Using Big Data in Economics: Examples from Intergenerational Research

June 27, CEMFI

Jan Stuhler
Universidad Carlos III de Madrid

Introduction

- Income inequality has been increasing in many developed countries
- Substantial interest in studying the extent, changes and determinants of income inequality
- One important dimension is the extent of intergenerational mobility or social mobility in a society

Cross-sectional vs. intergenerational inequality

From Solon (1999):

> *Imagine two societies, society A and society B. The distribution of earnings [and] the degree of cross-sectional inequality is the same in both societies. At first glance, the two societies appear to be equally unequal. But now suppose that, in society A, one's relative position in the earnings distribution is exactly inherited from one's parents. If your parents were in the 90th percentile of earnings in their generation, it is certain that you place in the 90th percentile in your own generation. [...] In contrast, in society B, one's relative position in the earnings distribution is completely independent of the position of one's parents. [...] Unlike society A, society B displays complete intergenerational mobility. Although societies A and B have the same measured inequality within a generation, the two societies are tremendously different in the character of their inequality.*

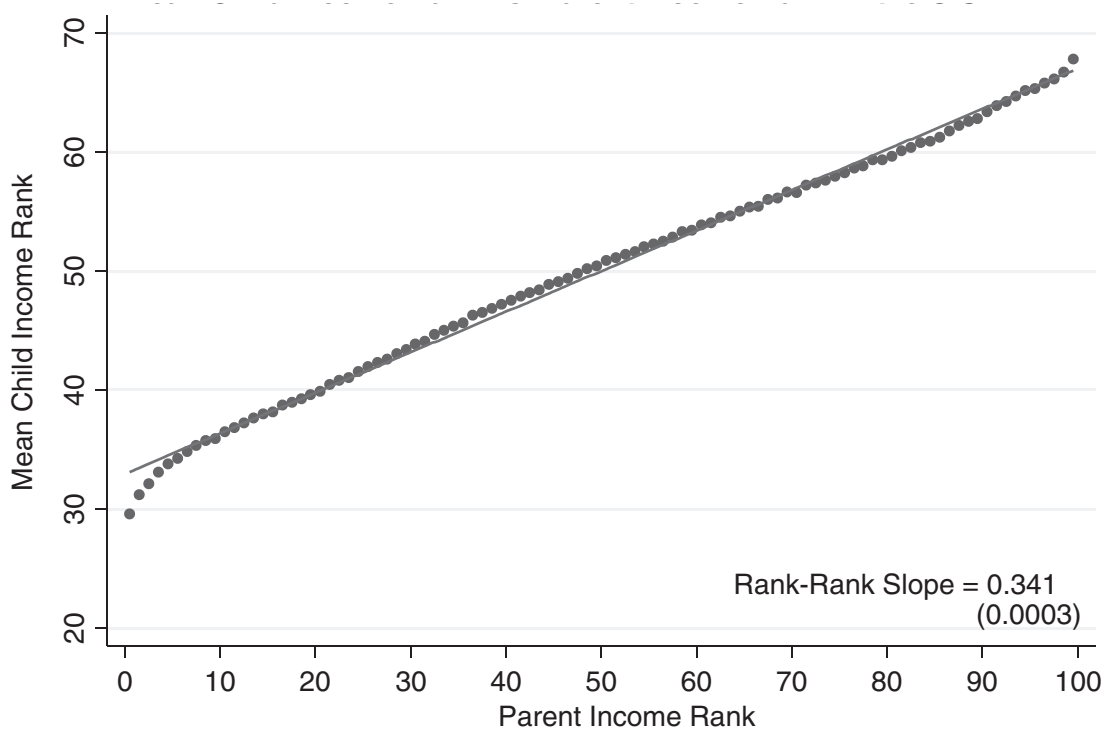
Measuring intergenerational dependence

- Intergenerational mobility is often summarized in the *intergenerational elasticity* of income (IGE)
- IGE defined as the slope coefficient in the regression of log incomes of offspring y^* on log income of parents x^*,

$$y^* = \beta x^* + \varepsilon$$

- Ideally want to measure long-run or lifetime income, but applications typically based on short snapshots of annual income. Problem: estimated β very sensitive to the age at which incomes are measured (Nybom and Stuhler, 2016).
- More recent research often considers income ranks instead of log incomes ($\rightarrow$ rank-rank regression or rank correlation).

Figure: Mean Child Income Rank vs Parent Income Rank in the US



Source: Chetty, Hendren, Kline and Saez (2014)

Content

Today's topics:

1. Chetty, Hendren, Kline and Saez (2014), "Where is the Land of Opportunity? The Geography of Intergenerational Mobility in the United States", Quarterly Journal of Economics
2. Güell, Rodriguez and Telmer (2015), "The Informational Content of Surnames, the Evolution of Intergenerational Mobility and Assortative Mating", Review of Economic Studies
3. Collado, Ortuno-Ortin, and Stuhler, "Kinship Correlations and Intergenerational Mobility". Working Paper

Paper #1:

- Where is the Land of Opportunity? The Geography of Intergenerational Mobility in the United States
  Raj Chetty, Nathan Hendren, Patrick Kline and Emmanuel Saez, *Quarterly Journal of Economics (2014)*

Intergenerational Mobility in the US: Chetty et al (2014)

Chetty, Hendren, Kline and Saez (2014):

- Use tax data from the US Internal Revenue Service (IRS)
- Match tax records of parents with the tax records of their children to study intergenerational mobility in the Unites States
- Core sample of nearly 10 million children born between 1980 and 1982 (14- to 16-year-olds), tracked until age 30.

Income definitions:

- Parent's income: average total family income over 5 years, 1996-2000
- Children's income: measured over two years, 2011-2012

From Science Mag:

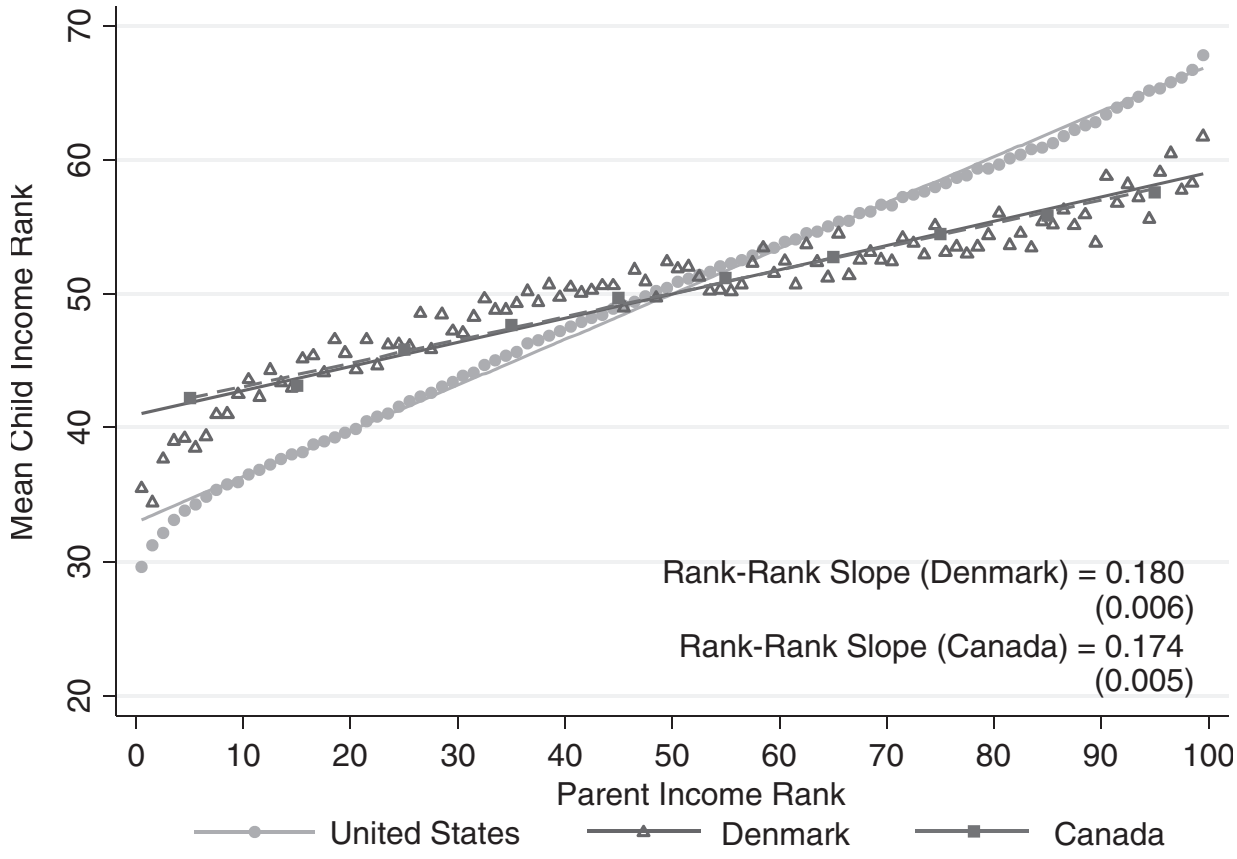
MACARTHUR FOUNDATION

How Two Economists Got Direct Access to IRS Tax Records

By Jeffrey Mervis | May. 22, 2014 , 2:00 PM

Raj Chetty of Harvard University and Emmanuel Saez of University of California (UC), Berkeley, created a **big media splash** last summer with a study showing that social mobility—the income status of adult children relative to their parents—correlates with where the children grew up. The study, based on an analysis of millions of U.S. tax records that had been largely off-limits to researchers, has fed the public perception that the American dream of equal opportunity for all may be fading. It also bolstered the reputations of the two young superstars—each has received the top prize for economists under 40 and a MacArthur "genius" award. And it has left their colleagues wondering how they pulled off such a feat.

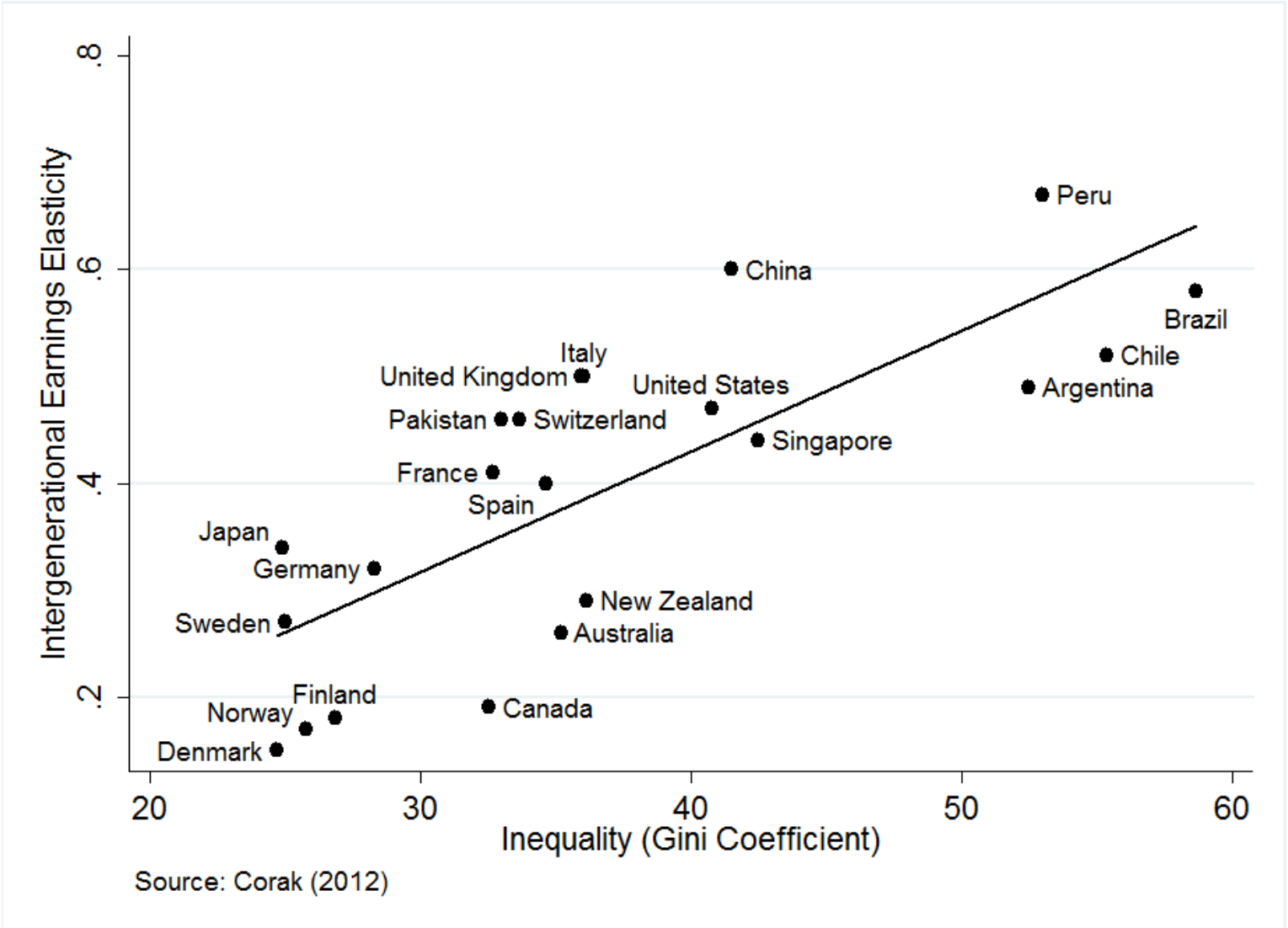
Figure: Rank-Rank Slope, Cross-Country Comparisons

