International Journal of Statistics and Probability; Vol. 12, No. 1; January 2023
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Regression: Identifying Good and Bad Leverage Points

Rand R. Wilcox[1], Lai Xu[1]

[1] Dept of Psychology, University of Southern California, USA Correspondence: Rand R. Wilcox, Dept of Psychology, University of Southern California, USA

Received: October 17, 2022 Accepted: December 1, 2022 Online Published: December 20, 2022

doi:10.5539/ijsp.v12n1p1 URL: https://doi.org/10.5539/ijsp.v12n1p1

Abstract

When dealing with regression, a well known concern is that a few bad leverage points can result in a poor fit to the bulk of the data. This is the case even when using various robust estimators, which is known as contamination bias. Currently, a relatively effective method for detecting bad leverage points is based in part on the least median of squares regression estimator. This note suggests a modification of this method that is better able to detect bad leverage points. The modification also provides a substantially better technique for dealing with contamination bias.

Keyword: linear models, outliers, robust regression

1. Introduction

Let $(Y_i, \mathbf{X}_i)$, $i = 1, \ldots, n$, denote a random sample from some unknown multivariate distribution where $\mathbf{X}_i$ is a vector of length p. Consider the usual linear regression model where it is assumed that the typical value of Y, given $\mathbf{X}$, is

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p, \tag{1}$$

where $\beta_0, \ldots, \beta_p$ are unknown parameters. Let $r_1, \ldots, r_n$ denote the usual residuals based on some estimate of the unknown parameters $\beta_0, \ldots, \beta_p$. The point $(Y_i, \mathbf{X}_i)$ is a regression outlier if $|r_i|$ is unusually large. The point $(Y_i, \mathbf{X}_i)$ is called a leverage point if $\mathbf{X}_i$ is an outlier among $\mathbf{X}_1, \ldots, \mathbf{X}_n$.

There are two types of leverage points: good and bad. Roughly, a good leverage point is one that is reasonably consistent with (1) when (1) reflects the true regression line for the bulk of the data. A bad leverage point is a leverage point that is a regression outlier. A serious concern is that bad leverage points can result in a poor fit to the bulk of the points resulting in a poor reflection of the nature of the association among most of the participants. This is not surprising when dealing with the ordinary least squares (OLS) regression estimator because it has a breakdown point of only $1/n$, where the breakdown point refers to the minimum proportion of points that must be altered to make an estimate arbitrarily large or small.

A seemingly simple solution is to replace the OLS estimator with a robust estimator having a reasonably high breakdown point. But even when using an estimator with a breakdown point of .5, the highest possible value, Wilcox (2022) demonstrates that the better-known robust estimators can still be seriously impacted by bad leverage points. That is, a high breakdown point guarantees that a few outliers cannot result in estimates that are arbitrarily large or small. But it does not guarantee that a few outliers cannot have a substantial impact on the estimates of the slopes resulting in a misleading understanding of the association. This property is known as contamination bias. A simple solution would be to remove all leverage points, but good leverage points result in good fit to the bulk of the data and have the added advantage of lowering the standard error. A more satisfactory approach is to remove bad leverage points and retain good leverage points.

A classic approach to identifying leverage points is via Mahalanobis distance, but it suffers from masking, meaning that the very presence of outliers causes them to be missed (e.g., Rousseeuw & van Zomeren,1990). This result stems from the fact that the sample mean and covariance matrix have a breakdown point of only $1/n$. A well-known alternative is based on the so-called hat matrix (e.g., Montgomery et al., 2012, p. 213). But Rousseeuw and van Zomeren (1990) point out that this approach also suffers from masking, basically because the hat matrix has a breakdown point of only $2/n$.

A major advance was derived by Rousseeuw and van Zomeren (1990), which is called method RZ henceforth. The method is based in part on the least median of squares regression estimator, which has a breakdown point of .5. The goal here is to suggest a slight modification of method RZ that has the potential of substantially increasing the likelihood of detecting bad leverage points. In addition, the modification provides an improved method for dealing with the contamination bias associated with robust estimators that have a high breakdown.

