International Journal on Advanced Science Engineering Information Technology

Vol.10 (2020) No. 6
ISSN: 2088-5334

Measuring the Impact of Environmental Sustainability on Tuberculosis Rates Using the Two-Stage Least Squares Method in the Polled Model

Suhad Ali Shaheed Al-Temimi[a,1], Rawaa Salh Al-Saffar[a]

[a] *Statistics Department, Administration and Economics College, Mustanisiryah University, Iraq*
Email: [1]dr.suhadali@uomustansiriyah.edu.iq

***Abstract*— Iraq has witnessed several changes that directly and negatively affected the nature of society and the environment, such as an increase in the population over the past two decades. Hence, an increase in demand for food, energy, housing, and water means an increase in solid and liquid waste as a result of weak environmental awareness. Tuberculosis is one of the infectious diseases, and Iraq is witnessing a noticeable increase in the rate of infections at the governorates level. The explanatory variables that were chosen are among the variables of the sustainable environment adopted by the Ministry of health in Iraq. Therefore, it was essential to know their effect on the phenomenon under study (number of tuberculosis cases). The results of estimating the model parameters using the two-stage least squares method and the transformations method show that the explanatory variables significantly affect the dependent variable, as explained above. This study focused on the effect of some sustainable environment variables (the population, the number of health institutions, the proportion of the population that uses clean drinking water, and the percentage of the population with access to sanitation) on tuberculosis rates based on the polled model at the governorates level for the period (2013–2017). The two-stage least squares method was used to estimate model parameters. The results showed that increasing environmental awareness represented by sustainable environment variables positively impacts lower rates of tuberculosis at the governorates level.**

***Keywords*—environmental sustainability; tuberculosis incidence rate; the two-stage least squares method; polled model.**

I. INTRODUCTION

The drinking water purification plants are considered one of the essential elements for achieving environmental sustainability. In Iraq, these plants are characterized by inefficiency and lack of capacity as they receive more water than their design capacity. Besides, there are not enough sewerage networks to treat rainwater, liquid waste from hospitals, factories, and production waste. The rise in the population and urban expansion in Iraq has exacerbated these problems over the past two decades. One of the most critical repercussions of this is embodied in the spread of diseases and epidemics at the level of all provinces in Iraq. The most important of which is tuberculosis [1]. A polled model was studied for cross-section data, time series, and parameters estimation using the two-stage least squares method. It aims to demonstrate the positive effect of environmental sustainability on reducing tuberculosis rates.

The parameters' efficiency characterizes this estimation method after the model has been converted to the reduced formula to get rid of the accidental parameters and the model is fully diagnosed. Variance components models have been used to integrate cross-sectional data and time series [2]. These models are considered very useful because they provide more information about the regression parameters through the differences between the cross-section and the period using the variance method for data analysis. Another study concerns integrating time-series and cross-section data (TSCS) and the properties of the fixed-effect and the general least squares method with an indication of the properties of the estimators [3].

Several time series and cross-section data integration models, estimation, and testing mechanisms for parameters have been studied using the general least squares method and balanced and unbalanced integration data with an application example provided [4]. Ujjan et al. presented the different spatial data features and their analysis and the characteristics of a strong relationship to social sciences applications [5]. Ujjan et al. also studied the essential concepts and issues related to spatial arrangement and measurement, the role of a spatial weight's matrix and different spatial correlation measures and classification of spatial process models, a spatial dependence as a source of disturbance, and spatial regression models. The effect of exchange rate systems on inflation and economic growth in Italy for the period (1961-1998) has been examined [6]. The researcher has used the two-stage least squares method (2 SLS) to test the internal effect of exchange rate systems.

Garofalo concluded that the relationship between exchange rate systems and economic performance might have adverse effects, thereby weakening the results. A two-stage general spatial least squares method (GS2SLS) has been proposed to estimate the spatial autoregression model in the presence of autoregression of the disturbances (random error). It is subject to arrive at a scientific method to infer the spatial regression type that contains spatial regression in the dependent variable, exogenous variables, and disturbance limits [7]. The researchers generalized the estimator of the generalized momentums method of the spatial autoregression parameter in the term of random error, and proved the consistency of their estimation, and determined the approximate distribution through the normal distribution.

