Common Correlated Effects Estimation of Hybrid Panel Data Models

Hassan Hopoop Razaq[1] and Mohammed Sadiq Abdul Razzaq[2]
[1]Department of Economics, College of Administration and Economics, Thi-Qar University, 64001 Thi-Qar, Iraq
[2]Department of Statistics, College of Administration and Economics, Baghdad University,10001 Baghdad, Iraq
hasanhopop@utq.edu.iq, dr_aldouri@coadec.uobagdad.edu.iq

Keywords: Common Correlated Effects, Hybrid Coefficients Model, Unbalanced Panel Data, Mean Group Estimator, Pooled Estimator, Half Jackknife Panel Estimator.

Abstract: In this research, a model of panel data models was reviewed which is hybrid coefficients model which is characterized by a portion of the regression coefficients being fixed slopes while the other portion of the coefficients are random slopes meaning that they have a normal distribution with an unknown mean and variance. Several methods were used to estimate the parameters of this model in the case of unbalanced panel data. these estimation methods depend on the common correlated effects estimator which is composed of three estimators, common correlated effect mean group estimator(CCEMG), common correlated effect pooled (CCEP) and half jackknife panel (HJP)estimator to estimate the parameters of the hybrid coefficients model represented by the first fixed slope and the mean of the random slope coefficient. Monte Carlo experiments and different sample sizes (NT) are small, medium and large, with different variance levels to compare between estimation methods, the simulation results showed that the (CCEP) is the best estimation method because it has the less average mean absolute error (AMAE). The (CCMG) is the best method after (CCEP).

1 INTRODUCTION

Since the beginning of the fifties of the twentieth century, applied studies have taken a new path characterized by the increasing use of theoretical hypotheses that are in the form of primary information derived from outside the scope of the sample and provided by the theory behind the phenomenon studied or previous studies and combining them with sample data whether cross – sectional data or time series data, to obtain estimates of the parameters of the model studied that more efficient than the estimators based on time series or cross sectional alone. methods used to combine time series and cross-sectional data at the single equation level involve using the cross-sectional data to estimate some model parameters and then introducing these estimators as constant value constraints into the time series regression equation.

Panel data consists of cross-sectional data and time series data. the behavior of data categories is observed over a period of time. These categories maybe (countries, companies,…,etc) [1]. Observation from the same category is correlated and correlation is the distinguishing feature of these data. They are also known as longitudinal data. panel data have group effects (cross sectional) or have time effects or both. panel data are divided into two types, balanced panel data which contain equal time periods for all cross sections but if the time period differs across cross sections, it is called unbalanced panel data. the main objective of this research is to estimate the parameters of hybrid coefficients model for unbalanced panel data and to compare the estimation methods to show the best estimation methods.

2 PANEL DATA REGRESSION MODELS

Regression analysis is one of the important topics of statistics, which is widely used in all fields of science and knowledge because it describes the relation between variables in the form of an equation it can be known in general as the analysis that specializes in studying the effect of one or more variables called the independent variable or independent variables on one variable called the dependent variable [2]. This is for the purpose of estimating or predicting the value of the dependent variable given the information of the independent variable or variables.

Accordingly, the regression model is used to arrive at mathematical model that explains the quantitative relationship between the dependent and the independent variables. Therefore, it can be defined as a statistical method used to analyze data that contains two or more variables when the goal is to discover the nature of this relationship.

In general, regression models can be classified according to several factors [3]:

1) The number of the variables. There is common division of regression models according to the number of independent variables used in the models. if the model contains one independent variable with the dependent variable, the model is called a simple regression model. however, if the number of independent variables is more than one, it is called a multiple regression model.
2) The form of the relation between the variables. Here we can distinguish between two types of regression models, the linear regression model and nonlinear regression model.
3) Level of measurement for variables. This classification depends on the type of data for the dependent variable, whether it is descriptive or quantitative data. here, it can be divided in two types of regression models, logistic regression models it is used when the data of the dependent variable is descriptive. the second is the traditional regression model, which is used when the data of the dependent variable is quantitative data. if the quantitative data is time series data, then time series regression models are used, and if the quantitative contains cross sections, then cross – sectional data regression models are used. but if the dependent variable and the independent variables are time series and cross – sectional data, then panel data models are used because the independent variables and the dependent variable have two dimensions, which are the cross – section and time.

