Group project

Team member: Qihui Feng; Junling Wang;

Chen Zhang; Yuchen Zhou

Class section: STAT 512 - 005

Professor: Chuanhai Liu

Due date : May 4, 2019

**Introduction:**

It helps to save money when bargaining with the dealer if we know what is a reasonable price. And it makes sense to use a regression model to explain the relationship between cars information and a reasonable selling price. Fortunately, Cardekho[1] provides a dataset publicly about second-handed cars conditions and ideal selling prices. Therefore, we are interested to find if it is possible to estimate the selling price based on the information of the used cars’ conditions and the present prices that dealers make.

To verify our model, we make several assumptions: we expect that the more distance that a car has been driven the less selling price it will have; on the other hand, the higher present prices are the higher selling prices are.

**Methods (\***all tests have significance value of 0.05**)**

1. A brief description: We have a dataset with sample size of 301 and 9 variables including a response variable. Each observation is one second-handed selling information. Here are the names, types, measurement units and variable definition:

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Type | Measurement Units | Variable Definition |
| Car\_Name | Categorical | N/A | The name of the car |
| Year | Discrete | Year | The year in which the car was bought |
| Selling\_Price | Continuous | Lakh Rupee | The price the owner wants to sell the car at |
| Present\_Price | Continuous | Lakh Rupee | The current ex-showroom price of the car |
| Kms\_Driven | Continuous | Km | The distance completed by the car in km |
| Fuel\_Type | Categorical | N/A | Fuel type of the car |
| Seller\_Type | Categorical | N/A | Whether the seller is a dealer or an individual |
| Owner | Discrete | Times | The number of owners the car has previously had |

1. Description of preliminary exploraory analyses:

We firstly excluded variable Car\_Name. It is a categorical variable with 98 different values. We are sacrificing too many degree of freedom to include this variable.

We then made a pairwise scatterplot in figure 1.0. We found three data points which are far away from the most of the data set. Therefore we drop all of them. We then found the relationship between continuous variables doesn’t look normal enough, we decide to do a transformation.

We wanted to make a decision among Power Transformation, Boxcox transformation[3] and Yeo and Johnson transformation[4]. We wanted a more normal relationship between variables and we don’t have negative values, therefore, we chose Boxcox.

The predictors has a range larger than the regressor, therefore, we transform predictors first. The summary of the predictors transformation is shown in Result 1. The Likelihood ratio test for all logarithmic transformations has a p-value of around 0. The Likelihood ratio test for no transformations also has a p-value of around 0. Therefore, the variable of Present\_Price should be in log scale and the power of the Kms\_Driven variable should be 0.33.

Then we focus on the transformation of the response variable. The summary table shown in Result 2 of the response variable transformation indicates that the estimated power is 0.06, which falls in the 95% confidence interval. The likelihood ratio tests for logarithmic transformation and no transformation both have a p-value of around 0, which all supports that 0.06 power should be used to transform the response variable.

After transformation we can see a clear linear relationship between present

1. Model building process:

We start with a model with the lowest complexity, which is a mean function with polynomial degree of 1 with no interaction term. Then we manually do backward subset selection. We found that no variables need to be dropped according to f-partial test.

We think the simple mean function is good enough for our

i). The simplest model already has a good performance in diagnostics analysis and result which we can see them in part d) and e) in the section.

ii). The k value of this model is already big enough which is 23. It is almost 10% of the overall degree of freedom. If we keep adding the complexity of this model, we are going to sacrifice more degree of freedom.

1. Diagnostic methods:

We make assumptions that:

i) The variance are constant

ii) Residuals are not curvature

We will test our assumption by plotting the residual plots.

