Journal of Information Systems Engineering and Management
2025, 10(31s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

A Comprehensive Framework for Residual Analysis in Regression and Machine Learning

Vibhu Verma
Principal Data Scientist,
GWU, Capital One,
NY, USA
vvibhu1@gmail.com

ARTICLE INFO

Received: 30 Dec 2024

Revised: 12 Feb 2025

Accepted: 26 Feb 2025

ABSTRACT

Residual analysis is one of the most crucial methodologies in statistical modeling and machine learning. Generally, it tends to be an important tool in the evaluation of the precision of a model, diagnosing violations of assumptions, and refinement. This paper critically reviews residuals, their mathematical underpinning foundations, and how they feature in model performance evaluation. Key diagnostic methods that have been explored in this paper include heteroscedasticity, non-linearity, autocorrelation, and influential outliers. Further, we have to develop a new case based on the decomposition of residuals and SHAP values for the analysis of unexplained sales trends. This study underlines how residual patterns can indicate hidden deficiencies in the model and how model improvement can obtain better results. We conclude the length in optimizing model performance and future research directions in residual-based diagnostics.

Keywords: Machine Learning, Residual Analysis

Introduction

Residuals play a central role in the assessment of statistical and machine learning models. In any given predictive framework, the residuals defined by the differences of the observed value and the prediction of the model carry a good deal of rich information beyond merely an error measure. The residuals, when the model is well-specified, should have a random pattern around zero. Such randomness would be indicative that this model captured the systematic part of the relationship between the independent variables and the response variable. The presence of visible systematic patterns of residuals-for instance, trend or curvature or clumping-might indicate important omission of relationships or variables, non-fulfilment of some assumptions-like homoscedasticity or normality-, over- or underfitting in the model performance.

Residuals are an important part in regression and machine learning with regard to the strength assessment of a statistical model. From residuals, being the differences between the observed values and the predicted ones, analysts can diagnose problems, improve the performance of a model, and gain further insights. The main roles played by residual analysis in model evaluation and refinement are discussed below.

1. Model Fit and Assumption Checking

Some of the basic assumptions which a regression model requires are linearity-the relationship between predictors and response is linear-independence, homoscedasticity-the errors are uncorrelated with each other and with constant variance- and normality. Residual analysis helps to check these assumptions:

Linearity: a plot of residuals versus fitted values should show no non-random pattern to suggest the failure of a model, in that part of the underlying relationship may have gone into the addition of polynomial terms or nonlinear transformation.

Homoscedasticity: A "fanning out" pattern in residuals, with increasing variance along fitted values, is indicative of heteroscedasticity. This violates important assumptions of regression and may result in statistical inference that is not trustworthy. Possible remedies include transformation-a common example being logarithmic-and weighted least squares.

Independence & Autocorrelation: In time series models, residuals must not have a pattern over time; otherwise, it means there is autocorrelation, in which the previous value takes priority over the current one. Therefore, the model would need further modification to incorporate either autoregressive elements or lagged predictors.
Normality: Most of the statistical tests in hypothesis testing and confidence interval estimation require residuals to be normally distributed. To check this assumption, one may analyze residual histograms and Q-Q plots or make use of a statistical test such as the Shapiro-Wilk test.
2. Outlier and Influence Diagnostics
Not all data points are equal in their contribution to the predictions of a model; points may be influential. Being able to identify such influential observations is key to ensuring models are robust: Outliers: Points with large residuals can indicate that something is out of the ordinary-perhaps an anomaly, or an error in collecting the data. Depending upon context, such outliers could be excluded, transformed, or be subject to robust regression methods.
High-Leverage Points: Although not outliers, some points have extreme values of the predictor variables and might, therefore, have undue influence on the regression line. These can be obtained from leverage statistics obtained from the hat matrix. Cook's Distance: It measures how much an observation influences regression estimates. A high Cook's Distance would suggest that a particular observation's deletion would result in a massive effect on the model. 3. Informed Model Improvement
One of the most powerful diagnostics to suggest model improvement is residual analysis: Unmodeled Structures: A systematic residual pattern may reflect an important model omission, such as a missing quadratic or interaction effect. The addition of new features or polynomial terms could help this model.
Feature Engineering: When residuals vary systematically with an omitted variable, this is a sign that there may be more predictors. Such a pattern helps feature selection and engineering find all relevant factors.
Transformation: A log, square root, or Box-Cox transformation often resolves skewness in residuals and is better interpretable than the model itself.
4. Modern Applications in Machine Learning
Residuals are not only helpful within conventional regression, but they also play a central role in machine learning, most especially in: Ensemble Learning: Techniques such as gradient boosting fit models iteratively to residuals or pseudo-residuals, refining their predictions at every step. This can yield very accurate models that progressively reduce the error.
Deep Learning: ResNets model residual functions explicitly. Instead of learning the target function, they are learned how to iteratively update their predictions. It helps address the vanishing gradient problem by making it easier to construct a deep, yet efficient network. This will avoid errors in the target variable from modeling and also help in class balance and generalized improvement over segments of data.
5. Case Studies and Business Applications
Besides theoretical diagnostics, residual analysis has great practical significance in many industrial segments. The financial segment involves analysts in an attempt to refine forecasts, detect anomalies in the markets while adjusting their risk weights in a portfolio; correcting bathymetric (depth) readings includes sensor calibration in the environmental science. Residuals are useful in checking any predictive model, whether systematic underestimation or overestimation of a target condition occurs to inform clinical decisions. Retail & Business Analytics: Unexplained residuals may point to some of the exogenous factors at play, including economic fluctuations or marketing campaigns. It uses a method for residual decomposition, such as SHAP values, which might be known to the businesses if something happens when the performance goes up or down.

Theoretical Background

Consider a simple model of linear regression where one wants to forecast some response variable y on grounds of some predictor variable x. There, there is a modeled assumption which claims that the response observed is just a sum, plus residual error: which should be independent and identically distributed with means equal to zero and constant variances.
The fitted values from the model are the predicted responses for each observation. The residual for each observation is the difference between the observed value and the fitted value. One of the key properties of the OLS method is that the sum of all residuals is zero, meaning the model fits the data without any overall bias.
However, the residual variances are not constant for all observations; they are a function of the leverage of each observation. Leverage refers to the degree to which each particular observation influences its own fitted value. The

higher the leverage, the more influence an observation has on the fit, and the smaller its residual variance is likely to be.

