Available online at http://scik.org
J. Math. Comput. Sci. 11 (2021), No. 6, 8173-8196
https://doi.org/10.28919/jmcs/5852
ISSN: 1927-5307

THE PERFORMANCE OF COUNT PANEL DATA ESTIMATORS: A SIMULATION STUDY AND APPLICATION TO PATENTS IN ARAB COUNTRIES

AHMED H. YOUSSEF, MOHAMED R. ABONAZEL*, ELSAYED G. AHMED

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

Abstract: This paper provides four estimators of count panel data (CPD) models; fixed effects Poisson (FEP), random effects Poisson (REP), fixed effects negative binomial (FENB), and random effects negative binomial (RENB). In FEP and FENB models, we used conditional maximum likelihood (CML) estimation method. While for REP and RENB models, we used maximum likelihood (ML) estimation method. We conducted a Monte Carlo simulation study to compare the behavior of these estimators in the four models. The results of simulation show that the best estimator is FENB compared to other estimators (FEP, REP, and RENB), because it has minimum values for Akaike information criterion (AIC) and Bayesian information criterion (BIC), especially when the model or the data has an overdispersion problem. Moreover, a real dataset has been used to study the effect of some economic variables on the number of patents for seven Arab countries over the period from 2000 to 2016. Application results indicate that the RENB is the suitable model for this data, and the important (statistically significant) variables that effect on the number of patents is the gross domestic product per capita.

*Corresponding author
E-mail address: mabonazel@cu.edu.eg
Received April 12, 2021

Keywords: conditional maximum likelihood estimation; fixed effects panel; Hausman test; negative binomial panel data; Poisson panel data; random effects panel.

2010 AMS Subject Classification: 62J12, 62P20.

1. INTRODUCTION

Recently, panel data or longitudinal data sets have become one of the most exciting fields in econometrics literature due to new sources of data which observes the cross-sections of individuals over time. This allows constructing and testing more realistic behavioral models that could not be identified using a single cross-section or a single time-series data set. Therefore, panel data analysis is a core field in modern econometrics and multivariate statistics. Thus, panel data sets have become widely available, where there are many of the contributions and recent studies which have analyzed panel data, e.g. Baltagi [1] stated that the panel data refers to the pooling of observations on a cross-section of households, countries, firms, etc., over several time periods.

According to Vijayamohanan [2], the panel data refers to a data set containing observations on multiple phenomena over multiple time periods, where it has two dimensions; the spatial dimension (cross-sectional) and temporal dimension (time series). Greene [3] pointed out that the analysis of panel data is one of the important topics and common in economics, because it allows great flexibility in modeling differences in behavior across individuals and provide rich sources of information and rich environment for the development of estimation techniques. Furthermore, the researchers are uses time-series cross-sectional data to examine issues that could not be studied in either cross-sectional or time-series alone. Also, the analysis of panel data allows the model builder to learn about economic processes considering both heterogeneity across individuals, firms, countries, etc., and dynamic effects that are not visible in cross sections.

Abonazel [4] explained that pooling cross-sectional and time series data (panel data) achieves a deep analysis for the data and gives a richer source of variation, which allows for more efficient estimation of the parameters and more effective in identifying and estimating effects that are simply not detectable in cross-sectional or time series data. Also, panel data sets are more effective in studying complex issues of dynamic behavior.

Panel data models have become increasingly popular among applied researchers due to their heightened capacity for capturing the complexity of human behavior as compared to cross-sectional or time-series data models. Therefore, we will discuss the most popular models in panel data modeling, which is the fixed effects and random effects models.

In general, the fixed effects model has different intercepts, where the intercept is differing from unit to unit and fixed over time. The general form of the fixed effects model is [5, 6, 7]:

$$y_{it} = \alpha_i + \mathrm{x}'_{it}\beta + u_{it}, \quad i = 1, 2, \dots, N;\ t = 1, 2, \dots, T, \tag{1}$$

where y_{it} is the response variable for individual i at time t, x_{it} is the vector of explanatory variables, α_i is a scalar constant (the intercept) include the unobserved effect for special variables to the i^{th} individual over time, β is the vector of the regression coefficients, and u_{it} is the error term of the model.

In fixed effects model, the individual effects α_i are treated as fixed constants over time where individual effects are parts of the intercept, however in random effects model puts the individual effects into the error term and treat the individual effects, like u_{it} as random variables. The random effects model assumes that the unit's error term is not correlated with the predictors and the variation across entities is assumed to be random, in addition to the random effects model assumes that there is one constant term (α) for all across unites, and the differences of the intercept term can be captured in the error term, hence the error term become have new assumptions [6, 8]. The random effects model is given by:

$$y_{it} = \alpha + \mathrm{x}'_{it}\beta + \varepsilon_{it}, \quad i = 1, 2, \dots, N;\ t = 1, 2, \dots, T, \tag{2}$$

where $\varepsilon_{it} = \nu_i + u_{it}$; this means that the error term of the model consists two components, ν_i and u_{it}, where ν_i denotes the unobservable individual effects, which are unobservable factors affecting y and which do not vary over time or the unit's unobserved ability that is not included in the regression, such as managerial skills, level of intelligence, and the unobservable entrepreneurial of unit. While u_{it} denotes the disturbances, which varies with units and time and can be thought of as the usual disturbance in the regression or represents the other variables influencing y but which vary both over time and units.

The unobservable individual effects (ν_i) and the disturbances (u_{it}) are assumed to be independently distributed across units, where ν_i is uncorrelated with each independent variable included in the model [1].

On the other side, if the dependent variable of the panel data model takes non-negative integer value such as (0, 1, 2, …), in this case the model is called the count panel data (CPD) model. Actually, the CPD analysis is a data type used with increasing frequency in empirical research in economics, social sciences, and medicine, etc., for example, the number of patents in some countries of the world over several years, the number of deaths from covid-19 in the countries of the world in multiple time periods, and the number of accidents in several areas over several years. In econometrics literature, commonly used models that fit this data are Poisson and negative binomial models, where there are many economic studies that discussed these models in panel data modeling, e.g. [9, 10, 11, 12, 13].

Count regression models are varied depending on the types of data, where the count data is treated as dependent variable, so linear estimation methods, such as least squares that are designed to deal with continuous variable, are not appropriate for count data. Since the linear regression model assumes that the dependent variable follows the normal distribution, then it is not suitable for the count data. In addition to, the linear regression model may produce negative estimates for the response variable which is incorrect for the count data. So, the Poisson and negative binomial distributions are the basis of count data analysis.

