Original Research Article: full paper

(2023), «EUREKA: Physics and Engineering»
Number 1

HANDLING MULTI-COLLINEARITY USING PRINCIPAL COMPONENT ANALYSIS WITH THE PANEL DATA MODEL

Ahmed Hassen Youssef
Department of Applied Statistics and Econometrics[1]

Engy Saeed Abozaid
Department of Applied Statistics and Econometrics[1]

Shereen Hamdy Abdel Latif ✉
Department of Applied Statistics and Econometrics[1]
shereen_hamdy_@cu.edu.eg

[1]*Cairo University*
5 Dr. Ahmed Zewail str., Orman, Giza, Egypt, 12613

✉**Corresponding author**

Abstract

When designing a statistical model, applied researchers strive to make the model consistent, unbiased, and efficient. Labor productivity is an important economic indicator that is closely linked to economic growth, competitiveness, and living standards within an economy. This paper proposes the one-way error component panel data model for labor productivity. One of the problems that we can encounter in panel data is the problem of multi-collinearity. Therefore, multi-collinearity problem is considered. This problem has been detected. After that, the principal component technique is used to get new good unrelated estimators. For the purposes of our analysis, the multi-collinearity problem between the explanatory variables was examined, using principal component techniques with the application of the panel data model focused on the impact of public capital, private capital stock, labor, and state unemployment rate on gross state products. The analysis was based on three estimation methods: fixed effect, random effect, and pooling effect. The challenge is to get estimators with good properties under reasonable assumptions and to ensure that statistical inference is valid throughout robust standard errors. And after application, fixed effect estimation turned out to play a key role in the estimation of panel data models. Based on the results of hypothesis testing, the real data result showed that the fixed effect model was more accurate compared to the two models of random effect and pooling effect. In addition, robust estimation was used to get more efficient estimators since heteroscedasticity has been confirmed.

Keywords: fixed effect estimation, Hausman test, panel data, principal component analysis, robust regression.

DOI: 10.21303/2461-4262.2023.002582

1. Introduction

The use of panel data was first introduced by [1] in an analysis of public opinion, using market research gathered over time [2], also used to study the behavior of firms and wages of people over time. Periods of time are often years, but the span between periods can be longer or shorter than a year.

Panel data sometimes referred to as longitudinal data, is a dataset in which the behavior of entities is observed across time. The term «panel data» refers to the pooling of observations on a cross-section of, say, firms, countries, etc., over several periods [3]. A panel data regression differs from a traditional cross-section or time-series regression in that it has a double subscript on its variables if we have T periods. ($t = 1, 2, ..., T$) and N the number of individuals ($n = 1, 2, ..., N$), then with panel data we will have total observation units of $N \times T$ [4].

One of the major benefits from using panel data as compared to cross-section data on individuals is that it enables us to control for individual heterogeneity. Not controlling for these unobserved individual specific effects leads to bias in the resulting estimates.

Multicollinearity increases the standard errors of the coefficients. Increased standard errors in turn mean that coefficients for some independent variables may be found not to be significantly different from 0.

Unfortunately, data may be suffered from multi-collinearity. It is occurs when independent variables in a regression model are correlated. This correlation is a problem because independent

variables should be independent. If the degree of correlation between variables is high enough, it can cause problems when to fit the model and interpret the results. If the degree of correlation between variables is high enough, it can cause problems in the case of fitting the model, it makes it hard to interpret coefficients, and reduces the power of the model to identify statistically significant independent variables. These are serious problems. By overinflating the standard errors, multi-collinearity makes some variables statistically insignificant when they should be significant. Without multi-collinearity (with lower standard errors), those coefficients might be significant, [5].

One of the important steps in statistical analysis is to detect and deal with the multi-collinearity problem throughout the examination of the correlation matrix. One of the most important techniques used to handle multi-collinearity is the principal component analysis method.

Then three estimation methods were implemented to estimate the one-way error component model: fixed effect, random effect, and pooling effect estimation method. After carrying out the chow test and Hausman test, it was found that the fixed effect model is the most appropriate model, which focused on the impact of public capital, private capital stock, labor, and state unemployment rate on gross state products. And it plays a key role in the estimation of panel data models.

The challenge is developing an estimator's assumptions with good properties under reasonable assumptions and ensuring that statistical inference is valid throughout robust standard errors.

The use of panel data models comes from the fact that data used in many social sciences usually combines time series and cross-sections of units [6].

[7] pointed out that modality known as «fixed effects models» (FE) or «covariance models» allows the intercept to differ across cases, but not overtime (time-invariant). FE assumes that the slope coefficients are constant while the intercept varies across cross-sectional units. This type of approach considers individuality by letting the intercept vary across cases while slope coefficients are assumed to be constant across firms.

