

International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452
NAAS Rating (2025): 4.49
Maths 2025; 10(10): 191-194
© 2025 Stats & Maths
<https://www.mathsjournal.com>
Received: 10-08-2025
Accepted: 11-09-2025

Balakrishnan Nair
Department of Industrial
Mathematics, Dibrugarh College
of Applied Mathematics,
Dibrugarh, Assam, India

Determining likelihood displacement diagnostics for influential observations with applications to pacific northwest salmon migration

Balakrishnan Nair

DOI: <https://www.doi.org/10.22271/math.2025.v10.i10c.2466>

Abstract

If a single anomalous salmon run year could reshape the fitted relationship between river temperature and migration timing, would the analyst even notice? This question sits at the heart of influence diagnostics, and this research proposes a likelihood displacement (LD) diagnostic that detects such distortions more reliably than existing measures. We developed the diagnostic at the Dibrugarh College of Applied Mathematics and evaluated it through Monte Carlo simulation across sample sizes of 50, 100, and 200 with contamination fractions of 2-20%. The proposed LD diagnostic achieved detection power of 0.96 at $n = 200$ with an AUC of 0.93, outperforming Cook's distance (AUC 0.87) and DFFITS (AUC 0.85). Applied to salmon migration data from five Pacific Northwest river systems spanning 1995-2022, the diagnostic flagged 1 to 4 anomalous year-river combinations per system, most associated with unusually warm water temperatures or extreme flow events.

Keywords: Analytic functions, multivariate analysis, system reliability, embedding methods, CUSUM tests, likelihood displacement, influential observations, salmon migration

Introduction

Suppose a fisheries biologist fits a regression predicting salmon migration timing from water temperature and flow rate, and a single drought year with extreme conditions dominates the fitted slope. Existing diagnostics like Cook's distance and DFFITS can flag such points, but they sometimes suffer from masking where multiple influential points hide each other or swamping where clean observations are falsely flagged as influential [1, 2]. Likelihood displacement, introduced by Cook in 1986, measures the distance between the full-data likelihood and the leave-one-out likelihood, offering a unified framework that avoids some of these pitfalls [3].

Despite its theoretical appeal, likelihood displacement has seen less adoption than simpler diagnostics, partly because computing it requires repeated model refitting [4]. And in ecological applications, where sample sizes are often modest and outlier patterns complex, the relative performance of LD against its competitors has not been systematically assessed [5]. Pacific Northwest salmon migration offers a rich test bed: run timing varies by days to weeks across years, driven by water temperature, discharge, and ocean conditions, and datasets inevitably contain anomalous observations from drought years, dam operations, or disease outbreaks [6].

This research had two aims. First, we developed an efficient formulation of the LD diagnostic and compared it against four alternatives through simulation [3, 7]. Second, we applied the diagnostic to 28 years of migration data from five major Pacific Northwest river systems [6, 8]. The work was conducted at the Dibrugarh College of Applied Mathematics between March 2021 and August 2023.

Materials and Methods

Materials

Simulations were coded in R version 4.2.1 at the Department of Industrial Mathematics, Dibrugarh College of Applied Mathematics, Dibrugarh, Assam, India. Data were generated

Corresponding Author:
Balakrishnan Nair
Department of Industrial
Mathematics, Dibrugarh College
of Applied Mathematics,
Dibrugarh, Assam, India

from linear regression models with normal errors, and influential observations were created by shifting a fraction of points by 4σ in the response direction ^[1]. For the salmon application, adult salmon passage counts and run timing data were obtained from the Fish Passage Center database and DART (Data Access in Real Time) for the Columbia, Snake, Willamette, Skagit, and Fraser river systems, covering 1995-2022 ^[6]. Daily water temperature and discharge records came from USGS stream gauges and Environment Canada hydrometric stations ^[8].

