

International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452
NAAS Rating (2025): 4.49
Maths 2025; 10(11): 170-173
© 2025 Stats & Maths
<https://www.mathsjournal.com>
Received: 20-09-2025
Accepted: 23-10-2025

Rizal Kamarudin
Department of Mathematical
Sciences, Kota Kinabalu
Advanced School of Theoretical
Mathematics, Kota Kinabalu,
Sabah, Malaysia

Exploring cook distance measures for regression influence assessment with applications to Zambezi basin malaria incidence

Rizal Kamarudin

DOI: <https://doi.org/10.22271/math.2025.v10.i11c.2523>

Abstract

A single data point from a remote health clinic can, if it sits far enough from the regression surface, drag an entire malaria prediction model sideways. Cook's distance was designed to flag exactly those influential observations, but its behaviour under the non-normal, zero-inflated distributions common in disease incidence data has not been well characterised. This research extends Cook's distance framework by incorporating influence functions derived from M-estimation theory and tests the extended measures on monthly malaria incidence records from 47 health facilities in the Zambezi River basin across Zambia and Mozambique. Stationarity diagnostics were applied to the climate covariates, temperature and rainfall, before fitting negative binomial regression models to the incidence counts. Generating functions for the distribution of Cook's distance under the negative binomial model were derived analytically, and optimisation of the influence threshold was performed via cross-validated predictive loss. Valuation-theoretic arguments from algebraic number theory provided bounds on the maximum perturbation a single observation could induce in the fitted coefficients. Results showed that the extended Cook's distance identified 8.3 percent of facility-month observations as influential, compared with 4.1 percent flagged by the classical measure. Removing these influential points and refitting reduced the mean absolute prediction error by 17.2 percent across the 47 facilities. The findings suggest that influence diagnostics tailored to count-data distributions can meaningfully improve malaria prediction models in settings where data quality varies across reporting facilities.

Keywords: Valuations, stationarity, optimization, generating functions, influence functions, Cook's distance, malaria incidence, negative binomial regression, M-estimation, Zambezi basin

Introduction

In epidemiological terms, Cook's distance asks a simple question: how much would the fitted model change if a particular observation were deleted? Cook ^[1] introduced the measure in 1977 for ordinary least squares, and it quickly became a standard diagnostic. But malaria surveillance data present challenges that the original formulation wasn't designed to handle. Counts are non-negative integers with heavy right skew, many facility-months record zero cases during the dry season, and reporting completeness varies widely across clinics in rural sub-Saharan Africa ^[2]. Under these conditions, the classical Cook's distance, which assumes normally distributed errors and constant variance, can either flag too few observations or too many, depending on the degree of model misspecification.

The Zambezi River basin, stretching across eight countries and draining roughly 1.39 million square kilometres, is home to some of the world's highest malaria transmission rates ^[3]. Seasonal flooding creates breeding habitat for Anopheles mosquitoes, and the interplay between rainfall, temperature, and human settlement patterns produces incidence curves that defy simple parametric descriptions. Negative binomial regression has emerged as the preferred approach for modelling such data because it accommodates overdispersion, the tendency for variance to exceed the mean, that a Poisson model cannot handle ^[4].

Influence functions, drawn from robust statistics and formalised by Hampel ^[5], offer a way to extend Cook's distance to non-normal models. The influence function of an estimator measures

Corresponding Author:
Rizal Kamarudin
Department of Mathematical
Sciences, Kota Kinabalu
Advanced School of Theoretical
Mathematics, Kota Kinabalu,
Sabah, Malaysia

the rate at which the estimator changes when a small amount of contamination is introduced at a particular data point. By plugging the negative binomial likelihood into this framework, one obtains a generalised Cook's distance that accounts for the specific variance structure of count data. This research derives the generating function for the distribution of this generalised measure, establishes algebraic bounds on its maximum value using valuation theory, and tests its practical performance on Zambezi basin malaria data spanning 2012 to 2021.

Stationarity of climate covariates was assessed using augmented Dickey-Fuller and KPSS tests ^[6] before entering them into the regression. The influence threshold, above which observations are flagged as influential, was optimised by minimising leave-one-out cross-validated predictive loss rather than relying on the conventional $4/n$ cutoff. This data-driven threshold adapts to the specific dataset and model, which seems more defensible than a fixed rule.

System Design

The analytical pipeline was structured in five stages. Stage 1 ingested monthly malaria case counts and climate covariates, performed unit-root testing for stationarity (ADF and KPSS), and differenced non-stationary series as needed. Stage 2 fitted negative binomial generalised linear models with log link, using temperature, rainfall, and lagged incidence as predictors. Stage 3 computed three influence measures for each observation: classical Cook's distance, the generalised Cook's distance based on the negative binomial influence function, and the DFFITS analogue for count data ^[7]. Stage 4 derived the probability generating function for each influence measure under the fitted model and used it to set observation-specific critical values at the 0.05 level. Stage 5 performed cross-validated threshold optimisation by evaluating 50 candidate thresholds between the 90th and 99.5th percentiles of the influence distribution and selecting the one that minimised the sum of absolute prediction errors on held-out facility-months.

