Dyadic Regression

An Introduction to the Econometrics of Networks

CEMFI, November 28th to December 2nd, 2022

Bryan S. Graham

University of California - Berkeley

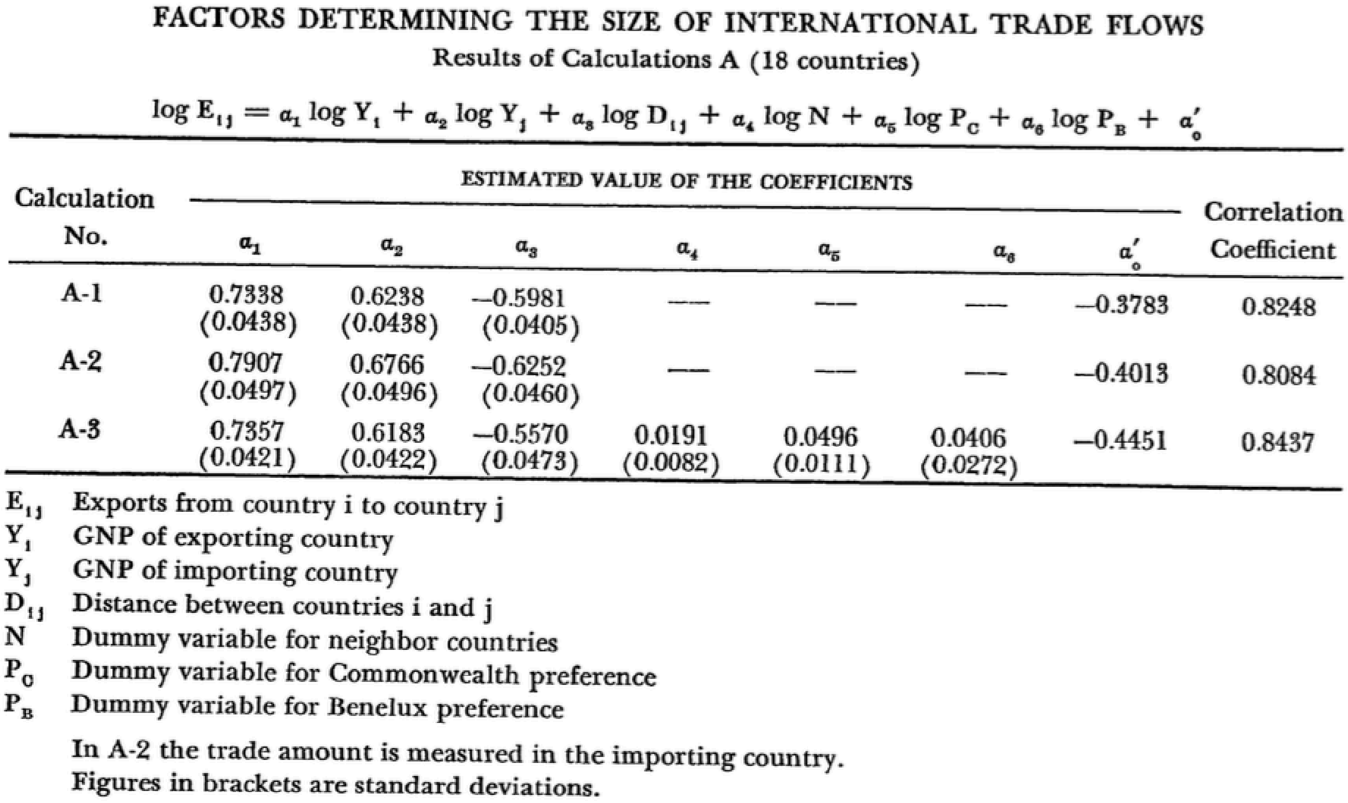
Dyadic regression analyses are abundant in social science research (see below).

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 are used in practice.

Tinbergen (1962, SWE, Table VI-1)



FACTORS DETERMINING THE SIZE OF INTERNATIONAL TRADE FLOWS

Results of Calculations A (18 countries)

$$\log E_{ij} = \alpha_1 \log Y_i + \alpha_2 \log Y_j + \alpha_3 \log D_{ij} + \alpha_4 \log N + \alpha_5 \log P_C + \alpha_6 \log P_B + \alpha'_0$$