The rest of the paper is organized as follows: section 2 provides Poisson panel models. In section 3 presents negative binomial panel models. Section 4 will be devoted to determining the settings of the simulation through design of Monte Carlo experiment and how the data is generated, where presents the main steps for making the Monte Carlo simulation study. Section 5 offers the results of the simulation study. In section 6, the empirical study on patents for seven Arab countries is presented. Finally, section 7 contains concluding remarks.

2. POISSON PANEL MODELS

The most common probability models for modelling CPD is Poisson panel model. In the Poisson distribution is the mean and the variance are the same, the higher the value of the mean of the distribution, the greater the variance or variability in the data [14]. The Poisson panel model assumes that the dependent variable (y_{it}) has a Poisson distribution. The probability mass function of y_{it} with parameter λ_{it} can be expressed as:

$$f(y_{it};\lambda_{it}) = \frac{[exp(-\lambda_{it})]\,(\lambda_{it})^{y_{it}}}{y_{it}!},\ i = 1,2,\dots,N;\ t = 1,2,\dots,T, \qquad (3)$$

where y_{it} represents a variable consisting of count values and $\lambda_{it} > 0$. λ_{it} is the expected or predicted mean of the count variable y_{it}, and the subscripts (i) and (t) indicates that the model describes each observation in the data. In the model (3), the mean and the variance of y_{it} must be equal, i.e. $E(y_{it}) = var(y_{it}) = \lambda_{it}$.

The Poisson panel model has one parameter (λ_{it}) which it must be positive. It is convenient to specify λ_{it} as an exponential function of the independent variables. The exponential form ensures that λ_{it} remains positive for all possible combinations of parameters and independent variables.

2.1 Fixed Effects Poisson Model

In the FEP model, all characteristics that are not time-varying are captured by the individual effects (α_i). The intercept (constant term) is merged into α_i, hence the explanatory variables (x_{it}) do not contain an intercept [15]. The conditional probability function of the FEP model as:

$$f(y_{it}|\mathrm{x}_{it},\alpha_i,\beta) = \frac{[exp(-\alpha_i\lambda_{it})]\,(\alpha_i\lambda_{it})^{y_{it}}}{y_{it}!},\ i = 1,2,\dots,N;\ t = 1,2,\dots,T, \qquad (4)$$

where $\lambda_{it} = exp(\mathrm{x}_{it}'\beta)$. The last equality specifies an exponential functional form. To estimate the parameters of the model (4), it can use the CML estimation method that developed by Hausman et al. [16]. Since y_{it} and $\sum_{t=1}^{T} y_{it}$ are follow the Poisson distribution, then the conditional joint density function (CJDF) for the i^{th} observation is:

$$f(y_{i1},\dots,y_{iT}|\textstyle\sum_{t=1}^{T} y_{it}) = \frac{(\sum_{t=1}^{T} y_{it})!}{(\sum_{t=1}^{T}\lambda_{it})^{\sum_{t=1}^{T} y_{it}}} \times \prod_{t=1}^{T} \frac{\lambda_{it}^{\,y_{it}}}{y_{it}!},$$

when taking the logarithm of CJDF and summing over all individuals, the conditional log-likelihood is:

$$lnL = \sum_{i=1}^{N}\{ ln(\sum_{t=1}^{T} y_{it})! - \sum_{t=1}^{T} lny_{it}! + \sum_{t=1}^{T}[y_{it}\mathrm{x}_{it}'\beta - y_{it}ln \sum_{t=1}^{T} exp(\mathrm{x}_{it}'\beta)]\},$$

it can obtain the estimated parameters for the FEP model by solving:

$$\sum_{i=1}^{N} \sum_{t=1}^{T} \mathrm{x}_{it}' \left(y_{it} - \frac{\sum_{t=1}^{T} y_{it}}{\sum_{t=1}^{T} \lambda_{it}} \lambda_{it} \right) = 0.$$

2.2 Random Effects Poisson Model

In the REP model, the individual effects (unobserved heterogeneity) are expressed as ν_i instead of α_i, while the intercept is included and merged into x_{it}. The individual effect ν_i must follow a specified distribution in order to estimate the parameters of the REP model. Therefore, many researchers assumed that the individual effect in the REP model has a gamma distribution with parameters (γ, γ), see e.g. [5, 14, 16, 17, 18].

The REP model assumes that the response variable (y_{it}) has a Poisson distribution and the individual effect has a gamma distribution, then ML estimation method should be used to estimate the parameters of the REP model. The ML function for the it^{th} observation is:

$$f(y_{it}|\nu_i, \mathrm{x}_{it}) = \prod_{t=1}^{T} \left(\frac{\lambda_{it}{}^{y_{it}}}{y_{it}!}\right) \left[\frac{\gamma}{\gamma+\sum_{t=1}^{T} \lambda_{it}}\right]^{\gamma} \left[\frac{\Gamma(\sum_{t=1}^{T} y_{it}+\gamma)}{\Gamma(\gamma)}\right] [\gamma + \sum_{t=1}^{T} \lambda_{it}]^{-\sum_{t=1}^{T} y_{it}},$$

and the log-maximum likelihood function is:

$$lnL = \sum_{i=1}^{N}\{ \sum_{t=1}^{T}(y_{it}\mathrm{x}_{it}'\beta - lny_{it}!) + \gamma ln\gamma - \gamma\, ln[\gamma + \sum_{t=1}^{T} exp(\mathrm{x}_{it}'\beta)] + l\,n[\Gamma(\sum_{t=1}^{T} y_{it} + \gamma)] - ln[\Gamma(\gamma)] - \sum_{t=1}^{T} y_{it}\, ln\, [\gamma + \sum_{t=1}^{T} exp(\mathrm{x}_{it}'\beta)]\},$$

thus, it can obtain the estimated parameters of this model by solving:

$$\sum_{i=1}^{N} \sum_{t=1}^{T} \mathrm{x}_{it}' \left[y_{it} - \lambda_{it} \left(\frac{\bar{y}_i+\gamma/T}{\bar{\lambda}_i+\gamma/T}\right)\right] = 0.$$

3. NEGATIVE BINOMIAL PANEL MODELS

The negative binomial model is one of the basic models for count data analysis. This model has found a widespread use in the fields of health, social, economic, and physical sciences when the response variable comes in the form of non-negative integers or counts [19].

In general, the negative binomial panel model introduced as a generalized version of Poisson model that allows the variance of the dependent variable to differ from its mean. The negative binomial panel model is a two-parameter model; with mean λ_{it} and dispersion parameters ϕ_i.