The paper is organized as follows. Section 2 reviews method RZ and provides some background on why it can be unsatisfactory. Section 3 describes a proposed modification of method RZ. Section 4 reports simulations results on how well the modified method compares to method RZ. Section 5 illustrates the proposed method using data aimed at predicting the reading ability of children. The illustration demonstrates that the proposed modification can make an

important difference compared to using method RZ.

2. A Review of Method RZ

Method RZ begins by estimating the slopes and intercept with the least median of squares (LMS) estimator that was first proposed by Hampel (1975). That is, choose values for the parameters that minimize

$$\text{MED}(r_1^2, \ldots r_n^2), \tag{2}$$

the median of the squared residuals. This was the first regression estimator to achieve the maximum possible breakdown point, .5.

Let M_r be the median of $r_1^2, \ldots, r_n^2$, the squared residuals based on the LMS fit, and let

$$\hat{\tau} = 1.4826\left(1 + \frac{5}{n-p-1}\right)\sqrt{M_r}.$$

Method RZ declares the point $(Y_i, X_{i1}, \ldots, X_{ip})$ a regression outlier if

$$|r_i|/\hat{\tau} > 2.5. \tag{3}$$

Rousseeuw and van Zomeren (1990) discuss two approaches to detecting outliers among the explanatory data. The first uses a robust analog of Mahalanobis distance, which is based on robust measures of location and scatter that have a breakdown point of .5. Two such estimators were discussed: the minimum volume ellipsoid (MVE) estimator and the minimum covariance determinant (MCD) estimator. The MVE method searches for the half of the data that has the smallest volume. The mean and covariance matrix are computed based on this half of the data and the covariance matrix is rescaled to achieve consistency under normality. (The R function cov.mve in the R package MASS performs the calculations.) MCD searches instead for the half of the data for which the determinant of the covariance matrix is minimized. Here, the MCD estimator is computed using results in Hubert et al. (2012). Note that using an analog of Mahalanobis distance implicitly assumes that $\mathbf{X}$ has an elliptically contoured distribution.

Another approach to detecting leverage points, mentioned by Rousseeuw and van Zomeren, is based on a projection method, the basic idea of which stems from Donoho and Gasko (1992). Rousseeuw and van Zomeren conclude that in terms of detecting outliers, the choice between these two approaches is not important. Here, some results comparing the MCD method and an alternative projection method are included.

Here, when dealing with a single explanatory variable, $p = 1$, the well-known MAD-median rule is used to detect outliers. Let M denote the median of $X_1, \ldots, X_n$ and let MAD denote the median absolute deviation statistic. That is, MAD is the median of $|X_1 - M|, \ldots, |X_n - M|$, which has a breakdown point of .5. Then X_i is flagged as an outlier if

$$\frac{|X_i - M|}{\text{MAD}/.6745} > K, \tag{4}$$

where K is taken to be $\sqrt{\chi^2_{.975,1}}$, the square root of the .975 quantile of a chi-squared distribution with one degree of freedom. When dealing with a normal distribution, MAD/.6745 estimates the population standard deviation. Then X_i is declared a bad leverage point if both (4) and (3) are true.

To provide a glimpse of what motivated this paper, consider again $p = 1$ and the situation where X and Y are independent and both have normal distributions. First consider $n = 20$ and suppose three bad leverage points are added, namely (X, Y)=(3,3), (4,4) and (5,5). Figure 1 shows a boxplot of 3000 LMS estimates of the slope for this particular situation. As is evident, LMS is highly impacted by the regression outliers. That is, altering 13% of the data can substantially impact the estimate of the slope. The right boxplot is where $n = 30$ and the same bad leverage points are added. LMS improves as expected but concerns remain that the estimates tend to be much larger than zero.