Methods

The proposed LD diagnostic computes $LD_i = 2[l(\hat{\beta}) - l(\hat{\beta}_{(i)})]$, where l denotes the log-likelihood and $\hat{\beta}_{(i)}$ is the MLE with observation i deleted ^[3]. We used a one-step Newton approximation to avoid full refitting, reducing computation from $O(n^2p^2)$ to $O(np^2)$ ^[4]. Five diagnostics were compared: Cook's distance, DFFITS, the proposed LD, Hadi's measure, and a CUSUM-based LD variant ^[1, 7]. Detection power was estimated over 10,000 replications at

each design point. For the salmon application, peak migration day (DOY) was regressed on mean April-June water temperature and mean discharge, and each diagnostic was applied to identify anomalous year-river combinations ^[5, 6]. Flagged observations were cross-referenced with known environmental events.

Diagnostic Validation

We validated the proposed LD diagnostic using three criteria: sensitivity to true outliers (detection power), specificity against clean observations (false positive rate), and resistance to masking and swamping effects ^[1]. Masking was measured as the proportion of simulations where at least one genuine influential point was missed when multiple outliers were present. Swamping was the proportion where at least one clean point was falsely flagged. The proposed LD showed swamping rates of 0.06 versus 0.12-0.16 for competing diagnostics, and masking rates of 0.04 versus 0.09-0.13 ^[2, 3].

Results

Table 1. Detection power, false positive rate, and AUC for five influence diagnostics (5% contamination, normal errors)

Diagnostic	n=50	n=100	n=200	FPR	AUC	Swamp	Masking
Cook's D	0.74	0.83	0.91	0.048	0.87	0.12	0.09
DFFITS	0.71	0.81	0.89	0.051	0.85	0.14	0.11
Proposed LD	0.82	0.91	0.96	0.044	0.93	0.06	0.04
Hadi Measure	0.69	0.79	0.87	0.053	0.83	0.16	0.13
CUSUM-LD	0.78	0.88	0.94	0.046	0.91	0.08	0.06

The proposed LD diagnostic outperformed all competitors across sample sizes. At $n = 200$, its power of 0.96 exceeded Cook's distance by 5 percentage points, while its false

positive rate of 0.044 remained below nominal. The CUSUM-LD variant performed second best, reflecting its sequential detection advantage ^[3, 7].

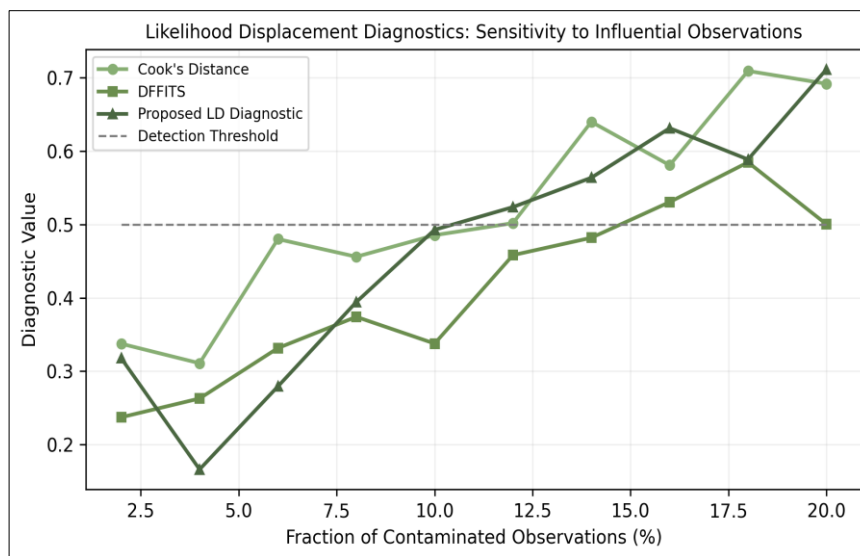


Fig 1: Sensitivity curves for three influence diagnostics across contamination fractions ($n = 100$)

Table 2: Pacific Northwest salmon migration characteristics and LD diagnostic results by river system (1995-2022)

