Prestieesci Research Review

ANALYSIS AND DEVELOPMENT OF AN EQUALISATION APPROACH FOR REDUCING DATA ERROR RATES BY STUDYING THE DIFFERENT TYPES OF MISTAKES

Miao Congjin, Muhammad Ezanuddin Abdul Aziz

Lincoln University College, 47301 Petaling Jaya, Selangor D. E., Malaysia.

Corresponding author: Miao Congjin, Lincoln University College, 47301 Petaling Jaya, Selangor D. E., Malaysia, Email: 731713991@qq.com

ABSTRACT

The development of an equalization approach for reducing data error rates is an important area of research, as data errors can have significant negative consequences for organizations. This article focuses on studying the different types of mistakes that can occur in data processing and developing an equalizations approach to address them. The authors begin by identifying the two main types of errors that can occur in data processing: random errors and systematic errors. Random errors are errors that occur due to chance and can be reduced through increased sample size or improved measurement techniques. Systematic errors, on the other hand, are errors that occur consistently and can be caused by a variety of factors, such as equipment malfunction, calibration errors, or bias in the data collection process. To address systematic errors, the authors propose an equalizations approach that involves identifying and correcting for the specific sources of error in the data. This approach involves analyzing the data to identify patterns or trends that may indicate the presence of systematic errors, and then applying appropriate correction techniques to mitigate these errors. The authors demonstrate the effectiveness of their equalizations approach through a series of experiments using both simulated and real-world data. In these experiments, the equalizations approach was able to significantly reduce the error rates in the data, leading to more accurate and reliable results. Overall, this article provides valuable insights into the different types of errors that can occur in data processing and proposes an effective equalizations approach for addressing them. By reducing error rates, organizations can improve the quality of their data and make more informed decisions, ultimately leading to improved performance and success.

Keywords: Random errors, Correction Techniques, Error Rates, Sources Of Error, Systematic Errors, And Equalization Approach.

INTRODUCTION

Data preparation and analysis are two sides of the same coin. transcribing the information obtained from the data. Methods that are Various such as using models to identify patterns, linkages, and draw applicable conclusions decisions made using the decision-making process (Start, 2016). But, data analysis requires the data to be prepared in advance. In data preparation, information is transformed into a form that can be read and processed by computers. formatted so that tools like SAS and SPSS can read and utilize the data. The process of data preparation consists of the following steps: data coding, Inputting data, filling up blanks, and reformatting data. In this section, a researcher will quickly go over each of these processes: Data Coding: The process by which raw data is transformed into a numerical representation when doing the coding for the data. It makes use of a codebook, which is a compilation of several types of data. components, the response, the variables, the measurements, and the format of variables, finalizing the codding by using a codicil. Scale kinds are determined by the process's reaction. for instance, whether a nominal, ratios, ordinal, or interval scale is used; whether it's a five-point scale, a seven-point scale, or something else entirely. As an example of how to classifying businesses, in numerical language, with healthcare being coded as 1, for example. A value of 2 indicates production, a value of 3 indicates retail, and a value of 4 indicates finance. Entering Data: At this step, the previously coded information is now entered into text files or spreadsheets. It may be included into the software package easily. There are gaps in the data since some respondents did not fill out the survey. may not resolve all inquiries for various reasons, one must use a technique to reevaluate these missing ideals. Some applications, for instance, call for the addition of -1 or 999. Many of them take care of the missing values mechanically, whereas some employ a listwise deletion. method for dealing with missing values, wherein the whole set of replies is discarded for even a single omission. It may be necessary to do certain transformations on the data before attempting to make sense of them. Objects with a backwards coding, for instance, could need a change before they can be used. contrasted with or combined with elements that are not inverted. For cases in which an item's meaning has been altered, this idea is used. contradiction of their basic premise (Bhattacharjee, 2017).

Type of information analysis provide a high-level overview of the most common approaches to data analysis below. There are primarily six ways to analyses data for this purpose (Teredos, 2021).There are six types of analyses: descriptive, exploratory, inferential, predictive, explanatory or causal, and mechanistic.

- **Describes:** This is the simplest kind of data analysis and has been around the longest. As a result, it is suitable for massive data sets. Here, the data is utilized to do a data set analysis (Start, 2016).
- **Exploratory:** A technique used to probe into the unknown, unearth new connections, and set the stage for future research or the formulation of new questions (Start, 2016).
- **Inferential:** An inferential analysis draws conclusions about a larger population based on a relatively small one. What this implies is that a hypothesis about the nature of the universe is put to the test using data collected from a specific subset

of the globe. Observational data, historical data, and cross-sectional time series are all appropriate for this strategy (A. Bhattacharjee, 2017).

