On the Large-sample Size Critical Values of the Maximum Absolute Internally Studentized Residuals

Tobias Ejiofor Ugah[1], Kingsley Chinedu Arum[1], Charity Uchenna Onwuamaeze[1], Everestus Okafor Ossai[1], Nnaemeka Martin Eze[1], Emmanuel Ikechukwu Mba[1], Caroline Ngozi Asogwa[2,*], Angela, Obayi Adaora[2], Ifeoma Christy Mba[3], Oluchukwu Chukwuemeka Asogwa[4], Ikenna Emmanuel Chimezie [4], Comfort Njideka Ekene-Okafor[5].

[1] Department of Statistics, Faculty of Physical Sciences, University of Nigeria, Nsukka, Nigeria

[2] Department of Computer Science, Faculty of Physical Sciences, University of Nigeria, Nsukka, Nigeria

[3] Department of Economics, Faculty of Social Sciences, University of Nigeria, Nsukka, Nigeria

[4] Department of Mathematics and Statistics, Alex Ekwueme Federal University Ndufu Alike, Ebonyi State, Nigeria

[5] Department of Computer Science/Mathematics, Faculty of Natural Sciences and Environmental Studies, Godfrey Okoye University, Nigeria.

Received April 6, 2024; Revised June 17, 2024; Accepted August 20, 2024

Cite This Paper in the Following Citation Styles

***(a)**: [1] Tobias Ejiofor Ugah, Kingsley Chinedu Arum, Charity Uchenna Onwuamaeze, Everestus Okafor Ossai, Nnaemeka Martin Eze, Emmanuel Ikechukwu Mba, Caroline Ngozi Asogwa, Angela, Obayi Adaora, Ifeoma Christy Mba, luchukwu Chukwuemeka Asogwa, Ikenna Emmanuel Chimezie, Comfort Njideka Ekene-Okafor, "On the Large-sample Size Critical Values of the Maximum Absolute Internally Studentized Residuals," Mathematics and Statistics, Vol.12, No.5, pp. 443-447, 2024. DOI: 10.13189/ms.2024.120505*

***(b)**: Tobias Ejiofor Ugah, Kingsley Chinedu Arum, Charity Uchenna Onwuamaeze, Everestus Okafor Ossai, Nnaemeka Martin Eze, Emmanuel Ikechukwu Mba, Caroline Ngozi Asogwa, Angela, Obayi Adaora, Ifeoma Christy Mba, luchukwu Chukwuemeka Asogwa, Ikenna Emmanuel Chimezie, Comfort Njideka Ekene-Okafor (2024). On the Large-sample Size Critical Values of the Maximum Absolute Internally Studentized Residuals, Mathematics and Statistics, 12(5), 443-447. DOI: 10.13189/ms.2024.120505*

Abstract The maximum absolute internally studentized residual is a regular diagnostic measure for identification of a single outlying observation in the response variable in linear regression models. However, due to the daunting and formidable nature of the probability density function of this statistic, exact critical values are tough to compute. The Bonferroni inequality and intensive simulations are the only tools for determining its critical values as a means for detecting a single outlying observation in a linear regression model. In this paper, we present a straightforward alternative technique for obtaining asymptotic critical values of this statistic. The technique can be applied to any linear regression model and is convenient for routine use. The asymptotic distribution of this statistic is derived and used in obtaining the upper bounds for its critical values. It is shown that the proposed technique does not depend on the number of independent variables or the number of regression parameters in the model. Thus, the computational cumbersomeness and tedium imposed by the complexity associated with the distribution of this statistic and the use of the Bonferroni inequality are circumvented. The main advantages of the proposed procedure are its computational simplicity and efficiency to handle large datasets in high dimension. The asymptotic critical values of this statistic obtained by the proposed method are almost identical to those obtained by other authors, even though the techniques and principles employed in this work are entirely different from that employed by them.

Keywords Critical Values, Bonferroni Inequality, Test Statistic, Studentized Residual, Hat Matrix, Leverage

1 Introduction

The use of regression techniques to analyze large datasets from medical, physical and social sciences as well as from other areas of science and technology is not uncommon. It is also not uncommon for a large dataset to contain an outlying observation. An outlying observation may be as a result of human errors such as during transcription of the data or the result of gross deviation from prescribed experimental procedures. The presence of an outlying observation in a dataset can obviously impair estimation efficiency and the meaningfulness of standard tests of hypotheses. Of immense concern to researchers is the detection of outlying observations in linear regression, because of the common application of regression models in many areas. It is therefore necessary that outlying observations be detected and scrutinized by data analysts.

The study of outlying observations in linear models is a very exciting statistical problem. Excellent book-length treatments of outliers include [1, 2, 3, 4, 5, 6, 7]. A good number of papers have been published and devoted exclusively to this subject such as [8, 9, 10, 11].

Consider the linear regression model

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

where $\mathbf{Y} = (y_1, y_2, ..., y_n)'$ is an $n \times 1$ vector of values of the response variable, $\boldsymbol{X} = (X_1', X_2', ..., X_n')'$ is an $n \times$ p matrix of explanatory variables, $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)'$ is a p $\times$ 1 vector of unknown parameters, $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)'$ is an $n \times 1$ vector of independent normal random variables with mean 0 and (unknown) variance σ^2. For $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ being the ordinary least squares estimator of $\boldsymbol{\beta}$, the vector of the ordinary least squares residuals $\mathbf{e}$ can be written as

$$\mathbf{e} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} = (\boldsymbol{I} - \mathbf{H})\mathbf{Y},$$

where $\mathbf{H} = (h_{ij}) = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is called the hat matrix or the leverage matrix. $\mathrm{e}_i = \mathrm{y}_i - \hat{\mathrm{y}}_i$ is called the ith residual, where $\hat{\mathrm{y}}_i$ is the predicted value of y_i. The residual mean square estimator of σ^2 is then

$$\hat{\sigma}^2 = \frac{\mathbf{e}'\mathbf{e}}{n-p}.$$

The ordinary least squares residuals e_i are important diagnostic tools in linear regression. They are employed as a basis for identifying outlying observations in the response variable in linear regression models. However, these residuals have several deficiencies which dwarf their usefulness as a means for detecting a single outlying observation in a linear regression model. Their variances are unequal and they are correlated. Standardized residuals r_i are preferable for diagnostic purposes. The ith standardized residual has a representation of the form

$$r_i = \frac{\mathrm{e}_i}{\hat{\sigma}\sqrt{1 - \mathrm{h}_{ii}}}. \tag{2}$$

The ith standardized residual above is sometimes called an internally studentized residual. r_i is an important component of most diagnostic measures in linear regression diagnostics. Standardized residuals r_i are more tractable and are preferable to ordinary residuals e_i for detecting outliers in the response variable in linear regression models. However, the probability density function of r_i is complex.

Define

$$\xi_i = \frac{r_i}{\sqrt{n - \mathrm{p}}}. \tag{3}$$

Ellenberg [12] derived the joint probability distribution of ξ_i and showed that the probability density function for any ξ_i is given by

$$f(\xi_i) = C\left(1 - \xi_i^2\right)^{\frac{(n-p-3)}{2}}, \qquad \xi_i^2 \le 1 \tag{4}$$

where

$$C = \frac{\Gamma\left(\frac{n-p}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n-p-1}{2}\right)}.$$

The test statistic

$$\mathrm{R}_n = \max|r_i|. \tag{5}$$

is called the maximum absolute internally studentized residual statistic. Many classical linear regression diagnostics are built around R_n. The standardized residuals r_i can be used to detect an outlying observation using the probability density function of R_n. When R_n is noticeably large, we declare the component of $\mathbf{Y}$ having that value to be outlying. The distribution of R_n is complex and intractable. Its exact critical values are very hard to compute and available ones are obtained using the first-order Bonferroni upper bound or intensive simulations (see Cook and Prescott [13]).

2 Critical Values of R_n

Many authors have employed different procedures to obtain critical values of R_n. Tietjen et al.[14], following the suggestion of Behnken and Draper [15], used R_n for a single outlier detection in simple linear regression models. Tietjen et al.[14] determined critical values of max $|r_i|$ using an intensive simulation study involving thousands of sampling experiments for a simple linear regression.

Following Ellenberg's [12] suggestion, Prescott [16] obtained a table of upper bounds for the critical values of R_n. Prescott's [16] table entries were calculated from the expression

$$\mathrm{U} = \sqrt{\frac{(n-p)F}{(n-\mathrm{p}-1+F}}, \tag{6}$$

where F is the $100(1 - \frac{\alpha}{n})$ percentage point of the F distribution with 1 and $(n - \mathrm{p} - 1)$ degrees of freedom. For a simple linear regression model (p=2), Prescott's results were found to be almost identical to those of Tietjen et al.[14].

Let r_0 denote an upper bound value of R_n. Lund [17] made use of the Bonferroni inequality to obtain r_0 from the expression

$$\int_{\xi_0}^{1} 2n\, f(\xi_i)\, \mathrm{d}\xi_i = \alpha, \tag{7}$$

where ξ_0 is a value of ξ_i (see equation(3)) and $r_0 = \xi_0\sqrt{n - \mathrm{p}}$.

Ugah et al. [18] set

$$\mathrm{Y}_i = |r_i| \tag{8}$$

and derived the distribution of $\mathrm{Y}_i = |r_i|$ in terms of y_i as

$$f_{|r_i|}(y_i) = \mathrm{H}\left(1 - \frac{y_i^2}{(n-\mathrm{p})}\right)^{\left(\frac{n-p-3}{2}\right)}, \qquad 0 < y_i < \sqrt{n-p} \tag{9}$$

where

$$\mathrm{H} = \frac{2\Gamma\left(\frac{n-\mathrm{p}}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\sqrt{\text{n-p}}\,\Gamma\left(\frac{n-p-1}{2}.\right)}.$$

Then using the Bonferroni inequality, they obtained r_0 by solving

$$\int\limits_{r_0}^{\sqrt{n-p}} n\, f_{|r_i|}(y_i)\, \mathrm{d}y_i = \alpha. \tag{10}$$

Lund [17] combined the concept of Bonferroni inequality and rigorous numerical integration to compute approximate critical values of R_n. The method used by Prescott [16] to obtain r_0 demands that the values of the F-distribution be known. The paucity of percentage points for the F-distribution makes Prescott's method difficult to implement. The method used by Tietjen et al.[14] to obtain r_0 requires an extensive Monte Carlo study. Ugah et al. [18] obtained the distribution of $|r_i|$ and used the concept of Bonferroni inequality and rigorous integration to obtain r_0. The values of r_0 obtained by [16, 17, 18] are identical. The different techniques employed by these authors to obtain r_0 are computationally very demanding, because they have daunting computational requirements, and may not be feasible to apply in large datasets in high dimension. Therefore, a simple and computationally efficient method for computing upper bounds of the critical value of R_n to handle outlying problems in large datasets in high dimention is needed.

Our objective is to propose a new easy-to-use procedure for obtaining r_0, especially for large datasets in high dimension. In this paper we propose an efficient procedure for obtaining r_0 to combat outlying problems in the response variable in linear regression in large datasets. The method we propose does not depend on the number of independent variables or parameters in the model. It does not involve numerical integration or simulation. The new approach is also devoid of mathematical rigor associated with the use of the Bonferroni inequality. It is very easy to apply and is in a very convenient form for immediate use by any researcher. The asymptotic critical values obtained by the proposed approach are shown to exhibit no substantial difference with the ones obtained by Houston [19], especially for large n ($n \geq 500$). The simplicity of the proposed method makes it very efficient and effective in large datasets in high dimension, and hence a very useful tool in applied regression diagnostics.