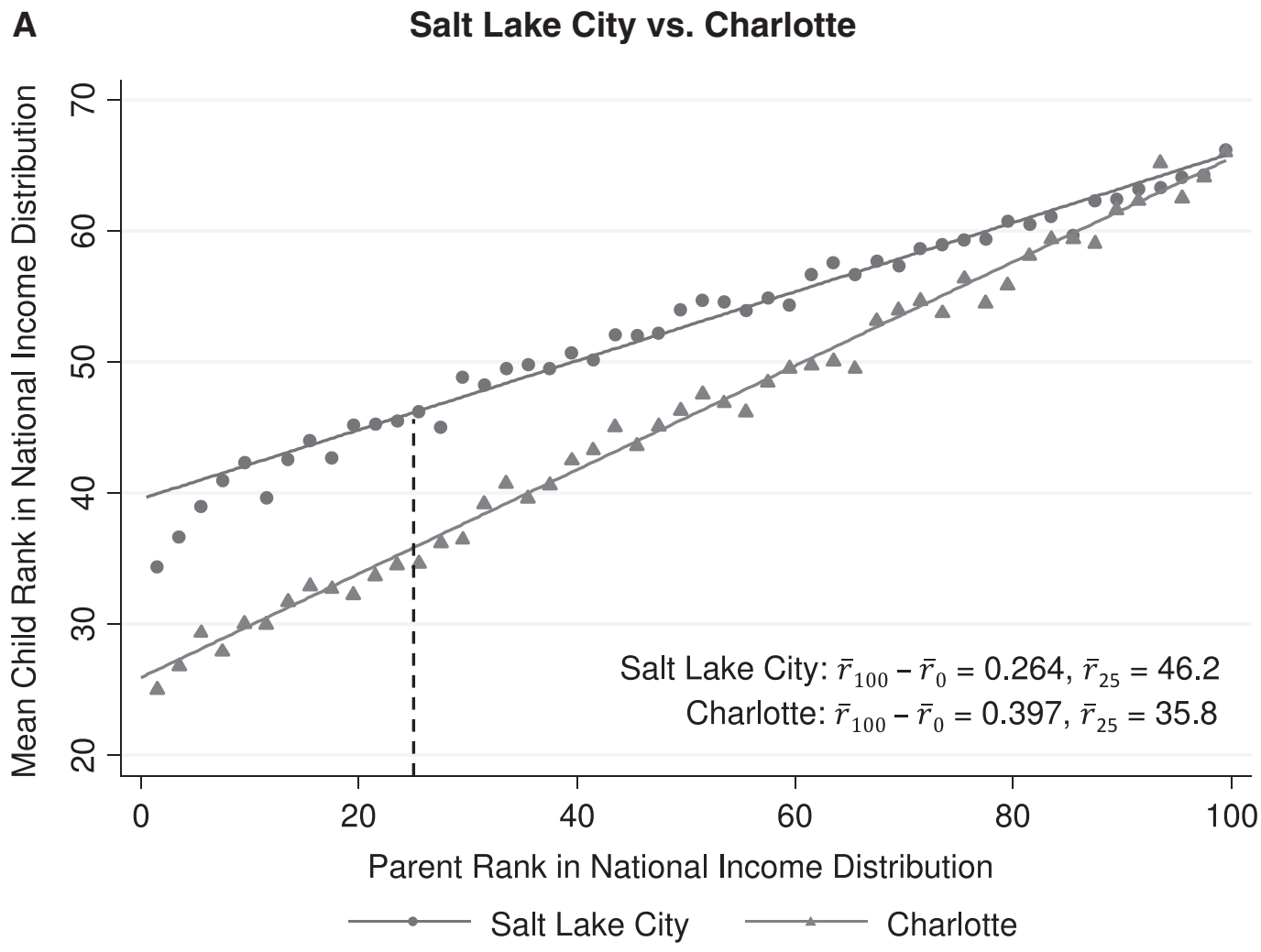


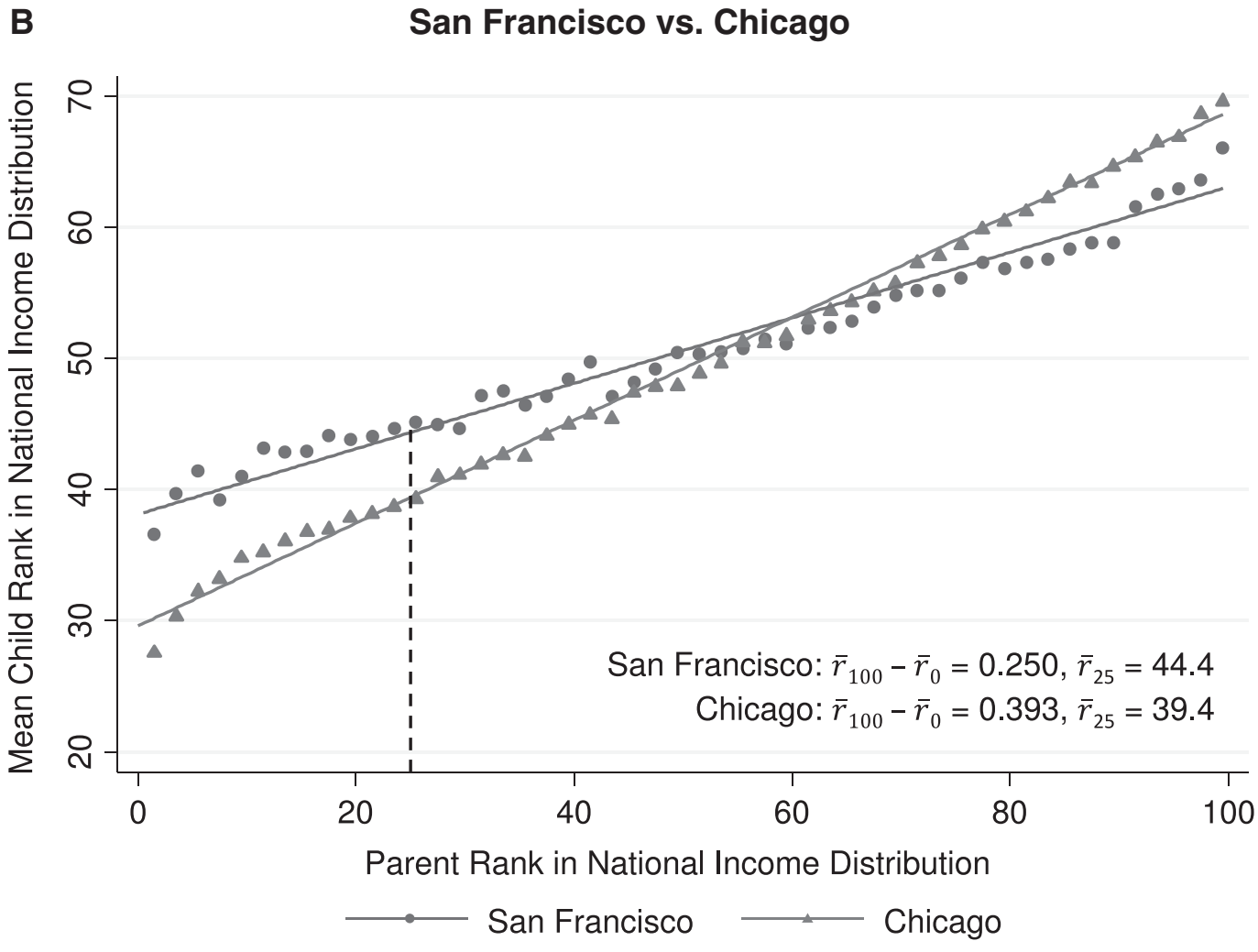
Source: Chetty, Hendren, Kline and Saez (2014)

The Great Gatsby curve

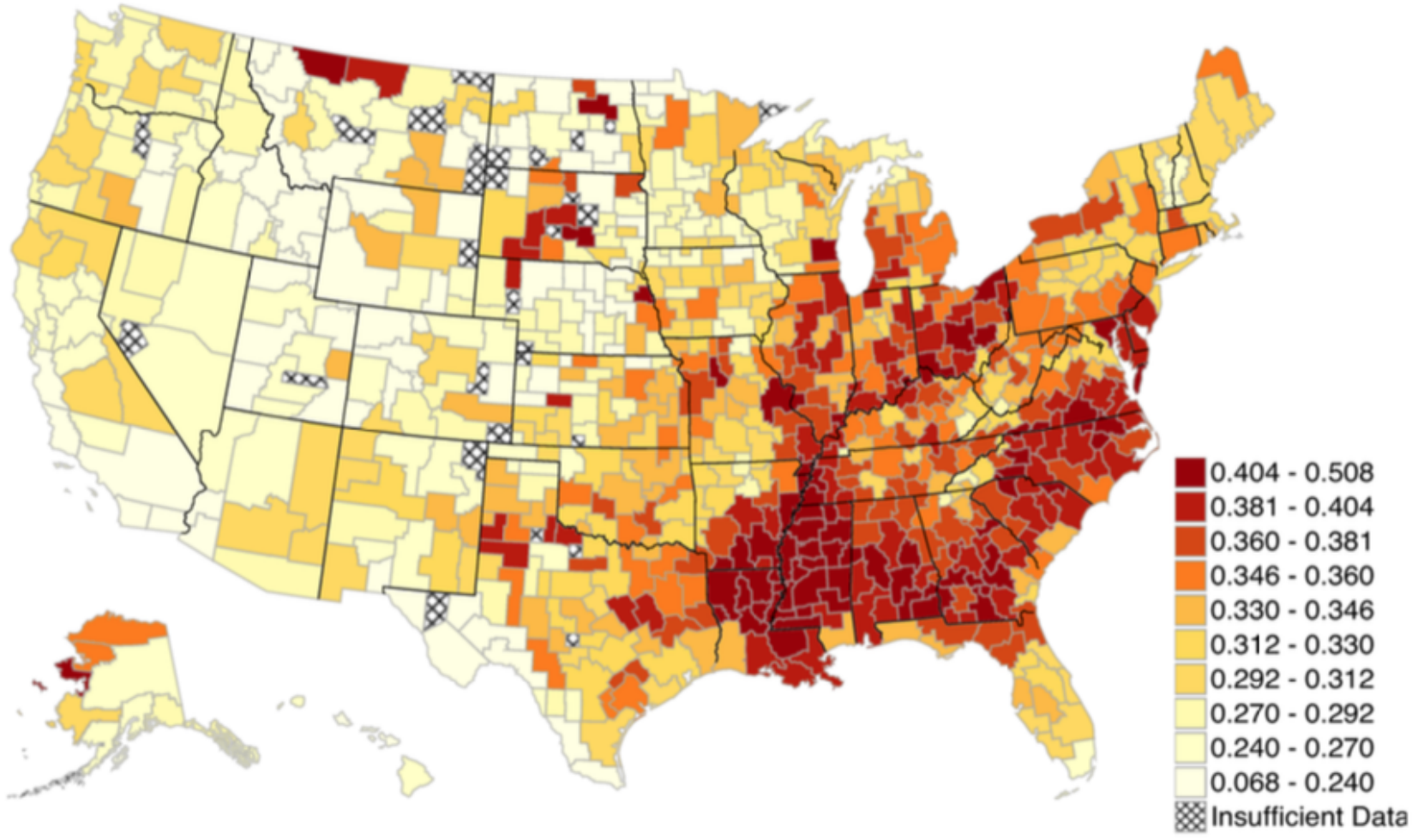
Figure: The Great Gatsby curve




A
Salt Lake City vs. Charlotte
Mean Child Rank in National Income Distribution
20
30
40
50
60
70
0
20
40
60
80
100
Parent Rank in National Income Distribution
Salt Lake City: $\bar{r}_{100} - \bar{r}_0 = 0.264$, $\bar{r}_{25} = 46.2$
Charlotte: $\bar{r}_{100} - \bar{r}_0 = 0.397$, $\bar{r}_{25} = 35.8$
Salt Lake City
Charlotte


B
San Francisco vs. Chicago
Mean Child Rank in National Income Distribution
20
30
40
50
60
70
0
20
40
60
80
100
Parent Rank in National Income Distribution
San Francisco: $\bar{r}_{100} - \bar{r}_0 = 0.250$, $\bar{r}_{25} = 44.4$
Chicago: $\bar{r}_{100} - \bar{r}_0 = 0.393$, $\bar{r}_{25} = 39.4$
San Francisco
Chicago

B. Relative Mobility: Rank-Rank Slopes $(\bar{r}_{100} - \bar{r}_0)/100$ by CZ



Corr. with baseline $\bar{r}_{25}$ = -0.68 (unweighted), -0.61 (pop-weighted)

Intergenerational Mobility in the 50 Largest Commuting Zones

(1) Upward mobility rank	(2) CZ name	(3) Population	(4) Absolute upward mobility	(5) P(child in Q5 | parent in Q1)	(6) Pct. above poverty line	(7) Relative mobility rank-rank slope
1	Salt Lake City, UT	1,426,729	46.2	10.8	77.3	0.264
2	Pittsburgh, PA	2,561,364	45.2	9.5	74.9	0.359
3	San Jose, CA	2,393,183	44.7	12.9	73.5	0.235
4	Boston, MA	4,974,945	44.6	10.5	73.7	0.322
5	San Francisco, CA	4,642,561	44.4	12.2	72.5	0.250
6	San Diego, CA	2,813,833	44.3	10.4	74.3	0.237
7	Manchester, NH	1,193,391	44.2	10.0	75.0	0.296
8	Minneapolis, MN	2,904,389	44.2	8.5	75.2	0.338
9	Newark, NJ	5,822,286	44.1	10.2	73.7	0.350
10	New York, NY	11,781,395	43.8	10.5	72.2	0.330
11	Los Angeles, CA	16,393,360	43.4	9.6	73.8	0.231
12	Providence, RI	1,582,997	43.4	8.2	73.6	0.333
13	Washington DC	4,632,415	43.2	11.0	72.2	0.330
14	Seattle, WA	3,775,744	43.2	10.9	72.0	0.273
15	Houston, TX	4,504,013	42.8	9.3	74.7	0.325
16	Sacramento, CA	2,570,609	42.7	9.7	71.3	0.257
17	Bridgeport, CT	3,405,565	42.4	7.9	72.4	0.359
18	Fort Worth, TX	1,804,370	42.3	9.1	73.6	0.320
19	Denver, CO	2,449,044	42.2	8.7	73.3	0.294
20	Buffalo, NY	2,369,699	42.0	6.7	73.1	0.368
21	Miami, FL	3,955,969	41.5	7.3	76.3	0.267
22	Fresno, CA	1,419,998	41.3	7.5	71.3	0.295

23	Portland, OR	1,842,889	41.3	9.3	70.5	0.277
24	San Antonio, TX	1,724,863	41.1	6.4	74.3	0.320
25	Philadelphia, PA	5,602,247	40.8	7.4	69.6	0.393
26	Austin, TX	1,298,076	40.4	6.9	71.9	0.323
27	Dallas, TX	3,405,666	40.4	7.1	72.6	0.347
28	Phoenix, AZ	3,303,211	40.3	7.5	70.6	0.294
29	Grand Rapids, Michigan	1,286,045	40.1	6.4	71.3	0.378
30	Kansas City, MI	1,762,873	40.1	7.0	70.4	0.365
31	Las Vegas, NV	1,568,418	40.0	8.0	71.1	0.259
32	Chicago, IL	8,183,799	39.4	6.5	70.8	0.393
33	Milwaukee, WI	1,660,659	39.3	4.5	70.3	0.424
34	Tampa, FL	2,395,997	39.1	6.0	71.3	0.335
35	Orlando, FL	1,697,906	39.1	5.8	71.5	0.326
36	Port St. Lucie, FL	1,533,306	39.0	6.2	71.2	0.303
37	Baltimore, MD	2,512,431	38.8	6.4	67.7	0.412
38	St. Louis, MO	2,325,609	38.4	5.1	69.0	0.413
39	Dayton, OH	1,179,009	38.3	4.9	68.2	0.397
40	Cleveland, OH	2,661,167	38.2	5.1	68.7	0.405
41	Nashville, TN	1,246,338	38.2	5.7	67.9	0.357
42	New Orleans, LA	1,381,652	38.2	5.1	69.5	0.397
43	Cincinnati, OH	1,954,800	37.9	5.1	66.4	0.429
44	Columbus, OH	1,663,807	37.7	4.9	67.1	0.406

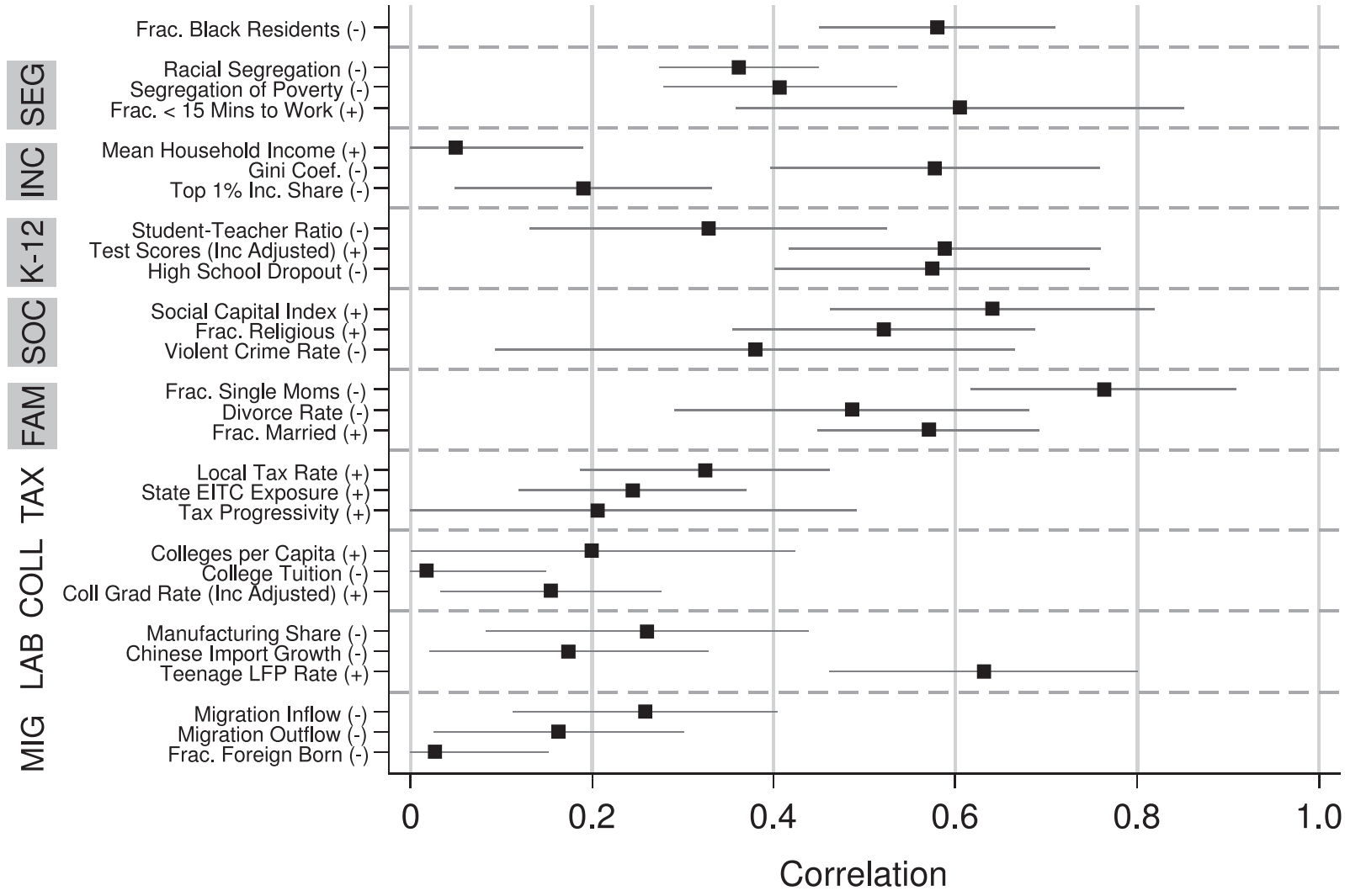
Main Findings

Findings: Large differences in intergenerational mobility, and intergenerational upward mobility, across areas in the US:

1. Upward mobility varies substantially at the regional level.
   - Upward mobility is lowest in the Southeast and highest in the Great Plains
   - The West Coast and Northeast also have high rates of upward mobility, though not as high as the Great Plains.
2. Substantial within-region variation as well.
3. On average, urban areas tend to exhibit lower levels of intergenerational mobility than rural areas

Next step: Does intergenerational mobility correlate with local economic or institutional factors?

