

# Design of Randomized Experiments to Measure Social Interaction Effects

Jinyong Hahn  
UCLA\*

Keisuke Hirano  
University of Arizona<sup>†</sup>

27 January 2009

## Abstract

We consider the use of randomized experiments to measure social interaction effects. Randomization at two levels—across groups and within groups—can resolve an omitted variables problem for a linear-in-means model of endogenous social interactions. We examine how the randomization should be carried out to estimate the coefficients of interest most precisely, and calculate the optimal treatment probabilities under different criteria.

## 1 Introduction

There has been considerable work in recent years on estimating models where individuals within a group influence each other's behavior. Brock and Durlauf (2001) and Moffitt (2001) survey recent studies of social interaction effects.

Models with social interaction effects present severe identification problems, as was pointed out in the seminal paper by Manski (1993). We consider a version of Manski's linear-in-means model of social interactions, and show how a randomization can be used to identify the model. Then we consider the experimental design issue: how should the randomization probabilities be chosen to obtain the most precise estimates of the coefficients of interest? We calculate the optimal treatment probabilities under certain criteria. We find that for some criteria, the design used by Duflo and Saez (2003) is nearly optimal.

---

\*Department of Economics, University of California, Los Angeles, Box 951477, Los Angeles, CA 90095-1477 (hahn@econ.ucla.edu)

<sup>†</sup>Department of Economics, University of Arizona, Tucson, AZ 85721 (hirano@u.arizona.edu)

## 2 Linear-in-Means Model

Suppose we observe  $N$  non-overlapping groups  $g = 1, \dots, N$ . For each group  $g$ , we sample  $M^g$  individuals. We assume that the  $M^g$  individuals form a random subset of the full group, and we assume that all the variables of interest are independent across groups.

Consider the following linear-in-means model of social interactions, which is a special case of the model introduced by Manski (1993):

$$y_{gi} = E_g[y_{gi}]\beta + x_{gi}\eta + \alpha_{gi} + \epsilon_{gi}, \quad (1)$$

where all variables are measured in deviations from sample means, and  $E_g[\cdot]$  denotes the mean within the  $g$ th group. The outcome of interest is  $y_{gi}$  for the  $i$ th individual in the  $g$ th group, and for simplicity, we assume that the covariate  $x_{gi}$  is a scalar.

The coefficient  $\beta$  on  $E_g[y_{gi}]$  is the *endogenous social effect* (Manski, 1993). Equation (1) implicitly rules out so-called exogenous social effects, which would arise if the group-mean of a covariate appeared as an explanatory variable in the equation. In other words, the covariate  $x_{gi}$  affects outcomes only through the individual  $i$ . The term  $\alpha_g$ , which is not observed by the econometrician, captures the presence of correlated effects within a group.

To obtain a reduced form equation, take the group mean of Equation 1 to obtain

$$E_g[y_{gi}] = E_g[x_{gi}]\frac{\eta}{1-\beta} + \frac{\alpha_g}{1-\beta}. \quad (2)$$

We assume that the system is in a social equilibrium, so that (2) holds, and we can substitute (2) into (1) to get

$$y_{gi} = x_{gi}\eta + E_g[x_{gi}]\frac{\beta\eta}{1-\beta} + \alpha_g^* + \epsilon_{gi}, \quad (3)$$

where  $\alpha_g^* = \alpha_g/(1-\beta)$ .

In general,  $\alpha_g^*$  may be correlated with  $E_g[x_{gi}]$ , so that a simple regression of the outcome on  $x_{gi}$  and  $E_g[x_{gi}]$  would not necessarily lead to reasonable estimates of the parameters  $\beta$  and  $\eta$ . Graham and Hahn (2005) suggest viewing this as a panel data model, where the  $\alpha_g^*$  are group fixed effects. Since one of the regressors of interest,  $E_g[x_{gi}]$ , does not vary within the group, they propose to use a version of the instrumental variables estimator of Hausman and Taylor (1981) to estimate the model. Below, we argue that randomization can also be used solve the problem, and then consider how the randomization should be implemented.

### 3 Benefit of Randomization

Recently a number of studies have used randomized experiments to measure social interaction effects. Some examples include Angelucci and De Giorgi (2008), Duflo and Saez (2003), and Miguel and Kremer (2003). We consider a two-level randomization scheme as follows:

1. Select a random subset of groups, using group treatment probability  $p$ .
2. Within each “treated” group, select individuals randomly with probability  $q$ . These individuals receive the treatment; other individuals in the group (and all individuals in untreated groups) do not receive the treatment.

As an example, Duflo and Saez (2003) used  $p = \frac{2}{3}$  and  $q = \frac{1}{2}$ . We will examine how the randomization identifies social interaction effects, then consider the choice of  $p$  and  $q$ .

Let  $D_g = 1$  for treated groups (and 0 otherwise), and  $L_{gi} = 1$  for treated individuals (and 0 otherwise). Consider the model

$$y_{gi} = \alpha_2 + \mu_2 L_{gi} + \delta_2 D_g + v_{gi}. \quad (4)$$

Note that for groups with  $D_g = 1$ ,  $Pr(L_{gi} = 1) = q$ , and for groups with  $D_g = 0$ ,  $Pr(L_{gi} = 1) = 0$ . Therefore, we may write the model as

$$y_{gi} = \alpha_2 + \mu_2 L_{gi} + \delta_2^* E_g[L_{gi}] + v_{gi},$$

with  $\delta_2^* = \frac{\delta_2}{q}$ . If we set

$$\begin{aligned} v_{gi} &= \alpha_g^* - E[\alpha_g^*] + \epsilon_{gi}, \\ \alpha_2 &= E[\alpha_g^*], \\ L_{gi} &= x_{gi}, \end{aligned}$$

this has the same form as Equation (3). Crucially, however, randomization ensures that  $(L_{gi}, E_g[L_{gi}])$  is independent of  $v_{gi}$ . Therefore the model is a pure random effects model, with random effects independent of the regressors. Standard OLS is consistent for the parameters  $\mu_2$  and  $\delta_2$ , or GLS can be used.

