Robust Regression Methods for the Analysis of Unreplicated Factorials

Román de la Vara and Víctor M. Aguirre *

Abstract—**The existing methods for analyzing unreplicated factorials that do not contemplate the possibility of outliers in experimental data have a poor performance for detecting the active effects when that possibility becomes a reality. We propose an iterative procedure based on robust regression which has a good performance in the presence or absence of contaminated data.**

Keywords: Active Effects, Atypical Data, Robust Regression, MM Estimator,L1 Regression

1 Introduction

In this paper we propose a new method that considers the possibility of outliers, which is based on robust regression techniques, particularly on MM-estimation [14]. According to a simulation study the new method has a better performance compared with existing methods.

The paper is organized as follows: in Section 2 we describe briefly the two existing methods that consider the possibility of faulty observations in factorial experiments namely [6] and [1]. In Section 3 we introduce some robust regression techniques and terminology in order to explain MM-estimation. In Section 4 we describe the new method. In Section 5 we illustrate the proposed method using an example where the outlier is not evident, the example is also analyzed with the two existing methods that consider outliers and two methods that do not consider that possibility. Except for the robust method, the other approaches fail to detect any of the significant effects. In Section 6 we use Monte Carlo simulation to compare the new method with the methods mentioned in Sections 2 and 5. The comparison is made under a common ground since all of the methods are calibrated to have an experiment-wise error rate (EER) of 5%. Finally, the conclusions are given in Section 6

*Román de la Vara, Center for Research in Math (CIMAT), Quality Engineering Department, Guanjuato, GTO, 36000, México. Email: delavara@cimat.mx. Víctor M. Aguirre, Autonomous Technology Institute of México (ITAM), Statistics Department, México, DF 01000, México. Email: aguirre@itam.mx. Professor Victor Aguirre was partially financed by Asociación Mexicana de la Cultura A. C. He conducted part of this research while on sabbatical leave at CIMAT. This version: February 17, 2009.

2 The Existing Methods

In this section we describe briefly the two existing methods that consider the possibility of atypical observations in experimental data.

2.1 Ranks Method

This method was introduced in [1] and then further studied in [2]. Let $\mathbf{y}$ be the vector of observations and consider its rank transformation $R(\mathbf{y})$. If there are ties then assign to each observation the average of the ranks that would correspond to them when there are no ties. Let $\mathbf{X}$ the $n \times n$ matrix containing the $n-1$ contrasts and a first column of ones. The effects based on ranks T_i^R $(i = 1, 2, ..., n)$ are given by

$$\begin{bmatrix} T_0^R \\ T_1^R/2 \\ \vdots \\ T_{n-1}^R/2 \end{bmatrix} = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t R(\mathbf{y}).$$

Let $T_{(i)}^R$ $(i = 1, 2, ..., n-1)$ be the ordered effects. A Shapiro Wilks type test for normality is applied to the effects T_i, where the test statistic, see [2] is given by .

$$W = \frac{\left(\sum_{i=1}^n m_i T_{(i)}^R\right)^2}{\sum_{i=1}^n m_i^2 \sum_{i=1}^n \left(T_{(i)}^R - \overline{T}\right)^2}$$

where: $m_i = \Phi^{-1}(p_i)$, Φ^{-1} is the inverse cumulative normal distribution, and p_i are probabilities spread in the zero-one interval, see details in [3]. The same reference gives the formula for p-value of the test. The test is rejected if the computed p-value is lower than $\alpha = 0.034$, where this critical value was obtained by simulation (see Section 6) in order to have an experimentwise error rate equal to 5%.

The rejection of the normality test implies that there is at least one active effect in the experiment. Significant effects are identified with the so called fourth-spread test that declares as active the effects falling outside the interval $[-2d_F, +2d_F]$, where d_F is the estimated interquartile range of the effects, see [2].

