Hindawi
Journal of Applied Mathematics
Volume 2021, Article ID 1478843, 5 pages
https://doi.org/10.1155/2021/1478843

Research Article

On the Upper Bounds of Test Statistics for a Single Outlier Test in Linear Regression Models

Tobias Ejiofor Ugah, Emmanuel Ikechukwu Mba, Micheal Chinonso Eze, Kingsley Chinedu Arum, Ifeoma Christy Mba, and Henrietta Ebele Oranye

University of Nigeria, Nsukka, Nigeria

Correspondence should be addressed to Micheal Chinonso Eze; chinonso.eze@unn.edu.ng

Received 21 June 2021; Revised 30 August 2021; Accepted 1 September 2021; Published 22 September 2021

Academic Editor: Saeid Abbasbandy

A bewildering large number of test statistics have been found for testing the presence of an outlier in multiple linear regression models. Exact critical values of these test statistics are not available, and approximate ones are usually obtained by the first-order Bonferroni upper bound or large-scale simulations. In this paper, we show that the upper bound values of two of these test statistics are algebraically the same. An application to real data for multiple linear regression is used to demonstrate the procedure.

1. Introduction

An outlier is a discordant observation. It is an observation that does not fit in with the pattern of the remaining observations. It differs markedly not only from other members of the set from which it occurs, but also from its fitted value. Such an observation usually has a large residual. Outliers meet data analysts at the point of data analysis and in data mining. Reference [1] pointed out that there various causes of outliers such as human errors, erroneous operation of computer systems, sampling errors, or standardization failures. Excellent books on outliers include [2–4].

Outliers usually have a major influence on the resulting parameter estimates, and their presence impacts adversely on the results of the statistical inference concerning the models. They can reduce the power of statistical tests during analysis. Reference [5] advised that there is the need for the analyst to identify outliers if they exist so that appropriate measures might be taken.

Outliers need to be identified and corrected or eliminated. The process of identification and correction of outliers is not a straightforward thing; rather, it requires marked ability, competence, circumspection, and a strict adherence to scientific objectivity (impartiality) of high degree. If identified outliers cannot be remedied, they need to be removed because they contaminate the information contained by the remainder of that set of data (see [1, 6]).

Test for an outlying observation in the response variable is usually based on the use of test statistics that depend on the standardized residuals. Different test statistics have been developed for testing of an outlier in a least squares analysis of linear regression models. However, exact critical values of some of these test statistics are not available and are not easy to obtain. The available approximate ones are based on the first-order Bonferroni upper bound or large-scale simulations.

Upper bounds for the critical values of test statistics for detecting the presence of a single outlier in linear regression have been developed by [7, 8]. Although formal distinctions exist in the principles invoked by [7, 8] in deriving these upper bounds, we show in this paper that these upper bounds derived by [7, 8] are algebraically the same.

The multiple linear regression model is

$$Y = X\beta + \varepsilon, \tag{1}$$

where $\mathbf{Y}$ is the $n \times 1$ observation vector, X is an $n \times$ p matrix of constants, β is a $p \times 1$ vector of unknown parameters to be

estimated, and ε is an $n \times 1$ vector of normally distributed errors. Assuming that $E(\varepsilon) = 0$ and $\mathrm{Var}(\varepsilon) = \sigma^2 1$, the least squares estimator of β in (1) is given by

$$\widehat{\beta} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}, \tag{2}$$

and the vector of residuals is

$$\mathbf{e} = \mathbf{Y} - \mathbf{X}\widehat{\beta} = \left(I - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right)\varepsilon. \tag{3}$$

The variance-covariance matrix **e**

$$\mathrm{Var}(\mathbf{e}) = \left(I - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathrm{X}'\right)\sigma^2. \tag{4}$$

If σ^2 is estimated using $\sigma\wedge^2 = \mathbf{e}'\mathbf{e}/(n-p)$, then the estimated variance-covariance matrix of **e** becomes

$$\widehat{\mathrm{Var}}(\mathbf{e}) = \left(I = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right)\sigma\wedge^2. \tag{5}$$

Residuals are important diagnostic tools in regression analysis as no regression analysis is complete without a thorough examination of them. They are versatile as most regression diagnostics are written in terms of them. They are used in checking model adequacy and the validity of model assumptions. A thorough examination of the residuals therefore provides valuable information concerning the appropriateness of assumptions that underlie statistical models and helps in pinpointing an appropriate model. Different types of graphic plots (representations) of residuals are used for diagnostic purposes.

