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)

Answer: Inferences from analysis of categorical variables -

i. **season** : Demand for bikes goes very down in Spring season
ii. **yr**: Demand for bikes is less in year 2018 and more in year 2019
iii. **holiday**: On holiday demand for bikes is less.
iv. **weathersit** : Demand for bikes is more in clear and cloudy weather, less in rainy weather, and no demand in winters.

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

Answer: If we don't use "drop_first" we will get a redundant feature. e.g.

If we have a variable gender, we don't need both a male and female dummy. Just one will be fine. If male=1 then the person is a male and if male=0 then the person is female.

So drop_first will drop the first column while creating dummy variable and we get n-1 columns for categorical variable with n categories.

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

Answer: 'temp' has the highest correlation with the target variable among the numerical variables.

('causal' and 'registered' total make target variable, thus have a higher correlation with target variable. So they also treated as target variables and not predictors.)

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

Answer: We can validate the model on below points:

i. Check the R-squared and adjusted R-squared, greater than 60% is good. In our case it is approximately 84%.
ii. p-value of all predicting variables is smaller than 0.05
iii. VIF value of all predicting variables is smaller than 10
iv. Error terms are normally distributed
v. Residuals have equal or almost equal variance.

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

Answer: T op 3 features contributing variables are:

i. **yr** : Demand increases in year 2019 (Coefficient=1941.1246)
ii. **weathersit_rainy** : Demand decreases in rainy weather (Coefficient=-1651.6616)
iii. **temp** : Demand increases with increase in temp (Coefficient=809.8933)

General Subjective Questions

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

Answer:

Definition:

Linear regression is a linear approach to modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables).

Types:

Linear regression models can be classified into two types depending upon the number of independent variables:

- Simple linear regression: This is used when the number of independent variables is 1.
- Multiple linear regression: This is used when the number of independent variables is more than 1.

Equation:

Simple linear regression: Y = β0 + β1X

Multiple linear regression: Y = β0 + β1X1 + β2X2 + ... + βnXn + E

The strength of the linear regression model can be assessed using 2 metrics:

1. R^2 or Coefficient of Determination

2. Residual Standard Error (RSE)

Assumptions:

i. There is some Linear relationship between X and Y
ii. Error terms are normally distributed (not X, Y)
iii. Error terms are independent of each other
iv. Error terms have constant variance (homoscedasticity)

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

Answer: Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

i. In the first one we will see that there seems to be a linear relationship between x and y.
ii. In the second one we can conclude that there is a non-linear relationship between x and y.
iii. In the third one we can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
iv. Finally, the fourth one show an example when one high-leverage point is enough to produce a high correlation coefficient.

Application:

The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

3. What is Pearson's R? (3 marks)

Answer: Pearson's correlation (also called Pearson's R) is a correlation coefficient commonly used in linear regression. The Pearson coefficient is a type of correlation coefficient that represents the relationship between two variables that are measured on the same interval or ratio scale. The Pearson coefficient is a measure of the strength of the association between two continuous variables.

One of the most commonly used formulas is Pearson's correlation coefficient formula.

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

Sample correlation coefficient

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

Population correlation coefficient

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

Assumptions:

i. For the Pearson r correlation, both variables should be normally distributed.

ii. There should be no significant outliers.
iii. Each variable should be continuous i.e. interval or ratios for example weight, time, height, age etc.
iv. The two variables have a linear relationship.
v. The observations are paired observations. That is, for every observation of the independent variable, there must be a corresponding observation of the dependent variable.
vi. Error terms have constant variance (homoscedasticity)

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

Answer: Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Why Scaling: Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To bring all features in the same standing, we need to do scaling so that one significant number doesn't impact the model just because of their large magnitude. So we need to scale features because of two reasons:

i. Ease of interpretation
ii. Faster convergence for gradient descent methods

Normalization/MinMax Scaling is used when we want to bound our values between two numbers, typically, between [0,1] or [-1,1]. While **Standardisation** transforms the data to have zero mean and a variance of 1, they make our data unitless.

The advantage of Standardisation over the other is that it doesn't compress the data between a particular range. This is useful, especially if there is are extreme data point (outlier).

MinMax Scaling x = (x-min(x))/(max(x)-min(x))

Standardisation x = (x-mean(x))/sd(x)

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

(3 marks)

Answer: **Variance Inflation Factor** (VIF) : VIF basically helps explaining the relationship of one independent variable with all the other independent variables.

VIF = 1/(1-(R2))

VIF greater than 10 is definitely high.

An infinite VIF shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

(3 marks)

Answer: Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

It is used to check following scenarios:

If two data sets —

i. come from populations with a common distribution
ii. have common location and scale
iii. have similar distributional shapes
iv. have similar tail behaviour

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

Below are the possible interpretations for two data sets.

a) Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis
b) Y-values < X-values: If y-quantiles are lower than the x-quantiles.
c) Image for post
d) X-values < Y-values: If x-quantiles are lower than the y-quantiles.
e) Image for post
f) Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis