وزارة التعليم العالي والبحث العلمي
جامعة ميسان
كلية التربية الاساسية

مجلة ميسان للدراسات الاكاديمية

العلوم الانسانية والاجتماعية والتطبيقية

ISSN (Paper)-1994-697X
(Online)-2706-722X

المجلد 23 العدد 49 السنة 2024

مجلة ميسان للدراسات الاكاديمية

العلوم الانسانية والاجتماعية والتطبيقية

كلية التربية الاساسية - جامعة ميسان - العراق

ISSN (Paper)- 1994-697X
(Online)- 2706-722X

مجلد (23) العدد (49) اذار (2024)

INTERNATIONAL STANDARD SERIAL NUMBER INTERNATIONAL CENTRE

OJS / PKP
www.misan-jas.com

IRAQI Academic Scientific Journals

doi

Crossref

ORCID

OPEN ACCESS

http://www.misan-jas.com/index.php/ojs

journal.m.academy@uomisan.edu.iq

رقم الايداع في دار الكتب والوثائق بغداد 1326 في 2009

ISSN (Paper) 1994-697X

ISSN (Online) 2706-722X

DOI: https://doi.org/10.54633/2333-023-049-004

Compare Robust Wilk's statistics Based on MM-estimator for the Multivariate Multiple Linear Regression

Thamer Warda Hussein [1], Abdullah A. Ameen [2]
[1] Department of Mathematics, College of Science, University of Basrah, Basra, Iraq
[2] Department of Mathematics, College of Science, University of Basrah, Basra, Iraq
Corresponding author: thamer93work@gmail.com
ORCID: https://orcid.org/0009-0006-0621-6373

Abstract:

In the realm of multivariate linear regression, the classical Wilks' statistic stands out as a widely employed method for hypothesis testing, yet it exhibits high sensitivity to the influence of outliers. Numerous authors have explored non-robust test statistics grounded in normal theories across diverse scenarios. In this investigation, we developed a robust variant of the Wilks' statistics, utilizing the MM-estimator. This approach relies on observation weights determined through Hampel and Huber weight functions. We conducted a comparative analysis between the proposed statistics and the conventional Wilks' statistic. Monte Carlo studies were employed to assess the performance of the test statistics across various datasets, particularly under normal distribution conditions. The study delves into the comparative effectiveness of two test statistics—classical Wilks' and the newly proposed robust statistics. Both exhibited type I error rates and test power close to expected significance levels. However, in scenarios involving data contamination, the proposed statistical method demonstrated superior performance. It emerged as the preferred approach when dealing with corrupted or affected data.

Keywords: MM-Estimator, Outliers, Robustness, P-Value, Wilk's Statistic.

1. Introduction:

Let's suppose we have a q-variate dependent (predictor) vector $\boldsymbol{X_i} = \left(x_{i1}, x_{i2}, \dots, x_{iq}\right)^T)$ and a p-variate independent (response) vector $\boldsymbol{Y_i} = \left(y_{i1}, y_{i2}, \dots, y_{ip}\right)$. The model for multivariate multiple linear regression is expressed as follows:

$$Y = XB + \Xi \qquad (1)$$

Where $Y = (\boldsymbol{Y_1}, \dots, \boldsymbol{Y_n})^T, X = (\boldsymbol{1_n}, (\boldsymbol{X_1}, \dots, \boldsymbol{X_n})^T), \boldsymbol{1_n}$ is a n-dimensional vector whose all entries are 1, B is $((q \times 1) \times p)$ slope matrix and Ξ is $(p \times n)$ errors matrix. Multivariate regression finds practical applications in diverse fields such as engineering, biology, psychology, finance, and many others. Recent research studies on multivariate regression include works by (Friedman and Breiman,1997), (McKean and Davis,1993), (Ollia and

Koivunen,2003). To test the null hypothesis H_0 that there is no significant relationship between the set of dependent variables $\boldsymbol{Y}$ and the set of independent variables $\boldsymbol{X}$, meaning all population regression coefficients are zero, various statistics have been employed. The most widely used is Wilks' statistic Λ, defined as:

$$\Lambda = \frac{|E|}{|E+H|} \tag{2}$$

Where, E and H are given by:

$$E = Y^T Y - \hat{B} X^T Y , \tag{3}$$

$$H = \hat{B} X^T Y - n\bar{y}\bar{y}^T, \tag{4}$$

where

$$\hat{B} = (X^T X)^{-1} X^T Y \tag{5}$$

In the scenario where $\Lambda \leq \Lambda_{\alpha,p,v_E,v_H}$, the null hypothesis H_0 is rejected. Here, $\Lambda_{\alpha,p,v_E,v_H}$ represents the critical values with degrees of freedom $p,\ v_E = n - q - 1$ and $v_H = q$, at a significance level α in Wilks' critical values table. A significant Wilks' statistic, where its associated p-value is below a predetermined significance threshold α, provides evidence to reject the null hypothesis. This implies that, collectively, there is at least one independent variable that significantly influences the set of dependent variables. Assuming that Y follows a multivariate normal distribution, classical statistics are highly sensitive to the impact of outliers (Moller,2005). Several robust estimators of location and scatter in multivariate data, designed to withstand the influence of potential outliers, have been introduced. These include the M-estimator proposed by (Maronna ,1976), the Minimum Covariance Determinant Estimator (MCD) by (Rousseeuw ,1984), and the S-estimator by (Davies ,1987), (Rousseeuw and Leroy,1988), (Lopuhaa and Hendrik, 1989). In high dimensions, (Woodruff and Rocke,1994) explored a robust estimator of location and scatter. The simulation study in section 4 will discuss the impact of outliers on the Wilks' statistic. Therefore, we introduce an alternative robust Wilks' statistic in comparison to the classical Wilks' statistic. The MM-estimator (MM), introduced by (Yohai,1987), is utilized for its robustness in estimating scatter and location matrices. To enhance efficiency while maintaining robustness, we propose re-weighted steps for the MM estimator, as summarized in section 2. Accuracy assessments of the proposed approximations are presented in section 3. Section 4 employs a simulation study to evaluate the performance of the proposed statistics and compare different test statistics in various cases, considering factors such as robustness, test power, and significance level.

2. Robust Estimators:

Estimating the multivariate parameters of the dataset is a prerequisite for constructing robust Wilks' statistics. The MM-estimator proposed by (Yohai,1987) is recognized as an exceptionally robust estimator for both the multivariate scatter matrix and the location matrix. MM-estimator is established through a three-stage process. Initially, a preliminary regression estimate is calculated. This initial estimate is designed to be consistent, robust and possess a high breakdown point in the second phase, an M-estimate of the scale of errors is calculated, utilizing residuals derived from the initial regression estimate. M-estimates are a type of robust estimator that minimizes a certain objective function, often based on a robust function called a psi-function. This step aims to estimate the scale of the errors, providing information about the variability in the data. Lastly, in the third

www.misan-jas.com ISSN (Paper)-1994-697X (Online)-2706-722X

phase, an M-estimate of the regression parameters is computed. This is done using a proper redescending psi-function, which is a type of robust function that assigns less weight to extreme residuals. Redescending psi-functions are crucial for maintaining robustness in the presence of outliers. The three-stage procedure ensures that the estimator is robust at each step of the parameter estimation process. Effective algorithms for computing MM estimates are available in widely-used programming languages such as R, Python, SAS, and MATLAB.

3. The Proposed Wilks' Statistics:

Due to the complicity of the classical Wilks' distribution which was introduced by (Parvin,1958), we will use Bartlett approach for the Wilks' statistics distribution which is defined by (Rencher, 2002):

$$-\left(v_E-\frac{1}{2}\left(p-v_H+1\right)\right)\ln(\Lambda)\simeq\chi^2_{pv_H} \tag{6}$$

(Todorov and Filzmoser, 2010) presented an alternative re-weighted Wilks' statistic which defined as:

$$\Lambda_R=\frac{|\,E_R\,|}{|\,E_R+H_R\,|}\,, \tag{7}$$

where H_R and E_R in our proposed approximation are given by:

$$H_R=Y^T\left(WX(X^TWX)^{-1}X^TW-\frac{J_W}{\sum^n{}_0\,w_i}\right)Y\,, \tag{8}$$

$$E_R=Y^T\;(W-WX\;(X^TWX)^{-1}X^TW\;)\;Y\,, \tag{9}$$

where $W=diag(w_i),\quad i=1,2,\ldots,n,$

$$J_w=w^Tw,\qquad w=(w_1,w_2,\ldots,w_n)^T.$$

In our present study, we propose the following recommendations:

1. Find the estimated location vector $\hat{\mu}$ and the estimated covariance matrix $\hat{\Sigma}$ by using the MM-estimate.

2. Calculate the weights w_i of the observation $\boldsymbol{Y_i}$ by using the Hampel weight function (Campbell, 1980) and Huber weight function those defined as:

$$w_i=\begin{cases}1, & MD(\boldsymbol{Y_i})\le d_0\\ \frac{d}{MD(\boldsymbol{Y_i})}\,, & MD(\boldsymbol{Y_i})>d_0,\end{cases} \tag{10}$$

where

$$d=d_0e^{-\frac{1}{2}\left(\frac{MD(Y_i)-d_0}{b_2}\right)},\;\; d_0=\sqrt{p}+\frac{b_1}{\sqrt{2}},\quad b_1=2,\;\; b_2=1.25,$$

and $MD(Y_i)$ is the Mahalanobis distances which given by:

www.misan-jas.com ISSN (Paper)-1994-697X (Online)-2706-722X

$$MD(Y_i) = \sqrt{(\boldsymbol{Y_i} - \hat{\mu}^{\boldsymbol{0}})^{\boldsymbol{T}} \left(\hat{\Sigma}^0\right)^{-1} (\boldsymbol{Y_i} - \hat{\mu}^0)} \quad (11)$$

And w_i with Huber weight function is given by:

$$w_i = \begin{cases} 1, & MD(\boldsymbol{Y_i}) \le \sqrt{\chi^2_{0.975}(p)}, \\ 0, & wtherwise \end{cases} \quad (12)$$

3. For $j = j + 1$ calculate the weighted estimated location matrix $\hat{\mu}$ and the weighted estimated covariance matrix $\hat{\Sigma}$ as following:

$$\hat{\mu}^j = \frac{1}{\sum_1^n w_i} \sum_1^n w_i \boldsymbol{Y_i} \quad (13)$$

$$\hat{\Sigma}^j = \frac{1}{\sum_1^n w_i - 1} \sum_1^n w_i \left(\boldsymbol{Y_i} - \hat{\mu}^j\right)\left(\boldsymbol{Y_i} - \hat{\mu}^j\right)^T \quad (14)$$

4. Iterate through steps two and three until the following criterion is met:

$$\frac{\left|\hat{\Sigma}^{j+1}/p\right|^p}{det\left(\hat{\Sigma}^{j+1}\right)} \le \frac{\left|\hat{\Sigma}^{j}/p\right|^p}{det\left(\hat{\Sigma}^{j}\right)}$$

Now, we will introduce robust versions of Wilks' statistics, denoted as of Λ_{R1} and Λ_{R2}, similar to Λ_R in (7). These statistics rely on MM-estimators with Huber and Hampel weight functions, and we will construct their approximate distribution.

$$-\left(v_{E_R} - \frac{1}{2}\left(p - v_{H_R} + 1\right)\right) ln\Lambda_{R_W} \simeq \chi^2_p v_{H_R} \quad (15)$$

The degrees of freedom for v_{E_R} , and v_{H_R} the robust Wilks' statistic Λ_R can be found as follows:

$$v_{H_R} = trace\left(WX(X^TWX)^{-1} X^TW - \frac{J_W}{\sum_1^n w_i}\right) \quad (16)$$

$$v_{E_R} = trace(W - WX(X^TWX)^{-1} X^TW) \quad (17)$$

Now, let's utilize the QQ-plots technique to assess the accuracy of Λ_{R1} and Λ_{R2} through simulations involving k=3000 samples from the multivariate normal distribution. We will consider various scenarios, including different values for p (number of dependent variables), q (number of independent variables), and n (sample size). The accuracy of the classical distribution of the k statistics will be compared to the approximate distribution of Λ_{R1} and Λ_{R2} using QQ-plots. Some of these plots will be included in (Fig.1). The standard cut-off values for a test, 95%, 97.5%, and 99%, are presented in the plots as vertical lines. The plots, give an evident that the approximations are accurate across all dimensions considered in the study $p \in 2,3,4$ and $q \in 2,3,5$, for both large and small sample sizes, and across varying correlation levels r and $r = 0$.

