Bike Sharing Assignment

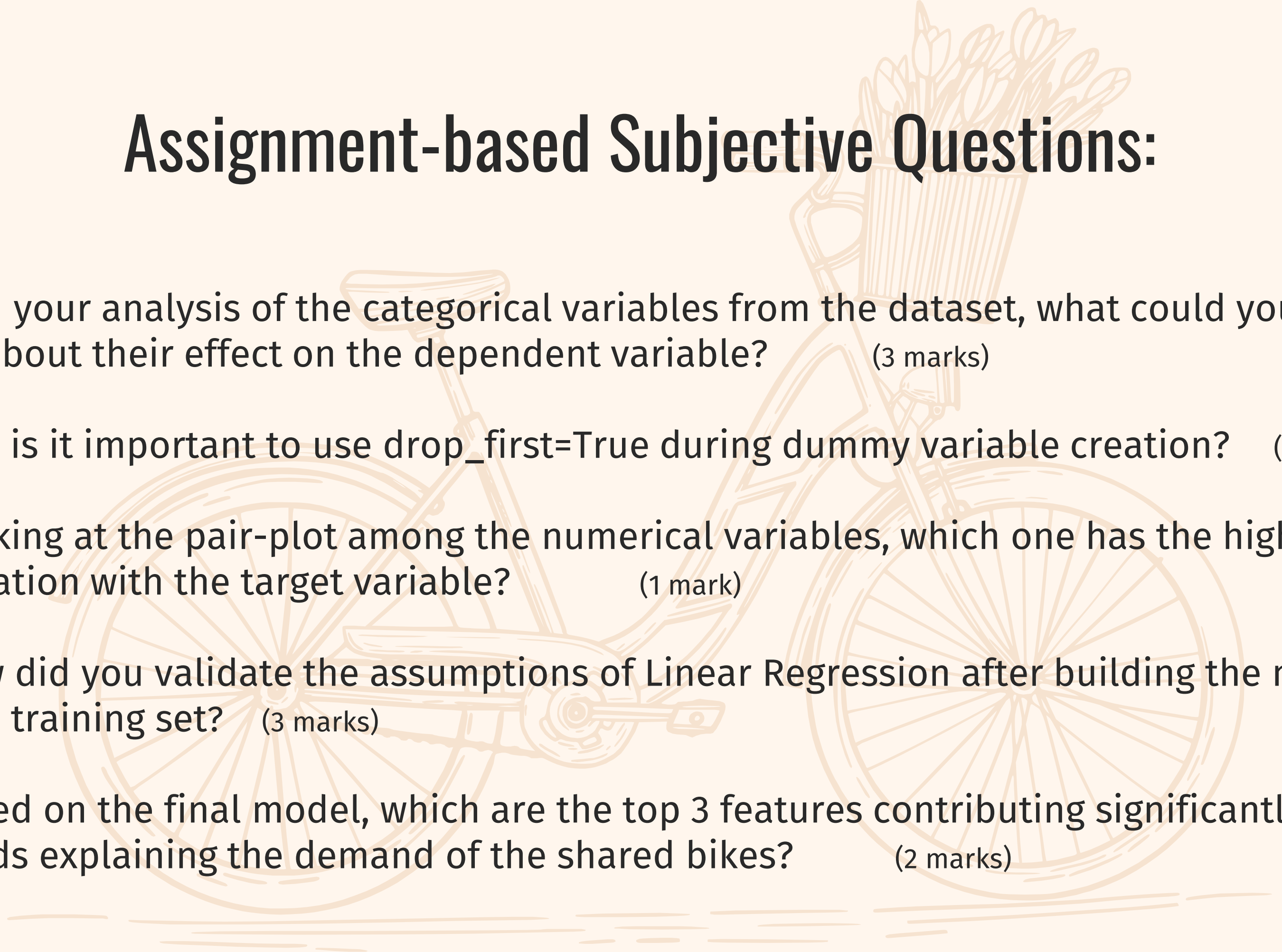
Linear Regression model based on Multiple Linear Regression

-By Vivek Kumar

Contents

- *Assignment-based Subjective Questions*
- *General Subjective Questions*

Assignment-based Subjective Questions:



1.From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

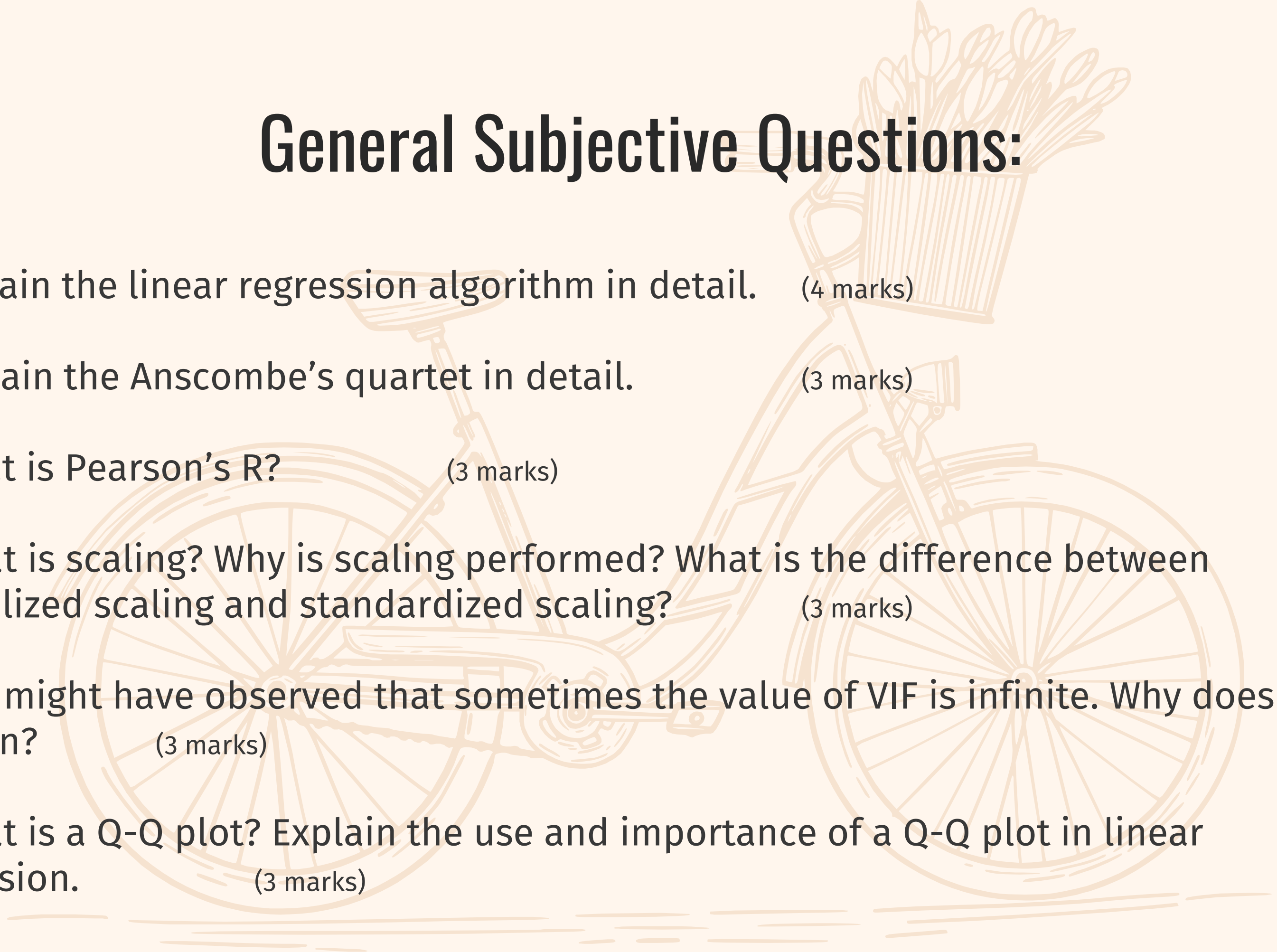
2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

General Subjective Questions:



1. Explain the linear regression algorithm in detail. (4 marks)

2. Explain the Anscombe’s quartet in detail. (3 marks)

3. What is Pearson’s R? (3 marks)

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

01

Assignment-based Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

1. Fall has the highest median, which is to be anticipated given that the weather is best for biking, followed by summer.

