LINEAR REGRESSION SUBJECTIVE QUESTIONS

By: Rishika Bansal

Assignment-based Subjective Questions

Question 1: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

- The demad of bike is less in the month of spring when compared with other seasons
- The demand bike increased in the year 2019 when compared with year 2018.
- Month Jun to Sep is the period when bike demand is high. The Month Jan is the lowest demand month.
- Bike demand is less in holidays in comparison to not being holiday.
- The demand of bike is almost similar throughout the weekdays.
- There is no significant change in bike demand with workign day and non working day.
- The bike demand is high when weather is clear and Few clouds however demand is less in case of Lightsnow and light rainfall.
- We do not have any dat for Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog , so we can not derive any conclusion. May be the company is not operating on those days or there is no demand of bike.

Question 2: Why is it important to use drop_first=True during dummy variable creation?

- Yes, we should. Let's imagine we are looking at a coin flip, and have a feature called is_head, we do not need a column is_tail because we already know it via is_head=False.
- Same applies to other features like your month, if jan to nov are false it is clear that it is december.
- Why is that important? Because more dummy features make it harder for the algorithm to fit or even worse make it easier to overfit.

Question 3: Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

- From the above pairplot we could observe that, **temp** has highest positive correlation with target variable cnt
- A positive correalation observed between cnt and temp (0.63)
- A Negative correlation observed for cnt with hum and windspeed (-0.099 and -0.24)

Question 4: How did you validate the assumptions of Linear Regression after building the model on the training set?

- The distribution plot of error term shows the normal distribution with mean at Zero.
- It seems like the corresponding residual plot is reasonably random.
- Also the error terms satisfies to have reasonably constant variance (homoscedasticity)

Question 5: Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

- Based on final model top three features contributing significantly towards explaining the demand are:
  - Temperature (0.552)
  - Weathersit_3 : Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds (-0.264)
  - Year (0.256)

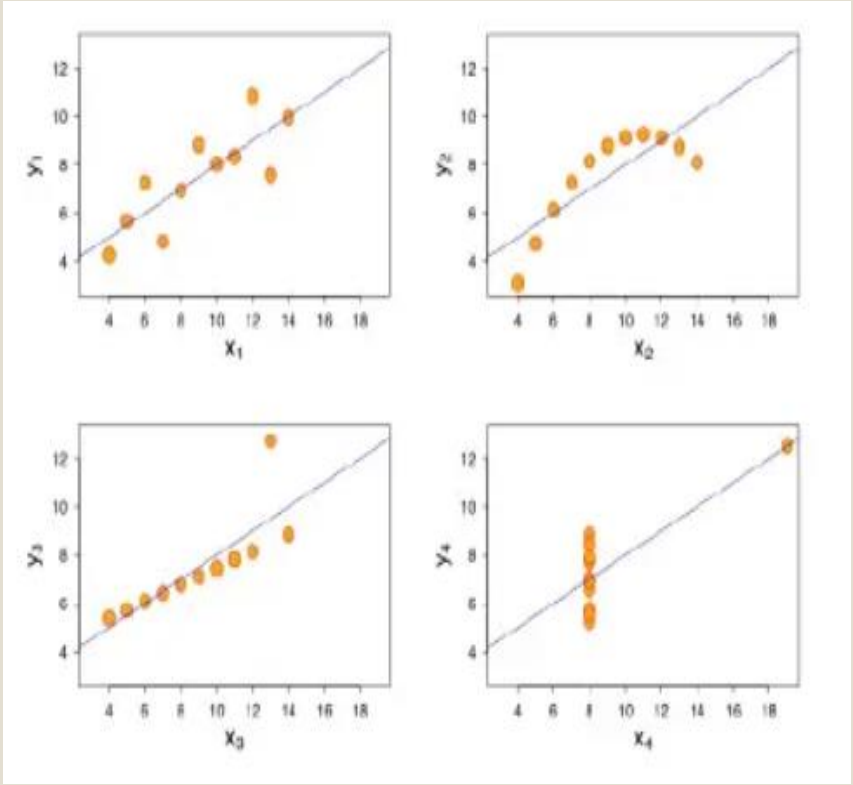
General Subjective Questions

Question 1: Explain the linear regression algorithm in detail.

