The Standardized World Income Inequality Database

Frederick Solt
University of Iowa

18 September 2013

The Problem: Quantity vs. Quality

- Goal: To make valid comparisons of levels and trends in income inequality across countries and over time
- Need: Comparable data for many countries and years
- Issue: Most of the available data is incomparable
  1. Income definition (e.g., market income, consumption)
  2. Unit of analysis (e.g., household, adult equivalent, individual)
  3. Harmonization (e.g., treatment of non-monetary income)
  4. Coverage (entire population vs., e.g., major cities, rural, adults)

The Problem: Quantity vs. Quality

- Goal: To make valid comparisons of levels and trends in income inequality across countries and over time
- Need: Comparable data for many countries and years
- Issue: Most of the available data is incomparable
  1. Income definition (e.g., market income, consumption)
  2. Unit of analysis (e.g., household, adult equivalent, individual)
  3. Harmonization (e.g., treatment of non-monetary income)
  4. Coverage (entire population vs., e.g., major cities, rural, adults)

The Problem: Quantity vs. Quality

- Goal: To make valid comparisons of levels and trends in income inequality across countries and over time
- Need: Comparable data for many countries and years
- Issue: Most of the available data is incomparable
  1. Income definition (e.g., market income, consumption)
  2. Unit of analysis (e.g., household, adult equivalent, individual)
  3. Harmonization (e.g., treatment of non-monetary income)
  4. Coverage (entire population vs., e.g., major cities, rural, adults)

The Problem: Quantity vs. Quality

- Goal: To make valid comparisons of levels and trends in income inequality across countries and over time
- Need: Comparable data for many countries and years
- Issue: Most of the available data is incomparable
  1. Income definition (e.g., market income, consumption)
  2. Unit of analysis (e.g., household, adult equivalent, individual)
  3. Harmonization (e.g., treatment of non-monetary income)
  4. Coverage (entire population vs., e.g., major cities, rural, adults)

The Problem: Quantity vs. Quality

- Goal: To make valid comparisons of levels and trends in income inequality across countries and over time
- Need: Comparable data for many countries and years
- Issue: Most of the available data is incomparable
  1. Income definition (e.g., market income, consumption)
  2. Unit of analysis (e.g., household, adult equivalent, individual)
  3. Harmonization (e.g., treatment of non-monetary income)
  4. Coverage (entire population vs., e.g., major cities, rural, adults)

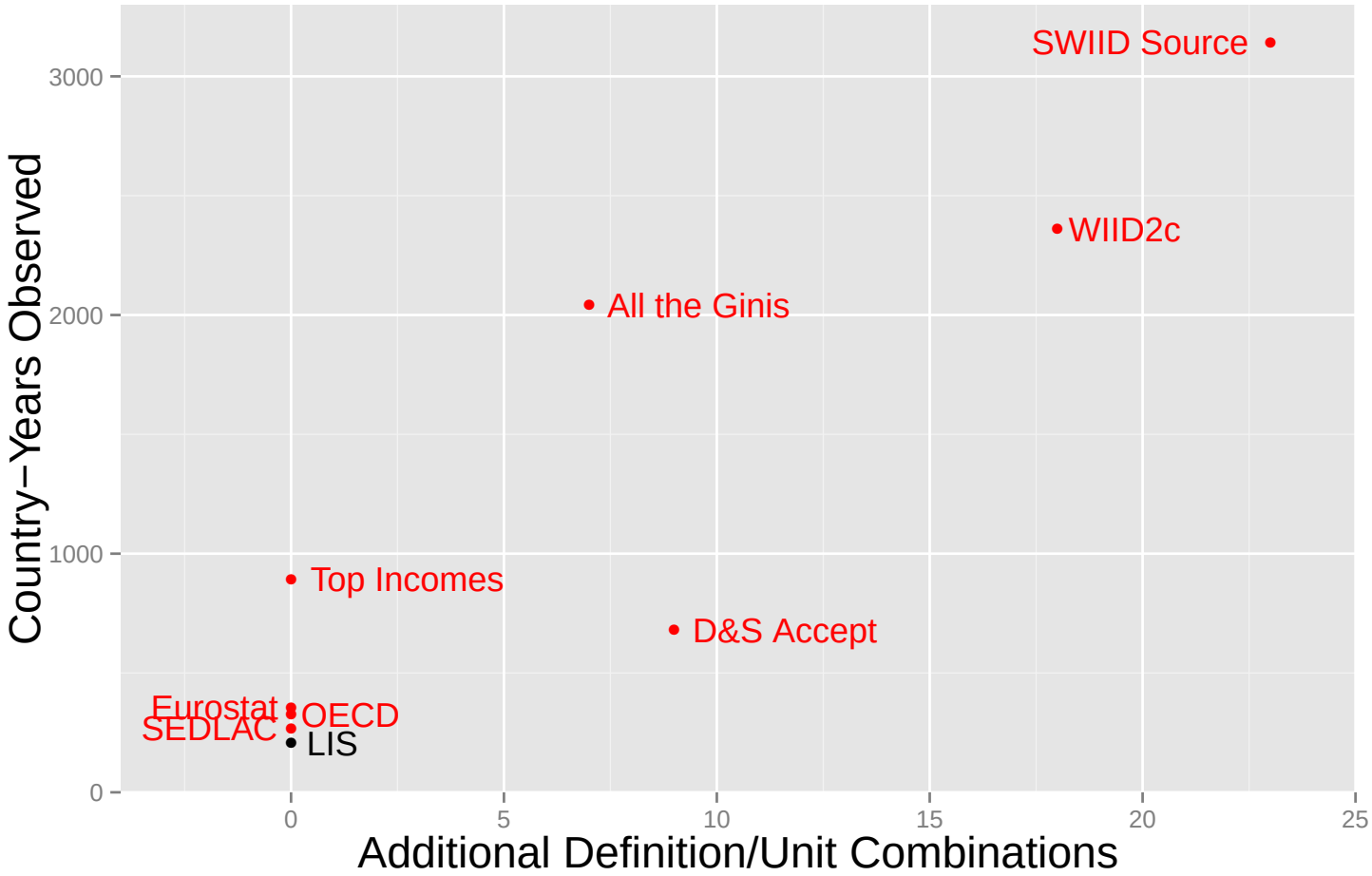
The Problem: Quantity vs. Quality

- Goal: To make valid comparisons of levels and trends in income inequality across countries and over time
- Need: Comparable data for many countries and years
- Issue: Most of the available data is incomparable
  1. Income definition (e.g., market income, consumption)
  2. Unit of analysis (e.g., household, adult equivalent, individual)
  3. Harmonization (e.g., treatment of non-monetary income)
  4. Coverage (entire population vs., e.g., major cities, rural, adults)

The Problem: Quantity vs. Quality

- Goal: To make valid comparisons of levels and trends in income inequality across countries and over time
- Need: Comparable data for many countries and years
- Issue: Most of the available data is incomparable
  1. Income definition (e.g., market income, consumption)
  2. Unit of analysis (e.g., household, adult equivalent, individual)
  3. Harmonization (e.g., treatment of non-monetary income)
  4. Coverage (entire population vs., e.g., major cities, rural, adults)

A Tradeoff



- More observations, more incomparability

Two (Not-So-Attractive) Options

- Comparable Data for a Few Countries and Years
  - Luxembourg Income Study (208)
  - *All the Ginis* (590)
  - SWIID Source Data (962)
- But this entails giving up on many comparisons and throwing away a lot of data
- Incomparable Data for Many Countries and Years
  - Just ignore incomparability, or
  - Make a fixed adjustment for different definitions and units
- But crudely reducing these differences to constants means any comparisons are extremely dubious

Two (Not-So-Attractive) Options

- Comparable Data for a Few Countries and Years
  - Luxembourg Income Study (208)
  - *All the Ginis* (590)
  - SWIID Source Data (962)
- But this entails giving up on many comparisons and throwing away a lot of data
- Incomparable Data for Many Countries and Years
  - Just ignore incomparability, or
  - Make a fixed adjustment for different definitions and units
- But crudely reducing these differences to constants means any comparisons are extremely dubious

Two (Not-So-Attractive) Options

- Comparable Data for a Few Countries and Years
  - Luxembourg Income Study (208)
  - *All the Ginis* (590)
  - SWIID Source Data (962)
- But this entails giving up on many comparisons and throwing away a lot of data

- Incomparable Data for Many Countries and Years
  - Just ignore incomparability, or
  - Make a fixed adjustment for different definitions and units
- But crudely reducing these differences to constants means any comparisons are extremely dubious

Two (Not-So-Attractive) Options

- Comparable Data for a Few Countries and Years
  - Luxembourg Income Study (208)
  - *All the Ginis* (590)
  - SWIID Source Data (962)
- But this entails giving up on many comparisons and throwing away a lot of data

- Incomparable Data for Many Countries and Years
  - Just ignore incomparability, or
  - Make a fixed adjustment for different definitions and units
- But crudely reducing these differences to constants means any comparisons are extremely dubious

Two (Not-So-Attractive) Options

- Comparable Data for a Few Countries and Years
  - Luxembourg Income Study (208)
  - *All the Ginis* (590)
  - SWIID Source Data (962)
- But this entails giving up on many comparisons and throwing away a lot of data

- Incomparable Data for Many Countries and Years
  - Just ignore incomparability, or
  - Make a fixed adjustment for different definitions and units
- But crudely reducing these differences to constants means any comparisons are extremely dubious

Two (Not-So-Attractive) Options

- Comparable Data for a Few Countries and Years
  - Luxembourg Income Study (208)
  - *All the Ginis* (590)
  - SWIID Source Data (962)
- But this entails giving up on many comparisons and throwing away a lot of data

- Incomparable Data for Many Countries and Years
  - Just ignore incomparability, or
  - Make a fixed adjustment for different definitions and units
- But crudely reducing these differences to constants means any comparisons are extremely dubious

Two (Not-So-Attractive) Options

- Comparable Data for a Few Countries and Years
  - Luxembourg Income Study (208)
  - *All the Ginis* (590)
  - SWIID Source Data (962)
- But this entails giving up on many comparisons and throwing away a lot of data

- Incomparable Data for Many Countries and Years
  - Just ignore incomparability, or
  - Make a fixed adjustment for different definitions and units
- But crudely reducing these differences to constants means any comparisons are extremely dubious

