Bima Journal of Science and Technology, Vol. 9(2B) Jul, 2025 ISSN: 2536-6041

E-ISSN: 3115-4662 DOI: 10.64290/bima.v9i2B.1287

# Comparative Assessment of the Classical Half-Slope Ratio and the Normalized Difference Ratio as a Robust Linearity Diagnostic Tools

Babangida Ibrahim Babura[1]*

[1]*Department of Applied Mathematics, Federal University of Technology, Babura, Kano Road, Babura, Jigawa State, Nigeria

Corresponding Author: baburabi.math@futb.edu.ng

## ABSTRACT

Robust quantitative diagnostics are crucial for assessing linearity in Exploratory Data Analysis (EDA), especially when data contain outliers. This work defines and compares two such diagnostics derived from median-based half-slopes calculated across partitioned bivariate data. We examine the established Classical Half-Slope Ratio (CHR), calculated as the ratio of the right to the left half-slope, which indicates linearity near unity but is unbounded. We contrast this with a proposed Normalized Difference Ratio (NDR), formulated as the normalized difference between the half-slopes. The NDR is inherently bounded within [-1, 1], precisely indicates linearity at zero, and directly signals the direction of data curvature through its sign (+/-). Illustrative examples confirm that NDR's magnitude quantifies the degree of non-linearity, while its sign offers clear guidance for data transformations. While both CHR and NDR are valuable outlier-resistant tools complementing visual analysis, the NDR's bounded, centered scale provides distinct advantages for comparative analysis, standardization, and potential algorithmic use.

**Keywords:** Linearity Diagnosis, Robust Statistics, Exploratory Data Analysis, Half-Slope Ratio, Normalized Difference Ratio, Outliers.

## INTRODUCTION

The assessment of linearity in the relationship between variables is a fundamental prerequisite and diagnostic step in numerous statistical modeling and data analysis tasks across diverse fields, from econometrics and environmental science to engineering and bioinformatics. Many classical statistical procedures, most notably Ordinary Least Squares (OLS) regression, rely heavily on the assumption of a linear underlying relationship between the response and predictor variables. Deviations from linearity can lead to biased estimates, incorrect inferences, and poor model predictions. Therefore, effective tools for diagnosing potential non-linearity are essential components of the data analysis toolkit.

Exploratory Data Analysis (EDA), as pioneered by Tukey (1977), emphasizes visual inspection and data-driven exploration prior to formal modeling. Scatter plots are the primary visual tool for assessing bivariate linearity, often supplemented by examining residual plots from initial model fits. While invaluable, visual methods can be subjective and may fail to detect subtle or complex non-linear patterns, particularly in large or high-dimensional datasets. Formal statistical tests for linearity exist, such as F-tests comparing linear models to polynomial alternatives, or the Ramsey RESET test. However, these classical tests often rely on stringent assumptions (e.g., normality and homoscedasticity of errors) and, like OLS itself, can be highly sensitive to the presence of outliers or data contamination.

In practice, datasets frequently contain atypical observations (outliers) that do not conform to the pattern exhibited by the majority of the data. The need to handle such

Bima Journal of Science and Technology, Vol. 9(2B) Jul, 2025 ISSN: 2536-6041
E-ISSN: 3115-4662 DOI: 10.64290/bima.v9i2B.1287

contamination without distorting the analysis led to the development of robust statistics (Ayinde et al., 2015). Robust methods aim to provide reliable results that are resistant to the influence of a bounded fraction of outliers. While robust regression techniques (e.g., M-estimation, S-estimation, MM-estimation, Least Trimmed Squares, Theil-Sen estimation) can provide parameter estimates less affected by outliers, diagnosing linearity within a robust framework remains important (McKean, 2004) . Diagnostics based on residuals from robust fits are possible but can sometimes be complex to implement or interpret (Ayinde et al., 2015) .

An intuitive approach, rooted in EDA principles, involves partitioning the data based on the predictor variable and comparing robust estimates of the slope across different segments (Walters et al., 2006) . Tukey's resistant line, for instance, partitions the data into three groups and uses median summary points to derive robust slope and intercept estimates. The comparison of the slope in the lower segment ($b_L$) versus the upper segment ($b_R$) forms the basis for linearity assessment (Khedidja & Moussa, 2022) . A direct quantitative diagnostic arising from this is the Classical Half-Slope Ratio (CHR), defined as $CHR = b_R/b_L$ . A CHR value near 1 suggests linearity, while deviations indicate specific types of curvature ( 1 for upward, 1 for downward) or non-monotonicity ( 0 ), directly guiding potential data transformations (re-expression). While robust and interpretable, CHR suffers from limitations: its scale is unbounded, making comparisons across datasets difficult, it is asymmetric, and it can be highly sensitive if the denominator slope ($b_L$) is close to zero.

This raises the question: can we formulate a robust linearity diagnostic that retains the intuitive appeal and robustness of comparing median-based half-slopes, but offers a standardized, bounded scale for easier interpretation and comparison? Such a diagnostic, ideally centered at zero for linearity and providing clear directional information, could be valuable for both interactive EDA and integration into more automated data analysis pipelines.

This paper introduces and evaluates a candidate for such a diagnostic, termed the Normalized Difference Ratio (NDR). Based on the same robust half-slopes $b_L$ and $b_R$, the NDR is defined via a normalized difference, specifically $NDR = (b_R - b_L)/|b_R| + |b_L|$ . We hypothesize that this formulation provides a robust measure bounded within $[-1,1]$, where $NDR = 0$ indicates linearity, the sign of NDR indicates the direction of curvature, and the magnitude $NDR$ reflects the degree of non-linearity relative to the overall slope magnitudes.

This work is presented to formalize definition of NDR as a linearity diagnostic tool. Its properties were presented and compare its behavior theoretically and empirically to an established CHR. The NDR was further established using simulated data under various conditions (including linearity, different types of non-linearity, and potential contamination) as well as illustrative real-world data examples. The subsequent sections were organized as follows: Section 2 details an estimator and properties of CHR. Section 3 formally introduce the NDR and analyzes its key properties. Section 4 outlines the methodology for our comparative simulation study. Section 5 presents and interprets the simulation results. Section 6 discusses the findings and their implications. Finally, Section 7 concludes the paper.

