http://doi.org/10.1590/1678-992X-2022-0209

ISSN 1678-992X

Scientia Agricola

Animal Science and Pastures

A practical method to account for outliers in simple linear regression using the median of slopes

Luis O. Tedeschi[1*], Michael L. Galyean[2]

[1]Texas A&M University – Dept. of Animal Science, TX 77843-2471 – College Station, Texas – USA.
[2]Texas Tech University – Dept. of Veterinary Sciences, TX 79409 – Lubbock, Texas – USA.
*Corresponding author <luis.tedeschi@tamu.edu>

Edited by: Gerson Barreto Mourão

Received September 28, 2022
Accepted April 27, 2023

ABSTRACT: The ordinary least squares (OLS) can be affected by errors associated with heteroscedasticity and outliers, and extreme points can influence the regression parameters. Methods based on the median rather than on the mean and variance are more resistant to outliers and extreme points. These methods could be used to obtain regression parameter estimates that reflect more accurately the genuine relationship between the Y and X variables, leading to better identification of outliers and extreme points by comparing the slopes and intercepts of both methods. The Theil-Sen (TS) regression computes all possible pairwise slopes and determines the median of slopes as the regression slope. Here, we illustrated the potential use of TS and frequently used robust regression (RR) techniques to single linear regression using synthetic datasets and a practical problem in animal science. Three synthetic datasets were created assuming the normal distribution of Y and X values: one was free of outliers, while the other two had one or two clusters of outliers but the same X values. The TS, OLS, and RR had nearly identical regression parameter estimates for the dataset without synthetic outliers. However, the intercept and slope estimates by the OLS method differed considerably from the TS and RR methods when one or two clusters of outliers were included. The TS approach could be used to indirectly determine the presence of outliers or extreme points by comparing the 95 % confidence interval of the TS and OLS parameter estimates.

Keywords: estimation, methods, relationship, robust, statistics

Introduction

The ordinary least squares (OLS) is the standard method of regression analyses, but the technique is subject to errors associated to heteroscedasticity and the presence of outliers. Concerning outliers, even a single spurious point in either X or Y axes can markedly affect the slope and intercept estimates (Wilcox, 2021). After checking for the correctness of data entry, careful data evaluation for outliers becomes an essential preliminary step in the regression analyses and data analytics in general (Tedeschi, 2022).

Datasets with multiple X and Y values allow determining all possible two-point slopes. If these two-point slopes are weighted by the square distance between their X values, the weighted average equals the slope of the OLS regression method. The approach of computing all possible slopes was used to address the presence of outliers in regression data by Theil (1992) and Sen (1968). However, rather than calculating a weighted average, these scientists computed the median of the slopes. Given that the median is insensitive to extreme (i.e., influential) points, its use to fit linear regression was proposed by Wald (1940) with subsequent derivative works (Walters et al., 2006); however, the Theil-Sen (TS) approach possibly provides a more robust nonparametric method to estimate the slope.

Since its inception in the late 1970s by John W. Tukey (1977), robust estimation has percolated through many fields of science. However, agricultural sciences seem reluctant or unaware of its applications and benefits. Although various approaches exist (Zacharias et al., 1996), robust regressions (RR) have been generally neglected in the development and evaluation phases in the modeling field (Tedeschi, 2006). This is partly because one cannot be sure whether the achieved model adequacy is a result of the model logistics *per se* or the RR estimation process is changing parameter estimates, consequently making model predictions look better than reality.

Our objective was to illustrate the potential use of TS and another frequently used RR technique (e.g., M and MM estimators) when applied to a common analysis, the regression of observed on predicted values, to assess the fit of prediction equations. Although resilient to the effects of outliers, the TS approach is seldom used in animal science research. However, it could offer an alternative to OLS, mainly when data are influenced by outliers, either lightly or heavily affecting the estimation of the intercept and slope. We also briefly discuss the application of RR methods in animal science research.

Materials and Methods

All data calculations and statistical analyses were conducted with R 4.2.2 (R Core Team, 2019). No animals were used in this research; thus, Institutional Animal Care and Use Approval was unnecessary.

Synthetic dataset analyses

Appendix 1 has the R code to generate the synthetic datasets. Synthetic datasets were created assuming the normal distribution and random number generation based on the Mersenne-Twister method with a seed of 12345. The independent (X) dataset was generated considering a sample size of 200, a mean of 5, and a standard deviation (SD) of 0.1 (Appendix 1). The error dataset was generated assuming a sample size of 200, a mean 0, and an SD of 0.05 (Appendix 1). The dependent variable (Y) was the sum of X and the error ($Y = X + error$). These settings produced a Pearson correlation between Y and X of 0.915 ($p < 0.001$), a slope of 1, and zero intercept.

