Assignment-based Subjective Questions

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

Ans.) When we do a descriptive Analysis with respect to categorical variable on the effects of target variable we can infer the following the insights from the plots.

1. Season: 3: fall has highest demand for rental bikes

2. I see that demand for next year has grown

3. Demand is continuously growing each month till June. September month has highest demand. After September, demand is decreasing

4. When there is a holiday, demand has decreased.

5. Weekday is not giving clear picture about demand.

6. The clear weathershit has highest demand

7. During September, bike sharing is more. During the year end and beginning, it is less, could be due to extereme weather conditions.

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

Ans.) When creating dummy variables from categorical features, it is important to use `drop_first=True` to avoid the dummy variable trap.

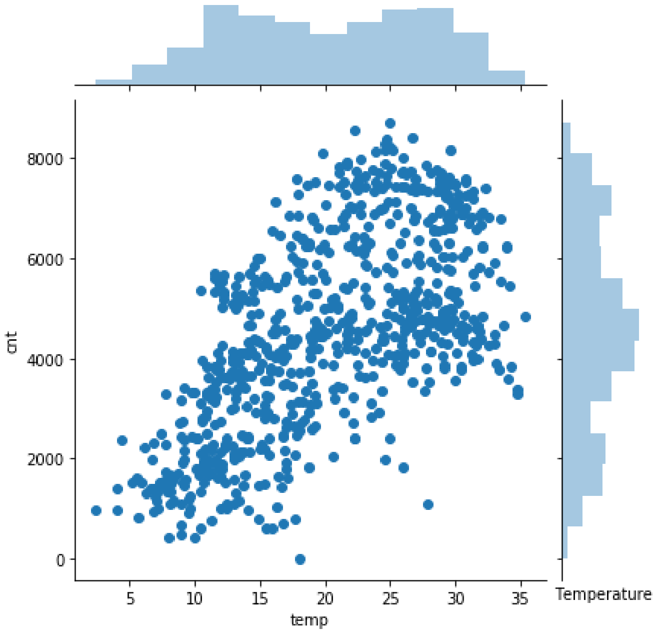
The dummy variable trap refers to the problem of including a dummy variable for each category of a categorical feature. For example, if we have a categorical feature called "**season**" with four categories **spring, summer, fall, winter**, we might create dummy variables called "season_summer", "season_spring", "season_fall" and “season_winter. However, if we include all four dummy variables in a regression model, we will encounter the dummy variable trap.

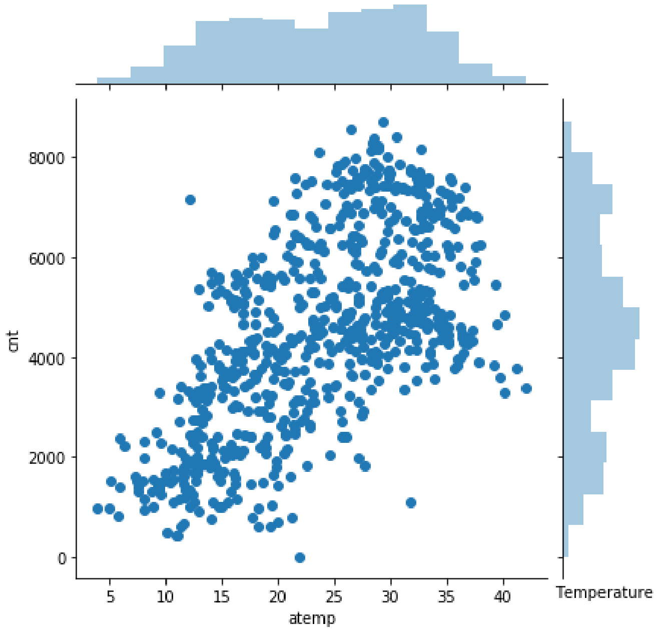
The dummy variable trap occurs because including all of the dummy variables can lead to perfect multicollinearity, which means that the independent variables are highly correlated with each other. This can lead to problems with the regression model, such as coefficients that are not statistically significant, unstable coefficient estimates, or incorrect predictions.