2. Median bike rentals are rising year over year, with 2019 having a higher median than 2018, which may be related to the growing popularity of bike rentals and people's increased environmental consciousness.

3. Because autumn months have a larger median, the overall spread in the month plot reflects the season plot.

4. Non-holiday rentals are higher than holiday rentals, which suggests that people prefer to utilise their own vehicles and spend time with their families on non-holiday days.

5. The overall median for all days is the same, but the difference between Saturday and WednesdayIt may be obvious that folks who have plans on Saturday may decide against renting bikes because it is a holiday.

6. The median of working and non-working days is almost equal, but the spreadlarger on days off from work since individuals may already have plans and not want to rentmotorcycles as a result.

7. The best conditions for hiring bikes are clear skies, moderate temperatures, and low humidity.temperature is lower and less.

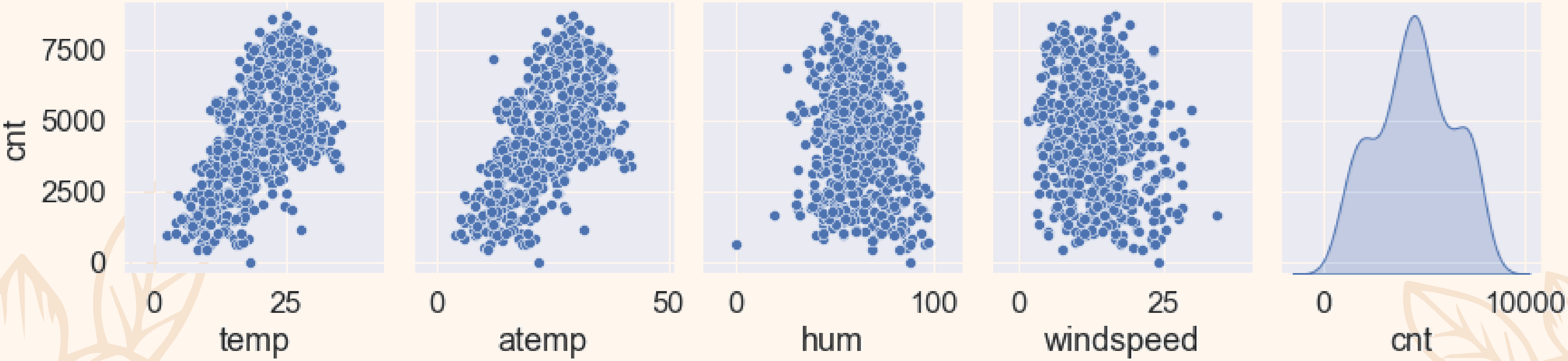
Q2. Why is it important to use drop_first=True during dummy variable creation?

Dummy variables with the value n-1 can represent a variable with n levels.
So, even without the first column, we can still express the data. If the first variable has a value of 1, then the variables from 2 to n must also have a value of 0.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

We have taken only the last row of pair plot because it answers the above question appropriately.

We can easily observe that the temp and atemp have very similar relation with cnt variable as there is a linear relation forming within those features.



