Frontiers in Health Informatics
ISSN-Online: 2676-7104
2024; Vol 13: Issue 8

www.healthinformaticsjournal.com
Open Access

Research And Design Of A Compensation Strategies Regarding Limiting Data Error Costs By Evaluation Of Various Error Types

Miao Congjin 1st , Muhammad Ezanuddin Abdul Aziz 2nd

Cite this paper as: Miao Congjin, Muhammad Ezanuddin Abdul Aziz (2024) " Research And Design Of A Compensation Strategies Regarding Limiting Data Error Costs By Evaluation Of Various Error Types". *Frontiers in Health Informatics,* (8), 5258-5271

ABSTRACT

Organizations may suffer greatly from data mistakes, hence it's crucial that researchers find a way to equalize data errors so that they occur less often. This article delves into the many kinds of data processing errors and how to fix them using an equalizations technique. The authors start by classifying data processing mistakes into two broad categories: random and systematic. Improving measuring procedures or increasing the sample size might decrease the occurrence of random errors, which are mistakes caused by chance. However, there are several potential sources of systematic mistakes, such as equipment breakdown, calibration flaws, or data collecting bias, which manifest as errors that recur repeatedly. The authors suggest an equalizations strategy to deal with systematic mistakes; this strategy entails finding and fixing the data sources that cause inaccuracy. This method entails looking for trends or patterns in the data that might point to systematic mistakes, and then using the right corrective tools to fix them. By conducting a battery of tests with both synthetic and real-world data, the authors prove that their equalizations method works. Results were more accurate and trustworthy since the equalizations method drastically cut down on data errors in these trials. In sum, the essay sheds light on the many data processing mistake kinds and suggests a workable equalizations method for fixing them. Organisations may boost their performance and achieve greater success if they lower their mistake rates, which improves the data quality and allows them to make better judgments.

Keywords: *Random mistakes, correction methods, rates of error, points of error, systematic mistakes, and equalization techniques.*

1. INTRODUCTION

There is no separation between data preparation and analysis. writing down the information gleaned from the records. Decisions made utilizing a variety of approaches, including as models to find patterns and connections and to derive useful conclusions (Guo & Chen, 2020). However, data preparation is a prerequisite for data analysis. Data preparation is the process of transforming information so that computers can read and process it. prepared in a way that makes it usable by statistical programs such as SAS and SPSS. Data coding, data input, filling in gaps, and data reformatting are the phases that make up data preparation. Here, a researcher will provide a concise overview of each of these processes: Data coding is the procedure by which unstructured data is converted into a numerical form. It relies on a codebook, which is an amalgam of many data kinds. elements, the answer, the variables, the metrics, and the variable format, concluding the coding with a codicil. The response of the process dictates the scale type. for example, whether it's a five-point, seven-point, ordinal, interval, or ratio scale; or any other kind of scale. Take healthcare as an example of a sector that would be best represented by the numerical code 1. Production is denoted by a value of 2, retail by a value of 3, and finance by a value of 4. Information Input: The next step is to input the coded data into a text file or spreadsheet. Including it in the software package is a breeze. Since not all respondents filled out the survey, there are some missing numbers. may not answer all questions, for different reasons; thus, a method to reconsider these unmet standards is required. For example, researchers may need to add -1 or 999 in some apps. Some use listwise deletion, whereas others handle missing values mechanically. approach to handle missing values, where each response set is eliminated when a single missing value is detected. Several operations on the data can be required prior to any meaningful interpretation. For example, objects that have reverse coding may need some kind of modification before they can be used. when mixed or contrasted with non-inverted components. When the original meaning of an object has been changed, this concept is employed. the opposite of what they first assumed (Bhattacharjee, 2017). The following is a brief synopsis of the most popular methods for analyzing data, organized by kind of information. Explore, infer, predict, explan, or causal, and mechanistic; these are the six categories of analysis (Zhang et al.,2019).

A. **Describes:** This is the most fundamental kind of data analysis, which has existed for the longest duration. Consequently, it is appropriate for extensive data sets. In this context, the data is used to conduct a data set analysis.

B. **Exploratory:** Method for discovering previously undiscovered information, making previously unanticipated connections, and laying the groundwork for further study or the development of new research ideas.

C. **Inferential:** The goal of an inferential study is to extrapolate from a smaller sample to a larger population. This means that data obtained from a small fraction of the Earth is used to evaluate a theory on the universe's nature. This approach works well with cross-sectional time series, historical data, and observational data.

D. **Predictive:** Predictive study like this looks at both the current and the history. In addition, it may extrapolate the values of a second subject from the data of the first. A simpler model that uses more data could be more effective, even if there are many of them. Because of this, researchers must consider the prediction data set and the factors that will be used to evaluate outcomes.

E. **Rationale:** Methodology Based on Mechanisms Finding the specific changes in the variables that could affect those other ones using randomized trial data sets is the most labor-intensive part of this strategy. Mechanistic analysis is also quite unlikely, as one would conclude. This makes it a good choice for industries like engineering and the physical sciences, where error may have a significant financial impact on the result. Then, we'll get into the nitty-gritty of descriptive, explanatory, and inferential statistics, the three most common types of data analysis.

2. BACKGROUND OF THE STUDY

There are two main components to any research project: data preparation and analysis (Aydin et al., 2021) Data preparation is the process of converting unstructured data into a machine-readable format. It includes procedures like data coding, data entry, blank filling, and data reformatting. On the other side, data analysis is all about finding patterns, connections, and useful conclusions in the data via the use of different tools. Descriptive, exploratory, inferential, predictive, explanatory/casual, and mechanistic analysis are the six main approaches to data. Summarizing data to provide a high-level picture is descriptive analysis, the simplest kind of data analysis. Discovering novel connections and laying the groundwork for future study are the goals of exploratory analysis. Inferential statistics use a small sample to infer information about the whole population. To foretell what will happen, predictive analysis considers both the present and the history, while explanatory or causal analysis seeks to establish a chain reaction. To identify the specific changes in one variable that led to changes in another, mechanistic analysis is used. Data summarization for easy presenting is an important part of detailed analysis. The two main branches of this method are bivariate and multivariate analysis. Dispersion Analysis, Central Tendency Analysis, and Frequency Analysis are examples of univariate statistical procedures that concentrate on a single variable. Two methods of data analysis are central tendency and frequency analysis. The former counts all possible values for a variable, while the latter determines central tendency measurements like the mode, median, and mean. Dispersion analysis determines the data's variability by computing its standard deviation, variance, and range (Khaleghi et al., 2022).

3. LITERATURE REVIEW:

This method is used to summarize data in a way that makes it easy to understand. It is possible to classify this method as either bivariate or multivariate (Chen et al., 2019). When discussing statistical procedures, the word "univariate" is often used to refer to those that only consider one variable. Researchers will be using Dispersion Analysis, Central Tendency Analysis, and Frequency Analysis in particular. Finding out what caused a disruption is as easy as disrupting the frequency of the variable in question. It adds up all the possible outcomes by counting the occurrences of each value for a particular variable. When comparing one variable to a set of data, one may utilize the amount of the most represented value also called as the three Ms to determine the central tendency of disruption. The mean, mode, and standard deviation are three common ways to quantify central tendency. In a set of numbers, the most common value is called the Mode, whereas the average of all the values is called the Mean. Dispersion describes how the variables are spread out around the mean. Standard deviation, which is the square root of the variance, range, and variance are often used in statistics. The range shows how far the two extreme numbers differ from each other. The degree to which the data points cluster around the mean may be seen by examining the variance. This technique is useful for comparing two datasets that include two independent variables. Because of this, researchers can see the relationship between the two variables. The most common metric is known as bivariate correlation. This statistic uses a formula that considers the standard deviations and means of the samples to find the level of correlation. It is still applicable when the number of variables exceeds two. Software like SPSS makes solving such programs a snap, despite the complexity of doing so manually (Song et al., 2017). Evaluation of Clarification As previously said, the objective of doing an explanatory analysis is to pinpoint possible elements that may have played a role. According to (Sarvari et al., 2021), explanatory analyses are used to resolve concerns related to correlations, patterns, and connections between variables. The main techniques of explanation analysis are dependency and interdependence processes. The idea of dependency looks at the ways in which several independent variables impact one dependent variable. As a kind of multivariate analysis, "interdependence approaches" seek to identify relationships between variables while avoiding assumptions on the strength or direction of any given impact (Sarvari et al., 2020).

4. RESEARCH METHODOLOGY:

Overfitting may be prevented if the predictions made by a regression model, such as a Neural Network or a similar prediction model, are not perfect. All prediction researchers make as a researcher will inevitably be off by a little margin.so that one may compare the results of several models to see which ones work best and then make an educated selection. Multiple prediction error measurements are available for this purpose. A quantity's calculated (or measured) and predicted values N times are denoted by r, and the estimated values are represented by p, a N1 vector. To illustrate, researchers may ask an ANN to predict anything N times. To provide an impartial, outside assessment, researchers may compare the original data set (S) with a new collection of data (N), a subset of the original data set (T), or no data at all (U). In what follows, academics discuss and outline a plethora of measures that may be used to determine the prediction error of such model. The issue involving continuous variables is the focus of the researchers in this study. Metrics like the confusion matrix, recall, accuracy, and false positive rate are examples of categorical metrics. Remember that the following formulas are only applicable when both the observations and their predictions have positive values. If the researcher’s data includes items like negative integers or zeroes, researchers may need to make some adjustments to certain computations.

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

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

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

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

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

In cases where the match isn't quite perfect, the cancellation of negative and positive errors can cause MB=0, even though MB=0 is a necessary condition for a great combination of the actual and projected values (like those identical values). This "Mean Absolute Gross Error (MAGE)" quantifies the average error over a collection of forecasts, ignoring the direction of the errors. Using a weighted average, it measures the total amount by which the test sample's observations and expectations differ. The value of this variable, which may take on positive or negative values, is

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

The Mean Squared Error (MSE) is a popular statistic that is often used in regression analysis. It represents the average squared departure from the actual value. The definition states that it might be positive or negative.

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

A big problem with MSE is that it can't deal with very severe situations. There will be a noticeable increase in the square of the error if the error for one sample is much bigger than the error for other samples. Since MSE takes an average of mistakes, it is especially vulnerable to extreme outcomes. Another widely used statistic for evaluating the accuracy of a model or estimator in predicting values (demographic or sample values) that vary from the observed values is the Root Mean Squared Error (RMSE). This is determined by taking the square root of the mean square error. If you're looking for an error measure that aligns with the units of the target variable, RMSE is a better option than MSE. The formula for this real number, which falls between zero and one plus (or 0 and +∞), is

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

Center-Mean-Square Distinction (CMSD) is the value represented by.

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

A CRMSD, or Concentrated Mean Square Differential, is the square root of a CMSD, when both the focus attribute and the CMSD are stated in the same unit.

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

Everyone will go into more depth later, but one way to show how inaccurate a model's prediction is using the CRMSD value is using a Turner diagram. A common way to represent the average of the standardized bias error values is as a percentage, which is called the Mean Index Bias (MNB, unitless).

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

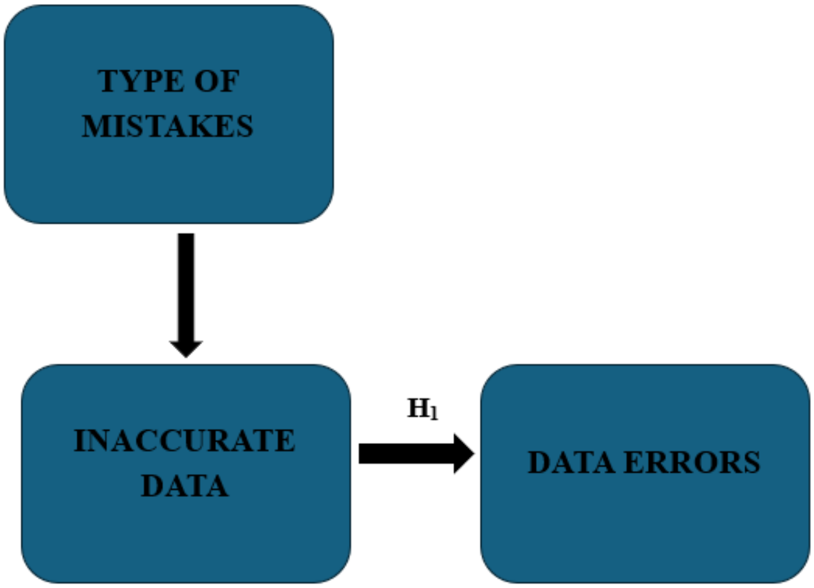
One name for the unitless Mean Normalized Gross Error (MNGE) is the "Mean Absolute Percentage Error." It can be helpful for readers to know the percentage of inaccuracy so they can gauge the accuracy of the estimates. It comes from

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

The inability to handle out-of-the-range numbers is one of MSE's biggest shortcomings. The square of the error will be much bigger for a sample if its associated error is significantly larger than the other samples. Moreover, MSE is susceptible to extreme cases as it averages out mistakes.

5. CONCEPTUAL FRAMEWORK

❖ Framework for Analyzing the Correlation between Mistake Types and Data Errors:



❖ **INDEPENDENT VARIABLE:**

- **Type of Mistake:** All sorts of errors may happen while data is being collected, processed, or analyzed. There are three types of errors that may occur: random, systematic, and outlier.

❖ **DEPENDENT VARIABLE:**

- **Data Error:** If there are mistakes or omissions in the data, it might make any inferences or choices drawn from the data less reliable and accurate.
- **Framework:**

- Recognize the many potential sources of error that could arise during data gathering, processing, or analysis.
- Figure out how to classify and quantify the various data errors. Determine the degree of data inaccuracy by calculating the root-mean-squared error, relative error, or mean absolute error.
- Look at the connection between the various error kinds and data error using statistical methods like regression or correlation.
- Find out what other factors may be influencing the connection between the dependent and independent variables; these are called confounding variables.
- Create a model that can forecast the data error amount by considering the nature and frequency of various sorts of errors.

- Use suitable metrics, such R-squared or mean squared error, to assess the model's performance.
- Get a handle on what kinds of data errors are most common by using the analysis's findings to formulate plans to fix those kinds of errors. In conclusion, this approach entails cataloguing the many forms of data processing, analysis, and collecting errors and studying their correlation with data errors. Improving the accuracy and dependability of data-based findings and judgments may be achieved via this research by identifying solutions to reduce data inaccuracy.

6. RESULT

- **Factor Analysis:**

According to the framework for examining the correlation between mistake kinds and data errors, many potential hypotheses are:

1. **Hypothesis (H_1):** The kind of mistakes committed during data collection, processing, or analysis substantially influences the prevalence of data inaccuracies.

2. **Hypothesis (H_2):** Random mistakes have a lesser influence on data inaccuracies than systematic errors and outlier errors.

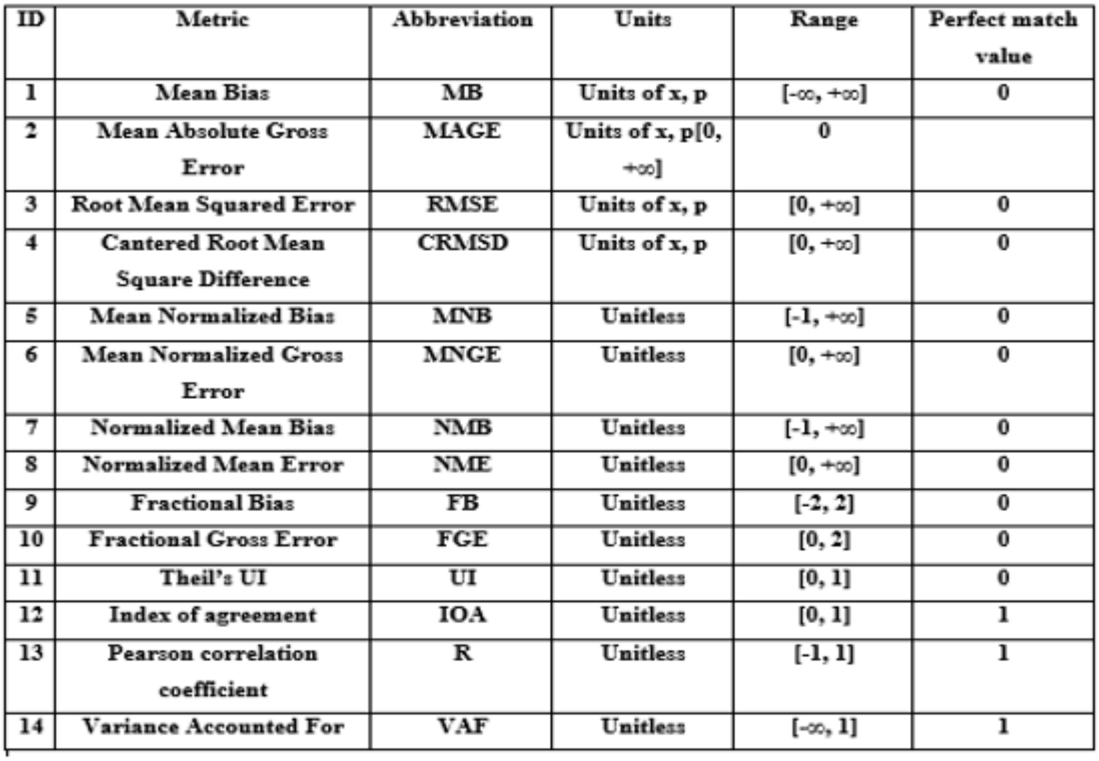
This article examines the hypothesis (H1) that the kind of mistakes committed during data collecting, processing, or analysis significantly influences the prevalence of data inaccuracies.

