espol Facultad de
Ciencias Sociales y Humanísticas

Endogeneity in the Linear-In-Means Model

A thesis presented for the degree of

Magíster en Ciencias Económicas

by

Pablo Andrés Estrada Cedeño

Guayaquil - Ecuador
September 18, 2021

Dedication

This thesis work is dedicated to my parents,

they have inspired me to always give the best that I have.

Pablo Estrada

Acknowledgements

Special thanks to Leonardo Sánchez, David Jacho, and Juan Estrada
for their invaluable contributions and comments.

Pablo Estrada

EVALUATION COMMITTEE

David Jacho Chávez, Ph.D.

Reviewer 1

Daniel Lemus, Ph.D.

Reviewer 2

Leonardo Sánchez Aragón, Ph.D.

Thesis Supervisor

DECLARACIÓN EXPRESA

“La responsabilidad del contenido de este Trabajo de Titulación, corresponde exclusivamente al autor, y al patrimonio intelectual de la misma **ESCUELA SUPERIOR POLITECNICA DEL LITORAL**”

PABLO ANDRES
ESTRADA
CEDENO

Digitally signed by PABLO
ANDRES ESTRADA CEDENO
Date: 2021.09.18 01:35:33
-04'00'

Pablo Estrada

Contents

Abstract

Linear-in-means models are widely used in different contexts to estimate peer effects. In these models, there are two potential sources of endogeneity: in the interaction network and the individual's characteristics. This paper proposes a General Three-Stage Least Square estimation modified to account for the endogeneity of the network and covariates in the linear-in-means model. The new procedure, called G3SLSX, modifies the G3SLS (Estrada et al., 2021) to recover the social and direct effects using a predetermined network and an exogenous variable as instrument. The Monte Carlo experiments show that G3SLSX has similar performance as G3SLS for the social effects. For the direct effects, G3SLSX outperforms G3SLS in the case of over-identification.

Keywords: Instrumental Variables; Linear-in-Means Models; Multiplex Networks

List of Figures

List of Tables

1 Introduction

The last two decades have witnessed a huge growth in the study of peer effects. In his seminal paper, Manski (1993) argues that peers may influence others because of three reasons: correlated effects, exogenous effects, and endogenous effects. Correlated effects refer to unobserved characteristics that are related to group selection. Exogenous effects point out to characteristics that relate to the background of the social group. On the other hand, the endogenous effects are related to the same outcome of their peers.

A common approach to estimate peer effects is to use the so called linear-in-means model (Equation 1). In Equation 1, the parameter β is a scalar that represents the peer effects, δ is a $k \times 1$ vector that indicates the contextual effects, and γ is a $k \times 1$ vector for the direct effects.

$$\mathbf{y} = \alpha\iota + \beta\mathbf{W}\mathbf{y} + \mathbf{W}\mathbf{X}\delta + \mathbf{X}\gamma + \mathbf{v} \tag{1}$$

Empirical applications of peer effects estimations are found in studies of criminal activities, student achievement, health, the workplace and the house market. For instance, Szumilo (2020) uses the linear-in-means model to find neighborhood price spillovers in the housing market. He analyzes how prices of buildings change when their own fundamental characteristics remain constant, but average prices of their neighborhoods change. Mas and Moretti (2009) study the productivity of a worker and how it is affected by his peers. Tumen and Zeydanli (2016) investigate spillovers in job satisfaction in UK. They estimate the correlation between the group-level and individual-level job satisfaction scores. Also, Advani and Malde (2018) provides a comprehensive survey of peer effects.

Researchers face two problems when dealing with the consistent estimation of peer effects. First, the problem of *correlated effects*, and second, the *reflection problem* associated with the endogenous effects. Depending on the context, the researcher can separate social effects from correlated effects when the agents do not have control over group selection. On the other hand, the challenge of the reflection problem consists of separating the effect of peers' outcomes from the peers' background characteristics.

In the linear-in-means model, Bramoullé et al. (2009) provide the network's characteristics to identify endogenous and exogenous effects. They start by assuming that there is no problem of correlated effects, in other words, the network in which agents interact is not related to unobserved characteristics. Then, they show that under intransitivity, peer effects and contextual effects are identified. They propose that the friends of my friends' outcomes and backgrounds provide an instrument to identify peer effects (Sacerdote, 2014).

The problem of correlated effects arises with the presence of endogenous networks or common shocks. Bramoullé et al. (2020) mentions four strategies to deal with this problem: random peers, random shocks, structural endogeneity and panel data. Natural experiments, where peers are randomly allocated, allow for the identification of the social causal effects. However, if peers are not random, researchers can use different sources of exogenous variation. For instance, they can combine randomized interventions and econometric methods for peer effects.

In some situations, natural experiment strategies are not feasible to obtain an exogenous network. Instead, we can use predetermined networks to recover social and contextual effects. In a recent work, Estrada et al. (2021) assumes that the adjacency matrix $\mathbf{W}$ is endogenous, and that there is a predetermined adjacency matrix $\mathbf{W_0}$. Instead of assuming that $E[\mathbf{v}|\mathbf{X}, \mathbf{W}] = 0$ as Bramoullé et al. (2009), they assume that $E[\mathbf{v}|\mathbf{X}, \mathbf{W_0}] = 0$. Using a linear projection of $\mathbf{W_0}$ on $\mathbf{W}$, they recover the peer effect β and contextual effects δ.

This paper aims to consistently estimate social and direct effects in the linear-in-means model with the most general case of endogeneity, when the network $\mathbf{W}$ and the background characteristics $\mathbf{X}$ are endogenous. Despite the interest in the estimation of peer effects, no one as far as we know has studied this problem. We show that by using an exogenous network $\mathbf{W_0}$ and variable $\mathbf{Z_2}$ as instruments for the endogenous network $\mathbf{W}$ and covariate $\mathbf{X_2}$, we can recover the parameters of the linear-in-means model.

We propose the G3SLSX procedure as a solution for the general case of endogeneity in the linear-in-means model. This procedure is based on the G3SLS from Estrada et al. (2021) and modifies the first step with a double linear projection of $\mathbf{WS}$ on $\mathbf{W_0S_{iv}}$ and $\mathbf{X}$ on $\mathbf{X_{iv}}$. Then, under two different Monte Carlo experiments, we showed that the proposed estimator performs as good as the G3SLS procedure and in some cases outperforms it.

