NAME: OPEOLUWA GOODNESS OMONIYI
STUDENT ID: 24075076
COURSE: MACHINE LEARNING AND NEURAL NETWORKS

EVALUATING THE PERFORMANCE OF A REGRESSION MODEL

1.0 INTRODUCTION

Regression is one of the foundational techniques in machine learning, used to model and predict continuous numerical outcomes. Whether estimating housing prices, forecasting sales, or predicting energy consumption, regression methods form the backbone of quantitative prediction tasks. However, building a regression model is only the first step; the ability to properly evaluate the model's performance is equally important.

A model may fit the training data closely yet perform poorly on new data. Another model may have acceptable accuracy but fail to capture underlying relationships. Others may violate statistical assumptions, leading to biased or misleading predictions.

To answer these questions, a rigorous framework for evaluation is needed. This tutorial explores in depth how to evaluate the performance of a regression and applies these concepts to a real dataset from the National Health and Nutrition Examination Survey (NHANES) 2017–2018. In this dataset, the goal is to predict blood mercury concentration, an important marker of environmental exposure, using metabolic, lipid, liver-function, and additional heavy-metal biomarkers.

The goal of this tutorial is to explain why regression evaluation matters and how it works, covering:

- Core error-based metrics (MAE, MSE, RMSE).
- Goodness-of-fit measures (R^2, Adjusted R^2).
- Residual analysis.
- Overfitting, and underfitting.

1.1 IMPORTANCE OF EVALUTION IN REGRESSION

A clear understanding of what constitutes a strong regression model is essential before selecting appropriate evaluation metrics. In practice, a good regression model should produce accurate predictions with minimal error on unseen data and demonstrate strong generalisation without evidence of overfitting or underfitting. It should capture meaningful underlying patterns in the data rather than noise, while performing reliably across the full range of predictor values. Additionally, the model should satisfy core statistical assumptions, including linearity, normally distributed residuals, and homoscedasticity, ensuring that inference and interpretation remain valid. Ultimately, an effective regression model must also be interpretable and practically useful within the context of the analysis, supporting clear and informed decision-making.

2.0 DATASET OVERVIEW

National Health & Nutrition Exam Survey (NHANES) 2017-2018

NHANES is a structured epidemiological dataset collected by the U.S. Centres for Disease Control and Prevention (CDC). The variables selected for this analysis are:

a) **Target Variable:**
Blood Mercury (μg/L): a biomarker of exposure to methylmercury, often arising from diet (especially fish) or environmental contamination.

b) **Predictor Variables:** The predictors include:

- Metabolic markers: Fasting glucose, Insulin.
- Lipid panel: HDL cholesterol, Total cholesterol
- Liver function markers: ALT, AST
- Other heavy metals: Lead, Manganese

Preprocessing steps

- Removed missing values: Ensured the dataset was clean and consistent for modelling.
- Log-transformed mercury: Reduced right-skew and stabilized variance for better model performance.
- Train–test split (80/20): Allowed the model to learn from one portion and be evaluated fairly on unseen data.

Models used

- **Linear Regression:** A simple, interpretable model that assumes a straight-line relationship between predictors and the target. It provides a baseline for comparison but may struggle with non-linear patterns.
- **Random Forest:** An ensemble of decision trees that captures complex, non-linear relationships and interactions between variables. It is more flexible and often more accurate, especially when the underlying patterns are not linear.

3.0 ERROR-BASED METRICS

Error-based metrics quantify how far predictions deviate from actual values. Where:

- x_i = true value
- $\hat{y}_i$ = predicted value
- n = total number of data points
- p = number of predictors.

a) **Mean Absolute Error (MAE):** It is a measurement of the typical absolute differences between observed data points and their predicted values\. It is easy to interpret and robust to outliers compared to MSE (Agrawal, 2025).

$$\mathrm{MAE} = \frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{n}} |\mathrm{x_i} - \mathrm{y_i}|$$

b) **Mean Squared Error (MSE):** It calculates the mean of the squared deviations between observed data points and predicted values. It is commonly used in regression analysis to evaluate the predictive accuracy of models.

$$\mathrm{MSE} = \frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{x_i} - \mathrm{y_i})^2$$

c) **Root Mean Squared Error (RMSE):** It is a metric used to quantify absolute error, calculated by squaring the individual errors to avoid cancellation of positive and negative values, like Mean Squared Error (MSE). It reflects the standard deviation of the residuals, where a residual denotes the difference between the predicted value and the observed value (D’Agostino, 2022).

