WORKING PAPER

ITLS-WP-12-07

Examining estimator bias and efficiency for pseudo panel data: A Monte Carlo simulation approach.

By

Chi-Hong Tsai[1], Waiyan Leong, Corinne Mulley & Geoffrey Clifton

[1] Corresponding author

April 2012

ISSN 1832-570X

INSTITUTE of TRANSPORT and LOGISTICS STUDIES

The Australian Key Centre in
Transport and Logistics Management

The University of Sydney

Established under the Australian Research Council's Key Centre Program.

NUMBER: Working Paper ITLS-WP-12-07

TITLE: **Examining estimator bias and efficiency for pseudo panel data: A Monte Carlo simulation approach.**

ABSTRACT: Pseudo panel data have been increasingly applied in empirical research as an alternative approach to a longitudinal analysis when genuine panel data are unavailable. However, conventional techniques are typically used to estimate pseudo panel data models without careful consideration to some unique properties of pseudo panel data. Ignoring properties such as time-varying cohort effects, a small number of constructed cohorts, large between-group variance, and trade-offs between cohort sizes and number of cohorts potentially lead to estimation bias or inefficiency not observed in genuine panel data.

This paper presents a Monte Carlo experiment with scenarios that are designed to generate, under conditions of limited observations, various data possessing pseudo panel data characteristics, and evaluates the performances of various estimators using the simulation results. The main research findings are that the large between-group variation of the exogenous variable and the variance of fixed group effects in pseudo panel data are the primary causes of estimation bias and inefficiency. Other factors including the cohort size and potential non-spherical errors have a smaller impact on the estimators' performances. An empirical application using Sydney Household Travel Survey data is also presented to illustrate the simulation findings.

KEY WORDS: ***Pseudo panel data; dynamic model; Monte Carlo simulation; estimator bias and efficiency.***

JEL Codes: C15, C23, C52, R41

AUTHORS: **Tsai, Leong, Mulley & Clifton**

CONTACT: INSTITUTE of TRANSPORT and LOGISTICS STUDIES (C37)
The Australian Key Centre in Transport and Logistics Management

The University of Sydney NSW 2006 Australia

Telephone: +612 9351 0071
Facsimile: +612 9351 0088
E-mail: business.itlsinfo@sydney.edu.au
Internet: http://sydney.edu.au/business/itls

DATE: April 2012

1. Introduction

In the last decade, the use of pseudo panel data has become more common in economic behaviour research. Pseudo panel data, introduced by Deaton (1985), are created from repeated cross-sectional data by grouping individuals into analyst-created cohorts based on some time-invariant variables which are observable for all individuals, such as birth year, gender, and household location (Verbeek, 1992). The pseudo panel approach is an alternative method of conducting a longitudinal study in the absence of genuine panel data[1] to identify the dynamics of economic behaviour.

The construction of pseudo panel data aims to produce cohorts that can be grouped together as a unique individual panel unit (Verbeek, 1992). Thus, groups may be defined as aggregations of similar cohorts across time, for example, a group consisting of cohorts of individuals born in a certain year. Pseudo panel data research reported in the literature (Gassner, 1998; J.M. Dargay and Vythoulkas, 1999; Dargay, 2002; Gardes et al., 2005; J. Dargay, 2007; Huang, 2007; Weis and Axhausen, 2009; Warunsiri and McNown, 2010; Bernard et al., 2011) has demonstrated that groups of created cohorts can be empirically estimated as if they were genuine panel data. Thus, conventional estimation techniques such as Fixed Effect (FE), Random Effect (RE) and Instrumental Variable (IV) estimators are commonly applied in pseudo panel data research.

However, it may be unwise to ignore the unique properties of pseudo panel data. As a consequence of how cohorts are created, the cross-sectional variance of the exogenous variables among cohorts across different groups (between-group variance) is usually larger than the variance of the exogenous variables among cohorts in the same group across time (within-group variance). This effect has an impact on the estimation efficiency for some estimators, especially the FE estimator that only takes account of within-group variation.

With pseudo panel data, the unobserved cohort effect is time-varying because each cohort, even within the same group, is composed of different individuals over time. Hence, non-spherical errors such as heteroscedasticity are likely to be introduced which cannot be controlled through conventional estimation techniques that incorporate only fixed individual or group effects. Moreover, since repeated cross-sectional surveys are not primarily concerned with understanding longitudinal questions at a disaggregate level, a rather small number of groups (G) and short time periods (T) are normally obtained as compared with aggregate genuine panel data. Therefore, in pseudo panel data construction, there is a trade-off between the cohort size (number of individuals in a cohort) and the total number of cohorts ($G*T$). Increasing the cohort size reduces the estimation bias, but it also decreases the estimation efficiency because the number of groups being estimated is reduced (Verbeek and Nijman, 1992).

Given the features identified above, applying estimation techniques developed for genuine panel data to limited-observation pseudo panel data, as is commonly practised in the literature, may lead to problematic estimation results and invalid policy interpretations. This paper tests the validity of this commonly used approach by employing a Monte Carlo simulation to investigate the performance of various estimators in static and dynamic pseudo panel models, whilst taking account the properties of pseudo panel data including: time-varying unobserved cohort effect; larger between-group variance than within-group variance; a small total number of cohorts; and the trade-off between cohort size and number of cohorts. An empirical demonstration using the Sydney Household Travel Survey (SHTS) is conducted to compare the estimators' performances in real data estimation with the Monte Carlo simulation results.

[1] Genuine panel data are longitudinal data that trace the same individuals or panel units over time and such data may be unavailable in some countries (Deaton, 1985).

2. Literature review

2.1 *Static panel data model estimation*

The theoretical background of the estimation techniques and model assumptions of genuine panel data underpin the knowledge of pseudo panel data models. Consider a simple static genuine panel data model as described in equation (1):

$$y_{it} = \beta_0 + \beta_1 x_{it} + u_{it}, \quad u_{it} = \alpha_i + \varepsilon_{it} \tag{1}$$

with *i* denotes the panel units (eg. firm, country, or household), *t* denotes the time period, u_{it} is the composite error term that includes the fixed individual effect α_i and the independent error term ε_{it}.

It is well known that the presence of α_i leads to the pooled Ordinary Least Squares (OLS) estimator being inefficient and producing unreliable statistical inferences. The conventional response is to use FE and RE estimators to incorporate unobserved individual effects (Hausman and Taylor, 1981). FE and RE are efficient under the assumption that ε_{it} is independent and identically distributed (*i.i.d.*).