To avoid the dummy variable trap, we can use the `drop_first=True` argument when creating dummy variables. This will drop the first category of the categorical feature and create dummy variables for the remaining categories, By dropping the first category and creating dummy variables for the remaining categories, we avoid the dummy variable trap and improve the performance of the regression model.

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

Ans.) By Looking at the pairplot among the numerical variables, we can infer that the temp and atemp has the highest correlation with the target variable 'cnt'.



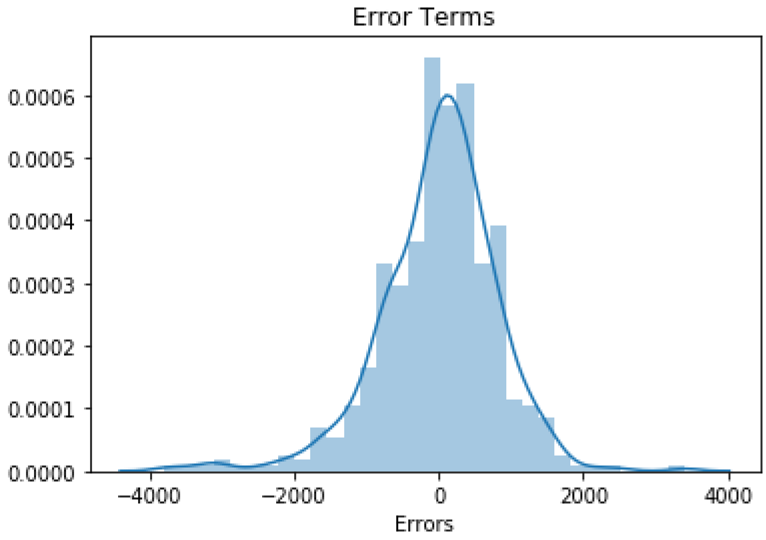


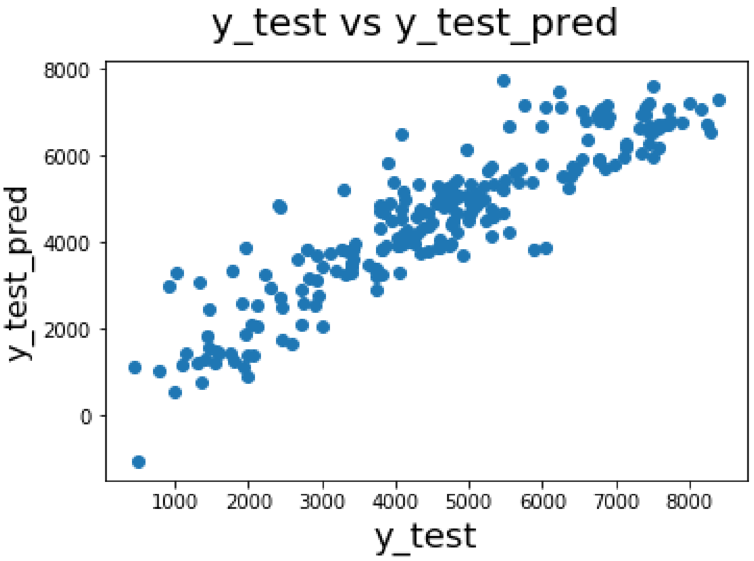
Q4.) How did you validate the assumptions of Linear Regression after building the model on the training set?