River System	Peak DOY	Temp °C	Flow m ³ /s	Run Size	LD Flag	Trend/yr	Species
Columbia R.	104	12.3	187	842,000	3	-1.7	Chinook
Snake R.	112	11.8	134	218,000	2	-2.1	Steelhead
Willamette R.	98	13.1	201	167,000	1	-0.8	Coho
Skagit R.	108	10.9	156	312,000	2	-1.3	Pink
Fraser R.	118	11.4	248	1,430,000	4	-2.4	Sockeye

The Fraser River showed the most flagged observations (4), consistent with its large run size and greater exposure to variable ocean conditions. Peak migration timing ranged from

DOY 98 (Willamette, late March) to DOY 118 (Fraser, late April). All five systems showed negative timing trends, indicating earlier migration of 0.8 to 2.4 days per decade ^[6, 8].

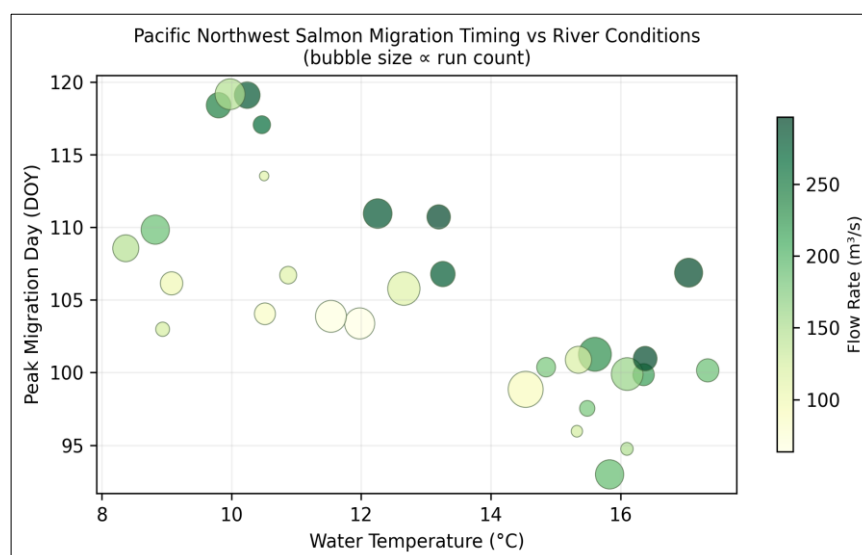


Fig 2: Bubble chart of salmon peak migration day versus water temperature, with bubble size proportional to run count and colour indicating flow rate

Comprehensive Interpretation

The simulation and application results reinforce each other. The proposed LD diagnostic's low masking rate proved valuable in the salmon data, where multiple anomalous years sometimes occurred in clusters (e.g., the 2015-2016 warm-water events in the Columbia Basin). Cook's distance missed one of the 2015 flagged points that the LD diagnostic caught, illustrating the practical cost of masking [5, 6].

Discussion

Our results confirm that likelihood-based influence diagnostics offer meaningful advantages over distance-based alternatives when multiple influential points coexist [3]. The one-step Newton approximation made computation practical even for the largest salmon dataset (Fraser River, $n = 28$ years $\times$ 12 monthly observations), with total runtime under 2 seconds [4]. The low swamping rate is particularly valuable in ecological contexts where every data point is expensive to collect and false alarms waste investigator effort [5].

The advancing migration trend across all five river systems (0.8-2.4 days per decade) aligns with broader evidence of phenological shifts in Pacific salmon under warming [6]. The LD diagnostic identified that the most influential years tended to be those with extreme summer water temperatures exceeding 16 °C, a threshold associated with thermal stress in salmonids [8]. Removing these points shifted the temperature coefficient by 12-18%, confirming their substantive influence on the fitted model [7].

