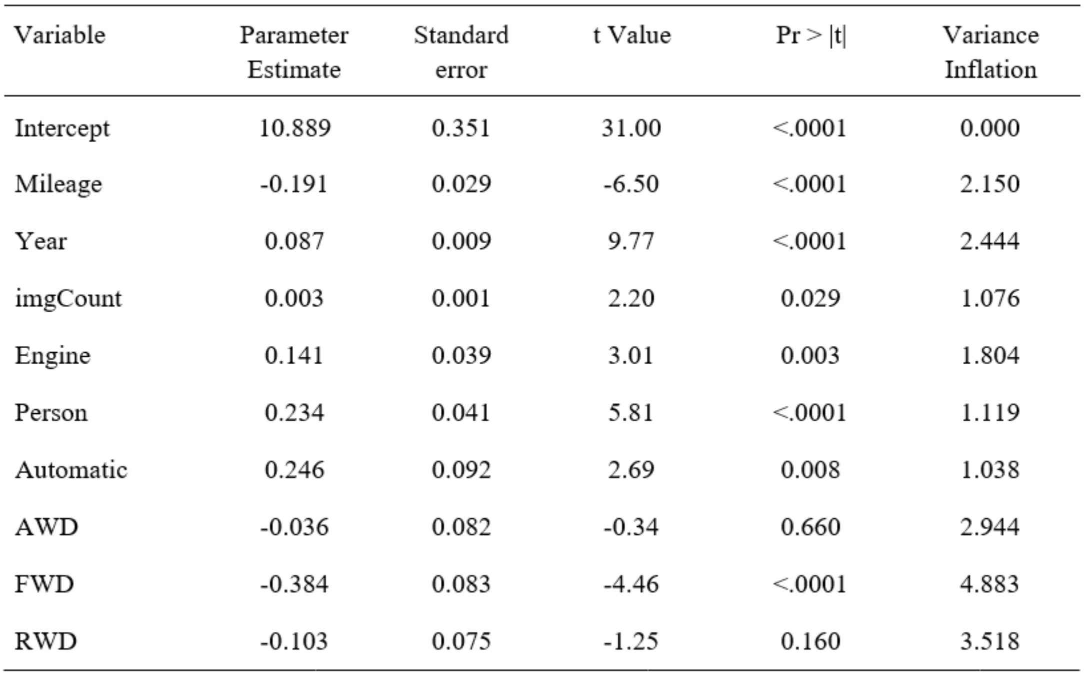
Adjusted R-Square Selection

Adjusted R-Square selection will calculate the adjusted R-Square of all the possible models, and the benefit of this method is that it would not be affected by increasing the number of variables.

Mallow's C_p

In the Mallows' C_p Selection, C_p value of all the possible models will be calculated. The selection criterion is that the C_p value is as small as possible, and the C_p value closer to the number of parameters will be considered as the best model.

After we do the model selection, the variance inflation factor decreases significantly.



Variable	Parameter Estimate	Standard error	t Value	Pr > |t|	Variance Inflation
Intercept	10.889	0.351	31.00	<.0001	0.000
Mileage	-0.191	0.029	-6.50	<.0001	2.150
Year	0.087	0.009	9.77	<.0001	2.444
imgCount	0.003	0.001	2.20	0.029	1.076
Engine	0.141	0.039	3.01	0.003	1.804
Person	0.234	0.041	5.81	<.0001	1.119
Automatic	0.246	0.092	2.69	0.008	1.038
AWD	-0.036	0.082	-0.34	0.660	2.944
FWD	-0.384	0.083	-4.46	<.0001	4.883
RWD	-0.103	0.075	-1.25	0.160	3.518

Table 8. Variance Inflation Factor after model selection

From table 8, the best selection model is:

$$ln\left(\widehat{price}\right) = 10.889 - 0.191ln(mileage) + 0.087year + 0.003imgCount \\ + 0.141engine + 0.234person + 0.246automatic - 0.036AWD \\ - 0.384FWD - 0.103RWD$$

Adjusted R^2 for this model is 0.7904, and $\hat{\sigma}^2$ is 0.06173.

2.9 Verify for assumption

Linear regression is an analysis that assesses whether one or more predictor variables explain the dependent variable. The regression has four key assumptions:

- The mean of error term is equal to zero.
- Linearity (The variance of residual is σ^2).
- No autocorrelation
- Normality of the error distribution

Assumption1: $E(\epsilon_i) = 0$

$$H_0: E(\epsilon_i) = 0$$
$$H_a: E(\epsilon_i) \neq 0$$

<table>
<tr><th colspan="5">Tests for Location: μ_0=0</th></tr>
<tr><th>Test</th><th colspan="2">Statistic</th><th colspan="2">p Value</th></tr>
<tr><td>Student’s t</td><td>t</td><td>-0.01106</td><td>Pr > |t|</td><td>0.9912</td></tr>
<tr><td>Sign</td><td>M</td><td>-1</td><td>Pr >= |M|</td><td>0.9431</td></tr>
<tr><td>Signed Rank</td><td>S</td><td>-82</td><td>Pr >= |S|</td><td>0.9182</td></tr>
</table>

Table 9. Assumption for The mean of error term is equal to zero.

We use Student's t, Sign, Signed Rank, three methods to verify the assumption. If the p-value is larger than α = 0.05, it will not reject the null hypothesis. From table 9, the p-value is greater than α----fail to reject the null hypothesis, which is $E(\epsilon_i) = 0$. Therefore, the mean of error term is equal to zero.

