The Missing Transfers: Estimating Mis-reporting in Dyadic Data

Margherita Comola
Paris School of Economics

Marcel Fafchamps
Stanford University

Idea

- We have data on link $\tau_{ij} = \{0, 1\}$ between i and j from both i and j
  - Example: did i make transfer to j
- Data is discordant: sometimes i reports, sometimes j reports, sometimes both
  - So we have two measures of the same thing: G_{ij} and R_{ij}
- Typical approach: we let $\tau_{ij} = \max\{G_{ij}, R_{ij}\}$
- We show that this underestimates the number of links
- We also show that this can bias inference and we propose a method to correct this

Discrepancies

- τ is true transfer
- Discrepancies between reports on τ made by giver and receiver
- Let $G = \{0, 1\}$ be report on τ made by giver
- Let $R = \{0, 1\}$ be report on τ made by receiver
- We only observe R and G

Under-reporting

- *Assume discrepancies are due to under-reporting only, i.e., if either i or j report τ, a transfer took place*
- Given this assumption, the data generation process is:

$$
\begin{aligned}
\Pr(G=1,R=0) &= \Pr(\tau=1,G=1,R=0) \\
&= \Pr(\tau=1)*\Pr(G=1|\tau=1)*\Pr(R=0|G=1,\tau=1) \\
\Pr(G=0,R=1) &= \Pr(\tau=1,G=0,R=1) \\
&= \Pr(\tau=1)*\Pr(G=0|\tau=1)*\Pr(R=1|G=0,\tau=1) \\
\Pr(G=1,R=1) &= \Pr(\tau=1,G=1,R=1) \\
&= \Pr(\tau=1)*\Pr(G=1|\tau=1)*\Pr(R=1|G=1,\tau=1) \\
\Pr(G=0,R=0) &= 1-\Pr(G=1,R=0)-\Pr(G=0,R=1) \\
&\quad -\Pr(G=1,R=1) \qquad (1)
\end{aligned}
$$

Under-reporting

- Assume under-reporting by i is (conditionally) independent of under-reporting by j, $\Pr(R|G,\tau) = \Pr(R|\tau)$.
- Reasonable if under-reporting results from reporting mistakes and omissions.
- We get:

$$\begin{aligned}
\Pr(G = 1, R = 0) &= \Pr(\tau = 1) * \Pr(G = 1|\tau = 1) * \Pr(R = 0|\tau = 1) \\
\Pr(G = 0, R = 1) &= \Pr(\tau = 1) * \Pr(G = 0|\tau = 1) * \Pr(R = 1|\tau = 1) \\
\Pr(G = 1, R = 1) &= \Pr(\tau = 1) * \Pr(G = 1|\tau = 1) * \Pr(R = 1|\tau = 1) \\
\Pr(G = 0, R = 0) &= 1 - \Pr(G = 1, R = 0) - \Pr(G = 0, R = 1) \\
&\quad - \Pr(G = 1, R = 1)
\end{aligned}$$

- 3 probabilities: $P(\tau = 1)$, $P(G = 1|\tau = 1)$ and $P(R = 1|\tau = 1)$.

Estimating mis-reporting

Here is an example using real data on transfers in one Tanzanian village:

$$
\begin{aligned}
\Pr(G=1, R=0) &= \Pr(\tau=1) * \Pr(G=1|\tau=1) * \Pr(R=0|\tau=1) \\
&= 0.0548 \\
\Pr(G=0, R=1) &= \Pr(\tau=1) * \Pr(G=0|\tau=1) * \Pr(R=1|\tau=1) \\
&= 0.0343 \\
\Pr(G=1, R=1) &= \Pr(\tau=1) * \Pr(G=1|\tau=1) * \Pr(R=1|\tau=1) \\
&= 0.0335
\end{aligned}
$$

Estimating mis-reporting

Straightforward algebra yields:

Table 4. MM estimates of under-reporting

in data: declared by i	0.09
in data: declared by j	0.07
in data: declared by i or j $\left(\tau_{ij}^{max}\right)$	0.12
in data: declared by i and j $\left(\tau_{ij}^{min}\right)$	0.03
$\Pr(\tau_{ij} = 1)$	0.18
$\Pr(G = 1 \vert \tau = 1)$	0.49
$\Pr(R = 1 \vert \tau = 1)$	0.38

Does it affect inference?

Imagine we want to estimate a model of the form:

$$\Pr(\tau_{ij} = 1) = \lambda(\beta_\tau X_\tau^{ij}) \tag{2}$$

- X_τ^{ij} is a vector of controls for dyad ij
- β_τ is a coefficient vector of interest
- λ is the logit function.

Does it affect inference?

