

## Assignment Part-II

### 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-1:

Regression is a technique to find the relationship between independent variable and dependent variable, Regression is a Parametric machine learning algorithm which means an algorithm can be described and summarized as a learning function.

Example-  $Y = f(x)$

Linear Regression also explains how change in dependent variable varies with a unit change of independent variable. In Simple Linear regression we can change one variable at a time where else in Multiple Linear regression we can change multiple variables at the same time.

$$f(X) = Y = B_0 + B_1 * X_1 + B_2 * X_2 + B_3 * X_3 \text{ (Linear Regression)}$$

The various assumptions that are made for linear regression are as follows:

1. Linearity
2. Outliers
3. Autocorrelation
4. Multicollinearity
5. Heteroskedasticity

**Linearity:** The relationship between independent and dependent variable must be linear in nature to perform Linear regression, if we build a linear model with a non-linear dataset, the model will fail to capture the linear trend mathematically also this will predict wrongly on unseen data.

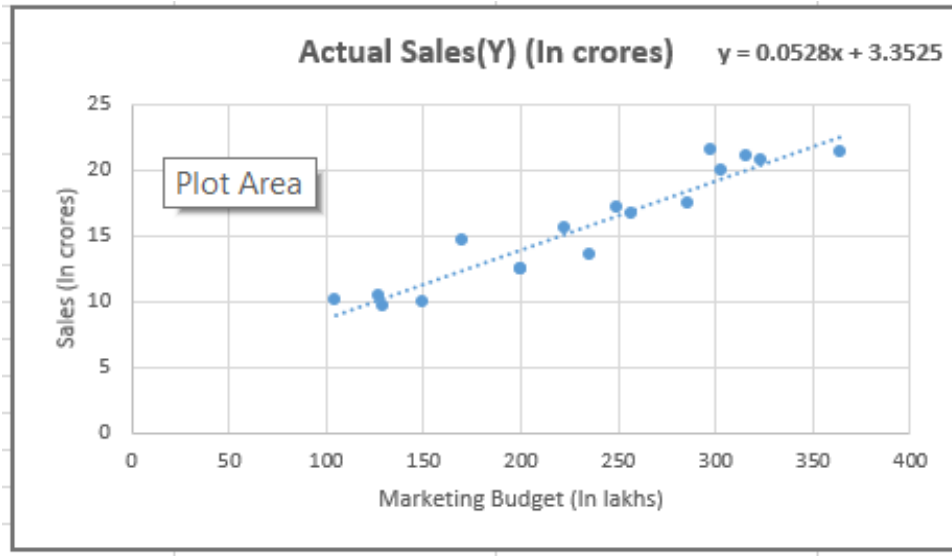
We can create the scatter plot of data to check the linearity and fit the best fit line which fits the given scatter plot in the best way, then we can calculate the residual.

To find the best fit line or regression line we have to minimize the cost function.

$$f(X) = \theta_0 + \theta_1 * X$$

$$\text{Residual} = |\text{actual value} - \text{predicted value}| \text{ OR}$$

$$E_i = Y_i - Y_{\text{pred}} \quad \text{where } Y_i = \text{actual value and } Y_{\text{pred}} : \text{Predicted value and } E_i: \text{error on data point}$$



The above scatter plot we can see the linearity of the data where we can fit the best fit line.

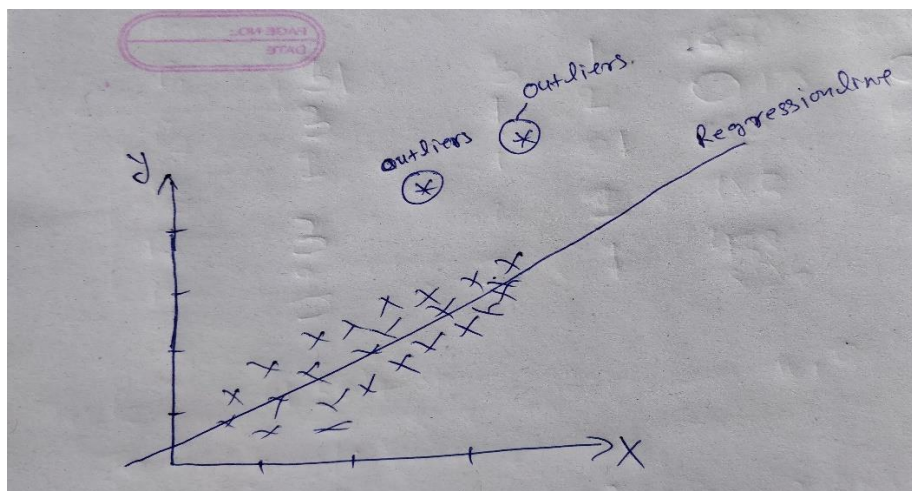
As we can see we start with scatter plot and then we found the best fit line, we also calculated the Residual on each point.

