IGE: The State of the Literature

Bradley Setzler

PhD Student, Department of Economics
Center for the Economics of Human Development
The University of Chicago
setzler@uchicago.edu

March 10, 2015

Table of Contents

Table of Contents

Definition of the IGE

- Notation:
  - Y^P is log parental lifetime earnings.
  - Y^C is log child adult lifetime earnings.
- The IGE is:

$$\beta \equiv \frac{\text{Cov}\left(Y^P, Y^C\right)}{\text{Var}\left(Y^P\right)}$$

- It is the probability limit of the ordinary least squares regression of Y^C on Y^P.
- The intergenerational correlation is:

$$\rho \equiv \text{Corr}\left(Y^P, Y^C\right) \equiv \frac{\text{SD}\left(Y^P\right)}{\text{SD}\left(Y^C\right)}\beta \iff \beta = \frac{\text{SD}\left(Y^C\right)}{\text{SD}\left(Y^P\right)}\rho$$

- Notice dependence of β on marginal distributions.

Questions the IGE Addresses

- The following are variations of the same question:
  - How similar are child and parent earnings?
  - How persistent is economic advantage across generations?
  - How economically immobile is society?
- β provides a univariate answer: $1 - \beta$ is earnings mobility.
- Low mobility means poor children are not rising out of poverty, which has negative normative connotations.
- At the same time, low mobility means successful families are remaining successful, which may have positive normative content (e.g., good parenting works).
- Related to other normatively-relevant measures (Gini).

Definition of the Gatsby Curve

- Suppose we know the IGE for J countries, so that we observe β_j, $j = 1, 2, \ldots, J$.
- Suppose we also know the Gini coefficient (γ), so that we observe γ_j, $j = 1, 2, \ldots, J$.
- Then, the Gatsby Curve is,

$$G \equiv \frac{\text{Cov}\left(\gamma_j, \beta_j\right)}{\text{Var}\left(\gamma_j\right)}$$

  or the linear regression of IGE on Gini.
- Provides a univariate characterization of the relationship between inequality and mobility.

Questions the IGE Cannot Address

- The following questions are not answered by β:
  - By what percentage would Y^C rise if Y^P rose by 1%? (the literal intergenerational earnings elasticity)
  - Why are child earnings (dis)similar to parent earnings?
  - What policy interventions would raise or lower β?
  - How would an increase in γ affect β (vice versa)?
- More generally, there are no clear policy implications of the IGE. What is the ideal IGE: 0.3? -0.3? 1.0? -1.0?
  - But the same can be said of all normatively-relevant measures, including inequality measures.
  - Regardless of what the "best" β is, we can agree that β is interesting and attempt to learn it.

Table of Contents

Solon (1999)

- Becker (1988) claimed that the IGE is around 0.2.
- In an influential handbook chapter, Solon (1999) surveys the literature and argues that the most credible estimates are around 0.4, matching his earlier (1992) estimate based on a classical measurement error correction.
- He concludes: *Learning that intergenerational earnings elasticities are larger than we used to think is a real step forward, but, as is so often the case in scholarly research, improving our answer to one question leads immediately to harder questions. Now that we know parental income is a fairly strong predictor of offspring's earnings, it becomes that much more important to find out which of the causal processes...are mainly responsible....*

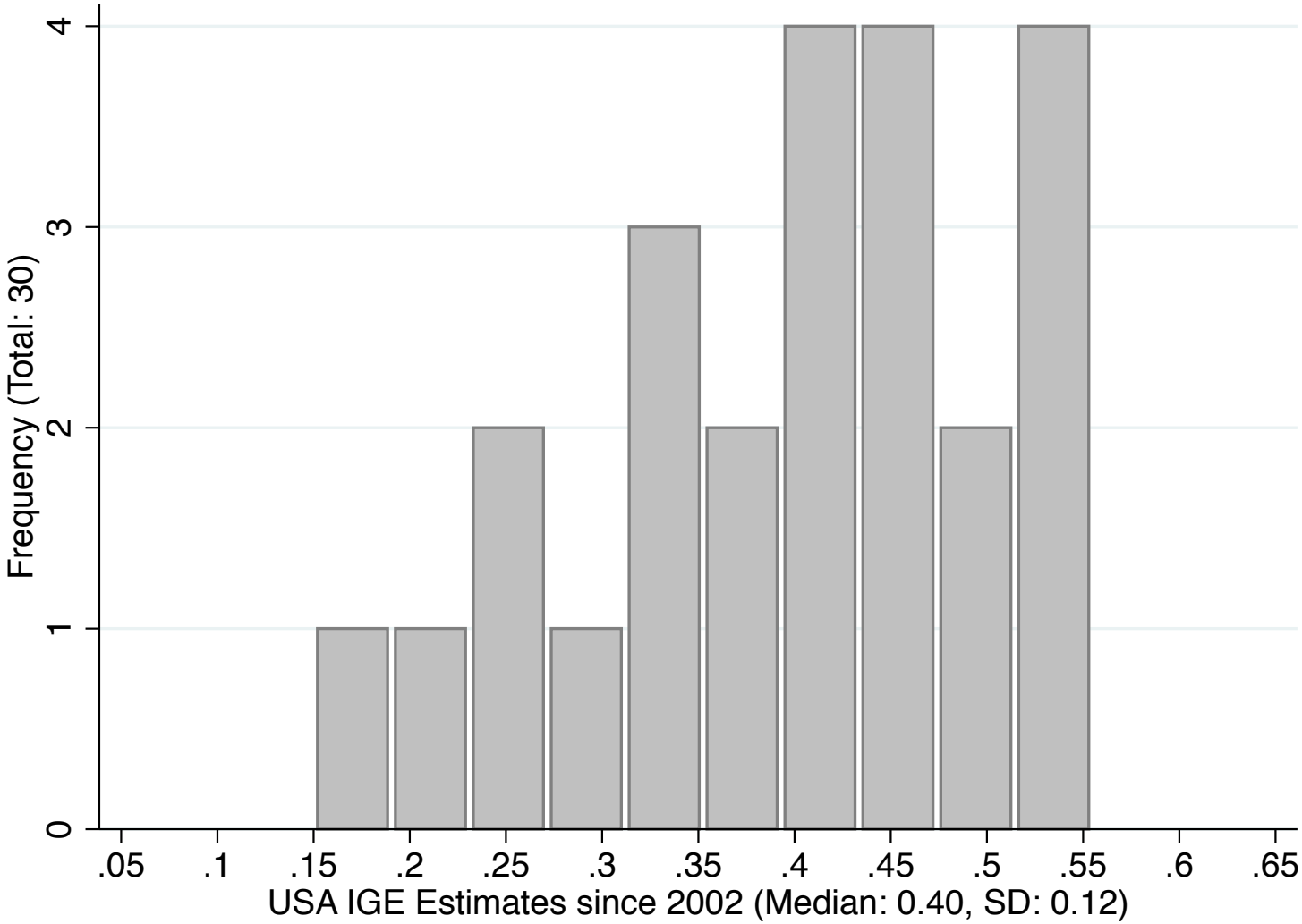
Mazumder (2005)

