Linear Regression Case study

Assignment-**based Subjective** Question

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Answer:

a. Season Count: Spring and summer have reported high usage of bikes.

b. Year count: demonstrates that over a period usage of bikes has increased.

c. Month count: demonstrates and re-affirm the observation of season that month between May – October when weather is clean and warm usage of bikes is higher.

d. Holiday count: it demonstrates that usage of bikes increases and decrease during holidays.

This is further re-affirmed during EDA where it is observed that usage of bike among casual users on Saturday and Sunday is significantly high.

Q2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

Answer: Dummy variables are created from single dataset and hence will have interdependency resulting in multicollinearity causing impact in linear regression. If we retain all variable in regression model, then intercept will have perfect multicollinearity. To prevent this, we drop first variable in order to simply the model without losing any information and coefficient created from linear regression are interpretable and statistically valid

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Answer: highest correlation observed from pair-plot is between ‘cnt’ vs ‘atemp’

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Answer: Assumption of linear regression is validated by residual analysis by evaluating "error term" in which we evaluate if error term is normally distributed. This is done to ensure homoscedasticity in data variable i.e. residual have constant variation at every level of predictor variable

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Answer:

a. Impact of weather conditions on boombikes revenue: variables such as temperature, humidity and wind speed have strong influence on boombikes customers.

b. Impact of time: Attributes associated with time example: day, weekday, working day, month, season etc. have significant impact on usage of bike by customer of boombikes. Time and weather to-gether have distinct impact on company business.

c. Impact of events: From the data it is visible that usage of bikes by casual users is highest during holidays. While registered users provide steady revenue during working and weekdays, causal users bring good influx of revenue during holiday.

General Subjective Questions

Q1. Explain the linear regression algorithm in detail? (4 marks)

Answer: Linear regression is a statistical method used to model the relationship between a dependent variable (what you want to predict) and one or more

independent variables (the predictors). The goal is to find the best-fitting straight line through the data points.

The Basic Idea

Imagine you have a dataset of houses, where you know the number of bedrooms and the price of each house. You want to predict the price of a new house based on the number of bedrooms it has. Linear regression helps you find a line that best fits the data points, so you can use this line to make predictions

The Equation

The equation of a simple linear regression line is:

$y=b0+b1 \cdot x$

- **(y)**: The dependent variable (e.g., house price).
- **(x)**: The independent variable (e.g., number of bedrooms).
- **(b_0)**: The intercept (the value of (y) when (x) is 0).
- **(b_1)**: The slope (how much (y) changes for a one-unit change in (x)).

2. Types of Linear Regression

- **Simple Linear Regression**: Involves one independent variable.
- **Multiple Linear Regression**: Involves more than one independent variable.

3. Assumptions of Linear Regression

For linear regression to provide reliable results, certain assumptions must be met:

- **Linearity**: The relationship between the independent and dependent variables should be linear.
- **Independence**: Observations should be independent of each other.
- **Homoscedasticity**: The residuals (errors) should have constant variance at every level of (x).
- **Normality**: The residuals should be approximately normally distributed.

4. Cost Function

The cost function measures how well the regression line fits the data. The most common cost function used in linear regression is the **Mean Squared Error (MSE)**, which is calculated as:

MSE=n1i=1∑n(yi−y^i)2

where:

- (n) is the number of observations.
- (y_i) is the actual value.
- (\hat{y}_i) is the predicted value.

5. Gradient Descent

To minimize the cost function and find the best-fitting line, we use an optimization algorithm called **Gradient Descent**. It iteratively adjusts the parameters (b_0) and (b_1) to reduce the MSE.

6. Evaluation Metrics

To evaluate the performance of a linear regression model, we use metrics such as:

- **R-squared**: Indicates the proportion of the variance in the dependent variable that is predictable from the independent variable(s).
- **Adjusted R-squared**: Adjusts the R-squared value based on the number of predictors in the model.
- **Root Mean Squared Error (RMSE)**: The square root of the MSE, providing a measure of the average error magnitude.

7. Applications

Linear regression is widely used in various fields, including:

- **Economics**: Predicting economic indicators.
- **Healthcare**: Estimating patient outcomes.
- **Marketing**: Forecasting sales and customer behavior.