Some non-parametric methods have been used in smoothing the function of a time-variable coefficients model for a non-parametric limit model of balanced longitudinal data [8]. It could be conducted through the local linear polynomial kernel (LLPK) technique and cubic smoothing splines (CSS) technique, as well as finding the coefficient estimates using the two-stage least squares method (2SLS) and by using (LLPK) and (CSS) techniques. The identification problem of spatial Durbin panel models has also been studied [9]. The study considered the identification of model parameters when estimating with two methods (QML) and (2SLS). The study arrived at estimating the Durbin spatial model through simulation.

There is an increase in tuberculosis cases at the governorate level in Iraq due to health and environmental impacts. In view of the novelty of the concept and standards for sustainable environmental awareness in Iraq, it is necessary to conduct research and studies that show how these standards play in reducing epidemic infections. The polled models of cross-section data and time series are essential models that explain the relationship between study variables. The research aims to study the impact of environmental sustainability standards in reducing the incidence of epidemic diseases transmitted through bacteria and viruses. The spread of diseases and epidemics, including tuberculosis, is mostly due to insufficient water purification plants and waste treatment plants due to the lack of urban planning that corresponds to the large population increase.

II. Materials and Methods

A. Polled Model

The polled model is one of the simplest longitudinal data models where the estimated regression parameters are constant for all time periods, as the changes are included in the random error term, and the polled model can be formulated as follows [9]

$$Y_{nt} = \mu_{nt} + \lambda Y_{nt} + X_{nt}\beta_n + \epsilon_{nt}t - 1 \qquad (1)$$

Where:

Y_nt: A vertical vector whose dimensions (N×1) represents the dependent variable at the cross-section (n=1,2,…,N) and at time (t=1,2,…,T).

λ: A dimensional vector (nt×1) Autoregressive parameter of .the dependent variable

X_nt: A matrix of dimensions (nt×k) of the explanatory variables of the n section and time period t

β: A dimensional vector (k×1) representing the response .parameters for explanatory variables

μ: Fixed effects of groups. It is a dimensional vector (n×1) and represents the constant term parameter

ϵ_{nt} : dimensional vector $(\mathrm{N} \times 1)$ which reflects the specifications of the error term of the model.

B. Two-Stage Least Squares Method (2 SLS).

This method is used to estimate the parameters of a model that is in a state of complete diagnosis. The estimation method for any model is in two stages [10].

1) First stage: at this stage, the endogenous variable is determined. The reduced formula and the ordinary least square method (OLS) are used after the necessary conditions are met to estimate the reduced formula, which is a diagnosis of model parameters and finding the model variables' estimated values.

2) Second stage: at this stage, the (OLS) method is used again in estimating the model parameters after substituting the values $(\hat{\mathrm{Y}}_{\mathrm{nt}})$ estimated in the first stage in place of the real values $(\mathrm{Y}_{\mathrm{nt}})$.

These two stages represent the approach used for the (2SLS) estimation method. To apply this method to estimate the pooled model in the stable state, as in equation (1). The incidental parameters are eliminated by using the transformation procedure by multiplying the model in equation (1) by the orthonormal transform standard represented by $(F_{\mathrm{T,T-1}})$.

This method is based on a conversion procedure to exclude spatial effects (and for a limited time) by reducing the number of observations to one observation per unit of the sample, that is $[(\mathrm{T}-1)\times(\mathrm{N}-1)]$ In the case of cross-section effects and time series in the model to get a fully diagnosed pooled model. The presence of the effect of cross-sections on the pooled model of longitudinal data leads to a bias of the model parameters and is also called the incidental parameters and is represented by (μ) at each cross-section as a result if the nature of the data, as the variable $(\mathrm{Y}_{\mathrm{nt}})$ is $(\mathrm{n}=1,2,\dots,\mathrm{N})$ of the observations at cross sections and $(\mathrm{t}=1,2,\dots,\mathrm{T})$ of the time series, and so the incidental parameters are established [11].

C. Transformation Procedure:

This Procedure is based on the orthonormal Eigenvector matrix for $(\mathrm{J_T})$ matrix which is $\left[F_{T,T-1}, \frac{1}{T}\,\iota_T\right]$, representing the Eigenvectors:

$$\mathrm{J_T} = I_T - \frac{1}{T}\,\iota_T\,\iota_T'$$

Where: I_T : Identity matrix and its dimensions $(\mathrm{T}\times\mathrm{T})$

ι_T : vertical vector of dimensions $(\times 1)$ consisting of unity .