ISSN (Paper)-1994-697X (Online)-2706-722X

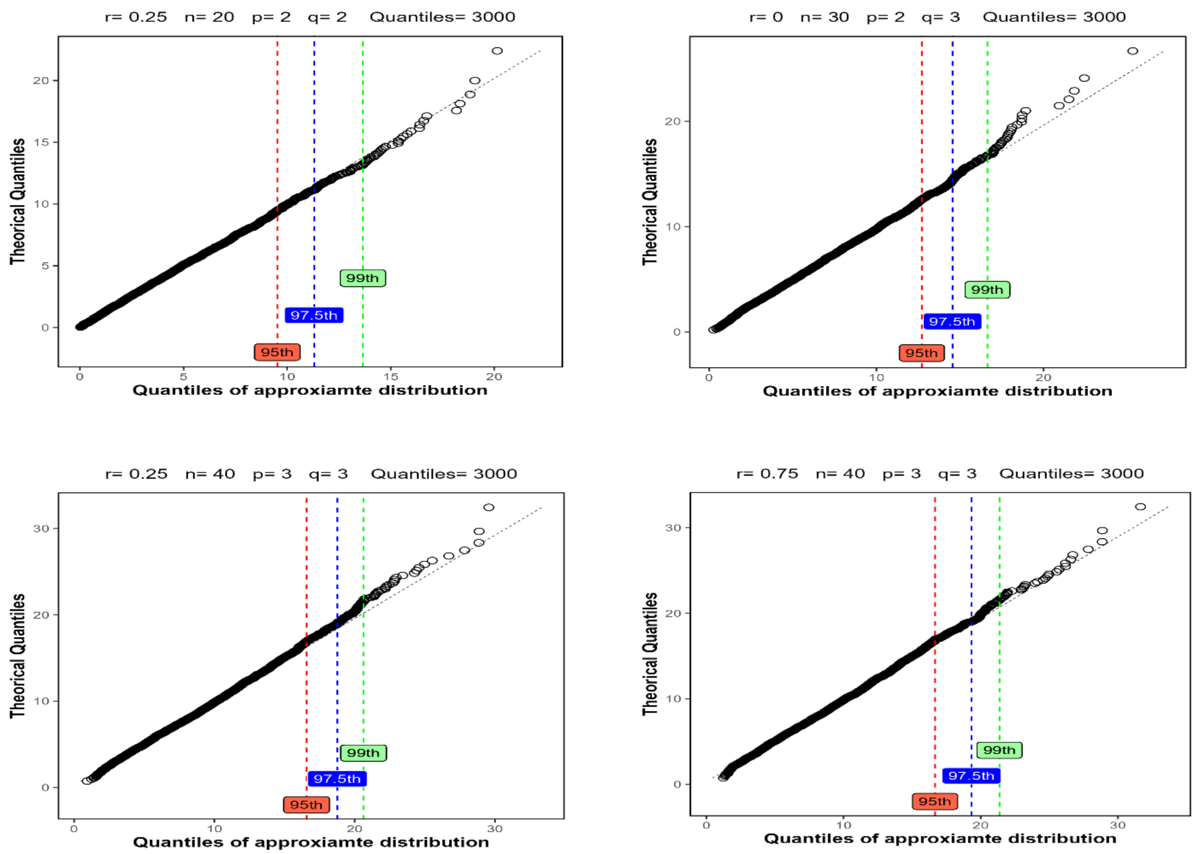


Figure 1. Λ_{R_1} ***and*** Λ_{R_2} QQ-plots

4.Monte Carlo Simulation:

The Monte Carlo process proves to be a valuable technique for evaluating the effectiveness of test statistics. We will assess the performance of these statistics based on both the type I error rate and the power of the test. By employing these measures, we will compare the behavior of robust statistics with the classical Wilks' statistic. This comparison will be conducted under two scenarios: one where the data includes outliers and another where the data follows a completely normal distribution. To study the type I error rate and the power of the test, we will consider various cases, including the number of independent variables $p = 2, 3, 4$, dependent variables $q = 2, 3, 5$, sample sizes $n = 20, 30, 40$, and correlations between components of the dependent variable Y. These correlations will be explored in terms of no correlation $r = 0$, medium correlation $r = 0.5$, and strong correlation $r = 0.75$.

4.1 Significance Level:

To assess and compare the rates of type I error for the test statistics under consideration, we generate observations of varying sizes n from both a multivariate normal distribution $\boldsymbol{Y_i} \sim N_p(\mathbf{0}, \boldsymbol{I})$ and a distribution containing outliers using the following model:

$$\boldsymbol{Y_s} \sim N_p(\mathbf{0}, \mathbf{I}), \qquad s = 1,2, \ldots, [m],$$

$$\boldsymbol{Y_t} \sim N_p(\mu^*, c\boldsymbol{I}), \qquad t = [m] + 1, \ldots, n$$

ISSN (Paper)-1994-697X (Online)-2706-722X

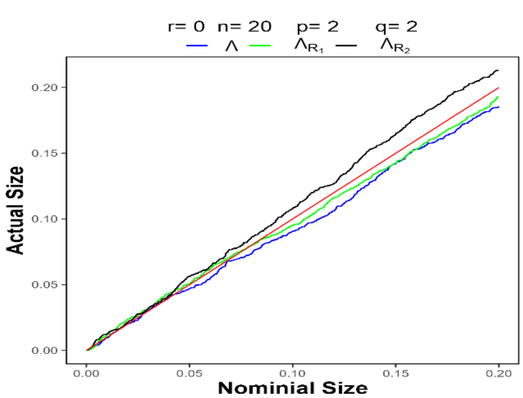
where $m = \frac{80n}{100}$, $\lceil m \rceil$ is the largest positive integer that is not less than m, $\mu^* = v^2 \sqrt{\chi^2_{p,0.001}}\, \mathbf{1}_p^T$, $v = 5, c = 0.0625$ and $\mathbf{1}_p$ is a vector whose all entries are 1. This is done under the null hypothesis $H_0 : B_1 = \mathbf{O}$, where B_1 is a matrix consisting of all rows of the coefficient matrix B except the first row. The classical Wilks' statistic Λ is compared to Bartlett's χ^2 formula in equation (7), and the robust Wilks' statistics proposed in this study are compared to the approximate distribution (Huber and Hampel) outlined in section 3. This simulation process is repeated $k = 3000$ times to calculate $\hat{\alpha} = \frac{T(k)}{k}$ (where $T(k)$, is the number of times the null hypothesis is rejected when it is true). The values of $\hat{\alpha}$ serve as estimates of the threshold significance level when the simulated critical values exceed the chosen significance level of 0.05. The nominal level interval is obtained from Fawcett and Salter's standard error formula (Salter and Facucett,1989). $\alpha \pm 2 \times \sqrt{\frac{\alpha(1-\alpha)}{k}}$ providing a standard deviation interval around the nominal level. We adopt P-value plots proposed by (Davidson and McKinnon,1998), as they offer a comprehensive representation of how the test statistics adhere to the approximate distribution under the null hypothesis across simulated samples. (Fig.2) depict the P-value plots, showing that the statistics Λ , Λ_{R_1} and Λ_{R_2} closely align with the 45° line.

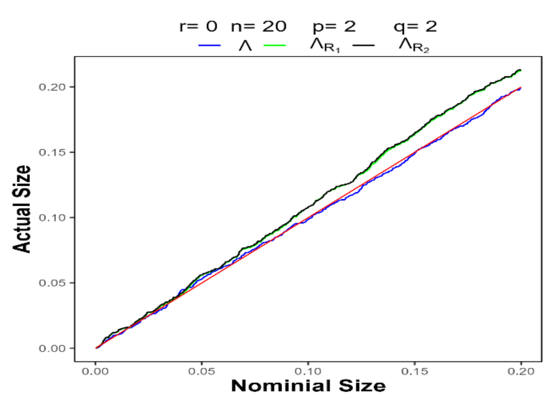
4.2 Power of the test:

To assess the power of the test for the considered statistics and make comparisons between them, we generate observations of varying sizes n from both a multivariate normal distribution $\boldsymbol{Y_i} \sim N_p(\mathbf{0}, \boldsymbol{I})$ and a distribution containing outliers using the model:

$$\boldsymbol{Y_s} \sim N_p(\mathbf{0}, \mathbf{I}), \qquad s = 1,2,\dots,\lceil m \rceil,$$
$$\boldsymbol{Y_t} \sim N_p(\mu^*, c\boldsymbol{I}), \qquad t = \lceil m \rceil + 1,\dots,n$$

where $m = \frac{80n}{100}$, $\lceil m \rceil$ is the largest positive integer that is not less than m, $\mu^* = v^2 \sqrt{\chi^2_{p,0.001}}\, \mathbf{1}_p^T$, $v = 5, c = 0.0625$ and $\mathbf{1}_p$ is a vector whose all entries are 1. This is done under the alternative hypothesis. We use Davidson and MacKinnon's method to compare the resulting power-size curves. (Fig.3) illustrates that the statistics Λ , Λ_{R_1} and Λ_{R_2} closely align with each other in terms of power size, considering normally distributed datasets without outliers. In cases where the data is corrupted and exhibits no or low correlation between dependent variables, Λ_{R_1} and Λ_{R_2} outperform Λ. However, in scenarios of high correlation, Λ , Λ_{R_1} and Λ_{R_2} show similar performance, as depicted in (Fig.4 and Fig.5).





ISSN (Paper)-1994-697X (Online)-2706-722X

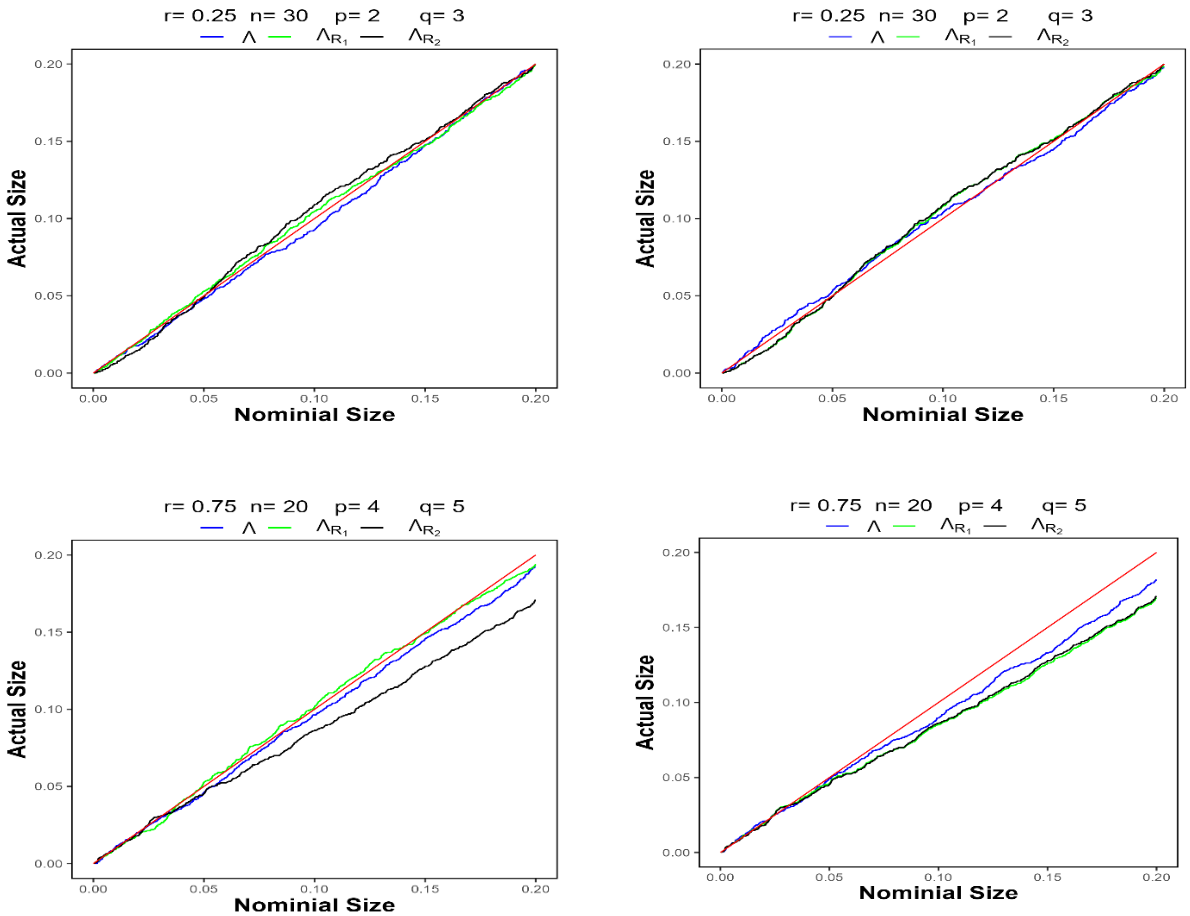
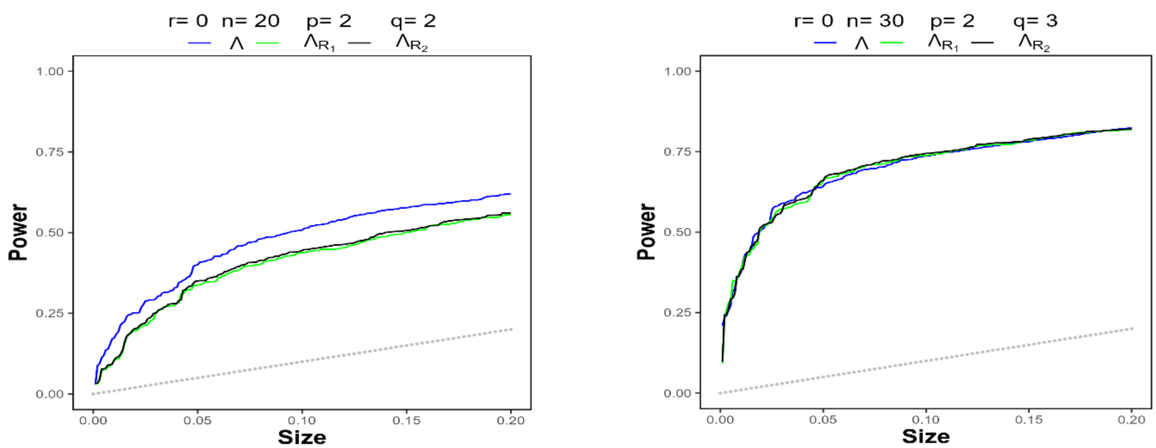


Figure 2. P-value plots of Λ, Λ_{R_1} *and* Λ_{R_2}



ISSN (Paper)-1994-697X (Online)-2706-722X

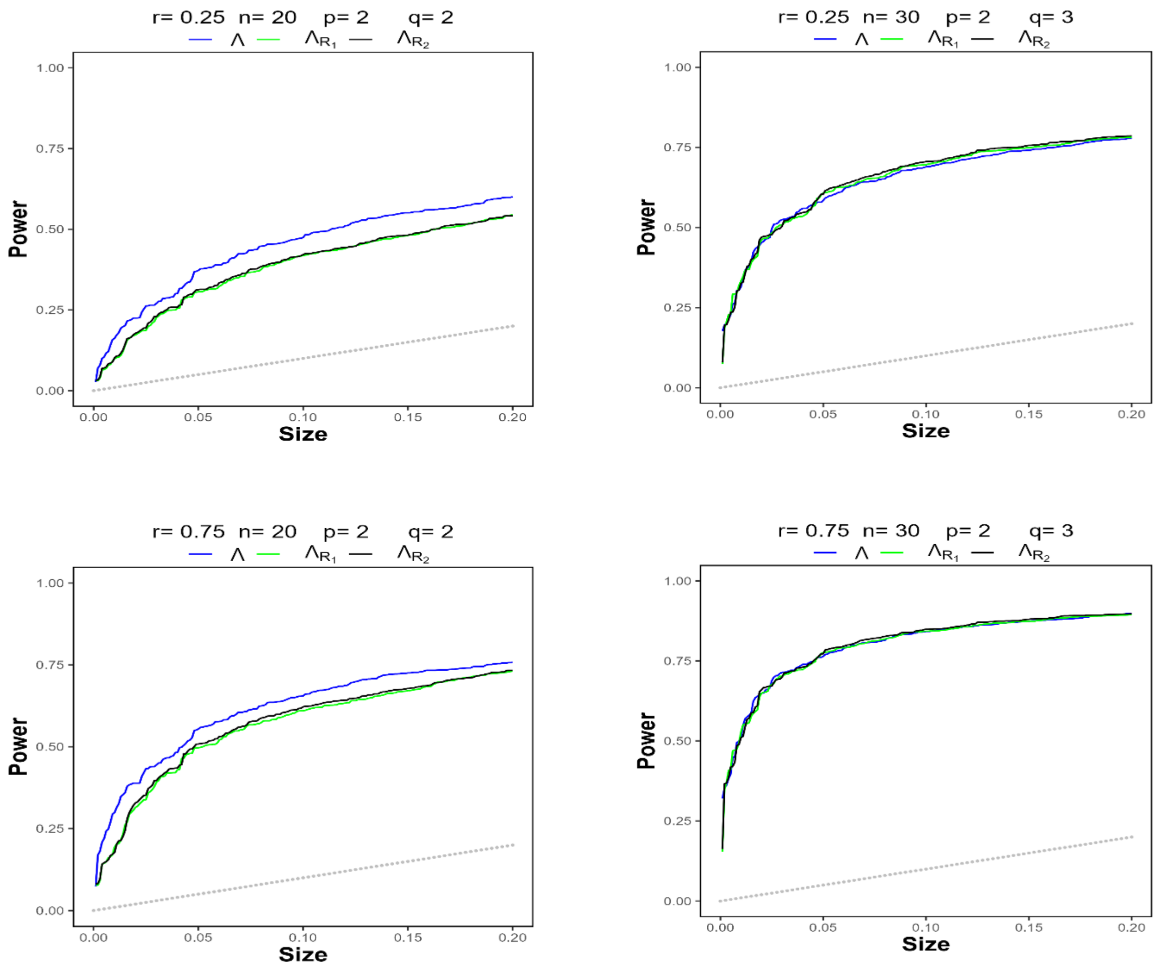
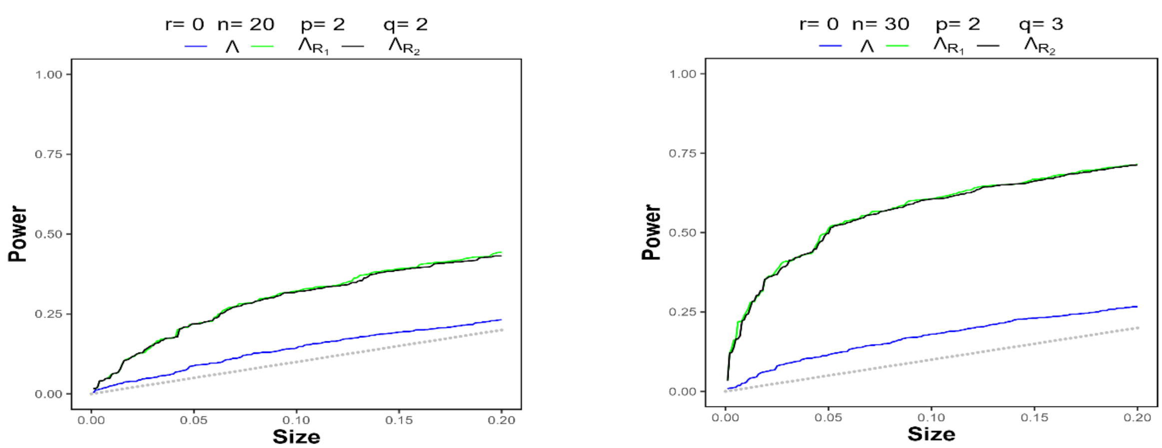


Figure 3. Curves of size power, in case of the data is fully normal



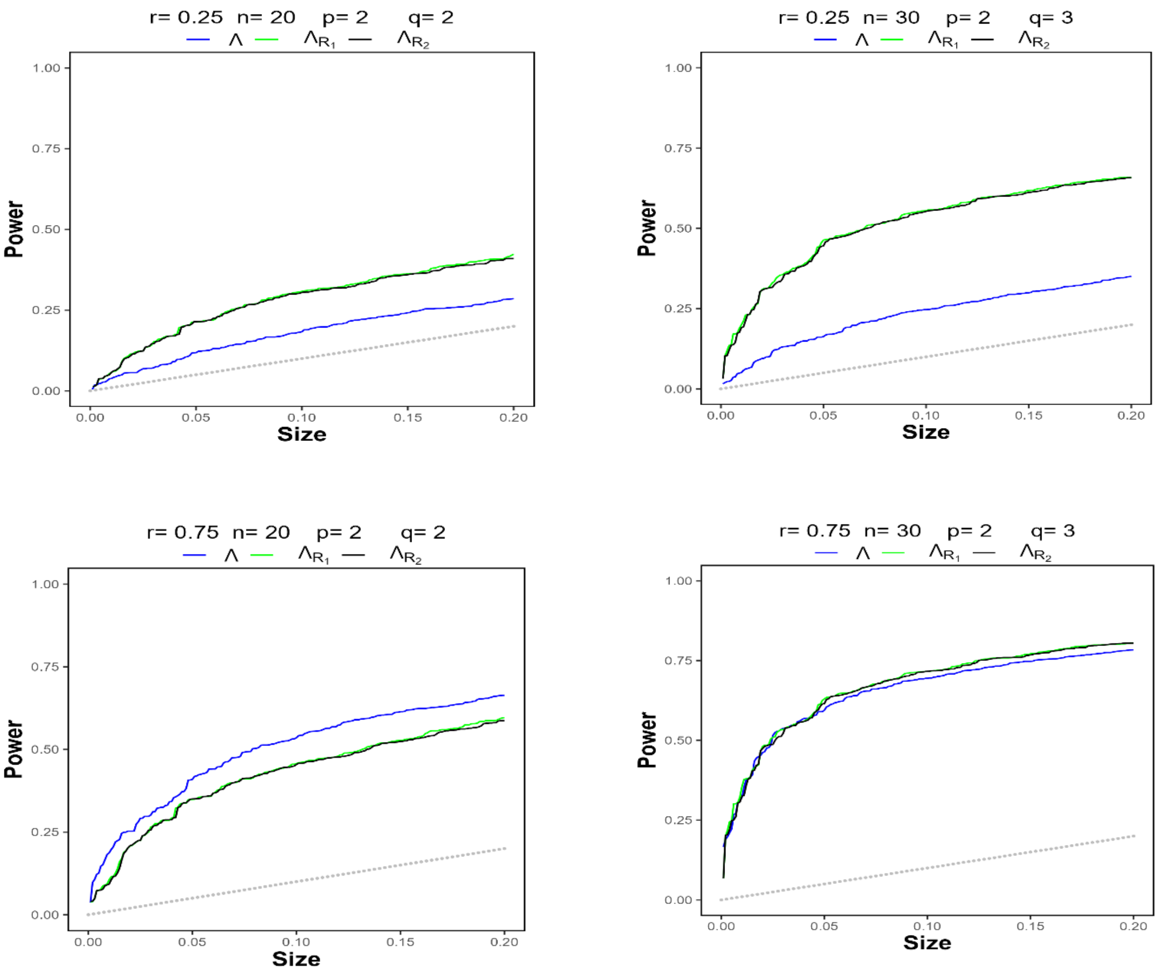
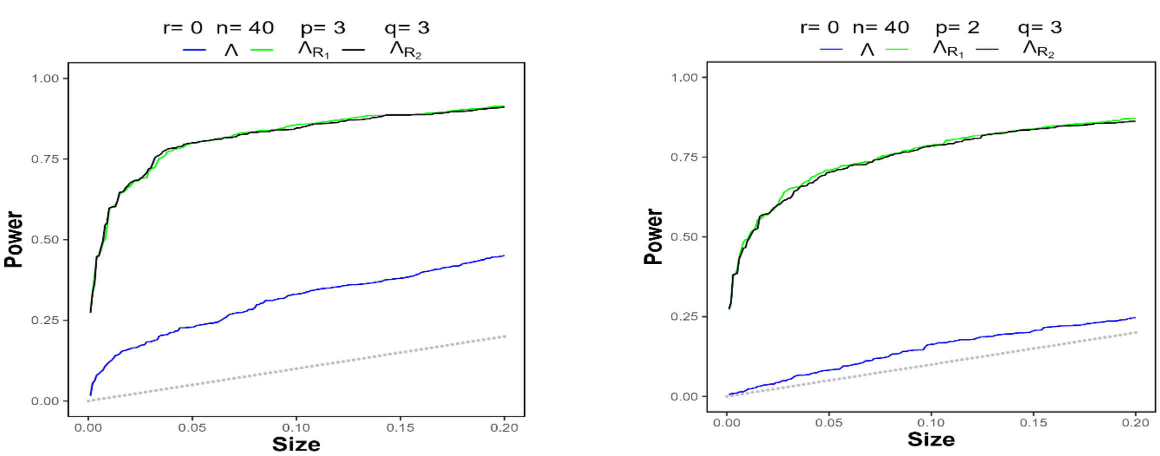


