Building an Efficient Model by Using Panel Data

Obaid M. Al-Zawbaee, Nozad H. Mahmood

Department of Business Administration, Cihan University-Sulaimaniya,
Sulaimaniya, 46001, Kurdistan Region, Iraq

***Abstract*—The panel data models have gained great attention because they consider the effect of changing time series and the cross-sectional units. It has a higher number of degrees of freedom and is, therefore, more efficient. Our research aims to build the most efficient model using panel data through comparison among six statistical models and to determine the most efficient model through the mean squares of error (RMSE) and the coefficient of determination (R^2). Using the EViews-12 package to apply to two data sets to determine the most efficient model. The more critical findings are that the hypothesis assumed by the research states that the accuracy of the adopted models varies. Furthermore, it was found by building six types of panel data models that the best model is the (two-way fixed effect model) because it achieves the lowest value of (RMSE) and the largest value of R^2.**

***Keywords*—Panel data, Time series, Cross section data, Fixed effect model, Random effect model.**

I. Introduction

The panel data models have gained great attention because they consider the effect of changing time series and the effect of changing the difference between the cross-sectional units, in addition to the fact that the panel models do not require the condition of the size of the time series required for analysis. In addition, the panel analysis is superior to the analysis of time series alone or cross-sectional data alone because it includes more informational content than time series or cross-sectional ones, thus, the possibility of obtaining higher-confidence estimates. On the other hand, the problem of autocorrelation between variables is less severe than in time series data. It has a higher number of degrees of freedom and is, therefore, more efficient (Baltagi and Baltagi, 2008; Dielman, 1998; Lee and Huang, 2002; Mohammed et al., 2022).

Our research aims to build the most efficient model using panel data through comparison among six statistical models and to determine the most efficient model through the mean squares of error (RMSE) and the coefficient of determination (R^2). It was applied to two data sets to determine the most efficient model.

The importance of the research comes through the application of statistical theory to determine the most efficient model for panel data by comparing six types of models. The study is based on the deductive analytical method, which employs the foundations of statistical theory and analyzes panel data to arrive at the most efficient model.

Several researchers investigated the topic of the study, and the most critical research and studies on the subject were:

The study (2020) tried to find out what causes economic growth in the Arab Gulf countries and where that growth comes from by using panel data. The results showed that the best model is the two-way fixed effect model and that more needs to be done to train workers and expand the current employment policy (Alyasari and Alzawbaee, 2020).

Another study (2014) was conducted to identify the determinants of economic growth in the Gulf countries using the VECM model. The variables of foreign investments, exports, and local investments were used between 1972 and 2012. The study concluded that the model works for every country except for the Sultanate of Oman (Altaee, et al., 2014).

Another study (2006) that came out used Panel VAR to find that the Arabian Gulf countries were using more electricity because their economies were growing (Al-Iriani, 2006).

A. Panel data models

Panel data models are better than cross-sectional data models or time series data on their own in several situations, including:

It has more information, more degrees of freedom, better efficiency, and less linear multiplicity among the variables.

It can solve the problem of heterogeneity in cross-sectional and time-series data with specific variance (Yaffee, 2003; Wooldridge, 2010; Whang, 2020).

Cihan University-Erbil journal of Humanities and Social Sciences (CUEJHSS)
Volume VII No. 1 (2023) 6 pages
DOI: 10.24086/cuejhss.vol7n1y2023.pp 82-87
Received 1 November 2022; Accepted 19 December 2022; Regular research paper: Published: 10 March 2023
*Corresponding author's e-mail: nozad.mahmood@sulicihan.edu.krd

Basically, we have three primary types of panel data models (PLOS, Random Effects, and Fixed Effects), which further become six distinct types.

Pool ordinary least square (PLOS)

Its formula is:

$$y_{it} = \beta_0 + x_{it}\,\beta + \epsilon_{it} \quad (1)$$

Where:

Var(ϵ_{it}) = Б2, E(Ƹ(Ƹ) = 0

We use the ordinary least square (OLS) method to estimate the parameters of the above model (Allaoui and Chkir, 2006; Begg et al., 2014).

