The Econometrics Journal

ROYAL ECONOMIC SOCIETY · FOUNDED 1890 ·

Econometrics Journal (2023), volume **00**, pp. 1–23.
https://doi.org/10.1093/ectj/utad025

Spatial differencing for sample selection models with 'site-specific' unobserved local effects

ALEXANDER KLEIN[†] AND GUY TCHUENTE[‡]

[†]*School of Economics, University of Kent, Park Wood Rd, Canterbury CT2 7FS, UK, CEPR, and CAGE.*
Email: A.Klein-474@kent.ac.uk

[‡]*Purdue University, NISER, GLO, Krannert Building, 403 Mitch Daniels Blvd, West Lafayette, IN, 47907-2056, USA.*
Email: gtchuent@purdue.edu

First version received: 23 June 2022; final version accepted: 8 June 2023.

Summary: This paper proposes an estimator which combines spatial differencing with a two-step sample selection estimator. We derive identification, estimation, and inference results from 'site-specific' unobserved effects. These effects operate at a spatial scale that cannot be captured by administrative borders. Therefore, we use spatial differencing. We show that under justifiable assumptions, the estimator is consistent and asymptotically normal. A Monte Carlo experiment illustrates the small sample properties of our estimator. We apply our procedure to the estimation of a female wage offer equation in the United States and the results show the relevance of spatial differencing to account for 'site-specific' unobserved effects.

Keywords: *Sample selection*, *site-specific unobserved heterogeneity*, *spatial difference*.

JEL codes: *C13*, *C31*, *J16*.

1. INTRODUCTION

Spatial differencing offers an identification strategy in the situations when researchers are limited to cross-section data, suspect unobserved omitted variables, and lack suitable instrumental variables.

It has been used in the context of linear regressions (e.g., Holmes, 1998; Black, 1999; Gibbons and Machin, 2003).[1] However, little is known about its use (or its application) in the context of nonlinear models. Therefore, we extend spatial differencing in this direction.

In the context of nonexperimental data, bias due to unobserved omitted variables can be a serious concern. Researchers have three main options: (a) they can use proxies, (b) instrumental variables, or (c) differencing the data across time or space.

(a) Proxies reduce the bias if they manage to capture the effect of omitted variables. However, proxies can be imperfect and their inclusion may even exacerbate the bias problem.[2]

[1] Duranton et al. (2011) also use spatial differencing, but they complement it with instrumental variable estimation.

[2] See Todd and Wolpin (2003) for details on the use of proxy, and Oster (2019) for a rigorous treatment of the evaluation of the robustness to omitted variables.

(b) Instrumental variables may help to alleviate the bias. However, as discussed in Todd and Wolpin (2003), the 'quasi-experimental' local average treatment effect (LATE) obtained in the instrumental variable model may not correspond to the *ceteris paribus* effect and thus may not correspond to the deep structural parameter of interest (see, for example, Heckman et al., 1997; Heckman and Vytlacil, 2005; Deaton and Cartwright, 2018).

(c) Differencing across time and/or space helps researchers to control for individual-specific unobserved heterogeneity. Differencing across time is widely used as panel data techniques have been developed over the years. Differencing across space is another option, but only linear spatial differencing methods have been developed so far.

Our paper develops an estimator for a sample selection model which uses spatial differencing even in the presence of a nonlinear element—in our case Mill's ratio. We derive the variance–covariance matrix of the proposed estimator that accounts for spatial differencing and sample selection, and examine its asymptotic distribution. The new estimator and the standard errors are easy to implement.

1.1. *Motivation for the estimator*

Suppose we have cross-sectional data on individuals with information on wages, educational attainment, and various socio-economic variables. Individuals self-select whether to work or not. We are interested in estimating returns to education with a sample selection model. We are concerned that there are unobserved variables in local labour markets which will bias our estimates, unless we control for them. We can use location-specific fixed effects such as county dummies or region dummies. However, local labour markets might operate at a geographically finer level than counties or regions, and they can be very heterogeneous. This means that the county or region dummies will not fully control for such heterogeneity. There is also no a priori reason to believe that the local labour markets will operate solely within the existing administrative borders.

As a result, even though location-specific dummies will control for some unobserved effects pertaining to the location, they will not control for all unobserved heterogeneous location effects at the finer spatial level, causing the estimated parameters to be inconsistent.[3] Spatial differencing across individuals living in the same neighbourhood, however, can eliminate this fine-location unobserved heterogeneity. The issue at stake, though, is that in a sample selection model, Mill's ratio cannot be differenced out using the assumptions and methods applied in linear models, since it is a nonlinear function containing the location-specific effects. Our methodology, however, accommodates this case.

1.2. *Main results*

We show that under justifiable assumptions that the spatially close individuals have similar unobserved heterogeneity and a similar inverse Mill's ratio derivative, spatial differencing eliminates the unobserved effects. The parameters of interest are then estimated using a two-step approach. We derive asymptotic properties and propose a correction of standard errors which accounts for the two-step nature of our estimation and spatial differencing.

[3] It is possible to envision a case when there is a very large number of observations at a finer geographical level which would then allow to estimate fixed effects at the finer geographical level.

Spatial differencing, while removing unobserved spatial effects, also induces a correlation in the error terms. We take that correlation into account and derive the asymptotic properties of our estimator using similar arguments to those present in the derivation of the asymptotic behaviour of clustered standard errors: the number of locations goes to infinity and the size of the location is assumed random and almost surely bounded.[4]

We find that the results about the asymptotic behaviour of our estimator also extends to a linear model without sample selection. This has important implications and researchers need to be cautious about it. Indeed, we show that the consistency of the estimator applied to the spatially differenced data requires (i) a large number of locations and (ii) a limited number of individuals in each location. Monte Carlo simulations also suggest that it would be better if the number of individuals in the site was small as well.

1.3. Contribution to literature

We contribute to spatial differencing literature by extending it to a sample selection model.[5]

Our paper also relates to sample selection literature which addresses a challenge posed by the presence of individual-specific unobserved heterogeneity in both the outcome and the selection equations. In such cases, the identification of the parameters is complicated by the nonlinearity or the incidental parameter problem (see Chamberlain, 2010; Fernández-Val and Weidner, 2016). This literature examines it in the context of cross-section or panel data and the solutions are based on either a full model specification or on a differencing procedure. Wooldridge (1995) uses a Mundlak approach to specify the individual-specific unobserved heterogeneity in both equations. He also imposes a special functional form on the selection mechanism. Kyriazidou (1997), however, does not impose strong restrictions on the selection equation functional form, and uses a nonparametric approach to difference out the unobserved fixed effect. Rochina-Barrachina (1999) for panel data and Ahn and Powell (1993) for cross-section data similarly rely on differencing to identify the parameters of the model, but they also impose additional distributional or functional form assumptions to the selection equation or nonparametric equation.

The main difference between our paper and the literature on the selection correction in panel data context is a different cluster asymptotic since, in each cluster, there is a finer common site-specific unobserved heterogeneity shared by individuals in that cluster. As a result, the outcomes of the individuals in our model are not independent while they are in the panel data case. Therefore, our asymptotic results are derived using an asymptotic distribution theory for a large number of clusters with heterogeneous, random, and bounded cluster sizes. Indeed, in each cluster, there is a finer common site-specific unobserved heterogeneity shared by individuals in that cluster.

The clustered dependence created by the finer site-specific unobserved heterogeneity also relates our asymptotic discussion to the literature on clustering at variance level (see Wooldridge, 2010, for a textbook treatment). The asymptotic distribution in that literature is derived using either a large or a fixed number of clusters. The fixed number of clusters leads to non-normal asymptotic distribution, and discussion about recent contributions can be found in Hansen and Lee (2019). An asymptotic distribution theory for a large number of clusters was first derived by White (1984), and has been investigated by several authors allowing either fixed cluster size or heterogeneous cluster. Recent developments include Hansen and Lee (2019), who propose conditions on the relation between the cluster sample sizes and the full sample in a regular asymptotic, or Djogbenou et al. (2019), who derive asymptotic with varying cluster sizes and

[4] Locations in our model are equivalent to clusters and we use ‘location’ and ‘cluster’ interchangeably.

[5] Additional references to already cited research are Belotti et al. (2018) and Druckenmiller and Hsiang (2018).

carry out a cluster wild bootstrap. Our results complement this literature by extending the many cluster asymptotic to a sample selection model.

In the empirical application, we are interested in estimating a sample selection model of female wages in the decades of the twentieth century prior to World War II, times when the US economy underwent substantial social and economic transformation that affected labour markets and employment opportunities for women (Goldin and Katz, 2010; Goldin, 2021). We have reasons to be concerned that individual unobserved heterogeneity is present in the main as well as selection equation. However, prior to World War II, the United States population censuses recorded nominal wages only in 1940, hence no wage information is available in the censuses before 1940. This poses a problem because we cannot use panel data techniques discussed earlier. Fortunately, the 1940 population census provides detailed geographical information about individual's location: state, county, and city/town. This allows us to use the spatial differencing technique proposed in this paper. We estimate a sample selection model with ordinary least squares (OLS), Heckman's two-step estimator, and the spatially differenced estimator proposed in this paper. The results show that not controlling for unobserved individual heterogeneity using spatial differencing biases the results of the standard Heckman sample selection estimator.

