A Report

on

Regression model to estimate the selling price of a car for "CarDekho" Company

Prepared by

Aviral Jain (2019B3A70317P)

Bhavya Patel (2019B3A30534P)

Prakhar Gupta (2019B3A70516P)

Hari Sankar (2019B3A70564P)

Prepared under

Dr NVM Rao, Professor

Department of Economics and Finance, BITS Pilani

Problem statement

CarDekho is a company that buys second-hand cars from individuals. It comes up with a valuation for the cars after a thorough inspection. This inspection includes noting down the age of the car, the distance travelled, mileage, engine, maximum power, torque and number of seats. Given historical data, come up with a model to estimate the selling price of a car.

Q1. **Model Development:**

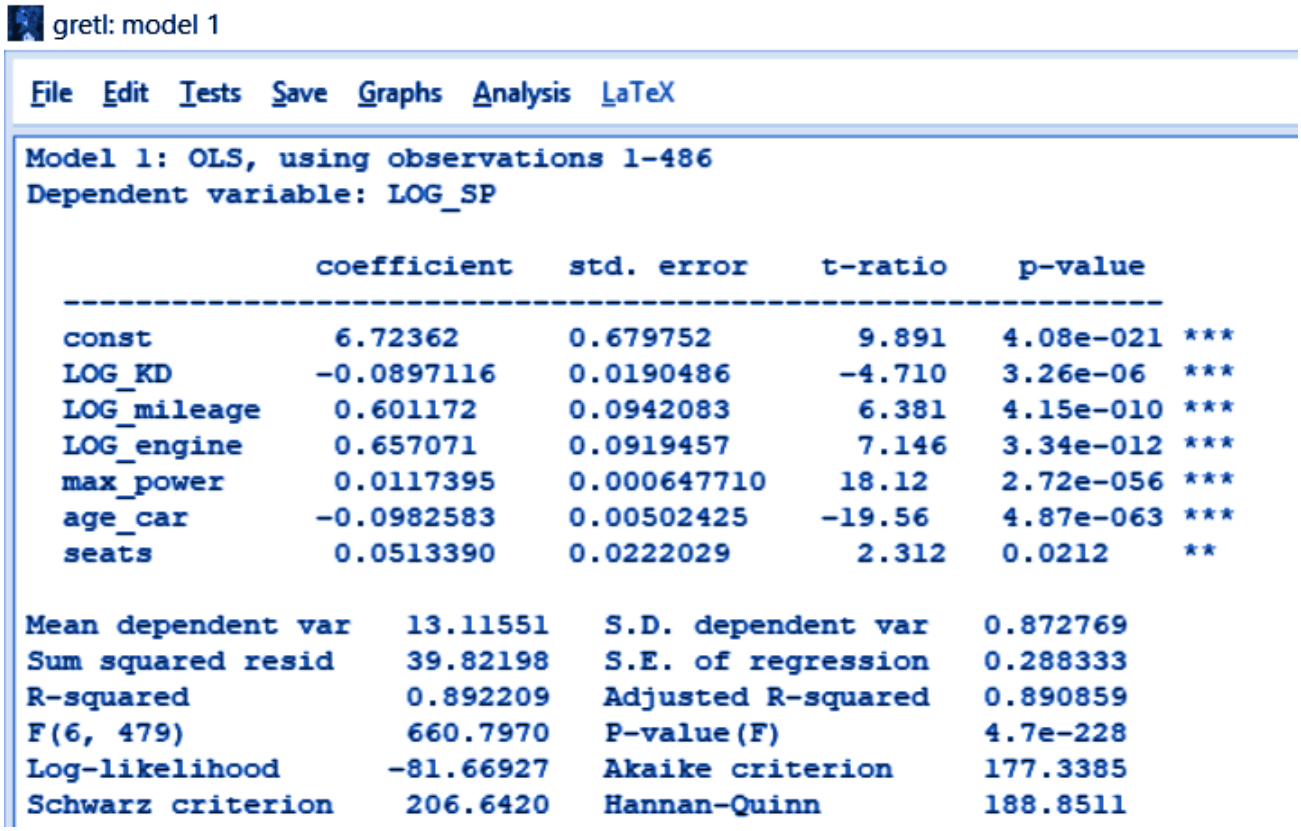
- **Regression Model:**

$$LOG\_SP = \beta_1 + \beta_2 LOG\_KD + \beta_3 LOG\_MIL + \beta_4 LOG\_ENG + \beta_5 MAXPWR + \beta_6 AGE + \beta_7 SEATS +$$

1. LOG_SP= natural logarithm of the selling price
2. LOG_KD= natural logarithm of km driven
3. LOG_MIL= natural logarithm of mileage
4. LOG_ENG= natural logarithm of engine capacity (in cc)
5. MAXPWR= maximum power (in bhp)
6. AGE= age of the car (in years)
7. SEATS= number of seats
8. β_i's = coefficient estimates, where i= 1,2,3,.....,7.
9. ϵ= error term assumed to be normal, identically and independently distributed.

- Data cleaning - If the number of cases of missing values is extremely small; then, we may drop those values from the analysis. As the number of values missing, in this case, is 1.625% of the sample and are limited to 13 rows out of 500 rows (2.6%) we have decided to drop them. Some of the values in the mileage column were in km/kg which were converted to km/l by multiplying them with 0.8 (specific gravity of petrol). We have also removed outliers (ex: high selling price value = 1000000) to obtain better results.
- We have excluded the torque variable with rpm because there is an exact relation between torque, rpm and power given by H = T x rpm/5252, H is horsepower, T is pound-feet, rpm is how fast the engine is spinning, and 5252 is a constant that makes the units jibe. We have included other variables after further reading as we felt that they have a significant impact on the selling price of the car. As km driven or the age of the car increases the car gets worn out and its selling price will decrease. If max power increases the car will travel at higher speeds and hence more people will desire it. If mileage is higher it becomes fuel-efficient, meaning people will have to spend less on petrol costs later and its initial selling price will be higher. More no. of seats more people can be seated comfortably and can travel together from one place to another and hence its selling price will be more.