The mean of the negative binomial panel model is understood in the same manner as the Poisson mean, but the variance of the negative binomial has a much wider scope than is allowed by the Poisson model. When the variance of count data exceeds the mean, i.e. if $var\ (y_{it}) > E(y_{it})$, then we speak about overdispersion. But if $var\ (y_{it}) < E(y_{it})$, then this is called underdispersion. The Poisson model does not allow for overdispersion or underdispersion. Hence, we used the negative binomial model instead of the Poisson model [19].

3.1 Fixed Effects Negative Binomial Model

The FENB model assumes that for a given unit i, the response variable (y_{it}) is independent over time and $\sum_{t=1}^{T} y_{it}$ has a negative binomial distribution with parameters θ_i and $\sum_{t=1}^{T} \lambda_{it}$. These assumptions imply that:

$$\sum_{t=1}^{T} y_{it} \sim NB\left(\theta_i \sum_{t=1}^{T} \lambda_{it}\ , (\theta_i \sum_{t=1}^{T} \lambda_{it})(1+\theta_i)\right),$$

where $\theta_i = \alpha_i/\phi_i$, Hausman et al. [16] showed that the CJDF of the FENB model for the i^{th} observation is:

$$f(y_{i1}, \dots, y_{iT} | \sum_{t=1}^{T} y_{it}) = \frac{\Gamma(\sum_{t=1}^{T} \lambda_{it})\Gamma(\sum_{t=1}^{T} y_{it}+1)}{\Gamma(\sum_{t=1}^{T} \lambda_{it} + \sum_{t=1}^{T} y_{it})} \times \left[\prod_{t=1}^{T} \frac{\Gamma(\lambda_{it}+y_{it})}{\Gamma(\lambda_{it})\Gamma(y_{it}+1)}\right],$$

where $\Gamma(\cdot)$ is the gamma function. In order to estimate the parameters of this model, Hausman et al. [16] used the CML estimation method. Thus, it can obtain the CML estimation of this model by maximizing the following log-conditional maximum likelihood function:

$$lnL = \sum_{i=1}^{N} \{ ln\Gamma(\sum_{t=1}^{T} \lambda_{it}) + ln\Gamma(\sum_{t=1}^{T} y_{it} + 1) - ln\Gamma(\sum_{t=1}^{T} \lambda_{it} + \sum_{t=1}^{T} y_{it}) +$$
$$\sum_{i=1}^{T} [ln\Gamma(\lambda_{it} + y_{it}) - ln\Gamma(\lambda_{it}) - ln\Gamma(y_{it} + 1)]\}.$$

3.2 Random Effects Negative Binomial Model

For the RENB model, Hausman et al. [16] assumed that the dependent variable (y_{it}) specified to be independent and identically distributed negative binomial, and $1/(1+\delta_i)$ is distributed as beta with parameters (a, b), where $\delta_i = \nu_i/\phi_i$, i.e. $1/(1+\delta_i) \sim Beta(a, b)$. The mean and the variance of the response variable (y_{it}) are $\lambda_{it}\delta_i$ and $\lambda_{it}\delta_i(1+\delta_i)$, respectively.

To estimate the parameters of RENB model, it can use the ML estimation method. Then the joint density function for the ij^{th} observation is:

$$f(y_{it}|\mathrm{x}_{it}) = \frac{\Gamma(\alpha+b)\,\Gamma\left(\alpha+\sum_{t=1}^{T}\lambda_{it}\right)\Gamma\left(b+\sum_{t=1}^{T}y_{it}\right)}{\Gamma(\alpha)\,\Gamma(b)\,\Gamma\left(\alpha+b+\sum_{t=1}^{T}\lambda_{it}+\sum_{t=1}^{T}y_{it}\right)} \times \prod_{t=1}^{T}\left[\frac{\Gamma(\lambda_{it}+y_{it})}{\Gamma(\lambda_{it})\Gamma(y_{it}+1)}\right].$$

The ML estimation of the RENB model can be obtained by maximizing the following log-maximum likelihood function:

$$lnL = \sum_{i=1}^{N}\{\, ln\,\Gamma(\alpha+b) + ln\,\Gamma[\alpha+\textstyle\sum_{t=1}^{T}\lambda_{it}] + \; ln\,\Gamma(b+\sum_{t=1}^{T}y_{it}) - ln\,\Gamma(\alpha) - ln\,\Gamma(b) -$$
$$ln\,\Gamma[\alpha+b+\textstyle\sum_{t=1}^{T}\lambda_{it}+\sum_{t=1}^{T}y_{it}] + \sum_{i=1}^{T}[ln\,\Gamma(\lambda_{it}+y_{it}) - ln\,\Gamma(\lambda_{it}) - ln\,\Gamma(y_{it}+1)]\}.$$

4. SIMULATION DESIGN

We will use the Monte Carlo simulation for making a comparison between the behavior of FEP, REP, FENB, and RENB estimators of the four CPD models above. We used R language to conduct our Monte Carlo simulation [20, 21]. Several studies have been relied upon when conducting a Monte Carlo simulation study such as [4, 22, 23, 24, 25, 26].

4.1 In Case of Moderate and Large Samples

The simulation study was carried out in the moderate and large samples based on the following:

1. The values of N were chosen to be 30, 50, 100, 200, 300, and 500 to represent moderate and large samples for the number of individuals.
2. The values of T were chosen to be 10, 15, 40, 50, 100, and 200 to represent different size for the time period.
3. The values of $\beta_1, \beta_2,$ and β_3 were chosen to be 1.
4. The response variable (y_{it}) is generated from the negative binomial distribution with different values of the dispersion; where ϕ were chosen to be $0.5,\ 1,$ and 5.
5. The individual effects (α_i) were generate as independent normally distribution with mean -1 and standard deviation 0.5, where α_i is differing from unit to unit and fixed over time.
6. We generate the explanatory variables using random numbers following the uniform distribution from -1 to 1.
7. For all experiments we ran 1000 replications and all the results for all separate experiments are obtained by precisely the same series of random numbers.

We can note that the generated model in our simulation is FENB model with three cases of the dispersion parameter (ϕ). In the first case the dispersion parameter $\phi < 1$ (i.e., $\phi = 0.5$), therefore we speak about underdispersion. While in the second case the dispersion parameter $\phi = 1$, this is called equidispersion. In the third case the dispersion parameter $\phi > 1$ (i.e., $\phi = 5$), thus we speak about overdispersion.

