Linear Regression Assignment

Name: Bhumika Nautiyal

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

- **Ans-** I have done analysis with the help of boxplot. We can conclude many things from graph like:

1. Demand increases in the fall season. Spring is the season with the lowest demand.
2. The demand for bike has increased from the previous year i.e 2018 -2019.
3. September is the month with highest demand. During December and January the demand is less, reason could be extreme cold.
4. The demand decreases during holidays. The reason could be people prefer to spend time at home..
5. Weekdays have tough competition between them but Thursday, friday, Saturday and Sunday have somewhat more demand.
6. Most bike are shared during cloudy season.

- But the good thing for the business is that demand has increased from previous year.

Q2. Why is it important to use drop_first=True during dummy variable creation?

- Ans: drop_first=True is important to use as it helps in reducing the extra coloumn created during the creation of dummy variable. This reduces the number of dummy variables to N-1 for a variable with N categories.
- ***Example:*** Consider a categorical variable "Color" with three categories: Red, Green, and Blue. Without drop_first=True, we get three dummy variables:
- Color_Pink(1 if Pink, else 0)
- Color_Yellow (1 if Yellow, else 0)
- Color_Blue (1 if Blue, else 0)
- Here, if Color_Pink and Color_Yellow are both 0, then Color_Blue must be 1, introducing multicollinearity.
- With drop_first=True, we might drop the first dummy variable and only use:
- Color_Green (1 if Yellow, else 0)
- Color_Blue (1 if Blue, else 0)

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans: “temp” has the highest correlation with the target variable.

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans: I can validate the assumption of Linear regression based on some points :

1. Firstly, error terms should be distributed normally.
2. There should be less multicollinearity among variables.
3. But linearity should be there among the variables

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

- Ans: The the top 3 features contributing significantly towards explaining the demand of the shared bikes:
  1. Sep
  2. Temp
  3. Winter

General Subjective Questions

Q 1. Explain the linear regression algorithm in detail.

Ans. Linear regression is a fundamental algorithm in statistics and machine learning used for predicting a quantitative response variable based on one or more predictor variables. It assumes a linear relationship between the input(s) (predictor variables) and the output (response variable).

Linear regression can be categorized into two main types:

1. Simple Linear Regression (SLR) for one predictor and
2. Multiple Linear Regression (MLR) for two or more predictors.

This can also be represented mathematically, **Y = mX + c.**

Y stands for dependent variable we are trying to predict. X variable stands for independent variable , which will help in prediction. m is the slope of regression line and c is constant, which is also known as Y-intercept.

Objective

The objective of linear regression is to find the values of the coefficients that minimize the sum of the squared differences (residuals) between the observed values and the values predicted by the model. This method of estimation is known as Ordinary Least Squares (OLS).

Assumptions

Linear regression makes several key assumptions about the data:

1. **Linearity:** There is a linear relationship between the dependent and independent variables.
2. **Independence:** Observations are independent of each other.
3. **Homoscedasticity:** The residuals have constant variance at every level of the independent variable(s).
4. **Normality:** The residuals of the model are normally distributed

Types of Linear Regression

1. **Simple Linear Regression:** Involves a single independent variable to predict the dependent variable. The relationship between the two variables is represented by the equation of a straight line.

   $Y = \beta 0 + \beta 1 X + \epsilon$

2. **Multiple Linear Regression:** Uses two or more independent variables to predict the dependent variable. The relationship is modeled by a hyperplane in a multidimensional space.

   $Y = \beta 0 + \beta 1 X 1 + \beta 2 X 2 + ... + \beta n X n + \epsilon$

2. Explain the Anscombe’s quartet in detail.

Ans: Anscombe's quartet comprises four datasets 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. This quartet was created by the statistician Francis Anscombe in 1973 to demonstrate both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties.

Key Properties :

The four datasets in Anscombe's quartet are constructed to have the following properties in common:

Mean of x: The mean of the x values is approximately the same across the four datasets.

Mean of y: The mean of the y values is approximately the same across the four datasets.

Variance of x: The variance of the x values is the same for the first three datasets and slightly different for the fourth.

