



# The estimation and interpretation of coefficients in panel gravity models of migration

Michael P. Cameron<sup>1,2</sup> · Jacques Poot<sup>2,3</sup>

Received: 17 May 2018 / Accepted: 14 November 2018 / Published online: 27 November 2018  
© The Author(s) 2018, corrected publication 2019

## Abstract

In this paper, we demonstrate that the conventional ordinary least squares and fixed effects estimators of classical gravity models of migration are biased, and that the interpretation of coefficients in the fixed effects gravity model is typically incorrect. We then present a best linear unbiased (BLU) estimator for gravity models of migration, and illustrate its application with inter-regional data from New Zealand. The results demonstrate that the standard ordinary least squares and fixed effect models lead to biased coefficients on population. Alternative estimates that are BLU are provided for a data generating process with fixed origin and destination effects. The coefficients on population must be interpreted in this model as growth rate effects rather than level effects. Our findings also have significance for other types of spatial interaction modelling.

**Keywords** Gross migration flows · Gravity model · Fixed effects · New Zealand

**JEL Classification** O15 · R23

## 1 Introduction

Greater availability of temporal gross flow data in many types of spatial interaction (trade, migration, tourism, etc.) in recent years has led to origin and destination fixed effects (FE) models becoming the standard approach for the estimation of gravity mod-

---

✉ Jacques Poot  
h.j.poot@vu.nl

<sup>1</sup> School of Accounting, Finance and Economics, University of Waikato, Private Bag 3105, Hamilton 3240, New Zealand

<sup>2</sup> National Institute of Demographic and Economic Analysis, University of Waikato, Private Bag 3105, Hamilton 3240, New Zealand

<sup>3</sup> Department of Spatial Economics, Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands

els, following Anderson and van Wincoop (2003). In a panel setting, the coefficients of the spatial interaction (dyadic) variables are estimable and consistent when origin and destination FE are interacted with period FE (e.g., Baldwin and Taglioni 2006), without recourse to the inclusion of systemic variables which are referred to as multi-lateral resistance terms in the trade literature (Anderson and van Wincoop 2003) and balancing factors in the migration literature (Alonso 1978). Moreover, when compared with ordinary least squares (OLS), FE models eliminate bias arising from standard error clustering when some variables apply only to the origin or to the destination, and not to both (Feenstra 2004). Peeters (2012) has critiqued the OLS and standard FE estimators for their dissimilarity in coefficient point estimates, but despite this critique and those of others (e.g., Etzo 2011), FE gravity models continue to be widely used (e.g., Lewer and Van den Berg 2008; Aldashev and Dietz 2014).

In this paper, we first demonstrate that both the conventional OLS estimator of the classical gravity model of migration and the conventional FE estimator are biased estimators of a spatial interaction data generating process (DGP) with origin and destination time invariant effects. We then show that comparisons of the population size coefficients between OLS and panel FE gravity models are not straightforward. We demonstrate that there should be no expectation that the signs and magnitudes of the coefficients of time-varying origin and destination variables on these models be similar. We then introduce a Best Linear Unbiased (BLU) Estimator for the FE DGP that can be used in gravity model specifications. We illustrate the issue of gravity model estimation with the example of five-yearly internal migration flows between regions of New Zealand over the period 1991–2013.

## 2 Model specifications

We commence with the classical gravity model, in which migration is positively related to the population of origin and destination areas, and inversely related to some measure of distance between them. Although this model can be augmented to include additional factors that differ among regions and over time (Lewer and Van den Berg 2008), we assume the classical specification that is expressed in log-linear form and supplemented with origin and destination fixed effects:

$$\ln M_{ij,t} = \alpha + \beta_1 \ln P_{i,t} + \beta_2 \ln P_{j,t} + \beta_3 \ln D_{ij,t} + \gamma_i + \varphi_j + \varepsilon_{ij,t}; \quad i \neq j \quad (1)$$

