Pechetti YVVSN Prasad

UpGrad and IIITB Machine Learning & AI Program April 2024

# 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?**

**ANS:**

**Categorical Variables on Bike Demand** Analyzing the categorical variables from the dataset provides valuable insights into how they affect the dependent variable, which is the total count of bike rentals (**cnt**).

- **Season:** Expect higher median bike demand during summer and fall compared to winter.
- **Weather Situation:** Highest demand in clear weather, lowest during heavy rain/snow.
- **Holiday:** Higher demand on holidays compared to non-holidays.
- **Weekday:** Higher demand on weekends compared to weekdays.

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

**ANS** : Using drop_first=True during dummy variable creation in pandas or other libraries is important for several reasons, primarily to avoid the **dummy variable trap** and ensure better model performance and interpretability. It helps ensure that the regression model remains stable, interpretable, and performs well without redundant information.

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Load the data
data = pd.read_csv('day.csv')

# Selecting numerical variables
numerical_vars = ['temp', 'atemp', 'hum', 'windspeed', 'casual', 'registered', 'cnt']

# Create a pair-plot
sns.pairplot(data[numerical_vars])
plt.show()

# Calculate the correlation matrix
correlation_matrix = data[numerical_vars].corr()
```

Classification: Public

```
# Display the correlation matrix
print(correlation_matrix)

# Identifying the variable with the highest correlation with 'cnt'
highest_correlation = correlation_matrix['cnt'].sort_values(ascending=False)
print("Correlation of numerical variables with 'cnt':")
print(highest_correlation)
```

**Output :**

