Journal of Statistics Applications & Probability
An International Journal

http://dx.doi.org/10.18576/jsap/090208

A Comparative Study of Robust Estimators for Poisson Regression Model with Outliers

Mohamed Reda Abonazel and Omnia Mohamed Saber*

Department of Applied Statistics and Econometrics, Faculty of Graduate Studies for Statistical Research, Cairo University, Giza, Egypt

Received: 18 Apr. 2019, Revised: 1 Oct. 2019, Accepted: 4 Oct. 2019
Published online: 1 Jul. 2020

Abstract: The present paper considers Poisson regression model in case of the dataset that contains outliers. The Monte Carlo simulation study was conducted to compare the robust (Mallows quasi-likelihood, weighted maximum likelihood) estimators with the nonrobust (maximum likelihood) estimator of this model with outliers. The simulation results showed that the robust estimators give better performance than maximum likelihood estimator, and the weighted maximum likelihood estimator is more efficient than Mallows quasi-likelihood estimator.

Keywords: Count regression models, Generalized linear models, Mallows quasi-likelihood, M-estimation, Weighted maximum likelihood

1 Introduction

Poisson regression model is a basic count data model in the generalized linear models (GLMs). The maximum likelihood (ML) estimator for this model has been provided by McCullagh and Nelder [1].

Outliers are one of the significant statistical issues, but most people do not know how to deal with them. Most parametric statistics (like means, standard deviations, and correlations) are sensitive to outliers. Outliers can mess up your analysis. It is well known that the ML estimator for GLMs is very sensitive to outliers, see [2,3,4]. In different regression models, it is necessary to use a robust estimator to detect outliers and to provide resistant stable results in the presence of outliers, see [5,6,7]. However, despite the fair amount of the present pieces of literature, robust inference for GLMs seems to be very limited.

Preisser and Qaquish [8] considered a class of robust estimation for Poisson regression models based on quasi-likelihood in the general framework of generalized estimating equation, but Cantoni and Ronchetti [9] showed that these estimators are not robust. They also proposed a robust approach based on the natural generalizations of quasi-likelihood function. Hosseinian and Morgenthaler [10] proposed another robust estimator for Poisson regression models based on weighted maximum likelihood.

The present paper investigates the efficiency of two robust estimators of Poisson regression model, and compares them with maximum likelihood estimator.

This paper is organized as follows: Section Two provides Poisson regression model and ML estimator. Section Three presents the robust estimators for this model. Section Four displays the simulation results. Section Five involves the concluding remarks.

2 Poisson Regression Model and ML Estimator

The Poisson distribution is discrete probability of count of the events which randomly occur in a given interval of time. Density function of this distribution is

$$f(Y=y|\mu)=\frac{e^{-\mu}\mu^{y}}{y!}; \qquad y=0,1,2,\ldots \tag{1}$$

* Corresponding author e-mail: mabonazel@hotmail.com

In this distribution, the mean must equal the variance. Thus, the usual assumption of homoscedasticity would not be appropriate for Poisson data, see [11]. Based on a sample $\{y_1,...,y_n\}$, we can write the model in terms of the mean of the response $(E(y_i)=\mu)$:

$$y_i = E(y_i) + \varepsilon_i; \quad i = 1,2,\ldots,n. \tag{2}$$

To describe the relationship between the mean of the response variable and the linear predictor, the log-link function is used:

$$\mu_i = e^{x_i^T \beta}, \tag{3}$$

where $x_i = (x_{i1},\ldots,x_{ip})$ is the covariates vector (independent variables), and $\beta = (\beta_1,\ldots,\beta_p)^T$ is a vector of regression coefficients. The log-likelihood function for n independent Poisson observations with probabilities is given in equation (1) (Hilbe [12]):

$$\log L(\beta) = \sum_{i=1}^{n} \{y_i log(\mu_i) - \mu_i - log(y_i!)\}, \tag{4}$$

where μ_i is defined in equation (3). However, maximizing the log-likelihood has no closed-form solution, so numerical search procedures are used to find the ML estimates. Iteratively reweighted least squares can again be used to obtain these estimates. McCullagh and Nelder [1] showed that this algorithm is equivalent to Fisher scoring and leads to ML estimates. Using the iterative reweighted least squares algorithm discussed by [13], the weighted least squares (or ML) estimator is

$$\hat{\beta}_{ML} = \left(X^T \hat{W} X\right)^{-1} X^T \hat{W} \hat{Z}, \tag{5}$$

where $\hat{Z} = (\hat{z}_1,\ldots,\hat{z}_n)^T$ with $\hat{z}_i = log(\hat{\mu}_i) + \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i}$, $\hat{W} = diag(\hat{\mu}_i)$, and

$$X = \begin{pmatrix} x_{11} & \ldots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \ldots & x_{np} \end{pmatrix}.$$

3 Robust Estimators

This section tackles two robust estimators of the model. These estimators are based on the maximum likelihood method with different weigh matrices.

