University of Kent

School of Economics Discussion Papers

Spatial differencing for sample selection models

Alex Klein and Guy Tchuente

December 2016

KDPE 1701

Spatial differencing for sample selection models

Alex Klein, Guy Tchuente

School of Economics, University of Kent

This paper offers an identification strategy in the situation when researchers work with cross-sectional data, face unobserved heterogeneity causing endogeneity problem, lack instrumental variables and, on top of it, face sample selection problem. To accomplish that, we take advantage of recent advances of spatial econometrics. What motives us to consider the case of cross-sectional data which data generating process involves sample selection and seemingly unsolvable problem of endogeneity and no instrumental variables?

Recent decades have witnessed a rise of panel data sets which was accompanied by the proliferation of estimation techniques attempting to take advantage of the time and cross section dimension to identify the causal effect of regressors on the variables of interest. Similarly, considerable advances were made in the areas of weak instrumental variable estimation techniques and imperfect instruments. All of this offers researchers various identification strategies which help them to identify vast variety of empirical models even in the situations when strong instrumental variables are not available or exclusion restrictions would not necessarily hold. But what if panel data sets or instrumental variables are not readily available to researchers?

There are three broad possibilities. One is to dispense of causality claim and consider the regression results as sophisticated correlations. Second solution is offered by the literature identifying causal effect with higher moments. Third solution is spatial differencing in which empirical model takes advantage of the spatial dimension of the data to control for unobserved heterogeneity that might render estimator biased and inconsistent. Our paper contributes to that literature. Spatial differencing has been used only in the context of linear regressions so far. We extend this approach to cross-section data with sample-selection. Specifically, we offer a solution to the problem of differencing out spatial unobserved effects when nonlinear element - in our case Mill’s ratio - is present, propose estimation procedure, and derive formula for estimating standard errors.

Spatial Differencing for Sample Selection Models

Alex Klein[*] Guy Tchuente[†]

University of Kent

December 2016

Abstract

This paper uses spatial differencing to estimate parameters in sample selection models with unobserved heterogeneity. We show that under the assumption of smooth changes across the space of unobserved site-specific heterogeneity and selection probability, key parameters of a sample selection model are identified. A simple estimation procedure is proposed and the formula for the estimator of the standard error is derived.

Keywords: Sample selection, spatial difference, Instrumental variable.

JEL: C13, C31.

1 Introduction

Identification of causality in empirical models is very difficult when only cross-sectional data exist and no instrumental variables are readily available. Indeed, despite the increasing availability of panel data which allow to control for unobserved heterogeneity, panel data are still not always available to applied economists. Furthermore, despite the

[*] School of Economics, e-mail: A.Klein-474@kent.ac.uk

[†] Corresponding author: School of Economics, University of Kent, e-mail: g.tchuente@kent.ac.uk, Keynes College, University of Kent, Canterbury, Kent, CT2 7NP.

advances of weak instrumental variable and partial identification techniques, which allow to use imperfect instruments, there are many situations when no instrumental variables are available *at all.* This paper offers an identification strategy when researchers face exactly this situation and, in addition, need to estimate a sample selection model.

The literature is not completely silent what to do in these situations. One solution is offered by the literature identifying causal effect with higher moments e.g. (Lewbel (1997), Klein and Vella (2010), Lewbel (2012)). Another is due to spatial differencing literature in which researchers can take advantage of the spatial dimension of the cross-sectional data to control for unobserved heterogeneity, (e.g. Duranton, Gobillon, and Overman (2011), Black (1999) or Holmes (1998)). Our paper contributes to this literature by offering an identification strategy based on spatial differencing. So far, it was applied only in the context of linear regressions. We extend this approach to cross-section data with sample selection. We make two contributions. First, we provide a solution to the problem of differencing out spatial unobserved effects when a nonlinear element - in our case the Mill's ratio - is present. To our knowledge, this is the first paper which proposes a sample selection estimator in the context of spatial differencing. Second, we contribute to the literature which attempts to identify cross-sectional models with endogenous regressors when no instrumental variables are available. The paper uses standard two-step approach of Heckman (1974; 1979) and offers a correction of standard errors.