## MATERIALS AND METHODS

### The Classical Half-Slope Ratio (CHR)

Tukey (1977) came up with the Classical Half-Slope Ratio (CHR), which is a well-

E-ISSN: 3115-4662 DOI: 10.64290/bima.v9i2B.1287

known diagnostic tool used in Exploratory Data Analysis to check the linearity of the relationship between two variables and facilitate data adaption.

**Determination of CHR Value**

CHR value is obtainable based on the following steps:

***Data Partitioning*:** The data points $(x_i,y_i)$ were put in order by their x-values. Then, the points were split into three groups (Left, Middle, and Right) with about the same number of points in each category.

***Summary Points*:** Find the median x-value and median y-value for each group. This gives us three strong summary points: $(x_L,y_L)$ for the Left group, $(x_M,y_M)$ for the Middle group, and $(x_R,y_R)$ for the Right group.

***Half-Slopes Calculation*:** Then, two "half-slopes" were determined:

- The Left Half-Slope ($b_L$) is the slope of the line that goes from the Left summary point to the Middle summary point given by:

$$b_L=\frac{y_M-y_L}{x_M-x_L}$$

- The Right Half-Slope ($b_R$) is the slope of the line that goes from the Middle to the Right summary points given by:

$$b_R=\frac{y_R-y_M}{x_R-x_M}$$

Were $x_M\neq x_L$ and $x_R\neq x_M$.

***CHR Calculation*:** The CHR is then found by dividing the right half-slope by the left half-slope:

$$CHR=\frac{b_R}{b_L}$$

This definition requires. $b_L\neq 0$

**Interpretation**

The CHR value is interpreted as follows:

$CHR\approx 1$**:** Suggests that $b_L\approx b_R$, means that the slope is the same across the whole data range, which is a strong sign of linearity.

$CHR>1$**:** Implies $b_R>b_L$ (assuming $b_L>0$). The trend is steeper in the right of the median than to the left. This indicates an upward-bending curve (e.g., quadratic $y=x^2$).

$CHR<1$ (and $CHR>0$)**:** Implies that $b_R<b_L$ (having that $b_L>0$). The trend is not as steep on the right side of the data as it is on the left. This means that the curve bends down, (e.g., logarithmic trend $y=logx$ or root $y=\sqrt{x}$).

$CHR<0$**:** Indicates the sign of $b_L$ is opposite to that of $b_R$. This means that the relationship in the data range is not always the same, like a peak or a valley.

The extend of deviation from 1 indicate the strength of the non-linearity in the data while the direction guides appropriate data transformations to achieve linearity.

**Properties**

**Robustness:** CHR is effectively resistant to outliers in both y and x direction and within each group since it uses medians to determine summary points to estimate half-slopes..

**Scale:** Because CHR is a ratio, it doesn't have any dimensions. But its scale has no limits and might go anywhere from $-\infty$ to $+\infty$. This can make comparisons across different datasets difficult and can lead to extreme values if $b_L$ is close to zero.

**Asymmetry:** The definition $b_R/b_L$ is asymmetric. If the alternative $b_L/b_R$ were used, the interpretation relative to 1 would be inverted.

**Sensitivity to** $b_L\approx 0$**:** The CHR can be highly unstable or undefined if the left half-slope $b_L$ is zero or very close to zero.

Bima Journal of Science and Technology, Vol. 9(2B) Jul, 2025 ISSN: 2536-6041
E-ISSN: 3115-4662 DOI: 10.64290/bima.v9i2B.1287

**The Normalized Difference Ratio (NDR)**

The CHR is a helpful and strong diagnostic tool, but its unbounded size might make it hard to understand and compare different datasets or models. To solve this problem, we proposed the Normalized Difference Ratio (NDR), a strong linearity test with a standardized, limited scale that is centered at zero and keeps the intuitive idea of comparing strong half-slopes.

**Definition**

Let $b_L$ and $b_R$ be the robust left and right half-slopes, respectively, calculated as defined in Section 2, based on partitioning the data sorted by the predictor variable $x$ into three groups and finding the slopes between the median summary points of these groups. The Normalized Difference Ratio (NDR) is defined as:

$$NDR = \frac{b_R - b_L}{|b_R| + |b_L|} \quad (1)$$

In the specific case where both half-slopes are zero, $b_L = b_R = 0$, the expression becomes indeterminate $(0/0)$. Since this corresponds to a perfectly linear flat relationship identified by the robust summary points, we define $NDR = 0$ in this instance. For all other cases where at least one slope is non-zero, the denominator $|b_R| + |b_L|$ is strictly positive, and the NDR is well-defined by Equation 1.

**Justification of the Formulation**

We chose the NDR formulation because it has a number of useful qualities for a linearity diagnostic of dataset namely:

- **Centering at Zero for Linearity:** The difference between the right and left half-slopes, $b_R - b_L$, is the numerator in NDR expression. This difference became zero if and only if $b_L = b_R$, and thus the diagnostic will naturally center at $NDR = 0$ for perfect linearity.

**Directional Information:** The sign of the numerator shows the direction of the curvature. If $b_R - b_L > 0$ (i.e., $NDR > 0$) implies the slope increases from left to right (upward curve), while $b_R - b_L < 0$ (i.e., $NDR < 0$) implies the slope decreases (downward curve).

**Normalization and Boundedness:** $|b_R| + |b_L|$ in NDR formulation is a normalization factor reflecting the sum of the absolute magnitudes of the slopes scales the difference $b_R - b_L$. This specific normalization was chosen because:

a. It is non-negative and became zero when both slopes are zero.

b. It is symmetric with respect to $b_L$ and $b_R$.

c. It guarantees the NDR is bounded within $[-1,1]$ inherent from triangle inequality, $|b_R - b_L| \leq |b_R| + |-b_L| = |b_R| + |b_L|$. Consequently dividing by the left hand side of the expressed inequality gives $\frac{|b_R - b_L|}{|b_R| + |b_L|} \leq 1$.

That is $|NDR| = \frac{|b_R - b_L|}{|b_R| + |b_L|} \leq 1$.

**Simplicity and Interpretability:** The operations in NDR formulation are basic arithmetic operations with absolute values applied directly to the half-slopes, maintaining a relatively simple structure.