Synthetic outliers – Synthetic outliers were created assuming the normal distribution and random number generation based on the Mersenne-Twister method with a seed of 12346. A pre-determined number of X values (i.e., 10) were randomly selected from a specific X range, and random values obtained from a third normal distribution were added to the Y values to emulate outliers (Appendix 1). This cluster of points (i.e., 10) had the same X values, but their Y values were increased (or decreased) depending on the mean used in the second normal distribution. For the one outlier cluster simulation, the X range was between 5.18 and 5.4 (Appendix 1), and the third normal distribution had a mean –0.8 and SD of 0.02 (Appendix 1); thus, the adjusted Y values were decreased from the original values. For the two outlier cluster simulation, in addition to the cluster generated for the one outlier cluster simulation, the X range was between 4.8 and 4.9 (Appendix 1), and the fourth normal distribution had a mean of 1 and SD of 0.02 (Appendix 1); thus, the adjusted Y values were increased from the original values.

Real dataset analyses

A comparison between OLS and TS regressions was conducted using the dataset gathered by Hales et al. (2022) to predict dietary concentrations (Mcal kg^{-1} of dry matter) of metabolizable energy (ME) from digestible energy (DE), commonly used in animal science research. Their dataset contained 134 observations from 34 papers published from 1975 to 2020 with bulls, steers, and heifers, using open-circuit respiration calorimetry systems obtained from respiration chambers or headboxes. The dietary concentrations of DE (1.84 to 3.88 Mcal kg^{-1}), crude protein (7.88 % to 24.08 %), neutral detergent fiber (15.65 % to 68.81 %), ether extract (1.94 % to 8.71 %), and starch (0 % to 56.85 %) were either measured and reported by the studies or calculated from various sources as described by Galyean et al. (2016).

Theil-Sen and robust regressions

Theil-Sen regression – The original development of the TS approach was published by Theil (1950a, 1950b, 1950c) and Sen (1968). The TS computes the slope (b_{TS}) of the regression between Y on X as the median of the *n* – 1 slopes between two consecutive pairs of Y and X, as shown in Eq. [1]. Alternatively, the TS can be computed with all possible combinations of Y and X as shown in Eq. [2], following the repeated median approach developed by Siegel (1982), who included preceding and succeeding Y and X pairs. Regardless of the method used, X_j and X_i must be different to avoid division by zero. The first approach (Eq. [1]) results in *n* ´ (*n* + 1) / 2 –1 data points, whereas the second approach (Eq. [2]) results on *n* × *n* data points. In our study, we adopted Eq. [1] to compute the slope and Eq. [3] to compute the intercept (a_{TS}) of the TS regression. The TS parameter estimates were obtained with the *mblm* package version 0.12.1 of R.

$$b_{TS} = \underset{\substack{i=1\ldots n-1\\ j=i+1\ldots n}}{M}\left(\frac{Y_J - Y_i}{X_J - X_i}\right) \quad [1]$$

$$b_{TS} = \underset{i=1\ldots n}{M}\left[\underset{j=1\ldots n}{M}\left(\frac{Y_J - Y_i}{X_J - X_i}\right)\right] \quad [2]$$

$$a_{TS} = \underset{i=1\ldots n}{M}\left(Y_i - b_{TS} \times X_i\right) \quad [3]$$

The correlation and determination coefficients for the TS regression were computed using the known relationship between the correlation coefficient and slopes of Y regressed on X (Y dependent and X independent variables), and X regressed on Y (X dependent and Y independent variables) when using OLS statistics. The correlation coefficient (r_{YX}) calculation is shown in Eq. [4], and the determination coefficient is r_{YX}^2. The slope of Y regressed on X gives the sign of the correlation. Thus, to compute the r_{YX} and r_{YX}^2 for the TS regression, we estimated the slopes (b_{YX} and b_{XY}) using Eq. [1], and applied them to Eq. [4].

$$r_{YX} = sign(b_{YX}) \times \sqrt{b_{XY} \times b_{YX}} \quad [4]$$

where: b_{XY} is the slope of X regressed on Y and b_{YX} is the slope of Y regressed on X.

Robust regression – The adoption of robust analysis and regression was partly slow because of complications and divergence in defining what *robust* meant, and many methods (i.e., 68) were devised (Andrews et al., 1972), resulting in confusion about which one to use. Many techniques and algorithms still exist, but the scientific community has gained much more information about them, and some have consolidated more towards the M-estimator classes of techniques (Wilcox, 2021). Several robust analyses and regression methods are available in commonly used statistical software packages; therefore their limited adoption is more related to a lack of information than how to apply them. Robust regression is a category of methods used to deal with outliers and extreme data points.

Robust regression methods characterize the location and scale of data points to ensure that changes in the data points caused by outliers have a relatively small effect on the regression parameters (Wilcox, 2021). Several RR approaches exist, namely least median squares, least trimmed squares, least trimmed absolute value, and many more location estimators that rely on different influence functions. Commonly used location estimators are the M-estimators with varying influence functions (e.g., Huber, Andrews, Hampel, and biweight, to list a few), R-estimators, S-estimators, and the MM-estimator, among many others (Wilcox, 2021). The main difference among these location estimators regards their breakdown values. The breakdown value refers to the quantitative robustness of a method (i.e., how robust a method is to increasing contamination in the data). The greater the breakdown value, the more robust the method. The breakdown value of the TS estimator is approximately 0.293 (Dietz, 1989). The M-estimator is substantially more efficient than the OLS method as it can handle a few outliers (Wilcox, 1998). However, the M-estimator is not the most effective method compared to the MM-estimator. Although the MM-estimator struggles with contamination bias, it can have the highest breakdown value of 0.5 under normality and satisfactory efficiency for small samples compared with other robust estimators (Wilcox, 2021). Thus, we used the M- and MM-estimators for our comparisons and they were computed with the *rlm* function in the *MASS* package of R.

