Süleyman Demirel Üniversitesi
Fen Bilimleri Enstitüsü Dergisi
Cilt 27, Sayı 1, 84-89, 2023

Süleyman Demirel University
Journal of Natural and Applied Sciences
Volume 27, Issue 1, 84-89, 2023

DOI: 10.19113/sdufenbed.1141519

Vulnerability of the Tukey M Robust Regression Method Against Multicollinearity

Filiz KARADAĞ[1], Hakan Savaş SAZAK[*2]

[1,2] Ege University, Faculty of Science, Department of Statistics, 35100, İzmir

(Alınış / Received: 06.07.2022, Kabul / Accepted: 14.02.2023, Online Yayınlanma / Published Online: 25.04.2023)

Keywords
Condition index,
Correlation,
Least squares,
M estimators,
Variance inflation factors

Abstract: In this study, we investigate whether the Tukey M robust regression method provides a solution for the data sets suffering from multicollinearity problem. It is observed that high values of variance inflation factors (VIF) which is a sign of the multiple linear link among the explanatory variables, cannot be controlled by the robust methods which work through the residual values. The reason for this fact is that multicollinearity and high values of VIF which is a result of multicollinearity do not produce extreme residuals. For this reason, the robust methods cannot provide a solution for the high VIF problem. This fact is shown by an extensive simulation study. In the simulation study, the explanatory variables were derived from trivariate normal distribution for three different correlation values. In this study, we also used two real-life data examples and we observed that the results support the findings of the simulation study. For all these reasons, we can conclude that specialized methods should be utilized in the case of multicollinearity.

Tukey M Dayanıklı Regresyon Yönteminin Çoklu Doğrusal Bağlantıya Karşı Zafiyeti

Anahtar Kelimeler
Koşul indeksi,
Korelasyon,
En küçük kareler,
M tahmin edicileri,
Varyans şişirme faktörü

Öz: Bu çalışmada, Tukey M dayanıklı regresyon metodunun, çoklu doğrusal bağlantı problemine sahip veri setleri için bir çözüm sunup sunmadığı araştırılmıştır. Çalışmada açıklayıcı değişkenler arasında çoklu doğrusal bağlantı göstergesi olan yüksek VIF (varyans şişirme faktörü) değerlerinin, artık değerler üstünden çalışan dayanıklı metotlarla kontrol edilemediği gözlenmiştir. Bunun sebebi çoklu doğrusal bağlantının ve bunun sonucu olan yüksek VIF değerlerinin ekstrem artık değerler üretmiyor olmasıdır. Dolayısıyla dayanıklı metotlar yüksek VIF problemine bir çözüm sunamamaktadır. Bu durum kapsamlı bir simülasyon çalışması ile gösterilmiştir. Simülasyon çalışmasında üç farklı korelasyon değeri için üç değişkenli normal dağılıma sahip açıklayıcı değişkenler üretilmiştir. Çalışmada ayrıca iki gerçek hayat veri örneği kullanılmış ve sonuçların simülasyon bulgularını desteklendiği görülmüştür. Tüm bu sebeplerden dolayı çoklu doğrusal bağlantı durumunda özel yöntemlerin kullanılması gerektiği sonucunu çıkarabiliriz.

1. Introduction

Multicollinearity can be defined as the high linear relationship among two or more explanatory variables. It is crucial to understand the causes and the extent of the multicollinearity. Thus, it should be determined whether the cause of the multicollinearity is the nature of the variables or the consequence of the data collection method which can be helpful in finding the remedies for this problem [1].

In a multiple regression analysis, first, it should be detected if multicollinearity exists because it has many adverse effects on the regression analysis. In the case of the existence of multicollinearity, the regression coefficients, extra sum of squares, the variability of the estimated regression coefficients, the fitted values, the predictions and the simultaneous tests of beta can be negatively affected [2]. Additionally, even if a definite statistical relationship exists between the dependent variable and the set of the predictor variables, many of the estimated regression coefficients individually may be statistically insignificant [3].

The measurement of the marginal effect of the explanatory variables is not easy since the marginal

* Corresponding author: hakan.savas.sazak@ege.edu.tr

contribution of a predictor variable in reducing the error sum of squares can be affected by other variables which are already in the regression model. This is due to the fact that under multicollinearity, the explanatory variables that are already included in the model contain almost the same information [3]. The most well-known effect of the multicollinearity is its capability to inflate the variances of the estimators of the regression coefficients which also constitutes a barrier to establish the regression model correctly [2].

