Question-1:

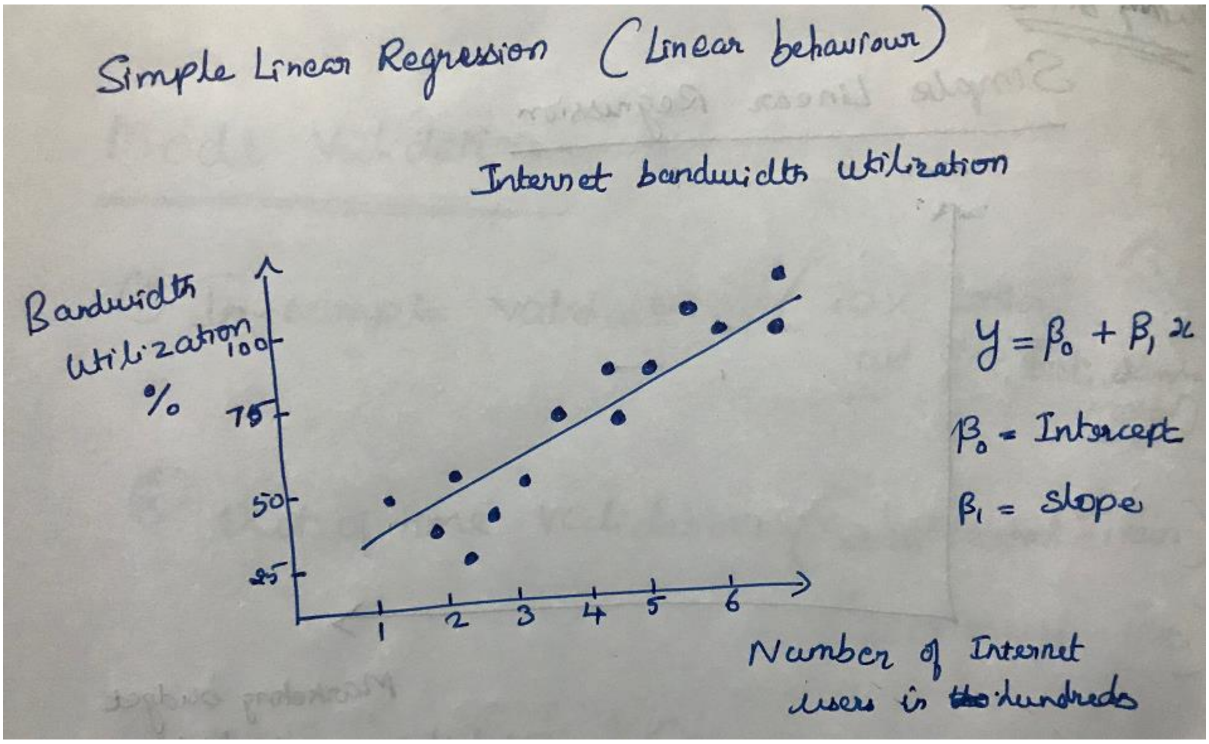
List down at least three main assumptions of linear regression and explain them in your own words. To explain an assumption, take an example or a specific use case to show why the assumption makes sense.

Answer:

Assumption 1: Linearity → Linear relationship of Independent and Dependent variable

The dependent/output variable is a variable predicted.

The independent variables should have a linear relationship with dependent/output variables.



Assumption 2: Multicollinearity → Variable correlation must be handled effectively.

Multicollinearity is actually

> Relationship between one variable and another variable
> (or)
> Relationship between one variable and 2 or more variable

As mentioned above, Multicollinearity can be identified as 2 scenarios.

> 2 variables having very high correlation.

Example: 2 variables representing height of people in centimetre and feet. You can eliminate 1 of the 2 in the regression model.