Comparison among methods – The methods (OLS, TS, and M-, and MM-estimators) were compared with 500 synthetic datasets with two clusters of outliers described above (Appendix 1). Subsequently, the Pearson correlation of the 500 intercepts was estimated, and slopes between methods were obtained.

Statistical analyses

Confidence intervals and significance of intercepts and slopes – For the TS approach, the 95 % confidence interval (CI_{95}) for the medians of the slopes and intercepts were computed using the Wilcoxon Rank Sum (WRS) test via the *wilcox.test* function of the *stats* package of R. In a preliminary analysis, the WRS CI_{95} values were similar to those estimated by the algorithm described by Conover (1999) that estimates the lower and upper ranks of the medians using the following equation: $rank = n \times q \pm Z_{0.05} \times \sqrt{n \times q \times (1-q)}$, where n is the sample size, q is the quartile of interest (0.5 for median), and $z_{0.05}$ is the z-critical value (i.e., 1.96). Wilcox (2021, Ch. 4) discussed other functions to compute CI_{95} for medians. For the OLS and RR, the *t*-test was used to calculate the CI_{95} and the *p*-value of the intercept and slope.

Outliers – The existence of outliers and influential points in the Y values was determined using the Cook's Distance (CD) (Kutner et al., 2005), using the *cooks. distance* function of *stats* Package of R, and the studentized residuals (Kutner et al., 2005), using the *studres* function of the *MASS* package of R. Points that had CD greater than four times the mean of CD were deemed influential points, and points that had studentized residues above 3 and below -3 were considered outliers. Additionally, the Tukey's boxplot (Tukey, 1977) was used to identify potential outliers (points outside of 1.5 times the interquartile range [IQR], *i.e.*, the end of the whiskers).

Results and Discussion

Synthetic dataset analyses

Table 1 has the regression statistics, and Figures 1 through 3 show the scatter points of Y and X values, with regression lines depicted for the various methods. Figure 4 shows their respective boxplots, depicting potential outliers.

Without synthetic outliers – A plot of the Y and X values of the synthetic dataset is shown in Figure 1. As shown in Table 1, the OLS regression had an intercept of –0.0802 ($p_{(H0=0)}$ = 0.617), a slope of 1.0168 ($p_{(H0=1)}$ = 0.599), r^2 of 0.837, and mean square error (MSE) of 0.0023. The CD reported 11 values (shown in open circles in Figure 1); however, no potential outlier was identified using the studentized residue and the boxplot (Figure 4). The TS regression had an intercept of –0.222, a slope of 1.0455, r^2 of 0.9613, and MSE of 0.00232. The CI_{95} using the WRS test for the intercept was –0.2296 and –0.2168 ($p_{(H0=0)}$ < 0.001), and for the slope, it was 1.0198 and 1.0412 ($p_{(H0=1)}$ < 0.001). The RR using the M-estimator had an intercept of –0.2033 ($p_{(H0=0)}$ = 0.176),

Table 1 – Regression statistics of different regression methods for the synthetic datasets[1]

Regression	Intercept	P-value (H_0=0)	Slope	P-value (H_0=1)	r^2	MSE
Dataset without outliers						
OLS	– 0.0802	0.617	1.0168	0.599	0.837	0.0023
TS	– 0.222	—	1.0455	—	0.9613	0.00232
RR-M	0.2033	0.176	1.0415	0.166	0.8566	0.00181
RR-MM	–0.242	0.123	1.0493	0.116	0.8602	0.00168
Dataset with 1 cluster of outliers						
OLS	3.363	< 0.001	0.322	< 0.001	0.0425	0.0267
TS	0.163	—	0.968	—	0.732	0.033
RR-M	0.264	0.129	0.947	0.128	0.6	0.022
RR-MM	–0.194	0.218	1.04	0.206	0.845	0.0018
Dataset with 2 clusters of outliers						
OLS	6.96	< 0.001	– 0.384	< 0.001	0.026	0.0634
TS	0.517	—	0.898	—	0.594	0.0823
RR-M	0.706	< 0.001	0.86	< 0.001	—	0.00295
RR-MM	–0.279	0.08	1.056	0.076	0.84	0.0021

[1] OLS = ordinary least-squares, TS = Theil-Sen, RR-M = robust regression using the M-estimator, RR-MM = robust regression using the MM-estimator, MSE = mean square error.

a slope of 1.0415 ($p_{(H0 = 1)}$ = 0.166), r^2 of 0.8566, and MSE of 0.00181. The RR using the MM-estimator had an intercept of –0.242 ($p_{(H0 = 0)}$ = 0.123), a slope of 1.0493 ($p_{(H0 = 1)}$ = 0.116), r^2 of 0.8602, and MSE of 0.00168.

