Dyadic Regression

Econometric Methods for Networks,

University of Oslo, September 11th to 13th, 2017

Bryan S. Graham

University of California - Berkeley

Dyadic Regression

$\exists$ Large empirical literature which models outcomes for all $\binom{N}{2}$ dyads in a random sample of N agents as a function of agent-level covariates.

Gravity models, analysis of wars between nation-states, risk-sharing links etc.

Considerable confusion and controversy associated with such analyses.

Jackson and Nei (2015, PNAS)

Table 2. Logistic regressions of dyadic war on dyadic trade

	(1)	(2)
Dyad trade	−1,974.37***	
	(383.69)	
Lagged dyad trade		−1,150.24***
		(248.29)
Observations	36,832	35,658

SEs in parentheses. Logit regression of dyad in conflict on dyadic trade, including decade dummies and dyads within 1,000 km of each other. Dyad at war if involved on opposite sides of an MID 5. Dyad trade is normalized by the minimum of the two countries' GDPs. Conflict data from COW. Trade and GDP data from ref. 32. Distance data from ref. 33. ***$P < 0.01$.

Fafchamps and Gubert (2007, AERPP)

TABLE 1—LINKS AND INCOME CORRELATION

	Coefficient estimate	Dyadic t-value
Income correlation		
Correlation of *i* and *j*'s incomes[a]	1.083	1.44
Geographic proximity		
Same sitio = 1[b]	2.647	8.84
Difference in distance to road if same sitio	−0.121	−3.90
Difference in:		
Dummy = 1 if primary occupation of head is farming	0.028	0.23
Number of working members × number of activities	0.003	0.06
Age of household head	−0.010	−2.52
Health index 1–4 (1 = good health, 4 = disabled)	0.027	0.46
Years of education of household head	−0.010	−0.59
Total wealth[a]	−0.113	−2.37
Village dummies	Included but not shown	
Intercept	−5.995	−15.41
Number of observations	10,264	

Notes: The dependent variable = 1 if *i* cites *j* as the source of mutual insurance, 0 otherwise. Estimator is logit. All *t*-values based on standard errors corrected for dyadic correlation of errors.

[a] Instrumented variables—see text for details.

[b] Small cluster of 15–20 households.

Tinbergen (1962, SWE)

FACTORS DETERMINING THE SIZE OF INTERNATIONAL TRADE FLOWS

Results of Calculations A (18 countries)

$$\log E_{ij} = a_1 \log Y_i + a_2 \log Y_j + a_3 \log D_{ij} + a_4 \log N + a_5 \log P_C + a_6 \log P_B + a'_0$$

Calculation No.	ESTIMATED VALUE OF THE COEFFICIENTS							Correlation Coefficient
	a_1	a_2	a_3	a_4	a_5	a_6	a'_0	
A-1	0.7338 (0.0438)	0.6238 (0.0438)	−0.5981 (0.0405)	—	—	—	−0.3783	0.8248
A-2	0.7907 (0.0497)	0.6766 (0.0496)	−0.6252 (0.0460)	—	—	—	−0.4013	0.8084
A-3	0.7357 (0.0421)	0.6183 (0.0422)	−0.5570 (0.0473)	0.0191 (0.0082)	0.0496 (0.0111)	0.0406 (0.0272)	−0.4451	0.8437

E_{ij} Exports from country i to country j
Y_i GNP of exporting country
Y_j GNP of importing country
D_{ij} Distance between countries i and j
N Dummy variable for neighbor countries
P_C Dummy variable for Commonwealth preference
P_B Dummy variable for Benelux preference

In A-2 the trade amount is measured in the importing country.
Figures in brackets are standard deviations.

Dyadic Regression (continued)

Let $Y_{ij} = Y_{ji}$ be an undirected outcome of interest associated with dyad $\{i, j\}$ (directed case poses few additional challenges).

Let X_i be a vector of agent-level covariates.

Let U_i be unobserved agent-level heterogeneity.

The dyadic regression function (symmetric in its two arguments) is

$$g\left(x, x'\right) = \mathbb{E}\left[Y_{ij} \middle| X_i = x, X_j = x'\right]$$

Dyadic Regression: Nonparametric DGP

Specialize to binary case with $Y_{ij} = D_{ij}$ (general case is straightforward).