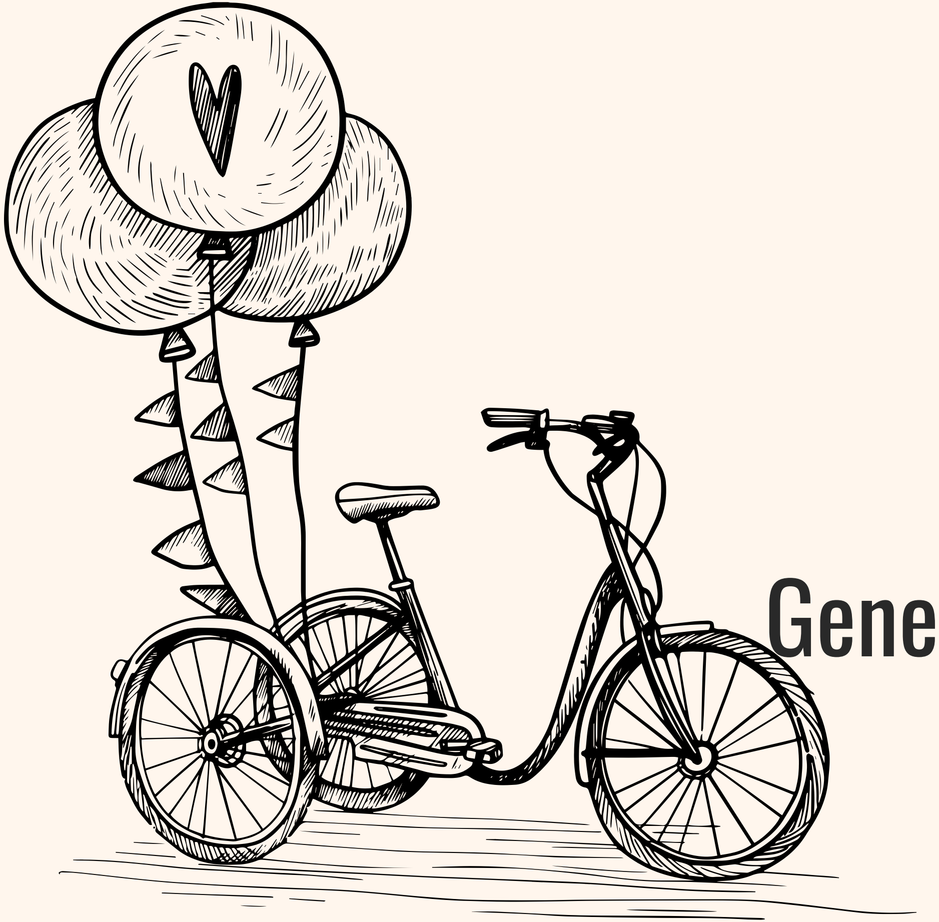
Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

1. Residual errors follow normal distribution
2. Maintains linear relation between dependant variable (y_test and y_predicted)

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

The top three features have been highlighted in the below table from our last model: Model 8

Sl.No.	Feature	coef	std err	t	P>|t|	[0.025	0.975]
1.	const	0.0417	0.018	2.337	0.020	0.007	0.077
2.	temp	0.5428	0.021	25.540	0.000	0.501	0.585
3.	windspeed	-0.1756	0.027	-6.559	0.000	-0.228	-0.123
4.	mnth_Sept	0.0978	0.017	5.761	0.000	0.064	0.131
5.	season_summer	0.0903	0.011	8.152	0.000	0.069	0.112
6.	season_winter	0.1205	0.011	10.863	0.000	0.099	0.142
7.	yr_2019	0.2366	0.009	26.806	0.000	0.219	0.254
8.	holiday_Yes	-0.0928	0.028	-3.312	0.001	-0.148	-0.038
9.	weathersit_Good/Clear	0.0939	0.009	10.226	0.000	0.076	0.112



02

General Subjective Questions

Q1. Explain the linear regression algorithm in detail?

A machine learning technique called linear regression is based on the supervised learning paradigm. It determines the connection between independent (Target) and dependent (Predictor) variables that best fits the provided data. In order to determine the best linear connection between the independent and dependent variables, it constructs the best straight-line fitting to the available data. The Sum of Squared Residuals Method is primarily used.

There are two forms of linear regression:

i. **Simple Linear Regression:** It uses a straight line to depict the connection between a dependent variable and just one independent variable. On the scatter plot of these two points, the straight line is drawn.

