Linear Regression

Assignment questions & Answers

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)

Based on my analysis of categorical variables against the count of bike rentals, the below are my findings:

- The median of bike rental count is more on a non-holiday compared to holiday days.
- The count of bike rentals has increased from the year 2018 to 2019 in a very significant number implying the popularity of the bike rentals over the years.
- A clear weather contributes more to the count of bike rentals as people preferred to use other means of transport when the weather is not good.
- Count of bike rentals is high during falls , then summer comes as second and followed by winter and spring.
- January, February, March, November and December shows a dip in the bike rentals compared to other months, may be due to extreme weather conditions.
- Medians of both working and non working days for bike rentals remain almost similar.

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

- While creating dummy variables for a categorical variable with n levels we will create n-1 dummy variables by using drop_first= True. This is because if we have 10 columns , and we create 9 columns and if all of them have 0 value, it is obvious that the 10th column is having value 1. Hence we can reduce the number of columns so that it will reduce the multicollinearity and redundancy in the model hence giving the best fit model.

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

- temp and atemp has the highest correlation of 0.63 with the count of bike rentals from the analysis of pairplot.

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

We can validate the assumptions of Linear Regression after building the model using the following methods:

- Linear relation : We can check the linearity between the target variable and the independent variables by plotting a pairplot between them.
- Error term has mean zero: We can calculate the error or residual by taking the difference of actual and predicted values of y and then calculate the mean.
- Homoscedasticity check: Homoscedasticity ,ie the variance of error terms are equal or almost same across the regression line can be verified by plotting between the error terms and predicted value and see if it is not following any particular pattern or not.
- Normal distribution of error terms: This can be checked by plotting a histogram and see the distribution.
- Multicollinearity: If there is correlation between independent variable, it can be verified using correlation plot or by calculating the VIF.

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

By looking at the final model and after analysing the coefficients of the predictor variables,

- Temperature (temp) contributes significantly to the demand of shared bikes as it has a positive coefficient of 0.5527.
- The year variable also contributes significantly as we have seen that the demand for bike rentals increases over the year. The year 2019 has a positive coefficient of 0.2332.
- The weathersit_rain variable is negatively correlated with the demand for bike rentals as the weather becomes unfavourable during rainy season , the demand for bike rentals declines as expected. It has a negative coefficient of -.2785.

General Subjective Questions

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

Linear Regression is a supervised model of machine learning which explains linear relationship between a dependent or output variable and independent or predictor variables. The output of a Linear regression will be continuous variables.

Linear Regression is of 2 types:

1.Simple Linear Regression

It establishes a linear relationship between target variable and 1 independent variable. The equation of Linear relationship is given by y=B0+B1x, where y is the dependent variable,B0 is the intercept and B1 is the

Slope or coefficient.

2. Multiple Linear Regression

This model represents relationship between target variable and multiple predictor variables.The equation is given by y=B0+B1x1+B2x2+…+Bnxn, Where B1,B2,Bn etc are the coefficients of each predictor variables.

The model fits the best fit line by reducing the sum of squares of residuals, where residual is the difference between the actual data points and the predicted data points. R squared is an important term in analysing the

significance of the model as it implies how well the variance in the data is explained by that particular model.

R2= 1-RSS/TSS

Where RSS – Residual sum of squares and TSS – Total sum of squares.

The below are the assumptions of Linear Regression model:
Linear relationship between X and y.
Error terms are normally distributed with mean 0.
Error terms are independent of each other
Error terms have constant variance.

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

Anscombe's quartet consists of four datasets that have similar statistical information (mean,variance,R2 etc) But when plotted shows different graphical representation. Each dataset consist of 11 (x,y) data points. They were constructed in 1973 by statistician Francis Anscombe to demonstrate both the importance of graphing data when analyzing it, and the effect of outliers and other influential observations on statistical properties.

Key Points of Anscombe's Quartet:

1. Identical Statistical Properties:

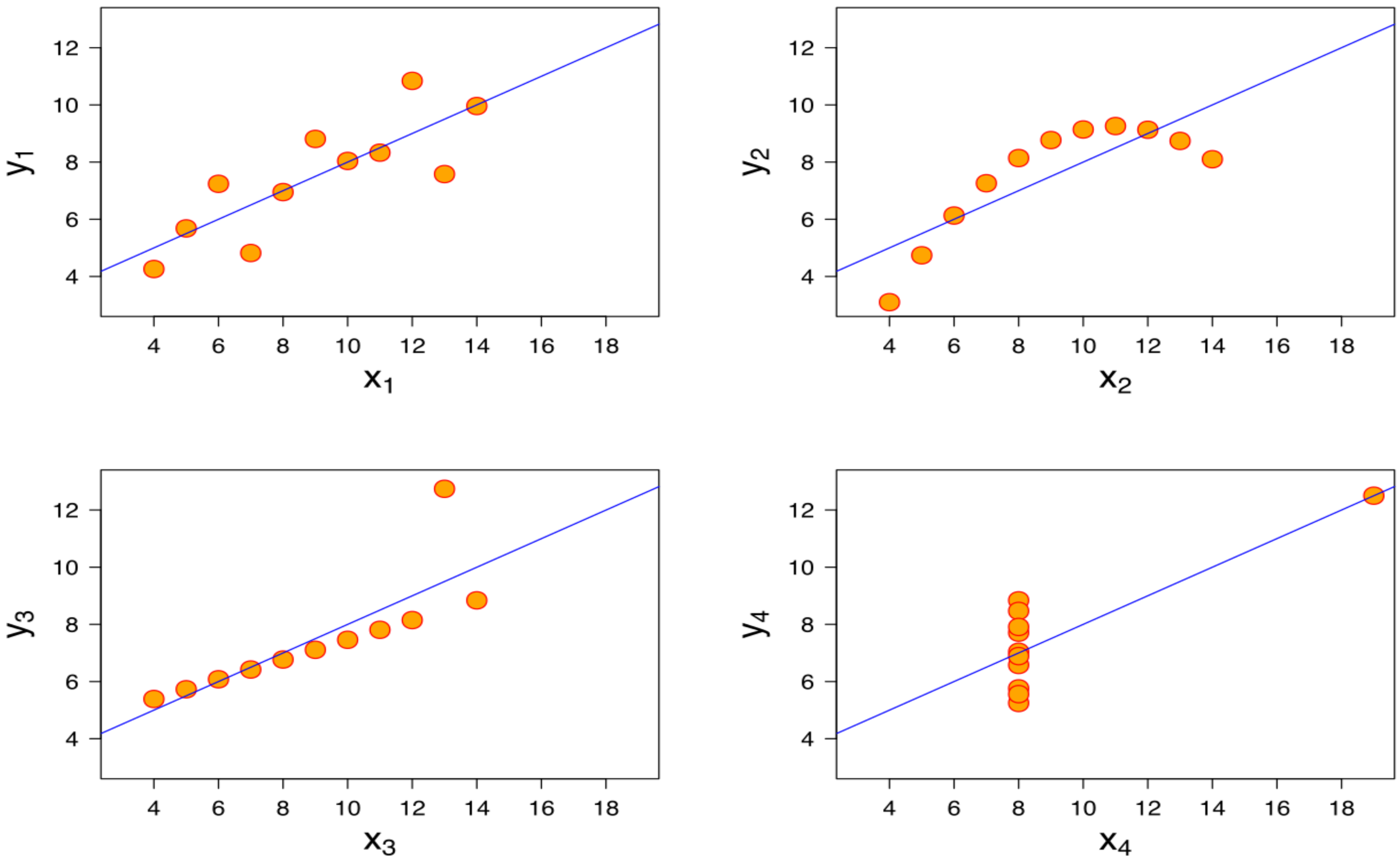
- All four datasets have the same mean for the x and y values.
- They have the same variance for x and y.
- Each dataset has the same correlation coefficient between x and y.
- The linear regression line (y = mx + c) is nearly the same for all four datasets.

2. Visual Differences:

    - Dataset 1: Shows a typical linear relationship between x and y, which would be expected based on the regression analysis.
    - Dataset 2: The data is more curvilinear, indicating a non-linear relationship despite the linear regression line.
    - Dataset 3: Contains an outlier, which heavily influences the regression line, leading to misleading interpretations if only the statistics are considered.
    - Dataset 4: All x-values are the same except for one, creating a vertical line. The single differing point (an outlier) forces the regression line to fit in a misleading way.