2.2 Bayesian Method

This method is reported in [6] which is an extension of [5]. Basically the method calculates for each effect the posterior probability of being an active effect and for each observation the posterior probability of being an outlier. Large posterior probabilities, say greater than p_c (where its value will be determined later in section 6) correspond to active effects or outlier observations. In the 16 runs factorial experiment there are $2^{15} \times 2^{16}$ possible combinations of active effects and outlying observations, a enormous amount of work is required for the direct calculation of the posterior probabilities. For this reason [6] proposed an iterative procedure for approximating these probabilities.

Let $a_{(r_1,r_2)}$ be the event that one particular combination of r_1 active effects and r_2 outlier observations occurs. The posterior probability of this event is denoted by $p\left(a_{(r_1,r_2)}\left|\mathbf{y}\right.\right)$, the formula is given [6]. The posterior probability p_i that a particular effect i is active is then

$$p_i = \sum_{(r_1,r_2):\ i\ active} p\left(a_{(r_1,r_2)}\left|\mathbf{y}\right.\right)$$

and the posterior probability q_j that the observation j is faulty is

$$q_j = \sum_{(r_1,r_2):\ j\ outlier} p\left(a_{(r_1,r_2)}\left|\mathbf{y}\right.\right),$$

The iterative procedure is as follows: initially compute the posterior probabilities that each effect is active assuming that there are no atypical data, and then in the second step, taking into account the significant effects detected in the first step, compute the posterior probabilities that each data point is atypical. In the third step, assuming the atypical data found in the second step, compute the posterior probabilities for each effect of being active, and so on . See [8] for an algorithm programming this method. This procedure generally converges in three to six steps.

This method requires the prior determination of the following parameters: α_1 the probability of active effect, α_2 the probability of an outlier observation, γ an expansion factor of the error standard deviation due to an active effect, and k the expansion factor for outliers. The prior values suggested by [6] are $\alpha_1 = 0.2$, $\alpha_2 = 0.05$, $\gamma = 2.5$, $k = 5$, but of course one can try different values. In this paper we calibrated the Bayesian method using the recommended values. The paper [6] also suggests to declare an effect as active if the posterior probability is greater than 0.5, but in the simulation study (see Section 6) we found that this critical probability must be $p_c = 0.91$ for an experimentwise error rate equal to 5%.

3 Robust Regression Techniques

The proposal is based on robust regression known as MM-estimation that was introduced in [14]. We prefer MM-estimation over other robust techniques because we want to obtain estimators with high breakdown point and high efficiency at the same time, see [12]. The former property protects against contaminated data and the later guarantees an estimator variance comparable to the variance of the least squares estimator under normal error data.

Let $r_i(\theta) = y_i - \widehat{y}_i(\theta)$, $i = 1, 2, ..., n$, minimizing the function $\rho(r) = |r|$ gives the $\mathbf{L}_1$ regression estimator, which is robust against response outliers, see [11], page 10, but still it has a breakdown equal $1/n$ because it is affected by leverage points. Since in experimental design we deal mainly with response outliers, $\mathbf{L}_1$ regression estimator will be useful in the first step of our proposed method for detecting active effects.

The function ρ that we use satisfies the conditions: C1. It is symmetric and continuously differentiable and $\rho(0) = 0$; C2. There exists $c > 0$ such that ρ is strictly increasing on $[0, c]$ and constant on $[0, \infty]$. In order to obtain robust estimates ρ increases more slowly than the quadratic function.

S-estimates were introduced in [14]. They are defined similar to M-estimates, but they minimize in an implicit way the residuals dispersion $S(\theta)$,which is a solution to the equation $(1/n)\sum \rho(r_i/S) = K$, where the ρ function satisfies the same conditions for M-estimates and $K = E_\Phi[\rho]$, where Φ is the standard normal distribution. [11] show that for C3: $K/\rho(c) = 0.5$ the breakdown point for S-estimates is 50% but its efficiency is hardly 28.7% (see Table 19 on page 142 of [14]).

