Student Name: Mohit Patil

Registered E-mail ID: mohitz4418@gmail.com

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:

Categorical variables require special attention in regression analysis because, unlike continuous variables, they cannot by entered into the regression equation just as they are. Instead, they need to be recoded into a series of variables which can then be entered into the regression model.

There are a variety of coding systems that can be used when recoding categorical variables. In our analysis we used One Hot Encoding i.e. Dummy Variables.

In our analysis there are 4 categorical features season, mnth, weekday, whethersit to analyse them we firstly converted them from number to its individual category and them applying the one hot encoding on it.

We have "cnt" as our target variable which is Total Count of Rental Bikes. The count of bikes varies for every month, season and according to whether type.

In our model the mostly the total count bikes rented based on whether type – Heavy snow rain, Season – spring, and in January month. The season, mnth and whethersit this categorical column affected on our model.

We handled the categorical variables in following steps:

1. Take the categorical variable and transform it using one-hot encoding.
2. Fit those one-hot encodings to a regression model (ignoring your other features).
3. Replace the categorical variable in your original dataset with the beta coefficients that you found in step 2.
4. Proceed as you would.

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

Answer:

If a categorical variable has 3 variables, we can represent same amount information using 2 dummy variables if you have k categorical levels in a categorical variable, we only need k-1 dummy columns and to do that we remove the first column of dummy variable data frame using **drop_first = True**.

We drop this one column to achieve efficiency. Because of we can represent same amount of information using 2 variables so why do we need that 3rd variable.

So actually, we are removing a redundant column from it. Which helps to achieve model optimization.

When we convert the categorical variables to dummies, indirectly we are giving importance to each value in a categorical column by making each value as a column.

So, coming to your question, why to drop one variable in regression is because the importance or value of that left-over variable can be found by remaining variables. So to avoid redundancy we are dropping a column.

For example, if there are Red, Blue and green are categorical variables. When we convert these into dummies Red, blue and green will be the extra columns created. So, if the particular sample has green value, the Red, Blue will be indicated as Zero. So automatically it indicates the sample belongs to green. If the value of green column can be explained by Red and blue column, then there is no need of green column.

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

Answer:

In pair-plot among the numerical variables the **“registered”** variable has highest correlation with target variable.

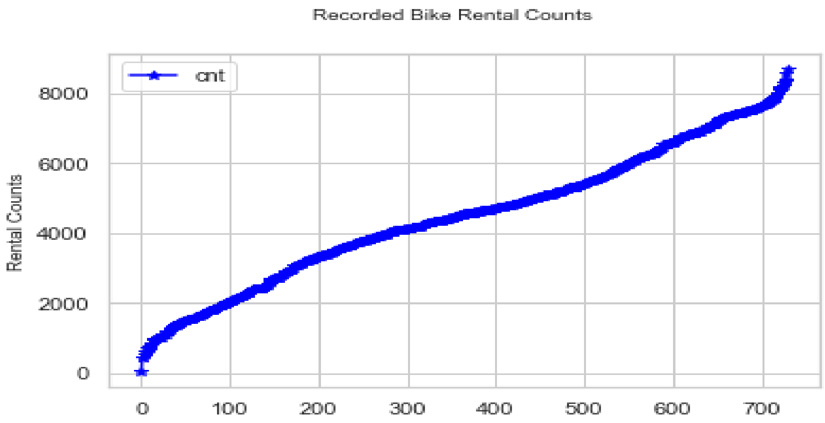
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Answer:

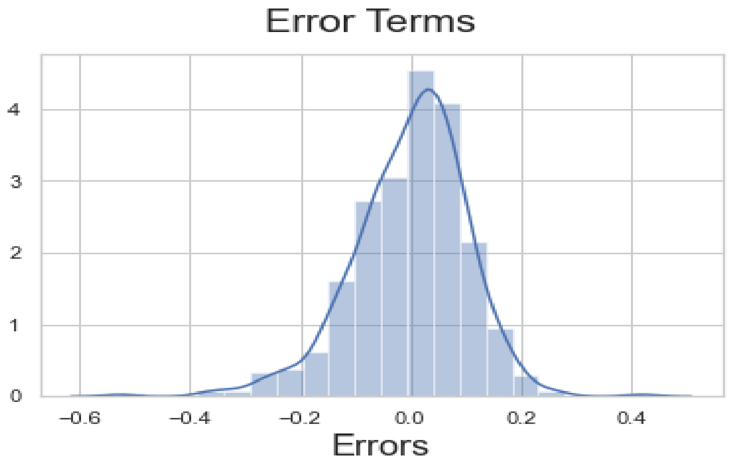
Let’s understand each assumption one by one:

a) There is a linear relationship between X and Y:

To verify this assumption, we can plot a scatter plot or pair plot and if X and Y should display some sort of a linear relationship, otherwise, there is no use of fitting a linear model between them.



b) Error terms are normally distributed with mean zero:



There is no problem if the error terms are not normally distributed if you just wish to fit a line and not make any further interpretations.

