**Predicting Used Car Prices**

Final Report, by Tina Kirk

## I. Introduction

Used car prices can vary widely and are dependent on several factors. The data for this analysis was found on kaggle.com [1], originally collected by Car Dekho. It is comprised of 4340 records of used car purchases from India from 1995 to 2020. There were almost 1500 different models of cars listed, but only 29 makes. The selling prices listed were in Indian rupees. Other columns were year (assuming model year but could be year of purchase), kilometers driven, number of owners, fuel type (Diesel, Petrol, CNG, Electric, LPG), seller type (Dealer, Individual, Trustmark Dealer), and transmission type (Automatic, Manual). For selling price, I made a column to convert to US dollars for ease of understanding. Qualitative columns were transformed into factor variables: make to makef, fuel to fuelf, seller_type to seller_typef, and transmission to transmissionf.

The purpose of this project is to arrive at the best model for predicting used car prices (in US dollars), given a subset of these variables.

## II. Analysis

The columns of “selling price” (in rupees) and “name” were not used as they have replacements of “selling price” (in US dollars) and “make,” respectively. The correlation matrix, without make, is shown below in Fig. 1. Selling price vs. make is shown in Fig. 2.

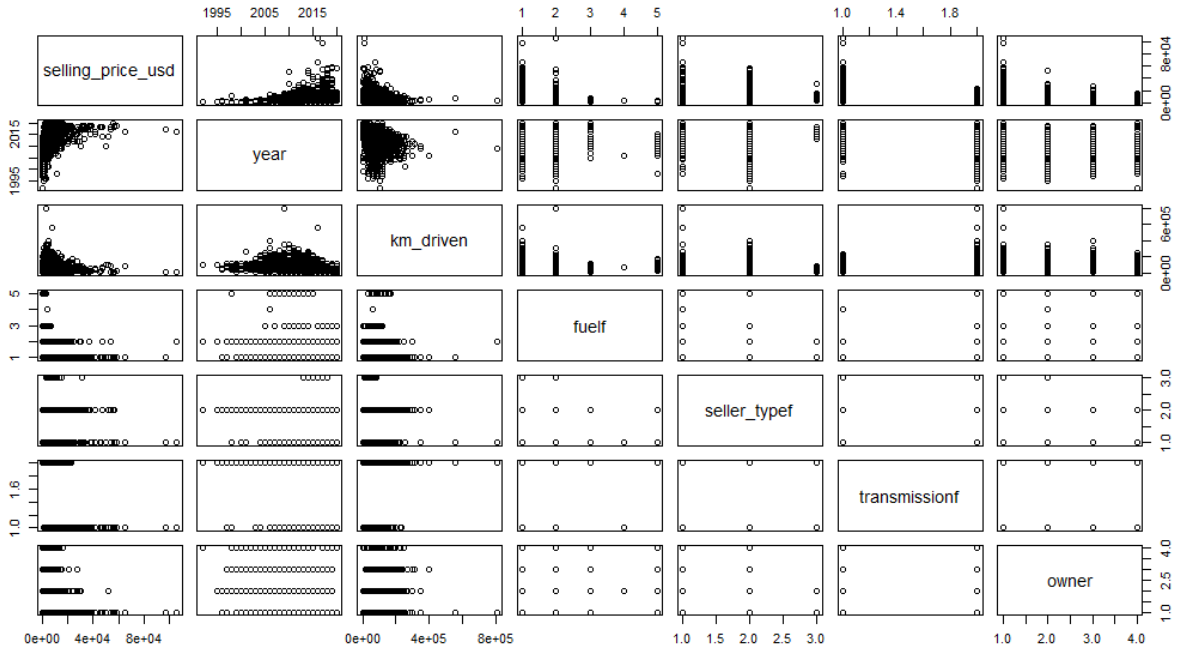



Fig. 1 Correlation matrix with selling_price_usd, year, km_driven, fuelf, seller_typef, transmissionf, and owner

We can see some trends in the plot in Fig. 1.

- As the year increases, the price of the used car tends to increase. This makes sense if it is the model year OR the year purchased, because on the one hand, a later model year shows a newer car, and on the other hand, a later purchase year for any make or model, with inflation, may cause the price to increase.
- As the kilometers driven increases, the price tends to decrease.
- Diesel and petrol used cars get the highest prices, with CNG, electric, and PNG cars at the lowest.
- Cars sold by dealers and individuals appear to have higher prices than Trustmark Dealers.
- Cars with automatic transmissions sell for more than those with manual transmissions.
- The more owners a car has had, the lesser the selling price tends to be.

