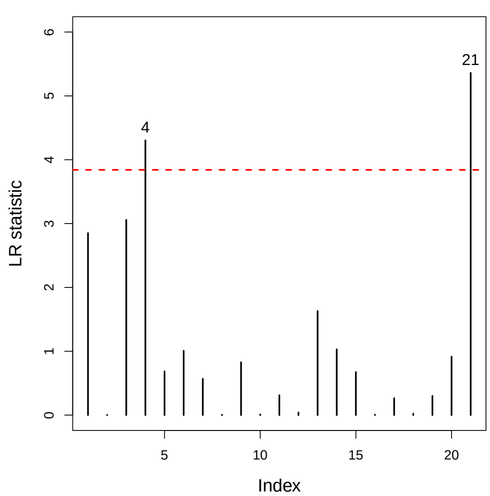
SUPPLEMENT TO "ON THE MEAN-SHIFT OUTLIER MODEL FOR LAD REGRESSION"

FELIPE OSORIO

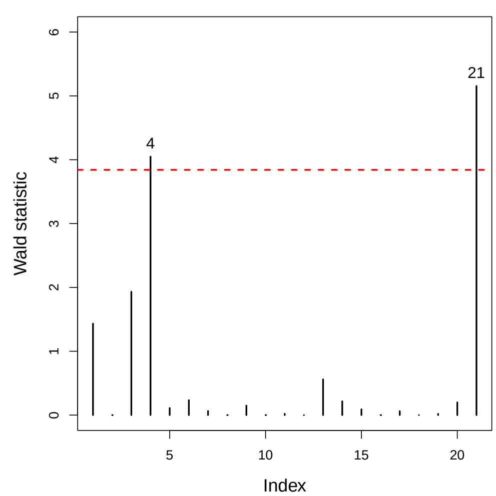
ABSTRACT. In this supplementary material, we present aditional results obtained from the analysis of two real datasets and supplementary figures related to the real-life examples discussed in Section 4 of the main paper.

1. REAL DATA EXAMPLES

Brownlee's Stack loss data. The Brownlee's dataset (Brownlee, 1965) corresponds to 21 days of operation of an ammonium oxidation plant as a stage in nitric acid production and has been studied by numerous authors. In particular, the influence diagnostic in LAD regression has been analyzed in Sun and Wei (2004). Table 1 presents the estimation of the regression coefficients considering the stackloss (y), ten times the percentage of unconverted ammonium that escapes from the absorption column as a function of the following variables: Air flow (x_1), which measures the rate of operation of the plant; Temperature (x_2), which corresponds to the inlet temperature of water for cooling in an absorption tower; and Acid (x_3), which is proportional to the acid concentration in the tower. Table 1 also presents the parameter estimates with some cases removed.



(a)



(b)

FIGURE 1. Test statistics of the null hypothesis $H_0 : \gamma = 0$ for Brownlee's stack loss data: (a) likelihood ratio LR_i and (b) Wald W_i statistics.

Figure 1 displays the LR and Wald statistics for testing the hypothesis $H_0 : \gamma = 0$ which allows observations 4 and 21 to be identified as outliers. We should note that, although these observations have a small impact on the regression coefficient estimates, they have a significant effect on the scale parameter estimates. Note that these observations are also highlighted in the residual plots presented in Figure 2 (a)