### 4 Design Issues

Consider OLS estimation of the model in Equation (4). How should  $(p, q)$  be chosen to obtain the most precise estimates of the parameters of interest? To derive concrete results, we consider the

homoskedastic case, where  $V(\alpha_g) = 0$ .<sup>1</sup>

The second moment matrix of  $(1, L_{gi}, D_g)'$  is

$$\begin{bmatrix} 1 & pq & p \\ pq & pq & pq \\ p & pq & p \end{bmatrix},$$

and the variance-covariance matrix of the OLS estimator  $(\widehat{\alpha}_2, \widehat{\mu}_2, \widehat{\delta}_2)$  is

$$\begin{bmatrix} \frac{1}{1-p} & 0 & \frac{1}{p-1} \\ 0 & \frac{1}{pq(1-q)} & \frac{1}{pq-p} \\ \frac{1}{p-1} & \frac{1}{pq-p} & \frac{1-pq}{p(1-p)(1-q)} \end{bmatrix}.$$

In general, different choices of  $(p, q)$  will affect the precision of estimates for the key parameters  $\mu_2$  and  $\delta_2$  differently.

We consider the following cases:

1. Minimize  $V(\widehat{\mu}_2)$ : This leads to the optimization problem

$$\min_{p,q} V(\widehat{\mu}_2) = \min_{p,q} \frac{1}{pq(1-q)}.$$

The trivial solution is  $p = 1$  and  $q = \frac{1}{2}$ , which is not surprising since we are only concerned with the individual response and not the interaction effect.

2. Minimize  $V(\widehat{\delta}_2)$ : The variance of the estimated coefficient on the group variable is

$$V(\widehat{\delta}_2) = \frac{1-pq}{p(1-p)(1-q)}.$$

For any fixed  $p$ , this can be shown to be a strictly increasing function of  $q$  over  $q \in [0, 1]$ . However, if  $q = 0$ , then by construction, we must have  $p = 0$  as well. Thus a minimum does not exist.

3. Minimize a weighted average of variances: consider the weighted criterion function

$$\omega V(\widehat{\mu}_2) + (1-\omega)V(\widehat{\delta}_2),$$

---

<sup>1</sup>The analysis could be extended to consider other estimators and alternative assumptions about the group random effects, but the homoskedastic case is a natural benchmark.

where  $\omega \in (0, 1)$  is a weight. The criterion can be simplified to

$$\frac{\omega}{pq} + \frac{1}{p(1-q)} + \frac{1-\omega}{1-p}.$$

In the case where  $\omega = \frac{1}{2}$ , this is minimized by setting

$$q = \sqrt{2} - 1 = 0.41421,$$

and

$$p = \frac{1}{2\sqrt{2} + 2} \left( 2\sqrt{2} - \sqrt{2\sqrt{2} + 3} + 3 \right) = 0.70711.$$

4. Sequential optimization: suppose that, for any given  $p$ , we choose  $q$  to minimize the variance of the individual-level coefficient:

$$\min_q \frac{1}{pq(1-q)}.$$

Then we choose  $p$  to minimize the variance of the group-level coefficient:

$$\min_p \frac{1-pq}{p(1-p)(1-q)}.$$

Straightforward calculations yield the solution

$$(p, q) = \left( 2 - \sqrt{2}, \frac{1}{2} \right) = (0.58579, 0.5).$$

In closing, we note that approach 3, and especially approach 4, yield optimal treatment probabilities close to the case  $(p, q) = (\frac{2}{3}, \frac{1}{2})$  used in recent studies such as Duflo and Saez (2003). Our analysis can be easily modified to consider alternative criteria and alternative assumptions about the data-generating process.

## References

- Angelucci, M., and De Giorgi, G., 2008, “Indirect Effects of an Aid Program: How do Cash Transfers Affect Ineligibles’ Consumption?,” forthcoming, *American Economic Review*.
- Brock, W. A., and Durlauf, S. N., 2001, “Interactions-based Models,” in *Handbook of Econometrics*, vol. 5, ed. J. Heckman and E. Leamer, Amsterdam: North Holland, pp. 3297-3380.
- Duflo, E., and Saez, E., 2003, “The Role of Information and Social Interactions in Retirement Plan Decisions: Evidence from a Randomized Experiment,” *Quarterly Journal of Economics* 118, 815-842.

Graham, B., and Hahn, J., 2005, "Identification and Estimation of the Linear-in-Means Model of Social Interactions," *Economics Letters* 88(1), 1-6.

Hausman, J. A., and Taylor, W. E., 1981, "Panel Data and Unobservable Individual Effects," *Econometrica* 49(6), 1377-1398.

Manski, C. F., 1993, "Identification of Endogenous Social Effects: The Reflection Problem," *Review of Economic Studies* 60(3), 531-542.

Miguel, E., and Kremer, M., 2003, "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities," *Econometrica* 72(1), 159-217.

Moffitt, R. A., 2001, "Policy Interventions, Low-Level Equilibria and Social Interactions," in *Social Dynamics*, ed. S. Durlauf and P. Young, Cambridge: MIT Press.