- **Answer:** A straight line is used by a linear regression technique to create a relationship between independent and dependent variables. This approach is only appropriate for numerical variables.
- When using linear regression, the following actions are taken:
- Test and training data are separated into two sections of the dataset.
- The training data is then separated into dependent target datasets and independent features.
- The training dataset is utilised to create a linear model. The gradient descent approach is used internally by the Python APIs to find the best-fit line's coefficients. By minimising the cost function, which is commonly represented by the residual sum of squares, the gradient descent algorithm works.
- The predicted variable changes from a line to a hyperplane when numerous characteristics are present.
- The predicted variable takes the following form:
- Y= BO+B1x1+ß2x2+B3x3+…..+ Bnxn
- The predicted variable is than compared with test data and assumptions are checked.

Question 2: Explain the Anscombe's quartet in detail.

- **Answer:** Four distinct datasets make up Anscombe's quartet, which exhibit very comparable simple descriptive statistics yet noticeably diverse distributions when represented graphically. The mean, sample variances for x and y, correlation coefficient, linear regression line, and R-Square value are some examples of these fundamental statistics. The quartet provides as an example of how various datasets with similar statistical properties can display striking differences when visualised. The accompanying graphs are shown:



Question 2: Explain the Anscombe's quartet in detail.

- A simple linear relationship is shown in the opening plot (top left).
- The correlation coefficient is insignificant since the second plot (top right) does not follow a normal distribution and hence shows a non-linear relationship.
- Although there is a clear regression line in the third plot (bottom left), it displays linearity. Outliers in the data are thought to be the cause of this disparity.
- The fourth plot (bottom right) does not show a linear connection, but the statistics are affected by the presence of outliers.

In order to ensure reliable and accurate results, it is advised to visualise data and remove outliers before analysis.

Question 3: What is Pearson's R?

- **Answer:** The strength of a relationship between two variables is measured by Pearson's R. It is calculated by dividing the covariance of two variables by the sum of their standard deviations. Its range of values is +1 to -1.
- There are several types of correlation coefficient formulas.
- One of the most commonly used formulas in stats is Pearson's correlation coefficient formula.

$$\rho_{X,Y} = \frac{\mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y]}{\sqrt{\mathrm{E}[X^2] - [\mathrm{E}[X]]^2}\,\sqrt{\mathrm{E}[Y^2] - [\mathrm{E}[Y]]^2}}.$$

- • A value of 1 denotes a complete linear positive correlation. It implies that if one variable rises, the others will follow suit.
- • Zero indicates there is no association.
- • A result of -1 indicates a completely negative connection. It implies that if one variable rises, another will fall.

Question 4: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer: To keep a variable within a certain range, variable scaling is used. In the case of linear regression analysis, it acts as pre-treatment. Scaling is mostly used to speed up gradient descent computation. If the step sizes are selected to assure accuracy, the gradient descent process can become quite time-consuming when the data comprises both tiny variables (with values in the range of 0-1) and large variables (with values in the range of 0-1000).

Normalised Scaling	Standardized scaling
Called min max scaling, scales the variable such that the range is 0-1	Values are centred around mean with a unit standard deviation
Good for non- gaussian distribution	Good for gaussian distribution
Value id bounded between 0 and 1	Value is not bounded
Outliers are also scaled	Does not affect outliers

Question 5: You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Answer: The R-squared statistic can be used to quantify the degree of correlation between a particular predictor variable and the other predictor variables in a linear regression model. Regressing the relevant predictor against the remaining predictor variables yields this statistic.

The formula for VIF is

$VIFi = 1/ 1- R_i^2$

Basically, if R square is 1 than VIF becomes infinite. It means that there is perfect correlation between the features.