Following this line, [14] proposed the MM-estimates that combine high breakdown point with high efficiency (see also [15]). He defines MM-estimates in three stages:

1. Take an initial estimator $\widehat{\boldsymbol{\theta}}_0$ with high breakdown point, possibly 0.5.
2. Compute the residuals $r_i = y_i - \widehat{\boldsymbol{\theta}}_0' \mathbf{x}_i$ $(i = 1, 2, \ldots, n)$ and compute the S-estimate which is the solution of $(1/n)\sum \rho_0(r_i/S) = K$ using a function ρ_0 satisfying the assumptions C1-C3.
3. Let ρ_1 be another function satisfying assumptions C1-C3 such that $\rho_1(u) \leq \rho_0(u)$; $\sup\rho_1(u) = \sup\rho_0(u) = \rho(c)$. Then the MM-estimate $\widehat{\boldsymbol{\theta}}_1$ is defined as any solution that minimizes $(1/n)\times \sum \rho_1(r_i/S)$ which verifies $S(\widehat{\boldsymbol{\theta}}_1) \leq S(\widehat{\boldsymbol{\theta}}_0)$.

We can choose the functions ρ_0 and ρ_1 as follows: take a function ρ satisfying conditions C1-C3 and let $0 < c_0 < c_1$. Define $\rho_0(r) = \rho(r/c_0)$ and $\rho_1(r) = \rho(r/c_1)$. Choose

c_0 so that $K/\rho(c_0) = 0.5$ which determines a breakpoint of 50% and c_1 determines the asymptotic efficiency, [14].

For example, in our method we consider the Tukey´s bisquare function

$$\psi_c(u) = \begin{cases} \frac{u}{c}\left[\left(\frac{u}{c}\right)^2 - 1\right]^2 & if \quad |u| \le c \\ 0 & if \quad |u| > c, \end{cases}$$

where c is the tuning constant for high breakdown point and efficiency. The corresponding ρ function is obtained by integrating ψ_c and dividing by $c/6$, which is

$$\rho_c(u) = \begin{cases} \left(\frac{u}{c}\right)^6 - 3\left(\frac{u}{c}\right)^4 + 3\left(\frac{u}{c}\right)^2 & if \quad |u| < c \\ 1 & if \quad |u| \ge c. \end{cases}$$

Taking $c_0 = 1.548$ the resulting estimate has a breakdown value of 50% and choosing $c_1 = 4.687$ has an efficiency of 95% with the normal distribution. For an efficiency of 99.3% the tuning constant value is $c_1 = 7.7$.

4 The Proposed Robust Method

In unreplicated factorial experiments with 16 runs the saturated model has 15 effects, that is, including the constant term the number of parameters is equal to the number of observations. In order to achieve robustness, typically robust methods assume that the number observations is at least two times the number of parameters, and this is far from being the case in unreplicated experiments. Therefore we propose an iterative method for estimating the 15 effects in the 16 runs factorial using MM estimation. The proposed method consists of the following three steps:

1. Base Model. From 15 effects 1365 models with 4 terms can be constructed. The base model is selected from this subset according to hierarchy principle: ignoring the quadruple interaction, we consider the models where the presence of double or triple interactions implies the inclusion of some simple effects that compound these interactions; additionally the models include at most one triple interaction. Thus, for a 16 runs factorial experiment and 4 independent terms there are 165 models in the subset of interest. $\mathbf{L}_1$ regression is applied to each model and the sum of absolute residuals is computed. The base model minimizes this sum. We suggest $\mathbf{L}_1$ regression because this regression is robust against outliers on the response, see [11], page 10.

2. MM estimation. The base model is iteratively augmented with a fifth term and it is fitted by using MM estimation, where the fifth term represents each time one of the 11 effects not present in the base model. Thus after 11 iterations we complete the set of 15 robust estimated coefficients (effects). Because it is an iterative application of MM-estimation a high efficiency of 99.3% is recommended for good performance in the normal error case. This efficiency correspond to $c_1 = 7.7$ in the last stage of MM-estimation procedure.