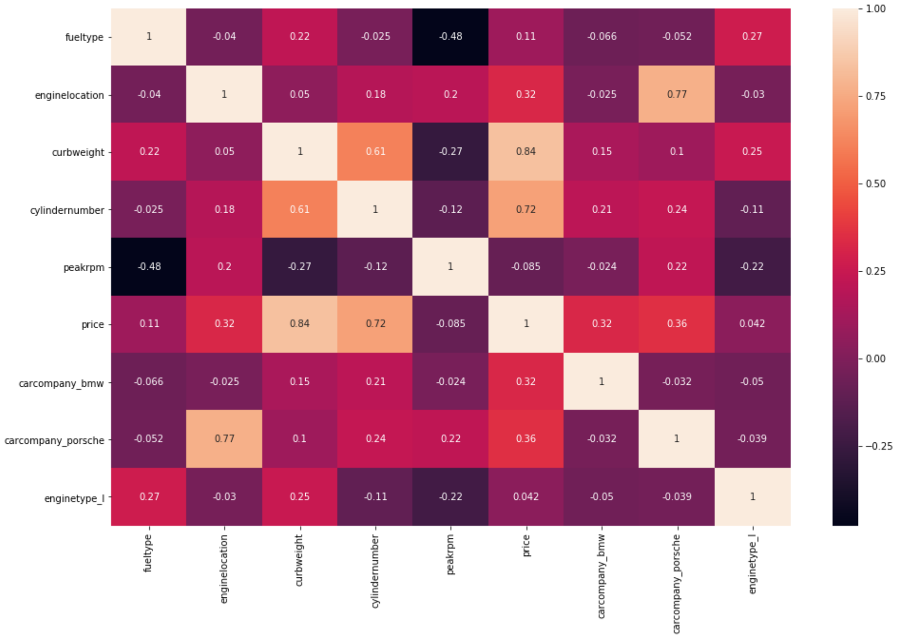
Height in Centimetre	Height in Feet
167	5.48
168	5.51
170	5.58
172	5.64
175	5.74
175	5.74

(or)

2 or more variables derives from another variable.
Example 3: Illiteracy rate calculated from Illiterate count and Total population.

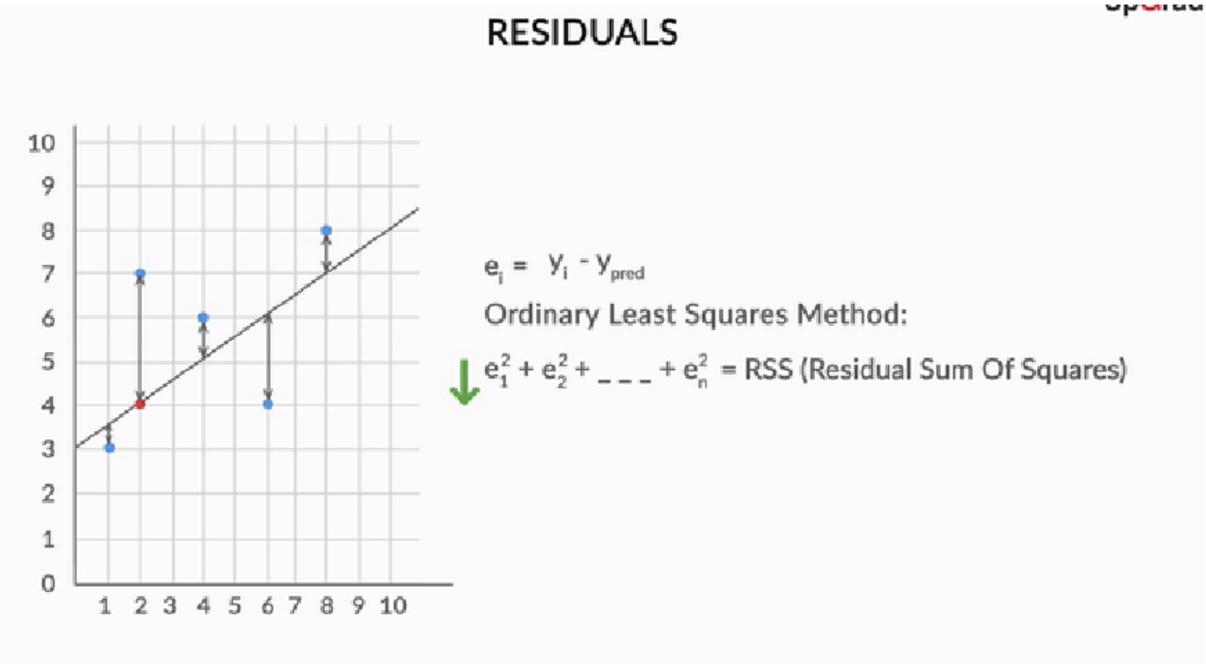
Year	Illiterate people	Total Population	Illiteracy Rate
2010	4632	109340	4.24
2011	5049	112771	4.48
2012	4438	114298	3.88
2013	4977	118679	4.19
2014	5163	122025	4.23
2015	4378	126353	3.46
2016	4804	128833	3.73
2017	4911	132207	3.71

Correlation can be found and eliminated using variable’s VIF value (>3.33) and correlation plot. Below is an example of correlation plot.



Assumption 3: Homoscedasticity → Best Fit line does have more variation in Residual value.

Best fit line of the linear regression should not have a wide range of variance in the residual Error value. This will deviate the linear regression model.



RESIDUALS

$e_i = y_i - y_{pred}$

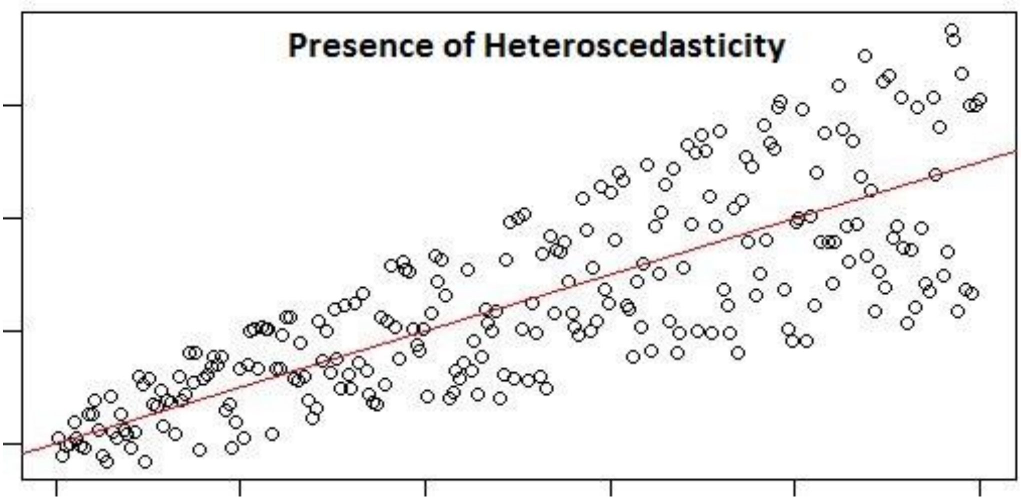
Ordinary Least Squares Method:

$e_1^2 + e_2^2 + \_\_\_ + e_n^2$ = RSS (Residual Sum Of Squares)

Error value (ei) should not have more variation.

Homoscedasticity is referring to the situation where error term values are almost same across all values of the variables.

Heteroscedasticity means unequal scatter, which means that the variance of residuals should not increase with the fitted values of the response variable. The presence of heteroscedasticity implies that you want your points to look like a funnel. This must be rectified by rebuilding models with additional variables.



****

Question-2:

By now you have seen multiple **model evaluation metrics** used for regression models, such as r-squared, adjusted r-squared, RMSE, the residual plot etc.

In this question, you are required to **explain at least three regression model evaluation metrics** in your own words.

1. For the final model that you have built, explain each evaluation metric with its intuition (i.e. what and how it measures) and relate the intuition to its mathematical formula. You may use figures or examples to explain if needed. Limit your answer to 1000 words for this part.

Answer:

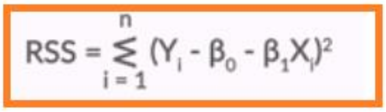
From the final model derived, regression model evaluation metrics were identified as below:

R-squared:	0.888
Adj. R-squared:	0.881
RMSE:	0.086326264

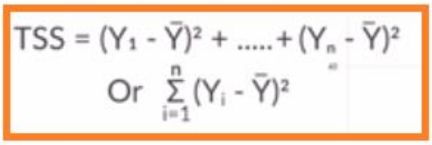
R-squared value is 0.888.