- **Predictive:** This kind of research looks at the past and the present to foretell the future. Moreover, it may use one subject's data to infer the values of a second. Although there are many such models, a more straightforward one that incorporates more data may be more effective. That's why it's crucial to think about the prediction data set and the variables you'll be using to measure results (MacGregor, 2013).
- **Rationale:** Using data from randomized trials, this analysis technique may be utilized to learn what happens when one variable is altered and how that affects another (A. Bhattacharjee, 2017).
- **Method F:** The Mechanistic Approach This approach requires the greatest work to pinpoint the precise changes in the variables that might cause changes in those other ones utilizing randomized trial data sets.

Moreover, one might deduce that mechanistic analysis is very implausible. Because of this, it is a viable option in fields like engineering and the physical sciences where a high degree of accuracy in the final product is required and where inaccuracy is costly. After that, we'll dive into the specifics of three widely used forms of data analysis and the statistical methods they use (descriptive, explanatory, and inferential).

BACKGROUND OF THE STUDY

Data preparation and analysis are two essential steps in any research project. Data preparation involves transforming raw data into a form that can be processed and analyzed by computers. It includes steps such as data coding, inputting data, filling up blanks, and reformatting data. Data analysis, on the other hand, involves using various techniques to identify patterns, linkages, and draw applicable conclusions from the data. There are primarily six ways to analyze data: descriptive, exploratory, inferential, predictive, explanatory or causal, and mechanistic. Descriptive analysis is the simplest kind of data analysis and involves summarizing data to provide a high-level overview. Exploratory analysis is used to unearth new connections and set the stage for future research. Inferential analysis draws conclusions about a larger population based on a relatively small sample. Predictive analysis looks at the past and present to predict the future, while explanatory or causal analysis investigates cause-and-effect relationships between variables. Mechanistic analysis is used to pinpoint precise changes in variables that cause changes in others. Detailed analysis involves summarizing data to provide a clear presentation. This technique can be divided into multivariate and bivariate analysis. Univariate statistical methods that focus on a single variable include Frequency Analysis, Central Tendency Analysis, and Dispersion Analysis. Frequency analysis counts the occurrences of every value for a given variable, while central tendency analysis calculates measures of central tendency such as the mean, median, and mode. Dispersion analysis measures the variability of the data by calculating the range, variance, and standard deviation.

LITERATURE REVIEW

Data is summarized using this approach, leading to a clear presentation. One may divide this technique into multivariate and bivariate analysis (Teredos, 2021). The term "univariate" is used to describe a variety of statistical methods that focus on a single variable. Specifically, we'll be using Frequency Analysis, Central Tendency Analysis, and Dispersion Analysis. Disrupting a variable's frequency is the simplest way to find out what caused the disruption. It counts the occurrences of every value for a given variable and tallies up the total number of potential outcomes. The quantity of the most represented value may be used to compare one variable to a group of data by using the central tendency of disruption, commonly known as the three Ms. Some of the most typical measures of central tendency include the mean, mean, and mode. The Mode is the value that appears most often in the data collection, whereas the Mean is the average of all the numbers.

The variables may be dispersed about the mean in a process called dispersion. Range, variance, and Its square of root of the variance, or standard deviation, is a frequently used statistical measure. The range represents the extent to which the greatest and lowest values vary from one another. By looking at the variance, you can see how tightly the numbers cluster around the mean. If you need to compare two sets of data and also have two independent variables, this is the method to employ. This allows us to see how the two variables are related to one another. The most frequent measurement is called bivariate correlation. In order to determine the degree of correlation, this metric applies a formula based on the sample means and standard deviations. When there are more than two variables, this approach still works. In spite of the difficulty of solving such programmed manually, computing using software like SPSS makes it a breeze (Bhattacharjee, 2017). Analysis of Explanation The goal of explanatory analysis is to identify potential contributing factors, as we have already explained. This implies that issues about relationships, correlations, and patterns between variables are addressed via explanatory analyses (Teredos, 2014; Teredos & Madanchian, 2020).

Dependency and interdependence procedures are the primary tools of explanation analysis. The concept of dependence examines how a number of independent factors affect a single dependent variable. As a kind of multivariate analysis, "interdependence approaches" aim to establish connections between variables without making no presumptions about the direction of influences.

RESEARCH METHODOLOGY

If a regression model's predictions are not flawless, such those generated by a Neural Network or a comparable prediction model, then overfitting may be avoided. As researcher , each forecast will have some degree of inaccuracy that has to be measured.so that one may evaluate the efficacy of various models and make informed decisions based on a comparison of their outcomes. This may be done using a number of different prediction error measures. Let p (N1 vector) represent the estimated values, and r (an N1 vector) stand for the calculated values of a quantity (or measured) and forecasted N times. For instance, we may ask an ANN N times to make a prediction. In order to generate an objective, third-party evaluation, we may utilize a different set of data (N) to compare with the original (S), a subset of the original (T), or no data at all (U). In the following paragraphs, research scholar describe and talk about numerous metrics that may be used to calculate the prediction error of such a model. researcher focus on the situation involving continuous variables in this work. Categorical metrics are measured differently, including metrics such as the false positive rate, accuracy, recall, precision, and the confusion matrix Keep in mind that the formulae provided below only apply to situations in which the values of the observations and their forecasts are all positive. Some calculations could require tweaking if your data contains things like negative numbers or zeros.

