International Journal of Mathematics and
Computer Science, **11**(2016), no. 1, 43–49

Linear Regression Residuals Recomputed

John N. Haddad

Department of Mathematics and Statistics
Notre Dame University - Louaize
Zouk Mosbeh, Lebanon

email: john.n.haddad@ndu.edu.lb

(Received October 22, 2015, Accepted December 10, 2015)

Abstract

The estimation of the parameters of the simple linear regression model in the presence of an outlier is considered. Two statistics are obtained by maximizing the log-likelihood function, as constrained by the outlying observation. A simple function of these statistics is shown to produce an unbiased estimate of the variance of the residual errors. This estimation procedure allows for the residual value to be computed from an unbiased estimating equation. A simulation study is carried out to show the viability of the proposed estimate.

1 Introduction

The basic simple linear regression model can be expressed as

$$y_i^* = \alpha + \beta\, x_i^* + \epsilon, \quad i = 1, \ldots, n\,, \tag{1}$$

where the errors ϵ_i are assumed to be independently and identically distributed normal random variables with mean zero and constant variance σ^2. The unknown parameters α and β are respectively the intercept and the slope of the regression line. On replacing α by $\bar{y^*} - \beta\bar{x^*}$ in the the model specified by Equation (1), one has

$$y_i = \beta\, x_i + \epsilon_i \tag{2}$$

Key words and phrases: Linear regression, Residuals, Linear constraints, Nonlinear programming, Outliers.
AMS (MOS) Subject Classifications: 62M10.
ISSN 1814-0432, 2016, http://ijmcs.future-in-tech.net

where $y_i = y_i^* - \bar{y^*}$ and $x_i = x_i^* - \bar{x^*}$, $\bar{y^*}$ and $\bar{x^*}$ being the sample averages of the y_i^*'s and the x_i^*'s, respectively. Then, letting $\beta_{\ell s}$ denote the least-squares estimate of β, the i^{th} residual is given by $e_i = y_i - \beta_{\ell s}\, x_i$, $i = 1, \ldots, n$ and it can be shown that $\sum_{i=1}^{n} e_i^2/(n-2)$ is an unbiased estimate of σ^2.

In the context of parameter inference, the presence of outlying observations will create large residuals and consequently inflate the estimate of the variance. Numerous textbooks on linear regression present substantive discussions on this well-studied problem; see, for instance, Neter et al. (1996) and Wonnacott & Wonnacott (1990). Thus, more robust techniques such as those described in detail in Huber & Ronchetti (2009) ought to be utilized in lieu of the least-squares approach.

It will be assumed in the sequel that the data contains a single "troublesome" outlying observation that may be due to an intervention of some sort, such as a recording error or a breakdown in the system under consideration. Intervention analysis has been customarily used in the context of time series ARMA modelling, as can be seen from Box and Tiao (1976) or Abraham (1980).

The intervention at the j^{th} location can be viewed as causing y_j to be away from the actual mean by $q\,\sigma$ units where $q \geq 3$. The resulting constraint is then incorporated in the log-likelihood of the sample, which leads to a non-linear optimization problem whose optimal solution is derived in Section 2. The likelihood approach under constrained maximization is discussed for instance in Chen et al. (2008). Section 3 includes a simulation study that demonstrates the viability of the proposed procedure. A concluding discussion is presented in Section 4.

