Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

Application of Negative Binomial Regression in Analyzing Health Data of Air Pollution: A Comparative Study

Deepali N.Bramhpurkar
Statistics
Shri Jagdisprasad Jhambarmal Tibrewala University, Jhunjhunu, Rajasthan.

Dr.Swati Desai
Statistics
Shri Jagdisprasad Jhambarmal Tibrewala University, Jhunjhunu, Rajasthan.

Dr.Sangita Patil
Statistics
Mit Arts Commerce & Science College Alandi, Pune

Abstract - This study investigates the application of Negative Binomial regression for analyzing hospital admission data in the presence of air pollution. Daily hospital visits were modeled as count responses and examined using Generalized Linear Models implemented in MATLAB. The performance of the Negative Binomial approach was compared with Ordinary Least Squares and transformation-based methods, including logarithmic and square-root transformations. Model evaluation was carried out using confidence interval length, coverage probability, and goodness-of-fit measures. The results reveal substantial overdispersion in the health data, for which traditional methods provide inefficient and unstable inference. In contrast, the Negative Binomial model demonstrates superior reliability by producing adaptive confidence intervals and balanced coverage, making it a robust framework for environmental health assessment.

Keywords - ***Negative Binomial Regression, GLM, Overdispersion, Hospital Admissions, Air Pollution, Confidence Interval***

INTRODUCTION

Hospital admission data are commonly recorded as count outcomes and are widely used in environmental epidemiology to assess the impact of air pollution on public health. In the present study, daily hospital visits were obtained by aggregating individual patient admission records over 727 consecutive days. The corresponding environmental data included daily measurements of Air Quality Index (AQI), PM2.5, and PM10 concentrations.

Preliminary exploratory analysis of the dataset revealed a mean daily hospital visit count of 21.32 and a variance of 58.43, yielding a dispersion ratio of 2.74. Since the variance substantially exceeds the mean, the data exhibit clear overdispersion. This violates the fundamental assumption of constant variance underlying traditional Ordinary Least Squares (OLS) regression. When applied to overdispersed count data, OLS may produce inefficient estimates, misleading standard errors, and unreliable confidence intervals.

To address non-normality and heteroscedasticity, transformation-based methods such as logarithmic and square-root transformations are often employed. Although these techniques attempt to stabilize variance, they introduce interpretational challenges and may result in biased back-transformed estimates, particularly in the presence of zero counts and high variability.

Generalized Linear Models (GLMs) provide a principled framework for modeling non-normal response variables. In particular, the Negative Binomial regression model extends the Poisson model by incorporating a dispersion parameter, allowing the variance to exceed the mean. This flexibility makes it especially suitable for environmental health data characterized by heterogeneity and unobserved risk factors.

Previous methodological studies have demonstrated, through simulation experiments, that the Negative Binomial GLM outperforms OLS and transformation-based methods in factorial experimental settings involving overdispersed responses. However, validation of these findings using real-world environmental health datasets remains limited.

The present study extends prior simulation-based evidence by applying OLS, logarithmic transformation, square-root transformation, and Negative Binomial GLM approaches to real hospital admission and air pollution data. All analyses were implemented in MATLAB, enabling direct comparison of confidence interval performance, expected interval length, and coverage probability through Monte Carlo simulation. The primary objective is to identify the most reliable and efficient inferential framework for modeling pollution-related hospital admissions under overdispersion.

OBJECTIVES

1.To examine the presence of overdispersion in daily hospital admission data associated with air pollution exposure.

2.To apply and compare different regression approaches, namely Ordinary Least Squares (OLS), logarithmic transformation, square-root transformation, and Negative Binomial Generalized Linear Models, for modeling pollution-related hospital visits.

3.To evaluate the efficiency of confidence intervals obtained from each method by computing the Expected Length of Confidence Intervals (ELOCI).

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

4.To assess the reliability of statistical inference by estimating coverage probabilities through Monte Carlo simulation.

5.To investigate the association between air pollution indicators (AQI, PM2.5, and PM10) and hospital admissions using real-world health data.

6.To validate previous simulation-based findings by extending them to an applied environmental health context.

7.To identify the most appropriate modeling framework for overdispersed health count data based on empirical performance measures.

METHODOLOGY

3.1 Data Sources and Preparation

The study utilizes two real-world datasets: hospital admission records and air pollution measurements. Individual patient admission data were obtained from hospital records and aggregated to generate daily counts of hospital visits based on the date of admission. Air pollution data included daily observations of Air Quality Index (AQI), PM2.5, and PM10 concentrations.

Both datasets were merged using the date variable to construct a unified time-series dataset. Observations containing missing or inconsistent values were excluded from the analysis. All statistical analyses were implemented using MATLAB.

3.2 Exploratory Data Analysis

Descriptive statistics were computed to examine the distributional properties of daily hospital visits. The presence of overdispersion was assessed using the variance-to-mean ratio. A ratio substantially greater than unity was interpreted as evidence of overdispersion, justifying the application of Negative Binomial regression.