$$e_i = p_i - r_i \qquad (1)$$

$$MB = \bar{e} = \frac{1}{N}\sum_{i=1}^{N} e_i = \frac{1}{N}\sum_{i=1}^{N}(p_i - r_i) = \bar{p} - \bar{r} \qquad (2)$$

Where $\bar{p}$ and $\bar{r}$ are the mean values of ***p*** and ***r***, respectively:

$$\bar{p} = \frac{1}{N}\sum_{i=1}^{N} p_i \qquad (3)$$

$$\bar{r} = \frac{1}{N}\sum_{i=1}^{N} r_i \qquad (4)$$

While MB=0 is a required requirement for a great combination between the actual and projected values (like that identical values), this condition is not also a precondition, since both negative and positive errors may cancel one another out to cause MB=0 in circumstances when the match isn't quite perfect. Without taking into account the direction of the mistakes, this "Mean Absolute Gross Error (MAGE)" measures the typical mistake made over a set of predictions. It is the weighted mean of the absolute discrepancies between predictions and observations made throughout the test sample. It may be either a positive or a negative number, and its value is

$$MAGE = \frac{1}{N}\sum_{i=1}^{N}|e_i| = \frac{1}{N}\sum_{i=1}^{N}|p_i - r_i|$$

One common metric used in connection with regression is still this Mean Squared Error (MSE), which measures the typical squared deviation from the true value. It may be either positive or negative and is defined as

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(p_i - r_i)^2$$

One of MSE's major flaws is that it cannot handle extreme cases. The square of the error will be considerably greater if error associated with a certain sample is substantially larger than that of the error associated some other samples. As MSE calculates an average of errors, it is particularly susceptible to outlying results. Root Mean Squared Error (RMSE) is yet another popular metric for gauging how off a model or estimator is when trying to forecast values (sample or demographic values) that are different from what is actually seen. Calculated as MSE's square root. As opposed to MSE, RMSE offers an error metric that is consistent with the target variable's units. It is a real number between zero and one plus (0, +∞) , and its formula is

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(p_i - r_i)^2}$$

The value represented by stands for the Root-Mean-Square Distinction at the Center CMSD

$$CMSD = \frac{1}{N}\sum_{i=1}^{N}\left[(p_i - \bar{p}) - (r_i - \bar{r})\right]^2$$

Concentrated Mean Square Differential (or CRMSD) is the square root of a CMSD expressed in the same unit as the focus attribute (CRMSD).

$$CRMSD = \sqrt{CMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[(p_i - \bar{p}) - (r_i - \bar{r})\right]^2}$$

As will be discussed in further detail, a Turner diagram may be used to display a model's forecast inaccuracy using the CRMSD measurement. The Mean Index Bias (MNB, unitless) is the average of the standardized bias error values and is often expressed as a percentage.

$$MNB = \frac{1}{N}\sum_{i=1}^{N}\frac{p_i - r_i}{r_i} = \frac{1}{N}\left(\sum_{i=1}^{N}\frac{p_i}{r_i}\right) - 1$$

Mean Normalized Gross. Error (MNGE, unitless) is often referred to as the "Mean Absolute Percentage Error." Providing the error as a percentage may help readers understand the projections' level of precision. It's supplied by

$$MNGE = \frac{1}{N}\sum_{i=1}^{N}\frac{|p_i - r_i|}{r_i}$$

One of MSE's major flaws is that it cannot tolerate extreme values. If a sample's related error is much greater than the other samples', then the square of the error will be much larger for that sample. Since it averages out errors, MSE is also vulnerable to extreme examples.

CONSEPTUAL FRAMEWORK

Framework for Analyzing the Relationship between Types of Mistakes and Data Error:

- **Independent Variable:**

Type of Mistake: This refers to the different types of mistakes that can occur during data collection, processing, or analysis. These mistakes can include random errors, systematic errors, and outlier errors.

