Identification of peer effects via a root estimator☆

Xiaodong Liu

Department of Economics, University of Colorado Boulder, UCB 256, Boulder, CO 80309, USA

HIGHLIGHTS

- A root estimator is proposed to identify peer effects in a linear-in-means model.
- The identification does not rely on variation of group sizes or intransitivity.
- The root estimator is consistent and asymptotic normal under heteroskedasticity.
- The root estimator performs well in finite samples.

ARTICLE INFO

ABSTRACT

By exploiting the correlation structure of individual outcomes in a network, we show that a carefully constructed root estimator can identify peer effects in linear social interaction models, when identification cannot be achieved via variation of group sizes or intransitivity of network connections. We establish the consistency and asymptotic normality of the root estimator under heteroskedasticity, and conduct Monte Carlo experiments to investigate its finite sample performance.

© 2017 Elsevier B.V. All rights reserved.

1. Introduction

Tremendous progress has been made in understanding identification of peer effects since the seminal work by Manski (1993) (see Blume et al., 2011, for a review). When individuals are randomly assigned into groups and social networks are formed within each group, identification of peer effects can be achieved via either variation of group sizes (see, e.g., Lee, 2007; Graham, 2008) or exclusion restrictions based on network topology (Bramoullé et al., 2009). However, if all groups are of the same size and every individual is equally influenced by all the other group members, then peer effects cannot be identified by the above methods. In this paper, we propose a root estimator that can be used to identify peer effects in such situations.

The idea of the root estimator traces back to Ord (1975), where it is used to estimate models of spatial interaction. In a recent paper, Jin and Lee (2012) generalize the original root estimator in Ord

(1975) to estimate a more general class of spatial models. In this paper, we show that a carefully constructed root estimator, based on both linear and quadratic moment conditions of the error term, can identify peer effects in a linear-in-means model with equal-sized groups. We establish the root-n consistency and asymptotic normality of the root estimator under heteroskedasticity, and conduct Monte Carlo experiments to investigate its finite sample performance.

Of course, the usefulness of the root estimator is not limited to the specific network structure considered in this paper. Here, we focus on the linear-in-means model with equal-sized groups for two reasons. First, identification of peer effects in this situation cannot be achieved by existing methods. Second, data with equal-sized groups are not uncommon in the real world. For example, the capacity of college classes are often fixed over time. For a popular class with full enrollment every semester, its students in each semester form equal-sized groups.

The rest of the paper is organized as follows. Section 2 presents the linear-in-means model and discusses its identification issues. Section 3 introduces the root estimator and studies its asymptotic properties. Section 4 provides simulation results on the finite sample performance of the proposed estimator. Section 5

concludes. The proofs are collected in the online appendix. Throughout the paper, let $\mathbf{I}_n$ denote an $n \times n$ identity matrix, $\boldsymbol{\iota}_n$ denote an $n \times 1$ vector of ones, and $\operatorname{diag}_{i=1}^n \{d_i\}$ denote an $n \times n$ diagonal matrix with the ith diagonal element being d_i.

2. Linear-in-means social interaction model

Consider a sample of n *equal-sized* groups with m ($m \geq 2$) individuals in each group. Then, in a linear-in-means social interaction model, the outcome, $y_{i,g}$, of individual i in the gth group is given by

$$y_{i,g} = \lambda_0 \bar{y}_{i,g} + \mathbf{x}_{i,g} \boldsymbol{\beta}_0 + \bar{\mathbf{x}}_{i,g} \boldsymbol{\gamma}_0 + u_{i,g}, \tag{2.1}$$

where $\mathbf{x}_{i,g}$ is a p-dimensional row vector of exogenous individual characteristics, and $u_{i,g}$ is a possibly heteroskedastic error term. In this model, $\bar{y}_{i,g} = \frac{1}{m-1} \sum_{j=1,j\neq i}^m y_{j,g}$ is the average outcome of the individuals (other than i) in the gth group, with its coefficient λ_0 representing the *endogenous effect*; $\bar{\mathbf{x}}_{i,g} = \frac{1}{m-1} \sum_{j=1,j\neq i}^m \mathbf{x}_{j,g}$ is the vector of average characteristics of the individuals (other than i) in the gth group, with its coefficient vector $\boldsymbol{\gamma}_0$ representing *exogenous (contextual) effects*. It has been one of the main interests in the social interaction literature to separately identified endogenous and exogenous peer effects.

In matrix form, model (2.1) can be written as

$$\mathbf{y}_g = \lambda_0 \mathbf{A}_m \mathbf{y}_g + \mathbf{X}_g \boldsymbol{\beta}_0 + \mathbf{A}_m \mathbf{X}_g \boldsymbol{\gamma}_0 + \mathbf{u}_g, \tag{2.2}$$

where $\mathbf{y}_g = (y_{1,g}, \ldots, y_{m,g})'$, $\mathbf{X}_g = (\mathbf{x}'_{1,g}, \ldots, \mathbf{x}'_{m,g})'$, $\mathbf{u}_g = (u_{1,g}, \ldots, u_{m,g})'$, and $\mathbf{A}_m$ is an adjacency matrix given by $\mathbf{A}_m = \frac{1}{m-1}(\boldsymbol{\iota}_m \boldsymbol{\iota}'_m - \mathbf{I}_m)$. We allow for heteroskedasticity of unknown form and assume $\mathbf{u}_g$ are independently distribution across g with $\mathrm{E}(\mathbf{u}_g|\mathbf{X}_g) = \mathbf{0}$ and $\mathrm{E}(\mathbf{u}_g\mathbf{u}'_g|\mathbf{X}_g) = \boldsymbol{\Sigma}_g \equiv \operatorname{diag}_{i=1}^m\{\sigma_{i,g}^2\}$. We assume $|\lambda_0| < 1$. Then, the reduced form of (2.2) is

$$\mathbf{y}_g = (\mathbf{I}_m - \lambda_0 \mathbf{A}_m)^{-1}(\mathbf{X}_g \boldsymbol{\beta}_0 + \mathbf{A}_m \mathbf{X}_g \boldsymbol{\gamma}_0 + \mathbf{u}_g)$$

$$= \sum_{k=1}^\infty \lambda_0^{k-1} \mathbf{A}_m^{k-1}(\mathbf{X}_g \boldsymbol{\beta}_0 + \mathbf{A}_m \mathbf{X}_g \boldsymbol{\gamma}_0 + \mathbf{u}_g). \tag{2.3}$$

In the current literature, identification of peer effects is usually achieved through either variation of group sizes (see, e.g., Lee, 2007; Graham, 2008) or exclusion restrictions based on network topology (Bramoullé et al., 2009).

Lee (2007) show that, for linear-in-means models, endogenous and exogenous effects can be identified if group sizes have sufficient variation. However, this identification strategy does not work in our case as all groups in the sample have the same size.

Bramoullé et al. (2009) show that if the matrices $\mathbf{I}_m, \mathbf{A}_m, \mathbf{A}_m^2$ are linearly independent, then the exogenous characteristics of indirect connections given by $\mathbf{A}_m^2 \mathbf{X}_g$ can be used as instrumental variables for the outcomes of direct connections $\mathbf{A}_m \mathbf{y}_g$ to identify the endogenous effect from the exogenous effect. The linear independence of $\mathbf{I}_m, \mathbf{A}_m, \mathbf{A}_m^2$ is satisfied, if intransitivity exists in a network such that two individuals, who share a common connection/friend, are not directly connected. In our case, the adjacency matrix $\mathbf{A}_m = \frac{1}{m-1}(\boldsymbol{\iota}_m\boldsymbol{\iota}'_m - \mathbf{I}_m)$ corresponds to a complete network where all individuals are directly connected. It is easy to see $\mathbf{A}_m^2 = \frac{1}{m-1}\mathbf{I}_m + \frac{m-2}{m-1}\mathbf{A}_m$ is linearly dependent on $\mathbf{I}_m$ and $\mathbf{A}_m$.

As the linear-in-means model (2.1) with equal-sized groups cannot be identified by the above methods, it is sometimes given as an example of the *reflection problem* (Manski, 1993), referring to the failure to separately identify endogenous and exogenous effects. In the following, we show that this model actually can be identified via a root estimator.

