Hausman and Taylor Estimator Analysis on The Linear Data Panel Model

Bernadhita H. S. Utami[1], Agus Irawan[2], Miswan Gumanti[3], Gilang Primajati[4]

[1,2,3]System Information, STMIK Pringsewu, Lampung, Indonesia

[4]Universitas Bumigora, NTB, Indonesia

Article Info	ABSTRACT
	Panel data modelling in the field of econometrics applies two main approaches, namely fixed effect estimators and random effects. The application of the Hausman and Taylor estimator to real data is used to test for fixed effects or random effects based on the idea that the set of estimated coefficients obtained from the fixed effect estimates is taken as a group. A good estimator is an estimator that is as close as possible to represent the characteristics of the population. The characteristics of a good estimator include unbiasedness, efficiency, and consistency. The purpose of this study is to identify the properties of the Hausman and Taylor estimator in the linear model of panel data. Based on the analysis using panel data, it is found that the Hausman and Taylor estimator on the random effects panel data is an estimator that is consistent and efficient even though it is not unbiased.

Corresponding Author:

Bernadhita H. S. Utami,
System Information, STMIK Pringsewu, Lampung.
Email: bernadhita.herindri.s@mail.ugm.ac.id

A. INTRODUCTION

Panel data modelling in the field of econometrics applies two main approaches, namely fixed effect estimators and random effects. In (Kojima et al., 2016), conducted research that investigate global demand characteristics of major vegetable oils for future research and policy analysis by using a cross-country panel data of 165 countries from 1991 to 2011 using panel data fixed effect model. In the fixed effect approach, the time factor cannot be observed for each unit of observation but is captured by a dummy variable. Furthermore, this unobservable time factor is treated as part of the error in the random effects model, assuming that its correlation with the regression is zero, (Baltagi and KhantiAkom, 1990). If this assumption is met, the random effect estimator has the advantage of having greater efficiency than the fixed effect estimator, (Frondel and Vance, 2010). The random effects panel data model assumes that all explanatory variables are uncorrelated with random individual effects, while the fixed effects panel data model assumes that all variables are correlated with random individual effect. To investigate whether modelling has any of the characteristics of these two approaches, Hausman in (Baltagi and Liu, 2012) sparked the idea hereinafter referred to as the Hausman test. Tests for fixed effects or random effects are based on the idea that the set of estimated coefficients derived from the fixed effect estimates is taken as a group. The null statistical hypothesis in Hausman's test is that there is no difference between the set of estimated coefficients and the specific set of individuals that cannot be observed. If the test results suggest rejecting H_0, the applied researcher generally proceeds to draw conclusions based on the fixed effect model estimates.

According to (Yusra et al., 2019), testing the equivalence of fixed and random effects using the Hausman and Taylor test. The test uses a direct model specification that can simultaneously estimate fixed effects and random effects, either based on Ordinary

Least Squares (OLS) or Generalized Least Squares (GLS) to test both coefficient equations for individual variables as well as the entire range of coefficients. Hausman test used for random effect model selection with fixed effect model. This test works by testing whether there is a relationship between errors in the model with one or more variables explanation in the model, (Agus Astapa et al., 2018). Hausman and Taylor estimators on panel data allow multiple explanatory variables to be correlated with individual effects. One of the main drawbacks of the fixed effect estimator is that it removes the effects of time-invariant variables. In contrast Hausman and Taylor estimators recapture estimates of time-invariant variables that are important in empirical applications, (Baltagi et al., 2003). (Baltagi and Bresson, 2012) concerning to introduces small sample properties and pretest estimation of a spatial Hausman and Taylor model.

The research of (Sarafidis and Wansbeek, 2010) demonstrate the Monte Carlo method using the Hausman and Taylor estimator on robust panel data suggested by (Arellano, 1993). The size of the Mean Square Error (MSE) depends on the type and level of observation. The results of this study yield an analysis that can be extended to the dynamic Hausman and Taylor model, in which a researcher can examine the sensitivity of using the difference transformation versus the mean or median transformation to eliminate individual effects. The development of the Hausman and Taylor estimator application was also introduced by Harding and Lamarche (2014). They are using the Hausman-Taylor instrument variable as a penalized estimation approach of quantile panel models. In addition, the results of (Maani et al., 2015), by using the Hausman and Taylor estimator, it can be identified that immigrants' labor market integration is significantly affected by the local concentration and resources of their ethnic group.