- We now assume that the three probabilities can be represented by three distinct logit functions:

$$\Pr(\tau = 1) = \lambda(\beta_\tau X_\tau) \quad (3)$$

$$\Pr(G = 1|\tau = 1) = \lambda_G(\beta_G X_G) \quad (4)$$

$$\Pr(R = 1|\tau = 1) = \lambda_R(\beta_R X_R) \quad (5)$$

- The main equation of interest is $\lambda(\beta_\tau X_\tau)$

Simulation analysis

Data generating process of the form

$$\Pr(\tau_{ij} = 1) = \lambda(\beta_{\tau 0} + \beta_{\tau 1} x_i + \beta_{\tau 2} x_j + \beta_{\tau 3} d_{ij} + \varepsilon_{\tau ij}) \tag{6}$$

- x_i and x_j are two uniformly distributed individual attributes (for instance wealth),
- d_{ij} is a uniformly distributed relational attribute (for instance geographic distance)

Simulation analysis

- Scenario 1: mis-reporting is purely random, *i.e.*,
  $\Pr(G_{ij} = 1) = \lambda(\beta_{G0} + \varepsilon_{Gij})$ and $\Pr(R_{ij} = 1) = \lambda(\beta_{R0} + \varepsilon_{Rij})$ with ε_{Gij}, $\varepsilon_{Rij} \backsim N(0, 1)$ and $E[\varepsilon_{Gij}\, \varepsilon_{Rij}] = 0$.
- Scenario 2: mis-reporting depends on individual attributes, *i.e.*,
  $\Pr(G_{ij} = 1) = \lambda(\beta_{G0} + \beta_{G1} x_i + \varepsilon_{Gij})$ and
  $\Pr(R_{ij} = 1) = \lambda(\beta_{R0} + \beta_{R2} x_j + \varepsilon_{Rij})$.
  respondents with high wealth more likely to report transfers
- Scenario 3: mis-reporting depends on relational attribute, *i.e.*,
  $\Pr(G_{ij} = 1) = \lambda(\beta_{G0} + \beta_{G3} d_{ij} + \varepsilon_{Gij})$ and
  $\Pr(R_{ij} = 1) = \lambda(\beta_{R0} + \beta_{R3} d_{ij} + \varepsilon_{Rij})$.
  transfers to proximate households are easier to recall.
- Scenario 4: both 2 and 3 i.e.
  $\Pr(G_{ij} = 1) = \lambda(\beta_{G0} + \beta_{G1} x_i + \beta_{G3} d_{ij} + \varepsilon_{Gij})$ and
  $\Pr(R_{ij} = 1) = \lambda(\beta_{R0} + \beta_{R2} x_j + \beta_{R3} d_{ij} + \varepsilon_{Rij})$.

Table 1. Simulation results

	(1)	(2)	(3)	(4)	(5)
	true model τ_{ij}	our estimator intercept only	our estimator with covariates	standard logit τ_{ij}^{max}	standard logit τ_{ij}^{min}
	Scenario 1:				
$\beta_{\tau 1}$	1.73	1.75	1.76	1.48	1.13
$\beta_{\tau 2}$	1.73	1.75	1.75	1.48	1.14
$\beta_{\tau 3}$	-1.73	-1.74	-1.75	-1.45	-1.09
	Scenario 2:				
$\beta_{\tau 1}$	1.73	2.3	1.72	1.92	1.83
$\beta_{\tau 2}$	1.74	2.12	1.72	1.77	2.21
$\beta_{\tau 3}$	-1.74	-1.83	-1.73	-1.51	-0.97
	Scenario 3:				
$\beta_{\tau 1}$	1.73	1.72	1.76	1.48	1.18
$\beta_{\tau 2}$	1.73	1.73	1.76	1.48	1.19
$\beta_{\tau 3}$	-1.74	-1	-1.75	-0.8	0.52

Table 1. Simulation results

	(1)	(2)	(3)	(4)	(5)
	true model τ_{ij}	our estimator intercept only	our estimator with covariates	standard logit τ_{ij}^{max}	standard logit τ_{ij}^{min}
	Scenario 2:				
$\beta_{\tau 1}$	1.73	2.3	1.72	1.92	1.83
$\beta_{\tau 2}$	1.74	2.12	1.72	1.77	2.21
$\beta_{\tau 3}$	-1.74	-1.83	-1.73	-1.51	-0.97
	Scenario 3:				
$\beta_{\tau 1}$	1.73	1.72	1.76	1.48	1.18
$\beta_{\tau 2}$	1.73	1.73	1.76	1.48	1.19
$\beta_{\tau 3}$	-1.74	-1	-1.75	-0.8	0.52
	Scenario 4:				
$\beta_{\tau 1}$	1.74	2.26	1.73	1.92	1.85
$\beta_{\tau 2}$	1.73	2.07	1.72	1.75	2.23
$\beta_{\tau 3}$	-1.73	-1.04	-1.72	-0.86	0.64