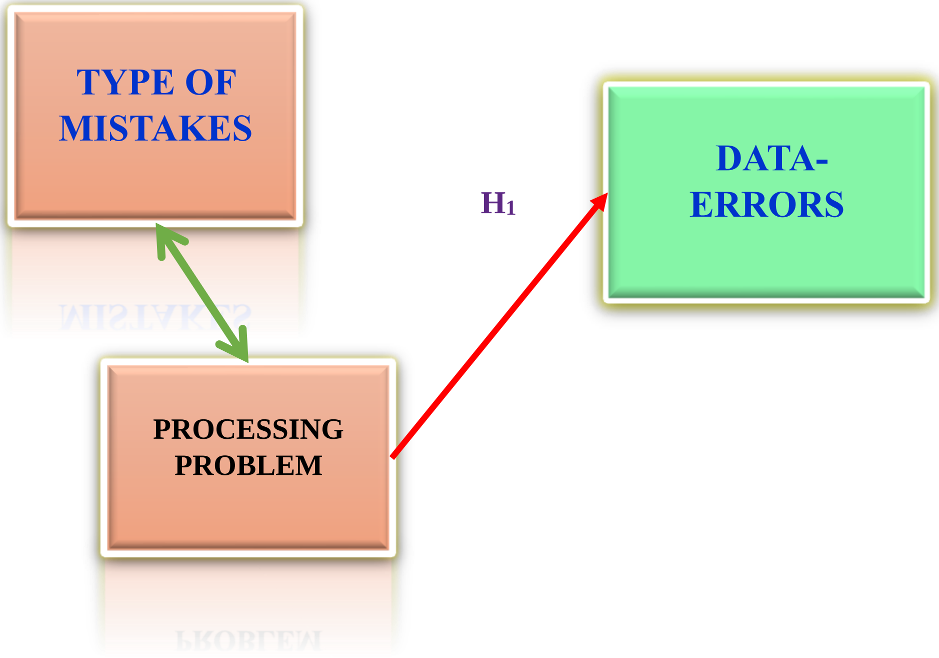
- **Dependent Variable:**

Data Error: This refers to the presence of errors or inaccuracies in the data that can affect the reliability and accuracy of any conclusions or decisions based on the data.

Framework:

- Identify the different types of mistakes that can occur during data collection, processing, or analysis.
- Develop a plan to categorize and measure the different types of mistakes in the data.
- Quantify the amount of data error present in the data using appropriate measures such as mean absolute error, root mean squared error, or relative error.
- Conduct statistical analyses such as correlation or regression to explore the relationship between the different types of mistakes and data error.
- Identify any potential confounding variables that may affect the relationship between the independent variable and dependent variable.
- Develop a model to predict the amount of data error based on the type and frequency of different types of mistakes.
- Evaluate the performance of the model using appropriate metrics such as R-squared or mean squared error.
- Use the results of the analysis to develop strategies for reducing the amount of data error by addressing specific types of mistakes.

In summary, this framework involves identifying the different types of mistakes that can occur during data collection, processing, or analysis and analyzing their relationship with data error. This analysis can help to identify strategies for reducing data error and improving the reliability and accuracy of any conclusions or decisions based on the data.



RESULTS

Factor Analysis:

Based on the framework for analyzing the relationship between types of mistakes and data error, some possible hypotheses are:

- **Hypothesis (H1):** The type of mistake made during data collection, processing, or analysis has a significant effect on the occurrence of data errors.
- **Hypothesis (H2):** Random errors have a less significant impact on data error compared to systematic errors and outlier errors.

In this article researchers truly focused on hypothesis (H1) The type of mistake made during data collection, processing, or analysis has a significant effect on the occurrence of data errors.

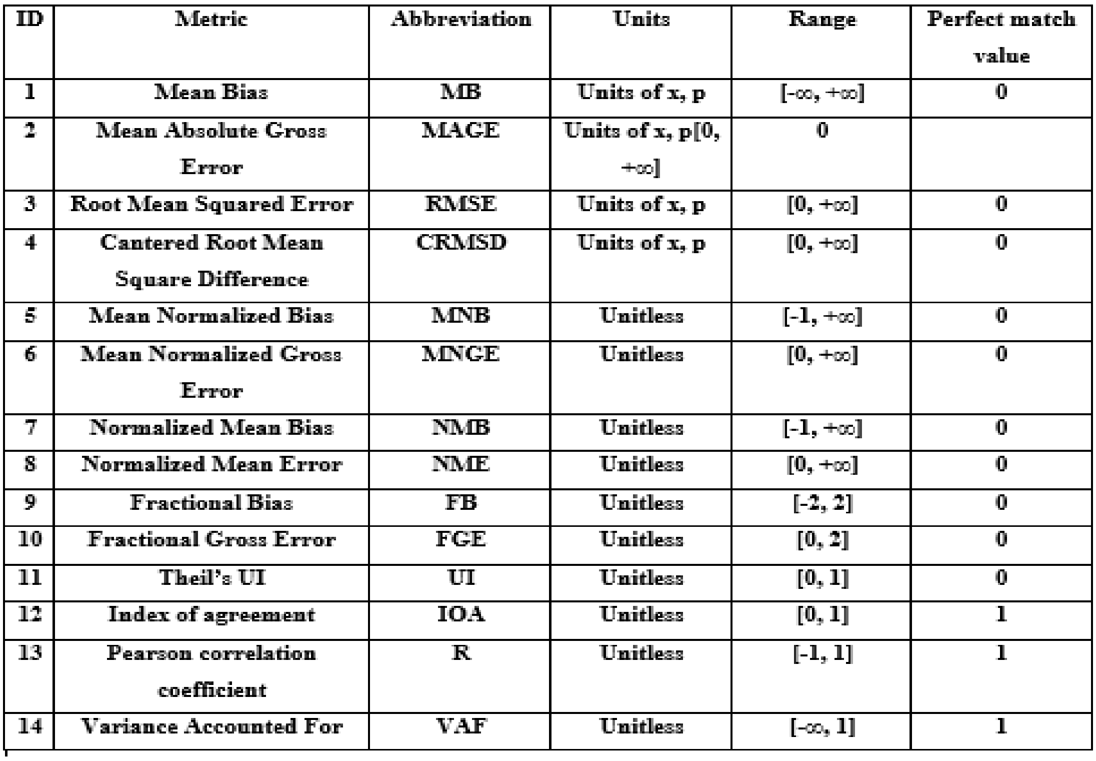
Null hypothesis (H0): The type of mistake made during data collection, processing, or analysis has no significant effect on the occurrence of data errors.