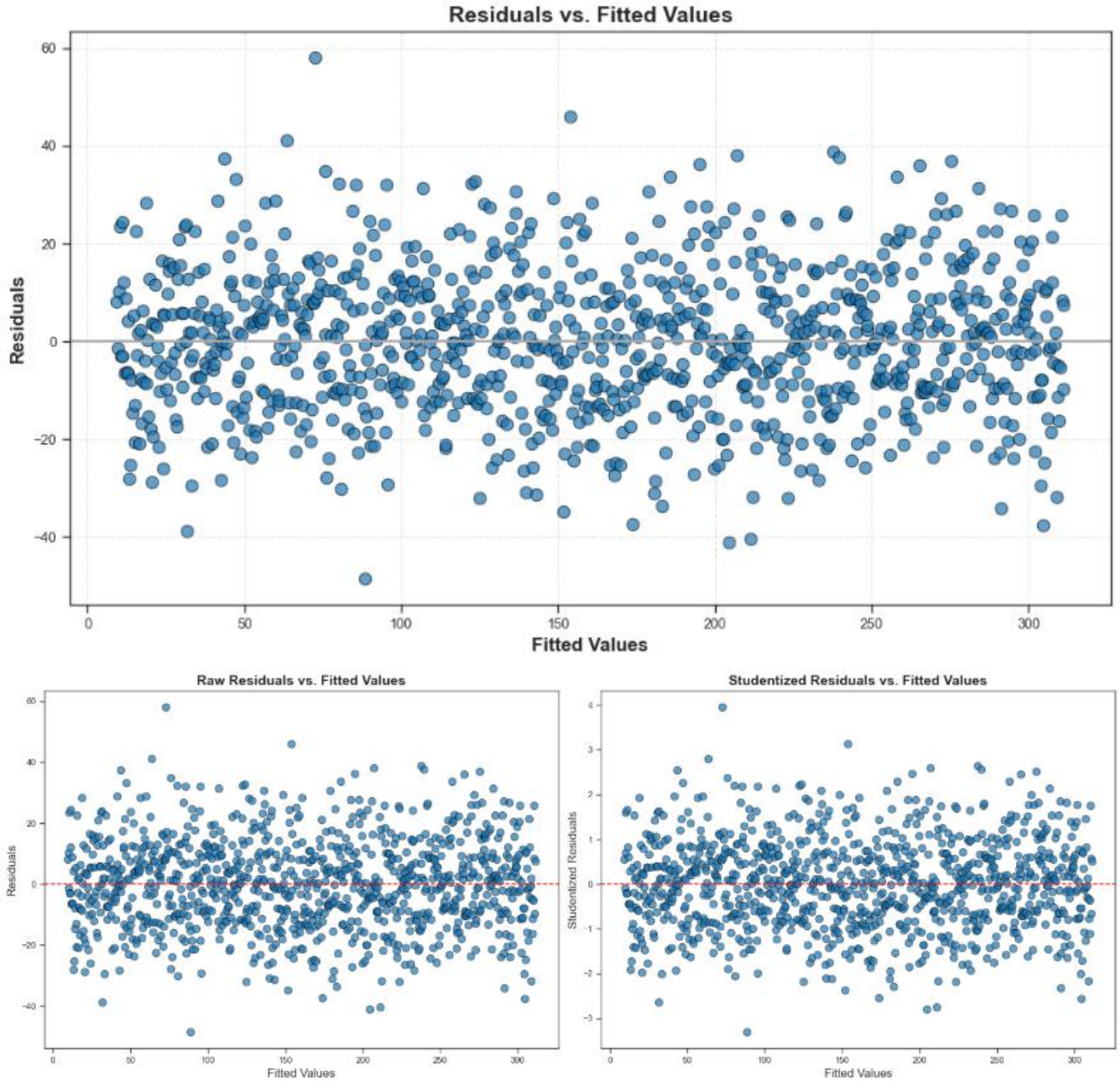
2.2 Standardization

Residuals are often standardized or studentized for diagnostics because residual variances might be dependent on leverage. Standardizing changes the residuals by dividing them by an estimate of the standard deviation of the error and accounting for leverage of the particular observation. This standardizes the residual in a way that gives a better sense of model fit across all the different observations.

Studentized residuals are similar, except that in the latter case we estimate the standard deviation of the error without using the particular observation. That is one common method for detecting outliers, because it identifies observations whose removal greatly reduces the estimate of model error.

2.3 The Role of the Hat Matrix

One of the most important matrices in the linear regression analysis is the hat matrix. It projects the observed responses onto the fitted responses, in essence projecting the observed data onto the space spanned by the predictors. The diagonal elements of this matrix show the influence each observation has on its own prediction, and the larger the value, the greater the leverage. Observations with high leverage have an inordinately large influence on the regression model, especially regarding the fitted coefficients and variances of the residuals.



Diagnostic Uses of Residual Analysis

3.1 Checking Model Fit

Residual analysis is a serious diagnostic tool that gives insights into whether a model is correctly specified and how it might be improved. In this section, we discuss several such diagnostic uses of these residuals. For each there is an explanation and Python code examples using dummy data. We further explain how one can check the model fit, homoscedasticity, linearity, serial correlation, outliers, influential observations, normality of residuals, perform support feature selection, compare various models, and at last diagnosis the bias-variance problem.

Summary

3.1 Model Fit Check

A well-fitted model will present residuals which randomly scatter around zero. The indication of any pattern in residual plot shows the existence of misspecification in model.

Summary

• Random Scatter: Residuals that are randomly scattered with no pattern would indicate that the model picked up the underlying trend.

• Systematic Patterns: Curved or clustered residuals may indicate missing variables, wrong functional forms-for example, a linear model for a nonlinear relationship- or omitted interactions.

Interpretation

If the plot shows a curved pattern that indicates that a linear model is insufficient for quadratic data. This diagnosis drives model improvement through either transformation or choosing a non-linear model.

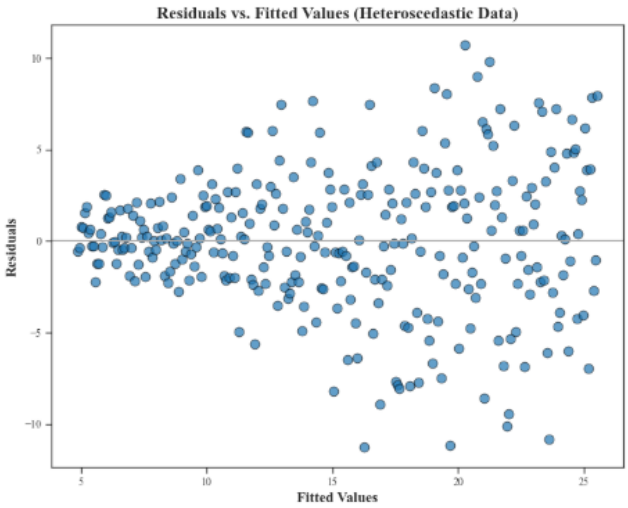
3.2 Detecting Heteroscedasticity

Heteroscedasticity occurs when the variance of residuals is not constant across all levels of an independent variable in a regression model. This violates the assumption of homoscedasticity in Ordinary Least Squares (OLS) regression, leading to unreliable standard errors, misleading hypothesis tests, and inefficient estimates.