**Properties and Interpretation**

The NDR as formulated exhibits the following properties relative to its proposed application as a linearity diagnostic tool:

**Range:** The NDR has bounded interval within the closed interval $[-1,1]$.

**Linearity Point:** $NDR = 0$ uniquely identifies the case where the robust half-slopes are equal ( $b_L = b_R$ ), corresponding to linearity as captured by this method.

Bima Journal of Science and Technology, Vol. 9(2B) Jul, 2025 ISSN: 2536-6041
E-ISSN: 3115-4662 DOI: 10.64290/bima.v9i2B.1287

- **Sign Interpretation:** The sign of NDR provides information about the direction of curvature relative to a linear trend:
  - $NDR>0$ : Indicates $b_R>b_L$ . This signals an upward-bending curve. Suggesting data transformations like decreasing powers/logs for y or increasing powers for x etc.
  - $NDR<0$ : Indicates $b_R<b_L$ , a case of a downward-bending curve. Guides towards data transformations like increasing powers for y or decreasing powers/logs for x.
- **Magnitude Interpretation:** The absolute value on NDR ( $|NDR|$), indicate the degree of non-linearity on a normalized scale from 0 (linear) to 1 (maximal deviation). It represents the magnitude of the difference between slopes relative to the sum of their individual magnitudes. Larger $|NDR|$ values signify stronger non-linearity.
- **Boundary Cases ( $NDR=\pm 1$ ):** The NDR value reaches its highest value of 1 under conditions representing maximal non-linearity within this framework. This occurs if $|b_R-b_L|=|b_R|+|b_L|$, which implies either:

a. One of the slopes is zero (but not both). E.g., if $b_L=0, b_R\neq 0$ , then $NDR= b_R/|b_R| =\text{sign}(b_R)$ . This corresponds to a data fit that is horizontal in one segment and sloped in the other.

b. The slopes have opposite signs and equal magnitude ($b_R=-b_L\neq 0$). E.g., if $b_R=c>0$ and $b_L=-c$ , then $NDR= \big(c-(-c)\big)\big/c+c$ . This corresponds to a symmetric V-shape or inverted V-shape non-monotonicity.

Thus, $|NDR|=1$ flags serious deviation from linearity.

- **Robustness:** The NDR demonstrate robustness properties from the underlying median-based half-slopes $b_L$ and $b_R$. Medians are known for their resistance to outliers, having a high breakdown point within their calculation group. While a formal derivation is beyond this scope, the NDR is expected to possess good qualitative robustness against vertical outliers and moderate robustness against leverage points affecting the median calculations. Its influence function is expected to be bounded due to the normalization and the bounded influence of medians, contrasting with the potentially unbounded influence function associated with ratio-based measures like CHR when the denominator slope approaches zero. The breakdown point of the overall procedure depends on the partitioning strategy and the median calculation within groups.

**Symmetry Property:** The NDR is anti-symmetric with respect to the order of the slopes: $NDR(b_R,b_L)=-NDR(b_L,b_R)$ . This reflects its nature as a normalized difference.

These properties suggest that the NDR provides a well-behaved, interpretable, and robust diagnostic measure suitable for assessing linearity on a standardized scale.

**Simulation Study Design**

To evaluate the performance and properties of the NDR and compare it with the CHR, we designed a Monte Carlo simulation study and some analyses on real-world datasets.

The primary objectives of the simulation study are:

- To assess the ability of NDR and CHR to distinguish linear from various non-linear relationships under ideal conditions (normally distributed errors, no outliers).
- To evaluate the robustness of both diagnostics to heavy-tailed error distributions and various types of data contamination (vertical outliers, leverage points).

E-ISSN: 3115-4662 DOI: 10.64290/bima.v9i2B.1287

- To compare the stability (bias and variance) of NDR and CHR values under these different conditions.

**Data Generating Processes (DGPs)**

We simulated bivariate data $(x_i,y_i)$, for $i=1,...,N$, according to the following models:

- **Linear (Null Model):** Consider $y_i=\beta_0+\beta_1 x_i+\epsilon_i$. for linearity in $(x_i,y_i)$. With parameter set at $\beta_0=5,\beta_1=2$.
- **Monotonic Non-linear Models:**
  - Quadratic (Upward Curve): Consider $y_i=5+2x_i+\beta_2 x_i^2+\epsilon_i$. In this case $\beta_2$ varied (e.g., 0.05, 0.1, 0.2) to have some control at the degree of non-linearity.
  - Square Root (Downward Curve): Consider $y_i=5+\beta_3\sqrt{x_i}+\epsilon_i$. With $\beta_3$ varied (e.g., 5, 10, 15). $x_i$ generated strictly positive.
  - Exponential (Upward Curve): Consider $y_i=\beta_4 exp(\beta_5 x_i)+\epsilon_i$. $\beta_4=5$, $\beta_5$ varied (e.g., 0.05, 0.1).
- **Non-monotonic Models:**
  - V-Shape: $y_i=5+\beta_v\vee x_i-c\vee+\epsilon_i$. Center $c$ typically set to median of $x_i$. $\beta_v$ varied (e.g., 1, 2, 3).
  - Sine Wave: $y_i=5+\beta_s sin\left(\frac{2\pi x_i}{P}\right)+\epsilon_i$. Amplitude $\beta_s$ and period $P$ varied relative to the range of $x$.

The predictor variable $x_i$ was generated either as a sequence $1,...,N$ or drawn from a Uniform(0, 20) distribution. Sample sizes ($N$) considered were $N\in\{21,51,101\}$.

**Error Distributions and Contamination**

For each DGP, the error term $\epsilon_i$ was generated from one of the following distributions, scaled to have a specific standard deviation $\sigma$ (e.g., $\sigma=1.5$):

- **Standard Normal:** $\epsilon_i\sim N(0,\sigma^2)$.

**Heavy-tailed:** Student's t-distribution with 3 degrees of freedom ( $\epsilon_i\sim t_3$ ), scaled appropriately.

**Contaminated Normal (for robustness):** $\epsilon_i\sim(1-\delta)N(0,\sigma^2)+\delta N$, with contamination proportion $\delta\in\{0.05,0.10\}$ and variance inflation $k=3$.

Additionally, specific outlier scenarios were simulated on data generated with standard normal errors:

**Vertical Outliers:** A proportion $\delta$ (e.g., 10

**Leverage Points:** A proportion $\delta$ of $x_i$ values were replaced with outlying values (e.g., $x_{max}+\text{range}$), paired with either their original $y_i$ (good leverage) or a shifted $y_i$ (bad leverage).

