Conner Cook
Spring 2021
Math 130
Final Project

Predicting Car Sales

Introduction

I have chosen my final project to be based on finding car value. I have used numerous variables to create a model for predicting the cars' value. Information needed for this model includes the kilometers driven, whether the car is used or brand new, what fuel the car takes, what year the car was made, if the car was sold at a dealership or by an individual, and the type of transmission that is in the car. I am interested in finding the car value because I will be buying cars for the rest of my life and I want to educate myself in making sound purchases. I want to see which aspects of the car make it worth more and if I want to spend more money for something or whether or not I can find a less expensive alternative.

Data Collection

I found my data on a website named Kaggle which is a data collection website. The following is a picture of my data:

	A	B	C	D	E	F	G	H	I
1	Car_Name	Year	Selling_Price	Present_Price	Kms_Driven	Fuel_Type	Seller_Type	Transmission	Owner
2	ritz	2014	3.35	5.59	27000	Petrol	Dealer	Manual	0
3	sx4	2013	4.75	9.54	43000	Diesel	Dealer	Manual	0
4	ciaz	2017	7.25	9.85	6900	Petrol	Dealer	Manual	0
5	wagon r	2011	2.85	4.15	5200	Petrol	Dealer	Manual	0
6	swift	2014	4.6	6.87	42450	Diesel	Dealer	Manual	0
7	vitara brezza	2018	9.25	9.83	2071	Diesel	Dealer	Manual	0
8	ciaz	2015	6.75	8.12	18796	Petrol	Dealer	Manual	0
9	s cross	2015	6.5	8.61	33429	Diesel	Dealer	Manual	0
10	ciaz	2016	8.75	8.89	20273	Diesel	Dealer	Manual	0
11	ciaz	2015	7.45	8.92	42367	Diesel	Dealer	Manual	0
12	alto 800	2017	2.85	3.6	2135	Petrol	Dealer	Manual	0

Analysis / Conclusion (The results of the tests are part of the conclusion)

This data set shows all of my variables. Some are categorical like the year of the car and the fuel type. The reason they are categorical variables is because they have a finite number of possible options. Others are numerical like selling price and kilometers driven. The reason they are numerical variables is because they have infinite possible options. Most variables are self explanatory and some need further explanation. My response variable or the variable I am trying to guess is Present Price. When the number says 5.59, this is equivalent to the car's cost multiplied by one thousand which would be $5,590. Kilometers driven is self explanatory. Year is a categorical variable with every year from 2003 to 2018 being a possibility. Fuel type has 3 options: Petrol, Diesel and CNG. Transmission has two options: either automatic or manual. Seller type is if a dealership sold the car or if an individual sold the car. Owner is how many people owned the car prior to you buying it meaning 0 is a brand new car and 2 means you will be the third person to own the car. Selling price is what the car manufacturer sold the car for. I can not include the car name which is its model since it is too powerful and will make the other variables not important.

Test #1

I am going to make a multiple linear regression model using all of the variables and I am going to test if they are a good predictor of the present value of a car.

Model #1

Here is my Mathematical model:

$y= b0 + b1x1 + b2x2 + b3x3 + b4x4 + b5x5 + b6x6 + b7x7 + b8x8 + b9x9 +b10x10 + \dots + b21x21$

What the model means:

All of the B values are coefficients I will find when I do my test in R. y is the response variable present price. x1 is how many kilometers the car has gone. x2 and x3 are categorical variables for fuel type. If the fuel is diesel, then x2 will equal one and x3 will be zero. If the fuel type is petrol, then x3 will be one and x2 will be zero. The fuel type CNG is the control. If the

fuel type is CNG, then x3 and x2 will be zero. Another categorical variable is x4 which is seller type. Dealership is the control meaning if a car is sold at a dealership, x4 is zero. If the car is sold by an individual then x4 would be one. x5 is very similar to x4 but instead of seller type x5 is the transmission. Automatic transmission is the control so if the car is manual, x5 will be one. x6 is equal to how many owners the car had before you bought it. x7 equals how much the car manufacturer sold it for. Now x8 through x21 are categorical variables for the year the control year is 2003 and x8 is 2004 and x21 is 2018. If the car was made in 2016 x19 would be one and the rest would be zero.