However, if non-spherical errors are present, then the Feasible Generalized Least Squares (FGLS) estimator (Parks, 1967) and Panel-Corrected Standard Error (PCSE) estimator (Beck and Katz, 1995) are preferred. The PCSE estimator is developed to correct the standard errors obtained from OLS estimation. Since OLS estimation is only inefficient (giving still unbiased estimator) in the presence of unobserved individual effects and non-spherical errors, a better approximation of the standard errors from OLS potentially gives unbiased and more efficient estimates. For the purposes of this paper, the FGLS estimator is not considered since it has been shown to under estimate the standard errors when *T* is not substantially larger than *N* (Beck and Katz, 1995) and this is typically the case with pseudo panel data because they are usually constructed from repeated cross-sectional individual surveys with a short time period.

2.2 *Dynamic panel data model estimation*

Genuine panel data models can also be used to capture dynamic economic behaviour through a dynamic panel data model. Various dynamic model forms have been developed and empirically employed according to the nature of economic behaviour assumed. The partial adjustment model as specified in equation (2) is a commonly applied model to take account of the effect of previous behaviour on current behaviour, where the lagged dependent variable, y_{it-1}, is used to represent the economic behaviour in the previous time period *t-1*.

$$y_{it} = \lambda y_{it-1} + \beta_1 x_{it} + u_{it}, \quad u_{it} = \alpha_i + \varepsilon_{it} \tag{2}$$

In estimating model (2), it is well-known that OLS is biased upwards and FE is biased downwards as a result of the endogeneity between y_{it-1} and the error term (Nickell, 1981). Endogeneity can be addressed through the Instrumental Variable (IV) estimator first developed by Anderson and Hsiao (1981). This method introduces instruments which are correlated with the explanatory variables but uncorrelated with error terms. The parameters are then estimated using two-stage least squares (2SLS). The lagged values of the endogenous variable such as y_{it-2} are regarded as an ideal instrument for y_{it-1}. However, Arellano and Bond (1991) showed that the 2SLS estimator is not efficient because the first-differenced transformation is likely to produce serial correlation and that the use of a Generalised Method of Moments

(GMM) estimates the parameters more efficiently by imposing moment conditions. Moreover, Kiviet (1995) demonstrated that the IV estimation methods may lead to small sample bias and large standard errors. Beck and Katz (2004) also showed that although OLS is biased in the dynamic model, it performs better than the IV approach in terms of Root Mean Square Errors (RMSE) which takes account of both bias and efficiency. Therefore, for dynamic panel data models, OLS may still be preferred to IV techniques.

The discussion above highlights that when comparing the performance of various estimators, bias is not the only factor that should be taken into consideration. An unbiased estimation may be obtained but at a high cost of inefficiency. Thus, both the parameter estimates and standard errors need to be evaluated to identify the most appropriate estimator.

2.3 Pseudo panel data estimation

The pseudo panel data model was first introduced by Deaton (1985) as follows:

$$\bar{y}_{gt} = \beta_0 + \beta_1 \bar{x}_{gt} + \bar{\alpha}_{gt} + \bar{\varepsilon}_{gt}, \quad \text{g}=1,\dots,\text{G}; \quad t=1,\dots,\text{T} \tag{3}$$

Compared to the genuine panel data model (equation (1)), equation (3) uses the subscript g instead of i to denote the created groups in the pseudo panel data. The variables $\bar{y}_{gt}$ and $\bar{x}_{gt}$ represent the way in which the observation of each variable is the mean value for all individuals classified into group g at time period t. The most critical point in equation (3) that distinguishes the pseudo panel data model from the genuine panel data model is that the average unobserved cohort effect $\bar{\alpha}_{gt}$ is time-varying, whereas the unobserved individual effect (α_i) is fixed in genuine panel data model (equation (1)). The result is that the time-varying cohort effects will not be eliminated through the demeaned transformation in the FE estimation, so the conventional FE estimator will be problematic whether in the static or dynamic pseudo panel model.

Deaton (1985) highlighted this point by emphasising that the sample cohort means in pseudo panel data sets are consistent but “error-ridden” estimates of the true population means which are unobservable. Deaton proposed using an errors-in-variable estimator to estimate this population relationship. Verbeek and Nijman (1992) found that with a sufficiently large cohort size (n_c), the time-varying $\bar{\alpha}_{gt}$ can be treated as constant over time as $\bar{\alpha}_g$, so that the pseudo panel data can be used as genuine panel data, using conventional estimation techniques. However, Verbeek and Nijman also found that reducing n_c improves the estimation efficiency because more groups can be created resulting in more observations in the pseudo panel data.

Table 1 summarises some recent pseudo panel studies and the estimation techniques used. With static models, most studies adopt FE as the estimator (Gassener, 1998; Gardes et al., 2005; Huang, 2007; Weis and Axhausen, 2009; Warunsiri and McNown, 2010). Following the results of Deaton (1985) and Verbeek and Nijman (1992), these studies ignore time-varying unobserved heterogeneity $\bar{\alpha}_{gt}$ because n_c is believed to be sufficiently large. In this case, FE is theoretically unbiased and consistent.

Dynamic models can also be estimated on pseudo panel data. As with genuine panel data, the lagged dependent variable is likely to be correlated with the error term and thus causes estimation bias. The IV estimator can be used to address the endogeneity problem (Dargay and Vythoulkas, 1999; Bernard et al., 2011) and results show that the IV estimator should be chosen over the FE estimator (Dargay and Vythoulkas, 1999). However, Dargay (2002), Dargay (2007), and Huang (2007) only reported their estimation results based on the FE estimator which is likely to be downward biased.

Although there is a substantial body of applied pseudo panel data research, we observe that there is much less discussion on how the key properties of pseudo panel data such as non-spherical errors and large between-group variation may impact the bias and efficiency of the estimators. A focus on a small number of cohorts is also important since many pseudo panel data sets have a relatively small *G* and *T*, with *G*<20 and *T*<20 arising from the requirements of cohort construction. As suggested by Verbeek and Nijman (1992), these conditions normally stipulate at least one hundred individuals in each cohort as sufficient for the time-varying unobserved cohort effect $\bar{\alpha}_{gt}$ to be treated as constant over time. These characteristics mean that there is no guarantee that the properties of estimators used in genuine panel data will all carry over to pseudo panel data.

3. Monte Carlo experiment

Monte Carlo simulations employed in this paper are designed to simulate data that share similar characteristics with actual data scenarios. These "real data" scenarios are conditioned on parts of the properties observed from pseudo panel data created from the Sydney Household Travel Survey (SHTS) which forms the empirical demonstration in Section 5. As with the real data, both static and dynamic models are studied in the Monte Carlo simulation.