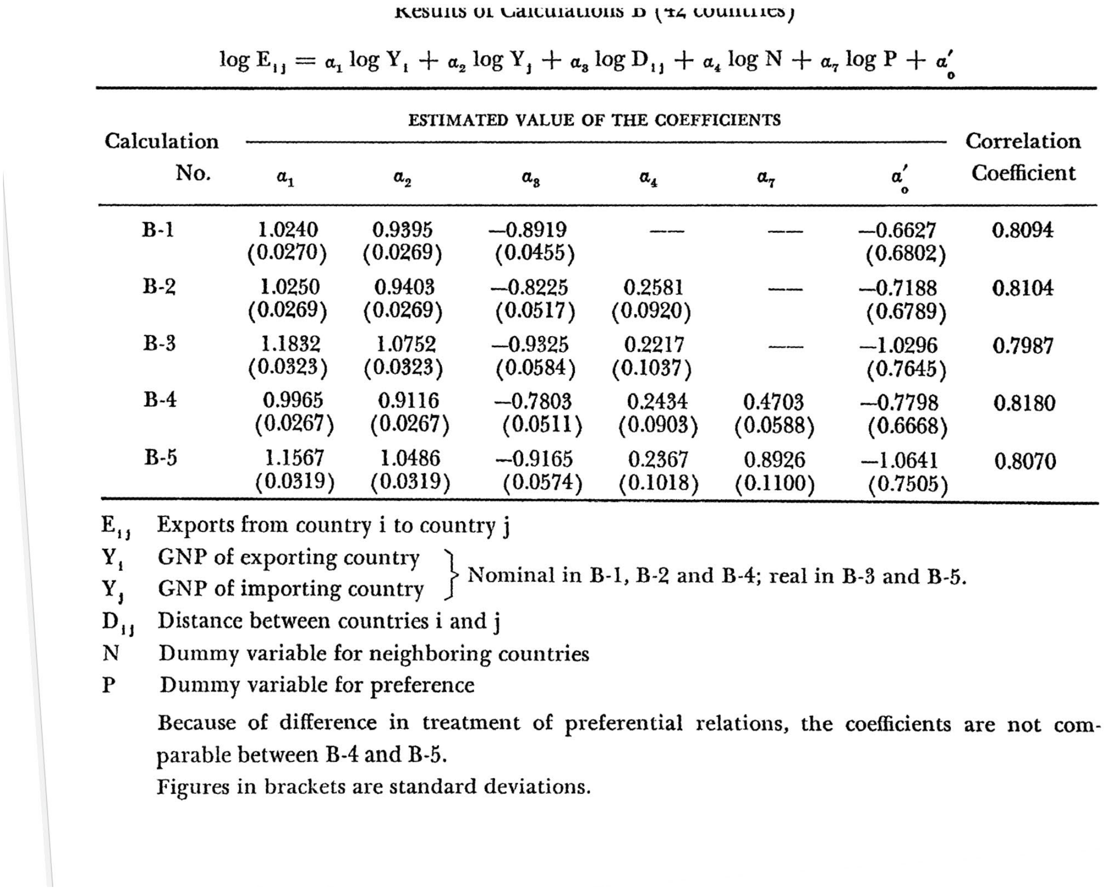
Calculation No.	ESTIMATED VALUE OF THE COEFFICIENTS							Correlation Coefficient
	α_1	α_2	α_3	α_4	α_5	α_6	α'_0	
A-1	0.7338 (0.0438)	0.6238 (0.0438)	–0.5981 (0.0405)	—	—	—	–0.3783	0.8248
A-2	0.7907 (0.0497)	0.6766 (0.0496)	–0.6252 (0.0460)	—	—	—	–0.4013	0.8084
A-3	0.7357 (0.0421)	0.6183 (0.0422)	–0.5570 (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)

$$\log E_{ij} = a_1 \log Y_i + a_2 \log Y_j + a_3 \log D_{ij} + a_4 \log N + a_7 \log P + a'_o$$

Calculation No.	ESTIMATED VALUE OF THE COEFFICIENTS						Correlation Coefficient
	a_1	a_2	a_3	a_4	a_7	a'_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 } Nominal in B-1, B-2 and B-4; real in B-3 and B-5.

Y_j GNP of importing country }

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)

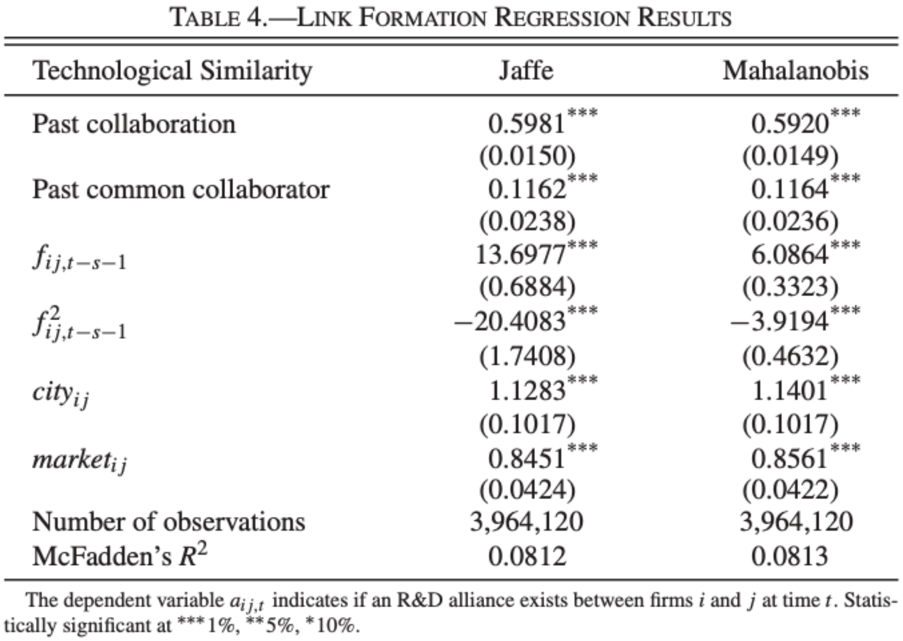


TABLE 4.—LINK FORMATION REGRESSION RESULTS

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)
$f_{ij,t-s-1}^2$	-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)

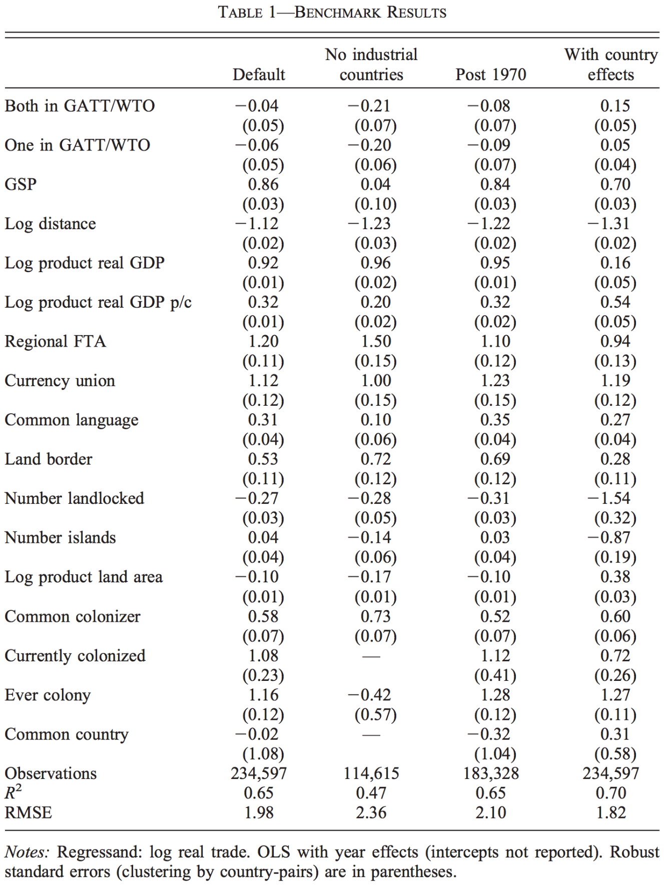
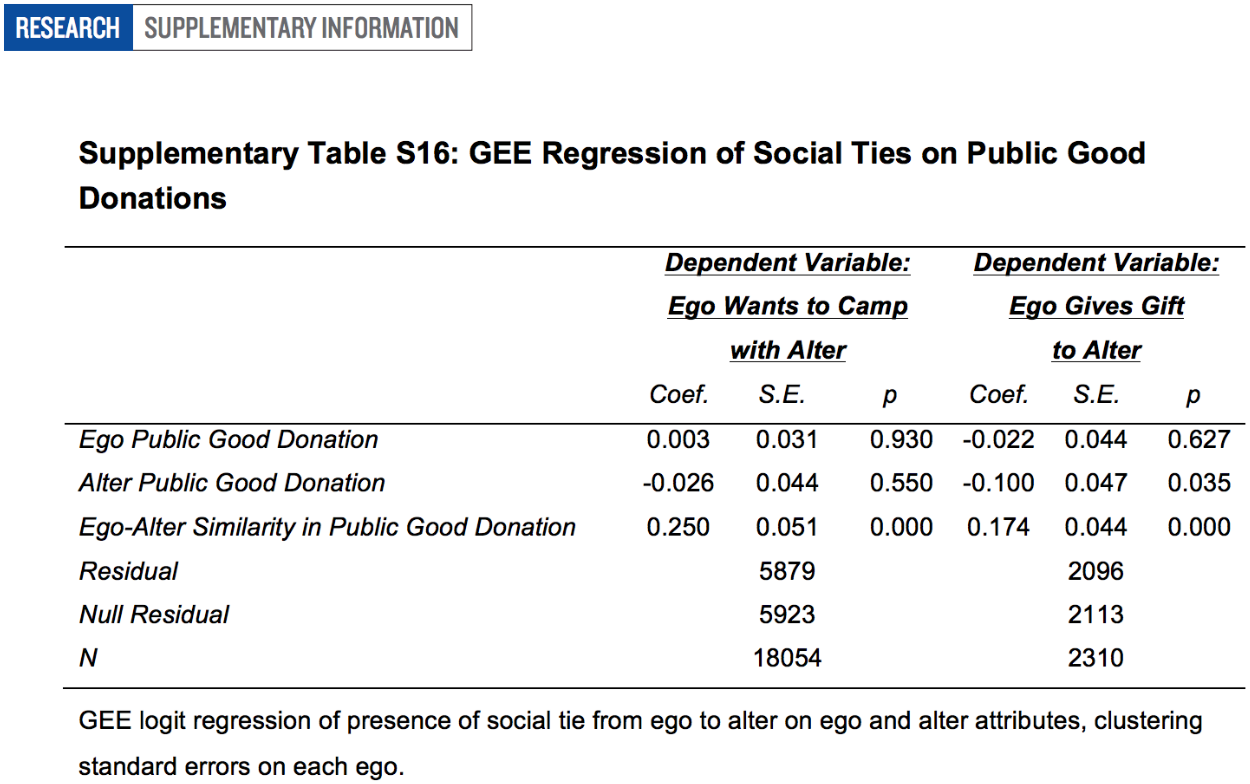


Table 1—Benchmark Results

	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)