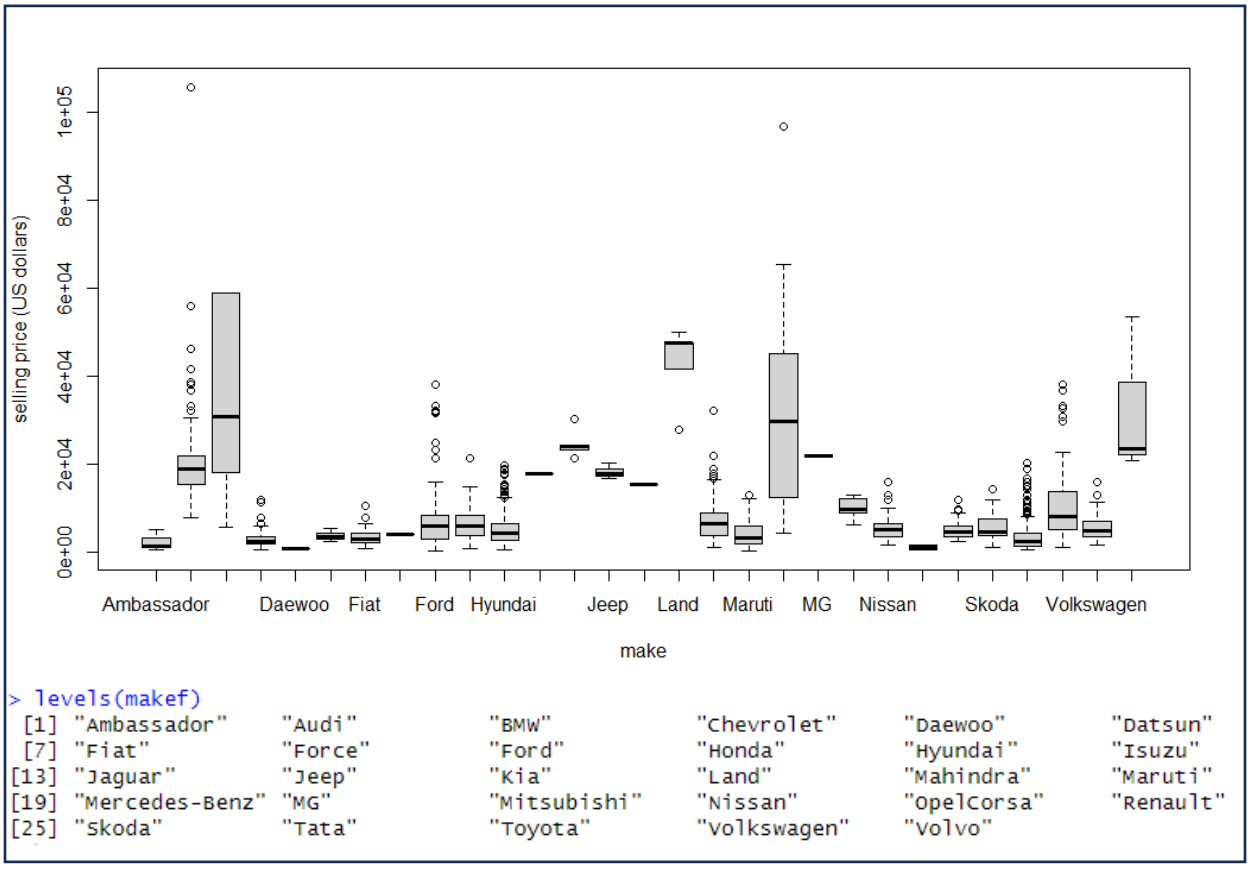
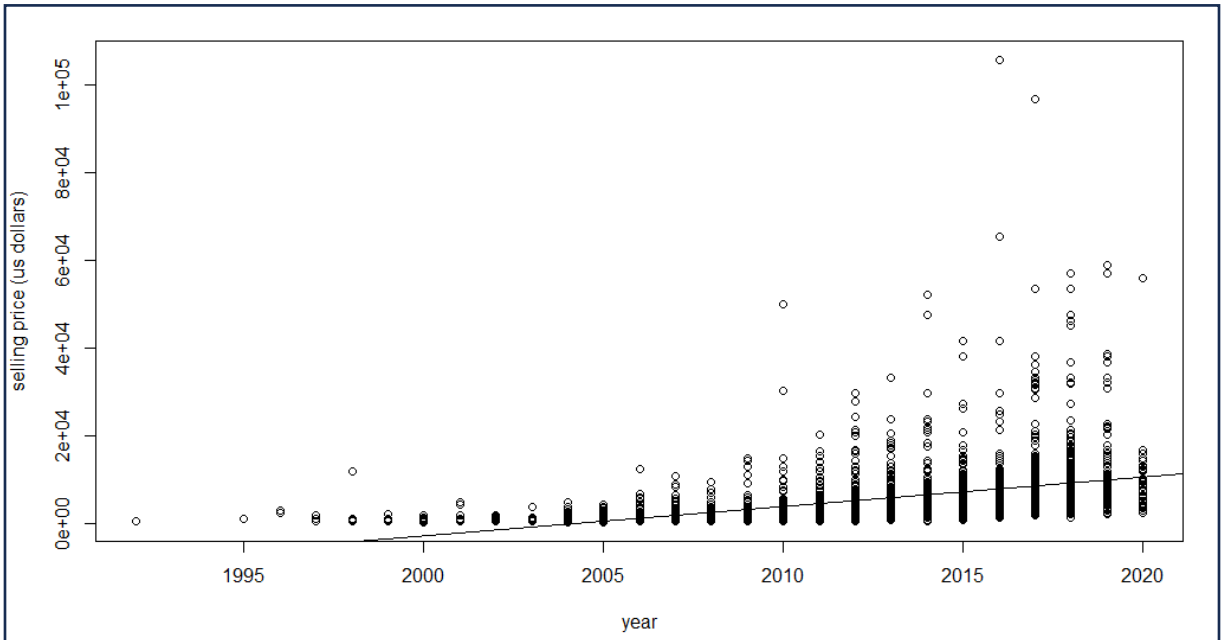