3 Materials and Methods

In developing the new approach, use is made of asymptotic properties of internally studentized residuals r_i. For most practical problems, especially for large samples, the internally studentized residuals r_i have approximately a standard normal (Gaussian) distribution (see [2, 4, 20]. Let

$$Z_i = |r_i|\,. \tag{11}$$

It can be shown that the distribution of $Z_i = |r_i|$ follows a standard half-normal distribution. If X follows a normal distribution with mean 0 and variance 1, $X \sim N(0,1)$, then $|X|$ follows a (standard) half-normal distribution (see [21, 22]. Since for large n the internally studentized residuals r_i have approximately a standard normal distribution, $r_i \sim N(0,1)$ approximately, consequently, the distribution of $Z_i = |r_i|$ follows a standard half-normal distribution for large n. The large sample probability density function of $Z_i = |r_i|$ in terms of z_i is

$$f_{Z_i}(z_i) = \sqrt{\frac{2}{\pi}} e^{-\frac{z_i^2}{2}}, \qquad 0 < z < \infty \tag{12}$$

while the cumulative density function (CDF) is given by

$$F_{Z_i}(z_i) = \mathrm{Erf}\left(\frac{z_i}{\sqrt{2}}\right), \qquad 0 < z < \infty \tag{13}$$

and Erfc[z] gives the complementary error function erfc(z). Equations (12) and (13) are feature-friendly and do not depend on n and p. Using equations (12) and (13) and the general knowledge of maximum in distribution theory, it can be shown that the probability density function of R_n is given by

$$f_{\mathrm{R}_n}(r_i) = n\left[F_{Z_i}(r_i)\right]^{n-1} f_{Z_i}(r_i), \qquad 0 < r < \infty, \tag{14}$$

3.1 Asymptotic Critical Values of $\mathbf{R}_n$

Houston [19] extended Table 1 in Lund [17] using the same propbability density function (used by Lund [17]) and a different integration procedure. In this paper, we present a straightforward method for obtaining large-sample critical values of R_n and extend the table from 1000 to 1500 data points.

Let r_0^* denote an asymptotic critical value of R_n obtained by using the new approach. To obtain r_0^*, it is necessary to evaluate

$$\int\limits_{r_0^*}^{\infty} n\left[F_{Z_i}(r_i)\right]^{n-1} f_{Z_i}(r_i)\, \mathrm{dr} = \alpha. \tag{15}$$

Equation (15) allows closed form integration and by that means, it can be shown that

$$\int\limits_{r_0^*}^{\infty} n\left[F_{Z_i}(r_i)\right]^{n-1} f_{Z_i}(r_i)\, \mathrm{dr}_i = 1 - \mathrm{Erf}\left(\frac{r_0^*}{\sqrt{2}}\right)^n. \tag{16}$$

Equating (16) to α and solving for r_0^* gives

$$r_0^* = \sqrt{2}\,\mathrm{Erf}^{-1}\left\{(1-\alpha)^{1/n}\right\}. \tag{17}$$

Equation (17) is amenable to evaluation and allows convenient computation of r_0^*. It depends on n and α. Thus, only n and α are needed to obtain r_0^*.

Houston [19] remarked that the approximate critical values of equation (5) become less dependent on the number of independent variables and/or regression parameters in the regression model as n increases, especially for $n > 100$. This observation is reflected in equation (17). The approximate critical values of R$_n$ for significance levels α = $0.1, 0.05$, 0.01 and $n = 500$ to 1000 are presented in Table 1, while values of r_0^* for significance levels α=$0.1, 0.05, 0.01$ and $n = 1000$ to 1500 are presented in Table 2. The computations were implemented by using *Mathematica* software version 12.1.