Ordinary residuals are not all that suitable for diagnostic purposes, and a standardized version of them is usually preferred. This is because the variances of the residuals are not homogeneous, and this makes them intractable. A standardized residual has a representation of the form

$$R_i = \frac{y_i - \widehat{y}_i}{\sigma\wedge^2\sqrt{1 - h_{ii}}}, \tag{6}$$

where $\widehat{y}_i$ is the predicted value of y_i and h_{ii} is the ith element of matrix $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, called the hat matrix. The ith transformed residual R_i is often called an internally studentized residual. They are tractable and are more versatile. They are used as a replacement of the ordinary residuals in regression diagnostics. Numerous graphical and numerical techniques for checking model assumptions using standardized residuals can be found in the regression literature. They are also fundamental building blocks for most of the known test statistics studied in the literature for outlier detection in linear models (see [9, 10]).

The test statistic

$$R_n = \max\left|\frac{y_i - \widehat{y}_i}{\sigma\wedge^2\sqrt{1 - h_{ii}}}\right| = \max|R_i|, \tag{7}$$

is called the maximum absolute internally studentized residuals. Reference [11], following the suggestion of [12], used a large-scale simulation study involving many thousands of sampling experiments to obtain approximate critical values of (7) for a simple linear regression. The approximate values obtained by [11] are almost the same with the values obtained by [13].

Reference [7] considered the test statistic

$$R_n^* = \max\left|\frac{e_i}{\bar{\sigma}_{e_i}}\right|, \tag{8}$$

where $\bar{\sigma}_{e_i}^2$ is the estimated average variance of the ordinary residuals. Reference [9] showed that the variance of the residuals is $(\mathrm{n} - \mathrm{p})\sigma^2/n$, so that the estimated variance of the ordinary residuals $\bar{\sigma}_{e_i}^2 = (\mathrm{n} - \mathrm{p})\sigma\wedge^2/n$.

Therefore,

$$R_n^* = \frac{n^{1/2}\max|e_i|}{\left(\sum e_i^2\right)^{1/2}}. \tag{9}$$

Reference [7] showed that the corresponding percentage point R_0^* of R_n^* is bounded above by

$$U = \sqrt{\frac{(n-p)F}{n-p-1+F}}, \tag{10}$$

where F is the $100(1 - \alpha/n)$ percentage point of the F distribution with degrees of freedom 1 and $n - p - 1$, n is the number of observations, and p is the number parameters estimated. Reference [7] results for simple linear regression were found to be almost identical to those of [11]. Reference [7] also suggested the use of (10) to obtain other critical values that are not in Table 1. The reference [7] result was not elaborate and extensive enough as the result in [8] because of the unavailability of the needed values of the F -distribution (see [8]).

Define

$$\xi_i = \frac{R_i}{\sqrt{n-p}}. \tag{11}$$

Reference [14] showed that the joint distribution of $\xi_i's$ has a multivariate Inverted-Students Function and that the probability density function for any ξ_i is a univariate Inverted-Students Function with probability density function given by

$$f(\xi_i) = C\left(1 - \xi_i^2\right)^{(n-p-3)/2}, \quad \xi_i^2 \le 1, \tag{12}$$

Table 1: Upper bounds of the critical values of R_n for detecting a single outlier in a simple linear regression model.