where  $M_{ij,t}$  is the instantaneous flow or force of gross migration from area  $i$  (the origin) to area  $j$  (the destination) at time  $t$ ,  $i, j = 1, 2, \dots, R$ ,  $P_{i,t}$  and  $P_{j,t}$  the corresponding population stocks in areas  $i$  and  $j$  respectively,  $D_{ij,t}$  is the distance between  $i$  and  $j$ , and  $\gamma_i$  and  $\varphi_j$  are time-invariant origin and destination-specific fixed effects. The  $\beta_k$  ( $k=1, 2, 3$ ) and origin and destination fixed effects are to be estimated, and  $\varepsilon_{ij,t}$  is assumed to be a white noise error term. If data were available on instantaneous flows of migration, the BLU estimator of Eq. (1) is OLS with  $2(R - 1)$  binary dummy variables representing the origin and destination fixed effects. While it is becoming increasingly clear that assumptions of spatially correlated and heteroscedastic errors

are more plausible for this DGP than white noise (e.g., Beenstock and Felsenstein 2016), the use of OLS does not even in that case render the estimates inconsistent and biased. We will therefore not pursue this issue further, but instead adopt robust estimation throughout, while also allowing for a spatial contiguity effect. However, considering explicitly the discrete time interval  $t \rightarrow (t + \delta)$  over which the migration flow is measured, Eq. (1) becomes:

$$\ln M_{ij,t \rightarrow (t+\delta)} = \alpha + \beta_1 \ln P_{i,t \rightarrow (t+\delta)} + \beta_2 \ln P_{j,t \rightarrow (t+\delta)} + \beta_3 \ln D_{ij,t \rightarrow (t+\delta)} + \gamma_i + \varphi_j + \varepsilon_{ij,t \rightarrow (t+\delta)}; \quad i \neq j \quad (2)$$

Hence, Eq. (1) is the cross-sectional limit case, i.e.  $\delta \rightarrow 0$ , of Eq. (2). Now assume, realistically, that the populations of areas  $i$  and  $j$  are growing at time-varying growth rates. Additionally, we assume that  $D_{ij,t \rightarrow (t+\delta)}$  is time-invariant ( $D_{ij}$ ). A major source of bias now arises from estimating regression Eq. (2) with cross-sectional or pooled ( $t=1,2,\dots,T$ ) migration flow data  $M_{ij,t \rightarrow (t+1)}$ , in which each migration flow matrix corresponds to a relatively long period, while the populations  $P_{it}$  and  $P_{jt}$  are measured at the start of the period. Because the size of the origin and destination populations change over the course of the period, the estimated coefficients of Eq. (2) are biased estimates of Eq. (1). The degree of bias will be relatively small for annual time-steps, but larger for longer time-steps that are common in the migration literature. To reduce this bias, it is convenient to substitute the geometric average of the population over the time period, i.e.:

$$P_{k,t \rightarrow (t+1)} = \sqrt{P_{k,t} P_{k,(t+1)}}; k \in \{i, j\} \quad (3)$$

Now, let  $g_{k,t \rightarrow (t+1)}$  refer to the time-varying growth rate of population  $k$ . Then:

$$P_{k,(t+1)} = P_{k,t}(1 + g_{k,t \rightarrow (t+1)}); k \in \{i, j\} \quad (4)$$

Substituting (4) into (3) and then (3) into (2) and using that for small  $x$ ,  $\ln(1 + x) \approx x$ , leads to:

$$\ln M_{ij,t \rightarrow (t+1)} = \alpha + \beta_1 \ln P_{i,t} + \beta_2 \ln P_{j,t} + \beta_3 \ln D_{ij} + \gamma_i + \frac{1}{2}\beta_1 g_{i,t \rightarrow (t+1)} + \varphi_j + \frac{1}{2}\beta_2 g_{j,t \rightarrow (t+1)} + \varepsilon_{ij,t \rightarrow (t+1)}; \quad i \neq j \quad (5)$$