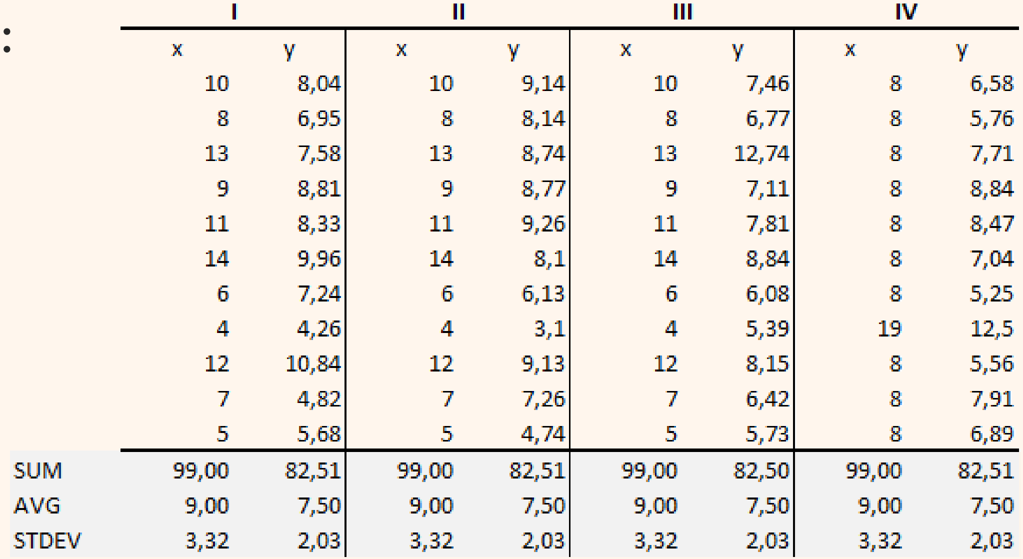
Formula for the Simple Linear Regression: $Y=\beta 0+\beta 1X1 +\epsilon$

- **ii. Multiple Linear Regression:** It shows the relationship between one dependent variable and several independent variables. The objective of multiple regression is to find a linear equation that can best determine the value of dependent variable Y for different values independent variables in X. It fits a ‘hyperplane’ instead of a straight line.

Formula for the Multiple Linear Regression: $Y=\beta0+\beta1X1+\beta2X2+\ldots+\beta pXp+\epsilon$

Q2. Explain the Anscombe's quartet in detail.

Francis Anscombe, a statistician, created Anscombe's Quartet. This technique maintains four datasets, each with eleven (x, y) pairings. The fact that both datasets share the same descriptive statistics is crucial to keep in mind. Regardless of the fact that their summary statistics are comparable, each graph has a unique narrative to tell. The statistics of the 4 datasets are briefly summarised below:



	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

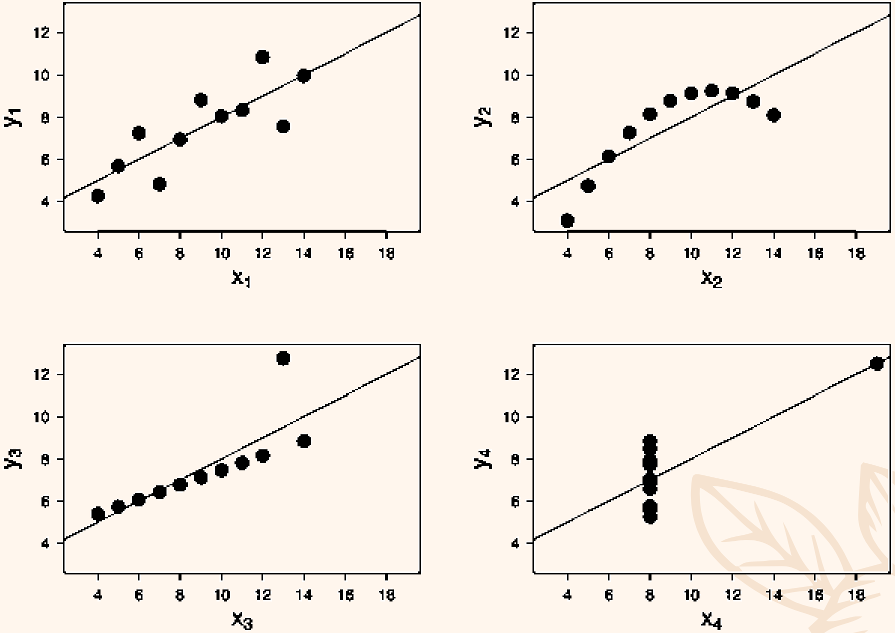
According to the summary statistics, x and y's means and variances were the same for all groups for both x and y:

- For each dataset, the average y is 7.50 and the average x is 9.
- Similarly, each dataset's x and y variances are 11 and 4.13, respectively.

For each dataset, the correlation coefficient (a measure of the strength of a link between two variables) between x and y is 0.816.