In the current decade, panel data models have gained great interest especially in economic and medical studies because they take into account the effect of change in time as well as the effect change in cross – sectional observations. many researchers studied panel data models, some of whom were interested in studying the properties of these models mathematically [4], [5]. Among those who were interested in applying these models in their studies [6], [7].

The panel data has the following properties [1], [8]:

1) it allows controlling individual variation that may appear in the case of cross – sectional or temporal data which leads to biased results;
2) it helps control some variables that remain constant among individuals but change over time, such as national policies and international agreements;
3) panel data contain more information content that cross – sectional or temporal data, and therefore it is possible to obtain estimates with higher confidence and the problem of common correlation between variables is less severe than time series data;
4) panel data distinguished from others in that it includes a greater number of degrees of freedom and is characterized by better efficiency as well as multicollinearity between variables and more information content if cross – sectional or temporal data are used;
5) panel data models provide a better possibility to study the adjustment dynamics that may be hidden by cross – sectional data, and they are also suitable for studying periods of economic conditions such as unemployment poverty, growth and others. on the other hand, it is possible through panel data to link the behaviors of the sample observations from one point in time to another;
6) panel data models contribute to reducing the possibility of the problem of neglected variables resulting from unobserved item characteristics that usually lead to biased estimates;
7) the importance of using panel data is that takes into account what is described as heterogeneity or unobserved difference specifics to the sample observations whether cross – sectional or temporal;
8) these models help prevent the common problem of heteroscedasticity.

3 THE MODEL

In simple and multiple regression model, the regression coefficients are fixed parameters so we seek to estimate these parameters using classical or Bayesian methods [9]. However, when the regression coefficients(slopes) are random parameters and not fixed, then they are called random coefficient (RCR) model [10].

In the research, we will discuss a model that contains both fixed and random regression coefficients, which is given by the following [11]:

$$Y_{it} = X_{1it}\beta_1 + X_{2it}\beta_{2i} + u_{it}, \quad (1)$$

where i=1,2,…,N represent cross section unit and t=1,2,…,T_i represent time series period.

Model (1) is unbalanced panel data, meaning that there are N individuals observed over varying time periods length (i=1,2, …, T_i). the model in (1) can be rewritten by stacking over time period t:

$$Y_i = X_{1i}\beta_1 + X_{2i}\beta_{2i} + u_i, \quad (2)$$

where $Y_i = (Y_{i1}, \dots, Y_{iT_i})'$,$X_{1i}, X_{2i}$ are matrices of rank $T_i \times K_1$ and $T_i \times K_2$ for explanatory variables and $\beta_1$is fixed slope of rank $K_1 \times 1$, β_{2i} is assumed random slope of rank $K_2 \times 1$, finally u_i is the random error of rank $T_i \times 1$. the model (2) has the following assumptions [12]:

Assumption 1: the random error has $E(u_i) = 0$ and $var(u_i) = \sigma_u^2 I_{T_i}$ $i = 1,2, \dots, N$.

Assumption 2: the slope coefficients β_{2i} are independent and distributed with:

$$E(\beta_{2i}) = \bar{\beta}_{2,} var(\beta_{2ij}) = \begin{cases} \Delta & i = j \\ 0 & i \neq j \end{cases} \; i,j = 1, \dots, N.$$

From assumption (2) can be written the random slope [13]:

$$\beta_{2i} = \bar{\beta}_2 + \mu_i, \quad (3)$$

where $\bar{\beta}_2 = (\bar{\beta}_{21}, \dots, \bar{\beta}_{2K_2})'$is a vector of constant parameters and $\mu_i = (\mu_1, \dots, \mu_{K_2})'$ with:

$$E(\mu_i) = 0, var(\mu_{ij}) = \begin{cases} \Delta & i = j \\ 0 & i \neq j \end{cases} \; i,j = 1, \dots, N,$$

where Δ is a K_2 diagonal matrix. we note that the random slope in assumption (3) is constant over time periods but changes over the cross sections due to the change in the random error.