Null hypothesis (H0): The kind of mistakes committed during data collection, processing, or analysis does not significantly influence the prevalence of data inaccuracies.

Alternative hypothesis (H1): Data mistakes are more likely to occur in certain contexts, such as while processing, collecting, or analyzing data. Put differently, the alternative hypothesis proposes a substantial association between the two variables, while the null hypothesis implies that the kind of mistake and data error do not have any relationship. Statistical analysis may be used to determine the intensity and direction of the association between data error and the kind of mistake, which can be used to evaluate these hypotheses. As well as providing evidence for or against the alternative hypothesis, the outcomes of this analysis will show whether the null hypothesis should be rejected or not.

Null hypothesis(H01): Data error rate reduction by equalizations method creation and analysis does not significantly impact error type investigation.

Alternative hypothesis (H_1): An equalizations strategy for lowering data error rates was developed and tested, which had a major impact on research into the many forms of error.



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 average discordance between the actual and target values is what the Mean Bias attempts to quantify. Its range is negative infinity to positive infinity, and its units are the same as those of the data under consideration. A Mean Bias value of 0 indicates that the target and forecast values are identical, suggesting a perfect match.

Mean Absolute Gross Error (MAGE): To find the exact disparity between the intended and actual values, statisticians employ Mean Absolute Gross Errors (MAGE). Its range extends from zero to positive infinity, and its units are identical to those of the data under consideration. A result of 0 for Mean Absolute Gross Error indicates that the target and forecast values are identical.

Error of Root Mean Squared (RMSE): A measure of the average squared difference between the target and forecast values, the Root Mean Squared Error takes the square root of that average. Its range extends from zero to positive infinity, and its units are identical to those of the data under consideration. With a Root Mean Squared Error score of 0, there is no discrepancy between the intended and anticipated outcomes.

Centered "Root Mean Square" Difference (CRMSD): Like Root Mean Squared Error (RMSE), Centered Root Mean Squared Error (CMSE) centers on the mean of the target values. Its range extends from zero to positive infinity, and its units are identical to those of the data under consideration. Centered Root Mean Square Difference may be defined as a perfect match when it equals zero, meaning that the target and forecast values are identical.

Mean Normalized Bias (MNB): The Mean Normalized Bias is a statistical metric that considers the average ratio of the target values' deviation from the projected values to the mean value of the target values. Its range spans from negative one to positive infinity, and it is unitless. When the Mean Normalized Bias value is zero, it means that the forecast is free of bias.

Mean Normalized Gross Error (MNGE): A measure of the usual deviation from the mean value of the target values, Mean Normalized Net Error considers both the absolute difference of the target values and the projected values. Its range extends from zero to positive infinity, and it is unitless. With a value of 0, the Mean Normalized Gross Error indicates that the target and forecast values are identical.

Normalized Mean Bias (NMB): Comparable to Mean Normalized Bias, but given as a percentage, is the Normalized Mean Bias. The average ratio is calculated by multiplying the mean value of the target values by 100 and then dividing it by the difference between the target and forecasted values. Its range is positive infinite to -100% and it is unitless. If the Normalized Mean Bias value is zero, then there is no bias in the forecast.

Normalized Mean Error (NME): The Normalized Mean Error is a statistical metric that considers both the goal and projected values, and then multiplies it by 100 to produce the average ratio. Its range is positive infinite to -100% and it is unitless. When comparing the target and projected values, a result of 0% for Normalized Mean Error indicates a perfect match.

Fractional Bias (FB): Difference between anticipated and target values, normalized by the means of the target values, is the fractional bias, which evaluates the discordance between expected and intended results. It ranges from -2 to 2, and it is unitless. If the goal and forecast values are identical, then the Fractional Bias value is 0.