Fig. 2 Selling price vs. make with levels of make listed

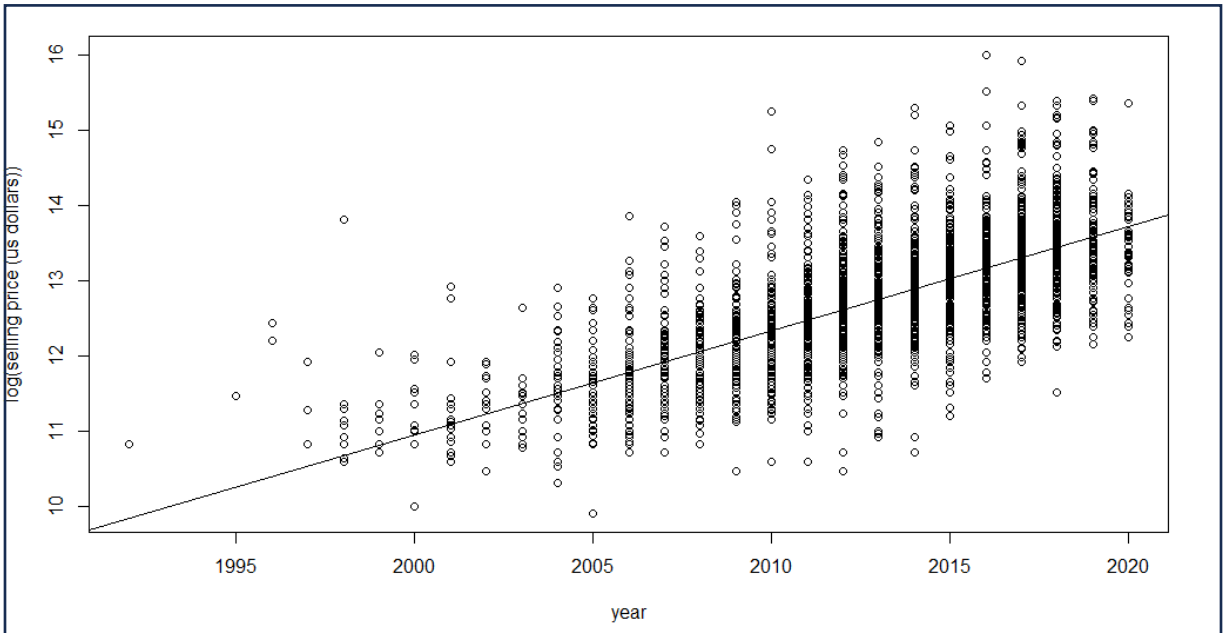
Fig. 2 shows that BMW, Land Rover, Mercedes-Benz, and Volvo tend to grab the higher prices, with some Audi models having outlier prices in the same region as the top 50% of both BMW and Mercedes-Benz. Some models of Ford, Mahindra, and Toyota can sell for higher prices than the other 21 makes.




```
> summary(model.year)

Call:
lm(formula = selling_price_usd ~ year)

Residuals:
   Min     1Q Median     3Q    Max
 -8150  -2739  -1127    578  97693

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.351e+06  4.533e+04  -29.82   <2e-16 ***
year         6.743e+02  2.252e+01   29.95   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6252 on 4338 degrees of freedom
Multiple R-squared:  0.1713,    Adjusted R-squared:  0.1711
F-statistic: 896.9 on 1 and 4338 DF,  p-value: < 2.2e-16
```

Fig. 3 Selling price vs. year, with summary of model

Fig. 3 shows a postive linear trend with selling price vs. year and the summary of the model. The summary shows that year is important in predicting selling price, when used alone, but the adjusted r-squared value is not very impressive, however, with only 17% of the variation in selling price account for by the model on year. K-S test and Levene's test showed residuals are not normal, nor do they have constant variance, respectively. Boxcox test showed that a log(y) transformation may be the best. Fig. 4 shows the results of the log(selling price) transformation.




```
lm(formula = selling_price_tr ~ year)

Residuals:
     Min       1Q   Median       3Q      Max
-2.17571 -0.40646 -0.01491  0.29168  3.14165

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.661e+02  4.371e+00  -60.87   <2e-16 ***
year         1.385e-01  2.171e-03   63.79   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6029 on 4338 degrees of freedom
Multiple R-squared:  0.484,     Adjusted R-squared:  0.4839
F-statistic:  4069 on 1 and 4338 DF,  p-value: < 2.2e-16
```