$F_{T,T-1}$: A matrix with dimensions $(\mathrm{T}\times(\mathrm{T}-1))$

Consisting of the Eigenvectors of matrix $\mathrm{J_T}$ .The Orthonormal represented by the amount ($F_{T,T-1}$) has several conditions that must be taken into consideration and on which it relies to conduct the appropriate orthonormal test, and the conditions are as follows [12], [13]:

$$\left.\begin{array}{l} F'_{T,T-1}F_{T,T-1} = I_{T-1} \\ \iota'_T\, F_{T,T-1} = 0 \\ \left[F_{T,T-1}, \frac{1}{\sqrt{T}}\,\iota_T\right]' \left[F_{T,T-1}, \frac{1}{\sqrt{T}}\,\iota_T\right] = I_T \\ F_{T,T-1}F'_{T,T-1} = J_T \\ F'_{T,T-1}\,\iota_T = 0 \\ J_T\, F_{T,T-1} = F_{T,T-1} \\ F_{T,T-1}F'_{T,T-1} + \frac{1}{T}\,\iota_T\,\iota'_T = I_T \end{array}\right\} \quad (2)$$

As the explanatory variables and the dependent variable in equation (1) are multiplied by the amount $(F_{T,T-1})$ and thus the matrix of the dependent variable (Y_{nT}) becomes with dimensions $(\mathrm{n} \times (\mathrm{T}-1))$ after it was $(\mathrm{n} \times \mathrm{T})$, as is the case with (X_{nT}) , and as follows :

$$\left.\begin{array}{c} \left[Y^*_{n1}, Y^*_{n2}, \dots, Y^*_{n,T-1}\right] = \left[Y^1_{n1}, Y^1_{n2}, \dots, Y^1_{nT}\right] \\ 2 \\ \left[X^*_{n1,k}, X^*_{n2,k}, \dots, X^*_{n,T-1,k}\right] = \left[X^n_{n1,k}, X^n_{n2,k}, \dots, X^n_{nT,k}\right] \end{array}\right\} \quad (3)$$

Where:

$X^n_{nT,k}$ represents the column ($\mathrm{k^{th}}$) of the matrix X^1_{nT} with dimension $(\mathrm{n} \times \mathrm{k_x})$.

As for the random error term (ϵ_{nt}), one of the essential characteristics of the orthonormal matrix $(\mathrm{J_T})$ is that it is a singular matrix, so when multiplied by the random error term (ϵ_{nt}), it becomes (σ^2_{0,I_T}). Accordingly, the elements of $(\epsilon_{nt}\mathrm{J_T})$) are linearly dependent, i.e., linked with each other, and (J_t) is an orthogonal matrix, and the trace of the matrix is equal to $(\mathrm{T}-1)$ [12].

$$\left[\epsilon^{*\prime}_{n1}, \epsilon^{*\prime}_{n2}, \dots, \epsilon^{*\prime}_{n,T-1}\right]' = \left(F'_{T,T-1} \otimes I_n\right)[\epsilon'_{n1}, \epsilon'_{n2}, \dots, \epsilon'_{nT}]' \quad (4)$$

Since (ε^*_{it}) that represents the element $\left(\mathrm{i^{th}}\right)$ of the random error vector $(\epsilon^*_{nt}\ )$ is (i,i,d) , then

$$E \quad = \sigma_0^2 I_{n(T-1)} \quad (5)$$

Thus (ε^*_{it}) is uncorrelated to all (i,t) values and is independent under the normal distribution assumption. As for the fixed spatial effects (μ) , one of the properties of $(F_{T,T-1}\ )$ if multiplied by a vector of constants, the results is equal to zero as mentioned in equation (2) – the second condition, and as a result:

$$[\mu_n, \dots\dots, \mu_n]' F_{T,T-1} \ = 0 \quad (6)$$

After applying the transformation procedure, the model in equation (1) becomes as follows:

$$Y^*_{nt} = \lambda Y^*_{nt} + X^*_{nt}\beta_n + \epsilon^*_{nt}\,, \qquad t = 1, \dots, T-1 \quad (7)$$