Table 2. Descriptive statistics (N=14042)

variable	dummy	mean	min	max	sd
τ_{ij}^{i}	yes	0.09			
τ_{ij}^{j}	yes	0.07			
τ_{ij}^{max}	yes	0.12			
τ_{ij}^{min}	yes	0.03			
wealth (i and j)	no	4.01	0	23.09	3.75
$wealth_i * wealth_j$	no	15.98	0	378.59	24.89
same education	yes	0.65			
same religion	yes	0.35			
blood link	yes	0.02			
neighbors	yes	0.40			
declared friends (i and j)	no	5.29	0	19	3.06

Table 3. Main results

	(1) τ_{ij}^{max}	(2) τ_{ij}^{min}	(3) $\Pr(\tau = 1)$	(4) $\Pr(G\|\tau)$	(5) $Pr(R\|\tau)$
$wealth_i$	0.062***	0.057***	0.045	-0.053*	0.055
	(0.021)	(0.019)	(0.051)	(0.028)	(0.079)
$wealth_j$	0.096***	0.051**	0.062	0.084	-0.058
	(0.030)	(0.026)	(0.041)	(0.060)	(0.045)
$wealth_i * wealth_j$	0.004	0.002	0.013**	-0.001	-0.003
	(0.003)	(0.003)	(0.006)	(0.003)	(0.006)
same education	-0.012	0.060	-0.052	0.173	-0.143
	(0.118)	(0.177)	(0.306)	(0.359)	(0.282)
same religion	0.434***	0.464***	0.367	0.212	0.216
	(0.099)	(0.145)	(0.282)	(0.296)	(0.273)
blood link	2.718***	2.627***	2.631***	1.003**	1.321***
	(0.252)	(0.246)	(0.601)	(0.459)	(0.354)
neighbors	1.063***	1.503***	0.683*	0.891***	0.674**
	(0.111)	(0.157)	(0.350)	(0.283)	(0.264)
declared friends$_i$				0.086***	
				(0.026)	

Estimating mis-reporting

Table 5. Estimates of under-reporting with covariates

	gifts
average fitted $\Pr(\tau_{ij} = 1)$	0.20
average fitted $\Pr(G = 1 \vert \tau = 1)$	0.38
average fitted $\Pr(R = 1 \vert \tau = 1)$	0.30

Robustness

- Robustness to assumption that errors uncorrelated between i and j?
- We calculate estimates of $\Pr(\tau_{ij} = 1)$ for different possible values of the correlation in under-reporting between i and j.
- Extremely high or low correlation values are irreconciliable with the data:
  - high positive correlation would imply little discordance, which is not what the data show;
  - high negative correlation would imply even more discordance than what is in the data.
- => There is a range of intermediate correlation values which are potentially consistent with the data.
- => Feasible estimates of $\Pr(\tau_{ij} = 1)$ vary between 13% and 27%.

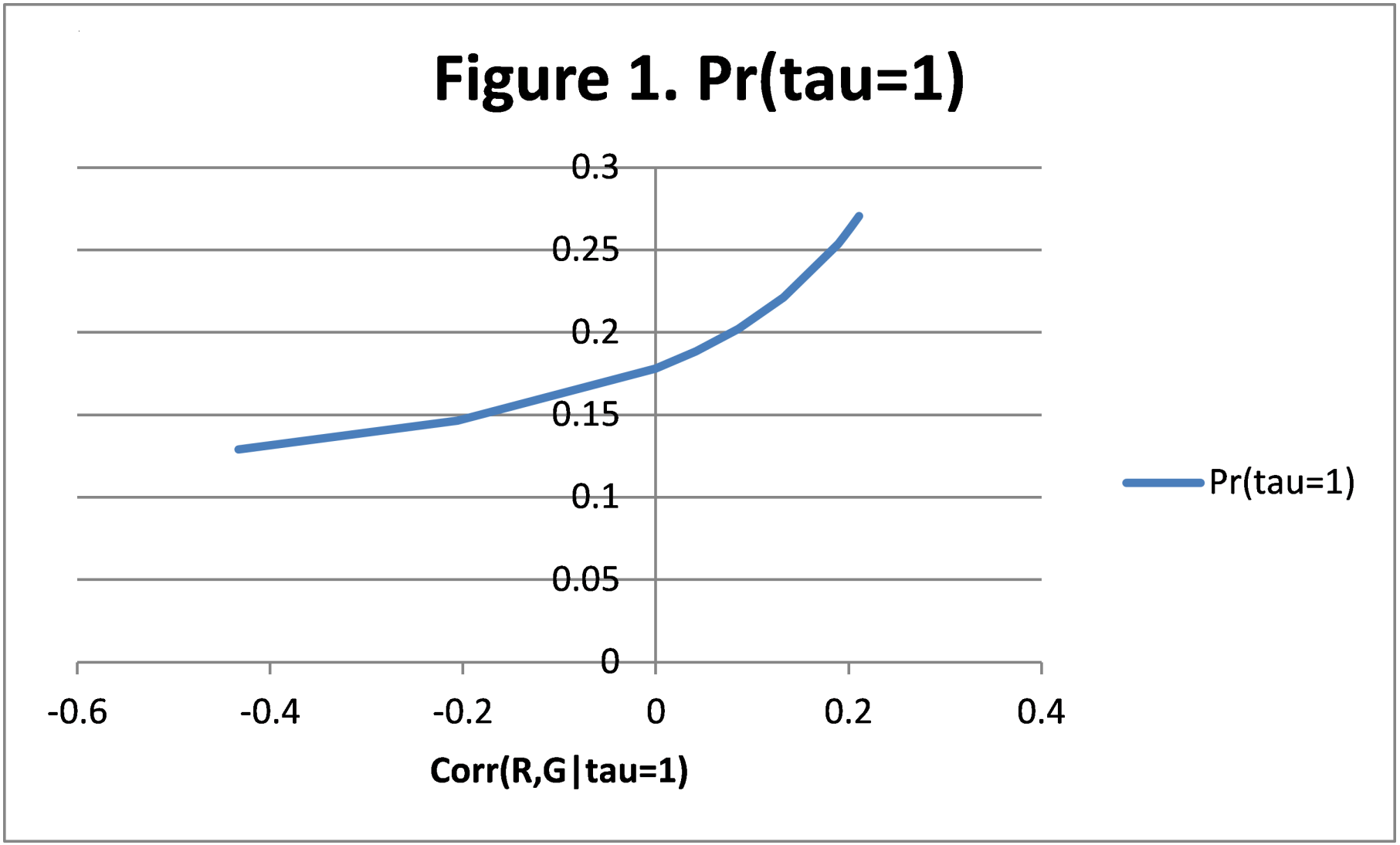
Figure 1. Pr(tau=1)
0.3
0.25
0.2
0.15
0.1
0.05
0
-0.6
-0.4
-0.2
0
0.2
0.4
Corr(R,G|tau=1)
Pr(tau=1)

Another illustration: to correct treatment effects and LATE estimates

- This example is taken from Fafchamps and Quinn (2015).
- Treatment aims to create new links.
- Link measure is remembering having talked to someone.
- Outcome is diffusion of business practice.

Effect of treatment on link formation

- Here network is undirected, but when i remembers talking to j, j does not always remember talking to i.
- Let $\tau = 1$ if i and j spoke to each other and 0 otherwise.
- Let $\lambda = \Pr(\tau = 1)$.
- Let $i = 1$ be shorthand for i reported talking to j.
- Let $\theta = \Pr(i = 1 | \tau = 1)$; $1 - \theta$ is under-reporting.
- We observe:
  - $P_1 \equiv \Pr(i = 1, j = 0) = \Pr(j = 1, i = 0)$
  - $P_2 \equiv \Pr(i = 1, j = 1)$

Effect of treatment on link formation

- $=>$ $P_1 = \lambda\theta(1-\theta)$ and $P_2 = \lambda\theta^2$.
- $=>$ $\theta = \frac{P_1}{P_1+P_2}$ and $\lambda = \frac{(P_1+P_2)^2}{P_2}$
- In the data of that paper, $(1-\theta) = 68.2\%$
- Likelihood of taking rises
  - from uncorrected value of 17.4%
  - to corrected value of 54.6%
- $=>$ Effect of treatment on likelihood of talking is much larger than estimated using individual reports.

LATE of treatment on outcome

- The ITT effect of treatment on outcome is 6.6% for diffusion of VAT and 4.4% for diffusion of bank account.
- The LATE effect of treatment is ITT/probability of talking.
- Without correction, LATE estimates are very large, and hard to believe.
- With correction, LATE estimates are 12% and 8%, which is more reasonable.

Note: standard errors when estimating dyadic regressions

- Dyadic observations are not independent.
- Standard errors must be adjusted, otherwise inference will be inconsistent.
- Apply the formula developed by Fafchamps and Gubert (2007), using the scores in lieu of X in formula below:

$$AVar(\widehat{\beta}) = \frac{1}{N-K}(X'X)^{-1}\left(\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N}\sum_{l=1}^{N}\frac{m_{ijkl}}{2N}X_{ij}u_{ij}u'_{kl}X_{kl}\right)(X'X)^{-1}$$

- There is an ado file on my website called ngreg that does this for you.