Fig. 4 Log(selling price) vs. year plot and summary

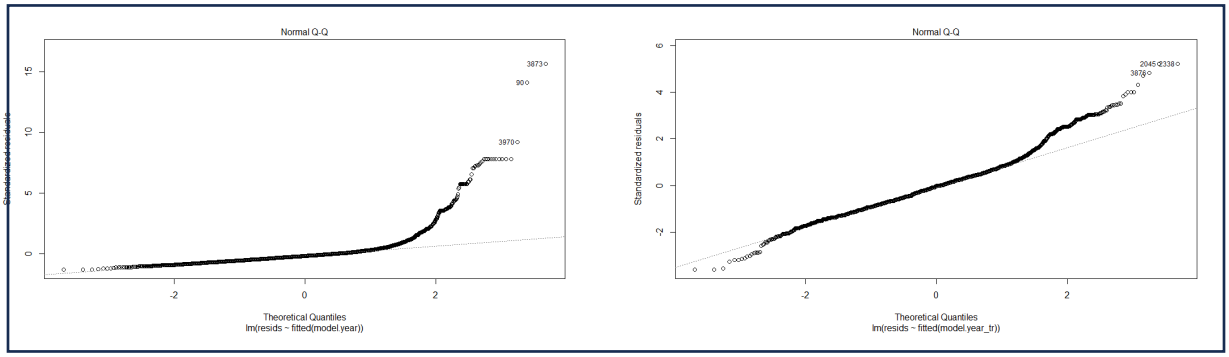


Fig. 5 Normal QQ plot without transformation (left), Normal QQ plot after log(y) transformation (right)

We can see from Fig. 5 that this transformation calmed down the upper end of the Normal QQ plot of the residuals quite a bit from the original. Since there are other variables to consider, transforming y as such may not work for them. K-S test and Levene's test show that the residuals are not normal and do not have constant variance. Boxcox test shows the best transformation might be log(y). After trying the log(y) transformation, we get an r-squared value of 48% but still had the same issues with non-normality and non-constant variance of the residuals. A further Boxcox test suggested a -1 power, so 1/log(y) was employed and that one gave us an r-squared of 51% with the K-S and Levene's test having higher p-values, but still not showing what we wanted.

Cars that have higher mileage tend to sell for less. In Fig. 6, we see the predictor km_driven is important in predicting selling price, when used alone, but the r-squared value is a paltry 3.7%. A log(selling price) transformation suggested by Boxcox test did not prove to be much additional help, with an r-squared value of only about 6%. Also suffering from the residuals lack of normality and constant variance as the original. It seems that since year would benefit from log(selling price) but km_driven would not, it might be a futile effort to try to transform each of these individually, so I waited to try any more transformations until the end.

```
> summary(model.km)

Call:
lm(formula = selling_price_usd ~ km_driven)

Residuals:
    Min      1Q  Median      3Q     Max
 -7126   -3484   -1565    1149   98148

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.858e+03  1.777e+02   44.23   <2e-16 ***
km_driven   -2.831e-02  2.194e-03  -12.91   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6740 on 4338 degrees of freedom
Multiple R-squared:  0.03698,   Adjusted R-squared:  0.03675
F-statistic: 166.6 on 1 and 4338 DF,  p-value: < 2.2e-16
```

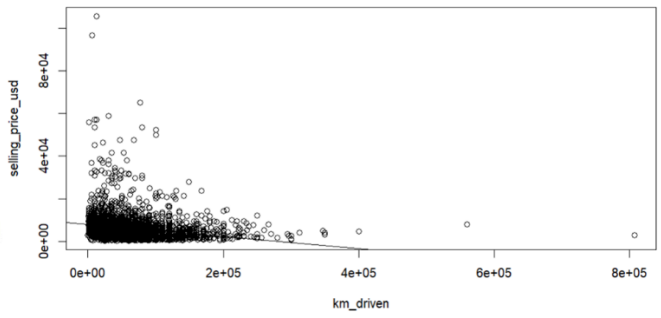
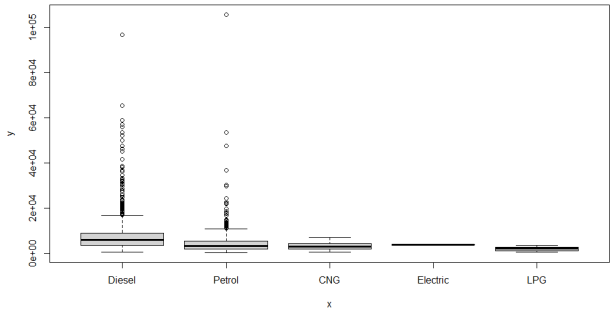