A common method for detecting heteroscedasticity is plotting residuals against fitted values. If the residuals exhibit a systematic pattern, such as a fan shape where variance increases with fitted values, heteroscedasticity is present. Formal statistical tests, such as the **Breusch-Pagan test** and **White's test**, can confirm this issue by checking if residual variance depends on predictor variables. The **Goldfeld-Quandt test** is another method that compares variance across subsets of data.

To address heteroscedasticity, analysts can apply **logarithmic** or **square root transformations** to stabilize variance. Alternatively, **robust standard errors** can be used to correct statistical inference. In cases where the variance structure is known, **Weighted Least Squares (WLS)** or **Generalized Least Squares (GLS)** regression can be effective solutions.

Detecting and correcting heteroscedasticity is crucial for ensuring accurate regression modeling. By analyzing residual patterns and applying appropriate corrective measures, models become more reliable and generalizable for real-world applications.



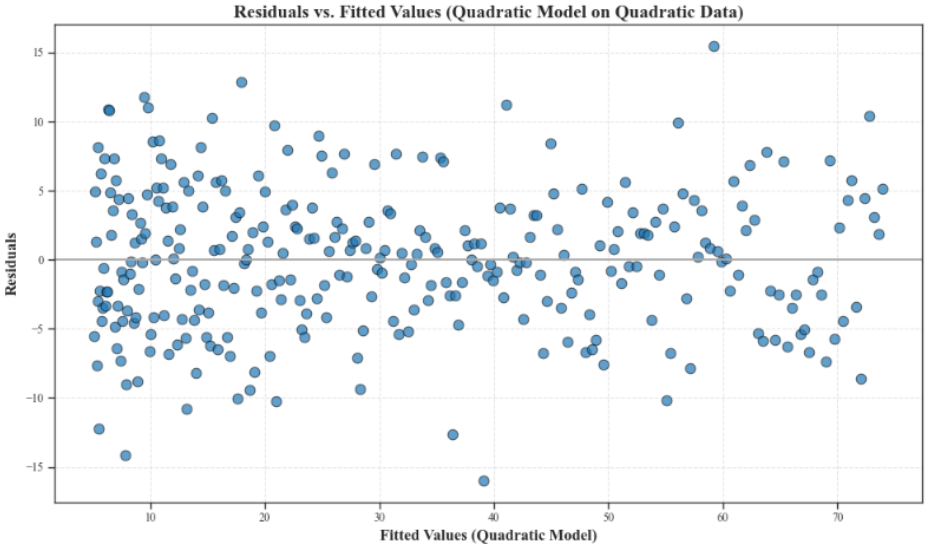
Identifying Non-Linearity

Non-linearity occurs when the relationship between independent and dependent variables is not well captured by a linear model. This can be diagnosed by analyzing residuals—if they exhibit a systematic pattern rather than random scatter, the model may be misspecified.

A **U-shaped** or **inverted-U pattern** in a residual vs. fitted values plot suggests that the model is missing a quadratic or higher-order term. This indicates that a simple linear model does not adequately describe the relationship. Another common sign of non-linearity is when residuals consistently increase or decrease over the range of predictor values, violating the assumption that residuals should be randomly distributed.

To address non-linearity, one approach is **feature engineering**, where polynomial terms (e.g., X2X^2, X3X^3) are introduced to better capture curved relationships. Another solution is to use **non-linear regression models**, such as **decision trees, random forests, or neural networks**, which do not assume a strict linear form. Additionally, **spline regression** and **Generalized Additive Models (GAMs)** allow for flexible, data-driven modeling of non-linear patterns.

Detecting and correcting non-linearity ensures that a model better represents real-world relationships, improving predictive accuracy and inference reliability.



Detecting Autocorrelation

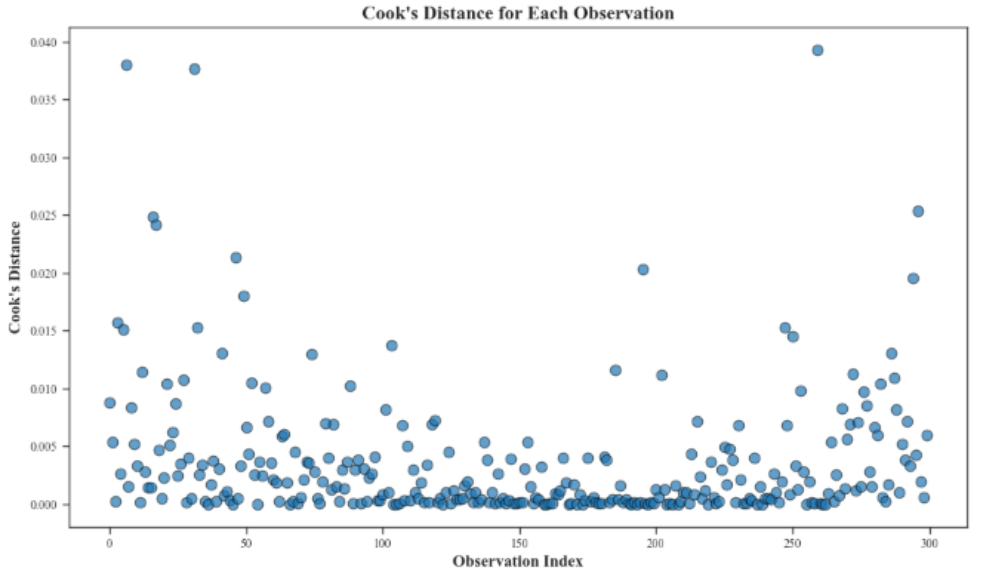
Autocorrelation occurs when residuals in sequential observations are correlated, meaning past errors influence future errors. This violates the assumption of independence in regression models and is common in time series data, leading to biased estimates and unreliable confidence intervals.

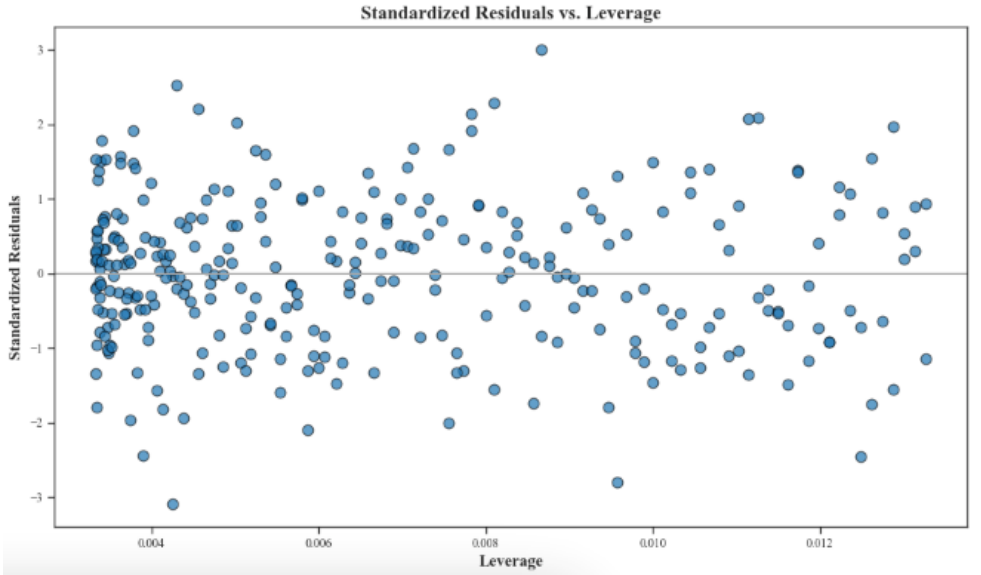
A **Residual vs. Time plot** is a simple way to detect autocorrelation. If residuals exhibit patterns, such as cycles or trends instead of random scatter, it suggests that errors are dependent over time. For instance, positive autocorrelation means residuals remain high or low for consecutive periods, while negative autocorrelation indicates alternating patterns.