There are many tools to detect multicollinearity. Checking the scatterplots and correlations between the explanatory variables can be useful but we should keep in mind that correlation and multicollinearity are not the same things, thus, there can still be multicollinearity even when all the correlations are low. A similar diagnostic tool is to examine the whole correlation matrix of the explanatory variables. This gives the opportunity to see all the correlations at once but as we mentioned before, this is not enough to determine the existence of multicollinearity. Fortunately, in addition to these, several multicollinearity detection methods have been developed [3]. Now let us sort the eigenvalues of the variance-covariance matrix of the p explanatory variables (Σ) in descending order as $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ (see Chatterjee and Hadi [4] for details). If at least one of the eigenvalues is close to zero, there is a serious multicollinearity [5].

Many other symptoms of multicollinearity can be observed including a small determinant of the correlation matrix, improbable signs or size of the estimators of the regression coefficients, unexpected magnitudes of the standard errors of the regression coefficients and large confidence intervals of the regression coefficients [2, 5]. The sum of the reciprocals of λ_k, k=1,2,...,p is also used as a multicollinearity diagnostic measure. If the sum of them is larger than 5p, then multicollinearity is present. This rule is given below

$$\sum_{k=1}^{p}\left(\frac{1}{\lambda_k}\right) > 5p. \qquad (1)$$

Another measure of multicollinearity is the condition index and the kth condition index is found by

$$CI_k = \sqrt{\frac{\lambda_{max}}{\lambda_k}}\ , \quad k = 1,2,\dots p. \qquad (2)$$

The greater the condition index, the higher the multicollinearity is. If the condition index is between 10 and 30, a moderate multicollinearity is expected while the condition index being greater than 30 indicates a high multicollinearity [4, 6, 7].

There is another detection method suggested by Marquaridt [5] which is called variance inflation factors (VIF). VIF are the diagonal elements of the inverse of the variance-covariance matrix of the explanatory variables after the correlation transformation. Usage of VIF is widely recognized for detecting the presence of multicollinearity. VIF can be accepted as a tool in measuring the amount of inflation in the variances of the regression coefficients when the explanatory variables are linearly related compared to the case when they are linearly independent [3].

VIF can be mathematically shown below

$$\text{VIF} = diag(r_{xx}^{-1}) = diag\left(\left(X^{*'}X^{*}\right)^{-1}\right) \qquad (3)$$

$$\text{VIF}_k = \frac{1}{1-R_k^2}\ for\ k = 1,2,\dots,p, \qquad (4)$$

where X^* is the matrix of the explanatory variables after correlation transformation and R_k^2 is the coefficient of the multiple determination of X_k on the remaining explanatory variables. The larger the VIF, the more variance of the estimators of the regression coefficients is inflated and so higher the severity of the multicollinearity is [3].

In order to handle the multicollinearity problem, there are many suggestions in the literature. One approach to dealing with multicollinearity is to collect more information or additional data but this may not be possible in most of the situations and even if it is possible it may not solve the problem if the additional data also possess the same problem. Removing one or more explanatory variables from the model, defining new predictors or respecification of the model are other remedies [8]. Using alternative estimation methods which are not influenced unfavorably by multicollinearity as Least Square (LS) is another remedy. One is the ridge regression method which was proposed by Hoerl and Kennard [9] as an alternative to the LS method. There is another method called principle component approach which is based on working with the eigenvalues and eigenvectors of the correlation matrix of the explanatory variables [4]. There are also some studies focused on using robust ridge regression methods but in this study, we investigate whether the Tukey M robust estimation for the regression coefficients provides a simpler solution for the adverse effects of multicollinearity in regression analysis [10, 11]. To do so, we conducted a simulation study including the LS method and the Tukey M robust estimation method for the regression coefficients and examine the effects of multicollinearity on the regression analysis based on them. As a classical robust estimation method we used the Tukey M estimators by utilizing MATLAB Robustfit module. Basically, we checked the differences in the variances of the regression coefficients produced by these methods. We will give more detailed information about the methods used in

this study in Section 2. Section 3 will present the simulation results and the related comments. Two real-life data examples are given in Section 4 for illustration. The final section includes discussion and some concluding remarks.

2. Material and Method

The general linear regression (GLR) model can be given as follows

$$Y_{n\times 1} = X_{n\times(p+1)}\beta_{(p+1)\times 1} + \varepsilon_{n\times 1} \quad (5)$$

Here, Y is the vector of the response variable, $X$is the matrix of the explanatory variables, β is the vector of the regression parameters, ε is the vector of the error term, n is the sample size and p is the number of slope parameters. The assumptions related to Eq. (5) are

$$E(\varepsilon) = 0\,,$$

$$Var(\varepsilon) = \sigma^2 I\,,$$

$$rank(X) = p+1\,,$$

where I is the identity matrix. Many estimators for the regression parameters were suggested in the literature. In this study, we include two of them, the LS estimators and one of the most commonly used robust estimators, the Tukey M estimators, for the regression parameters [12]. It is also reported by Yu and Yao [12] that the Tukey M estimators achieve both robustness and high efficiency for regression models. Here, we intend to observe the differences if any between the classical estimators and the Tukey M robust estimators for the regression parameters.