2.3.3 Diagnostic Calculation and Evaluation Metrics

For each generated dataset under each scenario, the half-slopes $b_L,b_R$ were calculated using the 3-group median method detailed earlier, followed by the calculation of CHR and NDR. We performed $M=2000$ Monte Carlo replications for each scenario (results presented are based on $M=500$ for brevity). Performance was evaluated based on:

**Distribution under Linearity (H0):** The empirical distribution (mean, variance, quantiles) of CHR and NDR when the true DGP is linear. We assessed the closeness of the median CHR to 1 and the median NDR to 0, and their variability (e.g., Interquartile Range - IQR).

**Distribution under Non-Linearity (H1):** The empirical distribution of CHR and NDR for each non-linear DGP. We assessed how well the distributions separate from the null distribution (e.g., comparing medians and IQRs). The proportion of NDR values having the correct sign (positive for upward curves, negative for downward) was calculated.

E-ISSN: 3115-4662 DOI: 10.64290/bima.v9i2B.1287

- **Robustness Evaluation:** Comparison of the bias (deviation of the median diagnostic value from its expected value under no contamination) and variance (e.g., IQR or Median Absolute Deviation - MAD) of CHR and NDR under different error distributions and outlier contamination schemes relative to the clean normal error case.

**Real-World Data Application**

A publicly available datasets known or suspected to exhibit non-linearity or contain outliers would be analyzed to illustrate the practical application of CHR and in comparison with NDR, Each dataset, scatter plots, CHR, and NDR values would be computed and interpreted. Specifically, we deploy practical application and comparative behavior of the CHR and NDR diagnostics, as applied to several well-known real-world datasets available in R programming packages namely: the 'cars' dataset , the 'faithful' geyser data (Azzalini 1990, ), and Anscombe's quartet (Anscombe 1973).

**Computational Implementation**

All simulations and calculations were performed using the R statistical programming environment (Version 4.3.1). Specific R packages used included 'dplyr' for data manipulation, 'ggplot2' for plotting, and 'stats' for median calculations.

## RESULTS

**Simulation Results**

In this section, we presents the results of the Monte Carlo simulation study designed to evaluate the performance of the proposed NDR measure and compare it with the CHR estimator under various conditions based on the methodology outlined in Subsection 2.3.

3.1.1 Performance under Different Model Structures (No Contamination)

We first examine the ability of the diagnostics to distinguish between linear and non-linear models using data generated with standard normal errors and no additional contamination.

Figure 1 displays the distribution of NDR values. In this case. the NDR distributions for the linear model are tightly centered around zero across all sample sizes (N=21, 51, 101), validating its hypothesized ability to correctly identify linearity under ideal conditions. For the monotonic non-linear models, the we can observe NDR ability to captures both the presence and direction of curvature. The upward-bending quadratic model' yields NDR distributions been centered above zero (median NDR ≈ 0.18), while the downward-bending 'a square root mode**l'** yields distributions clearly centered below zero (median NDR ≈ -0.20). The magnitude of the median NDR in these cases provides an indication of the average degree of non-linearity detected for these specific model parameters. Notably, for the non-monotonic V-shape model, the NDR distribution is concentrated near the boundary value of +1. This distinct result arises because the symmetric V-shape leads to $b_R \approx -b_L$ , mapping cleanly to NDR $\approx +1$ according to its definition, thereby clearly signaling a strong structural deviation from linearity different from the monotonic curves. The precision of the NDR estimates, indicated by the decreasing interquartile range (IQR) with increasing N, enhances the visual separation between the distributions for different model types, particularly improving the distinction between linear and mildly non-linear patterns at larger sample sizes.

Bima Journal of Science and Technology, Vol. 9(2B) Jul, 2025 ISSN: 2536-6041
E-ISSN: 3115-4662 DOI: 10.64290/bima.v9i2B.1287

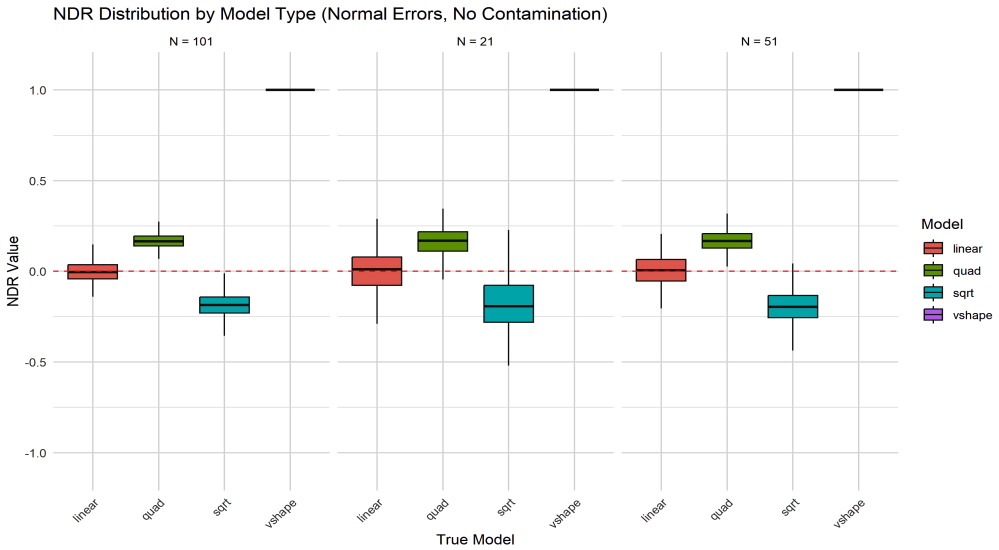


**Figure 1:** Distribution of NDR values by true model type and sample size (N), under normal errors with no contamination. The red dashed line indicates NDR = 0 (perfect linearity).

Figure 2 presents the corresponding distributions for the CHR on a log10 scale (where CHR=1 corresponds to 0). The CHR performs similarly well for the linear (centered at 1) and monotonic non-linear models (quadratic > 1, square root < 1), correctly identifying the direction relative to linearity. However, the visualization for the V-shape model illustrates a practical limitation. Since the expected CHR is negative ($\approx -1$), it cannot be displayed on the log scale necessary to view the range of the other models adequately. The resulting boxplot shows high variability and fails to convey the specific non-monotonic structure as clearly as the NDR's mapping to +1. While CHR's precision for linear and monotonic models also improves with N, its utility for simultaneous visual comparison across all tested model types is hampered by its unbounded scale and the sign issue with non-monotonic cases.