A formal statistical test for detecting autocorrelation is the **Durbin-Watson (DW) test**. The DW statistic ranges from 0 to 4:

- **~2**: No autocorrelation.
- **<2**: Positive autocorrelation (common in financial and economic time series).
- **>2**: Negative autocorrelation (less common but occurs in differenced data).

To address autocorrelation, models like **ARIMA (AutoRegressive Integrated Moving Average)** or **Generalized Least Squares (GLS)** can be used. Additionally, introducing lagged variables or using robust standard errors can help account for time-dependent structures.





Checking for Outliers and Influential Points

Outliers and high-leverage points can disproportionately impact a regression model, leading to biased estimates and misleading interpretations. **Outliers** are observations with large residuals, while **high-leverage points** have extreme predictor values that can significantly affect the regression line.

To detect them, **Cook's Distance** measures the influence of each observation on regression coefficients—values above 0.5 or 1 indicate potential issues. **Leverage plots** (based on the hat matrix) identify points with excessive influence on predictions.

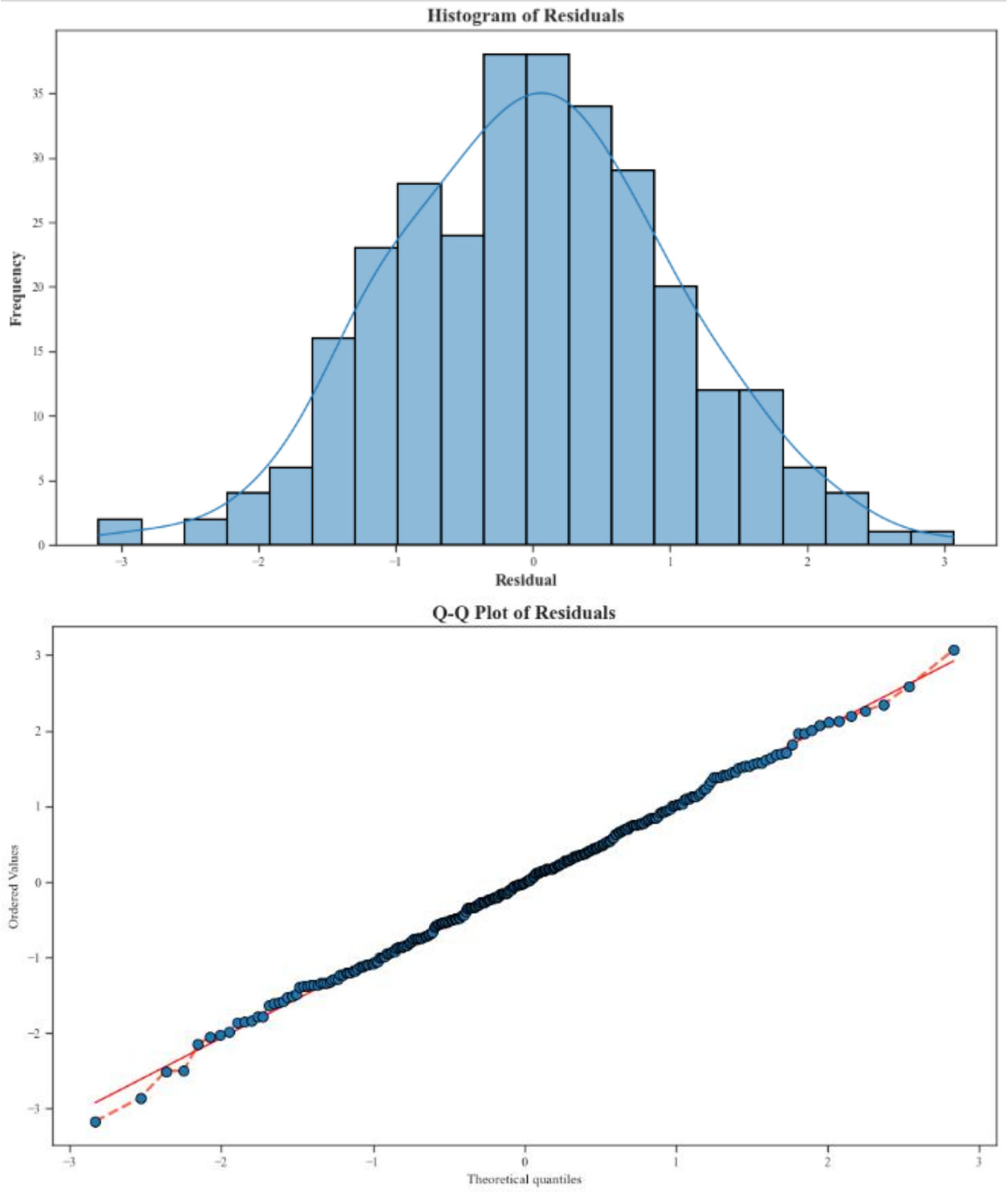
FIGURE

Assessing Normality of Residuals

The normality of residuals is a key assumption in regression, affecting hypothesis tests and confidence intervals. If residuals deviate significantly from normality, p-values and standard errors may be unreliable.

