LINEAR REGRESSION ASSIGNMENT

DECEMBER 2023

SHISHIR KHANAL

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:

Based on the categorical variable analysis conducted through boxplots and bar plots, several noteworthy trends emerge that shed light on their impact on the dependent variable:

- More people are booking during the fall season, and overall, the number of bookings has gone up a lot from 2018 to 2019 in every season.
- The majority of bookings happened in May, June, July, August, September, and October. The trend shows an increase from the beginning to the middle of the year, then starts to decrease towards the end of the year. The number of bookings for each month appears to have gone up from 2018 to 2019.
- Wednesday, Thursday, Friday, and Saturday have a higher number of bookings compared to the beginning of the week.
- Booking numbers increased notably during clear weather. Additionally, when comparing 2019 to the previous year 2018, there was an overall increase in bookings for every weather condition.
- Booking seems to be less in number, when it's not a holiday.
- The number of bookings seems to be more on working days compared to non-working days.
- The increase in bookings from the previous year to 2019 reflects positive progress in terms of business.

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

Answer:

'drop_first=True' is important to use, because it helps in reducing an extra column during the creation of dummy variables. This is necessary for reducing correlations among the dummy variables.

By setting drop_first=True, we automatically drop one of the dummy variables, and this serves as the reference category. This eliminates perfect multicollinearity in regression models, making the model more stable and interpretable.

Let's consider, we have three types of values in categorical column A, B, C, and we want to create dummy variable for that column. If one variable is not A and B, then It is obvious C. So we do not need 3rd variable to identify the C.

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

Answer:

From the observations made in the pair-plot involving numerical variables, it can be concluded that the variable **'temp'** exhibits the highest correlation with the target variable.

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

Answer:

Validating the assumptions of linear regression is an important step to ensure the reliability of the model. Here are some assumptions that I have used to validate model after building a linear regression model on the training set:

- **Normality of Residuals:**
  The residuals should be approximately normally distributed. I checked this assumption by plotting histogram of residuals (we can also use a Q-Q plot (quantile-quantile plot) to assess normality).

- **Multicollinearity:**
  If our linear regression model includes multiple independent variables, we need to check for multicollinearity, which occurs when independent variables are highly correlated. High correlation can destabilize the model coefficients and make interpretation challenging. I checked multicollinearity by using variance inflation factor (VIF) and correlation matrices to identify multicollinearity.

- **Independence of Residuals:**
  The residuals (differences between observed and predicted values) should be independent of each other. I checked this assumption by plotting the residuals against the predicted values or the independent variables. There is no visible pattern in these plots.

- **Linearity:**
  The relationship between the independent and dependent variables should be linear. I checked this assumption by plotting the observed values against the predicted values. A scatter plot with a straight line indicates a linear relationship.

- **Homoscedasticity (Constant Variance):**
  The variance of the residuals should be constant across all levels of the independent variables. I checked this assumption by plotting the residuals against predicted values, which helped identify heteroscedasticity. There is No a fan or cone-shaped pattern in the residuals.

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: Based on the final model, following are the foremost 3 features contributing significantly towards explaining the demand of the shared bikes:

1. temp
2. weathersit_Light_snowrain
3. year

General Subjective Questions:

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

Answer:

Linear regression is a statistical method used for modeling the linear relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data. It is a fundamental algorithm in supervised machine learning and statistics, commonly employed for predictive analysis and understanding the relationships between variables.

In this algorithm, the model finds the best fit linear line between the independent and dependent variable. This best fit is characterized by minimizing the total prediction error across all data points. The error, in this context, represents the distance between the data points and the regression line.

Linear Regression Line:

A regression line shows the relationship between dependent and independent variables. A positive linear relationship occurs when both variables increase, while a negative linear relationship happens when the dependent variable decreases as the independent variable increases.

Linear Regression is generally classified into two types:

- Simple Linear Regression/ Univariate Linear regression
- Multiple Linear Regression

1. Simple Linear Regression/ Univariate Linear regression:

When we try to find the relationship between a single independent variable (input) and a corresponding dependent variable (output) then it is known as simple linear regression/ univariate linear regression. This can be expressed in the form of a straight line.

The equation of a line can be written as:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- Y represents the output variable or dependent variable.

- ε (Epsilon) is the error term.
- β0 and β1 are two unknown constants that represent the intercept and coefficient respectively. To create a model, we must "learn" the values of these coefficients. And once we have the value of these coefficients, we can use the model to predict the target variable.