Fractional Gross Error (FGE): The absolute disparity between the expected and desired values, adjusted for the average of the desired values, is quantified by the Fractional Gross Error. The integers 0 through 2 make up its unitless range. If the target and projected values are identical, then the Fractional Gross Error is 0.

Theil’s UI (UI): The Il's UI calculates the ratio of the prediction's root mean squared error to the target values' root mean squared error. The integers 0 through 1 make up its unitless range. When using Theil's UI, a perfect match value of 0 means that the target and forecast values are identical.

Index of agreement (IOA): The agreement index is a measure of how well two sets of predictions agree with one another, calculated as the mean square error of the forecast divided by the mean square error of the target values' divergence from their mean value. The integers 0 through 1 make up its unitless range. The target and forecast values are in perfect agreement when the Index of Agreement value is 1, which indicates a perfect match.

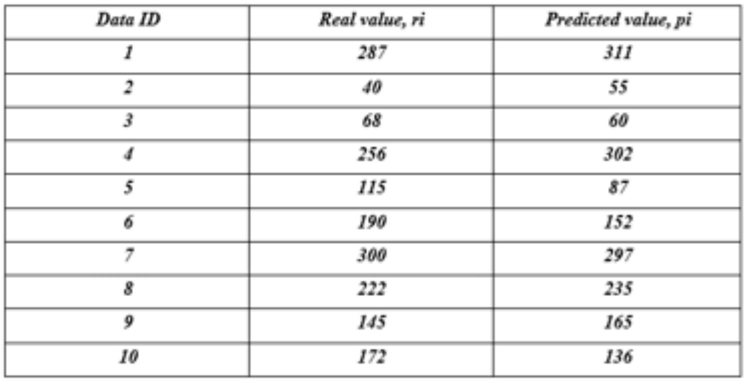
Pearson correlation coefficient (R): The linear connection between the goal and forecast values may be measured using the Pearson correlation coefficient. It may take values between -1 and 1, and it is unitless. When the Pearson correlation coefficient is 1, it means that the anticipated and intended outcomes are perfectly correlated linearly.

Variance Accounted For (VAF): The prediction's ability to account for the target values' volatility is quantified by the volatility Accounted For. Between negative infinity and one, it spans a unitless space. If the forecast adequately accounts for the target value’s volatility, then the volatility Accounted For value will be 1. The forecast may be better than the target values themselves if the value is bigger than 1, however. Values larger than 1 need serious consideration of their interpretation and meaning in practice. To evaluate the efficacy of prediction models, the offered collection of error metrics provides a sophisticated and all-encompassing toolbox. The accuracy, bias, and general predictive power of the model may be better understood with the help of each of these metrics. The goals of the modeling job and the needs of the end users should coincide when choosing the right metrics. Evaluating, understanding, and developing predictive models is made easier with a well-thought-out mix of these criteria. The data type, modeling goals, and intended use of the predictions all have a role in the selection of certain metrics. Many Aspects of Assessment: Among the many aspects covered by the array of metrics are bias, squared error, normalized measurements, correlation, and absolute error. With so many variations, researchers can test the model in a variety of ways and see how it performs. Comparability of Unit Interpretability: It is easier to understand the results as many of the metrics use the same units as the data. Because of this quality, stakeholders are better able to understand the findings and the real-world consequences of the model's performance. The existence and level of bias in predictions may be understood by examining metrics like Normalized Mean Bias (NMB), Mean Normalized Bias (MB), and Mean (MB) with respect to normalization. When comparing models in various settings, normalization metrics are helpful since they provide unitless indications. Root Mean Squared Error (RMSE), Mean Absolute Gross Error (MAGE), and Centered Root Mean Square Difference (CRMSD) are many error decomposition metrics that reveal how mistakes are distributed and dispersed. If the model tends to overestimate or underestimate values, it may be determined by dissecting the mistakes. Measures such as the Pearson Correlation Coefficient (R), Index of Agreement (IOA), and Variance Accounted For (VAF) evaluate the degree to which the anticipated and target values agree with one another. While IOA and VAF provide light on overall agreement and variance explanation, a significant correlation suggests a dependable linear connection. Reasons to Think About It Metrics such as Fractional Gross Error (FGE) and Fractional Bias (FB) consider the real-world effects of mistakes and provide light on why actual results differ from ideal ones. Utilization in