2 The Proposed Estimation Procedure

Assuming that the ϵ_i's are independently and identically normally distributed errors in the model specified by Equation (2), one has

$$\ell(\beta, \sigma) \propto -n \ln(\sigma) - \sum_i (y_i - \beta\, x_i)^2/2\sigma^2 \tag{3}$$

where $\ell(\beta, \sigma)$ denotes the log-likelihood function. Then, the maximum likelihood estimates of β and σ are the values that maximize $\ell(\beta, \sigma)$ or, equivalently, those that minimize $-\ell(\beta, \sigma)$. However, the minimization process has to be constrained by the outlier restriction, that is,

$$\frac{y_j - \beta\; x_j}{\sigma} \;=\; q\,, \tag{4}$$

where j is the location of the outlier and q is larger than or equal to 3 in magnitude. Thus, Equation (3) and the constraint given in Equation (4) form a nonlinear programming problem, which can be expressed in terms of the following Lagrangian function:

$$(n-1)\ln(\sigma) + \sum_{i \neq j} \frac{(y_i - \beta\, x_i)^2}{2\sigma^2} + \lambda_j \left(\frac{(y_j - \beta\, x_j)}{\sigma} - q \right), \tag{5}$$

where $\sum_{i \neq j}$ indicates that the sum runs from $i = 1$ to $i = n$, excluding the case $i = j$.

The existence of an optimal solution is discussed for instance in Winston (2004). In this instance, the solution must satisfy the following equations, which are obtained by differentiating Equation (5) with respect to σ and β:

$$(n-1)/\sigma - \sum_{i \neq j} (y_i - \beta\, x_i)^2/\sigma^3 \; - \; \lambda_j\, (y_j - \beta\, x_j)/\sigma^2 = 0 \tag{6}$$

and

$$- \sum_{i \neq j} x_i (y_i - \beta\, x_i)/\sigma^2 \; - \; (\lambda_j\, x_j/\sigma) = 0\,. \tag{7}$$

On solving Equations (6), (7) and (4) simultaneously by substituting λ_j as determined from Equation (7) in the left-hand side of Equation (6) multiplied by σ^3, expanding the resulting expression, noting that the terms in β^2 cancel out, then expressing β in terms of q and σ by making use of Equation (4) and simplifying, one obtains the following quadratic estimating equation:

$$(n-1)\hat{\sigma}^2 - q\, \mathcal{S}_1 \hat{\sigma} - \mathcal{S}_2 = 0, \tag{8}$$

where $\mathcal{S}_1 = \sum_{i=1}^{n} (x_i/x_j^2)(y_i\, x_j - y_j\, x_i)$ and $\mathcal{S}_2 = \sum_{i=1}^{n} ((y_i\, x_j - x_i\, y_j)/x_j)^2$. Thus, the optimal solution lying in the feasibility region is given by

$$\hat{\sigma} = \frac{q\, \mathcal{S}_1 + \sqrt{(q\, \mathcal{S}_1)^2 + 4(n-1)\, \mathcal{S}_2}}{2(n-1)} \tag{9}$$

and

$$\hat{\beta} = (y_j - q\, \hat{\sigma})/x_j\,. \tag{10}$$

Since the Kuhn-Tucker conditions are satisfied, this solution is indeed optimal.

Usually, the value of q, which is needed to determine σ in Equation (8), has to be determined. This can be achieved by approximating it by the standardized residual of the j^{th} observation wherein the least-squares estimates of β and σ are obtained from the other $n-1$ data points.

Since, as proved in the Appendix, the estimator

$$\tilde{\sigma}^2 = (\mathcal{S}_2 - \mathcal{S}_1^2\, x_j^2 / \sum x_i^2)/(n-2) \tag{11}$$

is unbiased for σ^2, it can be utilised in lieu of the estimator of σ^2 obtained from in Equation (9). This approach first provides an estimate of σ^2; then, q can readily be evaluated from Equation (8) and, finally, β is estimated from Equation (10). Accordingly, one may regard the standardized residual of the j^{th} observation as being recomputed.