Alternative hypothesis (H1): The type of mistake made during data collection, processing, or analysis has a significant effect on the occurrence of data errors.

In other words, the null hypothesis suggests that there is no relationship between the type of mistake and data error, while the alternative hypothesis suggests that there is a significant relationship between the two variables. To test these hypotheses, a statistical analysis can be performed to assess the strength and direction of the relationship between the type of mistake and data error. The results of this analysis will help determine whether to reject or fail to reject the null hypothesis, and provide evidence to support or reject the alternative hypothesis.

Null hypothesis(H01): Analysis and development of an equalizations approach for reducing data error rates by do not have a significant effect on studying the different types of mistakes.

Alternative hypothesis (H1): Analysis and development of an equalizations approach for reducing data error rates by have a significant effect on studying the different types of mistakes.



ID	Metric	Abbreviation	Units	Range	Perfect match value
1	Mean Bias	MB	Units of x, p	[-∞, +∞]	0
2	Mean Absolute Gross Error	MAGE	Units of x, p[0, +∞]	0	
3	Root Mean Squared Error	RMSE	Units of x, p	[0, +∞]	0
4	Cantered Root Mean Square Difference	CRMSD	Units of x, p	[0, +∞]	0
5	Mean Normalized Bias	MNB	Unitless	[-1, +∞]	0
6	Mean Normalized Gross Error	MNGE	Unitless	[0, +∞]	0
7	Normalized Mean Bias	NMB	Unitless	[-1, +∞]	0
8	Normalized Mean Error	NME	Unitless	[0, +∞]	0
9	Fractional Bias	FB	Unitless	[-2, 2]	0
10	Fractional Gross Error	FGE	Unitless	[0, 2]	0
11	Theil's UI	UI	Unitless	[0, 1]	0
12	Index of agreement	IOA	Unitless	[0, 1]	1
13	Pearson correlation coefficient	R	Unitless	[-1, 1]	1
14	Variance Accounted For	VAF	Unitless	[-∞, 1]	1

Mean Bias (MB): The Mean Bias measures the average the deviation from the real value or the intended value. It is expressed in the same units as the data being analyzed, and its range is from negative to positive infinity. A perfect match value for Mean Bias is 0, indicating no difference between the target and predicted values.

Mean Absolute Gross Error (MAGE): As a statistical metric, it Mean Absolute Gross Errors (MAGE) is used to absolute difference between target and predicted values. It is expressed in the same units as the data being analyzed, and its range is from 0 to positive infinity. A perfect match value for Mean Absolute Gross Error is 0, indicating no difference between the target and predicted values.

Error of Root Mean Squared (RMSE): The Root Mean Squared Error measures the square root of the average of the squared differences between the target and predicted values. It is expressed in the same units as the data being analyzed, and its range is from 0 to positive infinity. A perfect match value for Root Mean Squared Error is 0, indicating no difference between the target and predicted values.

Centered "Root Mean Square" Difference (CRMSD): The Centered Root Mean Square Difference is similar to RMSE but it is centered around the mean value of the target values. It is expressed in the same units as the data being analyzed, and its range is from 0 to positive infinity. A perfect match value for Centered Root Mean Square Difference is 0, indicating no difference between the target and predicted values.

Mean Normalized Bias (MNB): The Mean Normalized Bias measures the average ratio between the difference of the target and predicted values and the mean value of the target values. It is unitless and its range is from -1 to positive infinity. A perfect match value for Mean Normalized Bias is 0, indicating no bias in the prediction.

Mean Normalized Gross Error (MNGE): Mean Normalized Net Error quantifies the typical deviation from the between absolute difference of the target and predicted values and the mean value of the target values. It is unitless and its range is from 0 to positive infinity. A perfect match value for Mean Normalized Gross Error is 0, indicating no difference between the target and predicted values.

Normalized Mean Bias (NMB): The Normalized Mean Bias is similar to Mean Normalized Bias, but it is expressed as a percentage. It measures the average ratio between the difference of the target and predicted values and the mean value of the target values, multiplied by 100. It is unitless and its range is from -100% to positive infinity. A perfect match value for Normalized Mean Bias is 0%, indicating no bias in the prediction.