The application of the Hausman and Taylor estimator to real data has been widely used. In the research of (Baltagi et al., 2014), used the Hausman and Taylor estimator setting to analyze SME data and trade, exports and FDI. The results of the analysis allow concurrent testing to control for the effect of common exogenous determinants and estimate the direct relationship between exports and FDI in terms of bilateral effects that are not observed and do not vary over time. This approach provides consistent parameter estimates and is superior to the corresponding fixed effects model. In addition, the analyze yielded parameter estimates for time-invariant variables such as distance that control for bilateral heterogeneity. (Egger and Pfaffemayr, 2001) generates a derived Hausman and Taylor estimator that allows to test the spatial dependence on homoscedastic or heteroscedastic disorders. The results contributed to examining the large sample nature of the estimator and demonstrating the small sample performance through Monte Carlo simulations. In 2013, (Cng et al., 2013) using the Hausman and Taylor estimator in application of 1995 to 2011 panel data that included 18 of Vietnams major country partners and provided by Vietnams authorities and international organizations.

A good estimator is an estimator that is as close as possible to represent the characteristics of the population, (Chapman and Robbins, 1951). The characteristics of a good estimator include unbiasedness, efficiency, and consistency, (Chan et al., 2008). The purpose of this study is to identify the properties of the Hausman and Taylor estimator in the linear model of panel data. The identification of the characteristics of this estimator is useful for adding broader insights in the field of spatial panel data as introduced by (Millo and Piras, 2012) (Xu and Li, 2020).

B. LITERATURE REVIEW

1. Data Panel Random Effect Model

In the last four decades, longitudinal data analysis has become an object of research interest by researchers. In the study of economics, this analysis is called panel data analysis and has developed into a major subfield of econometrics, (Greene, 2018). According to (Baltagi and Bresson, 2012), Some of the advantages obtained in analysis using panel data include panel data providing more informative, more varied data, lower collinearity between variables, more degrees of freedom, and more efficiency. The second advantage of studying repeated cross-sectional data from observations is that researchers can study the dynamics of change better because panel data can detect and measure unobservable effects in both cross-sectional and pure time-series data. Another advantage is that panel data allows researchers to study more complex behavioural models because autocorrelation between individuals can be resolved compared to cross-sectional analysis or pure time series analysis, (Greene, 2018). The basic framework of the panel data regression model can model differences in behavior between individuals which is expressed in the following equation:

$$y_{it} = x'_{it}\beta + z'_i\alpha + \varepsilon_{it} \tag{1}$$

with as many as K independent variable on x_{it}.

The element of heterogeneity derived from individual effects is expressed by $z'_i\alpha$, where z_i contains constants and sets of individual or group-specific variables, which can be observed, such as race, gender, location, and so on; or unobserved, such as family-specific characteristics, individual heterogeneity in skills or preferences, and so on, all of which are held to be constant

over time t, (Gardebroek and Lansink, 2003). This model is a classical regression model. If z_i observed for all individuals, then the entire model can be treated as an ordinary linear model and fits the least squares. Complications arise when c_i unobserved, which will be the case in most applications, (Baltagi and Liu, 2012).

The fixed effects model in panel data analysis allows unobserved individual effects to be correlated with the included variables, (Elhorst, 2014). The model of the difference between the units as a parametric shift of the regression function. This model may be viewed as valid only for cross-sectional units in this study, not for additional ones outside the sample. For example, a comparison between countries might include a complete set of countries that it is reasonable to assume that the model is constant. As before, to see the model formulation in the block of observations T for group i, y_i, X_i, $u_i i$, and ε_i. For observation T obtained

$$\eta_{it} = \varepsilon_{it} + u_i \tag{2}$$

and $\eta_i = [\eta_{i1}, \eta_{i2}, \ldots, \eta_{iT}]'$ hereinafter referred to as the error component model.