3.3 Statistical Models

Let Yt denote the number of hospital visits on day t

Four modeling approaches were considered.

3.3.1 Ordinary Least Squares (OLS):

The OLS model is specified as:

$$Yt=Xt\beta+\varepsilon t$$

where Xt : represents pollution variables and

εt denotes random error.

3.3.2 Logarithmic Transformation(LOG):

To reduce heteroscedasticity, a logarithmic transformation was applied:

$$\log(Yt+1)=X_t\beta+\varepsilon_t$$

The constant term prevents undefined values for zero counts.

3.3.3 Square-Root Transformation(SQRT):

A square-root transformation was also considered:

$$\sqrt{Yt}= X_t\beta+\varepsilon_t$$

This transformation stabilizes variance for moderate count levels.

3.3.4 Negative Binomial GLM:

The Negative Binomial model was specified within the Generalized Linear Model framework:

$$Yt\sim NB(\mu t,k)$$

$$\log(\mu t) = Xt\beta +\varepsilon t$$

where k - is the dispersion parameter.

Parameter estimation was carried out using the Iteratively Reweighted Least Squares algorithm.

3.4 Confidence Interval Construction:

For each method, 95% confidence intervals for the mean response were constructed.

OLS intervals were obtained using standard normal-theory methods.

Log and square-root intervals were computed on the transformed scale and back-transformed.

Negative Binomial intervals were constructed using Wald-type procedures on the log scale.

3.5 Performance Evaluation:

Model performance was evaluated using:

Expected Length of Confidence Intervals (ELOCI)

Coverage Probability

Monte Carlo simulation with 1000 replications was conducted to estimate coverage probabilities.

3.6 Simulation Procedure:

Monte Carlo samples were generated from the fitted Negative Binomial model using estimated dispersion parameters. For each simulated dataset, all four models were refitted and confidence intervals were recalculated. Coverage probability was computed as the proportion of intervals containing the true mean response.

RESULT AND DISCUSSION

Analysis using MATLAB code of real data set:

1. Mean = 21.32
2. Variance = 58.43
3. Dispersion = 2.74
4. Estimated dispersion k = 71.453

Overdispersion:

The variance of hospital admissions exceeded the mean, confirming overdispersion. The presence of overdispersion justifies the application of the Negative Binomial model, which explicitly accounts for extra-Poisson variation.

PM2.5 and AQI showed significant positive association with admissions.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

GLM-NB provided realistic standard errors.

OLS underestimated variance.

Individual Confidence Interval:

OLS: [19.44, 21.49], [19.81, 21.82], …

LOG: [-1, ∞]

SQRT: [10^{12}, 10^{13}]

NBGLM: [19.88, 21.10], [20.21, 21.41], …

The OLS confidence intervals are symmetric and reasonable but relatively wide. The log-transformed intervals contain negative lower bounds and infinite upper bounds, which are meaningless for count data. The square-root method produces unrealistically large values due to numerical explosion after back-transformation. The Negative Binomial intervals remain within realistic ranges and preserve the count nature of the response variable. This highlights the practical superiority of the GLM approach.

Coverage Probability Analysis (Corrected Simulation):

Method	Coverage
OLS	0.949
LOG	0.743
SQRT	0.891
NBGLM	0.916

The OLS method achieved coverage close to the nominal 95% level. However, this high coverage is mainly due to excessively wide confidence intervals, indicating conservative inference. The logarithmic transformation performed poorly, covering the true mean in only 74% of cases, demonstrating severe undercoverage. The square-root transformation showed moderate performance but still failed to reach the nominal level. The Negative Binomial GLM achieved coverage of approximately 92%, which is reasonably close to the target level, reflecting reliable inferential properties. These results demonstrate that NBGLM offers a better balance between precision and reliability.

Expected Length of Confidence Intervals (ELOCI):

Method	ELOCI
OLS	3.433
LOG	∞
SQRT	8.88×10^{7}
NBGLM	2.087

The OLS method produced wide confidence intervals, indicating low efficiency. The log-transformation method resulted in infinite interval lengths, reflecting numerical instability during back-transformation. Similarly, the square-root transformation produced extremely large interval lengths, indicating serious computational and statistical limitations. In contrast, the Negative Binomial model generated moderate and stable interval lengths. This confirms that NBGLM yields more efficient and interpretable confidence intervals compared to transformation-based approaches.

Comparative Performance of Methods:

Criterion	OLS	LOG	SQRT	NBGLM
Handles Overdispersion	✗	Partially acceptable	Comparative Performance of Methods	appropriate
Coverage Accuracy	High (Overwide)	Poor	Moderate	Good
CI Stability	Moderate	Poor	Poor	Good
Interpretability	Moderate	Low	Low	High
Overall Performance	Comparative Performance of Methods	✗	Comparative Performance of Methods	appropriate

CONCLUSION:

This study investigated the suitability of different regression approaches for modeling pollution-related hospital admission data characterized by overdispersion. Using real-world health and environmental datasets, Ordinary Least Squares, logarithmic transformation, square-root transformation, and Negative Binomial Generalized Linear Models were systematically compared through empirical analysis and Monte Carlo simulation.