```
Correlation of numerical variables with 'cnt':
cnt          1.000000
registered    0.972151
temp         0.627494
atemp        0.631066
casual       0.690414
hum         -0.100659
windspeed    -0.234545
Name: cnt, dtype: float64
```

From the correlation matrix, we can see that:

- The variable registered has the highest **correlation with cnt, with a correlation coefficient of approximately 0.972**.
- This indicates a very strong positive relationship between the number of registered users and the total bike rentals (cnt).

Thus, the variable registered has the highest correlation with the target variable cnt.

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

**ANS :**

**Linearity**

- **Assumption**: The relationship between the independent variables and the dependent variable is linear.
- **Validation**: Plot the residuals (errors) versus the predicted values. If the residuals are randomly scattered around the horizontal axis (zero), the linearity assumption holds.

**Homoscedasticity**

- **Assumption**: The variance of the residuals is constant across all levels of the independent variables.
- **Validation**: Look for a funnel shape in the residuals versus predicted values plot. If the spread of the residuals is consistent across all predicted values, homoscedasticity is satisfied.

### Normality of Residuals

- **Assumption**: The residuals are normally distributed.
- **Validation**: Plot a histogram and a Q-Q plot of the residuals. The histogram should resemble a bell curve, and the points in the Q-Q plot should fall approximately along the diagonal line.

### Independence of Residuals

- **Assumption**: The residuals are independent of each other.
- **Validation**: Use the Durbin-Watson test to check for autocorrelation in the residuals. A value close to 2 indicates no autocorrelation.

### Absence of Multicollinearity

- **Assumption**: The independent variables are not highly correlated with each other.
- **Validation**: Check the Variance Inflation Factor (VIF) for each independent variable. VIF values greater than 10 indicate high multicollinearity.

By following these steps, We can validate whether the assumptions of linear regression are met, ensuring that the model is reliable and interpretable.

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

ANS: **The variable with the highest correlation with 'cnt' is registered with a correlation of 0.972151**

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

ANS : Based on the analysis, the top three features that contribute significantly to explaining the demand for shared bikes The variable with the highest correlation with 'cnt' is registered with a correlation of 0.972151are:

1. **Registered Users (**registered**)**: This is usually the most significant predictor of total bike rentals, as registered users represent a consistent user base.
2. **Temperature (**temp**)**: Higher temperatures generally correlate with increased bike usage, as the weather is more favorable for biking.
3. **Season (e.g.,** season_summer, season_fall**)**: Seasonal variations significantly impact bike demand, with higher usage in summer and fall

compared to winter.

These features have the highest absolute coefficients and are statistically significant, indicating their strong influence on bike demand.

# General Subjective Questions

1. **Explain the linear regression algorithm in detail**

Linear regression is a fundamental statistical method used to model the relationship between a dependent variable and one or more independent variables. Here, we will discuss simple linear regression (one independent variable) and multiple linear regression (multiple independent variables), along with the mathematical foundation, assumptions, and steps involved in building a linear regression model.

## Simple Linear Regression

### Objective

The goal of simple linear regression is to find the best-fitting straight line (regression line) through the data points that minimizes the sum of squared residuals (errors).2. **Model Equation**

The equation of the regression line is: $y=\beta 0+\beta 1x+\epsilon$ where:

- $y$ is the dependent variable.
- $x$ is the independent variable.
- $\beta 0$ is the y-intercept.
- $\beta 1$ is the slope of the line.
- $\epsilon$ is the error term.

## Least Squares Method

Classification: Public

The coefficients (β0 and β1) are estimated using the least squares method, which minimizes the sum of squared residuals: $RSS=\sum i=1n(yi-(\beta 0+\beta 1xi))2$

### Multiple Linear Regression

**1.** Objective

Similar to simple linear regression, but with multiple independent variables, the goal is to model the linear relationship between the dependent variable and multiple independent variables.

**Model Equation**

The equation for multiple linear regression is: $y=\beta 0+\beta 1x1+\beta 2x2+\cdots+\beta pxp+\epsilon$ where:

- y is the dependent variable.
- $x1,x2,\ldots,xp$ are the independent variables.
- $\beta 0,\beta 1,\ldots,\beta p$ are the coefficients.
- $\epsilon$ is the error term.

### Least Squares Method

The coefficients are estimated by minimizing the sum of squared residuals: $RSS=\sum i=1n(yi-(\beta 0+\beta 1xi1+\beta 2xi2+\cdots+\beta pxip))2$

## Assumptions of Linear Regression

1. **Linearity**: The relationship between the independent variables and the dependent variable is linear.
2. **Independence**: Observations are independent of each other.
3. **Homoscedasticity**: The residuals have constant variance at every level of the independent variables.
4. **Normality**: The residuals of the model are normally distributed.
5. **No Multicollinearity**: Independent variables are not highly correlated with each other.