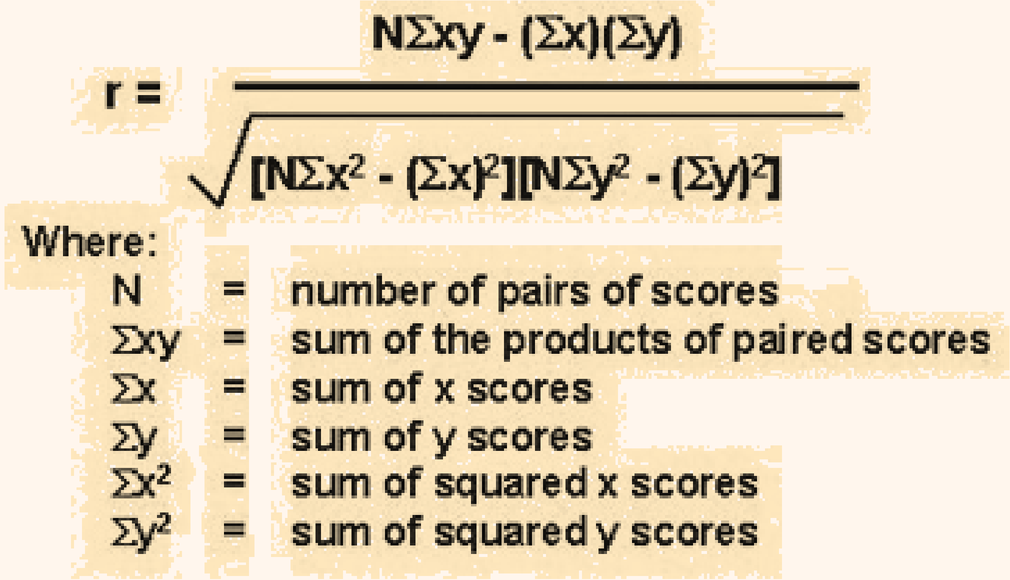
These four datasets display the identical regression lines when we plot them on an x/y coordinate plane, but each dataset tells a different narrative:

- Dataset I appears to have clean and well-fitting linear models.
- Dataset II is not distributed normally.
- In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.
- Dataset IV shows that one outlier is enough to produce a high correlation coefficient.



Q3. What is Pearson’s R?

Karl Pearson created the correlation coefficient known as Pearson's R, which is represented by the letter "r" and measures the strength of a linear link between two variables. It ranges from +1 to -1, with 1 denoting total linear positive correlation, 0 denoting no linear correlation, and -1 denoting entire linear negative correlation.



$$r = \frac{N\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2 - (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$

Where:

N	=	number of pairs of scores
Σxy	=	sum of the products of paired scores
Σx	=	sum of x scores
Σy	=	sum of y scores
Σx^2	=	sum of squared x scores
Σy^2	=	sum of squared y scores

- Statistically significant relationship between age and height.
- Relationship between temperature and ice cream sales.
- Relationship among job satisfaction, productivity, and income.
- Which two variables have the strongest co-relation between age, height, weight, size of family and family income.

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

The method of scaling involves normalising the data within a specific range. Our dataset frequently shows that distinct ranges for numerous variables. Scaling is therefore necessary to fit them all inside a single range.Normalization and Standardization are the two scaling techniques that are most frequently addressed. The values are generally scaled into a [0,1] range after normalisation. Data that has been standardised normally has a mean of 0 and a standard deviation of 1. (unit variance).

Formula of Normalized scaling:

$$x = \frac{x - min(x)}{max(x) - min(x)}$$

Formula of Standardized scaling:

$$x = \frac{x - mean(x)}{sd(x)}$$

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

The value of VIF is calculated by the below formula: $VIF_i = \frac{1}{1-R_i^2}$

- Where, ‘i’ refers to the ith variable.

- If R-squared value is equal to 1 then the denominator of the above formula become 0 and the overall value become infinite. It denotes perfect correlation in variables.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Q-Q plot is used in linear regression to determine if the points are roughly on the line. If they don't, it signifies that neither our errors nor our residuals are Gaussian (Normal).

Relevance of the Q-Q plot The points are as follows:

I. Equal sample sizes are not required.

II. II. Multiple distributional features can be examined at once. For instance, variations in position, scale, symmetry, and the existence of outliers.

III. III. Compared to analytical techniques, the q-q plot can shed more light on the nature of the difference.

Thank You!

- From: Vivek Kumar