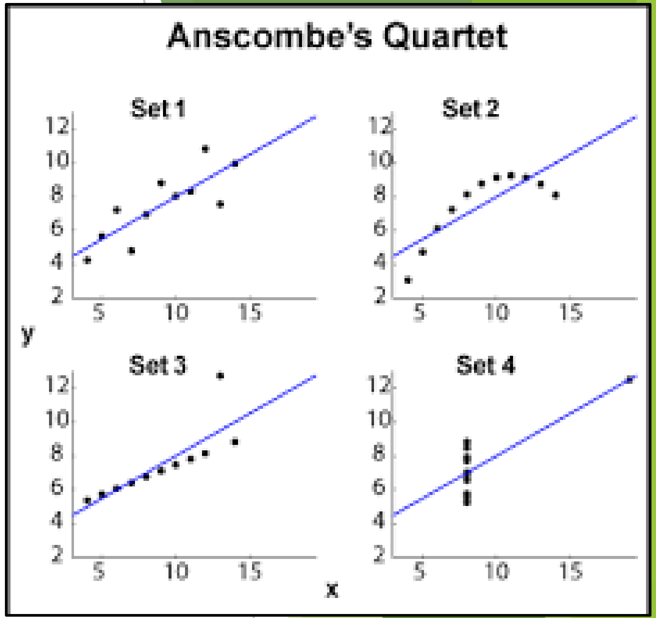
Variance of y: The variance of the y values is approximately the same across the four datasets.

Correlation: The correlation between x and y variables is the same for the first three datasets and still strong for the fourth.

Linear regression line: When a linear regression line (y = mx + b) is fit to each dataset, the slope (m) and intercept (b) are approximately equal across the four datasets.

Coefficient of determination (R^2): The R^2 value, which measures the proportion of the variance for a dependent variable that's explained by an independent variable in a regression model, is the same for all datasets.

Datasets
Dataset I follows a simple linear relationship approximately, corresponding to the intuition of how a scatter plot of two variables with a certain correlation would appear.
Dataset II demonstrates a curve (quadratic relationship) between x and y. Despite this nonlinear relationship, the simple linear regression statistics mirror those of Dataset I.
Dataset III shows a linear relationship similar to Dataset I but with an outlier that affects the slope of the regression line. Without the outlier, the relationship between x and y would be stronger and the slope steeper.
Dataset IV involves a set of x values that are nearly identical, which means the correlation and linear regression model are heavily influenced by a single outlier. Without this outlier, there would be no correlation between x and y.



Types of dataset(downloaded picture)

Implications

Anscombe's quartet is a compelling demonstration of the limitations of basic statistics and the dangers of relying solely on them to analyze data. It underscores the necessity of visual data analysis to detect underlying patterns, relationships, or anomalies that summary statistics alone cannot reveal. This is particularly relevant in the era of big data and complex datasets, where such nuances can have significant implications for data analysis and interpretation.

Conclusion

Anscombe's quartet remains a classic example in statistical education, emphasizing the critical role of data visualization and the caution one must exercise when interpreting statistical summaries. It teaches that statistical analysis is not merely about computation but also about understanding the data's underlying structure and context.

Q3. What is Pearson’s R?

- **Ans.** Pearson's R, also known as Pearson's correlation coefficient (denoted as �*r*), is a measure used in statistics to assess the strength and direction of a linear relationship between two continuous variables. The value of Pearson's R ranges from -1 to 1, where:

- **1** indicates a perfect positive linear relationship,
- **-1** indicates a perfect negative linear relationship, and
- **0** indicates no linear relationship between the variables.

Calculation

Pearson's R can be calculated using the formula:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

where:
•*Xi* and *Yi* are the individual sample points indexed with �*i*,
•¯*X*¯ and ¯*Y*¯ are the mean values of the *X* and *Y* samples, respectively,
•*n* is the number of pairs of scores.

Q3. What is Pearson’s R?

- Ans. Pearson's R, also known as Pearson's correlation coefficient (denoted as *r*), is a measure used in statistics to assess the strength and direction of a linear relationship between two continuous variables. The value of Pearson's R ranges from -1 to 1, where:

- **1** indicates a perfect positive linear relationship,
- **-1** indicates a perfect negative linear relationship, and
- **0** indicates no linear relationship between the variables.

Calculation

Pearson's R can be calculated using the formula:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

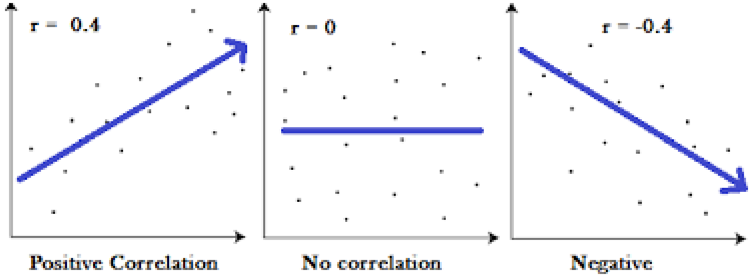
where:
•*Xi* and *Yi* are the individual sample points indexed with �*i*,
•¯*X*¯ and ¯*Y*¯ are the mean values of the *X* and *Y* samples, respectively,
•*n* is the number of pairs of scores.