2 Sample Selection Models with Spatial Correlation

We are interested in estimating the regression equation:

$$y_{ij} = x'_{ij}\delta + \gamma_{ja} + \gamma_j + \varepsilon_{ij} \tag{1}$$

where x'_{ij} is a vector of controls, γ_j is location fixed effect, γ_{ja} is a site-specific effect for site a which is at a finer spatial scale than location j, and ε_{ij} is the error term.[1]

[1] The site-specific component, γ_{ja}, is a simplification for γ_{ja_i}. We are implicitly assuming that the site-specific effects are the same for all individual sharing the same site.

Examples include the effect of local soil quality on agricultural productivity, impact of local taxation on growth of firms, or the effect of local amenities on house prices. We can control for γ_j with location dummy variables. However, they might not be enough to capture all unobserved heterogeneity related to location j as there can be considerable heterogeneity across locations j at very fine spatial scale. Furthermore, the standard location fixed effect γ_j relies upon an arbitrary specification of the comparison neighborhood group, as pointed out by Gibbons and Machin (2003), making it an imperfect control for the site-specific effect γ_{ja}. If γ_{aj} is correlated with x_{ij}, OLS estimates of δ will be biased. In absence of suitable instrumental variables for x_{ij}, spatial differencing offers a solution by differencing out the unobserved site-specific effects γ_{ja}. The presence of sample selection introduces nonlinearity into equation (1), making it challenging to apply spatial differencing.

We specify the model with sample selection as follow. Consider two latent dependent variables y^*_{1ij} and y^*_{2ij} on a cross-section. They both follow a regular linear model for individual i in a location j and are specified as follow:

$y^*_{1ij} = z'_{ij}\beta + \theta_{ja} + \theta_j + \varepsilon_{1ij}$ - selection equation

$y^*_{2ij} = x'_{ij}\delta + \gamma_{ja} + \gamma_j + \varepsilon_{2ij}$ - outcome equation

where ε_{1ij} and ε_{2ij} are independent identically distributed across individual error terms, θ_{ja} and γ_{ja} are a site-specific effects for site a affecting the selection and the outcome equation respectively and which are defined at a finer spatial scale than j.

We are interested in the estimation of the effect x_{ij} on an observed outcome y_{2ij}. The outcome is modeled in the form of a truncated sample selection model and is represented by equation (2).

$$y_{2ij} = \begin{cases} y^*_{2ij} \;\; if \;\; y^*_{1ij} > 0 \\ - \;\; if \;\; y^*_{1ij} \leq 0 \end{cases} \tag{2}$$

A consistent estimate of δ can be achieved by undertaking a Tobit regression assuming:[2]

Condition 1: $Cov[z_{ij}, \theta_{ja} + \theta_j + \varepsilon_{1ij}] = 0$, z_{ij} is exogenous;

[2] Identification required an exclusion restriction i.e. a variable that affects y^*_{1ij} but not y^*_{2ij} otherwise it relies on the nonlinearity of the inverse Mills ratio.

Condition 2: $Cov[x_{ij}, \gamma_{ja} + \gamma_j + \varepsilon_{2ij}] = 0$, x_{ij} is exogenous;

Condition 3: errors $(\varepsilon_{1ij}, \varepsilon_{2ij})$ satisfy $\varepsilon_{2ij} = \rho \times \varepsilon_{1ij} + v_{ij}$ with $\varepsilon_{1ij} \sim \mathcal{N}(0, 1)$ and independent of v_{ij}.

In most applications, the condition 2 is unlikely to hold because there is a possibility that, within a location, there could be omitted variables driving both the average site-specific effect and some observed characteristic of interest. The standard way to deal with the correlation between x_{ij} and γ_{ja} would be to find a suitable instrument for the x_{ij} and estimate an IV Tobit or IV two-stage Heckit. However, it is often difficult to find a variable correlated with x_{ij} and uncorrelated with local conditions. The exclusion restriction is likely to be violated and yield inconsistent estimates for δ. Another option is to use the finer location fixed effect and estimate the model using the classic Heckman two-stage procedure, but this will in practice lead to a proliferation of variables and reduced degrees of freedom.

2.1 Identification via Spatial Differencing

An alternative to the IV two-stage Heckit estimation technique is spatial differencing. Duranton, Gobillon, and Overman (2011), Black (1999) or Holmes (1998) use spatial differencing in the case of linear models to solve endogeneity problems arising from unobserved site effect γ_{ja}. We investigate the application of this spatial differencing technique when IVs are not available.

