Predicting Price of Motorcycle

Zootopia
Yushi Pan, Yanran Qiu, Chuanchuan Liu, Jingbin Cao

Introduction

- Business problem: predict price of motorcycle

- Dataset: Motorcycle Dataset from Kaggle

This dataset contains information about used motorcycles listed on www.bikewale.com
The seven columns in the given dataset are as follows*: name, selling price, year, seller type, owner, km driven, and ex showroom price.*

- Outline:
  - ❖ Part I: Clean Data and Visualize Data
  - ❖ Part II: Model Selection & Data Transformation
  - ❖ Part III: Check Model
  - ❖ Part IV: Test Model

Part I: Clean Data and Visualize Data

summary( )

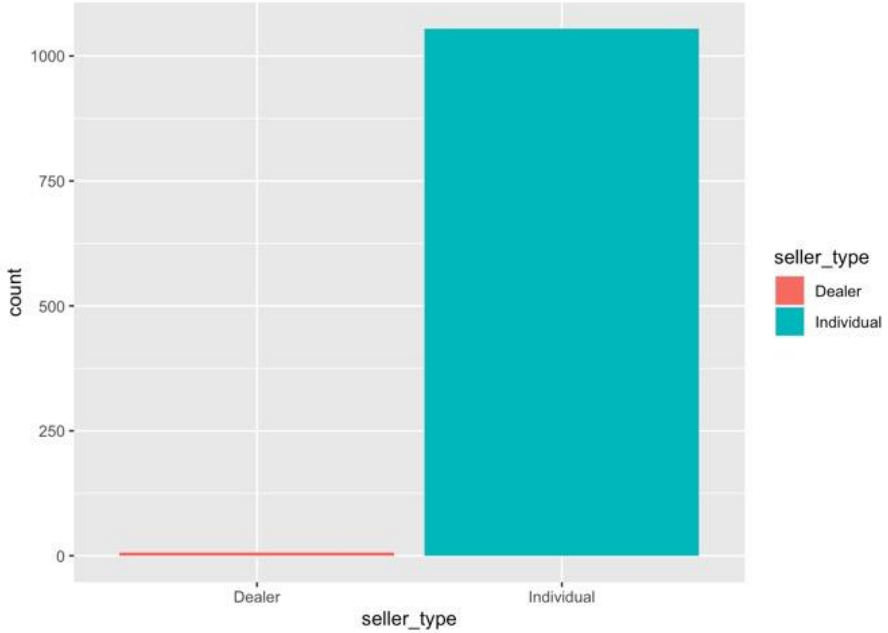
```
##      name           selling_price         year      seller_type
##  Length:1061        Min.   :  5000   Min.   :1988   Length:1061
##  Class :character   1st Qu.: 28000   1st Qu.:2011   Class :character
##  Mode  :character   Median : 45000   Median :2015   Mode  :character
##                     Mean   : 59638   Mean   :2014
##                     3rd Qu.: 70000   3rd Qu.:2017
##                     Max.   :760000   Max.   :2020
##
##     owner            km_driven      ex_showroom_price
##  Length:1061        Min.   :   350   Min.   :  30490
##  Class :character   1st Qu.: 13500   1st Qu.:  54852
##  Mode  :character   Median : 25000   Median :  72752
##                     Mean   : 34360   Mean   :  87959
##                     3rd Qu.: 43000   3rd Qu.:  87032
##                     Max.   :880000   Max.   :1278000
##                                      NA's   :435
```

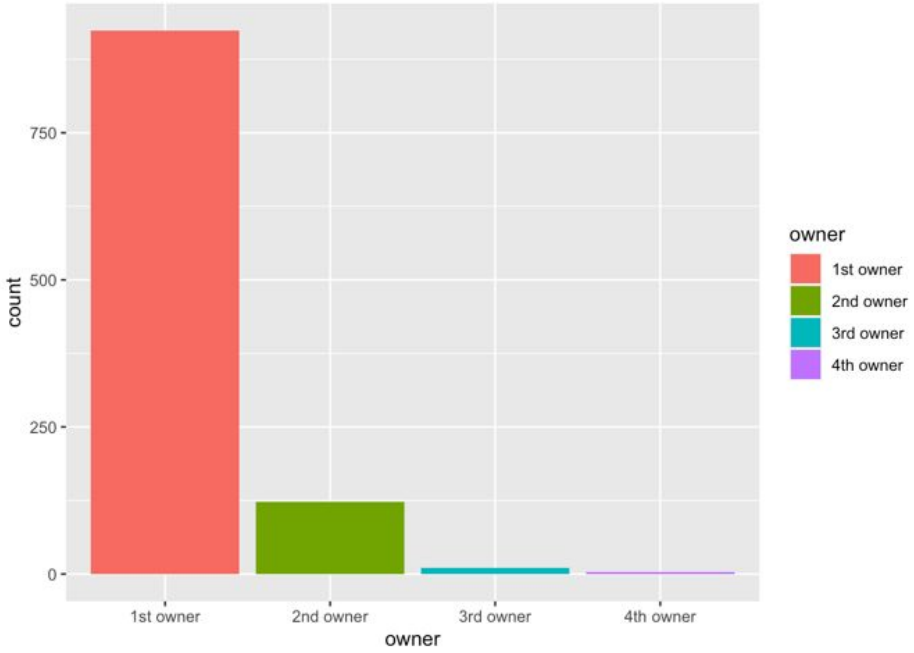
Qualitative independent variables: year, km_driven

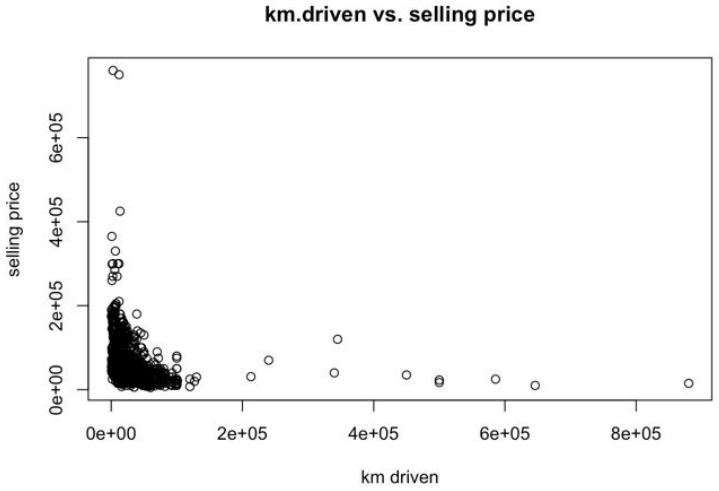
Quantitative independent variables: seller_type, owner