Since the ratio between RSS (Residual sum of square) and TSS (Total sum of square) is

(1 - .888) = .112. It clearly states that error distribution is very less and Linear regression has achieved good best of line.

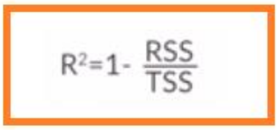


$$RSS = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$



$$TSS = (Y_1 - \bar{Y})^2 + ...+ (Y_n - \bar{Y})^2$$

$$\text{Or } \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$



$$R^2 = 1 - \frac{RSS}{TSS}$$

$$adjusted - r - squared = 1 - \frac{(1-R^2)(n-1)}{n-k-1}$$

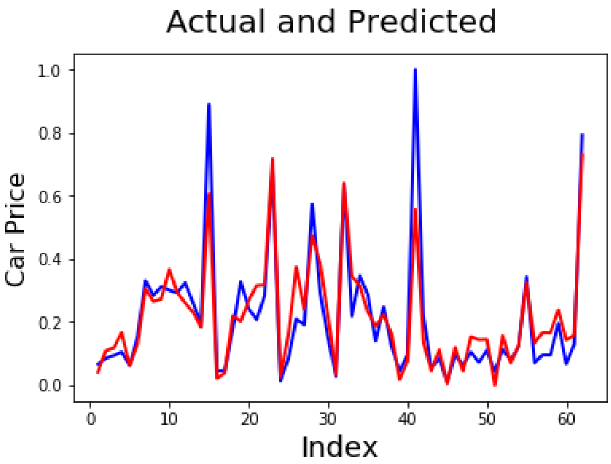
n → is the total number of data points and

k → is the number of predictors.

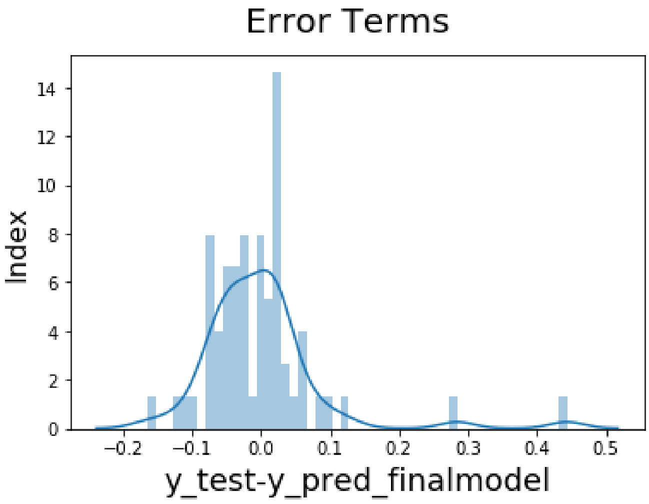
Adjusted R-squared value from the significant variables is 0.881 which is almost equal to R-Squared value 0.888. so, the model is good.

RMSE (Root Mean Squared Error) is 0.086 which is very good.

Model Evaluation – Actual Vs Predicted



Error Distribution



2. Compare the advantages and disadvantages of any three-evaluation metrics. If you do not think there's any advantage or disadvantage of a certain metric, mention that. Limit your answer to 1000 words for this part.

Answer:

1) R-Squared 2) Adjusted R-Squared 3) RMSE 4) MSE

Adjusted R-Squared gives the best result than R-Squared because,

In case of multivariate analysis, the R-Squared is calculated from all the variable in the model.

Whereas the Adjusted R-Squared is calculated from the variable which are significant in the model.

So, in Multivariate – Linear regression analysis Adjusted R-Squared value has very high priority.

RMSE gives the best result than MSE because,

RMSE (root-mean-square error) serves to aggregate the magnitudes of the errors in predictions into a single measure of predictive power.

RMSE is the square root of the (MSE) mean squared error, thus RMSE is very sensitive to outlier's error than MSE.