Fig. 6 Selling price vs. km_driven with summary of model

All fuel types except electric appear to important in predicting selling price when used alone. The r-squared value still an unimpressive 8%. Still the same issues with residuals as above. Boxcox test suggests a log(selling price) transformation. See Fig. 7 for plot and summary.




```
lm(formula = selling_price_usd ~ fuelf)

Residuals:
    Min      1Q  Median      3Q     Max
 -7408   -3076   -1262    1132  101545

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    7941.8      142.0  55.941  < 2e-16 ***
fuelfPetrol   -3848.7      201.5 -19.102  < 2e-16 ***
fuelfCNG      -4651.9     1051.2  -4.425 9.86e-06 ***
fuelfElectric -4261.8     6588.8  -0.647    0.518
fuelfLPG      -5949.8     1380.9  -4.309 1.68e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6587 on 4335 degrees of freedom
Multiple R-squared:  0.08067,   Adjusted R-squared:  0.07982
F-statistic: 95.09 on 4 and 4335 DF,  p-value: < 2.2e-16
```

Fig. 7 Selling price vs. fuel type and summary of model

Finishing up selling price vs. the rest of the individual predictors, it was found that seller type, transmission type, and number of owners were all individually important predictors, however each with low r-squared values of 6%, 28%, and 4%, respectively. Each of them also had residuals that were both non-Normal and had a non-constant variance. Each of their individual Boxcox test showed that a log(selling price) transformation would be best. Fig. 8 shows the plots of selling price vs. these last three predictors.