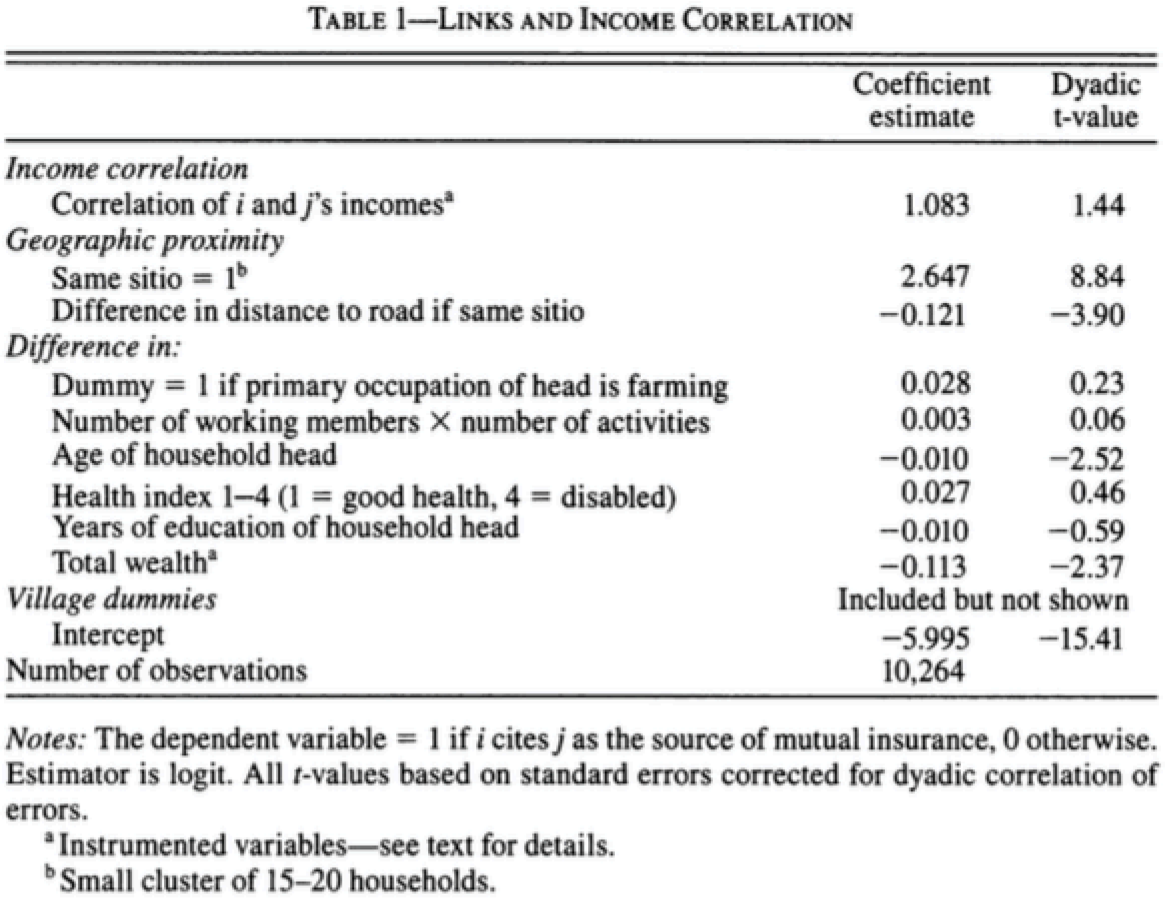


TABLE 1—LINKS AND INCOME CORRELATION

	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).

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).

   - Frequentist theory for these approaches is very new (e.g., Graham, 2020a,b; Menzel, 2021).

Dyadic Regression: Notation & Setup

Let $i \in \mathbb{N}$ index agents in an infinite population of interest. Observable attribute $X_i \in \mathbb{X} = \{x_1, \ldots, x_L\}$.

Attribute partitions the population into $L = |\mathbb{X}|$ subpopulations which I will refer to as "types".

Let $\mathbb{N}(x) = \{i : X_i = x_l\}$ is the index set for type l agents.

Dyadic Regression: Notation & Setup

Associated with each ordered pair of agents is the scalar directed outcome $Y_{ij} \in \mathbb{Y} \subseteq \mathbb{R}$.

I will refer to agent i as the "ego" of the directed dyad and agent j as its "alter".

In the context of the trade example the ego agent is the exporting country, the alter the importing one.

The *adjacency matrix* $\left[Y_{ij}\right]_{i,j\in\mathbb{N}}$ collects all such outcomes into an infinite random array.

Within-type or X-Exchangeability

From the standpoint of the econometrician, the indexing of agents within subpopulations homogenous in X_i is arbitrary: agents of the same type are exchangeable.

Exchangeability of agents within subpopulations homogenous in X_i induces a particular form of exchangeability on the adjacency matrix.

This form of exchangeability, in turn, induces a particular form of dependence across the rows and columns of $\left[Y_{ij}\right]_{i,j\in\mathbb{N}}$.

The structure of this dependence allows for the formulation of LLNs and CLTs.

Within-type or X-Exchangeability (continued)

Let $\sigma_x : \mathbb{N} \to \mathbb{N}$ be any permutation of a finite number of the agent indices which satisfies the restriction

$$\left[X_{\sigma_x(i)}\right]_{i\in\mathbb{N}} = [X_i]_{i\in\mathbb{N}} . \tag{1}$$

Condition (1) constrains index permutations to occur among agents of the same type.

A network is *relatively exchangeable* with respect to X (or X-exchangeable) if

$$\left[Y_{\sigma_x(i)\sigma_x(j)}\right]_{i,j\in\mathbb{N}} \overset{D}{=} \left[Y_{ij}\right]_{i,j\in\mathbb{N}} \tag{2}$$

for all permutations σ_x satisfying (1).

X-exchangeablility is a natural generalization of joint exchangeability, as introduced in the context of Aldous-Hoover Theorem.

Graphon & Dyadic Regression Function

Let α, $\{(U_i, X_i)\}_{i\geq 1}$ and $\left\{\left(V_{ij}, V_{ji}\right)\right\}_{i\geq 1, j\geq 1}$ be (sequences of) i.i.d. random variables, additionally independent of one another.

Consider the random array $\left[Y_{ij}^{*}\right]_{i,j\in\mathbb{N}}$ generated according to the rule

$$Y_{ij}^{*} = \tilde{h}\left(\alpha, X_i, X_j, U_i, U_j, V_{ij}\right) \tag{3}$$

with $\tilde{h} : [0,1] \times \mathbb{X} \times \mathbb{X} \times [0,1]^3 \rightarrow \mathbb{Y}$ a measurable function.