Sample size	$\alpha = 0.10$ R_n	$\alpha = 0.05$ R_n	$\alpha = 0.01$ R_n
4	1.4131	1.4139	0.4142
5	1.6974	1.7147	1.7286
6	1.8838	1.9270	1.9751
7	2.0142	2.0799	2.1667
8	2.1125	2.1961	2.3178
9	2.1911	2.2883	2.4398
10	2.2562	2.3643	2.5407
12	2.3602	2.4840	2.6988
14	2.4414	2.5760	2.8186
16	2.5079	2.6502	2.9136
18	2.5641	2.7121	2.9919
20	2.6126	2.7651	3.0575
30	2.7869	2.9516	3.2812
60	3.0508	3.2247	3.5869

where

$$C\frac{\Gamma((n-p)/2)}{\Gamma(1/2)\Gamma((n-p-1)/2)}. \tag{13}$$

Reference [8], following the suggestion of [7], made use of the results of [14]. Reference [8] used the first-order Bonferroni inequality to obtain the upper bounds R_0 of the critical values of R_n. Reference [8] obtained ξ_0 from

$$\int_{\xi_0}^{1} 2nf(\xi_i)\mathrm{d}\xi_i = \alpha, \tag{14}$$

where $\xi_0 = R_0/\sqrt{n-p}$ and then obtained R_0 using the relationship between R_0 and ξ_0 given by the equation

$$R_0 = \xi_0\sqrt{n-p}, \tag{15}$$

for sample sizes up to $n = 100$, regression parameters $p = 25$ and $\alpha = 0.10, 0.05$, and 0.01. With that, [8] is claimed to have produced the most elaborate upper bound values R_0. For $p = 2$, the upper bounds R_0^* computed by [7] using (10) and the upper bound values R_0 computed by [8] using (14) are extremely very close.

2. Demonstration of the Sameness of Upper Bounds

In this section, we show that the upper bounds R_0^* and R_0 are algebraically identical. From (10), we let

$$U = \sqrt{\frac{(n-p)F}{n-p-1+F}}. \tag{16}$$

We determine the distribution of U as follows:

$$\Pr(U < u) = \Pr\left(\sqrt{\frac{(n-p)F}{n-p-1+F}} < u\right) = \Pr\left(F < \frac{u^2(n-p-1)}{n-p-u^2}\right), \tag{17}$$

so that

$$f_U(u) = f_F\left[\left(\frac{u^2(n-p-1)}{n-p-u^2}\right), 1, n-p-1\right]\frac{2(n-p-1)(n-p)u}{(p+u^2-n)^2}, \tag{18}$$

with distribution domain or range given by $(0, \sqrt{n-p})$. Explicitly, we have

$$f_U(u) = H\left(1-\frac{u^2}{n-p}\right)^{(n-p-3)/2} \quad 0 < u < \sqrt{n-p}, \tag{19}$$

where

$$H = \frac{2\Gamma((n-p)/2)}{\Gamma(1/2)\sqrt{n-p}\Gamma((n-p-1)/2)}. \tag{20}$$

Then, using the first Bonferroni inequality, one can obtain the upper bounds R_0^* by solving

$$\int_{R_0^*}^{\sqrt{n-p}} nf_U(u)\mathrm{d}u = \alpha. \tag{21}$$

Now, from (11), we have

$$\Pr(R_i < r) = \Pr(\xi_i\sqrt{n-p} < r) = \Pr\left(\xi_i < \frac{r}{\sqrt{n-p}}\right), \tag{22}$$

so that

$$f_{R_i}(r) = f_\xi\left[\left(\frac{r}{\sqrt{n-p}}\right)\right]\frac{1}{\sqrt{n-p}}, \tag{23}$$

with distribution domain or range given by $(-\sqrt{n-p}, \sqrt{n-p})$. Explicitly, we have

$$f_{R_i}(r) = D\left(1-\frac{r^2}{n-p}\right)^{(n-p-3)/2} \quad -\sqrt{n-p} < r < \sqrt{n-p}, \tag{24}$$

where

$$D = \frac{\Gamma((n-p)/2)}{\Gamma(1/2)\sqrt{n-p}\Gamma((n-p-1)/2)}. \tag{25}$$

Let

$$\begin{aligned} Y_i &= |R_i|, \\ \Pr(Y_i < y) &= \Pr(|R_i| < y) = \Pr(-y < R_i < y). \end{aligned} \tag{26}$$