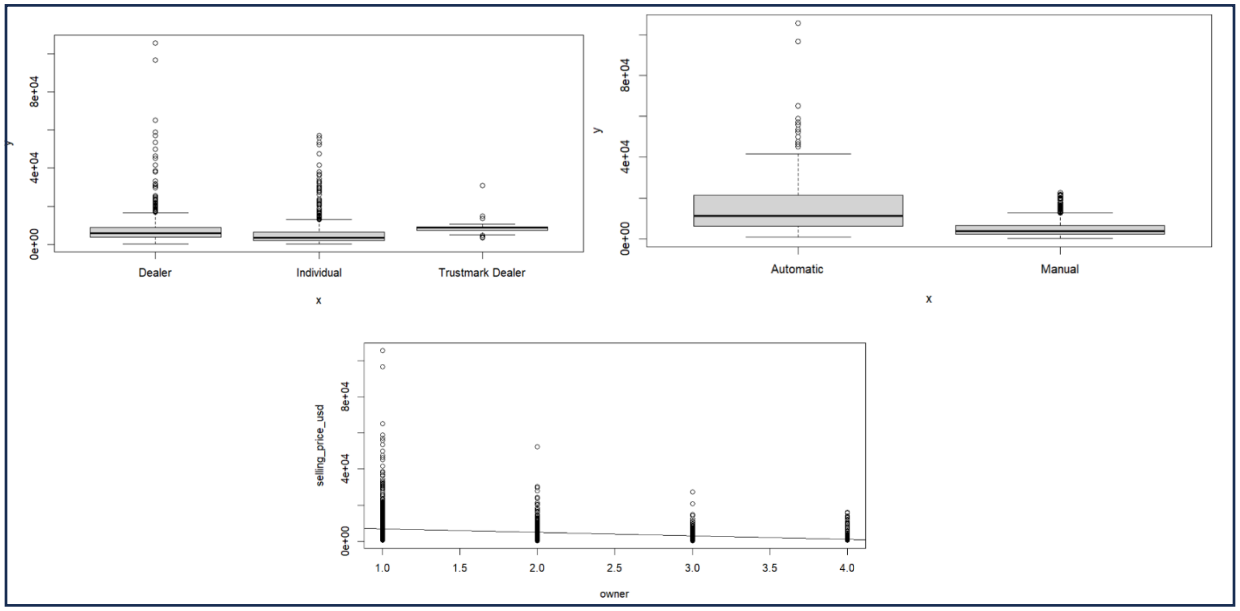



Fig. 8 Selling price vs. seller type, transmission type, and owner

After performing a best subset selection algorithm with all the variables, partially shown in Fig. 9, the full model has adjusted r-squared of 68.44%. With least important variable, owner, removed, adjusted r-squared is unchanged. It was much more difficult to figure out which variable to remove next, due to all the different factors of the variables. So I did an anova on the model with all variables except owner, shown in Fig. 10.

```
          makefAudi makefBMW makefChevrolet makefDaewoo makefDatsun makefFiat makefForce makefFord
1  ( 1 )  " "       " "      " "           " "        " "        " "       " "        " "
2  ( 1 )  " "       " "      " "           " "        " "        " "       " "        " "
3  ( 1 )  " "       "*"      " "           " "        " "        " "       " "        " "

23 ( 1 )  "*"       "*"      "*"           " "        "*"        " "       " "        "*"
24 ( 1 )  "*"       "*"      "*"           " "        "*"        " "       " "        "*"
25 ( 1 )  "*"       "*"      "*"           " "        "*"        " "       " "        "*"
26 ( 1 )  "*"       "*"      "*"           " "        "*"        " "       " "        "*"
```

```
          makefHonda makefHyundai makefIsuzu makefJaguar makefJeep makefKia makefLand makefMahindra
1  ( 1 )  " "        " "         " "        " "         " "       " "      " "       " "
2  ( 1 )  " "        " "         " "        " "         " "       " "      " "       " "
3  ( 1 )  " "        " "         " "        " "         " "       " "      " "       " "

23 ( 1 )  "*"        " "         " "        "*"         "*"       " "      "*"       "*"
24 ( 1 )  "*"        " "         "*"        "*"         "*"       " "      "*"       "*"
25 ( 1 )  "*"        " "         "*"        "*"         "*"       " "      "*"       "*"
26 ( 1 )  "*"        " "         "*"        "*"         "*"       " "      "*"       "*"
```

```
          makefMaruti makefMercedes-Benz makefMG makefMitsubishi makefNissan makefOpelCorsa
1  ( 1 )  " "         " "                 " "     " "            " "         " "
2  ( 1 )  " "         " "                 " "     " "            " "         " "
3  ( 1 )  " "         " "                 " "     " "            " "         " "

23  ( 1 ) " "         "*"                 "*"     "*"            " "         " "
24  ( 1 ) " "         "*"                 "*"     "*"            " "         " "
25  ( 1 ) " "         "*"                 "*"     "*"            " "         " "
26  ( 1 ) " "         "*"                 "*"     "*"            " "         " "

          makefRenault makefSkoda makefTata makefToyota makefVolkswagen makefVolvo year km_driven
1  ( 1 )  " "         " "        " "       " "         " "            " "        " "  " "
2  ( 1 )  " "         " "        " "       " "         " "            " "        "*"  " "
3  ( 1 )  " "         " "        " "       " "         " "            " "        "*"  " "

23  ( 1 ) "*"         " "        "*"       "*"         " "            "*"        "*"  "*"
24  ( 1 ) "*"         " "        "*"       "*"         " "            "*"        "*"  "*"
25  ( 1 ) "*"         " "        "*"       "*"         " "            "*"        "*"  "*"
26  ( 1 ) "*"         " "        "*"       "*"         " "            "*"        "*"  "*"

          fuelfPetrol fuelfCNG fuelfElectric fuelfLPG seller_typefIndividual
1  ( 1 )  " "         " "      " "          " "      " "
2  ( 1 )  " "         " "      " "          " "      " "
3  ( 1 )  " "         " "      " "          " "      " "

23  ( 1 ) "*"         "*"      " "          " "      " "
24  ( 1 ) "*"         "*"      " "          " "      " "
25  ( 1 ) "*"         "*"      " "          "*"      " "
26  ( 1 ) "*"         "*"      "*"          "*"      " "

          seller_typefTrustmark Dealer transmissionfManual owner
1  ( 1 )  " "                           "*"                 " "
2  ( 1 )  " "                           "*"                 " "
3  ( 1 )  " "                           "*"                 " "

23  ( 1 ) "*"                           "*"                 " "
24  ( 1 ) "*"                           "*"                 " "
25  ( 1 ) "*"                           "*"                 " "
26  ( 1 ) "*"                           "*"                 " "
```