Table 1. Values of r_0^* for detecting a single outlier in linear regression ($500 \leq n \leq 1000$).

n	0.1	0.05	0.01
500	3.7058	3.8844	4.2638
520	3.7158	3.8939	4.2725
540	3.7253	3.9031	4.2809
560	3.7345	3.9119	4.2890
580	3.7433	3.9203	4.2968
600	3.7518	3.9285	4.3043
620	3.7600	3.9364	4.3116
640	3.7679	3.9440	4.3186
660	3.7756	3.9513	4.3254
680	3.7830	3.9585	4.3319
700	3.7902	3.9654	4.3383
720	3.7972	3.9721	4.3445
740	3.8040	3.9786	4.3505
760	3.8106	3.9850	4.3564
780	3.8170	3.9911	4.3621
800	3.8233	3.9971	4.3676
820	3.8294	4.0030	4.3730
840	3.8353	4.0087	4.3782
860	3.8411	4.0142	4.3834
880	3.8467	4.0196	4.3884
900	3.8522	4.0249	4.3933
920	3.8576	4.0301	4.3980
940	3.8628	4.0352	4.4027
960	3.8680	4.0401	4.4073
980	3.8730	4.0449	4.4117
1000	3.8779	4.0497	4.4161

Table 2. Values of r_0^* for detecting a single outlier in linear regression ($1000 \leq n \leq 1500$).

n	0.1	0.05	0.01
1000	3.8779	4.0497	4.4161
1020	3.8827	4.0543	4.4204
1040	3.8875	4.0588	4.4246
1060	3.8921	4.0633	4.4287
1080	3.8966	4.0676	4.4327
1100	3.9011	4.0719	4.4367
1120	3.9054	4.0761	4.4405
1140	3.9097	4.0802	4.4443
1160	3.9139	4.0843	4.4481
1180	3.9180	4.0882	4.4518
1200	3.9221	4.0921	4.4554
1220	3.9260	4.0960	4.4589
1240	3.9300	4.0997	4.4624
1260	3.9338	4.1034	4.4658
1280	3.9376	4.1071	4.4692
1300	3.9413	4.1106	4.4725
1320	3.9450	4.1142	4.4758
1340	3.9486	4.1176	4.4790
1360	3.9521	4.1211	4.4821
1380	3.9556	4.1244	4.4853
1400	3.9590	4.1277	4.4883
1420	3.9624	4.1310	4.4913
1440	3.9658	4.1342	4.4943
1460	3.9690	4.1374	4.4973
1480	3.9723	4.1405	4.5001
1500	3.9755	4.1436	4.5030

4 Discussion

The approach considered herein is based on the asymptotic properties of the maximum absolute internally studentized residual. A table of asymptotic critical values of R_n is extended from 1000 to 1500 for α = 0.1, 0.01 and 0.05 using a simple procedure. The new method is shown to be independent of the number of regression parameters or independent variables in the model. Our results are almost identical to those of Houston [19] (for $n = 500$), even though the techniques and principles employed in this work and that invoked by Houston [19] are entirely different. Ease of computation is one of the main merits of the new approach over computationally intensive methods. This is a great advantage and is particularly appreciated when large datasets in high dimension are involved. In particular, it can be very computationally efficient in large datasets in high dimension, where more sophisticated methods are hard to apply because of their high computational demands (or requirements). The new result in this work also provides a platform and an approach for describing asymptotic distributions and properties of other diagnostics that consist of the internally studentized residuals. Above all, tables of r_0^* generated using the new approach are very simple and lack the bulkiness and awkwardness associated with tables of upper bounds of R_n (see Tables 1-2 in [19]). These tables of r_0^* can be used for any linear regression diagnostics involving R_n irrespective of the number of parameters or independent variables in the model.

REFERENCES

[1] Atkinson, A., Riani, M., “Influence and Outliers,” Robust Diagnostic Regression Analysis, Springer, 2000, pp. 1-14.

[2] Chatterjee, S., Had, A. S., “Regression Diagnostics: Detection of Model Violations,” Regression Analysis by Examples, Wiley, 2006, pp. 93-123.

[3] Barnett V., Lewis T., “Introduction,” Outliers in Statistical Data, 2nd ed, John Wiley and Sons, 1978, pp. 6-15.

