Mathematics and Statistics 9(6): 1004-1010, 2021
DOI: 10.13189/ms.2021.090616

http://www.hrpub.org

An Asymptotic Test for A Single Outlier in Linear Regression Models

Ugah Tobias Ejiofor[1], **Mba Emmanuel Ikechukwu**[1,*], **Eze Micheal Chinonso**[1] ,
Arum Kingsley Chinedu[1], **Mba Ifeoma Christy**[2], **Urama Chinasa**[2] ,
Comfort Njideka Ekene-Okafor[3]

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

[2]Department of Economics, Faculty of Social Sciences, University of Nigeria, Nsukka, Enugu State, Nigeria

[3]Department of Computer Sciences/Mathematics, Faculty of Natural Sciences and Environmental Studies, Godfrey Okoye University
Enugu State, Nigeria

Received July 24, 2021; Revised September 26, 2021; Accepted October 17, 2021

Cite This Paper in the following Citation Styles

***(a)**: [1] Ugah Tobias Ejiofor, Mba Emmanuel Ikechukwu, Eze Micheal Chinonso, Arum Kingsley Chinedu, Mba Ifeoma Christy, Urama Chinasa, Ekene-Okafor Comfort, "An Asymptotic Test for A Single Outlier in Linear Regression Models," Mathematics and Statistics, Vol.9, No.6, pp. 1004-1010, 2021. DOI: 10.13189/ms.2021.090616*

***(b)**: Ugah Tobias Ejiofor, Mba Emmanuel Ikechukwu, Eze Micheal Chinonso, Arum Kingsley Chinedu, Mba Ifeoma Christy, Urama Chinasa, Ekene-Okafor Comfort, (2021). An Asymptotic Test for A Single Outlier in Linear Regression Models. Mathematics and Statistics, 9(6), 1004-1010. DOI: 10.13189/ms.2021.090616*

Abstract It is not uncommon to find an outlier in the response variable in linear regression. Such a deviant value needs to be detected and scrutinized to find out why it is not in agreement with its fitted value. Srikantan [1] has developed a test statistic for detecting the presence of an outlier in the response variable in a multiple linear regression model. Approximate critical values of this test statistic are available and are obtained based on the first-order Bonferroni upper bound. The exact critical values are not available and a result of that, tests carried out on the basis of this approximate critical values may not be very accurate. In this paper, we obtained more accurate and precise critical values of this test statistic for large sample sizes (herein called asymptotic critical values) to improve on the tests that use these critical values. The procedure involved using the exact probability density function of this test statistic to obtain its asymptotic critical values. We then compared these asymptotic critical values with the approximate critical values obtained. An application to simulation results for linear regression models was used to examine the power of this test statistic. The asymptotic critical values obtained were found to be more accurate and precise. Also, the test performed better under these asymptotic values (the power performance of this test statistic was found to better when the asymptotic critical values were used).

Keywords Asymptotic, Bonferroni upper Bound, Critical Values, Internally Studentized Residuals, Outlier, Regression, Residual, Test Statistic

1 Introduction

If a value of the response variable deviates considerably from its fitted value than others values deviate from their fitted values, we call such a value an outlier. One of the most popular definitions of an outlier has been given by [2]). He described an outlier as an observation which deviates so much from other observations as it were generated by a different mechanism. Stefansky [3] defined an outlier as one that does not fit in with the pattern in the dataset. According to [4], an outlying value is a value that deviates markedly from other values in the dataset. Johnson et al. [5] defined an outlier as an observation which is inconsistent with the remainder of observation in dataset from which it occurs. An outlying data is caused by such factors as human errors, the erroneous operation of computer systems, sampling errors or standardization failures (see [6]). According to Domańsk [6], numerous statistical methods for outlier detection have been proposed. He recommended that circumspection (cautiouness), double checking, recalculation, etc may help.

Outliers may have adverse effects on data analysis. They may increase variance and reduce the power of statistical tests during data analysis. Rousseeuw et al. [7] point out that they can greatly bias regression analysis. Outliers have a potentially large influence on the results of the statistical inference concerning the models. According to him, the presence of outliers in a dataset can cause inflated error rates and considerable distortions of parameter estimates. Rajarathinam et al. [8] emphasized the importance of identifying outliers in a dataset if they exist so that appropriate measures might be taken.