Figure: Correlates of Spatial Variation in Upward Mobility



Intergenerational mobility across regions within countries

Many follow-up papers (check out www.equality-of-opportunity.org)

Descriptive:

- Changes in intergenerational mobility over time (Chetty et al, AER P&P)

Causal:

- Exploit intraregional mobility of movers to identify the causal impact of a region on intergenerational mobility:
- Chetty and Hendren (2018a) "The Impacts of Neighborhoods on Intergenerational Mobility I: Childhood Exposure Effects." Quarterly Journal of Economics
- Chetty and Hendren (2018b) "The Impact of Neighborhoods on Intergenerational Mobility II: County-Level Estimates." Quarterly Journal of Economics

Paper #2:

- The Informational Content of Surnames, the Evolution of Intergenerational Mobility and Assortative Mating
  Maia Güell, José Rodríguez Mora, and Christopher Telmer
  The Review of Economic Studies (2015)

Introduction: Names

To study intergenerational mobility we need one of three things:

1. Be in data heaven ($\rightarrow$ previous and the next paper)
2. Construct family lineages from cross sections of censuses, either manually or through automated linking algorithms
   (Long and Ferrie, 2007, 2013; Abramitzky et al. (ABEFP), 2019; Craig, Eriksson, Niemesh, 2019; Buckles et al., 2019; Feigenbaum, 2019)
3. Use the informational content of names
   (Olivetti and Paserman, 2015; Güell, Rodríguez Mora and Telmer, 2015; Clark, 2012; Clark and Cummins, 2014; Barone and Mocetti, 2019)

Authors	*Year*	*Publication*	*Names*	*Method*	*Data*	*Main Application*
Clark	2012	Working Paper	Surnames	Name Frequencies	Repeated cross-section of surname frequencies	Multigenerational mobility in Sweden
Clark	2012	Working Paper	Surnames	Grouping	Repeated cross-section of rare surnames	Multigenerational mobility in England
Collado, Ortuño and Romeu	2012	Reg. Science and Urban Econ.	Surnames	Grouping (by region)	Single cross-section across areas	Intergenerational consumption mobility in Spain
Collado, Ortuño and Romeu	2013	Working Paper	Surnames	Grouping	Repeated cross-section of surname averages	Multigenerational mobility in Spanish provinces
Clark	2014	Princeton University Press	Surnames	Grouping	Repeated cross-section of rare surnames	Inter- and multigenerational mobility in various countries
Clark and Cummins	2014	Economic Journal	Direct and Surnames	Grouping	Repeated cross-section of rare surnames	Multigenerational wealth mobility in England
Güell, Rodríguez and Telmer	2015	Review of Economic Studies	Surnames	R2	Single cross-section	Intergenerational mobility level and trends in Catalonia
Clark and Diaz-Vidal	2015	Working Paper	Surnames	Grouping	Repeated cross-section of surname averages	Multigenerational and assortative mobility in Chile
Olivetti and Paserman	2015	American Economic Review	First names	Two-sample Two-stage IV	Repeated cross-section	Historical mobility trends in United States
Barone and Mocetti	2016	Working Paper	Surnames	Two-sample Two-stage IV	Repeated cross-section of surname averages	Multigenerational mobility in Florence, Italy (1427-2011)
Nye et al	2016	Working Paper	Surnames	Name Frequencies	Repeated cross-section of name frequencies	Intergenerational mobility in Russia
Durante, Labartino and Perotti	2016	Working Paper (R&R AEJ:Policy)	Surnames	Name Frequencies	Single cross-section of surname frequencies	Family connections at Italian universities
Feigenbaum	2018	Economic Journal	Direct, First and Surnames	R2, Grouping		Historical mobility level in Iowa, United States
Güell, Pellizzari, Pica, and Rodríguez	2018	Economic Journal	Surnames	R2	Single cross-section across areas	Cross-regional variation in mobility in Italy
Olivetti, Paserman and Salisbury	2018	Explorations in Economic History	First names	Two-sample Two-stage IV	Repeated cross-section	Multigenerational mobility in United States

Note: The table lists selected intergenerational mobility research that use first or surnames to overcome the lack of direct parent-child links. The year indicates the year of article publication, and does therefore not reflect the time at which the study was created.

Name-based estimators: Applications

Name-based estimators have been instrumental in some of the most active research areas in the literature:

1. Intergenerational mobility in the very long run (i.e., *multi*generational mobility)
2. Intergenerational mobility in countries and historical time periods for which intergenerational panel data with direct parent-child links are not available
3. Variation in intergenerational mobility across regions

The informational content of names

Why are names informative?

- Both surnames and first names are informative about socioeconomic status and intergenerational transmission:
- Children *inherit* both their surname and other factors that influence their socioeconomic status
- Parents *choose* first names for their children. Choices correlate with parental socioeconomic status

The informational content of names

Güell, Rodríguez Mora and Telmer (2015) present a model to show

- (i) that surnames have informational content, and
- (ii) that this informational content of surnames (ICS) increases monotonically in intergenerational persistence.

Idea:

- Both socioeconomic status and surnames transmitted from one generation to the next $\rightarrow$ Surnames will explain some of the variation in socioeconomic status
- Socioeconomic status more strongly transmitted from one generation to the next $\rightarrow$ Surnames will explain larger share of variation in socioeconomic status

The informational content of names

Explain the economic status of individual i with (sur)name j by vector of surname dummy variables, $Surname_j$

$$y_{ij} = \beta'(Surname)_j + \gamma' X_{ij} + \varepsilon_{ij}, \tag{1}$$

where X_{ij} may include region of birth, year of birth, ethnicity.

Then estimate placebo regression: randomly reassign surnames to individuals (while maintaining their marginal distribution),

$$y_{ij} = \beta'(Fake\ surname)_j + \gamma' X_{ij} + \varepsilon_{ij}. \tag{2}$$

Informative content of surnames (ICS) defined as

$$ICS \equiv aR^2 - aR_P^2$$

Table: Informational Content of Names (Güell et al, 2015)

Table 2: ICS. Baseline population.

LHS: years of education	(1)	(2)	(3)	(4)	(5)	(6)
CatalanDegreeSurname2		1.706	1.015	1.707		
		(0.011)	(0.012)	(0.011)		
Surname Dummies			Yes		Yes	
Fake Surnames Dummies				Yes		Yes
Adjusted R^2	0.2652	0.2735	0.2980	0.2735	0.2955	0.2653
Surnames jointly significant*			Yes	No	Yes	No
(p-value)			0.000	0.534	0.000	0.601

Notes: All regressions include age and place of birth dummies. Fake-surnames have the same distribution as Surnames and are allocated randomly. (*) F-test if Surname dummies are jointly significant. Standard errors in parenthesis. Population: Male Spanish citizens living in Catalonia aged 25 and above, with frequency of first surname larger than one. Number of observations: 2,057,134. Number of surnames: 30,610. Source: 2001 Catalan Census (Idescat).

Table: Informational Content of Names (Santavirta and Stuhler, 2019)

	Son's occupational status				Son's schooling			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
First name dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Demographic dummies		Yes	Yes	Yes		Yes	Yes	Yes
Region of birth (county)			Yes				Yes	
Region of birth (parish)				Yes				Yes
AR2	0.058	0.161	0.174	0.254	0.088	0.244	0.267	0.358
Implied ICF	0.058	0.059	0.048	0.026	0.088	0.080	0.064	0.036
Bootstrapped 95% CI	[0.037, 0.080]	[0.041, 0.076]	[0.032, 0.062]	[0.014, 0.038]	[0.062, 0.114]	[0.062, 0.098]	[0.049, 0.079]	[0.026, 0.047]
N	11,505	11,505	11,505	11,505	11,505	11,505	11,505	11,505

Figure: From Güell et al. (2018)

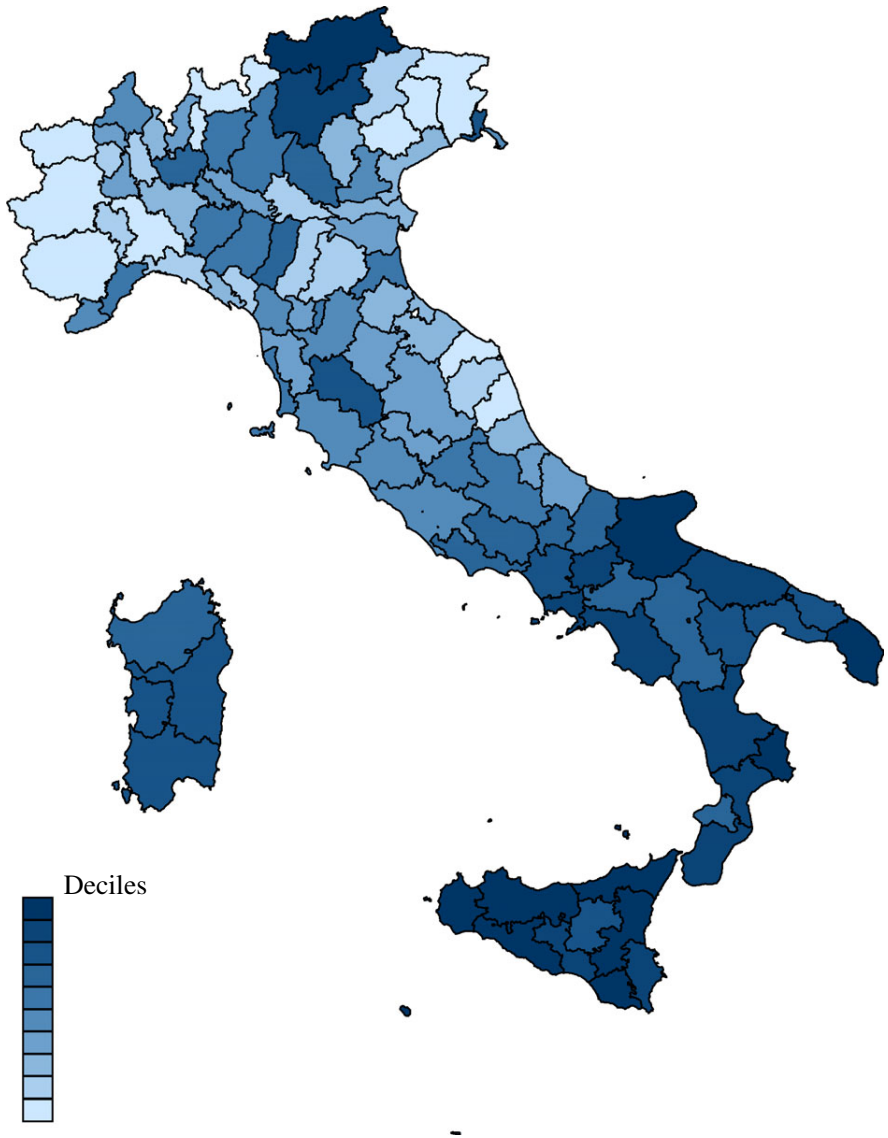


Fig. 2. *Social Mobility (ICS-30) across Italian Provinces*
Notes. Darker blue implies lower mobility. Colour figure can be viewed at wileyonlinelibrary.com

Paper #3:

- Kinship Correlations and Intergenerational Mobility
  Ignacio Ortuño-Ortin, Lola Collado, and Jan Stuhler
  Working Paper (2019)

Motivation

How persistent are socioeconomic inequalities between families?

- How strongly are advantages transmitted from one generation to the next? How similar are siblings or spouses?

What are the causal mechanisms (e.g. nature vs nurture)?

- For example, could genetic transmission explain the observed persistence of inequalities in the very long run?

This paper

Questions:

1. How persistent are socioeconomic inequalities?
2. What are the causal mechanisms (e.g. nature vs nurture)?

What we do:

- Study intergenerational and assortative processes using "horizontal" kins such as siblings, uncle-nephew and distant siblings *in-law*
- Estimate a model of intergenerational transmission that is (more) generalthan previous models

This paper

Questions:

1. How persistent are socioeconomic inequalities?
2. What are the causal mechanisms (e.g. nature vs nurture)?

What we do:

- Study intergenerational and assortative processes using "horizontal" kins such as siblings, uncle-nephew and distant siblings *in-law*
- Estimate a model of intergenerational transmission that is (more) generalthan previous models

Related empirical literature

1. Evidence on intergenerational persistence, e.g. Solon (1999), Black and Devereux (2011), Jäntti and Jenkins (2014), and sibling correlations, e.g. Björklund and Jäntti (2012)
2. Evidence on multigenerational persistence, e.g. Clark and Cummins (2012), Clark (2014), Lindahl et al (2014), Braun and Stuhler (2018), Adermon et al (2018), Barone and Mocetti (2019) + many recent

Inter- and multigenerational persistence

Intergenerational (or parent-child) correlation in (log) income

$$y_{it} = \alpha + \beta y_{it-1} + \varepsilon_t$$

For the US, literature finds $\hat{\beta} \approx 0.4$ (Solon, 1990s) or $\hat{\beta} \approx 0.5$ (e.g. Mazumder, 2016).

More recent literature on **multi**generational persistence:

- High persistence on surname level (e.g. Clark, 2014).
  For example, across six centuries in Florence (Barone and Mocetti, 2019).
- Other studies observe direct family links, but fewer generations
  Lindahl et al (2014), Braun and Stuhler (2018)

Inter- and multigenerational persistence

Contrast between inter- and multigenerational correlations can be rationalized by **latent** transmission model, such as

$$y_t = \delta z_t + u_t$$
$$z_t = \gamma z_{t-1} + v_t$$

where y_t is observed outcome and z_t is latent factor

- $\gamma \approx 0.8$ (Clark, 2014)
- $\gamma \approx 0.6$ (Braun and Stuhler, 2018; Neidhöfer and Stockhausen, 2019)

Data: Swedish registers and Spanish Census

(1) Swedish register data:

- 1/3 of Swedish population born between 1932 and 1967, plus their siblings, parents and children
- family links up to three (four) generations (censoring/selectivity)

(2) Spanish Census from Cantabria, with full name of each person:

- newborns in Spain receive two surnames, with first=father's and second=mother's (first) surname
- child generation born 1956-1976 (71,479 males, 68,830 females), identify relatives via surnames
- only education

▸ Identifying relatives in the Spanish Census

Data: Swedish registers and Spanish Census

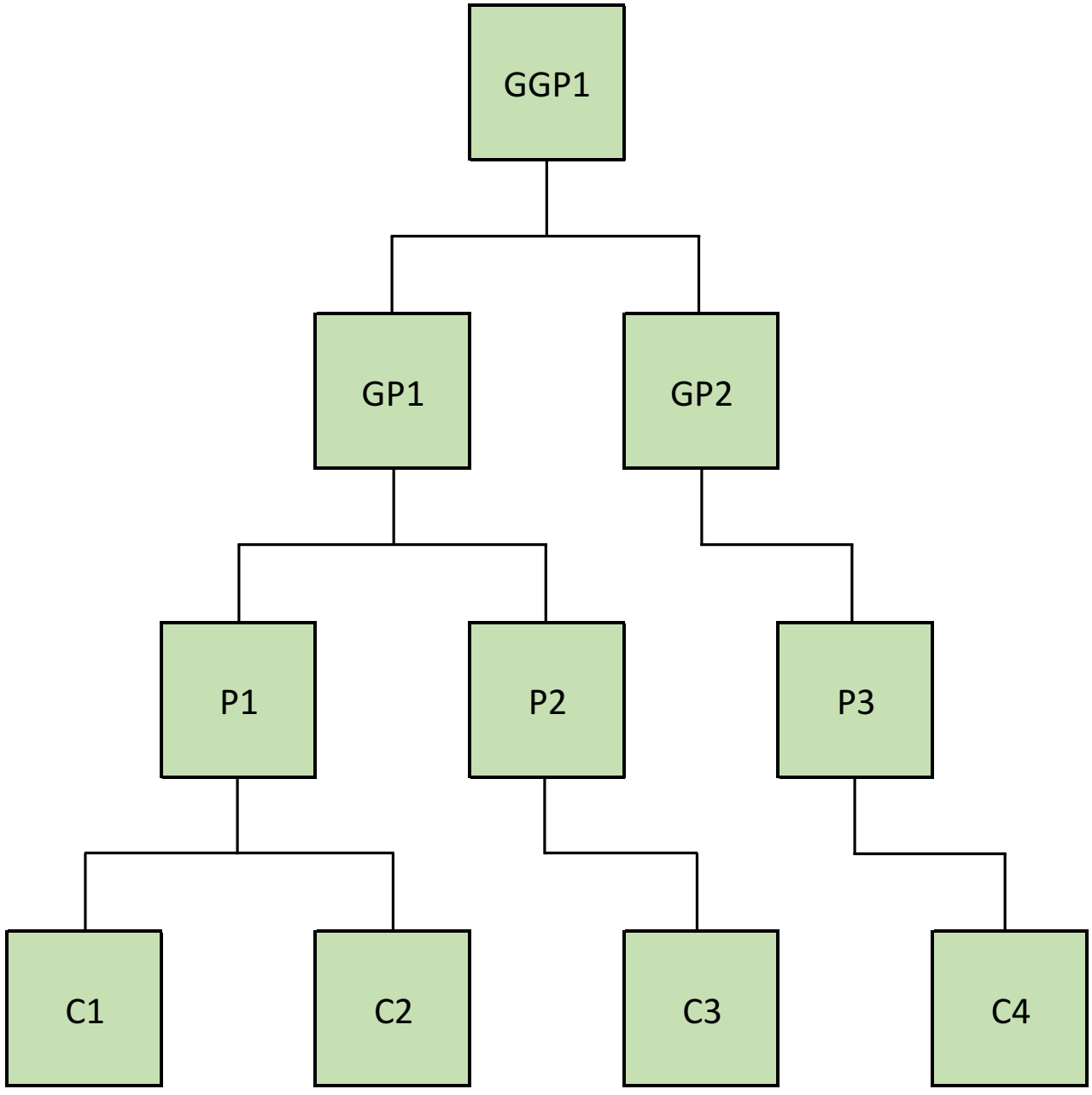
(1) Swedish register data:

- 1/3 of Swedish population born between 1932 and 1967, plus their siblings, parents and children
- family links up to three (four) generations (censoring/selectivity)

(2) Spanish Census from Cantabria, with full name of each person:

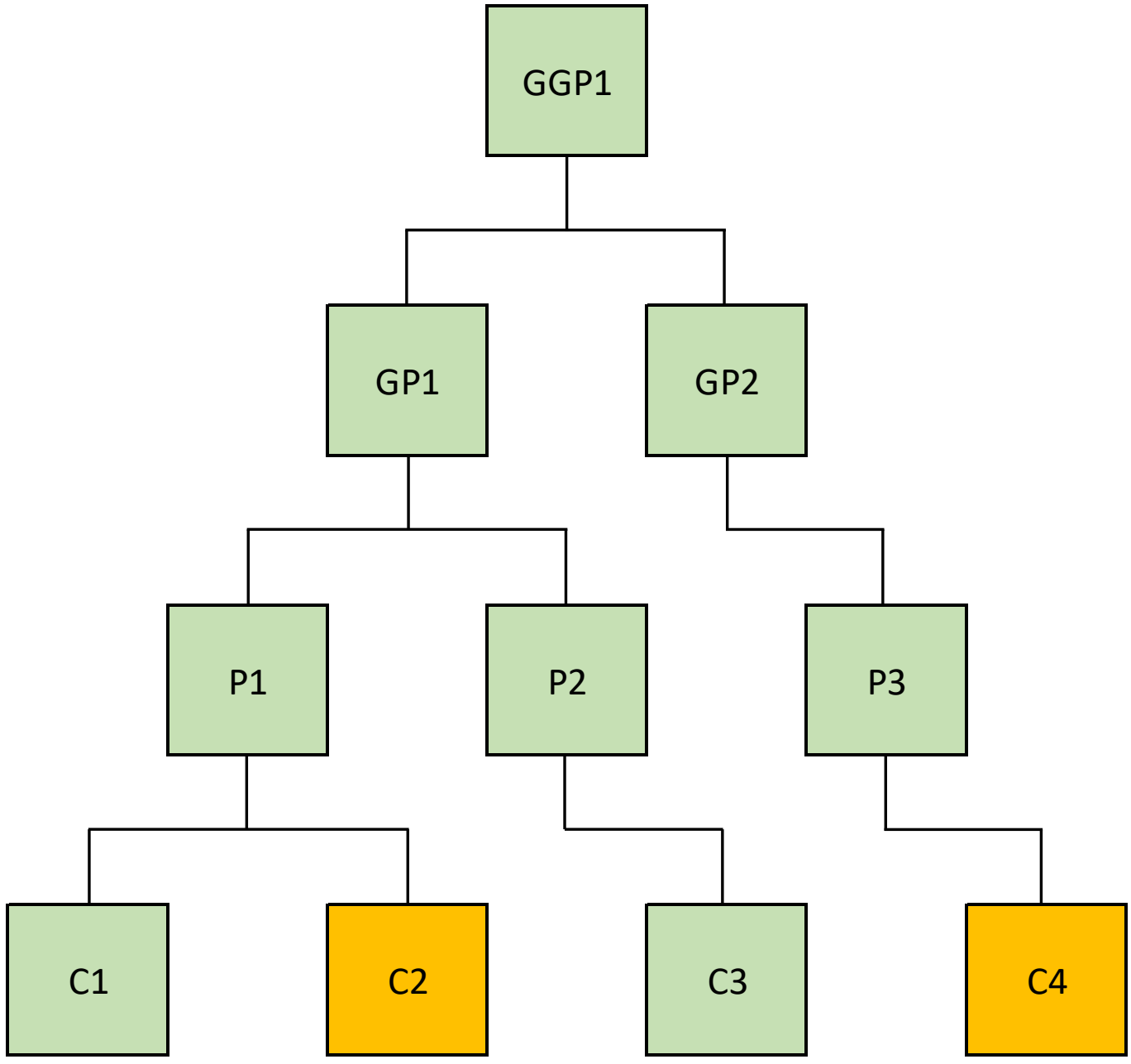
- newborns in Spain receive two surnames, with first=father's and second=mother's (first) surname
- child generation born 1956-1976 (71,479 males, 68,830 females), identify relatives via surnames
- only education

▸ Identifying relatives in the Spanish Census

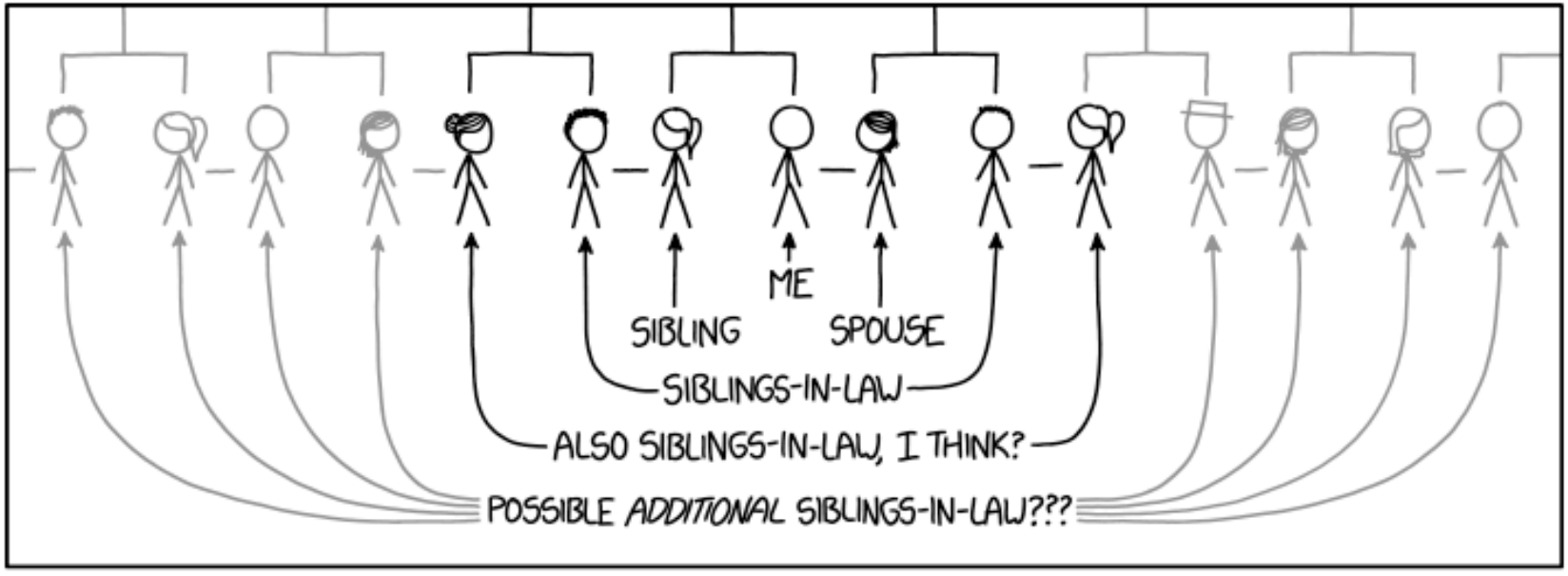

GGP1
GP1
GP2
P1
P2
P3
C1
C2
C3
C4

	kinship	kinship type	# correlations
a–x	spouses	direct, horizontal	1
x–b	siblings	direct, horizontal	3
ax–by	cousins	direct, horizontal	10
ax–a	child-parent	direct, vertical	4
ax–b	child-uncle/aunt	direct, vertical	8

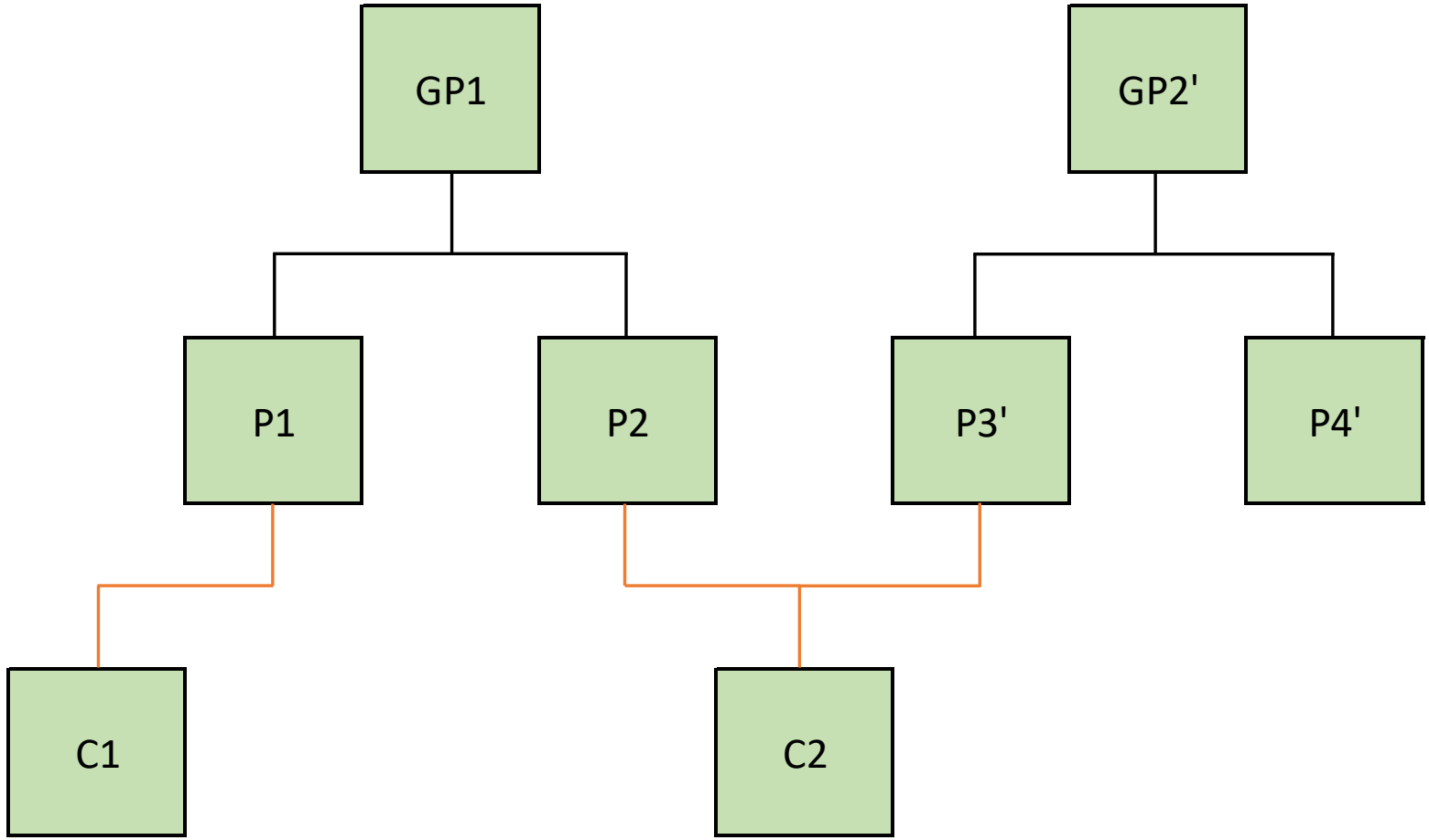
outcome	n	kinship	# families	# pairs	observed
educyrs	1	husband-wife	399,861	413,062	0.491
	2	Brothers	49,327	59,749	0.438
	3	Sisters	44,924	53,787	0.418
	4	Brothers-Sisters	87,548	111,545	0.375
	5	MCousins-FB	31,353	70,137	0.167
	6	FCousins-FB	29,581	63,032	0.135
	7	MFCousins-FB	53,357	144,100	0.143
	8	MCousins-MS	36,602	82,049	0.172
	9	FCousins-MS	34,025	73,649	0.158
	10	MFCousins-MS	62,522	170,577	0.157
	11	MCousins-FBMS	62,210	156,747	0.161
	12	FCousins-FBMS	58,410	140,522	0.142
	13	MFCousins-FMMF	60,335	148,691	0.143
	14	MFCousins-MMFF	60,200	148,631	0.147
	15	Father-son	320,020	396,304	0.380
	16	Father-daugther	306,933	376,255	0.321
	17	Mother-son	342,038	306,470	0.366
	18	Mother-daugther	327,809	400,337	0.347
	19	Uncle-nephew-BF	177,515	280,067	0.254
	20	Uncle-niece-BF	172,660	266,289	0.218
	21	Uncle-nephew-BM	198,086	312,019	0.241
	22	Uncle-niece-BM	191,862	295,580	0.209
	23	Aunt-nephew-SF	182,561	285,618	0.234
	24	Aunt-niece-SF	176,859	270,325	0.217
	25	Aunt-nephew-SM	209,942	333,141	0.251
	26	Aunt-niece-SM	203,208	316,625	0.235

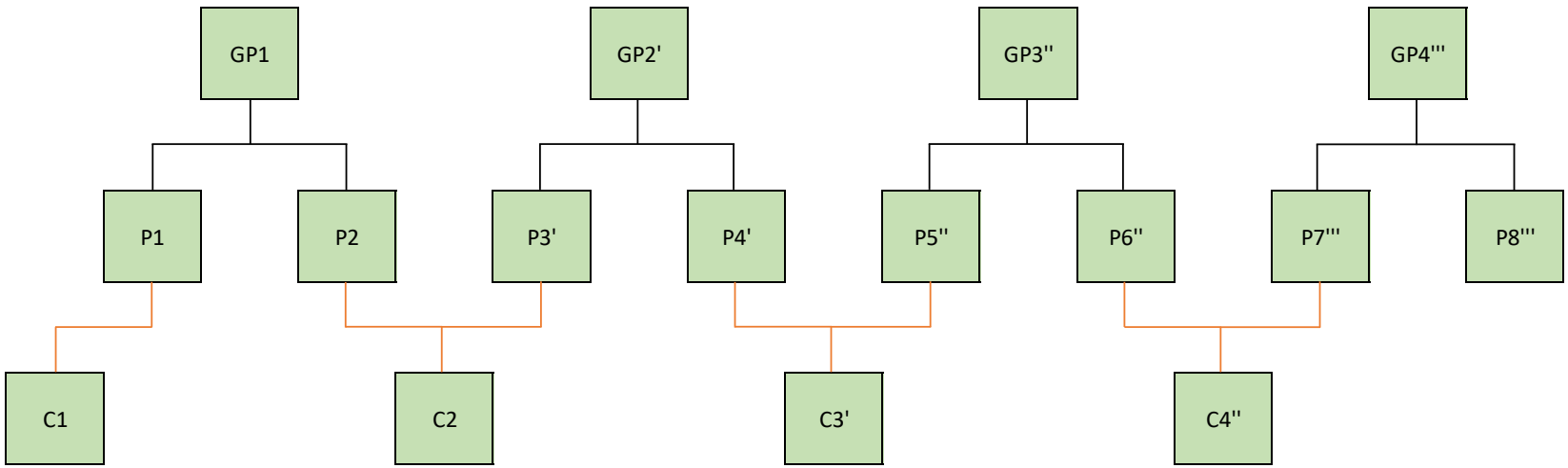

GGP1
GP1
GP2
P1
P2
P3
C1
C2
C3
C4

Sibling-in-Law



People complain that "<X>th cousin <Y> times removed" is hard to understand, but to me the most confusing one is sibling-in-law, because it chains across both sibling and marriage links and I don't really know where it stops.


GP1
GP2'
P1
P2
P3'
P4'
C1
C2


GP1
GP2'
GP3''
GP4'''
P1
P2
P3'
P4'
P5''
P6''
P7'''
P8'''
C1
C2
C3'
C4''

	kinship	kinship type	# correlations
a–x	spouses	direct, horizontal	1
x–b	siblings	direct, horizontal	3
ax–by	cousins	direct, horizontal	10
ax–a	child-parent	direct, vertical	4
ax–b	child-uncle/aunt	direct, vertical	8
a–b	siblings in-law (degree 1)	affinity, horizontal	4
a–y	spouse of sib-in-law (dg 1)	"	3
x–c	sibling of sib-in-law (dg 1)	"	4
a–c	siblings in-law (degree 2)	"	8
a–z	spouse of sib-in-law (dg 2)	"	4
x–d	sibling of sib-in-law (dg 2)	"	10
a–d	siblings in-law (degree 3)	"	16
a–w	spouse of …	"	…
ax–y	child-sibling in law of parents (dg 1)	affinity, vertical	8
...	…	"	…

- 205 moments (but some duplicates $\rightarrow$ 141 unique moments).
- Minimize difference between theoretical moments ρ_i and sample moments $\hat{\rho}_i$, $min_{\{...\}} \sum_i w_i(\rho_i - \hat{\rho}_i)^2$.