4.2 In Case of Small Samples

In this section, we will study the behavior of the four estimators in case of small samples. The data were generated by the same method in the case of moderate and large samples with the difference in cross section size to be 5, 10, 15, and 20 and time series to be 15 and 20, and the dispersion parameter is one.

The Monte Carlo experiment has been designed to compare the small, moderate, and large samples performances of ML estimators of REP and RENB models and CML estimators of FEP and FENB models based on AIC [27] and BIC [28].

5. SIMULATION RESULTS

The results of the Monte Carlo simulation study for the moderate and large samples have been provided in tables from 1 to 6, while figures from 1 to 4 displays the small samples results. Each table represents AIC and BIC values (rounded to integer) for different values of T and ϕ. Tables from 1 to 6 present the estimation results (AIC and BIC) of FEP, REP, FENB, and RENB estimators for different values of N.

In tables from 1 to 3, when the dispersion parameter equal 0.5 or 1, we find that the AIC and BIC values of FENB estimator have smallest values than the FEP, REP, and RENB estimators. For example, in table 1 when $\phi = 1$ and $T = 10,$ the AIC value of FENB is 388, but the AIC values of FEP, REP, and RENB are 413, 579, and 556, respectively. While the BIC value of FENB is 403, but the BIC value of FEP, REP, and RENB are 424, 598, and 575, respectively. And when the dispersion parameter is increasing to 5, then AIC and BIC values of FENB estimator are decreasing dramatically and still AIC and BIC values of FENB estimator is the smallest. For example, in table 1 when $\phi = 5$ and $T = 10,$ the AIC value of FENB is 364, but the AIC values of FEP, REP, and RENB are 596, 779, and 615, respectively. While the BIC value of FENB is 378, but the BIC values of FEP, REP, and RENB are 607, 797, and 633, respectively. So, the results of tables from 1 to 3 showed that the FENB estimator is better than FEP, REP, and RENB estimators in case of moderate samples (N = 30, 50, 100).

For the results of large samples (N = 200, 300, 500), in tables 4 to 6, showed that when the time periods (T) is increasing from 10 to 15 or 200 and the dispersion parameter (ϕ) is increasing from 0.5 to 1 or 5, the AIC and BIC values of FENB estimator is still smaller than the AIC and BIC values of other estimators. For example, in table 5 when $N = 300$, $\phi = 5$, and $T = 10$, the AIC value of FENB is 3511, but the AIC values of FEP, REP, and RENB are 5883, 7684, and 6026, respectively. While the BIC value of FENB is 3535, but the BIC values of FEP, REP, and RENB are 5901, 7714, and 6057, respectively. Whereas when $N = 300$, $\phi = 5$, and $T = 50$, the AIC value of FENB is 23092, but the AIC values of FEP, REP, and RENB are 38116, 40751, and 30809, respectively. While the BIC value of FENB is 23122, but the BIC values of FEP, REP, and RENB are 38139, 40789, and 30847, respectively. So, the results of tables from 4 to 6 showed that the FENB estimator is better than FEP, REP, and RENB estimators in case of large samples.

Figures from 1 to 4 display the AIC and BIC values of different estimators for small samples in N and T. These figures showed that in case of increasing the number of units (N), the AIC and BIC values of all estimators are increased. But still the FENB estimator is better than other estimators in case of small samples, even if the dispersion parameter equal one.

Table 1: AIC and BIC values of different estimators when $N = 30$

Criterion	Estimator	*T* =10	*T* =15	*T* =40	*T* =50	*T* =100	*T* =200
$\phi = 0.5$							
AIC	FEP	442	722	2075	2420	6090	10811
	REP	619	912	2319	2658	6392	11147
	FENB	429	705	2016	2369	5883	10512
	RENB	605	894	2260	2606	6183	10848
BIC	FEP	454	735	2090	2435	6108	10831
	REP	637	933	2344	2685	6422	11181
	FENB	444	722	2037	2390	5907	10539
	RENB	623	915	2285	2633	6213	10881
$\phi = 1$							
AIC	FEP	413	664	2122	3118	5687	12793
	REP	579	847	2359	3378	5978	13139
	FENB	388	619	1970	2847	5262	11553
	RENB	556	806	2225	3134	5600	12034
BIC	FEP	424	676	2138	3133	5705	12813
	REP	598	868	2384	3405	6008	13173
	FENB	403	635	1990	2869	5286	11580
	RENB	575	826	2251	3161	5630	12068
$\phi = 5$							
AIC	FEP	596	1037	2571	3869	6440	15335
	REP	779	1247	2810	4135	6723	15673
	FENB	364	621	1692	2371	4260	9504
	RENB	615	968	2261	3159	5389	11869
BIC	FEP	607	1049	2586	3885	6458	15355
	REP	797	1267	2836	4161	6753	15707
	FENB	378	637	1712	2393	4284	9530
	RENB	633	989	2287	3185	5419	11902

Table 2: AIC and BIC values of different estimators when $N = 50$

Criterion	Estimator	*T* =10	*T* =15	*T* =40	*T* =50	*T* =100	*T* =200
$\phi = 0.5$							
AIC	FEP	786	1055	3255	4199	8974	18073
	REP	1086	1364	3638	4609	9462	18615
	FENB	760	1030	3178	4098	8716	17604
	RENB	1059	1338	3560	4508	9202	18143
BIC	FEP	798	1069	3272	4216	8994	18095
	REP	1107	1387	3666	4638	9495	18652
	FENB	776	1049	3200	4122	8742	17633
	RENB	1080	1361	3588	4537	9234	18179
$\phi = 1$							
AIC	FEP	721	1133	3363	4707	9710	18455
	REP	996	1449	3745	5138	10209	18994
	FENB	671	1052	3141	4321	8898	17171
	RENB	951	1373	3551	4795	9477	17868
BIC	FEP	734	1147	3379	4724	9729	18476
	REP	1017	1473	3773	5167	10242	19030
	FENB	688	1071	3163	4345	8924	17199
	RENB	972	1396	3579	4824	9510	17904
$\phi = 5$							
AIC	FEP	973	1493	5297	5876	13395	23832
	REP	1272	1824	5733	6307	13886	24376
	FENB	574	939	3114	3714	8208	15316
	RENB	998	1448	4238	4903	10538	19055
BIC	FEP	986	1507	5314	5893	13414	23853
	REP	1293	1848	5761	6336	13918	24412
	FENB	591	957	3137	3737	8234	15345
	RENB	1019	1472	4266	4932	10570	19091

Table 3: AIC and BIC values of different estimators when $N = 100$

Criterion	Estimator	*T* =10	*T* =15	*T* =40	*T* =50	*T* =100	*T* =200
$\phi = 0.5$							
AIC	FEP	1456	2173	7069	8833	17302	37893
	REP	2016	2791	7868	9696	18261	39011
	FENB	1417	2116	6863	8570	16848	36765
	RENB	1976	2735	7661	9432	17803	37880
BIC	FEP	1470	2188	7088	8852	17324	37916
	REP	2040	2818	7900	9729	18297	39050
	FENB	1436	2138	6888	8596	16877	36797
	RENB	2001	2761	7693	9464	17839	37920
$\phi = 1$							
AIC	FEP	1541	2440	7428	9623	19331	38778
	REP	2109	3073	8234	10472	20300	39869
	FENB	1414	2244	6836	8836	17832	35839
	RENB	1994	2899	7708	9773	18980	37298
BIC	FEP	1555	2456	7447	9642	19352	38802
	REP	2133	3099	8265	10504	20336	39909
	FENB	1433	2266	6862	8863	17861	35870
	RENB	2018	2926	7740	9806	19016	37337
$\phi = 5$							
AIC	FEP	2035	2900	9162	11617	26275	53460
	REP	2643	3553	9984	12476	27288	54587
	FENB	1213	1821	5739	7338	15927	32642
	RENB	2070	2828	7734	9695	20405	40924
BIC	FEP	2049	2916	9181	11637	26297	53484
	REP	2667	3580	10015	12508	27324	54627
	FENB	1233	1843	5764	7364	15956	32674
	RENB	2095	2854	7765	9728	20442	40964

Table 4: AIC and BIC values of different estimators when $N = 200$

Criterion	Estimator	*T* =10	*T* =15	*T* =40	*T* =50	*T* =100	*T* =200
$\phi = 0.5$							
AIC	FEP	2738	4471	13920	17758	36276	72970
	REP	3845	5725	15530	19439	38217	75201
	FENB	2661	4346	13516	17250	35247	70841
	RENB	3765	5594	15122	18927	37186	73065
BIC	FEP	2755	4489	13941	17780	36300	72995
	REP	3873	5755	15565	19475	38257	75244
	FENB	2683	4370	13544	17279	35279	70875
	RENB	3793	5624	15157	18964	37225	73108
$\phi = 1$							
AIC	FEP	2909	4832	14085	18282	39330	81501
	REP	4038	6117	15662	19961	41285	83724
	FENB	2690	4429	13009	16874	36129	74712
	RENB	3836	5749	14709	18708	38447	77729
BIC	FEP	2926	4850	14106	18304	39353	81526
	REP	4066	6147	15697	19998	41325	83767
	FENB	2712	4453	13037	16903	36161	74746
	RENB	3864	5779	14744	18744	38486	77772
$\phi = 5$							
AIC	FEP	3710	6338	18914	25275	50986	107382
	REP	4889	7677	20574	27026	52961	109621
	FENB	2245	3846	11703	15432	31654	65768
	RENB	3882	5984	15813	20520	40428	82363
BIC	FEP	3727	6356	18935	25296	51010	107407
	REP	4917	7707	20608	27062	53001	109664
	FENB	2268	3870	11731	15461	31686	65802
	RENB	3910	6014	15848	20556	40467	82406

Table 5: AIC and BIC values of different estimators when $N = 300$

Criterion	Estimator	*T =10*	*T =15*	*T =40*	*T =50*	*T =100*	*T =200*
$\phi = 0.5$							
AIC	FEP	4202	6800	20854	26174	54488	110267
	REP	5889	8696	23289	28693	57410	113565
	FENB	4083	6597	20223	25445	52884	107178
	RENB	5768	8489	22651	27961	55793	110467
BIC	FEP	4220	6819	20877	26197	54513	110294
	REP	5919	8728	23326	28731	57452	113610
	FENB	4107	6623	20253	25476	52917	107214
	RENB	5798	8521	22688	28000	55835	110512
$\phi = 1$							
AIC	FEP	4431	7448	23003	28309	59628	119295
	REP	6146	9362	25432	30864	62582	122638
	FENB	4071	6850	21078	25968	54681	109616
	RENB	5811	8817	23707	28796	58194	114064
BIC	FEP	4449	7467	23025	28332	59653	119322
	REP	6176	9394	25469	30902	62624	122683
	FENB	4095	6875	21108	25998	54715	109652
	RENB	5841	8849	23744	28834	58236	114109
$\phi = 5$							
AIC	FEP	5883	9147	29922	38116	77376	157627
	REP	7684	11133	32446	40751	80353	160991
	FENB	3511	5616	18192	23092	47518	96938
	RENB	6026	8748	24631	30809	60771	121400
BIC	FEP	5901	9166	29945	38139	77401	157654
	REP	7714	11165	32483	40789	80394	161036
	FENB	3535	5642	18221	23122	47551	96974
	RENB	6057	8781	24668	30847	60812	121445

Table 6: AIC and BIC values of different estimators when $N = 500$

Criterion	Estimator	*T* =10	*T* =15	*T* =40	*T* =50	*T* =100	*T* =200
$\phi = 0.5$							
AIC	FEP	6987	11517	34800	44163	90093	188100
	REP	9790	14666	38819	48395	94901	193680
	FENB	6790	11182	33758	42842	87566	182370
	RENB	9590	14324	37757	47064	92357	187937
BIC	FEP	7007	11537	34823	44187	90119	188128
	REP	9823	14700	38858	48436	94946	193728
	FENB	6816	11210	33789	42874	87601	182408
	RENB	9622	14359	37797	47105	92402	187984
$\phi = 1$							
AIC	FEP	7415	12082	37232	47737	95237	198462
	REP	10264	15275	41235	52005	100090	204052
	FENB	6822	11102	34245	43772	87761	182291
	RENB	9718	14374	38580	48461	93483	189719
BIC	FEP	7435	12103	37256	47761	95264	198491
	REP	10297	15310	41275	52045	100134	204100
	FENB	6848	11129	34277	43805	87797	182329
	RENB	9751	14409	38620	48502	93527	189767
$\phi = 5$							
AIC	FEP	9209	15020	46861	61780	128929	273759
	REP	12152	18304	51005	66135	133911	279428
	FENB	5610	9280	28915	37965	79252	165683
	RENB	9679	14416	39209	50474	101376	207999
BIC	FEP	9228	15040	46884	61804	128955	273788
	REP	12185	18339	51044	66176	133955	279476
	FENB	5636	9308	28947	37998	79287	165721
	RENB	9712	14451	39248	50514	101420	208046