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{x_i} - \mathrm{y_i})^2} = \sqrt{\mathrm{MSE}}$$

3.1 Goodness-of-Fit Measures

a) R-squared (R²): also known as the coefficient of determination, is utilised to evaluate the goodness of fit of a regression model (GeeksforGeeks, 2025). It measures the proportion of the variation in the dependent variable that can be explained by the independent variables within the model. R² is a valuable metric for assessing the overall performance and explanatory capability of a regression model.

$$\mathrm{R^2} = 1 - \frac{\mathrm{SSR}}{\mathrm{SST}}$$

Where:

- R2 is the R-Squared.
- SSR represents the sum of squared residuals between the predicted values and actual values.
- SST represents the total sum of squares, which measures the total variance in the dependent variable.

b) **Adjusted R²:** R² has a major drawback: when you add more features, it can only stay the same or increase, even if the new feature is irrelevant. This can misleadingly suggest a better model. Adjusted R² solves this by penalizing extra predictors. If a new feature adds little value, the penalty causes the adjusted R² to drop; if it's genuinely useful, adjusted R² increases. This makes adjusted R² a more reliable indicator of model performance.

$$\text{Adjusted } R^2 = 1 - (1 - R^2) \cdot \frac{n-1}{n-p-1}$$

4.0 ANALYSIS OF REGRESSION MODEL PERFORMANCE

Metric	Linear regression	Random forest
MAE	0.6236	0.6242
MSE	0.6434	0.6425
RMSE	0.8021	0.8015
R-squared (R²)	0.0632	0.0645
Adjusted R²	0.0497	0.0510

Linear Regression Performance

The linear regression demonstrates very poor predictive performance, indicating that the selected predictors explain only a minimal portion of the variance in blood mercury levels. With an R² score of 0.0632 and an adjusted R² of 0.0497, the model accounts for just over 6% of the variability in the outcome, meaning it is barely more informative than predicting the mean mercury value for all observations. The error metrics, MAE (0.6236), MSE (0.6434), and RMSE (0.8021), further reflect substantial prediction errors, suggesting that the assumed linear relationships between these biomarkers and blood mercury are inadequate. Overall, the results show that the selected biomarkers have a very weak linear relationship with blood mercury, making the model barely more informative than predicting the mean.

Random Forest Regression Performance

The Random Forest regression model performs almost identically to the linear regression model, with only a negligible improvement in metrics (R² rising from 0.0632 to 0.0645). Such a small change indicates that even a non-linear model cannot extract meaningful predictive patterns from the available biomarkers. The model still explains only about 6.5% of the variance in blood mercury levels, showing that the issue lies not in the modelling technique but in the limited predictive value of the selected features. Therefore, simply switching to a more advanced algorithm cannot resolve the problem.

4.1 Residual Analysis and Diagnostic Plots

Even when numerical performance metrics look acceptable, a regression model may still violate key assumptions. To check this, residual analysis is performed. Residuals represent the difference between actual and predicted values:

$$e_i = y_i - \hat{y}_i$$

Examining residuals helps reveal whether the model assumptions, linearity, constant variance, and unbiased errors, are being met.

Residual vs. Fitted Plot

It is used to evaluate how well a regression model fits the data by examining the pattern of residuals (errors) across predicted values. It helps identify problems that violate regression assumptions.

- **Nonlinearity:** If residuals show curves or patterns instead of being randomly scattered, it suggests the model is missing non-linear relationships (UVA Library, 2015).
- **Heteroscedasticity (non-constant variance):** If the spread of residuals increases or decreases across fitted values (e.g., a funnel shape), it indicates uneven variance, which can affect model reliability.
- **Systematic bias:** If residuals consistently sit above or below zero in certain regions, the model may be biased, meaning it systematically over- or under-predicts for specific ranges.