Horizontal approach: Summary

Advantages of "horizontal" compared to "vertical" approach:

1. Socioeconomic outcomes measured within same generation, at approximately same age and time
   Vertical approach: Distant ancestors tend to have basic education; are mostly farmers
2. Can use modern "big data" registry data (with direct family links)
   Vertical approach relies on historical sources, surname-based estimators
3. Can consider many more kinship moments
   Can consider more detailed intergenerational models

The model

Outcome y_t^i of child i in generation t

$$y_t^k = \beta^k \tilde{y}_{t-1}^k + z_t^k + x_t^k + u_t^k$$
$$z_t^k = \gamma^k \tilde{z}_{t-1}^k + e_t^k + v_t^k$$

where $k = \{m, f\}$ denotes male or female children, and $\{\tilde{y}_{t-1}^k, \tilde{z}_{t-1}^k\}$ weighted parental averages,

$$\tilde{y}_{t-1}^k = \alpha_y^k y_{t-1}^m + (1 - \alpha_y^k) y_{t-1}^f$$
$$\tilde{z}_{t-1}^k = \alpha_z^k z_{t-1}^m + (1 - \alpha_z^k) z_{t-1}^f$$

The x_t^k and e_t^k are shared by siblings of the same gender, correlated between siblings of different genders.

Assortative mating

Assortative mating in both observed and latent variable.

Consider the linear projection

$$\begin{pmatrix} z^f_{t-1} \\ y^f_{t-1} \end{pmatrix} = \begin{pmatrix} r^m_{zz} & r^m_{zy} \\ r^m_{yz} & r^m_{yy} \end{pmatrix} \begin{pmatrix} z^m_{t-1} \\ y^m_{t-1} \end{pmatrix} + \begin{pmatrix} w^m_{t-1} \\ \varepsilon^m_{t-1} \end{pmatrix}$$

where w^f_{t-1} and ε^f_{t-1} might be correlated, but uncorrelated with z^f_{t-1} and y^f_{t-1}, and where $r^f_{sd}(s, d = y, z)$ are functions of correlations and standard deviations of z^f, z^m,y^f, y^m.

Model summary

The **baseline model** is comparatively general, allowing for:

1. Direct (β^k) and indirect (γ^k) transmission
2. Two-parent structure (*cannot* be reduced to one parent)
3. Assortative mating in two dimensions (in y_t and z_t)
4. Correlated shocks among siblings (x_t^k and e_t^k)
5. Gender asymmetries in all parameters

In total we have 21 unknown parameters.

Swedish data: **Education**

Education (baseline)

Education (years of schooling, demeaned by gender and cohort):

- 141 distinct moments (up to sibling-in-laws of 5th order)
- Correlations weighted by family size

Baseline specification:

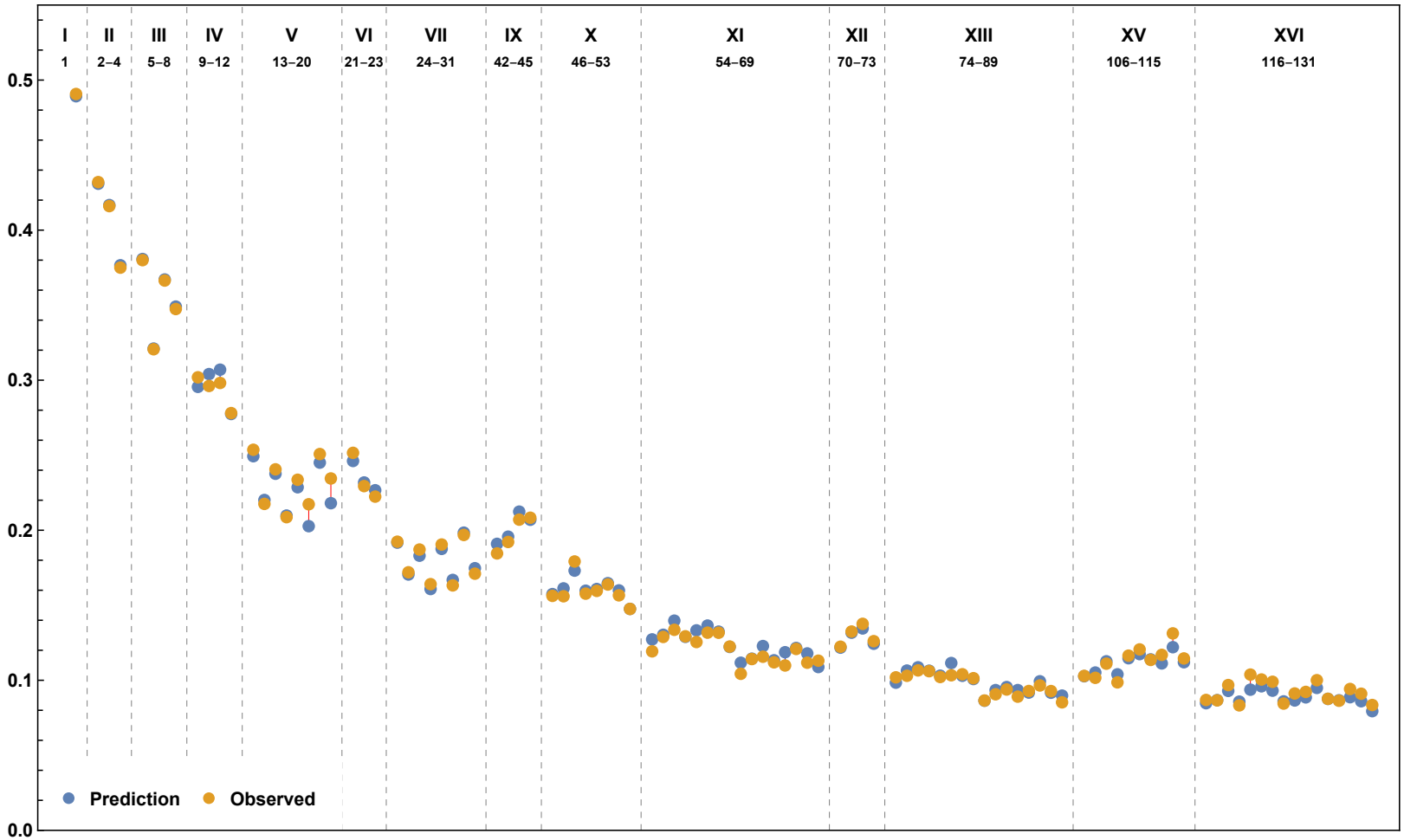
- All correlations up to in-laws of 3rd order, exclude cousins
- 105 moments

Table: Estimated and Calibrated Moments in Swedish registers (Education)

Kinship type			Data		Calibration		Kinship type			Data		Calibration	
#	##	name	number of pairs (1)	sample correlation (2)	predicted correlation (3)	percent error (4)	#	##	name	number of pairs (1)	sample correlation (2)	predicted correlation (3)	percent error (4)
I	1	Husband-Wife	413,062	0.491	0.489	-0.3		...					
II	2	Brother	387,028	0.432	0.431	-0.3	XII	72	MFMS	299,602	0.138	0.135	-2.2
	3	Sister	431,698	0.416	0.417	0.3		73	FMMS	273,809	0.126	0.124	-1.4
	4	Brother-Sister	800,127	0.375	0.377	0.5	XIII	74	M-MMMS	160,726	0.102	0.098	-3.5
III	5	Father-Son	396,304	0.380	0.381	0.2		75	M-MMFS	174,261	0.103	0.106	3.4
	6	Father-Daughter	376,255	0.321	0.321	0.1		76	M-MFMS	158,401	0.107	0.109	1.9
	7	Mother-Son	306,470	0.366	0.367	0.2		77	M-MFFS	160,105	0.106	0.106	0.3
	8	Mother-Daughter	400,337	0.347	0.349	0.5		78	M-FMMS	147,949	0.102	0.103	0.9
IV	9	Brother in-law (HS)	602,262	0.302	0.296	-2.1		79	M-FMFS	156,876	0.103	0.111	7.9
	10	Brother-Sister in-law (WB)	578,269	0.296	0.304	2.7		80	M-FFMS	133,588	0.104	0.103	-1.0
	11	Brother-Sister in-law (HS)	650,127	0.298	0.307	3.0		81	M-FFFS	131,756	0.101	0.101	-0.5
	12	Sister in-law (WB)	596,540	0.278	0.277	-0.2		82	F-MMMS	152,751	0.087	0.086	-0.1
V	13	Nephew-Uncle (BF)	280,067	0.254	0.249	-1.7		83	F-MMFS	165,828	0.091	0.093	3.1
	14	Niece-Uncle (BF)	266,289	0.218	0.220	1.2		84	F-MFMS	151,100	0.094	0.095	1.7
	15	Nephew-Uncle (BM)	312,019	0.241	0.238	-1.2		85	F-MFFS	153,065	0.089	0.093	4.9
	16	Niece-Uncle (BM)	295,580	0.209	0.210	0.5		86	F-FMMS	140,585	0.093	0.092	-1.1
	17	Nephew-Aunt (SF)	285,618	0.234	0.229	-2.1		87	F-FMFS	150,162	0.097	0.099	2.9
	18	Niece-Aunt (SF)	270,325	0.217	0.203	-6.7		88	F-FFMS	126,129	0.093	0.092	-1.2
	19	Nephew-Aunt (SM)	333,141	0.251	0.245	-2.3		89	F-FFFS	124,968	0.085	0.090	5.3
	20	Niece-Aunt (SM)	316,625	0.234	0.218	-7.0	*XIV*	*90*	*M-MMM-M*	*84,025*	*0.094*	*0.082*	*-13.4*
VI	21	Brother in-law (WWS)	252,232	0.252	0.246	-2.2		*91*	*M-MMF-M*	*100,261*	*0.101*	*0.086*	*-15.2*
	22	Sister in-law (HHB)	226,795	0.229	0.232	1.1		*92*	*M-MFM-M*	*93,237*	*0.105*	*0.090*	*-13.9*
	23	Brother-Sister in-law (HWBS	464,081	0.222	0.227	1.9		*93*	*M-FMM-M*	*80,486*	*0.097*	*0.085*	*-12.0*
VII	24	Nephew-Aunt in-law (BF)	231,767	0.192	0.192	-0.3		*94*	*M-MMM-F*	*79,690*	*0.087*	*0.073*	*-16.7*
	25	Niece-Aunt in-law (BF)	221,287	0.172	0.171	-0.8		*95*	*M-MMF-F*	*95,733*	*0.094*	*0.076*	*-19.9*
	26	Nephew-Aunt in-law (BM)	254,534	0.187	0.183	-2.2		*96*	*M-MFM-F*	*89,364*	*0.093*	*0.080*	*-13.6*
	27	Niece-Aunt in-law (BM)	241,873	0.164	0.161	-2.0		*97*	*M-MFF-F*	*95,020*	*0.095*	*0.075*	*-20.4*
	28	Nephew-Uncle in-law (SF)	227,403	0.190	0.188	-1.5		*98*	*M-FMM-F*	*76,514*	*0.095*	*0.076*	*-20.0*
	29	Niece-Uncle in-law (SF)	215,068	0.163	0.167	2.2		*99*	*M-FMF-F*	*89,054*	*0.088*	*0.079*	*-9.8*
	30	Nephew-Uncle in-law (SM)	264,524	0.197	0.198	0.8		*100*	*M-FFM-F*	*77,332*	*0.094*	*0.076*	*-19.0*
	31	Niece-Uncle in-law (SM)	251,782	0.171	0.175	2.1		*101*	*M-FFF-F*	*80,067*	*0.082*	*0.072*	*-12.9*
VIII	*32*	*Male Cousins (B)*	*70,137*	*0.208*	*0.159*	*-23.8*		*102*	*F-MMM-F*	*76,344*	*0.080*	*0.064*	*-20.3*

	33	*Male Cousins (S)*	*82,049*	*0.215*	*0.160*	*-25.4*		*103*	*F-MMF-F*	*91,080*	*0.090*	*0.066*	*-26.6*
	34	*Male Cousins (BS)*	*156,747*	*0.202*	*0.152*	*-24.8*		*104*	*F-MFM-F*	*84,736*	*0.092*	*0.070*	*-23.4*
	35	*Female Cousins (B)*	*63,032*	*0.169*	*0.126*	*-25.5*		*105*	*F-FMM-F*	*72,410*	*0.082*	*0.068*	*-17.4*
	36	*Female Cousins (S)*	*73,649*	*0.197*	*0.124*	*-37.0*	XV	106	XMMMM	288,374	0.103	0.103	-0.2
	37	*Female Cousins (BS)*	*140,522*	*0.177*	*0.118*	*-33.2*		107	XMMMF	312,703	0.102	0.105	3.5
	38	*Male-Female Cousins (B)*	*144,100*	*0.179*	*0.141*	*-20.9*		108	XMMFM	311,795	0.111	0.113	1.4
	39	*Male-Female Cousins (S)*	*170,577*	*0.196*	*0.141*	*-28.2*		109	XMMFF	162,928	0.099	0.104	5.4
	40	*Male-Female Cousins (BS)*	*148,691*	*0.179*	*0.133*	*-25.5*		110	XMFMM	308,163	0.116	0.115	-1.5
	41	*Male-Female Cousins (SB)*	*148,631*	*0.184*	*0.135*	*-26.5*		111	XMFMF	166,250	0.121	0.117	-2.6
IX	42	XMMM	461,883	0.185	0.191	3.5		112	XMFFM	304,684	0.114	0.114	0.3
	43	XMMF	500,448	0.192	0.196	1.8		113	XFMMM	278,416	0.117	0.111	-4.9
	44	XMFM	481,006	0.207	0.212	2.6		114	XFMFM	149,478	0.131	0.122	-7.0
	45	XFMM	447,263	0.208	0.207	-0.6		115	XFFMM	143,733	0.115	0.112	-2.2
X	46	MMM	362,409	0.156	0.157	0.8	XVI	116	MMMM	230,313	0.087	0.085	-2.5
	47	MMF	393,579	0.156	0.161	3.4		117	MMMF	251,223	0.087	0.087	0.2
	48	MFM	375,442	0.179	0.173	-3.4		118	MMFM	248,811	0.097	0.093	-4.0
	49	MFF	391,389	0.158	0.160	1.2		119	MMFF	259,925	0.083	0.086	3.0
	50	FMM	353,470	0.160	0.161	0.8		120	MFMM	245,814	0.104	0.094	-9.7
	51	FMF	378,720	0.164	0.165	0.5		121	MFMF	265,220	0.101	0.096	-4.5
	52	FFM	341,316	0.157	0.160	2.1		122	MFFM	241,998	0.099	0.093	-6.0
	53	FFF	351,350	0.148	0.148	0.0		123	MFFF	248,449	0.084	0.086	1.7
XI	54	M-MMM	202,632	0.119	0.127	6.7		124	FMMM	224,873	0.091	0.087	-5.1
	55	M-MMF	219,007	0.129	0.130	1.1		125	FMMF	246,186	0.092	0.089	-3.9
	56	M-MFM	192,819	0.134	0.140	4.6		126	FMFM	237,791	0.100	0.095	-5.1
	57	M-MFF	199,811	0.129	0.129	-0.3		127	FMFF	247,495	0.088	0.088	0.0
	58	M-FMM	183,670	0.125	0.133	6.2		128	FFMM	223,661	0.086	0.087	0.3
	59	M-FMF	196,631	0.132	0.136	3.6		129	FFMF	240,328	0.094	0.089	-5.8
	60	M-FFM	160,857	0.132	0.132	0.6		130	FFFM	213,155	0.091	0.086	-5.5
	61	M-FFF	164,528	0.122	0.122	-0.2		131	FFFF	220,553	0.084	0.079	-4.9
	62	F-MMM	192,818	0.104	0.112	7.1	*XVII*	*132*	*MMMMS*	*176,790*	*0.071*	*0.066*	*-8.3*
	63	F-MMF	208,008	0.114	0.114	0.2		*133*	*MMMFS*	*199,041*	*0.075*	*0.071*	*-6.0*
	64	F-MFM	183,929	0.116	0.123	6.1	*XVIII*	*134*	*MMMMM*	*153,057*	*0.047*	*0.046*	*-3.2*
	65	F-MFF	191,177	0.112	0.113	1.2		*135*	*FFFFF*	*144,976*	*0.054*	*0.043*	*-20.1*
	66	F-FMM	175,507	0.110	0.119	8.0	*XIX*	*136*	*MMMMMS*	*117,473*	*0.047*	*0.035*	*-25.5*
	67	F-FMF	187,178	0.121	0.122	0.5		*137*	*MMMMFS*	*135,096*	*0.042*	*0.038*	*-9.9*
	68	F-FFM	151,606	0.112	0.118	5.5	*XX*	*138*	*MMMMMM*	*106,844*	*0.031*	*0.025*	*-21.9*
	69	F-FFF	155,658	0.113	0.109	-3.7		*139*	*FFFFFF*	*100,871*	*0.043*	*0.023*	*-46.5*
XII	70	MMMS	278,938	0.122	0.122	-0.5	*XXI*	*140*	*MMMMMMS*	*82,523*	*0.032*	*0.019*	*-40.0*
	71	MMFS	310,160	0.132	0.132	-0.6		*141*	*MMMMMFS*	*96,840*	*0.027*	*0.021*	*-24.4*

Figure: Baseline Fit in Swedish Registers



Panel A: Intergenerational Processes

Parameters:

β^m	β^f	γ^m	γ^f		
0.144	0.129	0.664	0.565		
σ^2_{ym}	σ^2_{yf}	σ^2_{zm}	σ^2_{zf}	σ^2_{um}	σ^2_{uf}
4.648	4.465	2.070	1.560	1.978	2.329
α_{ym}	α_{yf}	α_{zm}	α_{zf}		
0.390	0.020	0.658	0.773		

Parent-child correlations in z:

Father-Son	Father-Dau	Mother-Son	Mother-Dau
0.586	0.600	0.527	0.508

Ancestor correlations in y and z:

	Father-Son	Grandf-...	GGrandf-...	GGGrandf-Son
in y	0.381	0.209	0.121	0.071
in z	0.586	0.343	0.201	0.118

Panel B: Sibling Processes

Parameters:

σ^2_{xm}	σ^2_{xf}	σ_{xmxf}	σ^2_{em}	σ^2_{ef}	σ_{emef}
0.178	0.246	0.069	0.657	0.711	0.625

Variance Shares:

in y	3.8%	5.5%	1.5%	14.1%	15.9%	13.7%
in z	-	-	-	31.7%	45.6%	34.8%

Sibling correlations in z:

Brothers	Sisters	Brother-Sister
0.678	0.824	0.711

Baseline: Intergenerational results

- **Intergenerational transmission** (vertical)
  - Little direct transmission ($\beta^k \approx 0.1$)
  - Strong latent transmission ($\gamma^k \approx 0.6$)
  - Parent-child correlation substantially larger in latent than educational advantages: $Corr(z_t^k, z_{t-1}^k) \approx 0.55$ vs. $Corr(y_t^k, y_{t-1}^k) \approx 0.35$
- **Siblings** (horizontal)
  - Siblings share observable (captured in sibling correlation) and latent (not fully captured) advantages.
  - Shared latent component quite important.
  - Sibling correlations in latent factor ≈ 0.7 vs. ≈ 0.4 in years of schooling

Baseline: Intergenerational results

- **Intergenerational transmission** (vertical)
  - Little direct transmission ($\beta^k \approx 0.1$)
  - Strong latent transmission ($\gamma^k \approx 0.6$)
  - Parent-child correlation substantially larger in latent than educational advantages: $Corr(z_t^k, z_{t-1}^k) \approx 0.55$ vs. $Corr(y_t^k, y_{t-1}^k) \approx 0.35$
- **Siblings** (horizontal)
  - Siblings share observable (captured in sibling correlation) and latent (not fully captured) advantages.
  - Shared latent component quite important.
  - Sibling correlations in latent factor ≈ 0.7 vs. ≈ 0.4 in years of schooling

Panel C: Assortative Processes

Parameters:

r^m_{zz}	r^m_{zy}	r^m_{yz}	r^m_{yy}	$\sigma^2_{\omega m}$	$\sigma^2_{\epsilon m}$
0.663	-0.008	0.696	0.143	0.673	2.919
r^f_{zz}	r^f_{zy}	r^f_{yz}	r^f_{yy}		
0.747	0.112	0.662	0.249		

Spousal correlations in y and z:

ρ_{ymyf}	ρ_{zmzf}	ρ_{ymzf}	ρ_{zmyf}
0.489	0.754	0.540	0.580

Panel D: Variance Decomposition

%	y	z	x	Cov(y,z)
Male	0.013	0.445	0.038	0.038
Female	0.016	0.349	0.055	0.030

Baseline: Results

- **Assortative mating** (horizontal)
  - Strong sorting in latent factor
  - Little additional sorting by education
  - Spousal correlations substantially higher in latent than in educational advantages: $Corr(z^m_{t-1}, z^f_{t-1}) = 0.79$ *vs.* $Corr(y^m_{t-1}, y^f_{t-1}) = 0.49$
- **Gender asymmetries**
  - Shared sibling component in latent factor z similar for same- and mixed-gender siblings
  - Shared sibling component in education y lower for mixed-gender siblings

Baseline: Results

- **Assortative mating** (horizontal)
  - Strong sorting in latent factor
  - Little additional sorting by education
  - Spousal correlations substantially higher in latent than in educational advantages: $Corr(z^m_{t-1}, z^f_{t-1}) = 0.79$ *vs.* $Corr(y^m_{t-1}, y^f_{t-1}) = 0.49$
- **Gender asymmetries**
  - Shared sibling component in latent factor z similar for same- and mixed-gender siblings
  - Shared sibling component in education y lower for mixed-gender siblings

Robustness and out-of-sample fit

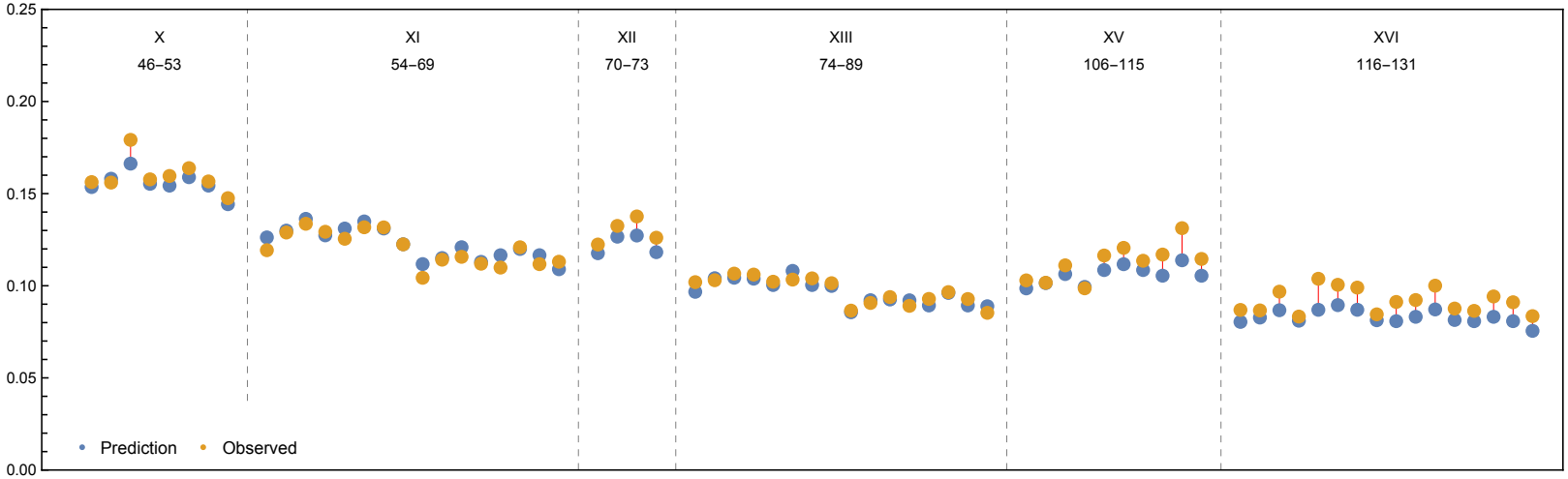
Good **in-sample** fit:

- Across vertical and horizontal moments
- Across consanguine ("blood") and affine ("in-law") relations
- Mean absolute error across 105 kinship types = 1.9 percent

Mostly good **out-of-sample** fit. Robustness test:

- Drop moment groups 10+ (including distant kins)
- Reduces set of empirical moments from 105 to 35

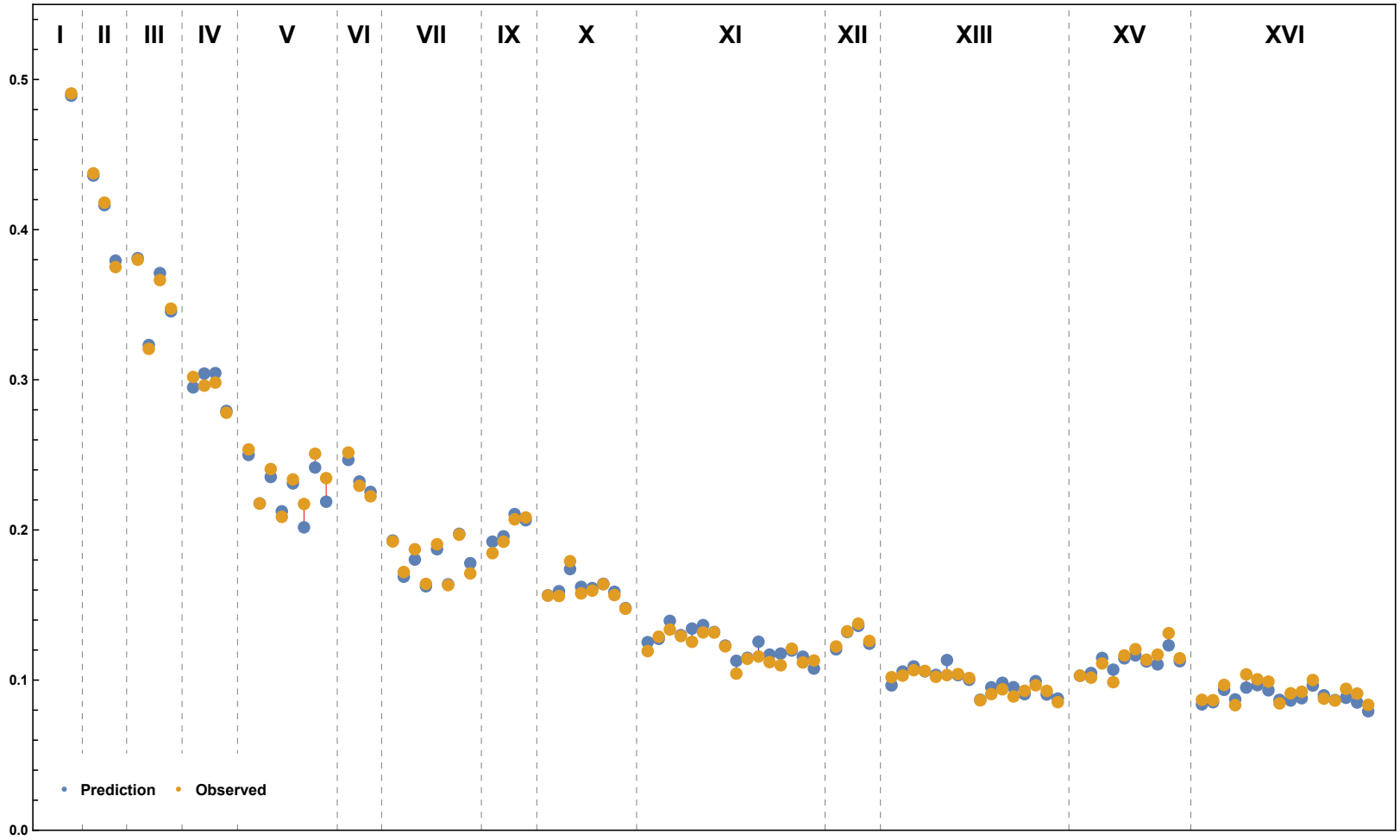
Figure: Out-of-Sample Fit in Swedish Registers



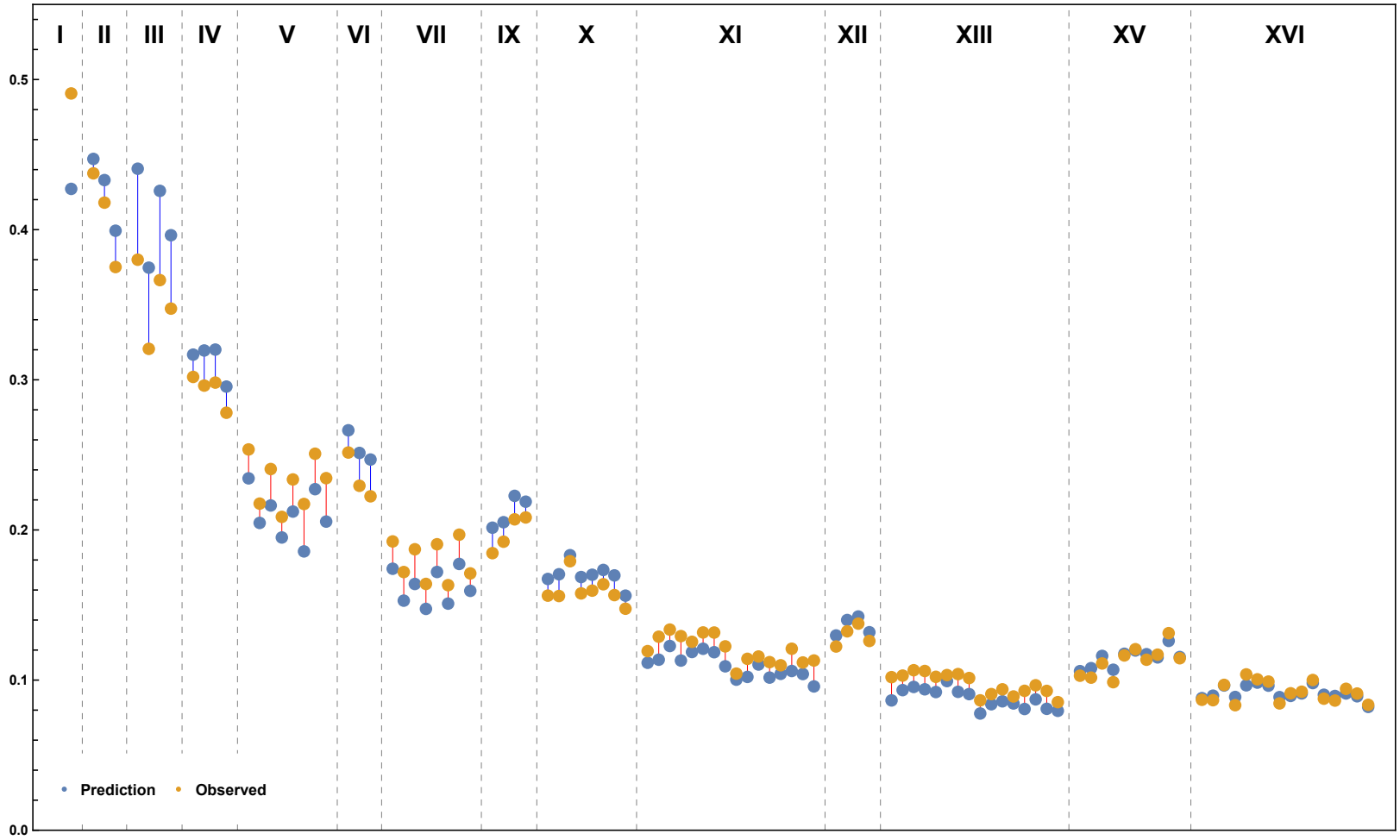
Restricted models

Can more restricted models fit the data?

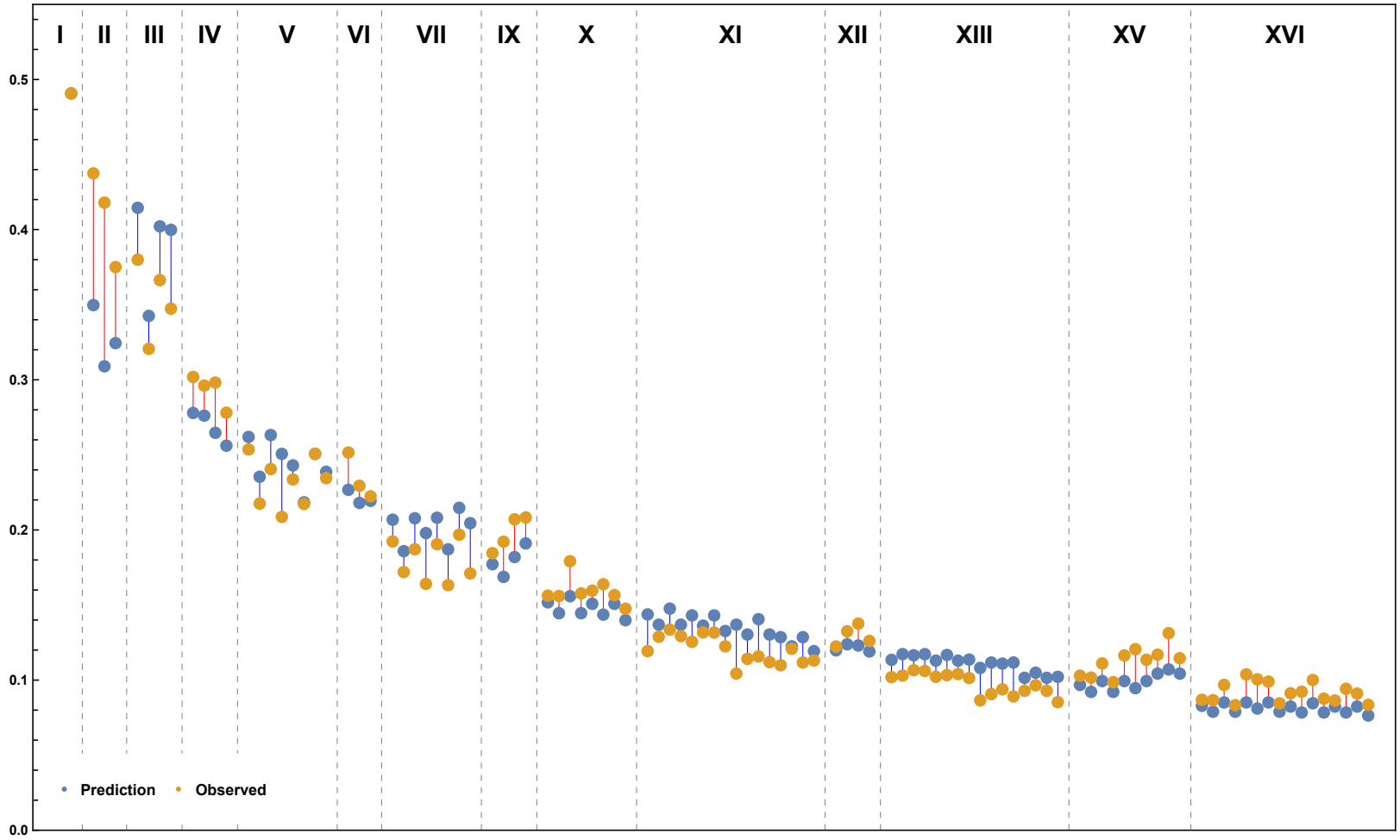
1. No direct transmission ($\beta = 0$)
2. No latent transmission ($\gamma = 0$)
3. No shared sibling component
4. Assortative mating only in observables