This paper begins by examining the sources of endogeneity in the linear-in-means model (Section 2). Then, it describes the G3SLSX procedure (Section 3). Section 4 describes the Monte Carlo experiments and results. Finally, Section 5 show the conclusion.

2 Sources of Endogeneity

In the linear-in-means model, there are two possible sources of endogeneity: the interaction network $\mathbf{W}$ and the agent's characteristics $\mathbf{X}$. This leads to identify four cases.

- Case 1: $\mathbf{W}$ and $\mathbf{X}$ are both exogenous. In this case, Bramoullé et al. (2009) use a G2SLS procedure to estimate the social effects.

- Case 2: $\mathbf{W}$ is endogenous and $\mathbf{X}$ is exogenous. If a predetermined exogenous network $\mathbf{W_0}$ is available, we can use the G3SLS procedure proposed by Estrada et al. (2021).

- Case 3: $\mathbf{W}$ is exogenous and $\mathbf{X}$ is endogenous. Including an instrument $\mathbf{Z}$ for the endogenous regressor $\mathbf{X}$ in the G2SLS procedure will estimate consistently the direct effects.

- Case 4: $\mathbf{W}$ and $\mathbf{X}$ are both endogenous. Using an exogenous network $\mathbf{W_0}$ and instrument $\mathbf{Z}$, a modified G3SLS estimation can recover the social and direct effects.

In case 1, the network $\mathbf{W}$ and regressor $\mathbf{X}$ are both exogenous. Therefore, $E[\mathbf{v}|\mathbf{X}, \mathbf{W}] = 0$, i.e., there are no correlated effects. Bramoullé et al. (2009) assume that the network shows intransitivity to identify social effects, meaning that $\mathbf{I}, \mathbf{W}$, and $\mathbf{W^2}$ are linearly independent. Then, they estimate the model via a G2SLS procedure. First, equation 1 is estimated using

2SLS with the instrument set $\mathbf{Z} = [\iota, \mathbf{W^2X}, \mathbf{WX}, \mathbf{X}]$. Second, with the social parameters obtained, they calculate the optimal instrument matrix $\mathbf{Z}^* = [\iota, E[\mathbf{Wy}|\mathbf{X}, \mathbf{W}], \mathbf{WX}, \mathbf{X}]$, and estimate a 2SLS procedure in the linear-in-means model.

Estrada et al. (2021) propose a G3SLS procedure that estimates social effects in the second case, when $\mathbf{W}$ is endogenous and $\mathbf{X}$ is exogenous. The first step consists of estimating by OLS: $\mathbf{WS} = \mathbf{W_0S}\Pi + U$, where $\mathbf{S} = [\mathbf{y}, \mathbf{X}]$. In the second step, they estimate by 2SLS: $\mathbf{y} = \mathbf{D_0}\psi^* + \mathbf{e}$, where $\mathbf{D_0} = [\iota, \mathbf{X}, \mathbf{WS}]$, $\psi^* = (\alpha, \gamma, \theta^*)$, and $\theta^* = (\beta^*, \delta^*)$. In this step, the instrument matrix is defined as $\mathbf{Z} = [\iota, \mathbf{X}, \mathbf{W_0^2X}, \mathbf{W_0X}]$. Finally, in the third step, they estimate by 2SLS: $\mathbf{y} = \mathbf{D}\psi^* + \mathbf{e}$, using the optimal instrument matrix $\mathbf{Z}^* = [\iota, E[\mathbf{W_0y}|\mathbf{X}, \mathbf{W_0}], \mathbf{W_0X}, \mathbf{X}]\widehat{\Gamma}$, where $\widehat{\Gamma} = [I_{k+1}, \widehat{\Pi}]$.

For case 3 and 4, the matrix $\mathbf{X}$ is endogenous. We will focus on case 4 that is the most general case of endogeneity. We can solve the problem of endogeneity using instruments for both $\mathbf{W}$ and $\mathbf{X}$. First, we rewrite the linear-in-means model in matrix form using $\mathbf{S} = [\mathbf{y}, \mathbf{X}]$ as follows:

$$\mathbf{y} = \alpha\iota + \mathbf{WS}\theta + \mathbf{X}\gamma + \mathbf{v} \tag{2}$$

In equation 2, the $n \times n$ network $\mathbf{W}$ and $n \times k$ matrix $\mathbf{X}$ are endogenous so that $E[\mathbf{v}|\mathbf{X}] \neq 0$ and $E[\mathbf{v}|\mathbf{W}] \neq 0$. However, we can define an $n \times k_1$ exogenous regressor $\mathbf{X_1}$, $n \times k_2$ endogenous regressor $\mathbf{X_2}$, and its $n \times l_2$ instrument $\mathbf{Z_2}$. So we construct the $n \times k$ matrix $\mathbf{X} = [\mathbf{X_1}, \mathbf{X_2}]$ and the $n \times l$ matrix $\mathbf{X_{iv}} = [\mathbf{X_1}, \mathbf{Z_2}]$, where $k = k_1 + k_2$ and $l = k_1 + l_2$. Also, we define the endogenous network $\mathbf{W}$ and the predetermined network $\mathbf{W_0}$. Assumptions 2, 2, and 2 allow to recover the structural parameters of equation 2.

There exists an $n \times n$ network matrix $\mathbf{W_0}$ and an $n \times l$ instrument matrix $\mathbf{X_{iv}}$ such that $E[\mathbf{v}|\mathbf{X_{iv}}, \mathbf{W_0}] = 0$

Let Λ be the $l \times k$ matrix of coefficients from the regression,

$$\mathbf{X} = \mathbf{X_{iv}}\boldsymbol{\Lambda} + \mathbf{U_2} \tag{3}$$

where the $n \times k$ matrix of errors $\mathbf{U_2}$ is such that $E[\mathbf{U_2}|\mathbf{X_{iv}}] = \mathbf{O}$, $E[\mathbf{X_{iv}^T X_{iv}}]$ is positive definite, and rank$(\mathbf{\Lambda}) = k$.