3. Root estimator

3.1. Asymptotic identification

We assume that we observe an independently distributed sample of $(\mathbf{y}_g, \mathbf{X}_g)$ of size n from a population of equal-sized groups. Therefore, in the asymptotic analysis, we keep the group size m fixed and let the number of groups n go to infinity. Let $\mathbf{u}_g(\boldsymbol{\theta}) = \mathbf{y}_g - \lambda\mathbf{A}_m\mathbf{y}_g - \mathbf{X}_g\boldsymbol{\beta} - \mathbf{A}_m\mathbf{X}_g\boldsymbol{\gamma}$, where $\boldsymbol{\theta} = (\lambda, \boldsymbol{\beta}', \boldsymbol{\gamma}')'$. The root estimator of $\boldsymbol{\theta}$ is based on the linear moment functions $\mathbf{f}_{1,g}(\boldsymbol{\theta}) = [\mathbf{X}_g, \mathbf{A}_m\mathbf{X}_g]'\mathbf{u}_g(\boldsymbol{\theta})$, and the quadratic moment function $f_{2,g}(\boldsymbol{\theta}) = \mathbf{u}_g(\boldsymbol{\theta})'\mathbf{A}_m\mathbf{u}_g(\boldsymbol{\theta})$. The quadratic moment function exploits the correlation structure of individual outcomes in a network. Let $\mathbf{f}_{1,\infty}(\boldsymbol{\theta}) = \lim_{n\to\infty} n^{-1}\sum_{g=1}^n \mathrm{E}[\mathbf{f}_{1,g}(\boldsymbol{\theta})]$ and $f_{2,\infty}(\boldsymbol{\theta}) = \lim_{n\to\infty} n^{-1}\sum_{g=1}^n \mathrm{E}[f_{2,g}(\boldsymbol{\theta})]$. For $\boldsymbol{\theta}$ to be asymptotically identified, the moment equations $\mathbf{f}_\infty(\boldsymbol{\theta}) \equiv [\mathbf{f}_{1,\infty}(\boldsymbol{\theta})', f_{2,\infty}(\boldsymbol{\theta})]' = \mathbf{0}$ need to have a unique solution at the true parameter vector $\boldsymbol{\theta}_0 = (\lambda_0, \boldsymbol{\beta}'_0, \boldsymbol{\gamma}'_0)'$ (Definition 5.2 in Davidson and MacKinnon, 1993). As we assume $\mathrm{E}(\mathbf{u}_g|\mathbf{X}_g) = \mathbf{0}$ and $\mathrm{E}(\mathbf{u}_g\mathbf{u}'_g|\mathbf{X}_g) = \boldsymbol{\Sigma}_g \equiv \operatorname{diag}_{i=1}^m\{\sigma_{i,g}^2\}$, it follows that $\mathrm{E}[\mathbf{f}_{1,g}(\boldsymbol{\theta}_0)] = \mathbf{0}$ and $\mathrm{E}[f_{2,g}(\boldsymbol{\theta}_0)] = \operatorname{tr}(\mathbf{A}_m\boldsymbol{\Sigma}_g) = 0$. Hence, $\boldsymbol{\theta}_0$ is a solution of $\mathbf{f}_\infty(\boldsymbol{\theta}) = \mathbf{0}$. What is left to show is that $\boldsymbol{\theta}_0$ is the only solution.

As $\mathbf{A}_m^2 = \frac{1}{m-1}\mathbf{I}_m + \frac{m-2}{m-1}\mathbf{A}_m$, it follows from the reduced form (2.3) that $\mathrm{E}(\mathbf{A}_m\mathbf{y}_g|\mathbf{X}_g)$ is linearly dependent on $\mathbf{X}_g$ and $\mathbf{A}_m\mathbf{X}_g$, such that $\mathrm{E}(\mathbf{A}_m\mathbf{y}_g|\mathbf{X}_g) = \mathbf{X}_g\mathbf{c}_1 + \mathbf{A}_m\mathbf{X}_g\mathbf{c}_2$, where $\mathbf{c}_1, \mathbf{c}_2$ are $p \times 1$ vectors of constants. Then, $\mathbf{f}_{1,\infty}(\boldsymbol{\theta}) = \mathbf{0}$ implies

$$\lim_{n\to\infty} n^{-1}\sum_{g=1}^n \mathrm{E}([\mathbf{X}_g, \mathbf{A}_m\mathbf{X}_g]'[\mathbf{X}_g, \mathbf{A}_m\mathbf{X}_g])$$

$$\times \begin{bmatrix} (\lambda_0 - \lambda)\mathbf{c}_1 + \boldsymbol{\beta}_0 - \boldsymbol{\beta} \\ (\lambda_0 - \lambda)\mathbf{c}_2 + \boldsymbol{\gamma}_0 - \boldsymbol{\gamma} \end{bmatrix} = 0.$$