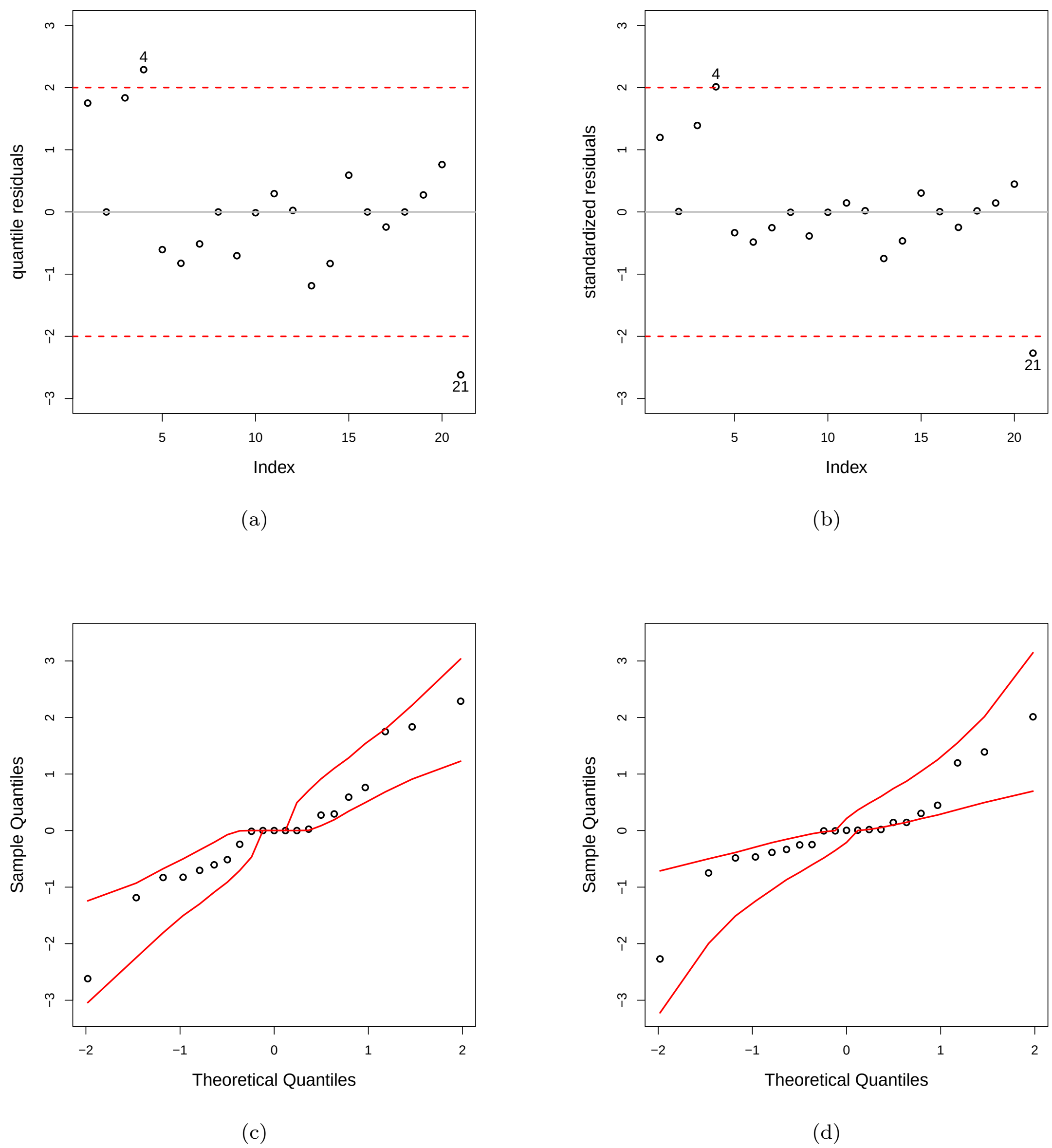


FIGURE 2. Plot of residuals for Brownlee's stack loss data: Index plot of (a) randomized quantile residuals, (b) standardized residuals, simulated envelopes for randomized quantile residuals (c), and standardized residuals (d).

and (b). These results are consistent with the findings presented in Flores (2015), who, using an extension of the concept of leverage, identified the role of observations 4 and 21. Furthermore, our results complement those presented in Sun and Wei (2004), who identified observation 17 as influential on the estimation of the coefficients. Despite the existence of extreme observations, the QQ-plot with simulated envelope presented in Figure 2 (c) and (d) shows that the LAD fit appears to be adequate.

TABLE 1. Parameter estimates and standard errors (in parenthesis) for Brownlee's stack loss data.

Deleted observations	Intercept β_0	Air flow β_1	Temperature β_2	Acid β_3	scale parameter ϕ	λ
none	-39.6899	0.8319	0.5739	-0.0609	2.8339	3.5529
	(13.0314)	(0.1477)	(0.4032)	(0.1712)		
2	-39.6940	0.8297	0.5776	-0.0603	2.9752	3.4433
	(12.6317)	(0.1519)	(0.3922)	(0.1704)		
3	-39.6519	0.8304	0.5809	-0.0621	2.5914	2.9333
	(10.8313)	(0.1255)	(0.3328)	(0.1414)		
4	-39.9865	0.8347	0.5637	-0.0569	2.4350	2.8997
	(10.6544)	(0.1230)	(0.3437)	(0.1397)		
17	-36.4245	0.7991	0.6962	-0.1057	2.9423	3.2325
	(14.9725)	(0.1345)	(0.3688)	(0.1907)		
21	-39.9864	0.8347	0.5637	-0.0569	2.3022	3.2062
	(11.8448)	(0.1483)	(0.4056)	(0.1554)		

TABLE 2. Percentage of relative change in parameter estimates for Brownlee's stack loss data.