We denote Δ_d to be a spatial difference operator. An example is a pair wise difference operator which takes the difference between each observation and another observation located at distance less than d from that observation.[3] We can also consider the neighbourhood of an individual $\mathcal{N}_{id}$, and define the spatial difference operator as the difference between the individual outcome and the average outcome of his/her neighbour. This operator is similar to a fixed-effect, the difference being that the neighbourhood can overlap. We call this operator the fixed-effect difference operator. A further possibility is to use a kernel as in Kyriazidou (1997) to weight neighbours in $\mathcal{N}_{id}$ according to how far they are, in term of characteristics. This operator is the kernel difference operator.

[3] d is a number chosen by the researcher which define the neighborhood of an individual.

We define the pairwise spatial difference operator for any variable A with the observation k in a neighbourhood d of i as follows:

$$\Delta_d A_{ij} = A_{ij} - A_{kj}$$

For the spatial difference operator Δ_d, $\Delta_d y_{2ij} = y_{2ij} - y_{2kj}$ with k an observation in the neighbourhood d of i. It follows,

$$
\begin{aligned}
E[\Delta_d y_{2ij} | x, z, y^*_{1ij} > 0, y^*_{1kj} > 0, \gamma_d, \theta_d] &= E[y_{2ij} - y_{2kj} | x, z, y^*_{1ij} > 0, y^*_{1kj} > 0, \gamma_d, \theta_d] && (3)\\
&= E[y_{2ij} | x, z, y^*_{1ij} > 0, \gamma_d, \theta_d] - E[y_{2kj} | x, z, y^*_{1kj} > 0, \gamma_d, \theta_d] && (4)\\
&= x'_{ij}\delta + \gamma_{aj} + \gamma_j + \rho\lambda(z'_{ij}\beta + \theta_{ja} + \theta_j)\\
&- [x'_{kj}\delta + \gamma_{aj} + \gamma_j + \rho\lambda(z'_{kj}\beta + \theta_{ja} + \theta_j)]\\
&= \Delta_d x'_{ij}\delta + \Delta_d\gamma_{aj} + \rho\Delta_d\lambda(z'_{ij}\beta + \theta_{ja} + \theta_j) && (5)
\end{aligned}
$$

where $\lambda(c) = \phi(c)/\Phi(c)$ is the inverse Mills ratio.

Going from equation (3) to (4) follows from the linearity of expectation and from the mean independence of y_{2ij} and y^*_{1kj} conditional on $\{x, z, y^*_{1ij} > 0, \gamma_d, \theta_d\}$ and of mean independence of y_{2kj} and y^*_{1ij} conditional on $\{x, z, y^*_{1kj} > 0, \gamma_d, \theta_d\}$.[4] The second term $\Delta_d\gamma_{aj}$ (site-specific difference) and the third term $\rho\Delta_d\lambda(z'_{ij}\beta + \theta_{ja} + \theta_j)$ (sample selection term) in equation (5) present a challenge for identification. The conditions to consistently estimate the model are the following.

Assumption 1: The site-specific unobservable effect is homogenous in the a neighbourhood of the individual i.e. $\Delta_d\gamma_{ja} = 0$ for d small enough.

Under Assumption 1 we have:

$$E[\Delta_d y_{2ij} | x, z, y^*_{1ij} > 0, y^*_{1kj} > 0, \gamma_d, \theta_d] = \Delta_d x'_{ij}\delta + \rho\Delta_d\lambda(z'_{ij}\beta + \theta_{ja} + \theta_j) \qquad (6)$$

Here we apply the same identification condition as Duranton, Gobillon, and Overman (2011) allowing to difference out the site-specific unobserved effect γ_{ja}. The sample selection term depends on the unobservable site-specific and location effects $\theta_{ja} + \theta_j$. Because of nonlinearity simple differencing will not work as in the case of γ_{ja}. Assumption 2 helps us to deal with this challenge:

[4] θ_d and γ_d are respectively the location and the site-effect of the selection and outcome equation of all neighbours in neighbourhood d of i.

Assumption 2:

(i) $\Delta_d \theta_{ja} = 0$ for d small enough.