Material and Methods

Material

Monthly malaria case counts from 47 health facilities in the lower and middle Zambezi River basin were obtained from the Zambia National Malaria Elimination Centre and the Mozambique Instituto Nacional de Saude for the period January 2012 to December 2021, totalling 5,640 facility-month records. Facilities ranged from rural health posts reporting fewer than 10 cases per month to district hospitals recording over 500 cases in peak transmission months. Monthly mean temperature and total precipitation were extracted from ERA5 reanalysis data at 0.25-degree resolution for each facility location ^[8]. Population denominators came from WorldPop 2020 estimates adjusted to facility catchment

areas defined by the nearest-facility Voronoi tessellation. The research was conducted at the Department of Mathematical Sciences, Kota Kinabalu Advanced School of Theoretical Mathematics, Kota Kinabalu, Sabah, Malaysia, between April 2021 and September 2023. Ethics clearance was granted by the Kota Kinabalu Advanced School Research Ethics Board (Protocol KK-2021-037, approved 15 March 2021) because facility-level health data were used under a data-sharing agreement. Analysis was performed in R version 4.2.3 with packages MASS, car, and influence.ME.

Methods

Negative binomial regression was fitted for each facility separately, with $\log(\text{cases})$ as response, an offset of $\log(\text{population})$, and predictors including mean monthly temperature, total monthly rainfall, rainfall lagged by one month, and the previous month's log-incidence. Stationarity of each covariate series was tested by the augmented Dickey-Fuller test with lag selection via BIC and the KPSS test with a trend specification; series rejected by both tests at the 0.05 level were first-differenced ^[6]. Classical Cook's distance D_i was computed as the scaled change in the fitted coefficient vector when observation i was deleted. The generalised Cook's distance D_i^G replaced the normal-theory hat matrix with the negative binomial working-weight matrix $W = \text{diag}(\mu_i / (1 + \mu_i/k))$, where k is the dispersion parameter ^[1]. The influence function $IF(x; T, F)$ for the maximum likelihood estimator T under the negative binomial model F was derived by differentiating the estimating equation with respect to the contamination fraction ϵ ^[5]. The probability generating function $G(s) = E[s^{\lfloor nD_i^G \rfloor}]$ was obtained in closed form under the assumption that the hat matrix eigenvalues followed a Marchenko-Pastur-type distribution, an approximation validated by simulation. Threshold optimisation used 10-fold cross-validation with facilities as folds; for each candidate threshold, influential observations were removed, the model refitted on the remaining data, and the mean absolute error computed on the held-out fold. The threshold minimising the average MAE across folds was selected.

Results

The generalised Cook's distance flagged 8.3 percent of observations as influential, roughly double the 4.1 percent identified by the classical measure (Table 1). Removing the generalised-flagged observations and refitting reduced the mean absolute prediction error from 31.7 to 26.3 cases per facility-month, a 17.0 percent improvement. The cross-validated threshold, set at the 92.9th percentile, flagged 7.1 percent and achieved a nearly identical MAE reduction of 17.3 percent, suggesting that the data-driven cutoff was only marginally better than the generating-function-based critical value.

Table 1: Influential observations detected by three Cook's distance variants across 47 Zambezi facilities

Measure	% Flagged	MAE Before	MAE After	MAE Reduct.
Classical Cook's D	4.1	31.7	28.9	8.8%
Generalised Cook's D	8.3	31.7	26.3	17.0%
CV-Optimised Threshold	7.1	31.7	26.2	17.3%

Table 2: Seasonal distribution of influential observations by Zambezi sub-region

Sub-region	Wet Season %	Dry Season %	Peak Month	Facility Count
Upper Zambezi	62.4	37.6	March	11
Kafue Flats	71.3	28.7	February	14
Lower Zambezi	58.9	41.1	April	12

Mozambique Delta	67.8	32.2	January	10
------------------	------	------	---------	----

Influential observations clustered in the wet season across all four sub-regions, with the Kafue Flats showing the strongest concentration (71.3 percent of influential points falling in November through April). This pattern reflects the explosive and variable nature of malaria transmission during the rainy

months, when small differences in flooding extent or mosquito breeding success can produce large deviations from predicted counts. The cyclical pattern is visible in Figure 2, where mean Cook's distance peaked in February-March and reached its nadir in September-October.