Multi-collinearity defined by [8] when the predictors are uncorrelated, all eigenvalues of the design matrix are equal to one and the design matrix is full rank. If at least one eigenvalue is different from one, especially when equal to zero or near zero, then non-orthogonally exists, meaning that multi-collinearity is present.

The high R-squared (say > 0.8) may indicate the problem of multi-collinearity mentioned by [9].

Principal component analysis (PCA) calculates an uncorrelated set of variables (components or PCs). These components are ordered so that the first few retain most of the variation present in all the original variables. This new system summarizes the total data variation in decreasing order so that the first new variable has the largest variation, the second has the second largest, and so on. These new variables are the principal components were proposed by [10].

To choose between fixed or random effects, there were two tests to check: Hausman Test, and the Breusch-Pagan Lagrange Multiplier test proposed by [11].

The chow test shows that the best method is the common effect of the fixed effect. The next step is to determine whether the common effect is better than the random effect, then the Lagrange Multiplier Test is required, and the Hausman test shows that the best method is the random effect of the fixed effect mentioned by [4].

The properties of the FE estimator and its robust variance-covariance matrix also showed that tests based on these robust standard errors are consistent if $N \to \infty$, regardless of the relative size of N and T even in cases where the data is equicorrelated. The fixed effect always gives consistent estimates, but they may not be the most efficient studied by [12].

For the within estimator, [13] suggested a simple method for obtaining robust estimates of the standard errors that allow for a general variance-covariance matrix on the ν_{it}.

[3] proposed the hetroskedasticiy problem without detecting the multicolinearity problem between the explanatory variables, therefore our main aim in this research is to detect multi-colinearity problem with principal component analysis on the panel data and this is not applied before with panel data. Principal component analysis technique applied to EEG data as [10] but the advantage of this research is to apply the PCA technique to panel data and get new unrelated variables and get efficient estimators. [4] estimated the regression model with panel data can be done through

three approaches, our research also used this approach to determine the most appropriate method that can be used to detect the multicolinearity problem.

The practical significance of the work is very important for economics sector because the stability and change are essential elements of social reality and economic progress. The processes of GSP can be differentiated through the years. Therefore, sample selectivity and biases due to omitted variables can be controlled with panel data.

Cross-sectional surveys are a way to provide information on specific concerns at a specific period, but they don't provide any informΦation about the stability that is currently in place. Retrospective inquiry can yield limited information about change, although this is frequently hampered by «recall bias». However, reliable data on change is necessary to determine if phenomena like poverty are long-lasting or transient. These issues can be resolved through panel data studies, which also offer a crucial tool for successful policy design.

The aim of this study is to determine and get new good efficient estimators than its existing competitive estimators in panel data applications. The PCA Technique has been used to model, for example the effect of public capital stock and private capital stock and some of predictors on gross state product.

To achieve this aim, the following objectives are accomplished:

– to investigate the correlation matrix and VIF between the explanatory variables to detect if there is a multicolinearity problem between the predictors or not;

– to investigate the PCA technique to solve the multicolinearity problem by reducing the main predictors by new unrelated factors;

– to estimate model and selection method of panel data regression: common effect, fixed effect, and random effect model based on some of tests to determine the appropriate method that was used;

– to investigate the robust standard errors and their corresponding *t* values by using Breusch-Pagan test;

– to compare the standard errors in [14] and our robust standard errors to see the efficiently of the robust standard errors.

2. Materials and Methods

2. 1. The one-way fixed effects model

A study on the regression model, each variable cannot be measured or always observed due to at least one slacked variable will always be. To establish more accurate models and to make an accurate analysis, it is important to control the effect of these slacked variables on the model to be used. For the panel data, the fixed effect model assumes that differences between individuals can be accommodated from different intercepts (i. e., the constant coefficient is considered a constant) as it is mentioned by [15]:

$$y_{it} = \alpha_i + \beta_1 x_{it,1} + \beta_2 x_{it,2} + \cdots + \beta_p x_{it,p} + U_{it}. \quad (1)$$

The fixed-effects model consists of (1) and assumptions which $x_{it,p}$, $p = 1,2,...,P$ are non-stochastic variables and y_{it} are response random variables, α_i which are unknown intercept for each entity. U_{it} considered the error term and contains knowledge about α_i in cross-sectional regression models. The fixed-effects model allows α_i to be correlated with the regressors. Including α_i as intercepts, everyone has a different intercept term and the same slope parameters. β is estimated and the individual effects are then calculated per cross-section (or time) using the «Within» method [3].

Fixed Effect remove the effect of those time-invariant characteristics so can assess the net effect of the predictors on the outcome variable. Another important assumption of the fixed effect model is that those time-invariant characteristics are unique to the individual and should not be correlated with other individual characteristics. α_i is individual intercepts (fixed for given n).

