Dyadic Regression

Bonn/Mannheim Summer School on the Econometrics of Peer Effects and Social Interactions

Annweiler, Germany, July 15-19, 2019

Bryan S. Graham

University of California - Berkeley

Dyadic regression analyses are abundant in social science research (see below).

In economics they date (at least) to Tinbergen's (1962) pioneering analysis of trade flows.

While frequently used by empirical researchers, dyadic regression analysis lacks inferential foundations.

Widely varying approaches to hypothesis testing used in practice.

Tinbergen (1962, SWE, Table VI-1)

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.

Year: 1958, N = 18, N(N-1) = 306 (estimation by OLS)

Tinbergen (1962, SWE, Table VI-4)

Results of Calculations B (42 countries)

$$\log E_{ij} = \alpha_1 \log Y_i + \alpha_2 \log Y_j + \alpha_3 \log D_{ij} + \alpha_4 \log N + \alpha_7 \log P + \alpha'_o$$

Calculation No.	ESTIMATED VALUE OF THE COEFFICIENTS						Correlation Coefficient
	α_1	α_2	α_3	α_4	α_7	α'_o	
B-1	1.0240 (0.0270)	0.9395 (0.0269)	−0.8919 (0.0455)	—	—	−0.6627 (0.6802)	0.8094
B-2	1.0250 (0.0269)	0.9403 (0.0269)	−0.8225 (0.0517)	0.2581 (0.0920)	—	−0.7188 (0.6789)	0.8104
B-3	1.1832 (0.0323)	1.0752 (0.0323)	−0.9325 (0.0584)	0.2217 (0.1037)	—	−1.0296 (0.7645)	0.7987
B-4	0.9965 (0.0267)	0.9116 (0.0267)	−0.7803 (0.0511)	0.2434 (0.0903)	0.4703 (0.0588)	−0.7798 (0.6668)	0.8180
B-5	1.1567 (0.0319)	1.0486 (0.0319)	−0.9165 (0.0574)	0.2367 (0.1018)	0.8926 (0.1100)	−1.0641 (0.7505)	0.8070

E_{ij} Exports from country i to country j

Y_i GNP of exporting country
Y_j GNP of importing country
} Nominal in B-1, B-2 and B-4; real in B-3 and B-5.

D_{ij} Distance between countries i and j

N Dummy variable for neighboring countries

P Dummy variable for preference

Because of difference in treatment of preferential relations, the coefficients are not comparable between B-4 and B-5.

Figures in brackets are standard deviations.

Year: 1959, N = 42, N(N-1) = 1,722 (estimation by OLS)

König et al (2019, RESTAT)

TABLE 4.—LINK FORMATION REGRESSION RESULTS

Technological Similarity	Jaffe	Mahalanobis
Past collaboration	0.5981^{***}	0.5920^{***}
	(0.0150)	(0.0149)
Past common collaborator	0.1162^{***}	0.1164^{***}
	(0.0238)	(0.0236)
$f_{ij,t-s-1}$	13.6977^{***}	6.0864^{***}
	(0.6884)	(0.3323)
$f^2_{ij,t-s-1}$	-20.4083^{***}	-3.9194^{***}
	(1.7408)	(0.4632)
$city_{ij}$	1.1283^{***}	1.1401^{***}
	(0.1017)	(0.1017)
$market_{ij}$	0.8451^{***}	0.8561^{***}
	(0.0424)	(0.0422)
Number of observations	3,964,120	3,964,120
McFadden's R^2	0.0812	0.0813

The dependent variable $a_{ij,t}$ indicates if an R&D alliance exists between firms i and j at time t. Statistically significant at $^{***}1\%$, $^{**}5\%$, $^{*}10\%$.

Rose (2004, AER)