RSS will tell us how fit our data point on given regression line.

$$e_i = e_1^2 + e_2^2 + e_3^2 + \dots + e_n^2 = \text{RSS (Residual Sum of Squares)}$$

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

**Outliers:** A point that falls outside the data set is called outliers which means in statistics an outlier is a data point that significantly differs from the other data points in a sample.



As we can see clearly in above picture there is two data point significantly differ from the other data point.

This will also effect the model if there are good no of outliers best way to split the data set in two sample and create 2 different model one with outliers and other without outliers.

Outliers also affect the coefficient with changing the sign, Outliers increase the Residual which increase the error.

IQR(Inter quartile range) and Box Plot is the technique to identify the outliers.

$IQR = Q3 - Q1$  (where Q3 third quartile, Q1 first quartile)

Data set should be in between this range bellow formula

Outliers range =>  $(Q1 - 1.5 * IQR)$  to  $(Q3 + 1.5 * IQR)$

**Multicollinearity:** If there is correlation between predictor variable we can say there is a multicollinearity in between, which mean independent variable are correlated to each other, in the sample if we found the multicollinearity then it will lead the confusion of true relationship between dependent and independent variables.

When independent variable are correlated, the regression coefficient of a correlated variable depends on which other predictors are available in the model. If this happens, we'll end up with an incorrect conclusion

To check for multicollinearity, we can look for Variance Inflation Factor (VIF) values. A VIF value of 5 or less indicates no multicollinearity.

if  $VIF > 5$  which means it's highly correlated and we can drop that variable as they are correlated to other independent variable which can difficulty to estimate the true relation between dependent and independent variable.

---

## 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.

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.

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-2:

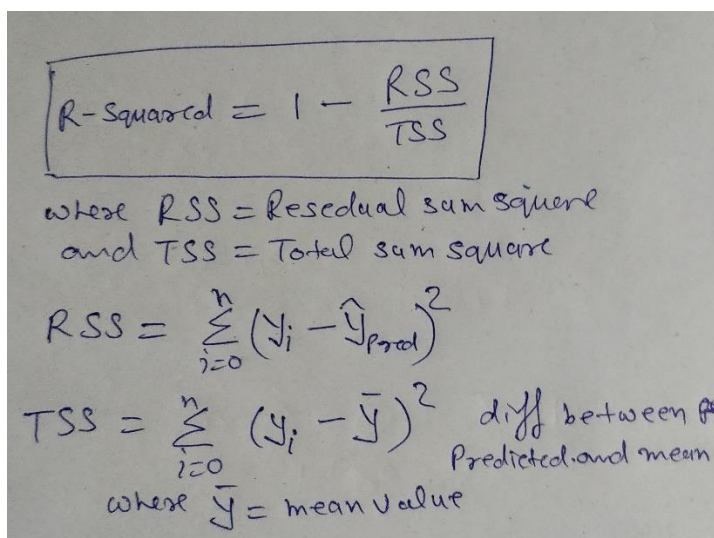
Regression model works on constructive evaluation principle, we build a model, evaluate from metrics, and then make improvements and continue until we achieve a desirable accuracy, evaluation metrics explain the performance of a model.

So Far we have seen following metric:

- 1- R-Squared
- 2- Adjusted R-Squared
- 3- RMSE
- 4- VIF
- 5- P-Value
- 6- Residual.

**R-Squared:** R-squared is evaluation metric through which we can measure how good the model is higher the R-square better the accuracy.

For example: Let say after evaluation we got R-squared = 0.81 which mean we are able to explain 81% of variance in data, also we can say the accuracy of a model is 81%.



Handwritten mathematical formulas and definitions:

$$R\text{-Squared} = 1 - \frac{RSS}{TSS}$$

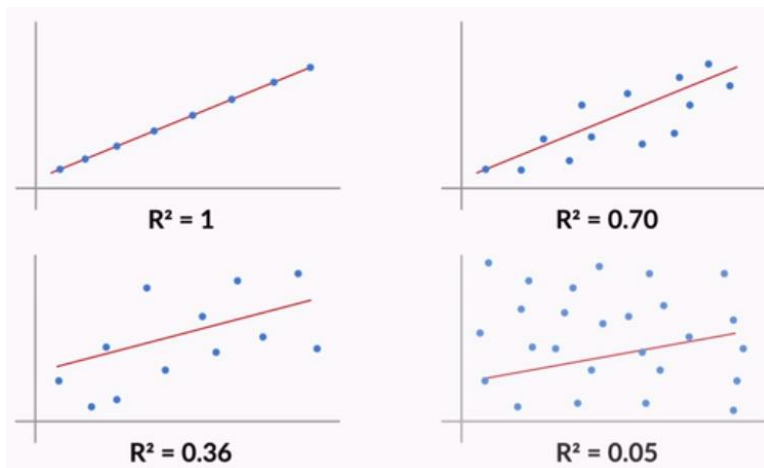
where  $RSS = \text{Residual sum square}$   
and  $TSS = \text{Total sum square}$