Most of the panel data applications utilize a one-way error component model for the disturbances, with $U_{it} = \mu_i + \upsilon_{it}$, where μ_i denotes the unobservable individual-specific effect and υ_{it} denotes the remainder disturbance. In this case, μ_i are assumed to be fixed parameters to be estimated, and the remainder disturbances are stochastic with υ_{it} independent and identically distributed.

The x_{it} is assumed independent of the υ_{it} for all i and t. Or $U_{it} = \lambda_t + \upsilon_{it}$, λ_t denotes the unobservable disturbance depending on only time (this implies a unique intercept coefficient for each time).

No overall intercept is usually included in the model. Under a fixed model, consistency does not require, that the individual intercepts (whose coefficients are the $\alpha_i's$), and U_{it} are uncorrelated, and only $E(x_{it}, U_{it}) = 0$ must hold.

According to the random effect model, it will estimate panel data where interference variables may be interconnected between time and between individuals, and the difference between intercepts is accommodated by the error terms of each unit, i. e., the random effect model is considered by (1) where μ_i or λ_t are random. Also, the pooled effect model is considered by (1) where $\alpha_i = \alpha$ which means that it is constant for all units. The fixed effects estimator will always give consistent estimates, but they may not be the most efficient, the random effects estimator is inconsistent if the appropriate model is the fixed effects model, and the random effects estimator is consistent and most efficient if the appropriate model is random effects model.

2. 2. Principal component analysis

PCA is a data reduction method as it replaces a set of correlated variables with a set of uncorrelated principal components, which represent unobserved characteristics of the population, and is suiTable for dimensionality reduction allowing for the extraction of data features through variance maximization [10].

PCA is particularly useful when the data at hand are large (i. e., multiple variables), big (i. e., multiple observations per variable), and highly correlated. With such high-dimensional data, the goal is to identify a reduced set of features that represent the original data in a lower-dimensional subspace with a minimal loss of information.

PCA is used for studying one Table of observations and variables with the main idea of transforming the observed variables into a set of new variables, the new variables are constructed as weighted averages of the original variables, and called the principal components, or factors.

The first principal component explains the largest proportion of the total variance. If the first few principal components explain a substantial proportion of the total variance, they can be used to represent the original items, thus reducing the number of variables required in models. While principal components analysis is easy to implement.

2. 3. Tests to determine an appropriate model

The following tests can be used to choose the best model to manage the data panel:

− chow test: The Chow test is used to identify whether a model has a fixed effect or a common effect, used most effectively when estimating panel data;

− hausman test: A statistical test to determine whether the model is Fixed Effect is the Hausman test, or Random Effect might be more fitting.

To determine which of the estimation method is appropriate, two tests are carried out, the chow test and Hausman Test.

2. 3. 1. Chow test

Chow test is a test to determine the model of whether Common Effect (CE) or Fixed Effect (FE) is most appropriately used in estimating panel data.

(Null hypnoses: the model is pooling effect estimation) versus alternative (fixed effect estimation is appropriate).

Stages after Chow test: if the Chow test selects a fixed effect, a random effect must then be conducted before a Hausman test is used to determine if the effect is fixed or random. However, if the Chow test chooses a common effect, then the Lagrange multiplier test must be performed to determine whether to select a common effect or a random effect.

2. 3. 2. Hausman test

Objectives of Hausman Test: The Hausman test is a test used to compare fixed effects and random effects and determine which is more effective. If the post-Chow test stage has been reached

and the fixed effect is the chosen outcome, the Hausman test should now be conducted. It is necessary to complete the procedures in order, therefore analyzing fixed effects first, and then moving on to random effects. Hausman test to show if the fixed effect estimation or random effect estimation is appropriately. The Hausman test statistics is:

$$H = (\hat{\beta}_{RE} - \hat{\beta}_{FE})'(V(\hat{\beta}_{RE}) - V(\hat{\beta}_{FE}))(\hat{\beta}_{RE} - \hat{\beta}_{FE}).$$

The statistic H is distributed as χ^2 under the null hypothesis, with degrees of freedom equal to the number of parameters for the time-varying regressors. In the concrete case of panel models, it is known that the FE estimator is consistent in the RE model as well as in the FE model. In the FE model, it is even efficient, in the RE model it has good asymptotic properties it mentioned by [14].

The conclusion that we must make when finished doing the Hausman test:

If Result: $H0$: Select RE ($p > 0.05$);

$H1$: Select FE ($p < 0.05$).

After selecting the appropriate model, it is necessary to take a test to detect heteroscedasticity.