Let we assume the strict exogeneity

$$
\begin{aligned}
E(\varepsilon_{it}|X_i) &= E(u_i|X_i) = 0, && (3) \\
E(\varepsilon_{it}^2|X_i) &= \sigma_\varepsilon^2, && (4) \\
E(u_i^2|X_i) &= \sigma_u^2, && (5) \\
E(\varepsilon_{it}u_j|X_i) &= 0, \forall i, t, j, && (6) \\
E(\varepsilon_{it}\varepsilon_{js}|X_i) &= 0, if t \neq s, or i \neq j, && (7) \\
E(u_i u_j|X_i, X_j) &= 0, i \neq j. && (8) \\
&&& (9)
\end{aligned}
$$

For group i, y_i, X_i, $u_i i$, and ε_i of T observations, we assume the error component models are

$$\eta_{it} = \varepsilon_{it} + u_i \tag{10}$$

and

$$\eta_i = [\eta_{i1}, \eta_{i2}, \ldots, \eta_{iT}]'. \tag{11}$$

Thus,

$$
\begin{aligned}
E(\eta_{it}^2|X_i) &= \sigma_\varepsilon^2 + \sigma_u^2, && (12) \\
E(\eta_{it}\eta_{is}|X_i) &= \sigma_u^2, t \neq s, && (13) \\
E(\eta_{it}\eta_{is}|X_i) &= 0, \forall t, s, if i \neq j. && (14) \\
&&& (15)
\end{aligned}
$$

For the T observations for unit i, we have matrices variance

$$\sum = \begin{pmatrix} \sigma_\varepsilon^2 & \sigma_u^2 & \sigma_u^2 & \sigma_u^2 & \sigma_u^2 \\ \sigma_u^2 & \sigma_\varepsilon^2 + \sigma_u^2 & \ldots & \vdots & \sigma_u^2 \\ \sigma_u^2 & \sigma_u^2 & \sigma_u^2 & \sigma_u^2 & \sigma_\varepsilon^2 + \sigma_u^2 \end{pmatrix} = \sigma_\varepsilon^2 I_T + \sigma_u^2 i_T i_T' \tag{16}$$

where i_T is a $T \times 1$ column vector of 1s.

C. RESEARCH METHOD

This research is pure research which aims to develop theory and find the characteristics of Hausman and Taylor estimator. First of all, the writer defines a linear regression model on panel data. Next, the author shows the difference between the fixed effect model and the random effect model. In the random effects model, the author states that the estimator used is the Hausman and Taylor estimator. After that, the author analyzes the best characteristics of the estimator, namely unbiased, minimum variance, consistent, and efficient. Next, the author performs a data simulation. To find out the estimator has a minimum variance, in practice it is necessary to compare the Hausman and Taylor estimator with other estimators such as the Generalized Least Quare estimator and the Feasible Generalized Least Square estimator.

D. RESULT AND DISCUSSION

Given a linear model of random effect panel data

$$y_{it} = x'_{it}\beta + (\alpha + u_i) + \varepsilon_{it}. \tag{17}$$

components u_i is an element of random heterogeneity of observations i-th and is constant with time t. The estimator using the ordinary least square estimation method is obtained as follows.

$$b = (X'X)^{-1}Xy \tag{18}$$

$$= \left[\sum_{i=1}^{n} X'_i X_i\right]^{-1} \sum_{i=1}^{n} X'_i(X_i\beta + w_i) \tag{19}$$

$$= \beta + \left[\sum_{i=1}^{n} X'_i X_i\right]^{-1} \sum_{i=1}^{n} X'_i w_i \tag{20}$$

Using the Gauss-Markov Theorem, we have the OLS estimator is unbiased estimator, (Ahn and Schmidt, 2019). Unbiased estimator statistically means that the expected value of the estimator is the same as the parameter value, (Chan et al., 2008). Practically speaking, unbiased estimator means that the estimator value actually represents the population parameter value. The unbiasedness of this estimator is not affected by the number of samples observed. Furthermore, the OLS estimator is a consistent estimator, because if the deviation is formed on each segment, it is obtained

$$y_{it} - \bar{y}i = (x_{it} - \bar{x}_{it})'\beta + \varepsilon_{it} - \bar{\varepsilon}_i. \tag{21}$$

Thus, when the sample is large,

$$plim\left[\frac{e'_{means}e_{means}}{nT}\right] = \frac{\sigma_\varepsilon^2 + \sigma_u^2}{T}. \tag{22}$$

Then the estimation is continued by using the Generalized Least Square method and obtained

$$\hat{\beta} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y \tag{23}$$

used to test the orthogonality of the general and regressor effects. This test is based on the idea that under the no correlation hypothesis, both the LSDV and FGLS estimators are consistent, but LSDV is inefficient, whereas under the alternative, LSDV is consistent, but FGLS is not.