A graph generated according to (3) is X-exchangeable.

Graphon & Dyadic Regression Function (continued)

Here α is a mixing parameter analogous to the one appearing in de Finetti's (1931) original representation theorem.

I will depress the dependence of $\tilde{h}$ on α, defining the notation $h\left(X_i, X_j, U_i, U_j, V_{ij}\right) \overset{def}{\equiv} \tilde{h}\left(\alpha, X_i, X_j, U_i, U_j, V_{ij}\right)$.

The function $h : \mathbb{X} \times \mathbb{X} \times [0,1]^3 \to \mathbb{Y}$ will be referred to as a **graphon**.

Graphon & Dyadic Regression Function (continued)

Crane and Towsner (2018), in an extension of the Aldous-Hoover representation result, show that for any X-exchangeable random array $\left[Y_{ij}\right]_{i,j\in\mathbb{N}}$ there exists another array $\left[Y_{ij}^{*}\right]_{i,j\in\mathbb{N}}$ generated according to (3) such that the two arrays have the same distribution:

$$\left[Y_{ij}\right]_{i,j\in\mathbb{N}} \overset{D}{=} \left[Y_{ij}^{*}\right]_{i,j\in\mathbb{N}}. \tag{4}$$

We can therefore use (3) as an ‘as if’ non-parametric data generating process for $\left[Y_{ij}\right]_{i,j\in\mathbb{N}}$; this will facilitate a variety of probabilistic calculations (e.g., computing conditional expectations, variances and, especially, covariances).

Sampling

Let $i = 1, \ldots, N$ index a simple random sample from the target population.

For each of the N sampled units the econometrician observes X_i and for each of the $\binom{N}{2}$ sampled dyads she observes $\left(Y_{ij}, Y_{ji}\right)$.

Composite Likelihood

Let $\left\{ f_{Y_{12}|X_1,X_2}(Y_{12}|\, X_1, X_2; \theta) : \theta \in \Theta \subseteq \mathbb{R}^{\dim(\theta)} \right\}$ be a parametric family of distributions for the conditional distribution of Y_{12} given X_1 and X_2.

We might model trade from exporter i to importer j given covariates as a Poisson random variable:

$$f_{Y_{12}|X_1,X_2}\left(y_{ij} \middle| X_i, X_j; \theta \right) = \exp\left[-\exp\left[W_{ij}'\theta \right] \right] \frac{\left\{ \exp\left[W_{ij}'\theta \right] \right\}^{y_{ij}}}{y_{ij}!} \qquad (5)$$

with $y_{ij} = 0, 1, 2, \ldots$ and $W_{ij} \overset{def}{\equiv} w\left(X_i, X_j \right)$ a known $J \times 1$ vector of functions of X_i and X_j.

Composite Likelihood (continued)

If $X_i = (\ln \mathtt{GDP_i}, \mathtt{LAT_i}, \mathtt{LONG_i})'$, then setting

$$W_{ij} = \begin{pmatrix} \ln \mathtt{GDP_i} \\ \ln \mathtt{GDP_j} \\ \ln \left[\left(\mathtt{LAT_i} - \mathtt{LAT_j} \right)^2 + \left(\mathtt{LONG_i} - \mathtt{LONG_j} \right)^2 \right]^{1/2} \end{pmatrix}$$

results in a basic gravity trade model specification.

Composite Likelihood (continued)

(5) only specifies the marginal distribution of Y_{ij} given X_i and X_j.

The econometrician is not asserting, for example, that

$$\begin{aligned} f_{Y_{12},Y_{13}|X_1,X_2,X_3}(y_{12},y_{13}|\,X_1,X_2,X_3;\theta) = & f_{Y_{12}|X_1,X_2}(y_{12}|\,X_1,X_2;\theta) \\ & \times f_{Y_{12}|X_1,X_2}(y_{13}|\,X_1,X_3;\theta)\,; \end{aligned}$$

since doing so would imply independence of Y_{12} and Y_{13} given covariates.

Composite Likelihood (continued)

Formulating a conditional likelihood for the entire adjacency matrix $\mathbf{Y} \overset{def}{\equiv} \left[Y_{ij}\right]_{1\leq i,j\leq N, i\neq j}$ given $\mathbf{X} \overset{def}{\equiv} [X_i]_{1\leq i\leq N}$ would require an explicit specification of the dependence structure across dyads sharing agents in common.

In contrast $f_{Y_{12}|X_1,X_2}(Y_{12}|\, X_1, X_2; \theta)$, which is a model for the marginal distribution of Y_{12} alone, does not require modeling this dependence.

Composite Likelihood (continued)

Let $l_{ij}\left(\theta\right) = \ln f_{Y_{12}|X_1,X_2}\left(Y_{ij}\middle| X_i, X_j; \theta\right)$ and consider the estimator which chooses $\widehat{\theta}$ to maximize:

$$\widehat{\theta} = \arg\max_{\theta \in \Theta} \frac{1}{N\left(N-1\right)} \sum_{i=1}^{N} \sum_{j \neq i} l_{ij}\left(\theta\right). \tag{6}$$

Because its summands are not independent of one another – at least those sharing indices in common are not – (6) does not correspond to a log-likelihood function for $\mathbf{Y}$ given $\mathbf{X}$.

Instead it corresponds to what is sometimes called a composite log-likelihood.