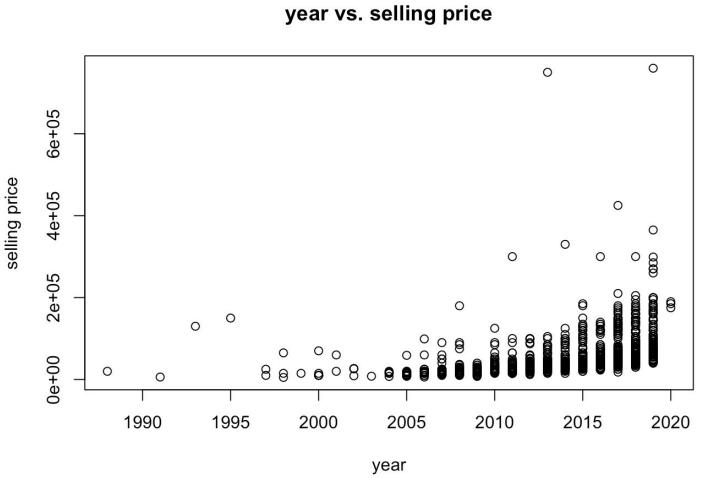
Dependent variable: selling_price

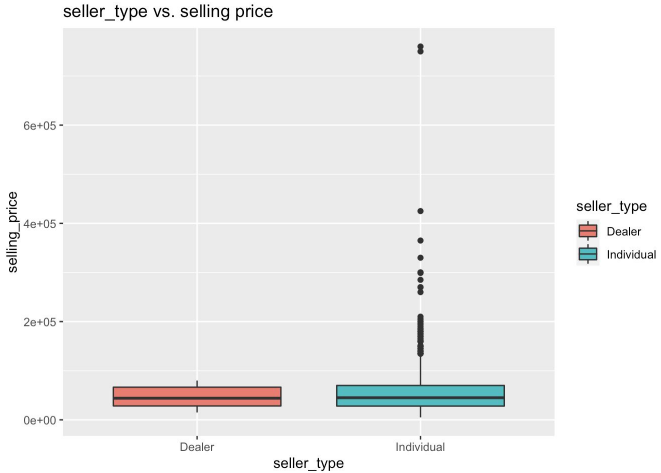
Visualization of Data

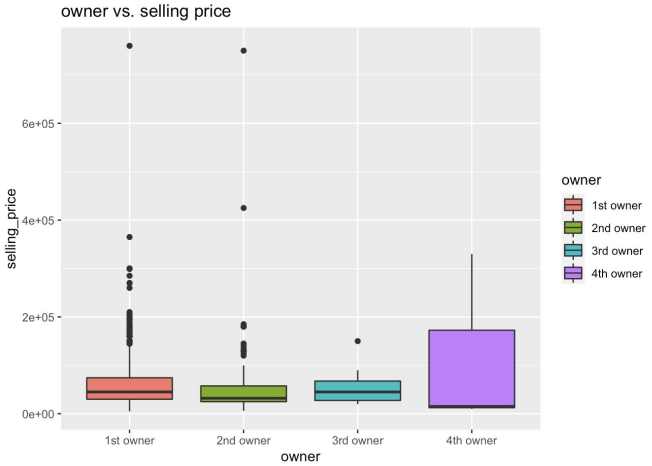




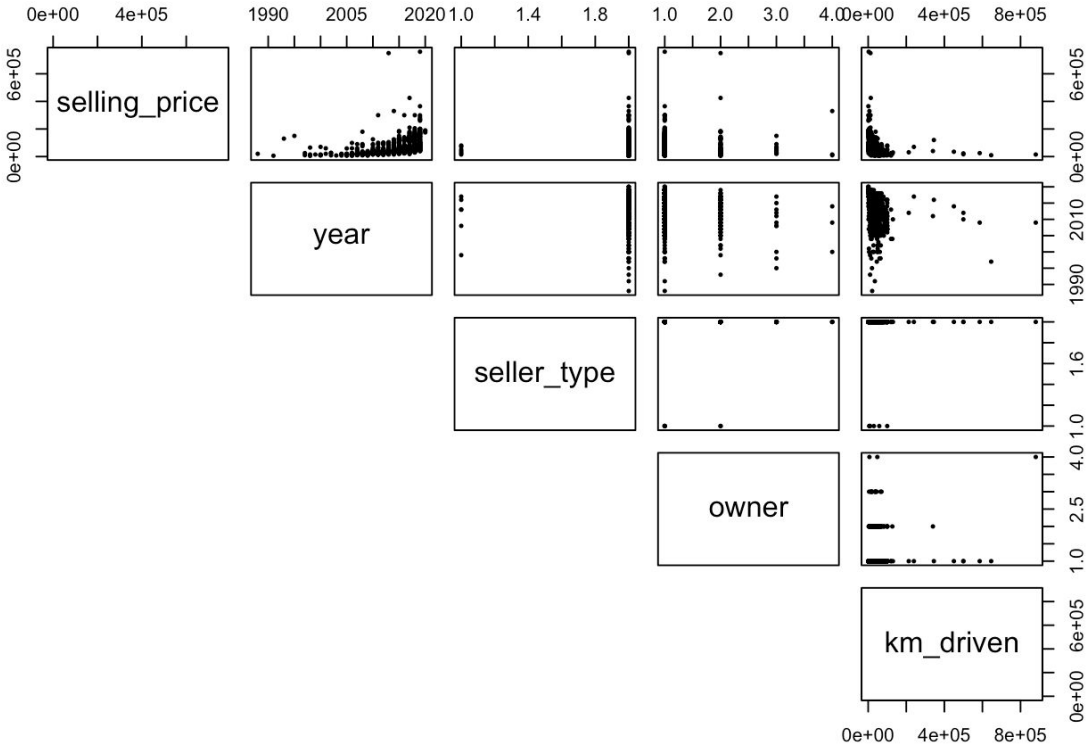

km.driven vs. selling price
selling price
0e+00
2e+05
4e+05
6e+05
0e+00
2e+05
4e+05
6e+05
8e+05
km driven