- We first formed a linear-linear model and got Adj R^2 as 0.74. Then seeing the large variances in data of some variables, we tried to take the natural log of those variables to make a different model and got an improved Adj R^2. Then, we plotted the graphs of each explanatory variable with the dependent variable to get a better idea of whether to log or inverse or add it to the model without transformation. Finally, we came up with the above model which had the Maximum Adj R^2 value = 0.890859.

Q2. **Find the coefficients and interpret them.**



(Note: Above Screenshot is being used for Both Q2 and Q3.)

Estimator	Value	Interpretation
β_2	-0.0897116	For every unit percentage increase in the value of KM, there will be a 0.0897116% decrease in the value of SP while holding constant values of other explanatory variables(Ceteris Paribus).
β_3	0.6011717	For every unit percentage increase in the value of MIL, there will be a 0.6011717 % increase in the value of SP while holding constant values of other explanatory variables(Ceteris Paribus).

β_4	0.657071	For every unit percentage increase in the value of ENG, there will be a 0.657071 % increase in the value of SP while holding constant values of other explanatory variables(Ceteris Paribus).
β_5	0.0117395	For every unit increase in the value of MAXPWR, there will be a 1.17395% increase in the value of SP while holding constant values of other explanatory variables(Ceteris Paribus).
β_6	-0.0982583	For every unit decrease in the value of AGE, there will be a 9.82583% increase in the value of SP while holding constant values of other explanatory variables(Ceteris Paribus).
β_7	0.051339	For every unit increase in the value of SEATS, there will be a 5.1339% increase in the value of SP while holding constant values of other explanatory variables(Ceteris Paribus).
R^2	0.8922	89.22% of the total sample variation of the dependent variable (LOG SP) is being explained by the regression model (i.e, by the values of the regressors). R-squared is the statistical measure of how close the data are to the fitted regression line.
Adj R^2	0.8909	Indicates how well terms fit a curve or line and adjusts for the number of terms in a model. 0.8909 value indicates that explanatory variables added are useful (statistically significant). The adjusted R-squared increases only if the new term improves the model more than would be expected by chance.

