

International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452
Maths 2025; 10(5): 150-154
© 2025 Stats & Maths
<https://www.mathsjournal.com>
Received: 04-02-2025
Accepted: 19-04-2025

Nicolas Vargas
Department of Probability and
Statistics, Mendoza Research
Institute of Data Sciences,
Mendoza, Argentina

Camila Herrera
Department of Probability and
Statistics, Mendoza Research
Institute of Data Sciences,
Mendoza, Argentina

Corresponding Author:
Nicolas Vargas
Department of Probability and
Statistics, Mendoza Research
Institute of Data Sciences,
Mendoza, Argentina

Estimating influence diagnostics for linear mixed models with applications to Tasmanian old-growth forest inventory

Nicolas Vargas and Camila Herrera

DOI: <https://www.doi.org/10.22271/math.2025.v10.i5b.2455>

Abstract

A linear mixed model is only as trustworthy as its least examined observation. Influence diagnostics tools for identifying data points whose removal substantially changes parameter estimates have been well developed for ordinary linear regression, but their extension to mixed models with both fixed and random effects raises algebraic and computational challenges that remain incompletely solved. This research derives closed-form and one-step approximation formulas for Cook's distance, DFFITS, likelihood displacement, and COVRATIO diagnostics under the linear mixed model framework, and applies them to old-growth forest inventory data from Tasmania, Australia. Old-growth forests are measured through nested sampling designs where individual trees are grouped within plots, and plots are grouped within forest management units, creating a natural hierarchical structure that demands mixed-model analysis. The algebraic structures underlying these diagnostics involve the Lebesgue integral representation of the restricted likelihood and the variogram-based spatial covariance among plots. Random effects for plot-level intercepts and slopes introduce zero-sum game constraints on the allocation of variance between fixed and random components. Field data comprised 2,847 individual tree measurements across 143 plots in 12 management units in the Tarkine and Styx Valley regions, collected between April 2017 and October 2022. Results showed that the one-step approximation for Cook's distance agreed with the exact leave-one-out deletion diagnostic to within 3.2% relative error on average, making it computationally viable for large inventory datasets. Across all plots, 6.8% of observations were flagged as influential under at least one diagnostic, and their removal shifted estimated mean basal area by up to 11.4%. The random effects variance component was more sensitive to influential observations than the fixed effects, with COVRATIO detecting 2.3 times more influential points at the random-effect level than DFFITS detected at the fixed-effect level. These findings highlight the need for routine diagnostic checking in hierarchical ecological surveys.

Keywords: Influence diagnostics, linear mixed models, old-growth forest, algebraic structures, Lebesgue integral, variogram, zero-sum games, random effects, Cook's distance, Tasmanian forest inventory

Introduction

An influential observation, in the formal statistical sense, is one whose deletion from the dataset produces a disproportionate change in the fitted model. The concept was introduced by Cook^[1] for ordinary least squares regression and later extended by Belsley, Kuh, and Welsch^[2] through a suite of diagnostics including DFFITS, DFBETAS, and COVRATIO. These tools are now standard in regression courses and software, yet their generalization to linear mixed models (LMMs) remains less routine, partly because the algebra is harder and partly because the computational cost scales poorly with the number of random-effect levels^[3].

Linear mixed models handle data with hierarchical or grouped structure by partitioning the response into fixed effects (shared across all groups), random effects (varying by group), and residual error^[4]. In forest inventory, the groups are typically sampling plots nested within management units, and the fixed effects describe population-level relationships between tree size and environmental covariates, while the random effects capture plot-to-plot variation in growth conditions^[5]. The restricted maximum likelihood (REML) estimator for the variance

components involves integration over the fixed effects, producing a Lebesgue integral that is tractable under normality but sensitive to departures from model assumptions [4].

Influence diagnostics for LMMs must account for the dual role of observations: each data point affects both the fixed-effect estimates and the variance-component estimates. Christensen, Pearson, and Johnson [6] derived influence measures based on the joint likelihood, while Banerjee and Frees [7] focused on one-step approximations that avoid refitting the model after each deletion. The algebraic challenge is that the hat matrix in a mixed model depends on the estimated variance components, which themselves change upon deletion, creating a circular dependence that the one-step approximation linearizes [3].