3.1 Simulation models

For simulation, the following static model in equation (4) and partial adjustment dynamic model in equation (5) for pseudo panel data are used:

$$\bar{y}_{gt} = \beta_0 + \beta_1 \bar{x}_{gt} + \bar{u}_{gt} \tag{4}$$

$$\bar{y}_{gt} = \lambda \bar{y}_{gt-1} + \beta_1 \bar{x}_{gt} + \bar{u}_{gt} \tag{5}$$

where

$$\bar{u}_{gt} = \bar{\alpha}_g + \bar{\omega}_{gt} + \bar{\varepsilon}_{gt}$$

$$\bar{\omega}_{gt} \sim N(0,\ (\sigma_\alpha / \sqrt{n_c})^2)$$

$$n_c \sim N(\bar{n}_c, \sigma_{n_c}^2)$$

$$\bar{\varepsilon}_{gt} \sim N(0, 1^2)$$

$$\bar{\alpha}_g \sim N(0, \sigma_\alpha^2)$$

The composite error term $\bar{u}_{gt}$ includes three elements: the fixed group effect $(\bar{\alpha}_g)$ for a given created group in the pseudo panel data set, the time varying cohort effect within groups $(\bar{\omega}_{gt})$, and *i.i.d.* disturbances $(\bar{\varepsilon}_{gt})$. σ_α^2 is an experimentally controlled variable to simulate the variance in the fixed group effect. Throughout the paper, $\bar{\alpha}_g$ is assumed to be uncorrelated with the exogenous variable $(\bar{x}_{gt})$. Allowing $\bar{\omega}_{gt}$ to be drawn from a random distribution allows time variation in the cohort effect since within-group cohorts across time are not created from the same individuals. $\bar{\omega}_{gt}$ is assumed to be normally distributed with a mean of zero, and its variance is positively related to σ_α^2 but negatively related to the square root of cohort size (n_c). This assumption is a convenient way of allowing the variance of $\bar{\omega}_{gt}$ to be smaller than the variance of $\bar{\alpha}_g$ and to allow a larger n_c to reduce the variance of $\bar{\omega}_{gt}$. To simulate the unequal cohort sizes found in real life data, n_c is assumed to be normally distributed across all the created cohorts, with a mean and variance bounded in the experiment by values computed from the real pseudo panel data constructed from the SHTS (Section 5).

The simplified models in equation (4) and (5) with one single exogenous variable $(\bar{x}_{gt})$ and one lagged dependent variable $(\bar{y}_{gt-1})$ are employed. This simplification is to avoid the confounding results from possible interactions between multivariate exogenous variables, and to allow the simulation results to be compared with previous studies in the related literature.

3.2 *Experiment Design*

As highlighted in Section 1, the features of pseudo panel data requiring further examination include: time-varying unobserved cohort effects ($\overline{\omega}_{gt}$); larger between-group variance in the exogenous variable than within-group variance; small total number of cohorts (*C*); and the trade-offs between n_c and *C*. $\overline{\omega}_{gt}$ has been incorporated in the simulation models as specified in Section 3.1. The other properties are examined through the following scenarios as summarised in Table 2. Each scenario is replicated for one thousand times in the simulation experiment.

Scenarios 3/4/6 are designed to allow the variance of the exogenous variable to be larger between groups than within groups, as compared to being identically distributed between and within groups as in Scenario 1/2. Scenario 5 reverses Scenarios 3/4/6. The magnitude of σ_α^2 is believed to have an impact on the estimation results because it is the factor that causes endogeneity and non-spherical errors, so a larger and a smaller effect are linked to each of the exogenous variable scenarios.

Scenario 6 is designed for a larger *C*, with a correspondingly smaller n_c, compared with Scenario 3. Note that *C* and n_c are related, given that the total number of sampled individuals is fixed. The simulation data in the experiment is made more similar to real data by conditioning the values of the between-group variance, the within-group variance, the number of cohorts and the cohort size on the SHTS pseudo panel data (discussed in more detail in Section 5 below).

Throughout the six scenarios *T* is kept constant at thirteen, and *G* is only changed in Scenario 6 for the purpose of investigating the trade-off between cohort size and number of cohorts, rather than examining the consistency properties of the estimators. Consistency, although an important measurement of a estimator's performance, is not the primary focus in this analysis because the aim of this simulation experiment is to examine estimator properties within the constraints of real data estimation where there is little flexibility in expanding the number of panel units (whether in pseudo or genuine panel data) or number of time periods.

3.3 *Estimators and performance measurements*

The properties of estimators commonly used in pseudo panel studies are compared in this Monte Carlo experiment. For static models, OLS is used as a reference estimator to be compared with FE, RE and PCSE estimators. Previous pseudo panel data research did not use the PCSE estimator but it is included in this analysis because PCSE is able to correct the standard error estimation problem arising from non-spherical errors. For dynamic models, the System Generalized Method of Moments (GMM) (Blundell and Bond, 1998) is also employed given its ability to incorporate the endogeneity between the lagged dependent variable and error terms, and given that it has been suggested as an appropriate estimator for pseudo panel data from previous Monte Carlo experiments (McKenzie, 2004; Inoue, 2008). However, as seen in Table 1, it is not commonly used. The GMM method applied in this analysis employs the second lag of the dependent variable ($\overline{y}_{gt-2}$) as an instrumental variable.

The evaluation of the performance of the estimators examined in the Monte Carlo experiment is based on the four measurements of bias, relative efficiency, overconfidence and Root Mean Squared Error (RMSE).

Two fundamental criteria to measure an estimator's performance are its bias and efficiency. Bias refers to the expectation of the difference between the value of the parameter estimate and its assumed value in the experimental design. The magnitude and direction of bias of each parameter can be identified from this measurement by comparing the assumed value of the experiment with the estimated value of the parameters. An efficient estimator by definition is an unbiased estimator with the least variance. If none of the evaluated estimators can be shown to have the minimum variance among all possible estimators, the measurement of relative efficiency can be used to compare the performance of the applied estimators. In principle, a more efficient estimator requires fewer observations to achieve the same statistical power, and has smaller standard errors of estimates when the same number of observations is applied to the estimation procedures being compared, and thus generates more reliable statistical inferences.

When evaluating standard errors, apart from efficiency, it is also important to examine whether the estimated standard errors are over-estimated or under-estimated. To evaluate the degree of overconfidence in the estimated standard errors, an indicator, developed in Beck and Katz (1995) and defined in equation (6), may be used. The numerator in equation (6) refers to the true sample variability for an experimental run of *R* replications of estimating $\tilde{\beta}$, whereas the dominator is the average reported standard errors of $\tilde{\beta}$ estimates over all replicates *r*. An overconfidence level larger than one hundred indicates that the estimator under-estimates the standard errors.

$$\text{Overconfidence} = 100 * \frac{\sqrt{\sum_{r=1}^{R} (\tilde{\beta}^{(r)} - \bar{\tilde{\beta}})^2}}{\sqrt{\sum_{r=1}^{R} (s.e.(\tilde{\beta}^{(r)}))^2}} \tag{6}$$

The Root Mean Square Error (RMSE) is an overall performance measure widely used to choose the most appropriate estimator (Judson and Owen, 1999) since it takes account of both bias and efficiency, albeit with equal weighting. RMSE is specified in equation (7).

$$\text{RMSE (Root Mean Square Error)} = \sqrt{BIAS(\tilde{\beta})^2 + Var(\tilde{\beta})} \tag{7}$$

The performance measurements summarised above provide an evaluation framework for the choice of estimators. However, there may not be a superior estimator that out-performs on all criteria. The choice of estimators may then depend on the researcher's primary concerns and willingness to trade-off the accuracy of the parameters with the validity of the confidence intervals as suggested by Reed and Ye (2011).

4. Analysis of Monte Carlo simulation

4.1 Simulation results for static models

The results for Scenarios 1 to 3 are presented in Table 3. The simulation results from Scenario 1, which assumes an identically distributed $\bar{x}_{gt}$ across time and groups and a large variance for the fixed group effect ($\sigma_\alpha = 0.5$), show that there is no substantial bias in the small pseudo panel data set (*G*=12, *T*=13). The FE and RE estimators, not surprisingly, are more efficient than OLS and PCSE in the presence of unobserved group effects $\bar{\alpha}_g$but with a slightly lower overconfidence level. In scenario 2 where σ_α^2 is reduced, it can be seen that the FE and RE estimators are not necessarily more efficient than OLS and PCSE. This is because $\bar{\alpha}_g$ which makes OLS and PCSE inefficient now has a much smaller impact on the estimation process. In these two scenarios, PCSE used to correct the non-spherical errors does not substantially improve the efficiency of OLS. This indicates that the simulation data generated for pseudo panel data do not possess strong non-spherical errors when $\bar{x}_{gt}$ is identically distributed.

Scenario 3 simulates data with a larger between-group cross sectional variance ($\sigma^2_{B,x}$) for the exogenous variable $\bar{x}_{gt}$, compared with its within-group time variance ($\sigma^2_{W,x}$). Comparing Scenario 3 with Scenario 1, the most striking difference is that the standard error of the FE estimator substantially increases from 0.084 to 0.419, whilst the standard errors of other estimators have a relatively minor increase. This result shows that when there is more between-group variation in $\bar{x}_{gt}$, the FE estimator which only takes account of within-group variation will be inefficient. Looking at the OLS, RE, and PCSE estimators, it can be seen that the PCSE estimator has substantially improved the efficiency of the OLS estimator as the standard error of β_1 drops from 0.181 to 0.162, showing that the non-spherical errors are more influential in this case than in Scenarios 1 and 2 where $\bar{x}_{gt}$ is identically distributed. However, judging from the high overconfidence level, both OLS and PCSE estimators tend to under estimate standard errors although they have the lowest RMSE. It is also important to note that the biases in Scenario 3 for all estimators are increased by more than 200% as compared with Scenario 1, although the magnitude of the biases are small in both cases. For Scenario 3, the RE estimator appears to be the most appropriate estimator because it is more efficient than FE whilst giving reliable standard errors as shown by the overconfidence indicator (overconfidence=103.153).

In summary, there is no one superior estimator in static models under the scenario where $\bar{x}_{gt}$ is identically distributed. In contrast, when larger between-group cross sectional variance relative to within-group time variance is present ($\sigma^2_{B,x} > \sigma^2_{W,x}$), the FE estimator is particularly inefficient and the RE estimator is suggested as the preferred estimator. The importance of using the overconfidence indicator is demonstrated because one may otherwise mistakenly ignore the possible under-estimated standard errors in Scenario 3 and favour the OLS or PCSE estimator.

4.2 Dynamic model simulation results

The dynamic model simulation results for Scenarios 1/2/3/4 are summarised in Table 4. Looking at Scenario 1 and Scenario 2 where $\bar{x}_{gt}$ is identically distributed across time and groups, λ_1 clearly shows an upward bias when using OLS and a downward bias when using the FE estimator. This effect is the well-known Nickell bias (Nickell, 1981). The OLS bias in λ_1 is reduced in Scenario 2 when σ_α is lowered to 0.2 but the bias remains the same in FE. This is because the bias of OLS comes from the interaction of λ and σ_α, whereas the bias of FE results from the correlation between the transformed lagged dependent variable and the transformed error terms (Baltagi, 2005). In contrast, β_1 does not show obvious biases for all estimators in either scenario. This suggests that the endogeneity problem in dynamic models only makes the estimate of λ_1 problematic but does not have a strong impact on β_1. In these two scenarios, FE performs better than other estimators when σ_α is large, but FE may not be the favoured estimator when σ_α is small. It is also important to note that GMM appears to be inefficient given the relatively large standard errors for both parameters. This concurs with the finding in Kiviet (1995) where he demonstrated that IV estimation methods may lead to small sample bias and large standard errors.

Where $\bar{x}_{gt}$ has a larger between-group variance ($\sigma^2_{B,x}$) than within-group variance ($\sigma^2_{W,x}$) (Scenario 3 and Scenario 4), the bias and standard errors of λ_1 are both increased as compared with Scenario 1 and Scenario 2. β_1 also becomes more biased for all estimators suggesting that the large between-group variation scenarios are likely to induce bias in exogenous variable estimates which were unbiased when $\bar{x}_{gt}$ was assumed to be identically distributed. As with the static model simulations, the standard errors of β_1 obtained from the FE estimator are increased by more than 400% suggesting severe inefficiency in the FE estimator. Although FE is the least biased estimator with the best overconfidence level, the inefficient standard errors will enlarge the confidence intervals which make the statistical inference unreliable. As before, no estimator is superior to all the others and the choice of estimators depends on the primary concerns of research. In this case, FE is preferred if the research focus is the unbiasedness of parameter estimates but if the research concern is the reliability of the statistical inference, OLS or RE is favoured given the lowest combined RMSE of λ_1 and β_1 estimates.