Composite Likelihood (continued)

While an appropriately specified composite log-likelihood typically delivers a valid estimating equation, accurate inference is more challenging

The unmodeled dependence structure in the data needs to be explicitly taken into account at the inference stage.

Limit Distribution

A mean value expansion of the first order condition associated with the maximizer of (6) yields,

$$\sqrt{N}\left(\widehat{\theta}-\theta_0\right)=\left[-H_N\left(\bar{\theta}\right)\right]^{+}\sqrt{N}S_N\left(\theta_0\right)$$

with $\bar{\theta}$ a mean value between $\widehat{\theta}$ and θ_0 which may vary from row to row and the $+$ superscript denoting a Moore-Penrose inverse.

Here $S_N\left(\theta_0\right)$ is the "score" vector

$$S_N\left(\theta\right)=\frac{1}{N}\frac{1}{N-1}\sum_i\sum_{j\neq i}s_{ij}\left(Z_{ij},\theta\right) \tag{7}$$

with $s\left(Z_{ij},\theta\right)=\partial l_{ij}\left(\theta\right)/\partial\theta$ for $Z_{ij}=\left(Y_{ij},X_i',X_j'\right)'$ and $H_N\left(\theta\right)=\frac{1}{N}\frac{1}{N-1}\sum_i\sum_{j\neq i}\frac{\partial^2 l_{ij}(\theta)}{\partial\theta\partial\theta'}$.

Limit Distribution (continued)

If the Hessian matrix $H_N\left(\bar{\theta}\right)$ converges in probability to the invertible matrix Γ_0, as I will assume, then

$$\sqrt{N}\left(\widehat{\theta}-\theta_0\right)=-\Gamma_0^{-1}\sqrt{N}S_N\left(\theta_0\right)+o_p\left(1\right)$$

so that the asymptotic sampling properties of $\sqrt{N}\left(\widehat{\theta}-\theta_0\right)$ will be driven by the behavior of $\sqrt{N}S_N\left(\theta_0\right)$.

As with the composite log-likelihood criterion function, the summands of $\sqrt{N}S_N\left(\theta_0\right)$ are not independent of one another.

A standard central limit theorem cannot be used to demonstrate asymptotic normality of $\sqrt{N}S_N\left(\theta_0\right)$.

Limit Distribution (continued)

Fortunately $S_N(\theta_0)$, although not a U-Statistic, has a dependence structure similar to one. This insight can be used to derive the limit properties of $\sqrt{N}\left(\widehat{\theta}-\theta_0\right)$.

Begin by re-writing $S_N(\theta_0)$ as

$$S_N = \binom{N}{2}^{-1} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \frac{s_{ij}+s_{ji}}{2}, \tag{8}$$

where $s_{ij} \overset{def}{\equiv} s\left(Z_{ij},\theta_0\right)$ and $S_N \overset{def}{\equiv} S_N(\theta_0)$.

While (8) has the cursory appearance of a U-Statistic it is, in fact not one: Y_{ij}, which enters s_{ij}, varies at the dyad level, hence S_N is not a function of N i.i.d. random variables.

Limit Distribution (continued)

Let $\mathbf{U} = [U_i]_{1 \leq i \leq N}$; the projection of S_N onto the *observed* covariate matrix $\mathbf{X}$ and the *unobserved* vector of unit-specific effects $\mathbf{U}$ equals:

$$V_N \overset{def}{\equiv} \mathbb{E}\left[S_N \middle| \mathbf{X}, \mathbf{U}\right] = \binom{N}{2}^{-1} \sum_{i<j} \frac{\bar{s}_{ij} + \bar{s}_{ji}}{2} \tag{9}$$

with $\bar{s}_{ij} \overset{def}{\equiv} \bar{s}\left(X_i, U_i, X_j, U_j\right)$ and

$$\bar{s}\left(X_i, U_i, X_j, U_j\right) \overset{def}{\equiv} \mathbb{E}\left[s\left(Z_{ij}, \theta_0\right) \middle| X_i, U_i, X_j, U_j\right].$$

Limit Distribution (continued)

The projection (9) *is* a U-statistic of order two: specifically it is a summation over all $\binom{N}{2}$ dyads that can be formed from the i.i.d. sample $\{(X_i, U_i)\}_{1 \leq i \leq N}$.

Unusually our U-statistic is defined in terms of a combination of both *observed* $\{X_i\}_{1 \leq i \leq N}$ and *unobserved* $\{U_i\}_{1 \leq i \leq N}$ random variables.

The projection error $T_N = S_N - V_N$ consists of a summation of $\binom{N}{2}$ conditionally uncorrelated summands; hence $\mathbb{V}(T_N) = \binom{N}{2}^{-1} \mathbb{E}\left(\mathbb{V}\left(\left.\frac{s_{12}+s_{21}}{2}\right| X_1, U_1, X_2, U_2\right)\right) = O\left(N^{-2}\right)$.

We also have that T_N and V_N are uncorrelated by construction.

Limit Distribution (continued)

Although we cannot numerically compute V_N – even if θ_0 is known – because the $\{U_i\}_{1\leq i\leq N}$ are unobserved, we can use the theory of U-statistics to characterize its sampling properties as $N \to \infty$.