- Using administrative data, Mazumder (2005) finds that, as the length of the panel of earnings increases, so that parents and children are observed over additional earnings periods, the IGE estimate rises as high as 0.6.
- Mazumder finds similarly high estimates after using more sophisticated (ARMA) measurement error corrections.
- He concludes that the IGE is at least 0.5 in the US.

Source	Data	IGE	Avg
Solon (1992)	PSID	0.29-0.41	1Yr
Solon (1992)	PSID	0.41	5Yr[a]
Solon (1992)	PSID	0.53	IV[b]
Mazumder (2005)	SIPP	0.53	6Yr[c]
Mazumder (2005)	SIPP	0.61	15Yr[c]
Mazumder (2005)	NLSY	0.44	3Yr[d]
Chetty, et al (2014)	IRS	0.34	5/2Yr[e]

[a] Average of log-earnings of parent only.
[b] Father's years of education as instrument.
[c] Log of average earnings of child and parent.
[d] Log of average earnings of parent only.
[e] Log of 5 (2) year avg. of parent (child) earnings.

No Consensus: 30 IGE Estimates in 2002-2013



The Gatsby Curve by Corak (2013)

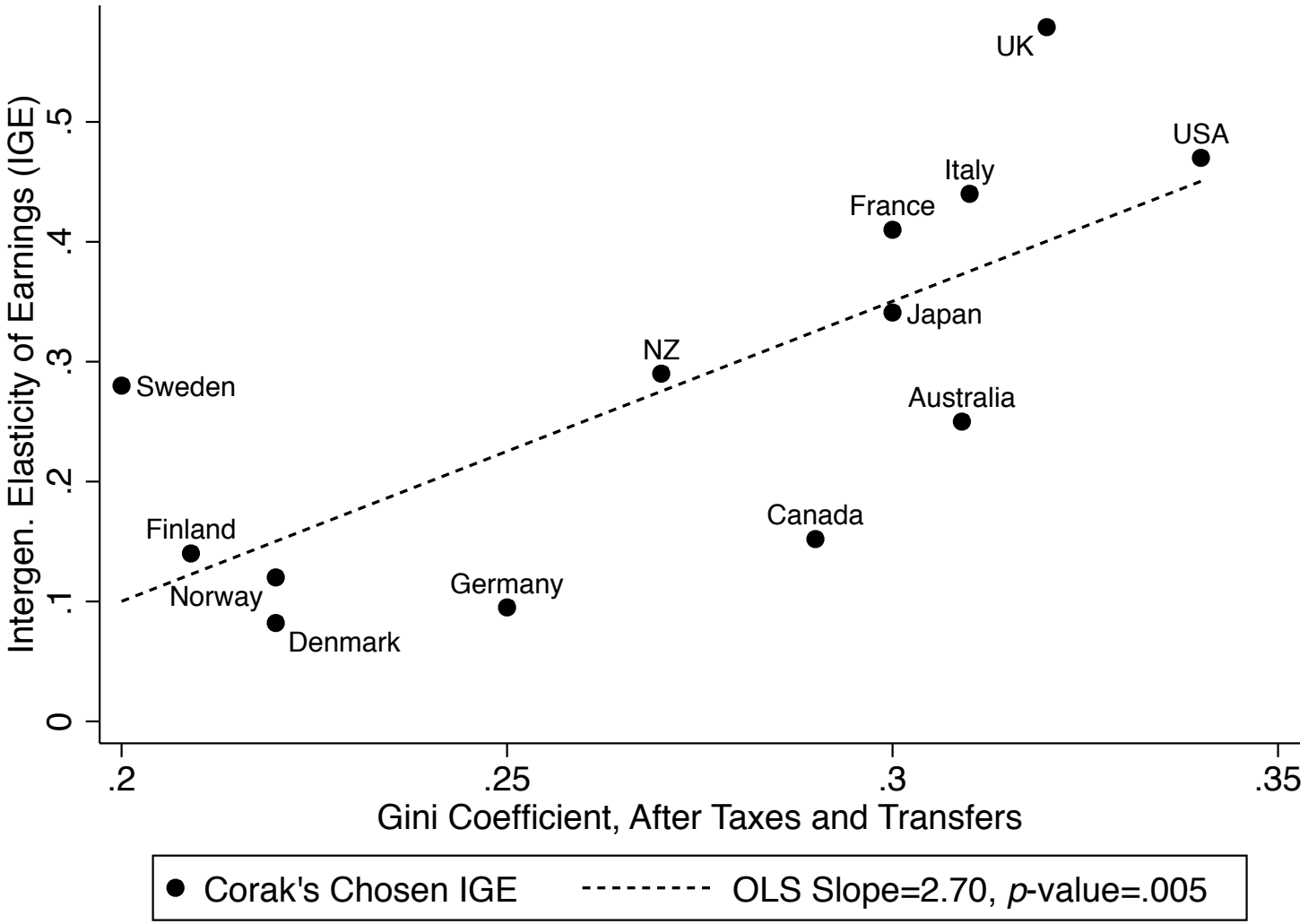


Table of Contents

Open Issues in the IGE Literature

- To understand why a consensus has not been reached regarding the IGE in the US, it is important to understand measurement error.
- If we observed lifetime earnings of parents and children, we could estimate β directly.
- Instead, we observe a set of parent earnings,

$$\left\{ \hat{Y}_{i,a}^{P} \right\}_{a \in \left\{ a_1, a_2, \ldots, a_{T_P} \right\}}.$$

and a set of child earnings,

$$\left\{ \hat{Y}_{i,a}^{C} \right\}_{a \in \left\{ a_1, a_2, \ldots, a_{T_C} \right\}}.$$

- How can we construct β from these measures?

Three Issues

- If $\hat{Y}_{i,a}^t \neq Y_{i,a}^t$, then there is *measurement error* in $\hat{Y}$ at age a.
- If the set of ages $\{a_1, a_2, \ldots, a_T\}$ is not representative, then, e.g., $\frac{1}{T}\sum Y_{i,a}^t \neq Y_i^t$, so the *age-earnings profile* needs to be accounted for in order to map $\left\{ Y_{i,a}^t \right\}$ into Y_i^t.
- Finally, if $Y_{i,a}^t \leq 0$ for any a, then $y_{i,a}^t \equiv \log Y_{i,a}^t$ is not real-valued, so non-positive earnings cause difficulty in constructing y_i^t, and the IGE is no longer defined.
- See my notes, "A Concise Overview of the IGE", for more details on these issues and literature review.

Issues Potentially Solved by Tax Data

- Tax data is less noisy than survey data. Why?
  - penalties for dishonesty,
  - rewards to difficult to sample populations (tax refunds),
  - reporting by third party (e.g., employer),
  - presumption of confidentiality.
- This solves the measurement error issue, at least as well as we can imagine solving it.
- Also, tax records cover many years. Why?
  - for most people, taxes must be paid each year,
  - tax records stay with the person even when changing jobs or location.
- This solves the age-earnings profile issue, at least for sufficiently old populations.
- However, does not solve the zero-earnings issue.

Table of Contents

Table of Contents

Population-level Earnings Data

- Previously, the only researcher access to government earnings data was the SIPP Census and SOI random sample of IRS data, representing around 0.1% of the population.
- For the first time, the US has allowed economists to access the national government's tax data.
  - Chetty, et al. partnered with IRS economists to create the IRS Databank.
  - Non-trivial 6-months of work to convert the existing tax records structure into a statistics-friendly platform, merge households/match spouses.
  - Result: a complete individual-level panel that contains one row per person per year for every person listed on a tax form during 1996-present.

IRS Databank Contents

- Included data:
  - All information from 1040 forms (income, taxes, many transfers, tuition/education expenses, retirement contributions, household structure, alimony, moving expenses, health expenses, teenage motherhood, geographic location).
  - All information from W-2 forms (reported wages from third party, the employer).
  - Other tax forms like 1098-T (college/educational information) and 1099 (which is like W-2 but for other sources of income).
- Updated as new tax data becomes available.
- Initial size 4TB (more than 5 billion rows of data), data managed with SAS.

Table of Contents

Short and Early Child Earnings Data

- Despite all of the available information, the IRS Databank is not perfect for studying the IGE.
- The panel is only around 16 years long.
- Children can only be matched to parents if they were dependents at some point during those 16 years.
- So to maximize the panel length for children, they need to be around 14-16 years old in the first year of tax records (1996). Resulting N around 10 million.
- So child earnings are observed up to at least age 30, but not past age 32.
- Can do very little controlling for cohort.
- 16 years for parents may be too short, and different parents are observed at different ages.

Table of Contents

Commuting Zone-level Data

- Although there is a great deal of information included in the IRS Databank regarding income, taxes, and transfers, there is very little non-financial information.
- Using the zip codes, Chetty et al. construct commuting zone (CZ) level covariates, which can be thought of as community characteristics. (There are 741 CZ's.)
- Includes information such as: racial composition, family structure composition, poverty and inequality measures, labor force characteristics, immigration, local and state government taxes and expenditures, cost of living, social capital, crime rates, and school characteristics.