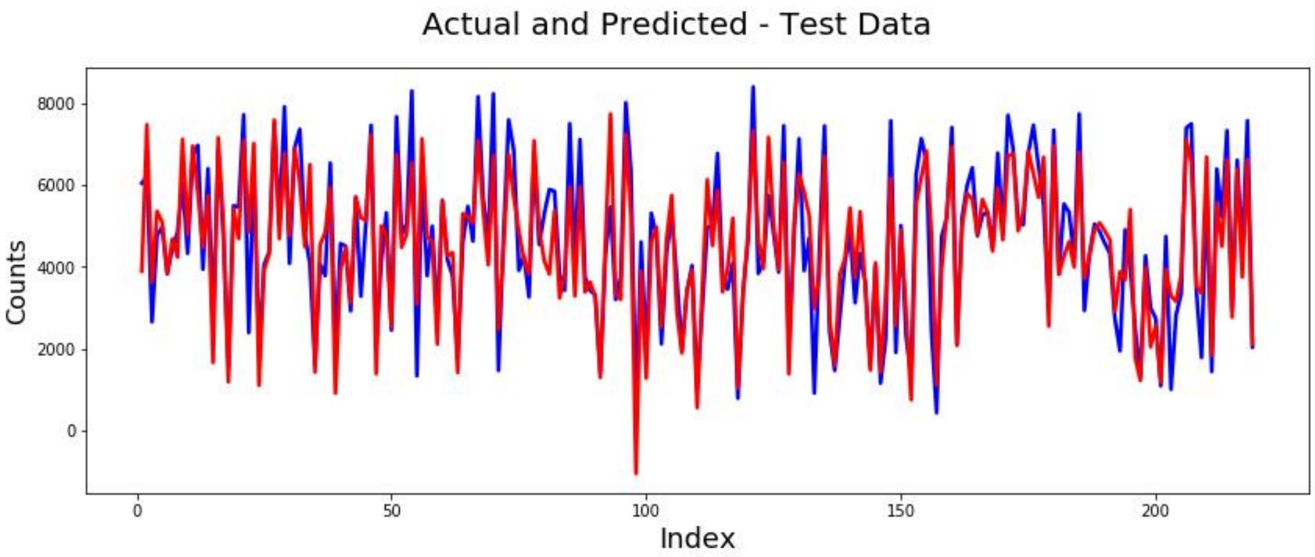
Ans.) After building a linear regression model on the training set, it is important to validate the assumptions of the model using the test set. Here are the main assumptions to be checked:

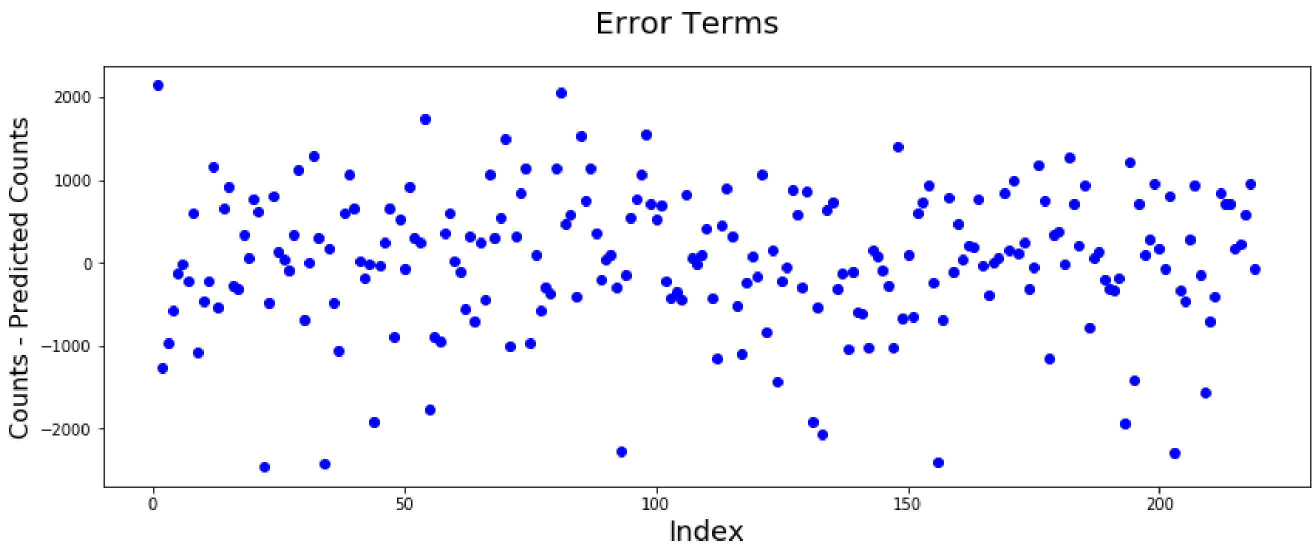
1.) Linearity: Check for a linear relationship between the independent and dependent variables by plotting the residuals against the predicted values. If there is a clear pattern in the plot, the relationship may not be linear.

