Subjective Questions Answers

- **K Srinivas (DSC40 batch)**

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?

Ans – a) 'Weekday' Variable - We see that weekday value = 0, i.e., Sunday has highest number of rentals, followed by Saturday.

b) 'holiday' Variable – We observe that, whenever there is a holiday rentals are higher, which seems obvious as customer must rent bikes along with their family/friends and go for biking.

c) 'weathersit' variable- Here, we see that rentals are tend to be higher if the situation of the weather is 1, i.e., Clear.

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

Ans - We generally, drop first dummy variable because-

1. To increase efficiency and drop redundant category, for e.g., if a variable has 3 categorical levels.

Category_1	Category_2	Category_3
0	1	0
0	0	1
1	0	0

As we can see, for these encoding two dummy variables will be enough as such:

- Category_1 – 00
- Category_2 – 01
- Category_3 – 10

Hence, if we have k categorical levels in a variable, we need k-1 dummy variables.

2. To avoid multi-collinearity.

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

Ans - We observe that 'temp' and 'atemp' has the highest correlation with the target variable 'cnt'.

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

Ans – We can validate the assumptions by –

1. By plotting the error terms and checking the **distribution is normal and the mean is around zero.**
2. By plotting a scatter plot of all the error terms and observing that there is no pattern forming. That is, **Error terms are independent of each other.**

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

Ans – Based on the final model, top 3 features are:

- 'atemp'
- 'yr'
- 'hum'

General Subjective Questions:

Q1) Explain the linear regression algorithm in detail.

Ans - Linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables.

- One Explanatory variable is called Simple Linear regression
- For more than one explanatory variable is called multiple Linear regression.

Formulation

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^{\mathsf{T}} \boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \ldots, n,$$

Where:

Y = target variable

βo = Constant

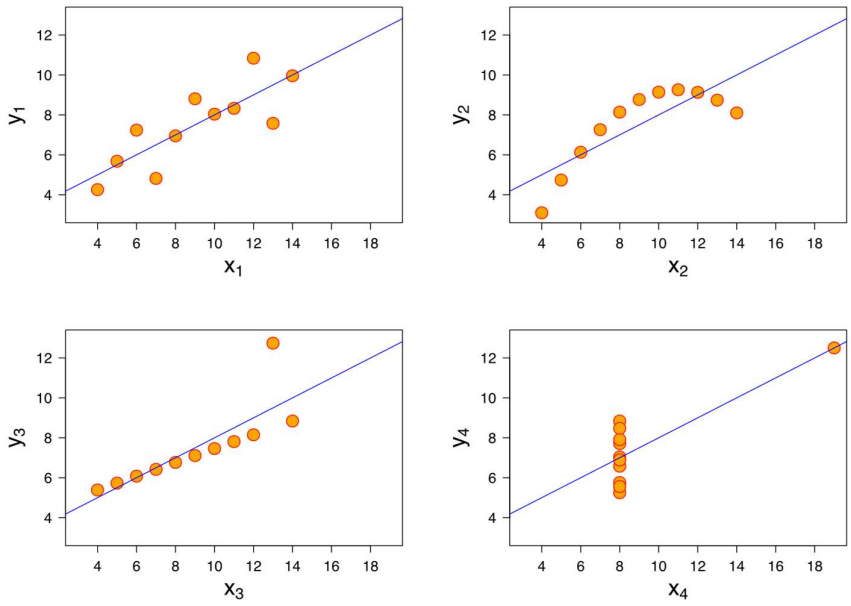
β1, β2…. = the explanatory variables.

Assumption we make for LR:

- Linear relation between X and Y
- Error terms are normally distributed
- Error terms are independent of each other
- Error terms have constant variance (Homoscedasticity)

Q2) - Explain the Anscombe's quartet in detail.

Ans) - **Intended** to counter the impression among statisticians that:
"Numerical calculations are exact, but graphs are rough."



Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed.

It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analysing and model building, and the effect of other observations on statistical properties. There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x, y points in all four datasets.

Description of each plot:
1. The points lie nearly on a straight line and are normally distributed.
2. Points lie nearly on a smooth curve, not a straight line.
3. The calculated regression is off-setted by one outlier.

4. One *high - leverage* (far off from the distribution) point produces a regression line even though rest of the distribution seems independent of y.

Q3) - What is Pearson's R?

Ans - In a Bivariate analysis, is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1.

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

where:

- cov is the covariance
- σ_X is the standard deviation of X
- σ_Y is the standard deviation of Y

Meaning of the Pearson's R:

1. A correlation coefficient of 1 means that for every positive increase in one variable, there is a positive increase of a fixed proportion in the other. For example, shoe sizes go up in (almost) perfect correlation with foot length.
2. A correlation coefficient of -1 means that for every positive increase in one variable, there is a negative decrease of a fixed proportion in the other. For example, the amount of gas in a tank decrease in (almost) perfect correlation with speed.
3. Zero means that for every increase, there isn't a positive or negative increase. The two just aren't related.

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

S. No.	Normalized scaling	Standardized scaling
1.	Minimum and maximum value of the features are used.	Mean and standard variation are used.
2.	Scales values between [0,1] or [-1,1]	Not Bounded by range
3.	It is affected by outliers	Not affected by outliers
4.	It is useful if we don't know the distribution of the features	Useful when the features have Gaussian or Normal distribution

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

Ans – As we have the formula of

$$VIF_i = \frac{1}{1-R^2}$$

For VIF to be infinity, value of R should be equal to 1. Indicating that, variables have a perfect correlation between them.

To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

Ans - In statistics, a Q–Q (quantile-quantile) plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other.

If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x.
A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.