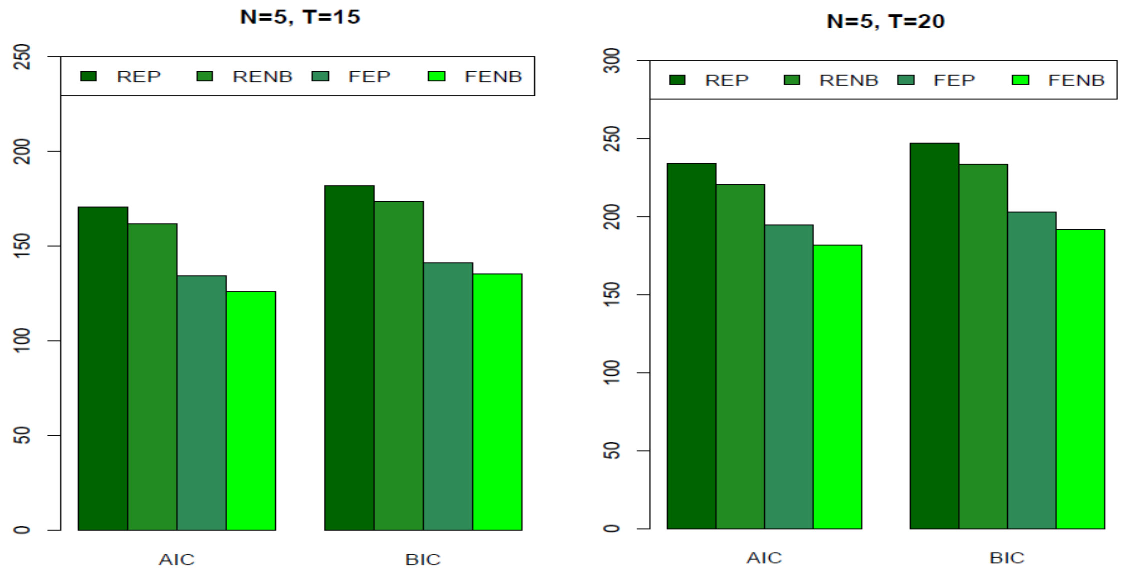

Fig. 1: AIC and BIC values of different estimators when $N = 5$

Fig. 2: AIC and BIC values of different estimators when N = 10

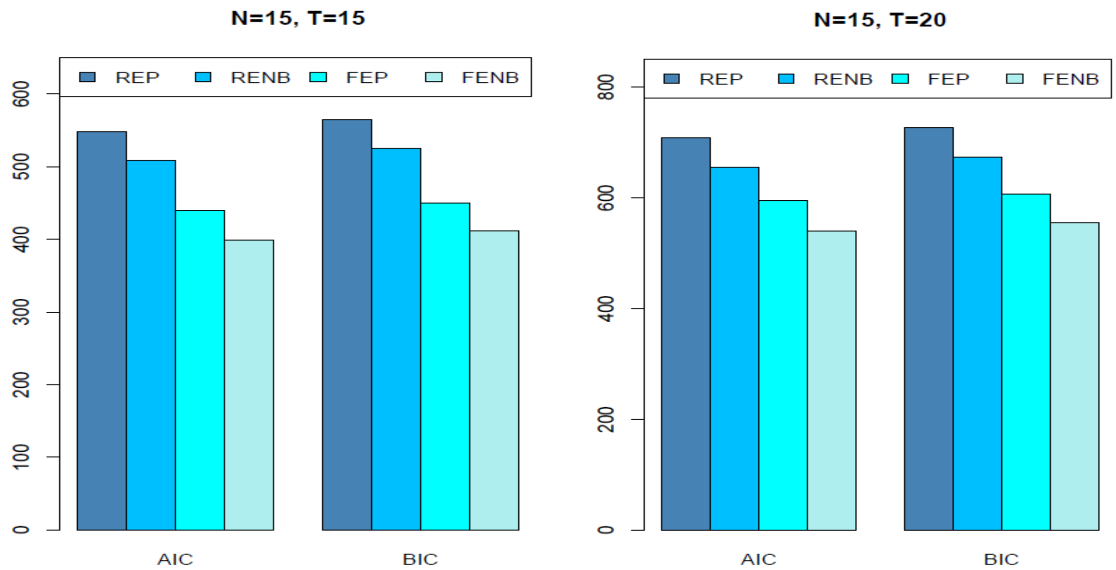

Fig. 3: AIC and BIC values of different estimators when N = 15

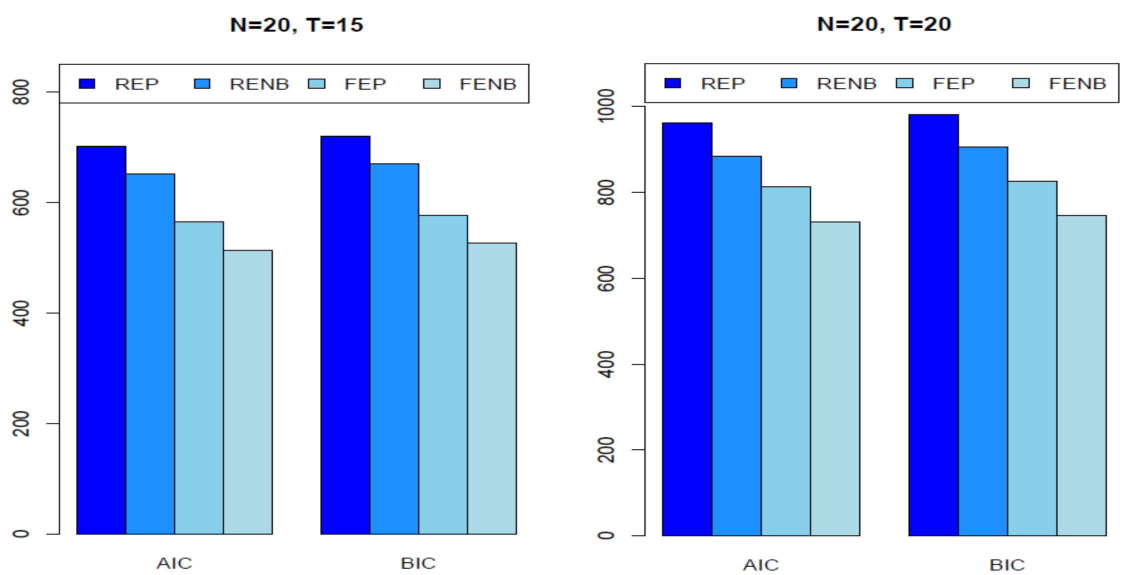

Fig. 4: AIC and BIC values of different estimators when N = 20

6. EMPIRICAL STUDY: PATENTS IN ARAB COUNTRIES

There are many economic studies are interested with patent applications, e.g. [9, 13, 16, 29, 30, 31, 32]. In our application, we will follow the same methodology presented by Youssef et al. [13], their methodology is summarized the estimation steps and how to select the appropriate model for the data based on the Hausman test and the goodness-of-fit measures (AIC and BIC). Youssef et al. [13] estimated the number of patents for seventeen high-income countries in the world over the period from 2005 to 2016, while in this application, the sample was chosen based on the available data on the number of patents in Arab countries in the World Bank website. Our sample contains seven Arab countries: Egypt, Algeria, Jordan, Morocco, Saudi, Tunisia, and Yemen over the period from 2000 to 2016.