Importance of Anscombe's Quartet:

- Visual Exploration: The quartet illustrates the crucial role of visualizing data before jumping to conclusions based on summary statistics. By plotting the data, one can identify patterns, outliers, or structures that simple statistics might miss.
- Misleading Conclusions: If one relies solely on statistical summaries without visual inspection, one might draw incorrect or oversimplified conclusions about the data.
- Teaching Tool: Anscombe's Quartet is widely used in statistics education to teach the importance of graphical analysis and to caution against the over-reliance on summary statistics.



y_1
12
10
8
6
4
4
6
8
10
12
14
16
18
x_1
y_2
12
10
8
6
4
4
6
8
10
12
14
16
18
x_2
y_3
12
10
8
6
4
4
6
8
10
12
14
16
18
x_3
y_4
12
10
8
6
4
4
6
8
10
12
14
16
18
x_4

3. What is Pearson’s R? (3 marks)

Pearson’s R also knows as Pearson’s correlation coefficient is a measure of linear relation between predictor and target variable. It quantifies the strength and direction of the linear relationship between these two variables.

The value of Pearson’s R ranges between -1 and 1. ‘1’ indicates a prefect positive relation between 2 variables and ‘-1’ indicates a perfect negative relation between them and 0 indicates no relation between the variables.

Pearson’s correlation coefficient is calculated by taking the ratio of covariance of two variables to the product of their standard deviations.

R= COV(X,Y)/sigma X *sigma Y

In the context of linear regression, Pearson’s R represents how well the dependent variable depends on the predictor variable. A high value of R indicates that the independent variable is a good predictor for that target or dependent variable. But correlation doesn’t always implies causation is a fact that we should keep in mind.

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

Scaling in Machine learning refers to the process of adjusting the range of features used in the model to a common range so that they can be compared on a common scale. It is particularly used in algorithms where the distance between the data points really matters like Gradient descent optimization technique.

Scaling is performed to:

Improve Model Performance: Algorithms like gradient descent converge faster when features are scaled because the optimization process is more efficient when the features are within the same range.

Ensure Equal Contribution: Scaling ensures that all features contribute equally to the model. Without scaling, features with larger ranges could disproportionately influence the model's predictions.

Enhance Interpretability: It makes it easier to interpret the results of the model, particularly in distance-based algorithms.

Difference Between Normalized Scaling and Standardized Scaling :

Normalized Scaling:

Normalization scales the data to a fixed range, typically [0, 1] or [-1, 1]. It adjusts the values to be within a specific range without affecting the relative differences between data points.

X= X-Xmin/Xmax-Xmin

Normalization is useful when you need to bound the values within a certain range, for instance, in algorithms that need bounded inputs like neural networks.

Standardized Scaling:

Standardization scales the data based on the mean and standard deviation, resulting in features with a mean of 0 and a standard deviation of 1.

Formula:

$X = X - Xmean/std(X)$

Standardization is preferred when features have different distributions or when the model assumes that the data is normally distributed (e.g., in linear regression or logistic regression).

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

The Variance inflation factor (VIF) is a measure to detect multicollinearity in a regression model. It quantifies how much the variance of a regression coefficient is inflated due to the presence of multicollinearity among the independent variables.

A VIF becomes infinite when perfect multicollinearity exists, that is one of the independent variable is a perfect linear combination of one or more independent variables.

$VIF = 1/1 - R^2$

An infinite VIF occurs if the correlation between an independent variable and combination of other independent variables is perfect. In this case, the model cannot distinguish between the perfectly correlated variables, and the variance of the affected variable's coefficient is infinitely inflated. Mathematically, this happens when the determinant of the matrix (X'X) used to compute VIF is zero, leading to a division by zero.

- Implication and Solution:

Implication: An infinite VIF indicates severe multicollinearity, which can destabilize the regression coefficients, making them highly sensitive to changes in the model. This undermines the reliability of the model and can lead to incorrect interpretations.

Solution: To address infinite VIF, you may need to:

- Remove one of the perfectly correlated variables from the model.
- Combine the correlated variables into a single feature through techniques like Principal Component Analysis (PCA).
- Rethink the model design to avoid including redundant predictors.

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

A quantile-quantile (Q-Q) plot is a graphical tool which is used to compare the quantiles of two distributions to determine if a set of data is likely to come from a theoretical distribution, such as a normal, exponential, or uniform distribution.

Q-Q plots can also be used to check assumptions in linear regression models, such as whether the model's residuals are normally distributed.

In Q-Q plots the sample data is plotted on Y axis and X axis depends on what sample data it is compared to.

If both the distributions are same, the points on the plot fall on a straight line with equation y=x.

Q-Q plots can be used to determine if 2 datasets come from similar distribution or not. This is useful in Linear regression if training and test data are received separately.

The advantages of Q-Q plot are as follows:

- They can be used to measure the shift in location, scale ,outliers etc
- They provide graphical assessment of goodness of fit rather than relying on statistical information.
- They don’t need the sample sizes to be equal for comparison.