

Peer Review

Review of: "Statistically Significant Linear Regression Coefficients Solely Driven by Outliers in Finite-Sample Inference"

Sunil Parakramaweera¹

1. Independent researcher

Statistically Significant Linear Regression Coefficients Solely Driven by Outliers in Finite-Sample Inference – by Felix Reichel

REVIEW

Going by the Abstract, it seems that you have completed the following tasks to achieve the main objective stated in the heading of this paper.

1. Investigate the impact of Outliers on the significance of OLS regression coefficients
2. Demonstrate that a single Outlier can cause an insignificant coefficient to become significant, and compare it with robust regression that reduces the effects of outliers on estimates
3. Approximate the influence of a single Outlier on estimated regression coefficients
4. Discuss diagnostic statistics to detect 'influential observations' in OLS regression
5. Relate the issue of the normality assumption in Simple Linear Regression (SLR)
6. Address the dangers of solely relying on p-values
7. Discuss existing regression methods and their relation to the assumptions of the Gauss-Markov theorem

In relation to (1) – (7), my observations are as follows:

- a. It is not clear to me your **definition** of an Outlier. What is an Outlier?
- b. You don't suggest a **new** criterion, different from existing criteria, to detect outliers. And there is no 'universal' rigid mathematical criterion to detect outliers. It is a 'subjective decision' of the

researcher to treat an observation in the sample as an ‘Outlier’.

- c. It is well known that Huber’s Robust Regression is less sensitive to outliers than OLS Regression. On pages 14-16, you confirm that with figs. 2-4. You could drop that single outlier from the data and run OLS regression again to get a better model and more accurate regression coefficient estimates. That outlier stays away from other observations in the top right corner, in figs. 2 and 3.
- d. The presently available criteria to detect ‘influential observations’ that exist can be found in any statistics book. For example, using a confidence interval with $|R_i|$ or $|e_i|$ (eq.14) and the critical t – value (eq.17).
- e. In your paper, you have highlighted only the influence of Outliers on estimated reg. coefficients. There is another critical issue that affects the estimates of reg. coefficients, i.e., ‘multicollinearity’ among data columns in matrix X . That issue can lead to distorted leverage values h_{ii} (eq.15) because of a singular or near-singular matrix ($X^T X$). That issue has been addressed by Ridge Regression with Tikhonov’s Regularization (eq. 27). In any ‘multiple regression’ analysis, both issues (robust and ridge) must be addressed for accurate estimates of reg. coefficients.

I suggest that you rewrite the paper along the lines (a) – (e) that cover both issues, accommodating them in a more general MLR model.

Attachments: available at <https://doi.org/10.32388/M92TY3>

Declarations

Potential competing interests: No potential competing interests to declare.