In our study, the dependent variable is the number of patent applications, and three explanatory variables: GDPC, IMPO, and UNEM; where GDPC is the gross domestic product per capita (U.S. Dollar), IMPO is the information and communication technology goods imports (percentage of total goods imports), and UNEM is unemployment rate (percentage of total labor force).

We repaired the data before estimating the parameters of CPD models. The data contains some missing values in the number of patent and IMPO, these missing values were estimated using the mean-imputation method [21, 33]. We performed a unit root test for all variables, and the results indicated that the data are stationary in the level [34]. The variance inflation factor (VIF) is calculated to check the multicollinearity problem of the explanatory variables, the results indicated that the data not have multicollinearity problem because all values of VIF less than five. For more details on how to deal with the multicollinearity problem in regression models, see e.g. [20, 35, 36].

We estimated the parameters in fixed effects models using CML method, while the ML estimation method was used to estimate the random effects models. Table 7 presents the results of FEP and REP models, the two models are statistically significant because the P-value of the Wald test is less than 0.05. Based on the results of Hausman test, the P-value of chi-squared is greater than 0.05, then we can accept the null hypothesis, this means that REP model is more appropriate.

Table 8 presents the results of CML estimates of FENB model and ML estimates of RENB model. The two (FENB and RENB) models are statistically significant because the P-value of the Wald test is less than 0.05. Since the P-value of Hausman test is greater than 0.05, then the RENB model is more appropriate.

Table 7: Estimates of Poisson panel models

Variable	**Fixed Effects Poisson Model**			**Random Effects Poisson Model**		
	Estimate	Z-value	P-value	Estimate	Z-value	P-value
GDPC	.0001355	48.07	0.001	.0001354	48.05	0.001
IMPO	-6.586289	-8.38	0.001	-6.58876	-8.38	0.001
UNEM	-.7792349	-2.31	0.021	-.7814001	-2.32	0.021
Intercept	--------	--------	--------	5.002777	13.15	0.001
Wald Test	$\chi^2 = 2350.23$, df = 3, P-value (χ^2) < 0.001			$\chi^2 = 2348.58$, df = 3, P-value (χ^2) < 0.001		
Hausman Test	$\chi^2 =$ 0.23, df = 3, P-value (χ^2) = 0.8896					

Table 8: Estimates of negative binomial panel models

Variable	**Fixed Effects NB Model**			**Random Effects NB Model**		
	Estimate	Z-value	P-value	Estimate	Z-value	P-value
GDPC	.0000551	2.96	0.003	.0000578	3.30	0.001
IMPO	-5.762925	-1.28	0.201	-4.332509	-0.98	0.329
UNEM	.7660229	0.54	0.592	.6834497	0.47	0.637
Intercept	.9827058	3.21	0.001	.9091168	2.98	0.003
Wald Test	$\chi^2 = 9.48$, df = 3, P-value (χ^2) = 0.0235			$\chi^2 = 11.50$, df = 3, P-value (χ^2) = 0.0093		
Hausman Test	$\chi^2 =$ 2.26, df = 3, P-value (χ^2) = 0.3225					

Based on the results from tables 7 and 8, we can conclude that REP and RENB models are more fit to this data than FEP and FENB models. Then we should use AIC and BIC to determine the best model (REP or RENB model). Table 9 shows that the RENB model has minimum values of AIC and BIC, and then the RENB model is the best model to fit the data.

Table 9: Goodness-of-fit measures of random effects models

Measure	Random Effects Poisson	Random Effects Negative Binomial
Log likelihood	-2701.233	-666.526
AIC	5412.465	1345.051
BIC	5426.361	1361.726

In the RENB model, we find that GDPC is statistically significant because the P-value of Z-value for this variable is less than 0.05, while IMPO and UNEM variables are not statistically significant.

7. CONCLUSION

In this paper, we used the Monte Carlo simulation for making a comparison study between the four estimation methods of CPD models. Furthermore, we examined the effect of some economic variables on the number of patent applications in seven Arab countries by applying four CPD models. We can summarize the main conclusions of our Monte Carlo simulation and the empirical study in the following points:

1. When the dispersion parameter equal one, the FENB estimator is better than FEP and REP estimators according to AIC and BIC values. Moreover, in case of increasing dispersion parameter value, the AIC and BIC values of FENB estimator is decreasing dramatically and the AIC and BIC values of FENB estimator is smaller than FEP, REP, and RENB estimators.
2. When the values of the number of units or time period are increased, the values of AIC and BIC of all CPD estimators are increasing in all simulation situations.
3. In general, simulation results indicated that the AIC and BIC values of FENB estimator is smaller than the AIC and BIC values of FEP, REP, and RENB estimators for all cases of the simulation. Thus, the FENB estimator is better than FEP, REP, and RENB estimators.
4. In our application, we examined the effect of some economic variables on the number of patents in seven Arab countries over the period from 2000 to 2016 by applying four CPD models to explore the main variables that effect on the number of patent applications in these countries. Based on the Hausman test and model-selection criteria (AIC and BIC), we found

that the RENB estimator is the appropriate for this data, because it has minimum AIC and BIC values. RENB results indicated that the GDP per capita has a positive significant effect on the number of patents in Arab countries, and the other variables have not significant effect.

In future work, we plan to study the efficiency of ML estimators in case of outliers [19, 21, 26] or missing data [21, 33] in CPD models. Moreover, we can study the impact of the COVID-19 pandemic [37] or the food and non-food expenditures [38, 39] on the number of patents in the Arab countries using modern CPD models.

CONFLICT OF INTERESTS

The author(s) declare that there is no conflict of interests.

REFERENCES

[1] B. Baltagi, Econometric analysis of panel data, Wiley, New York, (2008).