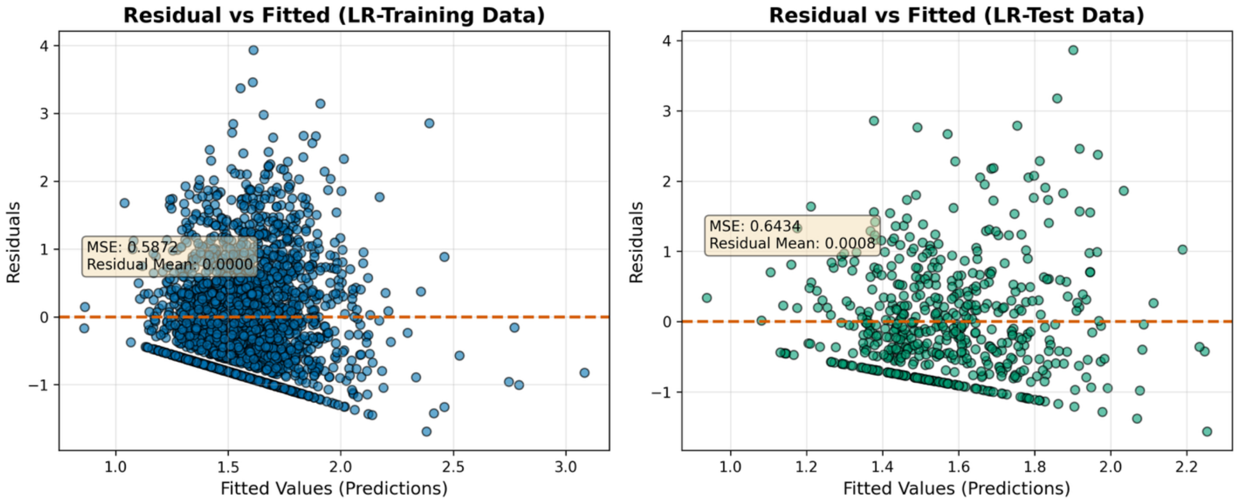

Figure 1: Residual Analysis of Linear Regression Results

The residual versus fitted plots for the linear regression model (figure 1), with a training MSE of 0.5872 and a test MSE of 0.6434, confirm the model's poor predictive performance and highlight violations of key linear assumptions. The minimal increase in error from training to test data indicates no severe overfitting; however, the extremely low explanatory power (R² ≈ 0.06) suggests the residuals are not randomly scattered and likely exhibit patterns such as heteroscedasticity or non-linearity signs that the model fails to capture the underlying relationship between the biomarkers and blood mercury concentration.