1. Inferential methods

We test our linear regression model and the corresponding coefficient estimates based on the following two test statistics:

i) F-test of comparing MSR and MSRes for each variable

ii) T-test for estimators

We want to investigate if any variable somehow explain part of the response variable using the F-test. And we also want to investigate our problem raised in the introduction: whether the regression coefficient of present\_price is positive and DMS\_Driven is negative. We will conduct a T-test to verify our assumption

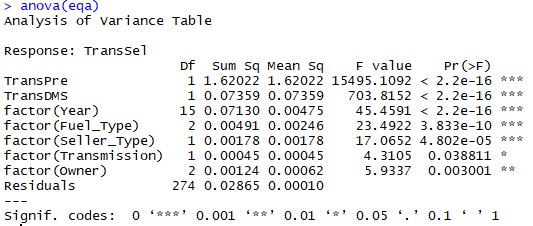
Results

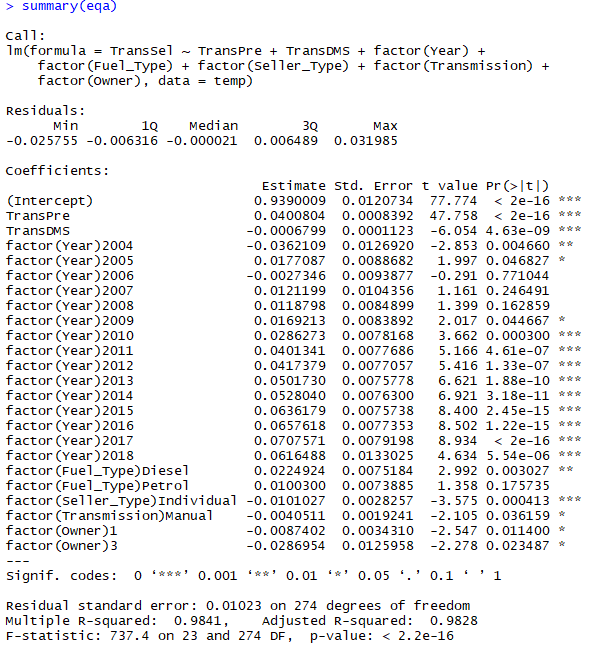
1. Best-subset table:

|  |  |
| --- | --- |
| Model | Partial F test comparing with the first model |
| TransSel ~ TransPre + TransDMS + factor(Year) + factor(Fuel\_Type) + factor(Seller\_Type) + factor(Transmission) + factor(Owner) | N/A |
| TransSel ~ TransPre + TransDMS + factor(Year) + factor(Fuel\_Type) + factor(Seller\_Type) + factor(Transmission) | 0.003 |
| TransSel ~ TransPre + TransDMS + factor(Year) + factor(Fuel\_Type) + factor(Seller\_Type) + factor(Owner) | 0.043 |
| TransSel ~ TransPre + TransDMS + factor(Year) + factor(Fuel\_Type) + factor(Transmission) + factor(Owner) | 0.00034 |
| TransSel ~ TransPre + TransDMS + factor(Year) + factor(Seller\_Type) + factor(Transmission) + factor(Owner) | 1.16e-10 |
| TransSel ~ TransPre + TransDMS + factor(Fuel\_Type) + factor(Seller\_Type) + factor(Transmission) + factor(Owner) | 2.2e-16 |
| TransSel ~ TransPre + factor(Year) + factor(Fuel\_Type) + factor(Seller\_Type) + factor(Transmission) + factor(Owner) | 4.629e-9 |
| TransSel ~ TransDMS + factor(Year) + factor(Fuel\_Type) + factor(Seller\_Type) + factor(Transmission) + factor(Owner) | 2.2e-16 |

Here are all the models we try in a backward subset selection process. The F-partial tests indicate that none of the parameters needs to be dropped from the original model.

b. and c. Anova table and parameter estimate:





We can see from the ANOVA test, all the parameters have P value less that 0.05. In the summary, residual standard error is 0.01. Comparing to 1, which is most of the values after transformation, the residual standard error is only 1%. R square values is close to one. Our mean function performans well in both of the tests.

d. In an agreement with our initial expectation

The summary table shows that the estimate coefficients for Present\_Price and Kms\_Driven are 0.04 and -0.0006, respectively. In addition, both of the corresponding p-values are around 0, which is far less than 0.05 that is the confidence level that we specified in this model. We have strong evidence that the coefficients for Present\_Price is significantly larger than 0 and Kms\_Driven are significantly smaller than 0. Our conclusion makes our assumption valid.