3.1 Mallows Quasi-likelihood Estimator

The quasi-likelihood (QL) estimators of GLMs (as in [14,1,15]) share the same nonrobustness properties. The QL estimator is the solution of the system of estimating equations:

$$\sum_{i=1}^{n} \frac{\partial}{\partial \beta} Q(y_i,\mu_i) = \sum_{i=1}^{n} \frac{y_i - \mu_i}{Var(\mu_i)} \mu_i' = 0, \tag{6}$$

where $\mu_i' = \frac{\partial}{\partial \beta}\mu_i$ and $Q(y_i,\mu_i)$ is the QL function. The solution of equation (6) is an M-estimator (Huber [16]) defined by the scorer function $\tilde{\psi}(y_i,\mu_i) = \frac{y_i - \mu_i}{Var(\mu_i)} \mu_i'$. Cantoni and Ronchetti [9] developed this estimator based on robust deviances that are natural generalization of QL functions, considering a general class of M-estimators of Mallows's type, where the influence of deviations on y and X are bounded separately. In other words, their robust estimator are based on the same class of robust estimators similar to what Preisser and Qaqish [8] proposed in the more general setup of generalized estimating equations. The Mallows quasi-likelihood (MQL) estimator is the solution of the estimating equations:

$$\sum_{i=1}^{n} \left[\psi_c(r_i)\, w(x_i) \frac{\mu_i'}{[Var(\mu_i)]^{\frac{1}{2}}} - \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}[\psi_c(r_i)]\, w(x_i) \frac{\mu_i'}{[Var(\mu_i)]^{\frac{1}{2}}} \right] = 0, \tag{7}$$

where $r_i = \frac{y_i - \mu_i}{[Var(\mu_i)]^{\frac{1}{2}}}$ are the Pearson residuals, $w(x_i) = \sqrt{1-h_i}$; where h_i is the ith diagonal element of the hat matrix, and $\psi_c(\cdot)$ is the Huber function defined by

$$\psi_c(r_i) = \begin{cases} r_i & |r_i| \leq c; \\ c\, sign(r_i) & |r_i| > c. \end{cases}$$

Cantoni and Ronchetti [9] illustrated that this estimator, which can be explicitly computed, does not require numerical integration, and the constant c is typically chosen to ensure a given level of asymptotic efficiency.

3.2 Weighted Maximum Likelihood Estimator

Hosseinian and Morgenthaler [10] introduced a robust estimator for the binary regression and generalized this estimator to Poisson regression based on a weighted maximum likelihood with weights that depend on μ_i and two constants c_1 and c_2. The estimation equation for the weighted maximum likelihood (WML) estimator is given by:

$$\sum_{i=1}^{n} \frac{\mu_i^{'}}{\mu_i} w(\mu_i)(y_i - \mu_i) x_i = 0, \tag{8}$$

where the weight function $w(\mu_i)$ is

$$w(\mu_i) = \begin{cases} 1 & \frac{\upsilon}{c_1} < \mu_i < c_1\upsilon; \\ \frac{c_1\mu_i}{\upsilon} & \mu_i < \frac{\upsilon}{c_1}; \\ \frac{c_2\upsilon - \mu_i}{\upsilon} & c_1\upsilon < \mu_i < c_2\upsilon; \\ 0 & otherwies, \end{cases}$$

where υ is the median of μ_i values, $c_1 = 2$, and $c_2 = 3$. To solve the system of equations in (8), Hosseinian [17] and Hosseinian and Morgenthaler [10] used the Newton-Raphson method.

4 Monte Carlo Simulation Study

In this section, we investigate the performance of the above-mentioned estimators through a simulation study. We compare between the nonrobust estimator (ML) and the robust estimators (MQL and WML). R software is used to perform our Monte Carlo simulation study. For further information on how to make Monte Carlo simulation studies using R, see [18, 19].

The simulated model is carried out based on equations (2) and (3) with the following simulation settings:

1. The number of independent variables are $p = 2$ and 6, where the independent variables are generated from uniform (-1, 1), and the vector of true regression coefficients is $\beta = 1$.
2. The values of sample size were chosen to be 75, 100, 200, 300 and 500 to represent moderate and large samples.
3. The percentages of outliers ($\tau\%$) in the response variable were chosen to be 5, 10, 20, 30 and 35.
4. The outliers generated from Poisson distribution with mean equal to 4 $IQR\left(e^{X\beta}\right)$; where IQR is the interquartile range.
5. For all experiments, we ran 1000 replications and all the results of all separate experiments are obtained by the same series of random numbers.

To compare the performance of the three estimators with different n, p, and $\tau\%$, we compute the average of mean squared error (MSE) and mean absolute error (MAE) for $\hat{\beta}$:

$$MSE = \frac{1}{1000}\sum_{l=1}^{1000}\left(\hat{\beta}_l - \beta\right)^2; \; MAE = \frac{1}{1000}\sum_{l=1}^{1000}\left|\hat{\beta}_l - \beta\right|,$$

where $\hat{\beta}_l$ is the vector of estimated values at l^{th} experiment of 1000 Monte Carlo experiments, while β is the vector of true coefficients.

The results are presented in Tables 1 to 5. Specifically, Table 1 reveals the MSE and MAE for ML, MQL, and WML estimators in case $n = 75, p = 2, 6$, and different $\tau\%$. While in cases $n = 100, 200, 300$, and 500, they are presented in Tables 2 to 5, respectively.

From Tables 1 to 5, we can summarize the effects of the main simulation factors on MSE and MAE values for all estimators (robust and nonrobust) as follows:

Table 1: MSE and MAE values of the estimators when $n = 75$.

Estimator	$\tau\%$				
	5	10	20	30	35
P=2, MSE					
ML	0.1189	0.1927	0.3536	0.5265	0.6406
MQL	0.0533	0.0615	0.1036	0.1802	0.2670
WML	0.0389	0.0400	0.0554	0.0837	0.1089
P=2, MAE					
ML	0.2753	0.3420	0.4660	0.5691	0.6327
MQL	0.1851	0.1973	0.2550	0.3412	0.4136
WML	0.1591	0.1594	0.1865	0.2293	0.2559
P=6, MSE					
ML	0.1135	0.1873	0.3241	0.4476	0.5171
MQL	0.0372	0.0516	0.0959	0.2242	0.3193
WML	0.0331	0.0376	0.0552	0.1068	0.1649
P=6, MAE					
ML	0.2694	0.3477	0.4602	0.5413	0.5836
MQL	0.1534	0.1783	0.2414	0.3634	0.4381
WML	0.1443	0.1528	0.1811	0.2405	0.2920

Table 2: MSE and MAE values of the estimators when $n = 100$.

Estimator	$\tau\%$				
	5	10	20	30	35
P=2, MSE					
ML	0.0767	0.1284	0.2395	0.3269	0.3704
MQL	0.0352	0.0453	0.0856	0.1620	0.2240
WML	0.0270	0.0272	0.0389	0.0572	0.0755
P=2, MAE					
ML	0.2169	0.2843	0.3840	0.4487	0.4727
MQL	0.1505	0.1706	0.2378	0.3284	0.3851
WML	0.1312	0.1315	0.1572	0.1905	0.2185
P=6, MSE					
ML	0.0727	0.1280	0.2325	0.3396	0.3904
MQL	0.0222	0.0278	0.0476	0.1033	0.1573
WML	0.0186	0.0205	0.0264	0.0439	0.0560
P=6, MAE					
ML	0.2135	0.2848	0.3846	0.4703	0.5043
MQL	0.1186	0.1320	0.1715	0.2485	0.3033
WML	0.1077	0.1130	0.1271	0.1595	0.1776

- As n increases, MSE and MAE of all estimators reduce.
- As $\tau\%$ increases, MSE and MAE of all estimators increase.
- As p increases, MSE and MAE of all estimators reduce in the case of small n and low $\tau\%$. However, the values of MSE and MAE increase when n and $\tau\%$ increase.

In all simulation cases, it is noticeable that the values of MSE and MAE for MQL and WML estimators are smaller than those of MSE and MAE for ML estimator. In other word, we can conclude that MQL and WML estimators are more efficient than ML estimator.