Table 5 summarises the results of using Scenario 3 as a reference scenario to compare the estimators from Scenarios 5 and 6. To further investigate the relationship between the distribution of $\overline{x}_{gt}$ and estimators' performances, Scenario 5 assumes that $\sigma^2_{W,x}$ is larger than $\sigma^2_{B,x}$. Compared with scenario 3, it can be seen that the biases of β_1 are moderately reduced for all estimators, whilst the biases of λ_1 do not change noticeably. Furthermore, all the standard errors are decreased, particularly for the FE estimator where the standard error drops from 0.440 to 0.176 for β_1. Given the lowest combined RMSE and the best overconfidence level, FE is the preferred estimator in this case.

Scenario 6 is used to evaluate the trade-off between cohort sizes (n_c) and number of cohorts (C) as discussed in Section 2.3. In this scenario, n_c is specified as n_c ~N(50, 15^2) as opposed to n_c ~N(150, 50^2) in previous scenarios. The results show that the biases are moderately increased for the OLS, RE, and PCSE estimators, but are less evident for the FE and GMM estimators. However, similar to Verbeek and Nijman (1992), standard errors are reduced for all estimators by around 40 percent to 45 percent, and this is because of the increase in the number of groups. If the RMSE for $\lambda_1$and β_1 are added up to form an overall measurement of error, then using a smaller n_c and a larger C as in Scenario 6 results in better results across all estimators as compared with Scenario 3. This finding suggests that reducing the average cohort size will improve the overall statistical inference from the estimation results at a relatively low cost of increasing the bias.

5. Empirical application

To illustrate the impact of pseudo panel data properties on the performance of various estimators, this section presents a pseudo panel data set constructed from the Sydney Household Travel Survey (SHTS). The SHTS is a repeated cross-sectional data set which has been undertaken continuously since 1997/1998. There are now thirteen consecutive years of data available. This continuous survey captures individuals' travel behaviour and socio-demographic characteristics which collectively provide important information for travel mobility analysis.

The constructed pseudo panel from the SHTS grouped public transport users based on their birth year and household distance to Sydney Central Business District (CBD). This process succeeded in reducing intra-cohort variation and increasing between-group variation to ensure each cohort can be treated as if it is an independent panel unit over time as suggested by Verbeek (1992).

Two data sets, one with a large average cohort size and another with a small average cohort size, are created to examine how n_c and C affect the estimation results. The first pseudo panel data set contains twelve groups classified by birth year and by distance to CBD for each of the thirteen years of available annual data. This grouping yields C=156 cohorts in total with n_c averaging 143. The second data set has a smaller average n_c of 86 individuals with twenty groups created by the same rules applied to the first data set but with a lower aggregation level.

A simple dynamic public transport demand model (equation (5)) is built to identify the relationship between public transport demand and its key explanatory variable of trip price. The public transport demand is defined as average number of public transport trips per person at a cohort level (*pttrip*). The price variable (*price*) is created from each respondent's reported public transport ticket price and ticket type for each public transport trip, calculated by dividing the total reported ticket price by the estimated average number of trips for each individual for each ticket type using a locally based ticket journey multiplier (CitiRail, 2010).

The estimators examined in the Monte Carlo simulations are applied to this dynamic public transport demand model. The analysis of the estimation results in Table 6 focuses on the comparison between the estimators’ performances and not on policy application arising from parameter significance or lack thereof. It is not surprising that the price variable is insignificant for all estimators, because it is a simplified model with only one single exogenous variable and thus most of the variations in *pttrip* are captured by its lagged variable (*L.pttrip*).

Comparing the performance of the various estimators, it is clearly seen that the FE estimator, with smaller parameter values and larger standard errors, behaves differently to the other estimators. This corresponds to the simulation results which suggest that FE tends to be biased downwards and inefficient when $\bar{x}_{gt}$ has larger between-group variation than within-group variation. Note that in both pseudo panel data sets the price variable does have a larger between-group variance ($\sigma^2_{B,x}$>$\sigma^2_{W,x}$ in Table 6), corresponding to scenarios 3/4/6 of the simulation experiments. On the other hand, with the sole exception of the GMM estimator for *L.pttrip*, all the estimators are slightly more efficient in the second data set as a result of the increased number of groups being estimated, although the average cohort size is reduced, confirming the simulation outcomes in section 4.2.

6. Conclusion

This paper examines the estimation performance of OLS, FE, RE, PCSE, and GMM, using Monte Carlo simulation experiments using scenarios based on the typical properties of pseudo panel data. The static model simulation results suggest that the variance of the fixed group effect does not lead to severe bias of β_1. Instead, the distribution of $\bar{x}_{gt}$ is more likely to lead to estimation bias, as well as causing inefficiency for all estimators, especially the FE estimator.

For dynamic models, there is no unambiguously superior estimator when $\bar{x}_{gt}$ has a larger between-group variation. FE appears to be the least biased but with the largest standard errors, whereas OLS, RE, and PCSE are more efficient but with larger biases. However, the biases of OLS, RE, and PCSE can be decreased by reducing σ^2_α (scenario 4). This suggests that OLS, RE and PCSE are potentially unbiased and efficient estimators if σ^2_α can be minimized, possibly through better model specifications.

The trade-off between cohort size n_c and total number of cohorts C in pseudo panel data construction is also investigated. The findings indicate that using a data set with a smaller n_c but a larger C effectively improves the estimation efficiency, at a relatively low cost of a slight increase in bias.

This paper has highlighted the importance of understanding the nature and properties of panel data before deciding which estimator to use in empirical applications. Future research may conduct simulation experiments with multivariate models which possess more complex interactions and correlations between regressors and the composite error term, to investigate whether the recommendations of this paper require further revision.

References

Anderson, T.W. and Hsiao, C. 1981, "Estimation of dynamic models with error components", *Journal of the American Statistical Association*, vol.76, no.375, pp. 598-606.

Arellano, M. and Bond, S. 1991, "Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations", *Review of Economic Studies*, vol.58, no.2, pp. 277-297.

Baltagi, B.H. 2005, *Econometric Analysis of Panel Data*, Fourth Edition, Wiley, Chichester, pp. 147-148.

Beck, N. and Katz, J.N. 1995, "What to do (and not to do) with Time-Series Cross-Section Data", *The American Political Science Review*, vol.89, no.3, pp. 634-647.

Beck, N. and Katz, J.N. 2004, "Time-Series-Cross-Section issues: dynamics", Working paper, Document ID: 36, The Society for Political Methodology, Washington University, St. Louis.

Bernard, J.-T., Bolduc, D. and Yameogo, N. D. 2011, "A pseudo-panel data model of household electricity demand", *Resource and Energy Economics*, vol.33, no.1, pp. 315-325.