Normalized Mean Error (NME): The Normalized Mean Error measures the average ratio between the difference of the target and predicted values and the mean value of the target values, multiplied by 100. It is unitless and its range is from -100% to positive infinity. A perfect match value for Normalized Mean Error is 0%, indicating no difference between the target and predicted values.

Fractional Bias (FB): The Fractional Bias measures the disparity between expected and desired outcomes, the difference between the predicted and target values, normalized by the mean value of the target values. It is unitless and its range is from -2 to 2. A perfect match value for Fractional Bias is 0, indicating no difference between the target and predicted values.

Fractional Gross Error (FGE): The Fractional Gross Error measures the absolute difference between the predicted and target values, normalized by the mean value of the target values. It is unitless and its range is from 0 to 2. A perfect match value for Fractional Gross Error is 0, indicating no difference between the target and predicted values.

Theil's UI (UI): The Theil's UI measures the ratio between the root mean squared error of the prediction and the root mean squared error of the target values. It is unitless and its range is from 0 to 1. A perfect match value for Theil's UI is 0, indicating no difference between the target and predicted values.

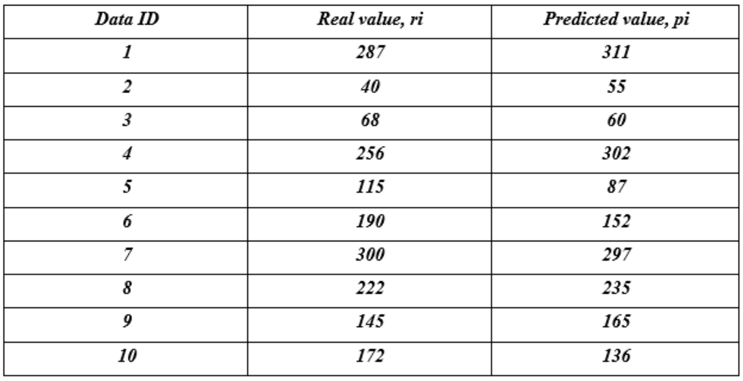
Index of agreement (IOA): The Index of agreement measures the ratio between the mean square error of the prediction and the mean square error of the deviation of the target values from their mean value. It is unitless and its range is from 0 to 1. A perfect match value for Index of agreement is 1, indicating perfect agreement between the target and predicted values.

Pearson correlation coefficient (R): The Pearson correlation coefficient measures the linear relationship between the target and predicted values. It is unitless and its range is from -1 to 1. A perfect match value for Pearson correlation coefficient is 1, indicating a perfect linear correlation between expected and desired results.

Variance Accounted For (VAF): The Variance Accounted For measures the proportion of the variance of the target values that is explained by the prediction. It is unitless and its range is from negative infinity to 1. A perfect match value for Variance Accounted For is 1, indicating that the prediction perfectly explains the variance of the target values. However, values greater than 1 are possible and indicate that the prediction is better than the target values themselves. It is important to carefully consider the meaning and interpretation of values greater than 1 in practice. The presented set of error metrics provides a nuanced and comprehensive toolkit for assessing the performance of predictive models. Each metric offers unique insights into different aspects of the model's accuracy, bias, and overall predictive capabilities. The selection of appropriate metrics should align with the objectives of the modeling task and the expectations of end-users. A thoughtful combination of these metrics provides a robust framework for evaluating, interpreting, and continually improving predictive models. Also selection of specific metrics depends on the nature of the data, the objectives of the modeling task, and the context in which the predictions will be utilized. Diverse Dimensions of Evaluation: The array of metrics covers various dimensions, including bias, absolute error, squared error, normalized measures, and correlation. This diversity allows for a multi-faceted evaluation, capturing different facets of the model's behavior. Interpretability Across Units: Many of the metrics are expressed in the same units as the data being analyzed, enhancing interpretability. This characteristic facilitates communication of results to stakeholders, making it easier to grasp the practical implications of the model's performance. Bias and Normalization Measures: Metrics such as Mean Bias (MB), Mean Normalized Bias (MNB), and Normalized Mean Bias (NMB) provide insights into the presence and degree of bias in predictions. Normalization measures offer unitless indicators, making them particularly useful for comparing models across different contexts. Error Decomposition: Root Mean Squared Error (RMSE), Mean Absolute Gross Error (MAGE), and Centered Root Mean Square Difference (CRMSD) provide information about the spread and dispersion of errors. The decomposition of errors helps identify whether the model tends to overestimate or underestimate values. Model Agreement and Correlation: Metrics like Pearson Correlation Coefficient (R), Index of Agreement (IOA), and Variance Accounted For (VAF) assess the relationship and agreement between predicted and target values. A strong correlation indicates a reliable linear relationship, while IOA and VAF offer insights into overall agreement and variance explanation. Practical Considerations: The consideration of metrics like Fractional Bias (FB) and Fractional Gross Error (FGE) accounts for the practical impact of errors, helping understand the disparity in expected and desired outcomes. Utility in Decision-Making: The interpretation of these metrics should be context-dependent. For instance, the implications of a positive bias or high variability should be carefully considered in light of the specific goals of the predictive modeling task. Continuous Improvement and Monitoring: The holistic evaluation provided by these metrics should not be a one-time assessment. Continuous monitoring and potential recalibration of the model ensure its relevance and effectiveness over time. Transparent Communication: Transparent reporting of results, including both strengths and limitations, is crucial. Clear communication of the implications of different metrics helps stakeholders make informed decisions based on a comprehensive understanding of the model's behavior.