TABLE 1—BENCHMARK RESULTS

	Default	No industrial countries	Post 1970	With country effects
Both in GATT/WTO	−0.04	−0.21	−0.08	0.15
	(0.05)	(0.07)	(0.07)	(0.05)
One in GATT/WTO	−0.06	−0.20	−0.09	0.05
	(0.05)	(0.06)	(0.07)	(0.04)
GSP	0.86	0.04	0.84	0.70
	(0.03)	(0.10)	(0.03)	(0.03)
Log distance	−1.12	−1.23	−1.22	−1.31
	(0.02)	(0.03)	(0.02)	(0.02)
Log product real GDP	0.92	0.96	0.95	0.16
	(0.01)	(0.02)	(0.01)	(0.05)
Log product real GDP p/c	0.32	0.20	0.32	0.54
	(0.01)	(0.02)	(0.02)	(0.05)
Regional FTA	1.20	1.50	1.10	0.94
	(0.11)	(0.15)	(0.12)	(0.13)
Currency union	1.12	1.00	1.23	1.19
	(0.12)	(0.15)	(0.15)	(0.12)
Common language	0.31	0.10	0.35	0.27
	(0.04)	(0.06)	(0.04)	(0.04)
Land border	0.53	0.72	0.69	0.28
	(0.11)	(0.12)	(0.12)	(0.11)
Number landlocked	−0.27	−0.28	−0.31	−1.54
	(0.03)	(0.05)	(0.03)	(0.32)
Number islands	0.04	−0.14	0.03	−0.87
	(0.04)	(0.06)	(0.04)	(0.19)
Log product land area	−0.10	−0.17	−0.10	0.38
	(0.01)	(0.01)	(0.01)	(0.03)
Common colonizer	0.58	0.73	0.52	0.60
	(0.07)	(0.07)	(0.07)	(0.06)
Currently colonized	1.08	—	1.12	0.72
	(0.23)		(0.41)	(0.26)
Ever colony	1.16	−0.42	1.28	1.27
	(0.12)	(0.57)	(0.12)	(0.11)
Common country	−0.02	—	−0.32	0.31
	(1.08)		(1.04)	(0.58)
Observations	234,597	114,615	183,328	234,597
R^2	0.65	0.47	0.65	0.70
RMSE	1.98	2.36	2.10	1.82

Notes: Regressand: log real trade. OLS with year effects (intercepts not reported). Robust standard errors (clustering by country-pairs) are in parentheses.

Apicella, Marlowe, Fowler & Christakis (2011, Nature)

RESEARCH SUPPLEMENTARY INFORMATION

Supplementary Table S16: GEE Regression of Social Ties on Public Good Donations

	Dependent Variable: Ego Wants to Camp with Alter			***Dependent Variable: Ego Gives Gift to Alter***		
	Coef.	*S.E.*	*p*	*Coef.*	*S.E.*	*p*
Ego Public Good Donation	0.003	0.031	0.930	-0.022	0.044	0.627
Alter Public Good Donation	-0.026	0.044	0.550	-0.100	0.047	0.035
Ego-Alter Similarity in Public Good Donation	0.250	0.051	0.000	0.174	0.044	0.000
Residual		5879			2096	
Null Residual		5923			2113	
N		18054			2310	

GEE logit regression of presence of social tie from ego to alter on ego and alter attributes, clustering standard errors on each ego.

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.

How to Conduct Inference?

Dyads present an ironic situation in that dyadic data sets, with 100,000 cases (or often considerably more), may seem ideal for hypothesis testing. Yet, the structure of dyadic data complicates the assessment to statistical significance. Because dyadic observations are not independent events, the usual tests of significance result in overconfidence, even when the model itself appears to be correctly specified (Erikson, Pinto & Rader, 2014, p. 457).

How to Conduct Inference? (continued)

Dyadic observations are not independent. This is due to the presence of individual-specific factors common to all observations involving that individual. It is thus reasonable to assume that $\mathbb{E}\left[u_{ij}u_{ik}\right] \neq 0$ for all k and $\mathbb{E}\left[u_{ij}u_{kj}\right] \neq 0$ for all k. By the same reasoning, we also have $\mathbb{E}\left[u_{ij}u_{jk}\right] \neq 0$ and $\mathbb{E}\left[u_{ij}u_{ki}\right] \neq 0$. Provided that regressors are exogenous,...OLS...yields consistent coefficient estimates but standard errors are inconsistent, leading to incorrect inference (Fafchamps and Gubert, 2007, p. 330).

Existing suggestions

1. Permutation approaches: quadratic assignment procedure (QAP) of Hubert (1985, PM), Krackhardt (1988, SN)

2. Integrated likelihood/MCMC: p_2 model of van Duijn, Snijders and Zijlstra (2004, SN), Zijlstra, van Duijn and Snijders (2009, BJMSP), Krivitsky, Handcock, Raftery and Hoff (2009, SN) – emerging frequentist theory.

Existing suggestions (continued)

3. Pairwise/composite likelihood: Bellio and Varin (2005, SM) – no frequentist theory.

4. Dyadic cluster-robust s.e.: Fafchamps and Gubert (2007, JDE), Cameron and Miller (2014, WP), Aronow, Samii and Assenova (2015, PA), Tabord-Meehan (2018, JBES) – emerging frequentist theory.

Dyadic Regression: Notation & Setup

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

- will focus on binary case with $Y_{ij} = D_{ij} \in \{0, 1\}$

Let X_i be a vector of agent-level covariates.

