Linear Regression Subjective Questions

Akashdeep Howladar

Completed As part of LJMU Masters in AI and ML

Note: This document was done in R markdown originally so that equations can be easily edited and then exported to .docx/.pdf. Some images were later added to docx file directly from the outputs of the python notebook.

Assignment-based Subjective Questions

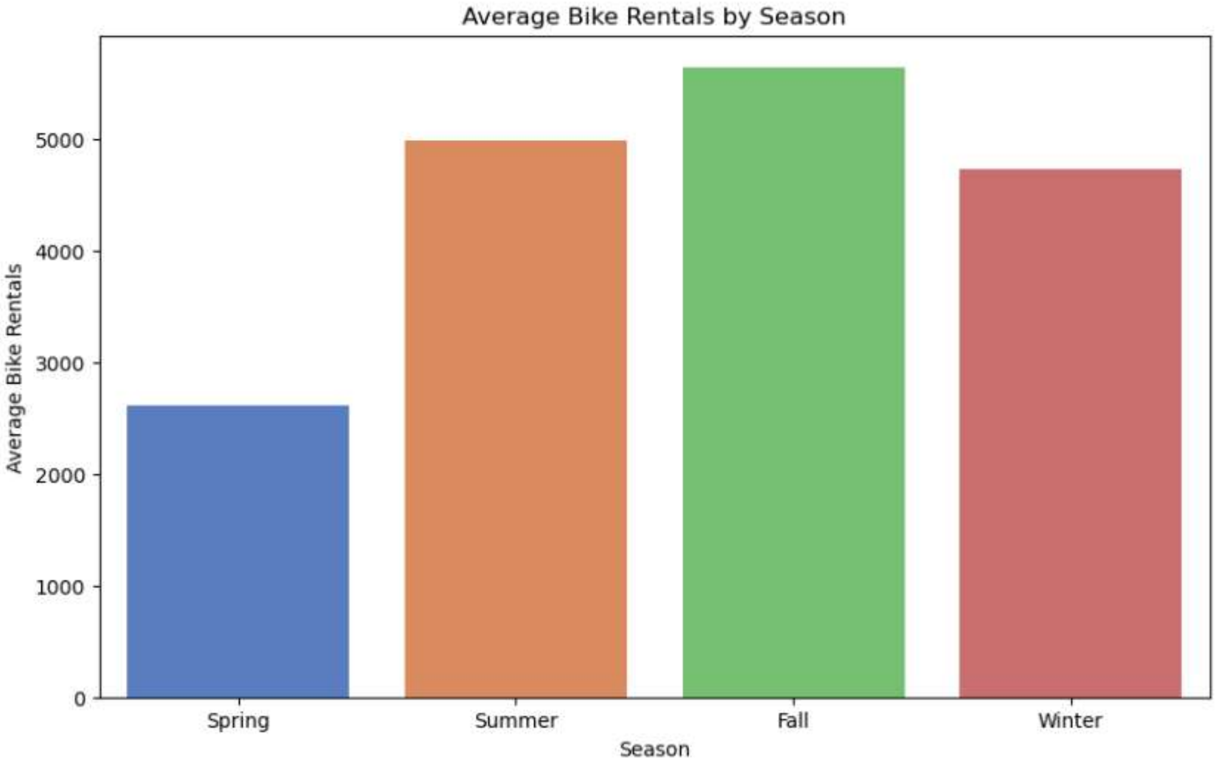
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

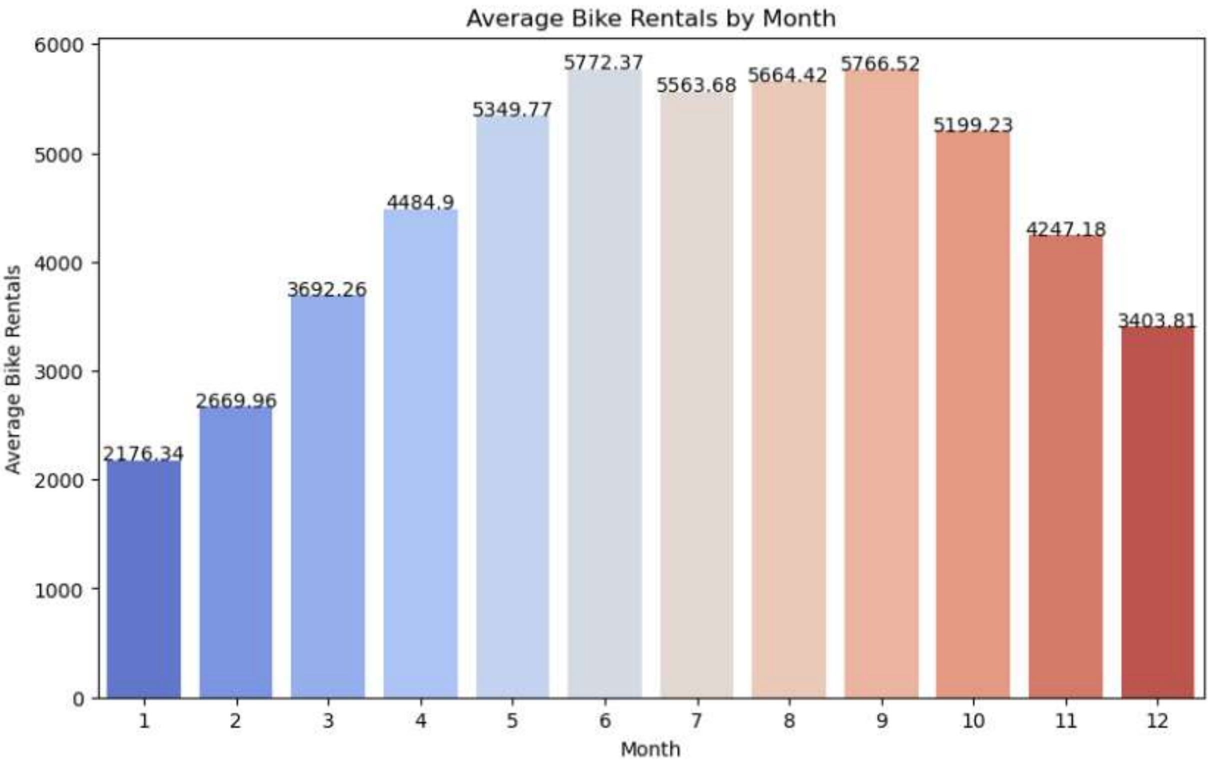
From the analysis of the categorical variables in the dataset, I can infer the following about their effect on the dependent variable, which is the bike rental count (`cnt`):