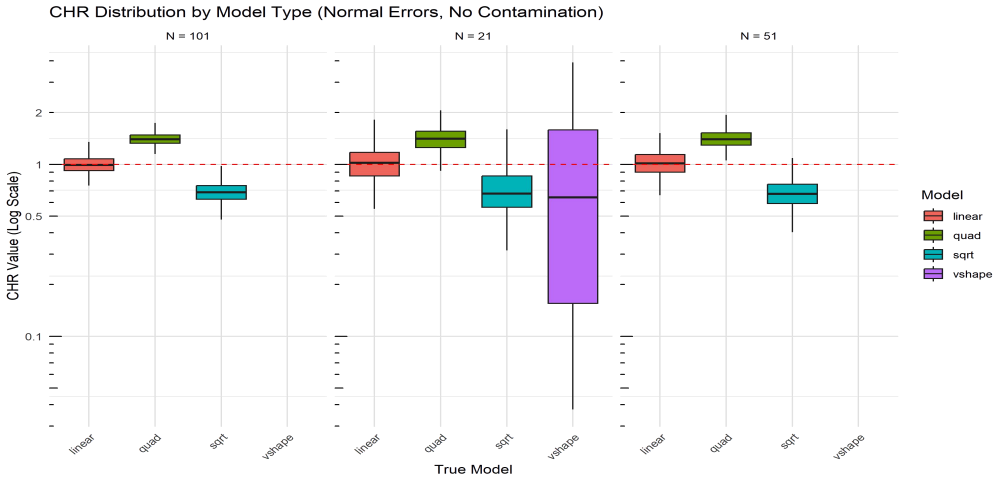


**Figure 2:** Distribution of CHR values (log10 scale) by true model type and sample size (N), under normal errors with no contamination. The red dashed line indicates CHR = 1.

Bima Journal of Science and Technology, Vol. 9(2B) Jul, 2025 ISSN: 2536-6041
E-ISSN: 3115-4662 DOI: 10.64290/bima.v9i2B.1287

Examining Figures 1 and 2, the NDR's main benefit in these settings is its restricted [-1,1] scale. This lets you see how it behaves in linear, monotonic non-linear, and non-monotonic circumstances at the same time, without having to change the scale, which can hide some results (such negative CHR values).

**Robustness to Error Distribution**

Robustness against non-normal errors is a key desideratum. Figure 3 investigates this by comparing the NDR distribution for the linear model when errors follow a standard normal versus a heavy-tailed t-distribution with 3 degrees of freedom ($t_3$).

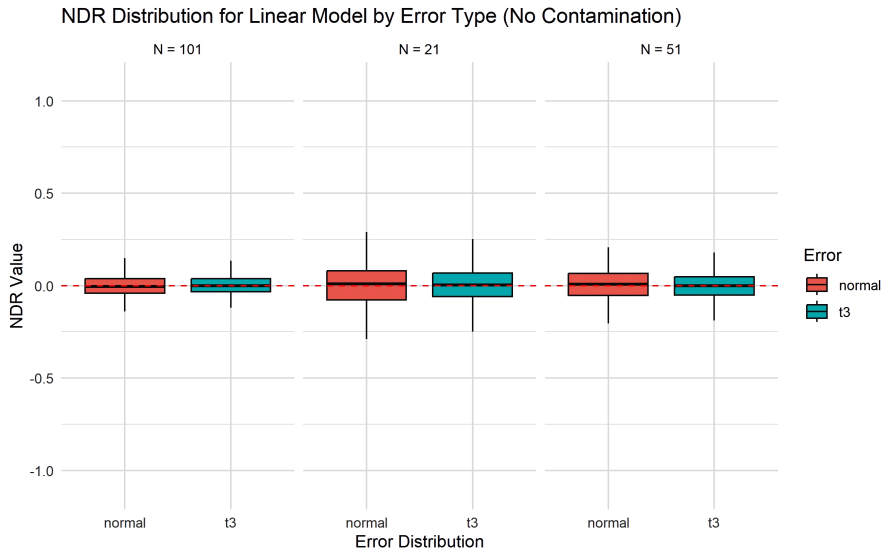


**Figure 3:** Distribution of NDR values for the linear model under Normal vs. t(3) error distributions (no contamination), faceted by sample size (N). The red dashed line indicates NDR = 0.

The results in Figure 3 are compelling. The median NDR remains tightly centered around zero for both error distributions across all sample sizes. This stability highlights the robustness of the diagnostic's central tendency, stemming directly from the use of medians in calculating the half-slopes, which are known to be resistant to the influence of extreme values present in heavy-tailed distributions. While the median remains stable, the variability of the NDR estimate, as indicated by the slightly larger IQR for the $t_3$ case, increases somewhat. This implies that while NDR reliably indicates linearity on average even with heavy-tailed noise, individual estimates may fluctuate more. Nevertheless, the core ability to identify the underlying linear structure is preserved. A similar pattern of median stability but potentially increased variability under heavy tails would be expected for CHR due to its reliance on the same median-based slopes.

**Robustness to Contamination**

Robustness is an important property for evaluation of NDR and conventionally resistance assessing performance is done under explicit data contamination process. The simulation design included scenarios with 10% vertical outliers and 10% bad leverage points applied to the linear model with normal errors. Based on the properties of median-based estimators we consider:

**Vertical Outliers:** We consider both NDR and CHR having inherit robustness from

Bima Journal of Science and Technology, Vol. 9(2B) Jul, 2025 ISSN: 2536-6041
E-ISSN: 3115-4662 DOI: 10.64290/bima.v9i2B.1287

median estimation with high breakdown point. Therefore, the diagnostics should remain centered near their ideal linearity values (0 and 1, respectively), although variability might increase. This reflects the median's resistance to extreme Y-values that do not have unusual X-values.

- **Bad Leverage Points:** These pose a greater problem due to extreme X-values and deviation from the pattern of the bulk of the dataset. Lverage points can potentially influence the median calculation within one of the three data partitions, causing biase for both NDR away from 0 and CHR away from 1. Quantifying the magnitude of this bias and the inflation of variance, especially comparing NDR and CHR under identical leverage contamination, remains a key task for fully characterizing their relative robustness.