Q3. **Conduct F-test and t-tests.**

1) F-Test:

- H_0: Overall Model is not significant.
  Mathematically, $\beta_2=\beta_3=\beta_4=\beta_5=\beta_6=\beta_7=0$
- H_1: Overall Model is significant.
  Mathematically, At least one $\beta_i \neq 0$, where i = 2,3,….,7.

(**Note:** See F-stat value in the table shown in Q2)
F_{cal} = 660.797
F_{tab} (df = 6,479) = 2.11749825 at 95% CI**.**

Since $F_{cal} > F_{tab}$
=> We reject H_0 at 95% CI.
=> **Overall Model is significant.**

2) t-Test:

(Following are the tests that need to be performed individually. Pl. consider it as 7 different hypotheses.)

- H_0: Individual parameters are not significant.
  Mathematically, $\beta_i = 0$. where i = 1,2,3,….,7.
- H_1: Individual parameters are significant.
  Mathematically, $\beta_i \neq 0$, where i = 1,2,3,….,7 .

(**Note**: See individual t-stat value in table shown in Q2)
t-tab (df = 484) =1.96487746 at 95% CI.

Every $|t_{cal}|$ value is greater than $|t_{tab}|$.
=> We reject all H_0 at 95% CI.
=> **All individual parameters are significant (i.e, $\beta_i \neq 0$. where i = 1,2,3,….,7)**

Q4. **Conduct tests for multicollinearity, heteroscedasticity and autocorrelation.**

❖ Tests for multicollinearity:

- **H_0: Multicollinearity does not exist.**
- **H_1: Multicollinearity exists.**

★ To check, we performed the tests shown below:

A. As the VIF values of all the coefficients lie between 1-5 we can safely conclude that our model does not have multicollinearity

```
Variance Inflation Factors
Minimum possible value = 1.0
Values > 10.0 may indicate a collinearity problem

      LOG_KD    1.743
 LOG_mileage    2.249
  LOG_engine    4.838
   max_power    3.912
     age_car    2.054
       seats    1.806

VIF(j) = 1/(1 - R(j)^2), where R(j) is the multiple correlation coefficient
between variable j and the other independent variables
```

B. Referring to the t-test and F-test performed in the above question we can conclude that **$F_{cal} > F_{tab}$, we** H_0 **reject at 95% CI and** every $|t_{cal}| > |t_{tab}|$ and we reject all H_0 at 95% CI. Hence, we can say multicollinearity does not exist in our model.

Inference - Multicollinearity is not present.

❖ Testing for heteroscedasticity: White test(with cross terms):

- **H_0: homoscedasticity Var (u_i) =σ^2**
- **H_1: heteroscedasticity Var (u_i) =σ_i^2**

W_{cal}=57.3783

$X^2_{tab}(df=27) = 46.963$

Since $W_{cal} > X^2_{tab}$

We reject H_0 at 90% CI

Inference - Heteroscedasticity is present.