2.) Homoscedasticity: Check for equal variances of the residuals across all levels of the independent variables by plotting the residuals against the predicted values. If there is a clear pattern in the plot, the variances may not be equal.


Error Terms
0.0000
0.0001
0.0002
0.0003
0.0004
0.0005
0.0006
−4000
−2000
0
2000
4000
Errors


y_test vs y_test_pred
8000
6000
4000
2000
0
y_test_pred
1000
2000
3000
4000
5000
6000
7000
8000
y_test





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

Ans.) Season, holiday, temp are the top significant contributing towards the demand of the shared bikes.

- Company should focus on expanding business during Spring.
- Company should focus on expanding business during September.
- Based on previous data it is expected to have a boom in number of users once situation comes back to normal, compared to 2019.
- There would be less bookings during Light Snow or Rain, they could probably use this time to serive the bikes without having business impact.

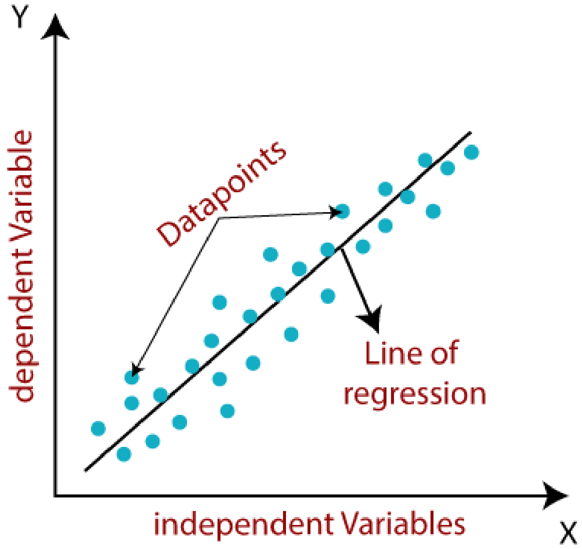
General Subjective Questions

Q1.) Explain the linear regression algorithm in detail.

Ans.) Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as **sales, salary, age, product price,** etc.

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

Linear regression is a statistical modeling technique used to establish a linear relationship between a dependent variable and one or more independent variables. The goal of linear regression is to find a line (or plane in higher dimensions) that best fits the data, so that we can use it to make predictions for new observations.



The algorithm for linear regression involves the following steps:

1. Data preparation: Collect data for the dependent variable and one or more independent variables. Clean and preprocess the data as needed.

2. Model specification: Define the model by specifying the dependent variable and the independent variables. The model can be represented by the equation y = β0 + β1x1 + β2x2 + ... + βnxn, where y is the dependent variable, x1, x2, ..., xn are the independent variables, and β0, β1, β2, ..., βn are the coefficients or parameters that determine the slope and intercept of the line.