Where:
$\boldsymbol{X}^*_{nt} = (X^{*\prime}_{n1}, \dots\dots, X^{*\prime}_{nT})$, and the instrumental variables that represent the desired set of explanatory variables in addition to the dependent variable is $\mathbf{Z}^*_{n,T-1} = \left(\mathbf{Y}^*_{n,T-1}, \mathbf{X}^*_{nt}\right)^n$

The diagnosis of parameters can be accessed by the rank condition so that the amount $[\mathrm{x^*_{nT}}]$ is of a full rank. Lee and Yu [14] have demonstrated that, in general, if the condition is met, the parameters (β_0) in (X^*_{nt}) are linearly independent of the random part represented by (ϵ^*_{nt})) , that is :

$$trZ^*_{nt}\left(\boldsymbol{\epsilon}^{*\prime}_{n,T-1}(\theta)\boldsymbol{\epsilon}^{*n}_{n,T-1}(\theta)\right) = 0 \quad (8)$$

The diagnosis by the rank condition has been proven by equation (8), and therefore the two-stage least squares (2SLS) method can be used to estimate model parameters and thus $(\bar{\theta}_{2sl,nT}\,)$:

$$\hat{\theta}_{2sl,nT} = \left[\left(\mathbf{X}^*_{n,T-1}\right)'\left(\mathbf{X}^*_{n,T-1}\right)^n\right]^{-1} \times \left(\mathbf{X}^*_{n,T-1}\right)' \quad (9)$$

D. Consistency and Asymptotically Distribution for 2SLS method

The proof of asymptotical properties of the estimator $\left(\hat{\theta}_{2sl,nT}\right)$ is performed after proving the diagnostic conditions in equation (8), and therefore the estimator according to the method $(2\ \mathrm{SLS})$ is a consistent estimator and has asymptotical characteristics of the natural distribution based on the central limit theory and the presence of the variance matrix $(\sigma_0{}^2)$, and thus

$$\sqrt{n(T-1)}\left(\hat{\theta}_{2sl,nT} - \theta_0\right)\left(0, P\lim_{n\to\infty}\Sigma^{-1}_{nT,2sl}\right) \quad (10)$$

Whereas $\left(\Sigma^{-1}_{nT,2sl}\right)$ represents the matrix of variance and covariance that we get from:

$$\Sigma^n_{nT,2sl} = \frac{1}{\mathrm{n(T-1)}}\left(\boldsymbol{X}^*_{n,T-1}\right)'\mathbf{H}_{Q,nT}\left(,\boldsymbol{X}^*_{n,T-1}\right) \quad (11)$$

(H_{nT}) represents the covariance matrix of the explanatory in equation (12) and thus (H_{nT}) will be according to the following formula:

$$\mathrm{H}_{nT} = \frac{1}{nT}\sum_{t=1}^{T}\left(\mathbf{X}^*_{n,T-1}\right)'\left(\mathbf{X}^*_{n,T-1}\right) \quad (12)$$

III. Results and Discussion

According to what was presented in the theoretical part, data issued by the Ministry of planning for the period (2013-2017) represented a five–year time series (t=5) and at the level of fifteen governorates representing cross-sections (n=15) except for the Kurdistan region. The data related to the research included the following:

- Number of cases of tuberculosis (dependent variable $\mathrm{Y_{nt}}$)
- The proportion of the population using drinking water (explanatory variable X^n_{1nt})
- The proportion of the population with access to a sewage system (explanatory variable X^n_{2nt})
- Population (explanatory variable X^n_{3nt})
- Number of health institutions (explanatory variable X^n_{4nt})

To determine whether the effect of cross-sections of the longitudinal data is fixed or random, the F test, the χ^2 test and $\mathrm{E_{views}}$–9 programs were used to test the null hypothesis, which states that the cross-section effects are fixed, and the results are as follows:

TABLE I
Detection Of Cross-Sections Effect If Fixed Or Random

	Redundant Fixed Effects Tests		
		Equation: Untitled	
	Test Null Hypothesis: cross-section fixed effects		
Prob.	d.f.	Statistic	Effects Test
0.1664	(4,70)	1.308761	Cross-section F
0.9663	4	6.6698	Cross-section Chi-square