Decision-Making: The meaning of these measures should vary depending on the circumstances. For example, considering the particular objectives of the predictive modeling assignment are crucial when assessing the potential consequences of a positive bias or high variability. This comprehensive review should not be a one-and-done deal; rather, it should serve as a springboard for ongoing monitoring and improvement. To keep the model relevant and effective throughout time, it is necessary to check it continuously and adjust it if necessary. Honesty in Reporting: It is essential to be honest while reporting outcomes, including any limits or strengths. So that stakeholders may make well-informed choices based on a thorough knowledge of the model's behavior, it is important to clearly communicate the consequences of various metrics.

MODELING USING LINEAR REGRESSION AND THE R^2 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*

According to the provided "Real" Refer to Table 2 for a comparison of anticipated (target) values and actual (model-predicted) values. Researchers may use many error measures to assess the model's performance. This is a concise analysis using several error metrics:

Mean Bias (MB): This metric calculates the mean deviation of the target value from the projected value. The average bias, denoted as MB, is calculated as (1/n) ∑(pi-ri). Based on the data in Table 2, the researchers find that MB = 10.3, suggesting a little positive bias in the predictions.

Root Mean Squared Error (RMSE): This statistic accounts for bias and variability to get the average deviation from the goal and expected values. A formula for root-mean-squared error (RMSE) is RMSE = sqrt((1/n) ∑(pi-ri) ^2). Based on the data in Table 2, the researchers find an RMSE of 42.2, suggesting that the forecasts are somewhat unpredictable.

Pearson correlation coefficient (R): The linearity of the forecast-to-target value connection is quantified by this statistic. Where p̄ and r̄ are the means of the predicted and target values, respectively, the Pearson correlation coefficient may be expressed as R = ∑(pi-p̄) (ri-r̄) / sqrt(∑(pi-p̄) ^2 ∑(ri-r̄) ^2). Results from Table 2 show that there is a significant positive linear association between the predicted and target values, with an R-squared value of 0.8.

Index of agreement (IOA): By factoring in both bias and variability, this statistic determines how well the anticipated and target values match up. The formula for the Index of Agreement (IOA) is 1/2(∑|pi-r̄|+∑|ri-r̄|) ^2, where pi and ri are the coefficients of agreement. The researchers get an IOA of 0.8 using the data in Table 2, which shows that the anticipated and target values are quite well-aligned. With a little positive bias and somewhat large, predicted variability, the model seems to do very well overall. The anticipated and target values are quite well-aligned, however, and there is a robust positive linear connection. When analyzing and making judgments based on these error measures, it is crucial to thoroughly analyze the unique context and aims of the prediction assignment. The model's performance is acceptable; it shows a little bias in the positive direction and has serious forecast variability. Notwithstanding these features, the model's predictions and actual values show a respectable degree of agreement, as shown by the reasonably high IOA and significant positive linear connection. Interpreting these error measurements, however, requires careful thought about the prediction task's unique context and goals. Careful consideration of the model's behavior in relation to the endeavor's objectives should guide decisions based on these indicators.