We apply MM-estimation using the procedure *rlm* in the R system loading the *MASS* package. $\mathbf{L}_1$ regression is part of the procedure *rq* in the quantreg package (quantile regression). In MM-estimation the initial and final scale are selected by an S-estimator with Tukey´s biweight function with tuning constant $c_0 = 1.548$. For the initial estimate the sample size was $psamp = 7$ (see details in *rlm* documentation). The final estimator is an M-estimator with Tukey´s biweight function and constant $c_1 = 7.7$ for an efficiency of 99.3% in each iteration.

3. Identifying Active Effects. The non sequential variant of [3] method is applied to the estimated robust effects from step 2. As was explained in Section 2 this variant consists in a normality test followed by an outlier test, as was proposed by [3], but both tests are applied at the same time to all effects. The former permits to control the global error rate (EER) and the second test identifies the active effects. By simulation (Section 6) we found that the critical value for the normality test must be $\alpha = 0.036$ for an $EER = 5\%$.

5 Aluminum Casting Example

This is a challenging case study on an aluminum casting process, it appeared in [9], there were five factors studied with 16 runs hence a 2^{5-1} fraction was used. The response was the fraction defective, the experimental results are given in Table 1. Notice that the responses go from 6% all the way to 100%, hence it is expected that some factors are significant. [9] analyzed the arcsine square root transformation of the response, here we analyze the response directly, the results are completely similar.

The data, the estimated effects computed with the original data, the effects computed from the rank transformation, the robust estimates of the effects, and the posterior probabilities from [6] are given in Table 1.

The Daniel plot, [7], of the original data (top-left Figure 1) shows no indication of significant effects. The same conclusion applies to the Daniel plot of the rank transformation (top-rigt Figure 1). A completely different picture emerges when the effects are estimated robustly according to the proposed procedure (bottom center Figure 1). Effects B, D and DB are clearly significant.

Figure 2 shows the barplots for the effects in Lenth's [10] method and the posterior probabilities of the Bayesian [6] method that takes into account the possibility of outliers. Both methods fail in detecting any active effects. The Bayesian method fails because it could not detect the

presence of the outlier in the data.

Table 2 shows the results of conducting a significant test on the five methods considered. We apply the methods with their critical points obtained by simulation in Section 6. Again only the robust procedure shows a significant result.

This case study is difficult because there is an outliying observation in the data that is not evident. In fact from a residual analysis [9] concluded that first observation (0.14) is an outlier. After removing this run and reanalyzing the data [9] finds that effects B, D and DB are clearly significant, the same result that robust method but having to detect the outlier first!

Name	**Data**	**Effects (original)**	**Effects (rank)**	**Effects (robust)**	Post Prob
(1)	0.14				0.457
A	0.98	0.045	-0.250	-0.028	0.025
B	0.36	-0.195	-0.750	**-0.149**	0.063
C	0.42	0.050	0.125	-0.020	0.026
D	1.00	-0.285	-3.750	**-0.195**	0.177
E	0.90	-0.005	2.375	0.048	0.024
AB	0.28	-0.090	-1.000	0.004	0.029
AC	0.14	-0.125	-1.875	-0.027	0.035
AD	0.22	-0.120	-2.250	-0.016	0.034
AE	0.26	0.170	3.625	0.051	0.050
BC	0.38	-0.115	-2.125	-0.020	0.033
BD	0.12	0.260	2.750	**0.182**	0.132
BE	0.30	0.160	3.375	0.045	0.046
CD	0.06	-0.055	0.125	0.017	0.026
CE	0.22	0.115	1.750	0.021	0.033
DE	0.38	0.180	3.875	0.039	0.054

