Assignment -based Subjective Questions

Question1:From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?(3marks)

Question2:Why is it important to use **drop\_first=True** during dummy variable creation?(2 mark)

Answer2:

The regression model contains dummy variables of categorical data after using one-hot encoding. The variables are highly correlated with each other which means one variable can predict from other variables. In the regression model, this variable creates a trap which is called the dummy variable trap. Including all variable result in redundant data.

The solution of the Dummy variable trap is to drop/remove one of the dummy variables. If there are p categories than p-1 dummy variable should use. The model should exclude one dummy variable.This Overall approach reduces Multi-colinearity in the dataset, which is one of the prime Assumption of Multiple Linear Regression.

Question3:Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?(1 mark)

Answer3: Temperature\_actuall(temp) and Temperature\_felt(atemp) has the highest correlation with target variable. Both has correlation value of 0.63 w.r.t target variable count\_rental\_bikes(cnt).

Question4:How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Question4: 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

**Question 1: Explain the linear regression algorithm in detail.(4 Marks)**

**Answer1:** Linear Regression is a machine learning algorithm which is based on **supervised learning** category. It finds a best linear-fit relationship on any given data, between independent (Target) and dependent (Predictor) variables. In other words, it creates the best straight-line fitting to the provided data to find the best linear relationship between the independent and dependent variables. Mostly it uses **Sum of Squared Residuals**Method.

Linear regression is of the 2 types:

a. **Simple Linear Regression:**It explains the relationship between a dependent variable and only one independent variable using a straight line. The straight line is plotted on the scatter plot of these two points.

**Formula for the Simple Linear Regression:**

Y=β0+β1X1 +ϵ

b. **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=β0+β1X1+β2X2+…+βpXp+ϵ

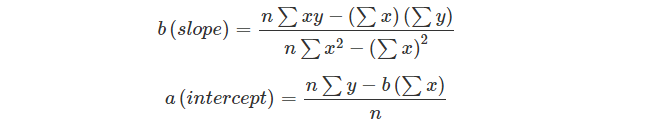
The equation of the best fit regression line Y = β₀ + β₁X can be found by the following two methods:

· Differentiation

· Gradient descent

We can use statsmodels or SKLearn libraries in python for the linear regression.

Where a and b given by the formulas:



Here,

X or X is the dependent variable we are trying to predict

Y or Y is the dependent variable we are using to make predictions.

b or β1 is the slope of the regression line which represents the effect X has on Y

a or β0 is a constant, known as the Y-intercept. If X = 0,Y would be equal to b.

**Question 2: Explain the Anscombe’s quartet in detail.**

**Answer2:**Anscombe’s Quartet was developed by statistician **Francis Anscombe**. This is a method which keeps four datasets, each containing eleven (x, y) pairs. The important thing to note about these datasets is that they share the same descriptive statistics. Each graph tells a different story irrespective of their similar summary statistics. Below is the glimpse of the statistics of the 4 datasets:

Table

Description automatically generated

The summary statistics show that the means and the variances were identical for x and y across the groups:

Mean of x is 9 and mean of y is 7.50 for each dataset.

·Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset

· The correlation coefficient (how strong a relationship is between two variables) between x and y is 0.816 for each dataset

When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset is telling a different story:

Chart, scatter chart

Description automatically generated

· 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.

This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

Additionally, Anscombe’s Quartet warns of the dangers of outliers in data sets.

**Question 3: What is Pearson’s R?**

**Answer:**Pearson’s R was developed by [**Karl Pearson**](https://en.wikipedia.org/wiki/Karl_Pearson) and it is a correlation coefficient which is a measure of the strength of a linear association between two variables and it is denoted by ‘r’. it has a value between +1 and −1, where 1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation.

Mathematically, Pearson’s correlation coefficient is denoted as the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a “product moment”, that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier product-moment in the name.

**Formula:**

Text

Description automatically generated

**Example:**

· 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.

* r > 5 < 8 means there is a moderate association
* r > 8 means there is a strong association

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

Answer4:

It 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.

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 solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

It is important to note that **scaling just affects the coefficients** and none of the other parameters like **t-statistic, F-statistic, p-values, R-squared**, etc.

**Normalized scaling:**

* It brings all of the data in the range of 0 and 1.
* **sklearn.preprocessing.MinMaxScaler**helps to implement normalization in python.

**Minmax Scaling:x = (x-min(x))/(max(x)-min(x))**

**Standardization Scaling:**

* Standardization replaces the values by their Z- scores. It brings all of the data into a standard normal distribution which has mean (**μ)** zero and standard deviation one (**σ**).

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

* **sklearn.preprocessing.scale** helps to implement standardization in python.
* One disadvantage of normalization over standardization is that it **loses** some information in the data, especially about **outliers**.

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

Answer5:

The value of VIF is calculated by the below formula:

VIFi =1/(1-Square(Ri))

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.

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).

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

**Answer:**The Q-Q plot or quantile-quantile plot is a graphical technique for determining if two data sets come from populations with a common distribution.

A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that’s roughly straight. Here’s an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.

**Use of Q-Q plot in Linear Regression:**The Q-Q plot is used to see if the points lie approximately on the line. If they don’t, it means, our residuals aren’t Gaussian (Normal) and thus, our errors are also not Gaussian.

**Importance of Q-Q plot: Below are the points:**

I. The sample sizes do not need to be equal.

II. Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers.

III. The q-q plot can provide more insight into the nature of the difference than analytical methods.