We can now compare this with two common estimators of the classical gravity model. The first is the OLS estimator, which assumes the specification:

$$\ln M_{ij,t \rightarrow (t+1)} = \alpha^* + \beta_1^* \ln P_{i,t} + \beta_2^* \ln P_{j,t} + \beta_3^* \ln D_{ij} + \varepsilon_{ij,t \rightarrow (t+1)}^* \quad (6)$$

and the second is the FE estimator which assumes the specification:

$$\ln M_{ij,t \rightarrow (t+1)} = \alpha' + \beta_1' \ln P_{i,t} + \beta_2' \ln P_{j,t} + \beta_3' \ln D_{ij} + \gamma_i' + \varphi_j' + \varepsilon'_{ij,t \rightarrow (t+1)} \quad (7)$$

It is clear that both estimators are biased estimators of the parameters of the “true” DGP described by Eq. (5). Given that  $\varepsilon_{ij,t \rightarrow (t+1)}$ , the geometric average of  $\varepsilon_{ij,t}$  and  $\varepsilon_{ij,t+1}$ , is also a white noise error term (in the assumed absence of temporal autocorrelation), it is straightforward to directly estimate Eq. (5) by restricted least squares (RLS) or by replacing the beginning-of-period log populations,  $\ln P_{k,t}; k \in \{i, j\}$ , by the estimates of the log of the midpoint populations,  $\ln P_{k,t} + \frac{1}{2}g_{k,t \rightarrow (t+1)}; k \in \{i, j\}$ , and applying OLS. This simply requires the assumption that the time-varying population growth rates are exogenous. This assumption is plausible because there is no correlation between gross migration *levels* and net migration or population growth *rates*, even though gross inward (outward) migration is positively (negatively) correlated with net migration, see Vias (2001). Given the assumption of white noise errors, the RLS or OLS estimators of (5) are BLU (Greene 2017).

The (log) population at any point in time is a function of the initial population and the population growth rate:

$$\ln P_{k,t} = \ln P_{k,0} + \sum_{\tau=0}^{(t-1)} \ln(1 + g_{k,\tau \rightarrow (\tau+1)}); \quad k \in \{i, j\} \quad (8)$$

Substituting (8) into (5) gives:

$$\begin{aligned} \ln M_{ij,t \rightarrow (t+1)} = & \alpha + \beta_1 \ln P_{i,0} + \beta_2 \ln P_{j,0} + \beta_3 \ln D_{ij} + \gamma_i \\ & + \beta_1 \left[ \sum_{\tau=0}^{(t-1)} g_{i,\tau \rightarrow (\tau+1)} + \frac{1}{2}g_{i,t \rightarrow (t+1)} \right] \\ & + \varphi_j + \beta_2 \left[ \sum_{\tau=0}^{(t-1)} g_{j,t \rightarrow (t+1)} + \frac{1}{2}g_{j,t \rightarrow (t+1)} \right] + \varepsilon_{ij,t \rightarrow (t+1)}; \quad i \neq j \end{aligned} \quad (9)$$

Since  $P_{i,0}$  and  $P_{j,0}$  in Eq. (9) are fixed and time-invariant, they form part of the fixed effects in the regression. Therefore, the coefficients  $\beta_1$  and  $\beta_2$  in Eq. (9) should be interpreted as relating to the effect of the population growth rates  $g_i$  and  $g_j$ . Because the choice of the base year is arbitrary, this shows that the effect of population scale in Eq. (7) is also subsumed in the fixed effects and the estimated coefficients  $\beta'_1$  and  $\beta'_2$  reflect the impact of the population *growth rates*. This is in contrast with the OLS model of Eq. (6) where the corresponding coefficients relate to the effect of population *levels*. This highlights that the coefficients in the OLS model are *not* directly comparable to the coefficients from a FE model.