2. 4. Estimation of the Fixed Effect Models

Fixed effect models include Covariance Model, Individual Dummy Variable Model, least squares Dummy Variable Model, and within Estimation model. Unlike LSDV, the «within» estimation does not need dummy variables, but it uses deviations from group (or time) means. That it «within» estimation uses variation within each individual or entity instead of many dummies mentioned by [16]. The fixed effect estimator for β is obtained if to use the deviations from the individual means as variables. The model in individual means is with $\bar{y}_i = \Sigma_t\, y_{it}/T$ and $\bar{\alpha}_i = \alpha_i$, $\bar{u}_i = 0$:

$$\bar{y}_i = \alpha_i + \beta\bar{x}_i' + \bar{U}_i. \tag{2}$$

Subtraction from:

$$y_{it} = \alpha_i + \beta\bar{x}_{it}' + \bar{U}_{it}. \tag{3}$$

And gives:

$$y_{it} - \bar{y}_i = (x_{it} - \bar{x}_i)'\beta + (U_{it} - \bar{U}_i). \tag{4}$$

Where the intercepts vanish. Here the deviation of y_{it} from $\bar{y}_i$ is explained (not the difference between different individuals, $\bar{y}_i$ and $\bar{y}_j$. The estimator for β is called the within or fixed effect estimator. Within refers to the variability (over time) among observations of individual i.

In terms of the fixed effect approach, the observations of the exogenous variables x_{it} were assumed to be independent of the error term υ_{it} for all cross-sections or time periods. According to [16] this is an appropriate specification if one is focusing on a specific set of firms and inference is limited to that set of firms- that is, this is an appropriate specification form for most accounting research.

Fixed Effect Estimator formula:

$$\hat{\beta}_{FE} = (\sum_i \sum_j ((x_{it} - \bar{x}_i)(x_{it} - \bar{x}_i)')^{-1}(x_{it} - \bar{x}_i)(y_{it} - \bar{y}_i)'. \tag{5}$$

This expression is identical to the well-known formula β-hat $= (X'X)^{-1}(X'Y)$ for N (demeaned w. r. t individual i) data with T repeated observations.

2. 5. Test to detect heteroscedasticity

Heteroscedasticity in panel data suggests that heteroscedasticity should be of concern for several reasons, including the following:

1) one-way error component models have two sources of heteroscedasticity: the individual-specific and the remainder error term;

2) graphical evidence is inconclusive;

3) heteroscedasticity can be caused by variables that are not always related to size;

4) the loss in efficiency in fixed-effects.

A more formal, mathematical way of detecting heteroscedasticity is what is known as the Breusch-Pagan test: Let's test for heteroscedasticity that it is possible to find in the {lmtest} package. If the panel data has a heteroscedasticity problem, a robust regression estimation is recommended.

The Breush-Pagan test is a test for the heteroscedasticity of regression errors is Contrary to homoscedasticity, which means «differently distributed», heteroscedasticity refers to «identical scatter». An essential presumption in regression is homoscedasticity; if this presumption is broken, it is not possible to perform analysis.The test assumes the error variances are due to a linear function of one or more explanatory variables in the model. That means heteroscedasticity could still be present in the regression model, but those errors (if present) are not correlated with the explanatory variables.

The test statistic approximately follows a chi-square distribution proposed by [17]:

− the null hypothesis for this test is that the error variances are all equal;

− the alternate hypothesis is that the error variances are not equal.

The statistic obtained from the second-stage artificial regression is distributed Chi-squared with k_2 degrees of freedom. Therefore, if the Breush-Pagan Lagrange Multiplier test statistic is greater than the Chi-Squared critical value under k_2 degrees of freedom, let's reject the null and conclude heteroscedasticity is present.

2. 6. Regression with robust standard errors

Even when the homogeneity of variance assumption is violated the ordinary least squares (OLS) method calculates unbiased, consistent estimates of the population regression coefficients. In this case, these estimates won't be the best linear estimates since the variances of these estimates won't necessarily be the smallest. Worse yet the standard errors will be biased and inconsistent. It is possible to perform our regression analysis to correct the issue of incorrect standard errors so that our interval estimates, and hypothesis tests are valid. It is done by using heteroscedasticity-consistent standard errors or simply robust standard errors. The concept of robust standard errors was suggested by some dude named Halbert White.

The usual method for estimating coefficient standard errors of a linear model can be expressed with this somewhat intimidating formula proposed by [15]:

$$Var(\hat{\beta}) = (X^T X)^{-1} X^T \Omega X (X^T X)^{-1}. \quad (6)$$

Where x is the model matrix (i. e., the matrix of the predictor values) and $\Omega = \sigma^2 I_n$ which is shorthand for a matrix with nothing but σ^2 on the diagonal and 0's everywhere else.

2. 7. Productivity data

Productivity is commonly defined as a ratio between the output volume and the volume of inputs. Productivity is considered a key source of economic growth and competitiveness, as such, is basic statistical information for many international comparisons and country performance assessments. Productivity data are used to investigate the impact of product and labor market regulations on economic performance. The Source of this data set was provided by [14]. As the study of [15] used this dataset, including a panel of 48 observations from 1970 to 1986 and the data frame containing: State: the state, year: the year, pcap: public capital stock, hwy: highway and streets, water: water and sewer facilities, util: other public buildings and structures, pc: private capital stock, GSP: gross state product, emp: labor input measured by the employment in non-agricultural payrolls, and unkempt: state unemployment rate.