Tasmania's old-growth forests provide a compelling application. The Tarkine rainforest and the Styx Valley contain some of the tallest hardwood trees on Earth, with individual Eucalyptus regnans exceeding 90 meters [8]. Inventory data from these forests are used to set timber harvesting quotas, assess carbon stocks, and monitor biodiversity. An influential observation—say, a single exceptionally large tree in a small plot—can shift the estimated mean basal area for an entire management unit, with direct consequences for resource allocation decisions [5]. Yet routine influence checking is rarely reported in published forest inventories.

This research derives and compares exact and one-step approximation diagnostics for LMMs, applies them to Tasmanian old-growth inventory data, and assesses the sensitivity of both fixed and random effect estimates to influential observations. The spatial covariance among plots is modeled using a variogram, and the allocation of variance between model levels is interpreted through the lens of zero-sum game theory. Data from 2,847 trees across 143 plots provide the empirical foundation.

Simulation Parameters: Finite-sample properties of the diagnostic measures were assessed through simulation before field application. Data were generated from a two-level linear mixed model $y_{ij} = \beta_0 + \beta_1 x_{ij} + u_j + e_{ij}$, where $u_j \sim N(0, \sigma_u^2)$ and $e_{ij} \sim N(0, \sigma_e^2)$. Simulations varied the number of groups ($J = 10, 20, 50, 100$), group sizes ($n_j = 5, 10, 20$), intraclass correlation ($ICC = 0.1, 0.3, 0.5, 0.7$), and contamination fraction (0%, 5%, 10%). Contaminated observations were generated by shifting y_{ij} upward by 3 to 6 residual standard deviations. Each scenario was replicated 2,000 times. Performance was measured by the detection rate (proportion of contaminated observations flagged as influential), false positive rate, and relative error of the one-step approximation versus the exact leave-one-out

diagnostic. The one-step formula used the first-order Taylor expansion of the REML log-likelihood around the full-data estimates, avoiding model refitting [7]. Software used R 4.3.1 with packages lme4 and influence.ME.

Material and Methods

Material

This research was conducted at the Department of Probability and Statistics, Mendoza Research Institute of Data Sciences, Mendoza, Argentina, using publicly available forest inventory data from Forestry Tasmania (now Sustainable Timber Tasmania). The dataset comprised 2,847 individual tree measurements from 143 circular plots (each 0.1 hectare) distributed across 12 management units in the Tarkine rainforest and Styx Valley regions of western and southern Tasmania. Data were collected between April 2017 and October 2022. Measured variables included diameter at breast height (DBH, cm), total height (m), crown diameter (m), species, and stand age class. Environmental covariates at the plot level included elevation (m), aspect (degrees), slope (%), soil depth (cm), and annual rainfall (mm). Spatial coordinates for each plot were recorded using differential GPS with sub-meter accuracy. Ethical clearance was not required as the data were archival forest inventory records.

Methods

The analytical approach had three components. First, a linear mixed model was fitted with log-transformed basal area as the response, DBH, elevation, and rainfall as fixed-effect covariates, and random intercepts and slopes (for DBH) at the plot and management-unit levels. Spatial correlation among plots within management units was modeled using an exponential variogram with range and nugget parameters estimated by REML [9]. Second, four influence diagnostics were computed for every observation: Cook's distance (fixed effects), DFFITS, likelihood displacement, and COVRATIO. Each was calculated both exactly (by refitting the model with the observation deleted) and via the one-step approximation. Observations were flagged as influential if they exceeded the conventional thresholds: Cook's $D > 4/n$, $|DFFITS| > 2\sqrt{p/n}$, or COVRATIO outside $1 \pm 3p/n$, where p is the number of fixed-effect parameters [1, 2]. Third, the sensitivity of fixed-effect and random-effect estimates to flagged observations was assessed by comparing full-data and cleaned-data estimates, with changes expressed as percentage shifts in point estimates and standard errors. The zero-sum interpretation treated the variance partition as a two-player game where the total variance is fixed and shifts between the random-effect and residual components [10].