Blundell, R. and Bond, S. 1998, "Initial conditions and moment restrictions in dynamic panel data models", *Journal of Econometrics*, vol. 87, no. 1, pp. 115-143.

CitiRail 2010, "A compendium of CityRail travel statistics", Sydney, Australia.

Dargay, J. 2007, "The effect of prices and income on car travel in the UK", *Transportation Research Part A*, vol.41, no.10, pp. 949-960.

Dargay, J. M. 2002, "Determinants of car ownership in rural and urban areas: a pseudo-panel analysis", *Transportation Research Part E*, vol.38, no.5, pp. 351–366.

Dargay, J.M. and Vythoulkas, P.C. 1999, "Estimation of dynamic car ownership model: a pseudo-panel approach", *Journal of Transport Economics and Policy*, vol.33, no.3, pp. 287-302.

Deaton, A. 1985, "Panel data from time series of cross-sections ", *Journal of Econometrics*, vol.30, no.1-2, pp. 109-126.

Gardes, F., Duncan, G., Gaubert, P., Gurgand, M. and Starzec, C. 2005, "Panel and pseudo-panel estimation of cross-sectional and time series elasticities of food consumption: the case of U.S. and Polish data", *Journal of Business & Economic Statistics*, vol.23, no.2, pp. 242-253.

Gassner, K. 1998, "An estimation of UK telephone access demand using pseudo-panel data", *Utilities Policy*, vol.7, no.3, pp. 143-154.

Hausman, J.A. and Taylor, W.E. 1981, "Panel data and unobservable individual effects", *Econometrica*, vol. 49, no. 6, pp. 1377-1398.

Huang, B. 2007, "The use of pseudo panel data for forecasting car ownership", Doctoral Dissertation, University of London.

Inoue, A. 2008, "Efficient estimation and inference in linear pseudo-panel data models", *Journal of Econometrics*, vol.142, no.1, pp. 449-466.

Judson, R.A. and Owen, A.L. 1999, "Estimating dynamic panel data models: a guide for macroeconomists", *Economics Letters*, vol.65, no.1, pp. 9-15.

Kiviet, J.F. 1995, "On bias, inconsistency, and efficiency of various estimators in dynamic panel data models", *Journal of Econometrics*, vol.68, no.1, pp. 53-78.

McKenzie, D.J. 2004, "Asymptotic theory for heterogeneous dynamic pseudo-panels", *Journal of Econometrics*, vol.120, no.2, pp. 235-262.

Nickell, S. 1981, "Biases in dynamic models with fixed effects", *Econometrica*, vol.49, no.6, pp. 1417-1426.

Parks, R.W. 1967, "Efficient estimation of a system of regression equations when disturbances are both serially and contemporaneously correlated", *Journal of the American Statistical Association,* vol.62, no.318, pp. 500-509.

Reed, W.R. and Ye, H. 2011, "Which panel data estimator should I use?", *Applied Economics,* vol.43, no.8, pp. 985-1000.

Verbeek, M. 1992, "Pseudo Panel Data", in L. Matyas and P. Sevestre (eds.), *The Econometrics of Panel Data: Fundamentals and Recent Developments in Theory and Practice*, Kluwer Academic Publishers, pp. 303-315.

Verbeek, M. and Nijman, T. 1992, "Can cohort data be treated as genuine panel data?", *Empirical Economics,* vol.17, no.9, pp. 9-23.

Warunsiri, S. and McNown, R. 2010, "The returns to education in Thailand: a pseudo-panel approach", *World Development,* vol.38, no.11, pp. 1616–1625.

Weis, C. and Axhausen, K.W. 2009, "Induced travel demand: evidence from a pseudo panel data based structural equations model", *Research in Transportation Economics,* vol.25, no.1, pp. 8-18.

Table 1: Summary of selected pseudo panel studies

Author	Context of Study	Area	Grouping	Observations[1]	Model	Estimation technique
Gassner (1998)	Telephone Access	UK	Age	324 (G=27, T=12)	Static	FE; RE[2]

Dargay and Vythoulkas (1999)	Car ownership	UK	Age	165
				(G=16, T=12)
					Dynamic	FE;
						Pooled OLS;
						RE;
						RE-IV

Dargay (2002)	Car ownership	UK	Age	134, 152, 159[3] (G=15, T=14)	Dynamic	FE

Gardes et al. (2005)	Food consumption	US	Age;
			Education
				90
				(G=18, T=5)
					Static	Between;
						FE;
						First-differences

Dargay (2007)	Car travel	UK	Age	256 (G=16, T=20)	Dynamic	FE
Huang (2007)	Car ownership	UK	Age	254 (G=16, T=19)	Static; Dynamic	Pooled OLS; FE

Weis and Axhausen (2009)	Car travel	Switzerland	Age;
			Gender;
			Region
				838
				(G=140, T=7)
					Static;
					SEM[4]
						FE

Warunsiri and McNown (2010)	Return to education	Thailand	Age	220;
				(G=11, T=20)
				440[5]
				(G=22, T=20)
					Static	Pooled OLS;
						FE;
						IV

Bernard et al. (2011)	Electricity	Canada	Region; House size	100 (G=25, T=4)	Dynamic	IV-dummy

[1] Observations included in the estimation. The number of observations in some studies may be smaller than the product of T and C as a result of dropping cohorts less than 100 individuals.

[2] The RE estimator is employed but rejected after the Hausman's test.

[3] Households are grouped based on the geographical locations. This study created 134 cohorts in rural areas, 152 cohorts in urban areas, and 159 cohorts in other areas.

[4] Structure Equation Model.

[5] 220 cohorts from 2-year band age grouping and 440 cohorts from 1-year band age grouping.

Table 2: Scenario design for Monte Carlo experiments

Scenario	Variance in exogenous variable ($\bar{x}_{gt}$)	Assumed distribution of fixed group effects ($\bar{\alpha}_g$)	Size of data (G[1]; T)	Cohort Size (n_c)
1	$\sigma^2_{B,x}$= $\sigma^2_{W,x}$=1 E($\bar{x}_{gt}$)=0	$\bar{\alpha}_g$~N(0,0.5^2)	G=12; T=13	n_c~N(150, 50^2)
2	$\sigma^2_{B,x}$= $\sigma^2_{W,x}$=1 E($\bar{x}_{gt}$)=0	$\bar{\alpha}_g$~N(0,0.2^2)	G=12; T=13	n_c~N(150, 50^2)
3	($\sigma^2_{B,x}$, $\sigma^2_{W,x}$)=(0.5^2, 0.2^2); E($\bar{x}_{gt}$)=0	$\bar{\alpha}_g$~N(0,0.5^2)	G=12; T=13	n_c~N(150, 50^2)
4	($\sigma^2_{B,x}$, $\sigma^2_{W,x}$)=(0.5^2, 0.2^2); E($\bar{x}_{gt}$)=0	$\bar{\alpha}_g$ ~N(0,0.2^2)	G=12; T=13	n_c~N(150, 50^2)
5	($\sigma^2_{B,x}$, $\sigma^2_{W,x}$)=(0.2^2, 0.5^2); E($\bar{x}_{gt}$)=0	$\bar{\alpha}_g$ ~N(0,0.5^2)	G=12; T=13	n_c~N(150, 50^2)
6	($\sigma^2_{B,x}$, $\sigma^2_{W,x}$)=(0.5^2, 0.2^2); E($\bar{x}_{gt}$)=0	$\bar{\alpha}_g$ ~N(0,0.5^2)	G=36; T=13	n_c~N(50, 15^2)