3. The Proposed Modification

Before proceeding, it is noted that there are many robust regression estimators (e.g., Wilcox, 2022). No single estimator dominates in terms of efficiency. Here, the MM-estimator derived by Yohai (1987) is considered, which has excellent theoretical properties including a breakdown point of .5 and a relative efficiency of 95%. But this is not sufficient when the goal is to detect bad leverage points because, like many robust estimators, the MM-estimator can suffer from contamination bias (Wilcox, 2022, section 10.14.1.) Results in section 4 of this paper illustrate this concern. The proposed method deals with this issue.

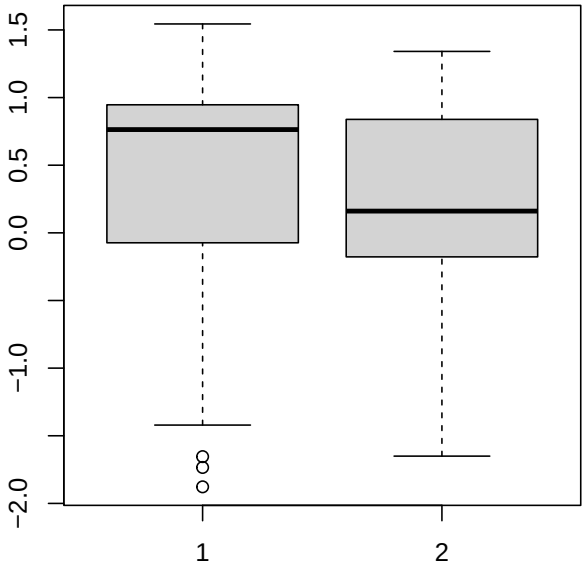

Figure 1. The left boxplot is based on 3000 LMS estimates of the slope, $n = 20$, when the true slope is zero and there three bad leverage points. The right boxplot is based on $n = 30$

It is noted that Gervini and Yohai (2002) suggested detecting bad leverage points by replacing the LMS estimator with an S-estimator. However, the S-estimator also suffers from contamination bias. In addition, Hössjer (1992) showed that S-estimators cannot achieve simultaneously both a high breakdown point and high efficiency under the normal model. Also, Davies (1993) reports results on the inherit instability of S-estimators.

The proposed modification is simple. First, remove all leverage points and estimate the slopes and intercepts using the remaining data. The focus here is on the LMS estimator and the MM-estimator. The basic idea is that bad leverage points cannot have an impact because all leverage points have been removed. Next, based on this fit, compute the residuals using all of the data yielding say $u_1, \ldots, u_n$. Next, check for outliers among $u_1, \ldots, u_n$ using the MAD-median rule. If u_i is an outlier and $(Y_i, \mathbf{X}_i)$ is a leverage point, decide that $(Y_i, \mathbf{X}_i)$ is a bad leverage point. When using this modified RZ method in conjunction with the LMS estimator, this will be called the RZ-LMS method henceforth. Using instead the MM-estimator is called the RZ-MM method.

When there are $p > 1$ explanatory variables, the modified versions of RZ use a projection method to detect leverage points that differs from the outlier detection method considered by Rousseeuw and van Zomeren (1990). There are in fact several related techniques that might be used (e.g, Wilcox, 2022, section 6.4). One advantage of this approach is that it does not assume that $\mathbf{X}$ has an elliptically contoured distribution. Moreover, it performs relatively well in terms of avoiding masking and swamping (declaring a point an outlier when in fact it is not all that unusual).

The complete details of the method used here are summarized in Wilcox (2022, section 6.4.9). For brevity, only an outline of the method is provided. The projection method begins by standardizing the marginal distributions. For each explanatory variable, subtract the median and divide this diffeerence by MAD/.6745. This is done so that the method is scale invariant. Next, consider any point, $\mathbf{X}_i$, which is assumed to be standardized for notational convenience. The method projects all of the data onto the line connecting $\mathbf{X}_i$ and the center of the data cloud. Any point among all n points is declared an outlier if its projected point is flagged an outlier among all n projections. This process is repeated for each i, $i = 1, \ldots, n$. Applying this method is easily done via the R function outpro in the R package WRS.