Through the result of the above table, the probability value for both (F and χ^2) tests is greater than (0.05). Therefore, the null hypothesis is accepted, which means that the cross – section effects are fixed. The above data has been entered in the pooled model of autoregression in equation (1) and a matrix (J_T) is a calculated based on the method of transformations. The matrix $(F_{T,T-1})$ which later represents the Eigen vectors matrix of the matrix (J_T) is multiplied by the model variables (1) to eliminate unwanted properties of the parameters, and then the model is estimated according to the method (2SLS). The results are obtained through Eviews–9 as shown below. The orthonormal Eigen vector matrix represented by the term $(F_{T,T-1}, \frac{1}{T}\,\iota_T)$ is as follows:

$$\left[F_{T,T-1}, \frac{1}{T}\,\iota_T\right] = \begin{bmatrix} 0.00000 & 0.894427 & 0.00000 & 0.00000 & 0.20000 \\ -0.50000 & -0.22360 & -0.50000 & -0.50000 & 0.20000 \\ -0.16667 & -0.22360 & 0.83333 & -0.16667 & 0.20000 \\ 0.83333 & -0.223607 & -0.16667 & -0.16667 & 0.20000 \\ -0.16667 & -0.223607 & -0.16667 & 0.83333 & 0.20000 \end{bmatrix}$$

As the matrix (F_T) is a matrix consisting of the Eigen vectors of the matrix (J_T) and it is:

$$[F_T] = \begin{bmatrix} 0.00000 & 0.894427 & 0.00000 & 0.00000 & 0.447214 \\ -0.50000 & -0.22360 & -0.50000 & -0.50000 & 0.447214 \\ -0.16667 & -0.22360 & 0.83333 & -0.16667 & 0.447214 \\ 0.83333 & -0.223607 & -0.16667 & -0.16667 & 0.447214 \\ -0.16667 & -0.223607 & -0.16667 & 0.83333 & 0.447214 \end{bmatrix}$$

While the matrix (J_T) is as follows:

$$J_T = \begin{bmatrix} 0.8 & -0.2 & -0.2 & -0.2 & -0.2 \\ -0.2 & 0.8 & -0.2 & -0.2 & -0.2 \\ -0.2 & -0.2 & 0.8 & -0.2 & -0.2 \\ -0.2 & -0.2 & -0.2 & 0.8 & -0.2 \\ -0.2 & -0.2 & -0.2 & -0.2 & 0.8 \end{bmatrix}$$

After applying the transformations method to the model variables represented by the dependent variable (Y_{nt}). The explanatory variables are presented as in equation (3). The identification by the Rank condition has been proved through equation (8). The (2SLS) method can then be used to estimate the model parameters as in equation (10). Thus, the vector of estimators in $\left(\hat{\theta}_{2sl,nT}\right)$ is as follows:

TABLE II
ESTIMATING THE PARAMETERS OF THE CLUSTER MODEL USING THE TWO-STAGE LEAST SQUARES METHOD

Parameters	Estimator	
	Estimate	T-test
λ	0.561279	1.02258
β_1	-0.050325	2.03032
β_2	-0.206504	1.7417
β_3	18.72192	1.83556
β_4	-1.281356	2.04889

According to Table (2), the value of the autoregression parameter (λ) is positive and equal (0.561279). Table 2 indicates the increase in the number of tuberculosis cases at the level of fifteen governorates except for the Kurdistan region when other explanatory variables exist. As for explanatory variable X_1 (the percentage of the population that uses potable drinking water), it is related negatively with the variable of the number of cases of tuberculosis, where the increase of one unit in the variable X_1 when other explanatory variables exist, the number of cases decreases by (0.05032). Also, for explanatory variable X_2 (the proportion of the population with a case to a sewage system, an increase of one unit in the variable X_2 when other explanatory variables exist, the number of cases decreases by (0.206504). An increase of one unit in X_3 (explanatory variable of the population) increases cases by (18.7219) when other explanatory variables exist. Finally, an increase of one unit in X_4 (explanatory variable of a few health institutions) leads to a decrease in the case by (1.281356) when other explanatory variables exist.