In the simple OLS specification of the gravity model [Eq. (6)], the expected sign on the coefficients  $\beta_1^*$  and  $\beta_2^*$  is positive. However, in the FE specification in Eqs. (5), (7) and (9), there is no a priori reason to believe that the estimates of coefficients of  $\beta_1$  and  $\beta_2$ , or of  $\beta'_1$  and  $\beta'_2$ , should both be positive. For a given propensity to migrate (e.g. linked to the age structure of the population), faster population growth will imply higher outward migration levels, i.e. a positive estimate of  $\beta_1$  and  $\beta'_1$ . This is mostly a demographic effect. On the other hand, faster population growth may not necessarily

imply relatively higher inward migration levels. The outcome would depend on a range of economic factors such as job growth, resource constraints and the corresponding prices (particularly of housing), and on the source of the population growth (natural increase, international migration, etc.). Consequently, the estimates of  $\beta_2$  and of  $\beta'_2$  are not necessarily positive.

### 3 Results

We test these ideas by means of data on migration flows between the sixteen regions of New Zealand. Inter-regional migration data were obtained from the Census of Population and Dwellings (1996, 2001, 2006 and 2013), based on self-reported region of residence 5 years previously. Responses that were unidentifiable or not elsewhere classified were distributed proportional to valid responses (including non-movers), while the few zero-count flows were increased by one (for estimation in the case where some flows are structurally zero, see e.g. Head and Mayer, 2014). This provides 960 observations of the dependent variable. Population numbers were taken from the Estimated Usually Resident Population at 30 June in each year. Distances between pair of regions were population-weighted straight line distances (based on the 2013 population distribution). Additional dummy variable controls were included for contiguity of regions, and for flows between the two main islands of New Zealand. We implement BLU, OLS, and FE regressions, equivalent to Eqs. (5), (6), and (7) respectively, as well as FE re-specified in growth rates as in Eq. (9).

Table 1 presents the results of our regression models, excluding control variables and fixed effects. Column (1) contains the results for the OLS specification, equivalent to Eq. (6), Column (2) is the FE specification of Eq. (7), Column (3) corresponds to the growth rates specification of Eq. (9), and Column (4) reports estimates of Eq. (5), which are BLU given the assumed DGP.

All coefficients in Table 1 are highly statistically significant. In terms of goodness of fit, the table reiterates the common finding in the literature that even when considering the population as the only driver of migration, these models still yield a relatively good description of migration patterns (see also Poot et al. 2016). However, the fit of the FE models is clearly superior, with a mean absolute percentage error (MAPE) of around 27%. It can be shown that time varying origin and destination FE can reduce the MAPE further to 24.2%, but the coefficients on population levels or growth rates are then of course no longer identified (this regression is therefore not included in Table 1). However, comparing the coefficients between the OLS (1) and FE (2) models highlights one substantial difference—the coefficient on the destination population is positive and statistically significant in the OLS specification, but negative and statistically significant in the FE specification. This change in coefficients could be construed as demonstrating a lack of robustness in the estimates. However, based on the exposition of our specification earlier, it is clear that the coefficients on population variables cannot be compared directly. The negative sign on destination population in the FE specification simply suggests that higher population *growth rates* in the destination are associated with smaller migration inflows. As described above, our results are consistent with inward migration being constrained, perhaps by unavailability of a

**Table 1** Regression results

Model	(1) OLS	(2) FE	(3) FE (growth rates)	(4) BLU (RLS)
$\ln P_i$	0.823*** (0.016)	0.929*** (0.183)	–	0.974*** (0.206)
$\ln P_j$	0.806*** (0.017)	–0.859*** (0.191)	–	–0.782*** (0.214)
$g_i$	–	–	1.051*** (0.228)	0.487*** (0.103)
$g_j$	–	–	–0.824*** (0.228)	–0.391*** (0.114)
$\ln D_{ij}$	–0.520*** (0.035)	–0.782*** (0.037)	–0.782*** (0.033)	–0.782*** (0.038)
R <sup>2</sup>	0.887	0.948	0.948	0.948
RMSE	0.488	0.336	0.337	0.337
MAPE (%)	42.6	26.8	27.0	27.0

N.B.: Robust standard errors in parentheses;  $n=960$ ; \*\*\* $p<0.01$ ; \*\* $p<0.05$ ; \* $p<0.1$

suitable quantity of affordable housing for migrants, or competition between internal and international in-migrants for the housing that is available. Moreover, this result is consistent with the model in Column (3), where the model is re-specified in growth rates. The results in column (4) for the BLU estimator demonstrate the degree of bias in the OLS and FE estimators for the DGP in Eq. (5). The coefficients on the origin population in columns (1) and (2), and on the destination population in column (2), are biased downwards.