Fig. 9 Best subset selection on full model

```
> anova(model.2)
Analysis of Variance Table

Response: selling_price_usd
                Df     Sum Sq    Mean Sq  F value    Pr(>F)
makef           28 1.0433e+11 3.7259e+09  250.361 < 2.2e-16 ***
year             1 2.8039e+10 2.8039e+10 1884.091 < 2.2e-16 ***
km_driven        1 5.0859e+08 5.0859e+08   34.174 5.413e-09 ***
fuelf            4 2.8925e+09 7.2312e+08   48.590 < 2.2e-16 ***
seller_typef     2 1.3031e+09 6.5154e+08   43.780 < 2.2e-16 ***
transmissionf    1 3.5189e+09 3.5189e+09  236.450 < 2.2e-16 ***
Residuals     4302 6.4023e+10 1.4882e+07
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fig. 10 Anova for selling price vs. all predictors, except owner

Next, km_driven was removed as it was the least important. The resulting adjusted r-squared value was a slightly lower 67.95%. Now, the rest of the variables seemed still be *very* important. Looking at the best subset selection results again, year hung in there all the way to the best model with 2 predictors. Of the qualitative variables, seller type was the first be eliminated altogether as the number of predictors decreased. This is how I decided to remove seller type next. The adjusted r-squared value only fell to 67.38%. Only planning to remove one more, I decided upon fuel type, as it was the least represented in models of lower predictor numbers. The resulting adjusted r-squared value is 66.21%. What we had left was make, year, and transmission. I removed one at a time and left the other two in to see what would happen. Removing make caused an adjusted r-squared value of 39.73%, removing year caused an adjusted r-squared value of 54.08%, and removing transmission caused an adjusted r-squared value of 64.45%. If I was going to remove another, I would remove transmission.

Analyzing it the way I did, the BIC and AIC values are both lowest for the model with selling price vs. make, year, km_driven, fuel, seller type, and transmission. This one also has the highest adjusted r-squared value. I found the lowest BIC and AIC, and largest adjusted r-squared with the regsubset summary, as shown in Fig. 11. Adjusted r-squared was highest for a model with 29 predictors (some factors of make, year, km_driven, some factors of fuel type, some factors of seller type, and transmission), BIC was lowest for a model with 20 predictors, and AIC was lowest for a model with 27 predictors. The models with 20 and 27 predictors use the same mix of variables as the model with 29.

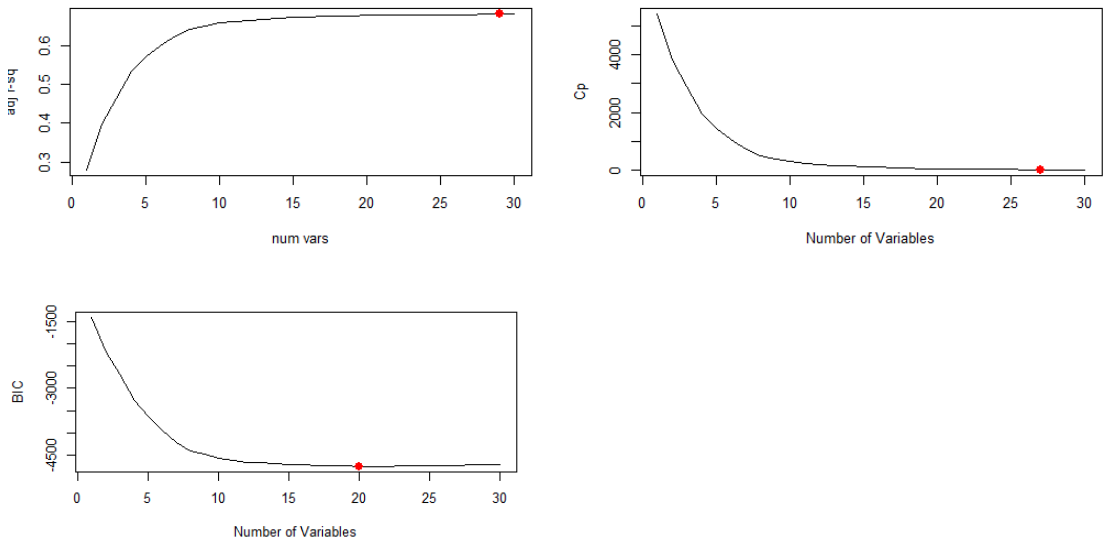



Fig. 11 Analysis of best subset selection

Boxcox test for the model chosen by either method shows that a log(selling price) transformation may help. The suggested transformation was implemented. Fig. 12 shows the Normal QQ plot of the residuals, the standardized residulas vs. fitted values, and the partial summary of the model, showing a much improved adjusted r-squared value of 79.24%, and the K-S test results.

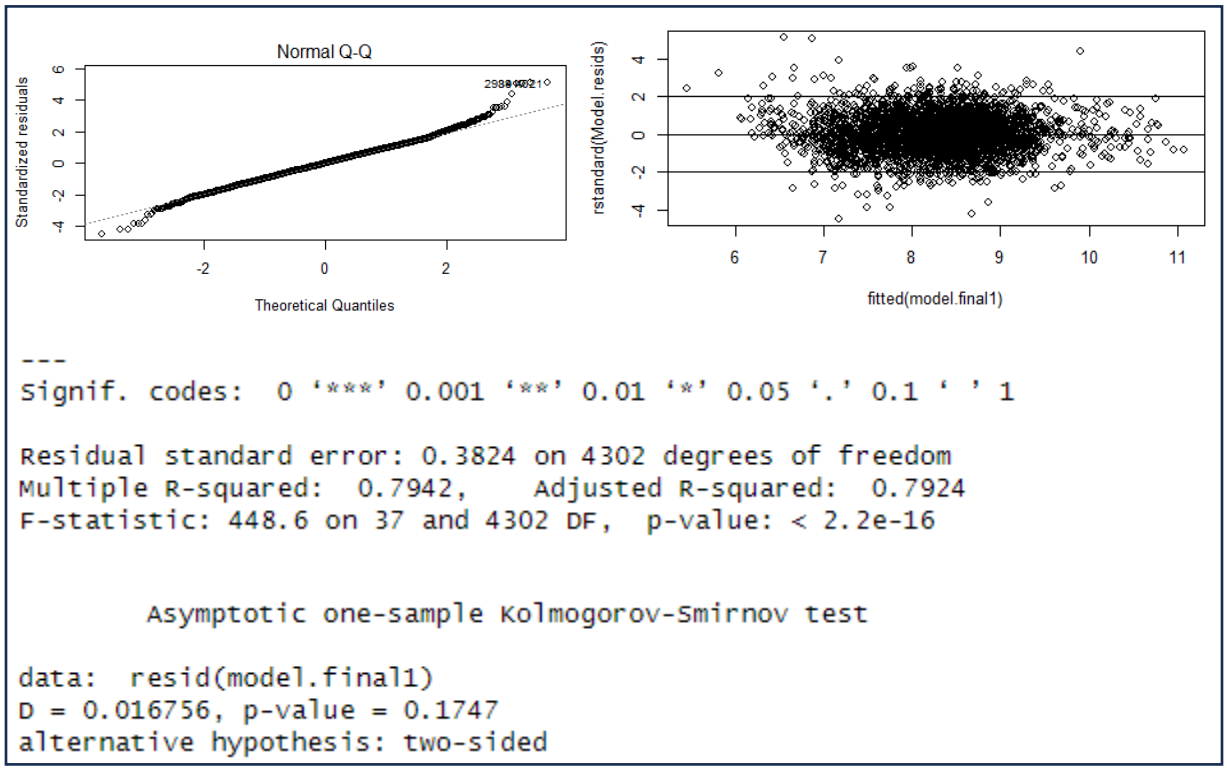



Fig. 12 Final model residual plots, summary, and K-S test results