The CI95 suggested that the TS regression intercept and slope differed from zero and one, respectively. However,these statistics must be interpreted carefully because they depend data continuity near the expected CI_{95} thresholds. In contrast, the *p* values for the RR using the M-estimator and the MM-estimator use a *t*-test to access the two-tail probability (i.e., CI_{95} for RR are calculated). Regardless of the significance of the intercepts and slopes, the regression patterns were very similar without any clear tendency to depart from the expected relationship of Y = X (Figure 1). The MSE were nearly identical between OLS and TS (0.0023 and 0.00232, respectively) and between the two RR estimators (0.00181 and 0.00168).

With one cluster of outliers – Figure 2 shows the Y and X values of the synthetic dataset with ten randomly selected data points to emulate a cluster of outliers below the Y = X line. As shown in Table 1, the OLS regression had an intercept of 3.363 ($p_{(H0 = 0)}$ < 0.001), a slope of 0.322 ($p_{(H0 = 1)}$ < 0.001), r^2 of 0.0425, and MSE of 0.0267. The one cluster of outliers greatly affected the OLS regression estimates, showing a significant departure from the original intercept (0) and slope (1). The CD and studentized residue methods identified all ten synthetic outliers as potential outliers and influential points (shown in red triangles in Figure 2). The box plot in Figure 4 also depicts the same 10 points below the minimum value for the whiskers. The TS regression had an intercept of 0.163, a slope of 0.968, r^2 of 0.732, and MSE of 0.033. The Wilcoxon CI_{95} for the intercept was 0.152 and 0.167 ($p_{(H0 = 0)}$ < 0.001), and for the slope, it was 0.866 and 0.897 ($p_{(H0 = 1)}$ < 0.001). Thus, the outliers did not impact the coefficients of the TS regression as they did the OLS regression. The RR using the M-estimator had an intercept of 0.264 ($p_{(H0 = 0)}$ = 0.129), a slope of 0.947 ($p_{(H0 = 1)}$ = 0.128), r^2 of 0.6, and MSE of 0.0022. The RR using the MM-estimator had an intercept of –0.194 ($p_{(H0 = 0)}$ = 0.218), a slope of 1.04 ($p_{(H0 = 1)}$ = 0.206), r^2 of 0.845, and MSE of 0.0018.

Regardless of the *p*-values of the intercepts and slopes, the regression patterns between the TS and the two RR estimators were similar without any significant departure from the expected relationship of Y = X (Figure 2). Notably, the MSE of the TS regression was slightly greater than the MSE of the OLS regression, whereas the MSE values for the two RR were considerably smaller. Nevertheless, the intercepts (0.163 and 0.264) and slopes (0.968 and 0.947) of the TS and M-estimator regressions were very close.

With two clusters of outliers – Figure 3 shows the Y and X values of the synthetic dataset in which 20 random data points were selected and modified to emulate two clusters of outliers above and below the Y = X line. These synthetic points were detected as outliers and influential points by the CD test, studentized residuals, and boxplot (Figure 4). As shown in Table 1, the OLS regression had an intercept of 6.96 ($p_{(H0 = 0)}$ < 0.001), a

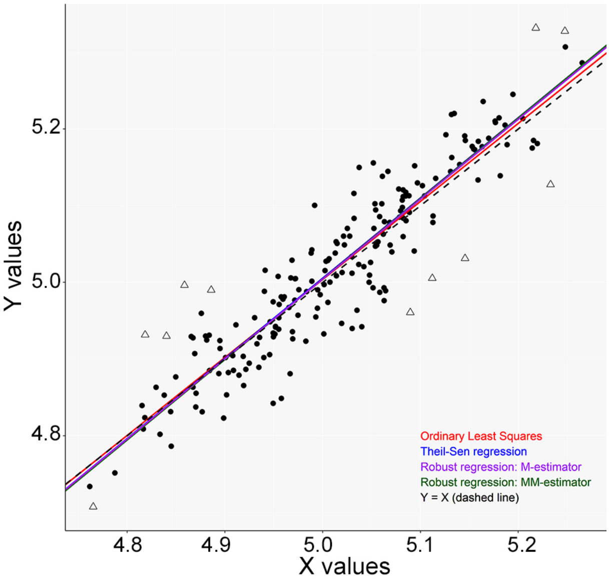

Figure 1 – Regressions using ordinary least squares (red line), Theil-Sen (blue line), and M-estimator (purple line) and MM-estimator (green line) robust regression, assuming a synthetic dataset with slope = 1 and intercept = 0 (black, dashed line) in which X ~ N(5, 0.1) and the random error ~ N(0, 0.05). Open triangles are influential points.