Deleted observations	Intercept β_0	Air flow β_1	Temperature β_2	Acid β_3	scale parameter ϕ	λ
2	0.0104	-0.2576	0.6400	-0.8621	4.9857	-3.0845
3	-0.0957	-0.1812	1.2229	1.9955	-8.5581	-17.4403
4	0.7473	0.3371	-1.7820	-6.5041	-14.0753	-18.3846
17	-8.2271	-3.9462	21.3122	73.5849	3.8253	-9.0189
21	0.7473	0.3371	-1.7820	-6.5041	-18.7613	-9.7593

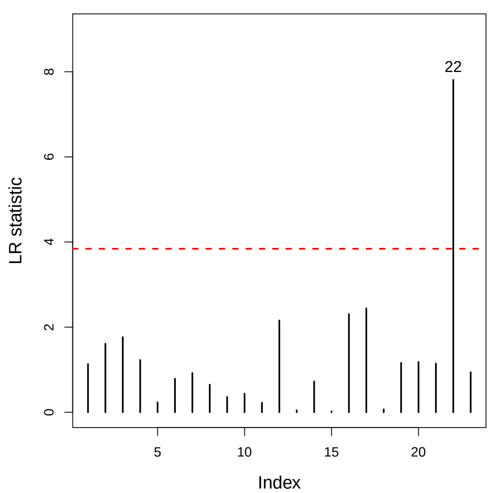
Aircraft data. This dataset is presented in (Rousseeuw and Leroy, 1987, pp. 154), and the aim is to model the cost of 23 single-engine aircraft (in unit of $100,000) as a function of the following explanatory variables: aspect ratio (x_1), lift-to-drag ratio (x_2), weight of the plane, in pounds (x_3), and maximal thrust (x_4). This dataset is available in the india R package and has been studied, using LAD regression, in Flores (2015), who identified cases 14 and 22 as extreme observations despite being different in nature. Indeed, observations 14 and 8 have a large effect on the estimation of the regression coefficients (see Table 4). The methodology proposed in this work identifies observation 22 as an outlier (see Figure 3 as well as Figure 4 (a) and (b)), a conclusion that is consistent with what is discussed in Flores (2015). Interestingly, although observation 22 has a moderate effect on the estimation of the coefficients, it has an important effect on the estimation of the scale parameter ϕ (Table 4). The detection of observation 22 as an outlier casts additional insight into the analysis developed in Flores (2015), which indicated that the role of observation 22 is somewhat unclear.

Table 3. Parameter estimates and standard errors (in parenthesis) for Aircraft data.

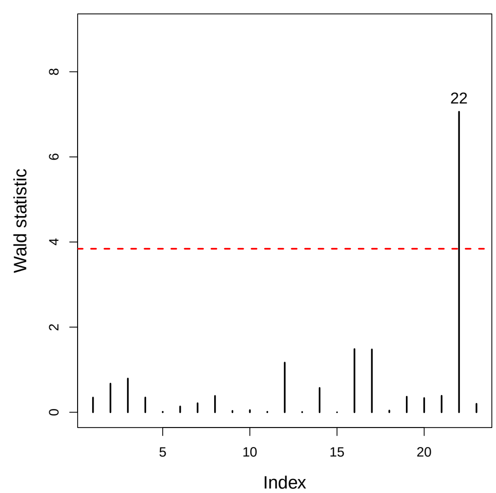
Deleted observations	Intercept β_0	Aspect ratio β_1	Lift-to-drag β_2	Weight β_3	Thrust	scale parameter ϕ	λ
none	1.9511	-3.0466	1.4736	0.0022	-0.0011	7.3612	10.5390
	(12.6828)	(2.2104)	(1.4879)	(0.0006)	(0.0006)		
8	1.6720	-4.7491	2.6527	0.0033	-0.0020	7.4759	13.1964
	(16.3237)	(2.7712)	(1.9408)	(0.0008)	(0.0008)		
14	2.4923	-3.9594	4.5387	0.0020	-0.0009	7.4502	11.9935
	(14.4340)	(2.6273)	(4.4304)	(0.0007)	(0.0007)		
22	9.3765	-4.1109	1.6351	0.0020	-0.0011	5.0505	10.7719
	(13.3904)	(2.2711)	(1.5405)	(0.0008)	(0.0007)		

Table 4. Percentage of relative change in parameter estimates for Aircraft data.