The structure of the paper is as follows. First, we expand the spatial differencing method in the case of the linear regression model to the case of sample selection. Then, we discuss identification assumptions, propose an estimation procedure, and derive the estimator of the corrected standard errors. We also conduct Monte Carlo simulations and then present an empirical application of our estimator.

2. SAMPLE SELECTION MODELS WITH SPATIAL CORRELATION

2.1. *Spatial differencing in a linear model*

In many economic applications, we are interested in estimating the following regression equation:

$$y_{ij} = x_{ij}'\delta + \gamma_j + \gamma_{j\alpha} + \varepsilon_{ij}, \tag{2.1}$$

where x_{ij}' is a vector of exogenous controls variables for i residing at location j (i can be, for example, a firm or an individual), γ_j is location fixed effect, $\gamma_{j\alpha}$ is an unobserved specific effect which is present at α, a finer spatial scale than location j, and ε_{ij} is the error term. We will call $\gamma_{j\alpha}$ a site-specific unobserved effect.[6]

An example of an application of this model can be found in the estimation of the effect of local taxation on the growth of firms. The impact of local taxation on the growth of firms can be affected by the location-specific effects such as county-specific effects. However, it also can be affected by finer-scale sites such as neighbourhoods, as highlighted in Duranton et al. (2011). 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, since there can be considerable heterogeneity at a finer spatial scale $\gamma_{j\alpha}$. Furthermore, standard location fixed effect γ_j relies on an arbitrary specification of the comparison neighbourhood group, as pointed out by Gibbons and Machin (2003), making it an imperfect control for a finer, site-specific effect $\gamma_{j\alpha}$. Including dummies for $\gamma_{j\alpha}$ is often not feasible because there can be too many of them, and they would suffer from the same issue of an arbitrary specification of the comparison neighbourhood group. If $\gamma_{j\alpha}$ is

[6] The site-specific component, $\gamma_{j\alpha}$, is a simplification for $\gamma_{j\alpha_i}$. We are implicitly assuming that the site-specific effects are the same for all its i.

correlated with x_{ij}, the OLS estimate of δ will be biased. In the absence of suitable instrumental variables for x_{ij}, the spatial differencing offers a solution by differencing out the unobserved site-specific effects $\gamma_{j\alpha}$.

Holmes (1998), Black (1999), Gibbons and Machin (2003), and Duranton et al. (2011) use spatial differencing in the case of linear models to solve endogeneity problems arising from unobserved site-specific effect $\gamma_{j\alpha}$. They take advantage of the fact that for sufficiently small distances between sites, their specific effect $\gamma_{j\alpha}$ changes smoothly, allowing thus to difference them out. This corresponds to the following assumption.

ASSUMPTION I1: *The site-specific unobservable effect is homogenous in the neighbourhood of the individual, i.e.,* $\Delta_d \gamma_{j\alpha} = 0$ *almost surely, for d small enough, where* Δ_d *is a spatial difference operator.*

2.2. *Spatial differencing in a sample selection model*

In several economic models, in addition to the site-specific fixed effects $\gamma_{j\alpha}$, the outcome of interest is not observed for a selected subsample. The selection can be the result of the decision of individuals, firms, or the researcher. The presence of sample selection then introduces nonlinearity to the model (2.1).

We specify the model with sample selection as follows. Consider two latent dependent variables y^*_{1ij} and y^*_{2ij} in a cross-section which follow a regular linear model for individual i in a location j:[7]

$$y^*_{1ij} = z'_{ij}\beta + \theta_{j\alpha} + \theta_j + \varepsilon_{1ij} \text{—selection equation,}$$
$$y^*_{2ij} = x'_{ij}\delta + \gamma_{j\alpha} + \gamma_j + \varepsilon_{2ij} \text{—outcome equation.}$$

Individual error terms are ε_{1ij} and ε_{2ij}; $\theta_{j\alpha}$ and $\gamma_{j\alpha}$ are site-specific effects for a site α in location j, affecting the selection and the outcome equation, respectively. The exogenous characteristics x_{ij} affect the outcome. They could be correlated with $\gamma_{j\alpha} + \gamma_j$, but not with ε_{1ij} and ε_{2ij}. The variables z_{ij} are exogenous variables determining selection, they can be a subset of x_{ij}. However, for identification purposes, some elements of z_{ij} are assumed to be absent from x_{ij}.

ASSUMPTION I2: ε_{1ij}, ε_{2ij} *are independent identically distributed normal random variables for all* i, j.[8]

The outcome is modelled in the form of a truncated sample selection model and is represented by (2.2).

$$y_{2ij} = \begin{cases} y^*_{2ij} & if \quad y^*_{1ij} > 0 \\ - & if \quad y^*_{1ij} \leq 0 \end{cases}. \tag{2.2}$$

Let us consider the following conditions.

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

[7] For the sake of clarity and simplicity of exposition, we will refer to i as individual for the rest of the paper.

[8] The Assumption I2 under which our results are derived imposes homoscedasticity and normality. These assumptions could be considered as strong. However, since our goal is to derive conditions under which the intuition of spatial differencing that applies to simple linear models could be generalised to sample selection models, we do not consider a semi-parametric model similar to Das et al. (2003) and Newey (2009) here. The semi-parametric model comes with additional identification challenges which could make the intuition behind the estimation of spatially differenced linear models in the context of sample selection models difficult.

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

Condition 3: and 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}.

It is possible to consistently estimate δ by Tobit regression under these three conditions.[9] In most applications, Conditions 1 and 2 are unlikely to hold, because there is a possibility that there could be a site-specific omitted variable affecting both the outcome and some observed characteristics of interest. Thus, it is possible that $Cov[x_{ij}, \gamma_{j\alpha} + \gamma_j] \neq 0$. The standard way to deal with the correlation between x_{ij} and $\gamma_{j\alpha}$ would be to find a suitable instrument for the x_{ij} and run an IV Tobit or IV two-stage Heckit. The very local nature of the site-specific effect means that it is not always evident to find a variable correlated with x_{ij} and uncorrelated with $\gamma_{j\alpha}$. The exclusion restriction is likely to be violated and IV two-stage Heckit will yield inconsistent estimates for δ.

Another option is to use the site-specific $\gamma_{j\alpha}$ fixed effects, and estimate the model using the classic Heckman two-stage procedure. This, however, has two main disadvantages. First, in practice it will lead to a proliferation of parameters and loss of degrees of freedom. Second, there is a conceptual issue: it relies on an arbitrary specification of the comparison sites (e.g., comparison neighbourhoods), which could cause a measurement error.

2.3. *Identification via spatial differencing*

This section investigates the application of this spatial differencing technique to the case of cross-section sample selection models. We denote Δ_d to be a spatial difference operator. One example is a pairwise difference operator which takes the difference between each observation and another observation located at distance less than d from that observation in location j, with individual i and k who are neighbours. The pairwise differencing of the variable A is:

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

Another example is the difference between the individual outcome and the average outcome of their neighbourhood $\mathcal{N}_{id}$. This operator is similar to the neighbourhood fixed effect operator, the difference being that the neighbourhoods can overlap. We call this operator the fixed effect difference operator. Let $\mathcal{N}_{id} = \{k, \quad in \quad neighbourhood \quad d\}$, the sample size of $\mathcal{N}_{id}$ is N_d, the differencing is given by:

$$\Delta_{df} A = A_{ij} - \frac{1}{N_d} \sum_{k \in \mathcal{N}_{id}} A_{kj}.$$

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 observable characteristics. This operator is the kernel difference operator:

$$\Delta_{dK} A = A_{ij} - \sum_{k \in \mathcal{N}_{id}} \psi(i, k) A_{kj},$$

where $\psi(i, k) = \frac{1}{h_{N_d}} K\left(\frac{(z'_{ij} - z'_{kj})\beta + (x'_{ij} - x'_{kj})\delta}{h_{N_d}}\right)$, K is a kernel density function while h_{N_d}, is a sequence of bandwidths. To illustrate our identification strategy and for the asymptotic derivation, we use

[9] Identification required an exclusion restriction, i.e., a variable that affects y^*_{1ij}, but not y^*_{2ij}. Otherwise, identification relies on the nonlinearity of the inverse Mill's ratio.

the pairwise spatial difference operator, while the fixed effect difference is used for the empirical application and for the Monte Carlo simulations.

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. Let $\xi_{ij} \equiv \{x_{ij}, z_{ij}, y^*_{1ij} > 0, \gamma_{id}, \theta_{id}\}$ with $\gamma_{id} = \{\gamma_{kj} \quad with \quad k \in \mathcal{N}_{id} \cup \{i\}\}$ and $\theta_{id} = \{\theta_{kj} \quad with \quad k \in \mathcal{N}_{id} \cup \{i\}\}$.

$$\begin{aligned} E[\Delta_d y_{2ij} | \xi_{ij}, \xi_{kj}] &= E[y_{2ij} - y_{2kj} | \xi_{ij}, \xi_{kj}] && (2.3) \\ &= E[y_{2ij} | \xi_{ij}] - E[y_{2kj} | \xi_{kj}] && (2.4) \\ &= x'_{ij}\delta + \gamma_{j\alpha} + \gamma_j + \rho\lambda\left(z'_{ij}\beta + \theta_{j\alpha} + \theta_j\right) \\ &\quad - \left[x'_{kj}\delta + \gamma_{j\alpha} + \gamma_j + \rho\lambda\left(z'_{kj}\beta + \theta_{j\alpha} + \theta_j\right)\right] \\ &= \Delta_d x'_{ij}\delta + \Delta_d\gamma_{j\alpha} + \rho\Delta_d\lambda\left(z'_{ij}\beta + \theta_{j\alpha} + \theta_j\right), && (2.5) \end{aligned}$$

where $\lambda(c) = \phi(c)/\Phi(c)$ is the inverse Mill's ratio while $\phi(c)$ and $\Phi(c)$ are respectively the density and distribution function of a normal random variable with mean zero and variance 1.