Two (Not-So-Attractive) Options

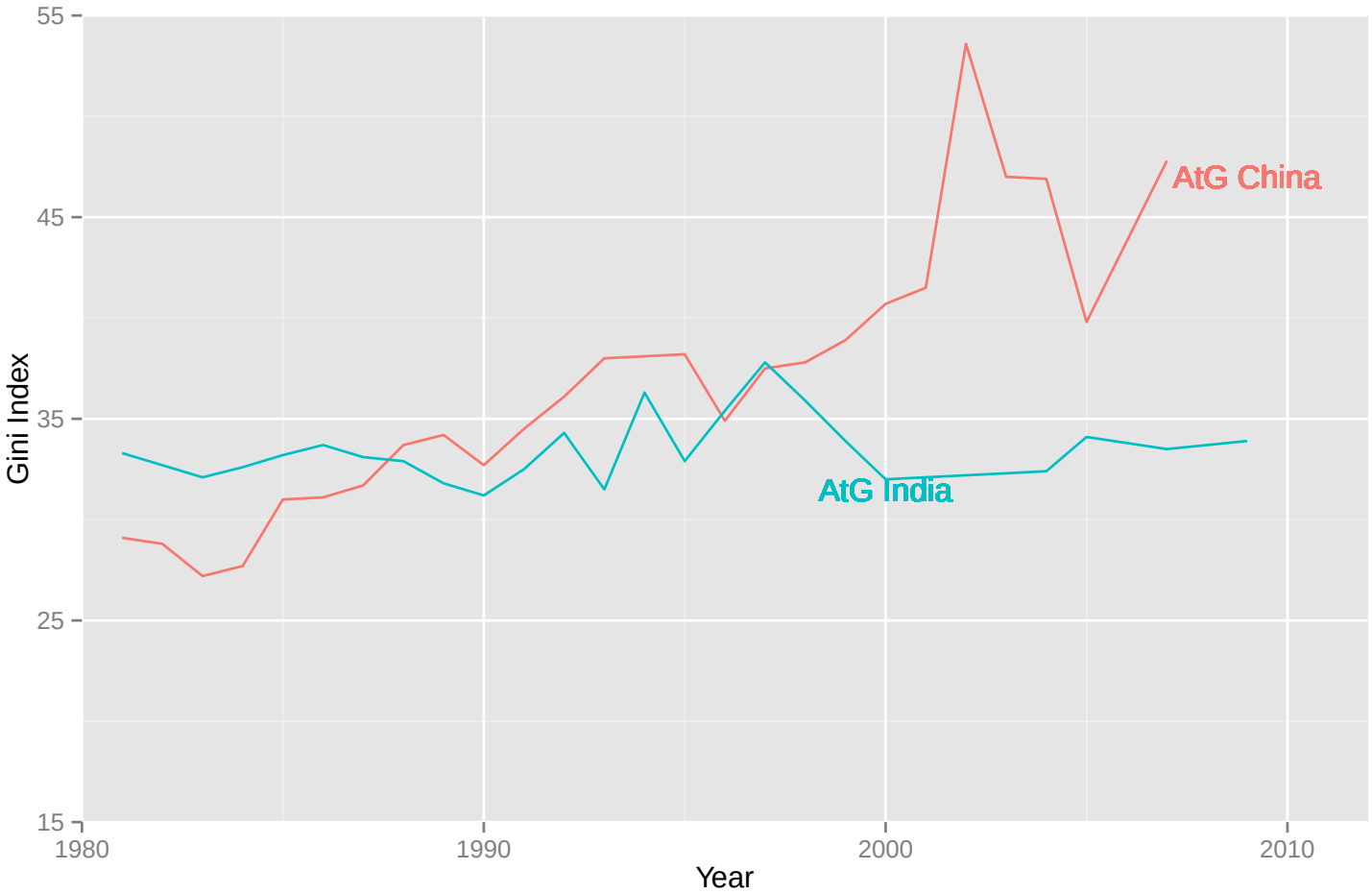
- Comparable Data for a Few Countries and Years
  - Luxembourg Income Study (208)
  - *All the Ginis* (590)
  - SWIID Source Data (962)
- But this entails giving up on many comparisons and throwing away a lot of data

- Incomparable Data for Many Countries and Years
  - Just ignore incomparability, or
  - Make a fixed adjustment for different definitions and units
- But crudely reducing these differences to constants means any comparisons are extremely dubious

Two (Not-So-Attractive) Options

- Comparable Data for a Few Countries and Years
  - Luxembourg Income Study (208)
  - *All the Ginis* (590)
  - SWIID Source Data (962)
- But this entails giving up on many comparisons and throwing away a lot of data
- Incomparable Data for Many Countries and Years
  - Just ignore incomparability, or
  - Make a fixed adjustment for different definitions and units
- But crudely reducing these differences to constants means any comparisons are extremely dubious

All the Ginis



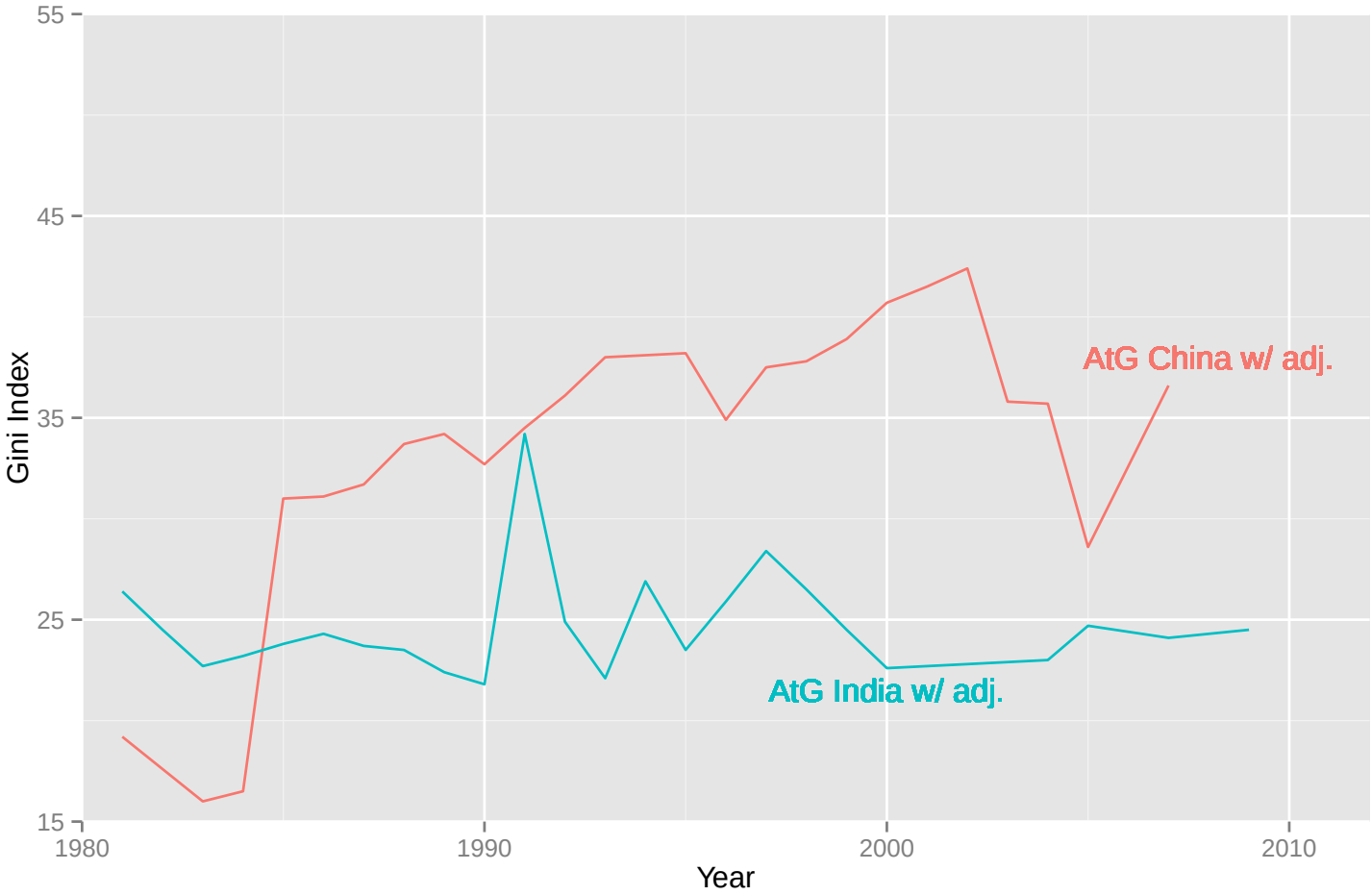
2013 *All the Ginis*

All the Ginis with Adjustments

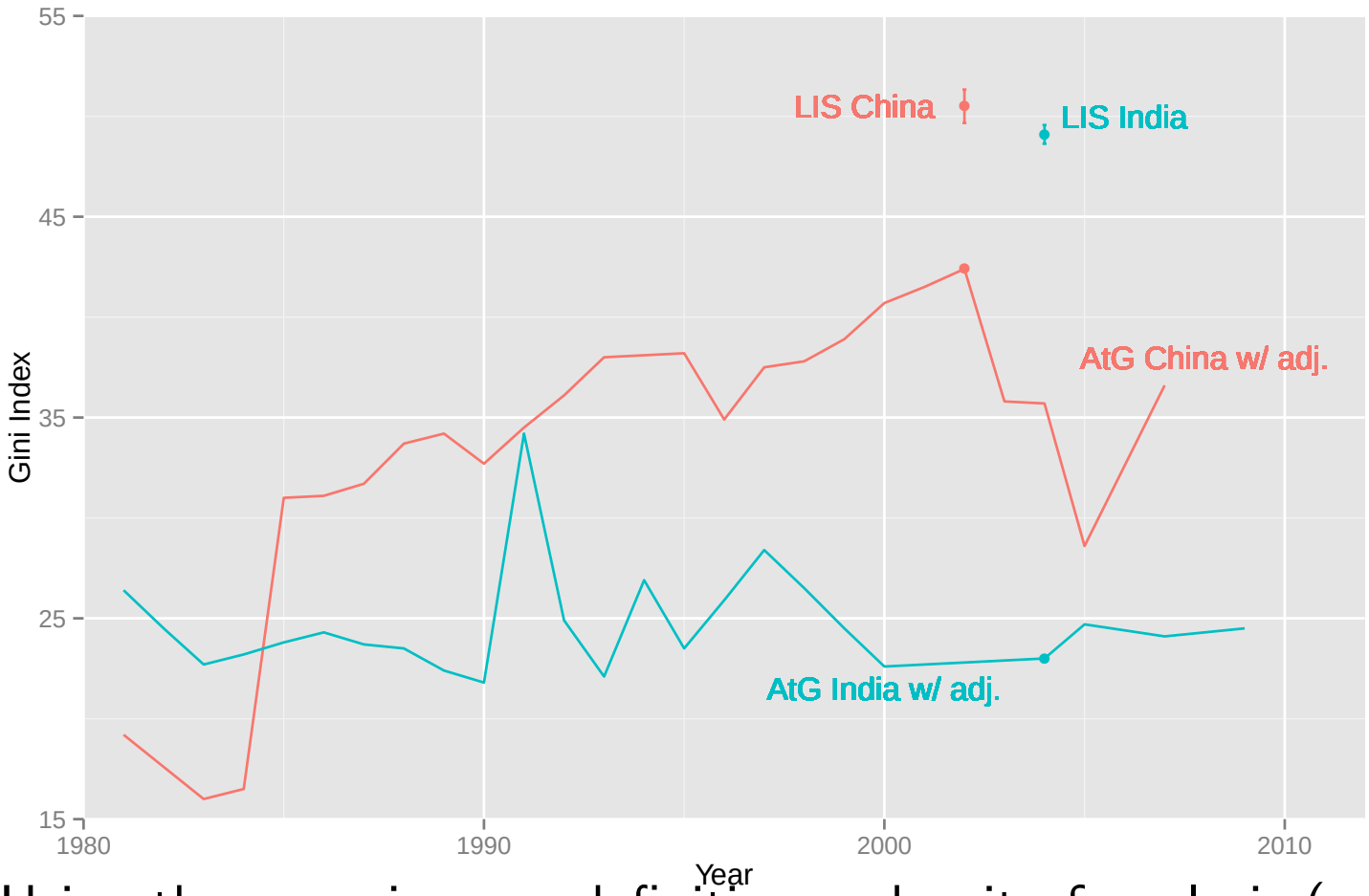
```
. reg Giniall Di Dhh Dg

      Source |       SS       df       MS              Number of obs =    1895
-------------+------------------------------           F(  3,  1891) =  255.12
       Model |  58090.3265     3  19363.4422            Prob > F      =  0.0000
    Residual |  143524.027  1891  75.8984808            R-squared     =  0.2881
-------------+------------------------------           Adj R-squared =  0.2870
       Total |  201614.354  1894  106.448972            Root MSE      =   8.712

------------------------------------------------------------------------------
     Giniall |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          Di |    1.737537   .4671214     3.72   0.000     .8214097    2.653665
         Dhh |   -1.469912    .513049    -2.87   0.004    -2.476113   -.4637103
          Dg |    11.15089   .4060547    27.46   0.000     10.35453    11.94725
       _cons |    32.17105   .4423607    72.73   0.000     31.30349    33.03862
------------------------------------------------------------------------------

. gen g = Giniall-1.737537*(1-Di)+1.469912*Dhh-11.15089*Dg
(6942 missing values generated)
```

- Use included dummy variables to calculate adjustments per Milanovic (2013, 8) . . .

All the Ginis with Adjustments



All the Ginis and the LIS



- Using the same income definition and unit of analysis (and harmonizing) can make a big difference

SWIID: The Logic

- How to maximize comparability for the widest possible coverage?
- This is a missing-data problem, fit for multiple imputation:
  - LIS gives high comparability, but with ~98% missing country-years
  - Other data informs imputations of these missing observations
  - 'Adjustment' modeled as varying across countries and over time, relying as much as possible on proximate years in the same country
  - Error terms in the models–the residual incomparability–is incorporated as uncertainty

SWIID: The Logic

- How to maximize comparability for the widest possible coverage?
- This is a missing-data problem, fit for multiple imputation:
  - LIS gives high comparability, but with ~98% missing country-years
  - Other data informs imputations of these missing observations
  - 'Adjustment' modeled as varying across countries and over time, relying as much as possible on proximate years in the same country
  - Error terms in the models–the residual incomparability–is incorporated as uncertainty

SWIID: The Logic

- How to maximize comparability for the widest possible coverage?
- This is a missing-data problem, fit for multiple imputation:
  - LIS gives high comparability, but with ~98% missing country-years
  - Other data informs imputations of these missing observations
  - 'Adjustment' modeled as varying across countries and over time, relying as much as possible on proximate years in the same country
  - Error terms in the models–the residual incomparability–is incorporated as uncertainty

SWIID: The Logic

- How to maximize comparability for the widest possible coverage?
- This is a missing-data problem, fit for multiple imputation:
  - LIS gives high comparability, but with ~98% missing country-years
  - Other data informs imputations of these missing observations
  - 'Adjustment' modeled as varying across countries and over time, relying as much as possible on proximate years in the same country
  - Error terms in the models–the residual incomparability–is incorporated as uncertainty

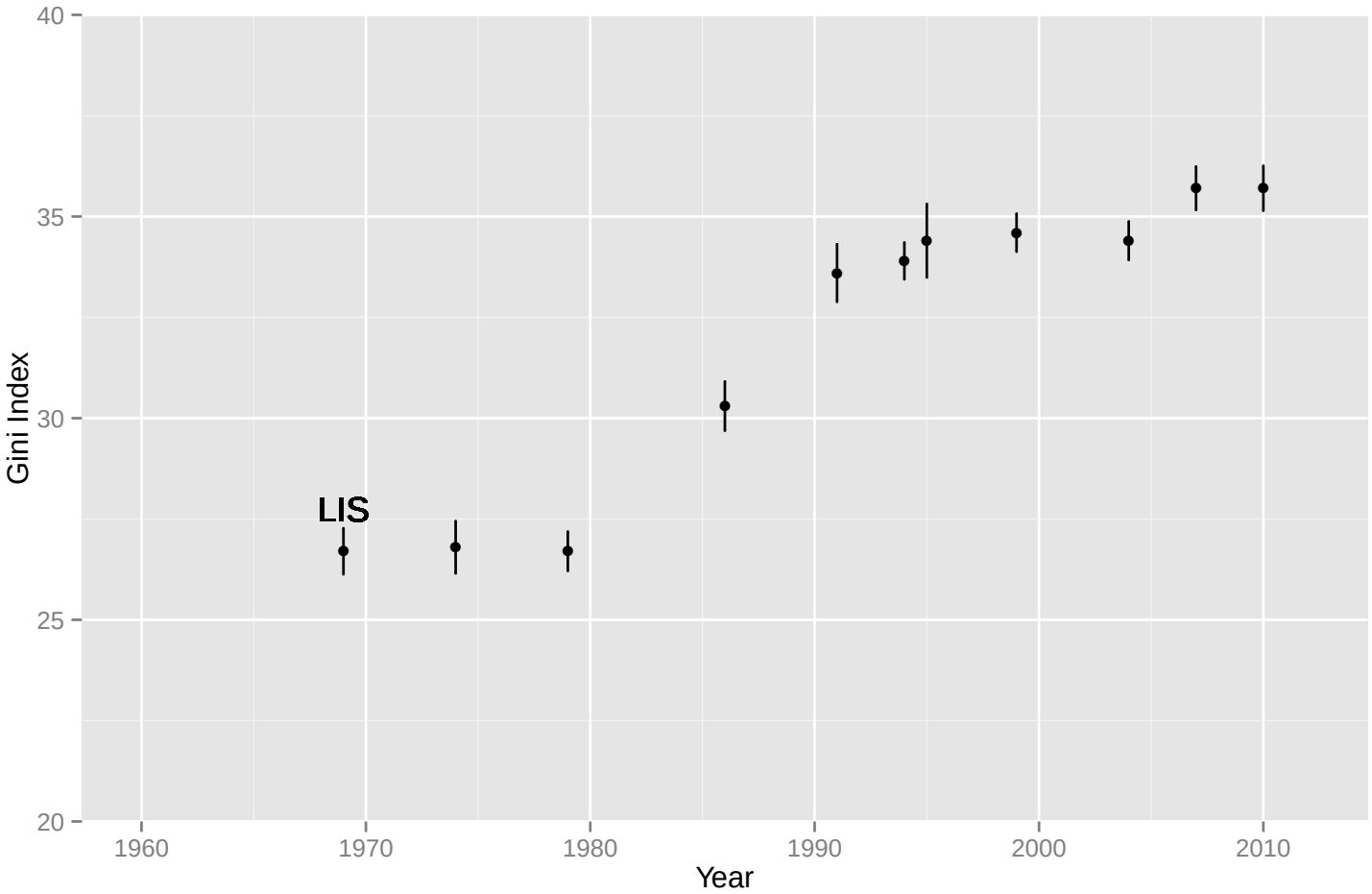
SWIID: The Logic

- How to maximize comparability for the widest possible coverage?
- This is a missing-data problem, fit for multiple imputation:
  - LIS gives high comparability, but with ~98% missing country-years
  - Other data informs imputations of these missing observations
  - 'Adjustment' modeled as varying across countries and over time, relying as much as possible on proximate years in the same country
  - Error terms in the models–the residual incomparability–is incorporated as uncertainty

SWIID: The Logic

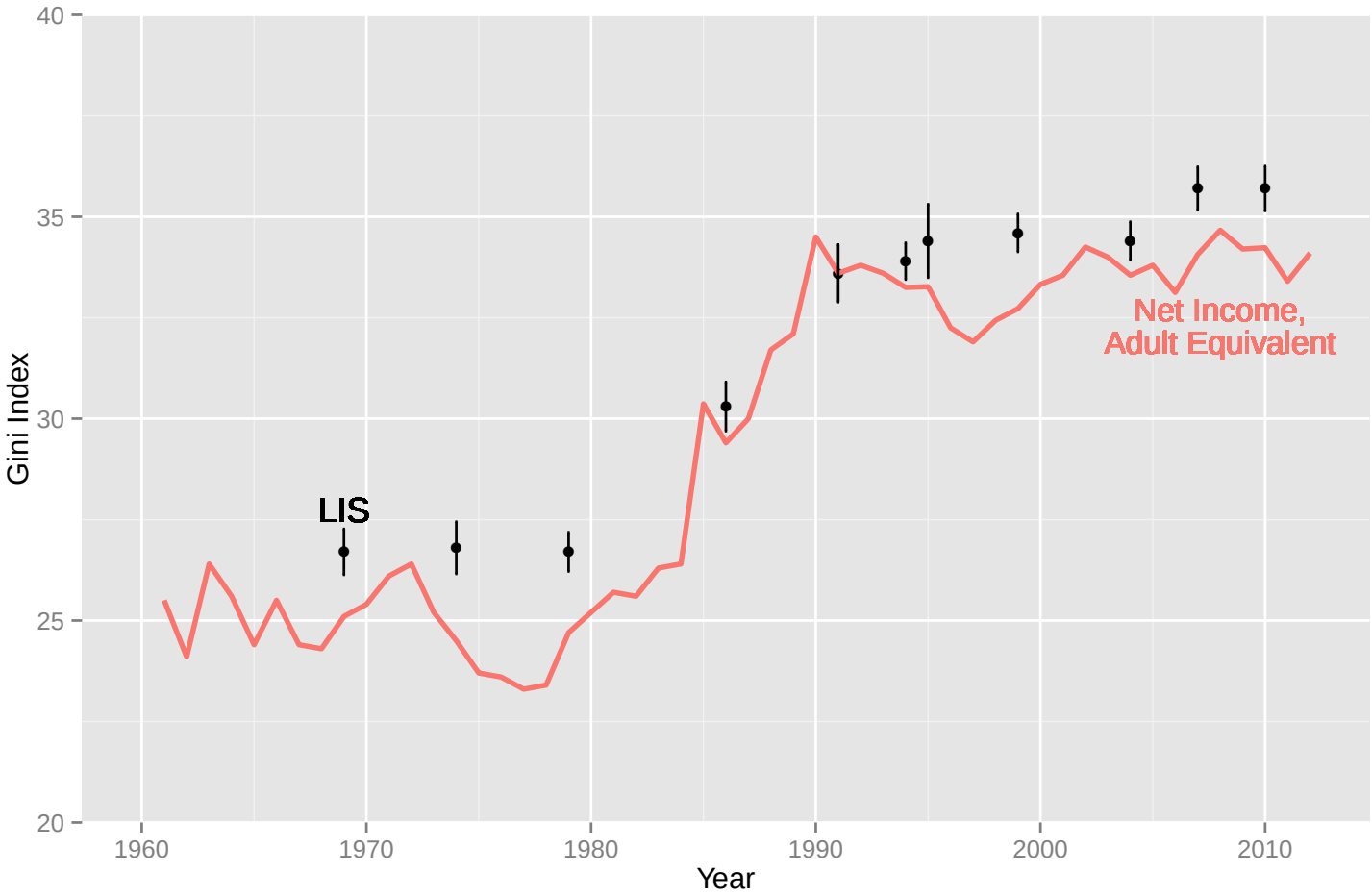
- How to maximize comparability for the widest possible coverage?
- This is a missing-data problem, fit for multiple imputation:
  - LIS gives high comparability, but with ~98% missing country-years
  - Other data informs imputations of these missing observations
  - 'Adjustment' modeled as varying across countries and over time, relying as much as possible on proximate years in the same country
  - Error terms in the models–the residual incomparability–is incorporated as uncertainty

An Illustration: Inequality in the United Kingdom



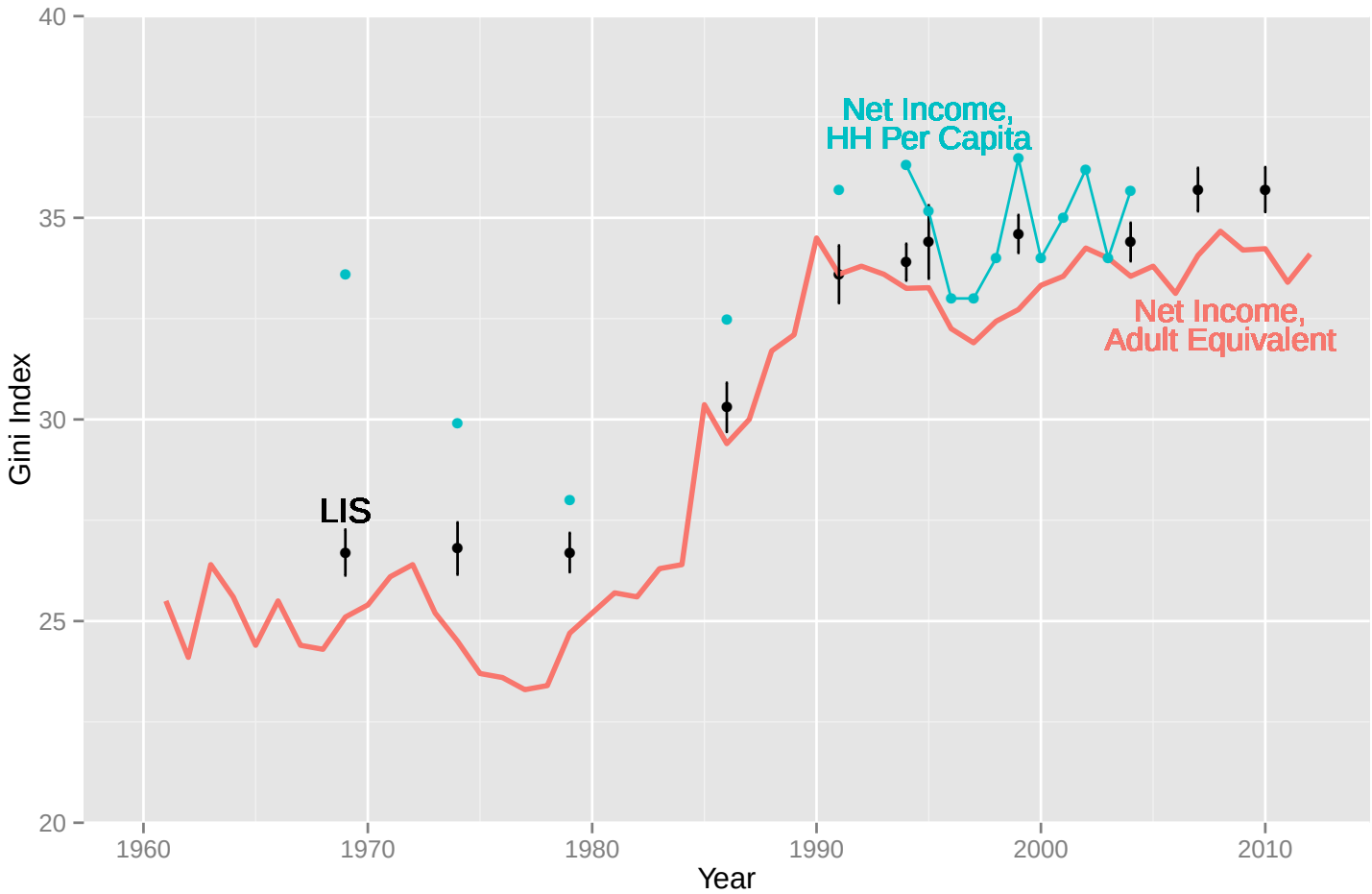
Luxembourg Income Study

An Illustration: Inequality in the United Kingdom



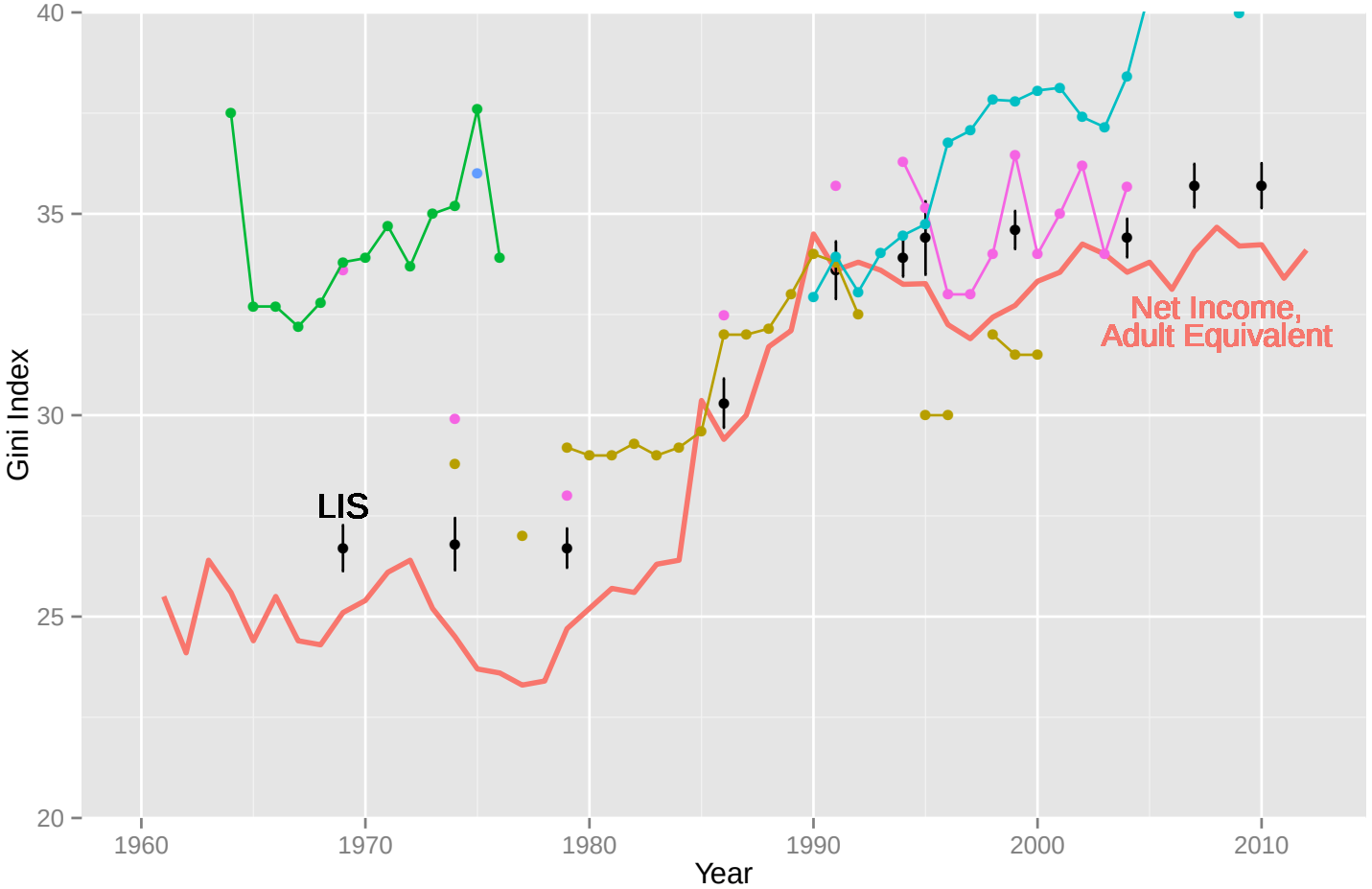
SWIID Source Data: A Complete Annual Series for the UK

An Illustration: Inequality in the United Kingdom



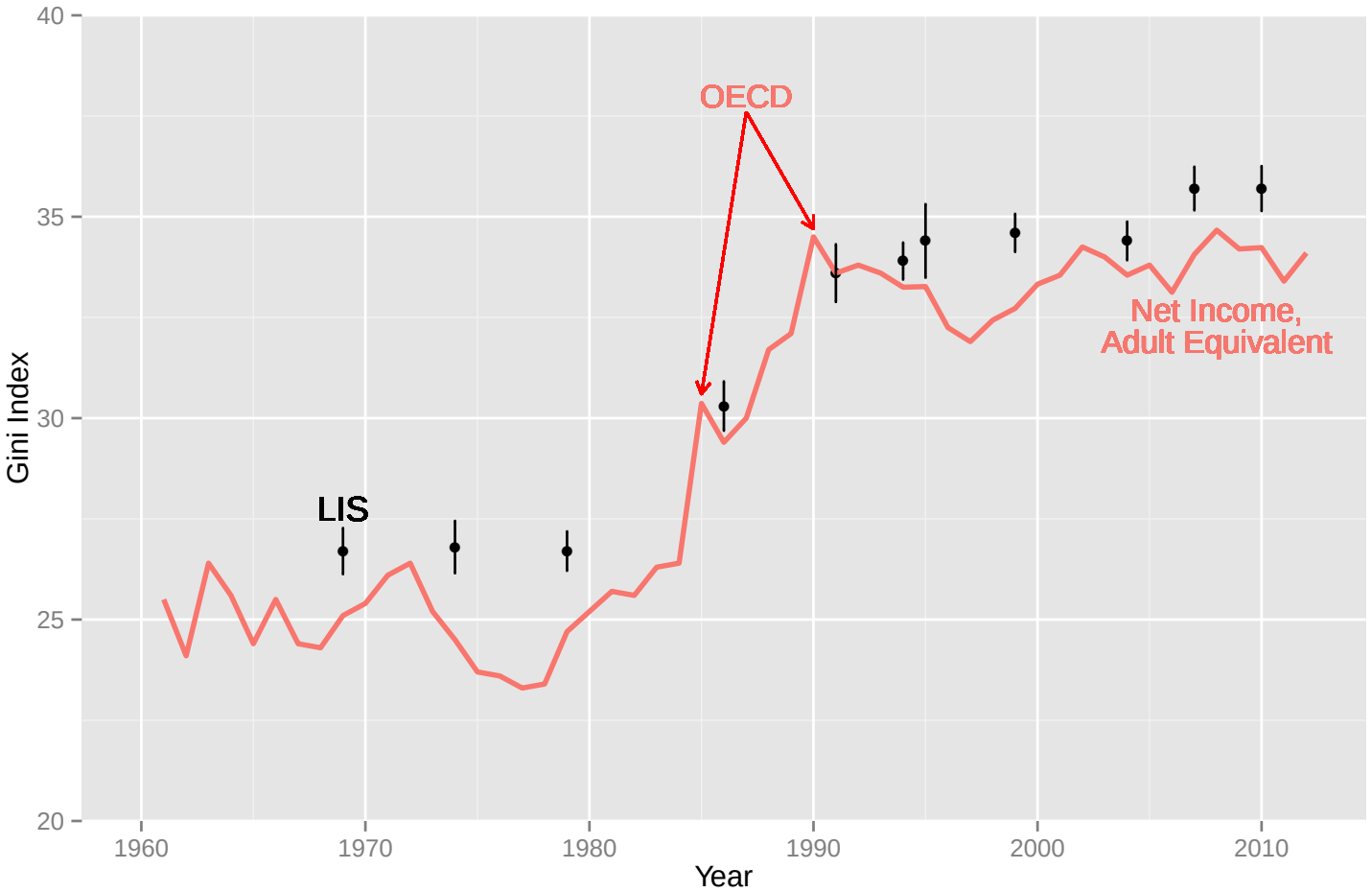
1. Counting across all countries, another series is actually more plentiful

An Illustration: Inequality in the United Kingdom



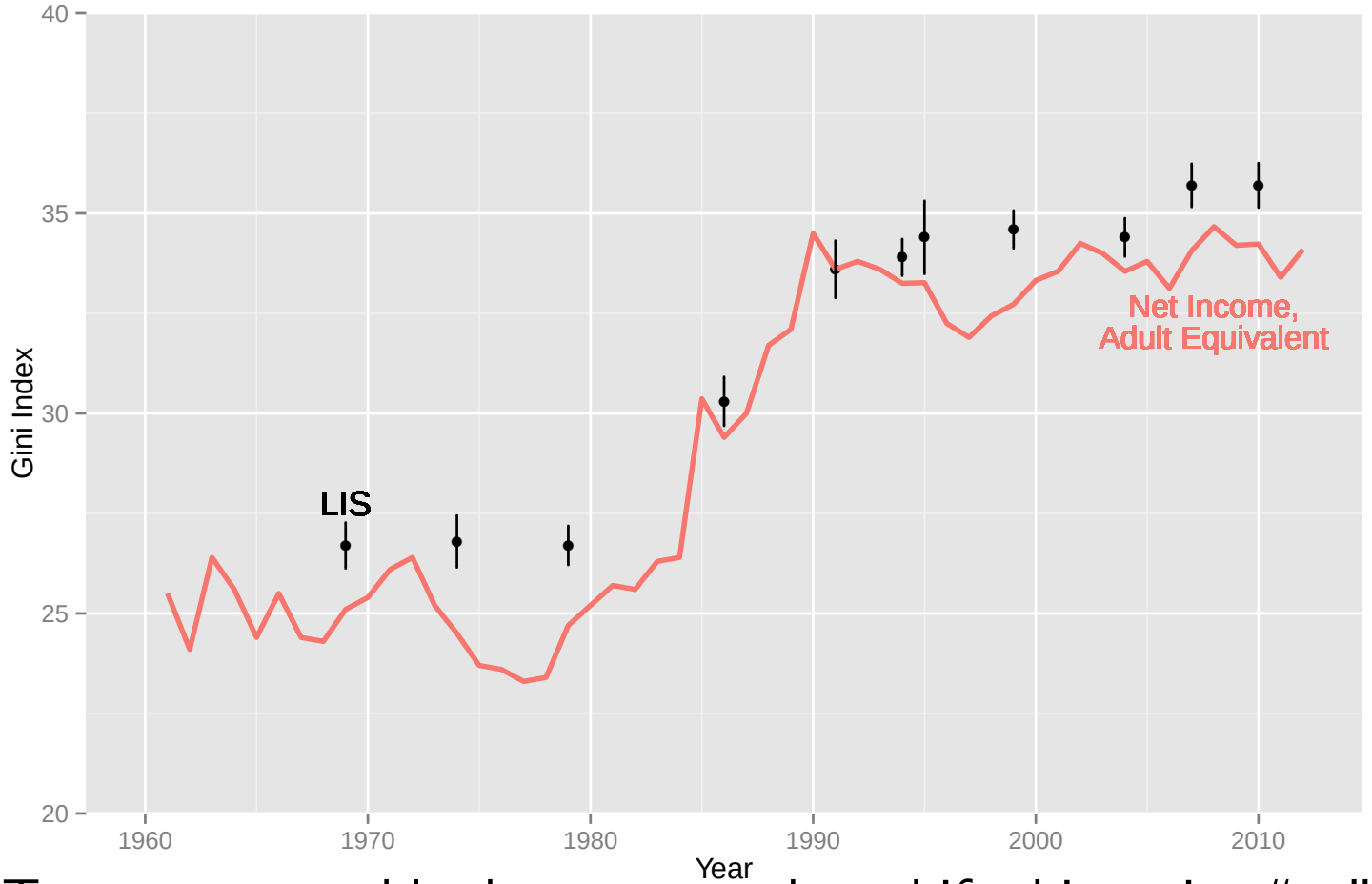
2. Using only the complete series discards a lot of other data

An Illustration: Inequality in the United Kingdom



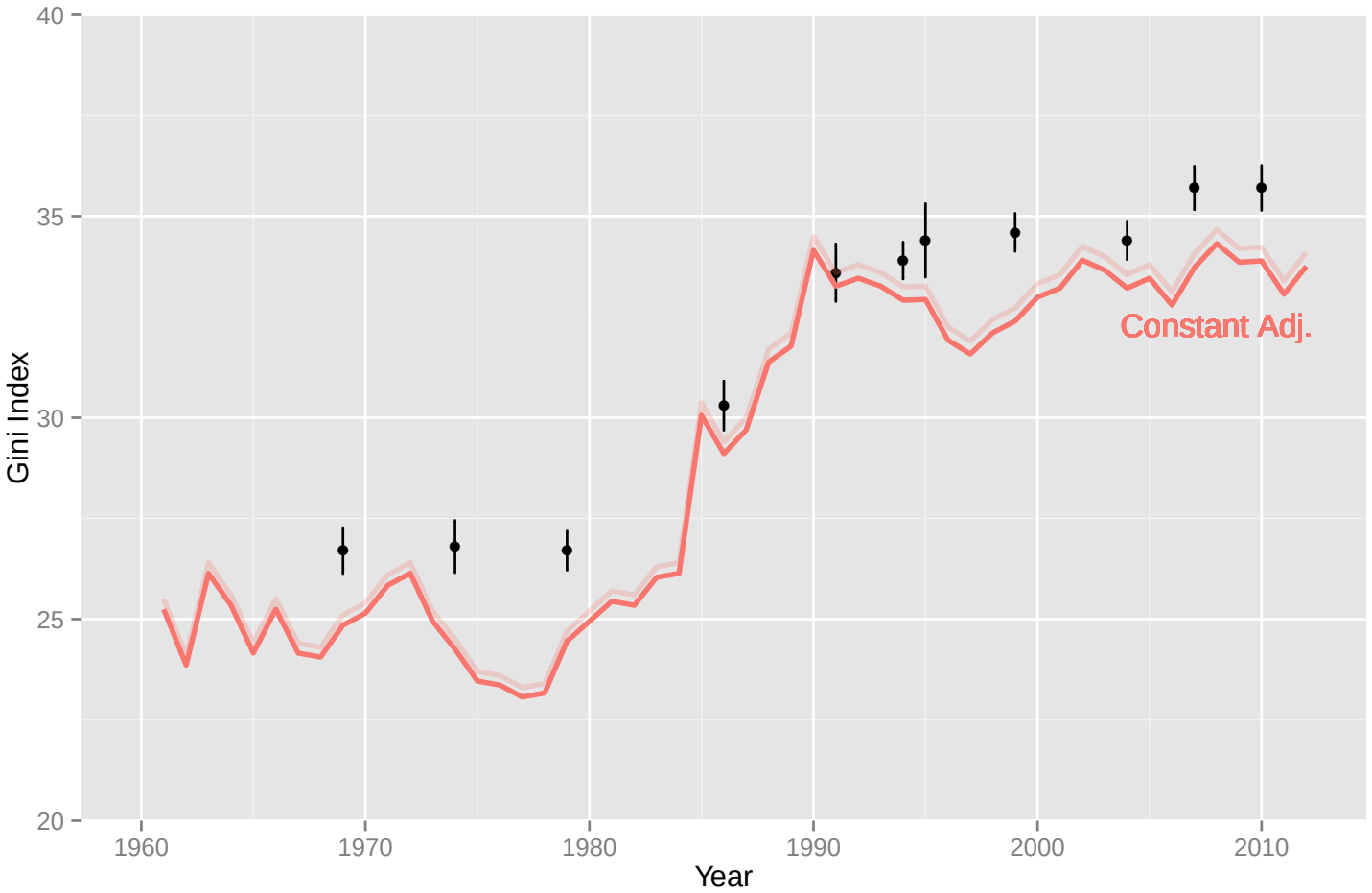
3. Harmonization is important . . .

An Illustration: Inequality in the United Kingdom



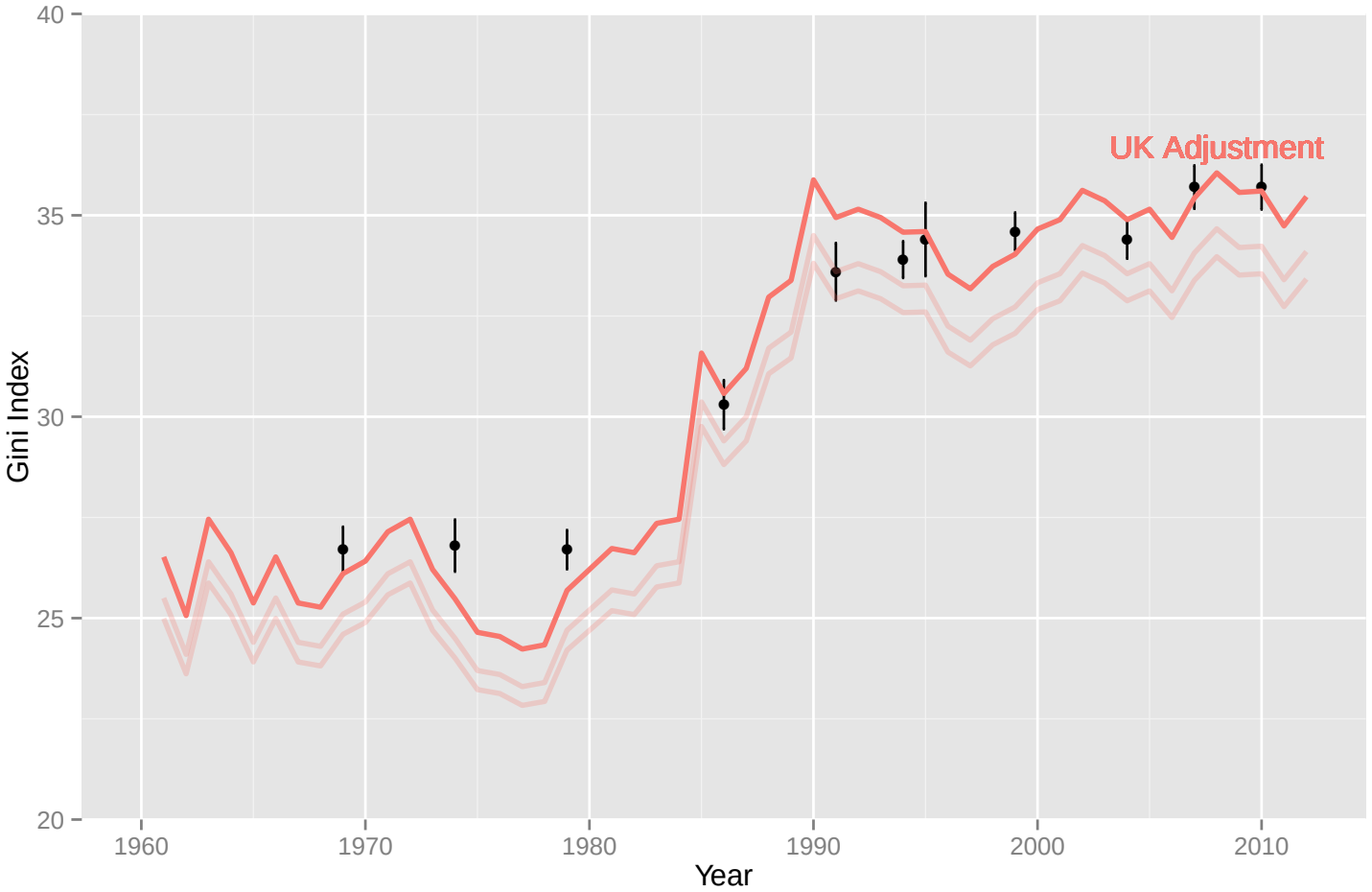
- To get comparable data, we need to shift this series "up": We need to multiply it by some factor greater than 1