year vs. selling price
selling price
0e+00
2e+05
4e+05
6e+05
1990
1995
2000
2005
2010
2015
2020
year


seller_type vs. selling price
selling_price
6e+05
4e+05
2e+05
0e+00
Dealer
Individual
seller_type
seller_type
Dealer
Individual


owner vs. selling price
selling_price
6e+05
4e+05
2e+05
0e+00
1st owner
2nd owner
3rd owner
4th owner
owner
owner
1st owner
2nd owner
3rd owner
4th owner

pairs(df, pch=20, cex=0.5, lower.panel = NULL)



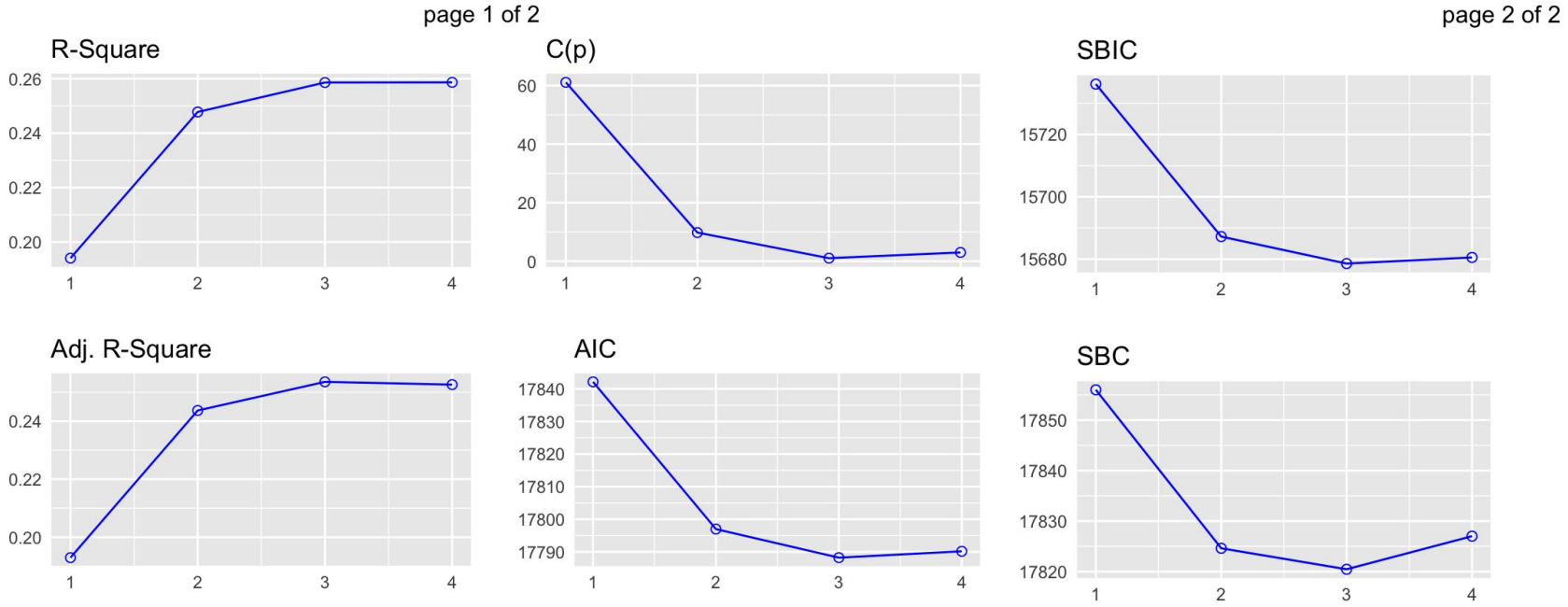
Feature encoding and split dataset

```
df$seller_type <- as.factor(df$seller_type)
df$owner <- as.factor(df$owner)
```

```
set.seed(123)
train_index <- sample(1:nrow(df), 0.7*nrow(df))
test_index <- setdiff(1:nrow(df), train_index)
train <- df[train_index,]
test <- df[test_index,]
```

Part II: Model Selection & Data Transformation

Best subsets regression



Best Subsets Regression

Model	Index	Predictors
	1	year
	2	year owner
	3	year owner km_driven
	4	year seller_type owner km_driven

Model 3 has the largest adjusted R-square value and the smallest MSE, Mallow's Cp, SBIC, SBC, and AIC. The best subset selected is (year, owner, km_driven).

Stepwise AIC regression

Selection Summary

Variable	AIC	Sum Sq	RSS	R-Sq	Adj. R-Sq
year	17842.173	287770361404.616	1.195327e+12	0.19403	0.19294
owner	17796.963	367485077758.314	1.115612e+12	0.24778	0.24370
km_driven	17788.197	3.83555e+11	1.099542e+12	0.25862	0.25358

Backward Elimination Summary

Variable	AIC	RSS	Sum Sq	R-Sq	Adj. R-Sq
Full Model	17790.143	1.099462e+12	383634937685.570	0.25867	0.25262
seller_type	17788.197	1.099542e+12	3.83555e+11	0.25862	0.25358

Stepwise Summary

Variable	Method	AIC	RSS	Sum Sq	R-Sq	Adj. R-Sq
year	addition	17842.173	1.195327e+12	287770361404.616	0.19403	0.19294
owner	addition	17796.963	1.115612e+12	367485077758.314	0.24778	0.24370
km_driven	addition	17788.197	1.099542e+12	3.83555e+11	0.25862	0.25358

Stepwise P-value regression

Selection Summary

Step	Variable Entered	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	year	0.1940	0.1929	61.0863	17842.1728	40190.8786
2	owner	0.2478	0.2437	9.7963	17796.9626	38906.5656
3	km_driven	0.2586	0.2536	1.0535	17788.1967	38651.5645