To go from (2.3) to (2.4) we use the linearity of expectation and the mean independence of y_{2ij} and y^*_{1kj} conditional on ξ_{ij}, as well as the mean independence of y_{2kj} and y^*_{1ij} conditional on ξ_{kj}, since we have assumed in Assumption I2 that ε_{1ij} and ε_{2ij} are independent and identically distributed (i.i.d.). The separation of the conditional set, ξ_{ij} and ξ_{kj}, is possible because we work with cross-sectional data. Such a separation is not possible for panel data. Indeed, in the context of panel data with individual effects and sample selection, when the differencing is used to remove the fixed effects, the conditional set cannot be separated as we have done to move from (2.3) to (2.4). To 'difference out' the unobserved heterogeneity, extra assumptions are imposed. For example, Kyriazidou (1997) imposes a 'conditional exchangeability' assumption that implies that the distribution of error terms are equal over time for all individuals in the sample. In the case of models with censoring, Lee (2001) discusses conditions under which first difference can be used, and imposes the linear implication of the 'conditional exchangeability' assumption. In a similar context using the first difference, Rochina-Barrachina (1999) imposes joint normality between the difference in the error of the outcome equation and the error in the selections equation in the two time periods.[10]

Estimating (2.5) presents two challenges for the identification of the parameter of interest δ and the sample selection parameter ρ: the site-specific difference $\Delta_d\gamma_{j\alpha}$, and the sample selection term $\rho\Delta_d\lambda(z'_{ij}\beta + \theta_{j\alpha} + \theta_j)$. As for the site-specific difference $\Delta_d\gamma_{j\alpha}$, under Assumptions I1 and I2, (2.5) becomes

$$E[\Delta_d y_{2ij} | \xi_{ij}, \xi_{kj}] = \Delta_d x'_{ij}\delta + \rho\Delta_d\lambda\left(z'_{ij}\beta + \theta_{j\alpha} + \theta_j\right). \qquad (2.6)$$

These assumptions allow us to difference out the site-specific unobserved effect $\gamma_{j\alpha}$, a strategy that was applied by Duranton et al. (2011).

As for the sample selection term $\rho\Delta_d\lambda(z'_{ij}\beta + \theta_{j\alpha} + \theta_j)$, we see that it depends on the unobservable site-specific and location effects $\theta_{j\alpha} + \theta_j$. Since that sample selection term is a nonlinear function, a simple spatial differencing will not always work, unlike the case of γ_{ja}. Therefore, the following assumption helps us to deal with this challenge:

ASSUMPTION I3: *(i) The site-specific unobservable selection effect is homogeneous in a neighbourhood of an individual, i.e., $\Delta_d\theta_{ja} = 0$ almost surely for d small enough. (ii) The following*

[10] See Dustmann and Rochina-Barrachina (2007) for a review on selection correction in panel data models.

equality holds for the inverse Mill's ratio in a neighbourhood of all individuals:

$$\frac{\lambda\big(z'_{ij}\beta+\theta_{j\alpha_i}+\theta_j\big)-\lambda\big(z'_{ij}\beta\big)}{\theta_{j\alpha_i}+\theta_j}=\frac{\lambda\big(z'_{kj}\beta+\theta_{j\alpha_k}+\theta_j\big)-\lambda\big(z'_{kj}\beta\big)}{\theta_{j\alpha_k}+\theta_j},$$

almost surely, for i and k in a neighbourhood d, and $\theta_{j\alpha_i}+\theta_j$ and $\theta_{j\alpha_k}+\theta_j$ are both different from 0.

Assumption I3(i) is similar to Assumption I1. It seems plausible that if that assumption holds for the outcome equation, it will hold for the selection equation as well.

Let $\lambda'(.)$ be the first derivative of the inverse Mill's ratio. There exist c_i and c_k which are, respectively, in the intervals formed by $[z'_{ij}\beta,\ z'_{ij}\beta+\theta_{j\alpha_i}+\theta_j]$, and $[z'_{kj}\beta,\ z'_{kj}\beta+\theta_{j\alpha_k}+\theta_j]$ such that $\lambda'(c_i)=\frac{\lambda(z'_{ij}\beta+\theta_{j\alpha_i}+\theta_j)-\lambda(z'_{ij}\beta)}{\theta_{j\alpha_i}+\theta_j}$ and $\lambda'(c_k)=\frac{\lambda(z'_{kj}\beta+\theta_{j\alpha_k}+\theta_j)-\lambda(z'_{kj}\beta)}{\theta_{j\alpha_k}+\theta_j}$ almost surely. Assumption I3(ii) is novel and one of the contributions of this paper. If the level of nonlinearity of $\lambda(.)$ is low, then the assumption will also hold. In the extreme case of local linearity of the inverse Mill's ratio, the Assumption I3(ii) perfectly holds.

2.4. *Similarity with nonlinear panel data assumptions*

Assumption I3(ii) imposes a condition on the behaviour of the nonlinear part of the (2.6). In the nonlinear panel data model, it is common to use assumptions on the functional form for identification. For example, Chamberlain (2010) derives identification results for panel data model with binary outcome showing that identification is possible only in the logistic case. Also, Bonhomme (2012) imposes compactness and a non-surjectivity assumption on the operator associated with the conditional distribution of the unobserved heterogeneity, and uses functional differencing to achieve identification and estimate the parameters of the model. Note that Assumption I3(i) are restrictions on the conditional distribution of unobserved heterogeneity similar to those in assumption 1(i) of Bonhomme (2012) while Assumption I3(ii) restricts the behaviour of the conditional distribution of the outcome in the same spirit as Bonhomme (2012) assumption 1(ii).

2.5. *Intuition and application*

The intuition of the Assumption I3(ii) is similar to the smoothness assumption imposed on the nonparametric selection control function in Ahn and Powell (1993) for cross-sectional data. In empirical applications, if individuals in a neighbourhood share the same finer-level unobserved heterogeneity (e.g., two neighbouring towns), the Assumption I3(i) should hold. Moreover, if the inverses Mill's ratio does not have substantial nonlinearity Assumption I3(ii) could hold. This is possible because there will exist many data sets for which $\lambda'(c_i)=\lambda'(c_k)$ in small neighbourhoods (e.g., i and k two neighbouring towns).[11]

The combination of Assumptions I3(i) and I3(ii) implies that

$$\lambda\left(z'_{ij}\beta+\theta_{j\alpha_i}+\theta_j\right)-\lambda\left(z'_{ij}\beta\right)=\lambda\left(z'_{kj}\beta+\theta_{j\alpha_k}+\theta_j\right)-\lambda\left(z'_{kj}\beta\right).$$

Thus, $\Delta_d\lambda(z'_{ij}\beta)=\Delta_d\lambda(z'_{ij}\beta+\theta_{j\alpha}+\theta_j)$.

[11] c_i, and c_k are, respectively, in the intervals formed by $[z'_{ij}\beta,\quad z'_{ij}\beta+\theta_{j\alpha_i}+\theta_j]$ and $[z'_{kj}\beta,\quad z'_{kj}\beta+\theta_{j\alpha_k}+\theta_j]$ such that $\lambda'(c_i)=\frac{\lambda(z'_{ij}\beta+\theta_{j\alpha_i}+\theta_j)-\lambda(z'_{ij}\beta)}{\theta_{j\alpha_i}+\theta_j}$ and $\lambda'(c_k)=\frac{\lambda(z'_{kj}\beta+\theta_{j\alpha_k}+\theta_j)-\lambda(z'_{kj}\beta)}{\theta_{j\alpha_k}+\theta_j}$.

THEOREM 2.1. *Let us consider the sample selection model presented in (2.2). Under Assumptions I1 to I3, the parameters δ and ρ are identified.*

Proof of Theorem 2.1: We have already shown that under the Assumptions I1 and I2 we can obtain (2.6). Applying Assumption I3 to (2.6) leads to the following equation

$$E[\Delta_d y_{2ij}|\xi_{ij}, \xi_{kj}] = \Delta_d x'_{ij}\delta + \rho\Delta_d\lambda\left(z'_{ij}\beta\right). \tag{2.7}$$

Thus, Assumptions I1 to I3 are sufficient for the identification of δ and ρ. □

We have derived the results using the pairwise spatial difference operator. However, the identification result also holds for other spatial difference operators. In the case of the average or kernel difference operator, the conditioning in (2.7) is on ξ_{kj} with $k \in \mathcal{N}_{id}$ for the average difference operator, and k is in the full sample for the kernel operator. Note that under Assumptions I1 and I3, any difference in the weighted average in a neighbourhood of the individual will enable us to remove the site-specific effect. The conditional expectation presented in (2.7) depends on exogenous observable variables and parameters of interest.

2.6. *Estimation and asymptotic properties*

In this section, we present an estimation procedure and derive the asymptotic properties of the proposed estimator. The estimation procedure involves two steps. In the first step, the probit model is estimated and the inverse Mill's ratio is predicted. In the second step, a spatial difference operator differences out both location and the site-specific unobserved heterogeneity. The model is then estimated using an ordinary least square estimator. When we have a full sample of N_f individuals, the estimation procedure is thus as follows:

Step 1: Estimate β by probit with location effect γ_j; and calculate $\hat{\lambda}_i = \lambda(z'_{ij}\hat{\beta})$.
Step 2: Estimate δ and ρ in the OLS regression

$$\Delta_d y_{2ij} = \Delta_d x'_{ij}\delta + \rho\Delta_d\lambda\left(z'_{ij}\hat{\beta}\right) + w_{ikj}. \tag{2.8}$$

Since we used spatial differencing and estimated $\lambda(z'_{ij}\hat{\beta})$ in Step 1, a particular structure of the variance–covariance matrix emerges. Therefore we also need to derive the correct estimator of standard errors which we will do in Section 2.7.

We will now show that the estimator obtained by this procedure is consistent and asymptotically normal. To derive the asymptotic properties, we use similar arguments as those used to derive the asymptotic properties of the clustered standard errors. Specifically, the population size of each location is assumed random, and bounded almost surely, and the law of large numbers is applied by letting the number of locations (clusters in case of clustered standard errors) go to infinity.

We consider a generic matrix of spatial difference Δ. The matrix form notation of (2.8) can be expressed as[12]

$$\Delta y_2 = \Delta x'\delta + \rho\Delta\lambda(z'\hat{\beta}) + \Delta\eta, \tag{2.9}$$

where η_{ij} are the same errors as in standard sample selection models.[13] Let us denote $\theta = (\delta, \rho)'$ and $W = [x', \lambda(z'\hat{\beta})]$. The simplified estimation (2.9) is

$$\Delta y_2 = \Delta W\theta + \Delta\eta,$$

[12] The variables without subscript represent vectors or matrices of all observations in the sample.

[13] We assume in the notation that $\lambda(z'\hat{\beta})$ is a vector with a typical element $\lambda(z'_{ij}\hat{\beta})$.

and the OLS estimator of θ is

$$\hat{\theta} = [(\Delta W)'\Delta W]^{-1}[(\Delta W)'\Delta y_2]. \tag{2.10}$$

The spatial nature of data implies that an observation k with n neighbours may appear in several 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. As a result, a particular structure of the covariance matrix emerges, and we need to take that into account when calculating the standard errors.

To proceed further, we need to introduce assumptions under which the asymptotic properties of our estimator are derived.

ASSUMPTION E1: *The selected sample size is N.*

- (i) *We observed $\{x_{ij}, z_{ij}\}$ independent and identically distributed random variables with $i = 1, ..., N$ and $j = 1, ..., J$.*
- (ii) *The number of individuals in a location j, N_j, is exogenous, random, identically distributed with $N_j < n_0$ almost surely, where n_0 is a scalar. Thus, $E(N_j) < \infty$.*
- (iii) *The outcomes and the latent variables are independent across locations, i.e., $j_1 \neq j_2$ the variables $y_{2ij_1} \perp y_{2ij_2}$ and $y^*_{1ij_1} \perp y^*_{1ij_2}$.*
- (iv) *$\theta_{j\alpha_i}$ is uncorrelated with z_{ij}.*

An implication of Assumption E1(i) in conjunction with Assumption I2 is that θ_j, γ_j are i.i.d. However, *within* a location j, there is a certain level of correlation among individuals which operates through $\gamma_{j\alpha_i}$. This means that our assumptions restrict how those within-location individual correlations occur.

Assumption E1(ii) restricts the location size to be bounded and implies that the number of locations has to grow to achieve a large sample size in our asymptotic calculation. This assumption is similar to those held in the literature on clustered samples asymptotic distributions, and it leads to an asymptotic distribution theory for a ‘large number of clusters’ similar to Wooldridge (2010), which assumes fixed cluster size. Moreover, Assumption E1(ii) corresponds to having all sampled clusters being small proportions of the population of clusters of interest. Thus, clustering is required based on Abadie et al. (2023).

This assumption corresponds to a specific case of assumption 1 in Hansen and Lee (2019), which allows for different cluster sizes ranging from fixed to infinite. We have, however, derived the asymptotic of our estimator under the more restrictive condition of Assumption E1(ii). The reason is that it can be proven that under a joint asymptotic (N, $J \to \infty$), assumption 1 in Hansen and Lee (2019) is equivalent to assuming that the size of the sample in each location is bounded. If we instead allow for a sequential asymptotic where the number of locations is fixed, and the sample size goes to infinity, then there exists at least one location with an infinite number of individuals, and the inequality used in the proof of Hansen and Lee’s (2019) theorem 1 becomes invalid.

To better illustrate our argument, let us consider the location sample size proposed by Hansen and Lee (2019): $N_j = N^\alpha$ with $0 \leq \alpha < 1$; we can prove that $1 - \alpha = \frac{ln(J)}{ln(N)}$. If we allow for a joint asymptotic, α is not define. If, on the contrary, we assume that the number of locations J is fixed, then, α goes to 1. In both cases, relying on Hansen and Lee’s (2019) assumption 1 seems not enough to warrant the desired asymptotic regularities.

Assumption E1(iv) implies that after controlling for location fixed effects, a consistent estimate of β can be achieved using a maximum likelihood procedure. It, however, allows for a correlation between $\gamma_{j\alpha_i}$ and $\theta_{j\alpha_i}$, making differencing important to avoid biased estimation of the parameters of interest.

ASSUMPTION E2: *z' and ΔW are full rank column, with each element having up to its fourth moment.*

THEOREM 2.2. *We consider the sample selection model presented in (2.2). Under Assumptions I1 to I3, E1, and E2.*

(i) $\hat{\theta} \rightarrow^p \theta$ *as* $N \rightarrow \infty$
(ii) $\sqrt{N}(\hat{\theta} - \theta) \rightarrow^d \mathcal{N}(0, \Theta)$ *with* $\Theta = C\Gamma C'$

where $C^{-1} = E((\Delta W_{ij})'\Delta W_{ij}), \Gamma = \frac{\rho^2}{s} E[(\Delta W_{ij})'\Omega_{ij}\Delta W_{ij}] + \frac{1}{s} E[(\Delta W_{ij})'\Delta e_{ij}\Delta e_{ij}(\Delta W_{ij})]$, *and* $\Omega_{ij} = [\lambda'(z'_{ij}\beta)]^2 z'_{ij} V_\beta z_{ij}$ *taking* V_β *as the first step probit variance–covariance matrix, and* $N/N_f \rightarrow^p s$.

Proof of Theorem 2.2: In the Appendix.[14] □

It is important to notice that the same type of asymptotic should be used in a linear model. In this respect, we complement Duranton et al. (2011) who propose a correction for the standard errors, but do not discuss the asymptotic properties of their estimators. Similarly, Black (1999) and Holmes (1998) use spatial differencing, but do not account for the fact that differencing will lead to a correlation between the pairs in which the same individual is present. Our asymptotic derivations do account for the presence of correlation between pairs and are valid, not only for a model with, but also without, sample selection (in our model, the absence of selection implies $\rho = 0$). They also have important practical implications: the consistency of the estimator requires a large number of locations γ_j, and a small number of individuals in each site $\gamma_{j\alpha}$.

2.7. *Estimator of variance*

This section derives a procedure to estimate the variance–covariance of the estimator in (2.10) which has a particular structure arising from (i) spatial differencing and (ii) a sample selection two-step estimation procedure.

We consider $B = \left[(\Delta W)'\Delta W\right]^{-1}$ and $\Sigma = Var[(\Delta W)'\Delta\eta]$ such that the conditional variance–covariance matrix of $\hat{\theta}$ is

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

Note that

$$\Sigma = (\Delta W)' Var(\Delta\eta)(\Delta W).$$

This means that we need a consistent estimator of $Var(\Delta\eta)$ to compute correct standard error for $\hat{\theta}$.

[14] The proof of asymptotic behaviour uses the pairwise difference. However, the result and the proof strategy are similar for all the differences proposed in this paper.

Let us consider that $Var(\Delta\eta) = V_1 + V_2$ with

$$V_1 = \Delta Var(e)\Delta'$$

$$= \frac{1}{s}\Delta R \Delta',$$

where R a diagonal matrix of dimension N_f (total number of observations), with $d_{ij} = \rho^2 - \rho^2 \lambda(z'_{ij}\beta)[z'_{ij}\beta + \lambda(z'_{ij}\beta)]$ as the diagonal elements.

$$V_2 = \frac{\rho^2}{s}\Delta Var\left[\lambda(z'\hat{\beta}) - \lambda(z'\beta)\right]\Delta'$$

$$= \frac{\rho^2}{s}\Delta D z V_\beta z' D \Delta',$$

where D is the square, diagonal matrix of dimension N_f with $\lambda(z'_{ij}\beta)[z'_{ij}\beta + \lambda(z'_{ij}\beta)]$ as the diagonal elements; z is the data matrix of selection equation; and V_β is the variance–covariance estimate from the probit estimation of the selection equation.