A good number of test statistics for a single outlier detection in a least squares analysis based on a linear regression model have been developed. However, exact critical values of these test statistics are not available. They are difficult to obtain owing to the complexity of the associated distributions. Approximate critical values of these test statistics are available and are obtained by using the first-order Bonferroni upper bound or large scale simulations. Little or nothing is known about asymptotic critical values of these test statistics. There is a clear need to study asymptotic critical values of these test statistics via their exact distribution. In this work, we implement a suggestion made by [9] to obtain asymptotic critical values of the test statistic found by [1].

The usual form of a multiple linear regression model is

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

where $\mathbf{Y}$ is the $n \times 1$ vector of observations, $\boldsymbol{X}$ an $n \times$ p matrix of constants, $\boldsymbol{\beta}$ is a p $\times$ 1 vector of unknown parameters to be estimated and $\boldsymbol{\varepsilon}$ an $n \times 1$ vector of normally distributed errors. Assuming that $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $Var(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{I}$, the least squares estimator of $\boldsymbol{\beta}$ in (1) is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

and vector of residuals is

$$\begin{aligned}\mathbf{e} =& \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} \\ =& (\boldsymbol{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\boldsymbol{\varepsilon}\end{aligned}$$

The variance-covariance matrix of $\mathbf{e}$ is

$$\text{Var}(\mathbf{e}) = (\boldsymbol{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\sigma^2$$

If σ^2 is estimated using $\hat{\sigma}^2 = \dfrac{\mathbf{e}'\mathbf{e}}{n-p}$, then estimated the variance-covariance matrix $\mathbf{e}$ becomes

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

and the estimate variance of the ith residual e_i is

$$\widehat{Var}(e_i) = s_i^2 = (1 - h_{ii})\hat{\sigma}^2 \tag{3}$$

where h_{ii} is the ith diagonal element of matrix $\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{\mathbf{-1}}\mathbf{X}'$, called the hat matrix (see [10])) and $s_i^2 = (1 - h_{ii})\hat{\sigma}^2$ is the ith diagonal element of $\widehat{Var}(\mathbf{e})$. The ordinary residuals are not all that appropriate for diagnostic purposes and a transformed (standardized) version of them is preferable. This is because, from (2) the variances of the residuals are not constant, but a function of $\mathbf{X}$ matrix which would suggest a standardization of the ith residual e_i. The standardization of e_i has a representation of the form

$$\begin{aligned} r_i =& \frac{e_i}{\sqrt{\{[\dfrac{\mathbf{e}'\mathbf{e}}{(n-p)}][\boldsymbol{I} - \mathbf{X_i}\left(\mathbf{X}'\mathbf{X}\right)^{\mathbf{-1}}\mathbf{X_i'}]\}}} \\ =& \frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}} = \frac{y_i - \hat{y}_i}{\hat{\sigma}\sqrt{1-h_{ii}}} \end{aligned} \tag{4}$$

where $\mathbf{X}_i$ the ith row of $\mathbf{X}$ and $\hat{y}_i$ is the predicted value of y_i. The ith standardized residual R_i is often called a studentized residual. The studentized residuals are basic building block for most of the case test statistics studied in the literature for outlier detection in linear models.

A good number of test statistics for a single outlier detection depends on the studentized residuals. Excellent journal-length treatments include [11], [12],[13], and [11]. Tietjen et al. [11] , following the suggestion of [14], proposed a test procedure for finding the presence of a single outlier in linear regression. They used a simulation study to determine the critical values of the test statistic. Prescott [12] showed that the critical values obtained by [11] are extremely close to those obtained using the first-order Bonferroni upper bound. Following [12] suggestion, [13] obtained elaborate tables of approximate for detecting a single outlier in linear regression critical values. According to [4], [13] has provided the most useful and comprehensive tabulation to date.