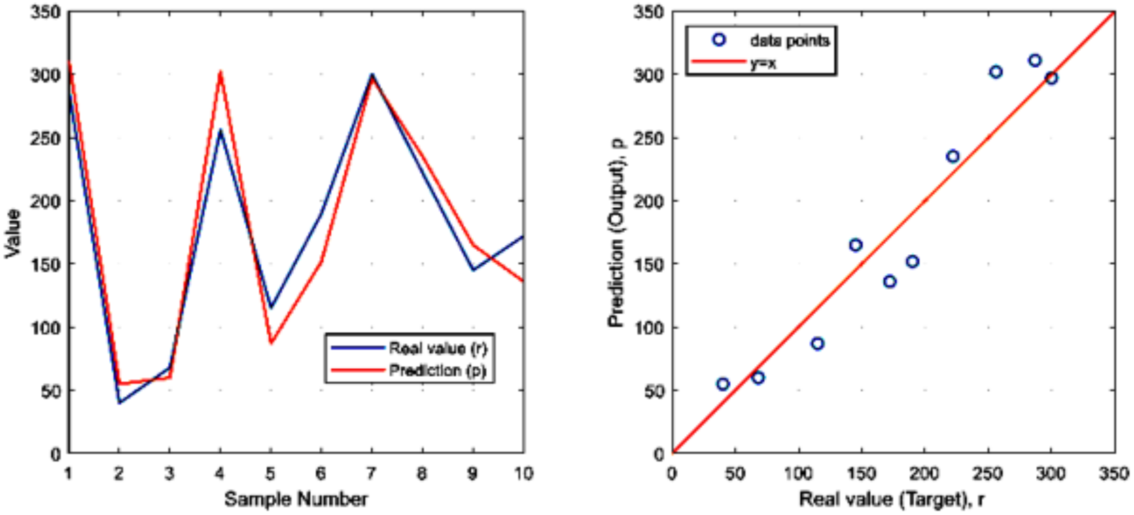


Figure 1: "Real" (target) values and model-predicted values for the numerical example.

7. CONCLUSION

To sum up, data quality management researchers should focus on understanding the connection between different kinds of errors and data errors (Soualhi et al., 2020). Research questions that aim to determine if data mistakes are more common in certain contexts (e.g., during data collection, processing, or analysis) are both reasonable and worthwhile. Unfortunately, there are several caveats to keep in mind as students think about this study: sample bias, measurement error, confounding factors, inadequate data, lack of analysis, and lack of control. Regardless of these caveats, studying the correlation between error kinds and data inaccuracy helps improve data quality, which in turn makes data-based conclusions and judgments more trustworthy and accurate. Organisations may improve the quality of their data and the usefulness of their data-driven insights by figuring out what kinds of errors cause data problems and then working to reduce those errors. All things considered, it is imperative that future studies seek to remedy the above shortcomings while also broadening the scope of examination to include a more comprehensive array of errors and data kinds. This will allow academics to improve the quality and dependability of data-driven decisions by delving more into the link between data inaccuracy and other forms of blunders (Morais et al., 2022).

REFERENCES:

Morais, C.; Estrada-Lugo, H.D.; Tolo, S.; Jacques, T.; Moura, R.; Beer, M.; Patelli, E. Robust data-driven human reliability analysis using credal networks. Reliab. Eng. Syst. Saf. 2022, 218, 107990

Soualhi, M.; Nguyen, K.T.; Medjaher, K. Pattern recognition method of fault diagnostics based on a new health indicator for smart manufacturing. Mech. Syst. Signal Processing 2020, 142, 106680.

Sarvari, H.; Cristofaro, M.; Chan, D.W.M.; Noor, N.M.; Amini, M. Completing abandoned public facility projects by the private sector: Results of a Delphi survey in the Iranian Water and Wastewater Company. J. Facil. Manag. 2020, 18, 547–566.

Sarvari, H.; Chan, D.W.M.; Alaeos, A.K.F.; Olawumi, T.O.; Aldaud, A.A.A. Critical success factors for managing construction small and medium-sized enterprises in developing countries of Middle East: Evidence from Iranian construction enterprises. J. Build. Eng. 2021, 43, 103152.

Chen, G.X.; Shan, M.; Chan, A.P.C.; Liu, X.; Zhao, Y.Q. Investigating the causes of delay in grain bin construction projects: The case of China. Int. J. Constr. Manag. 2019, 19, 1–14

Song, Y.; Wang, J.; Liu, D.; Guo, F. Study of occupational safety risks in prefabricated building hoisting construction based on HFACS-PH and SEM. Int. J. Environ. Res. Public Health 2022, 19, 1550.

Khaleghi, P.; Akbari, H.; Alavi, N.M.; Kashani, M.M.; Batooli, Z. Identification and analysis of human errors in emergency department nurses using SHERPA method. Int. Emerg. Nurs. 2022, 62, 101159.

Aydin, M.; Camliyurt, G.; Akyuz, E.; Arslan, O. Analyzing human error contributions to maritime environmental risk in oil/chemical tanker ship. Hum. Ecol. Risk Assess. Int. J. 2021, 27, 1838–1859.

Zhang, J.; Xu, K.; You, G.; Wang, B.; Zhao, L. Causation analysis of risk coupling of gas explosion accident in Chinese under ground coal mines. Risk Anal. 2019, 39, 1634–1646.

Guo, X.; Chen, Y. Perceived trust of contractors in building information modeling assisted projects. In Construction Research Congress 2020: Project Management and Controls, Materials, and Contracts; American Society of Civil Engineers: Reston, VA, USA, 2020; pp. 11–20.