Breusch and Pagan in (Greene, 2018) has introduced the Lagrange multiplier test to test the random effects model based on OLS residuals with statistical hypotheses:

$$H_0 : \sigma_u^2 = 0 \tag{24}$$

$$H_1 : \sigma_u^2 > 0 \tag{25}$$

From a practical point of view, the dummy variable approach in the random effects model has an impact on the loss of degrees of freedom. The random effects model is based on the assumption that object-specific effects are not observed (z_i) uncorrelated with the included variables (x_{it}). This assumption is the main weakness of this model. However, the random effects model allows the model to contain observed time-invariant characteristics, such as demographic characteristics, that are not captured in the fixed effects model. Hausman and Taylor estimators provide solutions to overcome weaknesses in the random effects model but still accommodate the possibility of a fixed effect model.

Pay attention to the linear random effect model

$$y_{it} = x'_{it}\beta + z'_i\alpha + \varepsilon_{it} \tag{26}$$

$$= x'_{1it}\beta_1 + x'_{2it}\beta_2 + z'_{1i}\alpha_1 + z'_{2i}\alpha_2 + \varepsilon_{it} + u_i \tag{27}$$

The Hausman and Taylor estimator defines the four sets of observed variables in the model as follows:

x_{1it} is a variable that varies with time and is not correlated with u_i,

z_{1i} is a time invariant variable and does not correlate with u_i,

x_{2it} is a variable that varies with time and is correlated with u_i,

z_{2i} is a time invariant variable and correlated with u_i.

There are five assumptions to obtain the Hausman and Taylor estimator on the random effects model:

1. $E[u_i|x_{1it}, z_{1i}] = 0$
   This means that there is no autocorrelation between random heterogeneity to time varying and time invariant.
2. $Var[u_i|x_{1it}, z_{1i}, x_{2it}, z_{2i}] = \sigma_u^2$
   This means that the variance of the heterogeneity effect is constant.
3. $Cov[\varepsilon_{it}, u_i|x_{1it}, z_{1i}, x_{2it}, z_{2i}] = 0$
   This means that error and heterogeneity are not correlated with time invariant and time varying variables.
4. $Var[\varepsilon_{it} + u_i|x_{1it}, z_{1i}, x_{2it}, z_{2i}] = \sigma^2 = \sigma^2 + \sigma_u^2$
   This means that the variance of the addition of error and heterogeneity is constant in the form of the sum of each error variance and heterogeneity variance.
5. $Corr[\varepsilon_{it} + u_i, \varepsilon_{it} + u_i|x_{1it}, z_{1i}, x_{2it}, z_{2i}] = \rho = \frac{\sigma_u^2}{\sigma_\varepsilon^2 + \sigma_u^2}$
   This means that the correlation of error and heterogeneity to time varying and time invariant variables is a comparison $Var[u_i|x_{1it}, z_{1i}, x_{2it}, z_{2i}]$ with $Var[\varepsilon_{it} + u_i|x_{1it}, z_{1i}, x_{2it}, z_{2i}]$.

The following are the steps to obtain a consistent and efficient Hausman and Taylor estimator:

1. Taking deviations from group means

$$y_{it} = x'_{1it}\beta_1 + x'_{2it}\beta_2 + z'_{1i}\alpha_1 + z'_{2i}\alpha_2 + \varepsilon_{it} + u_i \tag{28}$$

$$y_{it} - \bar{y}_{i.} = (x_{1it} - \bar{x}_{1i.})'\beta_1 + (x_{2it} - \bar{x}_{2i.})'\beta_2 + \varepsilon_{it} - \varepsilon_{i.} \tag{29}$$