Define

$$\xi_i = \frac{r_i}{\sqrt{n-p}} \tag{5}$$

Ellenberg [15] derived the join distribution of $\xi_i's$ and showed that it is a multivariate Inverted Students Function. and showed that the marginal probability density law for any ξ_i is a univariate Inverted-Students Function with probability density function given by

$$f(\xi_i) = C\left(1 - \xi_i^2\right)^{\frac{(n-p-3)}{2}}, \qquad \xi_i^2 \leq 1 \tag{6}$$

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

$$r_n = \max\left|\frac{e_i}{s_i}\right| = \max|r_i| \tag{7}$$

is called the maximum absolute internally studentized residuals. To obtain an upper bound r_0 of the critical value of r_n, [13] made use of the Bonferroni inequality and obtained ξ_0 from the equation

.

$$\int_{\xi_0}^{1} 2n\, f(\xi_i)\, \mathrm{d}\xi_i = \alpha \tag{8}$$

where $\xi_0 = \dfrac{r_0}{\sqrt{n-p}}$ and then obtained the upper bound r_0 of the critical value of the test statistic r_n using the relationship between r_0 and ξ_0 given by the equation

$$r_0 = \xi_0\sqrt{n-p} \tag{9}$$

for sample sizes up to $n = 100$, regression parameters $p = 25$ and $\alpha = 0.10, 0.05$ and 0.01. With that [13] is claimed to have produced the most comprehensive tabulation of the upper bound values r_0 of the critical values of r_n.

Let

$$\begin{aligned} t_i =& \frac{r_i^2}{n-p} \\ =& \frac{e_i^2}{\sum e_i^2(`1 - h_{ii})} \end{aligned} \tag{10}$$

It is well known from the general theory of least squares that t_i follows the beta distribution with Parameters $(\frac{1}{2}, \frac{n-p-1}{2})$ (see [15] and [16]. Let $f(t)$ denote the common distribution and $F(t)$ the cumulative distribution function of $t_i's.$ The cumulative distribution function $F(t)$ is the incomplete beta function with a representation of the form

$$F(t) = I_t\left(\frac{1}{2}, \frac{n-p-1}{2}\right) = \int_0^x \frac{t^{-\frac{1}{2}}(1-t)^{\frac{n-p-1}{2}-1}}{\text{Beta}\left(\frac{1}{2}, \frac{n-p-1}{2}\right)} \tag{11}$$

with $0 < t < 1$. It is verifiablet that

$$\tfrac{d}{dt}F(t) = f(t) \tag{12}$$

Srikantan [1] considered the test statistic

$$t_n = \max(t_1, t_2, t_3, ..., t_n) \tag{13}$$

[1] applied the first-order Bonferroni inequality and obtained upper bound t_0 of the critical value of t_n by evaluating the equation :

$$n\left[1 - I_t\left(\frac{1}{2}, \frac{n-p-1}{2}\right)\right] = \alpha \tag{14}$$

Srikantan [1] presented tables of upper bounds t_0 of the critical values of t_n in Tables 1 to 3 for sample sizes up to n = 20 and regression variables p = 1, 2 and 3. They were computed by solving equations (14) for $\alpha = 0.05$ and 0.01. In the section that follows, we consider obtaining asymptotic critical values of t_n.

2 Materials and Methods

Srikanta [1] proposed the test statistic t_n for detecting a single outlier in linear regression and determined approximate critical of t_n via the use of the first-order Bonferroni upper bound. In this paper, we consider obtaining asymptotic critical value x_0 of t_n by implementing a suggestion made by [9]. Chatterjee et al. [9] suggested that for most practical problems, especially when the sample size is large, that the lack of independence may be ignored. We implement this suggestion to obtain asymptotic critical value x_0 of t_n for large n=200. Then, we compare these asymptotic critical values x_0 with the upper bounds t_0 obtained by [1]. We also obtained the asymptotic critical value x_0 of t_n for small sample sizes up to n=20.

Since the determination critical values for a conventional statistical test requires the knowledge of the distribution of the test statistic when the null hypothesis is assumed to be true, we need the distribution t_n. Our approach to obtaining the distribution of t_n will be anchored on the theory on the distribution of maximum. Since $t_1, t_2, t_3, \ldots, t_n$ have identical beta distributed probability density function and we are assuming now that when the sample size is large $(n \to \infty)$ the lack of independence may be ignored (see [9]) , the distribution of t_n can be obtained. Let $F_{t_n}(x) = P(t_n \leq x)$. Using this assumption and the fact that $t_i's$ have identical beta distribution, it can be shown using a well known theory on the distribution of maximum that the probability distribution function of t_n , denoted herein by $f_{t_n}(x)$, is given in terms x by

$$f_{t_n}(x) = n\left[F_t(x)\right]^{n-1} f_t(x)$$

Explicitly we have

$$\begin{aligned} f_{t_n}(x) = \text{n}&\left[\frac{\text{Beta}\left(\text{x}, \frac{1}{2}, \frac{n-p-1}{2}\right)\Gamma\left(\frac{n-p}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n-p-1}{2}\right)}\right]^{n-1} \\ &\frac{(1-x)^{\left(\frac{n-p-1}{2}\right)-1}}{\sqrt{x}\text{Beta}\left(\frac{1}{2}, \frac{n-p-1}{2}\right)}, \quad 0 < x < 1 \end{aligned} \tag{15}$$

Let x_0 denote the asymptotic critical value of t_n at significance level α. To obtain x_0, it is neccessary to evaluate

$$\text{Prob}(t_n \geq x_0) = \int_{x_0}^{1} f_{t_n}(x)\text{dx} \tag{16}$$

and then set (16) equal to α and solve the resulting equation for x_0. That is, evaluating the equation

$$\int_{x_0}^{1} f_{t_n}(x)\text{dx}=\alpha \tag{17}$$

gives the asymptotic critical value x_0.

2.1 Evaluation of Asmptitic Critical values and Upper bounds of t_n

In Table 1 below, we present asymptotic critical values x_0 and upper bounds t_0 of the critical values of t_n. They were computed by solving equations (14) and (18), at significance level $\alpha = 0.05$ for sample sizes up to n = 100 and regression parameters p $=$ 2 and 3. It is observable from Table 1 that the asymptotic critical values x_0 and the upper bounds t_0 are overwhelmingly close. Figure 1 is a multiple graph of x_0, t_0 versus n for p=3 for large sample sizes. The closeness of the two curves is impressively overwhelming. The values of x_0 and t_0 for sample sizes n=20 and for regression parameters p $=$ 2 and 3 are also displayed in Table 2. It is observed from Table 2 that even for small sample sizes, the values of x_0

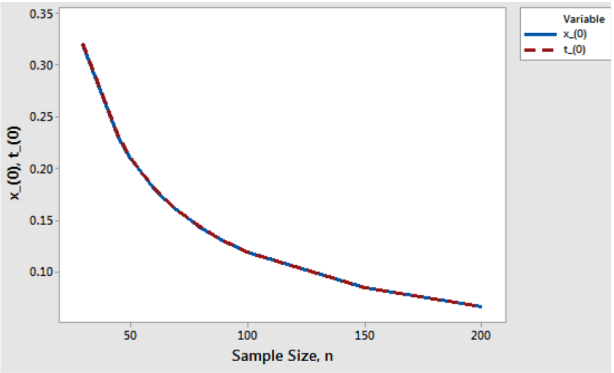

Figure 1. Graph of x_0, t_0 versus sample size $n = 200$; $p = 3$

and t_0 differ but not markedly. Figure 2 is a multiple graph of x_0, t_0 versus n verus for p=3 for small sample sizes. The same closeness between the the two curves as in Figure 1 is also observed. The indistinguishable pattern and closeness exhibited by these graphs (Figures 1 and 2) makes our observation in Tables 1 and 2 more salient and noticeable.

Table 1: Asymptotic critical values and upper bounds for the critical values of t_n

Sample	$\alpha = 0.05, p = 2$		$\alpha = 0.05, p = 3$	
size n	x_0	t_0	x_0	t_0
30	0.3100	0.3111	0.3199	0.3211
45	0.2251	0.2260	0.2300	0.2308
50	0.2069	0.2077	0.2109	0.2117
55	0.1916	0.1923	0.1950	0.1958
60	0.1786	0.1793	0.1815	0.1822
65	0.1673	0.1680	0.1699	0.1705
70	0.1576	0.1582	0.1598	0.1603
75	0.1489	0.1495	0.1509	0.1515
80	0.1413	0.1418	0.1430	0.1435
85	0.1344	0.1349	0.1360	0.1365
90	0.1282	0.1287	0.1296	0.1301
95	0.1226	0.1231	0.1239	0.1244
100	0.1176	0.1180	0.1187	0.1191
150	0.0838	0.0841	0.0844	0.0847
200	0.0658	0.0669	0.0661	0.0663

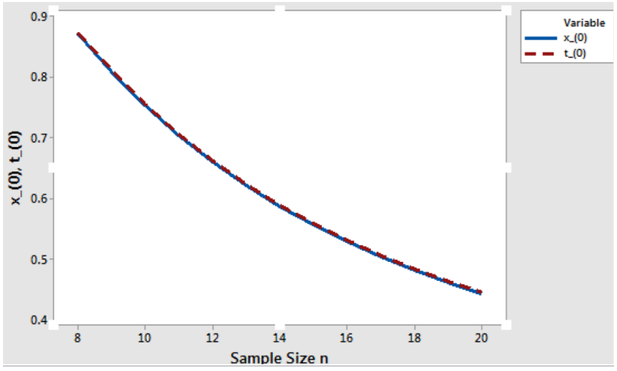

Figure 2. Graph of x_0, t_0 versus sample size $n = 20$; $p = 3$

Table 2: Asymptotic critical values and upper bounds for the critical values of t_n

Sample	$\alpha = 0.05, p = 2$		$\alpha = 0.05, p = 3$	
size n	x_0	t_0	x_0	t_0
8	0.8021	0.8038	0.8723	0.8737
9	0.7462	0.7481	0.8101	0.8125
10	0.6968	0.6987	0.7546	0.7564
11	0.6534	0.6553	0.7046	0.7065
12	0.6151	0.6170	0.6605	0.6624
13	0.5813	0.5831	0.6216	0.6234
14	0.5511	0.5530	0.5871	0.5889
15	0.5242	0.5259	0.5564	0.5582
16	0.4999	0.5011	0.529	0.5308
17	0.478	0.4797	0.5043	0.5060
18	0.4581	0.4597	0.482	0.4837
19	0.4399	0.4416	0.4617	0.4634
20	0.4232	0.4248	0.4432	0.4448

3 Simulation Study

(a)

In this section, some simulation results are presented to study and compare the power performances of the statistic t_n when using the asymptotic critical values x_0 and the upper bounds are used t_0 . All simulations were performed using the statistical software R.

Firstly, we consider a linear regression model with two regression coefficients or parameters:

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

where the true parameters were taken as $\beta_0 = 1$,$\beta_1 = 3$. The values of the explanatory variable X_1 were sampled from a univariate Gaussian distributed population with parameters $\mu = 5$ and $\sigma^2 = 2$ for $n = 30, 45$ and 50. The values X_1 were held constant throughout the simulations. The values of the error term were generated from a Gaussian distributed population with parameters $\mu = 0$ and $\sigma^2 = 1$ and varied throughout the simulations.

The simulation reported here was based on the following procedures: (a) A set of n values of the explanatory variable X_1 was sampled the normal distribution with parameter $\mu = 5$ and $\sigma^2 = 2$. (b) One set of $\hat{Y}$ values of size n was was determined according to the fixed equation $\hat{Y} = 1 + 3X$. (c) Ten thousand (10,000) sets of values of the error term ε values each of size n were generated according to a normal distribution with mean 0 and variance $\sigma = 1$. (d) Then 10,000 sets of Y values each of size n were generated by adding each set of the values of ε to a corresponding set of $\hat{Y}$ values.

