Multivariable Regression Analysis for prediction of factors affecting car mileage

Nikhil Sharad Giramkar
Department of Computer Engineering,
Modern Education Society's College of Engineering,
19, Late Prin. V.K. Joag Path,Wadia College Campus, Pune – 411001.

Abstract: Mileage is one of the most important factors for working class people while buying a new four-wheeler. There are certain factors which the engineers consider while manufacturing the vehicle. The design has a huge impact on mileage; hence they claim the mileage of their vehicles by keeping certain factors in their mind. In this paper, multivariable regression analysis has been done on a dataset consisting of specifications of four-wheeler petrol variant cars to predict the elements which affect the mileage in km/l. Descriptive statistics of all these elements has been provided. Pearson's correlation among them has also been established. A linear regression model with best fit has been applied on the collected sample. A regression equation has also been obtained which represents the model. Interpretation from the analysis and discussion of results has also been provided at the end.

Keywords: Car, Mileage, Variables, Multivariable regression, Regression model, Pearson's correlation.

I. INTRODUCTION

In the present study, it has been found that length, height and width (aerodynamics) of a car, its gross weight, fuel tank capacity, power (in terms of kW) and whether the four-wheeler has manual or automatic transmission are some basic elements which if changed will also alter the mileage in some extent. Muhammad Usman Ghani et.al used the technique of multivariable regression to calculate fuel economy. He suggested how the four-wheelers are precisely manufactured to give maximum efficiency while considering these factors [1].

Linear regression is a technique used to model a relationship between one dependent variable and one or more independent variables for a given dataset. Multivariable regression analysis is an extension to linear regression, where two or more independent variables hold a relationship with a dependent variable. Thus, if y is the dependent variable and $x_1, x_2, \ldots\ldots x_k$ be the independent variables then they can be represented in a form:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots\ldots \beta_k x_k + \varepsilon \qquad (1)$$

Where, β_0 is the intercept of the model, β_1, β_2 ...β_k are the coefficients of the independent variables and ε is the random error. We further assume that ε is normally and independently distributed for any given value of x_i [2]. The multivariable regression model is based on the following assumptions:

1) The random error term ε is normally distributed and has an expected value of zero and a constant variance σ^2.
2) The independent variables $x_1, x_2 \ldots\ldots x_k$ are known constants.
3) The coefficients $\beta_1, \beta_2, \ldots\ldots\ldots \beta_k$ are parameters hence, constants [3].

II. MATERIAL AND METHODS:

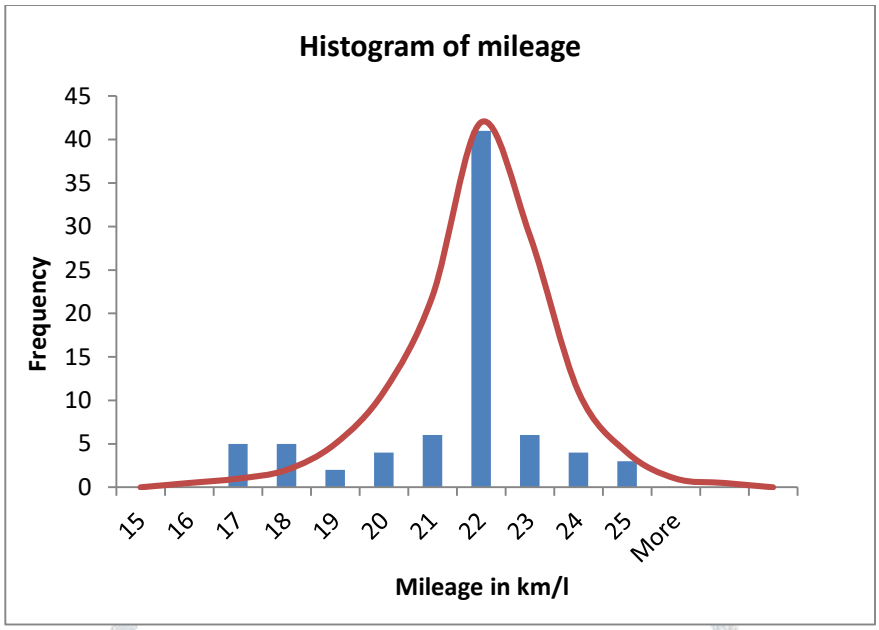
The present study was carried out using a seven-step modeling process [4]. The data was collected from the official website of Maruti Suzuki Arena, a car manufacturing company [5]. The specifications of 76 petrol variant cars such as name and model of the car, number of cylinders in a car, fuel tank capacity, length, height and width of the car, maximum power (in kW at 6000rpm), whether the car has manual or automatic transmission, top speed, gross weight, etc (as independent variables) and mileage of the car (as dependent variable) were collected and stored in an Excel sheet of Microsoft Excel 2007. The whole analysis was carried out using the "Data Analysis Toolpak" available in MS-Excel 2007 [6]. The dependent variable (mileage) was checked for normality in its distribution. Then certain descriptive statistics of both 'x' and 'y' variables were obtained including their mean, median, standard deviation, minimum, maximum and Interquartile Range. Pearson's correlation among all the independent variables was observed for multicollinearity in the sample. Then, a regression model was fit to the data. The summary obtained from regression was used to make interpretations. The results obtained were compared with previous research works.