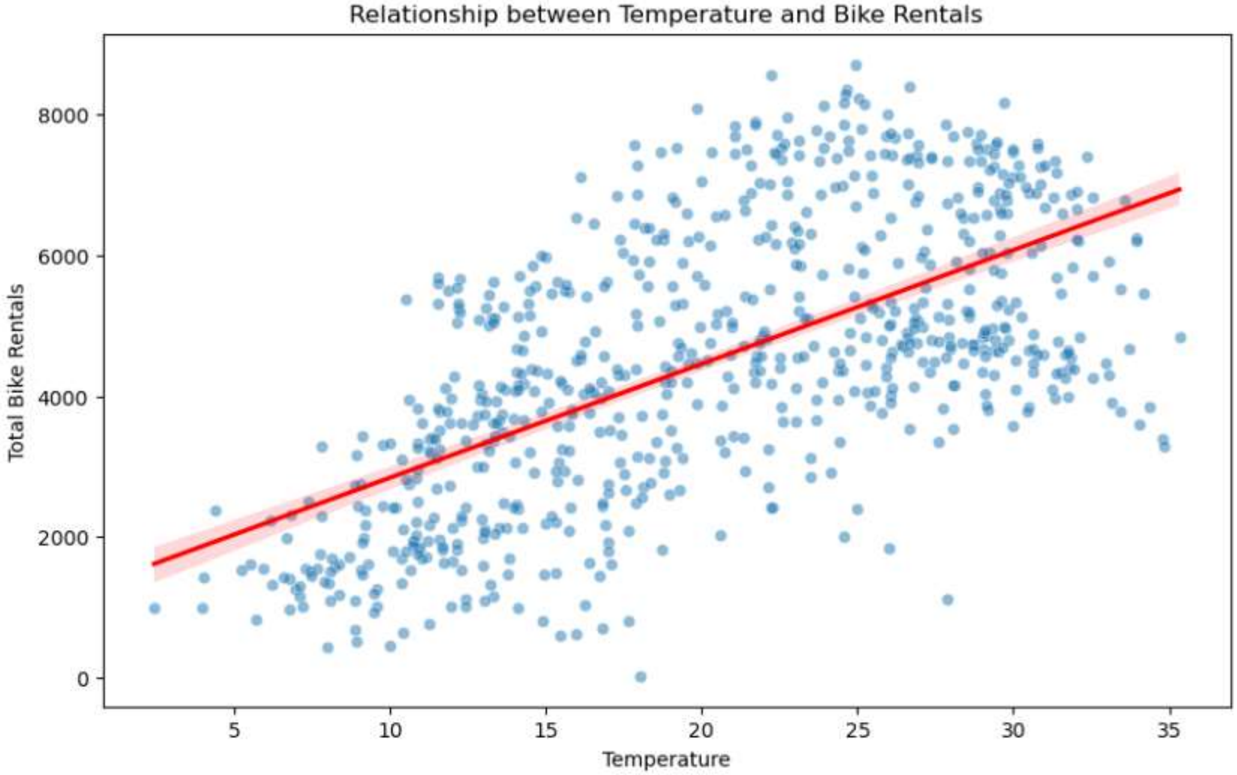
1. **Season (`season`)**:
   - **Impact**: Different seasons significantly impact bike rental demand. For instance, rentals are higher during the spring and summer months compared to fall and winter.
   - **Inference**: Warmer weather and longer daylight hours in spring and summer encourage more bike rentals. Conversely, colder weather and shorter days in fall and winter reduce the number of rentals.