In addition, the anova table shows that F values for all of the regression variables are large and the corresponding P-values are all around 0, which supports that the regression variables are linearly related to the responsible variable Selling\_Price. Therefore, we can estimate the selling price based on the used cars’ conditions and the present prices that dealers make.

**Discussion:**

The main problem we want to discuss is whether we can find a regression model to explain and predict selling price as precise as possible for our dataset. We also make some reasonable assumption to verify whether the regression coefficients are reasonable or not. According to all the diagnostic analysis and summary tables, we found our model explains current dataset pretty very. And all the assumption we made are valid in our model.

Comparing to another price prediction using linear regression method [2], we performed a data transformation before regression process. Since our scoring system are different (Chavda used cross validation test, we check R square and ANOVA table), it is hard to compare the results.

In a big-picture point of view, it is possible to add a little feature in a website application as a user premium service: predicting a reasonable selling price.

The limitation of our model so far is that it might only be useful for this dataset. From our observation, the source of data might only come from India. Therefore, it might be hard to use this model to explain selling\_price in the U.S. Also, the model can’t make a predict if the year the car was bought isn’t listed in the dataset.

**References**

1. Birla, N. (2018, June 24). Vehicle dataset from cardekho. Retrieved from <https://www.kaggle.com/nehalbirla/vehicle-dataset-from-cardekho>
2. Chavda, R. (2019, January). Price prediction. Retrieved May 4, 2019, from <https://www.kaggle.com/rajanchavda1/price-prediction>
3. Box, G. and Cox, D. (1964) An Analysis of Transformations. Journal of the Royal Statistical Society. Series B (Methodological), 26, 211-252.
4. Yeo, I.-K. and Johnson, R. A. (2000) A new family of power transformations to improve normality or symmetry. Biometrika, 87, 954--959.

**Appendix A**

pairs(car.data[,-1], pch=21)

length(unique(car.data$Car\_Name)) #98

length(unique(car.data$Year)) #16

length(unique(car.data$Transmission)) #2

length(unique(car.data$Owner)) #3

length(unique(car.data$Seller\_Type)) #2

length(unique(car.data$Fuel\_Type)) #3

#delete data might be outlier

which.max(car.data$Present\_Price)#87

which.max(car.data$Kms\_Driven)#65

which.max(car.data[-c(87),]$Present\_Price)#197

temp <- car.data[-c(65, 87, 197),]

summary(bc <- powerTransform(cbind(Kms\_Driven, Present\_Price)~ factor(Year) + factor(Fuel\_Type) + factor(Seller\_Type) + factor(Transmission) + factor(Owner), temp))

coef(bc, round = TRUE)

temp$TransDMS = temp$Kms\_Driven ^ 0.33

temp$TransPre = log(temp$Present\_Price, 2)

eqa = lm(Selling\_Price ~ TransPre + TransDMS + factor(Year) + factor(Fuel\_Type) + factor(Seller\_Type) + factor(Transmission) + factor(Owner), data = temp)

summary(bc <- powerTransform(eqa))

coef(bc, round = TRUE)

temp$TransSel = temp$Selling\_Price ^ 0.06

pairs(temp[,-1], pch=21) # optional

eqa = lm(TransSel ~ TransPre + TransDMS + factor(Year) + factor(Fuel\_Type) + factor(Seller\_Type) + factor(Transmission) + factor(Owner), data = temp)

plot(x = fitted(eqa), y = temp$TransSel) + lines(lowess(fitted(eqa), temp$TransSel), col = "blue")

plot(x = fitted(eqa), y = resid(eqa)) + lines(lowess(fitted(eqa), resid(eqa)), col = "blue")

plot(x = temp$TransDMS, y = temp$TransSel) + lines(lowess(temp$TransDMS, temp$TransSel), col = "blue")

plot(x = temp$TransPre, y = temp$TransSel) + lines(lowess(temp$TransPre, temp$TransSel), col = "blue")

summary(eqa)

eqa = lm(TransSel ~ TransPre + TransDMS + factor(Year) + factor(Fuel\_Type) + factor(Seller\_Type) + factor(Transmission) + factor(Owner), data = temp)