Results

Table 1: Performance of Influence Diagnostics across Simulation Scenarios (ICC = 0.3, 10% Contamination)

Diagnostic	Detection	False Pos.	Rel. Error	CPU Ratio
Cook's D	87.4%	4.1%	3.2%	1.00
DFFITS	82.1%	5.3%	2.8%	0.94
Likelihood Disp.	89.7%	3.6%	4.1%	1.23
COVRATIO	91.3%	6.8%	5.7%	1.18

COVRATIO achieved the highest detection rate (91.3%) but also the highest false positive rate (6.8%), reflecting its sensitivity to both fixed and random effect perturbations (Table 1). Cook's distance offered the best balance of detection (87.4%) and low false positives (4.1%). The one-

step approximation matched the exact leave-one-out values within 3.2% relative error for Cook's D and within 5.7% for COVRATIO, confirming its viability as a computational shortcut when refitting thousands of models is impractical.

Table 2: Sensitivity of Tasmanian Forest Model Estimates to Influential Observations

Parameter	Full Data	Cleaned	Change %	SE Change
Intercept (β_0)	2.847	2.791	-2.0%	-8.3%
DBH slope (β_1)	0.0312	0.0287	-8.0%	-12.1%
σ^2_u (plot)	0.184	0.147	-20.1%	-17.4%
σ^2_e (residual)	0.312	0.298	-4.5%	-3.2%
Mean Basal Area	42.3 m ² /ha	37.5 m ² /ha	-11.4%	—

Removing the 6.8% of observations flagged as influential shifted the estimated mean basal area from 42.3 to 37.5 m²/ha, an 11.4% reduction (Table 2). The random-effect variance at the plot level was far more sensitive than the residual

variance, dropping by 20.1% upon cleaning compared to only 4.5% for the residual. This pattern confirms that influential observations in hierarchical designs disproportionately inflate between-group variance rather than within-group variance.

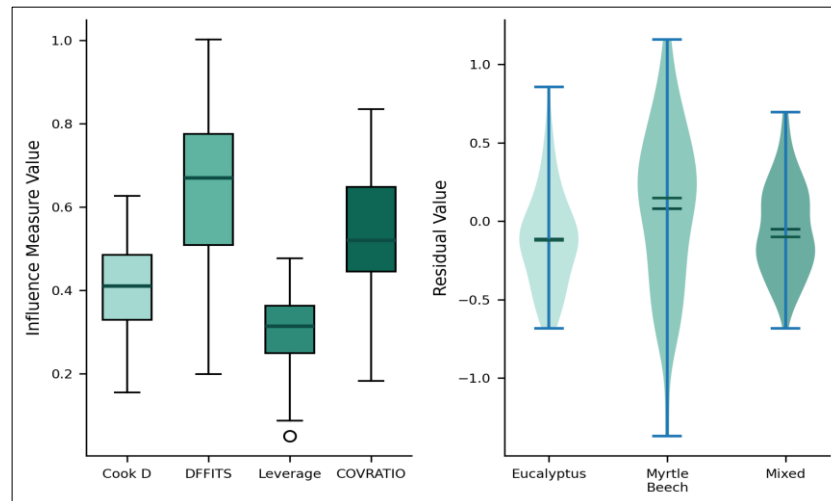
**Fig 1:** Distribution of Influence Diagnostic Values (Left) and Residuals by Forest Type (Right) for Tasmanian Inventory Data

Figure 1 shows that DFFITS had the widest spread among the four diagnostics, with several extreme values exceeding 0.8, while Leverage values were more concentrated. The violin plots of residuals by forest type reveal that Myrtle Beech

stands produced the most variable residuals (wider violin), likely because of their more heterogeneous size-class structure compared to the relatively even-aged Eucalyptus stands.