Let Π be the $(l+1) \times (k+1)$ matrix of coefficients from the regression,

$$\mathbf{WS} = \mathbf{W_0 S_{iv} \Pi} + \mathbf{U_1} \tag{4}$$

where the $n \times (k+1)$ matrix of errors $\mathbf{U_1}$ is such that $E[\mathbf{U_1}|\mathbf{W_0 y}, \mathbf{W_0 X_{iv}}, \mathbf{X_{iv}}] = \mathbf{O}$, $E[\mathbf{S_{iv}^T W_0^2 S_{iv}}]$ is positive definite, and rank$(\mathbf{\Pi}) = k+1$.

Assumptions 2, 2 and 2 are a modified version of those in Estrada et al. (2021). Assumption 2 includes not only the existence of an exogenous network, but also a variable that will serve as instrument. Assumption 2 is added to provide the identification of the parameter γ. Finally, assumption 2 provide the necessary conditions to recover the structural parameters of equation 2 with the modification that matrix $\mathbf{S_{iv}}$ is used instead of $\mathbf{S}$. A necessary condition for the identification of $\mathbf{\Lambda}$ is that $l \geq k$. Under these assumptions, the reduced form of the general case is:

$$\mathbf{y} = \alpha\iota + \mathbf{W_0 S_{iv}}\theta^* + \mathbf{X_{iv}}\gamma^* + \mathbf{e} \tag{5}$$

Equation 5 represents the reduced form of the fourth case described above, where we use an exogenous network $\mathbf{W_0}$ and an exogenous regressor $\mathbf{Z_2}$ to instrument the endogenous network $\mathbf{W}$ and regressor $\mathbf{X_2}$. In this equation, we know that $\mathbf{e} = \mathbf{U_1}\theta + \mathbf{U_2}\gamma + \mathbf{v}$, $\theta^* = \mathbf{\Pi}\theta$, and $\gamma^* = \mathbf{\Lambda}\gamma$.

With assumption 2 we are assuming that $\mathbf{Z_2}$ is a good instrument of $\mathbf{X_2}$. Additionally, assumption 2 implies that $\mathbf{W_0 Z_2}$ is a good instrument of $\mathbf{WX_2}$. Both implications allow us to recover the direct effects and the contextual effects, respectively. Finally, the social and direct effects in the linear-in-means model are estimated using a modified version of the G3SLS procedure. This procedure is described in detail in section 3.

3 The G3SLSX Procedure

In order to identify the social parameter θ, we need to modify the three-step estimation of Estrada et al. (2021) so it accounts for the endogeneity of $\mathbf{X_2}$. In the first step, we estimate the projection of $\mathbf{WS}$ on $\mathbf{W_0S_{iv}}$ and $\mathbf{X}$ on $\mathbf{X_{iv}}$. In the second step, we estimate by 2SLS the regression coefficients of $\mathbf{y}$ on $\mathbf{D_0}$ using as instruments $\mathbf{Z}$. Finally, in the third step, we calculate by 2SLS the regression coefficients of $\mathbf{y}$ on $\mathbf{D}$ using the optimal instrument matrix $\mathbf{Z}^*$.

Step 1

We calculate this step by OLS. Specifically, the estimator is:

$$\widehat{\mathbf{\Pi}} = (\mathbf{S_{iv}^T W_0^2 S_{iv}})^{-\mathbf{1}} \mathbf{S_{iv}^T W_0 WS} \tag{6}$$

Also,

$$\mathbf{X} = \mathbf{X_{iv}\Lambda} + \mathbf{U_2} \tag{7}$$

We also calculate this step by OLS:

$$\widehat{\mathbf{\Lambda}} = (\mathbf{X_{iv}^T X_{iv}})^{-\mathbf{1}} \mathbf{X_{iv}^T X} \tag{8}$$

The projection matrices $\widehat{\mathbf{\Pi}}$ and $\widehat{\mathbf{\Lambda}}$ will be used in step 2 and 3 to recover the structural parameters. The projection matrix $\widehat{\mathbf{\Pi}}$ differs from G3SLS since we project $\mathbf{WX}$ on $\mathbf{W_0X_{iv}}$ rather than $\mathbf{W_0X}$. This step relies on having consistent and strongly statistically significant projection matrices. Otherwise, we run into the problem of weak instruments (Stock and Yogo, 2005).

Weak instruments emerge when the instruments are weakly correlated with the endogenous variable. For the projection of $\mathbf{X}$ on $\mathbf{X_{iv}}$, we can use the common rule of thumb to avoid weak instruments. This rule of thumb verifies that the first-stage regression has an F-statistic of at least 10. Stock and Yogo (2005) suggested that having at F-statistic that exceeds 10 allows for reliance on 2SLS estimations. On the other hand, we should be cautious when using this rule on the regression of $\mathbf{WX}$ on $\mathbf{W_0X_{iv}}$.

Step 2

$$\mathbf{y} = \mathbf{D_0}\psi^* + \mathbf{e} \tag{9}$$

where $\mathbf{D_0} = [\iota, \mathbf{X_{iv}}, \mathbf{W_0S_{iv}}]$ and using $\mathbf{Z} = [\iota, \mathbf{X_{iv}}, \mathbf{W_0^2X_{iv}}, \mathbf{W_0X_{iv}}]$ as instrument. The vector ψ^* contains $[\alpha, \gamma^*, \theta^*]^T$.

$$\widehat{\psi}^*_{2SLS} = (\mathbf{D_0^TZ}(\mathbf{Z^TZ})^{\mathbf{-1}}\mathbf{Z^TD_0})^{\mathbf{-1}}\mathbf{D_0^TZ}(\mathbf{Z^TZ})^{\mathbf{-1}}\mathbf{Z^Ty} \tag{10}$$

Finally, we obtain the social parameter $\widehat{\theta} = (\widehat{\mathbf{\Pi}}^{\mathbf{T}}\widehat{\mathbf{\Pi}})^{-1}\widehat{\mathbf{\Pi}}^{\mathbf{T}}\widehat{\theta}^*$ and the direct effects $\widehat{\gamma} = (\widehat{\mathbf{\Lambda}}^{\mathbf{T}}\widehat{\mathbf{\Lambda}})^{-1}\widehat{\mathbf{\Lambda}}^{\mathbf{T}}\widehat{\gamma}^*$.