8. Advantages and Disadvantages

- **Advantages**:

    - Simple to understand and implement.
    - Provides interpretable results.
    - Computationally efficient.
- **Disadvantages**:
    - Assumes a linear relationship between variables.
    - Sensitive to outliers.
    - May not perform well with complex, non-linear data.

Q2. Explain the Anscombe's quartet in detail? (3 marks)

Answer: Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics, yet appear very different when graphed. It was created by the statistician Francis Anscombe in 1973 to demonstrate the importance of graphing data before analysing it and to show the effect of outliers and other influential observations on statistical properties.

It's called a "quartet" because it consists of four datasets. The term "quartet" typically refers to a group of four, often used in contexts like music (a group of four musicians) or literature (a series of four related works). In this case, Anscombe's quartet refers to the four distinct datasets that share similar statistical properties but differ significantly when visualized. This naming highlights the importance of considering multiple perspectives when analysing data.

Key Characteristics of Anscombe's Quartet

Each of the four datasets in Anscombe's quartet has the following identical statistical properties:

- **Mean of x**: 9
- **Mean of y**: 7.5
- **Variance of x**: 11
- **Variance of y**: 4.125
- **Correlation between x and y**: 0.816

- **Linear regression line**: (y = 3 + 0.5x)
- **Coefficient of determination (R^2)**: 0.67
- For all four datasets:

Property	Value	Accuracy
Mean of *x*	9	exact
Sample variance of *x*: s_x^2	11	exact
Mean of *y*	7.50	to 2 decimal places
Sample variance of *y*: s_y^2	4.125	±0.003
Correlation between *x* and *y*	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression:	0.67	to 2 decimal places

Despite these similarities, the datasets look very different when plotted on a graph:

1. **Dataset I**: Appears to be a simple linear relationship, where (y) could be modelled as Gaussian with a mean linearly dependent on (x).
2. **Dataset II**: Shows a clear relationship between the variables, but it is not linear. The Pearson correlation coefficient is not appropriate here.
3. **Dataset III**: The relationship is linear, but the regression line is affected by an outlier, which lowers the correlation coefficient.
4. **Dataset IV**: Contains a high-leverage point that produces a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

Importance of Anscombe's Quartet

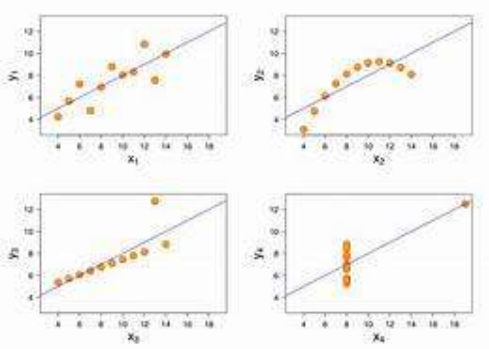
Anscombe’s quartet illustrates several important lessons in data analysis:

- **Graphing Data**: Always visualize your data before analysing it. Graphs can reveal patterns, relationships, and anomalies that simple statistics might miss.
- **Effect of Outliers**: Outliers can significantly affect statistical properties and the results of regression analysis.
- **Misleading Statistics**: Identical statistical properties do not guarantee similar data distributions. Relying solely on summary statistics can be misleading.

Visual Representation

Here is a visual representation of Anscombe’s quartet:

! Anscombe’s Quartet



The datasets are as follows. The *x* values are the same for the first three datasets

Anscombe's quartet

Dataset I		Dataset II		Dataset III		Dataset IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Conclusion

Anscombe's quartet is a powerful reminder of the importance of data visualization in statistical analysis. By plotting data, we can gain insights that are not apparent from summary statistics alone, leading to more accurate and meaningful interpretations.

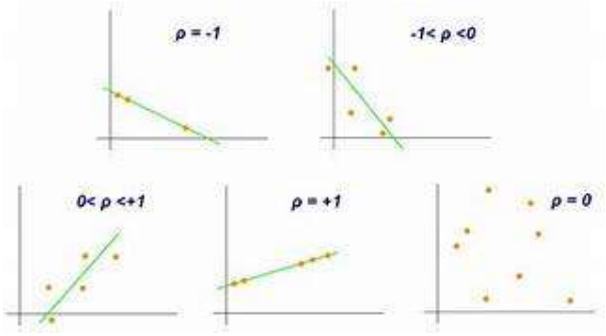
Q3. What is Pearson's R? (3 marks)

Answer: Pearson's R, also known as the Pearson correlation coefficient, is a statistic that measures the linear relationship between two variables. It 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.