Because of the symmetry of the distribution of R_i in (24), we obtain the distribution of $Y_i = |R_i|$ as follows:

$$\Pr(Y_i < y) = 2 \Pr(R_i < y). \tag{27}$$

Explicitly, we have

$$f_{Y_i}(y) = H\left(1 - \frac{y^2}{n-p}\right)^{(n-p-3)/2} \quad 0 < y < \sqrt{n-p}. \tag{28}$$

Then, using the first Bonferroni inequality, one can obtain R_0 by solving

$$\int_{R_0}^{\sqrt{n-p}} n f_{Y_i}(y)\mathrm{d}y = \alpha. \tag{29}$$

The sameness of (21) and (29) means that

$$\int_{R_0^*}^{\sqrt{n-p}} n f_{U_i}(u)\mathrm{d}u = \alpha \Rightarrow \int_{R_0}^{\sqrt{n-p}} n f_{Y_i}(y)\mathrm{d}y = \alpha, \tag{30}$$

implying that $R_0 = R_0^*$. This also means that R_n and R_n^* have distributions that are bounded by the same distribution.

Using (21) to obtain the upper bounds R_0^* averts the problem encountered by [7]. This is because (10) depends on the tabulated percentage points of F-distribution while (21) does not. Reference [8] remarked that implementation of the suggestion made by [7] was very difficult because the needed percentage points of the F-distribution were not available. Therefore, for any value of α, R_0^* can easily be obtained using (21) without recourse to a tabulated value of the F-distribution. It is also preferable to use (29) to obtain R_0 instead of (14). This is because, using (14) involves a kind of transformation from ξ_0 to R_0 as indicated in (15), but using (29) does not. The use of approximate critical values for detecting a single outlier in linear regression can be found in the work of [7, 8].

3. Table Construction

We use the Bonferroni inequality to obtain upper bound values for the 10 percent, 5 percent, and 1 percent critical values of the test statistic R_n. A table of the upper bounds of the critical values of R_n is presented in Table 1 for a simple linear regression and sample sizes up to 60. These were obtained by solving (29) using the Mathematica Software. It is to demonstrate numerically that the upper bound values of the two test statistics (7) and (8) are the same as what (30) shows. Equation (29) produces precise and accurate values of upper bound values of these test statistics ((7) and (8)). The values in Table 1 compare favorably with values obtained by [7] by solving (10), the values obtained by [8] by solving (14), and the approximate values obtained by [11] via simulation.

Table 2: Data set and standardized residuals for multiple regression of plant available phosphorus (Y) on inorganic phosphorus (X_1) and organic phosphorus (X_2).

Soil sample n	Y	X_1	X_2	R_i
1	64	0.4	53	0.13331
2	60	0.4	23	0.04388
3	71	3.1	19	0.40169
4	61	0.6	34	0.02905
5	54	4.7	24	-0.68481
6	77	1.7	65	0.79326
7	81	9.4	44	0.19893
8	93	10.1	31	0.80425
9	93	11.6	29	0.68538
10	51	12.6	58	-1.72895
11	76	9.4	37	-0.02256
12	96	23.1	46	-0.29759
13	77	23.1	50	-1.29842
14	93	21.6	44	-0.30278
15	95	23.1	56	-0.39542
16	54	1.9	36	-0.45870
17	168	26.8	58	3.17401
18	99	29.9	51	-0.85219

4. Application to Real Data

We now show that the upper bounds of the two test statistics are the same with an application to a real data set. The data in Table 2 is from [15]. Reference [15] carried out an investigation concerning the source from which corn plants obtain their phosphorus. It was carried out by chemically determining the concentrations of inorganic (X_1) and organic (X_2) phosphorus in the soils. They used eighteen soil samples in the experiment and measured the phosphorus content Y of the corn grown on $n = 18$ Iowa soils. The phosphorus content Y was used as the dependent variable in a multiple regression analysis with X_1 and X_2 as the independent variables. The content Y (phosphorus content) of the corn in soil sample number 17 was found to be considerably larger than the phosphorus content of the corn grown in the other soil samples (no explanation was given for its size) and produced a standardized residual of 3.18. Multiple linear regression analysis of the data set produced the result in Table 2.