Fixed effects model

In the POLS model, we assumed that the cross-sections (the colleges in our research) are homogeneous in terms of size and capabilities, but in fact, they are different, and for this purpose, it is logical to consider these differences, and the model that takes these differences into account is the fixed effects model one-way fixed effect model (Artelaris et al., 2006; Ghura, 1997).

$$y_{it} = (\alpha + u_i) + x_{it}\beta + \epsilon_{it} \quad (2)$$

(ƸƸƸ) is a random variable in the fixed effects model, and it is assumed to be associated with the independent variables where:

i=1,2 or =1,2,3,….

To consider the effects that occur between colleges and for each research college during the period, the two-way fixed effect model has been re-characterized as follows:

$$y_{it} = (\alpha + u_i + \gamma_t) + x_{it}\beta + \epsilon_{it} \quad (3)$$

Random effect model

It was assumed that the error term is normally distributed with a mean equal to zero and a constant variance for all cross sections within the fixed effects model. However, the truth may be different, so the random effects model is adopted, in which the error term is divided into two components that account for the changes that occur to the cross sections. As a result, the random effects model is one-way (a one-way random effect model), and its formula is as follows (Dielman, 1998; Khamfula, 2007; Bruno and Easterly, 1998):

$$y_{it} = \alpha + x_{it}\,\beta + (u_i + w_{it}) \quad (4)$$

Alternatively, dividing the error into three components that consider the changes that occur for the cross sections and the time series together. The two-way random effect model and its formula are: (Beddies, 1999; Breusch and Pagan 1980; Greene, 2003).

$$y_{it} = \alpha + x_{it}\,\beta + (u_i + \gamma_t + w_{it}) \quad (5)$$

Mixed model

Its formula:

$$Y = \beta_0 + \Sigma\,\beta iXi + \gamma + \epsilon \quad (6)$$

Where

Y=Dependent variable

β_0=Global intercept

βi=Fixed effect parameters

Xi=Fixed effect variables

γ=Random effect variance

ϵ=Residual variance.

II. Methodology

Based on the above information, this research aims to identify the most appropriate panel data models for estimating parameters for two distinct datasets (the number of graduated students) with three independent variables, and (the number of completed research) with two independent variables. Several vital criteria, including RMSE, R^2, and other tests to demonstrate the proper models among POLS, one-way random effect, two-way random effect, one-way fixed effects, two-way fixed effect, and mixed effect models, must be considered when deciding on a model that will improve consistency and efficiency. For examining our data, we utilize EViews-12 as a software package.

A. Practical application

There are two different data sets, as shown in Appendixes I and II. The number of colleges (nineteen) represents the cross sections, and the years (three) are the time series. Furthermore, the following variables were studied (Breihi and Abid, 2017):

The first group:

Y = Dependent Variable – The number of graduated students

X1 = Independent variable – The number of accepted students

X2 = Independent variable – The number of teachers

X3 = Independent variable – Number of technicians and administrators

The second group:

Y = Dependent variable – Number of completed research

X1 = Independent variable – Number of teachers

X2 = Independent variable – Number of technicians administrators

III. Results and Discussion

A. Results of the First Data Set

The first dataset includes one dependent variable, Y (The number of graduated students), and three independent variables, X1 (The number of accepted students), X2 (The number of teachers), and X3 (The number of technicians and administrators). We took logs for variables to reduce heteroscedasticity and normalize our data.

First, we have to test the stationary of the mean and variance of the times series. In other words, the variables' time series should not contain a unit root or be nonstationary. We tested the following hypotheses and used the (LLC, IPS, ADF-Fisher, and PP-Fisher) tests with all three methods (level, first difference, and second difference) to see how stable our data were, and the results are in Table I:

H_0: There is a unit root for the series

H_1: There is no unit root for the series (The series is stationary).

Since the *P*-values of all the tests (LLC, IPS, ADF-Fisher, and PP-Fisher) are less than α = 0.005 in all the unit

TABLE I
Unit Root Test for Stationary

Unit root test in	Test types	Series: Y, X1, X2, X3	Results
Level	Levin, Lin, and Chu t*	−3.3233	Stationary
		0.0004	
	Im, Pesaran. and Shin W-stat	−4.1263	
		0.0000	
	ADF-Fisher Chi-square	35.5937	
		0.0000	
	PP-Fisher Chi-square	45.5310	
		0.0000	
First difference	Levin, Lin, and Chu t*	−8.2446	Stationary
		0.0000	
	Im, Pesaran and Shin W-stat	−9.4116	
		0.0000	
	ADF-Fisher Chi-square	90.6706	
		0.0000	
	PP-Fisher Chi-square	130.4800	
		0.0000	
Second difference	Levin, Lin, and Chu t*	−6.52038	Stationary
		0.0000	
	Im, Pesaran, and Shin W-stat	−15.1934	
		0.0000	
	ADF-Fisher Chi-square	148.803	
		0.0000	
	PP-Fisher Chi-square	137.777	
		0.0000	

*=0.05

tests (level, first difference, and second difference), there is sufficient evidence to reject the null hypothesis and accept the alternative one. Hence, since the variables are stationary in time series, our data are appropriate for panel data analysis.

To select the appropriate model, Hauseman's test is used to choose between the random model effect and fixed model effect, which states the following hypothesis:

H_0: Random model rather than a fixed model is appropriate

H_1: Random model is not approporiate.

According to the results in Table II, since the *P* = 0.0001 and smaller than the significant level, we reject the null hypothesis and accept the alternative hypothesis. In other words, the fixed effect model is more appropriate than the random effect model.

Now let's decide between the fixed effect model and the PLOS model by using redundant fixed effects-likelihood ratio tests:

H_0: Pooled least square model rather than fixed model is appropriate

H_1: Pooled Least square model is not approporiate.

Table III illustrates that since *P*-values are less than the significant level, there is evidence to reject the null hypothesis and accept the alternative. In other words, it indicates that the fixed model is more appropriate for analyzing our data than the POLS.

However, the above tests are only used to compare the fixed effect model with each random effect and the POLS model. To select the most appropriate model for our data, we have to see the root-mean-square error (RMSE) and the coefficient of determination (R^2) criteria. The model with the least RMSE and the greatest indicates the best model.

TABLE II
Cross section Random-hausman Test

Test summary	χ^2 statistic	χ^2 df	Probability
Cross section random	27.791138	3	0.0000

TABLE III
Redundant Fixed Effects-Likelihood Ratio Tests

Effects test	Statistic	df	Probability
Cross section F	4.303988	18.35	0.0001
Cross section χ^2	66.539193	18	0.0000

Since the two-way fixed effect model has the least RMSE of (0.205828) and the greatest R^2 of (0.928603), compared to the other models, it is the best model to analyze our data set. As a result, ln_x1 is the only significant variable due to its small *P*-value. In other words, the number of students is the only predictor variable that influences the number of graduated students (Table IV).

B. Results of the Second Dataset

Our second dataset includes one dependent variable, Y (The number of completed research), and two independent variables, X1 (The number of teachers) and X2 (The number of technician administrators). First, we took logs for variables to reduce heteroscedasticity and normalize our data. The time series should be stationary in their mean and variance. For that purpose, we employed the LLC, IPS, ADF-Fisher, and PP-Fisher tests to see whether our datasets are stationary or not.

H_0: There is a unit root for the series

H_1: There is no unit root for the series (The series is stationary).

There is evidence to reject the null hypothesis and accept the alternative one as the p-values of all the tests (LLC, IPS, ADF-Fisher, and PP-Fisher) are less than the significant levels at unit tests (level, first difference, and second difference). Therefore, panel data analysis can be applied to our data since the variables are stationary over time (Table V).

We use the Hauseman's test to choose the appropriate model between the random model effect and fixed effect model within the following hypothesis:

H_0: Random model rather than fixed model is appropriate

H_1: Random model is not approporiate.

Based on Table VI, we reject the null hypothesis and accept the alternative hypothesis because the *P*-value is 0.0002. In other words, it is better to use the fixed effect model instead of the random effect model.

Now, let's employ the redundant fixed effects-likelihood ratio tests to choose between the fixed effect model and the PLOS model.

H_0: Pooled least square model rather than fixed model is appropriate

H_1: Pooled least square model is not appropriate.