THEOREM 2.3. *We consider the sample selection model presented in (2.2). Under Assumptions I1 to I3, E1 and, E2. The variance–covariance estimator of the $\hat{\theta}$ is given by*

$$V_{twostep} = B(\Delta W)'[\hat{V}_1 + \hat{V}_2](\Delta W)B', \tag{2.11}$$

where $\hat{V}_1 = \frac{1}{\tilde{s}}\Delta\hat{R}\Delta'$ and $\hat{V}_2 = \frac{\hat{\rho}^2}{\tilde{s}}\Delta\hat{D}z\hat{V}_\beta z'\hat{D}\Delta'$ with all unknown parameters replaced by their consistent estimates using the sample of selected individuals (N). Moreover, this is a consistent estimator of the asymptotic variance of $Var(\hat{\theta})$.

Proof of Theorem 2.3: In the Appendix. □

3. MONTE CARLO SIMULATION

In this section, we present the results of Monte Carlo simulations to (a) describe the behaviour of the estimator proposed in this paper and (b) offer empirical guidance for applied research. Regarding the latter, we will pay close attention to the implication of Assumption E1(ii) according to which it is important to have a large number of locations relative to the number of individuals in the sites. Monte Carlo experiments will offer empirical guidance as to when the number of locations is large enough.

The data is obtained using the following data-generating process. We assume that there are $J = 20, 30, 100$ nonoverlapping locations, and each location is divided into $s = 2, 4, 8$ sites. There are $n_j = 2, 5, 8, 10$ individuals sharing the same site. The latent variables are $y^*_{1ij} = z_{ij}\beta + \theta_{ijs} + \theta_j + \varepsilon_{1ij}$ and $y^*_{2ij} = x_{ij}\delta + \gamma_{ijs} + \gamma_j + \varepsilon_{2ij}$, where $\theta_{ija} = 10^{-5}j \times s$ and $\gamma_{ijs} = 5j \times s$ is the site-specific effect, while $\theta_j = 10^{-5}j$ and $\gamma_j = 10j$ are the location effects; for all i and j, $x_{ij} \sim \mathcal{N}(0, 1)$, $z_{ij} \sim U(0, 1)$ each drawn independently; $\delta = 1$, $\beta = 0.2$. The error terms in both equations for all i and j are generated as follows: $\varepsilon_{1ij} \sim \mathcal{N}(0, 1)$, $\varepsilon_{2ij} = \rho\varepsilon_{1ij} + v_{ij}$ where $v_{ij} \sim \mathcal{N}(0, 1)$ is independent of ε_{1ij} and $\rho = 0.7$. The results presented in Table 1 is for the parameter δ.

Table 1. Simulation results for site-specific spatial difference estimator of δ.

No. of locations	No. of sites	Site size	Mean bias	Coverage rate
20	2	2	0.009	93.90
		5	−0.011	98.34
		8	0.031	99.33
		10	−0.019	97.92
	4	2	0.007	92.30
		5	−0.009	98.05
		8	−0.078	97.06
		10	0.094	97.53
	8	3	0.000	88.70
		5	−0.007	97.89
		8	−0.024	99.37
		10	0.010	100.00
30	2	2	0.016	93.60
		5	−0.004	97.90
		8	0.060	100.00
		10	−0.014	98.31
	4	2	−0.003	90.60
		5	0.025	97.34
		8	0.019	98.64
		10	0.067	97.35
	8	2	0.008	87.70
		5	0.006	96.50
		8	−0.015	98.83
		10	−0.004	100.00
100	2	2	−0.002	87.70
		5	−0.008	97.59
		8	−0.010	98.86
		10	−0.039	100.00
	4	2	0.001	86.20
		5	0.008	95.50
		8	0.006	98.99
		10	0.073	99.11
	8	2	0.002	85.90
		5	−0.001	94.60
		8	0.006	99.17
		10	0.001	99.46

We summarise the main results of the simulations in the points below, but, in general, the ‘site-specific spatial difference’ estimator has the smallest mean bias and delivers a coverage rate below 95%.[15]

[15] There is room for improvement concerning our inference strategy. Cluster robust inference is part of a large and growing literature and our work gives some insight as to how differencing can be used in cross-sectional data. Future work will investigate the importance of heteroscedasticity, and small sample procedures such as bootstrap will be used to improve inference.

(a) The mean bias of the estimator increases with the number of individuals in the sites and decreases with the number of locations. For example, in a sample of 400 individuals which are spread across 100 locations with 2 sites and 2 individuals in each site, the mean bias is −0.002. However, with 600 individuals spread across 30 locations with 4 sites and 5 individuals in each site, the mean bias is 0.025. This result is in line with our asymptotic derivations.

(b) For a fixed number of locations, the bias increases with the number of individuals in sites. The empirical consequence of these results is that our estimator should be applied when the population size at the site level is small.

(c) The coverage rate of the estimator using the variance–covariance estimator in (2.11) and the normal asymptotic distribution suggests good coverage. However, they are larger than 95% in cases where the number of individuals in the sites is large (8 individuals in the sublocation).

4. EMPIRICAL ILLUSTRATION

In this section, we illustrate the importance of the spatial differencing methodology proposed in this paper with an empirical example. In particular, we estimate a sample selection model of female wages using a 1% sample from the 1940 population census of the United States (Ruggles et al., 2021).[16] This census provides information about employment status, nominal wages, educational attainment, and numerous demographic and socio-economic characteristics such as age, marital status, number of children, number of children less than five years old, and industry in which individuals are employed. In addition, it provides detailed geographic information about an individual's location: state, county, and town/city. We estimate a sample selection model as specified in (2.2). The outcome variable is log wage of female i residing in a county j and a town/city α.[17] Explanatory variables include potential experience, potential experience squared, marital status, number of children, education attainment, and industry dummy variables. Potential experience is defined as (age–years of schooling–6).[18] The selection equation determines whether a female works or not and in addition to the variables in the outcome equation, it also includes the number of children less than five years old. Education attainment is measured by the number of school years.

Female wages can be influenced by the unobserved characteristics of area-specific labour markets. As seen in (2.1), there can be location-specific effects γ_j and unobserved site-specific effects $\gamma_{j\alpha}$, and not controlling for them could bias the results. One solution would be to include geographical dummy variables: county fixed effects to control for γ_j and town fixed effects to control for $\gamma_{j\alpha}$. While they will control for county- and town-specific unobserved effects, there is a conceptual drawback to expect that they will fully control for $\gamma_{j\alpha}$.[19]

[16] Source: IPUMS dataset version 3.

[17] In 1940, most of the towns and cities resided within a county and had not yet developed into the metropolitan areas we know today, a development which happened after World War II.

[18] We follow Heckman et al. (2006) by using this formula to approximate the number of years of experience.

[19] There is also a practical drawback: including county and town/city fixed effects means including 86 county and 469 town dummies which greatly diminishes the degrees of freedom and might also cause incidental variable problems when estimating probit-selection equations. Our empirical application has a large number of observations because we use a sample from a full count US population census. However, we can envision empirical applications in which the inclusion of a large number of geographical fixed effects would be infeasible due to a small number of observations.

Table 2. Female wage equation US 1940 census.

	OLS	Heckman two-step		Heckman two-step		Spatial differencing
		Main	Selection	Main	Selection	
	(1)	(2)	(3)	(4)	(5)	(6)
Potential experience	0.0231^{***}	0.0627^{***}	0.0135^{***}	0.0600^{***}	0.0135^{***}	0.0847^{***}
	[0.005]	[0.001]	[0.001]	[0.001]	[0.001]	[0.0007]
(Potential experience) 2	-0.0007^{***}	-0.0010^{***}	-0.0004^{***}	-0.0009^{***}	-0.0004^{***}	– 0.0025
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.0025]
Number of children	-0.2197^{***}	-0.1267^{***}	-0.1203^{***}	-0.1012^{***}	-0.1208^{***}	-0.6820^{***}
	[0.013]	[0.008]	[0.006]	[0.007]	[0.007]	[0.0001]
Number of years of education	0.0139^{**}	0.0726^{***}	0.0025	0.0729^{***}	0.0011	0.0170^{***}
	[0.006]	[0.002]	[0.002]	[0.002]	[0.002]	[0.007]
Married (dummy variable)	-0.4319^{***}	-0.0570^{***}	-0.0917^{***}	-0.0500^{***}	-0.0977^{***}	-0.8033^{***}
	[0.052]	[0.009]	[0.014]	[0.009]	[0.014]	[0.0012]
Number of children age five and less			-0.1291^{***}		-0.1354^{***}	
			[0.020]		[0.020]	
Inverse Mill's ratio		0.4970^{***}		0.1903^{**}		4.6460^{***}
		[0.110]		[0.089]		[0.0106]
Town fixed effects	Yes	No	No	Yes	Yes	No
Observations	144,079	144,079		144,079		144,079

Notes: This table presents the results of estimating a wage equation where the dependent variable is the log of female wage. Data source: 1% sample from 1940 US census, IPUMS 2021. Robust standard errors are in brackets. *** $p < 0.01$, ** $p < 0.05$.