$$RSS = \sum_{i=0}^n (y_i - \hat{y}_{pred})^2$$
$$TSS = \sum_{i=0}^n (y_i - \bar{y})^2 \quad \text{diff between Predicted and mean}$$

where  $\bar{y} = \text{mean value}$

We can compute the RSS (Residual sum squared) with square sum of (actual – predicted)

In TSS (Total sum squared) we need to take squared sum of (predicted – mean value)



As we can see  $R^2 = 1$  which mean Residual is 0  $R\text{-Squared} = 1$

Best fit, as you can see as fit line getting poor the  $R\text{-Squared}$  value also get reduce and become close to zero  $R\text{-Squared}$  lies between 0 to 1

$0 < R\text{-Squared} \leq 1$

OLS Regression Results						
=====						
Dep. Variable:	price	R-squared:	0.839			
Model:	OLS	Adj. R-squared:	0.832			
Method:	Least Squares	F-statistic:	118.3			
Date:	Sun, 19 Aug 2018	Prob (F-statistic):	1.82e-51			
Time:	22:24:24	Log-Likelihood:	162.91			
No. Observations:	143	AIC:	-311.8			
Df Residuals:	136	BIC:	-291.1			
Df Model:	6					
Covariance Type:	nonrobust					
=====						
	coef	std err	t	P> t	[0.025	0.975]
-----						
const	0.0123	0.035	0.356	0.723	-0.056	0.081
enginesize	1.0919	0.047	23.197	0.000	0.999	1.185
stroke	-0.1666	0.059	-2.840	0.005	-0.283	-0.051
cylindernumber_two	0.2305	0.041	5.565	0.000	0.149	0.312
fuelsystem_idi	0.0752	0.024	3.104	0.002	0.027	0.123
car_company_bmw	0.1691	0.034	4.904	0.000	0.101	0.237
car_company_subaru	-0.0715	0.031	-2.295	0.023	-0.133	-0.010
=====						
Omnibus:	12.863	Durbin-Watson:	2.132			
Prob(Omnibus):	0.002	Jarque-Bera (JB):	28.627			
Skew:	0.295	Prob(JB):	6.08e-07			
Kurtosis:	5.111	Cond. No.	12.6			
=====						

**Model-7**

As you can see in final model we got R-Squared = 0.839

Which means we are able to explain 83% of variance in data, also we can say the accuracy of a model is 83%.

### Adjusted R-Squared :

Adjusted R-Squared Penalize the model which have too many feature or variable in that.

Let say if we have 2 model each with 3 features R-squared is 83.9% and adjusted-R-Squared 83.2%

Now if we add 3 more features in first model R-Squared will get increase but adjusted-R-Squared will penalize the model, and it will tell I am giving you a lower value as you have added new feature which could be cause a problem in modeling.

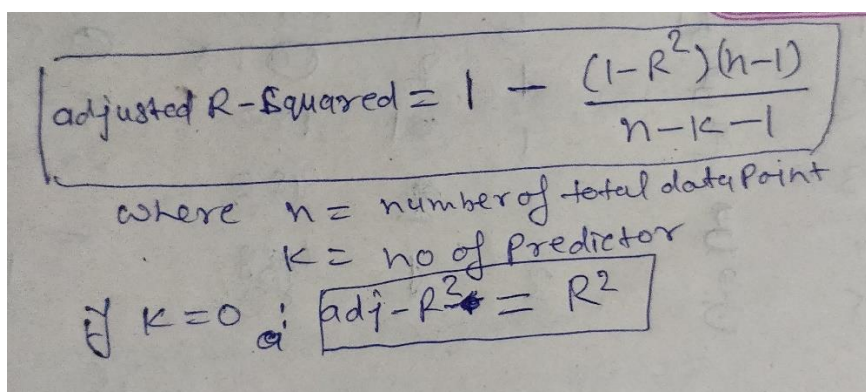
The adjusted R-squared increases only if the new term improves in the model, adjusted R-squared gives the percentage of variation explained by only those independent variables that in reality effect the dependent variable. Adjusted R-Squared also measure the goodness of model

As we can see in our final **model-7** R-Squared and adjusted R-squared is very close,

R-Squared = 0.839

Adjusted R-Squared = 0.832

and with this we assume none of the other variable need to add into the model as predictor.