An Illustration: Inequality in the United Kingdom



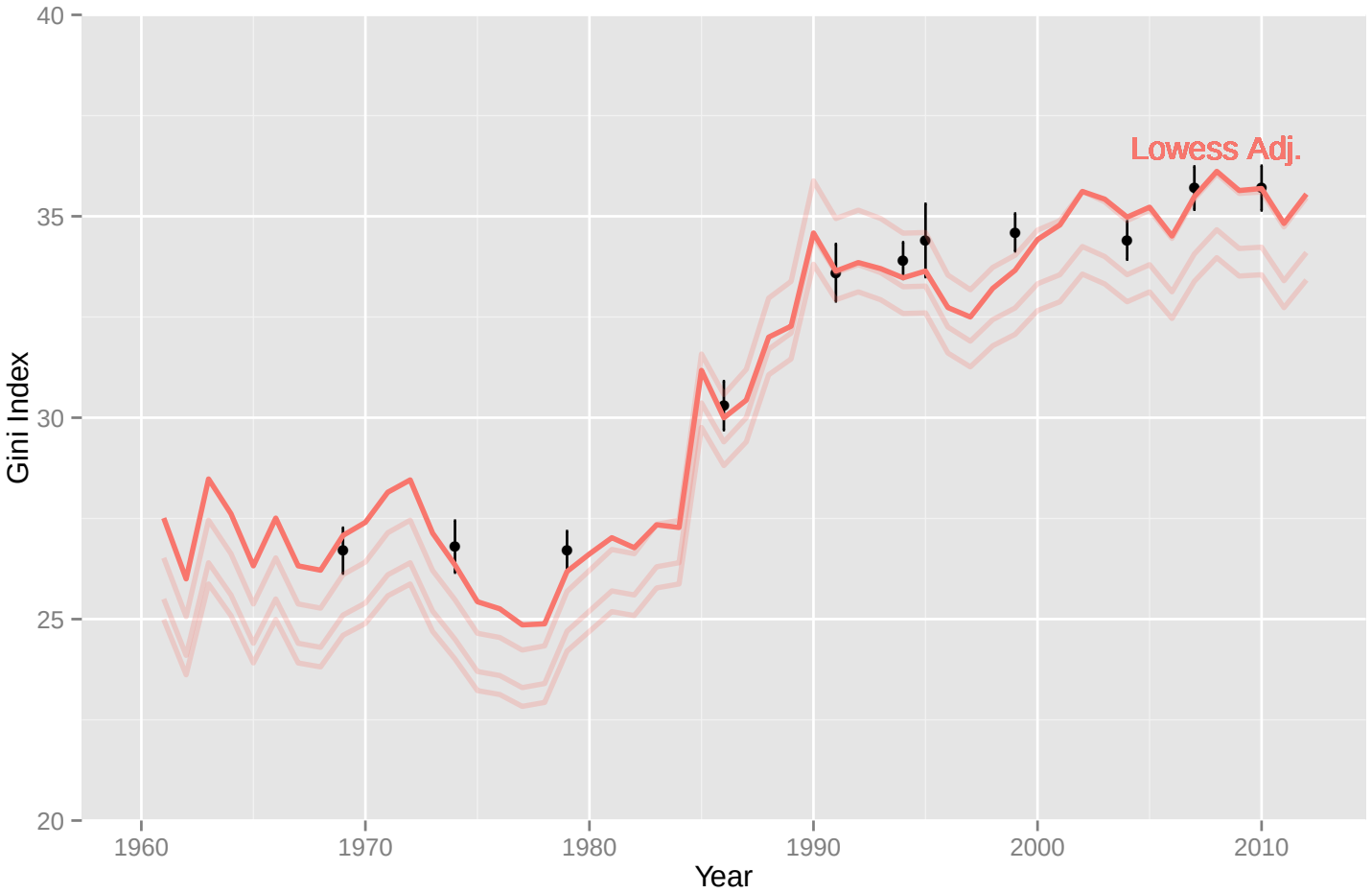
SWIID Source Data, Global Fixed Adjustment: $G \times .98$

An Illustration: Inequality in the United Kingdom



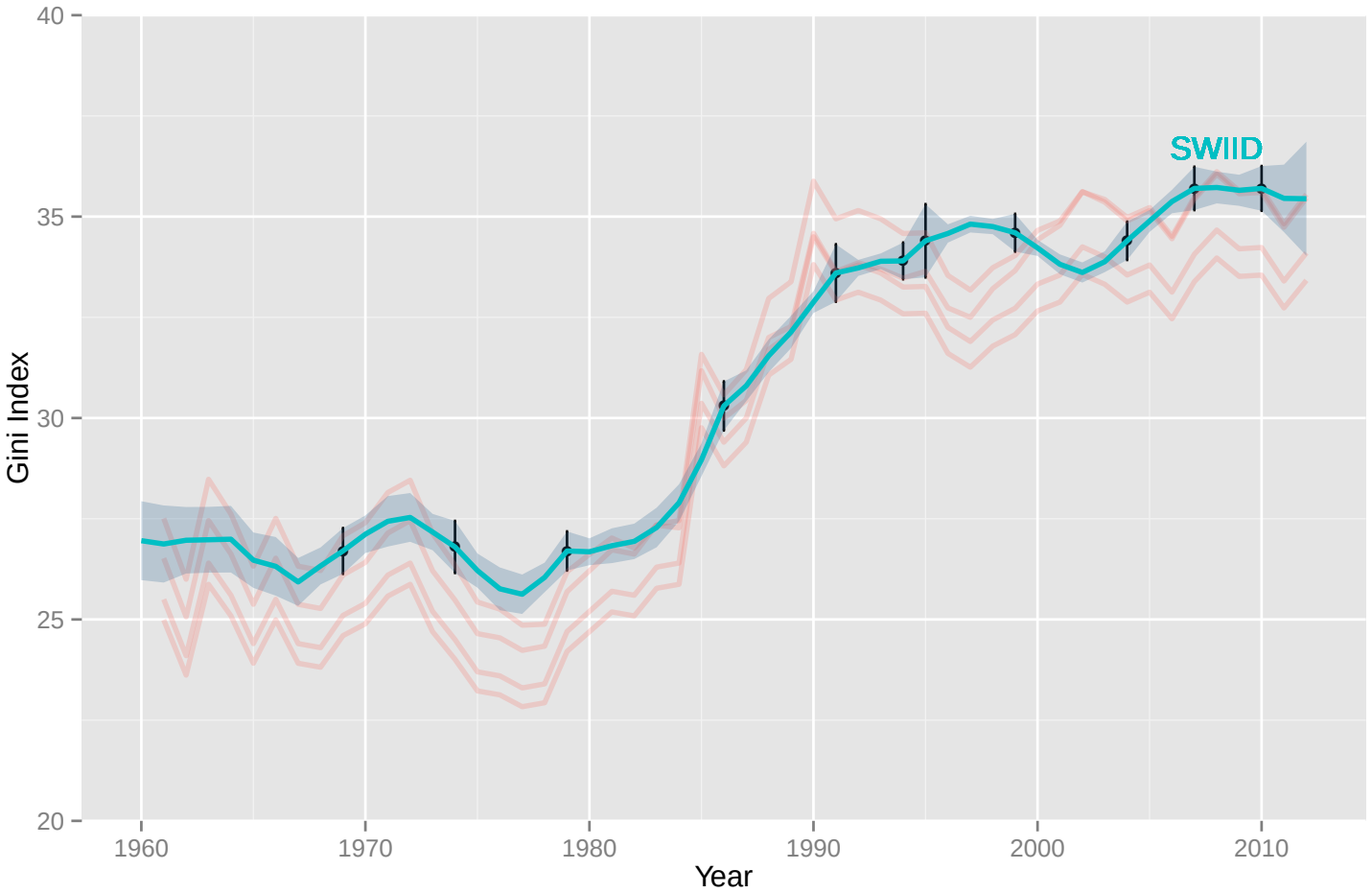
SWIID Source Data, UK Fixed Adjustment: $G \times 1.04$

An Illustration: Inequality in the United Kingdom



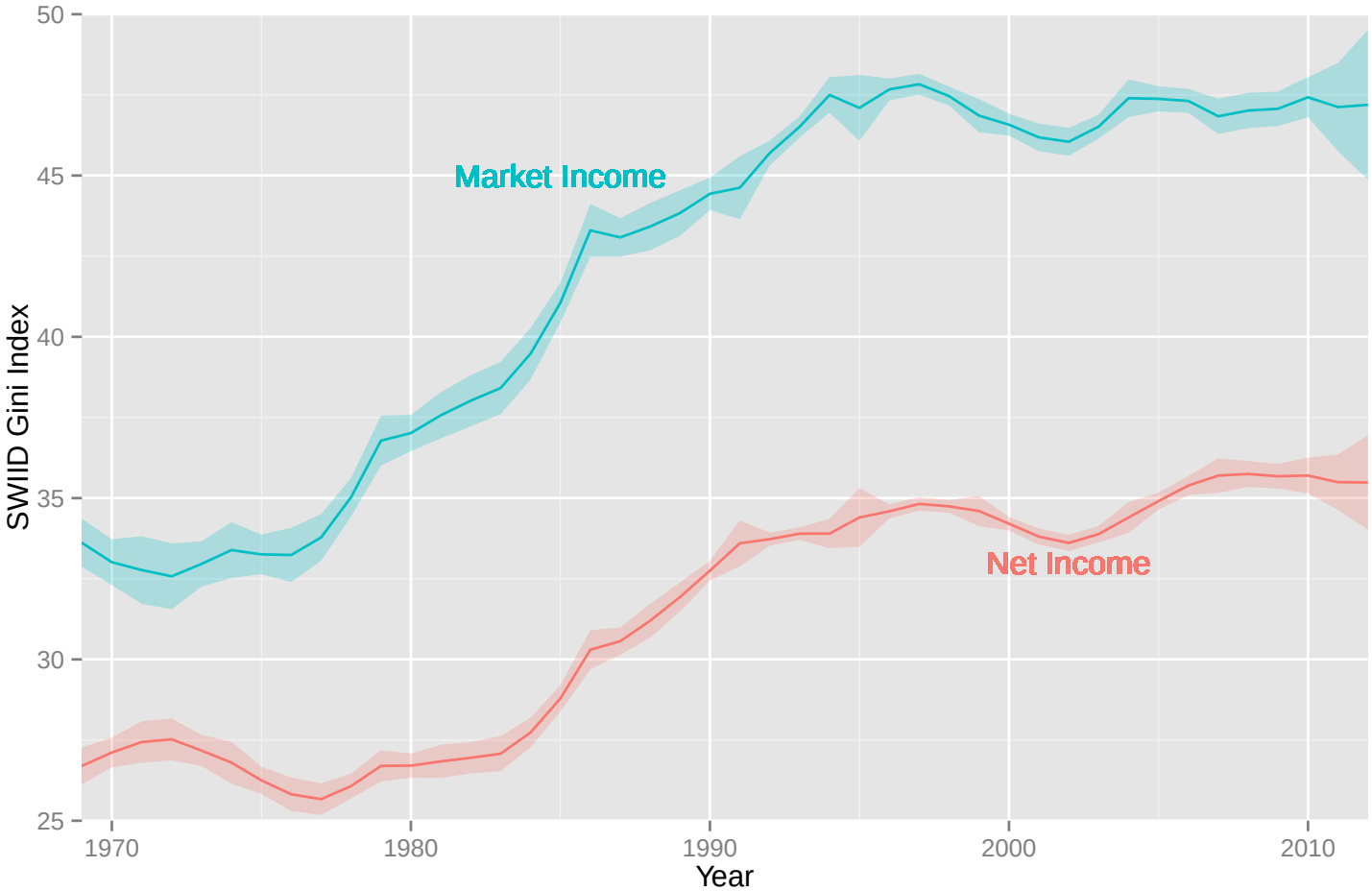
SWIID Source Data, Loess Estimate Adjustment: $G \times 1.00$ to 1.08