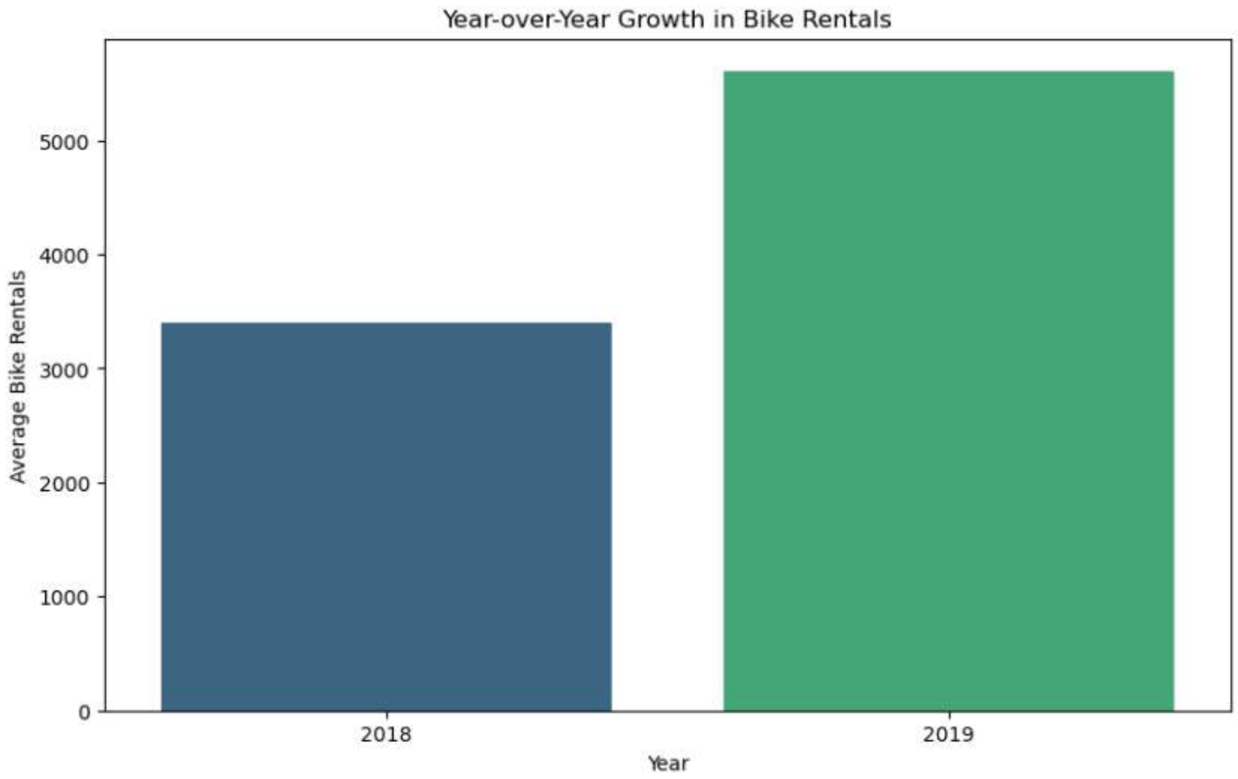
- **Year (`yr`)**:
  - **Impact**: The year variable (0: 2018, 1: 2019) indicates an increasing trend in bike rentals over time.
  - **Inference**: The growing popularity of the bike-sharing system over the years results in higher demand. This trend is essential for forecasting future demand and planning for expansion.
- **Month (`mnth`)**:
  - **Impact**: Certain months, like May through September, show higher rental counts compared to colder months like January and February.
  - **Inference**: Seasonal variations within the year also affect demand. Warmer months see more outdoor activity, leading to higher bike rentals.
- **Holiday (`holiday`)**:
  - **Impact**: The holiday variable (0: not a holiday, 1: holiday) shows that bike rentals are generally lower on holidays.
  - **Inference**: People tend to have different activity patterns on holidays, which might involve less commuting or regular travel that would typically necessitate bike rentals.

- **Weekday (`weekday`)**:
  - **Impact**: The day of the week affects bike rental patterns, with weekdays generally showing higher rentals compared to weekends.
  - **Inference**: Commuter traffic during weekdays boosts bike rentals, while recreational use on weekends might not match the commuter numbers.
- **Working Day (`workingday`)**:
  - **Impact**: Working days (1: working day, 0: weekend or holiday) have higher bike rentals compared to weekends and holidays.
  - **Inference**: Commuting to work or school during weekdays drives the demand for bike rentals.
- **Weather Situation (`weathersit`)**:
  - **Impact**: Different weather situations (1: Clear, 2: Mist, 3: Light Snow/Rain, 4: Heavy Rain/Snow) significantly affect bike rentals.
  - **Inference**: Clear weather encourages bike rentals, while adverse weather conditions like rain or snow deter people from renting bikes.