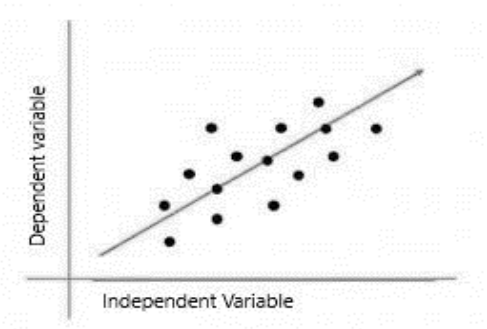
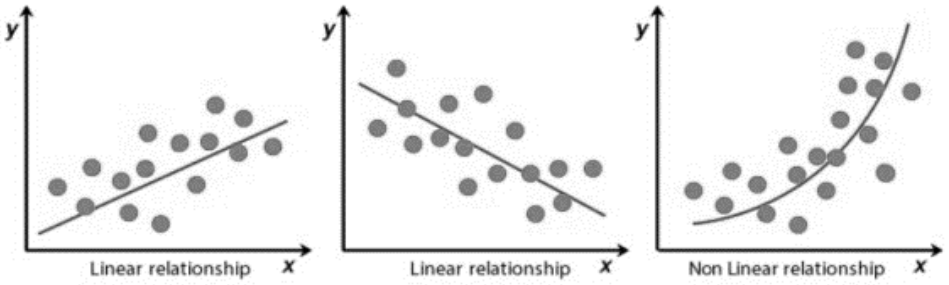


Fig 1: Graph representation of a simple linear regression model

Assumptions of Simple Linear Regression:

Linear regression is a powerful tool for understanding and predicting the behavior of a variable, however, it needs to meet a few conditions in order to be accurate and dependable solutions.

- **Linearity:** The relationship between the independent and dependent variables is assumed to be linear.



- **Independence:** Observations are assumed to be independent of each other. This means that the value of the dependent variable for one observation does not depend on the value of the dependent variable for another observation.

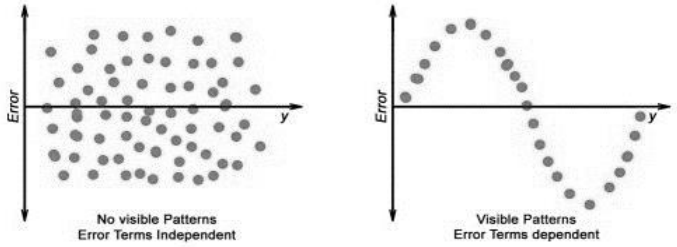
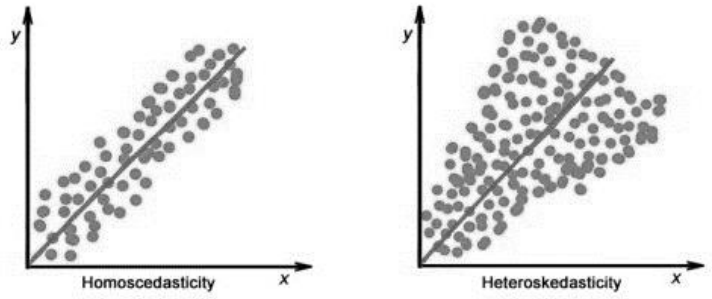


Fig 2: Graph representation

- **Homoscedasticity:** The variance of the residuals (the differences between observed and predicted values) should be constant across all levels of the independent variable(s).



- **Normality:** The residuals are assumed to be normally distributed. The mean of residuals should follow a normal distribution with a mean equal to zero or close to zero.