Table 1. Probability of detecting bad leverage points

n	RZ			RZ-LMS			RZ-MM		
	(3,3)	(4,4)	(5,5)	(3,3)	(4,4)	(5,5)	(3,3)	(4,4)	(5,5)
20	.213	.404	.445	.336	.695	.779	.401	.729	.856
30	.416	.627	.664	.502	.789	.823	.564	.891	.971
50	.622	.811	.841	.650	.858	.914	.761	.979	.996
75	.687	.881	.928	.729	.891	.945	.878	.995	.999

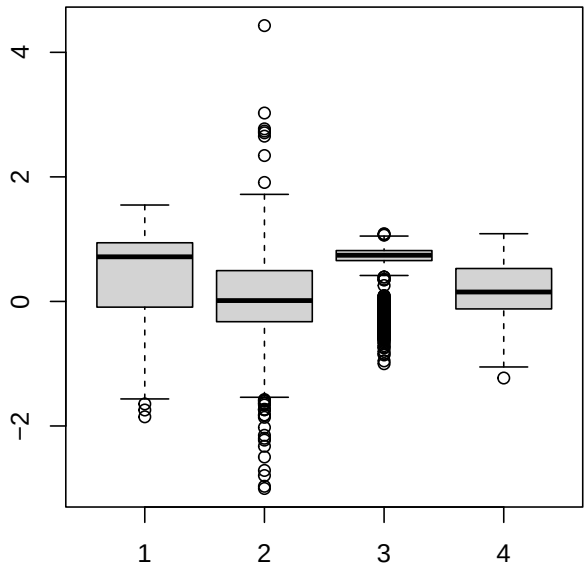

Figure 2. Boxplots based on 3000 estimates of the slope. 1 = LMS estimator, 2 = RZ-LMS, 3 = MM-estimator, 4 = RZ-MM, $n = 20$

4. Some Simulation Results

The first goal in the simulations is to determine the probability of detecting points that are in fact bad leverage points. First consider $p = 1$. Here, data are generated as described in section 2. That is, n values for X and Y were generated, where X and Y are independent. Then three bad leverage points were added: (X, Y)=(3,3), (4,4) and (5,5). The sample sizes are taken to be $n = 20$, 30, 50 and 75. Table 1 shows estimates of the probability that these three points are flagged as bad leverage points. This is done using method RZ, the RZ-LMS method and the RZ-MM methods. As is evident, the RZ-MM method dominates and in various situations offers a substantial advantage over the other two methods.

The next set of simulations is designed to provide some sense of the impact of removing bad leverage points on the estimates of the slope. The data were generated as done in Figure 1, in which case the slope is zero ignoring the leverage points. Figure 2 shows boxplots based 3000 estimates when $n = 20$. The left boxplot is based on the LMS estimator using all of the data. The next is based on the RZ-LMS estimator. The next boxplot is based on the MM-estimator using all of the data and the last is based on the RZ-MM method. Note that removing bad leverage points corrects the bias of the LMS estimator but at the expense of a much higher standard error. The same is true when using the MM-estimator. Put another way, failing to remove bad leverage points results in the MM-estimator routinely yielding a highly inaccurate estimate of the slope. That the MM-estimator has a smaller standard error than LMS is exactly what is expected based on known theoretical results previously mentioned. Figure 2 underscores the extent this is a practical issue.