If $\lim_{n\to\infty} n^{-1}\sum_{g=1}^n \mathrm{E}([\mathbf{X}_g, \mathbf{A}_m\mathbf{X}_g]'[\mathbf{X}_g, \mathbf{A}_m\mathbf{X}_g])$ has full column rank, then the solution of $\mathbf{f}_{1,\infty}(\boldsymbol{\theta}) = \mathbf{0}$ is given by

$$\boldsymbol{\beta} = \boldsymbol{\beta}_0 + (\lambda_0 - \lambda)\mathbf{c}_1 \quad \text{and} \quad \boldsymbol{\gamma} = \boldsymbol{\gamma}_0 + (\lambda_0 - \lambda)\mathbf{c}_2. \tag{3.1}$$

Substitution of (3.1) into $f_{2,\infty}(\boldsymbol{\theta}) = 0$ gives

$$(\lambda_0 - \lambda)\lim_{n\to\infty} n^{-1}\sum_{g=1}^n 2\operatorname{tr}(\mathbf{A}_m\mathbf{G}_m\boldsymbol{\Sigma}_g)$$

$$+ (\lambda_0 - \lambda)^2 \lim_{n\to\infty} n^{-1}\sum_{g=1}^n \operatorname{tr}(\mathbf{G}_m\mathbf{A}_m\mathbf{G}_m\boldsymbol{\Sigma}_g) = 0, \tag{3.2}$$

where $\mathbf{G}_m = \mathbf{A}_m(\mathbf{I}_m - \lambda_0\mathbf{A}_m)^{-1}$. Eq. (3.2) is quadratic in λ, and has two roots given by

$$\lambda = \lambda_0 + \frac{\lim_{n\to\infty} n^{-1}\sum_{g=1}^n \operatorname{tr}(\mathbf{A}_m\mathbf{G}_m\boldsymbol{\Sigma}_g)}{\lim_{n\to\infty} n^{-1}\sum_{g=1}^n \operatorname{tr}(\mathbf{G}_m\mathbf{A}_m\mathbf{G}_m\boldsymbol{\Sigma}_g)} \pm \frac{\left|\lim_{n\to\infty} n^{-1}\sum_{g=1}^n \operatorname{tr}(\mathbf{A}_m\mathbf{G}_m\boldsymbol{\Sigma}_g)\right|}{\lim_{n\to\infty} n^{-1}\sum_{g=1}^n \operatorname{tr}(\mathbf{G}_m\mathbf{A}_m\mathbf{G}_m\boldsymbol{\Sigma}_g)}.$$

To know which root is consistent, i.e., $\lambda = \lambda_0$, we need to know the sign of $\lim_{n\to\infty} n^{-1}\sum_{g=1}^n \operatorname{tr}(\mathbf{A}_m\mathbf{G}_m\boldsymbol{\Sigma}_g)$. As $|\lambda_0| < 1$ and $m \geq 2$, it follows that $\lim_{n\to\infty} n^{-1}\sum_{g=1}^n \operatorname{tr}(\mathbf{A}_m\mathbf{G}_m\boldsymbol{\Sigma}_g) = \lim_{n\to\infty} n^{-1}\sum_{g=1}^n (1-\lambda_0)^{-1}(m+\lambda_0-1)^{-1}\sum_{i=1}^m \sigma_{i,g}^2 > 0$. Hence, the consistent root is given by

$$\lambda = \lambda_0 + \frac{\lim_{n\to\infty} n^{-1}\sum_{g=1}^n \operatorname{tr}(\mathbf{A}_m\mathbf{G}_m\boldsymbol{\Sigma}_g)}{\lim_{n\to\infty} n^{-1}\sum_{g=1}^n \operatorname{tr}(\mathbf{G}_m\mathbf{A}_m\mathbf{G}_m\boldsymbol{\Sigma}_g)} - \frac{\left|\lim_{n\to\infty} n^{-1}\sum_{g=1}^n \operatorname{tr}(\mathbf{A}_m\mathbf{G}_m\boldsymbol{\Sigma}_g)\right|}{\lim_{n\to\infty} n^{-1}\sum_{g=1}^n \operatorname{tr}(\mathbf{G}_m\mathbf{A}_m\mathbf{G}_m\boldsymbol{\Sigma}_g)} = \lambda_0.$$

$$\tag{3.3}$$

With λ uniquely determined by (3.3), $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ can be identified by (3.1).