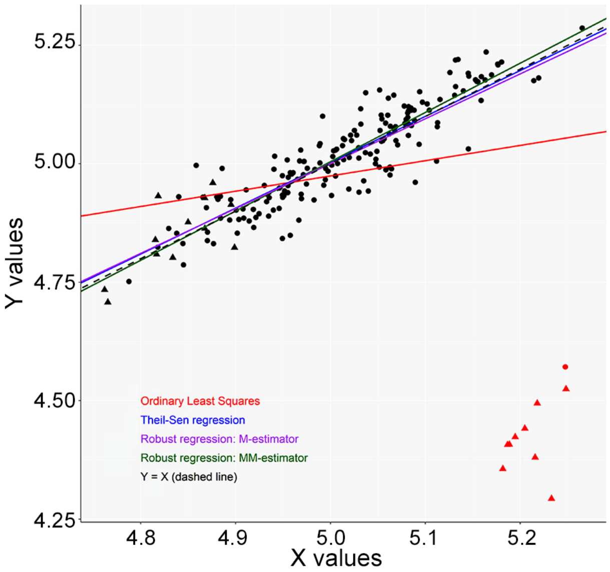

Figure 2 – Regressions using ordinary least squares (red line), Theil-Sen (blue line), and M-estimator (purple line) and MM-estimator (green line) robust regression, assuming the same dataset in Figure 1 but with 10 randomly created outlier points in a cluster (red symbols) and influential points (red triangles).

slope of −0.384 ($p_{(H0 = 1)}$ < 0.001), an r^2 of 0.026, and an MSE of 0.0634. As expected, the two clusters of outliers greatly impacted the OLS regression estimates, showing a significant departure from the original intercept (0) and slope (1), drastically increasing the MSE, and bringing the line farther apart from the Y = X line by assigning a negative slope. The TS regression had an intercept of 0.517, a slope of 0.898, an r^2 of 0.594, and an MSE of 0.0823. The CI$_{95}$ for the intercept was 0.504 and 0.521 ($p_{(H0 = 0)}$ < 0.001), and for the slope, it was 0.624 and 0.685 ($p_{(H0 = 1)}$ < 0.001). Interestingly Wilcoxon CI$_{95}$ for the slope was lower than the estimated slope (0.898). In this case, Conover (1999) CI$_{95}$ for the slope seemed more realistic (0.885 and 0.914) and yet did not include the value of 1 (p < 0.05). The RR using the M-estimator had an intercept of 0.706 ($p_{(H0 = 0)}$ < 0.001), a slope of 0.86 ($p_{(H0 = 1)}$ < 0.001), r^2 could not be determined, and MSE of 0.00295. The RR using the MM-estimator had an intercept of −0.279 ($p_{(H0 = 0)}$ = 0.08), a slope of 1.056 ($p_{(H0 = 1)}$ = 0.076), r^2 of 0.84, and MSE of 0.0021.

As with the one cluster of outliers, the TS and M-estimator RR had similar intercepts (0.517 and 0.706) and slopes (0.898 and 0.86), but the estimated MSE was considerably greater for the TS regression, suggesting that calculations are not being done similarly among these algorithms (i.e., packages). When the one and two clusters of outliers are compared, the TS and the M-estimator RR were very similar, but they were farther apart from the Y = X line than with the one cluster of outliers. On the other hand, the MM-estimator RR was more resilient to the two clusters of outliers, changing the intercept and slope slightly.

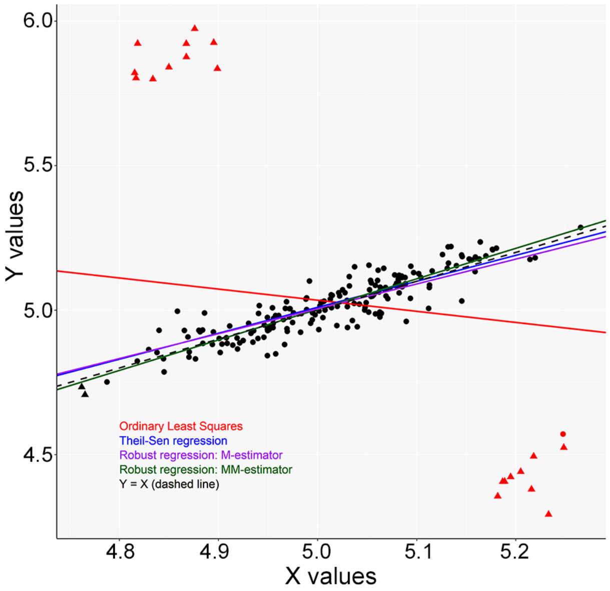

Figure 3 – Regressions using ordinary least squares (red line), Theil-Sen (blue line), and M-estimator (purple line) and MM-estimator (green line) robust regression, assuming the same dataset in Figure 1 but with 20 randomly created outlier points in two clusters (red points) and influential points (red triangles).

Relationship between methods

The Pearson correlation coefficients for the intercepts and slopes estimated by the OLS regression, TS regression, and the M-estimator and the MM-estimator RR for the 500 synthetic datasets with two clusters of outliers are shown in Table 2. The graphical representation of the scatter of the slopes is shown in Figure 5. As expected,