2. Obtain the Least Square Dummy Variable from fixed effect model of $\beta = (\beta'_1, \beta'_2)'$ based on x_1 and x_2.
3. Stack the group means of these residuals e_{it} in a full-sample-length data vector.

$$e^*_{it} = \bar{e}i. = \frac{1}{T}\sum_{t=1}^{T} y_{it} - x'_{it}b_w \tag{30}$$

where $t = 1, \ldots, T$ and $i = 1, \ldots, n$ so that the equation

$$y_{it} - \bar{y}_{i.} = (x_{1it} - \bar{x}_{1i.})'\beta_1 + (x_{2it} - \bar{x}_{2i.})'\beta_2 + \varepsilon_{it} - \frac{1}{T}\sum_{t=1}^{T} y_{it} - x'_{it}b_w \tag{31}$$

4. Residual variance in the regression is a consistent estimator of

$$\sigma_*^2 = \sigma_u^2 + \frac{\sigma_\varepsilon^2}{T} \tag{32}$$

It means $T \to \infty, \sigma_*^2 = \sigma_u^2$.

5. Defined weighted instrumental variable estimator

$$w'_{it} = (x'_{1it}, x'_{2it}, z'_{1i}, z_{2i}) \tag{33}$$

then

$$w^{*'}_{it} = w'_{it} - \hat{\theta}\bar{w}'_{i.} \tag{34}$$

where $\hat{\theta}$ is estimator of $\left(1 - \sqrt{\frac{\hat{\sigma}_\Delta^2}{\sigma_\varepsilon^2 + T\sigma_u^2}}\right)$.

The transformed data are collected in the rows data matrix W^* and in column vector y^*.

6. Time invariant variables in w_it and the group mean is the original variable then we have instrumental variable

$$v'_{it} = [(x_{1it} - \bar{x}_{1i.})', (x_{2it} - \bar{x}_{2i.})', z'_{1i}\bar{x}_{1i}]. \tag{35}$$

Thus, we have instrumental variable estimator for Hausman and Taylor model random effect is

$$\hat{\beta}_{IV} = \left((W^{*'}V)(V'V)^{-1}(VW^*)\right)^{-1}\left((W^{*'}V)(V'V)^{-1}(V'y^*)\right). \tag{36}$$

As a case study, the data set obtained from the Panel Study of Income Dynamics (PSID) during 1968-1972 involved 750 men aged 25 to 55 years to see the economic benefits of education. The variables of concern are experience (years of schooling), health,

race (as many as 70 respondents are citizens of color), union membership, unemployment (long unemployment in the previous year). Table 1 shows the estimation results using OLS and Hausman and Taylor estimation.

Table 1. Hausman and Taylor Estimation

	Variables	OLS	HT
x_1	Experience	0.013	0.021
	Health	-0.084	-0.028
	Unemployed	-0.001	-0.056
x_2	Experience		0.024
	Unemployed		-0.056
z_1	Race	-0.085	-0.028
	Union	0.045	0.122
	Schooling	0.067	
z_2	Schooling		0.124

The main focus of this study is the coefficient on schooling because Schooling and, possibly, Experienced, and Unemployed are correlated with latent effects, there may be serious bias in conventional estimates of this equation. The school variable was treated as endogenous (correlated with u_i) in both cases. The estimation result for the schooling variable was estimated to be 0.067, a value which the authors assumed was too small. As we saw earlier, even in the presence of a correlation between the measured effect and the latent effect, in this model, the Hausman and Taylor estimators provide a consistent estimator of the coefficients for variable time variables. In the HT column, estimates of instrumental variables are presented. Therefore, in future studies we can use it in the Hausman specification test to test the correlation between the included variables and latent heterogeneity.

E.   CONCLUSION AND SUGGESTION

Based on the results and discussion, it is found that the Hausman and Taylor estimator on the random effects panel data is an estimator that is consistent and efficient even though it is not unbiased.

ACKNOWLEDGEMENT

REFERENCES

Agus Astapa, I. G., Suwardika, G., and Suniantara, I. K. P. (2018). Analisis Data Panel Pada Kinerja Reksadana Saham. *Jurnal VARIAN*, 1(2):59–69.

Ahn, S. C. and Schmidt, P. (2019). Efficient Estimation of Models for Dynamic Panel Data. *Arizona State University*, pages 1–30.