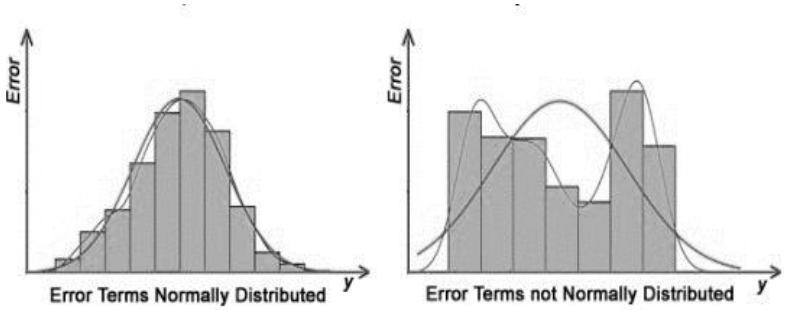
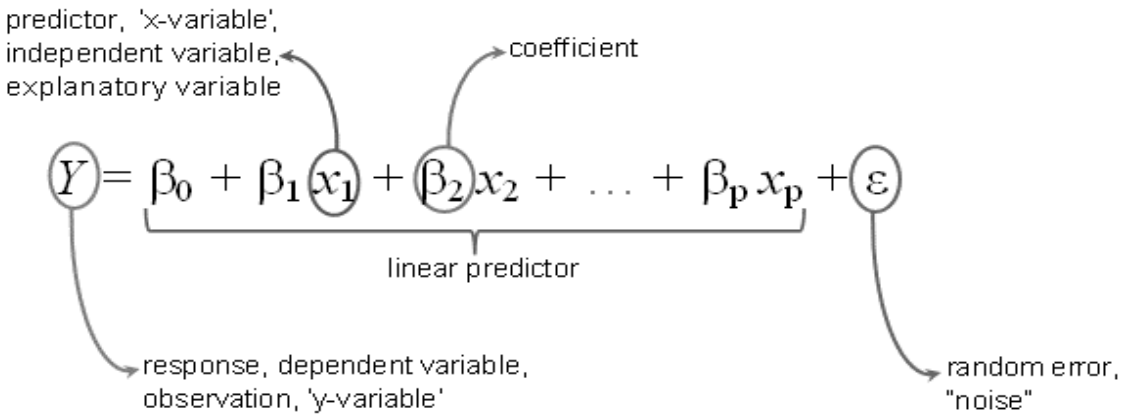


Fig 3: Graph representation

2. Multiple Linear Regression:

When we try to find the relationship between two or more independent variables (inputs) and the corresponding dependent variable (output) then it is known as multiple linear regression.

The equation that describes how the predicted value of y is related to p independent variables is called as Multiple Linear Regression equation:



Assumptions of Multiple Linear Regression:

All the four assumptions made for simple linear regression still hold true for multiple linear regression along with a few new additional assumptions such as **overfitting, multicollinearity, and feature Selection.**

Linear regression stands as a versatile and interpretable algorithm with widespread applications in finance, medicine, economics, biology, and engineering. Its simplicity, interpretability, and scalability driven by clear insights from coefficients, make it a foundational tool for statistical modeling, data analysis and machine learning. Despite its advantages, practitioners must be mindful of its sensitivity to assumptions and potential limitations in handling complex, nonlinear relationships.

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

Answer:

Anscombe's quartet is a group of datasets (x, y) that have the same mean, standard deviation, and regression line, but which are qualitatively different.

Anscombe's quartet tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.).

Let us take quartet consists of four different datasets, each containing 11 points, with two variables: x and y, such as x1 & y1, x2 & y2, x3 & y3, x4 & y4. The datasets and their graphical representation are shown in the following:

x1	y1
10	8.04
8	6.95
13	7.58
9	8.81
11	8.33
14	9.96
6	7.24
4	4.26
12	10.84
7	4.82
5	5.68
Mean	9	7.5
SD	3.3	2.0

x2	y2
10	9.14
8	8.14
13	8.74
9	8.77
11	9.26
14	8.1
6	6.13
4	3.1
12	9.13
7	7.26
5	4.74
Mean	9	7.5
SD	3.3	2.0

x3	y3
10	7.46
8	6.77
13	12.74
9	7.11
11	7.81
14	8.84
6	6.08
4	5.39
12	8.15
7	6.42
5	5.73
Mean	9	7.5
SD	3.3	2.0

x4	y4
8	6.58
8	5.76
8	7.71
8	8.84
8	8.47
8	7.04
8	5.25
19	12.5
8	5.56
8	7.91
8	6.89
Mean	9	7.5
SD	3.3	2.0

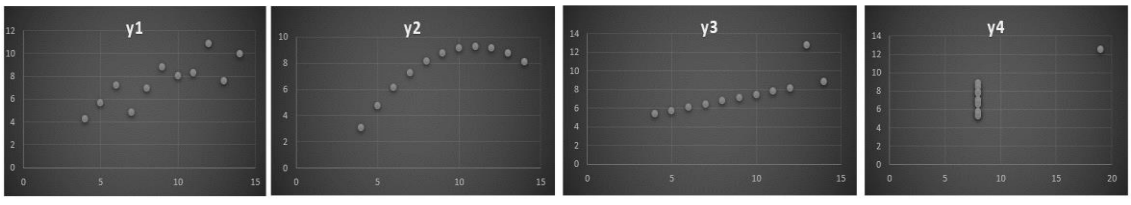


Fig 4: Graphical representation of Anscombe's quartet

Despite the variations in each dataset, they share identical summary statistics, including mean, standard deviations (SD), correlational coefficient, and linear regression line.