Graphically, we illustrate the relative efficiency (RE) of MQL and WML estimators for different $\tau\%$. The RE values are given by dividing the MSE of the estimator by the MSE of ML. The RE values of the estimators for $p = 2$ and $p = 6$ are shown in Figures 1 and 2, respectively.

Figure 1 indicates that RE values of WML are smaller than RE values of MQL for all n values. This suggests that the WML estimator is more efficient than the MQL estimator in different n and τ values. However, when n and $\tau\%$ increase ($n > 200$ and $\tau = 35\%$), the efficiency of WML increases. In Figure 2, the efficiency of the two estimators is close, but the WML estimator is still more efficient than the MQL estimator. Efficiency increases when $n < 200$ and $\tau = 35\%$.

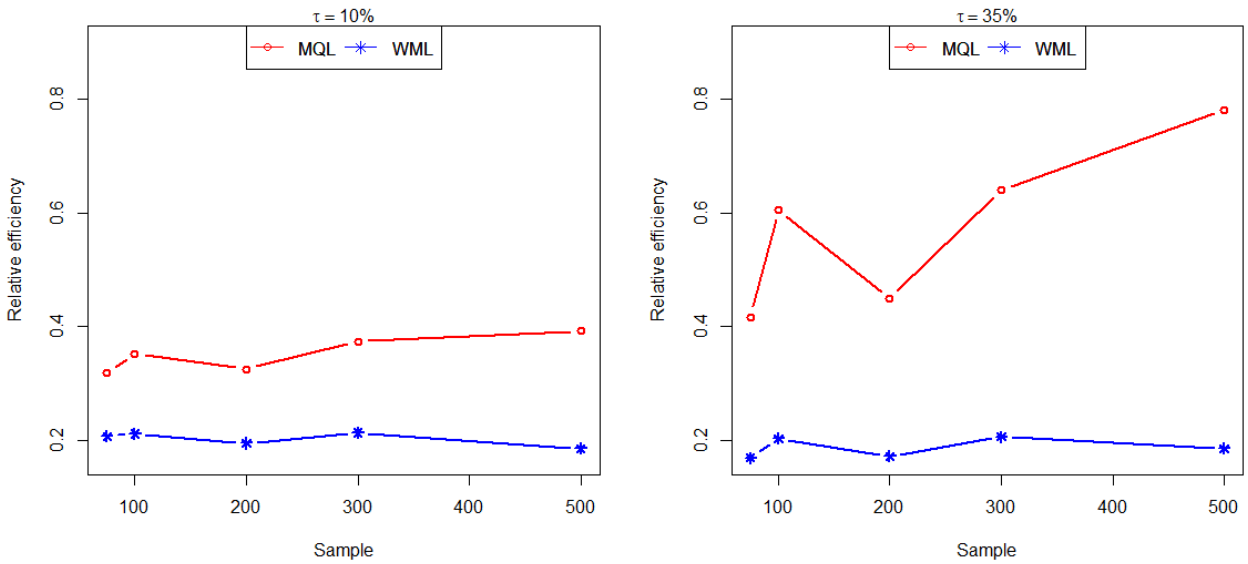

Fig. 1: Relative efficiency of the robust estimators when $p = 2$.