Handwritten formula for Adjusted R-Squared:

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

where  $n$  = number of total data point  
 $k$  = no of Predictor

if  $k=0$  :  $\text{adj-R}^2 = R^2$

### P-Value and VIF:

P-value will tell us how significant the variable and VIF will tell multicollinearity between the independent variable if VIF > 5, means highly correlated

P-Value will tell the probability of NULL Hypothesis being accept which means probability of fail to reject the Null Hypothesis.

Higher the P value higher the probability of fail to reject a NULL Hypothesis.

Lower the P-Value higher probability of Null Hypothesis being rejected.

In Our Observation if we see above final model all variable have less the 0.05 P value which means they are highly significant variable.

	Var	Vif
1	stroke	3.90
0	enginesize	3.86
3	fuelsystem_idi	1.16
4	car_company_bmw	1.10
2	cylindernumber_two	1.05
5	car_company_subaru	1.03

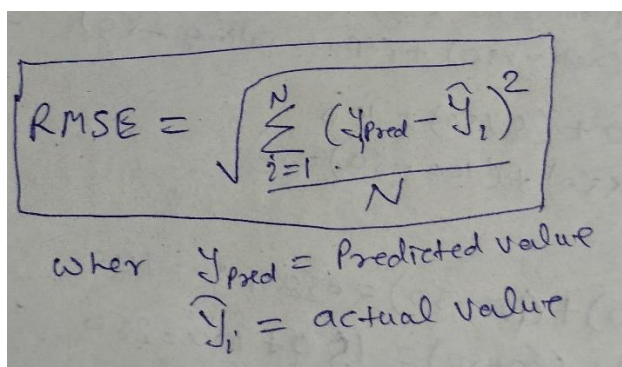
VIF value is less then 5 in our final model we can say there no correlation between independent variable.

### RMSE : (Root mean Square)

RMSE is a frequently used matric to measure of the differences between sample values predicted by a model and the values observed.

RMSE is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are.

It will evaluate the sum of mean squared error over the number of total sample observation.


$$RMSE = \sqrt{\frac{\sum_{i=1}^N (y_{pred} - \hat{y}_i)^2}{N}}$$

where  $y_{pred}$  = Predicted value  
 $\hat{y}_i$  = actual value

In Final **model-7** we calculated the RMSE: 0.08586514469292264 lower the RMSE better the model performance.

```
# Now let's check the Root Mean Square Error of our model.  
print('RMSE :', np.sqrt(metrics.mean_squared_error(y_test_m7, y_pred)))  
  
RMSE : 0.08586514469292264
```

**Durbin-Watson:** Durbin-Watson test evaluate the Autocorrelation it should lie between 0 to 4

In our above observation in final model we got Durbin-Watson: 2.132

### Summary:

There are some disadvantage if we not consider that our evaluation metrics, which will increase the in performance measurement. For example

All metrics like:

- 1- R-Squared
- 2- Adjusted R-Squared
- 3- RMSE
- 4- VIF
- 5- P-Value
- 6- Residual.
- 7- Durbin-Watson

Will tell us how good model is (best fit model) each metrics is important for model evaluation as I describe above.

In Linear Regression there are some assumption like:

1. Linearity
2. Outliers
3. Autocorrelation
4. Multicollinearity
5. Heteroskedasticity

which we need to check while measuring the matrix

1. RMASE has disadvantage RMSE will affected by the outliers. If we have outliers in our data sample



2. if sample population is not linear RMSE will affect as there would be no trend and which means standard deviation will vary.

3. Residual also affect by the outliers which will increase residual squared sum.

**stroke, engine size, fuelsystem\_idi, car\_company\_bmw, cylindernumber\_two, car\_company\_subaru**

These are the driving factor as per the **Model-7** on which the pricing of cars depends.

**Model-7 Observation:**

1. I can see a good model with R-Square = 0.839
2. I can see good adjusted R-Squared = 0.832
3. As R-Squared and adjusted R-Squared is very close, we can assume none of the other variable need to add into the model as predictor.
4. Model have Durbin-Watson:2.132 which is between 0 and 4, we can assume there is no Autocorrelation
5. As we can see  $VIF < 5$  which means no Multicollinearity
6. P value is less the 0.05 which means they are highly significant variable.