County fixed effect will control for any homogeneous effects operating within the county. It will, however, fail to sufficiently control for any heterogeneous unobserved effects operating in the county. Similarly, town/city fixed effects will suffice only if we assume that site-specific effects are confined to the towns'/cities' administrative borders. Indeed, there is no a priori reason to expect that local labour markets are spatially delineated *only* by the town/city, or county administrative boundaries. Furthermore, an area fixed effect approach, in general, depends on an arbitrary specification of a comparison group. All this means that county and town fixed effects can create a measurement error problem and lead to inconsistent parameters estimates. Therefore, we estimate (2.2) with the spatial differencing estimator developed in this paper where the spatial differencing is between a woman and the average of her neighbours. We chose a neighbourhood of a 15 mile radius. Historical circumstances of the United States in 1940 justifies this choice. Labour markets at that time were being gradually extended beyond the city limits (Lewis, 2002; Miller, 2018). However, they were still more local than today, very heterogeneous, and commuting was limited to small distances since suburbanisation and car ownership surged only from the 1950s (Jackson, 1987; Harris and Lewis, 2001).

The estimation results are presented in Table 2. Column (1) presents OLS estimates, columns (2)–(5) the estimates using the Heckman two-step estimator, and column (6) the estimates using spatial differencing and the estimator proposed in this paper. We have estimated two versions of Heckman's two-step estimator: one without town fixed effects, in columns (2) and (3), and the other with town fixed effects, in columns (4) and (5). Standard errors for OLS and Heckman's two-step estimator are clustered at town/city level. Standard errors of the proposed estimator are calculated using (2.11).

We see that in all specifications the estimated coefficients are statistically significant at a 1% level.[20] As for the coefficient signs, we see a quadratic potential experience profile which suggests a diminishing marginal returns to experience. Being married and having children lowered women's wages, while education attainment, as measured by the number of school years, increased

[20] The only exception is in column (1) where 'Number of years of education' is significant at a 5% level.

it. When we compare OLS estimates with the estimates using the Heckman two-step estimator, the estimated coefficients differ in magnitude.

When we compare the estimates between the Heckman two-step estimator with and without town/city fixed effects, respectively, we see that they are very similar and that the differences are largely confined to the second decimal place. This suggests that town/city fixed effects have very limited effects on the estimated parameters.

As discussed earlier, we are concerned that the unobserved site-specific effects are not fully controlled for. Therefore, in column (6) we have used spatial differencing and the estimator proposed in Section 2.6. The returns to potential experience increased when compared with the estimates of OLS and the Heckman two-step estimator, respectively. Returns to education increased when compared with OLS estimates and decreased relative to the estimates of the Heckman two-step estimator. The effects of marital status and the number of children increased in absolute terms and indicate a large negative effect of being married and having children, respectively. This finding is consistent with historical evidence which shows that being married and having children at that time posed some limitations to women's wage prospects (Kessler-Harris, 1982; Goldin, 2021). Overall, the change in the magnitude of the estimated coefficients suggests that the sample selection estimator without spatial differencing biases the estimates.

5. CONCLUSION

This paper has investigated a sample selection model with unobserved effect at a very fine location level. It proposes spatial differencing as an alternative identification strategy. We discuss the assumptions under which the parameters of the model are identified. The estimation of the parameters is done using a two-step estimation procedure. The spatial differencing and the two-step procedure lead to a novel estimator with properties that are also relevant for spatial differencing in linear models. To understand the behaviour of the new estimator, we derive an asymptotic distribution of the estimator using a theory for a large number of clusters. The derivation reveals two important implications for its empirical implementation: (i) the number of clusters needs to be large for inference to be based on a normal distribution. (ii) Each cluster should have a bounded number of individuals.

Monte Carlo experiments show that accounting for site-specific heterogeneity is crucial for identification. In particular, the estimator performs better with the increasing number of locations and fewer individuals in sites (thus, in location). The coverage rate of the test based on the corrected standard error has an empirical coverage around the theoretical one. We illustrate the importance of spatial differencing with an empirical example in which we estimate a sample selection model of female wages. The results confirm that not controlling for site-specific unobserved effects biases the results of the standard Heckman sample selection estimator.

REFERENCES

Abadie, A., S. Athey, G. W. Imbens and J. M. Wooldridge (2023). When should you adjust standard errors for clustering? *Quarterly Journal of Economics 138*(1), 1–35.

Ahn, H. and J. L. Powell (1993). Semiparametric estimation of censored selection models with a nonparametric selection mechanism. *Journal of Econometrics 58*(1–2), 3–29.

Belotti, F., E. Di Porto and G. Santoni (2018). Spatial differencing: Estimation and inference. *CESifo Economic Studies 64*(2), 241–54.

Black, S. E. (1999). Do better schools matter? Parental valuation of elementary education. *Quarterly Journal of Economics 114*(2), 577–99.

Bonhomme, S. (2012). Functional differencing. *Econometrica 80*(4), 1337–85.

Chamberlain, G. (2010). Binary response models for panel data: Identification and information. *Econometrica 78*(1), 159–68.

Das, M., W. K. Newey and F. Vella (2003). Nonparametric estimation of sample selection models. *Review of Economic Studies 70*(1), 33–58.

Deaton, A. and N. Cartwright (2018). Understanding and misunderstanding randomized controlled trials. *Social Science and Medicine 210*, 2–21.

Djogbenou, A. A., J. G. MacKinnon and M. Ø. Nielsen (2019). Asymptotic theory and wild bootstrap inference with clustered errors. *Journal of Econometrics 212*(2), 393–412.

Druckenmiller, H. and S. Hsiang (2018). Accounting for unobservable heterogeneity in cross section using spatial first differences. Technical report, National Bureau of Economic Research.

Duranton, G., L. Gobillon and H. G. Overman (2011). Assessing the effects of local taxation using microgeographic data. *Economic Journal 121*(555), 1017–46.

Dustmann, C. and M. E. Rochina-Barrachina (2007). Selection correction in panel data models: An application to the estimation of females' wage equations. *Econometrics Journal 10*(2), 263–93.

Fernández-Val, I. and M. Weidner (2016). Individual and time effects in nonlinear panel models with large *N*, *T*. *Journal of Econometrics 192*(1), 291–312.

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

Goldin, C. (2021). *Career and Family: Women's Century-Long Journey Toward Equity*. Princeton, NJ: Princeton University Press.

Goldin, C. and L. F. Katz (2010). *The Race between Education and Technology*. Cambridge, MA: Harvard University Press.

Hansen, B. E. and S. Lee (2019). Asymptotic theory for clustered samples. *Journal of Econometrics 210*, 268–90.

Harris, R. and R. Lewis (2001). The geography of North American cities and suburbs, 1900–1950: A new synthesis. *Journal of Urban History 27*(3), 262–92.

Heckman, J. J., H. Ichimura and P. E. Todd (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Review of Economic Studies 64*(4), 605–54.

Heckman, J. J., L. J. Lochner and P. E. Todd (2006). Earnings functions, rates of return and treatment effects: The mincer equation and beyond. In E. A. Hanushek and F. Welch (Eds.), *Handbook of the Economics of Education*, vol. *1*, 307–458. Amsterdam: North-Holland.

Heckman, J. J. and E. Vytlacil (2005). Structural equations, treatment effects, and econometric policy evaluation 1. *Econometrica 73*(3), 669–738.

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.

Jackson, K. T. (1987). *Crabgrass Frontier: The Suburbanization of the United States*. New York: Oxford University Press.

Kessler-Harris, A. (1982). *Out to Work: A History of America's Wage-Earning Women*. New York: Oxford University Press.

Kyriazidou, E. (1997). Estimation of a panel data sample selection model. *Econometrica 65*(6), 1335–64.

Lee, M.-J. (2001). First-difference estimator for panel censored-selection models. *Economics Letters 70*(1), 43–9.

Lewis, R. (2002). The changing fortunes of American central-city manufacturing, 1870–1950. *Journal of Urban History 28*(5), 573–98.

Miller, E. V. (2018). Gender differences in intercity commuting patterns in the Fox River valley, Illinois, 1912–1936. *Journal of Historical Geography 60*, 89–99.

Newey, W. K. (2009). Two-step series estimation of sample selection models. *Econometrics Journal 12*, S217–S229.

Oster, E. (2019). Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business and Economic Statistics 37*(2), 187–204.

Rochina-Barrachina, M. E. (1999). A new estimator for panel data sample selection models. *Annales d'Economie et de Statistique 55/56*, 153–81.

Ruggles, S., C. A. Fitch, R. Goeken, J. D. Hacker, M. A. Nelson, E. Roberts, M. Schouweiler and M. Sobek (2021). IPUMS ancestry full count data: Version 3.0. Dataset, IPUMS.

Todd, P. E. and K. I. Wolpin (2003). On the specification and estimation of the production function for cognitive achievement. *Economic Journal 113*(485), F3–F33.

White, H. (1984). *Asymptotic Theory for Econometricians*. Orlando, FL: Academic Press.

Wooldridge, J. M. (1995). Selection corrections for panel data models under conditional mean independence assumptions. *Journal of Econometrics 68*(1), 115–32.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*, Cambridge, MA: MIT Press.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Replication Package

Co-editor Petra Todd handled this manuscript.

APPENDIX

Proof of Theorem 2.2: The proof is written conditional on the set of number of individuals in the locations. Thus, when $E(N_j)$ is used, it can be considered as a constant.