# test if factor(Owner) can be drop

eqa1 = lm(TransSel ~ TransPre + TransDMS + factor(Year) + factor(Fuel\_Type) + factor(Seller\_Type) + factor(Transmission), data = temp)

anova(eqa1,eqa) # p-value =0.003 < 0.05, reject eqa1

# test if factor(Transmission) can be drop

eqa2 = lm(TransSel ~ TransPre + TransDMS + factor(Year) + factor(Fuel\_Type) + factor(Seller\_Type) + factor(Owner), data = temp)

anova(eqa2,eqa) # p-value =0.043 < 0.05, reject eqa2

# test if factor(Seller\_Type) can be drop

eqa3 = lm(TransSel ~ TransPre + TransDMS + factor(Year) + factor(Fuel\_Type) + factor(Transmission) + factor(Owner), data = temp)

anova(eqa3,eqa) # p-value =0.00034 < 0.05, reject eqa3

# test if factor(Fuel\_Type) can be drop

eqa4 = lm(TransSel ~ TransPre + TransDMS + factor(Year) + factor(Seller\_Type) + factor(Transmission) + factor(Owner), data = temp)

anova(eqa4,eqa) # p-value < 0.05, reject eqa4

# test if factor(year) can be drop

eqa5 = lm(TransSel ~ TransPre + TransDMS + factor(Fuel\_Type) + factor(Seller\_Type) + factor(Transmission) + factor(Owner), data = temp)

anova(eqa5,eqa) # p-value < 0.05, reject eqa5

# test if TransDMS can be drop

eqa6 = lm(TransSel ~ TransPre + factor(Year) + factor(Fuel\_Type) + factor(Seller\_Type) + factor(Transmission) + factor(Owner), data = temp)

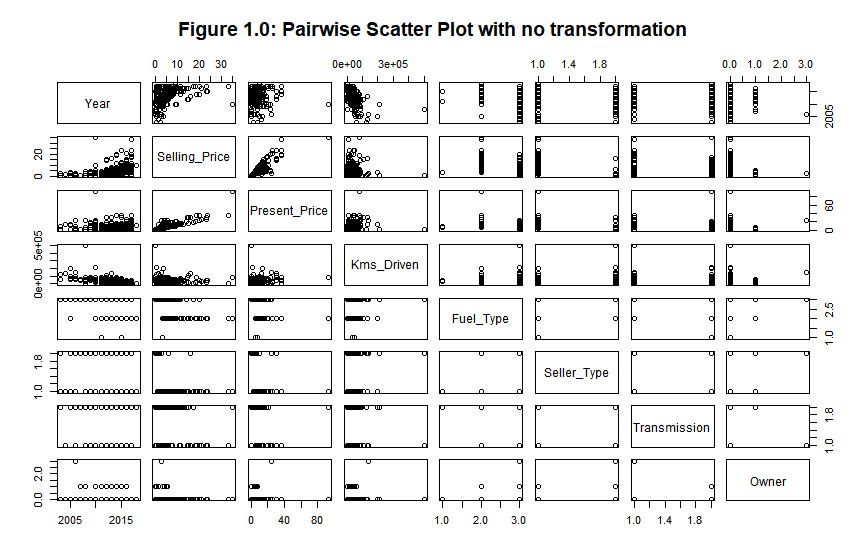
anova(eqa6,eqa) # p-value < 0.05, reject eqa6