Let U_i be unobserved agent-level heterogeneity.

Dyadic Regression: Notation & Setup (continued)

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

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

Here i and j denote two independent random draws from the population of interest.

Dyadic Regression: Nonparametric DGP

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.

May be possible to motivate this DGP via exchangeability arguments (e.g., Aldous-Hoover Theorem). See Menzel (2018).

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., Frank and Strauss, 1986, JASA; Fafchamps and Gubert, 2007, JDE).

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

Dyadic Regression: Parametric estimation

A prototypical specification for a binary outcome is

$$\text{logit}\left[g\left(X_i, X_j; \theta_0\right)\right] = \alpha + \left[t\left(X_i\right) + t\left(X_j\right)\right]' \beta + \omega\left(X_i, X_j\right)' \gamma$$

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 pseudo-likelihood 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 the logit kernel (cf., Cox and Reid, 2005, BM).

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 $s\left(Z_{ij};\theta\right)=\frac{\partial l(Z_{ij};\theta)}{\partial\theta}$ and $H_N\left(\theta\right)=\binom{N}{2}^{-1}\sum_{i<j}\frac{\partial^2 l(Z_{ij};\theta)}{\partial\theta\partial\theta'}$.

Dyadic Regression: Parametric estimation (continued)

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

...also not a U-Statistic (D_{ij} is a dyad-level random variable), 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, JDE) propose a now widely-used dyadic-clustered covariance estimator (cf., Cameron and Miller, 2014, WP; Aronow et al., 2017, PA).

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.

Fafchamps-Gubert (2007) Variance Estimate

Observe that

$$\Sigma_1 = \mathbb{E}\left[s\left(Z_{12};\theta_0\right)s\left(Z_{13};\theta_0\right)'\right]$$
$$\Sigma_2 = \mathbb{E}\left[s\left(Z_{12};\theta_0\right)s\left(Z_{12};\theta_0\right)'\right]$$

Natural analog estimates for these two terms are

$$\begin{aligned}\widehat{\Sigma}_1 =& \binom{N}{3}\sum_{i<j<k}\frac{1}{3}\left\{s\left(Z_{ij};\widehat{\theta}\right)s\left(Z_{ik};\widehat{\theta}\right)'\right.\\ &+ s\left(Z_{ij};\widehat{\theta}\right)s\left(Z_{jk};\widehat{\theta}\right)'\\ &\left.+ s\left(Z_{ik};\widehat{\theta}\right)s\left(Z_{jk};\widehat{\theta}\right)'\right\}\\ \widehat{\Sigma}_2 =& \binom{N}{2}^{-1}\sum_{i<j}s\left(Z_{ij};\widehat{\theta}\right)s\left(Z_{ij};\widehat{\theta}\right)'\end{aligned}$$

Degrees of freedom corrections? These are unbiased variance estimates when $s\left(Z_{ij};\widehat{\theta}\right)$ is replaced by $s\left(Z_{ij};\theta_0\right)$.

Fafchamps-Gubert (2007) Variance Estimate (continued)

Fafchamps and Gubert (2007) proposed the estimate

$$\hat{\Lambda}_{\text{FG}} = \frac{1}{N\left(N-1\right)^2} \sum_{i_1} \sum_{i_2 \neq i_1} \sum_{j_1} \sum_{j_2 \neq j_1} C_{i_1 i_2 j_1 j_2} s\left(Z_{i_1 i_2}; \widehat{\theta}\right) s\left(Z_{j_1 j_2}; \widehat{\theta}\right)'$$

with $C_{i_1 i_2 j_1 j_2} = 1$ whenever $\{i_1, i_2\}$ and $\{j_1, j_2\}$ share at least one index in common.

After much tedious algebra it is possible to show that the (brute-force) Fafchamps and Gubert (2007) estimator equals:

$$\hat{\Lambda}_{\text{FG}} = 4\widehat{\Sigma}_1 + \frac{2}{N-1}\left(\widehat{\Sigma}_2 - 2\widehat{\Sigma}_1\right),$$

which is a natural analog variance estimate (including "higher order" terms)!

Dyadic Regression: Variance estimation (continued)

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

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

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

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

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

When network is sparse these differences appear to be important.

Jack-knife Variance Estimate (continued)

The Jack-knife variance estimate is conservative (Efron and Stein, 1981).