The formula for Pearson's R is:

- r=[n(∑x2) −(∑x)2] [n(∑y2) −(∑y)2]n(∑xy) −(∑x) (∑y)
- Here, (n) is the sample size, (x) and (y) are the variables being compared, and (\sum) denotes the summation of the values1.
- Pearson's R is widely used in statistics to determine the strength and direction of the linear relationship between two continuous variables.
-



The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

Pearson correlation coefficient (r)	Correlation type	Interpretation	Example
Between 0 and 1	Positive correlation	When one variable changes, the other variable changes in the same direction.	Baby length & weight: The longer the baby, the heavier their weight.
0	No correlation	There is no relationship between the variables.	Car price & width of windshield wipers: The price of a car is not related to the width of its windshield wipers
Between 0 and –1	Negative correlation	When one variable changes, the other variable changes in the opposite direcIon	Elevation & air pressure: The higher the elevation, the lower the air pressure.

Pearson's correlation coefficient, while useful, has several limitations:

1. **Assumption of Linearity**: It assumes a linear relationship between the variables. If the relationship is non-linear, Pearson's R may not accurately represent the strength or direction of the relationship.
2. **Sensitivity to Outliers**: Pearson's R is highly sensitive to outliers, which can significantly affect the correlation value.
3. **Range of Data**: The correlation can be misleading if the range of data is restricted. A limited range can underestimate the true correlation.
4. **No Causation**: It does not imply causation. A high correlation between two variables does not mean that one variable causes the other.
5. **Homogeneity of Variance**: Pearson's R assumes homogeneity of variance (homoscedasticity). If the variances are not equal, the correlation may be inaccurate.
6. **Measurement Scale**: It requires interval or ratio scale data. It is not suitable for ordinal or nominal data.

Understanding these limitations is crucial for correctly interpreting the results of a Pearson correlation analysis

Q4. 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 used to adjust the range of data features so they fit within a specific scale. This is crucial in machine learning and data analysis because many algorithms perform better or converge faster when the data is scaled.

Scaling is performed for several reasons:

1. Improved Algorithm Performance: Algorithms like gradient descent, k-nearest neighbours, and support vector machines perform better when features are on a similar scale.
2. Equal Contribution: Ensures that all features contribute equally to the model, preventing features with larger ranges from dominating the learning process2.
3. Numerical Stability: Prevents numerical issues during calculations, such as overflow or underflow2.

Normalized Scaling vs. Standardized Scaling

- Normalized Scaling:
    - Definition: Also known as min-max scaling, it transforms data to fit within a specific range, typically 0 to 1.
    - Formula:

$$X_{norm} = \frac{X_{max} - X_{min}}{X - X_{min}}$$

    - Use Case: Useful when you need to ensure all features are on the same scale without changing the distribution shape.
- Standardized Scaling:
    - Definition: Also known as z-score normalization, it transforms data to have a mean of 0 and a standard deviation of 1.
    - Formula:

$$X_{std} = \frac{X - \mu}{\sigma}$$

    - Use Case: Useful when you need to center the data and ensure it has unit variance, which is important for algorithms that assume normally distributed data.

Scaling impacts specific machine learning algorithms:

1. Gradient Descent-Based Algorithms

- Algorithms: Linear Regression, Logistic Regression, Neural Networks
- Impact: Scaling helps in faster convergence of the gradient descent algorithm. When features are on different scales, the gradient descent updates can oscillate inefficiently, slowing down convergence. Standardization is often preferred here.

2. Distance-Based Algorithms

- Algorithms: K-Nearest Neighbours (KNN), K-Means Clustering
- Impact: These algorithms rely on distance metrics (e.g., Euclidean distance). If features are not scaled, features with larger ranges can dominate the distance calculations, leading to biased results. Normalization is commonly used to ensure all features contribute equally.

3. Support Vector Machines (SVM)

- Impact: SVMs aim to find the optimal hyperplane that separates classes. If features are not scaled, the hyperplane might be skewed towards features with larger ranges. Standardization helps in achieving a balanced hyperplane.