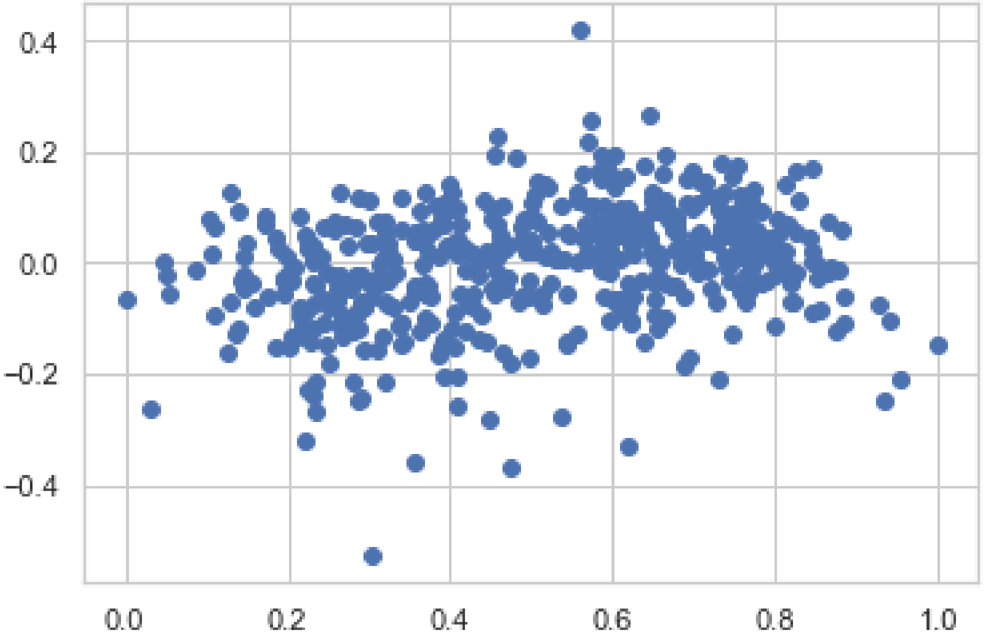
The assumption of normality is made, as it has been observed that the error terms generally follow a normal distribution with mean equal to zero in most cases

We can verify this assumption by plotting a distplot.

c) Error terms are independent of each other

The error terms should not be dependent on one another.

To verify this assumption, we can plot a scatterplot of error terms and of the points on scatter plot are scattered and not dependent on each other means it has no visible pattern then this assumption is verified.



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

Answer:

1. Heavy Snow Rain
2. yr
3. Spring

General Subjective Question.

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

Answer:

Linear regression is one of the very basic forms of machine learning where we train a model to predict the behaviour of your data based on some variables. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

Linear regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression model target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.

Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.

$$y = \theta_1 + \theta_2 . x$$

While training the model we are given:
x: input training data (univariate – one input variable(parameter))
y: labels to data (supervised learning)
When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best θ_1 and θ_2 values.
θ_1: intercept
θ_2: coefficient of x.

Once we find the best θ_1 and θ_2 values, we get the best fit line. So, when we are finally using our model for prediction, it will predict the value of y for the input value of x.

The key point in Simple Linear Regression is that the *dependent variable must be a continuous/real value*. However, the independent variable can be measured on continuous or categorical values.
Simple Linear regression algorithm has mainly two objectives:

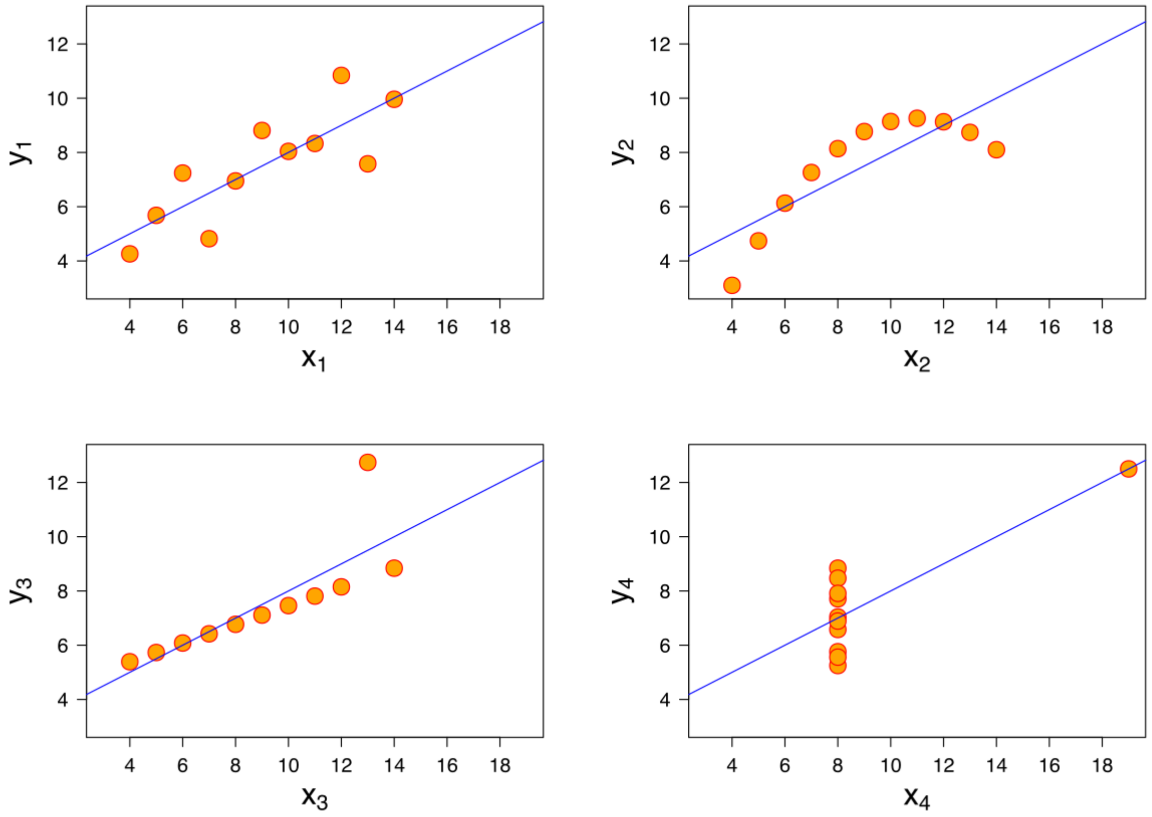
- Model the relationship between the two variables. Such as the relationship between Income and expenditure, experience and Salary, etc.
- Forecasting new observations. Such as Weather forecasting according to temperature, Revenue of a company according to the investments in a year, etc.