**Real-World Data Application Results**

Table 1 provided practical insights into the behavior of CHR and NDR to some selected real-world datasets application.

**Table 1:** CHR and NDR Estimates on Some Real World Datasets.

| | Dataset | $b_L$ | $b_R$ | CHR | NDR |
|---|---|---|---|---|---|
| 1 | cars | 4.20 | 5.20 | 1.24 | 0.11 |
| 2 | faithful | 0.10 | 0.04 | 0.38 | -0.44 |
| 3 | anscombe1 | 0.52 | 0.41 | 0.80 | -0.11 |
| 4 | anscombe2 | 1.06 | 0.12 | 0.11 | -0.80 |
| 5 | anscombe3 | 0.35 | 0.39 | 1.12 | 0.06 |

For the 'cars' dataset, both diagnostics suggested a mild upward curve ($CHR=1.24, NDR=0.11$), consistent with the physical expectation that stopping distance increases more than linearly with speed. The 'faithful' dataset showed in a more observable downward curve pattern in terms of slope change ($CHR=0.38, NDR=-0.44$), reflecting its known complex bivariate structure which is not captured by a single linear trend.

The Anscombe quartet data is considered to offer a more particularly illustrative comparisons (Anscombe 1973). Anscombe Set 1 is established to be linear, but both CHR value (0.80) and NDR value (-0.11) indicated a slight deviation, which can be attributed to the small sample size and specific data point placement influencing the median splits. But for Anscombe Set 2 (quadratic non-linearity), both CHR (0.11) and NDR (-0.80) strongly and correctly signaled a significant downward-curving trend in the slopes. However, the most notable is for Anscombe Set 3 (linear with a single prominent outlier), both diagnostics demonstrated excellent robustness: CHR (1.12) remained close to 1, and NDR (0.06) remained very close to 0. This highlights the ability of these median-based methods to resist the influence of isolated outliers and capture the underlying structure of the majority of the data, this is indeed a key advantage in practical data analysis.

Overall, the real-world applications align with the simulation findings, demonstrating the utility of both CHR and NDR in characterizing different data structures and the particular strength of their robustness in the presence of outliers.

A visual inspection of the data and overlaid half-slopes is illustrated in Figures 4 through 8. Supporting the application of CHR and NDR to the selected real-world datasets summary in Table 1.

Bima Journal of Science and Technology, Vol. 9(2B) Jul, 2025 ISSN: 2536-6041
E-ISSN: 3115-4662 DOI: 10.64290/bima.v9i2B.1287

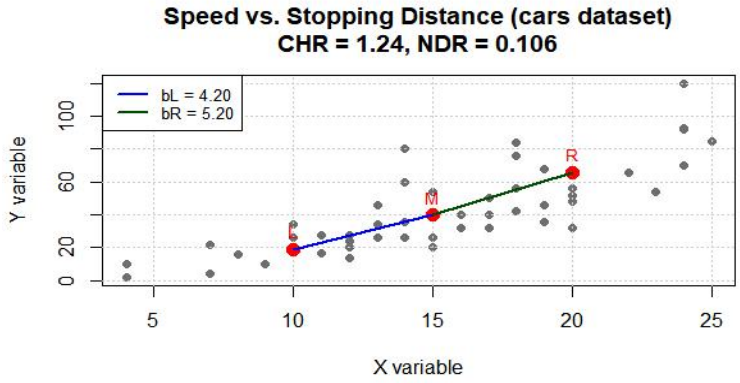


**Figure 4:** Scatter plot of the 'cars' dataset with overlaid median summary points (L, M, R) and half-slopes ($b_L$,$b_R$). CHR = 1.24, NDR = 0.11.

Interactions for the 'cars' dataset is plotted in Figure 4, we can see that the right half-slope ($b_R$=5.20) is visibly steeper than the left half-slope ($b_L$=4.20). This visual observation of an increasing slope is quantitatively supported by $CHR$=1.24 (greater than 1) and $NDR$=0.11 (positive). Both diagnostics suggest a mild upward curve, consistent with the physical expectation that stopping distance increases more than linearly with speed. The NDR value indicates a relatively modest deviation from perfect linearity.

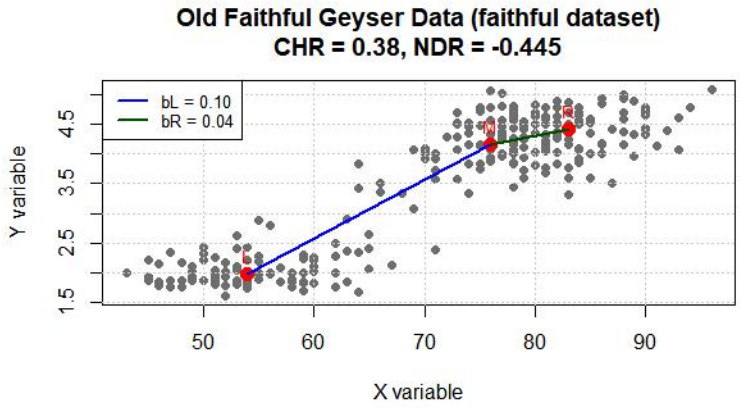


**Figure 5:** Scatter plot of the 'faithful' dataset with overlaid median summary points (L, M, R) and half-slopes ($b_L$,$b_R$). CHR = 0.38, NDR = -0.44.

The 'faithful' dataset plot in Figure 5 displays a more complex structure within two main clusters of data points. The half-slopes values, $b_L$=0.10 and $b_R$=0.04, indicate that the slope connecting the first cluster to the second ($b_L$) is relatively steeper than the slope within the second cluster ($b_R$). This results in $CHR$=0.38 and $NDR$=−0.44. The two diagnostics tools suggest a non-linear relationship, with the negative NDR indicating an overall decrease in the rate of change (a downward curve if a single smooth function were considered). The magnitude of the NDR (-0.44) signals a substantial deviation from linearity than observed in the 'cars' dataset.