4. Principal Component Analysis (PCA)

- Impact: PCA is sensitive to the variances of the features. If features are not scaled, PCA might give more importance to features with higher variance. Standardization ensures that each feature contributes equally to the principal components.

5. Tree-Based Algorithms

- Algorithms: Decision Trees, Random Forests, Gradient Boosting
- Impact: These algorithms are generally not affected by feature scaling because they are based on decision rules rather than distance metrics.

However, scaling can still be beneficial in some cases, especially when combining tree-based models with other algorithms in ensemble methods.

Practical Example

Imagine you are working with a dataset containing features like age (ranging from 0 to 100) and income (ranging from 0 to 100,000). Without scaling, income would dominate the distance calculations in KNN, leading to biased predictions. By normalizing these features, both age and income will contribute equally to the distance metric.

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Answer:

An infinite value of the Variance Inflation Factor (VIF) indicates perfect multicollinearity among the independent variables in a regression model. This occurs when one independent variable can be perfectly predicted by a linear combination of the other independent variables. Mathematically, this happens when the coefficient of determination

The Variance Inflation Factor (VIF) is a measure used to detect multicollinearity in regression analysis. Multicollinearity occurs when one predictor variable in a regression model can be linearly predicted from the others with a substantial degree of accuracy. VIF quantifies how much the variance of a regression coefficient is inflated due to multicollinearity. A VIF value becomes infinite when R^2i is equal to 1. This situation occurs when there is perfect multicollinearity, meaning that i^{th} the predictor can be perfectly predicted by a linear combination of the other predictors. In such cases, the denominator of the VIF formula (1 - R^2i) becomes zero, causing the VIF to approach infinity.

$$\mathrm{VIF_i} = \frac{1}{1 - R_i^2}$$

VIF become infinite because of following reasons: • Perfect Multicollinearity: when one predictor is an exact linear function of one or more other predictors perfect multicollinearity occurs. • Duplicate Variables: Including the same variable more than once creates perfect multicollinearity. • Linear

Combinations: If a predictor is a linear combination of other predictors, this will also result in perfect multicollinearity.

Q6. 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, is a graphical tool used to assess whether a dataset follows a particular theoretical distribution, often the normal distribution. In the context of linear regression, Q-Q plots are particularly useful for evaluating the assumption that the residuals (errors) of the regression model are normally distributed.

- **Plotting Procedure:**
  - **Quantiles:** For example, the median is the 0.5 quantile.
  - Theoretical Quantiles: These are the quantiles of the theoretical distribution which are compared with data against evaluation (e.g., normal distribution).
  - **Sample Quantiles:** These are the quantiles of the sample data.
- **Creating a Q-Q Plot:**
  - Sorting the data and determine its sample quantiles.
  - Calculating the corresponding theoretical quantiles from the specified theoretical distribution.
  - Plotting the sample quantiles against the theoretical quantiles.
- **Interpretation:**
  - **Straight Line:** If the data follows the theoretical distribution, the points will approximately lie on a straight line (the 45-degree reference line).
  - **Deviations:** Systematic deviations from the line suggest departures from the theoretical distribution.

- **Importance of a Q-Q Plot in Linear Regression:** In linear regression, one of the key assumptions is that the residuals (errors) are normally distributed. This assumption is critical because many statistical tests and confidence intervals rely on it.

- **Use and Importance**:
  - **Assessing Normality:**

1. **Residual Analysis**: By plotting the residuals of a regression model it can be determine if residuals follow normal distributions.

2. **Model Validation**: If the residuals deviate significantly from the normal line, it suggests that the model assumptions might be violated, which can affect the validity of the model's statistical tests and confidence intervals.

- **Detecting Outliers and Extreme Points:**
- **Identifying Skewness and Kurtosis:**
  1. **Skewness:** If the points form an S-shaped curve, it indicates skewness in the residuals.
  2. **Kurtosis:** If the points are concave up or down relative to the line, it indicates issues with the kurtosis
- **Improving Model Fit: Model Diagnostics**: Using a Q-Q plot to diagnose non-normality in residuals can guide transformations of the response variable or adding/removing predictors to improve model fit.