According to the results in Table VII, since p-values are less than the significant level, we should reject the null hypothesis. In other words, it indicates that the fixed model

TABLE IV
Model Selection Based on the Root Mean Square Error and R^2

Model	Variables	Coefficients	P	RMSE	R^2
Pooled least square	C	1.175536	0.1048	0.38846	0.74569
	LN_X1	0.714167	0.0000		
	LN_X2	0.355851	0.0025		
	LN_X3	−0.420507	0.0178		
One-way random effect	C	1.219720	0.1357	0.326556	0.738486
	LN_X1	0.648288	0.0000		
	LN_X2	0.324073	0.0104		
	LN_X3	−0.310958	0.1045		
Two-way random effect	C	1.221734	0.1261	0.303896	0.739990
	LN_X1	0.654766	0.0000		
	LN_X2	0.328251	0.0076		
	LN_X3	−0.324507	0.0790		
One-way fixed effect	C	4.218381	0.0026	0.216700	0.920862
	LN_X1	0.100896	0.0871		
	LN_X2	−0.021053	0.8890		
	LN_X3	0.001992	0.9886		
Two-way fixed effect	C	4.074158	0.0066	0.205828	0.928603
	LN_X1	0.114315	0.0478		
	LN_X2	0.045286	0.7400		
	LN_X3	−0.049089	0.7167		
Mixed effect	C	4.150965	0.0034	0.211177	0.924546
	LN_X1	0.106266	0.0091		
	LN_X2	0.013019	0.9266		
	LN_X3	−0.023961	0.8607		

RMSE: Root-mean-square error

TABLE V
Unit Root Test for Stationary

Unit root test in	Test types	Series: Y, X1, X2, X3	Results
Level	Levin, Lin, and Chu t*	−4.80283	Stationary
		0.0000	
	Im, Pesaran, and Shin W-stat	−4.43911	
		0.0000	
	ADF-Fisher Chi-square	35.2795	
		0.0000	
	PP-Fisher Chi-square	40.6645	
		0.0000	
First difference	Levin, Lin, and Chu t*	−15.2477	Stationary
		0.0000	
	Im, Pesaran, and Shin W-stat	−14.6944	
		0.0000	
	ADF – Fisher Chi-square	125.405	
		0.0000	
	PP – Fisher Chi-square	122.156	
		0.0000	
Second difference	Levin, Lin, and Chu t*	1.60566	Stationary
		0.094580	
	Im, Pesaran, and Shin W-stat	−11.5959	
		0.0000	
	ADF – Fisher Chi-square	102.900	
		0.0000	
	PP – Fisher Chi-square	55.2620	
		0.0000	

*=0.05

is more appropriate for analyzing our data than the POLS. Furthermore, to select the most appropriate model for our data, we must see the root mean square error (RMSE) and the coefficient of determination (R^2) criteria. The model with the least RMSE and the greatest indicates the best model.

TABLE VI
Cross section Random-hausman Test

Test summary	χ^2 statistic	χ^2 df	Probability
Cross section random	19.830685	3	0.0002

TABLE VII
Redundant Fixed Effects-likelihood Ratio Tests

Effects test	Statistic	df	Probability
Cross section F	2.177557	18.36	0.0231
Cross section χ^2	41.985030	18	0.0011

TABLE VIII
Model Selection Based on the Root-Mean-Square Error and R^2

Model	Variables	Coefficients	P	RMSE	R^2
Pooled least square	C	0.872235	0.3101	0.651922	0.280218
	LN_X1	0.450889	0.0012		
	LN_X2	0.187595	0.4776		
One-way random effect	C	1.079382	0.2702	0.555202	0.276965
	LN_X1	0.394404	0.014		
	LN_X2	0.193976	0.5081		
Two-way random effect	C	1.172297	0.228	0.531269	0.277686
	LN_X1	0.417743	0.009		
	LN_X2	0.144951	0.6161		
One-way fixed effect	C	2.884864	0.0023	0.4415872	0.655405
	LN_X1	−0.080635	0.0134		
	LN_X2	0.231174	0.0113		
Two-way fixed effect	C	2.726912	0.1061	0.422038	0.698344
	LN_X1	0.130205	0.0078		
	LN_X2	0.047467	0.9132		
Mixed effect	C	1.212313	0.21	0.434719	0.671675
	LN_X1	0.435653	0.0069		
	LN_X2	0.115461	0.6908		

The two-way fixed effect model is the best fit for our data, with an RMSE of (0.422038) and an R^2 (0.698344), respectively. LN_X1 (The number of teachers) is a statistically significant variable with a low *P*. RMSE: Root-mean-square error

Table VIII shows the two-way fixed effect model is the best fit for our data, with an RMSE (0.422038) and an R^2 (0.698344), respectively. LN_X1 (The number of teachers) is a statistically significant variable with a low *P*-value.