Diagnostic tools include:

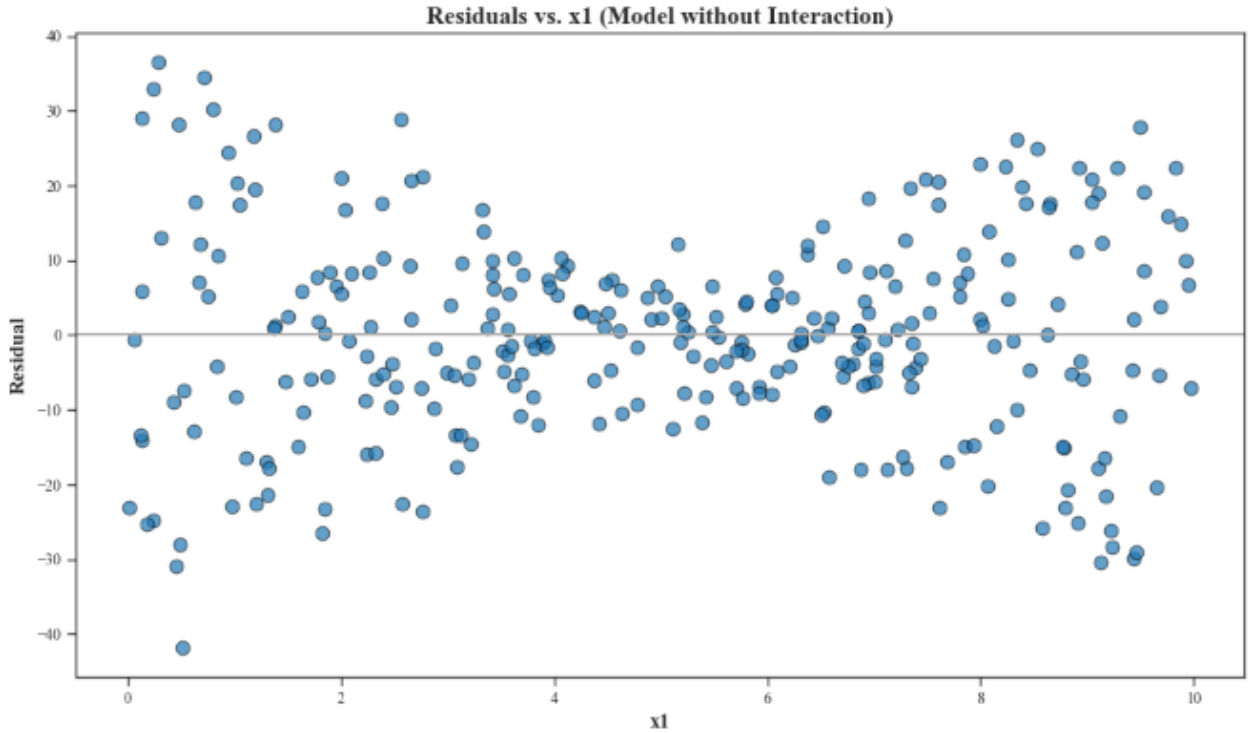
- **Histograms**: A bell-shaped distribution suggests normality.
- **Q–Q Plots**: Compare residual quantiles to a normal distribution—deviations from the diagonal indicate non-normality.
- **Shapiro-Wilk Test**: A formal test where a low p-value (<0.05) suggests non-normal residuals.

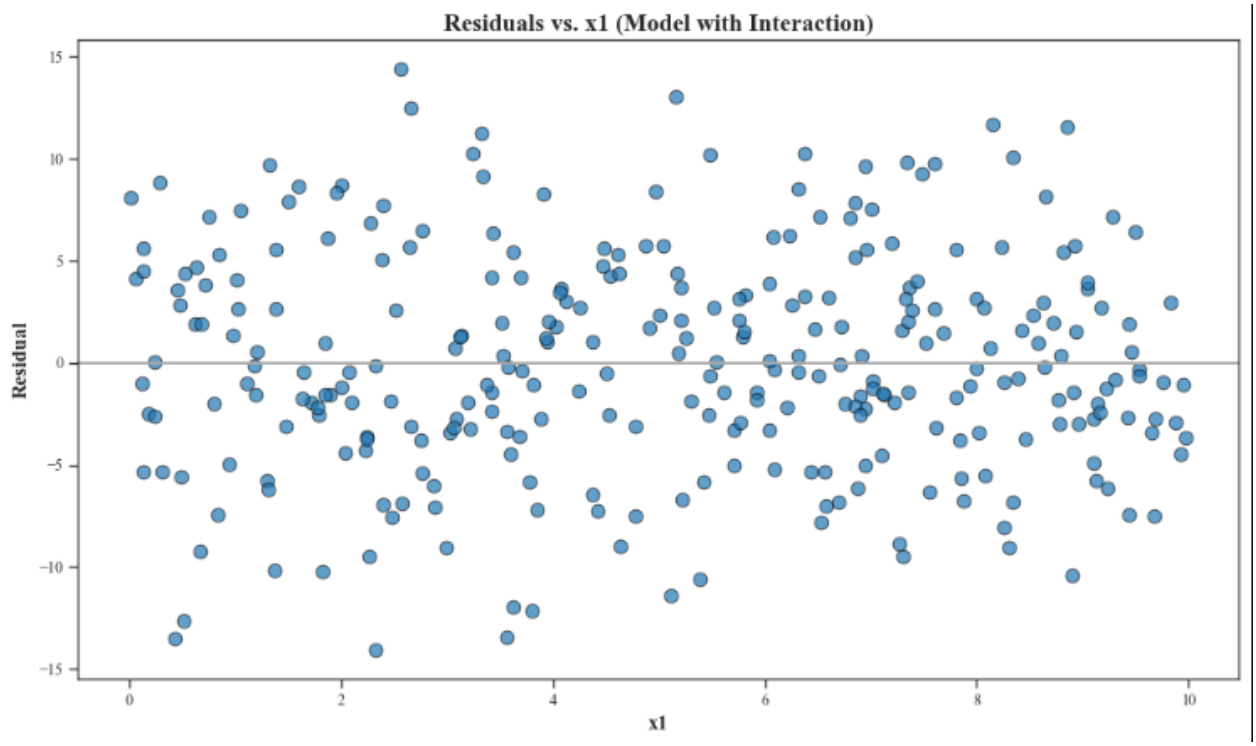


Feature Engineering & Model Improvement

Residual analysis helps identify missing features or interactions in a model. If residuals plotted against a predictor show a pattern (e.g., a trend or curve), it suggests that the predictor's effect may be **non-linear** or dependent on another variable.

Adding **polynomial terms** or **interaction effects** can improve model fit. For instance, if residuals display structure in a model without interactions but become randomly scattered after including an interaction term, it confirms that the added feature enhances predictive accuracy.