In each of these 10,000 samples of y each of size n, we introduced an outlier by subtracting a constant c from the minimum of each sample (we contaminate each sample by subtracting a constant c from the minimum value in that sample $(\min(y_i) - c), i = 1, 2, ..., n)$. We also considered introducing an outlier by adding a constand c to the maximum value in each dataset of the response variable $(\max(y_i) + c), i = 1, 2, ..., n)$. Then, we computed the values of the test statistic t_n by considering only the largest Z_i-value from each dataset. A total of 10,000 values of the test statistic t_n were computed. We measured the power of the test statistic as the percentage of correct rejection of the null hypothesis of no outlier.

For instance, for $c = 2$, $\alpha = 0.05$, $p = 2$, $n = 30$ the asymptotic critical value of t_n is $x_0 = 0.3100$ (see Table 1). Any of the 10,000 computed values of the test statistic t_n that is greater than or equal to the asymptotic critical value 0.3100 is said to be significant at five percent level and thus is declared as an outlie. From the simulation conducted, 2424 out of 10,000 (24.24%) these values were found to be greater than or equal to 0.3100 when $n = 30$ (see Table 3).

Similarly, for $c = 2$, $\alpha = 0.05$, $p = 2$, $n = 30$, we have that $t_0 = 0.3111$ for critical value of t_n computed using Bonferroni inequality (see Table 1). Any of the 10,000 computed values of t_n that is greater than or equal to the critical value 0.3111 is said to be significant at five percent level and thus is declared as an outlier. From the simulation conducted, 2398 out of 10,000 (23.98%) of these values were found to be greater than or equal to 0.3111 (see Table 3). We repeated the procedure for $p = 2$, c=3 and 4 and the results are displayed in Tables 3, 4, and 5.

Table 3: Percentage of outlier detection;p=2

n	$\min(y_i) - 2$		$\max(y_i) + 2$	
	% detect by x_0	% detect by t_0	% detect by x_0	% detect by t_0
30	24.24	23.98	20.81	20.55
45	12.02	11.78	13.18	13.33
50	16.27	16.11	35.04	34.75
60	17.39	17.07	13.18	12.99
70	9.61	9.43	15.75	15.54
80	9-53	9.38	15.51	15.33
90	12.53	12.4	16.9	16.66
100	21.61	21.46	35.37	34.94

Table 4: Percentage of outlier detection;p=2

n	$\min(y_i) - 3$		$\max(y_i) + 3$	
	% detect by x_0	% detect by t_0	% detect by x_0	% detect by t_0
30	62.31	61.86	49.87	49.54
45	31.21	30.84	34.74	34.36
50	46.75	46.46	78.93	78.61
60	44.21	43.86	33.44	33.18
70	25.79	25.48	38.32	37.96
80	26.47	26.15	38.00	37.49
90	32.95	32.59	50.91	50.51
100	63.95	63.54	78.84	78.61

Table 5: Percentage of outlier detection;p=2

n	$\min(y_i) - 4$		$\max(y_i) + 4$	
	% detect by x_0	% detect by t_0	% detect by x_0	% detect by t_0
30	92.1	91.93	81.86	81.54
45	62.98	62.56	67.44	67.11
50	82.96	82.68	98.29	98.19
60	77.65	77.33	66.56	66.15
70	57.11	56.80	71.14	70.89
80	58.07	57.80	71.12	70.90
90	65.59	63.41	87.50	87.31
100	95.96	95.87	98.93	98.89

(b)

Here, we consider a linear regression model with three regression coefficients:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

where the true parameters were taken as $\beta_0 = 1$, $\beta_1 = 3$ and $\beta_2 = 2$. The explanatory variables X_1 and X_2 were generated from a bivariate normal distribution with mean $\mu_1 = 5$, $\sigma_1^2 = 2$, $\mu_2 = 4$, $\sigma_2^2 = 1$ and and covariance $\sigma_{X_1,x_2} = 0.4$ for $n = 30, 45$ and 50. The variables X_1 and X_2 were held constant throughout the simulations. The values of the error term were generated from a Gaussian distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$ and varied throughout the simulations.