In our simulation study, it is possible to detect the multi-collinearity problem between the explanatory variables throughout the correlation matrix.

The following Cobb-Douglas production function relationship investigating the productivity of public capital in private production:

$$\ln GSP = \alpha + \beta_1 \ln PC + \beta_2 \ln P - cap + \beta_3 \ln L + \beta_4 Unemp + U. \quad (7)$$

3. Results and Discussion

3. 1. The correlation matrix between explanatory variables as follows

The following correlation matrix, matrix (8) contains the results of the correlation to detect the multi-collinearity problem between the explanatory variables as follows:

$$\text{Correlation matrix} = \begin{pmatrix} & x_1 & x_2 & x_3 & x_4 \\ x_1 & 1.0000000 & 0.8647919 & 0.9076301 & 0.1766757 \\ x_2 & 0.8647919 & 1.0000000 & 0.9722074 & 0.1843358 \\ x_3 & 0.9076301 & 0.9722074 & 1.0000000 & 0.1572794 \\ x_4 & 0.1766757 & 0.1843358 & 0.1572794 & 1.0000000 \end{pmatrix}. \quad (8)$$

Where, x_1 is lnPC, x_2 is ln$P$–cap, x_3 is ln$L$, and x_4 is Unemp.

From the above correlation matrix, it's clear that there is a high correlation between the independent variables.

To get rid of this problem, use the principal component technique and get four components as follows:

$$\begin{matrix} PC1 & PC2 & PC3 & PC4 \\ 9.211124e^{-01} & 3.892102e^{-01} & 0.008213451 & -9.644577e^{-06} \\ 3.883690e^{-01} & -9.201683e^{-01} & 0.049596218 & -5.276085e^{-05} \\ 2.686113e^{-02} & -4.249386e^{-02} & -0.998735206 & 8.604370e^{-04} \\ 6.262116e^{-06} & -8.231808e^{-06} & 0.000862045 & 9.999996e^{-01} \end{matrix}.$$

3. 2. The Relationship Between the Factors Must be Equal to or Approximate to Zero

The following matrix, contain the result of correlation between the first three components; $PC1$, $PC2$, and $PC3$ to ensure that multicollinearity is eliminated and that there is no longer any correlation between the factors with each other's:

$$\begin{matrix} & PC1 & PC2 & PC3 \\ PC1 & 1.000000e^{+00} & 7.195603e^{-08} & -1.574500e^{-0} \\ PC2 & 7.195603e^{-08} & 1.000000e^{+00} & 3.011797e^{-09} \\ PC3 & 1.574500e^{-08} & 3.011797e^{-09} & 1.000000e^{+00} \end{matrix} \quad (9)$$

and the importance of components.

The following matrix, contain the result of the standard deviations for $PC1$, $PC2$, and $PC3$, in the second row. The third row in the matrix showed the Proportion of variations, and the last row showed the Cumulative Proportions of the three components:

$$\begin{pmatrix} PC1 & PC2 & PC3 \\ 1.6965 & 0.9762 & 0.38347 \\ 0.7195 & 0.2382 & 0.03676 \\ 0.7195 & 0.9578 & 0.99452 \end{pmatrix}.$$

Then, calculate the first three factors from this component by multiplying this component with the original variables to get the factors.

The panel regression model became after solving the multi-collinearity problem as follows:

$$(GSP) \sim (PC1) + (PC2) + (PC3). \quad (10)$$

The algorithm of PCA, which consists of two parts, is included into the process of estimating PCs. On a balanced panel of PCs computed based on the quarterly series observed across the whole sample period, those variables with missing data are first projected (regressed) linearly.

Before a fresh set of PCs are calculated on the basis of the complete and projected series, this projection is employed in the second phase to fill in the missing observations. The process is iterated until it converges, or until the following PC estimations are sufficiently near to one another between iterations. In our instance, this happened four or five iterations into the process according to [18].

To determine which of the estimation effect method is appropriate, find the fixed effect estimation, random effect estimation, and pooling effect estimation results, **Table 1**. Then carry out the Hausman test, and F-test to know which is the appropriate effect for the model.

Table 1

Results of fixed, random, and pooling effect estimation methods

Coefficients	Fixed Effect Estimation	Random Effect Estimation	Pooling Effect Estimation
Intercept	–	–4.5267e+03	–5.9659e+03
Pca1	0.982363	1.0107e+00	1.0386e+00
Pca2	–1.216212	–1.2854e+00	–1.2949e+00
Pca3	–36.297449	-3.5640e+01	–2.2388e+01
R-squared	0.95841	0.96604	0.99118
Model	«Within»	«Random»	«Pooling»
Sum of squares	$1.3289e^{+11}$	$1.8384e^{+11}$	$3.9905e^{+12}$

3. 3. Tests to determine an appropriate model

Chow Test is a test to determine the most appropriately used in estimating panel data.