Table 1. Data, estimated effects and posterior probabilities

Method	**W**	**P-value**	**Calibrated Parameter**	**±2dF**	**Active Effects**
Lenth	—	—	$t_\alpha = 4.24$	—	None
Benski	0.98	0.9148	$\alpha = 0.065$	±0.51	None
Bayesian	—	—	$p_c = 0.91$	—	None
Ranks	0.96	0.5473	$\alpha = 0.034$	±8.00	None
Robust	0.85	0.0199	$\alpha = 0.036$	±0.10	B, D, BD

Table 2. Significance tests, aluminum casting example

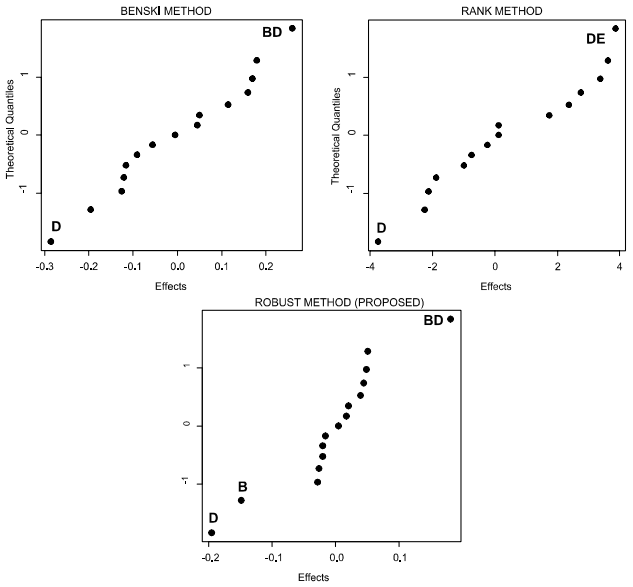

Figure 1. Daniels´s plot. Original data, rank and robust methods, aluminum casting example

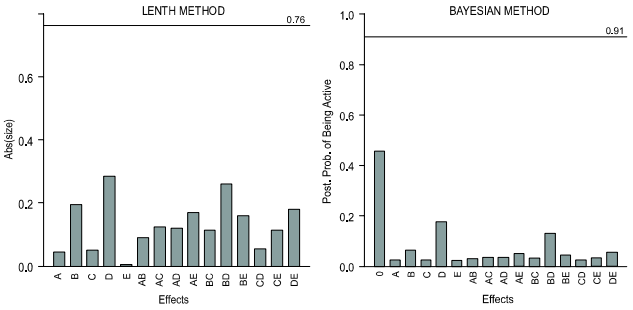

Figure 2. Barplots. Lenth and Bayesian methods, aluminum casting example

From this example we see the importance of methods that consider the possibility of outliers in data. In next section we report the results of a simulation study for a more general comparison of the methods.

6 Simulation Study

In this section we compare the performance of the robust method with existing methods that consider and do not consider the possibility of outliers in data. These methods are the ranks method [2] and the Bayesian method [6]. For reference, the study also includes the [10] and the non sequential variant of [3], as methods that do not contemplate the possibility of faulty observations.

The simulation study is divided in two parts: calibration of the methods and power study.

6.1 Calibration of the Methods

In order to achieve a fair comparison of methods all of them were calibrated to have an experimentwise error rate equal to 5% ($EER = 5\%$) when there are no active effects.

Table 3 shows the critical points that achieve the desired EER. The column MC stands for the number of Monte Carlo replications used to obtain the corresponding critical point.

Method	Outliers	Param	MC
1. Lenth (1989) (ref)	NO	$c = 4.24$	50000
2. Benski (ref)	NO	$\alpha = 0.065$	50000
3. Bayesian	YES	$p_c = 0.91$	3600
4. Ranks	YES	$\alpha = 0.034$	50000
5. Robust	YES	$\alpha = 0.036$	10000

Table 3. Calibration of the Methods

The calibrated value $c = 4.24$ for the [10] method is the same value reported by [13]. The calibrated values in the Table 7 are for 16 runs experiments (2^4, 2^{5-1}, 2^{6-2}, etc.). For other experimental sizes (say 8 or 32 runs) new calibrated values must be obtained.