By substituting (3) into (2), we get:

$$Y_i = X_{1i}\beta_1 + X_{2i}\bar{\beta}_2 + \varepsilon_i, \quad (4)$$

where $\varepsilon_i = X_{2i}\mu_i + u_i$, by rewriting the above model, we get:

$$Y = Q\bar{\theta} + \varepsilon, \quad (5)$$

where $Y = (Y_1, \dots, Y_N)', Q = (Q_1', \dots, Q_N')', \bar{\theta} = (\beta_1', \bar{\beta}_2')', Q_i = (X_{1i}, X_{2i}), \varepsilon = (\varepsilon_1', \dots, \varepsilon_2')'$.

4 ESTIMATION

In this section we will use different estimation methods to estimate the parameters of the hybrid coefficients model in the (5). These methods based on CCE estimator and are as follows:

4.1 CCE Mean Group Estimator

The common correlated effect estimator proposed by [14] to estimate panel data model with random regression coefficient. the CCEMG for estimate parameters in model (5) is given by [15]:

$$\hat{\bar{\theta}}_{CCEMG} = \frac{1}{N}\sum_{i=1}^{N} \hat{\theta}_{CCE,i}, \quad (6)$$

where $\hat{\theta}_{CCE,i}$is the (CCE) for each cross – sectional i which equal to:

$$\hat{\theta}_{CCE,i} = (Q_i' M_H Q_i)^{-1} Q_i' M_H Y_i, \quad (7)$$

where:

$$M_H = I_{T_i} - (H'H)^{-1}H',$$
$$H = (\bar{Q}, \bar{Y}),$$
$$\bar{Q} = (\bar{Q}_1, \dots, \bar{Q}_T)', \bar{Y} = (\bar{Y}_1, \dots, \bar{Y}_T),$$
$$\bar{Q}_t = \frac{1}{N}\sum_{i=1}^{N} Q_{it}, \bar{Y}_t = \frac{1}{N}\sum_{i=1}^{N} Y_{it},$$

The (CCEMG) is asymptotically distributed as [10]:

$$\sqrt{N}\left(\hat{\bar{\theta}}_{CCEMG} - \bar{\theta}\right) \rightarrow N(0, \hat{\gamma}_{CCEMG}),$$

where $\hat{\gamma}_{CCEMG}$ the variance of estimator in (6).

4.2 CCE Pooled Estimator

The CCEP estimator is calculated when the individuals slope coefficient β_{2i} are differ from cross–sectional to another. the CCEP is given by [14]:

$$\hat{\bar{\theta}}_{CCEP} = [\sum_{i=1}^{N} W_i Q_i' M_H Q_i]^{-1} \sum_{i=1}^{N} W_i Q_i' M_H Y_i, \quad (8)$$

where (W_i) represent weights for each i:

$$W_i = \frac{\sigma_i^{-2}}{\sum_{i=1}^{N} \sigma_i^{-2}}, \quad (9)$$

and

$$\sigma_i^2 = \frac{(Y_i - Q_i\hat{\theta}_{CCE,i})' M_H (Y_i - Q_i\hat{\theta}_{CCE,i})}{T_i}.$$

4.3 Half Jackknife Panel Estimator

This method was proposed by Dhaene and Jochmans (2012) [17]. the HJP estimator depend on CCE by dividing the time series within each cross – section into two halves and calculating the CCE for each half. the HJP is [16]:

$$\hat{\bar{\theta}}_{HJP} = 2\hat{\bar{\theta}}_{CCEMG} - \frac{1}{2}\left(\hat{\bar{\theta}}^{[1]}_{CCEMG} - \hat{\bar{\theta}}^{[2]}_{CCEMG}\right), \quad (10)$$

where $\hat{\bar{\theta}}_{CCEMG}$ is defined in (6), $\hat{\bar{\theta}}^{[1]}_{CCEMG}$ is calculated from time series t=1,2,…, $T_i/2$ in each cross–sectional i and $\hat{\bar{\theta}}^{[2]}_{CCEMG}$ is calculated from time series t=(T_i/2)+1,…,T_i.