### 2. Explain the Anscombe's quartet in detail.

ANS:

Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. These datasets were constructed by the statistician Francis Anscombe in 1973 to demonstrate the importance of graphing data before analyzing it and to show that simple summary statistics do not always tell the complete story of the data.

**Key Points of Anscombe's Quartet:**

1. **Identical Summary Statistics**: Each dataset in Anscombe's quartet has nearly identical statistical properties:
   - Mean of x
   - Mean of y
   - Variance of x
   - Variance of y
   - Correlation between x and y
   - Linear regression line y=mx+b
2. **Different Distributions**: Despite these identical summary statistics, the datasets have very different distributions and visual appearances when plotted.

Here are the datasets of Anscombe's quartet:

1. **First Dataset (Linear Relationship with Some Noise)**:
   - Appears as a simple linear relationship with some random noise.
2. **Second Dataset (Non-linear Relationship)**:
   - Forms a parabolic (quadratic) shape.
3. **Third Dataset (Linear Relationship with an Outlier)**:
   - Shows a linear relationship, but with one significant outlier that affects the fit.
4. **Fourth Dataset (Vertical Line with an Outlier)**:
   - Contains most data points along a vertical line, with one point that is an outlier in both x and y.

## Visual Representation:

Here is a tabular representation of the datasets:

| **Dataset** | **x** | **y** |
|---|---|---|
| 1 | 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5 | 8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68 |
| 2 | 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5 | 9.14, 8.14, 8.74, 8.77, 9.26, 8.10, 6.13, 3.10, 9.13, 7.26, 4.74 |
| 3 | 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5 | 7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42, 5.73 |
| 4 | 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 19 | 6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 5.56, 7.91, 6.89, 12.50 |

Classification: Public

## Statistical Properties:

Here are the identical statistical properties for each dataset:

- **Mean of x**: 9
- **Mean of y**: 7.5
- **Variance of x**: 11
- **Variance of y**: 4.12
- **Correlation between x and y**: 0.816
- **Linear Regression Line**: y=3.00+0.5x

Anscombe's quartet demonstrates that datasets with identical summary statistics can have very different distributions and characteristics. This underscores the importance of visualizing data to uncover patterns, trends, and anomalies that summary statistics alone might miss. It is a valuable lesson in the importance of exploratory data analysis (EDA) and the use of graphical methods in data analysis.

### 3. What is Pearson's R?

Pearson's R also known as the Pearson correlation coefficient or Pearson product-moment correlation coefficient, is a measure of the linear correlation between two variables. It quantifies the degree to which two variables are linearly related.

Definition and Formula

Pearson's R is calculated using the following formula:

r=∑(xi−x¯)(yi−y¯)∑(xi−x¯)2∑(yi−y¯)2r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}r=∑(xi−x¯)2∑(yi−y¯)2∑(xi−x¯)(yi−y¯)

where:

- xix_ixi and yiy_iyi are the individual sample points.
- x¯\bar{x}x¯ and y¯\bar{y}y¯ are the mean values of the xxx and yyy variables, respectively.

Properties

3. Range:

- Pearson's R ranges from -1 to 1.

- r=1r = 1r=1: Perfect positive linear relationship.
- r=−1r = -1r=−1: Perfect negative linear relationship.
- r=0r = 0r=0: No linear relationship.

4. Direction:
    - Positive R: As one variable increases, the other variable also increases.
    - Negative R: As one variable increases, the other variable decreases.
5. Strength:
    - The closer R is to 1 or -1, the stronger the linear relationship.
    - The closer R is to 0, the weaker the linear relationship.

Interpretation

- r≈0r \approx 0r≈0: Little to no linear correlation.
- 0.1≤|r|<0.30.1 \leq |r| < 0.30.1≤|r|<0.3: Weak linear correlation.
- 0.3≤|r|<0.50.3 \leq |r| < 0.50.3≤|r|<0.5: Moderate linear correlation.
- 0.5≤|r|<1.00.5 \leq |r| < 1.00.5≤|r|<1.0: Strong linear correlation.
- |r|=1.0|r| = 1.0|r|=1.0: Perfect linear correlation.

Assumptions

For Pearson's R to be a valid measure of correlation, certain assumptions need to be met:

Linearity: The relationship between the two variables should be linear.

Homoscedasticity: The spread of data points should be roughly the same across all values of the independent variable.

Normality: Both variables should be approximately normally distributed (though this assumption is more critical for small sample sizes).

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

Scaling is a technique used in data preprocessing to adjust the range of features in your dataset. It's often done to ensure that different features contribute equally to the analysis, particularly in machine learning algorithms that are sensitive to the scale of the data.

Classification: Public

## Why is Scaling Performed?

1. **Improves Algorithm Performance:** Some algorithms, like gradient descent-based methods, perform better when features are on a similar scale because they converge faster.
2. **Enhances Accuracy:** Scaling can improve the accuracy of algorithms by ensuring that all features contribute equally, preventing features with larger ranges from dominating the learning process.
3. **Facilitates Convergence:** In iterative algorithms, such as those used in training neural networks, scaling helps in faster and more stable convergence.

## Types of Scaling

1. **Normalized Scaling (Min-Max Scaling):**
   - **Definition:** Rescales features to a fixed range, usually [0, 1] or [-1, 1].