III. RESULTS AND DISCUSSION

Following results were obtained in the present study:

3.1 Normal Distribution

A histogram of mileage was built from the collected data and validated as per the central limit theorem [7]. The mileage was observed to have a normal distribution, which is shown by a bell curve (red line) in Graph 1.



Graph 1: Histogram of mileage denoting normal distribution of sample collected

The data collected was spread normally around the mean. Hence, the sample was enough to represent all types of petrol variant passenger cars like hatchbacks, sedans, SUVs and vans.

3.2 Descriptive Statistics

A brief summary of the sample dataset is provided in the form of descriptive statistics using various Excel formulae (Table 1). Following measures have been provided for each variable:

1) Measure of Frequency: Number/Count
2) Measure of Central Tendency: Mean, Median
3) Measure of dispersion: Standard deviation
4) Measure of position: Interquartile Range [8].

Table 1. Descriptive statistics of all independent and dependent variables

	Length	Height	Width	Max Power	Top speed	Gross Weight	No. of Cylinders	Fuel tank	Wheel base	Mile age
Number	76	76	76	76	76	76	76	76	76	76
Mean	3768.6	1597.7	1629.4	56.365	159.26	1333.28	3.4736	35.81	2445	20.873
Standard Deviation	240.10	89.401	95.173	11.361	22.517	157.809	0.5026	5.565	94.9912	1.9391
Median	3695	1564	1620	50	150	1315	3	35	2425	21.63
Maxi mum	4395	1825	1790	77	220	1740	4	48	2740	24.12
Mini mum	3445	1475	1475	35.3	140	1170	3	27	2350	16.11
Inter quartile Range	185	145	215	11	20	90	1	27	70	1.2025

3.3 Pearson's Correlation

Pearson's correlation amongst the 'x' variables was checked (Table 2) using the "Correlation" feature available in "Data Analysis Toolpak" of Microsoft Excel-2007. According to Nivea Thomas et.al. the coefficient of correlation is a value between 0 and 1. Generally, all variables have some correlation with each other. A value close to +1 indicates a strong relationship whereas, values close to 0 indicates negligible relationship among the variables [2]. It was found that there was a high correlation (>0.9) between wheelbase and length of the car. Thus, if we run a model with these 'x' variables, the results could have been misleading in determining the impact of each element on mileage. Therefore, to remove the multicollinearity, the wheelbase factor was dropped. The regression model was run with these corrections [9].

Table 2. The table describes the correlation amongst all the independent (X) variables.