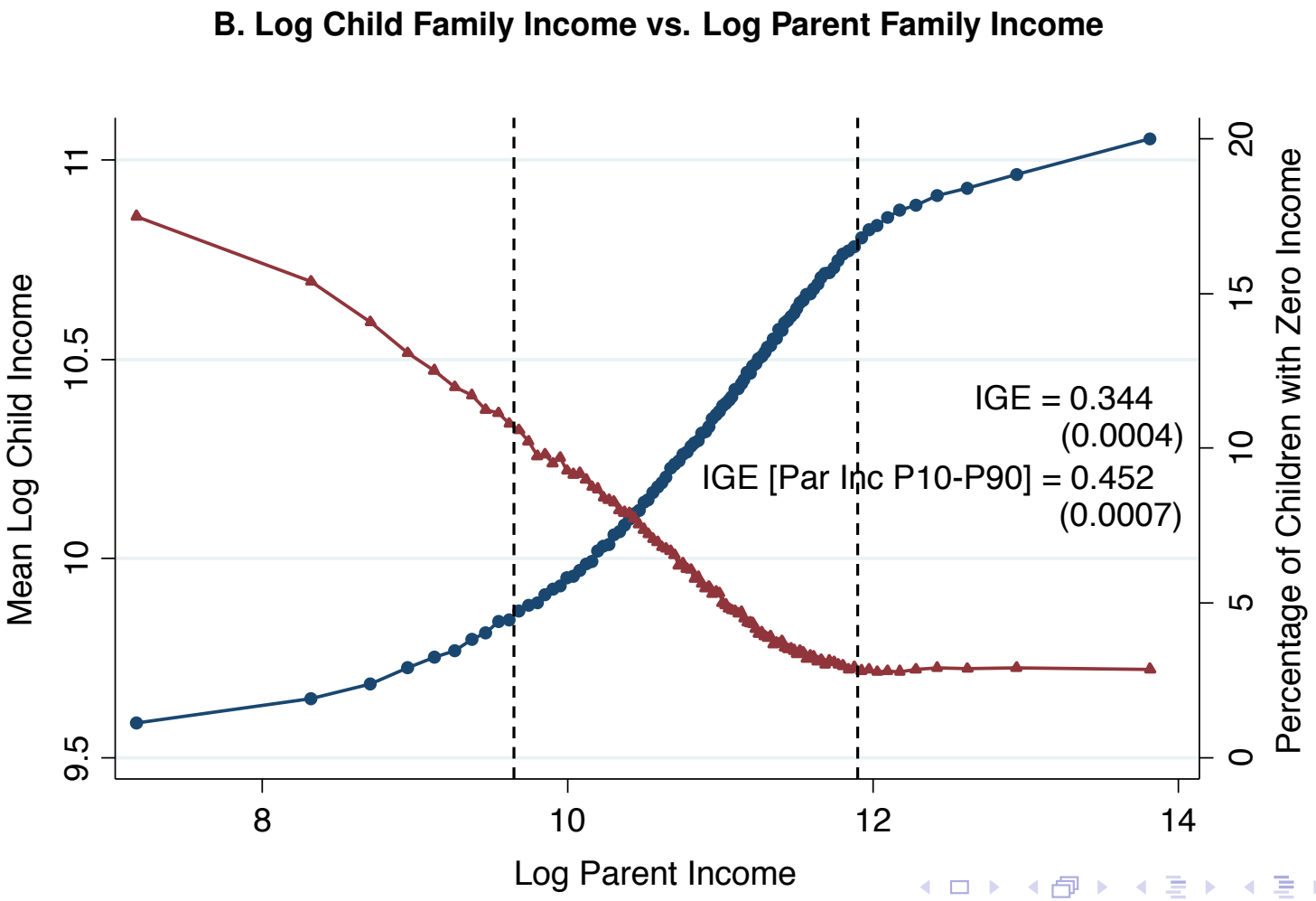
Table of Contents

Table of Contents

Non-linearity of the "IGE"

- Chetty, et al. find that the non-parametric regression of parent earnings on child earnings is non-linear, even in log units, suggesting that the IGE regression is misspecified.
  - Recall from our introduction: the IGE has always been known to be misspecified. It is meant only as a summary index of mobility, not as a true model.
  - This exercise is interesting, it is using goodness-of-fit to search for a better summary index of mobility.
- The relationship between log child earnings and log parent earnings is weaker at the bottom and top of the distribution, and this appears to be driven in part by the missing (zero) earnings of children near the bottom.
  - The trimmed middle 80% of the distribution has IGE around 0.45, while the untrimmed sample has IGE 0.34.

"IGE" Non-linearity



Rank-Rank IGE

- Consider transforming Y^C and Y^P into ranks (that is, marginal empirical CDF's); denote these by Q^C and Q^P.
  - They find that regressing Q^C on Q^P yields a linear relationship across the distribution.
  - This approximately linear relationship holds even locally among commuting zones.
  - And the approximate linearity holds in Canada and Denmark as well.
- But what happened to the zeros?
  - Apparently, all zeros were left as zeros in the regression.
  - Similarly, outliers at the top of the distribution become closer in distance.
  - Did they solve the problem by finding a better functional form, or did they just smooth the tails?