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

Answer:

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and 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 analysing it and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough.



- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.
- The second graph (top right) is not distributed normally; while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
- In the third graph (bottom left), the distribution is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.

- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyse 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 coefficient is 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.

Correlation is a technique for investigating the relationship between two quantitative, continuous variables, for example, age and blood pressure. Pearson's correlation coefficient (r) is a measure of the strength of the association between the two variables.

The Pearson's correlation coefficient varies between -1 and +1.

Questions a Pearson correlation answers -

- Is there a statistically significant relationship between age and height?
- Is there a relationship between temperature and ice cream sales?
- Is there a 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?

Correlation is a bi-variate analysis that measures the strength of association between two variables and the direction of the relationship. In terms of the strength of relationship, the value of the correlation coefficient varies between +1 and -1.

A value of ± 1 indicates a perfect degree of association between the two variables. As the correlation coefficient value goes towards 0, the relationship between the two variables will be weaker. The direction of the relationship is indicated by the sign of the coefficient; a + sign indicates a positive relationship and a - sign indicates a negative relationship.

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

Answer:

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data pre-processing step.

Since the range of values of raw data varies widely, in some machine learning algorithms, objective functions will not work properly without normalization.

For example, many classifiers calculate the distance between two points by the Euclidean distance. If one of the features has a broad range of values, the distance will be governed by this particular feature. Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance.

Another reason why feature scaling is applied is that gradient descent converges much faster with feature scaling than without it.

In machine learning, we can handle various types of data, this data can include multiple dimensions.

Feature standardization makes the values of each feature in the data have zero-mean (when subtracting the mean in the numerator) and unit-variance. This method is widely used for normalization in many machine learning algorithms.

The terms normalization and standardization are sometimes used interchangeably, but they usually refer to different things. Normalization usually means to scale a variable to have a value between 0 and 1, while standardization transforms data to have a mean of zero and a standard deviation of 1.

In scaling *(also called min-max scaling)*, you transform the data such that the features are within a specific range e.g. [0, 1].

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

where x’ is the normalized value.

The point of normalization is to change your observations so that they can be described as a normal distribution.

Normal distribution (Gaussian distribution), also known as the bell curve, is a specific statistical distribution where a roughly equal observations fall above and below the mean, the mean and the median are the same, and there are more observations closer to the mean.

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

Answer:

VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent 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). In general, one starts with the selection of all variables, and proceeds by repeatedly deselecting variables showing a high VIF. 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).

To avoid this problem
Try one of these:

1. Remove highly correlated predictors from the model. If you have two or more factors with a high VIF, remove one from the model.
2. Use Partial Least Squares Regression (PLS) or Principal Components Analysis, regression methods that cut the number of predictors to a smaller set of uncorrelated components.

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

Answer:

The Q-Q plot, or quantile-quantile plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal or exponential.

For example, if we run a statistical analysis that assumes our dependent variable is Normally distributed, we can use a Normal Q-Q plot to check that assumption. It's just a visual check, not an air-tight proof, so it is somewhat subjective. But it allows us to see at-a-glance if our assumption is plausible, and if not, how the assumption is violated and what data points contribute to the violation.

If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

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.

Few Advantages:

a) It can be used with sample sizes also

b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

It is used to check following scenarios:

If two data sets —

i. come from populations with a common distribution

ii. ii. have common location and scale

iii. iii. have similar distributional shapes

iv. 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.