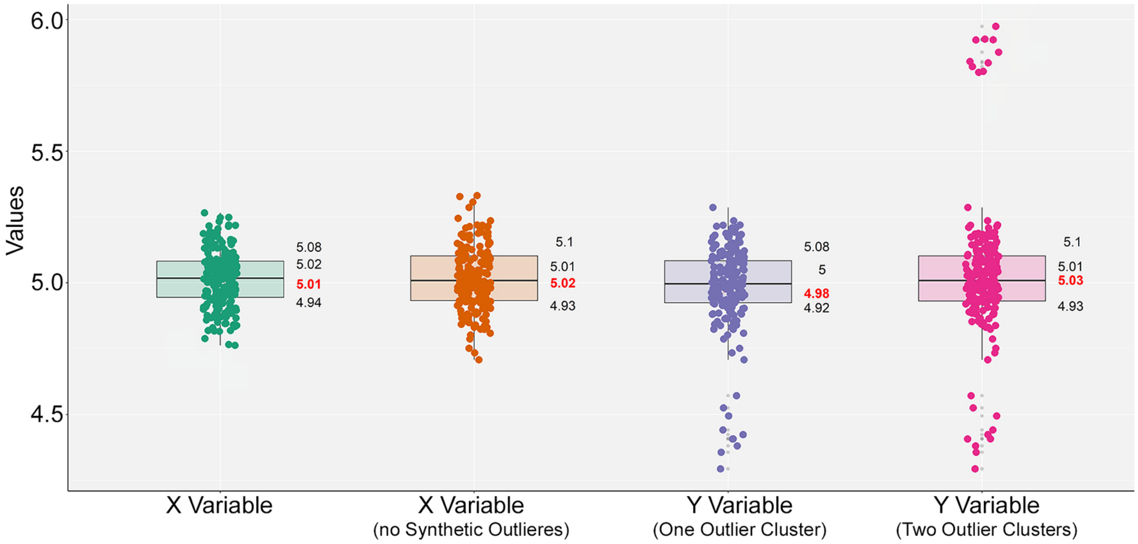

Figure 4 – Boxplots of the three synthetic datasets (Y and X values) showing the 25th (bottom line in of the box) and 75th (top line of the box) percentile, the median (50th percentile middle line in the box), and the average (red, star) values. Data points are jittered to minimize the overlapping of the data points. The data points outside the whisker limits are potential outliers and influential points.

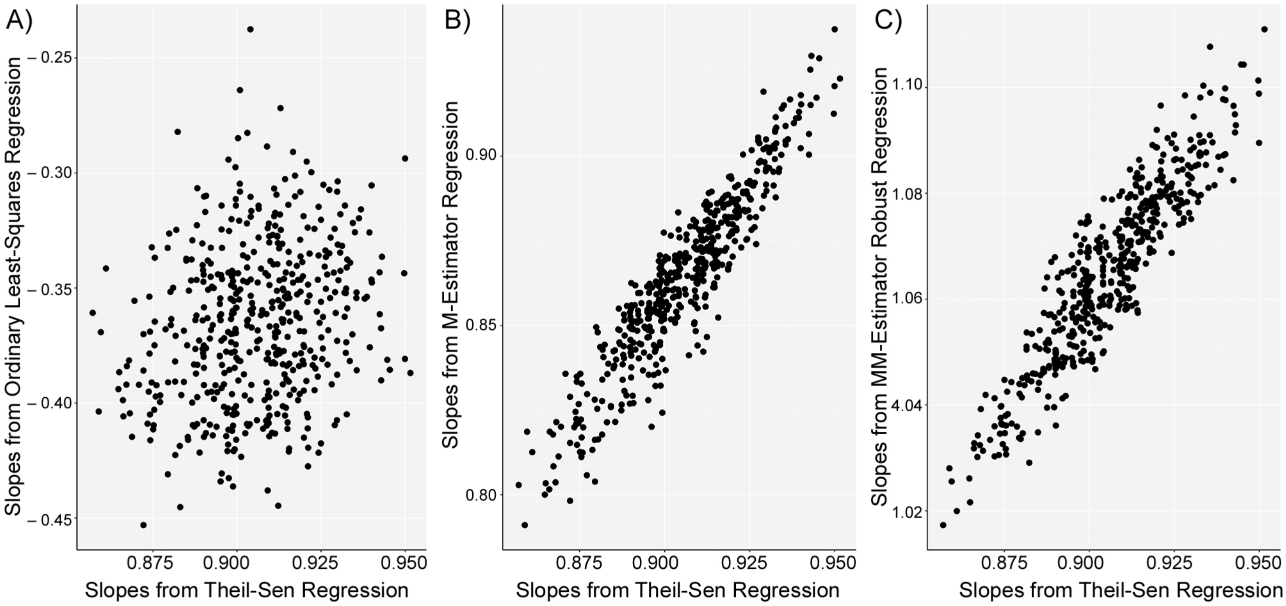

Figure 5 – Comparison between slopes determined using Theil-Sen regression with ordinary least-squares regression (A), M-estimator robust regression (B), and MM-estimator robust regression (C), using 500 simulated synthetic datasets with two outlier clusters.

Table 2 – Correlation of intercepts and slopes among methods[1]

Methods	Methods		
	TS	M-estimator	MM-estimator
		Intercepts	
OLS	0.224***	0.317***	0.006
TS		0.931***	0.917***
M-estimator			0.837***
		Slopes	
OLS	0.225***	0.318***	0.007
TS		0.931***	0.917***
M-estimator			0.837***

[1]Levels of significance: *** $p < 0.001$.

the parameter estimates of the TS regression were more closely related to the M- and MM-estimator RR, and the parameter estimates using the OLS regression were poorly correlated to both TS and RR (Table 2).