Rank-Rank Linearity across Countries

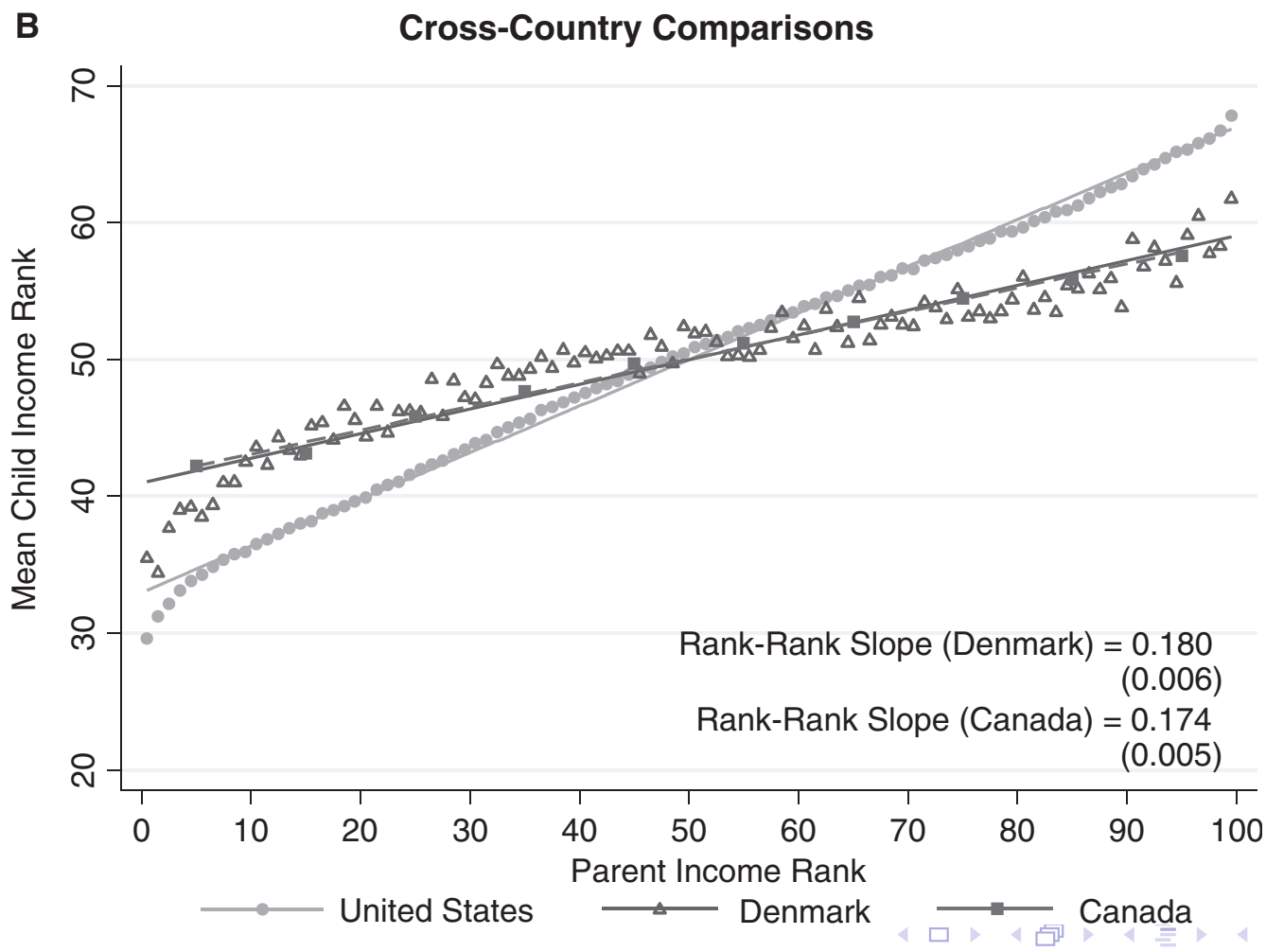


Table of Contents

Stability of Rank-Rank IGE after Age 30

- Recall: their data is limited in the oldest age of available child earnings. Does this affect the results?
  - To test, they use the randomized cross-sectional SOI samples from the IRS, which include much older children but also smaller samples and no panels.
  - They regress Q_a^C on $Q_{a'}^P$ for $22 \leq a \leq 40$, conditioning on a (but not a').
  - They find that the rank-rank slope peaks around age 31, then declines a bit.
  - They conclude that their data includes old enough children to trust the results.