This final model's K-S test shows Normality in residuals, due to high p-value. Levene's test failed due to the inclusion of the factors. If the factors are left out, Levene's test shows the residuals still do not have a constant variance. I am not sure how accurate this test is without the factors that are in the model. The Boxcox test for this final model showed a power of 0.5 might be better. After trying this on the final model, the adjusted r-squared was slightly lower.

The final model is (*keep in mind that makef, fuelf, seller_typef, and transmissionf are all factors, taking either a 1 or 0*):

-38.68 + 0.206(**Audi**) + 0.22(**BMW**) – 0.031(**Chevrolet**) – 0.013(**Daewoo**) – 0.032(**Datsun**) – 0.014(**Fiat**) – 0.012(**Force**) + 0.035(**Ford**) + 0.072(**Honda**) + 0.035(**Hyundai**) + 0.15(Isuzu) + 0.295(**Jaguar**) + 0.174(**Jeep**) + 0.047(**Kia**) + 0.309(**Land Rover**) + 0.062(**Mahindra**) + 0.014(**Maruti**) + 0.256(**Mercedes-Benz**) + 0.15(**MG**) + 0.213(**Mitsubishi**) + 0.034(**Nissan**) – 0.0009(**OpelCorsa**) + 0.0077(**Renault**) + 0.059(**Skoda**) – 0.05(**Tata**) + 0.13(**Toyota**) + 0.045(**Volkswagon**) + 0.221(**Volvo**) + 0.027(**year**) – 0.00000014(**km_driven**) – 0.071(**Petrol**) – 0.084(**CNG**) – 0.064(**Electric**) – 0.097(**LPG**) – 0.016(**Individual**) + 0.061(**Trustmark**) – 0.055(**Manual**)

### III. Conclusion

Best subset selection showed that the final model included all of the variables except owner. I was surprised that many predictors would be left in the end. With the K-S test showing non-Normality of residuals on all but the final model, I was also pleasantly surprised to see that. Levene's test kept giving an error with this model. So we still don't know if the condition of constant variances of the residuals is

met. From the plot, it sure looks like we have it! Boxcox showed that another slight transformation might work, but upon trying it, I found it did not really help.

Working with a dataset full of qualitative variables is not easy or straightforward, but is important. It would have been helpful to know what the year variable meant. We really would like to know the age of the car. If the year is the model year of the car, then we still don't know its age when this purchase took place. If the year is the used purchase year, then we don't know the model year. It seems more likely that it is the model year because with kilometers driven and the other information, the estimate of the used car value can be looked up on sites like Kelly Blue Book. The model looks large, but quickly diminishes as soon as the make and fuel type is known.

If I were going to do this project again, I would do an in-depth study of how to work with categorical variables. The number of items in the make category could have been diminished by making an "other" level and putting many of the car makes into it. It would be interesting to see how many cars were listed for each make in this dataset, as it is possible that some of the makes have very few cars accounted for and could be removed from the data altogether.

## References


1. Birla, Nehal. "Vehicle Dataset." Kaggle, January 14, 2023. https://www.kaggle.com/datasets/nehalbirla/vehicle-dataset-from-cardekho.