We will assume that

$$D_{ij} \Big| X_i, X_j, U_i, U_j \sim \text{Bernoulli}\left(h\left(X_i, X_j, U_i, U_j\right)\right)$$

for some function $h(\cdot)$, symmetric in its first and second, as well as its third and fourth, arguments.

Iterated expectations gives

$$g\left(x, x'\right) = \int\int h\left(x, x', u, v\right) f_{U|X}\left(u \middle| x\right) f_{U|X}\left(v \middle| x'\right) \mathrm{d}u\mathrm{d}v.$$

Dyadic Regression: Nonparametric DGP (continued)

Elements of $\mathbf{D} = \left[D_{ij}\right]$ are conditionally independent given $\mathbf{X}$ *and* the latent $\mathbf{U}$, but may be dependent conditional on $\mathbf{X}$ alone.

Captures types of dependence structures typically assumed in empirical work (e.g., Fafchamps and Gubert, 2007).

Will defer question of whether $g\left(x, x'\right)$ has a structural interpretation until later.

Dyadic Regression: Parametric estimation

A prototypical specification for binary outcomes is

$$\text{logit}\left[\pi\left(X_i, X_j; \theta\right)\right] = \alpha + \left[t\left(X_i\right) + t\left(X_j\right)\right]' \beta + \omega\left(X_i, X_j\right)' \gamma$$

for $\theta = (\alpha, \beta', \gamma')'$ with

1. $t(X)$ a vector of linear independent and known functions of X;

2. $\omega\left(X_i, X_j\right) = \omega\left(X_j, X_i\right)$ dyadic-specific regressors.

Dyadic Regression: Parametric estimation (continued)

Estimate θ_0 by maximizing the Bernoulli criterion function

$$L_N(\theta) = \binom{N}{2}^{-1} \sum_{i<j} l\left(Z_{ij}; \theta\right)$$

with $Z_{ij} = \left(X_i', X_j', D_{ij}\right)'$ and $l\left(Z_{ij}; \theta\right)$ equal to, for example, the logit kernel.

This can be done using standard software (see examples above).

Dyadic Regression: Parametric estimation (continued)

Under some basic conditions

$$\sqrt{N}\left(\widehat{\theta}_{\mathsf{DR}}-\theta_0\right)=\underbrace{\left[-H_N\left(\bar{\theta}\right)\right]^{+}}_{\text{Inverse Hessian}}\times\sqrt{N}S_N\left(\theta_0\right)$$

where

$$S_N\left(\theta\right)=\binom{N}{2}^{-1}\sum_{i<j}s\left(Z_{ij};\theta\right)$$

for $s\left(Z_{ij};\theta\right)=\frac{\partial l\left(Z_{ij};\theta\right)}{\partial\theta}$.

Dyadic Regression: Parametric estimation (continued)

$S_N(\theta)$ is not the sum of independent components.

...also not a U-Statistic, but it is "U-Statistic like".

A Hoeffding (1948) variance decomposition gives

$$\mathbb{V}\left(\sqrt{N}S_N(\theta_0)\right) = 4\Sigma_1 + \frac{2}{N-1}(\Sigma_2 - 2\Sigma_1)$$

where $\Sigma_p = \mathbb{E}\left[s\left(Z_{i_1 i_2};\theta_0\right) s\left(Z_{j_1 j_2};\theta_0\right)'\right]$ when the dyads $\{i_1, i_2\}$ and $\{j_1, j_2\}$ share $p = 0, 1, 2$ agents in common.

Dyadic Regression: Variance estimation

Fafchamps and Gubert (2007) propose a now widely-used dyadic-clustered covariance estimator (cf., Cameron and Miller, 2014; Aronow et al., 2017).

It turns out their estimator is equivalent to a natural analog estimate of $4\Sigma_1 + \frac{2}{N-1}(\Sigma_2 - 2\Sigma_1)$

Showing this involves tedious counting arguments.

Dyadic Regression: Variance estimation

The standard "econometrician's" estimate focuses on the leading term only:

$$\tilde{\Sigma}_1 = \frac{1}{N}\sum_{i=1}^{N} \hat{\bar{s}}_i(\theta)\,\hat{\bar{s}}_i(\theta)'$$

with $\hat{\bar{s}}_i(\theta) = \frac{1}{N-1}\sum_{j\neq i} s\left(Z_{ij};\theta\right)$.