Arellano, M. (1993). On The Testing of Correlated Effects with Panel Data. *Journal of Econometrics*, 59(1-2):87–97.

Baltagi, B. and Bresson, G. (2012). A Robust Hausman-Taylor Estimator. *Center fo Policy Research*, 140(1):1–46.

Baltagi, B. H., Bresson, G., and Pirotte, A. (2003). Fixed Effects, Random Effects or Hausman-Taylor? A Pretest Estimator. *Economics Letters*, 79(3):361–369.

Baltagi, B. H., Egger, P., and Kesina, M. (2014). Firm-Level Productivity Spillovers in China's Chemical Industry: A Spatial Hausman-Taylor Approach. *CESifo Working Paper*, 5114:1–37.

Baltagi, B. H. and KhantiAkom, S. (1990). On Efficient Estimation with Panel Data: An Empirical Comparison of Instrumental Variables Estimators. *Journal of Applied Econometrics*, 5(4):401–406.

Baltagi, B. H. and Liu, L. (2012). The Hausman-Taylor Panel Data Model with Serial Correlation. *Statistics and Probability Letters*, 82(7):1401–1406.

Chan, F. K., So, H. C., Zheng, J., and Lui, K. W. (2008). Best linear Unbiased Estimator Approach for Time-Of-Arrival Based Localisation. *IET Signal Processing*, 2(2):156–162.

Chapman, D. G. and Robbins, H. (1951). Minimum Variance Estimation Without Regularity Assumptions. *The Annals of Mathematical Statistics*, 22(4):581–586.

Cng, H. C., Th, ., Ngc, B., Th, B., and Linh, n. H. (2013). Trade Liberalization and Foreign Direct Investment in Vietnam : A Gravity Model Using Hausman - Taylor Estimator Approach M t cách ti p c n thông qua mô hình L c h p d n. *Journal of Science and Development*, 11(1):85–96.

Egger, P. and Pfaffemayr, M. (2001). Distance, Trade and FDI: A Hausman-Taylor SUR Approach. *WIFO Working Papers*, 164:1–27.

Elhorst, J. P. (2014). *Spatial Panel Data Models*, pages 37–93.

Frondel, M. and Vance, C. (2010). Fixed, Random, or Something in Between? A Variant of HAUSMAN's Specification Test for Panel Data Estimators. *Ruhr Economic Papers*, 160:1–13.

Gardebroek, C. and Lansink, A. O. (2003). Estimating Farm Productivity Differentials Using Panel Data: The Hausman-Taylor Approach. *Journal of Agricultural Economics*, 54(3):397–415.

Greene, W. H. (2018). *Econometric Analysis*. Pearson Education, Inc., New York, 8 edition.

Kojima, Y., Parcell, J., Cain, J., Kojima Charles, Y. H., Parcell, J., and Cain, J. (2016). A Global Demand Analysis of Vegetable Oils for Food and Industrial Use: A Cross-Country Panel Data Analysis with Spatial Econometrics. *Agricultural & Applied Economics Association*, pages 0–27.

Maani, S., Wang, X., and Rogers, A. (2015). Network Effects, Ethnic Capital and Immigrants' Earnings Assimilation: Evidence From a Spatial, Hausman-Taylor Estimation. *IZA Discussion Paper*, (9308).

Millo, G. and Piras, G. (2012). Splm: Spatial Panel Data Models in R. *Journal of Statistical Software*, 47(1).

Sarafidis, V. and Wansbeek, T. (2010). Cross-Sectional Dependence in Panel Data Analysis. *MPRA Paper*, 20367(75086):1–44.

Xu, Y. and Li, A. (2020). The Relationship Between Innovative Human Capital and Interprovincial Economic Growth Based on Panel Data Model and Spatial Econometrics. *Journal of Computational and Applied Mathematics*, 365:112381.

Yusra, I., Hadya, R., Begawati, N., Istiqomah, L., Afriyeni, and Kurniasih, N. (2019). Panel Data Model Estimation: The Effect Of Managerial Ownership, Capital Structure, and Company Size on Corporate Value. *Journal of Physics: Conference Series*, 1175(1).