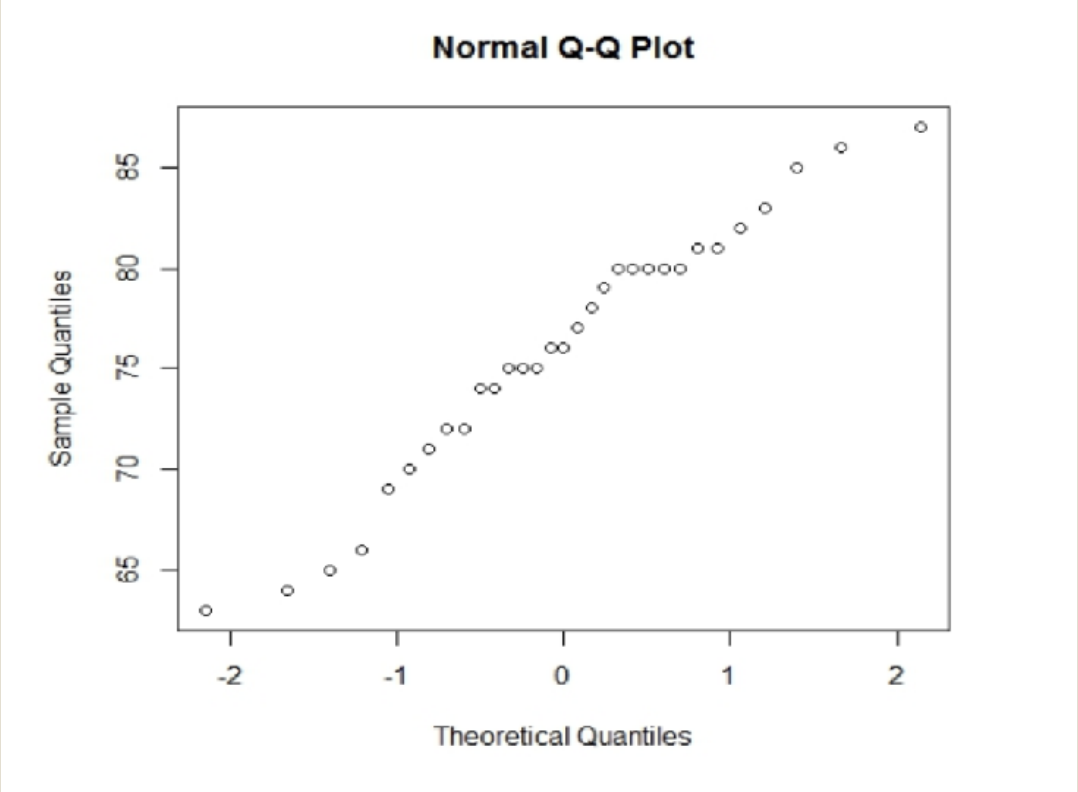
Question 6: What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Answer: A graphic representation that contrasts two sets of quantiles is referred to as a Q-Q plot, also known as a quantile-quantile plot. Identifying if the two datasets are drawn from the same underlying distribution is the main goal of the analysis. A completely straight line in this figure demonstrates that the datasets come from the same source, serving as a visual assessment of the data.

By placing your sample data in ascending order and charting them against quantiles obtained from a theoretical distribution, you can create Q-Q graphs. In order to match the size of your sample data, the number of quantiles is selected. Since many statistical methods assume normality, normal Q-Q plots are frequently used, however Q-Q plots can be made for any distribution.

Question 6: What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

In statistics, Q-Q plots are frequently utilised in an easy-to-understand way: by fitting a linear regression model and determining if the dots roughly line up with the line. The errors are also non-Gaussian if they wander from the line, which suggests that the residuals are not normally distributed. This indicates that the traditional confidence intervals and significance tests are invalid for small sample sizes if the estimator is not assumed to be Gaussian.



Thank you!