The philosophy of the LS method is obtaining the estimators by minimizing the sum of the squared errors. Theoretically, the LS method can be defined as follows

$$min \sum_{i=1}^{n} \varepsilon_i^2. \quad (6)$$

Since $\varepsilon = Y - X\beta$ from Eq. (5), we can also express Eq. (6) by using the matrix format as

$$\min\ (Y - X\beta)'(Y - X\beta). \quad (7)$$

Taking derivative of Eq. (7) w.r.t. β and equating it to zero gives the following estimator as the LS estimator of β

$$\hat{\beta} = (X'X)^{-1}X'Y. \quad (8)$$

The variance-covariance matrix of the LS estimator of β is

$$\mathrm{Var}(\hat{\beta}) = \sigma^2(X'X)^{-1}. \quad (9)$$

The M estimators were found by Huber [13]. The principle of the M estimation is the minimization of the sum of a selected ρ function of the errors instead of the sum of squares of them. More specifically, the M estimators are found by minimizing the following expression

$$\min \sum_{i=1}^{n} \rho(\varepsilon_i). \quad (10)$$

The M estimate for a given sample can be obtained by solving the equation given below

$$\sum_{i=1}^{n} \rho'(\varepsilon_i) = 0 \ \left(or \ \sum_{i=1}^{n} \psi(\varepsilon_i) = 0 \right). \quad (11)$$

We used the following bisquare function ψ in this study

$$\psi(u_i) = \begin{cases} u_i(1-u_i^2)^2 \ if\ |u_i| \le 1 \\ 0 \ if\ |u_i| > 1 \end{cases} \quad (12)$$

where $u_i's$ are the standardized residuals. There are many proposals in the standardization of the residuals. One needs to select a robust estimator of scale to do so. The most popular one is the re-scaled median absolute deviation (MAD). The procedure used in this study is given below which is the default option of the robustfit module of Matlab

$$u_i = \frac{r_{i,adj}}{k^* s}\,, r_{i,adj} = \frac{r_i}{\sqrt{(1-h_{ii})}},$$

$$k = 4.685\,, s = \frac{MAD(r_i)}{0.6745},$$

$$MAD(r_i) = median|r_i - median(r_i)|, \quad (13)$$

where k is the tuning constant, $r_i's$ are the raw residuals and $h_{ii}'s$ are the leverage values. The constant 0.6745 makes the scale estimation unbiased under normal distribution [13, 14].

3. Simulation Results

In order to compare the estimators mentioned in this paper, a simulation study was conducted including two different levels of sample sizes and several correlation levels for the explanatory variables. In this study, all the programs were written in Matlab for the GLR model given in Eq. (5) but for simplicity, the simulations were conducted for the model given below

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon. \quad (14)$$

In this model, β_0 is the intercept, β_1, β_2 and β_3 are the slope parameters and ε is the error term. We took $\beta_0 = 0$ and $\beta_1 = \beta_2 = \beta_3 = 1$ without loss of

generality. Simulations were conducted for *nn*=(10000/*n*) Monte Carlo runs and for the sample sizes *n*=50 and 100 with the correlation levels $\rho = 0$, 0.95 and 0.98. Since we have observed that high VIF values can be obtained at least with the correlation value of 0.95, we did not conduct simulations between 0 and 0.95. We simulated the samples with independent and identically distributed error terms from standard normal distribution and with the explanatory variables having trivariate standard normal distribution with several correlation levels as specified earlier. The simulation results are given in Tables 1 and 2. We conducted the simulations for two different levels of sample sizes to observe the possible effect of the increase in the sample size. Depending on the simulation results, first we should note that other than the natural effect of the sample size on the variance of the estimators, we did not observe any difference between the results for two different sample sizes. Second, as expected, we did not observe any bias for the estimators for any situation. Regarding to the correlation levels, we observe that as the correlation increases, the variances of the estimators of the slope parameters $(\beta_1, \beta_2, \beta_3)$ have a tendency to increase for both LS and robust estimators. For β_0, we do not observe any difference between the correlation levels. When we compare the LS and robust estimators in terms of their variances, the best performance is shown by the LS estimators even for high correlation levels. This fact shows that the Tukey M robust method cannot be a remedy for the multicollinearity problem. The reason of this fact is that the Tukey M robust estimators are based on the residuals but based on our observations, multicollinearity does not affect residuals.