Here we can show that

$$\hat{\Lambda}_{\mathrm{JK}} = 4\hat{\Sigma}_1 + \frac{4}{N-1}\left(\hat{\Sigma}_2 - \hat{\Sigma}_1\right).$$

Suggesting the corrected estimate (cf., Cattaneo, Crump and Jansson, 2014)

$$\hat{\Lambda}_{\mathrm{JK-C}} = \hat{\Lambda}_{\mathrm{JK}} - \frac{2\hat{\Sigma}_2}{N-1}$$

which identically equals the one proposed by Fafchamps and Gubert (2007).

Jack-knife Variance Estimate

To connect this estimate to the familiar Jack-knife idea let

$$\bar{s}_{-i}\left(\hat{\theta}\right) = \frac{2}{(N-1)(N-2)}\left[\sum_{j<k} s\left(Z_{jk};\hat{\theta}\right) - \sum_{j\neq i} s\left(Z_{ij};\hat{\theta}\right)\right].$$

With some algebra we get

$$\begin{aligned}\bar{s}_{-i}\left(\hat{\theta}\right) - \bar{s}\left(\hat{\theta}\right) &= \frac{2}{N-2}\left[\bar{s}\left(\hat{\theta}\right) - \hat{\bar{s}}_{1i}\left(\hat{\theta}\right)\right] \\ &= -\frac{2}{N-2}\hat{\bar{s}}_{1i}\left(\hat{\theta}\right)\end{aligned}$$

and hence that

$$\left(\frac{N-2}{N}\right)^2 \sum_i \left(\bar{s}_{-i}\left(\hat{\theta}\right) - \bar{s}\left(\hat{\theta}\right)\right)\left(\bar{s}_{-i}\left(\hat{\theta}\right) - \bar{s}\left(\hat{\theta}\right)\right)' = \frac{4\tilde{\Sigma}_1}{N}$$

cf. Frank and Snijders (1994)

Pigeonhole Bootstrap

Sample $k_1, k_2, \ldots, k_N$ with replacement from the index set $\{1, 2, \ldots, N\}$.

Compute the bootstrap mean (for example):

$$\bar{Y}^{(b)} = \binom{N}{2}^{-1} \sum_{i<j} Y_{k_i k_j}.$$

Calculate its variance as

$$\hat{\mathbb{V}}\left(\bar{Y}\right) = \frac{1}{B-1} \sum_{b=1}^{N} \left(\bar{Y}^{(b)} - \bar{Y}^{(\cdot)}\right)^2$$

with B the number of bootstrap networks and $\bar{Y}^{(\cdot)} = \frac{1}{B} \sum_{b=1}^{N} \bar{Y}^{(b)}$.

See Snijders and Borgatti (1999) and Owen (2007).

See Menzel (2018) for limitations and modifications.

Dyadic Regression: Asymptotic Normality

Let $r\left(X_i, X_j\right) = \left(1, t\left(X_i\right)' + t\left(X_j\right)', \omega\left(X_i, X_j\right)'\right)'$ and consider the decomposition:

$$\begin{aligned}
\sqrt{N} S_N\left(\theta_0\right) &= \sqrt{N}\binom{N}{2}^{-1} \sum_{i<j} \left\{D_{ij} - g\left(X_i, X_j; \theta_0\right)\right\} r\left(X_i, X_j\right) \\
&= \sqrt{N}\binom{N}{2}^{-1} \sum_{i<j} \left\{h\left(X_i, X_j, U_i, U_j\right) - g\left(X_i, X_j; \theta_0\right)\right\} r\left(X_i, X_j\right) \\
&+ \sqrt{N}\binom{N}{2}^{-1} \sum_{i<j} \left\{D_{ij} - h\left(X_i, X_j, U_i, U_j\right)\right\} r\left(X_i, X_j\right) \\
&= \sqrt{N} V_N + \sqrt{N} W_N
\end{aligned}$$

Dyadic Regression: Asymptotic Normality

V_N *is* a textbook U-statistic with

$$\sqrt{N}V_N \xrightarrow{D} \mathcal{N}(0, 4\Sigma_1)$$

...while $\mathbb{C}(V_N, W_N) = 0$ with $\mathbb{V}\left(\sqrt{N}W_N\right) = O\left(\frac{1}{N}\right)$

...so we can argue $\sqrt{N}S_N(\theta_0) \xrightarrow{D} \mathcal{N}(0, 4\Sigma_1)$

In practice it appears to be better to use the approximation

$$\sqrt{N}S_N(\theta_0) \overset{approx}{\sim} \mathcal{N}\left(0, 4\Sigma_1 + \frac{2}{N-1}(\Sigma_2 - 2\Sigma_1)\right)$$

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 (one) emerging practice in economics has a coherent foundation.