5 THE SIMULATION STUDIES

In this section, we will compare between CCEMG, CCEP and HJP for different sample sizes, Monte Carlo simulation was used to generate variables of the hybrid model. the program used to write the Monte Carlo study is in the R Language.
Rewrite the model:

$$Y_{it} = X_{1it}\beta_1 + X_{2it}\beta_{2i} + u_{it}, \quad (11)$$
i=1,2,…, N t=1,2,…,T_i,

in this study, the values of the explanatory variables were generated with a normal distribution with mean 0 and variance equal to one, allowing them to differ within each cross – sectional. the fixed slope coefficient is chosen equal to $\beta_1 = 2,4$ and 6, the random slope coefficient were generated from assumption (3): $\beta_{2i} = \bar{\beta}_2 + \mu_i$ with mean $\bar{\beta}_2$ equal to 0.1, 0.2 and 0.4 and the random error $\mu_i$were generated as normal distribution with mean 0 and variance equal to 25,30. The random error u_{it} were generated with normal distribution with mean equal to 0 and variance equal to 2,4 and 6. the values of N, T were chosen from different values (N=T= 6, 10, 12, 14 and 16). thus, the sample size in the case of balanced panel data is equal to (n=NT), while in the unbalanced panel data, the sample size is equal to $n = \sum_{i=1}^{N} T_i$. for each N cross sections there is cross section containing a time series that differs from the other sections to obtain unbalanced panel data. in this study the third cross section for each N was made to contain four individual of time series i.e (Ti=4 for i=3). Table 1 includes the results for AMAEwhen β_1 =2, $\beta_{2i} \sim N(0.1,25)$, Table 2 when β_1 = 4, $\beta_{2i} \sim N(0.2,30)$ and Table 3 when β_1 = 6, $\beta_{2i} \sim N(0.4,30)$.

5.1 Results

In this section we will calculated the (AMAE) for estimation methods CCEMG, CCEP and HJPE for compare between these estimators. the AMAE is given by:

$$AMAE(Y) = \frac{1}{R}\sum_{i=1}^{R} MAE(Y),$$

where R represent the number of replicates of the experiments equal to 1000. and MAE equal to:

$$MAE(Y) = \frac{1}{n}\sum_{i=1}^{n} |Y_i - \hat{Y}_i|.$$

Table 1: AMAE when$\beta_1 = 2, \beta_{2i} \sim N(0.1,25)$.

N=T	Estimators	σ_u^2		
		2	4	6
6	CCEMG	1.58890	2.44088	3.78405
	CCEP	.101954	.124670	1.46042
	HJP	8.81388	51.68297	18.55046
8	CCEMG	1.25650	1.62931	2.16993
	CCEP	0.93390	0.98810	1.20740
	HJP	3.76633	6.04728	17.28449
10	CCEMG	0.93546	1.23607	1.77261
	CCEP	0.63189	0.81115	0.95730
	HJP	12.99293	23.05623	24.81867
12	CCEMG	0.46961	1.32733	1.35627
	CCEP	0.58385	0.74383	0.89261
	HJP	2.48124	2.78504	5.79645
14	CCEMG	0.80965	0.88404	1.09568
	CCEP	0.55475	0.66024	0.77811
	HJP	2.23708	2.08343	3.11329
16	CCEMG	0.56952	0.67403	0.706467
	CCEP	0.68169	0.73699	0.78706
	HJP	1.61418	1.49426	2.94651

From Table 1 the best estimation method is CCEP as it has the lowest AMAE, except for N=T= 12 and σ_u^2 =2 where the results showed superiority the CCMG. The HJP has highest AMAE it was the worst estimation method. These results are illustrated in Figure 1.

Table 2: AMAE when$\beta_1 = 4, \beta_{2i} \sim N(0.2,30)$.

N=T	Estimators	σ_u^2		
		2	4	6
6	CCEMG	1.01183	3.05672	4.52806
	CCEP	1.15999	1.38826	1.60470
	HJP	8.01609	12.79155	22.28190
8	CCEMG	1.02024	1.86489	2.31472
	CCEP	1.03473	1.09711	1.18190
	HJP	5.08173	9.67312	10.27363
10	CCEMG	0.06092	1.24146	1.89807
	CCEP	0.75763	0.89222	1.08447
	HJP	33.79948	23.27028	15.88350
12	CCEMG	0.72896	1.06704	1.94116
	CCEP	0.75843	0.81948	0.97422
	HJP	2.39712	2.33144	4.08247
14	CCEMG	0.40919	1.02268	1.11696
	CCEP	0.69528	0.74736	0.76605
	HJP	7.22877	1.88920	3.76969
16	CCEMG	0.27739	0.90346	0.99819
	CCEP	0.51079	0.73086	0.76833
	HJP	1.01465	7.84661	1.70827

From Table 2 the best estimation method is CCEP as it has the lowest AMAE,except for N=T=8 N=T=10, N=T=12 and N=T=16 with σ_u^2 =2 where the results showed superiority the CCMG. The HJP has highest AMAE it was the worst estimation method. Figure 2 visualizes these AMAE results for easier comparison.