Figure 3 is the same as Figure 2, only $n = 50$. Now, when using LMS, removing bad leverage points provides little

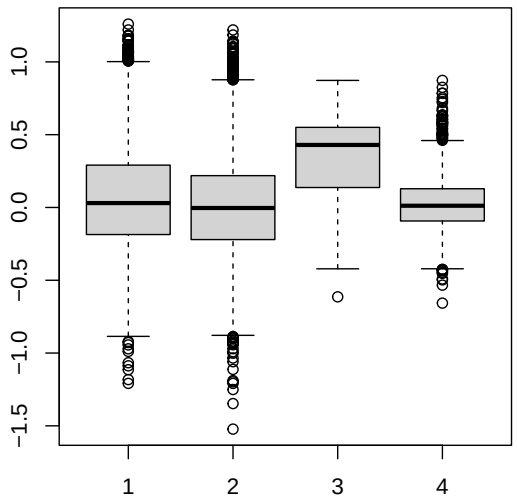

Figure 3. Same as Figure 2, only $n = 50$

Table 2. Probability of detecting bad leverage points, $p = 4$

n	RZ			RZ-LMS		
	(3,3,3,3)	(4,4,4,4)	(5,5,5,5)	(3,3,3,3)	(4,4,4,4)	(5,5,5,5)
20	.041	.041	.613	.613	.729	.775
30	.147	.147	.148	.677	.760	.806
50	.334	.346	.349	.683	.767	.815
75	.468	.512	.523	.692	.788	.835

improvement over using all of the data. The third boxplot illustrates that increasing the sample size did not correct the bias associated with the MM-estimator. This was expected based on results in Wilcox (2022). The last boxplot illustrates that in this case, removing bad leverage points corrects the bias of the MM-estimator and yields a smaller standard error.

The simulations in Table 1 were repeated with $p = 4$ explanatory variables. Now the three additional points were set equal to (3, 3, 3, 3), (4, 4, 4, 4) and (5, 5, 5, 5), respectively. Table 2 shows the results comparing RZ and RZ-LMS. The advantage of the RZ-LMS method over method RZ is even more striking versus $p = 1$. Note that for $n = 75$, the proportion of points that renders RZ relatively unsatisfactory is only $3/(75 + 3) = .038$.

The next set of simulations were based on the RZ-MM method. Now, one goal is to compare the approach of replacing the projection method for detecting outliers with the MCD method. Another goal is to compare the RZ-MM method when using the projection method for detecting outliers versus RZ-LMS. The results are in Table 3. For $n = 20$, MCD performs a bit better than the projection method. But otherwise the projection performed better despite generating data from a elliptically contoured distribution. Also, comparing the results in Table 2 to those in Table 3, RZ-MM (with the projection method) performs better than using the RZ-LMS.

5. An Illustration

Method RZ-MM is illustrated with data dealing with predicting reading ability in children stemming from an unpublished study by L. Doi. The explanatory variable is a measure of speeded naming for digits (RAN1T1) and the dependent variable is a measure of ability to identify words (WWISST2). The sample size is $n = 81$.

Figure 4 shows a scatterplot of the data. The solid line is the LMS regression line using all of the data. The dashed line is the MM regression line with bad leverage points removed via method RZ. The dotted line is based on the MM-estimator with bad leverage points removed using the RZ-MM method. Method RZ identifies a single bad leverage point in the lower left corner at (42, 85). The RZ-MM method identifies seven bad leverage points that include the six points on the

Table 3. Estimated probability of detecting bad leverage points, using RZ-MM-PRO and RZ-MM-MCD, $p = 4$

n	RZ-MM-PRO			Modified RZ-MM-MCD		
	(3,3,3,3)	(4,4,4,4)	(5,5,5,5)	(3,3,3,3)	(4,4,4,4)	(5,5,5,5)
20	.615	.730	.792	.646	.758	.821
30	.717	.836	.893	.695	.821	.876
50	.770	.904	.953	.747	.888	.933
75	.819	.942	.978	.806	.933	.973

right of Figure 4. This vividly illustrates the fact that the choice of method can make a substantial difference.