6.2 Power Study

Once that methods have been calibrated we proceed to the power study. For that purpose we generated experimental data assuming a model with three active effects of different sizes, and three scenarios with different contamination levels, the same active effects and scenarios proposed by [2]. To allow for outliers present in data, each replication of 16 observations was generated with the model

$$Y_i = \alpha_1 A + \alpha_2 AB + \alpha_3 C + \varepsilon_i \quad ; \quad i = 1, 2, \ldots, 16, \quad (1)$$

where A, B and C are the active effects and the error distribution is the normal mixture

$$\varepsilon_i \sim (1-\beta)N(0,1) + \beta N(0, K^2).$$

where the second distribution $N(0, K^2)$ occurs with probability β and generates the contaminated data. Three different contamination scenarios are denoted by the vector (β, K): $(0, 0)$, $(0.05, 5)$ and $(0.10, 10)$. For example, the vector $(0, 0)$ consists of normal data without outliers and the vector $(0.10, 10)$ is the most contaminated scenario, ten percent outliers with variance equal to 10. Note that the number of observation from the second distribution follows a binomial distribution with parameters (16,β). Hence for the scenario (0.10,10) there are typically between 1 to 3 outliers.

In terms of the model coefficients in equation (1) the corresponding sizes for the active effects A, AB and C are $\alpha_1 = 1$, $\alpha_2 = 0.5$ and $\alpha_3 = 2$, respectively.

Two measures of performance were computed: power and a merit statistic proposed by [4]. The power for an active effect is the percentage of experiments where it was correctly declared as a significant effect. The merit statistic is defined as follows: let s be the number of simulated experiments in one scenario, then in our case $15s$ is the total number of estimated effects in the simulation, and from these there are $3s$ active effects and $12s$ inert effects. Let

n^+ = Number of active effects declared as such in the simulation

n^- = Number of inert effects wrongly declared to be significant

then the merit statistics is given by

$$QG = \frac{n^+}{3s}\left(1 - \frac{n^-}{12s}\right) \times 100\%.$$

The metric QG falls in the interval [0%, 100%] and it is intended to be a measure of the global performance of the method taking into account all the active effects, where large values indicate a better performance.

Figure 3 shows the plots of the power for each effect and each method given in Table 3. As expected, all of the methods have a better performance when ($\beta = 0, K = 0$). Lenth's method was the worst in all respects. On the contrary, the proposed robust MM-estimation method was always at the top in the absence of contamination. In the contaminated scenarios the proposed robust method has the best performance, very far from the methods (Benski and Lenth) that do not contemplate the atypical data and far from the two existing method that contemplate this possibility Box-Meyer and Ranks). Ranks and Bayesian methods have a similar global performance on contaminated scenarios.

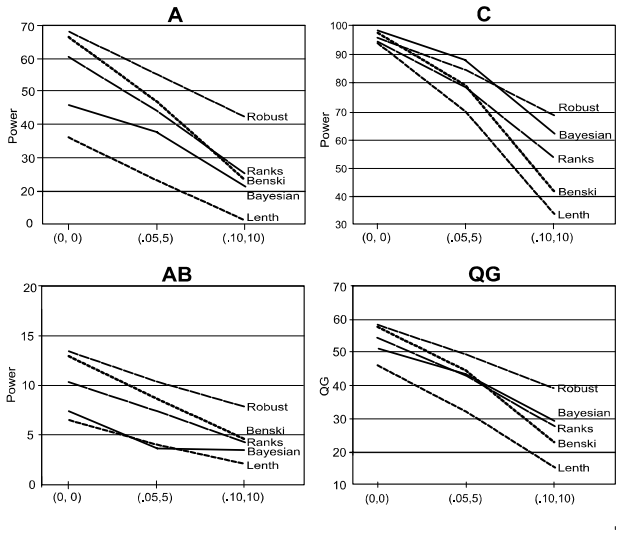

Figure 3. Power and QG of the Methods in Scenarios (0,0), (0.05,5) and (0.10,10)