- The first dataset shows simple linear relationship, y increases with x.

- The second dataset shows a linear trend, a single outlier affects the regression line, creating a misleading representation of the data.
- The third dataset shows perfect quadratic relationship, emphasizing nonlinear patterns.
- Fourth dataset introduces a new layer of complexity to the situation, one outlier contradicts the pattern.

Anscombe's Quartet warns against blindly trusting summary statistics or standard methods of analysis. It suggests us to look closely at our data, think about our assumptions, and use various analytical tools to get a full picture. This concept focuses on the importance of visualizing data, as graphs can reveal patterns and outliers that summary statistics alone may overlook.

Additionally, Anscombe's Quartet emphasizes the importance of exploratory data analysis (EDA). By thoroughly examining our data, conducting descriptive statistics, and visualizing relationships, we can find hidden insights and avoid misleading conclusions.

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

Answer:

Pearson's correlation coefficient(r) is the test statistics that measures the statistical relationship, or correlation, between two continuous variables. It is the most common way of measuring a linear correlation and its value lies between –1 and 1 that measures the magnitude and direction of the relationship between two variables.

Formula and Calculation:

Pearson's Correlation coefficient is represented as 'r', it measures how strong is the linear association between two continuous variables using the formula.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Where,

r = Pearson Correlation Coefficient

x_i = x variable samples y_i = y variable sample

$\bar{x}$ = mean of values in x variable $\bar{y}$ =mean of values in y variable

Assumptions:

The Pearson correlation coefficient is a good choice when all of the following are true:

- Both variables are quantitative
- The variables are normally distributed
- The data have no outliers
- The relationship is linear:
- Independent of case
- Homoscedasticity

Properties of Pearson's correlation coefficient(r):

- Range Limit: Coefficient values can range from +1 to -1, where +1 indicates a perfect positive relationship, -1 indicates a perfect negative relationship, and a 0 indicates no relationship exists.
- Pure number: It is independent of the unit of measurement.
- Symmetric: Correlation of the coefficient between two variables is symmetric. This means between X and Y or Y and X, the coefficient value of will remain the same.

Degree of Correlation:

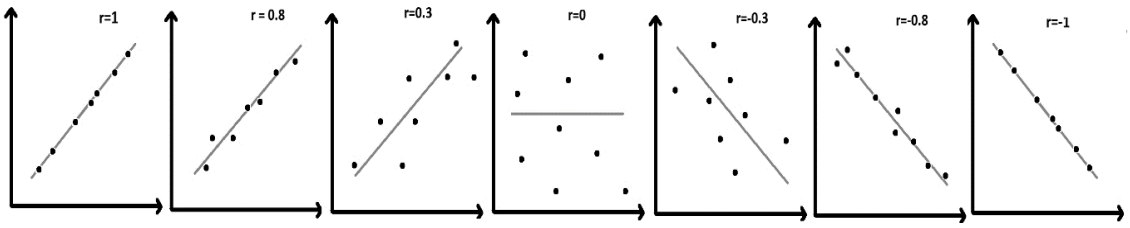


Fig 5: Degree of correlation

- Perfect: If the value is ± 1 or near ± 1, then it said to be a perfect correlation, as one variable increases, the other variable tends to also increase (if positive) or decrease (if negative).
- High degree: If the coefficient value lies between ± 0.50 and ± 1, then it is said to be a strong correlation.
- Moderate degree: If the value lies between ± 0.30 and ± 0.49, then it is said to be a medium correlation.
- Low degree: When the value lies below ± 0.29, then it is said to be a small correlation.
- No correlation: When the value is zero then it is said to be no correlation.

Pearson's correlation coefficient (r) is a widely used metric for measuring linear relationships between continuous variables, offers simplicity and standardization. It is commonly employed in various fields, such as statistics, data analysis, and machine learning, to assess the strength and direction of the correlation between two sets of data points, providing insights into their association. But caution is necessary due to its sensitivity to outliers and the assumption of linearity.

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

Answer:

Scaling is the step of data pre-processing in data analysis refers to the process of transforming numerical variables to a standard range or normalizing the data within a particular range. The primary reasons for scaling are to ensure that all variables contribute equally to the analysis and to make the interpretation of coefficients or feature importance in machine learning models more meaningful. It also helps in speeding up the calculations in an algorithm.