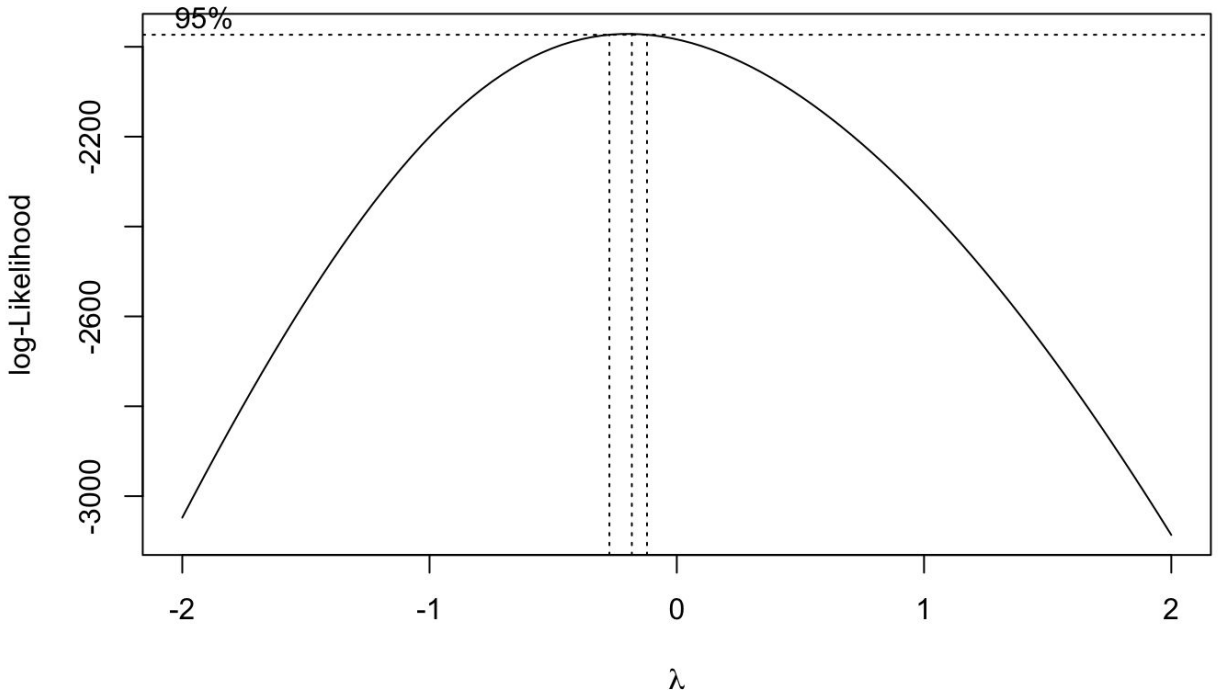
Elimination Summary

Step	Variable Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	seller_type	0.2586	0.2536	1.0535	17788.1967	38651.5645

Stepwise Selection Summary

Step	Variable	Added/ Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	year	addition	0.194	0.193	61.0860	17842.1728	40190.8786
2	owner	addition	0.248	0.244	9.7960	17796.9626	38906.5656
3	km_driven	addition	0.259	0.254	1.0530	17788.1967	38651.5645

Box-Cox transformation




```
             [,1]      [,2]
[1,] -0.1818182 -1971.339
[2,] -0.2222222 -1971.479
[3,] -0.1414141 -1972.253
[4,] -0.2626263 -1972.693
[5,] -0.1010101 -1974.196
```

$$Y' = Y^{-0.18}$$

Part III: Model Check

summary ()

```
##
## Call:
## lm(formula = selling_price ~ year + owner + km_driven, data = train)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.081950 -0.006937  0.002079  0.008788  0.040967
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.505e+00  2.616e-01  21.042  < 2e-16 ***
## year            -2.661e-03  1.298e-04 -20.498  < 2e-16 ***
## owner2nd owner   5.233e-04  1.552e-03   0.337  0.73609
## owner3rd owner -1.445e-02  4.452e-03  -3.245  0.00123 **
## owner4th owner -4.393e-02  1.380e-02  -3.184  0.00151 **
## km_driven        4.169e-08  1.103e-08   3.779  0.00017 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01378 on 736 degrees of freedom
## Multiple R-squared:  0.4352, Adjusted R-squared:  0.4314
## F-statistic: 113.4 on 5 and 736 DF,  p-value: < 2.2e-16
```