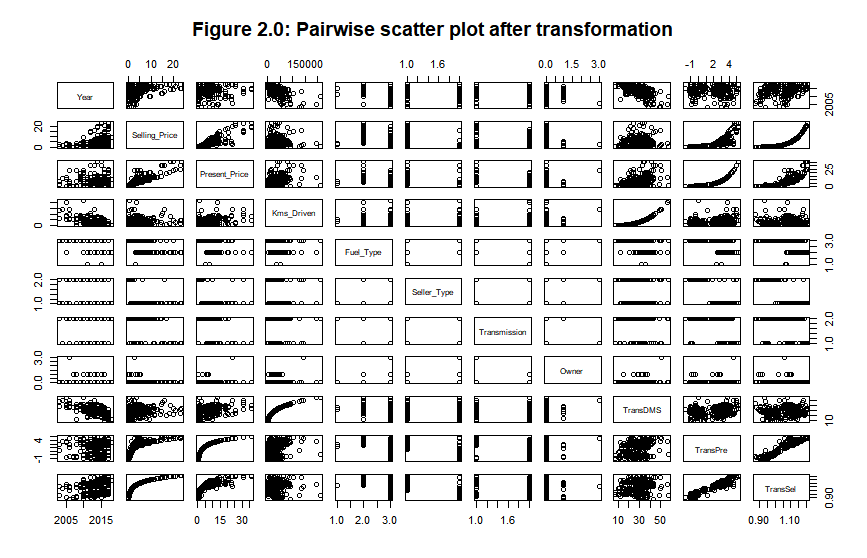
# test if Transpre can be drop

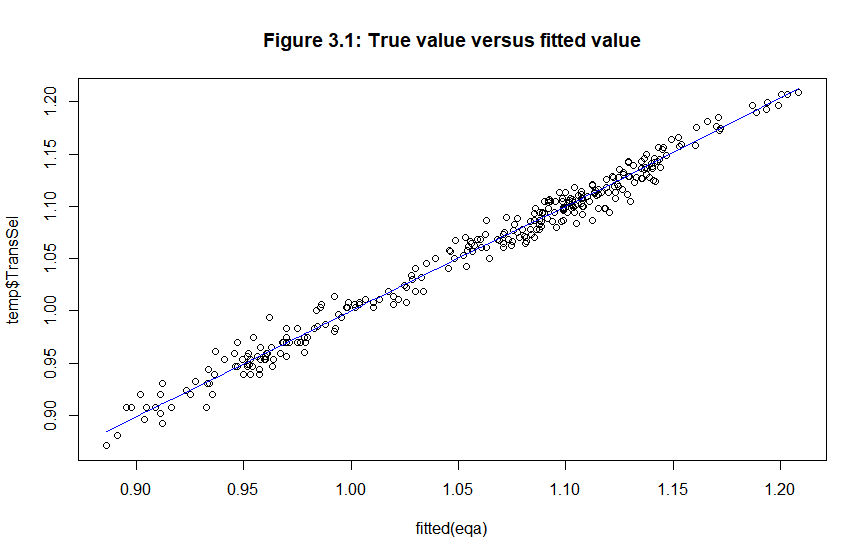
eqa7 = lm(TransSel ~ TransDMS + factor(Year) + factor(Fuel\_Type) + factor(Seller\_Type) + factor(Transmission) + factor(Owner), data = temp)

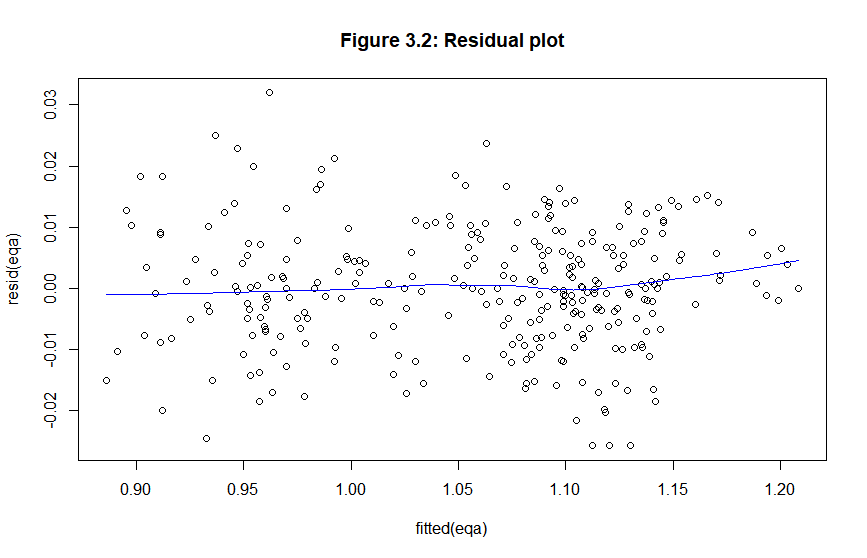
anova(eqa7,eqa) # p-value < 0.05, reject eqa7

