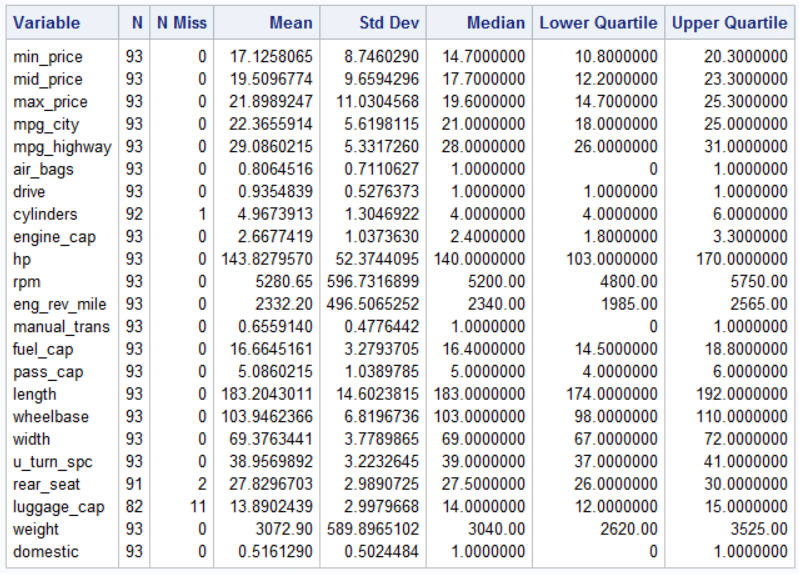
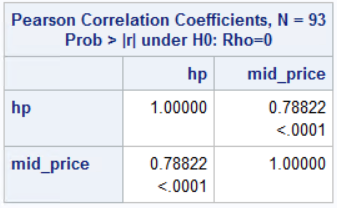
The 1993 car data provided as ‘93cars.dat’ is read into SAS. The data provides car and engine specifications of 93 cars and their price ranges (minimum, maximum and average).



The data has some missing values for the features – HP, rear-seat room, and luggage capacity. Relevant values are imputed for these missing data from relevant sources and applying some general knowledge about cars.

* number of cylinders for 1993 Mazda RX-7 1.3L twin-turbo charge Sports Edition = 4. source: <https://www.vehiclehistory.com/vehicle-engine-specifications/mazda/rx-7/1993>
* For sports cars with passenger capacity = 2, there is no rear seat and no boot-space
* For cars with passenger capacity > 6, there is an additional row of rear seat in-place of boot-space

On applying PROC CORR on the dataset for the two variables, we find that Horsepower and mid-range prices have a strong correlation (correlation coefficient = 0.79) and is significant (p-value < 0.001).



On applying PROC UNIVARIATE on mid-price to check for outliers and distribution of the variable. We find some outliers and they (prices greater than $38k) are removed from the assessment.

|  |  |
| --- | --- |
| C:\Users\Nimesh\AppData\Local\Microsoft\Windows\INetCache\Content.Word\cars_univariate1.png | C:\Users\Nimesh\AppData\Local\Microsoft\Windows\INetCache\Content.Word\cars_univariate2.png |

The cut-set of the data is now assessed for linear regression model. On applying PROC REG for model:

mid\_price = B0 + B1\*mpg\_city + B2\*air\_bags + B3\*manual\_trans + B4\*hp + B5\*domestic

Following are the results:

|  |  |
| --- | --- |
| C:\Users\Nimesh\AppData\Local\Microsoft\Windows\INetCache\Content.Word\cars_proc_reg1.png | C:\Users\Nimesh\AppData\Local\Microsoft\Windows\INetCache\Content.Word\cars_proc_reg1_2.png |

R2 =0.73 implies that all the 5 explanatory variables together explain 73% of the variance in the dependent variable mid\_price.

R2 lies in a range between 0 and 1, and with values closer to 1.00 indicating a very good model fit. R2 should be high for experimental data as most of the variables are under controlled environment. For industrial data (noisy data), even low r-square is acceptable, and there is always chance to improve. For the given data, R2 =0.73 is acceptable, but there is scope for betterment.

R2 of a model describes that how much of the variance in the dependent variable can be explained by the independent variables. This value increases as we keep on adding more explanatory variables. Some of them may not even be significant to the model. The R2 value can reach 1 by involving all the variables, but then there will be no variance in the residuals i.e. we will be fitting the model to the data. The model will give high accuracy to the dataset that was used to train the model, but will fail when we want to predict at some point other than of the dataset. To overcome this, a concept of adjusted R2 is formalized which imposes a penalty on any variable added to the model that has a very small explanatory power. Thus, on addition of more variables to the model, Adjusted R2 could go down and is therefore a more accurate measure of model fit. If the difference between R2 and adjusted R2 is large (i.e., more than 5% points), adjusted R2 is the more accurate indicator of model fit.

All coefficients that have a t-value greater than 1.96 (or a p-value less than 0.05) are statistically significant at the 95% confidence level. That implies that the corresponding variable has a non-zero impact on the dependent variable.

In the current model:

* mpg\_city (-0.372, |t-value|= 3.27) is significant at 0.01 significance level
* air\_bags (3.014, t-value= 4.28) is significant at <0.0001 significance level
* manual\_trans (|t-value|=0.57) is insignificant even at 0.1 significance level
* hp (0.081, t-value= 6.07) is significant at <0.0001 significance level
* domestic (-3.324, |t-value|= 3.3) is significant at <0.0001 significance level

From the model we can interpret that,

* For every increase of 10 HP of the engine, the mid-range price of the car increases by $810, controlling all other variables of the model.
* If the car is manufactured in U.S., then the mid-range price of the car will be $3,324 lesser than that of a car manufactured outside of U.S., controlling all other variables of the model.

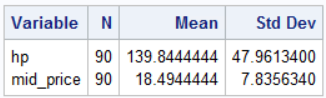
Of all the variables, HP is the most important variable as it affects the mid-price the most. We can say that by looking at the standardized estimates of the variables.

We cannot directly conclude which variable is more important, just by looking at the parameter estimates, because all the variables have different scale. Therefore, to compare the variables, all the variables are brought to same variance, with mean=0 and standard deviation =1. Higher the standardized coefficient, more important the variable is.

**Elasticity of midrange price with respect to HP (horsepower)**

PE = =  = β\*

where Δ indicates the change in either price (q) or hp (p). To compute average estimate of price elasticity, multiply the price coefficient with the average price and divide by the average HP.



Therefore,

*PE* = 0.613

Addition of non-linear term sq\_mpg\_city (mpg\_city\*mpg\_city), shows that it is significant in the model. R2 value has increased to 0.7573

|  |  |
| --- | --- |
| C:\Users\Nimesh\AppData\Local\Microsoft\Windows\INetCache\Content.Word\cars_proc_reg2.png | C:\Users\Nimesh\AppData\Local\Microsoft\Windows\INetCache\Content.Word\cars_proc_reg2_2.png |

Interaction between hp and city mpg (hp\_mpg\_city) should be included in the model. We can also add dummy-variables of car-type.

|  |  |
| --- | --- |
| C:\Users\Nimesh\AppData\Local\Microsoft\Windows\INetCache\Content.Word\cars_proc_reg3.png | C:\Users\Nimesh\AppData\Local\Microsoft\Windows\INetCache\Content.Word\cars_proc_reg3_2.png |

On addition of these additional variables and clean-up of some insignificant variables, the R2 value has increased to 0.7932