Deleted observations	Intercept β_0	Aspect ratio β_1	Lift-to-drag β_2	Weight β_3	Thrust β_4	scale parameter ϕ	λ
8	-14.3055	55.8792	80.0147	46.3897	102.4277	1.5586	25.2147
14	27.7405	29.9589	208.0016	-12.3773	-2.0541	1.2096	13.8009
22	380.5740	34.9341	10.9628	-8.5979	10.7095	-31.3906	2.2094



(a)



(b)

Figure 3. Test statistics of the null hypothesis $H_0 : \gamma = 0$ for Aircraft data: (a) likelihood ratio LR_i and (b) Wald W_i statistics.

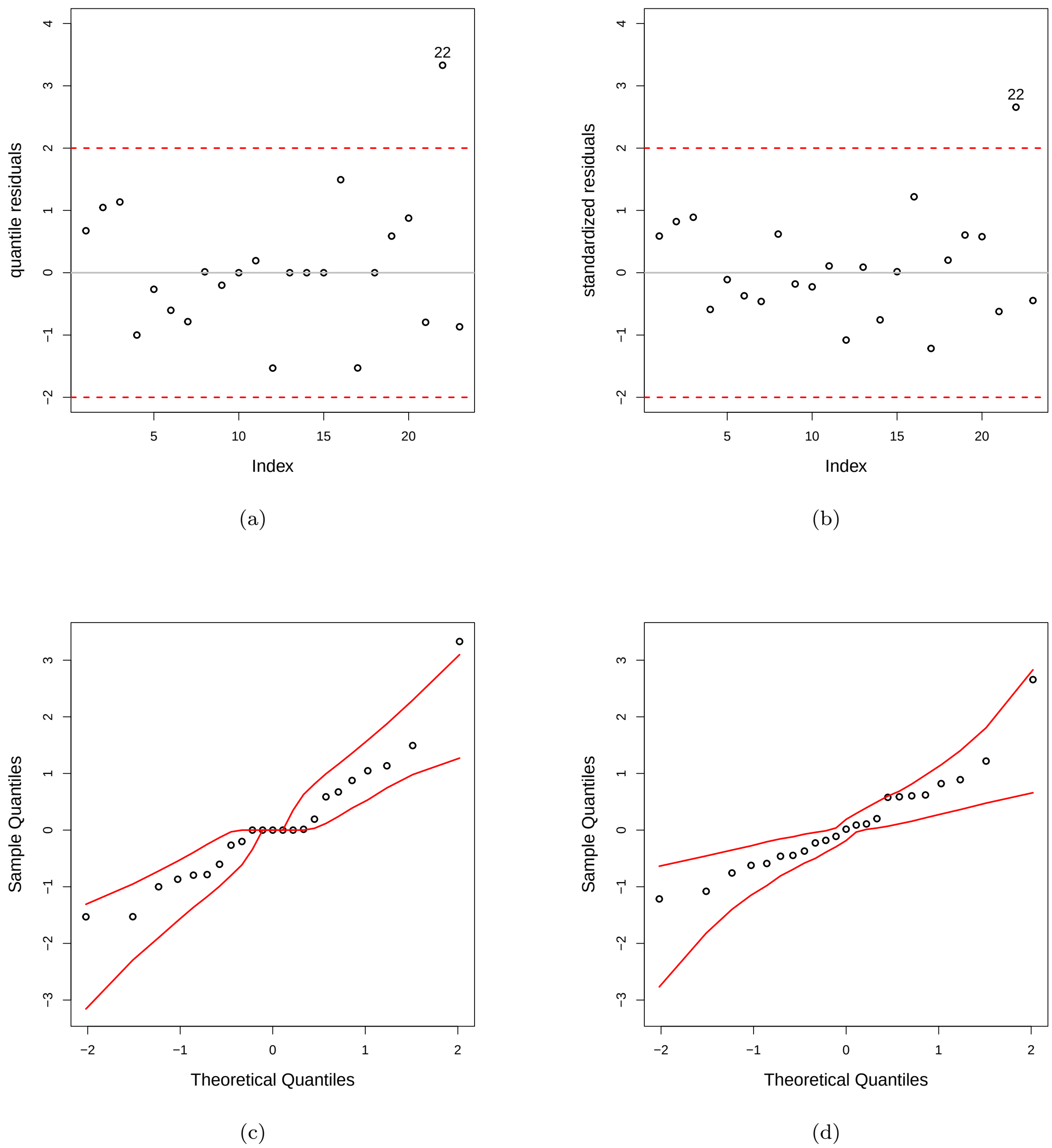
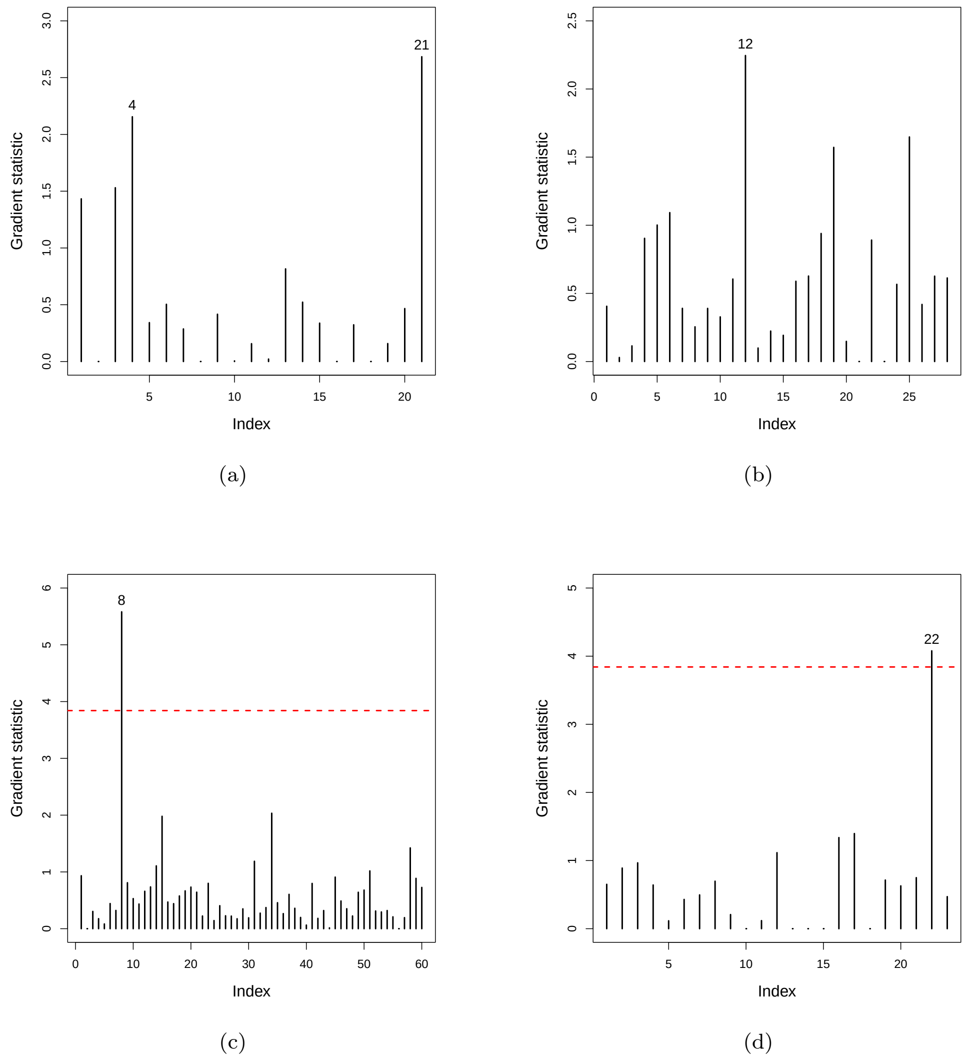


FIGURE 4. Plot of residuals for Aircraft data: Index plot of (a) randomized quantile residuals, (b) standardized residuals, simulated envelopes for randomized quantile residuals (c), and standardized residuals (d).

2. Additional results



(a) (b) (c) (d)

Figure 5. Index plot of gradient statistic T_i: (a) Brownlee's stack loss, (b) Skeena river sockeye salmon, (c) Martin Marietta and (d) Aircraft datasets.

References

Brownlee, K.A. (1965). *Statistical Theory and Methodology in Science and Engineering, 2nd Edition*. Wiley, New York.

Flores, S. (2015). Sharp non-asymptotic performance bounds for l_1 and Huber robust regression estimators. *TEST* **24**, 796-812.

Rousseeuw, P.J., Leroy, A.M. (1987). *Robust Regression and Outlier Detection*. Wiley, New York.

Sun, R.B., Wei, B.C. (2004). On influence assessment for LAD regression. *Statistics and Probability Letters* **67**, 97-110.

Orcid ID: https://orcid.org/0000-0002-4675-5201
Email address: faosorios.stat@gmail.com