Figure 4. Curves of size power, in case of the data contains outliers



www.misan-jas.com
ISSN (Paper)-1994-697X (Online)-2706-722X

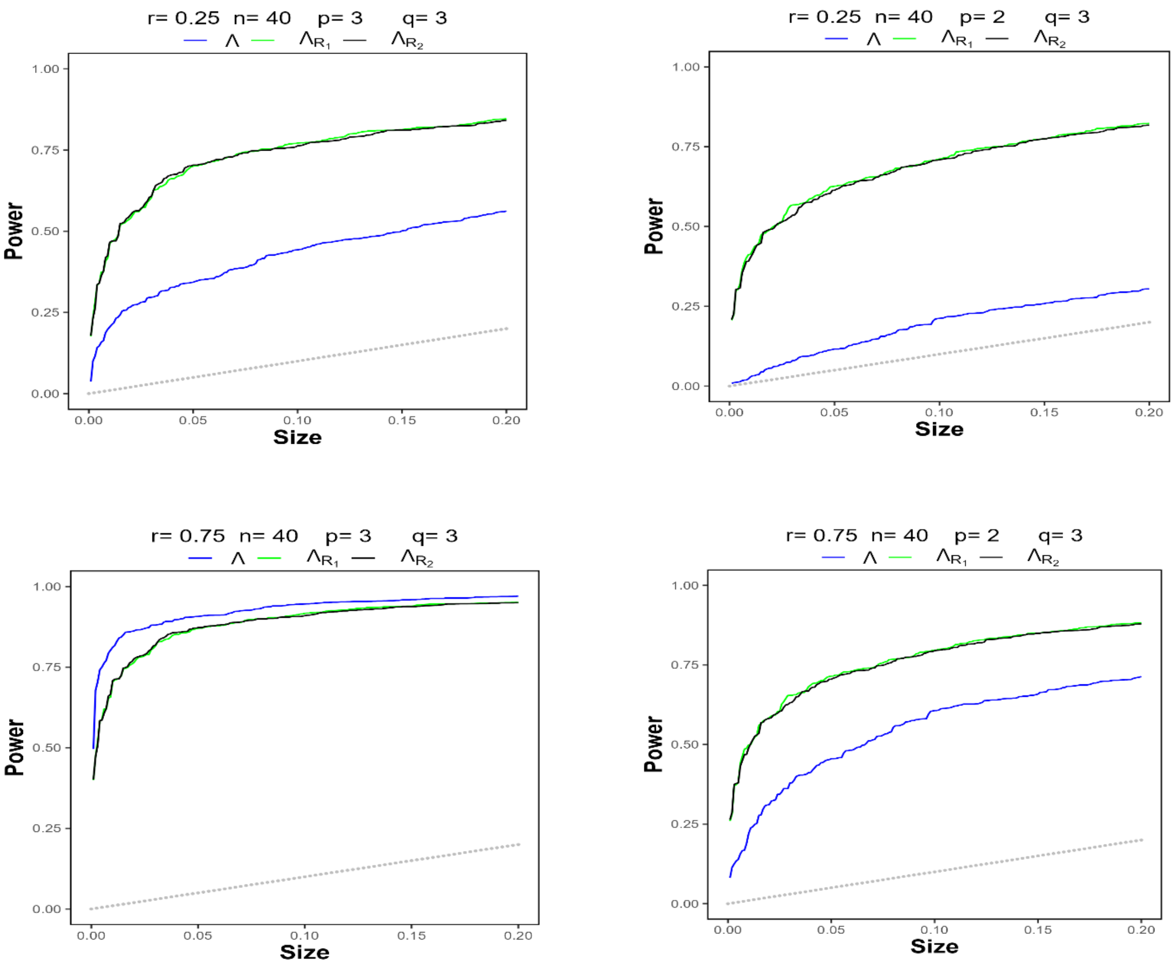


Figure 5. Curves of size power, in case of the data contains outliers

5. Conclusion:

We have presented improved robust versions of Wilks' statistic, denoted as, Λ , Λ_{R_1}, built upon the MM-estimator. Their approximate distributions were derived. The findings suggest that, under normal data distribution, the proposed statistics closely resemble the classical Wilks' statistic Λ. However, in the existence of a distribution with contaminants, the proposed statistics outperform the classical Wilks' statistic.