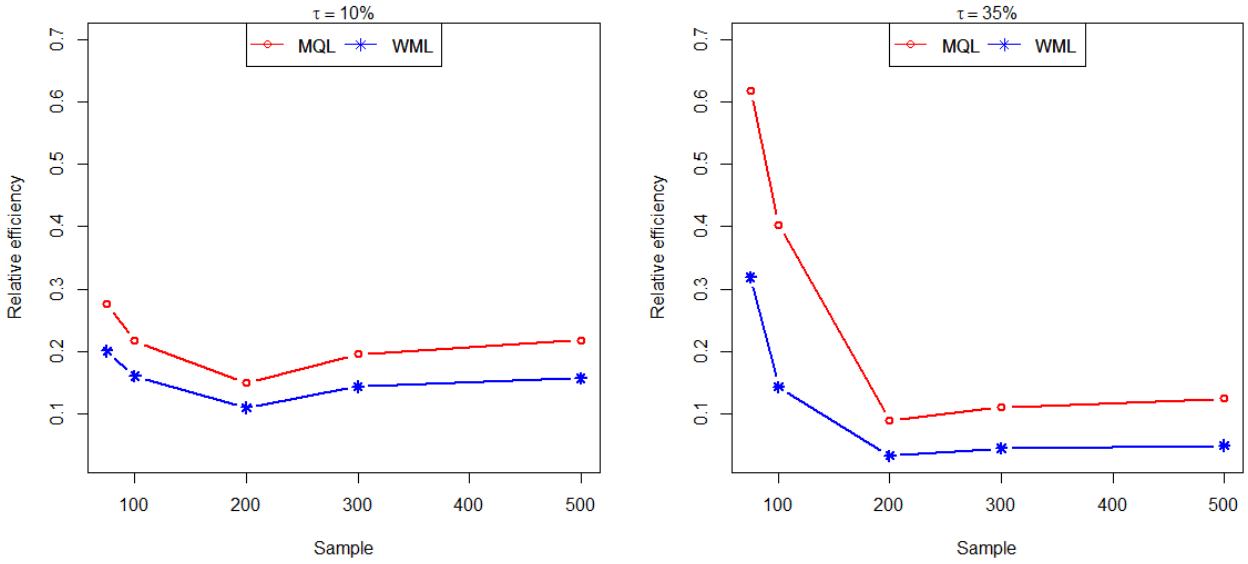

Fig. 2: Relative efficiency of the robust estimators when $p = 6$.

5 Conclusion

The present paper compared ML (nonrobust) with MQL and WML (robust) estimators for Poisson regression model with outliers. Our Monte Carlo simulation results indicated that the ML estimator is very sensitive to outliers, while MQL and WML estimators are more effective. In addition, WML is more efficient than MQL.

Table 3: MSE and MAE values of the estimators when $n = 200$.

Estimator	τ%				
	5	10	20	30	35
P=2, MSE					
ML	0.0402	0.0755	0.1496	0.2583	0.3020
MQL	0.0185	0.0246	0.0487	0.0918	0.1360
WML	0.0138	0.0148	0.0220	0.0377	0.0522
P=2, MAE					
ML	0.1599	0.2131	0.3029	0.4015	0.4384
MQL	0.1083	0.1258	0.1794	0.2511	0.3044
WML	0.0937	0.0966	0.1177	0.1554	0.1848
P=6, MSE					
ML	0.0495	0.1097	0.2903	0.5347	0.6927
MQL	0.0132	0.0164	0.0266	0.0478	0.0621
WML	0.0109	0.0121	0.0143	0.0199	0.0232
P=6, MAE					
ML	0.1764	0.2624	0.4310	0.5887	0.6785
MQL	0.0917	0.1020	0.1296	0.1724	0.1970
WML	0.0831	0.0869	0.0955	0.1109	0.1199

Table 4: MSE and MAE values of the estimators when $n = 300$.

Estimator	τ%				
	5	10	20	30	35
P=2, MSE					
ML	0.0256	0.0451	0.0848	0.1414	0.1711
MQL	0.0118	0.0169	0.0374	0.0732	0.1095
WML	0.0082	0.0096	0.0141	0.0239	0.0354
P=2, MAE					
ML	0.1262	0.1673	0.2309	0.2939	0.3306
MQL	0.0868	0.1045	0.1585	0.2266	0.2795
WML	0.0721	0.0780	0.0946	0.1234	0.1545
P=6, MSE					
ML	0.0264	0.0526	0.1203	0.2293	0.3103
MQL	0.0084	0.0103	0.0160	0.0264	0.0344
WML	0.0068	0.0076	0.0092	0.0119	0.0140
P=6, MAE					
ML	0.1295	0.1809	0.2773	0.3894	0.4583
MQL	0.0731	0.0808	0.1008	0.1291	0.1460
WML	0.0656	0.0693	0.0766	0.0870	0.0932

Table 5: MSE and MAE values of the estimators when $n = 500$.