By comparing the F-statistic and F-Table calculations, the basic refutation of the claim may be made. If the F count is higher (>) than the F Table and $H0$ is rejected, comparison is used, and the Fixed Effects Model is the most suiTable model to apply.

Regarding the Hausman test: After performing the Chow test and determining that the appropriate model is Fixed Effect, let's look at whether model – Fixed Effect or Random Effect – is the most appropriate.

Hausman test statistic has a degree of freedom equal to k, where k is the number of independent variables, and follows the Chi Square statistic distribution. $H0$ is rejected and a model of Fixed Effect is the proper model if the Hausman statistic value is more than the crucial value. In contrast, if the Hausman statistic value is lower than the critical value, a model of Random Effect is the proper model proposed by [4].

In the following sessions, Chow test and Hausman test results will be shown.

3. 3. 1. Chow test result

The hypotheses of the Chow Test are:

$H0$: Common Effect Model or pooled OLS.

$H1$: Fixed Effect Model.

F test for individual effects:

$$F = 87.428,\ df1 = 47,\ df2 = 765,\ p\text{-value} < 2.2e{-}16.$$

Alternative hypothesis: significant effects.

Chow-Test results showed the fixed effect is most appropriate.

3. 3. 2. Hausman test result

After completing the Hausman test,it is necessary to draw the following conclusion:

1. If the Hausman Test accepts $H0$ or a p-value greater than 0.05, the random effect technique is used. Lagrange multiplier testing is then used to assess whether Random effect or Common effect is still the preferred option.

2. If the Hausman Test returns an $H1$ value or a p-value of less than 0.05, the fixed effect technique is used.

The following hypotheses are tested by the Hausman tests:
*H*0: Random Effect Model.
*H*1: Fixed Effect Model.
Hausman Test will follow the distribution of Chi-squares as follow:

$$chisq = 7.7941,\ df = 3,\ p\text{-value} = 0.05046,$$

alternative hypothesis: one model is inconsistent.

The results of the Hausman test showed that; the *p*-value approximately ≃0.05, so the decision here does not be clear.

Based on the results of Chow and Hausman tests the decision is fixed effect model is the most appropriate model.

The results of the above **Table 2** are achieved by using the (lm) function in R and the output is called using the (summary) function on the model.

Table 2
Results of ordinary least square estimation

–	**Estimate**	**Std. Error**	***t* value**	***Pr*(>|*t*|)**
(Intercept)	–5.9659e+03	3.234e+02	–18.45	<2.2e–16***
pca1	1.0386e+00	3.567e–03	291.14	<2.2e–16***
pca2	–1.2949e+00	1.787e–02	–72.48	<2.2e–16***
pca3	–2.2388e+01	6.463e–01	–34.64	<2.2e–16***

The output talks about the residuals. Residuals are essentially the difference between the actual observed response values, and the response values that the model predicted. The Multiple R-squared.

Represents the percentage variation in the dependent variable (GSP) that is explained by the independent components (predictors). In our case, the R-squared value of 0.9912 means that 99 % percent of the variation in the variable 'GSP' is explained by the 'predictors'.

A small p-value indicates that it is unlikely observe a relationship between the predictor (GSP) and response variables due to chance. Typically, a *p*-value of 5 % or less is a good cut-off point. In our model example, the *p*-values are very close to zero. Three stars (or asterisks) represent a highly significant *p*-value.

A more formal, mathematical way of detecting heteroscedasticity is what is known as the Breusch-Pagan test: Let's test for heteroscedasticity that it is possible to find in the {lmtest} package [removed:].

3. 4. Breusch-Pagan test result

The existence of heteroscedasticity is a major concern in regression analysis and the analysis of variance, as it invalidates statistical tests of significance that assume that the modelling errors all have the same variance. Using the Breusch-Pagan test to check for heteroscedasticity, it examines the relationship between the estimated variance of the residuals from a regression and the values of the independent variables.

Results of Breusch-Pagan test:

$$BP = 105.44,\ df = 4,\ p\text{-value} < 2.2\text{e}{-16}.$$

A small chi-square value (along with an associated small p-value) indicates the null hypothesis is true (i. e. that the variances are all equal).If the p-value is less than the level of significance (*p*-value is less than $\alpha = 0.05$), then let's reject the null hypothesis. Since 2.2e–16 < 0.05. Thus, heteroscedasticity is present. To get the correct standard errors, it is possible to use the (vcovHC) function from the {sandwich} package.