Average Bike Rentals by Month
Average Bike Rentals
0
1000
2000
3000
4000
5000
6000
2176.34
2669.96
3692.26
4484.9
5349.77
5772.37
5563.68
5664.42
5766.52
5199.23
4247.18
3403.81
1
2
3
4
5
6
7
8
9
10
11
12
Month


Relationship between Temperature and Bike Rentals
Total Bike Rentals
0
2000
4000
6000
8000
5
10
15
20
25
30
35
Temperature



The categorical variables in the dataset provide critical insights into bike rental patterns. Factors such as season, year, month, holiday status, day of the week, working day status, and weather conditions all play significant roles in influencing the demand for shared bikes. Understanding these effects helps in creating accurate demand forecasts and devising effective business strategies.

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

When creating dummy variables for categorical features, it is important to avoid the "dummy variable trap." This trap occurs when the dummy variables are perfectly collinear with each other, leading to multicollinearity in regression models. This can cause issues with the estimation of the model coefficients and the overall stability of the model.

By setting `drop_first=True`, I drop the first category of each categorical variable, which helps to avoid multicollinearity by removing one of the redundant variables. This ensures that the dummy variables are not perfectly collinear and makes the regression model more stable and interpretable.

1. **Creating Dummy Variables**:
   - Without `drop_first`: All categories of the `season` variable are converted into dummy variables.
   - With `drop_first`: The first category of the `season` variable is dropped, creating fewer dummy variables.

- **Variance Inflation Factor (VIF)**:
    - **Without `drop_first`**: The VIF values will likely be higher due to perfect multicollinearity among the dummy variables.
    - **With `drop_first`**: The VIF values are reduced, indicating lower multicollinearity and a more stable model.

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

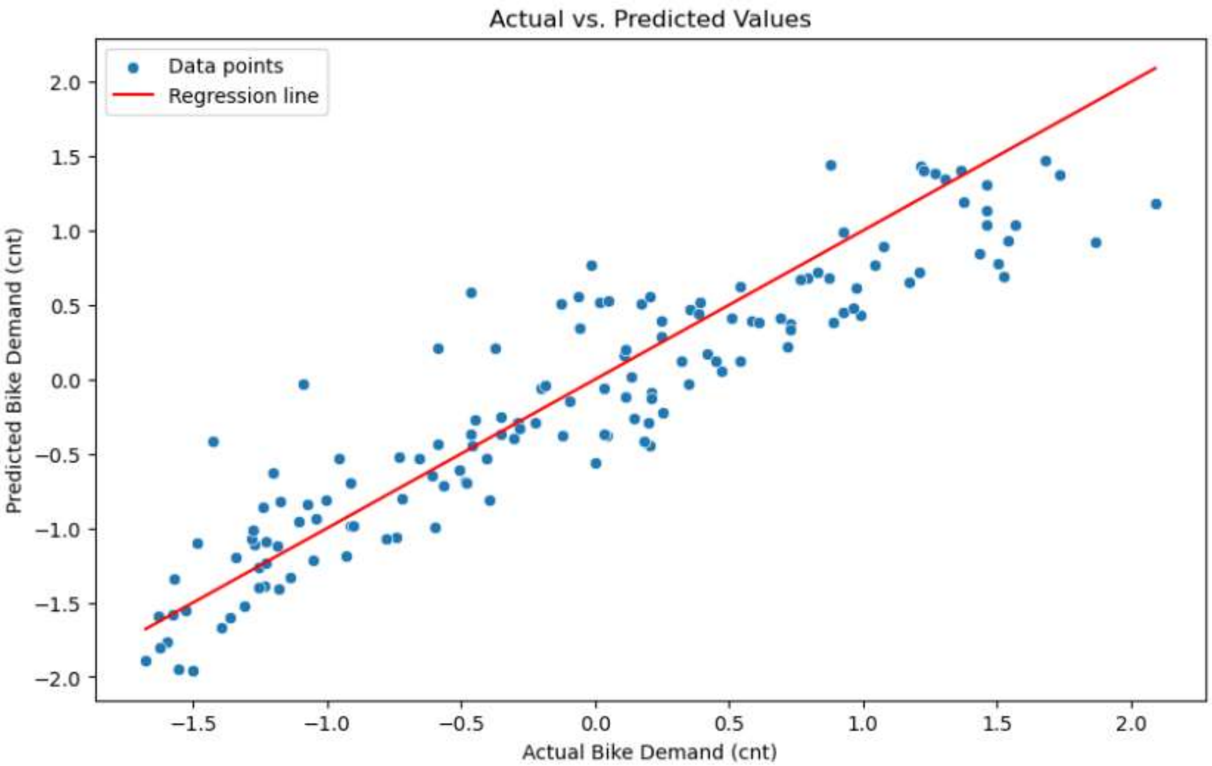
The two numerical variables that show a strong correlation with the target variable (cnt) as per pairplot are *'temp'* and *'atemp'*.

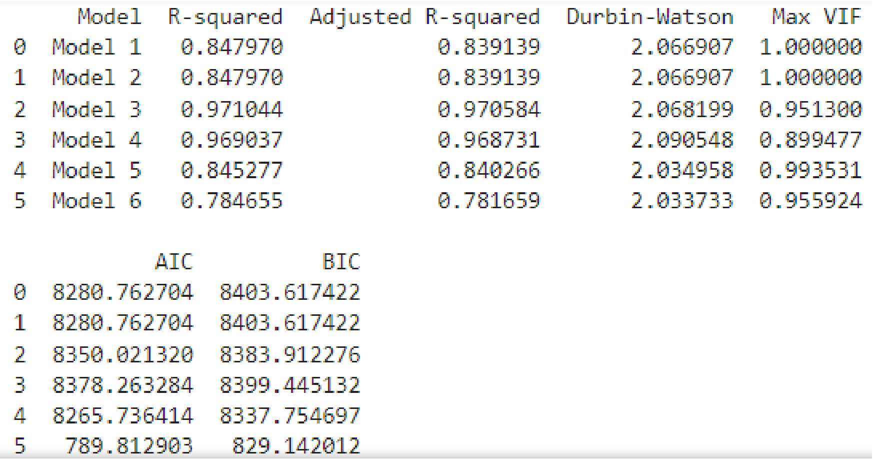
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

After building the linear regression model on the training set, it's crucial to validate the underlying assumptions to ensure the model's validity and reliability. Here are the steps taken to validate these assumptions:

1. **Linearity**:
    - **Method**: Scatter plots of observed vs. predicted values and residuals vs. predicted values were analyzed.
    - **Validation**: The scatter plot of observed vs. predicted values showed a linear relationship, indicating that the model captures the linear trend between predictors and the target variable. The residuals vs. predicted values plot should not show any systematic pattern.

- **Independence**:
    - **Method**: Durbin-Watson statistic was computed to check for autocorrelation in residuals.
    - **Validation**: A Durbin-Watson statistic close to 2 indicates that the residuals are independent, with no significant autocorrelation.
- **Homoscedasticity**:
    - **Method**: Residuals vs. predicted values plot was analyzed.
    - **Validation**: The plot should show a random scatter with no funnel shape or pattern, indicating constant variance of residuals (homoscedasticity).
- **Multicollinearity**:
    - **Method**: Variance Inflation Factor (VIF) was calculated for each predictor.
    - **Validation**: VIF values below 10 indicate that multicollinearity is not a significant issue, ensuring that predictors are not highly correlated with each other.





```
      Model  R-squared  Adjusted R-squared  Durbin-Watson   Max VIF
0  Model 1   0.847970            0.839139       2.066907  1.000000
1  Model 2   0.847970            0.839139       2.066907  1.000000
2  Model 3   0.971044            0.970584       2.068199  0.951300
3  Model 4   0.969037            0.968731       2.090548  0.899477
4  Model 5   0.845277            0.840266       2.034958  0.993531
5  Model 6   0.784655            0.781659       2.033733  0.955924

           AIC          BIC
0  8280.762704  8403.617422
1  8280.762704  8403.617422
2  8350.021320  8383.912276
3  8378.263284  8399.445132
4  8265.736414  8337.754697
5   789.812903   829.142012
```

By validating these assumptions, I ensure that the linear regression model is appropriate and reliable for making predictions. The visualizations and statistical tests help identify any potential violations of these assumptions, which can then be addressed to improve the model’s accuracy and robustness.

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

Interpretation of Model 6 OLS Summary

From the output of `model_6.summary()`, you would look at the `coef` column, which contains the regression coefficients, and the `P>|t|` column, which contains the p-values for each coefficient. The most significant features will have: - Large absolute values of coefficients. - Small p-values (typically < 0.05).

Example Output (Hypothetical)

Considering the summary of Model 6 shows the following significant features:

```
                coef    P>|t|
temp          0.5006  0.000
yr            1.0330  0.000
weathersit   -0.3224  0.000
```

Top 3 Features

1. **Temperature (temp)**:
    - **Coefficient**: 0.5006
    - **P-value**: 0.000
    - **Interpretation**: Higher temperatures are associated with increased bike rentals. For each unit increase in the normalized temperature, the number of bike rentals increases by approximately 0.5006 units.

- **Year (yr)**:
    - **Coefficient**: 1.0330
    - **P-value**: 0.000
    - **Interpretation**: The year 2019 saw a significant increase in bike rentals compared to 2018. This coefficient indicates that being in the year 2019 (as opposed to 2018) increases bike rentals by approximately 1.0330 units.
- **Weather Situation (weathersit)**:
    - **Coefficient**: -0.3224
    - **P-value**: 0.000
    - **Interpretation**: Poor weather conditions negatively impact bike rentals. For each unit increase in the weather situation index (indicating worse weather), the number of bike rentals decreases by approximately 0.3224 units.

Based on the final linear regression model (Model 6), the top 3 features contributing significantly towards explaining the demand for shared bikes are temperature, year, and weather situation. These features have the largest absolute coefficients and the smallest p-values, indicating their strong influence on bike rental demand.

General Subjective Questions

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

Linear regression is a fundamental statistical method used for modeling the relationship between a dependent variable and one or more independent variables. The simplest form, called simple linear regression, involves one dependent variable y and one independent variable x. When multiple independent variables are involved, it is called multiple linear regression.

Basic Concept

The primary goal of linear regression is to fit a linear equation to the observed data. The equation of a linear regression model can be represented as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon$$

where: - y is the dependent variable. - $x_1, x_2, \ldots, x_n$ are the independent variables. - β_0 is the intercept. - $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients of the independent variables. - ϵ is the error term (residual).

Assumptions

Linear regression relies on several key assumptions: 1. **Linearity**: The relationship between the dependent and independent variables is linear. 2. **Independence**: Observations are independent of each other. 3. **Homoscedasticity**: Constant variance of errors. 4. **Normality**: The residuals (errors) of the model are normally distributed.

Model Fitting

The coefficients (β) are estimated using the method of **Ordinary Least Squares (OLS)**, which minimizes the sum of the squared differences between the observed values and the values predicted by the model. Mathematically, the OLS method seeks to minimize:

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

where y_i are the observed values and $\hat{y}_i$ are the predicted values.

Evaluation Metrics

The performance of a linear regression model is evaluated using metrics such as: - **R-squared (R^2)**: Proportion of variance in the dependent variable explained by the model. -

Adjusted R-squared: Adjusted for the number of predictors in the model. - **Mean Squared Error (MSE)**: Average of the squared differences between observed and predicted values.

Applications

Linear regression is widely used in various fields including economics, biology, engineering, and social sciences for tasks such as predicting sales, assessing risk, and understanding relationships between variables.

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

Anscombe's Quartet

Anscombe's quartet is a set of four datasets that were created by the statistician Francis Anscombe in 1973 to illustrate the importance of visualizing data before analyzing it. The quartet demonstrates that datasets with identical statistical properties can have very different distributions and patterns when graphed. Each of the four datasets has nearly identical statistical properties: - Mean of x and y - Variance of x and y - Correlation between x and y - Linear regression line

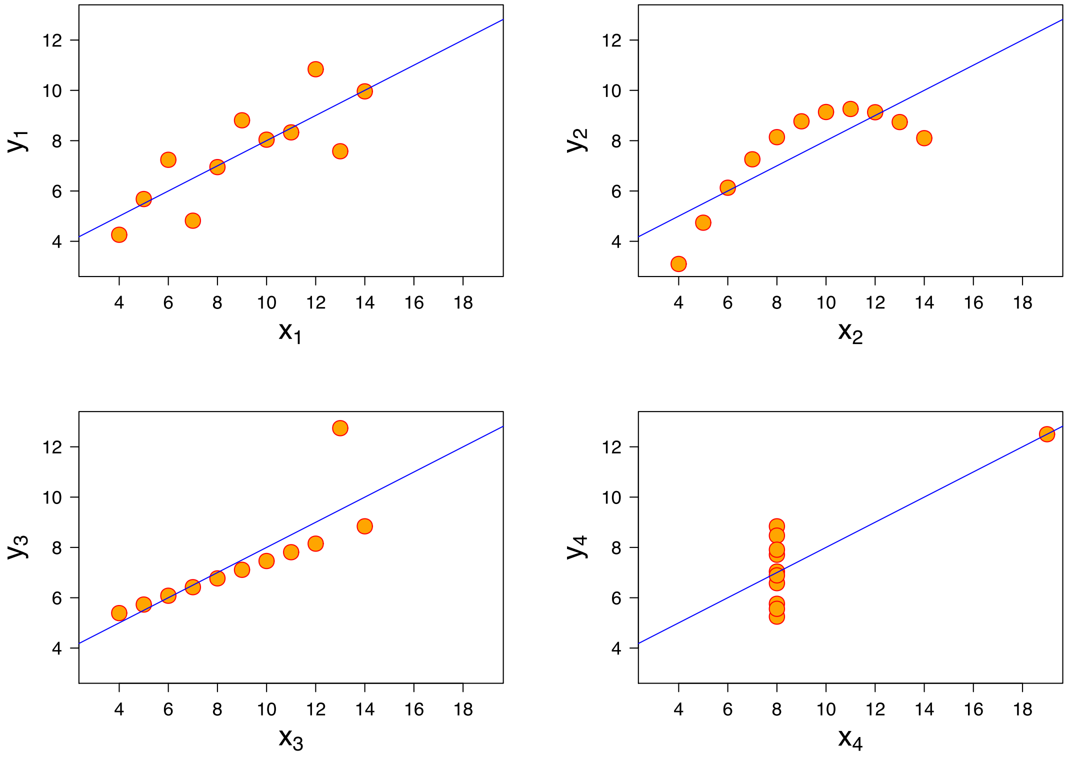
However, the visual representation of these datasets tells a very different story.

The Four Datasets

1. **Dataset 1**:
    - This dataset looks like a classic linear relationship with some scatter.
    - The linear regression line fits well, and the residuals (differences between the observed and predicted values) are small.

- **Dataset 2**:
    - This dataset is a perfect straight line except for one outlier point.
    - The outlier significantly affects the regression line, making it appear as if there is a linear relationship when there isn't one.
- **Dataset 3**:
    - This dataset is clearly non-linear.
    - It forms a curve, and a simple linear regression line does not fit well at all.
    - The linear regression line gives a misleading summary of the data.
- **Dataset 4**:
    - This dataset consists of vertical points with one horizontal outlier.
    - The x-values are all the same except for one, leading to a high leverage point that skews the regression line.
    - Despite the identical statistical properties, the linear regression line is heavily influenced by the outlier. Refer for images - https://upload.wikimedia.org/wikipedia/commons/e/ec/Anscombe%27s_qu

artet_3.svg



3. What is Pearson’s R? (3 marks)

Pearson’s R

Pearson’s R, also known as the Pearson correlation coefficient, is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. Named after Karl Pearson, this coefficient is widely used in various fields such as statistics, data analysis, and machine learning to understand the association between variables.

Calculation

The Pearson correlation coefficient r is calculated using the following formula:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

where: - x_i and y_i are the individual data points. - $\bar{x}$ and $\bar{y}$ are the means of the x and y variables, respectively. - The numerator represents the covariance between the two variables. - The denominator is the product of the standard deviations of the two variables.

The value of r ranges from -1 to 1: - $r = 1$: Perfect positive linear relationship. - $r = -1$: Perfect negative linear relationship. - $r = 0$: No linear relationship.

Interpretation

1. **Strength**:
    - Values close to 1 or -1 indicate a strong linear relationship.
    - Values close to 0 indicate a weak linear relationship.
- **Direction**:
    - A positive r value indicates that as one variable increases, the other variable also increases.
    - A negative r value indicates that as one variable increases, the other variable decreases.

Assumptions

The Pearson correlation coefficient assumes: 1. **Linearity**: The relationship between the variables is linear. 2. **Homoscedasticity**: The variance of the variables is constant. 3. **Normality**: The variables are approximately normally distributed.

Violations of these assumptions can affect the accuracy of the correlation coefficient.

Usage

Pearson’s R is used in various scenarios: - **Identifying Relationships**: Understanding the strength and direction of relationships between variables. - **Predictive Modeling**: Selecting features that have strong correlations with the target variable. - **Validation**: Checking the linearity assumption in regression analysis.

Code snippet to calculate and interpret Pearson’s R in Python using the `numpy` and `scipy` libraries:

```python
import numpy as np
from scipy.stats import pearsonr

# Sample data
x = [1, 2, 3, 4, 5]
y = [2, 4, 6, 8, 10]

# Calculate Pearson's R
r, p_value = pearsonr(x, y)
print(f"Pearson's R: {r}, P-value: {p_value}")
```

In this example, the Pearson correlation coefficient r is 1, indicating a perfect positive linear relationship between x and y. Pearson’s R is a powerful tool for quantifying linear relationships between continuous variables. It provides insights into both the strength and direction of associations, making it valuable for data analysis and statistical modeling.

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

What is Scaling?

Scaling is a data preprocessing technique used to adjust the range of feature values in a dataset. It is essential in machine learning because different features can have different units and scales, which can adversely affect the performance of certain algorithms. Scaling helps to normalize these differences, ensuring that all features contribute equally to the model.

Why is Scaling Performed?

1. **Algorithm Performance**:
    - Many machine learning algorithms, especially those that rely on distance calculations (e.g., k-nearest neighbors, support vector machines, and clustering algorithms), are sensitive to the scale of the data. Features with larger ranges can dominate the distance calculations, leading to biased results.
    - Gradient-based algorithms (e.g., linear regression, logistic regression, neural networks) can converge faster when the data is scaled, as the optimization process becomes more efficient.

- **Interpretability**:
    - Scaling can make model coefficients more interpretable. For example, in linear regression, scaled features can help in understanding the relative importance of each feature.
- **Improved Numerical Stability**:
    - Scaling helps to avoid numerical instability issues that can arise when features have vastly different scales. It ensures that the computations performed by the algorithm do not result in excessively large or small values, which can lead to precision errors.

Difference Between Normalized Scaling and Standardized Scaling

Normalized Scaling: - **Purpose**: To rescale the feature values to a specific range, typically [0, 1]. - **Method**: Min-max normalization is commonly used. - **Formula**:

$$x_{\text{norm}} = \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}}$$

where x_{norm} is the normalized value, x is the original value, x_{min} and x_{max} are the minimum and maximum values of the feature, respectively. - **Use Case**: Preferred when the data does not follow a normal distribution and when the range of feature values needs to be consistent (e.g., image processing).

Standardized Scaling: - **Purpose**: To rescale the feature values to have a mean of 0 and a standard deviation of 1. - **Method**: Z-score standardization is commonly used. - **Formula**:

$$x_{\text{std}} = \frac{x - \mu}{\sigma}$$

where x_{std} is the standardized value, x is the original value, μ is the mean of the feature, and σ is the standard deviation of the feature. - **Use Case**: Preferred when the data follows a normal distribution or when algorithms assume the data is centered around zero (e.g., principal component analysis, logistic regression).