Step 3

$$\mathbf{y} = \mathbf{D}\psi + \mathbf{e} \tag{11}$$

where $\mathbf{D} = [\iota, \mathbf{X}, \mathbf{WS}]$. Define $\widehat{\mathbf{D}} = [\iota, \mathbf{X_{iv}}\widehat{\mathbf{\Lambda}}, \mathbf{W_0S_{iv}}\widehat{\mathbf{\Pi}}]$. We calculate the optimal instrument:

$$\mathbf{E}[\mathbf{W_0y}|\mathbf{X_{iv}}, \mathbf{W_0}] = \mathbf{W_0}[\mathbf{I} - (\widehat{\pi}_{\mathbf{1}}\widehat{\theta})\mathbf{W_0}]^{\mathbf{-1}}\mathbf{D_x}\widehat{\psi}^*_{\mathbf{x}} \tag{12}$$

where $\mathbf{D_x} = [\iota, \mathbf{X_{iv}}, \mathbf{W_0X_{iv}}]$, and $\widehat{\psi}^*_{\mathbf{x}} = [\widehat{\alpha}, \widehat{\mathbf{\Lambda}}\widehat{\gamma}, \widehat{\pi}_{\mathbf{2}}\widehat{\theta}]$. Also, $\widehat{\pi}_1$ is the $1 \times (k+1)$ first row vector of the matrix $\widehat{\mathbf{\Pi}}$, and $\widehat{\pi}_2$ is the $l \times (k+1)$ partition of the matrix $\widehat{\mathbf{\Pi}}$. Using the optimal instrument, we construct the matrix to estimate the G2SLS as $\mathbf{Z}^* = [\iota, \mathbf{X_{iv}}, \mathbf{E}[\mathbf{W_0y}], \mathbf{W_0X_{iv}}]$. The main difference from Estrada et al. (2021) is the inclusion of $\widehat{\mathbf{\Lambda}}$ to recover $\widehat{\psi}^*_x$. In the case where $\mathbf{X_{iv}} = \mathbf{X}$, the matrix $\widehat{\mathbf{\Lambda}}$ is the identity matrix and the estimation is equivalent to the G3SLS.

Finally, to assess the performance of this procedure, we will evaluate it using two different Monte Carlo experiments described in the next section.

4 Monte Carlo

The Monte Carlo simulations follow two different data generating processes (DGPs). Design 1 presents unobserved degree heterogeneity in the network formation process (Graham, 2017). On the other hand, in Design 2, outcome and network are jointly determined through unobserved characteristics with homophily (Estrada et al., 2021).

For design 1, links are formed according to the rule $w_{ij}(\psi) = I\left\{x_i x_j + \psi(a_i + a_j) - u_{ij} \geq 0\right\}$ where ψ is a switch that activates the individual-level degree heterogeneity a_i. The variable u_{ij} comes from a logistic distribution with mean zero and scale parameter one. The exogenous variable a_i takes values of 1 or -1 with probability 0.5. Lastly, x_i is the observable exogenous characteristic.

For design 2, the links are formed based on the following rule. If $\varepsilon^*_{3;i} > \Phi^{-1}(0.95)$, then $w_{ij} = I\left[\left|\varepsilon^*_{3;i} - \varepsilon^*_{3;j}\right| < \widehat{F}^{-1}_{\varepsilon^*_3}(0.95)\right] \times (1 - w_{0;i,j}) + w_{0;i,j}$. Instead, if $\varepsilon^*_{3;i} < \Phi^{-1}(0.05)$, then $w_{ij} = I\left[\left|\varepsilon^*_{3;i} - \varepsilon^*_{3;j}\right| < \widehat{F}^{-1}_{\varepsilon^*_3}(0.95)\right] \times w_{0;i,j}$. Otherwise, $w_{ij} = w_{0;i,j}$. The variable $\varepsilon^*_{3;i}$ is the idiosyncratic error related to homophily. The key idea is that agents with larger values of ε_3 will be more likely to keep connections with similar agents that have larger values of ε_3.

The simulations generate 1,200 repetitions with $n \in \{50, 100, 200\}$. The number of endogenous variables x is $k = 1$ and the number of instruments x_{iv} is $l \in \{1, 2\}$. Thus, we can evaluate the estimator on the cases of just-identification ($l = 1$) and over-identification ($l = 2$). The exogenous variable x_{iv} was generated from a normal distribution with mean $\mu = 0$ and variance $\sigma = 3$. The endogenous variable x is allowed to be correlated to x_{iv} and the error e.

The parameter m that measures the importance of degree heterogeneity and homophily is set to $m \in \{10, 15\}$ for design 1, and $m \in \{2, 5\}$ for design 2. The higher the value of m, the stronger the endogeneity of the network $\mathbf{W}$. The density is set at $d = 0.01$ to allow for sparse networks. And the structural parameters are set to $\beta = 0.7$, $\delta = 1$, $\gamma = 1$, and $\lambda_l = 1/l$.

4.1 Discussion

The G3SLSX procedure accounts for the endogeneity of the network $\mathbf{W}$ and the variable $\mathbf{X}$. This procedure modifies the G3SLS introducing an instrument $\mathbf{Z_2}$ for the endogenous variables $\mathbf{X_2}$. In this section, we show that G3SLSX works as good as G3SLS and in some cases outperforms it. Specifically, in the case of over-identification, when $rank(\mathbf{X_{iv}}) > rank(\mathbf{X})$, the G3LSX consistently estimates the direct/background effects for the homophily design. In the case of the heterogeneity design, we show that G3SLSX is more efficient since it has a lower variance.

Figure 1 shows the Monte Carlo experiments for design 1. The box plots show the simulations for a sample size of $n \in \{100, 200\}$ and the parameter $m = 15$. When both $\mathbf{X_{iv}}$ and $\mathbf{X}$ have one variable, i.e., it is just-identified, social effects (peer and contextual) are consistent for G3SLS and G3SLSX, and both show a similar performance. For the background or direct effects, G3SLSX does a slightly better work than G3SLS.