## 4 Conclusion

Peeters (2012) critiqued the OLS and FE estimators for their dissimilarity in the point estimates of the population coefficients. However, as demonstrated in this paper, there is no a priori reason to expect these coefficients to hold the same sign. The coefficients in OLS and FE models must be interpreted differently. This misinterpretation is relatively common. For instance, the unexpected sign on employment in Aldashev and Dietz (2014), and the change in coefficient signs between OLS and FE models in Ramos and Surinach (2017) can be explained in this way. Moreover, we have shown that the standard OLS and FE models lead to biased coefficients on population compared with a BLU estimator of a DGP with fixed effects that correctly account for population change. These results have significance not only for the estimation and interpretation of the coefficients of gravity models in the migration literature, but also in the literature on trade and on other forms of spatial interaction.

**Acknowledgements** This research was funded by the Ministry of Business, Innovation and Employment (Grant Nos. UOWX1404 and MAUX1206) and by the Royal Society of New Zealand Marsden Fund (Grant No. UOW1302). The authors are grateful to Sialupapu Siameja, Tobias Dean, and Liam Bullen for excellent research assistance. The usual disclaimer applies.

**Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

## References

- Aldashev, A., Dietz, B.: Economic and spatial determinants of interregional migration in Kazakhstan. *Econ. Syst.* **38**(3), 379–396 (2014)
- Alonso, W.: A theory of movements. In: Hansen, N.M. (ed.) *Human Settlement Systems*, pp. 197–211. Ballinger, Cambridge, MA (1978)
- Anderson, J.E., van Wincoop, E.: Gravity with gravitas: a solution to the border puzzle. *Am. Econ. Rev.* **93**(1), 170–192 (2003)
- Baldwin, R., Taglioni, D.: Gravity for dummies and dummies for gravity equations. NBER Working Paper No. 12516. National Bureau of Economic Research, Cambridge, MA (2006)
- Beenstock, M., Felsenstein, D.: Double spatial dependence in gravity models; migration from the European Neighborhood to the European Union. In: Patuelli, R., Arbia, G. (eds.) *Spatial Econometric Interaction Modelling*. Springer, Berlin (2016)
- Etzo, I.: The determinants of the recent interregional migration flows in Italy: a panel data analysis. *J. Reg. Sci.* **51**(5), 946–966 (2011)
- Feenstra, R.C.: *Advanced International Trade: Theory and Evidence*. Princeton University Press, Princeton, NJ (2004)
- Greene, W.H.: *Econometric Analysis*, 8th edn. Pearson, New York (2017)
- Head, K., Mayer, T.: Gravity equations: workhorse, toolkit, and cookbook. In: Gopinath, G., Helpman, E., Rogoff, K.S. (eds.) *Handbook of International Economics*, vol. 4, pp. 131–195. Elsevier, Oxford (2014)
- Lewer, J.J., Van den Berg, H.: A gravity model of immigration. *Econ. Lett.* **99**, 164–167 (2008)
- Peeters, L.: Gravity and spatial structure: the case of interstate migration in Mexico. *J. Reg. Sci.* **52**(5), 829–856 (2012)
- Poot, J., Alimi, O., Cameron, M.P., Maré, D.C.: The gravity model of migration: the successful comeback of an ageing superstar in regional science. *Investig. Reg. J. Reg. Sci.* **36**, 63–86 (2016)
- Ramos, R., Surinach, J.: A gravity model of migration between ENC and EU. *Tijdschrift voor Economische en Sociale Geografie* **108**(1), 21–35 (2017)
- Vias, A.C.: Reevaluating the relationship between in-, out-, and net migration for non-metropolitan counties: an update on Beale's u-shaped curve. *Geogr. Anal.* **33**(3), 228–246 (2001)

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.