By Hari Thapliyal

Linear Regression Subjective Question

Question 1: What are the assumptions of linear regression regarding residuals?

There are 3 kinds of assumptions in linear regression

a. Assumption related to model (1 assumption)
b. Assumptions related to residuals (4 assumptions)
c. Assumptions related to estimator (2 assumptions)

Assumptions related to residuals are as below.

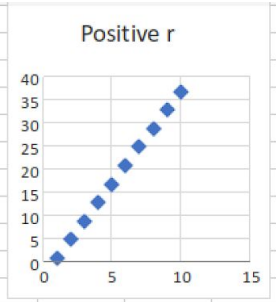
1. Error terms εi are normally distributed like residual (difference between y_predicted & y_actual). This is called **normality assumption**.
2. If you take the mean of residuals then its sum will be zero. It means error terms are normally distributed around zero. This is called **Zero means assumption**.
3. Error terms are independent. Features are not correlated to each other. The noise noticed on scatter plot does not have any pattern. it is random. Pairwise correlation between feature doesn't exist. This is called **independent error assumption**.
4. Error terms has constant variance i.e. Homoscedasticity. It means when X changes difference between y predicted and y actual does not change. It remains constant for all X.

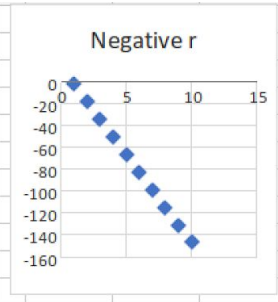
Question 2: What is the coefficient of correlation and the coefficient of determination?

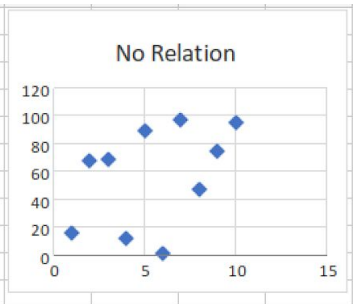
Coefficient of Correlation demonstrates the relationship between dependent variable (DV) (y) and independent variable (IV) X. In statistics it is represented by r. Using r you can tell whether two variables are related to each other or not, if yes then how strong is this relationship and what kind of relationship i.e. negative or positive.

- The value for r varies from between -1 to 1.
- If the value is 0 then it means there is no relationship between X and y. On scatter plot it will slow like random data spread across the area.
- If the value of r is negative sign then it means X & y shares reverse relationship, i.e. if X increases then y decreases and when X decreases then y increases. On scatter plot it will show a line with a negative slope.
- If the value of r is positive sign then it means X & y share positive relationship, i.e. if X increases y increases and when X decreases y decreases. On scatter plot it will slow a line with a positive slope.
- Any absolute value of r which is more than .8 means X & y has very strong relationship
- Any absolute value of r which is like .2 or less shows weak relationship between X & y.
- R shows relationship and NOT the causation.
- Formula for Sample's correlation is r(Xy) = S(Xy) / [S(X) * S(y)]. Where S is the standard deviation of the sample. X is IV and y is DV
- For linear regression we should use pearson correlation (R)
- For non-linear regression we should use spearman correlation (R). Otherwise results will be unreliable.

x	Positive r	Negative r	No Relation
1	1	-1	17
2	5	-17	69
3	9	-33	70
4	13	-49	13
5	17	-65	90
6	21	-81	2
7	25	-97	98
8	29	-113	48
9	33	-129	76
10	37	-145	96







Coefficient of Determination is output of regression. It is shown as R^2. It means this is the square of **Coefficient of Correlation (r) between X & y.** But unlike r, in the case of R square X can have multiple variables. It means multiple variables can predict the value of y. That is why it is also called "multiple"-correlation coefficient. Thus r indicates the strength of relationship between X (multiple predictor) and y variable.