anova()

```
## Analysis of Variance Table
##
## Response: selling_price
##             Df   Sum Sq  Mean Sq  F value    Pr(>F)
## year         1 0.100740 0.100740 530.2739 < 2.2e-16 ***
## owner        3 0.004295 0.001432   7.5353 5.702e-05 ***
## km_driven    1 0.002713 0.002713  14.2798 0.0001703 ***
## Residuals 736 0.139823 0.000190
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Calculate Residuals

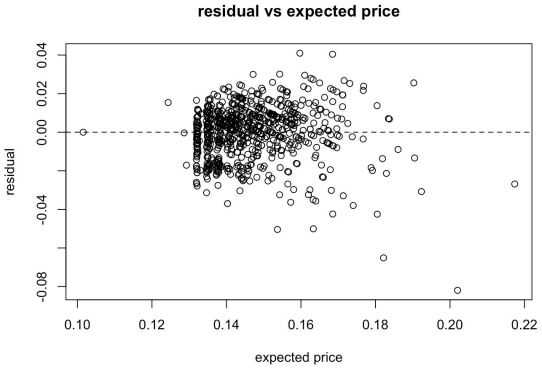
```
e = resid(model)
head(e)
```

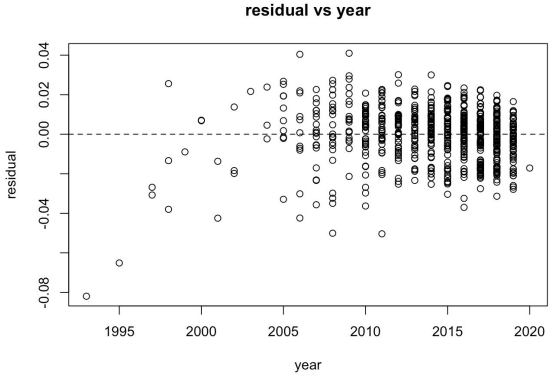
```
##           415           463           179           526           195           938
##  0.005123157  0.003921742  0.012618321  0.004259661  0.003363994 -0.005632955
```

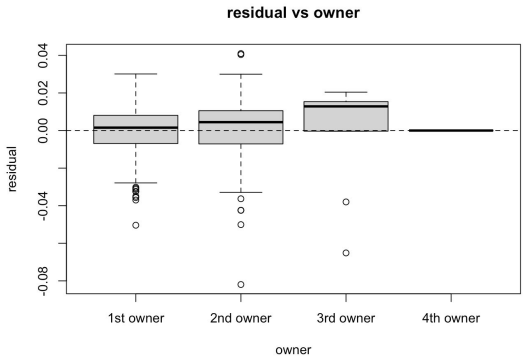
Check Residuals and assumptions in linear regression

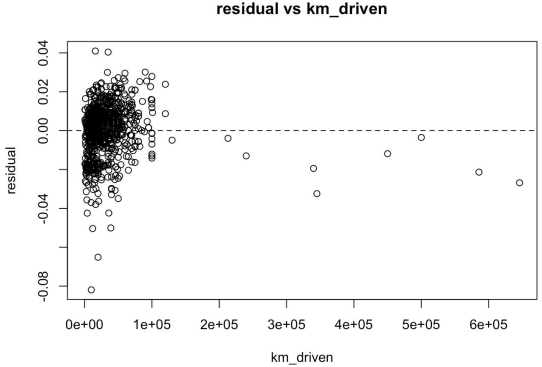
- linearity
- constancy of error variance
- independence of error terms
- normality of error terms
- absence of outliers

Residual Plot

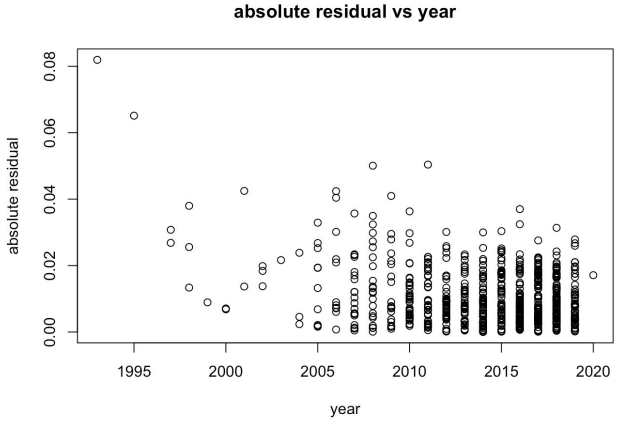


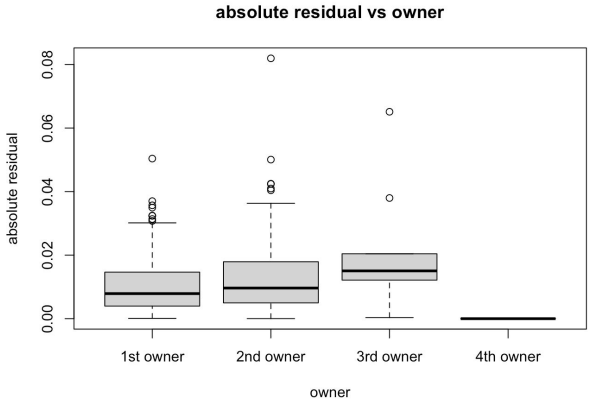


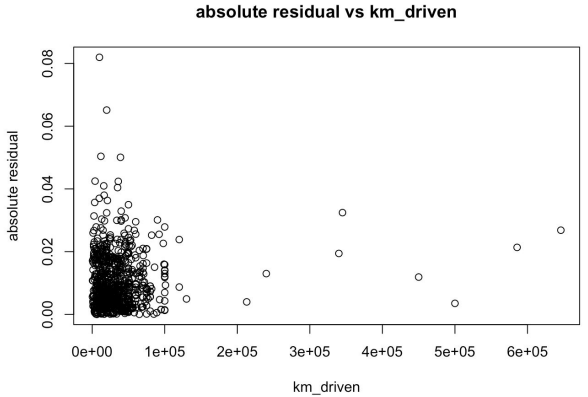




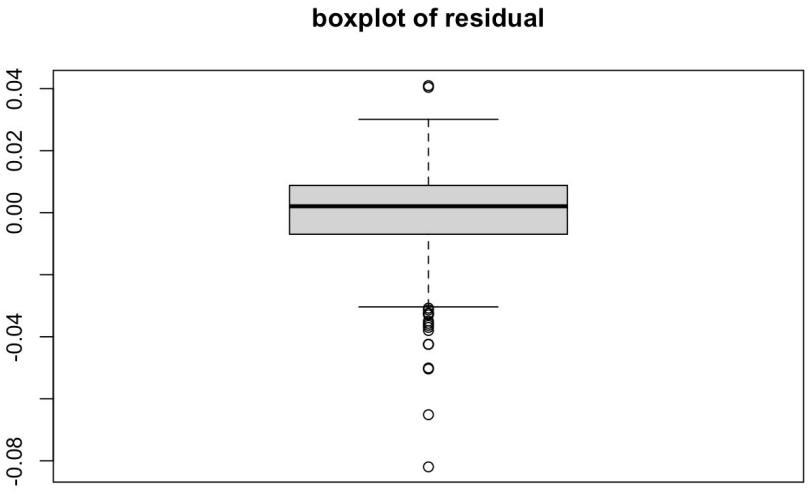
Absolute Residual Plot

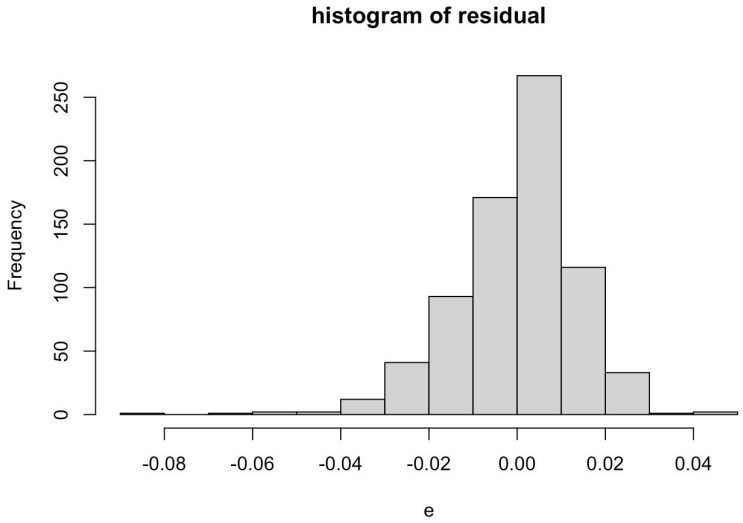






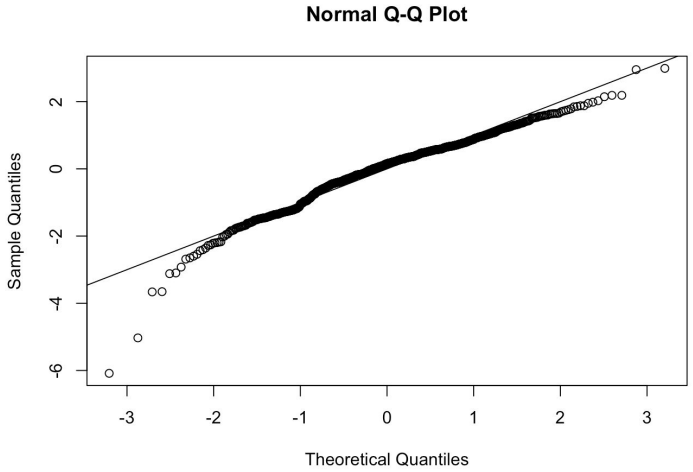
Distribution Plot of Residuals

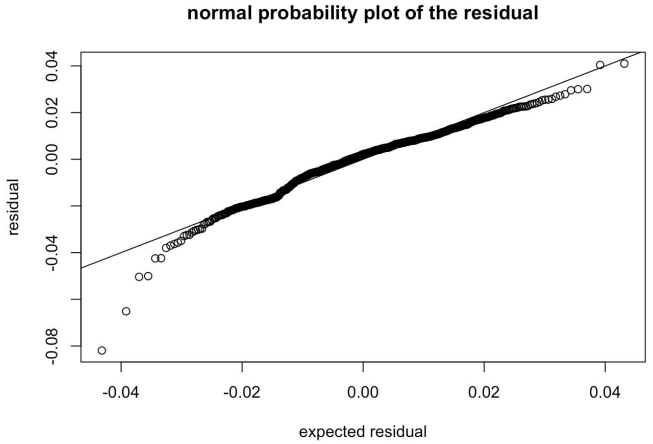




Normal Probability Plot

cor(res.exp, res.order) = 0.9805561
plot(res.exp, res.order)





Correlation Test for Normality

Breusch-Pagan Test

BP = 233.01
chi-squared (0.95, df=5) = 11.0705
chi-squared (0.99, df=1) = 15.08627
Based on BP test, we can conclude that the assumption of constancy of error variance is violated. **(NOTE!)**

F Test for Lack of Fit

sum (duplicated) = 189
F statistics = 1.7534
F (0.95, df1=736, df2=481) = 1.147556
F (0.99, df1=736, df2=481) = 1.215212

```
## Analysis of Variance Table
##
## Model 1: selling_price ~ year + owner + km_driven
## Model 2: selling_price ~ 0 + as.factor(year) + as.factor(owner) + as.factor(km_driven)
##   Res.Df      RSS  Df Sum of Sq      F    Pr(>F)
## 1    736 0.139823
## 2    481 0.072463 255   0.06736 1.7534 7.528e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on F Test for Lack of Fit, we can conclude there is lack of fit and the assumption of linearity is violated. **(NOTE!)**

Part IV: Test Model

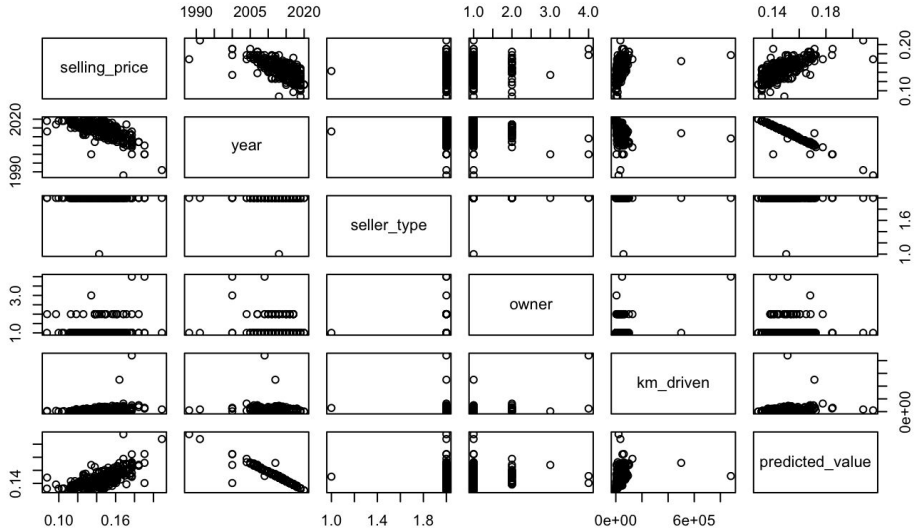
We have **319** Test data.

Transform the selling price of the test data set:

```
test$selling_price <- (test$selling_price)^-0.18
```

Predict the selling price from the dataset

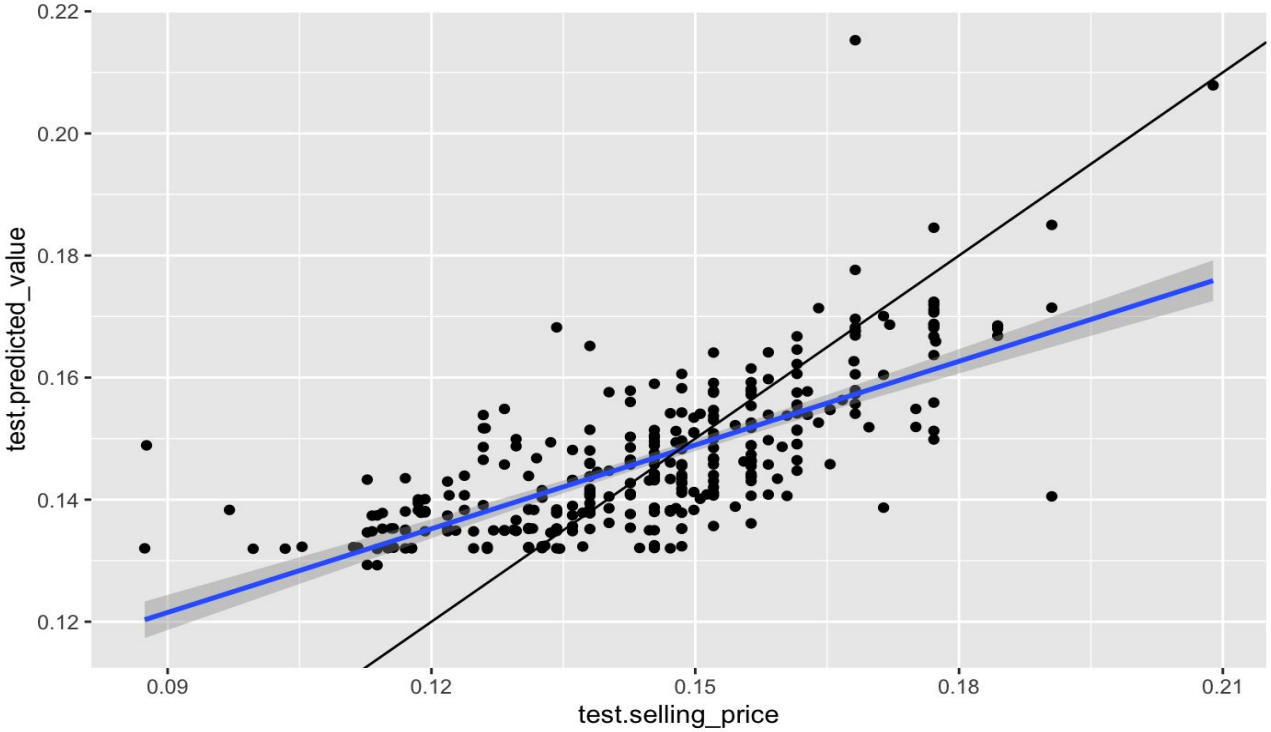
```
predictions <- predict(model, newdata=test,

interval="prediction")
```



	fit	lwr	upr
1	0.1319192	0.1048118	0.1590266
3	0.1350660	0.1079696	0.1621624
7	0.1352744	0.1081786	0.1623703
9	0.1571890	0.1300867	0.1842914
12	0.1408282	0.1136050	0.1680514
14	0.1319516	0.1048444	0.1590588
15	0.1293143	0.1021967	0.1564319
18	0.1582542	0.1310480	0.1854604
21	0.1636786	0.1365563	0.1908009

Comparing the data

	selling_price	year	seller_type	owner	km_driven	predicted_value
1	0.11382906	2019	Individual	1st owner	350	0.1319192
3	0.11703172	2018	Individual	1st owner	12000	0.1350660
7	0.13149931	2018	Individual	1st owner	17000	0.1352744
9	0.15635730	2010	Individual	1st owner	32000	0.1571890
12	0.15831117	2016	Individual	2nd owner	10000	0.1408282
14	0.09972115	2019	Individual	1st owner	1127	0.1319516

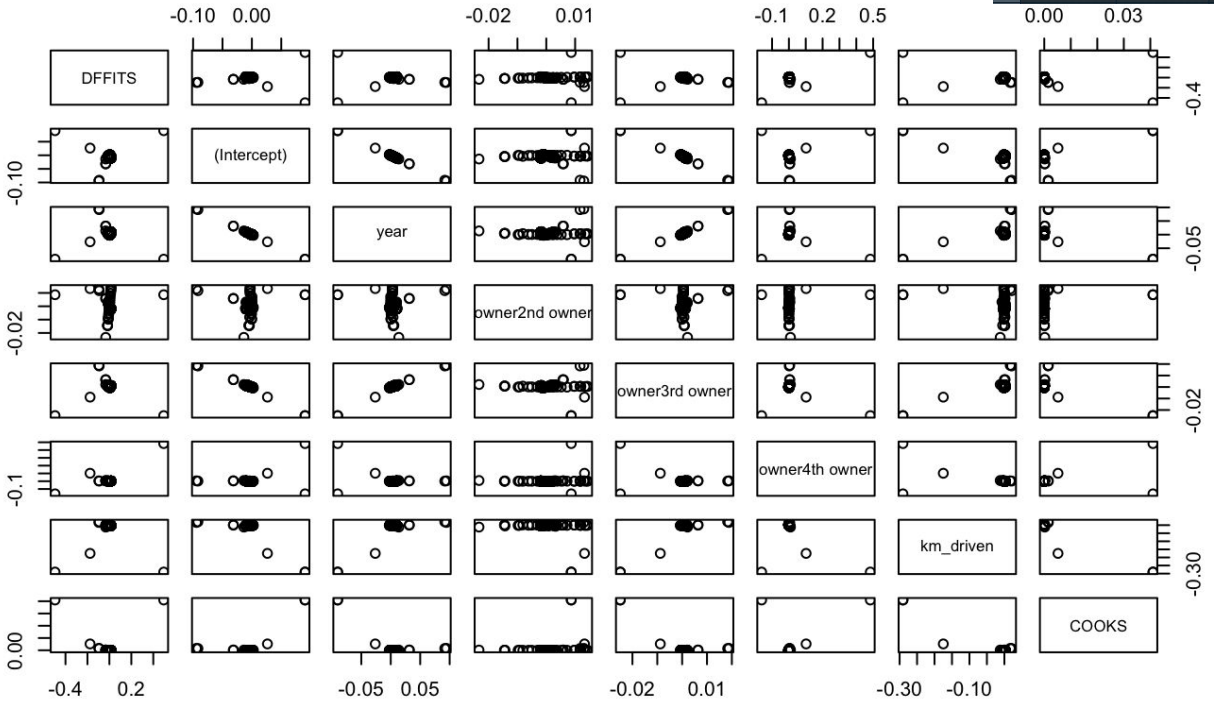


MSPE:0.00143

Out model is good enough to predict the price of a used motorcycle based on the year, owner, and driven distances.

Checking Influences

	DFFITS	(Intercept)	year	owner2nd owner	owner3rd owner	owner4th owner	km_driven	COOKS
1	-6.784859e+01	4.340580e+01	-4.353970e+01	8.692132e+00	-4.893052e+00	-1.345500e+01	1.695559e+01	4.164025e-01
3	8.277841e-03	-4.938085e-03	4.954818e-03	-1.378869e-03	5.132043e-04	9.866722e-04	-1.148152e-03	1.145663e-05
7	7.032188e-03	-4.337553e-03	4.351180e-03	-1.202331e-03	4.751577e-04	5.327688e-04	-4.350138e-04	8.268143e-06
9	2.568742e-03	1.741474e-03	-1.737424e-03	-7.110215e-04	-4.349891e-04	-2.321329e-04	-2.476203e-04	1.103261e-06
12	1.421058e-02	-2.591479e-03	2.594759e-03	1.338167e-02	4.197871e-04	1.541523e-03	-1.944906e-03	3.376405e-05
14	1.489017e-02	-9.582025e-03	9.611287e-03	-1.918774e-03	1.086876e-03	2.870989e-03	-3.573515e-03	3.706798e-05
15	1.503413e-02	-1.084640e-02	1.087312e-02	-1.605688e-03	1.388748e-03	2.627467e-03	-2.767308e-03	3.778873e-05
18	-3.412908e-03	-6.466808e-04	6.468580e-04	-3.118081e-03	1.185251e-04	9.597224e-05	-2.953711e-06	1.947544e-06
21	-3.492373e-03	-2.552738e-03	2.549866e-03	8.828124e-04	5.645713e-04	9.367554e-04	-7.160654e-04	2.039286e-06
22	7.783644e-03	-2.587719e-03	2.605753e-03	-1.851779e-03	3.207186e-05	5.864112e-04	-1.148997e-03	1.012941e-05

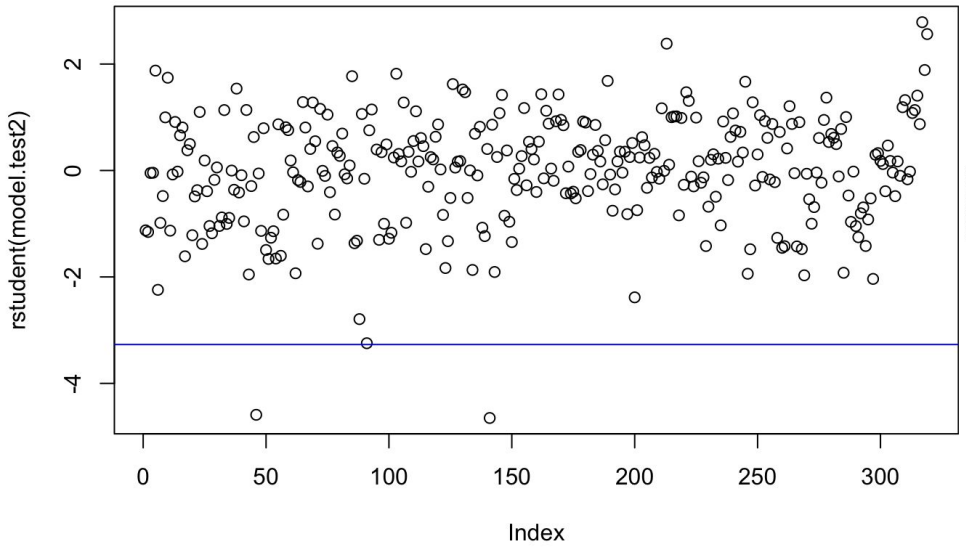


Original Output

DFFITS: 0.27429
DFBEATS: 0.11198

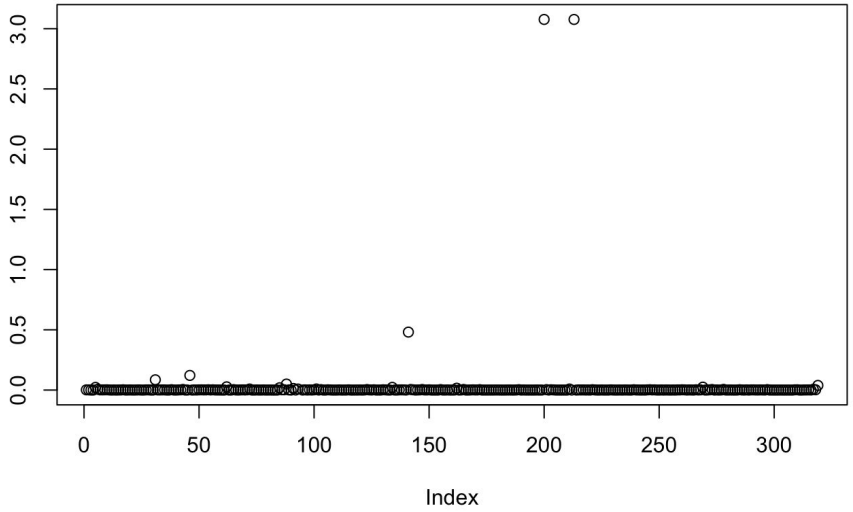
Since all of the values are very small, we can conclude that there are not influential data in our dataset.

Obtain the studentized deleted residuals and identify any outlying Y observations. Using Bonferroni outlier test procedure with a = 0.1.



if |ti| ≤ t(1 − 0.1/(2n), n − p − 1) = 3.27, we conclude no outliers. In the test data, we have only 2 outliers

Cook Distances



Although there are some outliers, but the values of them are below the critical values, so they do not have many influence on our model.

Improvement and Expansion

- Minimize the effect of the outliers
- Add interaction effect between predictors
- Collect data from varied sources to expand our dataset