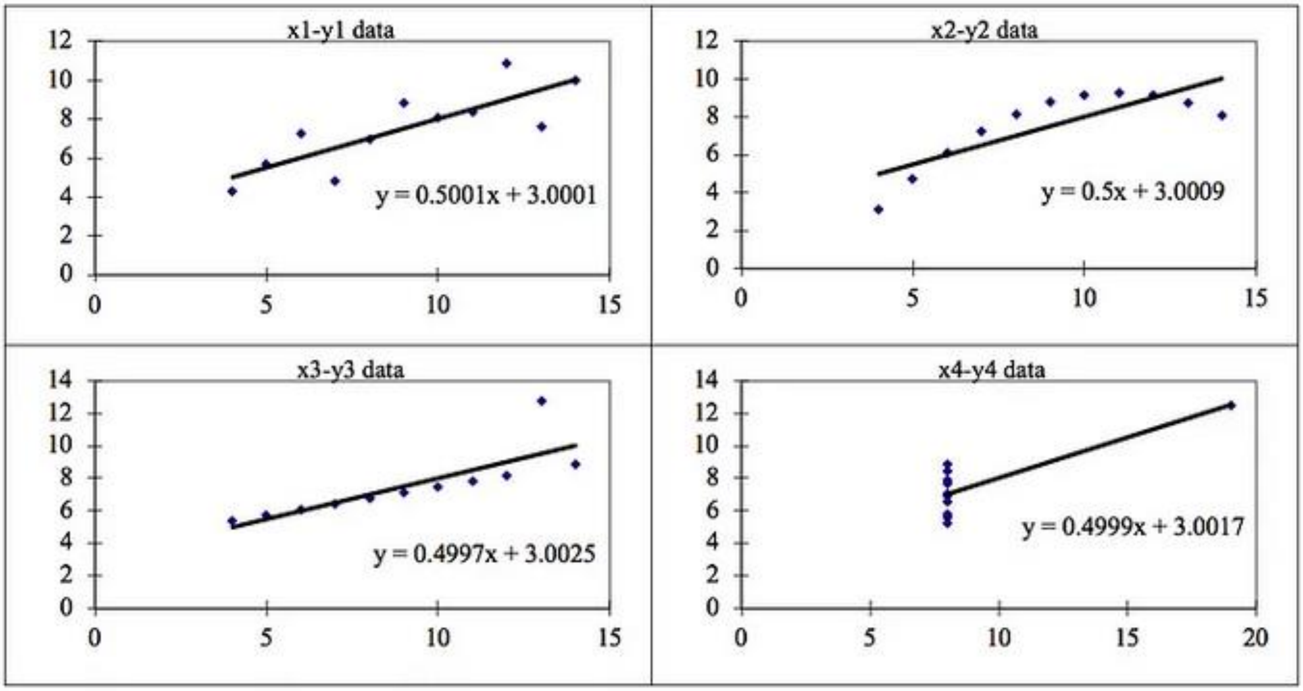
3. Parameter estimation: Estimate the values of the parameters β0, β1, β2, ..., βn that minimize the sum of squared errors between the predicted values and the actual values of the dependent variable.

4. Model evaluation: Evaluate the performance of the model by checking its assumptions and testing its accuracy on a validation dataset. This involves checking for linearity, normality, homoscedasticity, and independence of the residuals.

5. Prediction: Once the model is validated, use it to make predictions for new observations by substituting the values of the independent variables into the equation and calculating the corresponding value of the dependent variable.

Q2.) Explain the Anscombe’s quartet in detail.

Ans.) **Anscombe’s Quartet** can be defined as a group of four data sets which are **nearly identical in simple descriptive statistics**, but there are some peculiarities in the dataset that **fools the regression model** if built. They have very different distributions and **appear differently** when plotted on scatter plots.



The four datasets can be described as:

1. **Dataset 1:** this **fits** the linear regression model pretty well.

2. **Dataset 2:** this **could not fit** linear regression model on the data quite well as the data is non-linear.

3. **Dataset 3:** shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model

4. **Dataset 4:** shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model

Q3.) What is Pearson's R?

Ans.) Pearson's R, also known as the Pearson correlation coefficient, is a statistical measure of the linear relationship between two continuous variables. It is named after the British mathematician Karl Pearson, who developed the formula for calculating the coefficient.

Pearson's R ranges from -1 to 1, where a value of -1 indicates a perfect negative linear relationship, a value of 0 indicates no linear relationship, and a value of 1 indicates a perfect positive linear relationship between the two variables.

Pearson's R is calculated by dividing the covariance of the two variables by the product of their standard deviations. This gives a measure of the proportion of the total variability in one variable that can be explained by the other variable.

Pearson's R is widely used in statistical analysis to measure the strength and direction of the relationship between two variables. It is especially useful in regression analysis, where it is used to identify the independent variables that are most strongly related to the dependent variable. Pearson's R is also used in many other fields, such as social sciences, engineering, and finance, to investigate the relationships between different variables.

Q4.) What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans.) Scaling is a preprocessing technique that is commonly used in machine learning to normalize or standardize the range of data. It is an important technique because many machine learning algorithms rely on the similarity of scales across different variables, and scaling can help to improve the accuracy of these algorithms.