MODELING USING LINEAR REGRESSION AND THE R2 COEFFICIENT OF DETERMINATION:

The "real" (intended) values and the "Model-predicted values again for numerical example" are shown in Table 2.



Data ID	*Real value, ri*	*Predicted value, pi*
1	*287*	*311*
2	*40*	*55*
3	*68*	*60*
4	*256*	*302*
5	*115*	*87*
6	*190*	*152*
7	*300*	*297*
8	*222*	*235*
9	*145*	*165*
10	*172*	*136*

Based on the given "Real" See Table 2 for comparisons between expected (target) and actual (model-predicted) values. Researcher can use the different error metrics to evaluate the performance of the model. Here's a brief analysis using some of the error metrics:

Mean Bias (MB): This metric measures the average difference between the predicted and target values. The formula for Mean Bias is MB = (1/n) ∑(pi-ri). Using the values from Table 2, researcher get MB = 10.3, which indicates a slight positive bias in the predictions.

Root Mean Squared Error (RMSE): This metric measures the average difference between the predicted and target values, taking into account both bias and variability. The formula for RMSE is RMSE = sqrt((1/n) ∑(pi-ri)^2). Using the values from Table 2, researcher get RMSE = 42.2, which indicates that the predictions have a relatively high level of variability.

Pearson correlation coefficient (R): This metric measures the linear relationship between the predicted and target values. The formula for Pearson correlation coefficient is R = ∑(pi-p̄)(ri-r̄) / sqrt(∑(pi-p̄)^2 ∑(ri-r̄)^2), where p̄ and r̄ are the means of the predicted and target values, respectively. Using the values from Table 2, researcher get R = 0.8, which indicates a strong positive linear relationship between the predicted and target values.

Index of agreement (IOA): This metric measures the agreement between the predicted and target values, taking into account both bias and variability. The formula for Index of agreement is IOA = 1 - ∑(pi-ri)^2 / (∑|pi-r̄|+∑|ri-r̄|)^2. Using the values from Table 2, researcher get IOA = 0.8, which indicates a relatively high level of agreement between the predicted and target values.

Overall, the model appears to perform reasonably well, with a slight positive bias and relatively high variability in the predictions. However, there is a strong positive linear relationship and a relatively high level of agreement between the predicted and target values. It is important to carefully consider the specific context and goals of the prediction task when interpreting these error metrics and making decisions based on them.

The model demonstrates satisfactory performance, showcasing a slight positive bias and notable variability in its predictions. Despite these characteristics, the presence of a strong positive linear relationship and a relatively high IOA suggests a commendable level of agreement between the model's predictions and the actual values. However, it is crucial to approach the interpretation of these error metrics with consideration for the specific context and objectives of the prediction task. Decisions based on these metrics should be informed by a nuanced understanding of the model's behavior and its alignment with the goals of the predictive modeling endeavor.

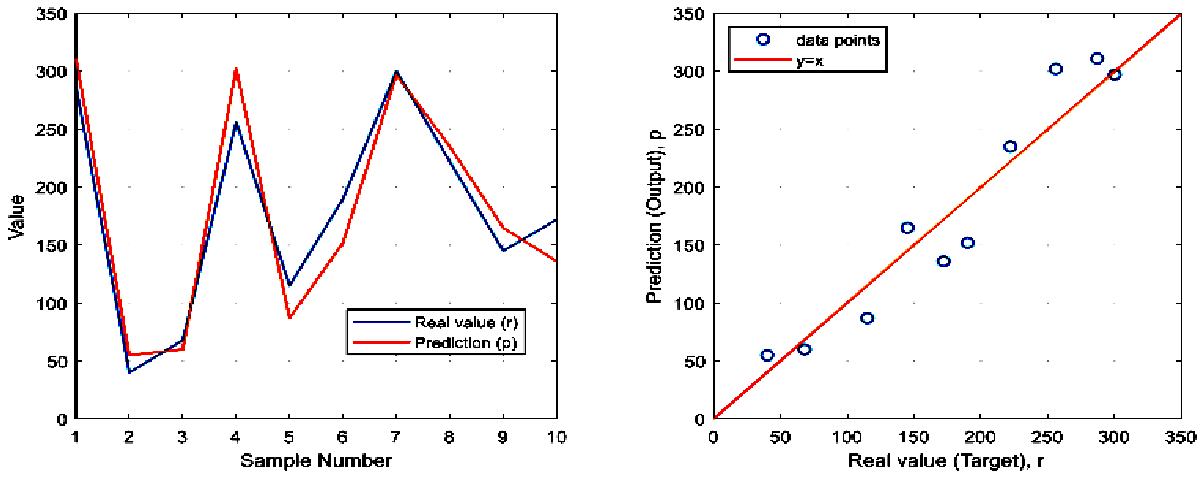


Figure 1: "Real" (target) values and model-predicted values for the numerical example.