Table 3: AMAE when $\beta_1 = 6, \beta_{2i} \sim N(0.4,30)$.

N=T	Estimators	σ_u^2		
		2	4	6
6	CCEMG	1.86835	3.32966	4.18577
	CCEP	0.93177	1.28543	1.54104
	HJP	10.41008	22.00085	30.92379
8	CCEMG	1.22375	1.54505	2.65452
	CCEP	0.81278	1.11075	1.29558
	HJP	3.49339	6.90341	13.76258
10	CCEMG	1.44390	1.09846	1.41257
	CCEP	0.71073	0.82516	0.88389
	HJP	14.04935	13.72030	18.08526
12	CCEMG	0.74377	1.18751	1.67669
	CCEP	.062070	.080228	0.87733
	HJP	1.37923	3.89849	4.67962
14	CCEMG	.076134	1.81599	1.08066
	CCEP	0.54181	0.72836	0.77666
	HJP	1.06013	3.45888	1.96723
16	CCEMG	0.73231	0.81515	1.50622
	CCEP	0.58370	0.74749	0.78464
	HJP	1.07678	1.15790	3.40047

From Table 3 the best estimation method is CCEP as it has the lowest AMAE for all values of N=T. the second place was taken by the CCEMG. the AMAE values decrease with increasing values of N, T and increase with increasing the variance levels. the HJP method has fluctuations in the AMAE values. The trends in AMAE for this scenario are depicted in Figure 3.

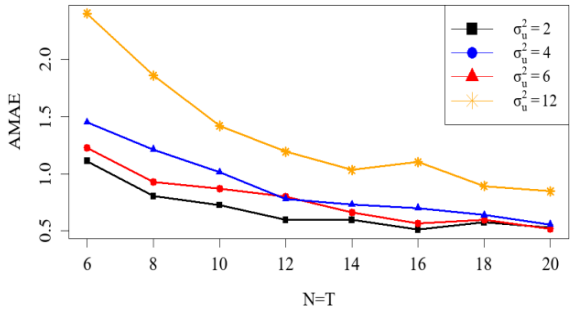

Figure 1: CCEP for $\beta_1 = 2, \beta_{2i} \sim N(0.1,25)$.

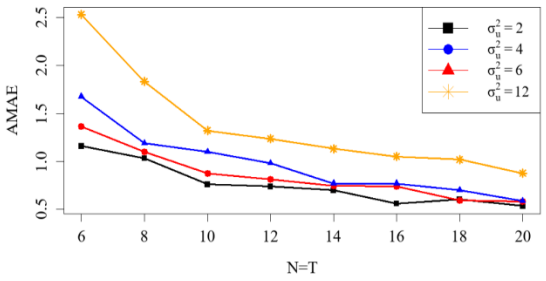

Figure 2: CCEP for $\beta_1 = 4, \beta_{2i} \sim N(0.2,30)$.

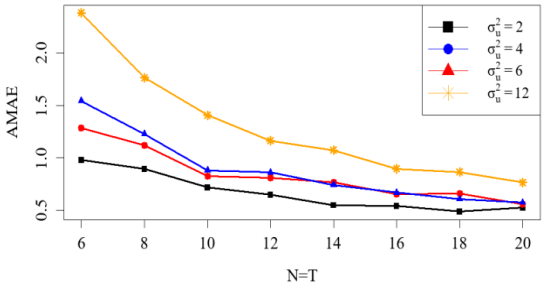

Figure 3: CCEP for $\beta_1 = 6, \beta_{2i} \sim N(0.4,30)$.