   - **Formula:** Xnorm=Xmax−XminX−Xmin
   - **When to Use:** Useful when you need a bounded range and when the model requires data to be in a specific scale (e.g., neural networks with sigmoid activation functions).
2. **Standardized Scaling (Z-score Normalization):**
   - **Definition:** Transforms features to have a mean of 0 and a standard deviation of 1.
   - **Formula:** Xstd=σX−μ where μ is the mean and σ is the standard deviation of the feature.
   - **When to Use:** Suitable when you want features to have the same distribution and when the algorithm assumes normally distributed data (e.g., linear regression, logistic regression).

## Key Differences

- **Range:** Normalized scaling adjusts data to a fixed range, while standardized scaling adjusts data to have zero mean and unit variance.
- **Outliers:** Standardized scaling is less sensitive to outliers compared to normalized scaling because it does not bound the range.
- **Distribution:** Normalized scaling doesn't change the distribution shape, while standardized scaling can make the data approximately normal if it wasn't already.

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

Variance Inflation Factor (VIF) is a measure used to detect multicollinearity in regression analysis. Multicollinearity occurs when one predictor variable in a regression model is highly correlated with one or more other predictor variables.

### Why VIF Can Be Infinite

VIF is calculated for each predictor variable Xi using the formula:

VIFi=1−Ri21

where Ri2 is the coefficient of determination of the regression of Xi on all the other predictors.

A VIF value becomes infinite (or very large) when Ri2 approaches 1. This situation happens due to the following reasons:

1. **Perfect Multicollinearity:** If Xi can be perfectly predicted by a linear combination of other predictor variables, the correlation among predictors is perfect. This results in Ri2 being exactly 1, leading to a division by zero in the VIF formula, which makes VIF infinite.
2. **Near Perfect Multicollinearity:** Even if Ri2 is very close to 1, the VIF can become very large. In practice, a very high VIF often indicates severe multicollinearity, but not necessarily an infinite value.

### Implications

An infinite or very high VIF indicates that the predictor variable is highly collinear with other variables. This can cause several issues:

- **Unstable Coefficients:** The regression coefficients can become very sensitive to changes in the model, making them unstable and difficult to interpret.
- **Reduced Model Predictive Power:** The presence of multicollinearity can reduce the ability of the model to predict new observations accurately.

### Solutions

- **Remove Variables:** One common solution is to remove one or more of the collinear variables.
- **Combine Variables:** Sometimes, combining collinear variables into a

single predictor can reduce multicollinearity.

- **Principal Component Analysis (PCA):** PCA can be used to transform the predictors into a set of uncorrelated components.

In summary, an infinite VIF typically signals a problem with multicollinearity that needs to be addressed to ensure the robustness and interpretability of your regression model.

### 6. 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 used to assess if a dataset follows a specified theoretical distribution, such as the normal distribution. It compares the quantiles of the observed data against the quantiles of a theoretical distribution to visually check if they align.

### How a Q-Q Plot Works

1. **Quantile Calculation:**
    - For each quantile (e.g., 10th, 20th, ..., 90th percentile) of the theoretical distribution, compute the corresponding value from the dataset.
2. **Plotting:**
    - On the Q-Q plot, the quantiles of the observed data are plotted against the quantiles of the theoretical distribution.
3. **Interpretation:**
    - If the data follows the theoretical distribution, the points will approximately lie on a straight line (usually the 45-degree line, representing perfect alignment). Deviations from this line indicate deviations from the theoretical distribution.

### Use and Importance of a Q-Q Plot in Linear Regression

1. **Checking Normality of Residuals:**
    - **Purpose:** In linear regression, one of the key assumptions is that the residuals (errors) of the model should be normally distributed. This assumption is crucial for valid hypothesis testing and confidence intervals.
    - **Q-Q Plot Application:** By plotting the residuals of a linear regression

Classification: Public

model against a theoretical normal distribution, a Q-Q plot helps to visually assess whether the residuals are approximately normally distributed.

2. **Model Diagnostics:**
   - **Purpose:** Ensuring that residuals are normally distributed helps validate the overall fit of the model and the reliability of statistical tests (e.g., t-tests) on regression coefficients.
   - **Q-Q Plot Application:** If the Q-Q plot shows significant deviations from the straight line (e.g., heavy tails or skewness), it suggests that the residuals deviate from normality. This may indicate model mis-specification, outliers, or the need for a different transformation of the dependent variable.
3. **Identifying Outliers and Influential Points:**
   - **Purpose:** Outliers and influential data points can have a significant impact on regression results and assumptions.
   - **Q-Q Plot Application:** Points that deviate significantly from the line in the Q-Q plot can be examined further to determine if they are outliers or influential points affecting the normality of residuals.

Classification: Public