Test #1 Statements

Significance of .05

Ho: b1=b2=b3=b4=b5=b6=b7=b8=b9 … =b21=0 (this means that the model has no predictive power)

Ha: at least one of the coefficient above is non zero (this means that at least one of the variables adds more predictive power)

Results of Test #1

```
Multiple R-squared:  0.864,     Adjusted R-squared:  0.8532
F-statistic: 80.28 on 22 and 278 DF,  p-value: < 2.2e-16
```

We must reject the null hypothesis that this model has no predictive power and accept the alternative hypothesis that at least one of the variables has predictive power because the p-value was approximately zero. I also found the R squared adjusted to be .8532 which is extremely strong due to the fact that R squared adjusted usually is a low-end answer. By low-end I mean it is the safe answer meaning if R squared adjusted is high this model is good at predicting the price of cars.

Test #2

Here are the results of the last model on R using the anova command:

```
Analysis of Variance Table

Response: Present_Price
               Df Sum Sq Mean Sq  F value    Pr(>F)
New_year       15 1176.2    78.4   7.1505 3.021e-13 ***
Kms_Driven      1  967.7   967.7  88.2444 < 2.2e-16 ***
Fuel_Type       2 4115.5  2057.8 187.6443 < 2.2e-16 ***
Seller_Type     1 2935.4  2935.4 267.6803 < 2.2e-16 ***
Transmission    1 1687.0  1687.0 153.8382 < 2.2e-16 ***
Owner           1   52.8    52.8   4.8107   0.02911 *
Selling_Price   1 8433.0  8433.0 768.9944 < 2.2e-16 ***
Residuals     278 3048.6    11.0
---
```

We can see by looking at the Pr(>F), on the right most column, which variables are known good predictors. A variable is a good predictor if the value is about less than .001. The variables I am going to put in my nested model will be year, seller type, transmission, kilometers driven, fuel type and selling price since their p value is close to zero.

Model #2

y= b0 + b1x1 + b2x2 + b3x3 + b4x4 + b5x5 + b6x6 + b8x8 + b9x9 +b10x10 + … + b21x21

Test #2 Statements

Significance level .01

Ho: b7=0 (This means that owner is not a good predictors)

Ha: b7 doesn't equal zero (This means that owner is a good predictor)

Test #2 Results

We must accept the null hypothesis that the variable owner is not a good predictor. The f value was .231 with a p value of 1. The R squared adjusted for the nested model was .8513.Which is less than the original model by a difference of .0022.

Conclusion

Here are my coefficients:

```
Coefficients:
                       Estimate 
(Intercept)            2.491e+00
New_year2004           6.721e+00
New_year2005           2.615e+00
New_year2006           3.871e+00
New_year2007          -2.078e+00
New_year2008          -5.673e-02
New_year2009           3.461e+00
New_year2010           3.584e+00
New_year2011          -9.279e-01
New_year2012          -1.225e-01
New_year2013          -1.010e+00
New_year2014          -1.607e+00
New_year2015          -3.303e+00
New_year2016          -3.697e+00
New_year2017          -4.842e+00
New_year2018          -5.586e+00
Kms_Driven             1.417e-05
Fuel_TypeDiesel       -2.748e+00
Fuel_TypePetrol       -1.268e+00
Seller_TypeIndividual  2.456e-01
TransmissionManual     2.866e-01
Selling_Price          1.660e+00
---
```

Experiment #1

To test my model, I am going to haphazardly select three cars on the data set and see how close the experimental present value is to the actual present value:

The first car I chose was on row 31 the actual value is 10.38 and the model got 9.31.

The second car I chose was on row 261 the actual value was 7 and the model got 6.05.

The third car I chose was on row 236 the actual value was 5.7 and the model got 5.12.

Experiment #2

To test the model against the real world, I am going to use my car which is a 2015 Honda Accord. The actual value is 10.8 and the model got 10.5 which is very close.

Final Conclusion

I created a good model for predicting cars' values since my final model had an R squared adjusted of .8513. My sample size was 301 cars so I had a considerable amount of data and I do not believe outliers will be a problem. With running my two experiments, I got good results even using my car which wasn't even part of the data set. Looking at the coefficients, I found some interesting results as to which variables change the price either

positively or negatively. For the year, it appears older cars are more expensive and that is probably due to the classic value. Kilometers driven has a small impact but surprisingly a positive impact. As for fuel type, it appears if CNG is the fuel type then the car would be more expensive. Diesel is the cheapest and petrol is in between. If the car was sold by an individual, then the car on average would cost about $200 more compared to being sold at a dealership. If the car is manual it is also more expensive than an automatic car. I do not know if power can be applied here but if it could I would guess my final model would have a large power for being within 20 percent of the actual value.