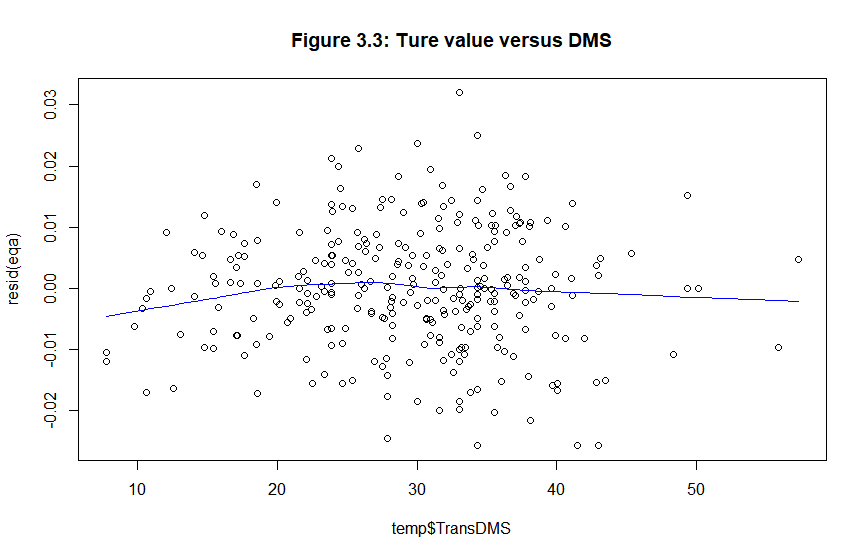
**Appendix B**

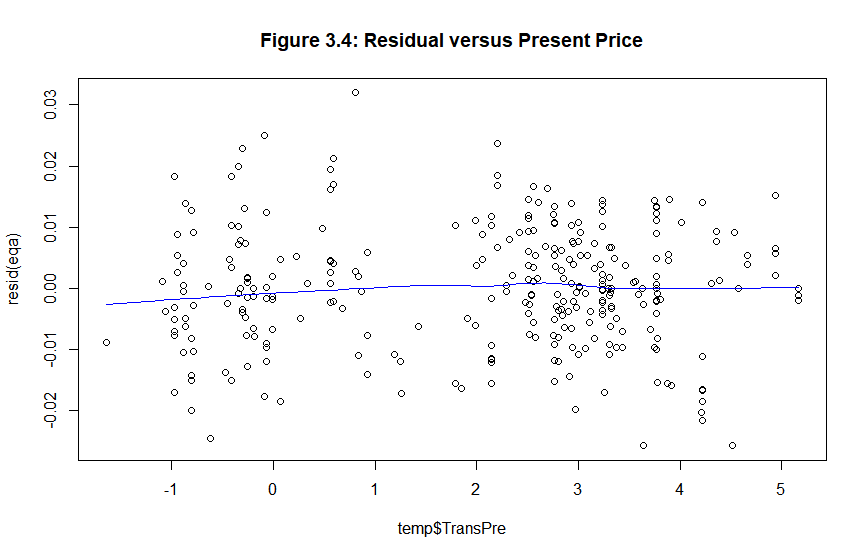




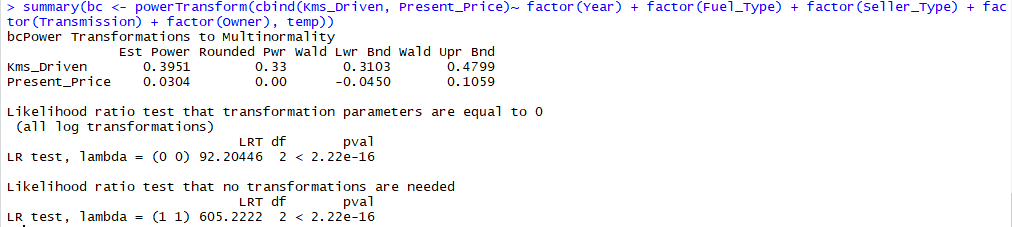








Result 1:



Result 2:

