Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 1 goes below this line> (Do not edit)

From my analysis:

1. Holiday: Negative effect on `cnt`, indicating lower bike demand on holidays as fewer people commute.
2. Working Day: Positive effect on `cnt`, suggesting higher bike usage on workdays due to commuting.
3. Holiday Weekday: An interaction variable showing reduced demand when holidays fall on weekdays.
4. Year (yr): Positive impact on `cnt`, reflecting an increase in bike demand over time, possibly due to the growing popularity or adoption of shared biking.
5. Weather Situation (weathersit): Clear weather (favorable conditions) has a positive effect on demand, while adverse weather (like rain or snow) reduces demand.

These categorical factors collectively indicate that bike demand is higher on workdays, in later years, and in clear weather.

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 2 goes below this line> (Do not edit)

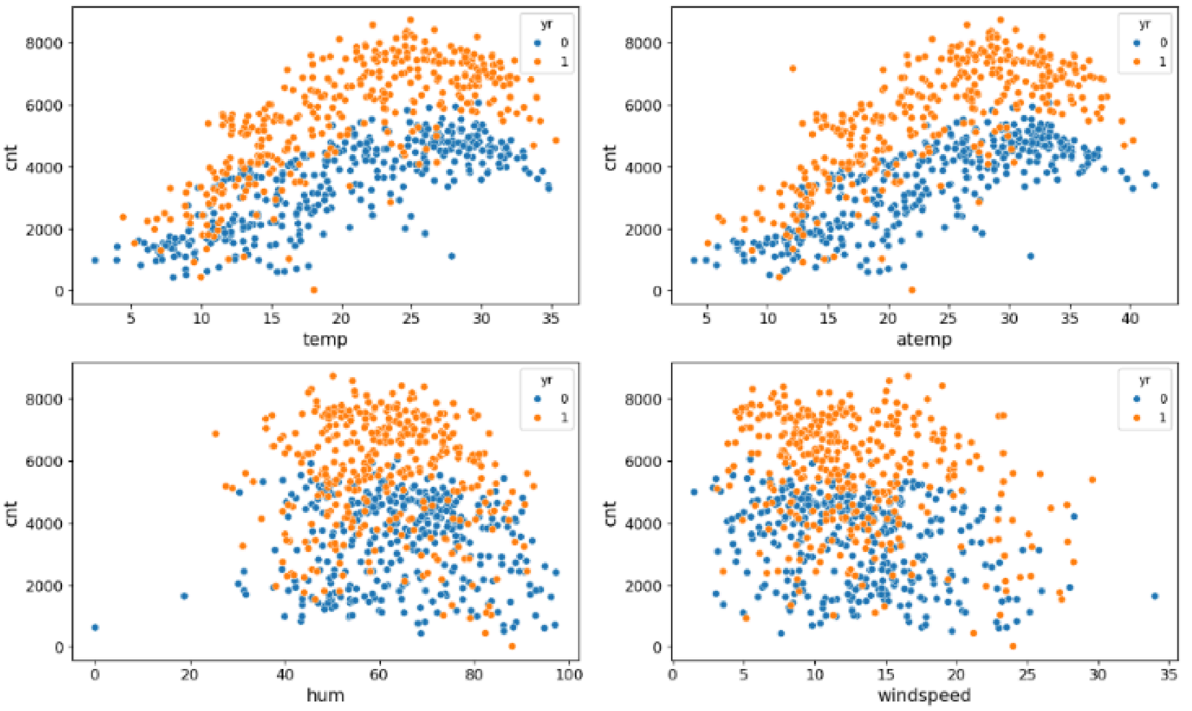
When creating dummy variables for categorical features, setting drop_first=True is crucial to prevent multicollinearity. This technique avoids the "dummy variable trap," where including all categories leads to perfect correlation among them. By omitting the first category, we establish a baseline category for comparison. Each remaining dummy variable then represents the effect of that category relative to the baseline. This approach improves model interpretability and ensures accurate coefficient estimates.

For Eg: Summer, Monsoon, Fall, Winter are four variables, when creating the dummy variables for these, without using drop_first = True, it would look like:
Summer - 1000
Monsoon - 0100
Fall - 0010
Winter - 0001

When creating the dummy variables for these, using drop_first = True, it would look like:
Monsoon - 100
Fall - 010
Winter - 001
where in, by default, we know Summer will be 000.

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest

correlation with the target variable? (Do not edit)
Total Marks: 1 mark (Do not edit)
Answer: <Your answer for Question 3 goes below this line> (Do not edit)



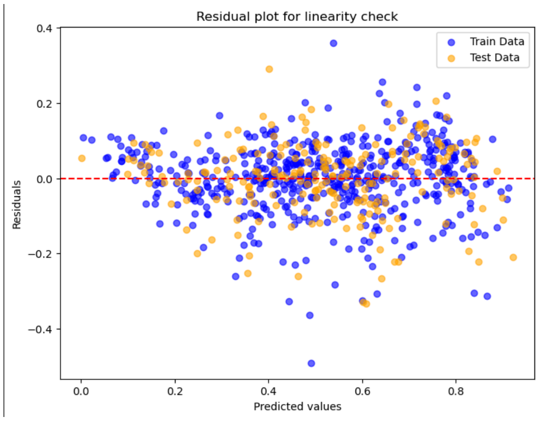
Based on the provided pair plots, it appears that the variable "temp" (temperature) has the highest correlation with the target variable "cnt" (count of rented bikes).

Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 4 goes below this line> (Do not edit)

Validating the assumptions of linear regression is essential to ensure that the model is reliable and the results are interpretable. After building the model on the training set, here are the main assumptions I have checked and the methods which I used to validate each:

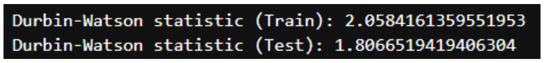
1. Linearity

- Check: Plot a scatter plot with the predicted values vs. the residuals.
- Validation: In the plot for Model 3, the residual plot shows that residuals are randomly scattered around zero, indicating that the model likely has a good linear fit without systematic patterns. The distribution of points suggests homoscedasticity, supporting the assumption of constant variance in errors across predicted values.



2. Independence of Errors

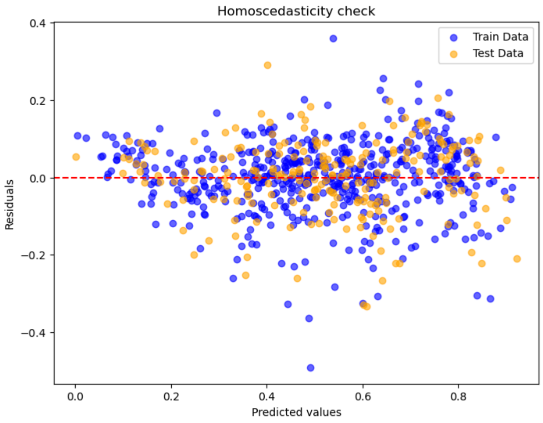
- Check: Calculate the Durbin-Watson test for autocorrelation in residuals.
- Validation: The Durbin-Watson statistics for both training (2.05) and test (1.81) data are close to 2, indicating minimal autocorrelation in the residuals. This suggests that the residuals are likely independent, satisfying the assumption of independence in the model.



```
Durbin-Watson statistic (Train): 2.0584161359551953
Durbin-Watson statistic (Test): 1.8066519419406304
```

3. Homoscedasticity (Constant Variance of Errors)

- Check: Plot a scatter plot with the predicted values vs. the residuals.
- Validation: The homoscedasticity check shows residuals scattered evenly around the horizontal axis without a clear pattern, suggesting constant variance across predicted values. This supports the assumption of homoscedasticity, indicating that error terms are similarly distributed across all levels of the predictor.

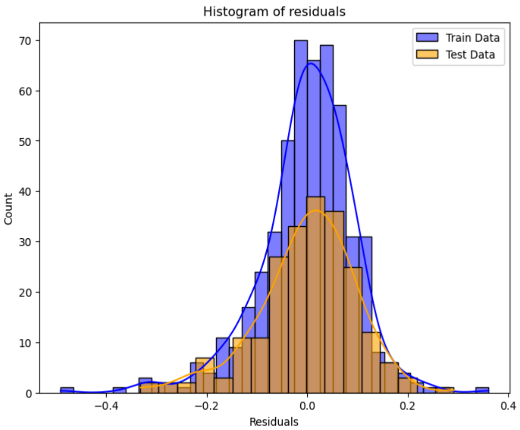


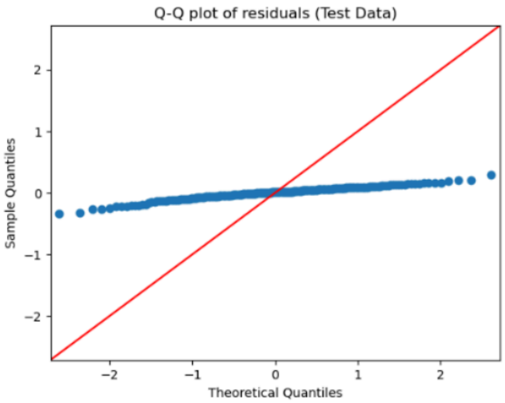
4. Normality of Errors

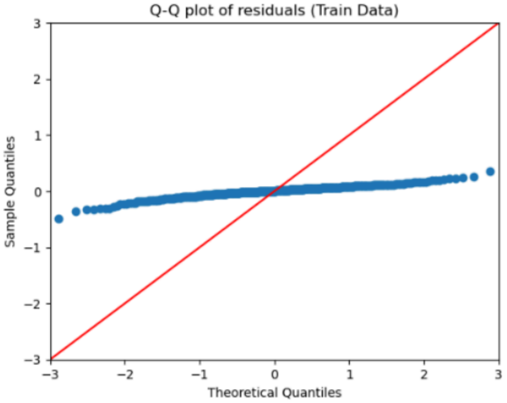
- Check: Use a Q-Q (quantile-quantile) plot or histogram of residuals to check if they follow a normal distribution.
- Validation: The histogram of residuals for both training and test data appears approximately normally distributed around zero, indicating that the normality assumption of residuals is likely met. This suggests that the model errors are symmetrically distributed, enhancing the reliability of inference.
- The Q-Q plot shows that the residuals for the training data deviate slightly from the theoretical line, indicating minor departures from normality, particularly in the tails. However, the points are

close to the line, suggesting that the normality assumption is reasonably satisfied for most of the data.

- The Q-Q plot for the test data shows residuals mostly aligning with the theoretical line, indicating that the residuals are approximately normally distributed. There is a slight deviation in the tails, but overall, the normality assumption is reasonably met.

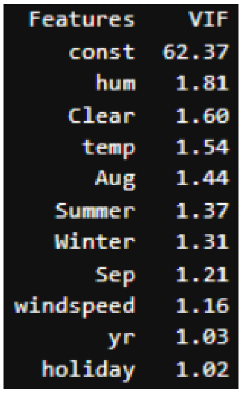






5. No Multicollinearity

- Check: Calculate the Variance Inflation Factor (VIF) for each predictor variable.
- Validation: A VIF above 5 (or in some cases, 10) suggests high multicollinearity, which can make coefficient estimates unreliable. Dropping correlated predictor can help reduce multicollinearity.



```
 Features    VIF
    const  62.37
      hum   1.81
    Clear   1.60
     temp   1.54
      Aug   1.44
   Summer   1.37
   Winter   1.31
      Sep   1.21
windspeed   1.16
       yr   1.03
  holiday   1.02
```

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 5 goes below this line> (Do not edit)

Linear Regression equation for Model 3:

cnt = 0.2075 + (0.2307 * yr) - (0.0897 * holiday) + (0.5378 * temp) - (0.2227 * hum) - (0.2113 * windspeed) + (0.1081 * Summer) + (0.1428 * Winter) + (0.0610 * Aug) + (0.1225 * Sep) + (0.0592 * Clear)

The top 3 features contributing significantly towards explaining the demand for shared bikes in this model are determined by the absolute values of their coefficients, as larger coefficients indicate a greater influence on the demand (cnt).

Based on the coefficients:

1. Temperature (temp): 0.5378 – This is the largest positive coefficient, indicating that an increase in temperature strongly increases the demand for bikes.
2. Year (yr): 0.2307 – This coefficient suggests that the demand for bikes has increased significantly over time (assuming yr represents different years).
3. Humidity (hum): -0.2227 – This is a negative coefficient, showing that higher humidity reduces bike demand.

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Linear regression is classified as a supervised learning algorithm that forecasts a value in the range of one or more features. As a term, linear regression is straightforward and in its simplest case, the model is linear in attempting to explain the linear relationship between a dependent and an independent variable. The outcome variables can be regarded as dependent variables, and the variable that is not affected can be seen as the independent or predictor variable.

1. Understanding the Linear Regression Model

In simple linear regression, where there’s only one input feature x, the model assumes a linear relationship:

$$y = \beta_0 + \beta_1 * x + \epsilon$$

- y is the target variable.
- x is the predictor variable.
- β_0 is the intercept.

- β_1 is the slope coefficient
- ϵ is the error term, which accounts for the randomness or other factors not explained by the model.

For multiple linear regression, with multiple predictors $x_1,\ x_2,\ x_3,\ \dots,\ x_n$:

$$y = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \dots + \beta_n * x_n$$

2. Objective of Linear Regression

The goal is to find the best-fitting line or hyperplane that minimizes the difference between the predicted values (y^) and the actual values of y. This is achieved by minimizing the sum of squared residuals, also known as Ordinary Least Squares (OLS).

The objective function is:

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

where:

- m is the number of observations.
- y_i is the actual target value.
- $\hat{y}_i$ is the predicted value.

3. Fitting the Model (Finding the Coefficients)

Linear regression solves for the coefficients $\beta 0,\ \beta 1,\ \dots,\ \beta n$ that minimize the sum of squared errors. This can be done using Gradient Descent, which iteratively updates coefficients by calculating the gradient of the loss function with respect to each coefficient.

4. Evaluating the Model

Common metrics used to evaluate linear regression models include:

- Mean Squared Error (MSE): The average of squared errors.
- Mean Absolute Error (MAE): The average of absolute errors.
- R-squared (R^2): Represents the proportion of variance explained by the model, with values closer to 1 indicating a better fit. Eg: A model having 0.89 as its R-squared value, means, the model can explain 89% of the variance present in the data.

5. Assumptions of Linear Regression

For linear regression to be reliable, certain assumptions about the data must be taken into consideration:

- Linearity: The relationship between predictors and target is linear.
- Independence: Observations are independent of each other.

- Homoscedasticity: The variance of residuals is constant across all levels of the independent variables.
- Normality of Errors: Residuals should follow a normal distribution.
- No Multicollinearity: Predictors should not be highly correlated with each other.

6. Interpretation of the Coefficients

Holding all other variables constant, each coefficient β_1 represents the estimated change in the target variable y for a one-unit change in x_1.

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

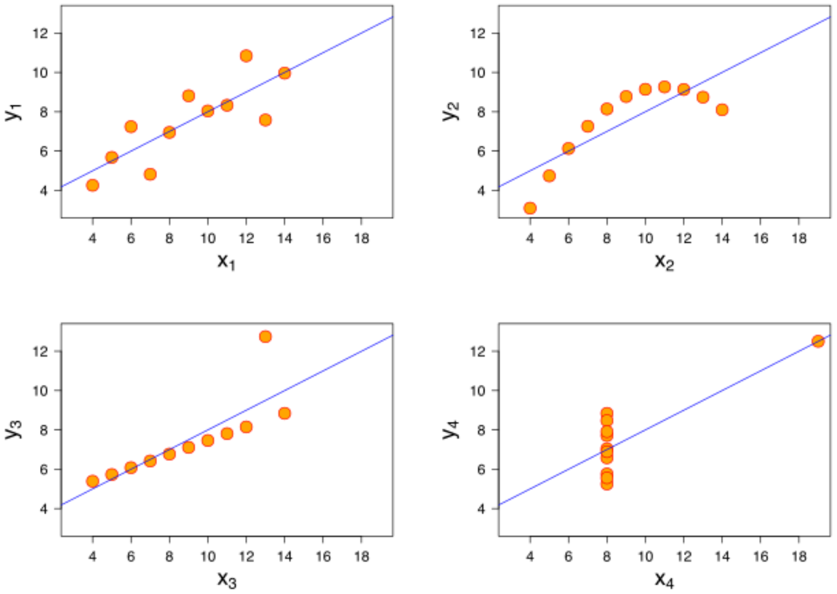
Anscombe's quartet is a set of four datasets that have nearly identical statistical properties, such as mean, variance, correlation, and linear regression line. However, when visualized, they reveal vastly different patterns and relationships. This quartet was created by Francis Anscombe in 1973 to demonstrate the importance of data visualization and the limitations of relying solely on summary statistics.

1. Identical Statistical Properties: Despite their distinct appearances, the four datasets share the same mean, variance, correlation coefficient, and linear regression line. This highlights how misleading summary statistics can be when used without visual exploration.

   Each of the four datasets has the following nearly identical statistical properties:

   - Mean of x: 9
   - Mean of y: 7.5
   - Variance of x: 11
   - Variance of y: 4.12
   - Correlation between x and y: ~ 0.82
   - Linear regression line: y= 3+0.5x

2. Diverse Visual Patterns:
   - Dataset I: This dataset exhibits a clear linear relationship between the x and y variables.
   - Dataset II: This dataset shows a quadratic relationship, where the y values increase and then decrease with increasing x values.
   - Dataset III: This dataset appears to have a strong linear relationship, except for one outlier point that significantly influences the regression line.
   - Dataset IV: This dataset has a constant x value, except for one outlier point. The y values vary randomly around a constant mean.



Anscombe's quartet underscores the importance of visualizing data before drawing conclusions. While summary statistics can provide a quick overview, they often fail to capture the nuances and underlying patterns within a dataset. By visualizing the data, we can uncover hidden relationships, identify outliers, and make more informed decisions.

Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Pearson's r, also known as the Pearson correlation coefficient or simply correlation coefficient, is a statistic that measures the strength and direction of a linear relationship between two continuous variables. It provides a way to understand how two variables move together.

The formula for Pearson's r

The formula for calculating Pearson's r between two variables X and Y is:

$$r = \sum(X_i - \bar{X})(Y_i - \bar{Y}) \,/\, \sqrt{\sum(X_i - \bar{X})^2(Y_i - \bar{Y})^2}$$

where:

- X_i, Y_i are individual data points of variables X and Y,
- $\bar{X}$, $\bar{Y}$ are the mean values of X and Y,
- $\sum(X_i - \bar{X})(Y_i - \bar{Y})$ is the covariance of X and Y,
- The denominator scales the result, normalizing it to be within the range of -1 to 1.

Pearson's r has values ranging from -1 to 1, indicating both the strength and direction of the relationship:

- r=+1: Perfect positive linear relationship – as one variable increases, the other also increases proportionally.
- r=−1: Perfect negative linear relationship – as one variable increases, the other decreases proportionally.
- r=0: No linear relationship – the variables are uncorrelated.

The closer r is to +1 or -1, the stronger the linear relationship between the variables. Values near 0 indicate a weak or no linear relationship.