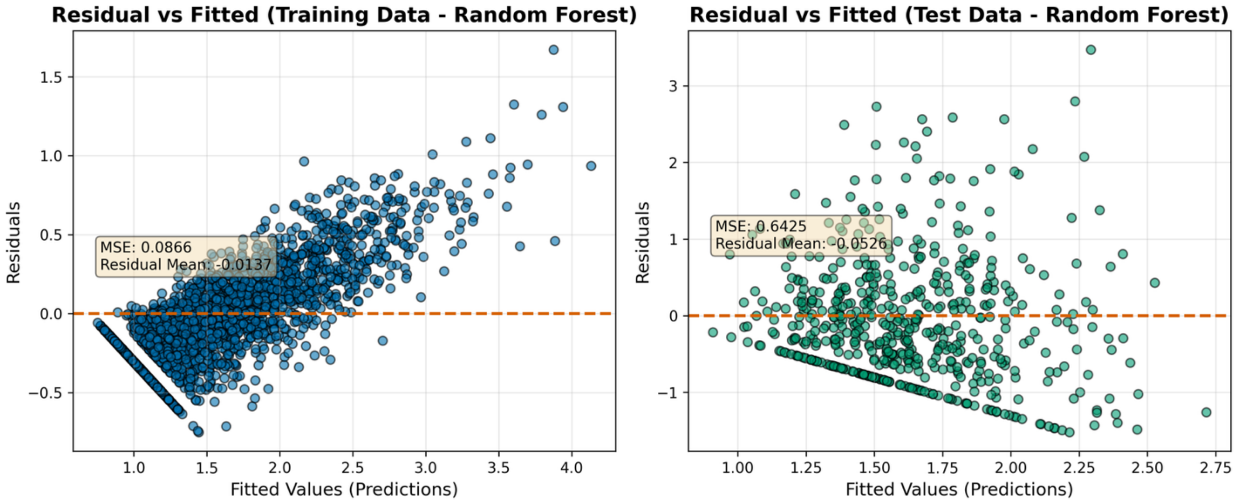

Figure 2: Residual Analysis of Random Forest

In figure 2, The Random Forest results demonstrate severe overfitting, with near-perfect training performance (MSE: 0.0866) but poor generalization to the test set (MSE: 0.6425). This indicates the model memorized training noise rather than learning a valid predictive signal for blood mercury. The identical failure of both simple and complex models confirms that the current biomarkers lack explanatory power, necessitating feature revision and outcome transformation.

5.0 Choosing the Right Evaluation Metrics

Selecting appropriate evaluation metrics is essential for assessing model performance effectively. MAE is preferred when interpretability matters, when the target distribution is skewed (such as mercury levels), or when outliers should not dominate the evaluation (Agrawal, 2025). RMSE is more suitable when large errors must be penalised heavily, making it valuable in engineering or forecasting contexts. R² helps compare models predicting the same outcome or evaluate the proportion of variance explained, while Adjusted R² offers a fairer comparison when models differ in complexity.

In practice, several key principles guide effective model evaluation. Never rely on a single metric, use a suite such as MAE, RMSE, and R² to form a complete picture. Prioritise metrics that match the practical impact of prediction errors and always visualise residuals, as patterns can reveal issues that numerical scores may overlook. Testing on unseen data remains essential for assessing generalisability, and simpler models should be preferred unless additional complexity clearly improves performance. Sensitivity analyses, such as evaluating the effect of outliers or applying transformations, help ensure that conclusions are robust and reliable.

5.1 CONCLUSION

Evaluating a regression model requires more than relying on a single metric; it involves combining error measures, goodness-of-fit statistics, and visual diagnostics. In this analysis of blood mercury prediction using metabolic, lipid, liver-function, and heavy-metal biomarkers from NHANES 2017–2018, both linear regression and Random Forest performed poorly. These results indicate a fundamentally weak linear and non-linear predictive relationship within the current biomarker set.

Crucially, a poor model outcome is still informative. The evaluation highlights that the selected biomarkers, as currently used, lack meaningful explanatory power for mercury levels, pointing to the need for transformations, additional exposure-related variables, and alternative modelling strategies. Comprehensive assessment using multiple metrics and diagnostic plots ensures that resulting models are not only statistically sound but also reliable and useful for real-world decision-making.

Biomarker Variable Definitions

- LBXGLU - Fasting Glucose (mmol/L)
- LBDINSI - Insulin (pmol/L) Both males and females 12 YEARS -150 YEARS
- LBDHDDSI - Direct HDL-Cholesterol (mmol/L)
- LBDTCSI - Total Cholesterol (mmol/L)
- LBDBPBSI - Blood lead (umol/L) Both males and females 6 YEARS -150 YEARS
- LBDTHGSI - Blood mercury, total (nmol/L)
- LBDBMNSI - Blood manganese (nmol/L) Both males and females 1 YEARS -150 YEARS
- LBXSASSI - Aspartate Aminotransferase (AST) (U/L)
- LBXSATSI - Alanine Aminotransferase (ALT) (U/L)

GitHub Repository

https://github.com/opeoluwa22/machine-learning-tutorial.git

REFERENCES

Agrawal, R. (2025). Know the best evaluation metrics for your regression model. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/05/know-the-best-evaluation-metrics-for-your-regression-model/

GeeksforGeeks. (2025). Regression metrics in machine learning.https://www.geeksforgeeks.org/machine-learning/regression-metrics/

D'Agostino, A. (2022). Assessing model performance for regression. Towards Data Science. https://towardsdatascience.com/assessing-model-performance-for-regression-7568db6b2da0/

CleverX. (2025). *What is model evaluation in machine learning*. CleverX. https://cleverx.com/blog/what-is-modal-evaluation-in-machine-learning#:~:text=Why%20is%20model%20evaluation%20important,model%20is%20useful%20and%20trustworthy

UVA Library. (2015). Understanding Diagnostic Plots for Linear Regression Analysis. UVA Library StatLab. https://library.virginia.edu/data/articles/diagnostic-plots