An Illustration: Inequality in the United Kingdom



SWIID Estimate and 95% Confidence Interval

An Illustration: Inequality in the United Kingdom



SWIID Estimates and 95% Confidence Intervals

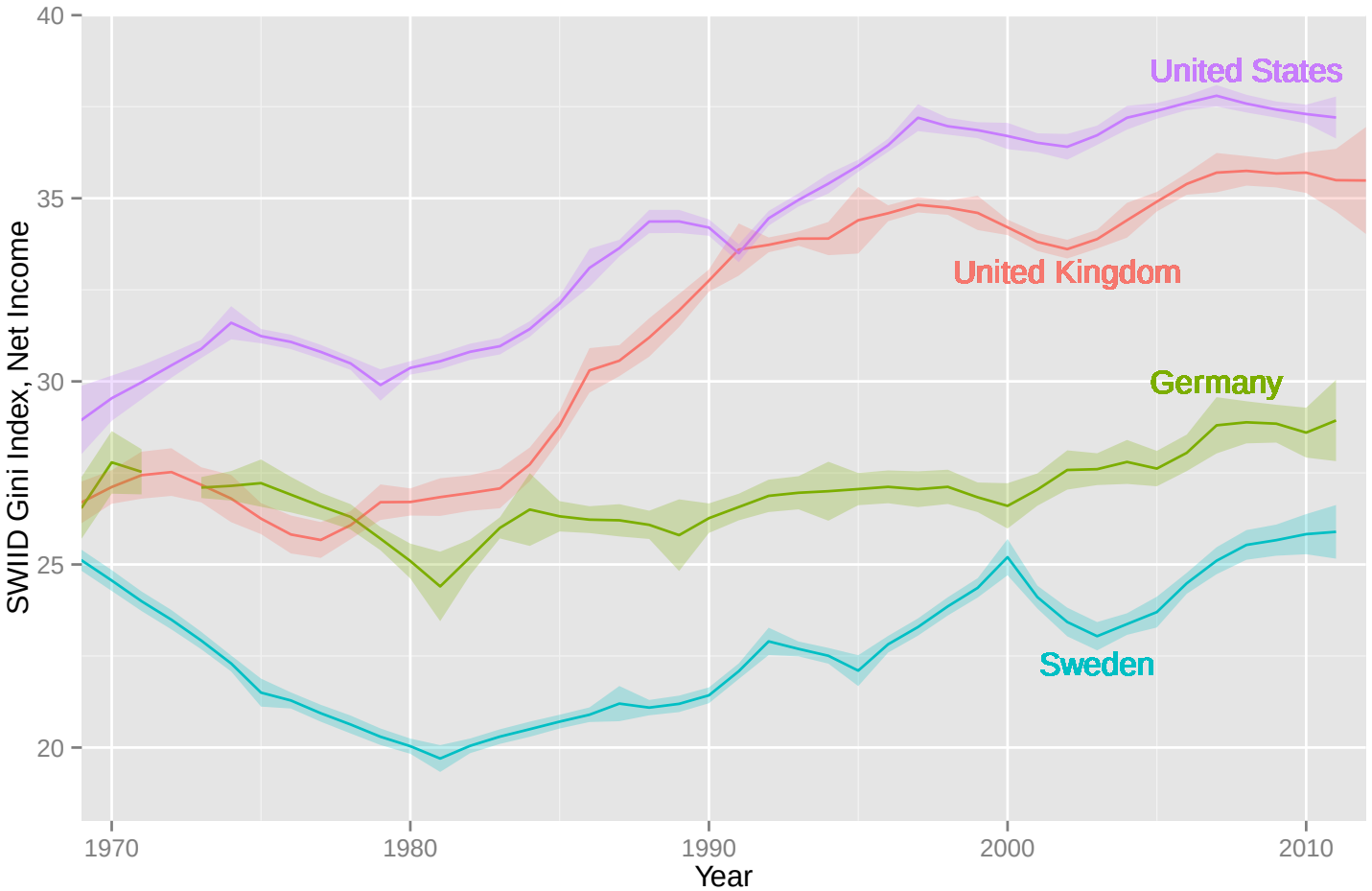
Constructing the SWIID

1. Collect all available Gini data
2. Categorize observations by combination of income definition and unit of analysis
3. Calculate all observed adjustment ratios $\rho_{abit} = \frac{G_{bit}}{G_{ait}}$
4. Predict unobserved $\hat{\rho}_{abit}$ using:
   a. Loess by country over time
   b. Regression by country-decade
   c. Regression by country
   d. Regression by 'region'
   e. Regression by advanced or developing world

Constructing the SWIID

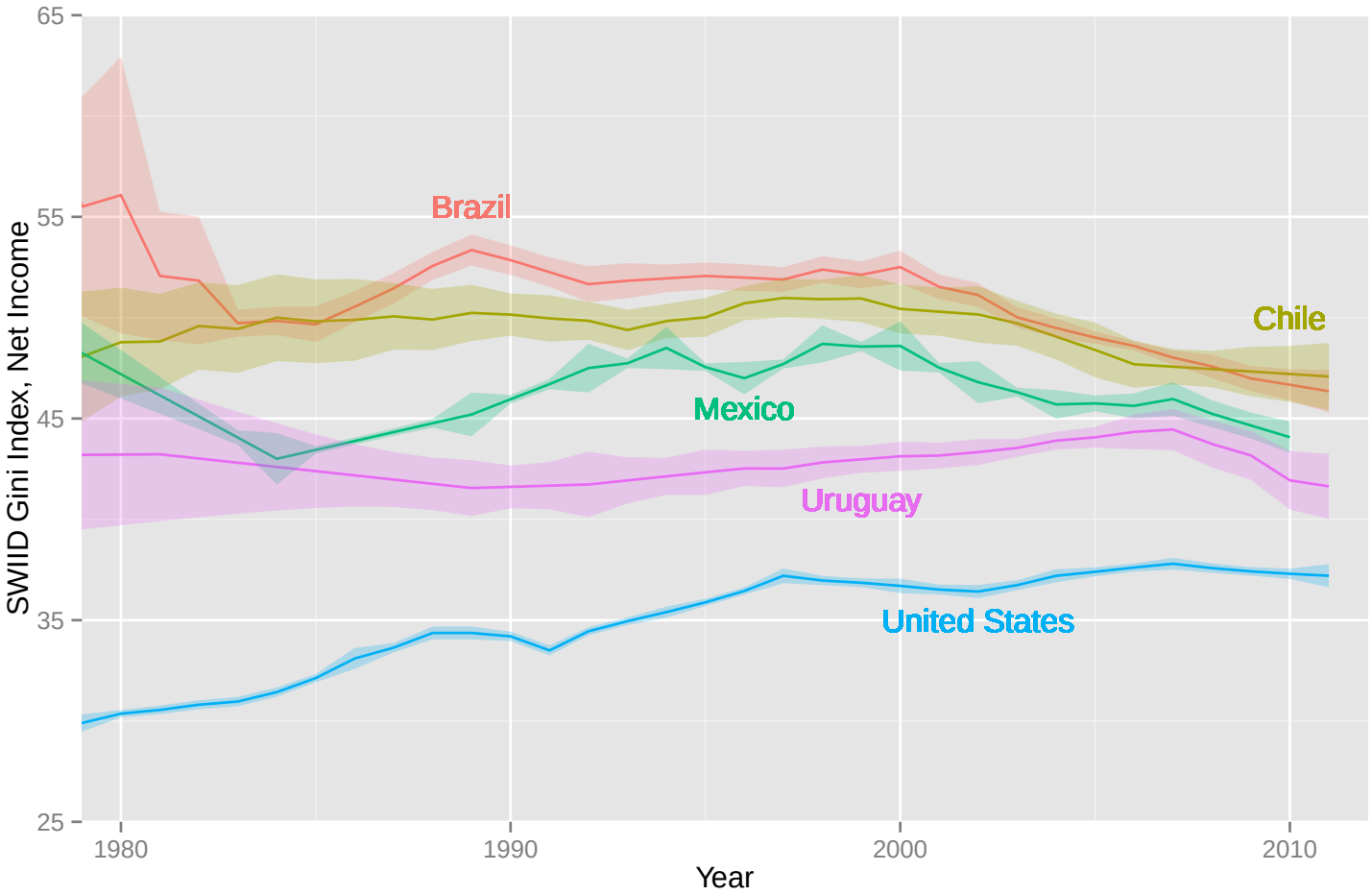
5. Make two-step predictions: $\hat{\rho}_{1bit} = \hat{\rho}_{abit} \times \hat{\rho}_{1ait}$
6. Use $\hat{\rho}$ and all observed Ginis to estimate standardized Ginis
7. If multiple estimates exist for a country-year, combine them:
   a. Start with best-fitting estimate for each country-year
   b. Incorporate additional estimates if they reduce error
8. Inform with estimates from surrounding country-years:
   a. Apply weighted moving-average smoother (with exceptions)
   b. Interpolate completely unobserved country-years