	Length	Height	Width	Wheel base	Max Power	Top speed	MT/AGS	Gross Weight	Fuel Tank	No. of Cylinders
Length	1.0000									
Height	0.2354	1.0000								
Width	0.7865	-0.1029	1.0000							
Wheel base	0.9121	0.2412	0.6598	1.0000						
Max power	0.8899	0.3495	0.8031	0.7548	1.0000					
Top Speed	0.6583	-0.1507	0.7530	0.3763	0.6628	1.0000				
MT/AGS	-0.1059	0.0562	-0.2040	-0.0955	-0.1388	-0.1167	1.0000			
Gross Weight	0.7297	0.6315	0.4974	0.6606	0.7551	0.3628	-0.0309	1.0000		
Fuel Tank	0.7434	0.3283	0.6283	0.6204	0.6875	0.4648	0.0111	0.8007	1.0000	
No. of Cylinders	0.6435	0.3889	0.5815	0.4412	0.8005	0.6109	-0.0380	0.6693	0.6226	1.0000

3.4 Regression model

Using the standard equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots\ldots \beta_k x_k \quad (2)$$

The independent variables were substituted in place of x_i (i=0, 1, 2, 3…k) and the dependent variable (mileage) was substituted in place of y. Thus, the equation formed was:

$$\text{Mileage} = \beta_0 + \beta_1(\text{length}) + \beta_2(\text{height}) + \beta_3(\text{width}) + \beta_4(\text{max. power}) + \beta_4(\text{top speed}) + \beta_5(\text{MT/AGS}) + \beta_7(\text{gross wt.}) + \beta_8(\text{fuel tank capacity}) + \beta_9(\text{no. of cylinders}) \quad (3)$$

The output of the regression model was obtained in the form of tables (Table 3-5) [10].
The regression statistics table (Table no. 3) gives the R square value of the model. It was found to be around 0.96, which denotes that the model has fit well to the sample.

Table 3. Regression statistics

Regression Statistics	
Multiple R	0.979288927
R Square	0.959006803
Adjusted R Square	0.953416822
Standard Error	0.418529021
Observations	76

The Significance-F value obtained in the ANOVA table (Table 4) is very low (3.21082E-42). This indicates that probability of the results being non-reliable is 3.21082×10^{-42} %. Hence, the results obtained in present study are reliable.

Table 4. ANOVA table

	df	SS	MS	F	Significance F
Regression	9	270.4612122	30.0512458	171.5581383	3.21082E-42
Residual	66	11.56099175	0.175166542		
Total	75	282.0222039			

The value of coefficients (β_i) and P-value for each element was obtained (Table 5). The value of coefficients is shown in yellow highlight while the P-values are shown in green and red highlight. The P-values highlighted in green color imply that the coefficients obtained have strong evidence (P-value < 0.05) while those highlighted in red color imply that the coefficients obtained have weak evidence (P-value > 0.05). The confidence interval was actually set to 95% hence, alpha was 0.05 (therefore, strength of variables were checked with 0.05 as a benchmark) [11]. The lower and upper bounds of the value of coefficient have been provided in last two columns.

Table 5. Coefficients and P-values table

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	11.81846808	3.048138709	3.87727371	0.000245914	5.732662861	17.90427331
LENGTH	0.003024888	0.000544152	5.558900272	5.27265E-07	0.001938453	0.004111323
HEIGHT	-0.004237504	0.00104138	-4.06912148	0.00012846	-0.006316687	-0.00215832
WIDTH	0.010129202	0.001323024	7.656097232	1.09248E-10	0.007487699	0.012770706
MAX POWER(kW)	-0.136494475	0.015722216	-8.681630934	1.60236E-12	-0.167884891	-0.105104059
TOP SPEED	0.039711414	0.003846967	10.32278414	2.07616E-15	0.032030696	0.047392132
MT/AGS	-0.192558355	0.105823082	-1.819625282	0.073351637	-0.403840952	0.018724242
GROSS WEIGHT	-0.000548694	0.000773886	-0.709011831	0.480814208	-0.002093807	0.000996418
FUEL TANK	-0.240023422	0.017869778	-13.43180754	1.49947E-20	-0.275701585	-0.204345259
NO. OF CYLINDERS	-0.36135452	0.193166046	-1.870693774	0.065822087	-0.747022969	0.02431393

Hence, the regression equation for the given sample can be represented by substituting value of coefficients in place of β as follows:

Mileage =11.81846808 + 0.003024888 (length) - 0.004237504 (height) + 0.010129202 (width) - 0.136494475 (max. power) + 0.039711414 (top speed) - 0.192558355 (MT/AGS) - 0.000548694 (gross wt.) - 0.240023422 (fuel tank capacity) - 0.36135452(number of cylinders) (4)

Therefore, for the given sample, the value of mileage in km/l can be predicted using the above equation. Where, following values should be given:

1) Value of length, height and width should be in millimeters;
2) Maximum power of the car in terms of kilowatt;
3) Top speed in km/hr;
4) Value of MT/AGS in terms of 1 or 0 (1 for manual transmission and 0 for automatic gear transmission);
5) Gross weight in terms of kilogram;
6) Fuel tank capacity in terms of liters and
7) Number of cylinders should be given to obtain mileage (in km/l).