[4] Chatterjee, S., Had, A. S., “Effects of an Observation on a Regression Equation,” Sensitivity Analysis in Linear Regression,” Wiley, 1988, pp. 71-182.

[5] Rousseeuw, P. J., Leroy, A. M., “Outlier Diagnostics,” Robust Regression and Outlier Detection, John Wiley, 1987, pp. 216-245.

[6] Hawkins, D.M., “Introduction,” Identification of Outliers, Springer, 1980, pp. 1-9.

[7] Belsley, D. A., Kuh, E., Welsch, R. E., “Detecting influential Observations and Outliers,” Regression Diagnostics: Identifying Influential Data and Sources of Collinearity, Wiley, 1980, pp. 1-63.

[8] Anthony, A., “Fast very robust methods for the detection of multiple outliers,” Journal of the American Statistical Association, vol. 89, no. 428, pp. 1329-1339. DOI: 10.2307/2290995

[9] Hadi, A. S.; Simonoff, J. S., “Procedures for the Identification of Outliers in Linear Models,” American Statistical Association, vol. 88, no. 424, pp. 1264–1272, 1992. DOI: 10.2307/2291266

[10] Hadi, A. S., “A new measure of overall potential influence in linear regression,” Computational Statistics and Data Analysis, vol. 14, no. 1, pp. 1–27, 1992. DOI: 10.1016/0167-9473(92)90078-T.

[11] Cook, R. D., “Detection of influential observations in linear regression,” Technometrics, vol. 19, no. 1, pp. 15–18, 1977. DOI: 10.1080/00401706.2000.10485981.

[12] Ellenberg J.H., “The joint distribution of the standardized least squares residual from general linear regression,” Journal of the American Statistical Association, vol. 68, no. 344, pp. 941–943, 1973. DOI: 10.2307/2284526.

[13] Cook, R. D., Prescot, P., “On the Accuracy of Bonferroni Significance Levels for Detecting Outliers in Linear Models,” Technometrics, vol. 23, no. 1, pp. 59-63, 1981. DOI: 10.1080/00401706.1981.10486237

[14] Tietjen, G.L., Moore, R.H., Beckman, R. J., “Testing for a single outlier in simple linear regression,” Technometrics, vol. 15, no. 4, pp. 717-721, 1973. DOI: 10.2307/1267383.

[15] Behnken, D.W., Draper, N. R., “Residuals and Their Variance Patterns,” Technometrics, vol. 14, no. 1, pp. 101-111, 1972. DOI: 10.1080/00401706.1972.10488887.

[16] Prescott, P., “An Approximate Test for Outliers in Linear Models,” Technometrics, vol. 17, no. 4, pp. 129-132, 1975. DOI: 10.1080/00401706.1975.10489282.

[17] Lund, R. E., “Tables for an Approximate Test for Outliers in Linear Models,” Technometircs, vol. 17, no. 1, pp. 473-476, 1975. DOI: 10.1080/00401706.1975.10489374.

[18] Ugah, T. E et al., “Upper Bounds of Test Statistics for a Single Outlier Test in Linear Regression Models,” Hindawi Journal of Applied Mathematics, vol. 2021, pp. 1-5, 2021. DOI: 10.1155/2021/1478843.

[19] Houston, B. F., “ Extended tables on an approximate test for outliers in linear models,” Journal of Statistical Computation and Simulation, vol. 21, no. 2, pp. 179-185, 1985. DOI: 10.1080/00949658508810812.

[20] Graybill, F.A., Iye, H.K., “Multiple Regression,” Regression Analysis: Concepts and Applications, pp. 199-349, 1991.

[21] Pewsey, A., “Large-sample inference for the general half-normal distribution,” Commun. stat. Theory Methods, vol. 31, no. 7, pp. 1045–1054,2002.

[22] Ahsanullah, M, Golam Kibria, B. M, Shakil M, “Normal distribution,” Normal and Student’s t Distributions and Their Applications, Atlantis Pres, pp. 7-50, 2014.