Remedy - We will first find $Var(u_i) = E(u_i)^2 = \sigma_i^2$ as a function of x_i (i.e. $f(x_i)$) for all heteroscedastic variables. Then we would divide the whole regression model with $\sqrt{f(x_i)}$ which will give us constant error variance (Weighted least squares method). By this, we would be able to remove the problem of Heteroscedasticity.

gretl: LM test (heteroskedasticity)

```
White's test for heteroskedasticity
OLS, using observations 1960-2445 (T = 486)
Dependent variable: uhat^2

                    coefficient     std. error    t-ratio   p-value
  ------------------------------------------------------------------
  const             -10.6367        14.9098       -0.7134   0.4760
  LOG_KD             -0.944171       0.619266     -1.525    0.1280
  LOG_mileage         2.52284        2.99785       0.8415   0.4005
  LOG_engine          2.71654        3.54971       0.7653   0.4445
  max_power           0.00736629     0.0192871     0.3819   0.7027
  age_car             0.0562908      0.145758      0.3862   0.6995
  seats               0.677672       0.625559      1.083    0.2792
  sq_LOG_KD           0.00710365     0.00737666    0.9630   0.3361
  X2_X3              -0.112251       0.0746765    -1.503    0.1335
  X2_X4               0.203493       0.0891221     2.283    0.0229   **
  X2_X5              -0.00149111     0.000604226  -2.468    0.0140   **
  X2_X6               0.000660385    0.00370425    0.1783   0.8586
  X2_X7              -0.0360518      0.0224046    -1.609    0.1083
  sq_LOG_mileage      0.443261       0.234559      1.890    0.0594   *
  X3_X4              -0.725186       0.392045     -1.850    0.0650   *
  X3_X5               0.00441323     0.00261273    1.689    0.0919   *
  X3_X6               0.0385068      0.0216273     1.780    0.0757   *
  X3_X7               0.123500       0.0817220     1.511    0.1314
  sq_LOG_engine      -0.120237       0.243645     -0.4935   0.6219
  X4_X5              -0.00209761     0.00291927   -0.7185   0.4728
  X4_X6              -0.0326329      0.0188747    -1.729    0.0845   *
  X4_X7              -0.132240       0.0928721    -1.424    0.1552
  sq_max_power        1.16158e-05    1.03280e-05   1.125    0.2613
  X5_X6               0.000170593    0.000152487   1.119    0.2638
  X5_X7               0.00154750     0.000740500   2.090    0.0372   **
  sq_age_car          0.00137010     0.000630522   2.173    0.0303   **
  X6_X7               0.00486474     0.00498539    0.9758   0.3297
  sq_seats            0.0129250      0.0145320     0.8894   0.3742

  Unadjusted R-squared = 0.118062

Test statistic: TR^2 = 57.378319,
with p-value = P(Chi-square(27) > 57.378319) = 0.000576
```

❖ Testing for the first-order autocorrelation: Breusch-Godfrey test

- **H_0: No autocorrelation ($\rho=0$)**
- **H_1: Autocorrelation ($\rho\neq0$)**

LM=0.275671

X^2_{tab}(df=1) = 2.71 at 99% CI

Since $LM_{cal} < X^2_{tab}$

We fail to reject H_0 at 99% CI.

Inference- No autocorrelation in the model.

The test is not reliable as it is not a time series data and hence we will not take its results into consideration.

gretl: autocorrelation

```
Breusch-Godfrey test for first-order autocorrelation
OLS, using observations 1960-2445 (T = 486)
Dependent variable: uhat

               coefficient    std. error     t-ratio    p-value
  ------------------------------------------------------------
  const          -0.00197240    0.680280      -0.002899   0.9977
  LOG_KD         -0.000353624   0.0190752     -0.01854    0.9852
  LOG_mileage    -0.00180383    0.0943436     -0.01912    0.9848
  LOG_engine      0.00174672    0.0920768      0.01897    0.9849
  max_power      -2.62010e-05   0.000650153   -0.04030    0.9679
  age_car         2.50830e-05   0.00502831     0.004988   0.9960
  seats           0.000155749   0.0222218      0.007009   0.9944
  uhat_1          0.0239529     0.0459879      0.5209     0.6027

  Unadjusted R-squared = 0.000567

Test statistic: LMF = 0.271287,
with p-value = P(F(1,478) > 0.271287) = 0.603

Alternative statistic: TR^2 = 0.275671,
with p-value = P(Chi-square(1) > 0.275671) = 0.6
```