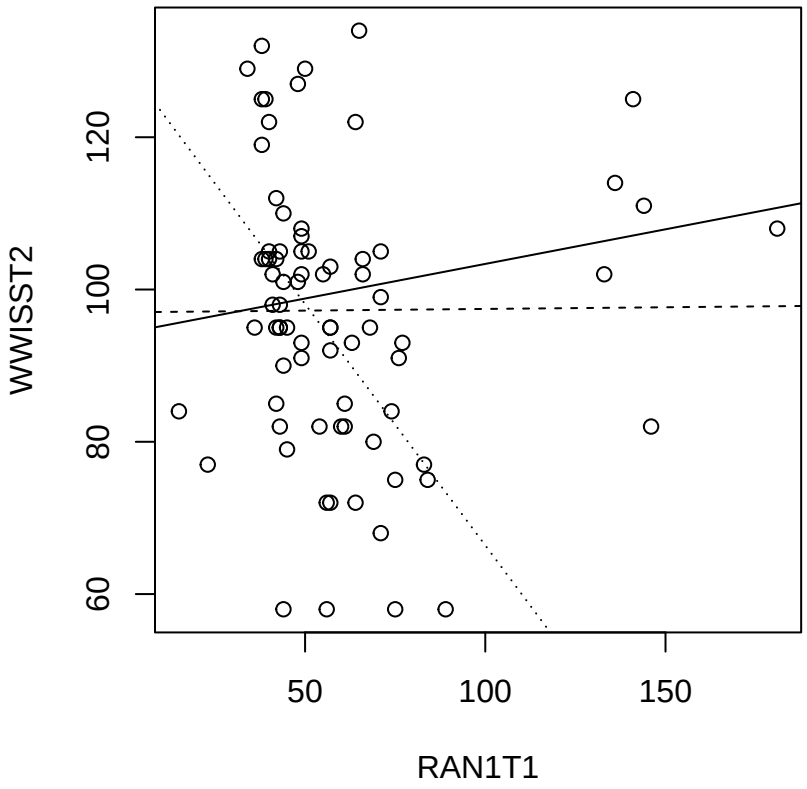

Figure 4. Scatterplot of the reading data. The solid is the LMS estimate using all of the data. The dashed line is LMS ignoring bad leverage points identified by RZ. The dotted line is the MM regression line ignoring bad leverage points based on the RZ-MM

To add perspective, Figure 5 shows a plot of the data with the bad leverage points, based on RZ-MM, removed. The solid line is a non-parametric estimate of the regression line based on the smoother derived by Cleveland (1979). The dashed line is the MM regression line. Both methods are in close agreement about the nature of the association, and as is evident they paint a decidedly different understanding about the nature of the association compared to method RZ. Also, the smoother supports the assumption that the regression line is reasonably straight. The hypothesis that the slope of the MM regression is zero was tested using the percentile bootstrap method in Wilcox (2022, section 11.1.3). This method allows the error term to be heteroscedastic. The p-value is less than .001. The same result was obtained when the MM-estimator is replaced by the robust estimator derived inTheil (1950) and Sen (1968). In contrast, if method RZ is used to identify bad leverage points, the p-value is .992 using the MM-estimator.