The substitution of the true value of Δy_2 in (2.10) yields the following equality

$$\hat{\theta} = \theta + [(\Delta W)'\Delta W]^{-1}[(\Delta W)'\Delta\eta].$$

Let us assume that $y_{2ij} = x'_{ij}\delta + \gamma_{j\alpha} + \gamma_j + \rho\lambda(z'_{ij}\beta + \theta_{j\alpha_k} + \theta_j) + e_{ij}$ with $E(e_{ij}|\xi_{ij}) = 0$. Thus, $\Delta y_{2ij} = \Delta x'_{ij}\delta + \rho\Delta\lambda(z'_{ij}\beta + \theta_{j\alpha} + \theta_j) + \Delta e_{ij}$. Under the identification Assumptions I1 to I3 have

$$\Delta y_{2ij} = \Delta x'_{ij}\delta + \rho\Delta\lambda\left(z'_{ij}\beta\right) + \Delta e_{ij}.$$

The second step regression equation is equivalent to

$$\Delta y_{2ij} = \Delta x'_{ij}\delta + \rho\tilde{\Delta}\lambda\left(z'_{ij}\beta\right) + \Delta\left[\rho\left(\lambda\left(z'_{ij}\beta\right) - \lambda\left(z'_{ij}\hat{\beta}\right)\right) + e_{ij}\right],$$

$\hat{\beta}$ is estimated by maximum likelihood probit in the first step with variance–covariance matrix V_β. For all i, j conditional on z_{ij}, we have the following. Given that $\lambda(.)$ is twice differentiable, the continuous mapping theorem implies that $\lambda(z'_{ij}\beta) - \lambda(z'_{ij}\hat{\beta})$ goes to zero in probability and is asymptotically normal.

If we assume that N_f is the full sample while N is the selected sample.We assume that $N/N_f \to s$. We, therefore, have

$$\sqrt{N_f}\left(\lambda\left(z'_{ij}\beta\right)-\lambda\left(z'_{ij}\hat{\beta}\right)\right)\to^{d}\mathcal{N}\left(0,\Omega_{ij}\right), \tag{A.1}$$

where $\Omega_{ij}=[\lambda'(z'_{ij}\beta)]^2 z'_{ij}V_\beta z_{ij}$.

We are interested in the limiting distribution of $\sqrt{N}(\hat{\theta}-\theta)$.

$$\sqrt{N}(\hat{\theta}-\theta)=N[(\Delta W)'\Delta W]^{-1}\frac{1}{\sqrt{N}}[(\Delta W)'\Delta\eta]$$

$$=\left[\frac{\sum_{i=1}^{N}(\Delta W_i)'\Delta W_i}{N}\right]^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}(\Delta W_i)'\Delta\eta_i.$$

While W_i are i.i.d., ΔW_i are not independent because an individual is allowed to appear in many pairs. The dependence structure is driven by the operator Δ. If Δ is such that each individual appears only in one pair, then the classical central limit theorem (CLT) and law of large number (LLN) could be applied. However, if individuals are allowed to appear in several pairs, then we need to apply CLT and LLN accounting for correlation.

$$\frac{\sum_{i=1}^{N}(\Delta W_i)'\Delta W_i}{N}=\frac{\sum_{j=1}^{J}\sum_{k=1}^{N_j}(\Delta W_{kj})'\Delta W_{kj}}{N}$$

$$=\frac{1}{J}\sum_{j=1}^{J}\frac{1}{E(N_j)}\sum_{k=1}^{N_j}(\Delta W_{kj})'\Delta W_{kj}+o_p(1).$$

Let us consider $Y_j=\frac{1}{E(N_j)}\sum_{k=1}^{N_j}(\Delta W_{kj})'\Delta W_{kj}$, these variables are i.i.d., moreover, note that by Assumption E1(ii) we can apply the law of large $N/J=(N_1+N_2+....+N_J)/J$ converges to $E(N_j)$ as J goes to ∞.

Under the assumption that all second moments of the variables in W exist (Assumption E2), $H_J=\frac{1}{J}\sum_{j=1}^{J}\frac{1}{E(N_j)}\sum_{k=1}^{N_j}Y_j$ is a matrix.

Thus, the law of large number applies to H_J if and only if it applies to all is elements.

Let a_j be a typical element of the matrix $\frac{1}{E(N_j)}\sum_{k=1}^{N_j}Y_j$. Let t and m be two variables from the set of variables forming W. For example, we can consider $t=x_1$ the first column of the random variable x.

If $t\neq m$ then,

$$E|a_j|\leq\frac{1}{E(N_j)}\sum_{k=1}^{N_j}E|\Delta t_{kj}\Delta m_{kj}| \tag{A.2}$$

$$=\frac{1}{E(N_j)}\sum_{k=1}^{N_j}E|(t_{kj}-t_{ij})(m_{kj}-m_{ij})| \tag{A.3}$$

$$\leq\frac{4}{E(N_j)}\sum_{k=1}^{N_j}E|t_{kj}m_{kj}| \tag{A.4}$$

$$\leq\frac{4}{E(N_j)}\sum_{k=1}^{N_j}\sqrt{E(|t_{kj}|^2)E(|m_{kj}|^2)} \tag{A.5}$$

$$\leq M_0 \tag{A.6}$$

with M_0 a constant.

The result is obtained by using successively the triangular inequality, the identical distribution of variable in W, the Cauchy–Schwarz's inequality and the existence of moment up to its fourth (which means that the second moment exists).

If $t = m$ we have,

$$E|a_j| \leq \frac{1}{E(N_j)} \sum_{k=1}^{N_j} E|\Delta t_{kj} \Delta t_{kj}| \tag{A.7}$$

$$= \frac{1}{E(N_j)} \sum_{k=1}^{N_j} E| \left(t_{kj} - t_{ij}\right)^2 | \tag{A.8}$$

$$\leq \frac{2}{E(N_j)} \sum_{k=1}^{N_j} E|t_{kj} t_{ij}| + E\left(t_{kj}^2\right) \tag{A.9}$$

$$\leq \frac{1}{E(N_j)} \sum_{k=1}^{N_j} \left(E|t_{kj}|\right)^2 + E\left(t_{kj}^2\right) \tag{A.10}$$

$$\leq M_0. \tag{A.11}$$

Thus, the LLN implies that

$$\frac{\sum_{i=1}^{N} (\Delta W_i)' \Delta W_i}{N} \rightarrow^p E((\Delta W_{ij})' \Delta W_{ij}) = C^{-1},$$

as J, N go to ∞.

We can also show that

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} (\Delta W_i)' \, \Delta \eta_i = \frac{\rho}{\sqrt{N}} \sum_{i=1}^{N} (\Delta W_i)' \, \Delta \left(\lambda \left(z_{ij}' \beta\right) - \lambda \left(z_{ij}' \hat{\beta}\right)\right) + \frac{1}{\sqrt{N}} \sum_{i=1}^{N} (\Delta W_i)' \, \Delta e_{ij}.$$

We consider $\Lambda_j = \sum_{k=1}^{N_j} (\Delta W_{kj})' \Delta(\lambda(z_{kj}' \beta) - \lambda(z_{kj}' \hat{\beta}))$, and $E_j = \sum_{k=1}^{N_j} (\Delta W_{kj})' \Delta e_{kj}$. Conditional on $\hat{\beta}$, Λ_j, and E_j are i.i.d. random variables. We assume that the number of individuals in a group is i.i.d. with finite mean $E(N_j)$.

Given all locations are assumed to be disjointed,

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} (\Delta W_i)' \Delta \eta_i = \frac{\rho}{\sqrt{N}} \sum_{j=1}^{J} \Lambda_j + \frac{1}{\sqrt{N}} \sum_{j=1}^{J} E_j.$$

We have $E(E_j) = 0$ for each j. Moreover,

$$\begin{aligned} Var\left(E_j\right) &= E\left[\sum_{k=1}^{N_j}\left(\Delta W_{kj}\right)'\Delta e_{kj}\left(\sum_{k=1}^{N_j}\left(\Delta W_{kj}\right)'\Delta e_{kj}\right)'\right] \\ &= E\left[\sum_{k=1}^{N_j}\left(\Delta W_{kj}\right)'\Delta e_{kj}\Delta e_{kj}\left(\Delta W_{kj}\right)\right] \\ &= E\left(N_j\right)E\left[\left(\Delta W_{kj}\right)'\Delta e_{kj}\Delta e_{kj}\left(\Delta W_{kj}\right)\right]. \end{aligned}$$

Under Assumption E2, $Var(E_j)$ is finite, because all variables have up to the fourth moments. Indeed, if we consider a typical element of $E[(\Delta W_{kj})'\Delta e_{kj}\Delta e_{kj}(\Delta W_{kj})]$, formed by the variables t and m,

$$\begin{aligned} E[(t_{kj} - t_{ij})(m_{kj} - m_{ij})\Delta e_{kj}(\Delta W_{kj})] &\leq 4E[t_{kj}m_{kj}(\Delta e_{kj})^2] \\ &\leq 4E|t_{kj}m_{kj}\Delta e_{kj}^2| \\ &\leq 4\sqrt[4]{E(|t_{kj}|^4)E[(\Delta e_{kj})^2]E(|m_{kj}|^4)E[(\Delta e_{kj})^2]} \\ &\leq M_0. \end{aligned}$$

It should be noted that $E[(\Delta e_{kj})^2] = 2E(e_{kj}^2) < \infty$.