For the case of over-identification, G3SLS and G3SLSX again consistently estimate social effects. In the case of the direct effects, even though both G3SLS and G3SLSX perform well, the latter is more efficient because of its lower variance. Finally, when the variable $\mathbf{X}$ is exogenous, both G3SLS and G3SLSX correctly estimate social effects, but G3SLS presents a lower variance in this case. We also calculated descriptive statistics for all the combinations of parameters in design 1. Tables 1, 2, and 3 in the Appendix show the mean, median, and standard deviation for all the experiments with unobserved heterogeneity.

Figure 2 shows the results of the experiments for design 2. The box plots show the results of the simulations with sample size of $n \in \{100, 200\}$ and the parameter $m = 5$. Regardless of the number of variables for X_{iv}, i.e., $l = 1$ or $l = 2$, G3SLS and G3SLSX show similar performance. However, standard deviations are higher for G3SLSX than G3SLS when $l = 1$.

For the background effects, G3SLSX outperforms the other two estimators. In the case of just-identification, even though G2SLS works better than G3SLS, the estimator proposed (G3SLSX) outperforms the previous two. For the over-identification case, G3SLSX clearly better estimates the direct effects compared to G2SLS and G3SLS. When the variable $\mathbf{X}$ is exogenous, as expected, G3SLSX has a higher variability due to the extra step that incurs. In the Appendix, tables 4, 5, and 6 show the mean, median, and standard deviation for all the experiments with homophily-related unobserved characteristics.

5 Conclusion

In this paper, we have presented a modification to the G3SLS procedure proposed by Estrada et al. (2021). This modified procedure, called G3SLSX, allows for the estimation of social and direct effects in the presence of endogenous networks and covariates in the linear-in-means model. Previous methodologies have focused on the reflection problem (Bramoullé et al., 2009) and correlated effects (Estrada et al., 2021). This paper goes a step forward, proposing a procedure that accounts for the most general case of endogeneity.

We showed that G3SLSX performs well in two different Monte Carlo designs with individual degree of heterogeneity and homophily. For both designs, we presented performance under over-identification and just-identification. Background effects are consistently estimated only with G3SLSX when the network presents homophily in the over-identification case.

We are aware that the main limitation is the reliability of $\mathbf{Z_2}$ and $\mathbf{W_0Z_2}$ as good instruments for $\mathbf{X_2}$ and $\mathbf{WX_2}$. In the case of weak instruments, the procedure could lead to inconsistent estimations. Further work will look into tests to address weak instruments in these contexts.

References

Advani, A. and Malde, B. (2018). Methods to identify linear network models: a review. *Swiss Journal of Economics and Statistics*, 154(1):12.

Bramoullé, Y., Djebbari, H., and Fortin, B. (2009). Identification of peer effects through social networks. *Journal of Econometrics*, 150:41–55.

Bramoullé, Y., Djebbari, H., and Fortin, B. (2020). Peer Effects in Networks: A Survey. *Annual Review of Economics*, 12(1):603–629. _eprint: https://doi.org/10.1146/annurev-economics-020320-033926.

Estrada, J., Huynh, K. P., Jacho-Chavez, D. T., and Sanchez-Aragon, L. (2021). On the Identification and Estimation of Endogenous Peer Effects in Multiplex Networks.

Graham, B. S. (2017). An Econometric Model of Network Formation With Degree Heterogeneity. *Econometrica*, 85(4):1033–1063.

Manski, C. F. (1993). Identification of Endogenous Social Effects: The Reflection Problem. *The Review of Economic Studies*, 60(3):531–542. Publisher: [Oxford University Press, Review of Economic Studies, Ltd.].

Mas, A. and Moretti, E. (2009). Peers at Work. *American Economic Review*, 99(1):112–145.

Sacerdote, B. (2014). Experimental and Quasi-Experimental Analysis of Peer Effects: Two Steps Forward? *Annual Review of Economics*, 6(1).

Stock, J. and Yogo, M. (2005). *Identification and Inference for Econometric Models*. Cambridge University Press, New York. Pages: 80-108.

Szumilo, N. (2020). Prices of Peers: Identifying Endogenous Price Effects in the Housing Market. *The Economic Journal*, (ueaa129).

Tumen, S. and Zeydanli, T. (2016). Social interactions in job satisfaction. *International Journal of Manpower*, 37(3):426–455. Publisher: Emerald Group Publishing Limited.

Tables

Table 1: Peer effects ($\beta = 0.7$) for simulations with heterogeneity.

				OLS	G2SLS	G3SLS	G3SLSX
n	l	m					
50	1	10	mean	1.056	1.324	0.836	0.807
			median	1.051	1.374	0.858	0.883
			std	0.054	0.630	1.351	1.456
		15	mean	1.062	1.451	0.901	0.876
			median	1.057	1.426	0.929	0.916
			std	0.058	0.531	1.523	1.562
	2	10	mean	1.060	1.345	0.877	0.770
			median	1.056	1.395	0.866	0.837
			std	0.060	0.674	1.393	1.353
		15	mean	1.064	1.429	0.870	0.898
			median	1.060	1.441	0.919	0.879
			std	0.063	0.662	1.572	1.324
100	1	10	mean	1.048	1.337	0.738	0.737
			median	1.047	1.413	0.800	0.815
			std	0.036	0.675	1.057	1.119
		15	mean	1.053	1.500	0.924	0.825
			median	1.051	1.490	0.941	0.919
			std	0.039	0.452	1.245	1.187
	2	10	mean	1.052	1.438	0.690	0.748
			median	1.049	1.459	0.746	0.767
			std	0.038	0.606	1.082	1.167
		15	mean	1.056	1.597	0.743	0.773
			median	1.052	1.517	0.821	0.793
			std	0.039	0.501	1.199	1.331
200	1	10	mean	1.045	1.381	0.651	0.729
			median	1.044	1.445	0.692	0.708
			std	0.025	0.827	0.731	0.746
		15	mean	1.049	1.605	0.750	0.802
			median	1.048	1.529	0.793	0.819
			std	0.026	0.426	0.864	0.850
	2	10	mean	1.048	1.515	0.680	0.683
			median	1.046	1.503	0.693	0.725
			std	0.025	0.527	0.805	0.874
		15	mean	1.051	1.591	0.810	0.753
			median	1.049	1.553	0.813	0.781
			std	0.026	0.303	0.891	0.969

Table 2: Contextual effects ($\delta = 1$) for simulations with heterogeneity.