3.5 Interpretation of Results:

From the above results, the impact of each independent variable (x) on the mileage (y) of car was depicted as follows:

1) **Intercept (β_0)**: This coefficient gives the value of mileage in km/l when all other variables are zero. This value does not have significance since a situation where all 'x' variables are zero is not possible.
2) **Length (β_1)**: Every one millimeter increase in length of the car will increase the mileage of the car by 0.003 km/l, all other variables being kept at the same level.
3) **Height (β_2):** Every one millimeter increase in height of the car will decrease the mileage of the car by 0.004 km/l, all other variables being kept at the same level.
4) **Width (β_3):** Every one millimeter increase in width of the car will increase the mileage of the car by 0.01 km/l, all other variables being kept at the same level.
5) **Max. Power (β_4):** Every one kilowatt (kW) increase in maximum power of the car will decrease the mileage of the car by 0.136 km/l, all other variables being kept at the same level.
6) **Top speed (β_5):** Every one kilometer per hour (km/hr) increase in top speed of the car will increase the mileage of the car by 0.039 km/l, all other variables being kept at the same level.
7) **MT/AGS (β_6):** When the vehicle has manual transmission then, the mileage of car is lower by 0.192 km/l as compared to automatic gear transmission vehicle, other variables being kept at the same level. The P-value of this variable is greater than 0.05 (alpha) hence, the probability that the value of coefficient might be zero cannot be ruled out. Therefore, there can be no statistical impact of type of transmission on mileage of car.
8) **Gross Weight (β_7):** Every additional kilogram in gross weight of car will decrease the mileage of car by 0.0005 km/l, all other variables being kept at the same level. The P-value of this variable is greater than 0.05 (alpha) hence, the probability that the value of coefficient might be zero cannot be ruled out. Therefore, there can be no statistical significance of gross weight on mileage of car
9) **Fuel Tank (β_8):** Every one liter increase in fuel tank capacity of the car will decrease the mileage of the car by 0.24 km/l, all other variables being kept at the same level.
10) **No. of Cylinders (β_9):** Increase in every one cylinder in the engine of the car will decrease the mileage of car by 0.36km/l, all other variables being kept at the same level. The p-value of this variable is greater than 0.05 (alpha) hence, the probability that the value of coefficient might be zero cannot be ruled out. Therefore, there can be no statistical significance of no. of cylinders on mileage of car.

3.6 Discussion

The results obtained in present study show that gross weight of the vehicle has a negative impact on mileage of the car (β_7 = - 0.000548694). Knittel et.al. found that a 10 percent decrease in weight of the car will increase the fuel economy (mileage) of the car by 4.26 percent. He also found that fuel economy increases by 2.6 percent for every 10 percent reduction in horsepower [12]. Similar results were found in the present study depicting that power has a negative impact on mileage of the car (β_4 = - 0.136494475).

Another element which also affects mileage is the aerodynamic drag. The aerodynamic drag depends on the design of the vehicle. More the drag, less is the fuel economy. Shamsul et. al. found that 15% reduction in aerodynamic drag at highway speed of 55mph can result in about 5–7% in fuel saving [13]. Therefore, to reduce this drag, the design of the vehicle must be optimum. A few modifications in a car result in decrease in drag. By installing front splitters, side skirts and diffusers, a decrease in aerodynamic drag was observed [14]. In present study, these modifications lead to increase in overall length and width of the vehicle. Hence, increase in length and width of the car decreases the aerodynamic drag which leads to increase in mileage. Similarly, in the present study, it was observed that length and width of the car had a positive impact on mileage as per the results (β_1 = + 0.003024888 and β_3 = + 0.010129202).

An increase in ground clearance of the vehicle increases the aerodynamic drag [15]. Thus, to reduce the drag, a car must have minimum ground clearance. In present study, a decrease in ground clearance of the car decreases the overall height. Similarly, in present study it was found in results that height of the car has a negative impact on mileage (β_2 = - 0.004237504).

The fuel tank capacity in liters had a negative impact on mileage (β_8 = - 0.240023422) as it leads to an additional increase in weight of the car.

The cylinders in the engine produce power. More the number of cylinders, more is the power generated. This leads to an increase in fuel consumption. The result also shows that number cylinders have a negative impact on mileage of the car (β_9 = - 0.36135452) but it is less significant.

Therefore, overall results obtained in this study are not contradicting the interpretations made by Knittle, Jason and Debojyoti [12, 13 and 15]. The positive/negative impact of the independent variables in the study can be estimated from the sign (+/-) of their value of coefficients (β).

The significance of these factors can be predicted from their p-value. The value of mileage in km/l for the sample data can also be predicted using the regression equation obtained.

IV. ACKNOWLEDGMENT

The author wishes to thank the team of Maruti Suzuki Arena, M/S, India for their kind support and cooperation. Thanks to Prof. Sudhir Dhanure, Sinhgad Institute of Technology and Sciences, Narhe, Dist: Pune and all my teachers of Modern Education Society's College of Engineering, Pune for their consistent support and guidance during the research work.

V. REFERENCES:

[1] Muhammad Usman Ghani et.al. 2018, Factors Effecting Fuel Consumption of SI Engine – A Case Study, Professional Trends in Industrial and System Engineering (PITSE), pp 602-610.

[2] Nivea Thomas et.al. 2016, Regression Modeling for Prediction of Construction Cost and Duration, Applied Mechanics and Materials. Vol. 857: pp 195-199.

[3] Regression model assumptions: https://www.jmp.com/en_us/statistics-knowledge-portal/what-is-regression/simple-linear-regression-assumptions.html

[4] Albright and Winston 2015, Business Analytics: Data Analysis and Decision Making, Fifth Edition, Cengage Learning, pp 13-15.

[5] Official website of Maruti Suzuki Arena: https://www.marutisuzuki.com/channels/arena/all-cars

[6] Data Analysis Toolpak: https://support.microsoft.com/en-us/office/load-the-analysis-toolpak-in-excel-6a63e598-cd6d-42e3-9317-6b40ba1a66b4

[7] Albright and Winston 2015, Business Analytics: Data Analysis and Decision Making, Fifth Edition, Cengage Learning, pp 323.

[8] Types of Descriptive statistics: https://baselinesupport.campuslabs.com/hc/en-us/articles/204305665-Types-of-Descriptive-Statistics

[9] Richard Williams 2015, Multicollinearity, University of Notre Dame, https://www.google.co.in/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwigya2V_sHqAhVTXHwKHZIMDbEQFjAPegQICRAB&url=https%3A%2F%2Fwww3.nd.edu%2F~rwilliam%2Fstats2%2Fl11.pdf&usg=AOvVaw296xDOuMG_10INbizoqUN

[10] Microsoft Office Support: https://support.microsoft.com/en-us/office/use-the-analysis-toolpak-to-perform-complex-data-analysis-6c67ccf0-f4a9-487c-8dec-bdb5a2cefab6

[11] Analysis of Discrete Data: https://online.stat.psu.edu/stat504/node/19/

[12] Knittel et.al. 2009, Automobiles on Steroids: Product Attribute Trade-offs and Technological Progress in the Automobile Sector, National Bureau of Economics Research.

[13] Shamsul Anuar Shamsuddin et. al. 2014, Review of Research on Vehicles Aerodynamic Drag Reduction Methods, International Journal of Mechanical & Mechatronics Engineering IJMME-IJENS Vol:14 No:02, pp 36.

[14] Jason Moffat. 2016, Aerodynamic Vehicle Design and Analysis, Blackpool and the Fylde College, Foundation Degree in Motorsport Engineering, pp 8. Elink:https://www.researchgate.net/publication/310019902_Aerodynamic_Vehicle_Design_and_Analysis

[15] Debojyoti Mitra 2010, Design Optimization of Ground Clearance of Domestic Cars, International Journal of Engineering Science and Technology Vol. 2 (7), pp 2678-2680.