- R^2, is always positive and its value varies between 0 to 1
- 0 means y cannot be predicted using X
- 1 means y can be predicted using X without error
- Value between 0 to 1 means y can be predicted using X with some error
- Higher R^2, value like more than .8 shows that X & y are strongly correlated
- R^2, .8 means 80% variation of y (DV) can be explained using X (IV)

- What can be predicted using X is called y^ (y hat) and actual value of y is called y. Coefficient of determination shows square of correlation between y^ and y
- $R^2 = \{ (1 / N) * \Sigma [(xi - \bar{x}) * (yi - \bar{y})] / (\sigma x * \sigma y) \}^2$. Here N is the number of observations used to fit the regression model, Σ is the symbol of summation, xi is the x value for i^{th} observation, $\bar{x}$ is the mean x value, yi is the y value for i^{th} observation, $\bar{y}$ is the mean y value, σx is the standard deviation of x, and σy is the standard deviation of y.

Question 3: Explain the Anscombe's quartet in detail.

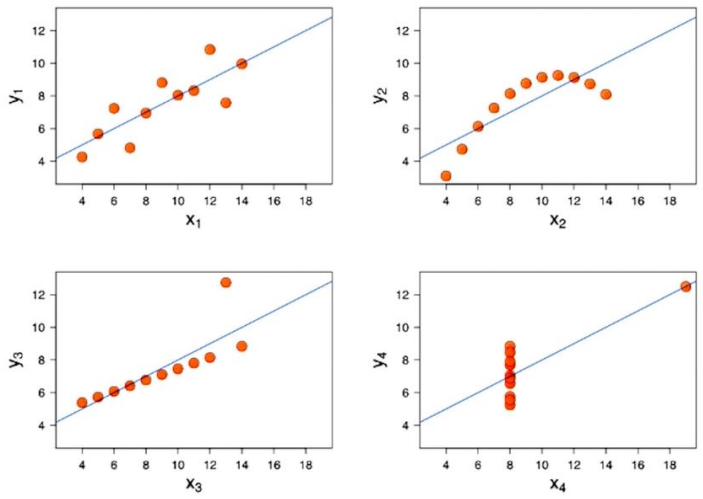
Regression analysis is a fantastic tool for creating model using which we can predict. But the biggest short coming of this tools are as following

1. It relies on those 3 kinds of assumptions and total 7 assumptions. Assumptions related to residual we discussed earlier. If these assumptions fail then model will fail.
2. It get influenced by outliers. So if outliers are there in the data. Model will fail to work as expected.
3. Linear regression model works only on linear relationship. If the relationship is not linear then model will fail.

Now without looking into all the data if we make the model we are bound to make a doom model. And sometimes data is so much it is not easy to look into the data. Therefore it is always a good idea to visualize our data in terms of X & y relationship. We can use python libraries like seaborn, matplotlib or tools like Tableau, PowerBI, Excel etc to visualize this. Now, during visualization if we observe that sometimes this relationship look pretty neat line, sometimes curve, sometimes random pattern on the chart. Then we cannot make model with this kind of data.

Anscombe's quartet helps us in visualizing the data and see the relationship between X & y. These 4 data set looks same without visualizing them.

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89



This graph is taken from https://youtu.be/Ftp3mmltV-k?t=53

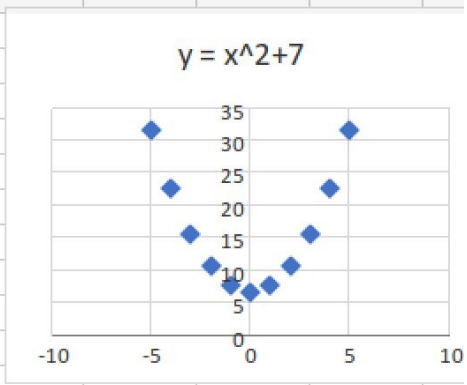
Question 4: What is Pearson's R?

R value is one of the very important measures in statistics. It is used to evaluate the relationship between the 2 variables, typically dependent (X) and dependee (y) or predictor (X) and predicted (y). If X & y shares linear relationship then y can be predicted using linear regression and the value of R will be between -1 to 1.