				OLS	G2SLS	G3SLS	G3SLSX
n	l	m					
50	1	10	mean	-0.263	-0.915	1.577	1.660
			median	-0.263	-0.912	0.943	0.946
			std	0.918	3.464	8.756	10.213
		15	mean	-0.298	-1.334	1.788	1.290
			median	-0.308	-1.186	0.745	0.952
			std	1.344	4.015	13.315	15.585
	2	10	mean	-0.491	-1.297	0.157	0.860
			median	-0.500	-1.424	0.471	0.738
			std	1.468	4.716	15.808	12.531
		15	mean	-0.523	-1.523	-0.284	1.075
			median	-0.489	-1.432	0.042	0.897
			std	2.187	6.430	24.199	18.590
100	1	10	mean	-0.269	-1.138	1.370	1.053
			median	-0.268	-1.167	0.889	0.684
			std	0.620	3.193	5.104	5.790
		15	mean	-0.301	-1.707	0.891	0.882
			median	-0.310	-1.551	0.589	0.493
			std	0.912	3.106	7.851	8.682
	2	10	mean	-0.476	-1.973	1.782	1.150
			median	-0.506	-1.634	0.860	0.672
			std	0.961	3.990	7.983	7.750
		15	mean	-0.488	-2.617	2.519	0.838
			median	-0.527	-1.962	1.204	0.596
			std	1.430	4.670	11.904	12.416
200	1	10	mean	-0.235	-1.466	1.270	1.207
			median	-0.251	-1.409	1.034	1.054
			std	0.431	3.682	3.009	3.408
		15	mean	-0.246	-2.317	0.943	1.154
			median	-0.247	-1.853	0.851	0.858
			std	0.632	2.676	4.269	4.787
	2	10	mean	-0.473	-2.464	1.158	1.299
			median	-0.469	-2.251	0.844	0.835
			std	0.633	2.915	4.432	4.598
		15	mean	-0.495	-2.716	0.928	1.023
			median	-0.505	-2.447	0.472	0.848
			std	0.937	2.818	6.282	6.970

Table 3: Direct effects ($\gamma = 1$) for simulations with heterogeneity.

				OLS	G2SLS	G3SLS	G3SLSX
n	l	m					
50	1	10	mean	0.770	0.566	1.910	1.558
			median	0.741	0.563	1.734	1.423
			std	0.645	1.528	1.755	1.980
		15	mean	0.761	0.374	1.968	1.517
			median	0.756	0.403	1.783	1.391
			std	0.961	1.963	2.599	3.035
	2	10	mean	0.171	0.347	0.337	1.535
			median	0.162	0.311	0.234	1.428
			std	1.278	3.429	14.331	2.816
		15	mean	0.163	0.269	0.144	1.483
			median	0.158	0.291	0.349	1.428
			std	1.924	5.418	21.803	4.040
100	1	10	mean	0.770	0.516	1.837	1.557
			median	0.743	0.484	1.809	1.569
			std	0.435	1.338	1.177	1.349
		15	mean	0.771	0.294	1.809	1.521
			median	0.748	0.324	1.806	1.526
			std	0.648	1.413	1.766	1.948
	2	10	mean	0.169	0.301	0.152	1.522
			median	0.146	0.229	-0.078	1.366
			std	0.863	2.521	6.361	1.909
		15	mean	0.162	0.436	-0.933	1.549
			median	0.133	0.257	-0.826	1.349
			std	1.294	3.591	10.031	2.997
200	1	10	mean	0.751	0.307	1.860	1.539
			median	0.749	0.372	1.864	1.494
			std	0.299	1.568	0.769	0.883
		15	mean	0.744	-0.079	1.817	1.548
			median	0.738	0.121	1.815	1.481
			std	0.447	1.248	1.124	1.252
	2	10	mean	0.203	0.430	-0.058	1.524
			median	0.199	0.344	-0.088	1.514
			std	0.570	1.584	3.463	1.233
		15	mean	0.212	0.427	-0.260	1.462
			median	0.213	0.413	-0.110	1.439
			std	0.855	2.105	5.127	1.866

Table 4: Peer effects ($\beta = 0.7$) for simulations with homophily.

				OLS	G2SLS	G3SLS	G3SLSX
n	l	m					
50	1	2	mean	0.786	0.743	0.670	0.557
			median	0.788	0.748	0.682	0.607
			std	0.039	0.045	0.076	0.217
		5	mean	0.851	0.791	0.652	0.570
			median	0.852	0.798	0.687	0.626
			std	0.054	0.076	0.153	0.278
	2	2	mean	0.801	0.759	0.671	0.595
			median	0.802	0.763	0.683	0.640
			std	0.040	0.045	0.084	0.309
		5	mean	0.871	0.811	0.647	0.653
			median	0.868	0.816	0.684	0.701
			std	0.056	0.097	0.177	0.448
100	1	2	mean	0.787	0.743	0.672	0.586
			median	0.788	0.744	0.677	0.603
			std	0.027	0.030	0.051	0.130
		5	mean	0.856	0.798	0.664	0.588
			median	0.856	0.801	0.674	0.616
			std	0.037	0.047	0.091	0.175
	2	2	mean	0.803	0.759	0.672	0.551
			median	0.802	0.761	0.678	0.597
			std	0.027	0.031	0.059	0.213
		5	mean	0.875	0.827	0.656	0.585
			median	0.873	0.831	0.675	0.640
			std	0.038	0.052	0.116	0.294
200	1	2	mean	0.817	0.766	0.668	0.576
			median	0.817	0.767	0.671	0.591
			std	0.023	0.025	0.047	0.111
		5	mean	0.900	0.851	0.656	0.583
			median	0.899	0.852	0.666	0.609
			std	0.031	0.038	0.090	0.158
	2	2	mean	0.836	0.789	0.669	0.563
			median	0.836	0.789	0.673	0.596
			std	0.025	0.025	0.050	0.167
		5	mean	0.918	0.897	0.652	0.577
			median	0.916	0.891	0.667	0.627
			std	0.035	0.048	0.099	0.235

Table 5: Contextual effects ($\delta = 1$) for simulations with homophily.