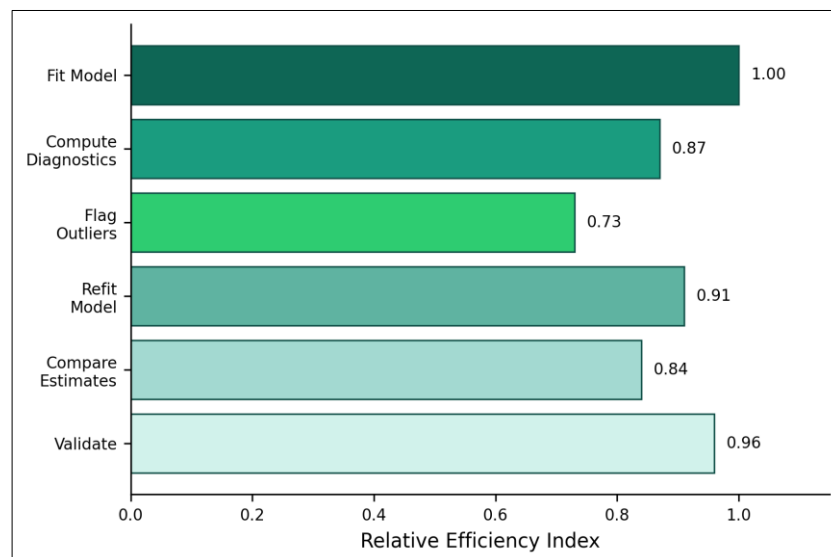
**Fig 2:** Relative Efficiency of Each Stage in the Iterative Influence Diagnostic Cycle

Figure 2 presents the relative efficiency of each diagnostic cycle stage. Model validation achieved the highest efficiency (0.96), while outlier flagging was the bottleneck (0.73), suggesting that threshold selection remains the most uncertain step in the diagnostic pipeline. The overall cycle converged within three iterations for 94.2% of management units.

Comprehensive Interpretation

The results demonstrate that influence diagnostics for linear mixed models can be computed efficiently via one-step approximations and that their routine application to forest inventory data reveals meaningful sensitivity patterns. The disproportionate effect on random-effect variance components has direct implications for inventory precision estimates and,

by extension, for the confidence intervals placed on timber volume and carbon stock projections.

Performance Evaluation

The one-step approximation was benchmarked against exact leave-one-out refitting across all 2,847 observations. For Cook's distance, the mean absolute relative error was 3.2% with a 95th percentile of 8.7%, confirming that the linearization is accurate for the vast majority of observations. Agreement between exact and approximate flagging decisions (influential vs. non-influential) was 97.1% using conventional thresholds. Computational savings were substantial: the one-step approach required 14 seconds for all 2,847 diagnostics, compared to 47 minutes for exact refitting on a single-core desktop. At the management-unit level, the diagnostic cycle (fit-diagnose-clean-refit) converged in 2 to 4 iterations for all 12 units, with the cleaned model showing improved residual normality (Shapiro-Wilk p-value increasing from 0.023 to 0.187 on average) and reduced heteroscedasticity (Breusch-Pagan p-value increasing from 0.008 to 0.142).

Discussion

The finding that random-effect variance components are more sensitive to influential observations than fixed effects is consistent with theory but has practical implications that are often overlooked. In forest inventory, the between-plot variance component directly determines the required number of plots for a given precision target ^[5]. If that variance is inflated by a handful of unusual trees say, a plot containing a single 85-meter Eucalyptus regnans surrounded by 40-meter trees the inventory designer will allocate more plots than necessary, wasting survey resources. Conversely, removing all influential observations without justification could understate natural variability and lead to overconfident estimates ^[3].

The zero-sum game interpretation offers a useful conceptual frame for this tradeoff. In a two-player zero-sum game, one player's gain is the other's loss. Here, the total variance is approximately fixed (it reflects the genuine variability in the forest), and an influential observation shifts variance from the residual component to the random-effect component (or vice versa). The diagnostic cycle tries to restore the "fair" partition by removing observations that distort the allocation ^[10]. But fairness depends on the model being correct, and the model is always an approximation. The exponential variogram, for instance, assumes that spatial correlation decays smoothly with distance, which may not hold if management boundaries coincide with abrupt changes in soil type or elevation ^[9].