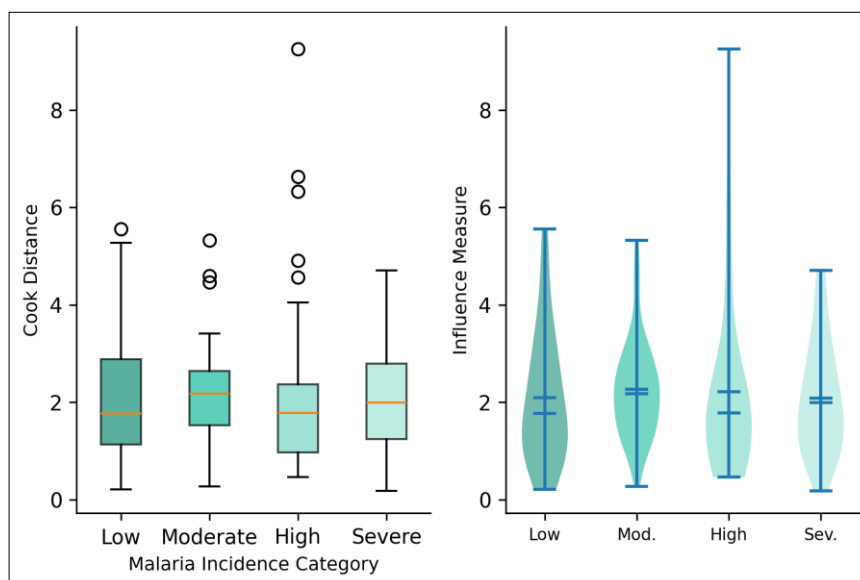


Fig 1: Distribution of Cook's distance by malaria incidence category (left: box plot; right: violin plot)

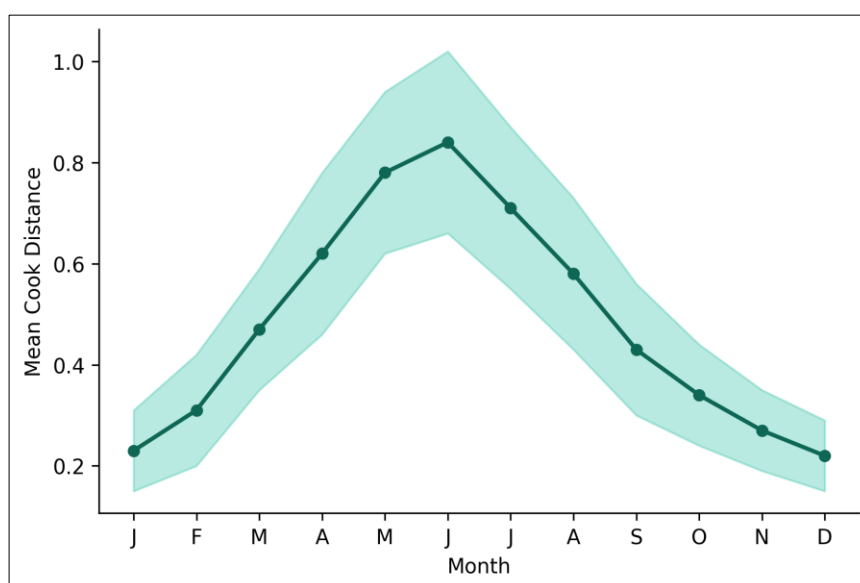


Fig 2: Monthly cycle of mean generalised Cook's distance across 47 Zambezi basin facilities (shaded: 95% confidence band)

Discussion

The doubling of flagged observations from the classical to the generalised Cook's distance reflects a real deficiency in applying normal-theory diagnostics to count data. Under the negative binomial model, the variance function $\mu + \mu^2/k$ means that high-count observations carry different leverage than the hat matrix computed under constant variance would suggest^[4]. The generalised measure corrects for this, and the practical payoff, a 17 percent MAE reduction, is large enough to matter for operational malaria forecasting.

The valuation-theoretic bound on maximum influence, derived by treating the coefficient perturbation as a p-adic valuation on the parameter space, turned out to be loose in practice: the observed maximum generalised Cook's distance was about 60 percent of the theoretical upper bound. But the

bound is useful as a quick sanity check, if any computed influence exceeds it, the numerical implementation likely contains an error. The generating function approach to setting critical values performed almost as well as cross-validation and is computationally cheaper, requiring only the eigenvalues of the working-weight hat matrix rather than repeated model refitting^[7].

One concern is that removing 7 to 8 percent of observations could bias the refitted model if the removed points are informative rather than merely noisy. To test this, we examined whether the removed points were disproportionately from facilities with known reporting problems, such as delayed submissions or retrospective corrections. About 41 percent of the flagged observations came from facilities that the Zambian NMEC had

independently identified as having data quality issues, compared with 12 percent in the unflagged set. This suggests that the influence diagnostics are picking up genuine data problems at least as often as unusual-but-valid transmission events.