Real dataset analyses

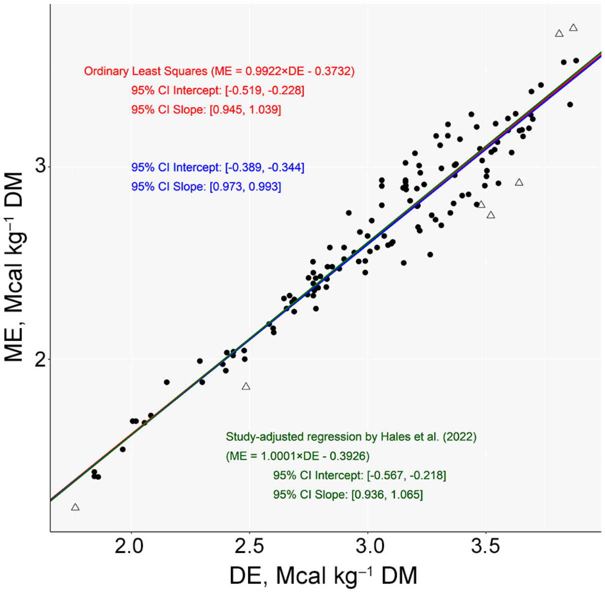

Figure 6 – Comparison of different regression methods to estimate dietary concentrations (Mcal kg^{-1} of dry matter) of metabolizable energy (ME) from digestible energy (DE), using the dataset gathered by Hales et al. (2022). Open triangles are influential points but not outliers. CI is the confidence interval.

Figure 6 shows the OLS and TS regressions to predict dietary concentrations (Mcal kg^{-1} of dry matter) of metabolizable energy (ME) from digestible energy (DE), using the dataset gathered by Hales et al. (2022). The CI_{95} of the intercepts and slopes of the OLS and TS overlap, suggesting they are not statistically different ($p > 0.05$) and no outlier is present, although influential points may exist (open triangles in Figure 5). The TS regression ($ME = 0.99 \times DE - 0.373$) is closer to that derived by Galyean et al. (2016) ($ME = 0.9611 \times DE - 0.2999$) than the one reported by Hales et al. (2022) ($ME = 1.0001 \times DE - 0.3926$), which was determined by the mixed-model regression with adjustment for random slope and intercept effects of the study. Nonetheless, the CI_{95} for the intercepts and slopes of these equations overlap. Concerning other applications in animal science, RR has been used in a limited number of studies to account for the existence of outliers (i.e., heteroscedasticity) in multi-breed genetic evaluations (Cardoso et al., 2007) to obtain smoother growth curves in pigs (Jiao et al., 2014), to determine energy partitioning in pigs (Strathe et al., 2010), and to perform a meta-analysis of the impact of growth hormones on carcass quality (Lean et al., 2018).

Practical applications

One practical application of the TS regression (or RR) is determining the existence and intensity of potential outliers in the dataset. For instance, one can check whether the CI_{95} of the slopes and intercepts between OLS and TS regressions overlap. If both overlap, then one would reject the null hypothesis at $p < 0.05$ that outliers exist in the dataset. This approach assumes that when outliers do not exist, the overlap of the distribution between slopes from OLS and TS regressions is expected to be high. In fact, when 1,000, 200-point synthetic datasets were created using the without synthetic outliers approach described above, the overlap between their distributions was 96.5 % for intercepts and 96.4 % for slopes. Thus, in the absence of outliers, one should expect that OLS and TS slopes CI_{95} or intercepts CI_{95} will overlap about 96 % of the time.

The main limitation of using the TS regression is that current algorithms are applied to single linear equations. Alternative approaches exist for multiple linear equations using OLS to estimate the slope of elemental subsets and then estimate the median of the slopes (Wilcox, 2021). However, this approach becomes ineffective as the sample size and the number of predictors (X variables) increase despite the adjustments (Wilcox, 2021). Similarly, nonlinearity may complicate TS regression because it would be more challenging to differentiate between outliers and the presence of curvilinearity in the response variable. Therefore, the M- or MM-estimator robust regressions should be used when the multiple regression analyses are needed to capture complex multi-dimensional relationships.

Our simulations confirmed that TS and RR (M- and MM-estimators) are resilient to outliers and influential points. They should be used as preliminary steps to certify that data are outlier-free to ensure that the parameter estimates reflect the proper relationship between the variables.

Authors' Contributions

Conceptualization: Galyean ML. **Data curation**: Tedeschi LO, Galyean ML. **Formal analysis**: Galyean ML. **Investigation**: Tedeschi LO, Galyean ML. **Methodology**: Tedeschi LO, Galyean ML. **Project administration**: Tedeschi LO. **Resources**: Tedeschi LO, Galyean ML. **Writing-original draft**: Galyean ML. **Writing-review & editing**: Tedeschi LO, Galyean ML.

References

Andrews DF, Bickel PJ, Hampel FR, Huber PJ, Rogers WH, Tukey JW. 1972. Robust Estimates of Location: Survey and Advances. Princeton University Press, Princeton, NJ, USA.