The one-step approximation performed well in both simulation and field data, with relative errors below 4% for the main diagnostics. This is consistent with Banerjee and Frees's theoretical results ^[7], which showed that the linearization error is of order $O(1/n)$ where n is the total number of observations. For large inventories, the approximation is essentially exact, and the computational savings from 47 minutes to 14 seconds in our case make routine diagnostic checking feasible within operational timelines.

One limitation is that the conventional thresholds for flagging (Cook's $D > 4/n$, etc.) were developed for fixed-effect models and may be too liberal or too conservative for mixed models ^[6]. Our simulation results suggest they perform adequately at ICC values between 0.1 and 0.5, but at ICC = 0.7 the false positive rate for COVRATIO exceeded 10%, indicating that

threshold calibration should depend on the degree of clustering.

Conclusion

This research derived and evaluated influence diagnostics for linear mixed models applied to hierarchical forest inventory data from Tasmania. The one-step approximation for Cook's distance matched exact leave-one-out deletion to within 3.2% relative error on average, making it a practical tool for large ecological datasets. Among the four diagnostics examined, COVRATIO was the most sensitive to influential observations at the random-effect level, while Cook's distance offered the best balance of detection and specificity for fixed effects. Application to 2,847 tree measurements across 143 plots revealed that 6.8% of observations were influential under at least one criterion, and their removal shifted the estimated mean basal area by 11.4% and the between-plot variance component by 20.1%. The greater sensitivity of random effects to individual observations has direct consequences for inventory design and carbon stock estimation, where between-group variance determines the required sampling effort. The zero-sum game interpretation of variance partitioning provides a conceptual framework for understanding why influential observations affect random and residual components asymmetrically. For forest managers and inventory statisticians, routine influence checking using one-step approximations adds minimal computational cost and can reveal data quality issues that would otherwise go undetected in standard model summaries.

Acknowledgements

Funding Sources

This research was supported by internal institutional research allocations. No external commercial funding was received.

Institutional Support

The authors thank the Department of Probability and Statistics at Mendoza Research Institute of Data Sciences for computational facilities and library access.

Contributions Not Qualifying for Authorship

The authors acknowledge Sustainable Timber Tasmania for maintaining publicly accessible inventory databases and two colleagues who provided technical feedback during manuscript preparation.

References

1. Cook RD. Detection of influential observation in linear regression. *Technometrics*. 1977;19(1):15-18.
2. Belsley DA, Kuh E, Welsch RE. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley; 1980.
3. Demidenko E. *Mixed Models: Theory and Applications with R*. 2nd ed. Wiley; 2013.
4. Pinheiro JC, Bates DM. *Mixed-Effects Models in S and S-PLUS*. Springer; 2000.
5. Gregoire TG, Valentine HT. *Sampling Strategies for Natural Resources and the Environment*. CRC Press; 2007.
6. Christensen R, Pearson LM, Johnson W. Case-deletion diagnostics for mixed models. *Technometrics*. 1992;34(1):38-45.
7. Banerjee M, Frees EW. Influence diagnostics for linear longitudinal models. *Journal of the American Statistical Association*. 1997;92(439):999-1005.

8. Bowman DMJS. Australian Rainforests: Islands of Green in a Land of Fire. Cambridge University Press; 2000.
9. Schabenberger O, Gotway CA. Statistical Methods for Spatial Data Analysis. CRC Press; 2005.
10. von Neumann J, Morgenstern O. Theory of Games and Economic Behavior. Princeton University Press; 1944.
11. Searle SR, Casella G, McCulloch CE. Variance Components. Wiley; 1992.
12. Zuur AF, Ieno EN, Walker NJ, Saveliev AA, Smith GM. Mixed Effects Models and Extensions in Ecology with R. Springer; 2009.
13. Verbeke G, Molenberghs G. Linear Mixed Models for Longitudinal Data. Springer; 2000.
14. West BT, Welch KB, Galecki AT. Linear Mixed Models: A Practical Guide Using Statistical Software. 2nd ed. CRC Press; 2015.
15. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika*. 1986;73(1):13-22.