(ii) The changes in the probability of selection are similar in a neighborhood of the individual i.e.

$$\frac{\lambda(z'_{ij}\beta + \theta_{ja_i} + \theta_j) - \lambda(z'_{ij}\beta)}{\theta_{ja_i} + \theta_j} = \lambda'(c_{ik}) = \frac{\lambda(z'_{kj}\beta + \theta_{ja_k} + \theta_j) - \lambda(z'_{kj}\beta)}{\theta_{ja_k} + \theta_j}$$

for i and k in a neighbourhood d small enough.

Under Assumption 2, the equation (6) becomes

$$E[\Delta_d y_{2ij} | x, z, y^*_{1ij} > 0, y^*_{1kj} > 0, \gamma_d, \theta_d] = \Delta_d x'_{ij}\delta + \rho \Delta_d \lambda(z'_{ij}\beta) \tag{7}$$

Assumptions 1 and 2 are sufficient for the identification of δ and ρ. We derive the results using the pair wise spatial difference operator, but they also hold for other spatial difference operators.

2.2 Estimation and Inference

The estimation of the model can be done using Heckman's two-step procedure after differencing. We proposed a procedure to correct for heteroscedasticity in the error term arising form the use of spatial differencing and Heckman's two-step procedure (see Heckman (1974) and, Heckman (1979)).

- **Step 1:** Estimate β by probit with location random effect calculate $\hat{\lambda}_i = \lambda(z'_{ij}\hat{\beta})$ and $\tilde{\Delta}_d \lambda(z'_{ij}\beta) = \lambda(z'_{ij}\hat{\beta}) - \lambda(z'_{kj}\hat{\beta})$
- **Step 2:** Estimate δ and ρ in the OLS regression

$$\Delta_d y_{2ij} = \Delta_d x'_{ij}\delta + \rho \tilde{\Delta}_d \lambda(z'_{ij}\beta) + w_{ikj} \tag{8}$$

The error term w_{ikj} is heteroscedastic, we derive an estimator to yield correct standard errors.

Consider a generic matrix of spatial difference Δ. The matrix form notation of equation (8) can be expressed in as

$$\Delta y_2 = \Delta x'\delta + \rho \Delta \lambda(z'\hat{\beta}) + \Delta \eta$$

where $\eta_{ij} = y_{2ij} - x'_{ij}\delta - \rho\lambda(z'_{ij}\hat{\beta})$ are the same error as in standard sample selection models. Let us denote $\theta = (\delta, \rho)'$ and $W = [x', \lambda(z'\hat{\beta})]$. Then the model becomes $\Delta y_2 = \Delta W\theta + \Delta\eta$ and the OLS estimate of θ is

$$\hat{\theta} = [(\Delta W)'\Delta W]^{-1}[(\Delta W)'\Delta y_2] \tag{9}$$

The spatial nature of our data implies that an observation k with n neighbours will have n pairs. This induces correlation in the error term $\Delta\eta$ for all n of these pairs because of the spatial differencing in the second step of the estimation procedure.

The variance-covariance matrix is

$$Var(\hat{\theta}) = B\Sigma B'$$

with $B = \left[(\Delta W)'\Delta W\right]^{-1}$ and $\Sigma = (\Delta W)'Var(\Delta\eta)(\Delta W)$.

We consider $Var(\Delta\eta) = V_1 + V_2$ with

$$\begin{aligned} V_1 &= \Delta Var\left[y_2 - x'\delta - \rho\lambda(z'\beta)\right]\Delta' \\ &= \Delta[I + \rho^2 R]\Delta' \end{aligned}$$

where R is a diagonal matrix of dimension N (total number of observations), with $d_{ij} = 1 - \lambda(z'_{ij}\beta)[z'_{ij}\beta + \lambda(z'_{ij}\beta)]$ as the diagonal elements.

$$\begin{aligned} V_2 &= \rho^2\Delta Var\left[\lambda(z'\hat{\beta}) - \rho\lambda(z'\beta)\right]\Delta' \\ &= \rho^2\Delta D z V_p z' D\Delta' \end{aligned}$$

where D is the square, diagonal matrix of dimension N with $1 - d_{ij}$ as the diagonal elements; z is the data matrix of selection equation; and V_p is the variance-covariance estimate from the probit estimation of the selection equation.