IV. CONCLUSIONS

The use of the transformation method in estimating the model is considered very important since it possesses an effective possibility in correcting the endogeneity within the model, which leads to obtaining reasonable estimates with approximate characteristics of the estimated model parameters. The explanatory variables that were chosen are among the variables of the sustainable environment adopted by the Ministry of health in Iraq. Therefore, it was essential to know their effect on the phenomenon under study (number of tuberculosis cases). The results of estimating the model parameters using the two-stage least squares method and the transformations method show that the explanatory variables significantly affect the dependent variable, as explained above.

REFERENCES

[1] R. A. Johnson and D. W. Wichern, Applied multivariate statistical analysis, vol. 5, no. 8. Prentice hall Upper Saddle River, NJ, 2002.

[2] H. K. Sharaf, M. R. Ishak, S. M. Sapuan, and N. Yidris, "Conceptual design of the cross-arm for the application in the transmission towers by using TRIZ–morphological chart–ANP methods," J. Mater. Res. Technol., vol. 9, no. 4, pp. 9182–9188, Jul. 2020, doi: 10.1016/j.jmrt.2020.05.129.

[3] A. Dubrov, "Applied Multivariate Data Analysis," Stat. Moscow, 1992.

[4] M. H. Forouzanfar et al., "Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015," Lancet, vol. 388, no. 10053, pp. 1659–1724, Oct. 2016, doi: 10.1016/S0140-6736(16)31679-8.

[5] R. M. A. Ujjan, Z. Pervez, K. Dahal, A. K. Bashir, R. Mumtaz, and J. González, "Towards sFlow and adaptive polling sampling for deep learning based DDoS detection in SDN," Futur. Gener. Comput. Syst., vol. 111, pp. 763–779, Oct. 2020, doi: 10.1016/j.future.2019.10.015.

[6] C. Prosser and J. Mellon, "The Twilight of the Polls? A Review of Trends in Polling Accuracy and the Causes of Polling Misses," Gov. Oppos., vol. 53, no. 4, pp. 757–790, Oct. 2018, doi: 10.1017/gov.2018.7.

[7] B. P. Scicluna et al., "Classification of patients with sepsis according to blood genomic endotype: a prospective cohort study," Lancet Respir. Med., vol. 5, no. 10, pp. 816–826, Oct. 2017, doi: 10.1016/S2213-2600(17)30294-1.

[8] B. Pölling, W. Sroka, and M. Mergenthaler, "Success of urban farming's city-adjustments and business models—Findings from a survey among farmers in Ruhr Metropolis, Germany," Land use policy, vol. 69, pp. 372–385, Dec. 2017, doi: 10.1016/j.landusepol.2017.09.034.

[9] L. Koufariotis et al., "Sequencing the mosaic genome of Brahman cattle identifies historic and recent introgression including polled," Sci. Rep., vol. 8, no. 1, p. 17761, Dec. 2018, doi: 10.1038/s41598-018-35698-5.

[10] J. M. Kieffer et al., "Evaluation of the psychometric properties of the EORTC chemotherapy-induced peripheral neuropathy questionnaire (QLQ-CIPN20)," Qual. Life Res., vol. 26, no. 11, pp. 2999–3010, Nov. 2017, doi: 10.1007/s11136-017-1626-1.

[11] Pearce, A., Tomalin, B., Kaambwa, B., Horevoorts, N., Duijts, S., Mols, F., ... & Koczwara, B. (2019). Financial toxicity is more than costs of care: the relationship between employment and financial toxicity in long-term cancer survivors. Journal of Cancer Survivorship, 13(1), 10-20.

[12] Sharaf, H. K., Ishak, M. R., Sapuan, S. M., Yidris, N., & Fattahi, A. (2020). Experimental and numerical investigation of the mechanical behavior of full-scale wooden cross arm in the transmission towers in terms of load-deflection test. Journal of Materials Research and Technology, 9(4), 7937-7946.

[13] Thompson, N. M., Widmar, N. O., Schutz, M. M., Cole, J. B., & Wolf, C. A. (2017). Economic considerations of breeding for polled dairy cows versus dehorning in the United States. Journal of dairy science, 100(6), 4941-4952.

[14] L. Lee and J. Yu, "Identification of Spatial Durbin Panel Models," J. Appl. Econom., vol. 31, no. 1, pp. 133–162, Jan. 2016, doi: 10.1002/jae.2450.