References:

Breiman, L., and Friedman, J. H. Predicting multivariate responses in multiple linear regression. Journal of the Royal Statistical Society Series B: Statistical Methodology 59, 1 (1997), 3–54. https://doi.org/10.1111/1467-9868.00054

Campbell, N. A. Robust procedures in multivariate analysis i: Robust covariance estimation. Journal of the Royal Statistical Society Series C: Applied Statistics 29, 3 (1980), 231–237. https://doi.org/10.2307/2346896

Davidson, R., and MacKinnon, J. G. Graphical methods for investigating the size and power of hypothesis tests. The Manchester School 66, 1 (1998), 1–26. https://doi.org/10.1080/01621459.1993.10594316

Davies, P. L. Asymptotic behaviour of s-estimates of multivariate location parameters and dispersion matrices. The Annals of Statistics (1987), 1269–1292. https://doi.org/10.1016/0047-259X(92)90028-E

Davis, J. B., and McKean, J. W. Rank-based methods for multivariate linear models. Journal of the American Statistical Association 88, 421 (1993), 245–251. https://doi.org/10.1111/1467-9957.00086
Lopuhaa, H. P. On the relation between s-estimators and m-estimators of multivariate location and covariance. The Annals of Statistics (1989), 1662–1683. https://doi.org/10.1214/aos/1176347386
Maronna, R. A. Robust m-estimators of multivariate location and scatter. The annals of statistics (1976), 51–67. https://doi.org/10.1214/aos/1176343347
Møller, S. F., von Frese, J., and Bro, R. Robust methods for multivariate data analysis. Journal of Chemometrics: A Journal of the Chemometrics Society 19, 10 (2005), 549–563. https://doi.org/10.1002/cem.962
Ollila, E., Oja, H., and Koivunen, V. Estimates of regression coefficients based on lift rank covariance matrix. Journal of the American Statistical Association 98, 461 (2003), 90–98. https://doi.org/10.1198/016214503388619120
Parvin, C. A. An Introduction to Multivariate Statistical Analysis, 3rd ed. T.W. Anderson. Hoboken, NJ: John Wiley, 742 pp., 99.95, hardcover.ISBN0 − 471 − 36091 − 0.ClinicalChemistry50, 5(052004), 981 − −982. https://doi.org/10.1373/clinchem.2003.025684
Rencher, John Wiley, S. Methods of Multivariate Analysis, Second Edition. Brigham Young University, 2002. https://doi.org/10.1002/0471271357
Rousseeuw, P. J. Least median of squares regression. Journal of the American statistical association 79, 388 (1984), 871–880. https://doi.org/10.1080/01621459.1984.10477105
Rousseeuw, P. J., and Leroy, A. M. A robust scale estimator based on the shortest half. Statistica Neerlandica 42, 2 (1988), 103–116. https://doi.org/10.1111/j.1467-9574.1988.tb01224.x
Salter, K., and Fawcett, R. A robust and powerful rank test of treatment effects in balanced incomplete block designs. Communications in Statistics-Simulation and Computation 14, 4 (1985), 807–828. https://doi.org/10.1080/03610918508812475
Todorov, V., and Filzmoser, P. Robust statistic for the one-way manova. Computational Statistics & Data Analysis 54, 1 (2010), 37–48. https://doi.org/10.1016/j.csda.2009.08.015
Woodruff, D. L., and Rocke, D. M. Computable robust estimation of multivariate location and shape in high dimension using compound estimators. Journal of the American Statistical Association 89, 427 (1994), 888–896. https://doi.org/10.1080/01621459.1994.10476821
Yohai, V. J. High breakdown-point and high efficiency robust estimates for regression. The Annals of statistics (1987), 642–656. https://www.jstor.org/stable/2241331

مقارنه إحصائيات ويلكس الحصينة استنادًا إلى مقدر MM للانحدار الخطي المتعدد متعدد المتغيرات

[1] ثامر ورده حسين، [2] عبد الله عبد القادر امين
[1] قسم الرياضيات، كليه العلوم، جامعه البصرة، البصرة، العراق
[2] قسم الرياضيات، كليه العلوم، جامعه البصرة، البصرة، العراق

المستخلص:
في الانحدار الخطي متعدد المتغيرات، تبرز إحصائية ويلكس الكلاسيكية كطريقة مستخدمة على نطاق واسع لاختبار الفرضيات، ومع ذلك فهي تظهر حساسية عالية لتأثير القيم المتطرفة. لقد درس العديد من المؤلفين إحصائيات اختبار غير حصينه ترتكز على نظريات عادية عبر سيناريوهات متعددة. في هذا الدراسة، قمنا بتطوير طريقه حصينه لإحصائيات ويلكس، باستخدام مقدر (MM) ويعتمد هذا النهج على أوزان المشاهدات التي يتم تحديدها من خلال داله وزن هامبل وهيوبر. لقد أجرينا تحليلاً مقارناً بين الإحصائيات المقترحة وإحصائيات ويلكس الكلاسيكية. تم استخدام طريقه مونت كارلو لتقييم أداء إحصاءات الاختبار عبر مجموعات البيانات المختلفة. في حال ان البيانات تتبع التوزيع الطبيعي أظهر الطرق المقترحة معدلات خطأ من النوع الأول قريبة من مستويات الأهمية المتوقعة وقوة اختبار مشابهه للطريقة الكلاسيكية. ومع ذلك، في السيناريوهات التي تنطوي على تلوث البيانات، أظهرت الطريقة الإحصائية المقترحة أداء متفوقا. حيث ان الاحصائيات الحصينة المقترحة هي النهج المفضل عند التعامل مع البيانات المتأثرة بالقيم المتطرفة.