IV. Conclusions

The most significant findings and suggestions from the study are:

1. The research has accepted the hypothesis that the accuracy of the selected models varies.
2. The data used for both groups is stationary in the mean; thus, differences are not necessary, and the logarithmic conversion was performed to increase the smoothing of the data and to overcome the assumption that the data are nonstationary in its variances.
3. The cross-sectional random-Hausman test and the redundant fixed effects-likelihood ratio tests confirmed the efficiency of the fixed effect model.
4. After utilizing six various models for panel data, it was determined that the best model is the two-way fixed effect model due to its large (R_2) and small (RMSE) values. RMSE= 0.205828, R^2 = 0.928603, and RMSE= 0.422038, R^2 = 0.968344 for the first and second groups, respectively.

References

Al-Iriani, M.A. (2006). Energy-GDP relationship revisited: An example from GCC countries using panel causality. Energy Policy, 34(17), 3342-3350.

Allaoui, Z., & Chkir, A. (2006). Mondialisation, Convergence et Croissance Économique: Une Analyse par les Données de Panel. MPRA Paper 7306. Germany: University Library of Munich.

Altaee, H.H.A., Saied, S.M., Esmaeel, E.S., & Adam, M.H.M. (2014). Financial development, trade openness and economic growth: Evidence from sultanate of Oman (1972-2012). Journal of Economics and Sustainable Development, 5(23), 64.

Alyasari, H.H.A., & Alzawbaee, O.M.M. (2020). Panel data evidence on the relationship between economic growth, electricity consumption and CO2 emissions in the GCC. The Scientific Journal of Cihan University-Slemani, 33, 47.

Artelaris, P., Arvanitidis, P., & Petrakos, G. (2006). Theoretical and Methodological Study on Dynamic Growth Regions and Factors Explaining their Growth Performance (No. DYNREG02). Economic and Social Research Institute (ESRI).

Baltagi, B.H., & Baltagi, B.H. (2008). Econometric Analysis of Panel Data. Vol. 4. Berlin: Springer.

Beddies, M.C.H. (1999). Investment, Capital Accumulation, and Growth: Some Evidence from the Gambia 1964–98. United States: International Monetary Fund.

Begg, D., Vernasca, G., Fischer, S., & Dornbusch, R. (2014). EBOOK: Economics: United States: McGraw Hill.

Breihi, F.K, & Abid, A.N. (2017). Measure the Relative Efficiency of the Faculties of the University of Anbar Using the Style Envelope data, Economics and Administration Studies Journal (EASJ) (formerly AL-DANANEER JOURNAL), 1(11), 1-41.

Breusch, T.S., & Pagan, A.R. (1980). The Lagrange multiplier test and its applications to model specification in econometrics. The Review of Economic Studies, 47(1), 239-253.

Bruno, M., & Easterly, W. (1998). Inflation crises and long-run growth. Journal of Monetary Economics, 41(1), 3-26.

Dielman, T.E. (1989). Pooled cross-sectional and time series data analysis: Dekker.

Ghura, M.D. (1997). Private Investment and Endogenous Growth: Evidence from Cameroon: United States: International Monetary Fund.

Greene, W.H. (2003). Econometric Analysis Pearson Education India. Delhi, India: Pearson Education India.

Khamfula, Y. (2007). Foreign direct investment and economic growth in EP and IS countries: The role of corruption. World Economy, 30(12), 1843-1854.

Lee, C.H., & Huang, B.N. (2002). The relationship between exports and economic growth in East Asian countries: A multivariate threshold autoregressive approach. Journal of Economic Development, 27(2), 45-68.

Mohammed, P.A., Obed, S.A., Ali, I.M., & Kadir, D.H. (2022). US Dollar/IQ Dinar currency exchange rates time series forecasting using ARIMA model. Cihan University-Erbil Scientific Journal, 6(1), 12-19.

Whang, Y.J. (2019). Econometric Analysis of Stochastic Dominance: Concepts, Methods, Tools, and Applications: Cambridge: Cambridge University Press.

Wooldridge, J.M. (2010). Econometric Analysis of Cross Section and Panel Data: Cambridge: MIT press.