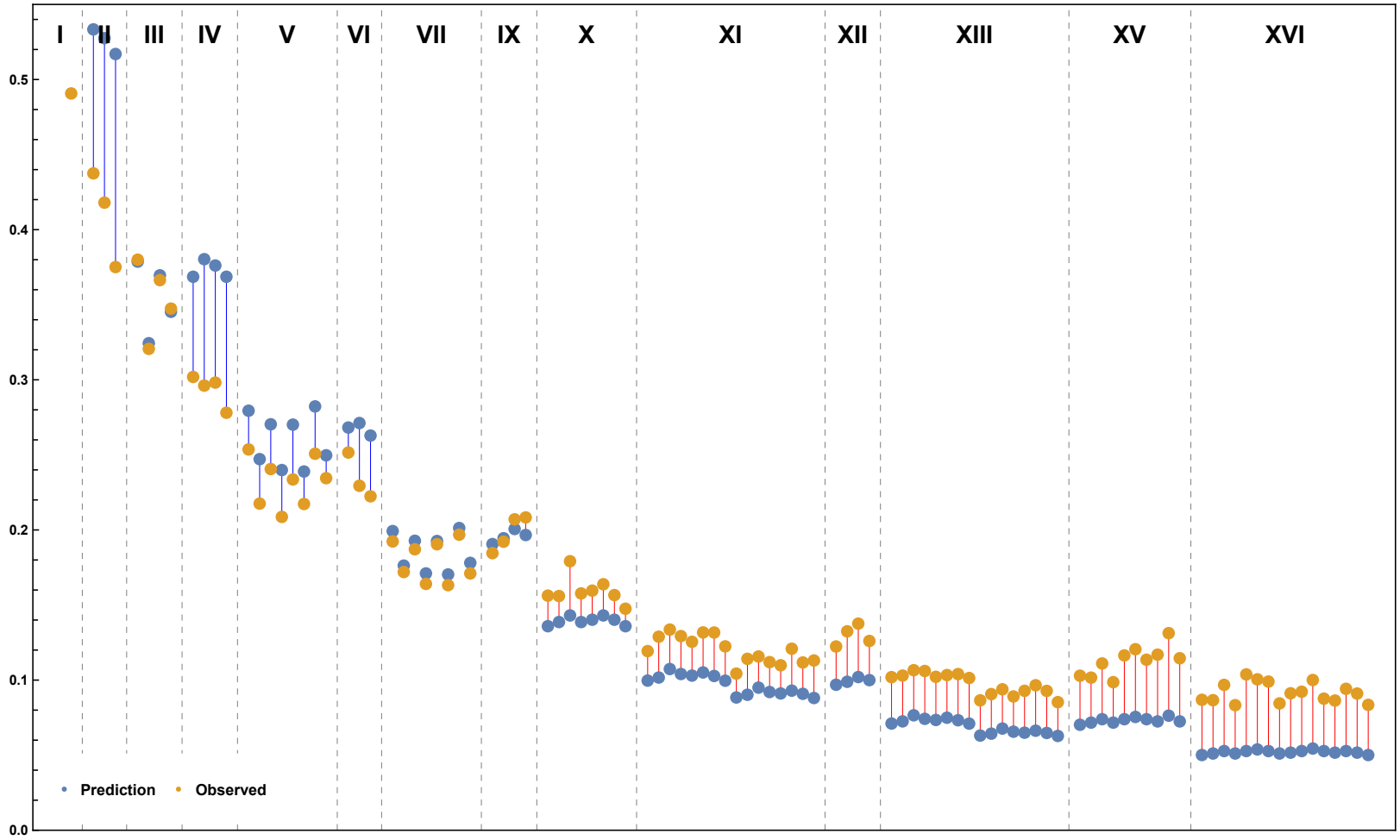
(a) Restricted model without direct transmission ($\beta = 0$)



(b) Restricted model without latent transmission ($\gamma = 0$)



(c) Restricted model without shared sibling component



(d) Assortative mating only in observables

Other applications

Sweden, income: ▸ Sweden Income

- Findings qualitatively similar, latent advantages more strongly transmitted than income itself
- However, vertical transmission of latent factors not as strong as for education

Spain, education: ▸ Spain Education

- Results qualitatively similar as in Sweden, but more persistence across all intergenerational, siblings, assortative dimensions
- Parent-child correlation in $z \approx 0.8$, spousal correlation ≈ 0.9

Other applications

Sweden, income: ▸ Sweden Income

- Findings qualitatively similar, latent advantages more strongly transmitted than income itself
- However, vertical transmission of latent factors not as strong as for education

Spain, education: ▸ Spain Education

- Results qualitatively similar as in Sweden, but more persistence across all intergenerational, siblings, assortative dimensions
- Parent-child correlation in $z \approx 0.8$, spousal correlation ≈ 0.9

Swedish data: The **Genetic Model**(s)

The standard genetic and two-factor models

Our baseline model:

- Quantifies transferability of socioeconomic advantages along intergenerational, sibling and assortative dimensions
- Otherwise remained agnostic about causal mechanisms

Question: Are *genes* an important component of the latent advantages captured by our model? Two exercises:

1. Standard genetic model (nested by our baseline model)
2. Two-factor model (with latent genetic and sociocultural factors)

The standard genetic and two-factor models

Our baseline model:

- Quantifies transferability of socioeconomic advantages along intergenerational, sibling and assortative dimensions
- Otherwise remained agnostic about causal mechanisms

Question: Are *genes* an important component of the latent advantages captured by our model? Two exercises:

1. Standard genetic model (nested by our baseline model)
2. Two-factor model (with latent genetic and sociocultural factors)

The standard genetic model

Standard genetic model of genetic inheritance with assortative mating used in Quantitative Genetics (Crow an Felsenstein, 1968).

Corresponds to our baseline model with following restrictions:

- $\beta^m = \beta^f = 0$
- $\gamma^m = \gamma^f = 1$, $\alpha_z^m = \alpha_z^f = 0.5$, $\sigma(z^m) = \sigma(z^f)$
- assortative mating in phenotype (e.g. education)

The standard genetic model and *education*

- The standard genetic model cannot fit kinship correlations in education
  Spouses must be far more similar in latent determinants of education than they are in "*phenotype*" education.
- However, the genetic model *appears* to fit if we were to consider only siblings, parents-children and uncles and aunts
- Need lots of data and many kinship moments to discriminate between genetic and other models!

Figure: Sample and Predicted Moments (Education, Genetic Model)

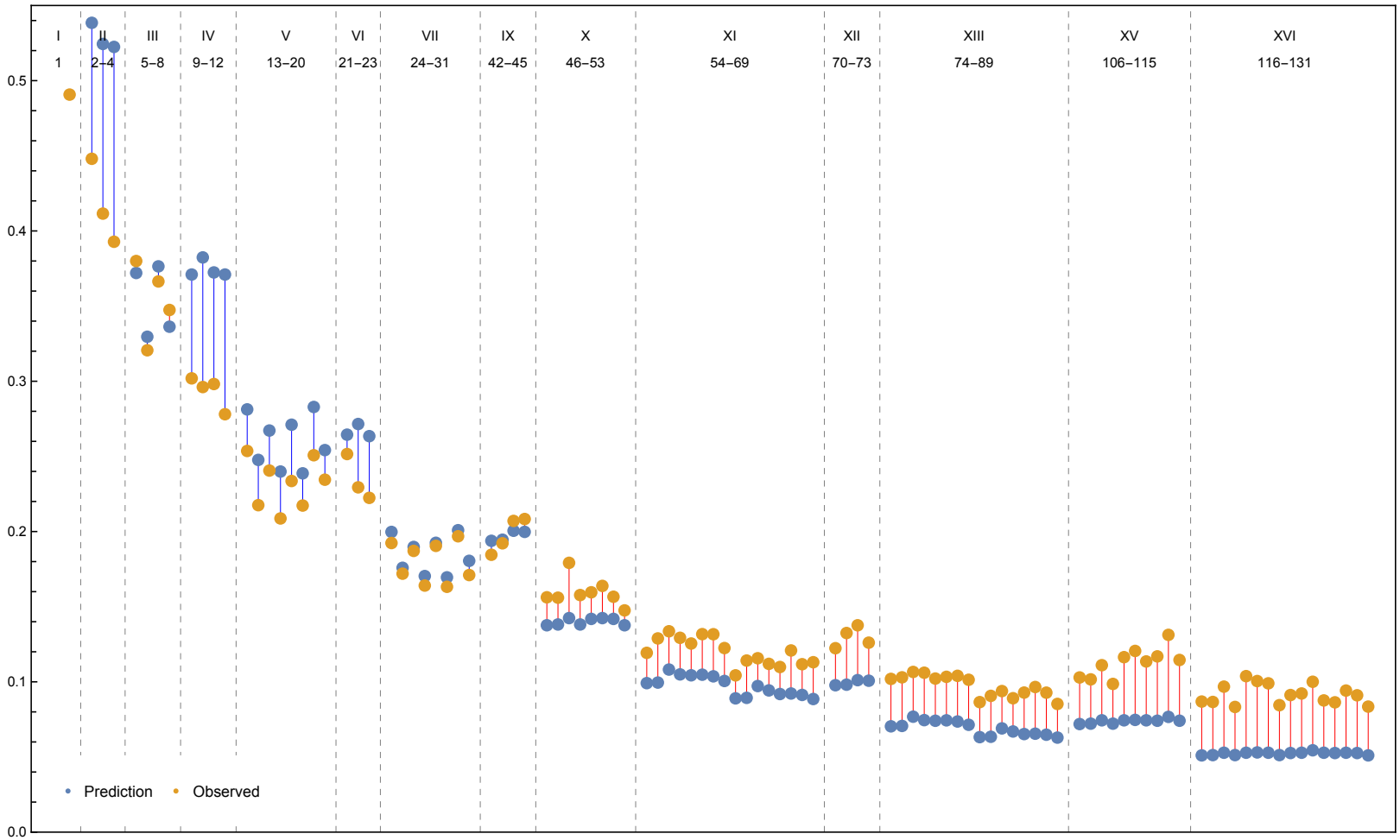
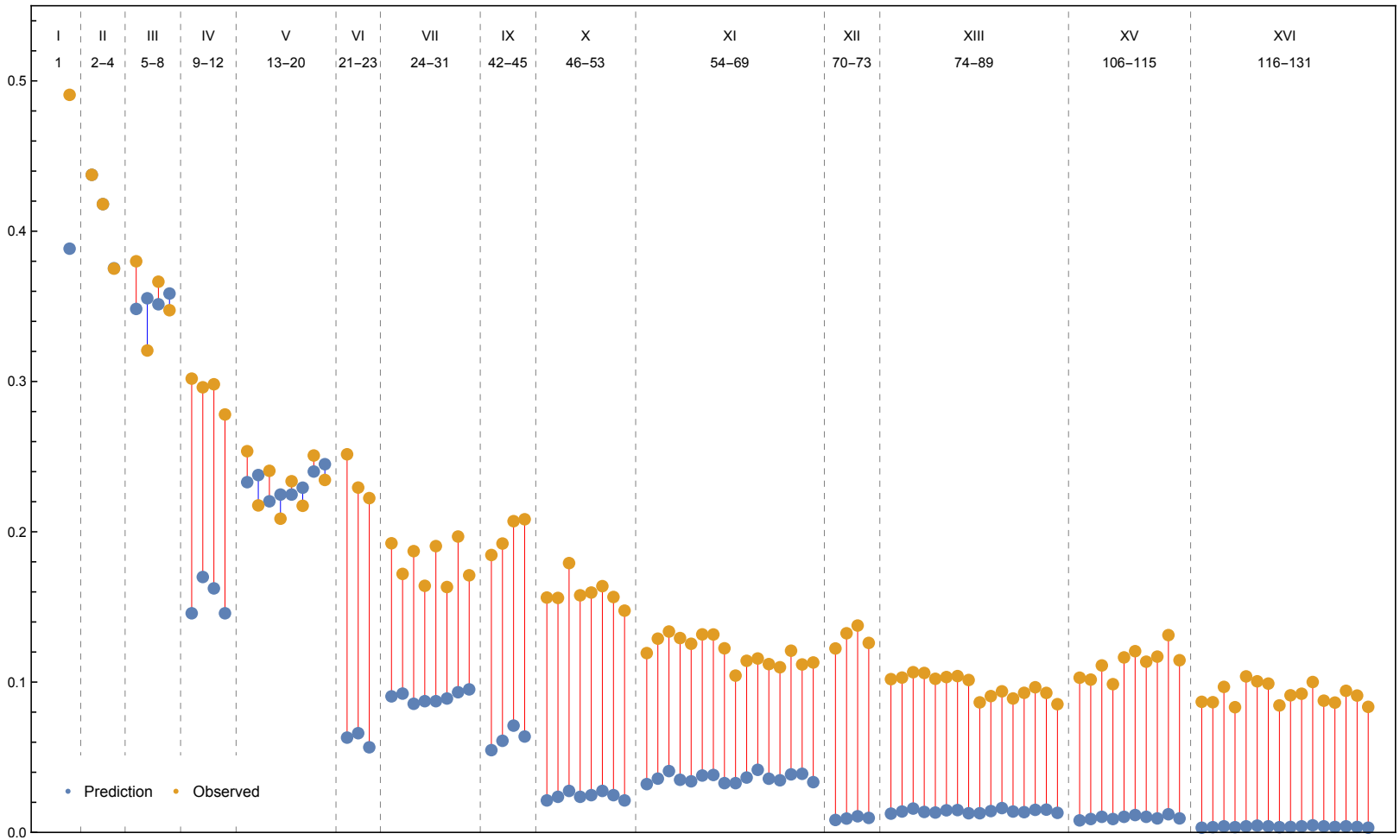


Figure: Sample and Predicted Moments (Education, Genetic Model, 15 Moments)



The two-factor model and *education*

Two-factor model: To quantify the relative contribution of genes, we decompose the the latent factor z_t^k into a genetic a genetic factor $z_t^{G,k}$, and a "cultural" factor $z_t^{C,k}$.

- Outcome y for individual from generation t and gender k

$$y_t^k = \beta^k \widetilde{y}_{t-1}^k + z_t^{G,k} + z_t^{C,k} + x_t^k + u_t^k$$

where $z_t^{G,k}$ follows the standard model of genetic inheritance (Crow and Felsenstein, 1968),

$$z_t^{G,k} = \frac{z_{t-1}^{G,m} + z_{t-1}^{G,f}}{2} + v_t^{G,k}$$

where $v_t^{G,k}$ is a white-noise error term.

- Do not need to impose that the "environments" of parents and offspring are independent as $z_t^{C,k}$ captures shared environments.

The two-factor model and *education*

Results based on the two-factor model: ▸ Results: Two Factor

- Genetic factor explains only 7% of the variance in years of schooling. Sociocultural factor explains 38% (31%) for males (females).

  Heritability estimate consistent with recent evidence from *genome-wide association studies* based on direct genetic information (e.g. Lee et al, 2018)

- Only negligible correlation between latent genetic factor and latent cultural factor.

- Little assortative mating in genes

  Consistent with evidence from polygenic scores (Domingue et al, 2014; Yengo et al 2018)

Summary

Main findings:

- Socioeconomic advantages are substantially more persistent than what one can observe from correlations in observable status (such as income or education)
- High degree of intergenerational persistence and very strong assortative mating in latent advantages
- A purely genetic model cannot explain the kinship pattern in education. (In Sweden, genes can explain 7%, sociocultural factors about 35% of the variance in education.)

Implications:

- What does this mean for observed cross-country differences in mobility, or between-group differences?

The end.

Interpretation

Does it matter if inequality is so highly persistent?

- Valuable input for debate on capitalism and inequality (e.g. Friedman, Becker, Piketty).
- Connects intergenerational inequality with group inequality, such as ethnic and racial inequalities (Borjas 1992, Margo 2017)? Individual-level and group-level persistence start to look similar.
- How to interpret prior parent-child evidence. Is mobility really higher in Canada than in US, or higher in Sweden than Spain?
- How effective is (social) policy across multiple generations, if observable status matters so little for intergenerational transmission?

Outlook

Remaining questions:

1. Consider more alternative models, such as the "grandparent effect" model ...
2. How sensitive are our results to violations of the steady-state assumption?
3. To which degree can we abstract from vertical moments?