[2] P.N. Vijayamohanan, Panel data analysis with Stata Part 1 fixed effects and random effects models, MPRA Paper, (2016), 76869.

[3] W.H. Greene, Econometric analysis, Pearson Education. New York, (2012).

[4] M.R. Abonazel, Generalized estimators of stationary random-coefficients panel data models: Asymptotic and small sample properties, Revstat Stat. J. 17 (2019), 493-521.

[5] A.C. Cameron, P.K. Trivedi, Regression analysis of count data, Cambridge University Press, (2013).

[6] C. Hsiao, Analysis of panel data, Cambridge University Press, (2014).

[7] D.N. Gujarati, D.C. Porter, Basic econometrics, Tata McGraw-Hill Education, (2009).

[8] J.M. Wooldridge, Econometric analysis of cross section and panel data, MIT press, (2010).

[9] R. Blundell, R. Griffith, F. Windmeijer, Individual effects and dynamics in count data models, J. Econ. 108 (2002), 113-131.

[10] S.S.P. Kumara, H.C. Chin, Study of fatal traffic accidents in Asia Pacific countries, Transport. Res. Record, 1897 (2004), 43-47.

[11] R.B. Noland, M.A. Quddus, Analysis of pedestrian and bicycle casualties with regional panel data, Transport. Res. Record, 1897 (2004), 28-33.

[12] J.F. Angers, D. Desjardins, G. Dionne, F. Guertin, Modelling and estimating individual and firm effects with count panel data, ASTIN Bull.: J. Int. Actuarial Assoc. 48 (2018), 1049-1078.

[13] A.H. Youssef, M.R. Abonazel, E.G. Ahmed, Estimating the number of patents in the world using count panel data models, Asian J. Probabil. Stat. 6 (2020), 24-33.

[14] R. Winkelmann, Econometric analysis of count data, Springer-Verlag Berlin Heidelberg, (2008).

[15] J.P. Boucher, M. Denuit, Fixed versus random effects in Poisson regression models for claim counts: A case study with motor insurance, ASTIN Bull.: J. Int. Actuarial Assoc. 36 (2006), 285-301.

[16] J.A. Hausman, B.H. Hall, Z. Griliches, Econometric models for count data with an application to the patents-R&D relationship, Econometrica, 52 (1984), 909-938.

[17] B.H. Baltagi, The Oxford handbook of panel data, Oxford University Press, (2015).

[18] Y. Croissant, G. Millo, Panel data econometrics with R, John Wiley and Sons Ltd, (2019).

[19] M.R. Abonazel, S.M. El-sayed, O.M. Saber, Performance of robust count regression estimators in the case of overdispersion, zero inflated, and outliers: simulation study and application to German health data, Commun. Math. Biol. Neurosci. 2021 (2021), Article ID 55.

[20] M.R. Abonazel, A practical guide for creating Monte Carlo simulation studies using R, Int. J. Math. Comput. Sci. 4 (2018), 18-33.

[21] M.R. Abonazel, Handling outliers and missing data in regression models using R: simulation examples, Acad. J. Appl. Math. Sci. 6 (2020), 187-203.

[22] A.H. Youssef, M.R. Abonazel, Alternative GMM estimators for first-order autoregressive panel model: an improving efficiency approach, Commun. Stat.-Simul. Comput. 46 (2017), 3112-3128.

[23] E. Plan, Modeling and simulation of count data, CPT: Pharmacometrics Syst. Pharmacol. 3(2014), 1-12.

[24] M.S. Sapuan, A.M. Razali, Z.H. Zamzuri, K. Ibrahim, Simulation on Poisson and negative binomial models of count road accident modeling, AIP Conf. Proc. 1784 (2016), 050009.

[25] M.R. Abonazel, Different estimators for stochastic parameter panel data models with serially correlated errors, J. Stat. Appl. Probab. 7 (2018), 423-434.

[26] M.R. Abonazel, O.M. Saber, A comparative study of robust estimators for Poisson regression model with outliers. J. Stat. Appl. Probab. 9 (2020), 279-286.

[27] H. Akaike, A new look at the statistical model identification, IEEE Trans. Autom. Control, 19 (1974), 716-723.

[28] G. Schwarz, Estimating the dimension of a model, Ann. Stat. 6 (1978), 461-464.

[29] B. Crépon, E. Duguet, Estimating the innovation function from patent numbers: GMM on count panel data, J. Appl. Econ. 12 (1997), 243-263.

[30] M. Cincera, Patents, R&D, and technological spillovers at the firm level: some evidence from econometric count models for panel data, J. Appl. Econ. 12 (1997), 265-280.

[31] P. Wang, L.M. Cockburn, M.L. Puterman, Analysis of patent data—a mixed-Poisson-regression-model approach, J. Bus. Econ. Stat. 16 (1998), 27-41.

[32] S. Gurmu, F. Pérez-Sebastián, Patents, R&D and lag effects: evidence from flexible methods for count panel data on manufacturing firms, Empirical Econ. 35 (2008), 507-526.

[33] E.A. Rady, M. Abdallah, M.R. Abonazel, A proposed generalized method of moment estimation of the panel vector autoregressive model with missing data, J. Comput. Theor. Nanosci. 18 (2021), 1-7.

[34] M.R. Abonazel, O.A. Shalaby, Using dynamic panel data modeling to study net FDI inflows in MENA countries, Stud. Econ. Econ. 44 (2020), 1-28.

[35] A.F. Lukman, B. Aladeitan, K. Ayinde, M.R. Abonazel, Modified ridge-type for the Poisson regression model: simulation and application, J. Appl. Stat. (2021), 1–13. https://doi.org/10.1080/02664763.2021.1889998.

[36] M.R. Abonazel, I.M. Taha, Beta ridge regression estimators: simulation and application, Commun. Stat.-Simul. Comput. (2021), 1-13. https://doi.org/10.1080/03610918.2021.1960373

[37] F.A. Awwad, M.A. Mohamoud, M.R. Abonazel, Estimating COVID-19 cases in Makkah region of Saudi Arabia: Space-time ARIMA modeling, PLOS ONE 16 (2021), e0250149.

[38] F.A. Awwad, S. Abdel-Rahman, M.R. Abonazel, Estimating equivalence scales and non-food needs in Egypt: Parametric and semiparametric regression modeling, PLOS ONE 16 (2021), e0256017.

[39] S. Abdel-Rahman, M.R. Abonazel, New measure of catastrophic health expenditures with application on rural Egypt. Middle East Develop. J. (2021), 1-26. https://doi.org/10.1080/17938120.2021.1958560