There seems to be no evident pattern in most of the **Fig. 1**. However, it does seem to look as if there is more variation in (residuals) with higher levels of GSP-hat.

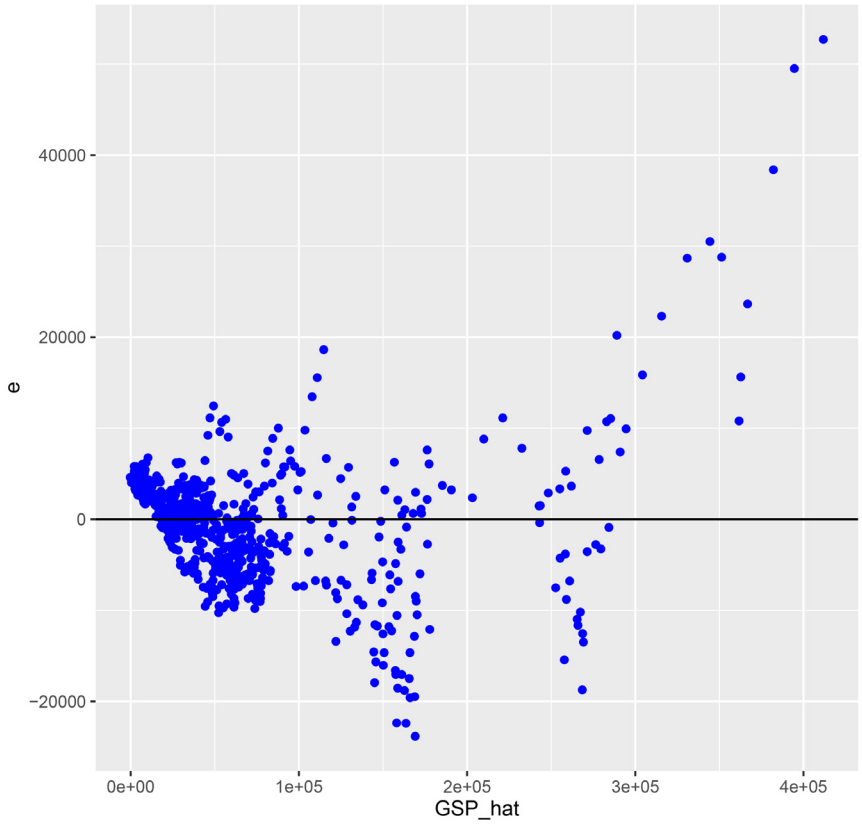

Fig. 1. The relationship between the residuals and GSP-hat

The model above results suffers from heteroscedasticity based on the results of the plot, and Breusch-Pagan test, it is possible to obtain heteroscedasticity robust standard errors and their corresponding *t* values.

Notice that the standard errors are smaller than before, the intercept and (pca3) variables are not statistically significant anymore. But it is possible to depend on the values of (Std. Error) of (pca1) and (pca2) in the original model and **Table 3**.

The model above results in **Table 2** suffers from heteroscedasticity based on the results of the plot, and Breusch-Pagan test, it is possible to obtain heteroscedasticity robust standard errors and their corresponding *t* values.

Table 3
Robust standard error estimation

–	**Estimate**	**Std. Error**	***t* value**	***Pr*(>|*t*|)**
(Intercept)	–5.9659e+03	5.0337e+02	–11.852	<2.2e–16***
pca1	1.0386e+00	9.6285e–03	107.865	<2.2e–16***
pca2	–1.2949e+00	2.7861e–02	–46.477	<2.2e–16***
pca3	–2.2388e+01	1.3785e+00	–16.241	<2.2e–16***

Its notice that from above **Table 3**: the standard errors are smaller than before at **Table 2**, the intercept and (pca3) variables are not statistically significant anymore. But it is possible to depend on the values of (Std. Error) of (pca1) and (pca2) in the original model in **Tables 2**, **3**.

3. 5. Limitations and developed direction of the study

The limitations of this study that cannot control for variables that vary over time (like GSP level or unemployment rate) and also low statistical power, and restricted time periods. The lack of real data sources on Egypt, and that is why there is missing values, and therefore it may be an obstacle to the lack of complete data availability.

The panel studies are now widely used in research across the social and life sciences. The challenge of panel data is the validity and reliability of measurement over time. Many social science researchers are of the opinion that repeating the same question time and time again impacts upon the validity of the measures.

An advantage of our research is when to compare our results of the estimator of standard errors with results of [14], let's get new estimators more efficient and smallest than [14], as to get the new unrelated components first before detecting the problem of heteroscedasticity by using Breush-Pagan test. Also, there is a limitation in that most of the reviewed studies assessed the heteroscedasticity without detecting the multicolinearity problem on the basis that it does not affect the goodness of fit of estimation. But when the problem was detected and resolved, it was found to affect the efficiency of the estimators.