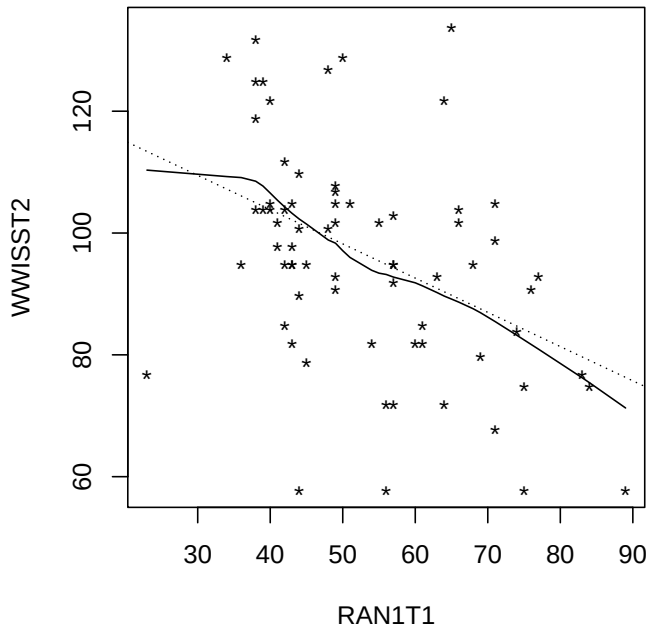

Figure 5. Scatterplot of the reading data with bad leverage points removed. The solid line is based on Cleveland's smoother. The dotted line is the MM regression line

6. Concluding Remarks

Despite the results reported here, perhaps there are situations where method RZ provides a substantial advantage over the RZ-MM method. The only certainty at the moment is that the RZ-MM method can provide a substantial advantage improvement over RZ in some situations.

There are several variations of the RZ-MM method. For example, there are several alternative methods for detecting multivariate outliers that have certain connections with the projection method. In addition, there are alternative robust regression estimators that provide a reasonable alternative to the MM-estimator. One difficulty is that no single robust estimator dominates in terms of efficiency. For example, there are situations where the MM-estimator has a smaller standard error than other robust estimators that might be used. But there are situations where other robust estimators have a smaller standard error than the MM-estimator. Summaries of these methods and details of how they compare are described in Wilcox (2022).

Finally, R functions for applying the methods used here are stored in the file Rallfun-v40.txt, which can be downloaded from https://osf.io/xhe8u/ The R function reglev.gen applies the RZ-MM method. It returns which points are leverage points as well as which, if any, are bad leverage points. The argument regfun can be used to control which robust estimator is used. The R function reg.reglev fits a regression line after any bad leverage points are removed. Again the argument regfun can be used to control which robust estimator is used. Setting the argument GEN=FALSE, method RZ is used instead. The R function regblp.ci computes a confidence interval for the slopes and intercept after bad leverage points are removed.

References

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association, 74*, 829–836. https://doi.org/10.2307/2286407

Davies, P. L. (1993). Aspects of robust linear regression. *Annals of Statistics, 21*, 1843–1899.

Donoho, D. L., & Gasko, M. (1992). Breakdown properties of the location estimates based on halfspace depth and projected outlyingness. *Annals of Statistics, 20*, 1803–1827.

Gervini, D., & Yohai, V. (2002). A class of robust and fully efficient regression estimators. *Annals of Statistics, 30*, 583–616.

Hampel, F. R. (1975). Beyond location parameters: robust concepts and methods (with discussion). *Bulletin of the ISI, 46*, 375–391.

Hössjer, O. (1992). On the optimality of S-estimators. *Statistics and Probability Letters, 14*, 413–419.

Hubert, M., Rousseeuw, P. J., & Verdonck, T. (2012). A deterministic algorithm for robust location and scatter. *Journal of Computational and Graphical Statistics, 21*, 618–637. https://doi.org/10.1080/10618600.2012.672100

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to Linear Regression. Analysis*. (5th Ed.). New York: Wiley.

Rousseeuw, P. J., & van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association, 85*, 633–639. https://doi.org/10.2307/2289995

Sen, P. K. (1968). Estimate of the regression coefficient based on Kendall's tau. *Journal of the American Statistical Association, 63*, 1379–1389.

Theil, H. (1950). A rank-invariant method of linear and polynomial regression analysis. *Indagationes Mathematicae, 12*, 85–91.

Wilcox, R. R. (2022). *Introduction to Robust Estimation and Hypothesis Testing*. 5th Edition. San Diego, CA: Academic Press.

Yohai, V. J. (1987). High breakdown point and high efficiency robust estimates for regression. *Annals of Statistics, 15*, 642–656.