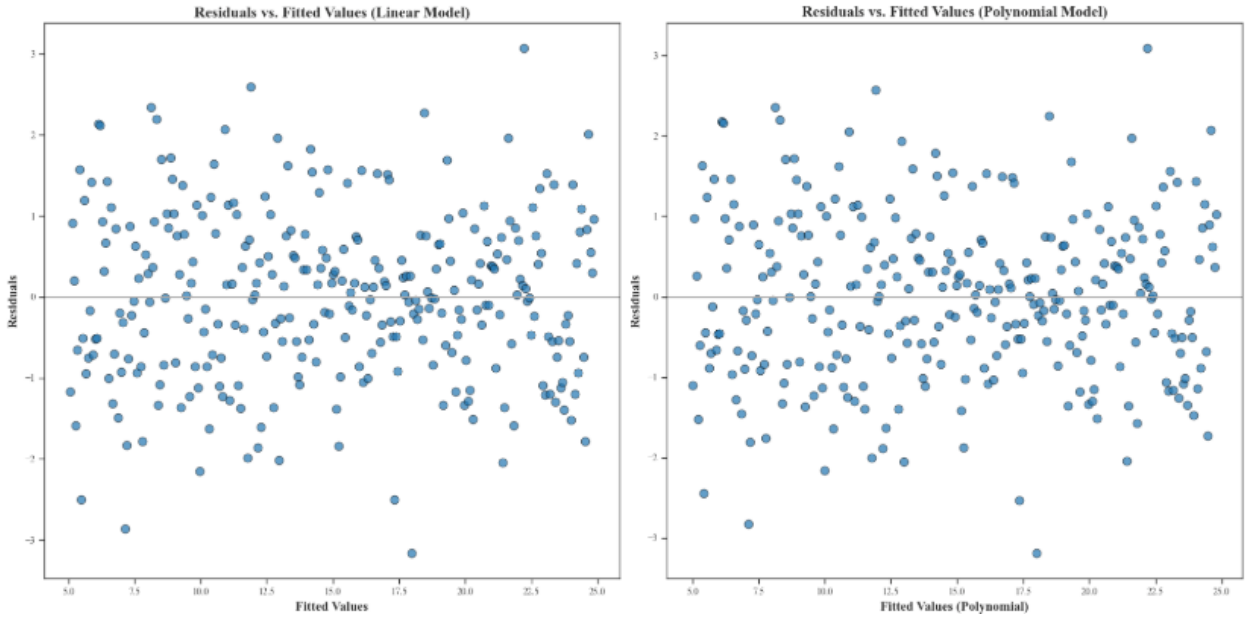




Comparing Multiple Models

Residual analysis helps evaluate and compare different models by assessing their predictive accuracy and assumption validity. **Residual-based metrics** such as **Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE)** quantify model performance—lower values indicate better fit.

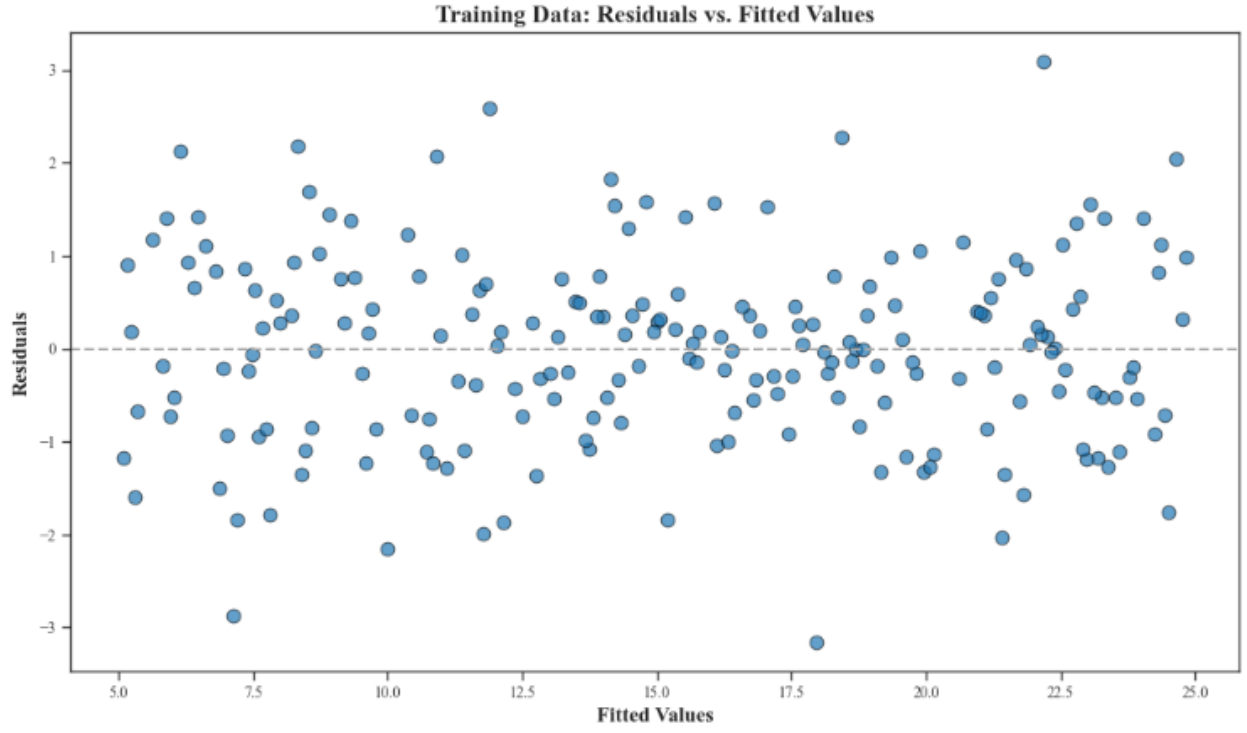
Residual plots also provide insights: a good model should have randomly scattered residuals with no discernible patterns. If one model shows a systematic structure while another has more evenly distributed residuals, the latter is preferred.