Intergenerational persistence

- Early literature (e.g. Becker and Tomes, 1986), estimates intergenerational elasticity of income

$$y_t = \beta y_{t-1} + \varepsilon_t$$

  with $\hat{\beta} \approx 0.15$ for U.S.
- However, these estimates turned out to be downward biased because of measurement error
  - Attenuation bias from classical measurement error (e.g. Atkinson 1980s, Solon, 1999) $\rightarrow \hat{\beta} \approx 0.4$ for U.S.
  - Lifecycle bias (e.g. Jenkins 1987, Nybom and Stuhler 2016, Mazumder 2016) $\rightarrow \hat{\beta} \approx 0.5$ for U.S. (?)

*Multi*generational persistence

More recent literature on persistence across multiple generations:

- High persistence of socioeconomic status on the **surname** level (e.g. Clark, 2014). For example, in historical data from Florence the average status of surnames still correlates across generations that are six centuries apart (Barone and Mocetti, 2019)
- Other studies observe direct family links, but fewer generations

Can be interpreted in **latent** transmission model in

$$y_t = \delta z_t + u_t$$
$$z_t = \gamma z_{t-1} + v_t$$

$\rightarrow \gamma \approx 0.75$ (Clark, 2014) or ≈ 0.6 (Braun and Stuhler, 2018)?

Spanish data: **Education**

Education (Spain)

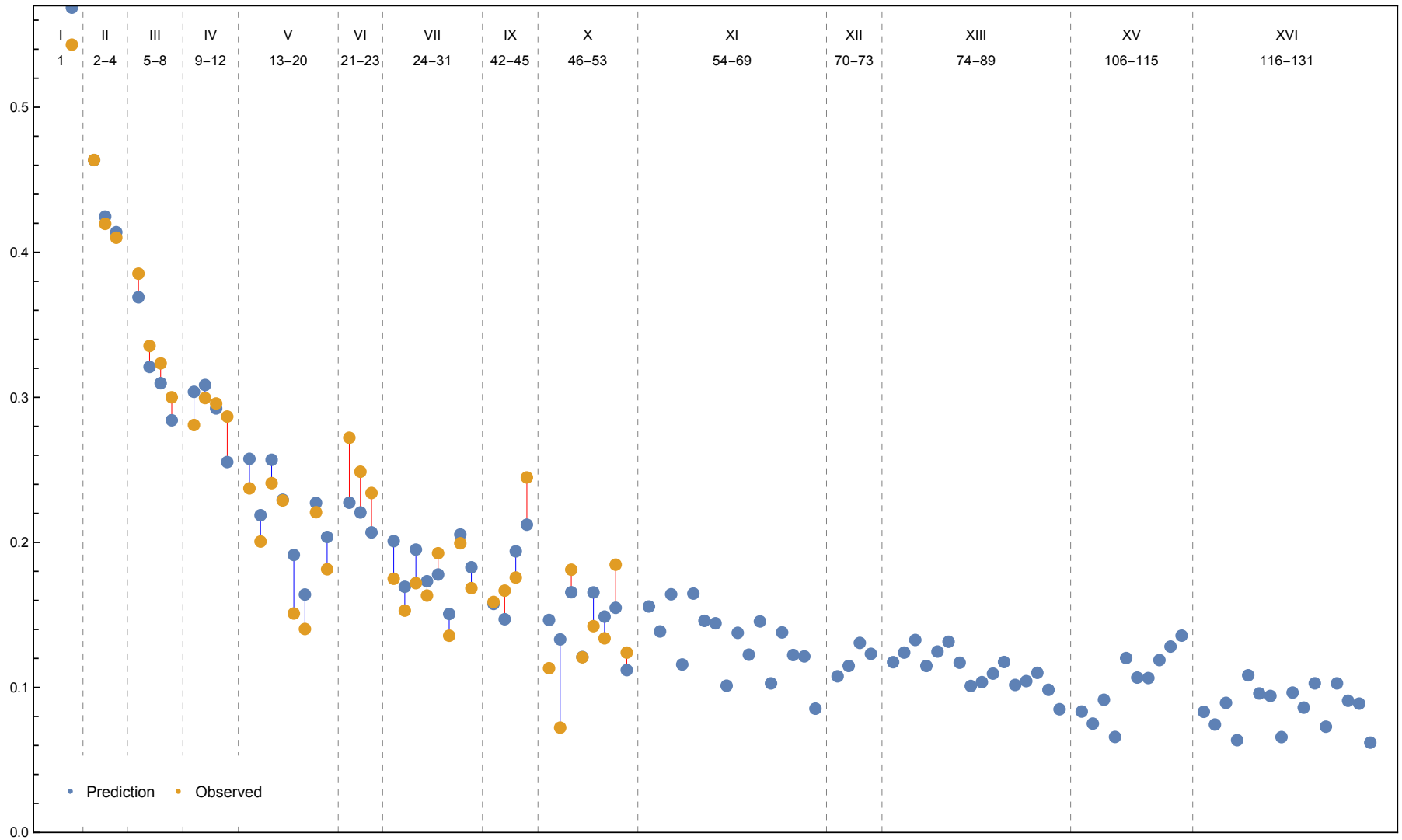
Education (years of schooling, demeaned by gender and cohort):

- Census from Cantabria
- 65 distinct moments: spouse, parent-child, siblings, nephew/niece-uncle/aunt, sibling-in-law up to second order
- # of moments should in principle suffice (-> robustness test in Swedish registers)

Main results:

- Results qualitatively similar as in Sweden
- Butt more persistence across all three dimensions: intergenerational, siblings, assortative
- Parent-child correlation in $z \approx 0.8$, spousal correlation ≈ 0.9

Figure: Fit in Spanish Census (Education)



Panel A: Intergenerational Processes

Parameters:

β^m	β^f	γ^m	γ^f		
0.027	0.111	0.915	0.842		
σ^2_{ym}	σ^2_{yf}	σ^2_{zm}	σ^2_{zf}	σ^2_{um}	σ^2_{uf}
13.579	13.213	6.519	2.779	5.162	7.003
α_{ym}	α_{yf}	α_{zm}	α_{zf}		
0.742	0.855	0.587	0.127		

Parent-child correlations in z:

Father-Son	Father-Dau	Mother-Son	Mother-Dau
0.760	0.827	0.732	0.883

Ancestor correlations in y and z:

	Father-Son	Grandf-...	GGrandf-...	GGGrandf-...
in y	0.369	0.271	0.205	0.156
in z				

Panel B: Sibling Processes

Parameters:

σ^2_{xm}	σ^2_{xf}	σ_{xmxf}	σ^2_{em}	σ^2_{ef}	σ_{emef}
1.650	2.644	2.089	0.558	0.001	0.018

Variance Shares:

in y	12.1%	20.0%	15.6%	4.1%	0.0%	0.1%
in z	-	-	-	8.6%	0.0%	0.4%

Sibling correlations in z:

Brothers	Sisters	Brother-Sister
0.674	0.784	0.667

Panel C: Assortative Processes

Parameters:

r^m_{zz}	r^m_{zy}	r^m_{yz}	r^m_{yy}	$\sigma^2_{\omega m}$	$\sigma^2_{\epsilon m}$
0.731	-0.139	0.418	0.357	0.381	8.369
r^f_{zz}	r^f_{zy}	r^f_{yz}	r^f_{yy}		
1.291	0.083	0.576	0.441		

Spousal correlations in y and z:

ρ_{ymyf}	ρ_{zmzf}	ρ_{ymzf}	ρ_{zmyf}
0.569	0.903	0.483	0.549

Panel D: Variance Decomposition

%	y	z	x	Cov(y,z)
Male	0.001	0.480	0.121	0.009
Female	0.010	0.210	0.200	0.009

Back

Swedish data: **Income**

Income (Sweden)

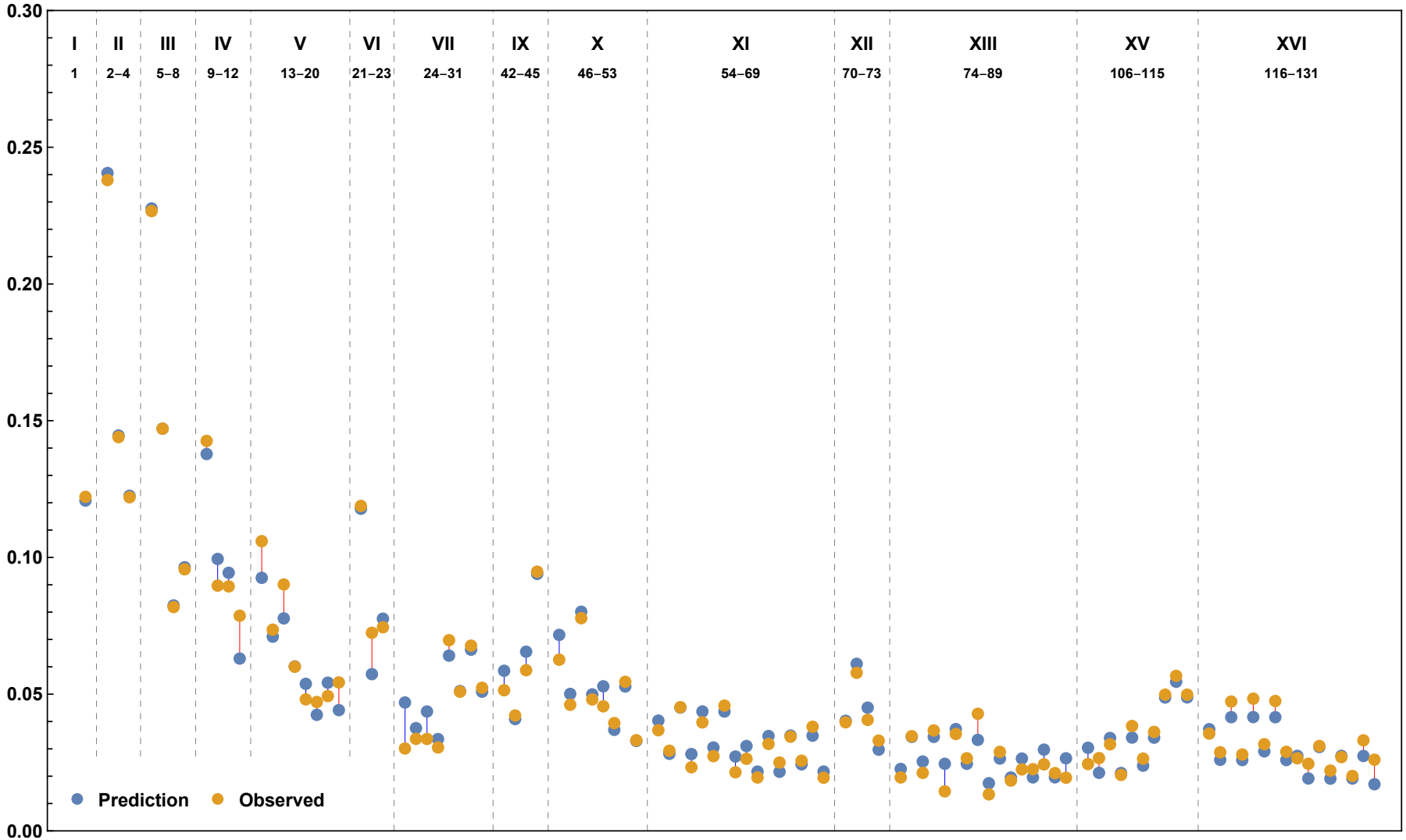
Educational attainment is key mediator for transmission of socio-economic advantages ("OED triangle", "Goldthorpe 2014). But do results generalize to other socioeconomic outcomes?

- *Ten-year* average of annual pre-tax **income**
- Measured around age 35 for children and around age 45 for parents
- 141 distinct moments, using 129 moments for calibration

Issues:

- Income correlations systematically lower for mixed or female pairs
- We do not model labor supply decisions, but model flexible enough to capture gender asymmetries

Figure: Sample and Predicted Moments (Income)



Income (Sweden): Results

Findings are qualitatively similar, but differ in magnitude:

Latent advantages more strongly transmitted than income itself, in all intergenerational, sibling and assortative processes:

- Father-son correlation in latent factor twice as large as in log income
- Sibling correlation in latent factor ≈ 0.8
- Spousal correlation in latent factor ≈ 0.65 (vs. ≈ 0.12 in log income)

However, the latent factors that determine educational attainment appear more strongly transmitted from one generation to the next than the latent factors that influence income. Back

Empirical Application: Spanish Census

Spanish Population Census 2001:

For the region of Cantabria we observe the full name of each person, and can use this information to identify kinship:

- Child generation born 1956-1976 (71,479 males, 68,830 females)
- A newborn in Spain receives two surnames, the first is the father's and the second the mother's (first) surname.
- Set of potential parents: couples born<1956, husband's and wife's surnames fit, age difference between parents and son>=16 years. Parents identified if only one couple in set (35% of cases).
- Siblings in child generation are identified.
- Set of potential siblings in parent generation: individuals sharing the same two surnames. Siblings identified if two individuals in set.
- Uncles nephews, and cousins are identified. Back

Table: Calibrated Parameters in Swedish Registers (Height)

Panel A: Intergenerational Processes

Parameters:

β^m	β^f	γ^m	γ^f		
0.000	0.000	1.000	1.000		
σ^2_{ym}	σ^2_{yf}	σ^2_{zm}	σ^2_{zf}	σ^2_{um}	σ^2_{uf}
1.000	1.000	0.731	0.731	0.163	0.237
α_{ym}	α_{yf}	α_{zm}	α_{zf}		
0.000	0.000	0.500	0.500		

Parent-child correlations in z:

Father-Son	Father-Dau	Mother-Son	Mother-Dau
0.605	0.605	0.605	0.605

Ancestor correlations in y and z:

	Father-Son	Grandf-...	GGrandf-...	GGGrandf-...
in y	0.470	0.285	0.172	0.104

in z						

Panel B: Sibling Processes

Parameters:

	σ^2_{xm}	σ^2_{xf}	σ_{xmxf}	σ^2_{em}	σ^2_{ef}	σ_{emef}
	0.107	0.032	0.000	0.000	0.066	0.035

Variance Shares:

in y	10.7%	3.2%	0.0%	0.0%	6.6%	3.5%
in z	-	-	-	0.0%	9.0%	4.8%

Sibling correlations in z:

Brothers	Sisters	Brother-Sister
0.605	0.695	0.653

0.605	0.695	0.653			
Panel C: Assortative Processes					
Parameters:					
r^{m}_{zz}	r^{m}_{zy}	r^{m}_{yz}	r^{m}_{yy}	$\sigma^{2}_{\omega m}$	$\sigma^{2}_{\epsilon m}$
0.000	0.210	0.000	0.287	0.687	0.917
r^{f}_{zz}	r^{f}_{zy}	r^{f}_{yz}	r^{f}_{yy}		
0.000	0.210	0.000	0.287		
Spousal correlations in y and z:					
ρ_{ymyf}	ρ_{zmzf}	ρ_{ymzf}	ρ_{zmyf}		
0.287	0.210	0.246	0.246		
Panel D: Variance Decomposition					
%	y	z	x	Cov(y,z)	
Male	0.000	0.731	0.107	0.000	
Female	0.000	0.731	0.032	0.000	

Panel A: Intergenerational Processes

Parameters:

β^m	β^f	γ^m	γ^f	
0.113	0.098	0.691	0.586	
σ^2_{ym}	σ^2_{yf}	σ^2_{zcm}	σ^2_{zcf}	σ^2_{zgm}
4.648	4.465	1.778	1.375	0.312
α_{ym}	α_{yf}	α_{zm}	α_{zf}	
0.447	0.000	0.574	0.657	

Within-person correlations in y and z:

	ρ_{yzc}	ρ_{yzg}	ρ_{zczg}
males	0.670	0.301	0.032
females	0.599	0.301	0.032

Parent-child correlations in z:

	Father-Son	Father-Dau	Mother-Son	Mother-Dau
in zc	0.578	0.579	0.537	0.509
in zg	0.512	0.512	0.512	0.512
in zc+zg	0.584	0.584	0.549	0.527

Ancestor correlations in y and z:

	Father-Son	Grandf-...	GGrandf-...	GGGrandf-...
in y	0.379	0.210	0.122	0.071
in zc+zg	0.584	0.342	0.200	0.117

Panel B: Sibling Processes

Parameters:

	σ^2_{xm}	σ^2_{xf}	σ_{xmxf}	σ^2_{em}	σ^2_{ef}	σ_{emef}
	0.118	0.174	0.000	0.679	0.729	0.651

Variance Shares:

in y	2.5%	3.9%	0.0%	14.6%	16.3%	14.3%
in z	-	-	-	38.2%	53.0%	41.6%

Sibling correlations in z:

	Brothers	Sisters	Brother-Sister
in zc	0.750	0.886	0.778
in zg	0.512	0.512	0.512

Panel C: Assortative Processes

Spousal correlations in y and z:

ρ_{ymyf}	ρ_{ymzcf}	ρ_{ymzgf}	ρ_{zcmyf}	ρ_{zcmzcf}	ρ_{zcmzgf}
0.491	0.518	0.096	0.548	0.703	0.079
ρ_{zgmyf}	ρ_{zgmzcf}	ρ_{zgmzgf}			
0.080	0.047	0.024			

Panel D: Variance Decomposition

%	y	z	zg	x
Male	0.009	0.383	0.067	0.025
Female	0.010	0.308	0.070	0.039

Back