This "Jackknife" estimate is biased (e.g., Efron and Stein, 1979).

It turns out that the Fafchamps and Gubert (2007) estimate is "bias-corrected" (albeit computationally inefficient).

When network is sparse these differences appear to be important.

Dyadic Regression

Applying some basic ideas/tools on exchangeable random graphs, network moments etc...

...puts dyadic regression on a much sounder inferential basis.

Potential to make a large empirical literature much more coherent.

It turns out that emerging standard practice in economics has a coherent foundation.

Average Partial Effects

Do import tariffs reduce trade flows?

1. draw agent i at random and exogenously assign her covariate value $X_i = x$

2. draw a second independent agent j at random and assign her covariate value $X_j = x'$.

The (ex ante) expected outcome associated with these assignments is

$$m^{\text{ASF}}\left(x, x'\right) = \int h\left(x, x', u, v\right) f_U\left(u\right) f_U\left(v\right) \mathrm{d}u\mathrm{d}v$$

Average Partial Effects: Identification

A simple identification result under "selection on observations" type assumptions follows if there is a proxy W_i for U_i such that:

1. $\mathbb{E}\left[D_{ij}\middle| X_i, X_j, U_i, U_j, W_i, W_j\right] = h\left(X_i, X_j, U_i, U_j\right)$; [*redundancy*]

2. $U_i \perp X_i | W_i = w,\ w \in \mathbb{W}$; [*conditional independence*]

3. a support condition holds. [*support*]

Dyadic proxy variable regression

Define the dyadic proxy variable regression (PVR) function as

$$q\left(x, x', w, w'\right) = \mathbb{E}\left[D_{ij} \middle| X_i = x, X_j = x', W_i = w, W_j = w'\right]$$

Under our two conditions (*and random sampling*)

$$\begin{aligned} q\left(X_i, X_j, W_i, W_j\right) &= \mathbb{E}\left[\mathbb{E}\left[D_{ij} \middle| X_i, X_j, U_i, U_j, W_i, W_j\right] \middle| X_i, X_j, W_i, W_j\right] \\ &= \mathbb{E}\left[h\left(X_i, X_j, U_i, U_j\right) \middle| X_i, X_j, W_i, W_j\right] \\ &= \int h\left(X_i, X_j, u, v\right) f_{U|W}\left(u \middle| W_i\right) f_{U|W}\left(v \middle| W_j\right) \mathrm{d}u\mathrm{d}v \end{aligned}$$

Double marginal integration

Putting things together we have

$$\begin{aligned}\mathbb{E}_{W_i}\left[\mathbb{E}_{W_j}\left[q\left(x,x',W_i,W_j\right)\right]\right] =& \int\left[\int h\left(x,x',u,v\right)\right. \\ & \left.\times f_{U|W}\left(u|\,w\right) f_{U|W}\left(v|\,w'\right)\mathsf{d}u\mathsf{d}v\right] \\ & \times f_W\left(w\right) f_W\left(w'\right)\mathsf{d}w\mathsf{d}w' \\ =& \int h\left(x,x',u,v\right) f_U\left(u\right) f_U\left(v\right)\mathsf{d}u\mathsf{d}v \\ =& m^{\mathsf{ASF}}\left(x,x'\right).\end{aligned}$$

A formal support condition is

$$\mathbb{S}\left(x,x'\right) \overset{def}{\equiv} \left\{w,w' \,:\, f_{W|X}\left(w|\,x\right) f_{W|X}\left(w'\middle|\,x'\right) > 0\right\} = \mathbb{W}\times\mathbb{W}.$$

Connection to Program Evaluation

When X_i is discretely-valued we can express the support conditioning a form similar to the overlap condition from program evaluation:

$$p_x(w)\, p_z(t) \geq \kappa > 0 \text{ for all } (w,t) \in \mathbb{W} \times \mathbb{W}$$

where $p_x(w) \overset{def}{\equiv} \Pr(X_i = x | W_i = w)$.

APE Wrap-up

Estimation and inference are straightforward for "flexible parametric" proxy variable regression functions.

Provides a framework for thinking about causal effects in dyadic settings (both experimental and observational).

When $X \in \{0, 1\}$ there are interesting connections to the program evaluation literature.

Dyadic regression wrap-up

Basic dyadic regression method extended in recent work by Graham (2017) and others.

More explicit "random effects" type estimators also possible.