Stability of Rank-Rank IGE after Age 30

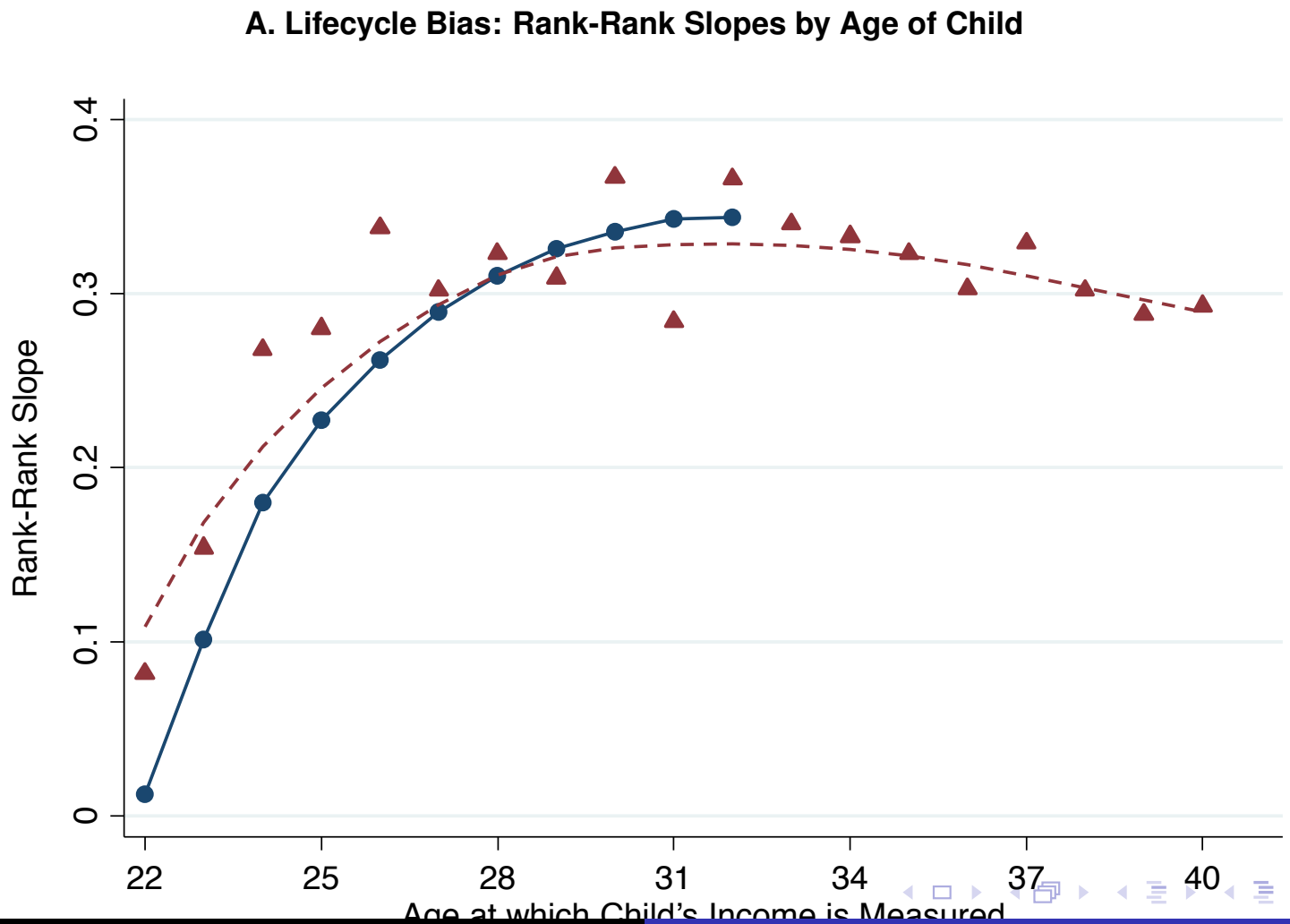


Table of Contents

Stability of Rank-Rank IGE across Cohorts

- Recall that they cannot study cohort effects in their main data, due to the limited availability of child cohorts.
- Instead, they investigate the rank-rank regression across cohorts of children born in the 1970's through early 1980's in the cross-sectional SOI samples.
- They find that the rank-rank IGE is about the same across those years (around 0.3), and that the rank-rank regression provides excellent fit in all three cases.