The two main types of scaling techniques are normalized scaling and standardized scaling. Normalized scaling is used to transform the data into a range between 0 and 1, with the minimum value of the dataset becoming 0 and the maximum value becoming 1. This is done by subtracting the minimum value of the dataset from each observation and then dividing the result by the range of the dataset. Normalized scaling is useful when the dataset has a known minimum and maximum value and the goal is to preserve the shape of the distribution.

Standardized scaling, on the other hand, involves transforming the data so that it has a mean of 0 and a standard deviation of 1. This is done by subtracting the mean value of the dataset from each observation and then dividing the result by the standard deviation of the dataset. Standardized scaling is useful when the dataset has an unknown range, or when the goal is to compare variables that are measured on different scales. Standardized scaling is also helpful when the algorithm used for analysis requires variables to have a normal distribution.

Scaling can have a significant impact on the performance of machine learning algorithms. If a dataset has variables measured on different scales, some variables may end up dominating the analysis, while others may have little to no impact. By scaling the variables, each feature contributes equally to the analysis, making it easier for machine learning algorithms to make accurate predictions.

In summary, scaling is a preprocessing technique used to transform the data into a similar scale, ensuring that each feature contributes equally to the analysis. Normalized scaling and standardized scaling are two common methods of scaling, with the former preserving the distribution of the data and the latter standardizing it to a mean of 0 and standard deviation of 1. Scaling is an important technique in machine learning and can have a significant impact on the accuracy of algorithms.

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

Ans.) The variance inflation factor (VIF) is a measure used to assess the degree of multicollinearity between predictor variables in regression analysis. Multicollinearity occurs when two or more predictor variables are highly correlated with each other, making it difficult to estimate the independent effect of each variable on the outcome. High VIF values indicate a high degree of multicollinearity, which can lead to unreliable regression coefficients and inaccurate predictions.

Sometimes, the VIF value for a predictor variable can be infinite. This occurs when there is perfect multicollinearity between two or more predictor variables, meaning that they are perfectly linearly related to each other. This results in the regression model being unable to estimate the coefficients of the variables separately, leading to a divide-by-zero error in the calculation of the VIF.

Perfect multicollinearity can arise for several reasons, such as data entry errors, using derived variables that are linear combinations of other variables, or using multiple variables that measure the same underlying construct. For example, if a regression model includes both temperature in Celsius and Fahrenheit scales, they will be perfectly correlated with each other, resulting in perfect multicollinearity and an infinite VIF value for one of the variables.

Detecting and resolving multicollinearity is important in regression analysis, as it can lead to biased estimates, reduced statistical power, and overfitting. If perfect multicollinearity is present in a model, it is important to identify the correlated variables and remove one of them from the analysis. One way to identify multicollinearity is to calculate the VIF values for each predictor variable and remove any variables with a high VIF value. If the VIF value is infinite, it indicates that there is perfect multicollinearity between the predictor variables and further investigation is needed to resolve the issue.

Q6.) What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans.) A Q-Q plot (quantile-quantile plot) is a graphical technique used to determine if a dataset is normally distributed. It compares the distribution of a sample to a theoretical normal distribution by plotting the observed values against the expected values.

In linear regression, a Q-Q plot is an important tool for assessing whether the residuals (the differences between the predicted and observed values) are normally distributed. Normality of residuals is an important assumption of linear regression, as violating this assumption can lead to biased or incorrect estimates of the model parameters.

The Q-Q plot helps to assess the normality assumption by comparing the distribution of the residuals to a normal distribution. If the residuals are normally distributed, the points on the Q-Q plot should form a straight line. If the residuals deviate significantly from a straight line, this suggests that the normality assumption may be violated and further investigation may be necessary.

The Q-Q plot is also useful for identifying outliers and skewness in the data. Outliers are points that deviate significantly from the expected values and may indicate errors or anomalies in the data. Skewness, or a lack of symmetry in the distribution, can affect the accuracy of the regression model and may require further analysis or transformation of the data.

Overall, a Q-Q plot is a valuable tool for assessing the normality assumption and identifying potential issues in the data. It allows the researcher to make informed decisions about the validity and reliability of the regression model and to identify any areas that may require further investigation or refinement.