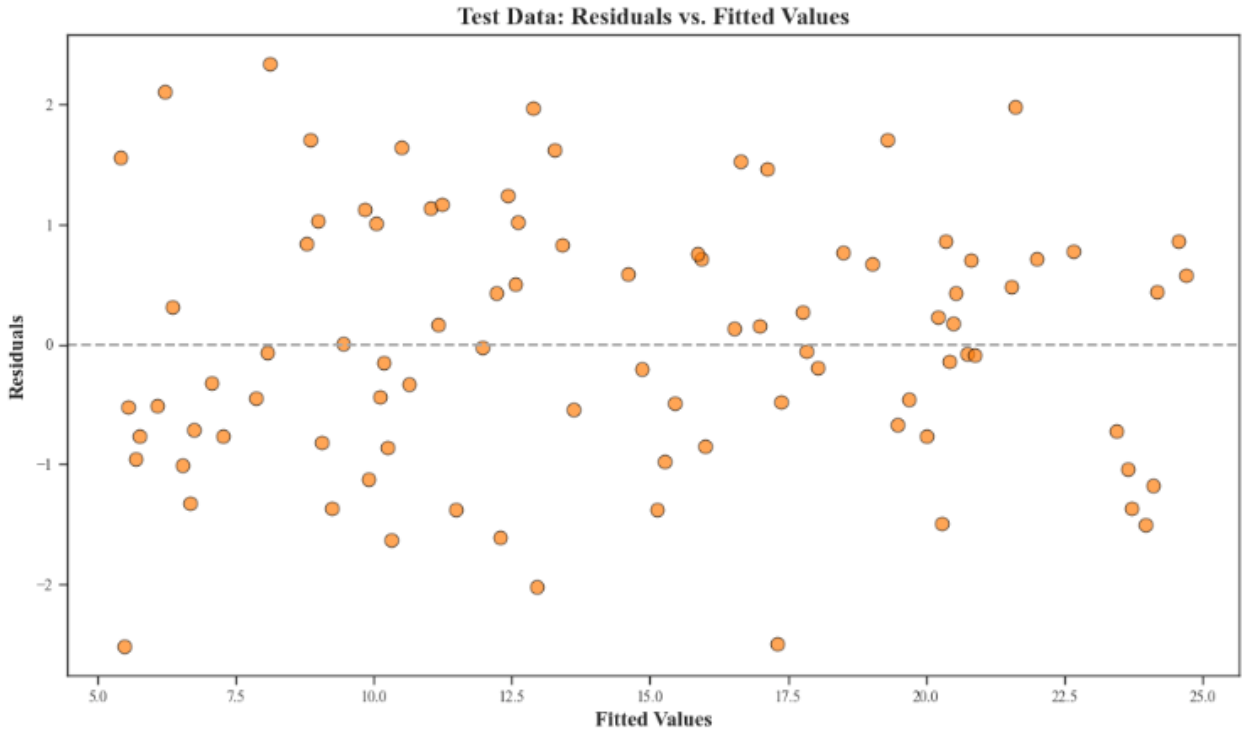


Bias–Variance Tradeoff Diagnosis

Residual analysis helps diagnose the **bias-variance tradeoff**, which affects a model's generalization ability. By comparing residuals on the **training** and **test** sets, we can assess whether a model underfits or overfits the data.

- **High bias (underfitting)**: If residuals exhibit large errors on both training and test sets, the model is too simplistic and fails to capture underlying patterns.
- **High variance (overfitting)**: If training errors are low but test errors are significantly higher, the model is too complex and sensitive to noise.





This section has detailed how residual analysis can be used diagnostically to:

- Check model fit and confirm randomness of residuals.
- Detect heteroscedasticity, non-linearity, autocorrelation, and outliers.
- Guide feature engineering and model improvement.
- Compare multiple models using quantitative and graphical diagnostics.
- Diagnose the bias-variance tradeoff via training/test comparisons.

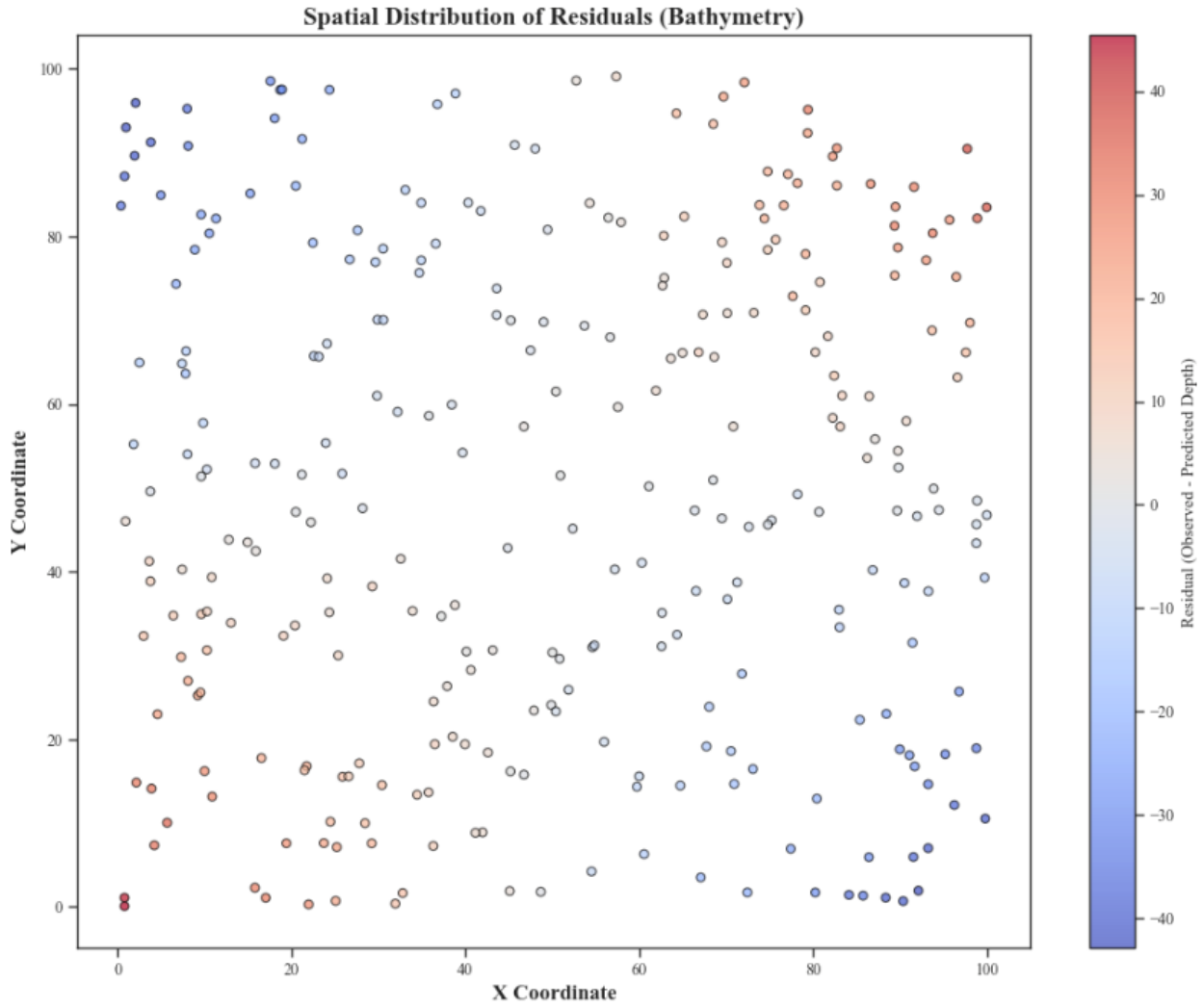
Applications and Case Studies of Residual Analysis

Residual analysis is a crucial component of model validation across various domains. While it is often discussed in theoretical contexts, its real-world applications span finance, environmental science, healthcare, and machine learning. By identifying patterns in residuals, analysts can refine models, detect missing variables, and improve predictive accuracy. This section explores several key applications and presents a case study using SHAP (Shapley Additive exPlanations) values to attribute sales impact changes.

4.1 Financial and Environmental Applications

In **finance**, residual analysis plays a key role in evaluating risk models and asset pricing models. Financial analysts use residuals to assess whether a model's predictions align with actual market behavior. For example, in stock return modeling, **heteroscedastic residuals** suggest that market volatility is time-dependent. This insight leads to the adoption of more sophisticated models, such as **GARCH (Generalized Autoregressive Conditional Heteroskedasticity)**, which explicitly accounts for changing variance over time. Similarly, in credit risk assessment, non-random residuals may indicate that borrower-specific factors are missing from the model, prompting further refinement.

In **environmental science**, residual analysis is widely used in spatial modeling and remote sensing. One notable example is **bathymetry modeling**, where researchers estimate water depth using remote sensing data. If residuals show spatial correlation, it suggests that the model has failed to capture certain environmental factors, such as variations in seafloor reflectance. This prompts the use of **geostatistical corrections** or the inclusion of additional variables, such as temperature or sediment composition, to improve depth predictions.