3 Simulation Study

A simulation study is carried out to assess the viability of the proposed procedure. This is achieved by letting $x_i = i$ and $y_i = 5\, x_i + \epsilon$ for $i = 1, \ldots, n$ where ϵ is a standard normal normal random variable, simulated using the rnorm(n) function of the R statistical package. An outlier is then generated at the 5^{th} location by replacing y_5 with $y_5 + \delta$, that is, by adding δ standard deviations to y_5. In this experiment, δ can take on the values 5 and 10, and small, moderate and large sample sizes are represented by setting n equal to 10, 50 and 150, respectively. The quantities of interest are, in order, β_{ls} and σ_{ls}, the ordinary least-squares estimates of β and σ, which are based on the entire data set; next, $\beta(j)$ and $\sigma(j)$, the ordinary least-squares estimates of β and σ, with the j^{th} observation removed from the data set, which enable one to obtain the value of standardized residual associated with the j^{th} observation, that is, $q(j) = (y_j - \beta(j)\, x_j)/\sigma(j)$, and then $\hat{\sigma}$ and $\hat{\beta}$, as specified in Equations (9) and (10), respectively; the last three values to be determined are $\tilde{\sigma}$ as given in Equation (11), $\tilde{q}$ evaluated from Equation (8), and $\tilde{\beta}$ estimated from Equation (4). The experiment is replicated 1000 times. The following table reports the average of those quantities, the numbers in parentheses being the sample standard errors. It can be seen that the proposed estimation methodology generally provides more accurate estimates than the other approaches.

Table 1. Performance of the proposed procedure versus the least-squares approach.

n	10	10	50	50	150	150
δ	5	10	5	10	5	10
β_{ls}	5.06543	5.12640	5.00054	4.98025	4.99874	4.99761
	(0.05150)	(0.05195)	(0.00473)	(0.01007)	(0.00190)	(0.00189)
σ_{ls}	1.85321	3.33449	1.21007	1.69371	1.07247	1.27694
	(0.31866)	(0.32748)	(0.12144)	(0.12969)	(0.06097)	(0.07161)
$\beta(j)$	5.00028	4.99699	4.99996	4.99543	5.00001	5.00002
	(0.05278)	(0.05421)	(0.00474)	(0.01031)	(0.00191)	(0.00190)
$\sigma(j)$	0.96050	0.99610	0.99230	0.98862	0.99266	0.99486
	(0.24927)	(0.24649)	(0.10083)	(0.10441)	(0.05652)	(0.05856)
$q(j)$	5.62282	11.08019	5.04708	9.96152	5.04088	9.91705
	(2.04506)	(3.47952)	(1.09748)	(1.43832)	1.06753)	(1.18170)
$\hat{\sigma}$	0.95811	0.96551	0.99228	0.98900	0.99212	0.99497
	(0.24751)	(0.24605)	(0.10082)	(0.10436)	(0.05643)	(0.05854)
$\hat{\beta}$	5.00226	4.99805	4.99998	4.99962	5.00001	5.00007
	(0.05289)	(0.05425)	(0.00474)	(0.01033)	0.00191)	(0.00190)
$\tilde{\sigma}$	0.97564	0.98078	0.99459	1.00922	0.99325	0.99709
	(0.25355)	(0.24971)	(0.10097)	(0.10251)	(0.05649)	(0.05846)
$\tilde{q}$	5.52234	10.90433	5.03529	9.74125	5.03383	9.89223
	(2.01795)	(3.42944)	(1.09462)	(1.34735)	(1.06642)	(1.17377)
$\tilde{\beta}$	5.00325	4.99844	4.99998	4.99921	4.99994	5.00005
	(0.05299)	(0.05425)	(0.00474)	(0.01032)	(0.00191)	(0.00190)

4 Concluding Remarks

It should be noted that one can readily adapt the proposed procedure to data sets that are devoid of outliers. In this case, the estimate of q will simply be the standardized residual. To verify that $\tilde{q}$ will match the corresponding standardized residual in absence of any outlier, the same simulation experiment was repeated with $\delta = 0$. Following 1000 replications carried out with samples of size 50, it was found that $e_5 = y_5 - \beta_{ls}\, x_5 = 0.03692$ and that $\tilde{q} = 0.03693$, these values being nearly identical. The remaining residuals can be obtained either by determining $\tilde{\beta}$ from $\tilde{\sigma}$ and $\tilde{q}$ and letting $\tilde{e}_i = y_i - \tilde{\beta}\, x_i$ or by recalculating $\tilde{\sigma}$, $\tilde{q}$ and $\tilde{\beta}$ for every point of the data set. The methodology advocated herein turns out to be robust in the presence of an outlier, the resulting estimates being at least as accurate as the ordinary least-squares estimates, whether or not the data set contains an outlier. Accordingly, it ought to be favoured over the conventional least-squares approach in practice.