We now show that the upper bound values R_0^* and R_0 are the same. We compute the upper value R_0^* of R_n^* using equation (10) which is given by

$$\sqrt{\frac{(n-p)F}{n-p-1+F}}, \tag{31}$$

where F is the $100(1-\alpha/n)$ percentage point of the F distribution with degrees of freedom 1 and $n-p-1$, n is the number of observations, and p is the number parameters estimated. For $\alpha = 0.01$, $n = 18$, $p = 3$, and $F = 19.76$, the upper bound for their critical value of R_n^* is

$$R_0^* = \sqrt{\frac{15 \times 19.76}{14 + 19.76}} = 2.96. \quad (32)$$

To obtain the upper bound for the critical value of R_n, we make use of equation (14). For $\alpha = 0.01$, $n = 18$, and $p = 3$, we apply equation (14) using the *Mathematica software* to obtain $R_0 = 2.96$. Thus, the upper bound values of the two test statistics are the same. Equation (21) or (29) gives the same value.

Finally, the observed value of 3.18 is found to be significant at the one percent level ($3.18 > 2.97$). Thus, the phosphorus content of the corn grown in soil sample number 17 should be regarded as an outlier, and the null hypothesis of no outlier in the data set is rejected at the 0.01 level.

5. Conclusions

In this article, we have shown that the upper bound values R_0 of the test statistic (7) and the upper bounds R_0^* of the test statistic (8) are identical. Although formal distinctions exist in the principles used by [7] in deriving R_0^* and those employed by [8] in deriving R_0, we have herein shown that they are algebraically the same. Having shown this, we recommend the use of (29) to compute the upper bounds of (7) or (8). It is more tractable than (10) and (14). Since (14) borders on some kind of transformation and (10) makes use of tabulated values of F-distribution, accuracy and precision may be lost when using them.

Data Availability

A real data set on regression with a single outlier has been analyzed and included in the paper.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] P. D. Domanski, "Study on statistical outlier detection and labelling," *International Journal of Automation and Computing*, vol. 17, no. 6, pp. 788–811, 2020.

[2] V. Barnett and T. Lewis, *Outlier in Statistical Datas*, Wiley and Son, Chichester, U.K., 1994.

[3] D. M. Hawkins, *Identification of Outliers*, Springer, Dordrecht, The Netherlands, 1980.

[4] P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection*, John Wiley and Sons, New York, USA, 1987.

[5] A. Rajarathinam and B. Vinoth, "Outlier detection in simple linear regression models and robust regression-a case study of wheat production data," *Statistics*, vol. 3, pp. 531–535, 2014.

[6] N. N. R. Ranga Suri, M. N. Murty, and G. Athithan, *Outlier Detection: Techniques and Applications: A Data Mining Perspective*, Springer, Cham, Germany, 2019.

[7] P. Prescott, "An approximate test for outliers in linear models," *Technometrics*, vol. 17, no. 1, pp. 129–132, 1975.

[8] R. E. Lund, "Tables for an approximate test for outliers in linear models," *Technometircs*, vol. 17, no. 4, pp. 473–476, 1975.

[9] D. F. Andrews and D. Pregibon, "Finding the outliers that matter," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 40, no. 1, pp. 85–93, 1978.

[10] F. J. Anscombe and J. W. Turkey, "The examination and analysis of residuals," *Technometrics*, vol. 5, no. 2, pp. 141–160, 1963.

[11] G. L. Tietjen, R. H. Moore, and R. J. Beckman, "Testing for a single outlier in simple linear regression," *Technometrics*, vol. 15, no. 4, pp. 717–721, 1973.

[12] D. W. Behnken and N. R. Draper, "Residuals and their variance patterns," *Technometrics*, vol. 14, no. 1, pp. 101–111, 1972.

[13] F. E. Grubbs, "Procedures for detecting outlying observations in samples," *Technometrics*, vol. 11, no. 1, pp. 1–21, 1969.

[14] J. H. Ellenberg, "The joint distribution of the standardized least squares residuals from a general linear regression," *Journal of the American Statistical Association*, vol. 68, no. 344, article 941943, pp. 941–943, 1973.

[15] G. W. Snedecor and W. G. Cochran, *Outlier in Statistical Methods*, The Iowa State University Press, Ames, 1967.