Example ranges for interpretation:

- 0.7 to 1.0 (or -0.7 to -1.0): Strong positive (or negative) correlation.
- 0.3 to 0.7 (or -0.3 to -0.7): Moderate positive (or negative) correlation.
- 0 to 0.3 (or -0.3 to 0): Weak or negligible correlation.

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Scaling is a data preprocessing technique used in machine learning and statistics to adjust the range of values in a dataset. This ensures that all features contribute proportionally to the model and prevents large-magnitude features from dominating smaller-magnitude features. Scaling helps in improving model performance, training speed, and convergence, especially in algorithms sensitive to feature magnitudes.

Scaling is performed because of the following:

1. Faster Convergence: Scaling helps gradient descent algorithms converge faster since features on a similar scale allow gradients to behave more predictably.
2. Prevents Dominance by Large Magnitude Features: If features have different scales (e.g., age in years vs. income in thousands), features with a larger magnitude may disproportionately influence the model, causing biases.

Normalization (or **min-max scaling**) rescales the features to a fixed range, usually between 0 and 1 or -1 and 1. This technique is helpful when the distribution is not Gaussian or when we want to keep values within a specific boundary.
Standardization (or **z-score scaling**) transforms features to have a mean of 0 and a standard deviation of 1. This makes the data follow a standard normal distribution (centered around zero with unit variance).

Difference Between Normalization and Standardization

Property	**Normalization (Min-Max Scaling)**	**Standardization (Z-Score Scaling)**
Range	Scales values between a fixed range (e.g., 0 to 1)	Centered around 0 with unit variance
Sensitive to Outliers	Yes	No (more robust)

Application	Non-Gaussian distributions; distance-based models	Gaussian-distributed features; linear models and SVMs

---

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

A Variance Inflation Factor (VIF) becomes infinite when there is perfect multicollinearity between a predictor and one or more other predictors in a regression model. Multicollinearity refers to a situation where two or more independent variables are highly correlated, which can lead to instability in the model's coefficient estimates.

The formula for calculating the VIF of a predictor variable X is:

$$VIF\ (X_{\mathrm{i}}) \ = \ 1 \, / \, 1 \ - \ R_{\mathrm{i}}^{2}$$

where R_{i}^2 is the coefficient of determination from regressing X_{i} on all the other predictors.

- When there is perfect multicollinearity, can be perfectly predicted by a linear combination of other predictors. This causes R_{i}^2 to be exactly 1.
- Since the formula includes $1 - R_{\mathrm{i}}^2$ in the denominator, an R_{i}^2 of 1 leads to division by zero, resulting in an infinite VIF.

VIF can be infinite due to the following reasons:

1. Perfect Linear Dependence: When one predictor is an exact linear combination of other predictors.
2. Dummy Variable Trap: In cases with categorical variables, including all dummy variables (without dropping one) for each category can lead to perfect multicollinearity and infinite VIF.

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

A Q-Q plot, or quantile-quantile plot, is a valuable tool in linear regression to assess the normality assumption of the model's residuals. In linear regression, we assume that the residuals (the differences between the observed and predicted values) are normally distributed.

A Q-Q plot helps us visually check this assumption. It compares the quantiles of the observed residuals to the quantiles of a theoretical normal distribution. If the residuals are indeed normally distributed, the points on the Q-Q plot should roughly follow a straight line. Deviations from this line indicate departures from normality.

By examining the Q-Q plot, we can identify potential issues like:

- Non-normality: If the points significantly deviate from the line, it suggests that the residuals may not be normally distributed. This can affect the validity of hypothesis tests and confidence intervals.
- Heavy tails: If the tails of the distribution are heavier than a normal distribution, it can indicate outliers or extreme values that might influence the model's performance.
- Skewness: If the plot shows a curved pattern, it suggests that the distribution is skewed, which can also impact the model's assumptions.

By using Q-Q plots, we can assess the validity of the normality assumption and take appropriate steps to address any issues, such as transforming the data or using alternative modeling techniques.