As mentioned above, the Pearson correlation coefficient, r, can take a range of values from -1 to +1. A value of 0 will show that there is no relationship between two variables.

A value greater than 0, shows positive relationship. Positive relationship means, that if the value of one variable increases then the value of other variable increases.

As for the value 0, the relationship between two variables is vice- versa, that is , if the value of one variable increases then the value of other variable decreases.

Here is a google image to show the 3 different relationship:



Limitations

There are some limitations also:

- It assumes that both variables are normally distributed.
- It is sensitive to outliers.
- It only measures linear relationships.

Q4.What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans: Scaling : It is a method used in data preprocessing to adjust the range of variable values. This technique is essential in many machine learning algorithms that are sensitive to the magnitude of values and might perform poorly if the individual features do not more or less look like standard normally distributed data (e.g., Gaussian with 0 mean and unit variance). Examples include gradient descent-based algorithms, algorithms that use distance measures like k-nearest neighbors (KNN) and support vector machines (SVM), and algorithms that use regularization.

Why scaling is performed?

- **Performance Improvement:** Many machine learning algorithms perform better or converge faster when features are on a similar scale.
- **Consistency of Units:** Features measured in different units do not contribute equally to the model fitting & function approximation and might end up creating a bias.
- **Numerical Stability:** Some algorithms, especially those involving matrix operations, can become numerically unstable if the scales of features are significantly different.

Normalized scaling VS standardized scaling

Normalized scaling	Standardized scaling
Normalization rescales the values into a range of [0, 1] or sometimes [-1, 1].	Standardization is not bounded to any range.
The main feature of this scale are Minimum and Maximum.	In standardized scaling, Mean and Standard deviation are used.
Formula , $\boldsymbol{x' = x - \min(x)/\max(x) - \min(x)}$**,** *where min means minimum, max means maximum, x is the original value and x′ is the normalized value.*	Formula, $\boldsymbol{z = x - \mu/\sigma}$ *where z is standardized value, x is original value, μ is mean of the feature column and σ is for standard deviation.*
Normalization is useful when you need to bound your values between two numbers, say, -1 and 1.	**Standardization** is useful when your data has varying scales and the algorithm you are using makes assumptions about your data.

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

Ans: Sometimes, the VIF can be infinite (or extremely large values approaching infinity). This situation occurs when the *Ri*^{2} for a variable is exactly 1 (or extremely close to 1), which means the predictor is perfectly (or almost perfectly) linearly correlated with other predictor(s) in the model. In mathematical terms, when *Ri*^{2}=1, the denominator in the VIF formula becomes 0, leading to an infinite VIF:

VIFi =1/ 1−*Ri*^{2} = 1/1-1= 1/0 =∞

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans: A Q-Q (quantile-quantile) plot is a graphical tool that helps to determine if a set of data plausibly came from some theoretical distribution. The distribution can be normal, uniform or exponential. In other words of linear regression analysis, Q-Q plots are primarily used to evaluate the normality of the residuals.

This plot compares the quantiles of the dataset to the qualities of a theoretical distribution in a plot to check if the data follows certain distribution or not.

Use and Importance in Linear Regression

1. **Examining the Distribution of Residuals:**
   1. One of the assumptions of linear regression is that the residuals (the differences between observed and predicted values) are normally distributed. If the Q-Q plot makes approximately straight line, then that means the residuals are distributed normally.
2. **Diagnosing Outliers:**
   1. Q-Q plots can also help in identifying outliers in the residuals. Points that deviate significantly from the straight line in a Q-Q plot may indicate outliers.
3. **Analyzing Homoscedasticity :**
   1. Although not its primary use, a Q-Q plot can sometimes give insights into whether the variance of the residuals is constant (homoscedasticity), another assumption of linear regression. Significant deviations from a straight line may indicate issues with homoscedasticity.
4. **Model Evaluation and Improvement:**
   1. By assessing the normality of residuals and identifying potential outliers or heteroscedasticity, Q-Q plots can inform necessary model adjustments. For instance, if residuals are not normally distributed, transformations of variables or a different modeling approach may be warranted.

THANK YOU