Decomposing V_N into a Hájek projection and a second remainder term yields:

$$V_N = V_{1N} + V_{2N}.$$

Limit Distribution (continued)

Continuing from the prior slide...

$$V_{1N} = \frac{2}{N}\sum_{i=1}^{N}\left\{\frac{\bar{s}_1^e\left(X_i,U_i\right)+\bar{s}_1^a\left(X_i,U_i\right)}{2}\right\} \tag{10}$$

$$\begin{aligned} V_{2N} = & \binom{N}{2}^{-1}\sum_{i<j}\left\{\frac{\bar{s}_{ij}+\bar{s}_{ji}}{2}\right. \\ & \left. -\frac{\bar{s}_1^e\left(X_i,U_i\right)+\bar{s}_1^a\left(X_i,U_i\right)}{2}-\frac{\bar{s}_1^e\left(X_j,U_j\right)+\bar{s}_1^a\left(X_j,U_j\right)}{2}\right\}. \end{aligned} \tag{11}$$

where $\bar{s}^e\left(x,u\right)=\mathbb{E}\left[\bar{s}\left(x,u,X_1,U_1\right)\right]$ and $\bar{s}^a\left(x,u\right)=\mathbb{E}\left[\bar{s}\left(X_1,U_1,x,u\right)\right]$.

The superscript ‘e’ denotes ‘ego’, while ‘a’ denotes ‘alter’.

Limit Distribution (continued)

Conveniently V_{1N} is a sum of i.i.d. random variables to which, after scaling by $\sqrt{N}$, a CLT may be applied.

Furthermore it can be shown that $\mathbb{V}\left(V_{2N}\right) = O\left(N^{-2}\right)$.

Limit Distribution (continued)

Putting these results together yields the asymptotically linear representation

$$\begin{aligned}\sqrt{N}\left(\widehat{\theta}-\theta_0\right) &= -\,\Gamma_0^{-1}\sqrt{N}\left(V_{1N}+V_{2N}+T_N\right)+o_p\left(1\right)\\ &= -\,\Gamma_0^{-1}\sqrt{N}V_{1N}+o_p\left(1\right)\\ &= -\,\Gamma_0^{-1}\frac{2}{\sqrt{N}}\sum_{i=1}^{N}\left\{\frac{\overline{s}_1^e\left(X_i,U_i\right)+\overline{s}_1^a\left(X_i,U_i\right)}{2}\right\}+o_p\left(1\right),\end{aligned}$$

and hence a limit distribution for $\sqrt{N}\left(\widehat{\theta}-\theta_0\right)$ of

$$\sqrt{N}\left(\widehat{\theta}-\theta_0\right)\overset{D}{\rightarrow}\mathcal{N}\left(0,4\left(\Gamma_0'\Sigma_1^{-1}\Gamma_0\right)^{-1}\right) \qquad (12)$$

where $\Sigma_1=\mathbb{V}\left(\frac{\overline{s}_1^e(X_1,U_1)+\overline{s}_1^a(X_1,U_1)}{2}\right)$.

Variance Estimation

An Anova decomposition gives

$$\begin{aligned}\mathbb{V}\left(S_N\right) &= \mathbb{V}\left(\mathbb{E}\left[S_N \middle| \mathbf{X}, \mathbf{U}\right]\right) + \mathbb{E}\left[\mathbb{V}\left(S_N \middle| \mathbf{X}, \mathbf{U}\right)\right] \\ &= \mathbb{V}\left(V_N\right) + \mathbb{V}\left(T_N\right) \\ &= \mathbb{V}\left(V_{1N}\right) + \mathbb{V}\left(V_{2N}\right) + \mathbb{V}\left(T_N\right). \end{aligned} \tag{13}$$

Variance Estimation (continued)

Let $p = 1, 2$ equal the number of agents dyads $\{i_1, i_2\}$ and $\{j_1, j_2\}$ share common and define the matrix Σ_p as

$$\Sigma_p \overset{def}{\equiv} \mathbb{C}\left(\frac{\bar{s}\left(X_{i_1}, U_{i_1}, X_{i_2}, U_{i_2}\right) + \bar{s}\left(X_{i_2}, U_{i_2}, X_{i_1}, U_{i_1}\right)}{2},\right. \tag{14}$$

$$\left.\frac{\bar{s}\left(X_{j_1}, U_{j_1}, X_{j_2}, U_{j_2}\right)' + \bar{s}\left(X_{j_2}, U_{j_2}, X_{j_1}, U_{j_1}\right)'}{2}\right)$$

$$= \mathbb{C}\left(\mathbb{E}\left[\left.\frac{s_{i_1 i_2} + s_{i_2 i_1}}{2}\right| X_{i_1}, U_{i_1}, X_{i_2}, U_{i_2}\right],\right.$$

$$\left.\mathbb{E}\left[\left.\frac{s_{j_1 j_2} + s_{j_2 j_1}}{2}\right| X_{j_1}, U_{j_1}, X_{j_2}, U_{j_2}\right]'\right). \tag{15}$$

Variance Estimation (continued)

Calculations analogous to those use in variance analyses for U-statistics yield

$$\mathbb{V}\left(V_{1N}\right) = \frac{4\Sigma_1}{N} \tag{16}$$

$$\mathbb{V}\left(V_{2N}\right) = \frac{2}{N\left(N-1\right)}\left(\Sigma_2 - 2\Sigma_1\right) \tag{17}$$

$$\mathbb{V}\left(T_N\right) = \frac{2}{N\left(N-1\right)}\Sigma_3, \tag{18}$$

such that, defining the notation $\Omega \overset{def}{\equiv} \mathbb{V}\left(\sqrt{N}S_N\right)$, from (13), (16), (17) and (18):

$$\Omega = 4\Sigma_1 + \frac{2}{N-1}\left(\Sigma_2 + \Sigma_3 - 2\Sigma_1\right). \tag{19}$$

Variance Estimation (continued)

Consistent with the form of the limit distribution given in (12), the variances of V_{2N} and T_N are of smaller order.