4.2 Healthcare and Dynamic Treatment Regimes

In **healthcare**, residual analysis is essential in validating models used for **dynamic treatment regimes (DTRs)**. These regimes guide sequential medical decisions, such as adjusting drug dosages based on patient response over time. Methods like **Q-learning**, a reinforcement learning technique, rely on residuals to ensure the model accurately captures patient reactions. If residuals show systematic patterns—such as clustering based on patient demographics—it may indicate that key predictors, such as genetic markers or coexisting conditions, are missing. This insight helps refine personalized treatment plans, leading to better patient outcomes.

Residual analysis is also used in **clinical trials** to verify the effectiveness of new treatments. In survival analysis, non-random residuals might suggest that a treatment's effect varies among subpopulations, prompting further investigation into potential interactions.

4.3 Machine Learning: Boosting and Deep Residual Networks

Modern machine learning techniques leverage residuals in fundamental ways.

1. **Gradient Boosting**: Many ensemble models, such as **XGBoost, LightGBM, and CatBoost**, build trees sequentially by fitting each new model to the residuals of the previous one. This approach ensures that new models correct the errors of earlier iterations, leading to improved predictions.
2. **Deep Residual Networks (ResNets)**: In deep learning, **ResNets** explicitly model residual functions. Instead of learning the full mapping from inputs to outputs, these networks learn the difference (residual) between the current prediction and the true target. This architecture prevents the problem of **vanishing gradients**, enabling the training of extremely deep neural networks that achieve state-of-the-art performance in image recognition, natural language processing, and other fields.

Residual analysis, therefore, is not just a diagnostic tool but an integral part of how modern machine learning models learn and improve.

4.4 Sales Impact Attribution Using SHAP Values

A powerful business application of residual analysis is found in **sales impact attribution**. Understanding the drivers of sales fluctuations is crucial for businesses, yet standard regression models may fail to explain all variations. This case study demonstrates how **SHAP values** can quantify the contribution of different factors to sales changes while identifying unexplained residual effects.

Case Study Setup:

- Sales data from **two time periods (July and December)** is used.
- Separate **Random Forest** models are trained for each period.
- **SHAP values** are computed to determine feature importance.

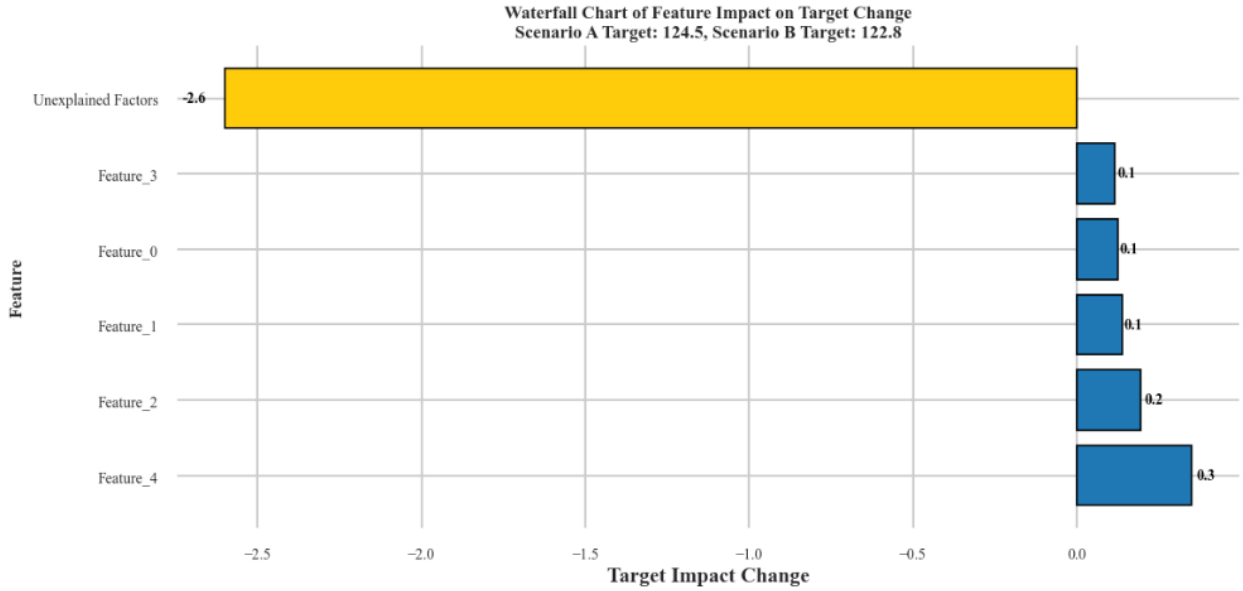
Methodology:

1. Compute the average SHAP values for each feature in **July** and **December**.
2. Calculate the difference in SHAP values between the two periods to determine **each feature's contribution to the sales change**.
3. The sum of all feature contributions provides an estimate of the total explained sales change.
4. Any difference between the **actual observed sales change** and the **sum of SHAP-based contributions** is classified as **residual (unexplained) impact**.

Insights Gained:

- If most of the sales change is explained by known factors (e.g., pricing, marketing spend, seasonal demand), it confirms that the model captures key drivers effectively.
- If a significant portion remains unexplained, it suggests **missing variables**, external shocks (e.g., macroeconomic changes), or data quality issues.

This approach provides businesses with a **transparent, interpretable** method for diagnosing model gaps and refining their sales forecasting strategies.



4.5 Summary of Applications

Residual analysis is an indispensable tool across industries. Whether ensuring the robustness of financial risk models, refining environmental predictions, improving healthcare treatments, or powering modern machine learning algorithms, residuals offer critical insights. Additionally, **business applications** such as SHAP-based sales attribution demonstrate how residual analysis can drive **data-driven decision-making**.

Across all these applications, understanding the "unexplained" portion of model predictions leads to **better models, improved forecasts, and actionable insights**. As data science continues to evolve, residual analysis will remain a key technique for model validation and enhancement.

REFERENCES

[1] Alevizos, E. (2020). *A combined machine learning and residual analysis approach for improved retrieval of shallow bathymetry from hyperspectral imagery and sparse ground truth data*. Remote Sensing, 12(21), 3489. https://doi.org/10.3390/rs12213489

[2] Ertefaie A, Shortreed S, Chakraborty B. Q-learning residual analysis: application to the effectiveness of sequences of antipsychotic medications for patients with schizophrenia. Stat Med. 2016 Jun 15;35(13):2221-34. doi: 10.1002/sim.6859. Epub 2016 Jan 10. PMID: 26750518; PMCID: PMC4853263.

[3] Ramosaj, B., & Pauly, M. (2018). *Consistent estimation of residual variance with random forest out-of-bag errors* [Preprint]. arXiv. https://arxiv.org/abs/1812.06270