6 CONCLUSIONS

In this paper, we represented different methods for estimating panel data model, these methods are CCEMG, CCEP and HJP used to estimate hybrid coefficients model with unbalanced panel data. we performed Monte Carlo simulation study. the simulation results showed that the best and most appropriate is CCEP as it has the smallest value of average mean absolute error. We also observe from the simulation results tables for the estimation methods that the AMAE values increase with increasing variance levels and decrease with increasing values of N, T. given the differences in variances, number of cross – sections and time series, the researchers recommend using the common correlated effect pooled method for hybrid panel data model.

REFERENCES

[1] B. Baltagi, Econometric Analysis of Panel Data, 3rd ed., England, 2005.

[2] F. N. Gumedze and T. T. Dunne, "Parameter estimation and inference in the linear mixed model," Linear Algebra and Its Applications, vol. 435, pp. 1920-1944, May 2011.

[3] L. Horvath and L. Trapani, "Statistical inference in a random coefficient panel data," London, pp. 1-46, 2016, [Online]. Available: http://www.elsevier.com/open-access/userlicense/1.0/.

[4] M. C. Bramati and C. Croux, "Robust estimators for the fixed effects panel data model," The Econometrics Journal, vol. 10, no. 3, pp. 521-540, 2007, [Online]. Available: https://doi.org/10.1111/j.1368-423X.2007.00220.x.

[5] L.-F. Lee and J. Yu, "Estimation of spatial autoregressive panel data models with fixed effects," Journal of Econometrics, vol. 154, no. 2, pp. 165-185, 2010, [Online]. Available: https://web.pdx.edu/~crkl/WISE/SEAUG/papers/Lee_Yu_JE10.pdf.

[6] C.-S. Chuang and L. Wang, "Semiparametric estimation of fixed-effects panel data varying coefficient models," Advances in Econometrics, vol. 25, pp. 61-91, 2009, [Online]. Available: https://www.scribd.com/document/972936943/Semiparametric-Estimation-of-Fixed-effects-Panel-d.

[7] H. Qin and C. Wang, "Semiparametric inference for varying coefficient panel data models with fixed effects," Journal of Statistical Planning and Inference, vol. 141, no. 1, pp. 369-382, 2011, [Online]. Available: https://mpra.ub.uni-muenchen.de/18850/1/QianWang20091110.pdf.

[8] C. Hsiao, Analysis of Panel Data, 4th ed., University of Southern California, 2022.

[9] P. A. V. B. Swamy, "Efficient inference in random coefficient regression model," Econometrica, vol. 38, no. 2, pp. 311-323, 1970.

[10] M. R. Abonazel, "Different estimators for stochastic parameter panel data models with serially correlated errors," Journal of Statistics Applications & Probability, vol. 7, no. 3, pp. 423-434, Nov. 2018.

[11] M. R. Abonazel, "Efficiency comparisons of different estimators for panel data models with serially correlated errors: A stochastic parameter regression approach," International Journal of System Science and Applied Mathematics, vol. 3, no. 2, pp. 37-51, Jul. 2018, doi: 10.11648/j.ijssam.20180302.14.

[12] K. P. Kalirajan, "On the estimation of a regression model with fixed and random coefficient," Journal of Applied Statistics, vol. 17, no. 2, pp. 237-244, Jun. 2011.

[13] C. Hsiao and M. H. Pesaran, "Random coefficient panel data models," IZA Discussion Paper no. 1236, pp. 1-38, Aug. 2004.

[14] A. Chudik and M. H. Pesaran, "Common correlated effects estimation of heterogeneous dynamic panel data models with weakly exogenous regressors," CESifo Working Paper no. 4232, pp. 1-60, May 2013.

[15] K. P. Kalirajan, "On the estimation of a regression model with fixed and random coefficients," Journal of Applied Statistics, vol. 17, no. 2, pp. 237-244, 1990, doi: 10.1080/75758283512.

[16] R. K. Crump, N. Gospodinov, and L. L. Gaffney, "A jackknife variance estimation for panel regressions," Federal Reserve Bank of New York Staff Reports, no. 1133, Oct. 2014, [Online]. Available: http://doi.org/10.59576/sr.1133.

[17] G. Dhaene and K. Jochmans, "Split-panel Jackknife Estimation of Fixed-effect Models," The Review of Economic Studies, vol. 82, no. 3, pp. 991-1030, 2015.