Preliminary data exploration revealed substantial overdispersion, with the variance exceeding the mean by more than twofold. This violated the fundamental assumptions of conventional linear regression and justified the use of distribution-based models. Although OLS achieved near-nominal coverage probability, it produced excessively wide confidence intervals, leading to inefficient and conservative inference. Transformation-based methods exhibited numerical instability, infinite or inflated confidence intervals, and severe undercoverage, thereby limiting their practical usefulness.

In contrast, the Negative Binomial GLM consistently demonstrated superior inferential performance. It generated stable and interpretable confidence intervals, achieved reasonably high coverage probabilities, and maintained moderate interval lengths by explicitly accounting for extra-Poisson variation. The balanced performance of the Negative

Binomial model highlights its theoretical and practical advantages for analyzing overdispersed health count data.

LIMITATIONS

- OLS-Does not explicitly model overdispersion.
- LOG-Cannot handle zero values directly without adding an arbitrary constant.
- SQRT-Back-transformation introduces bias in estimated mean responses. Only partially stabilizes variance for highly dispersed data
- NBGLM-Assumes a specific mean–variance relationship. More computationally intensive than OLS.

CITATION:

Analysis of 2^n Factorial Experiments with Negative Binomial Response Variables: A Comparative Study. international Conference on "Bridging Disciplines, Shaping the Future: Integrative Approaches to Global Challenges"; SJJTU/CONF/CSE/PAR/011/2025/257.

REFERENCES

[1] Agresti, A. (2015). *Foundations of Linear and Generalized Linear Models*. Wiley.
[2] Burnham, K. P., & Anderson, D. R. (2002). *Model Selection and Multimodel Inference* (2nd ed.). Springer.
[3] Cameron, A. C., & Trivedi, P. K. (2013). *Regression Analysis of Count Data* (2nd ed.). Cambridge University Press.
[4] Hilbe, J. M. (2011). *Negative binomial regression* (2nd ed.). CambridgeUniversityPress. https://doi.org/10.1017/CBO9780511973420
[5] Cameron, A. C., & Trivedi, P. K. (1998). *Regression analysis of count data*. Cambridge University Press.
[6] .Dobson, A. J., & Barnett, A. G. (2018). *An Introduction to Generalized Linear Models* (4th ed.). CRC Press.
[7] Dominici, F., Peng, R. D., Bell, M. L., Pham, L., McDermott, A., Zeger, S. L., & Samet, J. M. (2006). Fine particulate air pollution and hospital admissions. *Journal of the American Medical Association, 295*(10), 1127–1134. https://doi.org/10.1001/jama.295.10.1127
[8] Gentle, J. E. (2003). *Random Number Generation and Monte Carlo Methods*. Springer.
[9] Hilbe, J. M. (2011). *Negative Binomial Regression* (2nd ed.). Springer.
[10] Lawless, J. F. (1987). Negative binomial and mixed Poisson regression. *Canadian Journal of Statistics, 15*(3), 209–225. https://doi.org/10.2307/3314912
[11] McCullagh, P., & Nelder, J. A. (1989). *Generalized Linear Models* (2nd ed.). Chapman & Hall.
[12] Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to Linear Regression Analysis* (5th ed.). Wiley.
[13] Myers, R. H., Montgomery, D. C., & Anderson-Cook, C. M. (2016). *Response Surface Methodology* (4th ed.). Wiley.
[14] Peng, R. D., & Dominici, F. (2008). Statistical methods for environmental epidemiology. *Journal of the Royal Statistical Society: Series A, 171*(1), 1–22. https://doi.org/10.1111/j.1467-985X.2007.00545.x
[15] Pope, C. A., & Dockery, D. W. (2006). Health effects of fine particulate air pollution. *Journal of the Air & Waste Management Association, 56*(6), 709–742. https://doi.org/10.1080/10473289.2006.10464485
[16] Ripley, B. D. (2009). *Stochastic Simulation*. Wiley.
Robert, C. P., & Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer.
[17] Samoli, E., et al. (2013). Acute effects of air pollution on mortality. *Environmental Health Perspectives, 121*(1), 14–23. https://doi.org/10.1289/ehp.1104491
[18] 17.Schwartz, J. (2004). The effects of particulate air pollution on daily deaths. *Environmental Research, 94*(1), 7–13. https://doi.org/10.1016/S0013-9351(03)00018-6
[19] 18.Wald, A. (1943). Tests of statistical hypotheses concerning several parameters. *Transactions of the American Mathematical Society, 54*(3), 426–482. https://doi.org/10.1090/S0002-9947-1943-0012401-3