Stability of Rank-Rank IGE across Cohorts

Figure 1. Child Income Rank vs. Parent Income Rank by Birth Cohort

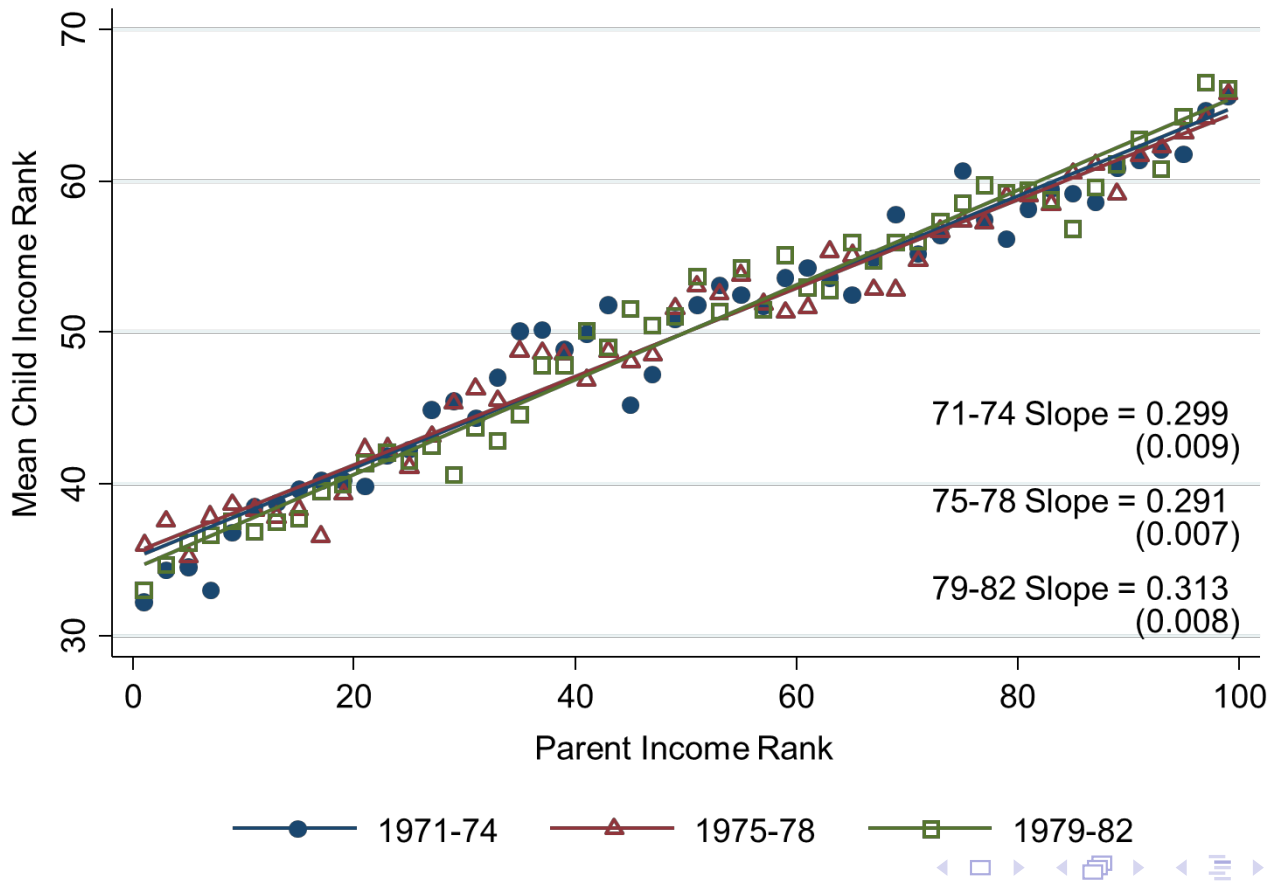


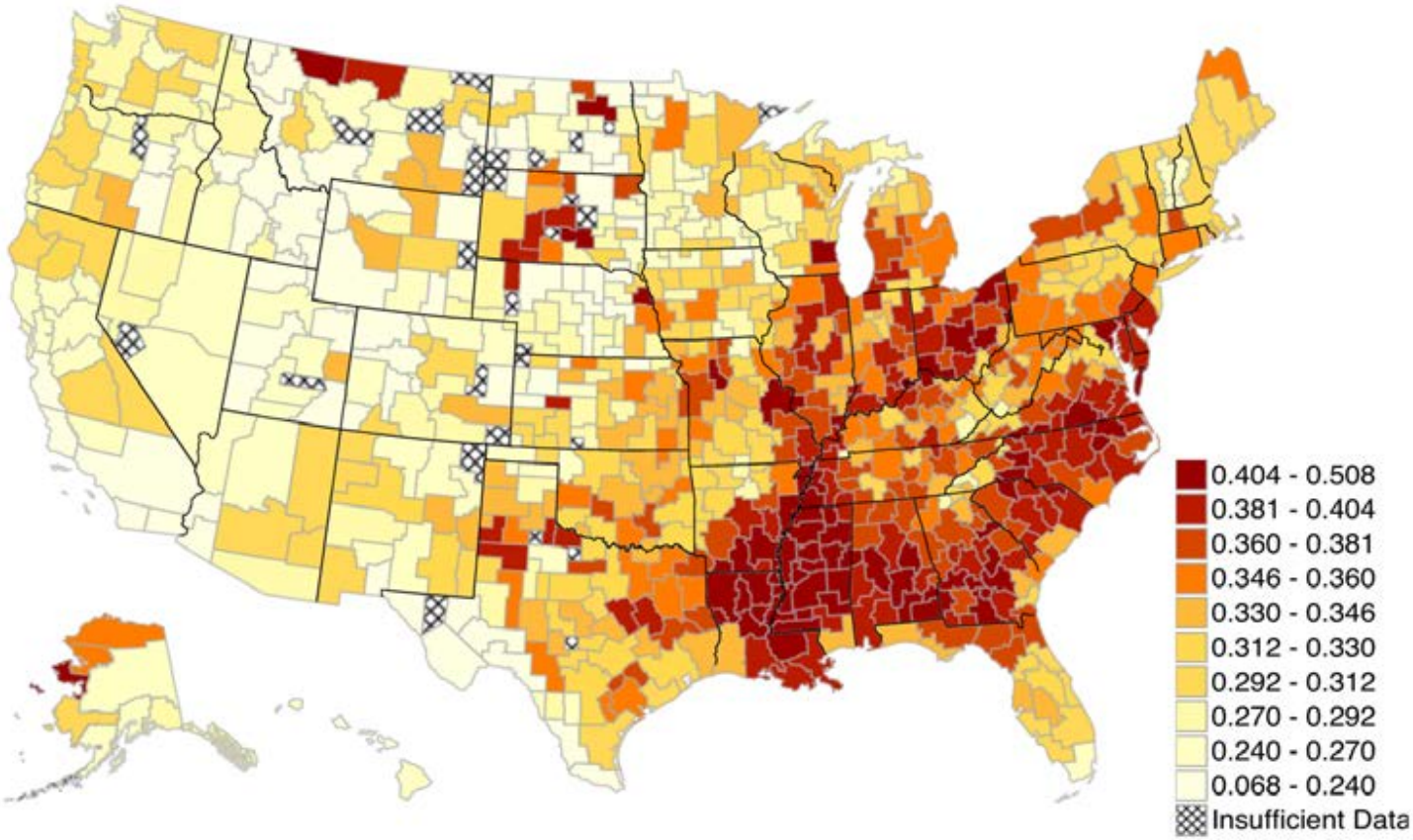
Table of Contents

Commuting Zone-level Correlates of Rank-Rank IGE

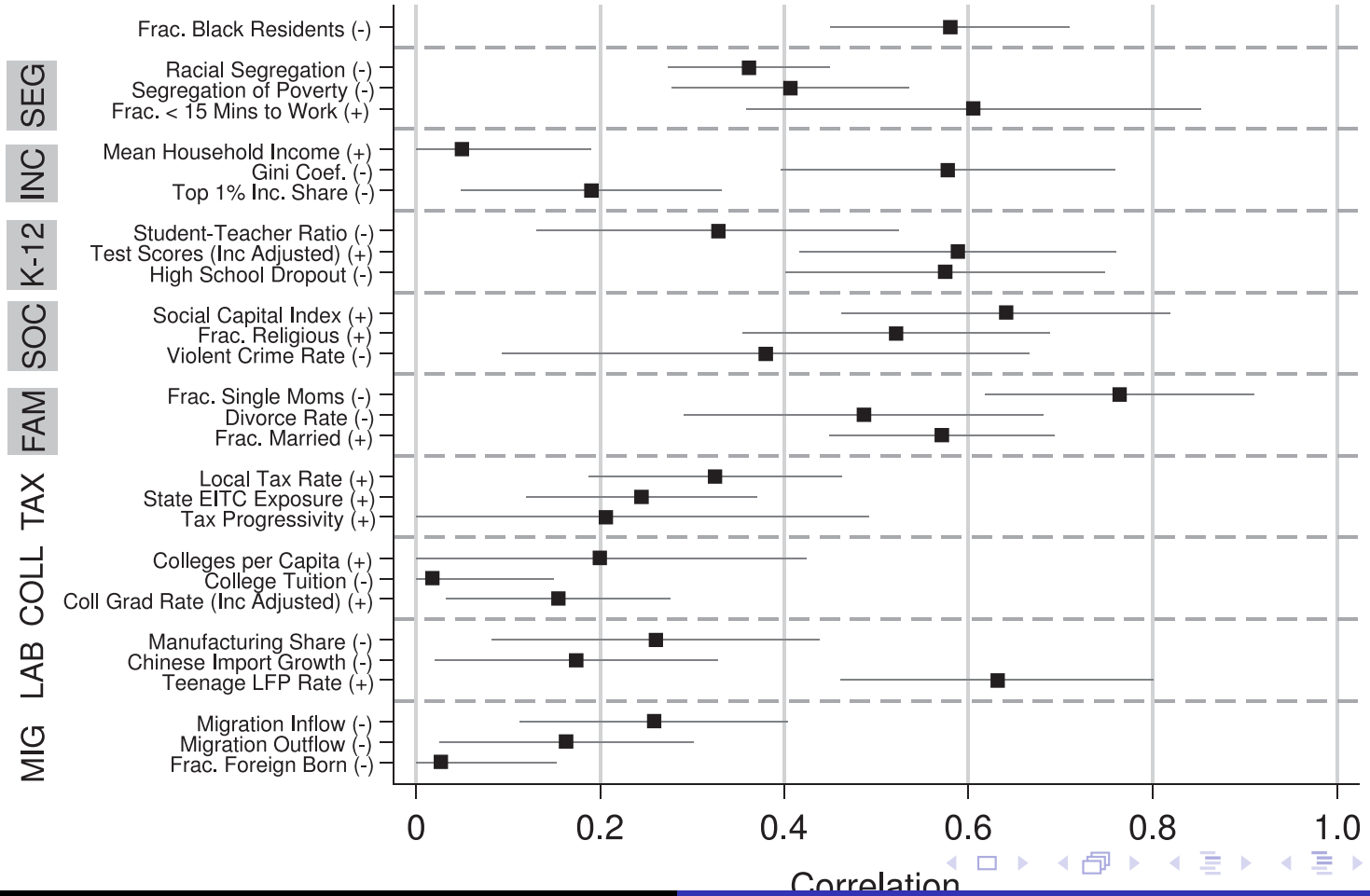
- The most discussed empirical result from the paper is the geographic heat map showing low rank-rank mobility in the American South and Midwest while the West Coast and Northeast have higher mobility.
- They find the IGE is positively associated with:
  - residential segregation by race or poverty status,
  - inequality (Gini coefficient, fraction in top 1%),
  - high school drop out rate,
  - the high school student-to-teacher ratio,
  - the inflow of migrants,
  - the violent crime rate, and,
  - the fraction of mothers who are single or divorced.

Geographic Correlates of Rank-Rank IGE

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



Other Correlates of Rank-Rank IGE



Conclusions

- There is no consensus on the IGE in the US, due to measurement error, age-earnings profiles, and zero-earnings individuals.
- Tax data could potentially solve the first two problems.
- Chetty, et al. are the first to use population administrative data in the US to study intergenerational mobility.
- They present interesting findings on intergenerational mobility and its correlates.
- However, the open issues are not completely resolved, more work to be done in this literature.

Bibliography

- Raj Chetty, Nathaniel Hendren, Patrick Kline, and Emmanuel Saez. Where is the land of Opportunity? The Geography of Intergenerational Mobility in the United States. The Quarterly Journal of Economics (2014), 129(4): 1553-1623.
- Raj Chetty, Nathaniel Hendren, Patrick Kline, Emmanuel Saez, and Nicholas Turner. Is the United States Still the land of Opportunity? Recent Trends in Intergenerational Mobility. The American Economic Review (2014), 104(5), 141-147.