EDITORIAL

The Role of Log-Transformation in Data Science and Statistical Analysis

Munawar Hussain Soomro,[1,2] Salma Memon,[1,2], Sanjeev Bista[3,4]

The importance and analytical power of log-transformation in data science lies in its ability to address several common data challenges, thereby, improving model accuracy and enriching analytical insights.[1] Log-transformation is a widely used technique in various fields (including statistics, economics, biology and environmental science) to address issues, such as skewed distribution, heteroscedasticity and multiplicative relationships between variables.[2, 3] Hereunder are some key reasons why log-transformation is important in data science.

The majority of real-world datasets are typically skewed or biased, with most of the observations falling at one end of the scale. Skewed data can breach statistical tests and model assumptions, leading to biased estimates and inaccurate inferences.[4] The log-transformation is, arguably, the most popular among the range of transformation options applied to convert skewed data into approximate normality.[1] However, log-transformation is only warranted if the original data follows a log-normal distribution or approximately so. If not, other transformation alternatives must be explored to meet the normality assumption.[1]

To normalize skewed distributions and stabilize variance, variables with non-normal distributions can be log-transformed using the natural logarithm: $X_{log} = \ln(X)$

Regression analyses then can be performed using the following general model: $Y = \beta_0 + \beta_1 X_{log} + \epsilon$, where Y represents the outcome variable, β_0 is the intercept, β_1 is the regression coefficient, and ϵ is the error term.

Variance stabilization and homogenization of error variance: In regression analyses, analysis of variance (ANOVA) and other parametric tests, an important assumption is homoscedasticity, meaning that the variance of errors remains constant across groups or values of independent variables. Violation of this assumption is known as heteroscedasticity and can produce biased standard errors, inaccurate p-values, and unreliable statistical inferences. [2, 5] The log-transformation of dependent variable or independent variable, or both is commonly used to stabilize variance and reduce heteroscedasticity, thereby improving the validity of inferential procedures.[1, 6] However, transformation may not always fully resolve the problem; therefore, researchers should assess residual variance after transformation and consider alternative approaches when heteroscedasticity persists.[7]

Linearization of relationships: Log-transformation can linearize nonlinear relationships between variables, making them easier to interpret and model.[8] In many cases, the relationship between variables may be multiplicative rather than additive. Taking the logarithm of one or both variables can transform multiplicative relationships into additive ones, enabling the use of linear regression models to estimate the effects of predictors on the outcome variable.[9]

Outlier Reduction: Extreme values or outliers have the potential to disproportionately influence statistical analyses and distort parameter estimates.[10] Log-transformations reduce the impact of outliers by compressing or shrinking large values and spreading out small values, which improve the robustness of statistical tests and reduce the influence of extreme observations on the results.[3]

Interpretation of transformed coefficients: Log-transformation is particularly useful when interpreting changes in the variables with exponential growth or decline, such as population, income or biomarker levels, furthermore, log-transformation facilitates the interpretation of percentage changes.[11] When the outcome variable is log-transformed in a regression model, regression coefficients correspond to proportional (or percentage) changes in the original scale of outcome when exponentiated.[11] This approach makes it easier to interpret the effects of predictors in regression models and to compare relative changes across different variables, groups, or time periods, even if their original scales differ widely.

Despite the fact that log transformation is an important element still it has several important limitations. First, it cannot directly handle zero or negative values, as the logarithm is undefined for these cases. Researchers must often add a constant before transformation, which can introduce arbitrariness and bias.[1, 3] Second, the interpretation of transformed

variables, as regression coefficients and statistical estimates become less intuitive because the coefficients represent multiplicative rather than additive effects and may complicate clinical or public health interpretation.[7] Additionally, the log-transformation may not always correct the heteroscedasticity or normalize the data, especially in situations when the distributions contain extreme outlier values.[3] However in some situations, the transformed data may still go against the presumptions needed for parametric analyses. Therefore, it is strongly advised that researchers should thoroughly evaluate the distribution of data and model assumptions. Furthermore, when necessary, consider the alternative statistical techniques.

In summary, log-transformation is a valid and useful tool to fix the common problems in data or relevant issues of uneven distributions, extreme values or outliers, heteroscedasticity, and nonlinear relationships. Furthermore, it can improve how well the statistical models work and make the results easier to understand. By changing the data into a form that fits close to the assumptions of statistical models, log-transformation may help researchers to make more accurate and meaningful conclusions. Therefore, it is always recommended to verify the assumptions following data transformation. To make the results understandable, practical and significant, it is also advised that researchers should try to bring the findings back to the original scale (before the log change) so that the results can be meaningful and useful in real life.

References:

1.Feng C, Wang H, Lu N, et al. Log-transformation and its implications for data analysis. *Shanghai Arch Psychiatry.* Apr 2014;26(2):105-109.

2.Manning WG. The logged dependent variable, heteroscedasticity, and the retransformation problem. *J Health Econ.* Jun 1998;17(3):283-295.

3.West RM. Best practice in statistics: The use of log transformation. *Ann Clin Biochem.* May 2022;59(3):162-165.

4.Knief U, Forstmeier W. Violating the normality assumption may be the lesser of two evils. *Behav Res Methods.* Dec 2021;53(6):2576-2590.

5.Ernst AF, Albers CJ. Regression assumptions in clinical psychology research practice-a systematic review of common misconceptions. *PeerJ.* 2017;5:e3323.

6.Zhou Y, Zhu Y, Wong WK. Statistical tests for homogeneity of variance for clinical trials and recommendations. *Contemp Clin Trials Commun.* Jun 2023;33:101119.

7.Lee DK. Data transformation: a focus on the interpretation. *Korean J Anesthesiol.* Dec 2020;73(6):503-508.

8.Chesnaye NC, van Diepen M, Dekker F, Zoccali C, Jager KJ, Stel VS. Non-linear relationships in clinical research. *Nephrol Dial Transplant.* Feb 4 2025;40(2):244-254.

9.Benoit K. Linear Regression Models with Logarithmic Transformations. *London School of Economics.* 2011;22(1):23-36.

10. Sullivan J, Warkentin M, Wallace L. So many ways for assessing outliers: What really works and does it matter? *J Business Research.* 2021;132:530-543.

11. Barrera-Gomez J, Basagana X. Models with transformed variables: interpretation and software. *Epidemiology.* Mar 2015;26(2):e16-17.

This editorial may be cited as: Soomro MH, Memon S, Bista S. The Role of Log-Transformation in Data Science and Statistical Analysis. Ann Allied Health Sci. 2025;12(1):1-2.

[1]Department of Paediatrics, Cumming School of Medicine, University of Calgary, Calgary, Canada
[2]Owerko Centre, Alberta Children's Hospital Research Institute, University of Calgary, Calgary, Canada
[3]Department of Cancer Epidemiology and Prevention Research, Alberta Health Services, Calgary, Alberta, Canada
[4]Department of Oncology, University of Calgary, Calgary, Alberta, Canada