Regarding the *QG* the statistic. The proposed robust method is systematically above the other methods for all scenarios. On the other side, Lenth's method is always below the other methods, indicating the worst performance. Benski's method has a competitive performance compared with Bayesian and ranks methods except in the most contaminated scenario. In conclusion our proposed robust method has a globally better performance in all scenarios.

7 Conclusion

We proposed a new method for detecting active effects in unreplicated factorial experiments, which considers the possibility of faulty observations. The new method is based in MM-estimation of the effects, a robust regression technique proposed by [14]. We illustrate the new method using a difficult example from the literature. The proposed method was the only method that detected a significant effect analyzing the original data. A simulation study was performed for comparing the new procedure with existing methods. The proposed robust method gave the best results in terms of power and global performance under both contaminated or uncontaminated scenarios.

References

[1] Aguirre-Torres V., "A simple analysis of unreplicated factorials with possible abnormalities," *Journal of Quality Technology* 25, pp. 183-187, 1993

[2] Aguirre-Torres V. and Pérez-Trejo M. E. "Outliers and the use of the rank transformation to detect active effects in unreplicated 2^f experiments," *Communications in Statistics* 30, pp. 637-663

[3] Benski, H. C. "Use of a normality test to identify significant effects in factorial designs," *Journal of Quality Technology* 21, pp. 174-178, 1989

[4] Benski, H. C. and Cabau, E. "Unreplicated experimental designs in reliability growth programs," *IEEE Trans. Reliability* 44, 199-205, 1995

[5] Box, G. E. P. and Meyer, R. "An analysis for unreplicated fractional factorials," *Technometrics* 28, pp. 11-18, 1986

[6] Box, G. E. P. and Meyer R. "Analysis of unreplicated factorials allowing for possible faulty observations," In *Design, Data and Analysis*. Mallow, C. Ed. Wiley, New York, 1987

[7] Daniel, C. "Use of half normal plots in interpreting factorial two-level experiments," *Technometrics*, 1, pp. 311-341, 1959.

[8] De la Vara, R. and Aguirre-Torres, V. . "R programming of Box-Meyer procedure with outliers," *Technical Report* DE-C07.2. Instituto Tecnológico Autónomo de México, México, D. F.; pp. 30 2007

[9] Kraber, S. "A case to test your metal (results)," *Stat-Teaser*. News from Stat-Ease, Inc. December, 1999

[10] Lenth, R. V. "Quick and easy analysis of unreplicated factorial experiments," *Technometrics* 31, pp. 469-473, 1989

[11] Rousseeuw, P. J. y A. M. Leroy. *Robust Regression and Outlier Detection*. Wiley, New York, 2003

[12] Simpson, J. R. and Montgomery, D. C. "A performance-based assessment of robust regression methods," *Communications in Statistics-Simulation and Computation* 27, 4, 1031-1049, 1998

[13] Ye, K. Q. and Hamada, M. "Critical values for the Lenth method for unreplicated factorial designs," *Journal of Quality Technology* 32, 1, 57–66, 2000

[14] Yohai, V. J. "High breakdown-point and high efficiency robust estimates for regression," *The Annals of Statistics* 15, 20, 642-656, 1987

[15] Yohai, V. J, Stahel, W. A. y Zamar, R. H. "A procedure for robust estimation and inference in linear regression," In *Directions in Robust Statitics and Diagnostics, Part II*. Eds. Stahel, W. A. y Weisber, S. Springer Verlag, New York, 1991