But if X & y share non-linear relationship then y cannot be predicted using linear regression but we need non linear regression. The value of R will be between -1 to 1

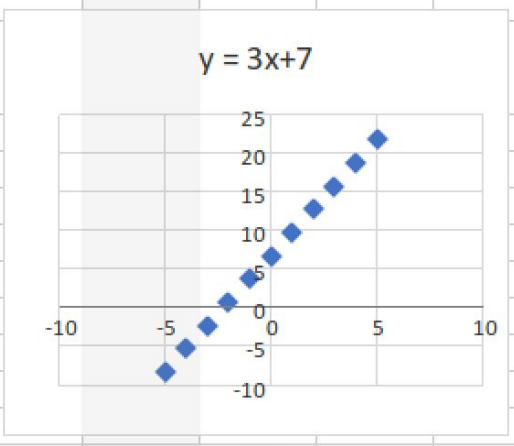
In case of **linear regression we use Perason's R** but his R is useless when the relationship is not linear. That time we use spearman's R or spearman's Rho calculator. It correctly predict the relation in the situation of non-linear relationship.

I	J	K	L	M	N
	Non Linear Relationship				
x	y = x^2+7	Rank X	Rank Y	d= Rank X - Rank Y	d^2
-5	32	11	1.5	9.5	90.25
-4	23	10	3.5	6.5	42.25
-3	16	9	5.5	3.5	12.25
-2	11	8	7.5	0.5	0.25
-1	8	7	9.5	-2.5	6.25
0	7	6	11	-5	25.00
1	8	5	9.5	-4.5	20.25
2	11	4	7.5	-3.5	12.25
3	16	3	5.5	-2.5	6.25
4	23	2	3.5	-1.5	2.25
5	32	1	1.5	-0.5	0.25
	Pearson Correlation			0	=CORREL(I3:I13,J3:J13)
	Spearman Correlation			0.711762	=CORREL(M3:M13,N3:N13)



We can clearly see pearson correlation in this above nonlinear regression case is zero. While spearman correlationship show a strong relationship between x & y

Linear Relationship			
x	y = 3x+7		
-5	-8		
-4	-5		
-3	-2		
-2	1		
-1	4		
0	7		
1	10		
2	13		
3	16		
4	19	Pearson Correlation	
5	22	1	



Question 5: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Features/columns/fields in the dataset can have various unit of measurement and various scales of measurement. For example quantity is measured in units but sales in Rupees. Sales can be in thousand but profit can be in hundreds. Because of this reason min-max of every feature is completely different. In this situation if we make a model using this kind of dataset then two major problems occurs. One, the significance of coefficient of lower scale variable e.g. quantity, looks abnormally different compared to high scale variable e.g. Sales. Second problem is related to the interpretation of the coefficient. To address this issue scaling is performed.

The process of scaling is performed only on numerical values and ordinal values. We generally do not perform this on dummy variables and features which has binary information like gender, success, exit etc.

In the process of scaling values of each column is compressed in such a way that it can be compared with other features. For example if quantity columns has min=10 and max=100 and sales columns has min=10,000 and max=99870 then both columns will be compressed in such a way that either both columns has values in 0 to 1 range or 1 to 3 or 4 sigma (depending upon outliers)

There are two methods of solving this different scale problem.

One, min-max scale or normalized scale: Using this technique every value of the columns is compressed between 0 and 1. The challenges of this approach is even outlier get compressed in this as 1 or around 1 for highest values & 0 or around zero for lowest values.

Formula for min-max scale is

X norm = (X - Xmin) / (X max - X min)

Second, Standardize scale: Using this technique mean & standard deviation of each column is computed and then for each value we compute the deviation from the mean. Values in the new column after this scale is applied can be between -3 to +3 (depending on outlier, can be some other also). But the mean of values of the column after applying this scale becomes zero (because sum is zero) and SD of the value becomes 1. So, after standard scaling, there is column with zero mean and 1 SD.

Formula for standardize scale is

z= x-u/sigma

Question 6: You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Answer

VIF = 1 / (1 - R^2)

VIF can become infinite only if 1-R^2 is 0

And this is possible only if R^2 is 1

So if any feature showing R^2 equal to 1, it means this is completely redundant feature, another feature in the dataset represent this. It is 100% reflection of other features and it does not have its own independence.

Very high VIF running in thousands or millions means R^2 of that feature is very high like 99.9999999

This is exactly call multicollinearity problem. As per the linear regression assumption if this kind of features is part of the model then it becomes extremely difficult to interpret the model coefficients. It is one of the failures of linear regression assumption.