The simulation reported here for the case of $p = 3$ was based on the following procedures: (a) A pair of the explanatory variables, (X_1, X_2) of size n was sampled from a bivariate Gaussian distribution with the aforementioned parameters. (b) One set of $\hat{Y}$ values of size n was was determined according to the fixed equation $\hat{Y} = 1 + 3X_1 + 2X_2$. (c) Ten thousand (10,000) sets of ε values each of size n were generated according to a normal distribution with mean 0 and variance $\sigma = 1$. Then 10,000 sets of Y values each of size n were generated by adding each set of values of ε of size n to a corresponding set of $\hat{Y}$ values.

In each of these 10,000 samples of Y each of size n, we introduced an outlier by subtracting a constant c from the minimum of each sample (we contaminate each sample by subtracting a constant c from the minimum value in that sample $(\min(y_i) - c), i = 1, 2, ..., n)$. We also considered introducing an outlier by adding a constand c to the maximum value in each dataset of the response variable $(\max(y_i) + c), i = 1, 2, ..., n)$. Then, we computed the values of the test statistic t_n. A total of

10,000 values of the test statistic t_n were computed. We measured the power of the test statistic as the percentage of correct rejection of the null hypothesis of no outlier.

For example. for $\alpha = 0.05$, $p = 3$, $n = 30$, we have that $x_0 = 0.3199$ for the asymptotic critical value of t_n (see Table 1). Any of the 10,000 computed values of the test statistic t_n that is greater than or equal to 0.3199 is said to be significant at five percent level and thus is declared as an outlier.

From the simulation conducted, 961 out of 10,000 (9.61%) of these values were found to be greater than or equal to 0.3199 (see Table 6).

Similarly, for $\alpha = 0.05$, $p = 2$, $n = 30$, we have that $t_0 = 0.3211$ for critical value of t_n computed using Bonferroni inequality (see Table 1). Any of the 10,000 computed values that is greater than or equal to the tabulated value 0.3211 is said to be significant at five percent level and thus is declared as an outlier. From the simulation conducted, 931 out of 10,000 (9.31%) of these values were found to be greater than or equal to 0.3211 (see Table 6). We repeated the procedure for $p = 3$, c=3 and 4. The results are displayed in Tables 6, 7 and 8.

Table 6: Percentage of outlier detection; p=3

n	$\min(y_i)-2$		$\max(y_i)+2$	
	% detect by x_0	% detect by t_0	% detect by x_0	% detect by t_0
30	9.61	9.31	18.40	18.14
45	24.74	24.47	20.52	20.36
50	17.79	17.53	12.84	12.66
60	11.56	11.34	11.94	11.79
70	14.56	14.37	14.22	14.04
80	11.71	11.52	14.48	14.29
90	15.41	15.14	10.03	9.86
100	14.61	14.44	9.66	9.45

Table 7: Percentage of outlier detection; p=3

n	$\min(y_i)-3$		$\max(y_i)+3$	
	% detect by x_0	% detect by t_0	% detect by x_0	% detect by t_0
30	25.68	25.35	46.71	46.37
45	61.27	60.94	53.74	53.33
50	44.84	44.40	35.50	33.17
60	31.77	31.40	31.00	30.74
70	36.11	35.91	37.93	37.71
80	30.58	30.32	37.32	37.06
90	48.36	47.99	27.51	27.20
100	41.50	41.22	27.31	27.05

Table 8: Percentage of outlier detection; p=3

n	$\min(y_i)-4$		$\max(y_i)+4$	
	% detect by x_0	% detect by t_0	% detect by x_0	% detect by t_0
30	55.04	54.59	76.00	75.72
45	91.70	91.54	86.01	85.78
50	78.95	78.65	65.54	65.23
60	64.32	63.93	63.18	62.83
70	68.77	68.48	72.05	71.77
80	64.01	63.77	71.76	71.53
90	84.51	84.33	58.58	58.52
100	77.04	76.72	60.89	60.67

The percentage of outlier detection (% detect) obtained by using x_0 and t_0 are displayed in Tables 3-8 below. Using the simulations considered herein, the percentage of correct outlier rejection is slightly higher when asymptotic critical values of the test statistic t_n are used than when the upper bounds of t_n are used as can be observed from Tables 3-8. This performance is consistent n=30,45, 50, 60, 70 and 80 and for c=2,3,and 4. Using these results, we may say that that the asymptotic criticals represent a slight improvement over the Bonferroni-inequality generated critical values. It can be shown that even for small sample sizes that the x_0 possesses the same properties over t_0.