Although the contribution of the $\frac{2}{N-1}(\Sigma_2 + \Sigma_3 - 2\Sigma_1)$ term to the variance of $\sqrt{N}S_N$ is asymptotically negligible, its contribution for finite N need not be.

Using a variance estimator which includes estimates of *both* the $4\Sigma_1$ and $\frac{2}{N-1}(\Sigma_2 + \Sigma_3 - 2\Sigma_1)$ variance terms may therefore result in tests with better size and power properties.

Variance Estimation: Analog Estimation

An analog estimate of Σ_1, the leading variance term, is

$$\hat{\Sigma}_1 = \binom{N}{3}^{-1} \sum_{i=1}^{N-2} \sum_{j=i+1}^{N-1} \sum_{k=j+1}^{N} \frac{1}{3} \left\{ \left(\frac{\hat{s}_{ij} + \hat{s}_{ji}}{2} \right) \left(\frac{\hat{s}_{ik} + \hat{s}_{ki}}{2} \right)' \right.$$

$$\left. \left(\frac{\hat{s}_{ij} + \hat{s}_{ji}}{2} \right) \left(\frac{\hat{s}_{jk} + \hat{s}_{kj}}{2} \right)' + \left(\frac{\hat{s}_{ik} + \hat{s}_{ki}}{2} \right) \left(\frac{\hat{s}_{jk} + \hat{s}_{kj}}{2} \right)' \right\}, \quad (20)$$

with $\hat{s}_{ij} \overset{def}{\equiv} s\left(Z_{ij}, \hat{\theta}\right)$.

Variance Estimation: Analog Estimation (continued)

Equation (20) is a summation over all $\binom{N}{3} = \frac{1}{6}N\left(N-1\right)\left(N-3\right)$ triads in the dataset.

Each triad ijk can be further divided into three pairs of dyads, $\{ij, ik\}$, $\{ij, jk\}$ and $\{ik, jk\}$, with each such pair sharing exactly one agent in common.

Equation (20) corresponds to the sample covariance of $\left(\widehat{s}_{ij} + \widehat{s}_{ji}\right)/2$ and $\left(\widehat{s}_{ik} + \widehat{s}_{ki}\right)/2$ across these $3\binom{N}{3}$ pairs of dyads.

Variance Estimation: Analog Estimation (continued)

To construct an estimate of $\mathbb{V}\left(\sqrt{N}S_N\right)$ separate estimates of Σ_2 and Σ_3 are not required, only their sum is needed. Using an ANOVA decomposition we can express this sum as

$$\begin{aligned}\Sigma_2+\Sigma_3 =&\mathbb{V}\left(\mathbb{E}\left[\left.\frac{s_{12}+s_{21}}{2}\right| X_1,U_1,X_2,U_2\right]\right)\\ &+\mathbb{E}\left[\mathbb{V}\left(\left.\frac{s_{12}+s_{21}}{2}\right| X_1,U_1,X_2,U_2\right)\right]\\ =&\mathbb{V}\left(\frac{s_{12}+s_{21}}{2}\right).\end{aligned}$$

This suggests the analog estimate

$$\widehat{\Sigma_2+\Sigma_3}=\binom{N}{2}^{-1}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\left(\frac{\widehat{s}_{ij}+\widehat{s}_{ji}}{2}\right)\left(\frac{\widehat{s}_{ij}+\widehat{s}_{ji}}{2}\right)'. \qquad (21)$$

Variance Estimation: Analog Estimation (continued)

From (19), (20) and (21) we get the variance estimate

$$\widehat{\mathbb{V}}\left(\sqrt{N}\left(\widehat{\theta}-\theta_0\right)\right)=\left(\widehat{\Gamma}'\widehat{\Omega}^{-1}\widehat{\Gamma}\right)^{-1} \tag{22}$$

where

$$\widehat{\Gamma} = H_N\left(\widehat{\theta}\right) \tag{23}$$

$$\widehat{\Omega} = 4\widehat{\Sigma}_1+\frac{2}{N-1}\left(\Sigma\widehat{{}_2+\Sigma}_3-2\widehat{\Sigma}_1\right). \tag{24}$$

Bootstrap

Rewriting our dyadic regression coefficient estimate in pseudo-U-Process form yields

$$\widehat{\theta} = \arg\max_{\theta \in \Theta} \frac{2}{N\left(N-1\right)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left\{ \frac{l_{ij}\left(\theta\right) + l_{ji}\left(\theta\right)}{2} \right\}.$$

Next let $\left\{V_i^b\right\}_{i=1}^{N}$ be a sequence of i.i.d. mean one random weights independent of the data.

One such sequence is drawn for each of $b = 1, \ldots, B$ bootstrap replications.

Bootstrap (continued)

In the b^{th} such replication we compute

$$\widehat{\theta}_b = \arg\max_{\theta\in\Theta} \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} V_i^b V_j^b \left\{ \frac{l_{ij}(\theta) + l_{ji}(\theta)}{2} \right\}.$$

The bootstrap distribution $\left\{\widehat{\theta}_b\right\}_{b=1}^{B}$ can then be used to approximate the sampling distribution of $\widehat{\theta}$.

Letting V_i^b be an exponential random variable with rate parameter 1 results in a (pseudo-) Bayesian bootstrap (Janssen, 1994).

If we let V_i^b equal the number of times agent i is sampled from the set $\{1, \ldots, N\}$ across N draws with replacement we get the proposal of Davezies et al (2019).