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 final model outcome** we can see below categorical variables have significant impact

a. Seasonal Impact - Season_2 and Season_4 (Summar and winter) have higher coefficient, 0.07 and 0.13, meaning season has a positive impact on cycle rentals
b. Weather impact – Weather situation 3 (Snow, rain, thunderstorm) has a negative impact, coefficient is -0.25
c. Temporal impact – Month_9 (September) and year has a positive impact on rental business, coefficients are 0.22 and 0.08

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

Answer

When we create dummy variables, a false value on rest of the variables itself signifies the first one. For example, season_1, season_2, season_3, season_4 are created, If season_2, season_3, season_4 are all 0, it means it is Season_1. So if we do not drop Season_1, it will create high collinearity.

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

Answer

The variable “registered” has the highest correlation 0.945 with target variable count.

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

Answer

1. **Multicollinearity Check (VIF Analysis) -** All VIF values greater than 5 were dropped
2. **Statistical Significance (P-values) -** All variables with p values higher than 0.05 were dropped.
3. **Normality of residuals –** Residuals are normally distributed
4. **Linearity & Homoscedasticity –** Actual and predicted values are having a linear relationship for both Train and Test data sets.
5. **R Square –** R2 score calculated on both test and train data set as 0.82 and 0.79. These values confirm there is no overfit

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. Temperature – Variable “temp” has the highest coefficient of +0.546 and highest t value 28
2. Year – Variable “year” has second highest coefficient of +0.229 and t value of above 26
3. Weather - Variable “weather_situation3” has highest negative correlation of -0.25

General Subjective Questions

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

Answer

The linear regression algorithm is a supervised machine learning method used to model the linear relationship between a dependent variable and one or more independent variables. Its objective is to find a "best-fit" line to predict the dependent variable for new data by minimizing the difference between predicted and actual values

There are 2 types of linear regressions

a. Simple Linear Regression - $y=\beta 0+\beta 1x+\epsilon$

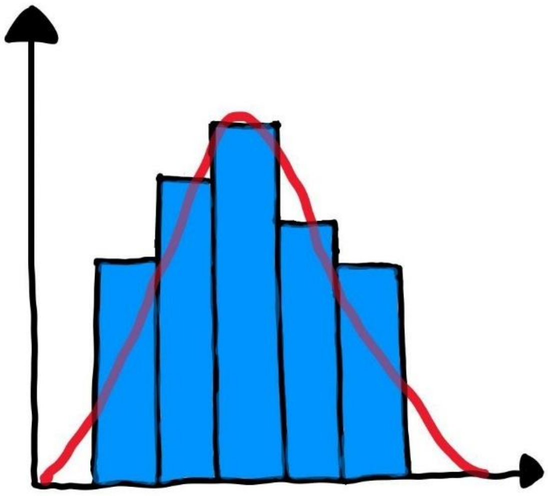
Where y is predicted dependent variable, $\beta0$ is intercept, $\beta1$ is coefficient (slope) , x is independent variable and ϵ is error term

b. Multiple Linear Regression - $y=\beta0+\beta1x1+\beta2x2+\cdots+\beta nxn+\epsilon$
Where y is predicted dependent variable, $\beta0$ is intercept, $\beta1$ and $\beta2$ etc are coefficients and x1,x2 are independent variables and ϵ is error term

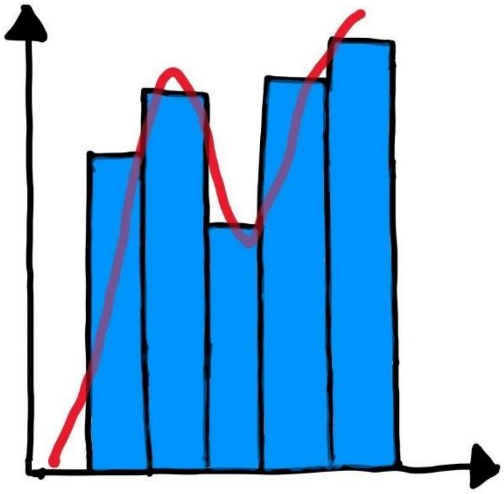
Apart from the formulation, there are some other aspects that remain the same:

1. The model now fits a hyperplane instead of a line
2. Coefficients are still obtained by minimising the sum of squared errors, the least squares criteria
3. For inference, the assumptions from simple linear regression still hold - zero-mean, independent and normally distributed error terms with constant variance

Assumptions - Reliable linear regression results depend on several assumptions about the data: a linear relationship, independent errors, constant variance of errors (homoscedasticity), normally distributed errors, and no multicollinearity among independent variables

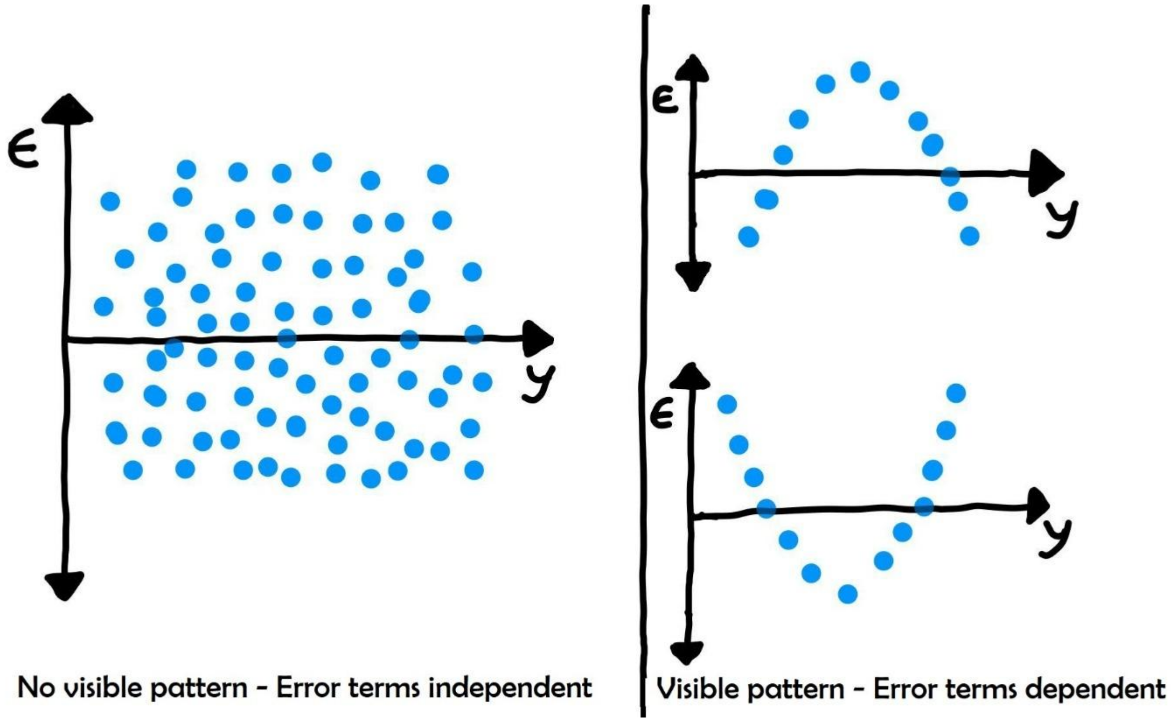


Error terms normally distributed



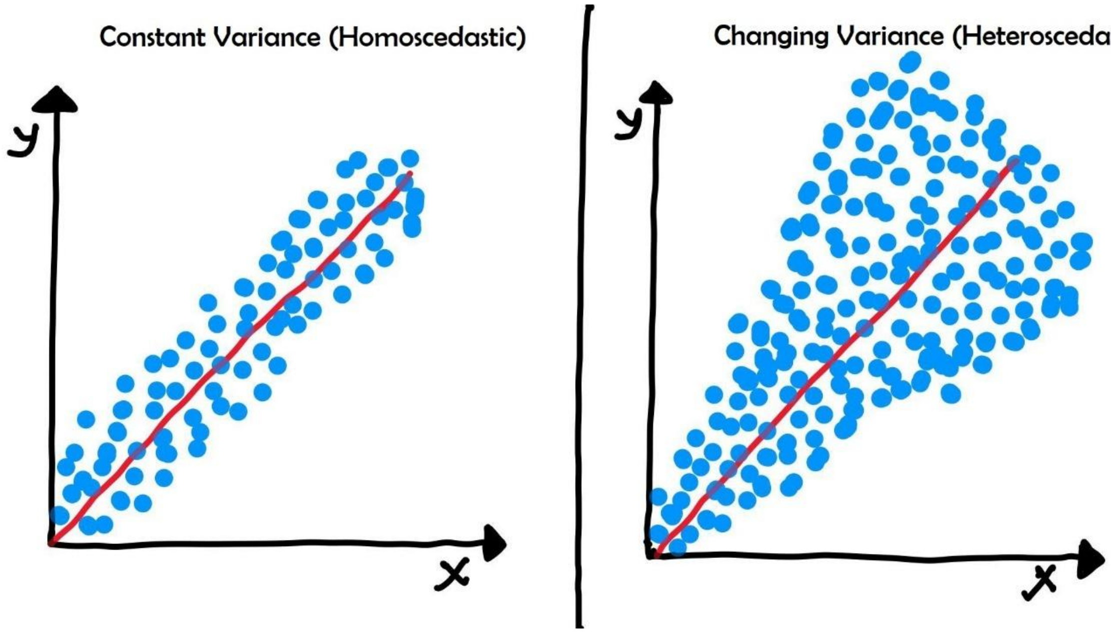
Error terms not normally distributed

The error terms should not be dependent on one another (like in a time-series data wherein the next value is dependent on the previous one).



Error terms have *constant variance* (homoscedasticity):

- The variance should not increase (or decrease) as the error values change.
- Also, the variance should not follow any pattern as the error terms change.



How to measure a linear regression model -

- **T-statistic:** You would look at the t-statistic for "hours studied." A high t-statistic would tell you that studying more hours has a statistically significant effect on the final exam score.
- **R-squared:** You would look at the R-squared value to see how much of the variation in exam scores can be explained by *both* attendance and hours studied. An R-squared of 0.70 means that 70% of the variation in test scores can be attributed to your model.
- **F-statistic:** The F-statistic would tell you if your model, which includes both "attendance" and "hours studied," is a better predictor of the exam score than a model that just uses the average score.

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

Answer

Anscombe's quartet is a famous collection of four datasets, created by statistician Francis Anscombe in 1973, which have nearly identical basic summary statistics (mean, variance, correlation, and linear regression line) but appear drastically different when plotted graphically. Its primary purpose is to demonstrate the vital importance of data visualization and the limitations of relying solely on numerical summaries during data analysis

Description of the Four Datasets

Despite the identical statistics, each dataset reveals a unique underlying pattern when visualized:

Dataset I: This dataset shows a roughly linear relationship between x and y with some natural variability. It is the only one of the four for which a linear regression model is appropriate and fits well with the assumptions of the algorithm.

Dataset II: This dataset exhibits a clear non-linear, parabolic relationship. A linear regression model applied to this data is misleading; a curved model (e.g., quadratic) would be much more appropriate.

Dataset III: This dataset appears to have a strong linear relationship for most data points, with the exception of a single, highly influential outlier. This one outlier is enough to skew the regression line and the correlation coefficient, making the summary statistics unrepresentative of the main body of the data.

Dataset IV: In this dataset, all data points except one share the exact same x-value, with varying y-values. A single high-leverage point with a different x-value forces the correlation and regression line to be what they are. Without that single point, there would be no discernible relationship between x and y at all.

Significance in Data Analysis

The "quartet" serves as a classic cautionary tale in statistics and data science, highlighting several key lessons:

- **Visualization is crucial:** Summary statistics can hide the true nature of data distributions, outliers, and underlying patterns. Plotting data is an essential step in exploratory data analysis (EDA).
- **Model validation is necessary:** Relying on statistical metrics like $R2$ alone can lead to the inappropriate use of models (e.g., applying linear regression to non-linear data).
- Assumptions must be checked: The datasets emphasize the importance of verifying the assumptions of a statistical model (such as linearity and the absence of influential outliers) before trusting its numerical outputs.

Anscombe's quartet effectively counters the "misguided belief" that "numerical calculations are exact, but graphs are rough". Both are necessary for robust and meaningful data interpretation

3. What is Pearson’s R? (3 marks)

Answer

Pearson's R, also known as the **Pearson correlation coefficient** (or the Pearson product-moment correlation coefficient, PPMCC), is a widely used statistical measure that quantifies the **strength and direction of the linear relationship between two continuous variables**. The coefficient is denoted by 'r' for a sample and the Greek letter 'ρ' (rho) for a population.

The value of Pearson's R always ranges from **-1 to +1**

- **Positive Correlation (r > 0):** Indicates a positive relationship, meaning that as one variable increases, the other variable tends to increase as well. A value of **+1** signifies a perfect positive linear correlation.

- **Negative Correlation (r < 0):** Indicates a negative, or inverse, relationship, meaning that as one variable increases, the other variable tends to decrease. A value of **-1** signifies a perfect negative linear correlation
- **No Correlation (r = 0):** Indicates that there is no linear relationship between the two variables. Changes in one variable do not predict changes in the other

The **absolute value** of *r* determines the strength of the relationship: the closer the absolute value is to 1, the stronger the linear correlation; the closer it is to 0, the weaker the correlation

Key Assumptions

Using Pearson's R effectively relies on several assumptions about the data:

- **Quantitative Variables:** Both variables must be continuous (interval or ratio level).
- **Linearity:** The underlying relationship between the variables should be linear, meaning it can be reasonably described by a straight line.
- **No Outliers:** Outliers can significantly skew the result of the coefficient and should be examined.
- **Normal Distribution:** Both variables are assumed to be approximately normally distributed.

Important Caveat: Correlation Does Not Imply Causation

A crucial point in interpreting Pearson's R is that a strong correlation between two variables does not necessarily mean one causes the other to change. It only indicates an association.

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

Answer

Scaling is a data preprocessing technique that transforms numerical features to a common range or distribution. It adjusts the magnitude of values without changing the relationships between data points.

Why is Scaling Performed?

Scaling is performed for several critical reasons, primarily to improve the performance, speed, and accuracy of machine learning algorithms. It helps prevent features with larger values from dominating algorithms that are sensitive to magnitude, such as those using gradient descent or distance calculations. Scaling also aids in faster convergence for gradient descent-based algorithms and helps meet the assumptions of certain models

Normalized Scaling (Min-Max Scaling) - Rescales data to a fixed range, typically [0, 1]. This scaling output is bounded (usually 0 to 1). Normalized scaling does not change shape of original distribution. Useful for algorithms requiring bounded inputs or when data distribution is not Gaussian.

Standardized Scaling (Z-score Scaling) - Transforms data to have a mean of 0 and a standard deviation of 1. This scaling output is unbounded and less sensitive. This centres data but preserves original shape/distribution. Useful when data has a Gaussian distribution or when algorithms assume normal distribution.

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

Answer

VIF formula is as below

VIF = 1 / 1- R^2

Mathematically speaking, In the above formula if R^2 becomes 1, VIF will become infinity. This situation will come if a model has a variable which is having perfect multicollinearity.

Infinite VIF values are often the result of specification errors in the model, such as:

- **Including the same variable twice:** For example, including a car's weight in both kilograms and pounds as separate predictors in the same model, since one is a perfect linear transformation of the other.
- **Including a variable that is a direct sum or difference of others:** For example, if a model includes "Count" and also separate variables for "Registered" and "Casual", where " Registered " + " Casual " = " Count ". This example is present in the “Bikes” dataset which was provided in assignment
- **Dummy variable trap:** When creating dummy variables for a categorical variable (e.g., for 'Gender': Male and Female), including both dummy variables and an intercept in the model results in perfect multicollinearity.

- Having more predictors than observations (data points): In cases where the number of variables exceeds the number of samples, it's possible for the model to perfectly "explain" the variance of one predictor with the others, leading to infinite VIFs for all variables

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

Answer

Quantile-Quantile (Q-Q) plot is a graphical tool used in statistics to compare two probability distributions by plotting their quantiles against each other. In data analysis, it is most frequently used to visually assess whether a sample dataset follows a specific theoretical distribution, typically the normal distribution

How a Q-Q Plot Works

The plot works by:

1. **Ordering the Data:** The observed data (e.g., residuals from a linear regression) are sorted in ascending order.
2. **Calculating Quantiles:** The position (quantile) of each data point is calculated.
3. **Generating Theoretical Quantiles:** The corresponding quantiles from a specified theoretical distribution (e.g., standard normal distribution, with a mean of 0 and standard deviation of 1) are generated.
4. **Plotting and Comparison:** The observed quantiles are plotted against the theoretical quantiles. A **45-degree reference line** is added; if the two distributions match, the points should fall approximately along this straight line

Use and Importance in Linear Regression

In linear regression, a key assumption is that the **model residuals are normally distributed** with a mean of zero and constant variance. The Q-Q plot is the primary visual diagnostic tool for verifying this assumption.

- **Assumption Validation:** The main use of the Q-Q plot is to check if this normality assumption holds true. If the residuals are normally distributed, the points on the Q-Q plot will form a straight line, parallel to the reference line.
- **Diagnosing Deviations:** The pattern of deviations from the line helps identify specific problems with the model or data distribution:

  - **S-shaped curves:** Suggest that the data has heavier or lighter tails than the normal distribution (kurtosis issues).
  - **Concave or Convex curves:** Indicate that the data is skewed (either left or right).
  - **Points far from the line at the ends:** Highlight potential outliers or extreme values in the data.

- **Guiding Transformations**: If the Q-Q plot shows significant deviations, it suggests that the model assumptions are violated. This may prompt data transformations (e.g., logarithmic or square root) of the variables to better fit the model's assumptions or using alternative statistical methods that do not require normality.
- **Complementary Tool:** While formal statistical tests (like the Shapiro-Wilk test) provide numerical p-values, Q-Q plots offer a visual and intuitive understanding of how and where the data deviates from the expected distribution, which is often more insightful than a single test statistic**.**