Estimator	$\tau\%$				
	5	10	20	30	35
P=2, MSE					
ML	0.0163	0.0294	0.0650	0.1070	0.1375
MQL	0.0075	0.0116	0.0318	0.0744	0.1072
WML	0.0052	0.0055	0.0089	0.0185	0.0256
P=2, MAE					
ML	0.1019	0.1375	0.2012	0.2599	0.2945
MQL	0.0697	0.0868	0.1494	0.2330	0.2804
WML	0.0575	0.0592	0.0760	0.1124	0.1339
P=6, MSE					
ML	0.0135	0.0261	0.0602	0.1134	0.1490
MQL	0.0044	0.0057	0.0084	0.0142	0.0186
WML	0.0036	0.0041	0.0047	0.0063	0.0074
P=6, MAE					
ML	0.0917	0.1279	0.1969	0.2731	0.3143
MQL	0.0532	0.0601	0.0728	0.0950	0.1075
WML	0.0483	0.0510	0.0546	0.0630	0.0679

References

[1] P. McCullagh and J. A. Nelder, *Generalized Linear Models*, 2nd Edition, Chapman & Hall, London UK, (1989).

[2] A. Agresti, *Categorical Data Analysis*, 2nd Edition, Wiley, New York USA, (2001).

[3] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw and W. A. Stahel, *Robust Statistics: The Approach Based on Influence Functions*, Vol. 196, John Wiley & Sons, (2011).

[4] A. Cameron and P. Trivedi, *Regression Analysis of Count Data*, 2nd Edition, Cambridge University Press, Cambridge UK, (2013).

[5] M. R. Abonazel and A. A. Gad, Robust partial residuals estimation in semiparametric partially linear model, *Communications in Statistics-Simulation and Computation*, Accepted paper (2018). https://doi.org/10.1080/ 03610918.2018.1494279.

[6] M. M. Elgohary, M. R. Abonazel, N. M. Helmy and A. R. Azazy, New robust-ridge estimators for partially linear model, *International Journal of Applied Mathematical Research*, **8**, 46-52 (2019)

[7] M.R. Abonazel and A. R. Rabie, The impact for using robust estimations in regression models: an application on the Egyptian economy, *Journal of Advanced Research in Applied Mathematics and Statistics*, **4**, 8-16 (2019).

[8] J. S. Preisser and B. F. Qaqish, Robust regression for clustered data with application to binary responses, *Biometrics*, **55**, 574-579 (1999).

[9] E. Cantoni and E. Ronchetti, Robust inference for generalized linear models, *Journal of the American Statistical Association*, **96**, 1022-1030 (2001).

[10] S. Hosseinian and S. Morgenthaler, *Weighted maximum likelihood estimates in Poisson regression*, International Conference on Robust Statistics, Italy, (2011).

[11] R. Winkelmann, *Econometric Analysis of Count Data*, 5th Edition, Springer Verlag, Berlin, (2008).

[12] J. M. Hilbe, *Negative Binomial Regression*, 2nd Edition, Cambridge University Press, Cambridge UK, (2011).

[13] A. Dobson and A. Barnett, *An Introduction to Generalized Linear Models*, 3rd Edition, Chapman and Hall/CRC, (2008).

[14] R. Wedderburn, Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method, *Biometrika*, **61**, 439-447 (1974).

[15] C. C. Heyde, *Quasi-Likelihood and its Application*, Springer-Verlag, New York USA, (1997).

[16] P. J. Huber, *Robust Statistics*, Wiley, New York USA, (1981).

[17] S. Hosseinian, *Robust Inference for Generalized Linear Models: Binary and Poisson Regression*. Doctoral thesis, EPFL Lausanne, Switzerland, (2009).

[18] M. R. Abonazel, A practical guide for creating Monte Carlo simulation studies using R, *International Journal of Mathematics and Computational Science*, **4**, 18-33 (2018).

[19] M. R. Abonazel, *Advanced statistical techniques using R: outliers and missing data*, presented at the 54th Annual Conference on Statistics, Computer Sciences and Operations Research, Cairo University, Egypt, (2019).

Mohamed R Abonazel is Associate professor of applied statistics and econometrics at Cairo University. He received the PhD degree in "applied statistics and econometrics" at Cairo University (Egypt). He is referee and editor of several international journals in the frame of applied statistics, applied mathematics, and econometrics. His main research interests are panel data models, time series models, spatial econometrics, non-parametric and semi-parametric regression, robust regression, model selection methods, missing data analysis, and count regression. https://orcid.org/0000-0001-6010-001X

Omnia M. Saber holds a Bachelor of Commerce (Statistics) from Al-Azhar University (Egypt). She is a master student (Applied Statistics and Econometrics) at Faculty of Graduate Studies for Statistical Research (FGSSR), Cairo University (Egypt). Her main research interests are generalized linear models, count data models, and robust regression.