Why is scaling performed?

Most of the time, collected data set contains features highly varying in magnitudes, range and units. When scaling is omitted, algorithms prioritize magnitude over units, leading to inaccurate modeling. To address this issue, we have to do scaling to bring all the variables to the same level of magnitude.

Scaling just impacts the coefficients and leaves other parameters such as t-statistic, F-statistic, p-values, R-squared, etc., unaffected.

There are two common types of scaling:

- Normalization/MinMax Scaling
- Standardization Scaling

Difference Between Normalized Scaling and Standardized scaling:

Normalization	Standardization
➢ Goal of normalization is to scale the data to a specific range, rescales values to a range between 0 and 1	➢ Goal of standardization is to transform the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).
➢ Normalization can be sensitive to outliers, especially if the range is influenced by extreme values	➢ Standardization is generally less sensitive to outliers since it uses the mean and standard deviation, which are less affected by extreme values
➢ Normalization retains the shape of the original distribution	➢ Standardization changes the shape of the original distribution
➢ May not preserve the relationships between the data points	➢ Preserves the relationships between the data points
➢ **sklearn.preprocessing.MinMaxScaler** helps to implement normalization in python	➢ **sklearn.preprocessing.scale** helps to implement standardization in python

Normalization	Standardization
➢ Useful when the distribution of the data is unknown or not Gaussian	➢ Useful when the distribution of the data is Gaussian or unknown
➢ MinMaxscaling x: (x – min)/(max – min)	Standardization: (x – mean)/standard deviation

The choice between normalization and standardization depends on the specific requirements of the analysis or machine learning algorithm being used, as well as the characteristics of the dataset.

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

Answer:

A variance inflation factor (VIF) is a measure of the amount of multicollinearity in regression analysis. Multicollinearity exists when there is a correlation between multiple independent variables in a multiple regression model. A large VIF on an independent variable indicates a highly collinear relationship to the other variables that should be considered or adjusted for in the structure of the model and selection of independent variables.

The formula for VIF is given by:

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

❖ Where **R^2** is a statistical measure that indicates how much of the variation of a dependent variable is explained by an independent variable in a regression model.

The VIF can become infinite when the R- squared (R^2) value is equal to 1. This situation arises when a perfect linear relationship exists between two independent variables in the model.

When $R^2=1$, the denominator in the VIF formula becomes zero, leading to an infinite VIF value. This indicates that the variance of the variable's estimated regression coefficient is inflated to an infinite degree due to perfect multicollinearity.

To address this issue, it's necessary to identify and handle multicollinearity in the regression model. This can involve removing highly correlated variables, combining them into a single variable, or using regularization techniques to mitigate the impact of multicollinearity on the model.

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

Answer:

A Q-Q plot, which stands for quantile-quantile plot, is a graphical tool used in statistics to assess whether a dataset follows a particular theoretical distribution, such as the normal distribution. It compares the quantiles of the observed data against the quantiles of a theoretical distribution, typically the normal distribution. Doing this helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential.

Here's how create Q-Q plot:

- **Sort the data:**
  Arrange the data in ascending order.
- **Calculate theoretical quantiles:**
  For each data point, compute the corresponding quantile from the theoretical distribution.
- **Plot the points:**
  Plot the observed quantiles against the expected quantiles.

To interpret a Q-Q plot, we need to look at the shape and pattern of the points. If the points lie on or close to a 45-degree line, it means that the data follow the reference distribution closely. If the points deviate from the line, it means that there are some differences between the data and the assumed distribution.

A Q-Q plot can be used in regression models to check some of the assumptions that are required for valid inference. We can also use a Q-Q plot to check if the residuals have a constant variance, which is an assumption for the homoscedasticity of the model. To do this, you need to create a Q-Q plot for the residuals of the model and compare them with the normal distribution.

So, in the context of linear regression, Q-Q plots are often used to check the normality of the residuals. Residuals are the differences between the observed values and the values predicted by the regression model. It is important to assess whether these residuals follow a normal distribution because many statistical tests and confidence intervals in linear regression assume normality.

statsmodels.api provides qqplot and qqplot_2samples to plot Q-Q graph for single and two different data sets respectively in python.

In summary, Q-Q plots are a valuable diagnostic tool in linear regression analysis. They help researchers and analysts assess the normality of residuals and identify potential issues with the model assumptions, which is important for making reliable inferences based on the regression analysis.

The End!!