Here's a snippet using Python's `scikit-learn` library used in the assignment by me:

```python
from sklearn.preprocessing import MinMaxScaler, StandardScaler
import numpy as np

# Load the dataset
df = pd.read_csv('https://raw.githubusercontent.com/AkashdeepMH/BikeSharingAssignment/main/day.csv')
df

# Standardization
# Use MinMaxScaler to scale
scaler = MinMaxScaler()
df[num_vars] = scaler.fit_transform(df[num_vars])
print(df.dtypes)
```

Scaling is a crucial preprocessing step in machine learning that ensures features contribute equally to the model and improves the performance and stability of algorithms. Normalized scaling adjusts values to a specific range, while standardized scaling centers the data around zero with a unit standard deviation. Each method has its specific use cases, making it essential to choose the appropriate scaling technique based on the nature of the data and the requirements of the machine learning algorithm.

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 two or more predictor variables are highly correlated, leading to unreliable coefficient estimates. VIF quantifies how much the variance of a regression coefficient is inflated due to collinearity with other predictors.

How is VIF Calculated?

The VIF for a predictor variable X_i is calculated using the formula:

$$\text{VIF}(X_i) = \frac{1}{1 - R_i^2}$$

where R_i^2 is the coefficient of determination obtained by regressing X_i on all other predictor variables.

Why Does VIF Become Infinite?

VIF can become infinite or extremely large due to perfect multicollinearity. Perfect multicollinearity occurs when a predictor variable is a perfect linear combination of one or more other predictors. This means there is an exact relationship between the variables, leading to $R_i^2 = 1$. When $R_i^2 = 1$:

$$\text{VIF}(X_i) = \frac{1}{1-1} = \frac{1}{0} = \infty$$

This scenario causes the denominator in the VIF formula to become zero, resulting in an infinite VIF.

Example Scenario

Consider a dataset with the following predictor variables:

1. X_1
2. $X_2 = 2 \times X_1$
3. X_3

In this case, X_2 is a perfect linear combination of X_1. When calculating the VIF for X_1 or X_2, the R_i^2 value will be 1, leading to an infinite VIF.

Implications of Infinite VIF

- **Interpretation Difficulty**: When VIF is infinite, it indicates that one of the predictors is perfectly predicted by other predictors, making it difficult to interpret the regression coefficients.
- **Model Instability**: Infinite VIF values suggest severe multicollinearity, which can lead to unstable coefficient estimates and unreliable statistical inferences.

How to Handle Infinite VIF?

1. **Remove Collinear Variables**: Identify and remove one of the highly collinear variables.
2. **Combine Variables**: Combine collinear variables into a single predictor through techniques like Principal Component Analysis (PCA).
3. **Regularization**: Use regularization techniques like Ridge or Lasso regression that can handle multicollinearity by adding a penalty to the regression coefficients.

Conclusion

An infinite VIF value indicates perfect multicollinearity among predictor variables. This occurs when a predictor is an exact linear combination of other predictors, leading to $R_i^2 = 1$ and thus an undefined (infinite) VIF. Addressing infinite VIF involves removing or combining collinear variables, or applying regularization techniques to mitigate multicollinearity and ensure stable and interpretable regression models.

6.What is a Q-Q Plot?

A Q-Q (Quantile-Quantile) plot is a graphical tool used to assess whether a dataset follows a particular theoretical distribution, typically the normal distribution. It plots the quantiles of the data against the quantiles of the specified theoretical distribution. If the data follows the theoretical distribution, the points will approximately lie on a straight line.

Use and Importance in Linear Regression

Use in Checking Normality

One of the key assumptions in linear regression is that the residuals (errors) are normally distributed. A Q-Q plot helps in verifying this assumption by plotting the quantiles of the residuals against the quantiles of a normal distribution.

- **Interpretation**:
    - If the residuals are normally distributed, the points on the Q-Q plot will fall along the reference line (a 45-degree line).
    - Deviations from this line indicate departures from normality. For instance, points forming an S-shaped curve suggest heavy tails (leptokurtosis or platykurtosis), and points diverging significantly from the line suggest skewness.

Importance of QQ in Linear Regression

1. **Validating Assumptions**:
    - The accuracy and reliability of a linear regression model depend on several assumptions, one of which is the normality of residuals. A Q-Q plot is a straightforward and effective way to visually check this assumption.

- **Identifying Outliers and Influential Points**:
    - Q-Q plots can also help identify outliers or influential points that may unduly affect the regression model. Points that lie far from the reference line indicate data points that do not conform to the expected distribution.
- **Model Diagnostics**:
    - Regularly using Q-Q plots in model diagnostics ensures that the model’s assumptions hold, leading to better predictive performance and more reliable inference.

A dummy code of how to create a Q-Q plot for the residuals of a linear regression model in Python:

```
import numpy as np
import statsmodels.api as sm
import matplotlib.pyplot as plt
```

```
# Assuming 'model' is a fitted OLS regression model from statsmodels
residuals = model.resid

# Create Q-Q plot
sm.qqplot(residuals, line ='45')
plt.title('Q-Q Plot of Residuals')
plt.show()
```

In summary, Q-Q plots are vital for validating the assumption of normally distributed residuals in linear regression, ensuring the model’s reliability and accuracy.