Similarly, we can show that $E(\Lambda_j) = O_p(1/\sqrt{N})$ because we assume that $\hat{\beta}$ is an $\sqrt{N}-$ consistent estimator of β. Thus we have,

$$\begin{aligned} Var\left(\Lambda_j\right) &= E\left[\left(\sum_{k=1}^{N_j}\left(\Delta W_{kj}\right)'\Delta\left(\lambda\left(z_{kj}'\beta\right)-\lambda\left(z_{kj}'\hat{\beta}\right)\right)\right)\left(\sum_{k=1}^{N_j}\left(\Delta W_{kj}\right)'\Delta\left(\lambda\left(z_{kj}'\beta\right)-\lambda\left(z_{kj}'\hat{\beta}\right)\right)\right)'\right] + O_p\left(1/N\right) \\ &= E\left[\sum_{k=1}^{N_j}\left(\Delta W_{kj}\right)'\Delta\left(\lambda\left(z_{kj}'\beta\right)-\lambda\left(z_{kj}'\hat{\beta}\right)\right)\left(\Delta W_{kj}\right)'\Delta\left(\lambda\left(z_{kj}'\beta\right)-\lambda\left(z_{kj}'\hat{\beta}\right)\right)'\right] + O_p\left(1/N\right) \\ &= E\left(N_j\right)E\left(\left(\Delta W_{kj}\right)'\Delta\left(\lambda\left(z_{kj}'\beta\right)-\lambda\left(z_{kj}'\hat{\beta}\right)\right)\Delta\left(\lambda\left(z_{kj}'\beta\right)-\lambda\left(z_{kj}'\hat{\beta}\right)\right)\left(\Delta W_{kj}\right)\right) + O_p\left(1/N\right) \\ &= E\left(N_j\right)E\left[\left(\Delta W_{kj}\right)'\Omega_{kj}\left(\Delta W_{kj}\right)\right] + O_p\left(1/N\right). \qquad \text{(A.12)} \end{aligned}$$

We need to show that $E[(\Delta W_{kj})'\Omega_{kj}(\Delta W_{kj})]$, with $\Omega_{kj} = [\lambda'(z_{kj}'\beta)]^2 z_{kj}' V_\beta z_{kj}$ is finite.

A typical element of this matrix is given by, $E[(t_{kj} - t_{ij})\Omega_{kj}(m_{kj} - m_{ij})]$. We can show the following using Cauchy–Schwarz's inequality.

$$\begin{aligned} E[(t_{kj} - t_{ij})\Omega_{kj}(m_{kj} - m_{ij})] &\leq 4E[t_{kj}m_{kj}\Omega_{kj}] \qquad \text{(A.13)} \\ &\leq 4E[|t_{kj}m_{kj}\Omega_{kj}|] \\ &\leq 4\sqrt[4]{E\left(|t_{kj}|^4\right)\left(E\left(\left[\lambda'\left(z_{kj}'\beta\right)\right]^2 z_{kj}'V_\beta z_{kj}\right)\right)^2 E\left(|m_{kj}|^4\right)}. \qquad \text{(A.14)} \end{aligned}$$

It remains to be proven that $E([\lambda'(z_{kj}'\beta)]^2 z_{kj}' V_\beta z_{kj}) < \infty$. The application of the Cauchy–Schwarz's inequality implies,

$$E\left(\left[\lambda'\left(z_{kj}'\beta\right)\right]^2 z_{kj}' V_\beta z_{kj}\right) \leq \sqrt{E\left[\lambda'\left(z_{kj}'\beta\right)\right]^4 E\left[\left(z_{kj}' V_\beta z_{kj}\right)^2\right]} \tag{A.15}$$

$$\leq \sqrt{E\left[\left(z_{kj}' V_\beta z_{kj}\right)^2\right]} < \infty. \tag{A.16}$$

This follows from noting that $|\lambda'(.)| \leq 1$ and the elements of z have up to their fourth moments.

The moment of a typical element $E[(t_{kj} - t_{ij})\Omega_{kj}(m_{kj} - m_{ij})] < \infty$. This proves that the variance is finite.

It is important to notice that conditional W, $\sum_{i=1}^N (\Delta W_i)'\Delta(\lambda(z_{ij}'\beta) - \lambda(z_{ij}'\hat{\beta}))$ and $\sum_{i=1}^N (\Delta W_i)'\Delta e_{ij}$ are independent random variables. Therefore,

$$\frac{1}{\sqrt{N}} \sum_{i=1}^N (\Delta W_i)'\Delta\eta_i \to^d \mathcal{N}(0, \Gamma), \tag{A.17}$$

where $\Gamma = \frac{\rho^2}{s} E[(\Delta W_{ij})'\Omega_{ij}\Delta W_{ij}] + \frac{1}{s} E[(\Delta W_{ij})'\Delta e_{ij}\Delta e_{ij}(\Delta W_{ij})]$.

$$\sqrt{N}(\hat{\theta} - \theta) \to^d \mathcal{N}(0, \Theta), \tag{A.18}$$

with $\Theta = C\Gamma C'$. This proofs the asymptotic normality of our two-step estimator.

We have proven that under Assumptions I1, I2, I3, E1, and E2,

$$\frac{\sum_{i=1}^N (\Delta W_i)'\Delta W_i}{N} \to^p E((\Delta W_1)'\Delta W_1) = C^{-1}.$$

Using similar arguments we can show that

$$\frac{\sum_{i=1}^N (\Delta W_i)'\Delta \eta_i}{N} \to^p E((\Delta W_1)'\Delta \eta) = 0.$$

Which means that $\hat{\theta}$ is a consistent estimator of θ. We have proven the estimator is both consistent and asymptotically normal. This ends the proof of Theorem 2.2. □

Proof of Theorem 2.3: The variance–covariance estimator of the $\hat{\theta}$ is given by

$$V_{twostep} = B(\Delta W)'[\hat{V}_1 + \hat{V}_2](\Delta W)B', \tag{A.19}$$

where $\hat{V}_1 = \frac{1}{\tilde{s}}\Delta\hat{R}\Delta'$ and $\hat{V}_2 = \frac{\hat{\rho}^2}{\tilde{s}}\Delta\hat{D}z\hat{V}_\beta z'\hat{D}\Delta'$ with all unknown parameters replaced by their estimates. Let us show that this is a consistent estimator of the asymptotic variance of $Var(\hat{\theta})$.

$$V_{twostep} = N \times B(\Delta W)'\left[\frac{\hat{V}_1}{N^2} + \frac{\hat{V}_2}{N^2}\right](\Delta W)N \times B'. \tag{A.20}$$

We have shown that under Assumptions I1, I2, I3, E1, and E2,

$$N \times B = \left[\frac{\sum_{i=1}^N (\Delta W_i)'\Delta W_i}{N}\right]^{-1} \to^p [E((\Delta W_1)'\Delta W_1)]^{-1} = C.$$

Note also that

$$(\Delta W)'\left[\frac{\frac{1}{\tilde{s}}\Delta\hat{R}\Delta'}{N^2}+\frac{\frac{\hat{\rho}^2}{\tilde{s}}\Delta\hat{D}z\hat{V}_\beta z'\hat{D}\Delta'}{N^2}\right](\Delta W)$$

$$=(\Delta W)'\left[\frac{\Delta\hat{R}\Delta'}{\tilde{s}N^2}+\frac{\hat{\rho}^2\Delta\hat{D}z\hat{V}_\beta z'\hat{D}\Delta'}{\tilde{s}N^2}\right](\Delta W)$$

$$(\Delta W)'\left[\frac{\Delta\hat{R}\Delta'}{\tilde{s}N^2}+\frac{\hat{\rho}^2\Delta\hat{D}z\hat{V}_\beta z'\hat{D}\Delta'}{\tilde{s}N^2}\right](\Delta W)$$

$$=(\Delta W)'\Delta\left[\frac{\hat{R}}{\tilde{s}N^2}+\frac{\hat{\rho}^2\hat{D}z\hat{V}_\beta z'\hat{D}}{\tilde{s}N^2}\right]\Delta'(\Delta W)$$

$$=(\Delta W)'\Delta\left[\frac{\hat{\rho}^2 I-\hat{\rho}^2\hat{D}}{\tilde{s}N^2}+\frac{\hat{\rho}^2\hat{D}z\hat{V}_\beta z'\hat{D}}{\tilde{s}N^2}\right]\Delta'(\Delta W)$$

$$=\frac{1}{N}(\Delta W)'\Delta\left[\frac{I-\rho^2 D}{\tilde{s}N}+\frac{\rho^2 DzV_\beta z'D}{\tilde{s}N}\right]\Delta'(\Delta W)+o_p(1)$$

$$=\frac{1}{N}(\Delta W)'\Delta\left[\frac{I-\rho^2 D}{\tilde{s}N}+\frac{\rho^2 DzV_\beta z'D}{\tilde{s}N}\right]\Delta'(\Delta W)+o_p(1).$$

Note also that, as N goes to ∞, $N\left(\frac{1}{N}(\Delta W)'\Delta\left[\frac{I-\rho^2 D}{\tilde{s}N}+\frac{\rho^2 DzV_\beta z'D}{\tilde{s}N}\right]\Delta'(\Delta W)+o_p(1)\right)$ goes to Γ in probability.

This implies that $NV_{twostep}$ converges to the asymptotic variance of $\hat{\theta}$. This ends the proof of Theorem 2.3. □