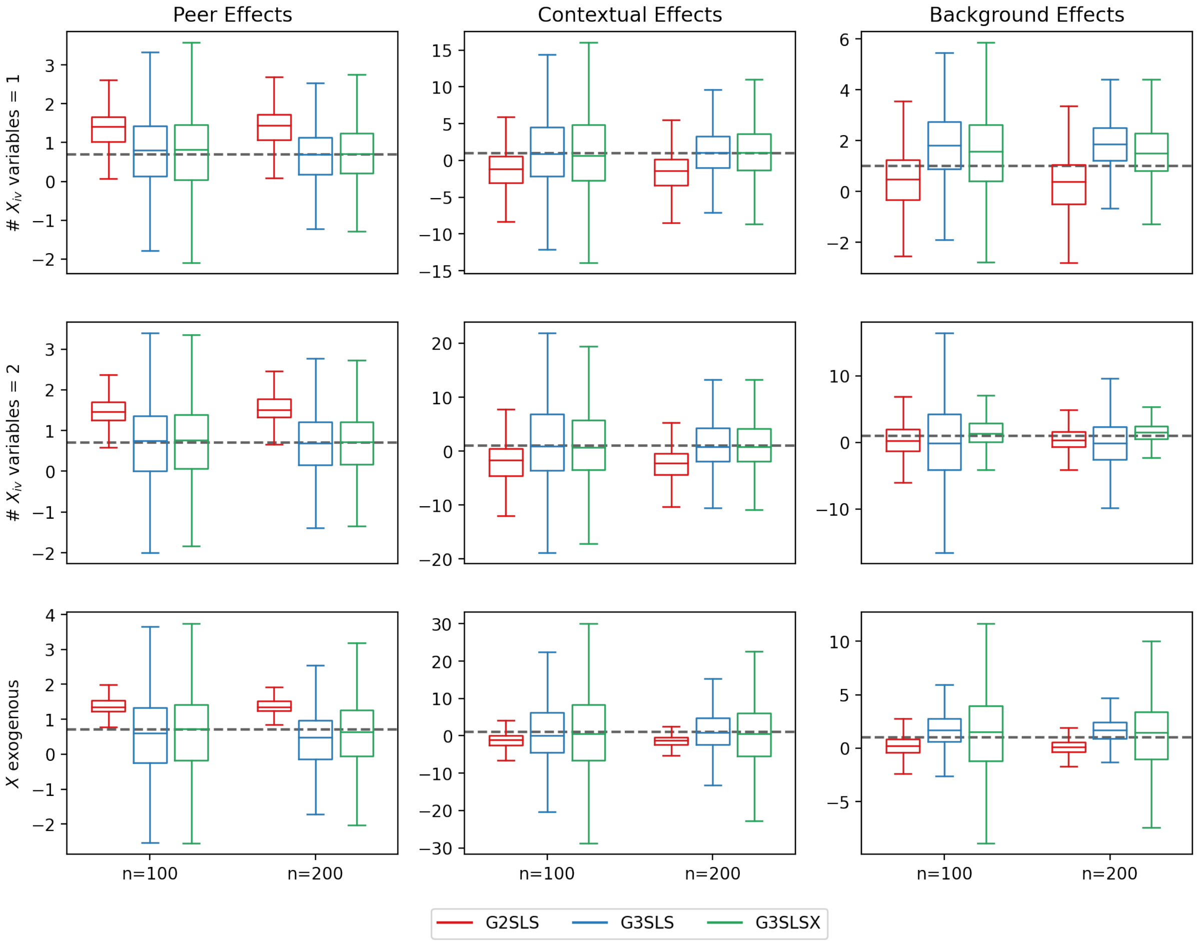
				OLS	G2SLS	G3SLS	G3SLSX
n	l	m					
50	1	2	mean	0.746	0.872	1.079	1.316
			median	0.745	0.879	1.074	1.252
			std	0.153	0.160	0.238	0.522
		5	mean	0.556	0.736	1.104	1.249
			median	0.541	0.728	1.091	1.189
			std	0.245	0.292	0.448	0.722
	2	2	mean	0.654	0.799	1.099	1.240
			median	0.652	0.800	1.086	1.169
			std	0.198	0.207	0.329	0.808
		5	mean	0.423	0.641	1.166	1.105
			median	0.413	0.634	1.167	1.051
			std	0.339	0.463	0.700	1.158
100	1	2	mean	0.741	0.874	1.085	1.290
			median	0.744	0.872	1.078	1.262
			std	0.104	0.108	0.167	0.352
		5	mean	0.541	0.719	1.114	1.276
			median	0.542	0.715	1.106	1.248
			std	0.161	0.191	0.311	0.474
	2	2	mean	0.646	0.798	1.094	1.330
			median	0.653	0.802	1.081	1.280
			std	0.127	0.135	0.224	0.536
		5	mean	0.398	0.569	1.125	1.234
			median	0.395	0.567	1.115	1.193
			std	0.226	0.259	0.444	0.771
200	1	2	mean	0.701	0.833	1.077	1.259
			median	0.701	0.830	1.076	1.238
			std	0.072	0.077	0.136	0.257
		5	mean	0.487	0.618	1.102	1.234
			median	0.482	0.616	1.097	1.200
			std	0.116	0.130	0.253	0.372
	2	2	mean	0.586	0.730	1.087	1.273
			median	0.588	0.733	1.081	1.244
			std	0.097	0.095	0.170	0.357
		5	mean	0.332	0.404	1.129	1.226
			median	0.337	0.410	1.109	1.215
			std	0.164	0.204	0.337	0.529

Table 6: Direct effects ($\gamma = 1$) for simulations with homophily.

				OLS	G2SLS	G3SLS	G3SLSX
n	l	m					
50	1	2	mean	1.120	1.186	1.271	1.042
			median	1.125	1.185	1.264	1.034
			std	0.117	0.124	0.153	0.263
		5	mean	1.025	1.115	1.296	1.054
			median	1.032	1.103	1.284	1.049
			std	0.226	0.251	0.278	0.386
	2	2	mean	0.046	0.026	-0.011	0.994
			median	0.047	0.035	-0.014	0.981
			std	0.117	0.117	0.167	0.380
		5	mean	0.075	0.049	-0.021	0.983
			median	0.076	0.049	-0.021	0.974
			std	0.238	0.262	0.393	0.588
100	1	2	mean	1.115	1.185	1.264	1.031
			median	1.117	1.181	1.261	1.019
			std	0.084	0.088	0.116	0.177
		5	mean	1.005	1.099	1.284	1.055
			median	1.000	1.083	1.274	1.049
			std	0.157	0.175	0.205	0.269
	2	2	mean	0.046	0.026	-0.011	1.042
			median	0.048	0.025	-0.016	1.031
			std	0.073	0.073	0.109	0.267
		5	mean	0.077	0.060	-0.014	1.043
			median	0.081	0.065	-0.012	1.020
			std	0.148	0.152	0.247	0.424
200	1	2	mean	1.122	1.179	1.261	1.024
			median	1.123	1.178	1.258	1.020
			std	0.051	0.055	0.077	0.117
		5	mean	1.032	1.087	1.273	1.037
			median	1.034	1.082	1.265	1.033
			std	0.100	0.105	0.143	0.192
	2	2	mean	0.059	0.039	-0.009	1.025
			median	0.058	0.038	-0.004	1.006
			std	0.056	0.054	0.083	0.162
		5	mean	0.098	0.088	-0.013	1.033
			median	0.090	0.085	-0.008	1.003
			std	0.111	0.113	0.193	0.264

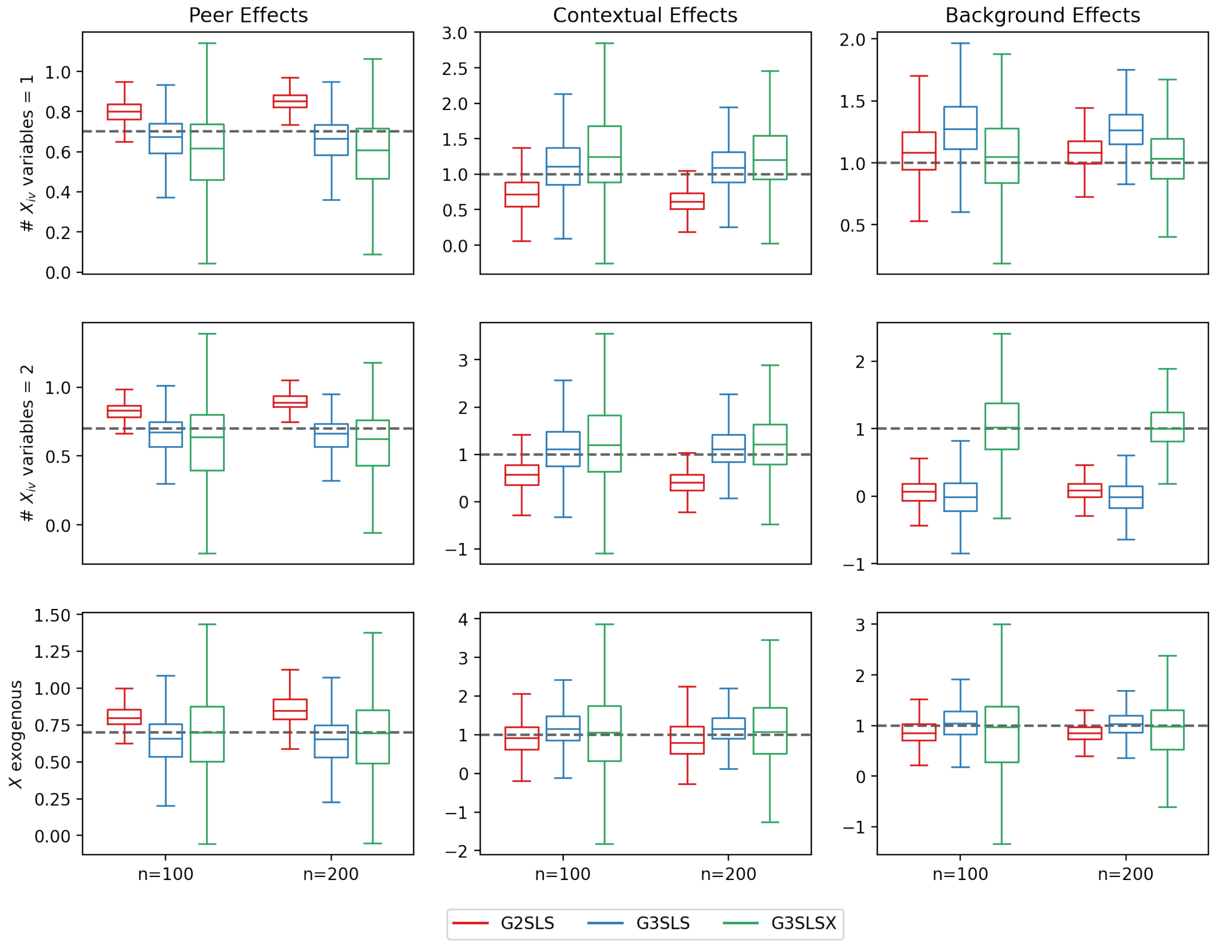
Figures

Figure 1: Design 1 - Unobserved Heterogeneity.



Note: G2SLS corresponds to the estimation of Bramoullé et al. (2009), G3SLS refers to the estimation proposed by Estrada et al. (2021), and G3SLSX is the modified version of G3SLS that includes covariates **X** as instruments.

Figure 2: Design 2 - Homophily.



Note: G2SLS corresponds to the estimation of Bramoullé et al. (2009), G3SLS refers to the estimation proposed by Estrada et al. (2021), and G3SLSX is the modified version of G3SLS that includes covariates **X** as instruments.