Assumption2：$Var(\epsilon_i) = \sigma^2$

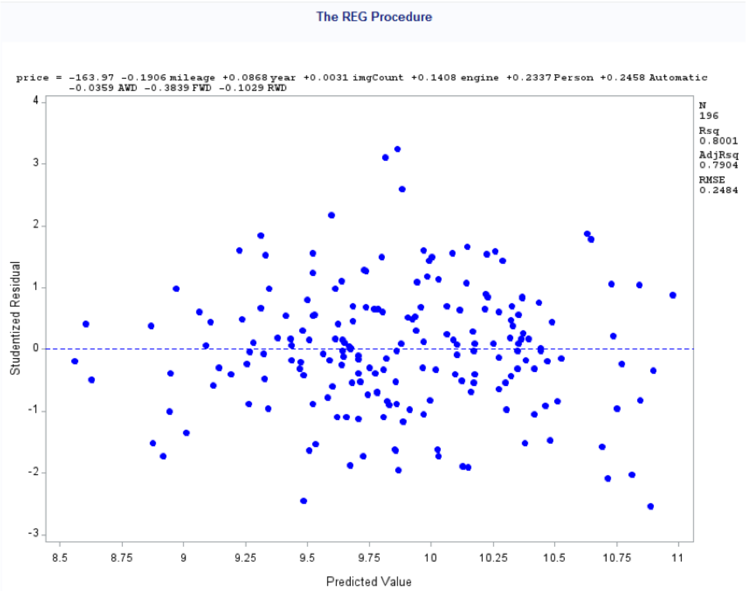


Figure 3. Studentized Residual of Predicted Value

To check the second assumptions, we should use residuals versus fitted values plot. Above is the plot from the regression analysis we did. The errors have constant variance, with the residuals scattered randomly around zero. Besides, the residuals plots do not have any patterns, therefore, the errors have constant variance.

Assumption3: $Cov(\epsilon_i, \epsilon_j) = 0 \quad \forall i \neq j$

$$\mathrm{H_0}: \rho = 0 \qquad \mathrm{H_0}: \rho = 0$$
$$\mathrm{H_a}: \rho < 0 \qquad \mathrm{H_a}: \rho > 0$$

Durbin-Watson Statistics			
Order	DW	Pr < DW	Pr > DW
1	1.9326	0.3142	0.6858

Table 10. Assumption for autocorrelation.

As usual, we use Durbin-Watson Statistics to test whether having autocorrelation. From table 10, both p-values for testing positive and negative autocorrelation are greater than $\alpha = 0.05$. So we fail to reject H_0. The hypothesis is valid.

Assumption4: $H_0: \epsilon_i \sim N(0, \sigma^2)$

$$H_0: \epsilon_i \sim N(0, \sigma^2)$$
$$H_a: \epsilon_i \nsim N(0, \sigma^2)$$

Tests for Normality

Test	Statistic		p Value	
Shapiro-Wilk	M	0.9913	Pr < W	0.2883
Kolmogorov-Smirnov	D	0.0460	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.0667	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.4257	Pr > A-Sq	>0.2500

Table 11. Assumption for Normality of the error distribution.

We use above 4 tests to verify the assumption. From table 11, we can notice that all the p-value is greater than α =0.05. We fail to reject the null hypothesis. Therefore, the error distribution is normal.

3. Results

This dataset includes 2 outliers and no influential point. All the variables in the best fitted model are significant except one wheel drive type called all-wheel-drive.

The main findings suggest that the model of year, the automatic transmission, for personal use, 4 wheel drive types, the number of description pages'photos and cylinders have positive correlation between the price and these factors. Moreover, there is a negative relationship between price and mileage; other wheel drive types.

4. Discussion

After verifying the assumptions, our hypothesis was supported.

In the past, people would have the new car experiences as the basis for the pricing of second-hand cars. Our defect is that the sample observations are not enough, so it cannot represent a very comprehensive information. But the advantage is that our data is very timely and real. It is possible to reflect the more authentic conditions of purchasing of used cars in San Jose. And we have studied other used cars in a similar way, for example: The automatic transmission will be relatively expensive.

5. Conclusion

In conclusion, if people want to buy a used car in San Jose State and they have a budget, they should choose the one with the mileage as low as possible, and that which has the latest year. They may ignore how many images are in the description pages, because they should contact the sellers, then go and see the seller in person.

If individuals would like to enjoy freeway in U.S., you should choose the second-hand car which has big engines, like the muscle car, though the price of the car will be higher as well.

Even though the car for personal use is expensive, do not risk buying the car which is not for personal use like rent cars, taxi cars, etc., you would never know how last driver drove the car, what it was used for, and what had happened to the car.

Nowadays most of the cars are automatic, and manual car is much cheaper. But we are going to San José, which is a busy metropolis, so driving a manual car will become a disaster .

There are various transmission systems you can choose, and all of them have benefits and drawbacks. From our perspectives, FWD is the best choice, thanks to its lower price and fuel economy. Also it does not snow in San José in winter, so it’s quite safe to drive. But, if you want something for fun, RWD is the option for you.

References

Carfax. (2017). Used cars for sale in San Jose, CA. Retrieved November 23, 2017, from https://www.carfax.com/Used-Cars-in-San-Jose-CA_c1023

Freund, R. J., Wilson, W. J., & Sa, P. (2006). *Regression analysis: Statistical modeling of a response variable*. Singapore: Elsevier