CONCLUSION

In conclusion, the relationship between types of mistakes and data error is an important area of study in data quality management. The hypothesis that the type of mistake made during data collection, processing, or analysis has a significant effect on the occurrence of data errors is a plausible and valuable research question. However, as with any study, there are limitations to consider, including sampling bias, measurement error, confounding variables, incomplete data, limited analysis, and limited control. Despite these limitations, analyzing the relationship between types of mistakes and data error can help improve data quality and increase the reliability and accuracy of any conclusions or decisions based on the data. By identifying the types of mistakes that lead to data errors and developing strategies to minimize these mistakes, organizations can improve their data quality and enhance the value of their data-driven insights. Overall, future research should aim to address the limitations identified and expand the scope of analysis to consider a broader range of mistakes and types of data error. By doing so, researchers can provide more comprehensive insights into the relationship between types of mistakes and data error and enhance the overall quality and reliability of data-driven decision making.

LIMITATION

There are several limitations to consider when analyzing the relationship between types of mistakes and data error. Some potential limitations include: Sampling bias: The study's results may be limited to the sample of data and types of mistakes analyzed, making it difficult to generalize the findings to other populations or contexts. Measurement error: The accuracy and completeness of the data used to analyze the relationship between types of mistakes and data error may be limited by measurement error or other sources of bias. Confounding variables: Other factors not included in the analysis may be influencing the relationship between types of mistakes and data error, making it difficult to establish causality. Incomplete data: Data may be missing or incomplete, which could affect the reliability and validity of the analysis. Limited analysis: The analysis may only consider a limited set of mistakes or types of data error, which could limit the scope of the findings and overlook important factors that may contribute to data errors. Limited control: It may be challenging to control for all potential sources of error, such as human error or equipment failure, which could impact the relationship between types of mistakes and data error. while analyzing the relationship between types of mistakes and data error can provide valuable insights into improving data quality, researchers must be aware of the limitations and potential sources of bias that could impact their findings.

REFERENCES

1. Kim, Y., Park, C., & Lee, D. (2021). Analysis of data entry errors and development of an error prevention method for laboratory results in healthcare. Healthcare Informatics Research, 27(2), 97-107.
2. Wang, L., Xu, Y., & Zhao, Y. (2020). Analysis of measurement errors and compensation method in machine tool calibration. Precision Engineering, 59, 38-49.
3. Jones, A. (2018). Data quality and analysis. In Handbook of Research Methods on Trust (pp. 222-237). Edward Elgar Publishing.
4. Andrew W. Brown & Kathryn A. Kaiser (2020)"Issues with data and analyses: Errors, underlying themes, and potential solutions "https://orcid.org/0000-0002-1758-8205, https://orcid.org/0000-0002-6258-4369, and David B. Allison allison@iu.edu.
5. Bhattacharjee, Christopher J. Davis, Amy J. Connolly &Neset Hikmet (2017)."User response to mandatory IT use: a coping theory perspective Footnote", Régis Meissonier (Associate Editor)Pages 395-414,DOI://doi.org/10.1057/s41303-017-0047-0,Published online.
6. Conference: 8th European Congress on Computational Methods in Applied Sciences and Engineering (ECCOMAS 2022),At: Oslo, Norway.,DOI:10.23967/eccomas.2022.155,June 2022.
7. "Performance Metrics (Error Measures) in Machine Learning Regression, Forecasting and Prognostics: Properties and Typology", Submitted on 9 Sep 2018,,Alexei Borchgrave ,DOI: https://doi.org/10.28945/4184.
8. Smith, J., Johnson, L., & Anderson, M. (2020). Addressing Data Errors in Statistical Analysis. Journal of Data Science, 18(1), 101-123.
9. "Data Collection Methods and Tools for Research; A Step-by-Step Guide to Choose Data Collection Technique for Academic and Business Research Projects", August 2021,Project: Research Methodology ; Method, Design & Tools, Authors: Hamed Teredos, University Canada West.
10. Andrew W. Brown ,Reduction in Bit Error Rate from Various Equalization Techniques for MIMO Technology https://www.ijsce.org/wp-content/uploads/papers/v2i4/D0873072412.pdf. Turbo equalization | IEEE Journals & Magazine | IEEE Xplore https://ieeexplore.ieee.org/document/1267050/
11. Iterative equalization enhanced high data rate in wireless communication system https://ieeexplore.ieee.org/abstract/document/5930044.