The corrected variance covariance estimate for $\hat{\theta}$ is then:

$$V_{twostep} = B(\Delta W)'[\hat{V}_1 + \hat{V}_2](\Delta W)B'$$

where $\hat{V}_1 = \Delta[I + \hat{\rho}^2\hat{R}]\Delta'$ and $\hat{V}_2 = \hat{\rho}^2\Delta\hat{D}z\hat{V}_p z'\hat{D}\Delta'$ with all elements replaced by their estimates.

3 Conclusion

This paper proposes the use of spatial differencing as an alternative solution when IV are not available. The paper discusses the assumptions under which the parameters of the model are identified. The estimation of the parameters is achieved using the classic Heckman's two-step estimation procedure. We also derive an easy to implement standard errors estimator that corrects for heteroscedasticity emerging from the use of two-step estimation and differencing.

References

Black, S. E. (1999): "Do Better Schools Matter? Parental Valuation of Elementary Education," *The Quarterly Journal of Economics*, 114(2), 577–599.

Duranton, G., L. Gobillon, and H. G. Overman (2011): "Assessing the Effects of Local Taxation using Microgeographic Data," *The Economic Journal*, 121(555), 1017–1046.

Gibbons, S., and S. Machin (2003): "Valuing English primary schools," *Journal of Urban Economics*, 53(2), 197–219.

Heckman, J. (1974): "Shadow Prices, Market Wages, and Labor Supply," *Econometrica*, 42(4), 679–694.

Heckman, J. J. (1979): "Sample Selection Bias as a Specification Error," *Econometrica*, 47(1), 153–161.

Holmes, T. J. (1998): "The Effect of State Policies on the Location of Manufacturing: Evidence from State Borders," *Journal of Political Economy*, 106(4), 667–705.

Klein, R., and F. Vella (2010): "Estimating a Class of Triangular simultaneous Equations Models without Exclusion Restrictions," *Journal of Econometrics*, 154(2), 154–164.

Kyriazidou, E. (1997): "Estimation of a Panel Data Sample Selection Model," *Econometrica*, 65(6), 1335–1364.

Lewbel, A. (1997): "Constructing Instruments for Regressions with Measurement error when no Additional Data are available, with an Application to Patents and R&D," *Econometrica*, 65(5), 1201–1213.

——— (2012): "Using Heteroscedasticity to Identify and Estimate Mismeasured and Endogenous Regressor Models," *Journal of Business & Economic Statistics*, 30(1), 67–80.

Recent Kent Discussion Papers in Economics

16/14: 'Appropriate Technology and Balanced Growth', Miguel León-Ledesma and Mathan Satchi

16/13: 'On the Relationship between Lifestyle and Happiness in the UK', Adelina Gschwandtner, Sarah L. Jewell and Uma Kambhampati

16/12: 'What drives firm profitability? A comparison of the US and EU food processing industry', Adelina Gschwandtner and Stefan Hirsch

16/11: 'Spillovers of Community-Based Health Interventions on Consumption Smoothing', Emla Fitzsimons, Bansi Malde and Marcos Vera-Hernández

16/10: 'Production and Endogenous Bankruptcy under Collateral Constraints', Miguel León-Ledesma and Jaime Orrillo

16/09: 'Government Spending Multipliers in Natural Resource-Rich Developing Countries', Jean-Pascal Nganou, Juste Somé and Guy Tchuente

16/08: 'Regularization Based Anderson Rubin Tests for Many Instruments', Marine Carrasco and Guy Tchuente

16/07: 'Estimation of social interaction models using regularization', Guy Tchuente

16/06: 'The Post-crisis Slump in Europe: A Business Cycle Accounting Analysis', Florian Gerth and Keisuke Otsu

16/05: 'The Revenue Implication of Trade Liberalisation in Sub-Saharan Africa: Some new evidence', Lanre Kassim

16/04: 'The rise of the service economy and the real return on capital', Miguel León-Ledesma and Alessio Moro

16/03: 'Is there a mission drift in microfinance? Some new empirical evidence from Uganda', Francis Awuku Darko

16/02: 'Early Marriage, Social Networks and the Transmission of Norms', Niaz Asadullah and Zaki Wahhaj

16/01: 'Intra-household Resource Allocation and Familial Ties', Harounan Kazianga and Zaki Wahhaj

15/21: 'Endogenous divorce and human capital production', Amanda Gosling and María D. C. García-Alonso

15/20: 'A Theory of Child Marriage', Zaki Wahhaj