2019, Volume 42, Issue 84, 1-8 / ISSN 2304-4306

E C O N O M Í A

revistas.pucp.edu.pe/economia

www.fondoeditorial.pucp.edu.pe

Probit Models for Grouped-data Migration Flows: A Theoretical Note

Coro Chasco[a, b,*], Patricio Aroca[c], Luc Anselin[d]

[a]Autonomous University of Madrid

[b]Nebrija University

✉ coro.chasco@uam.es [*]Corresponding author

[c]Escuela de Negocios, Adolfo Ibáñez University, Chile

✉ patricio.aroca@uai.cl

[d]University of Chicago

✉ anselin@uchicago.edu

Abstract

In this theoretical note, we propose the GProbit model as an alternative to gravity models to estimate grouped-data flows. This is a model based on the random utility theory, which is consistent with the principle of population behavior. Instead of migrant counts, the dependent variable of the GProbit model of flows consists of a number of observed proportions. It allows explaining the propensity to migrate from any origin to a destination, which is an interesting relative concept not affected by the size effect. For this reason, it is expected to have better fit and less problems of non-normality, as illustrated by an application for the internal migration flows of the Spanish regions.

Article History: Received: February 28 2019 / Revised: May 20 2019 / Accepted: May 25 2019
Keywords: Probit model; Gravity model; Proportions; Migration flows; Spain.
JEL Classification: C25, C31, C51, R23

Acknowledgements

This work was supported by Spanish Ministry of Economics and Competitiveness (ECO2015-65758-P). We also thank to Editor Gabriel Rodríguez and an anonymous Referee for useful comments in order to improve the paper.

https://doi.org/10.18800/economia.201902.001

1. Introduction

In many areas of economics, choices made by individuals are costly to collect or inaccessible. However, analysts may have access to the choice data aggregated across groups of individuals in the form of counts or shares. Regression-based spatial interaction macro-models have frequently been used to estimate group-data choices. This is the case of many applications of gravity models to migration or other spatial interaction processes (LeSage and Fischer, 2010). In these models, the (aggregated) observations are treated *as if* they were single entities. They specify the dependent variable as mere (log-transformed) aggregations of individual data, which can produce—among others—severe problems of non-normality and heteroscedasticity, enhancing spatial autocorrelation in the error terms. Moreover, a simple aggregation of individual choices does not necessarily lead to grouped or "herd" behavior (Schelling, 2006; Sen and Smith, 2012). Aggregation must result in models consistent with theory, which should be capable of identifying overall regularities in collective population behavior (Kanaroglou et al., 1996).

For this reason, we recommend following a different strand of the literature based on a choice-theoretic perspective. Although typically concerned with the identification of individual behavior, choice models have also been specified for grouped data when observations no longer consist of single individuals but sets of several persons who share similar characteristics (e.g. living in the same region). In these grouped-data choice models, the dependent variable consists of a number of observed proportions or relative frequencies (Gourieroux, 2000), which are estimated by a nonlinear weighted least squares method (Berkson, 1944, 1955, 1957; Amemiya, 1985). Grouped-data choice models can be easily generalized to spatial interaction models of migration or trade flows, as in Borjas (2006) and Aroca and Hewings (2002).

In this note, we briefly review the specification and estimation methods of the standard probit model for grouped-data flows.

2. From Individual to Grouped-data Choice Models

The random utility theory provides the framework to deal with individuals' decision (McFadden, 2001). Let be the U_{od} the utility that an individual gets from moving from region o to d, where o is the origin region while d represents any of the potential destination regions. Therefore, an individual will move from region o to d if $U_{od} \geq U_{oo}$, that is, when the utility of moving (U_{od}) is more profitable for this individual than the utility of staying (U_{oo}). The utility function of moving (U_{od}) has a non-stochastic part (V_{od}) and a random error term (ε_{od}):

$$U_{od} = V_{od} + \varepsilon_{od}. \tag{1}$$

This model stands that:

$$y^* = x'\theta + \varepsilon, \tag{2}$$

where $y^* = U_{od} - U_{oo}$ is a latent variable for which there is not a direct measure, but an indicator (y) that takes the value 1 if the individual moves or 0 when that individual decides to stay, conditional to a set of variables x which explain the migration decision. In **probabilistic terms**,

it goes like:

$$P(y=1) = P(y^* \geq 0) = P(U_{od} \geq U_{oo}) = P(V_{od} + \varepsilon_{od} \geq V_{oo} + \varepsilon_{oo}) \\ = P(\varepsilon_{oo} - \varepsilon_{od} \leq V_{od} - V_{oo}), \tag{3}$$

where the non-stochastic part in the indirect utility function (V_{od}), which is generally assumed to be linear and can be estimated by maximum likelihood (ML).

Choice data can be aggregated across groups of individuals in the form of counts or shares. Differently from the individual setup presented above, here we use the behavior of the whole population. Grouped data are obtained by observing the response of the individuals belonging to the same region or 'group' provided that they can share similar characteristics (e.g. spatial location, age, income class, etc.). In a theoretic model with no individual (spatial) interaction, adding up the independent probabilities for all the individuals who move from region o to j, will give the probability that a generic individual of region o ends up in d. This definition might change slightly depending upon the denominator of the share, which can be the total population in the origin region at the beginning of the period or the number of migrants departing from the origin region[1].

Assuming that each of the group components is large, by the law of the large numbers it can be concluded that the observed proportion (P) is close—or an estimation of—the population or theoretical proportion (π). Hence, we can treat this problem as a simple one of sampling from a Bernoulli population, in which the observed proportion is equal to the population proportion plus an error term (ε_{od}):

$$P_{od} = \pi_{od} + \varepsilon_{od}, \tag{4}$$

where the dependent variable consists of the n number of observed proportions of people moving from an origin o to a destination region d (M_{od}) over the total group of migrants moving out from o (M_o), that is $P_{od} = M_{od}/M_o$. By the central limit theorem, the error term ε_{od} is approximately normally distributed with $E(\varepsilon_{od}) = 0$; $\text{Var}(\varepsilon_{od}) = [\pi_{od}(1-\pi_{od})/M_o]$, being n_o the total number of migrants in region o.

3. The Probit Model of Migration Flows

The population proportion can be expressed as an indirect utility function, $\pi_{od} = F\left(x'_{od}\beta\right)$, for x_{od} a vector gathering a set of k factors which explains the migration decision and β contains a set of parameters. One of the functional forms most frequently used in application for F is the probit model, which by means of the Slutsky's theorem on convergence in probability, can be linearized (Gourieroux, 2000, section 4.2). The Cumulative Distribution Function (CDF) of the standard normal distribution is expressed as $\Phi\left(x'_{od}\beta\right)$. Since the CDF is strictly monotonic, it has an inverse form, $Z_{od} = \Phi^{-1}(P_{od})$, which by means of a Taylor series approximation for $\varepsilon_{od} = 0 \rightarrow P_{od} = \pi_{od}$ (Greene, 2003, section 21.4.6), leads to the probit model for grouped-data flows or "GProbit model of flows":

[1] From now on, we will consider the total number of migrants departing from the origin region as the denominator of this share.

$$Z_{od} = (X_d/X_o)\,\beta + u_{od}, \tag{5}$$

where $X_d/X_o = x_{od}$, being X_o and X_d the characteristics of spatial units o and d, respectively, and u_{od} is the error term.

Since the number of migrants moving from each origin region (M_o) is large, the random variable of the GProbit model of flows, u_{od}, is approximately normally distributed with $E[\,u_{od}|\,x_{od}] = 0$ and non-constant variance defined as:

$$\sigma^2_{od} = \frac{P_{od}\,(1 - P_{od})}{M_o \cdot [\phi\,[\Phi^{-1}\,(P_{od})]]^2}, \tag{6}$$

where ϕ is the Probability Distribution Function (PDF) of the standard normal distribution. Therefore, the GProbit model of flows is heteroskedastic by construction due to the different values adopted by the denominator of the ratio $P_{od} = M_{od}/M_o$, which is the flow rate of people living in an origin o who move to any destination d, including intra-regional flows, M_{oo} (for $d = o$).

Berkson (1944, 1955, 1957) proposed a simpler way to estimate grouped-data choice models by nonlinear Weighted Least Squares (WLS), which is a variation—for qualitative response models—of the MCSE or MIN χ^2 test of goodness of fit proposed in the literature (Amemiya, 1985, section 9.2.5 and 9.2.6). This method, which can also be applied to the GProbit model of flows, consists of finding parameter values minimizing a measure of the distance between the observed proportions (P_{od}) and the theoretical ones (π_{od}). It is solved in a two-step procedure because the weights are functions of the unknown parameters:

1. In the first step, the β parameters are estimated by Ordinary Least Squares (OLS), which produces consistent but inefficient estimates. This step provides the estimations of the dependent variable $\widehat{\Phi^{-1}(P_{od})} = \hat{Z}_{od}$ and the error variances, $\hat{\sigma}^2_{od}$:

$$\hat{\sigma}^2_{od} = \frac{\hat{P}_{od}\left(1 - \hat{P}_{od}\right)}{M_o \cdot \left[\phi\left(\hat{Z}_{od}\right)\right]^2} \tag{7}$$

2. In the second step, the estimated variances based on the first-step estimates $\hat{\sigma}^2_{od}$ are used as weights for the WLS. The MIN χ^2 estimator $\tilde{\beta}$ is defined as:

$$\tilde{\beta} = \left[\sum_{o=1}^{N}\sum_{d=1}^{N} \hat{\sigma}^{-2}_{od} x'_{od} x_{od}\right]^{-1} \sum_{o=1}^{N}\sum_{d=1}^{N} \hat{\sigma}^{-2}_{od} x'_{od} Z_{od}. \tag{8}$$

Hence, the Berkson's probit model of grouped-data flows can be expressed as follows:

$$Z^*_{od} = x^*_{od}\beta + u^*_{od}, \tag{9}$$

where $Z^*_{od} = Z_{od}/\hat{\sigma}_{od}$, $x^*_{od} = x_{od}/\hat{\sigma}_{od}$, and $u^*_{od} = u_{od}/\hat{\sigma}_{od}$.

However, since any other forms of heteroskedasticity are usually present in the error terms of spatial cross-section model, e.g. spatial group-wise heteroskedasticity (Chasco et al., 2018), the basic GProbit model of expression (5) should be estimated by OLS with a robust inference on the parameters (Anselin and Rey, 2014). Models (5) and (9) are linear models that can be estimated efficiently by standard methods like Ordinary Least Squares (OLS), Maximum Likelihood (ML) or whatever others.

4. An Empirical Illustration for Interregional Flows in Spain

We illustrate the performance of a GProbit model to estimate internal migration flows for the 17 NUTS 2 regions ("Autonomous Communities") in Spain, taken from the EVR register ("Estadística de Variaciones Residenciales") of the Spanish National Statistics Office (INE). Flows were constructed as the rate of emigrants moving from an origin region o to a destination region d over the total people of region o who have changed their residence during this period (including the intra-regional movements). We compare the performance and results of this model with the gravitational model using the conventional log transformation of flows for the dependent variable.

The distance matrix was formed using the log-transformed distance between the capital cities of the Spanish regions. We use six additional explanatory variables, which are the most significant from a set of more than 60 classical 'push' and 'pull' factors. They are: population, R&D expenditure per capita, average altitude, annual maximum temperature and annual atmospheric precipitation. All of them were defined as the ratio of the destination over the origin (D/O) values. We would expect a priori that flows are directly proportional to the D/O ratios of population and R&D expenditure and inversely proportional to the D/O ratios of housing price, altitude, maximum temperature and atmospheric precipitation. Data has been ordered according to the origin-centric scheme described by LeSage and Pace (2009) and Fischer and Wang (2011).

Ordinary least-squares (OLS) estimates are shown in Table 1. Regressions (1) and (2) model the interregional flows differently specified by the GProbit and gravity models, as observed proportions of people moving from o to d, M_{od}/M_o, and migrant counts M_{od}, respectively. Since the total group of migrants moving out from an origin region is the sum of the interregional plus intra-regional flows departing from this region $M_o = M_{od} + M_{oo}$, the GProbit model (2) allows estimating intra-regional migration rates directly as $M_{oo}/M_o = 1 - \sum_{d=1}^{n-1} M_{od}/M_o$. Hence, these proportions can be interpreted as a probability or 'propensity to migrate'. As regards the gravity model, the intra-regional flows (M_{oo}) must be estimated in a separate model with different explanatory variables due to the different nature of inter- and intra-regional flows (LeSage and Pace, 2008). They are presented in Table 1, regression (3).

All the coefficients are very significant. However, the adjusted R^2 takes a very low value, particularly for the gravity model estimation, which is in line with other previous analysis in the literature. Spanish interregional migration has long been resistant to traditional economic explanations: none of the considerable research on Spanish internal migration finds clear significance in even core variables of income and employment (Mulhern and Watson, 2009). The strong rigidity of the Spanish labor market, centrally controlled by the trade unions, and a very high national unemployment discourages internal migration (Bover and Velilla, 1999) and instead promotes migration to other countries.

Traditional measures of prediction accuracy are also presented in Table 1. Besides the adjusted-R^2 of the OLS estimations, we also report some traditional measures of prediction accuracy for the estimated variable of proportions or propensity to migrate, $\hat{P}_{od} = \hat{M}_{od}/\hat{M}_o$. First, we show the results of a bias indicator (RBIAS), which is the absolute difference between the observed and predicted values, divided by the predicted values. Positive values are indicative of predicted

Table 1
Estimation results for the interregional migration models.

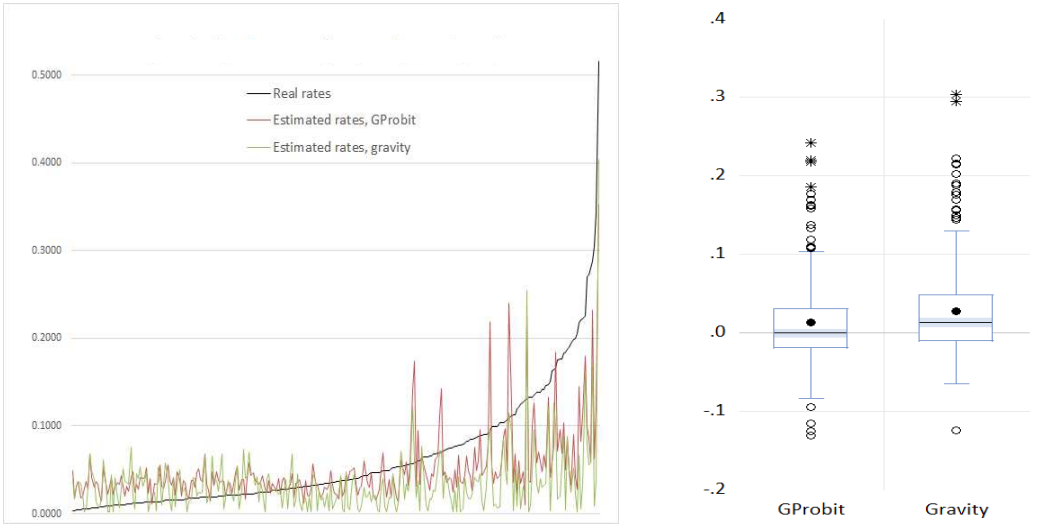
	GProbit model	Gravity model	
Dependent variable	$Z_{od} = \phi^{-1}(M_{od}/M_o)$	$\ln(M_{od})$	$\ln(M_{oo})$
	(1)	(2)	(3)
Constant	-1.820^{***}	7.088^{***}	13.044^{***}
Population D/O ratio	0.036^{***}	-	$0.4e^{-7***}$
Housing price D/O ratio	-	-0.481^{**}	-
R&D expenditure p.c. D/O ratio	0.073^{***}	0.137^{***}	-
Average altitude D/O ratop	-0.083^{***}	-0.245^{***}	-
Annual max. temperature D/O ratio	-	-	-0.088^{*}
Atmospheric precipitation D/O ratio	-0.081^{***}	-	-
O-D distance (log)	-0.158^{***}	-0.244^{**}	-
Adj. R-squared	0.312	0.094	0.847
Prediction accuracy measures for the propensity to migrate: $\hat{P}_{od} = \hat{M}_{od}/\hat{M}_o$:			
Bias indicator (RBIAS)	0.79	4.04	
Coeficient of variation (CV)	1.16	311.03	
Relative root mean sq. error (RRMSE)	0.16	0.35	

Note: A robust inference of the GPprobit model estimators have been computed.

overestimation, being zero the perfect situation of unbiasedness. Both models get positive values, though the gravity model has a RBIAS outcome (4.04) more than five times higher than the GProbit model (0.79).

Second, the coefficient of variation (CV) is a standardized measure of dispersion that is defined as the ratio of the standard deviation to the mean. In this context, it could be interpreted as a measure of efficiency of the estimates and homoskedasticity of the prediction errors. Hence, a completely efficient estimator will get a CV value of zero. As shown in Table 1, the GProbit estimation has a CV close to zero (1.16) and it is almost 300 times more efficient than the gravity model (311.03). Hence, the error terms are more homoscedastic for the GProbit than the gravity model estimation. Third, the relative root mean square error (RRMSE) constitutes a balance between bias and variability. It is the mean value of the square root of the squared difference between observed and predicted values, divided by the predicted values. Once again, zero is the best value and the GProbit model performs better (0.16) than the gravity model (0.35).

Figure 1 illustrates the results obtained by the prediction accuracy measures. The line graph with the real and estimated values of the flow rates shows that both models perform better in estimating flow rates closer to the average. However, they tend to overestimate lower rates while the higher ones are mainly underestimated. In fact, both models fail in estimating propensities to migrate above the average, particularly the gravity model. Additionally, the box plots for the difference between real and estimated migration rates show that this variable is closer to normality for the GProbit estimation, since it gets a mean and median values closer to zero, as well as a fewer upper outliers (more homoskedasticity) than the gravity model.



(a) Interregional migration rates (propensity to migrate).

(b) Difference between real and estimated rates.

Figure 1. Real, estimated and residual interregional flows, GProbit and gravity models.

5. Conclusions

The intent of this theoretical note is presenting an alternative to gravity models to model grouped-data flows of any kind (migration, transport, networks, etc.) based on the random utility theory. Logit and probit models for grouped-data are consistent with the theory of population behavior. Additionally, they have less problems of non-normality and heteroskedasticity, mainly because the dependent variable consists of a number of observed proportions (people moving from an origin to a destination region over the total group of migrants moving out from this origin) instead of migrant counts (as it is the case in the standard gravity models).

That is, the GProbit model of flows allows explaining the propensity to migrate from any origin to a destination, which is an interesting relative concept not affected by the size effect. Since it is a linear model, it can be expanded to include spatial autocorrelation[2] and heterogeneity effects[3], with some arrangements. This is something to be developed in a future work.

[2]Spatial autocorrelation arises when the aggregated flows from an origin to a destination are not independent from each other. As in the conventional spatial interaction model, the spatial GProbit model can adopt different specifications. For example, the spatial lag or SAR GProbit model can be expressed as follows: $Z_{od} = \rho_d W_d Z_{od} + \rho_o W_o Z_{od} + \rho_\omega W_\omega Z_{od} + \alpha\iota_N + X_d\beta_d + X_o\beta_o + \lambda D + \varepsilon_{od}$ and the spatial error GProbit model is $Z_{od} = \alpha\iota_N + X_d\beta_d + X_o\beta_o + \lambda D + \rho_d W_d u_{od} + \rho_o W_o u_{od} + \rho_\omega W_\omega u_{od} + \varepsilon_{od}$ for $W_d = I_n \otimes W$, $W_o = W \otimes I_n$, $W_\omega = W \otimes W$, n is the number of regions, W is the conventional (row-normalized) n-by-n spatial weight matrix, and ρ_o, ρ_d, ρ_ω are the spatial autoregressive parameters (LeSage and Pace, 2008).

[3]Unobservable heterogeneity could be modeled with a panel data Gprobit model as follows: $Z_{od,t} = \alpha_{o,t}\iota_{N\cdot t} + \alpha_{d,t}\iota_{N\cdot t} + X_d\beta_d + X_o\beta_o + \lambda D + \varepsilon_{od}$, for $t = 1, ..., T$, $\alpha_{o,t}$ origin-time specific factors and $\alpha_{d,t}$ destination-time specific factors.

References

Amemiya, T. (1985). *Advanced Econometrics*. Cambridge, Massachusetts (US): Harvard University Press.

Anselin, L., and Rey, S. (2014). *Modern Spatial Econometrics in Practice: A Guide to GeoDa, GeoDaSpace and PySAL*. Chicago, IL: Geoda Press LLC.

Aroca, P., and Hewings, G. (2002). Migration and regional labor market adjustment: Chile 1977-1982 and 1987-1992. *The Annals of Regional Science 36(2)*, 197-218.

Berkson, J. (1944). Application of the logistic function to bio-assay. *Journal of the American Statistical Association 39*, 357-365.

Berkson, J. (1955). Maximum Likelihood and Minimum χ^2 Estimates of the Logistic Function. *Journal of the American Statistical Association 50(269)*, 130-162.

Berkson, J. (1957). Tables for Use in Estimating the Normal Distribution Function by Normit Analysis: Part I. Description and Use of Tables Part II. Comparison Between Minimum Normit χ^2 Estimate and the Maximum Likelihood Estimate. *Biometrika 44(3/4)*, 411-435.

Borjas, G. (2006). Native Internal Migration and the Labor Market Impact of Immigration. *Journal of Human Resources 41(2)*, 221-258.

Bover, O., and Velilla, P. (1999). *Migrations in Spain: historical background and current trends*. Madrid: Banco de España.

Chasco, C., Le Gallo, J., and López, F. (2018). A scan test for spatial groupwise heteroscedasticity in cross-sectional models with an application on houses prices in Madrid. *Regional Science and Urban Economics 68*, 226-238.

Fischer, M., and Wang, J. (2011). *Spatial Data Analysis: Models, Methods and Techniques*. Springer Publishing Company.

Gourieroux, C. (2000). *Econometrics of Qualitative Dependent Variables*. Cambridge (UK): Cambridge University Press.

Greene, W. (2003). *Econometric Analysis*. Prentice Hall.

Kanaroglou, P., and Ferguson, M. (1996). Discrete Spatial Choice Models for Aggregate Destinations. *Journal of Regional Science 36(2)*, 271-290.

LeSage, J., and Fischer, M. (2010). Spatial Econometric Methods for Modeling Origin-Destination Flows. In M. Fischer and A. Getis (Eds.). *Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications*, pp. 409-433.

LeSage, J., and Pace, R. (2008). Spatial Econometric Modeling of Origin-Destination Flows. *Journal of Regional Science 48(5)*, 941-967.

LeSage, J., and Pace, R. (2009). *Introduction to Spatial Econometrics*. CRC Press.

McFadden, D. (2001). Economic Choices. *American Economic Review 91(3)*, 351-378.

Mulhern, A., and Watson, J. (2009). Spanish Internal Migration: Is there Anything New to Say? *Spatial Economic Analysis 4(1)*, 103-120.

Schelling, T. (2006). *Micromotives and Macrobehavior*. W. W. Norton & Company.

Sen, A., and Smith, T. (2012). *Gravity Models of Spatial Interaction Behavior*. Springer Science & Business Media.