For a more insightful comparisons, we deploy the Anscombe quartet provides particularly in the following way:

Bima Journal of Science and Technology, Vol. 9(2B) Jul, 2025 ISSN: 2536-6041
E-ISSN: 3115-4662 DOI: 10.64290/bima.v9i2B.1287

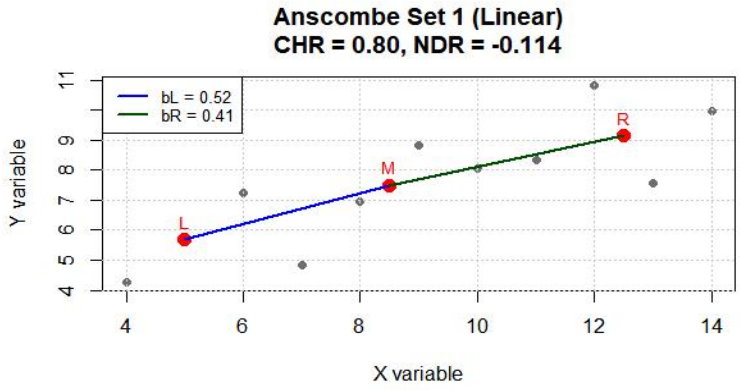


**Figure 6:** Scatter plot of Anscombe Set 1 (Linear) with overlaid median summary points (L, M, R) and half-slopes ($b_L$,$b_R$). CHR = 0.80, NDR = -0.11.

For Anscombe Set 1 of Figure 6, is established data set with approximately linear relation among variables. The visual impression is indeed one of linearity, though with some scatter deviation of points. The half-slopes $b_L$=0.52 and $b_R$=0.41 are relatively close. The resulting $CHR$=0.80 and $NDR$=−0.11 suggest a very slight downward curve. Considering the small sample size (N=11) and the specific arrangement of points in Anscombe's datasets (Anscombe 1973), minor deviations from perfect linearity indicators (CHR=1, NDR=0) are expected due to the sensitivity of the three-group median split to the individual point locations. The NDR value being close to zero show consistency with approximate linearity.

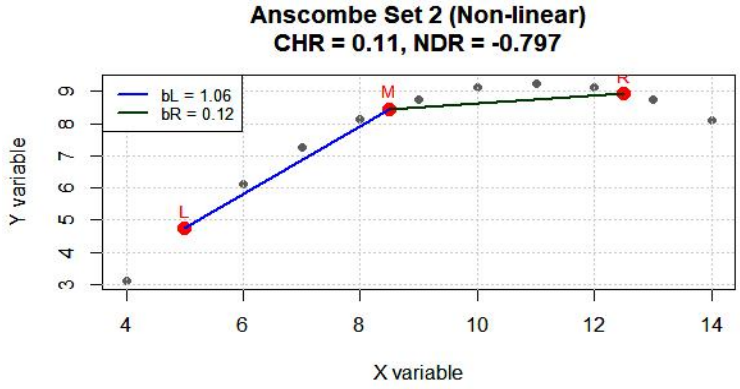


**Figure 7:** Scatter plot of Anscombe Set 2 (Non-linear) with overlaid median summary points (L, M, R) and half-slopes ($b_L$,$b_R$). CHR = 0.11, NDR = -0.80.

Anscombe Set 2 (Anscombe 1973) as visualized in Figure 7, displays a non-linearity pattern close to quadratic form. The left half-slope ( $b_L$=1.06 ) is positive and relatively steep, while the right half-slope ( $b_R$=0.12 ) is much flatter, indicating a significant decrease in the rate of change. This strong visual non-linearity is captured decisively by both diagnostics: $CHR$=0.11 (far from 1) and $NDR$=−0.80 (a large negative value). The NDR value, being close to -1, signals a significant deviation from

Bima Journal of Science and Technology, Vol. 9(2B) Jul, 2025 ISSN: 2536-6041
E-ISSN: 3115-4662 DOI: 10.64290/bima.v9i2B.1287

linearity and correctly identifies the downward-curving nature of the slope.

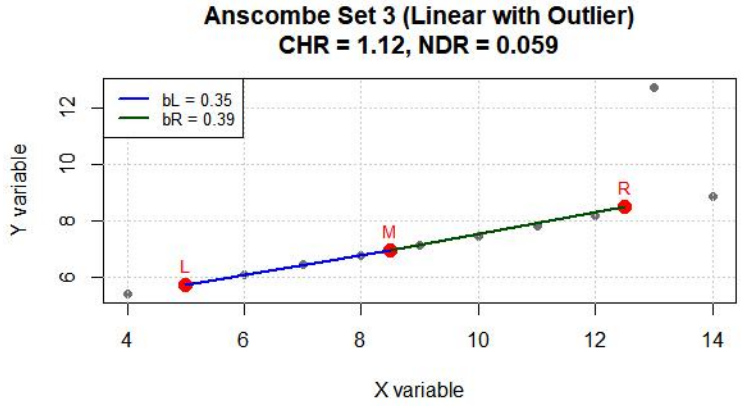


**Figure 8:** Scatter plot of Anscombe Set 3 (Linear with Outlier) with overlaid median summary points (L, M, R) and half-slopes ($b_L$,$b_R$). CHR = 1.12, NDR = 0.06.

Finally, Anscombe Set 3 (Anscombe 1973) in Figure 8 is characterized by an near perfect linear relationship between most points, but with one significant vertical outlier. Visually, the median summary points (L, M, R) and the estimated half-slopes ($b_L$=0.35,$b_R$=0.39) are clearly determined by the main linear trend, effectively handling the effect of outliers. This visual robustness is reflected in the diagnostic values: $CHR$=1.12 (very close to 1) and $NDR$=0.06 (very close to 0). This result is paramount as it empirically demonstrates the robustness of these median-based methods to isolated, extreme outliers, a critical advantage in practical data analysis where such points are common. Both CHR and NDR successfully identify the underlying linearity of the bulk of the data.

In summary, the application of CHR and NDR to these real-world datasets generally aligns with their expected behavior based on the data's known characteristics and visual inspection. They provide quantitative measures that correspond well with different data structures, and importantly, the Anscombe Set 3 example showcases their valuable robustness to outliers.