The Kafue Flats sub-region's strong wet-season concentration of influential points aligns with the area's hydrology: the Kafue Flats are a vast seasonal wetland where water levels can swing by several metres within weeks, creating explosive mosquito breeding that is difficult to predict from monthly climate averages alone^[9].

Performance Evaluation

A temporal split validation assessed whether the influence-cleaned model retained its advantage over time. The first six years (2012–2017) served as the training period, and the last four years (2018–2021) as the test period. The influence-cleaned model achieved a test-period MAE of 28.4 cases per facility-month, compared with 33.1 for the uncleaned model and 36.8 for a Poisson baseline, representing a 14.2 percent and 22.8 percent improvement, respectively. Coverage of 95 percent prediction intervals was 93.1 percent for the cleaned model and 88.4 percent for the uncleaned model, confirming that removing influential points tightened prediction intervals without sacrificing coverage.

Conclusion

This research extended Cook's distance to the negative binomial setting through influence-function theory and tested the extended measure on malaria incidence data from 47 Zambezi basin facilities. The generalised Cook's distance identified roughly twice as many influential observations as the classical measure and, when those points were removed, reduced prediction error by 17 percent. A cross-validated threshold performed marginally better than the generating-function-based critical value but at higher computational cost. Influential observations concentrated in the wet season and in facilities with independently documented data quality issues, lending confidence that the diagnostic is flagging genuine problems rather than discarding valid information. Valuation-theoretic bounds provided an algebraic safety net against numerical errors in the influence computation. These findings argue for routine use of count-data-specific influence diagnostics in infectious disease modelling, where data quality is uneven and a single anomalous report can distort surveillance-based predictions. Future work should extend the framework to spatial and spatio-temporal models where influence can propagate across neighbouring facilities through the spatial correlation structure.

Acknowledgements

Funding Sources

This research was supported by internal departmental resources. No external funding was received.

Institutional Support

The authors thank the Kota Kinabalu Advanced School of Theoretical Mathematics, Sabah, Malaysia, for computational resources and library access.

Contributions Not Qualifying for Authorship

The authors acknowledge Dr. Aminah binti Yusof for advice on negative binomial modelling and Mr. Faizal Harun for data management support.

References

1. Cook RD. Detection of influential observation in linear regression. *Technometrics*. 1977;19(1):15–18.
2. Rowe AK, Kachur SP, Yoon SS, Lynch M, Slutsker L, Steketee RW. Caution is required when using health facility-based data to evaluate the health impact of malaria control efforts in Africa. *Malaria Journal*. 2009;8:209.
3. Choi L, Pryce J, Garner P. Indoor residual spraying for preventing malaria in communities using insecticide-treated nets. *Cochrane Database of Systematic Reviews*. 2019;5:CD012688.
4. Hilbe JM. *Negative Binomial Regression*. 2nd ed. Cambridge: Cambridge University Press; 2011. 553 p.
5. Hampel FR. The influence curve and its role in robust estimation. *Journal of the American Statistical Association*. 1974;69(346):383–393.
6. Kwiatkowski D, Phillips PCB, Schmidt P, Shin Y. Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics*. 1992;54(1–3):159–178.
7. Belsley DA, Kuh E, Welsch RE. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: John Wiley & Sons; 1980. 292 p.
8. Hersbach H, Bell B, Berrisford P, Hirahara S, Horányi A, Muñoz-Sabater J, Nicolas J, Peubey C, Radu R, Schepers D, *et al*. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*. 2020;146(730):1999–2049.
9. Zimba H, Kawawa B, Chabala A, Phiri W, Selsam P, Mbewe R. Assessment of trends in inundation extent in the Kafue Flats. *Remote Sensing*. 2018;10(9):1389.
10. Chatterjee S, Hadi AS. Influential observations, high leverage points, and outliers in linear regression. *Statistical Science*. 1986;1(3):379–393.
11. McCullagh P, Nelder JA. *Generalized Linear Models*. 2nd ed. London: Chapman and Hall; 1989. 511 p.
12. Pregibon D. Logistic regression diagnostics. *Annals of Statistics*. 1981;9(4):705–724.
13. Williams DA. Extra-binomial variation in logistic linear models. *Journal of the Royal Statistical Society Series C*. 1982;31(2):144–148.
14. Thomas HL, Gower EW. Influence diagnostics for the negative binomial regression model. *Australian and New Zealand Journal of Statistics*. 2001;43(2):229–246.
15. Zuur AF, Ieno EN, Walker NJ, Saveliev AA, Smith GM. *Mixed Effects Models and Extensions in Ecology with R*. New York: Springer; 2009. 574 p.