G2SLS: Generalized 2SLS procedure for Stata

Nicolas Suarez Chavarria

Department of Economics, Stanford University

July 21, 2023

Preview

- I implement the generalized two-stage least squares procedure described in Bramoullé et al. (2009) to estimate peer effects models.

Preview

- I implement the generalized two-stage least squares procedure described in Bramoullé et al. (2009) to estimate peer effects models.
- I extend their original framework to estimate peer effects models using OLS and to allow for independent variables without peer effects.

Preview

- I implement the generalized two-stage least squares procedure described in Bramoullé et al. (2009) to estimate peer effects models.
- I extend their original framework to estimate peer effects models using OLS and to allow for independent variables without peer effects.
- Short application to showcase the `2gsls` package.

Outline

Motivation

- If we want to estimate a linear-in-means regression, there are no readily available packages to do so.

Motivation

- If we want to estimate a linear-in-means regression, there are no readily available packages to do so.
- Computing the mean outcomes and characteristics of peers with loops is hard and inefficient.

Motivation

- If we want to estimate a linear-in-means regression, there are no readily available packages to do so.
- Computing the mean outcomes and characteristics of peers with loops is hard and inefficient.
- To address this and the endogeneity problems in linear-in-means models, I developed the `2gsls` package.

Peer effects can be classified into 3 categories:

Peer effects can be classified into 3 categories:

- **Exogenous (or contextual) effects:** influence of exogenous peer characteristics on my outcomes.
- **Endogenous effects:** influence of peer outcomes on my outcomes.
- **Correlated effects:** individuals in the same reference group behave similarly because they face a common environment.

There are 2 main challenges when estimating a peer effects model:

There are 2 main challenges when estimating a peer effects model:

1. It is difficult to distinguish real social effects (endogenous and exogenous) from correlated effects.
2. Reflection problem: Individuals simultaneously determine each other's outcomes. This endogeneity makes it difficult to distinguish between endogenous and exogenous effects.

There are 2 main challenges when estimating a peer effects model:

1. It is difficult to distinguish real social effects (endogenous and exogenous) from correlated effects.
2. Reflection problem: Individuals simultaneously determine each other's outcomes. This endogeneity makes it difficult to distinguish between endogenous and exogenous effects.

Generalized Two-Stage Least Squares tackles these 2 problems:

There are 2 main challenges when estimating a peer effects model:

1. It is difficult to distinguish real social effects (endogenous and exogenous) from correlated effects.
2. Reflection problem: Individuals simultaneously determine each other's outcomes. This endogeneity makes it difficult to distinguish between endogenous and exogenous effects.

Generalized Two-Stage Least Squares tackles these 2 problems:

1. Adding network-level fixed effects controls for unobserved factors that affect individuals in the same group.
2. Using instrumental variables based on the network structure takes care of the endogeneity problem.

Context
Econometric framework

We start with a simple linear-in-means model:

$$y_i = \alpha + \beta \frac{1}{n_i} \sum_{j \in P_i} y_j + \gamma x_i + \delta \frac{1}{n_i} \sum_{j \in P_i} x_j + \varepsilon_i \tag{1}$$

- P_i are the peers of individual i.

We start with a simple linear-in-means model:

$$y_i = \alpha + \beta \frac{1}{n_i} \sum_{j \in P_i} y_j + \gamma x_i + \delta \frac{1}{n_i} \sum_{j \in P_i} x_j + \varepsilon_i \tag{1}$$

- P_i are the peers of individual i.
- β captures the endogenous peer effect.

We start with a simple linear-in-means model:

$$y_i = \alpha + \beta \frac{1}{n_i} \sum_{j \in P_i} y_j + \gamma x_i + \delta \frac{1}{n_i} \sum_{j \in P_i} x_j + \varepsilon_i \tag{1}$$

- P_i are the peers of individual i.
- β captures the endogenous peer effect.
- δ captures exogenous peer effects.

We can rewrite this more generally using matrices:

$$\boldsymbol{y} = \alpha\iota + \boldsymbol{G}\boldsymbol{y}\beta + \boldsymbol{X}\gamma + \boldsymbol{G}\boldsymbol{X}\delta + \varepsilon \tag{2}$$

We can rewrite this more generally using matrices:

$$\boldsymbol{y} = \alpha\iota + \boldsymbol{G}\boldsymbol{y}\beta + \boldsymbol{X}\gamma + \boldsymbol{G}\boldsymbol{X}\delta + \varepsilon \qquad (2)$$

- $\boldsymbol{G}$ is an N-by-N adjacency matrix representing the relationships between peers.

We can rewrite this more generally using matrices:

$$y = \alpha\iota + Gy\beta + X\gamma + GX\delta + \varepsilon \tag{2}$$

- G is an N-by-N adjacency matrix representing the relationships between peers.
- The i-th row of G captures the relationship of individual i with his peers.

- Bramoullé et al. (2009) developed a procedure to estimate equation (2).

- Bramoullé et al. (2009) developed a procedure to estimate equation (2).
- We will rewrite our model as follows:

$$
\begin{aligned}
y &= \begin{bmatrix} \iota & Gy & X & GX \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \\ \gamma \\ \delta \end{bmatrix} + \varepsilon \\
\Leftrightarrow y &= \tilde{X}\theta + \varepsilon
\end{aligned}
$$

- This model is identified if matrices I, G and G^2 are linearly independent.

We follow these steps:

1. We define our instrument $S = \begin{bmatrix} \iota & X & GX & G^2X \end{bmatrix}$ for $\tilde{X}$.

We follow these steps:

1. We define our instrument $S = \begin{bmatrix} \iota & X & GX & G^2X \end{bmatrix}$ for $\tilde{X}$.
2. We estimate our model using 2SLS:

$$\widehat{\theta}_{2SLS} = (\tilde{X}'P\tilde{X})^{-1}\tilde{X}'Py$$

with $P = S(S'S)^{-1}S'$.

We follow these steps:

1. We define our instrument $S = \begin{bmatrix} \iota & X & GX & G^2X \end{bmatrix}$ for $\tilde{X}$.
2. We estimate our model using 2SLS:

$$\widehat{\theta}_{2SLS} = (\tilde{X}'P\tilde{X})^{-1}\tilde{X}'Py$$

   with $P = S(S'S)^{-1}S'$.
3. We compute the predicted value of the outcome as:

$$\widehat{y}_{2SLS} = (I - \widehat{\beta}_{2SLS}G)^{-1}\left(\widehat{\alpha}_{2SLS} + X\widehat{\gamma}_{2SLS} + GX\widehat{\delta}_{2SLS}\right)$$

4. We build a new instrument for $\tilde{X}$:

$$\widehat{Z} = \begin{bmatrix} \iota & G\,\widehat{y}_{2SLS} & X & GX \end{bmatrix}$$

4. We build a new instrument for $\tilde{X}$:

$$\widehat{Z} = \begin{bmatrix} \iota & G\,\widehat{y}_{2SLS} & X & GX \end{bmatrix}$$

5. We get our final estimator using standard IV:

$$\widehat{\beta}_{G2SLS} = (\widehat{Z}'\tilde{X})^{-1}\widehat{Z}'y$$

$$V\left(\widehat{\beta}_{G2SLS}\right) = (\widehat{Z}'\tilde{X})^{-1}\widehat{Z}'\,D\,\widehat{Z}(\widehat{Z}'\tilde{X})^{-1}$$

where D is a diagonal matrix with the squared resids produced by $\widehat{\beta}_{G2SLS}$.

- Bramoullé et al. (2009) also present a version of this model with network-specific unobservable factors:

$$y = \sum_{l \in G} \alpha_l + Gy\beta + X\gamma + GX\delta + \varepsilon \tag{3}$$

where α_l is common to all individuals in the l-th component of the network.

- Bramoullé et al. (2009) also present a version of this model with network-specific unobservable factors:

$$y = \sum_{l \in G} \alpha_l + Gy\beta + X\gamma + GX\delta + \varepsilon \tag{3}$$

where α_l is common to all individuals in the l-th component of the network.

- We can transform this model by multiplying it by $(I - G)$ to get rid of these unobservable effects. ▸ G2SLS with FE details

- I extended the previous framework to allow for independent variables without peer effects:

$$\boldsymbol{y} = \alpha + \boldsymbol{G}\boldsymbol{y}\beta + \boldsymbol{X}_1\gamma + \boldsymbol{G}\boldsymbol{X}_1\delta + \boldsymbol{X}_2\psi + \varepsilon$$

- I extended the previous framework to allow for independent variables without peer effects:

$$\boldsymbol{y} = \alpha + \boldsymbol{G}\boldsymbol{y}\beta + \boldsymbol{X}_1\gamma + \boldsymbol{G}\boldsymbol{X}_1\delta + \boldsymbol{X}_2\psi + \varepsilon$$

- ψ captures the effects of our direct variables $\boldsymbol{X}_2$.

`g2sls` *depvar indepvars* [*if*] [*in*] , *adjacency(Mata matrix)* [`row` `fixed` `ols` `directvariables`(*varlist*) `level`(#)]

`g2sls` *depvar indepvars* [*if*] [*in*] , *adjacency(Mata matrix)* [`row` `fixed` `ols` `directvariables(`*varlist*`)` `level(#)`]

Options:

- `adjacency`: Mata matrix containing an N by N matrix of adjancency.
- `row`: row normalizes the adjacency matrix, so each row sums 1.
- `fixed`: adds component-level fixed effects.
- `ols`: reports OLS results instead of IV.
- `directvariables`: independent variables that will not have an exogenous effect.
- `level`: set confidence level for reported confidence intervals.

Application
Context

- Peer effects for college students in Chile between 2012 and 2019.

- Peer effects for college students in Chile between 2012 and 2019.
- 8 cohorts of approximately 500 students each from the Business and Economics school of the University of Chile.

- Peer effects for college students in Chile between 2012 and 2019.
- 8 cohorts of approximately 500 students each from the Business and Economics school of the University of Chile.
- Students are randomly assigned to their first semester classes. We define their peers as the students they share at least 1 class with.

Application
Context

- Peer effects for college students in Chile between 2012 and 2019.
- 8 cohorts of approximately 500 students each from the Business and Economics school of the University of Chile.
- Students are randomly assigned to their first semester classes. We define their peers as the students they share at least 1 class with.
- Our adjacency matrix will be block diagonal, with each cohort being represented by a block.

```
. describe gpa_first adm_score aff_action female major*

Variable        Storage   Display    Value
    name           type    format    label      Variable label
------------------------------------------------------------------------

gpa_first         float   %9.0g                 First semester GPA
adm_score         float   %9.0g                 Admission score
aff_action        byte    %9.0g                 Affirmative action
female            byte    %9.0g                 Female
major_econ        float   %9.0g                 Major in Economics
major_buss        float   %9.0g                 Major in Business
. list gpa_first adm_score aff_action female major* in 1/5
```

	gpa_first	adm_score	aff_ac~n	female	major_~n	major_~s
1.	.1698871	-1.262415	0	1	0	0
2.	.7442471	.44189	0	0	1	0
3.	-2.991099	.4029151	0	0	0	0
4.	.4959475	2.504061	0	0	1	0
5.	.7618809	2.822953	0	0	1	0

```
. g2sls gpa_first female aff_action adm_score, row adj(G)
                                              Number of obs =       4308
```

gpa_first	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
_cons	.0111257	.0778849	0.14	0.886	-.1415689	.1638204
gpa_first_p	.5676393	.4738957	1.20	0.231	-.3614408	1.496719
female	.1856059	.0177705	10.44	0.000	.1507666	.2204452
aff_action	.0935423	.0413983	2.26	0.024	.0123802	.1747044
adm_score	.3069133	.0187414	16.38	0.000	.2701704	.3436562
female_p	-.2034284	.1893837	-1.07	0.283	-.5747183	.1678614
aff_action_p	-.0598223	.1047165	-0.57	0.568	-.2651206	.1454761
adm_score_p	-.3068539	.0777929	-3.94	0.000	-.4593681	-.1543397

Application
IV model with fixed effects

```
. g2sls gpa_first female aff_action adm_score, row adj(G) fixed
                                                Number of obs =       4308
Controlling for component-level fixed effects
```

gpa_first	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
gpa_first_p	.0066238	1.45047	0.00	0.996	-2.837045	2.850293
female	.1870434	.0183427	10.20	0.000	.1510823	.2230045
aff_action	.0887381	.0427942	2.07	0.038	.0048394	.1726368
adm_score	.3074021	.0190128	16.17	0.000	.2701271	.3446771
female_p	.0508922	.427888	0.12	0.905	-.7879889	.8897733
aff_action_p	.0147204	.2325366	0.06	0.950	-.4411712	.470612
adm_score_p	-.1949845	.2778821	-0.70	0.483	-.7397767	.3498076

```
. g2sls gpa_first female aff_action adm_score, row adj(G) ols
                                              Number of obs =       4308
```

gpa_first	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
_cons	-.0233233	.0768411	-0.30	0.762	-.1739714	.1273248
gpa_first_p	-.7923793	.1912952	-4.14	0.000	-1.167417	-.4173421
female	.1847655	.0183224	10.08	0.000	.1488442	.2206869
aff_action	.1040999	.0424841	2.45	0.014	.0208091	.1873906
adm_score	.3188386	.016474	19.35	0.000	.286541	.3511363
female_p	-.0519749	.1807673	-0.29	0.774	-.406372	.3024222
aff_action_p	.0464968	.0974559	0.48	0.633	-.144567	.2375605
adm_score_p	-.106855	.043616	-2.45	0.014	-.1923649	-.0213452

Application
IV model with direct effects

```
. g2sls gpa_first female aff_action adm_score, row adj(G) directvariables(major_*)
                                                    Number of obs =      4308
```

gpa_first	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
_cons	-.5948202	.0914547	-6.50	0.000	-.7741186	-.4155219
gpa_first_p	-4.26719	.5560077	-7.67	0.000	-5.357252	-3.177128
female	.1861742	.0170568	10.91	0.000	.1527341	.2196143
aff_action	.0755271	.0389122	1.94	0.052	-.0007609	.1518151
adm_score	.2892999	.018163	15.93	0.000	.2536911	.3249087
female_p	.8655189	.2057867	4.21	0.000	.4620708	1.268967
aff_action_p	-.2825957	.096932	-2.92	0.004	-.4726325	-.0925588
adm_score_p	.1189513	.0754862	1.58	0.115	-.0290407	.2669433
major_econ	.6840927	.0416389	16.43	0.000	.6024591	.7657264
major_buss	.5122815	.0397691	12.88	0.000	.4343135	.5902495

We can use **`estimates store`** and **`estout`** to organize our results:

Variable	OLS			G2SLS		
GPA of peers	-0.7924***	-6.9507***	-6.6984***	0.5676	0.0066	-5.7565***
	(0.1913)	(0.3359)	(0.3198)	(0.4739)	(1.4505)	(1.2421)
Share of female peers	-0.0520	1.1091***	1.1682***	-0.2034	0.0509	1.0230**
	(0.1808)	(0.3564)	(0.3392)	(0.1894)	(0.4279)	(0.4088)
Share of peers in Aff. Action program	0.0465	0.6204***	-0.0977	-0.0598	0.0147	-0.1812
	(0.0975)	(0.1862)	(0.1804)	(0.1047)	(0.2325)	(0.2119)
Adm. Score of peers	-0.1069**	1.1104***	0.6738***	-0.3069***	-0.1950	0.4963**
	(0.0436)	(0.0869)	(0.0851)	(0.0778)	(0.2779)	(0.2329)
Female	0.1848***	0.1860***	0.1838***	0.1856***	0.1870***	0.1841***
	(0.0183)	(0.0180)	(0.0171)	(0.0178)	(0.0183)	(0.0175)
Affirmative Action program	0.1041**	0.0960**	0.0617	0.0935**	0.0887**	0.0603
	(0.0425)	(0.0415)	(0.0396)	(0.0414)	(0.0428)	(0.0400)
Admission score	0.3188***	0.3110***	0.2674***	0.3069***	0.3074***	0.2664***
	(0.0165)	(0.0162)	(0.0156)	(0.0187)	(0.0190)	(0.0184)
Major in Economics			0.6991***			0.7035***
			(0.0330)			(0.0446)
Major in Business			0.5415***			0.5419***
			(0.0297)			(0.0430)
Constant	-0.0233			0.0111		
	(0.0768)			(0.0779)		
Observations	4,308	4,308	4,308	4,308	4,308	4,308
Cohort level fixed effects	No	Yes	Yes	No	Yes	Yes

Concluding remarks

- I implement the generalized two-stage least squares in Stata to estimate peer effects models.
- The `g2sls` command allows for network fixed effects, OLS estimates with network-weighted variables and direct effects.
- **Future steps**: Implement a weak instruments tests for this context.

Thank you!

https://github.com/nicolas-suarez/

nsuarez@stanford.edu

References

- Bramoullé, Y., Djebbari, H., & Fortin, B. (2009). Identification of peer effects through social networks. Journal of econometrics, 150(1), 41-55.

Generalized Two-Stage Least Squares
Model with fixed effects

- We start by pre-multiplying equation (3) by $(I - G)$:

$$(I - G)y = (I - G)Gy\beta + (I - G)X\gamma + (I - G)GX\delta + \varepsilon$$

- We will rewrite our model as follows:

$$(I - G)y = \begin{bmatrix}(I - G)Gy & (I - G)X & (I - G)GX\end{bmatrix}\begin{bmatrix}\beta \\ \gamma \\ \delta\end{bmatrix} + \varepsilon$$

$$\Leftrightarrow (I - G)y = \tilde{X}\theta + \varepsilon$$

- This model is identified if matrices I, G, G^2 and G^3 are linearly independent.

We follow these steps:

1. We define our instrument $S = \left[(I-G)X \quad (I-G)GX \quad (I-G)G^2X\right]$ for $\tilde{X}$.
2. We estimate our model using 2SLS:

$$\widehat{\theta}_{2SLS} = (\tilde{X}'P\tilde{X})^{-1}\tilde{X}'P(I-G)y$$

   with $P = S(S'S)^{-1}S'$.
3. We compute the predicted value of the outcome as:

$$\widehat{y}_{2SLS} = (I-G)^{-1}(I-\widehat{\beta}_{2SLS}G)^{-1}(I-G)\left(X\widehat{\gamma}_{2SLS} + GX\widehat{\delta}_{2SLS}\right)$$

Generalized Two-Stage Least Squares
Model with fixed effects

4. We build a new instrument for $\tilde{X}$:

$$\widehat{Z} = \left[(I-G)G\,\widehat{y}_{2SLS} \quad (I-G)X \quad (I-G)GX\right]$$

5. We get our final estimator using standard IV:

$$\widehat{\beta}_{G2SLS} = (\widehat{Z}'\tilde{X})^{-1}\widehat{Z}'(I-G)y$$

$$V\left(\widehat{\beta}_{G2SLS}\right) = (\widehat{Z}'\tilde{X})^{-1}\widehat{Z}'\,D\,\widehat{Z}(\widehat{Z}'\tilde{X})^{-1}$$

where D is a diagonal matrix with the squared resids produced by $\widehat{\beta}_{G2SLS}$.

back