1 G= number of created groups in the pseudo panel data set

2 $\sigma^2_{B,x}$ and $\sigma^2_{W,x}$ are between-group cross-sectional variance and within group time variance of exogenous variable.

Table 3: Simulation results for static models

Model: $\bar{y}_{gt} = 0.2 + 0.8 * \bar{x}_{gt} + \bar{u}_{gt}$

	OLS	FE	RE	PCSE
Scenario 1: $\sigma^2_{B,x}$= $\sigma^2_{W,x}$=1; $\bar{\alpha}_g$~N(0,0.5^2)				
β_1	0.803	0.805	0.805	0.803
β_1_SE[1]	0.090	0.084	0.083	0.089
β_1_BIAS	0.003	0.005	0.005	0.003
β_1_CONF[2]	100.291	96.427	98.181	100.871
β_1RMSE	0.090	0.084	0.083	0.089
Scenario 2: $\sigma^2_{B,x}$= $\sigma^2_{W,x}$=1; $\bar{\alpha}_g$~N(0,0.2^2)				
β_1	0.798	0.798	0.798	0.798
β_1_SE[1]	0.082	0.084	0.082	0.081
β_1_BIAS	-0.002	-0.002	-0.002	-0.002
β_1_CONF[2]	98.469	97.757	98.789	99.436
β_1_RMSE	0.082	0.084	0.082	0.082
Scenario 3: ($\sigma^2_{B,x}$, $\sigma^2_{W,x}$)=(0.5^2, 0.2^2); $\bar{\alpha}_g$~N(0,0.5^2)				
β_1	0.791	0.783	0.789	0.791
β_1_SE[1]	0.181	0.419	0.265	0.162
β_1_BIAS	-0.009	-0.017	-0.011	-0.009
β_1_CONF[2]	168.135	102.474	103.153	185.097
β_1_RMSE	0.181	0.420	0.266	0.162

[1] Standard errors

[2] Overconfidence indicator

Table 4: Simulation results for dynamic models (scenario 1/2/3/4)

Model: $\bar{y}_{gt} = 0.2 * \bar{y}_{gt-1} + 0.8 * \bar{x}_{gt} + \bar{u}_{gt}$											
	OLS	FE	RE	PCSE	GMM		OLS	FE	RE	PCSE	GMM
Scenario 1: $\sigma^2_{B,x} = \sigma^2_{W,x} = 1$; $\overline{\alpha}_g \sim N(0,0.5^2)$											
λ_1	0.328	0.134	0.328	0.328	0.282	β_1	0.798	0.795	0.798	0.798	0.790
λ_1_SE	0.065	0.069	0.065	0.085	0.249	β_1_SE	0.093	0.088	0.093	0.091	0.106
λ_1_BIAS	0.128	-0.066	0.128	0.128	0.082	β_1_BIAS	-0.002	-0.005	-0.002	-0.002	-0.010
λ_1_CONF	127.425	94.475	127.503	96.520	73.281	β_1_CONF	95.755	96.530	95.753	97.339	88.312
λ_1_RMSE	0.143	0.096	0.143	0.154	0.262	β_1_RMSE	0.093	0.089	0.093	0.091	0.107
Scenario 2: $\sigma^2_{B,x} = \sigma^2_{W,x} = 1$; $\overline{\alpha}_g \sim N(0,0.2^2)$											
λ_1	0.219	0.134	0.219	0.219	0.215	β_1	0.799	0.795	0.799	0.799	0.787
λ_1_SE	0.065	0.069	0.065	0.081	0.242	β_1_SE	0.086	0.088	0.086	0.084	0.103
λ_1_BIAS	0.019	-0.066	0.019	0.019	0.015	β_1_BIAS	-0.001	-0.005	-0.001	-0.001	-0.013
λ_1_CONF	106.036	100.802	106.391	84.361	79.104	β_1_CONF	99.623	99.856	99.617	101.635	90.521
λ_1_RMSE	0.068	0.095	0.068	0.084	0.243	β_1_RMSE	0.086	0.088	0.086	0.084	0.104
Scenario 3: $(\sigma^2_{B,x}, \sigma^2_{W,x}) = (0.5^2, 0.2^2)$; $\overline{\alpha}_g \sim N(0,0.5^2)$											
λ_1	0.366	0.098	0.365	0.366	0.319	β_1	0.659	0.816	0.660	0.659	0.688
λ_1_SE	0.077	0.087	0.077	0.109	0.293	β_1_SE	0.198	0.440	0.198	0.197	0.365
λ_1_BIAS	0.166	-0.102	0.165	0.166	0.119	β_1_BIAS	-0.141	0.016	-0.140	-0.141	-0.112
λ_1_CONF	130.307	96.679	131.519	92.137	78.510	β_1_CONF	140.201	102.292	139.962	138.797	86.829
λ_1_RMSE	0.183	0.134	0.183	0.199	0.316	β_1_RMSE	0.243	0.441	0.243	0.243	0.382
Scenario 4: $(\sigma^2_{B,x}, \sigma^2_{W,x}) = (0.5^2, 0.2^2)$; $\overline{\alpha}_g \sim N(0,0.2^2)$											
λ_1	0.221	0.097	0.221	0.221	0.208	β_1	0.778	0.790	0.778	0.778	0.782
λ_1_SE	0.081	0.087	0.081	0.111	0.303	β_1_SE	0.187	0.441	0.187	0.187	0.325
λ_1_BIAS	0.021	-0.103	0.021	0.021	0.008	β_1_BIAS	-0.022	-0.010	-0.022	-0.022	-0.018
λ_1_CONF	104.494	96.374	104.557	75.845	79.011	β_1_CONF	114.732	97.619	114.623	113.817	81.836
λ_1_RMSE	0.083	0.135	0.083	0.113	0.303	β_1_RMSE	0.188	0.441	0.188	0.188	0.326