Table 1. The simulated values for *n*=50 with three levels of correlations

		$\rho = 0$		$\rho = 0.95$		$\rho = 0.98$	
Estimator:		LS	Robust	LS	Robust	LS	Robust
bias	β_0	-0.0005	-0.0011	0.0014	0.0011	0.0033	0.0034
	β_1	0.0030	0.0029	0.0035	0.0026	0.0043	0.0051
	β_2	-0.0008	-0.0006	-0.0012	0.0002	0.0012	0.0010
	β_3	-0.0026	-0.0024	-0.0021	-0.0029	-0.0054	-0.0060
nxvar	β_0	1.0684	1.1288	1.0691	1.1247	1.0541	1.1199
	β_1	1.0978	1.1678	15.1204	16.1211	37.1836	39.9254
	β_2	1.1155	1.1905	14.5919	15.5283	36.5866	39.0565
	β_3	1.1068	1.1724	14.8718	15.8277	38.1899	40.3686
nxmse	β_0	1.0685	1.1289	1.0692	1.1248	1.0547	1.1205
	β_1	1.0982	1.1682	15.1210	16.1214	37.1845	39.9267
	β_2	1.1155	1.1905	14.5920	15.5283	36.5866	39.0566
	β_3	1.1071	1.1727	14.8720	15.8281	38.1913	40.3704

Table 2. The simulated values for *n*=100 with three levels of correlations

		$\rho = 0$		$\rho = 0.95$		$\rho = 0.98$	
Estimator:		LS	Robust	LS	Robust	LS	Robust
bias	β_0	-0.0002	0.0000	-0.0019	-0.0020	-0.0005	-0.0002
	β_1	-0.0002	-0.0003	0.0020	0.0032	0.0064	0.0087
	β_2	0.0021	0.0019	0.0005	0.0002	0.0007	0.0003
	β_3	0.0012	0.0011	-0.0011	-0.0021	-0.0084	-0.0103
nxvar	β_0	1.0379	1.0940	1.0409	1.1026	1.0203	1.0832
	β_1	1.0559	1.1093	14.4161	15.2945	35.3720	37.2375
	β_2	1.0437	1.1041	14.0303	14.8097	35.9366	37.9109
	β_3	1.0308	1.0893	14.2437	15.1038	35.7273	37.6315
nxmse	β_0	1.0379	1.0940	1.0413	1.1029	1.0203	1.0832
	β_1	1.0559	1.1093	14.4165	15.2955	35.3761	37.2451
	β_2	1.0441	1.1045	14.0303	14.8097	35.9366	37.9109
	β_3	1.0310	1.0894	14.2438	15.1042	35.7343	37.6421

4. Applications

In this section, we give two real-life data examples for the illustration and comparison between the LS and Tukey M robust estimators.

4.1. Body Fat Data

The data set which is based on body fats was investigated in detail by Kutner et al. [3]. Body fat data set contains three explanatory variables (triceps skinfold thickness (X_1), thigh circumference (X_2) and midarm circumference (X_3), all in cm.) with a sample size of 20. The dependent variable is the body fat percentage (Y). We obtained the maximum condition index value as 53.33 and the VIF values as 708.84, 564.34 and 104.61. We also examined the scatter plots and the correlation matrix of the explanatory variables. All the information gathered here indicates that multicollinearity exists for this data set. Now we give the regression coefficient estimates and the standard errors of the regression estimators for the LS and robust estimators in Tables 3 and 4 respectively. In Table 3 we do not see much difference between the LS and robustfit estimates. Table 4 shows that the standard errors of the LS estimators are smaller than their counterparts. This result is very consistent with the simulations we have conducted. In the simulations we have found that the LS estimators have a better performance than the Tukey M robust estimators in terms of their variances.

Table 3: The regression estimates for "the body fat data set"

	LS	Robust
β_0	117.0847	122.5170
β_1	4.3341	4.4953
β_2	-2.8568	-2.9953
β_3	-2.1861	-2.2733

Table 4: The standard errors of the regression estimators

	LS	Robust
β_0	99.7823	108.4804
β_1	3.0166	3.2711
β_2	2.5884	2.8107
β_3	1.5811	1.7321

4.2. Longley Data

Another real-life data example with multicollinearity problem is the Longley econometric data set. It consists of a macroeconomic data set with 7 economic variables observed annually from 1947 to 1962 and includes 6 explanatory variables with 16 observations (GNP (Gross National Product) (X_1), number of people unemployed (X_2), number of people in the armed forces (X_3), population (≥ 14 years old) (X_4), years (X_5), number of people employed (X_6)). The dependent variable is the GNP Deflator percentage (Y) [15].