## DISCUSSION

Both simulation and real-world data application results provide substantial insights into the performance and practical utility of the Normalized Difference Ratio (NDR) as a newly proposed linearity diagnostic tool, especially when compared to an established Classical Half-Slope Ratio (CHR). Our findings under ideal simulation conditions (normal errors, no contamination) confirmed that both diagnostics effectively identify linearity and distinguish it from monotonic non-linear patterns obtainable from Figures 1 and 2. NDR values centered tightly around 0 for linear data, while CHR values centered around 1. For monotonic curves, both correctly reflected the direction of curvature. The real-world application to 'Anscombe Set 1' (Anscombe 1973) which is approximately linear and 'Anscombe Set 2' (Anscombe 1973) establish as quadratic corroborated these findings, with both diagnostics generally aligning with the known data structures. The 'cars' dataset also showed consistent detection of mild non-linearity by both methods, as seen in Figure 4 and Table 1.

E-ISSN: 3115-4662 DOI: 10.64290/bima.v9i2B.1287

Key differences highlighted in the simulation's framework, particularly an established boundedness for NDR's on $[-1, 1]$ scale, proved advantageous. This was evident for the case of non-monotonic V-shape models in the simulations, which NDR clearly flagged with values near +1. CHR's negative value for this case was obscured by the necessary log-scale visualization. The 'faithful' dataset in Figure 5, with its complex structure, yielded a strong negative NDR (-0.44), indicating a significant overall decline in slope estimate across the segments, a pattern also captured, though differently scaled, by CHR (0.38). The robustness analysis against heavy-tailed errors of Figure 3 supported NDR's utility. Its median value remained stable around 0 for linear data even under $t_3$ errors, demonstrating the precision of its median-based formulation. The application to 'Anscombe Set 3'in Figure 8 provided additional compelling real-world evidence of this robustness: accordingly, NDR (0.06) and CHR (1.12) yielded values near their respective linearity indicators, effectively resisting the effect of outliers and capturing the underlying linear trend of the majority data. This practical demonstration of robustness is a critical advantage of these methods.

NDR in practice offered several advantages. Its bounded scale simplifies interpretation and comparison across datasets. The sign of NDR provides immediate, unambiguous guidance for potential data transformations. This was consistent across both simulated and real datasets. For example, the positive NDR for the 'cars' dataset and the negative NDR for 'Anscombe Set 2' directly suggest the type of curvature present.

## CONCLUSION

In this study a Normalized Difference Ratio (NDR) for robust linearity diagnostic is proposed, The NDR's properties were explored and its performance with the established Classical Half-Slope Ratio (CHR) are compared according to simulations and real-world data applications. The key findings are:

Both NDR and CHR effectively identify linearity and non-linear patterns under ideal conditions, as shown in simulations framework and real-world datasets like Anscombe Sets 1 and 2.

The NDR's bounded $[-1,1]$ scale exhibit its clear advantage with linearity at 0 as compared to CHR. It also offers advantages in consistent interpretation and visualization across diverse data structures, including non-monotonic forms (simulation V-shape model) and complex real-world data (e.g., 'faithful' dataset), where CHR's visualization can be problematic.

NDR is robust to heavy-tailed error distributions in simulations, maintaining a stable median at zero for linear data.

Crucially, both NDR and CHR exhibited strong robustness to a significant outlier in the 'Anscombe Set 3' real-world data, correctly identifying the underlying linear trend.

The sign of NDR is balanced and consistent by directly indicating the direction of curvature, providing clear guidance for data transformations, as observed in both simulated curves and real datasets of 'cars' and 'Anscombe Set 2'.

In general, Compared to CHR, the NDR offers a combination of robustness, a standardized bounded scale, and straightforward sign-based directional interpretation. While further investigation into its behavior under severe leverage contamination can be explored, the NDR presents in this research a promising, practical, and robust tool for Exploratory Data Analysis, effectively complementing visual

E-ISSN: 3115-4662 DOI: 10.64290/bima.v9i2B.1287

inspection and aiding in informed model building decisions.

**Acknowledgement**

Appreciation goes to The Department of Applied Mathematics, Federal University of Technology Babura for providing state of the art Simulation Lab facilities which facilitate the extensive computational work of this research. However, The R Core Team are indeed essential enablers for both coding environment and real-life data sets access to validate all findings. Finally, the quality of this submission was greatly improved by adopting the quality feedbacks from the anonymous reviewers of the BIMA Journal.

## REFERENCES

Anscombe, F. J. (1973). Graphs in Statistical Analysis. *The American Statistician*, *27*(1), 17-21.

Ayinde, K. , Lukman, A. and Arowolo, O. (2015) Robust Regression Diagnostics of Influential Observations in Linear Regression Model. *Open Journal of Statistics*, **5**, 273-283. doi: 10.4236/ojs.2015.54029.

Azzalini, A., & Bowman, A. W. (1990). A look at some data on the Old Faithful geyser. *Applied Statistics*, *39*(3), 357-365.

Ezekiel, M. (1930). *Methods of Correlation Analysis*. Wiley.

Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (Eds.). (1983). *Understanding Robust and Exploratory Data Analysis*. Wiley zzalini, A., & Bowman, A. W. (1990). A look at some data on the Old Faithful geyser. *Applied Statistics*, *39*(3), 357-365.

Huber, P. J. (1981). *Robust Statistics*. Wiley.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. Wiley.

Kengwoung, T. M. (2014). A New Robust Method for Nonlinear Regression. *Journal of Biometrics & Biostatistics*, *05*(05). https://doi.org/10.4172/2155-6180.1000211.

Khedidja, D. D., & Moussa, T. (2022). Test for Linearity in Non-Parametric Regression Models. *Austrian Journal of Statistics*, *51*(1). https://doi.org/10.17713/ajs.v51i1.1047.

Maronna, R. A., Martin, R. D., & Yohai, V. J. (2006). *Robust Statistics: Theory and Methods*. Wiley.

McKean, J. W. (2004). Robust analysis of linear models. *Statistical Science*, *19*(4). https://doi.org/10.1214/088342304000000549.

Ramsey, J. B. (1969). Tests for Specification Errors in Classical Linear Least Squares Regression Analysis. *Journal of the Royal Statistical Society. Series B (Methodological)*, *31*(2), 350-371.

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. Wiley.

Sen, P. K. (1968). Estimates of the Regression Coefficient Based on Kendall's Tau. *Journal of the American Statistical Association*, *63*(324), 1379.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley.

Velleman, P. F., & Hoaglin, D. C. (1981). *Applications, Basics, and Computing of Exploratory Data Analysis*. Duxbury Press.

R Core Team (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org.