References

[1] B. Abraham, Intervention analysis and multiple time series, *Biometrika*, **67**, 1980, 73–78.

[2] G. E. P. Box, G. C. Tiao, Intervention analysis with applications to economic and environmental problems, *J. Amer. Stat. Assoc.,* **70**, 1975, 70–79.

[3] J. Chen, A. M. Variyath, B. Abraham, Adjusted empirical likelihood and its properties, *J. Comput. Graph. Stat.*, **17**, 2008, 426–443.

[4] P. J. Huber, E. M. Ronchetti, Robust Statistics, 2nd ed., Wiley, New York, 2009.

[5] J. Neter, M. H. Kutner, C. J. Nachtsheim, W. Wasserman, Applied Linear Statistical Models, 4th ed., Richard D. Irwin, Inc., Burr Ridge, Illinois, 1996.

[6] W. L. Winston, Operations Research: Applications and Algorithms, 4th ed., Duxbury, Belmont, California, 2003.

[7] R. J. Wonnacott, T. H. Wonnacott, Statistics for Business and Economics, 5th ed., Wiley, New York, 1990.

Appendix

Let

$$\mathcal{S}_1 = \sum (x_i/x_j)(y_i - x_i\, y_j/x_j)$$

and

$$\mathcal{S}_2 = \sum (y_i - x_i\, y_j/x_j)^2$$

where, throughout this appendix, the sums run from $i = 1$ to $i = n$, excluding the case $i = j$. Since

$$(y_i - x_i\, y_j/x_j) \sim \mathcal{N}(x_i(\beta - y_j/x_j)\,,\, \sigma^2)\,,$$

one has

$$\mathrm{E}(\mathcal{S}_1) = \Big(\sum x_i^2/x_j\Big)(\beta - y_j/x_j) \tag{12}$$

and

$$\mathrm{Var}(\mathcal{S}_1) = \sigma^2 \sum x_i^2/x_j^2\,.$$

Thus,

$$\mathrm{E}(\mathcal{S}_1^2) = \mathrm{Var}(\mathcal{S}_1) + (\mathrm{E}(\mathcal{S}_1))^2 = \Big(\sigma^2 + (\beta - y_j/x_j)^2 \sum x_i^2\Big) \sum x_i^2/x_j^2,$$

or, equivalently,

$$\mathrm{E}\Big(\mathcal{S}_1^2\, x_j^2 \Big/ \sum x_i^2\Big) = \sigma^2 + (\beta - y_j/x_j)^2 \sum x_i^2\,. \tag{13}$$

Moreover,

$$\mathrm{E}(\mathcal{S}_2) = \sum \mathrm{E}(y_i - x_i\, y_j/x_j)^2 = \sum \Big(\mathrm{Var}(y_i - x_i\, y_j/x_j) + (\mathrm{E}(y_i - x_i\, y_j/x_j))^2\Big)$$

so that

$$\mathrm{E}\,(\mathcal{S}_2) = (n-1)\,\sigma^2 + (\beta - y_j/x_j)^2 \sum x_i^2\,. \tag{14}$$

Finally, on subtracting Equation (13) from Equation (14), one obtains

$$\frac{\mathrm{E}\left(\mathcal{S}_2 \;-\; \mathcal{S}_1^2\, x_j^2 \Big/ \sum x_i^2\right)}{(n-2)} \;=\; \sigma^2\,.$$