Table 5: Simulation results for dynamic models (scenario 3/5/6)

Model: $\bar{y}_{gt} = 0.2 * \bar{y}_{gt-1} + 0.8 * \bar{x}_{gt} + \bar{u}_{gt}$											
	OLS	FE	RE	PCSE	GMM		OLS	FE	RE	PCSE	GMM
Scenario 3: $(\sigma^2_{B,x}, \sigma^2_{W,x})$=(0.5^2, 0.2^2); $\overline{\alpha}_g$~N(0,0.5^2)											
λ_1	0.366	0.098	0.365	0.366	0.319	β_1	0.659	0.816	0.660	0.659	0.688
λ_1_SE	0.077	0.087	0.077	0.109	0.293	β_1_SE	0.198	0.440	0.198	0.197	0.365
λ_1_BIAS	0.166	-0.102	0.165	0.166	0.119	β_1_BIAS	-0.141	0.016	-0.140	-0.141	-0.112
λ_1_CONF	130.307	96.679	131.519	92.137	78.510	β_1_CONF	140.201	102.292	139.962	138.797	86.829
λ_1_RMSE	0.183	0.134	0.183	0.199	0.316	β_1_RMSE	0.243	0.441	0.243	0.243	0.382
Scenario 5: $(\sigma^2_{B,x}, \sigma^2_{W,x})$=(0.2^2, 0.5^2); $\overline{\alpha}_g$~N(0,0.5^2); G=12; n_c~N(150, 50^2)											
λ_1	0.364	0.108	0.364	0.364	0.301	β_1	0.771	0.788	0.771	0.771	0.785
λ_1_SE	0.073	0.082	0.073	0.102	0.275	β_1_SE	0.173	0.176	0.173	0.170	0.201
λ_1_BIAS	0.164	-0.092	0.164	0.164	0.101	β_1_BIAS	-0.029	-0.012	-0.029	-0.029	-0.015
λ_1_CONF	132.126	95.155	133.168	95.122	79.092	β_1_CONF	106.426	100.119	106.556	108.047	90.671
λ_1_RMSE	0.180	0.123	0.179	0.193	0.293	β_1_RMSE	0.175	0.177	0.175	0.172	0.202
Scenario 6: $(\sigma^2_{B,x}, \sigma^2_{W,x})$=(0.5^2, 0.2^2); $\overline{\alpha}_g$~N(0,0.5^2); G=36; n_c~N(50, 15^2)											
λ_1	0.392	0.098	0.392	0.392	0.306	β_1	0.646	0.805	0.646	0.646	0.677
λ_1_SE	0.044	0.050	0.044	0.090	0.206	β_1_SE	0.108	0.252	0.108	0.124	0.215
λ_1_BIAS	0.192	-0.102	0.192	0.192	0.106	β_1_BIAS	-0.154	0.005	-0.154	-0.154	-0.123
λ_1_CONF	134.774	97.985	135.109	65.011	95.558	β_1_CONF	138.242	98.048	138.727	119.606	93.235
λ_1_RMSE	0.197	0.113	0.197	0.212	0.232	β_1_RMSE	0.188	0.252	0.188	0.198	0.248

Table 6: Estimation results for Sydney Household Travel Survey data

G=12; T=13; $\bar{n}_c$=143; σ_{n_c}=50 E($\bar{x}_{gt}$) = 1.72 $(\sigma^2_{B,x}, \sigma^2_{W,x})$=(0.50^2, 0.18^2)					
	OLS	FE	RE	PCSE	GMM
L.pttrip	0.881***	0.250**	0.881***	0.881***	0.895***
	(0.032)	(0.077)	(0.032)	(0.056)	(0.066)
price	-0.0122	-0.044	-0.0122	-0.0122	-0.0124
	(0.017)	(0.041)	(0.017)	(0.013)	(0.021)
G=20; T=13, $\bar{n}_c$=86; σ_{n_c}=38 E($\bar{x}_{gt}$) = 1.72 $(\sigma^2_{B,x}, \sigma^2_{W,x})$=(0.56^2, 0.21^2)					
L.pttrip	0.850***	0.258***	0.850***	0.850***	0.770***
	(0.031)	(0.066)	(0.031)	(0.050)	(0.127)
price	-0.011	-0.0283	-0.011	-0.011	-0.0206
	(0.015)	(0.037)	(0.015)	(0.009)	(0.018)

Note: Standard errors in parentheses; * p<0.05, ** p<0.01, *** p<0.001

Appendix 1: Notational Glossary

Variable/ parameter	Description	Units
α_i	Unobserved fixed individual effect in genuine panel data model	
$\bar{\alpha}_{gt}$	Combined unobserved cohort effect in pseudo panel data model	
$\bar{\alpha}_g$	Fixed group effect in pseudo panel data model	
β_0	Constant	
β_1	Parameter of exogenous variable	
$\tilde{\beta}$	Estimated parameter	
C	Total number of cohorts	
G	Number of created groups in pseudo panel data	
N	Number of panel units in genuine panel data model	
n_c	Cohort size (number of individuals in a cohort)	
$\bar{n}_c$	Average cohort size	
T	Number of time periods	
r	Replicate of simulation	
u_{it}	Composite error term in genuine panel data model	
$\bar{u}_{gt}$	Composite error term in pseudo panel data model	
x_{it}	Exogenous variable in genuine panel data model	
$\bar{x}_{gt}$	Exogenous variable in pseudo panel data model	
y_{it}	Dependent variable in genuine panel data model	
y_{it-1}	Lagged dependent variable in genuine panel data model	
$\bar{y}_{gt}$	Dependent variable in pseudo panel data model	
$\bar{y}_{gt-1}$	Lagged dependent variable in pseudo panel data model	
$\bar{y}_{gt-2}$	Second lagged dependent variable in pseudo panel data model	
ε_{it}	*i.i.d* error term in genuine panel data model	
$\bar{\varepsilon}_{gt}$	*i.i.d* error term in pseudo panel data model	
λ_1	Parameter of lagged dependent variable	
$\bar{\omega}_{gt}$	Time-varying cohort effect in pseudo panel data model	
σ_{α}^2	Variance of fixed group effect	
$\sigma_{n_c}^2$	Variance of cohort size	
$\sigma_{B,x}^2$	Between-group variance	
$\sigma_{W,x}^2$	Within-group variance	
pttrip	Average number of public transport trips per person	Trips
L.pttrip	Lagged variable of *pttrip*	Trips
Price	Average public transport price per trip	AU dollars