Yaffee, R. (2003). A Primer for Panel Data Analysis. New York: Connect-Information Technology at New York University.

Appendices

APPENDIX I
THE NUMBER OF GRADUATE STUDENTS, THE NUMBER OF ACCEPTED STUDENTS, AND THE TEACHING STAFF AND NUMBER OF TECHNICIANS AND ADMINISTRATORS

College and Faculty	**Y**	**X1**	**X2**	**X3**	**Y**	**X1**	**X2**	**X3**	**Y**	**X1**	**X2**	**X3**
College of Education for Girls	343	612	121	39	259	632	121	62	385	500	128	71
Faculty of Dentistry	21	105	58	65	36	98	62	55	43	75	70	71
Faculty of Administration and Economics - Ramadi	140	173	32	33	105	203	39	35	79	150	11	39
college of Literature	243	255	134	55	174	393	130	56	145	300	143	70
Faculty of Administration and Economics - Fallujah	68	250	29	49	55	112	33	52	75	120	15	28
College of Education - Al-Qaim	83	190	25	25	36	162	28	28	93	120	11	34
Faculty of Physical Education	63	99	22	20	85	89	27	21	84	85	51	46
Computer College	76	99	30	38	90	118	33	35	90	100	36	41
Faculty of Agriculture	188	301	147	73	195	402	164	59	183	390	174	59
College of Veterinary Medicine	22	41	30	59	25	50	32	63	31	50	16	40
College of General Medicine	58	134	96	110	62	125	93	86	40	100	115	105
College of Islamic Sciences – Fallujah	89	142	63	45	57	78	64	46	144	150	72	60
College of Islamic Sciences – Ramadi	149	274	82	27	158	158	84	30	108	200	84	30
College of Science	193	150	99	62	183	204	101	62	189	160	58	24
College of Law – Fallujah	93	96	25	43	42	68	23	45	54	57	23	45
Faculty of Law and Political Science	65	201	32	28	70	75	41	30	76	80	41	30
College of Engineering	149	205	127	82	161	196	140	85	185	185	142	104
College of Education for Human Sciences	664	698	158	41	301	646	174	43	476	300	150	55
College of Education for Pure Sciences	249	295	98	51	226	226	150	54	252	205	361	54

Y: Dependent variable – The number of graduate students, X1: Independent variable – The number of accepted students, X2: Independent variable Â Number of teachers, X3: Independent variable – Number of technicians and administrators

APPENDIX II
THE NUMBER OF RESEARCH COMPLETED, NUMBER OF TEACHER AND NUMBER OF TECHNICIANS ADMINISTRATORS

College and Faculty	**Y**	**X1**	**X2**	**Y**	**X1**	**X2**	**Y**	**X1**	**X2**
College of Education for Girls	37	121	39	143	121	62	109	128	71
Faculty of Dentistry	10	58	65	12	62	55	15	70	71
Faculty of Administration and Economics – Ramadi	33	32	33	35	39	35	40	11	39
College of Literature	65	134	55	25	130	56	50	143	70
Faculty of Administration and Economics – Fallujah	8	29	49	35	33	52	26	15	28
College of Education – Al-Qaim	17	25	25	12	28	28	15	11	34
Faculty of Physical Education	12	22	20	15	27	21	18	51	46
Computer College	79	30	38	24	33	35	34	36	41
Faculty of Agriculture	70	147	73	21	164	59	36	174	59
College of Veterinary Medicine	30	30	59	13	32	63	50	16	40
College of General Medicine	17	69	110	44	93	86	135	115	105
College of Islamic Sciences – Fallujah	37	63	45	13	64	46	62	72	60
College of Islamic Sciences – Ramadi	48	82	27	62	84	30	32	84	30
College of Science	51	99	62	57	101	62	60	58	24
College of Law – Fallujah	15	25	43	12	23	45	11	23	45
Faculty of Law and Political Science	9	32	28	10	41	30	20	41	30
College of Engineering	69	127	82	97	140	85	78	142	104
College of Education for Human Sciences	87	158	41	55	174	43	65	150	55
College of Education for Pure Sciences	64	98	51	10	150	54	51	361	54

Y: Dependent variable – Number of research completed, X1: Independent variable – Number of teachers, X2: Independent variable – Number of technicians administrators