Cardoso FF, Rosa GJM, Tempelman RJ. 2007. Accounting for outliers and heteroskedasticity in multibreed genetic evaluations of postweaning gain of Nelore-Hereford cattle1. Journal of Animal Science 85 4: 909-918. https://doi.org/10.2527/jas.2006-668

Conover WJ. 1999. Practical Nonparametric Statistics. 3rd ed. Wiley, New York, NY, USA.

Dietz EJ. 1989. Teaching regression in a nonparametric statistics course. The American Statistician 43: 35-40. https://doi.org/10.1080/00031305.1989.10475606

Galyean ML, Cole NA, Tedeschi LO, Branine ME. 2016. Board-invited review: efficiency of converting digestible energy to metabolizable energy and reevaluation of the California Net Energy System maintenance requirements and equations for predicting dietary net energy values for beef cattle. Journal of Animal Science 94: 1329-1341. https://doi.org/10.2527/jas.2015-0223

Hales KE, Coppin CA, Smith ZK, McDaniel ZS, Tedeschi LO, Cole NA, et al. 2022. Predicting metabolizable energy from digestible energy for growing and finishing beef cattle and relationships to prediction of methane. Journal of Animal Science 100: 1-11. https://doi.org/10.1093/jas/skac013

Jiao S, Maltecca C, Gray KA, Cassady JP. 2014. Feed intake, average daily gain, feed efficiency, and real-time ultrasound traits in Duroc pigs. I. Genetic parameter estimation and accuracy of genomic prediction. Journal of Animal Science 92: 2377-2386. https://doi.org/ 10.2527/jas.2013-7338

Kutner MH, Nachtsheim CJ, Neter J, Li W. 2005. Applied Linear Statistical Models. 5th ed. McGraw-Hill, New York, NY, USA.

Lean IJ, Golder HM, Lees NM, McGilchrist P, Santos JEP. 2018. Effects of hormonal growth promotants on beef quality: a meta-analysis. Journal of Animal Science 96: 2675-2697. https://doi.org/10.1093/jas/sky123

R Core Team. 2019. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. Available at: http://www.R-project.org [Accessed on: Feb 5, 2019]

Sen PK. 1968. Estimates of the regression coefficient based on Kendall's tau. Journal of the American Statistican Society 63: 1379-1389. https://doi.org/10.1080/01621459.1968.10480934

Siegel AF. 1982. Robust regression using repeated medians. Biometrika 69: 242-244. https://doi.org/10.1093/biomet/69.1.242

Strathe AB, Danfær A, Chwalibog A, Sørensen H, Kebreab E. 2010. A multivariate nonlinear mixed effects method for analyzing energy partitioning in growing pigs. Journal of Animal Science 88: 2361-2372. https://doi.org/10.2527/jas.2009-2065

Tedeschi LO. 2006. Assessment of the adequacy of mathematical models. Agricultural Systems 89: 225-247. https://doi.org/10.1016/j.agsy.2005.11.004

Tedeschi LO. 2022. ASAS-NANP Symposium: Mathematical Modeling in Animal Nutrition: The progression of data analytics and artificial intelligence in support of sustainable development in animal science. Journal of Animal Science 100: 1-11. https://doi.org/10.1093/jas/skac111

Theil H. 1950a. A rank-invariant method of linear and polynomial regression analysis. Part I. Proceedings of the Royal Netherlands Academy of Sciences 53: 386-392.

Theil H. 1950b. A rank-invariant method of linear and polynomial regression analysis. Part II. Proceedings of the Royal Netherlands Academy of Sciences 53: 521-525.

Theil H. 1950c. A rank-invariant method of linear and polynomial regression analysis. Part III. Proceedings of the Royal Netherlands Academy of Sciences 53: 1397-1412.

Theil H. 1992. A rank-invariant method of linear and polynomial regression analysis. p. 345-381. In: Raj B, Koerts J. eds. Henri theil's contributions to economics and econometrics: econometric theory and methodology. Springer, Dordrecht, The Netherlands. https://doi.org/10.1007/978-94-011-2546-8_20

Tukey JW 1977. Exploratory Data Analysis. Addison-Wesley, Reading, UK.

Wald A. 1940. The fitting of straight lines if both variables are subject to error. The Annals of Mathematical Statistics 11: 284-300. https://doi.org/10.1214/aoms/1177731868

Walters EJ, Morrell CH, Auer RE. 2006. An investigation of the median-median method of linear regression. Journal of Statistics Education 14: 1-22 https://doi.org/10.1080/10691898.2006.11910582

Wilcox R. 1998. A note on the Theil-Sen regression estimator when the regressor is random and the error term is heteroscedastic. Biometrical Journal 40: 261-268. https://doi.org/10.1002/(SICI)1521-4036(199807)40:3%3C261::AID-BIMJ261%3E3.0.CO;2-V

Wilcox RR. 2021. Introduction to Robust Estimation and Hypothesis Testing. 5th. Elsevier, Amsterdam, The Netherlands. https://doi.org/10.1016/C2019-0-01225-3

Zacharias S, Heatwole CD, Coakley CW. 1996. Robust quantitative techniques for validating pesticide transport models. Journal of the American Society of Agricultural and Engineering 39: 47-54