4 Conclusions

The principles employed in this work in deriving asymptotic critical values x_0 of the test statistic t_n involved using the exact distribution of this test statistic assuming lack of independence is ignored asymptotically(when $(n \Rightarrow \infty)$, while the principles employed by [1] in obtaining the upper bound t_0 of the critical value of t_n involved using the concept of the Bonferroni inequality. It was found that the asymptotic critical values x_0 agree to to all two decimal places with the upper bounds t_0 of the critical values of t_n obtained by [1]. They exhibit negligible difference from three decimal places upwards. We have observed that asymptotic critical values are consistently smaller than the upper bound t_0 of the critical value of t_n and that the percentage of correct rejection of the null hypothesis is higher when the asymptotic critical values x_0 are used. These two observations have serious implications when reference is made to the power performance of t_n . If the selection of a hypothesis test is to hinge on its ability to distinguish between the two hypotheses being tested (null and alternative hypotheses), then using x_0 is obviously preferred to using t_0. One may say that these asymptotic critical values represent an improvement over the Bonferroni-inequality generated critical values of t_n. In addition, the procedures or methodology used herein provides a very good alternative to the use of the Bonferroni inequality. We may therefore conclude by saying that the suggestion by [9] is a plausible one as the upshots of its implementation (for any value of n) are very favorable.

REFERENCES

[1] Srikantan K.S; “Testing for the Single Outlier in a Regression Model,” Sankhyā: The Indian Journal of Statistics, vol. 23, no. 3, pp. 251-260, 1975. DOI: 10.2307/25049159.”V”

[2] Hawkins D. M, “Introduction,” Identification of Outliers, Springer, pp.1-9, 1989.

[3] Stefansky W; “Rejecting Outliers in Factorial Designs,” Technometrics, vol. 14, no. 2, pp. 469-479, 1972. DOI:

10.1080/00401706.1972.10488930.

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

[5] Johnson R. A, Wirchern D.W, ”The Multivariate Normal Distribution”, Applied Multivariate Statistical Analysis, 2nd ed, Prentice-Hall, 1992, pp.149-209.

[6] Domański P .D; “Study on Statistical Outlier Detection and Labelling,” International Journal of Automation and Computing, vol.17, no. 2, pp. 788–811. 2020. DOI:10.1007/s11633-020-1243-2.

[7] Rousseeuw P J , Leroy A. M, “Introduction,” Robust Regression and Outlier Detection, John Wiley and Sons, 1987, pp.9-18

[8] Rajarathinam A; Vinoth B; “ Outlier detection in simple linear regression models and robust regression-A case study of wheat production data,” International Journal of Scientific Research, vol. 3, no. 2, pp .531-535, 2014. DOI:10.1.1.677.8865.

[9] Chatterjee S, Hadi A. S, “Effect of an Observation on a Regression Equation,” Sensitivity Analysis in Linear Regressin Analysis, John Wiley Sons, 1988, pp. 71-182.

[10] Jobson J. D, “ Influence, Outliers and Leverage,” Multivariate data analysis, 1st ed, Springer, 1991, pp.87-209.

[11] Tietjen G. L; Moore H; Beckman R.J; “ Residuals and Their Variance Patterns,” Technometrics, vol. 15, no. 4, pp. 717-721, 1973. DOI: 10.2307/1267383.

[12] 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.

[13] 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.

[14] 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.

[15] Ellenberg J. H; “The Joint Distribution of the Standardized Least Squares Residuals from a General Linear Regression”. Journal of the American Statistical Association, vol.68, no. 344, pp.941-943,1973. DOI: 10.2307/2284526.

[16] Muller K.E; Mok M; “The Distribution of Cook’s d Statistic,” Communications in Statistics – Theory and Methods, vol. 26, no.3, pp.525-546, 1997. DOI: 10.1080/03610927708831932.