We obtained the maximum condition index value as 15159.33 and the VIF values as 1214.57, 83.96, 12.16, 230.91, 2065.73 and 220.42. The processes followed in the previous example are also followed for this example. Tables 5 and 6 show the regression coefficient estimates and the standard errors of the regression estimators for the LS and robust estimators, respectively.

Table 5: The regression estimates for "the Longley data set"

	LS	Robust
β_0	2946.856	2686.2
β_1	0.2635	0.2589
β_2	0.0365	0.0355
β_3	0.0112	0.011
β_4	-1.7370	-1.7377
β_5	-1.4188	-1.2833
β_6	0.2313	0.2011

Table 6: The standard errors of the regression estimators

	LS	Robust
β_0	5647.977	6053.459
β_1	0.1082	0.1159
β_2	0.0302	0.0324
β_3	0.0155	0.0166
β_4	0.6738	0.7222
β_5	2.9446	3.156
β_6	1.3039	1.3976

According to Table 6, the results are consistent with the simulation results and the previous real-life data example. We observe that the standard errors of the Tukey M estimators are larger than the LS estimators.

5. Discussion and Conclusion

The main focus of this study is to investigate whether the Tukey M robust estimation method enables us to handle the regression analysis in the presence of multicollinearity. First, in the simulations we have found that one has to take at least a correlation value of 0.95 between the explanatory variables to obtain high VIF values. The most important result is that the classical robust estimators such as the Tukey M estimators cannot be a remedy for the multicollinearity problem. The real-life data examples also supported this fact. The reason of this fact is that the Tukey M robust regression method is focused on the residuals but we observed that multicollinearity does not cause an increase in the residuals (in magnitude) and thus the Tukey M robust regression estimators cannot cope with this problem. This shows that specialized methods should be utilized for the data sets possessing multicollinearity.

Declaration of Ethical Code

In this study, we undertake that all the rules required to be followed within the scope of the "Higher Education Institutions Scientific Research and Publication Ethics Directive" are complied with, and that none of the actions stated under the heading "Actions Against Scientific Research and Publication Ethics" are not carried out.

References

[1] Hocking, R.R., Pendleton, O.J. 1983. The regression dilemma. Commun. Stat. Theory Methods, 12(5), 497-527.

[2] Mansfield, E.R., Helms, B.P. 1982. Detecting multicollinearity. Am. Stat., 36, 158-160.

[3] Kutner, M.H., Nachtsheim, C.J., Neter, J., Li, W. 2004. Applied Linear Statistical Models, 5th edn. McGraw Hill, New York.

[4] Chatterjee, S., Hadi, A.S. 2012. Regression Analysis by Example, 5th edn. John Wiley and Sons, New Jersey.

[5] Marquaridt, D.W. 1970. Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. Technometrics 12(3), 591-612.

[6] Belsley, D.A., Klema, V.C. 1974. Detecting and assessing the problems caused by multicollinearity: A use of the singular-value decomposition. NBER Working Paper Series, 66.

[7] Belsley, D.A., Kuh, E., Welsch, R.E. 1980. Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. John Wiley and Sons, New York.

[8] Montgomery, D.C., Askin, R.G. 1981. Problems of nonnormality and multicollinearity for forecasting methods based on least squares. AIIE Trans. 13(2), 102-115.

[9] Hoerl, A.E., Kennard, R.W. 1970. Ridge regression. Biased estimation for nonorthogonal problems. Technometrics. 12(1), 55-67.

[10] Holland, P.W. 1973. Weighted Ridge Regression: Combining Ridge and Robust Regression Methods. NBER Working Paper Series. 11.

[11] Askin, R.G., Montgomery, D.C. 1980. Augmented Robust Estimators. Technometrics. 22, 333-341.

[12] Yu, C., Yao, W. 2017. Robust linear regression: A review and comparison. Communications in Statistics - Simulation and Computation. 46(8), 6261-6282.

[13] Huber, P.J. 1964. Robust estimation of a location parameter. Ann. Math. Stat. 35, 73-101.

[14] Elsaid, H., Fried, R. 2015. Tukey's M-estimator of the Poisson parameter with a special focus on small means. Stat. Methods Appl. 25, 191–209.

[15] Becker, R. A., Chambers, J. M., Wilks, A. R. 1988. The New S Language. Wadsworth & Brooks/Cole.