4. Conclusions

It is possible to conclude that throughout our paper; the problem of multi-collinearity has been solved by using a principal component technique in the case of the panel regression model, by using real data set was mentioned by [15].

Using plots of residuals, and the Breusch-Pagan test to detect the problem of heteroscedasticity, then run the robust standard error estimation to get the smallest standard errors for the model and get efficient estimators.

Conflict of interest

The authors declare that they have no conflict of interest in relation to this research, whether financial, personal, authorship or otherwise, that could affect the research and its results presented in this paper.

Financing

The study was performed without financial support.

Data availability

Data will be made available on reasonable request.

Acknowledgments

We express our gratitude to Faculty of Graduate Studies for Statistical Research – Cairo University.

References

[1] Lazarsfeld, P. F. (1940). «Panel» Studies. Public Opinion Quarterly, 4 (1), 122. doi: https://doi.org/10.1086/265373

[2] Andreß, H.-J. (2017). The need for and use of panel data. IZA World of Labor. doi: https://doi.org/10.15185/izawol.352

[3] Baltagi, B. H. (2005). Econometric analysis of panel data. John Wiley & Sons Inc.

[4] Zulfikar, R. (2018). Estimation Model and Selection Method of Panel Data Regression: An Overview of Common Effect, Fixed Effect, and Random Effect Model. INA-Rxiv. doi: https://doi.org/10.31227/osf.io/9qe2b

[5] Born, B., Breitung, J. (2014). Testing for Serial Correlation in Fixed-Effects Panel Data Models. Econometric Reviews, 35 (7), 1290–1316. doi: https://doi.org/10.1080/07474938.2014.976524

[6] Greene, W. (2012). Econometric analysis. Prentice Hall.

[7] Ramón Gil-García, J., Puron-Cid, G. (2014). Using panel data techniques for social science research: an illustrative case and some guidelines. CIENCIA Ergo Sum, 21-3, 203–216. Available at: https://www.redalyc.org/pdf/104/10432355004.pdf

[8] Adeboye, N. O., Fagoyinbo, I. S., Olatayo, T. O. (2014). Estimation of the Effect of Multicollinearity on the Standard Error for Regression Coefficients. IOSR Journal of Mathematics, 10 (4), 16–20. doi: https://doi.org/10.9790/5728-10411620

[9] Gujarati, D., Porter, C. (2008). Basic Econometrics. McGraw-Hill. Available at: https://cbpbu.ac.in/userfiles/file/2020/STUDY_MAT/ECO/1.pdf

[10] Costa, J. C. G. D., Da-Silva, P. J. G., Almeida, R. M. V. R., Infantosi, A. F. C. (2014). Validation in Principal Components Analysis Applied to EEG Data. Computational and Mathematical Methods in Medicine, 2014, 1–10. doi: https://doi.org/10.1155/2014/413801

[11] Katchova, A. (2013). Panel data models. Hentet, 4 (13).

[12] Hansen, C. B. (2007). Asymptotic properties of a robust variance matrix estimator for panel data when is large. Journal of Econometrics, 141 (2), 597–620. doi: https://doi.org/10.1016/j.jeconom.2006.10.009

[13] Arellano, M. (2009). PRACTITIONERS' CORNER: Computing Robust Standard Errors for Within-groups Estimators. Oxford Bulletin of Economics and Statistics, 49 (4), 431–434. doi: https://doi.org/10.1111/j.1468-0084.1987.mp49004006.x

[14] Baltagi, B. H. (2021). Econometric analysis of panel data. Springer Cham, 424. doi: https://doi.org/10.1007/978-3-030-53953-5

[15] Cook, L. M., Munnell, A. (1990). How does public infrastructure affect regional economic performance? New England Economic Review, 11–33. Available at: https://econpapers.repec.org/article/fipfedbne/y_3a1990_3ai_3asep_3ap_3a11-33.htm

[16] Baltagi, B. (2008). Econometric analysis of panel data. John Wiley & Sons Ltd.

[17] Breusch, T. S., Pagan, A. R. (1979). A Simple Test for Heteroscedasticity and Random Coefficient Variation. Econometrica, 47 (5), 1287. doi: https://doi.org/10.2307/1911963

[18] Stock, J. H., Watson, M. W. (2002). Forecasting Using Principal Components From a Large Number of Predictors. Journal of the American Statistical Association, 97 (460), 1167–1179. doi: https://doi.org/10.1198/016214502388618960

Received date 21.09.2022
Accepted date 19.12.2022
Published date 19.01.2023

How to cite: *Youssef, A. H., Abozaid, E. S., Abdel Latif, S. H. (2023). Handling multi-collinearity using principal component analysis with the panel data model. EUREKA: Physics and Engineering, 1, 177–188. doi: https://doi.org/10.21303/2461-4262.2023.002582*