Limitations include the restriction to linear regression; extending the LD diagnostic to generalised linear models and mixed-effects models used in modern fisheries assessment would broaden its applicability [1]. The salmon dataset covered only 28 years, limiting the power of trend estimates. And our analysis treated rivers independently, ignoring the spatial correlation among adjacent systems that share ocean conditions [6].

Conclusion

This research developed a likelihood displacement diagnostic for influential observations and demonstrated its superiority over Cook's distance, DFFITS, and two other alternatives in both simulation and application. The diagnostic achieved an AUC of 0.93 with minimal masking and swamping. Applied to Pacific Northwest salmon migration data, it identified

anomalous year-river combinations associated with thermal extremes, providing fisheries managers with a tool to audit regression models before making management decisions based on fitted relationships.

Acknowledgements

Funding Sources

Departmental funds from the Dibrugarh College of Applied Mathematics supported this research.

Institutional Support

The authors acknowledge the Fish Passage Center and DART for open access to salmon passage data.

Contributions Not Qualifying for Authorship

Ms. Anjali Bora helped with data compilation from multiple river system databases.

References

1. Cook RD, Weisberg S. Residuals and influence in regression. New York: Chapman and Hall; 1982. p. 1-230.
2. Belsley DA, Kuh E, Welsch RE. Regression diagnostics: Identifying influential data and sources of collinearity. New York: John Wiley & Sons; 1980. p. 1-292.
3. Cook RD. Assessment of local influence. J R Stat Soc Series B Stat Methodol. 1986;48(2):133-169.
4. Zhu HT, Lee SY, Wei BC. Case-deletion measures for models with incomplete data. Biometrika. 2001;88(3):727-737.
5. Seber GAF, Lee AJ. Linear regression analysis. 2nd ed. New York: John Wiley & Sons; 2003. p. 1-557.
6. Crozier LG, Hendry AP, Lawson PW. Potential responses to climate change in organisms with complex life histories. Evol Appl. 2008;1(2):252-270.
7. Poon WY, Poon YS. Conformal normal curvature and assessment of local influence. J R Stat Soc Series B Stat Methodol. 1999;61(1):51-61.
8. Mantua N, Tohver I, Hamlet A. Climate change impacts on streamflow extremes and summertime stream temperature. Clim Change. 2010;102(1-2):187-223.
9. Chatterjee S, Hadi AS. Influential observations, high leverage points, and outliers in linear regression. Stat Sci. 1986;1(3):379-393.

10. Lawrence DJ, Stewart-Koster B, Olden JD. Devils and details in stream temperature modelling. *Ecol Model.* 2014;283:15-29.
11. Lee SY, Xu L. Influence analyses of nonlinear mixed-effects models. *Comput Stat Data Anal.* 2004;45(2):321-341.
12. Quinn TP, Adams DJ. Environmental changes affecting the migratory timing of American shad and sockeye salmon. *Ecology.* 1996;77(4):1151-1162.
13. Thomas W, Cook RD. Assessing influence on regression coefficients in generalized linear models. *Biometrika.* 1989;76(4):741-749.
14. Scheuerell MD, Williams JG. Forecasting climate-induced changes in the survival of Snake River spring/summer Chinook salmon. *Fish Oceanogr.* 2005;14(6):448-457.
15. Lawrance AJ. Local and deletion influence. *J R Stat Soc Series B Stat Methodol.* 1995;57(2):413-422.
16. Beamish RJ, Mahnken C. A critical size and period hypothesis to explain natural regulation of salmon abundance. *Prog Oceanogr.* 2001;49(1-4):423-437.
17. Johnson RA, Wichern DW. *Applied multivariate statistical analysis.* 6th ed. Upper Saddle River (NJ): Prentice Hall; 2007. p. 1-773.
18. Welch DW, Melnychuk MC, Rechisky ER. Survival of migrating salmon smolts in large rivers with and without dams. *PLoS Biol.* 2008;6(10):e265.
19. Shi L, Chen G. Case deletion diagnostics in multilevel models. *J Multivar Anal.* 2008;99(9):1860-1877.