Inequality in Net Income in Four Rich Countries



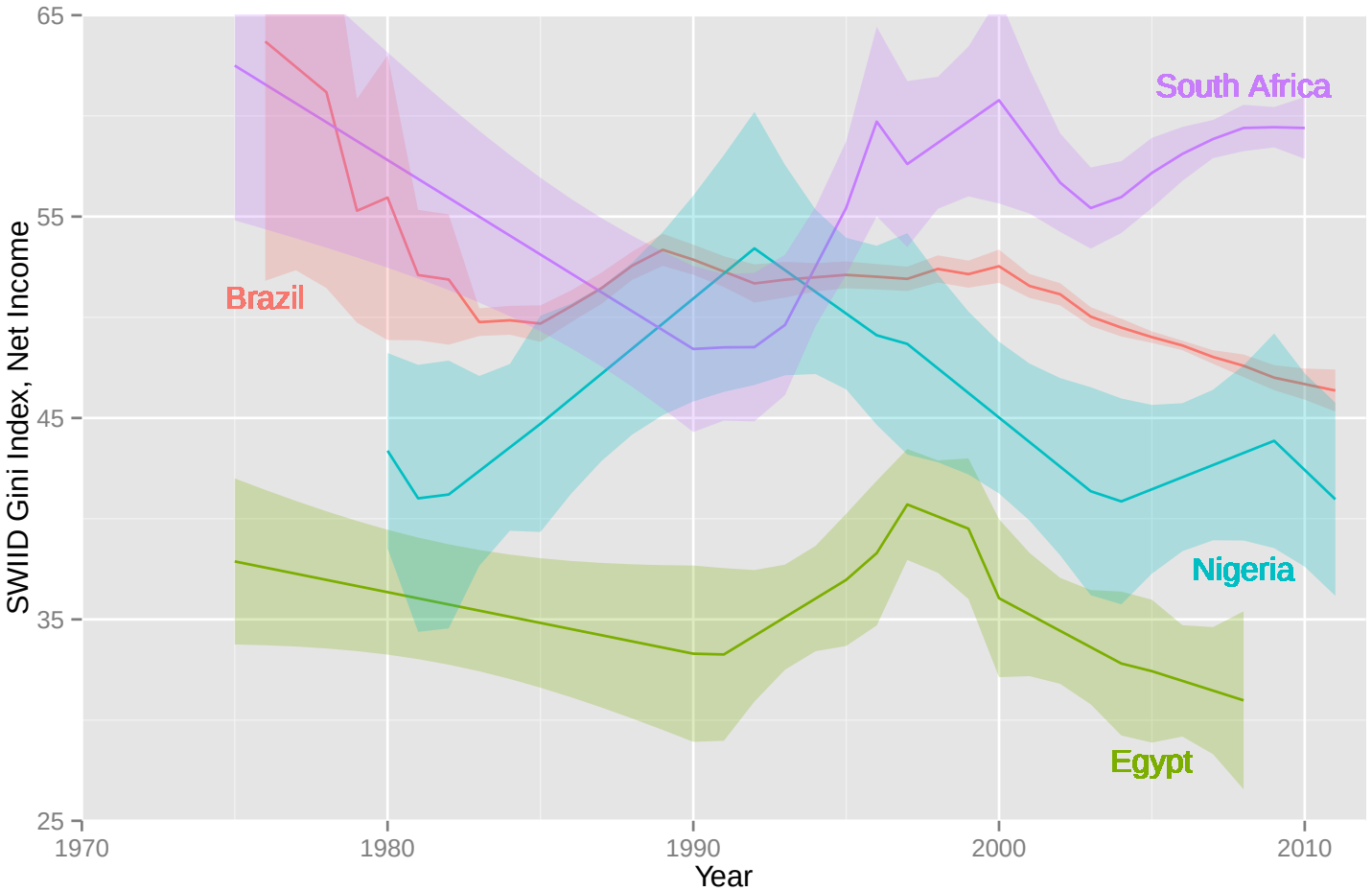
SWIID Estimates and 95% Confidence Intervals

Inequality in Net Income in Latin America



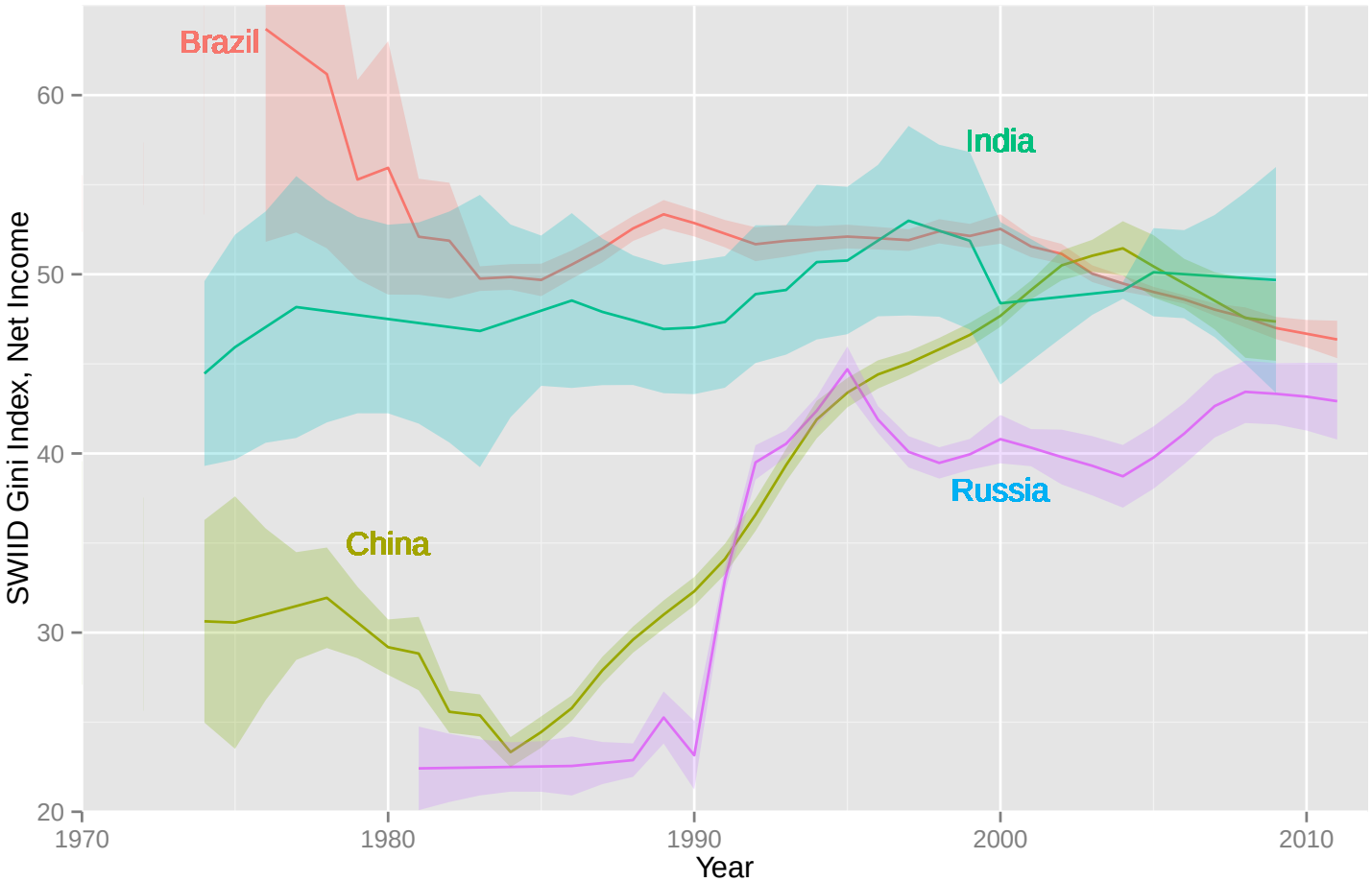
SWIID Estimates and 95% Confidence Intervals

Inequality in Net Income in Africa



SWIID Estimates and 95% Confidence Intervals

Inequality in Net Income in the BRICs



SWIID Estimates and 95% Confidence Intervals

Statistical Analysis with the SWIID

- Remember, the SWIID estimates are *estimates*
- Taking the standard errors into account is crucial to making well-grounded, defensible comparisons
- Fortunately, this is easier to do than it used to be . . .

Statistical Analysis with the SWIID

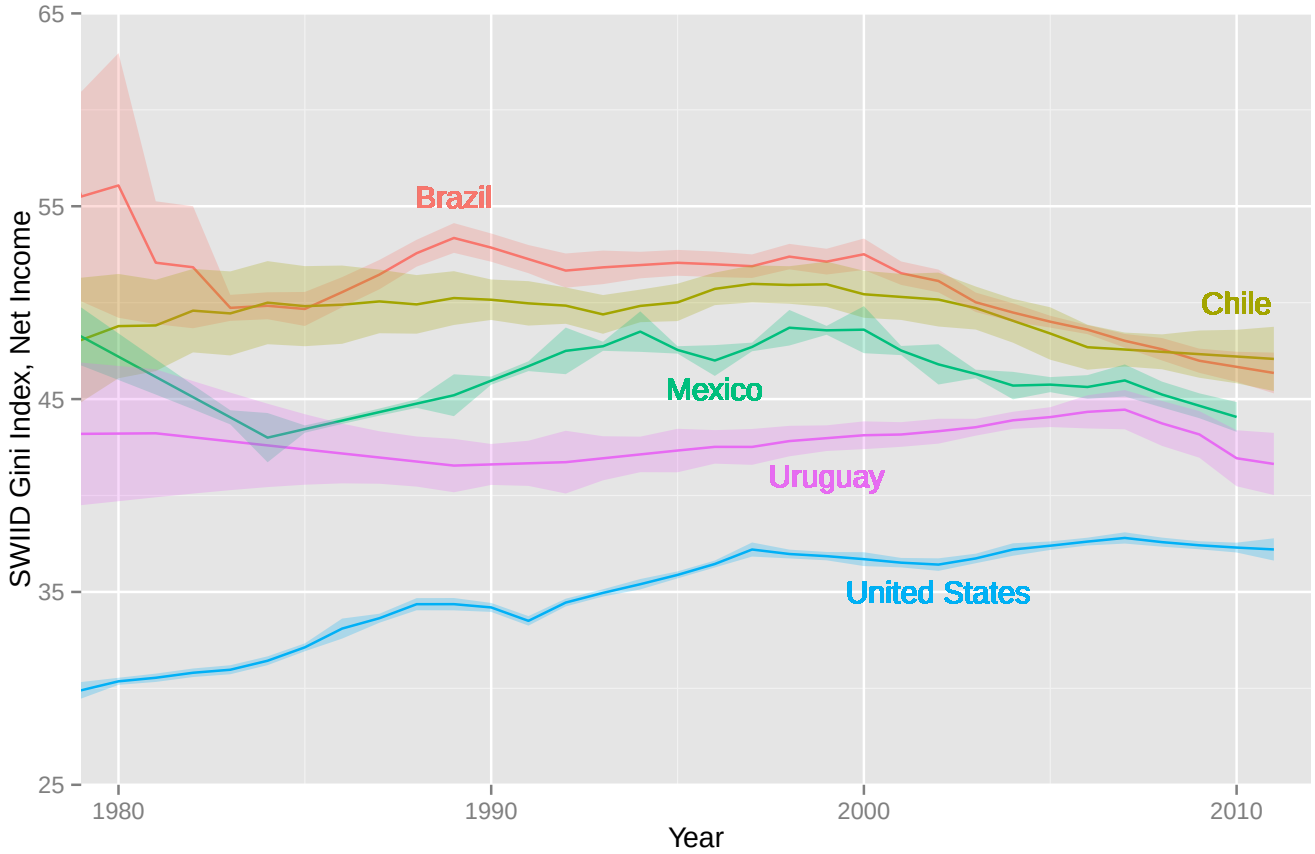
- Remember, the SWIID estimates are *estimates*
- Taking the standard errors into account is crucial to making well-grounded, defensible comparisons
- Fortunately, this is easier to do than it used to be . . .

Statistical Analysis with the SWIID

- Remember, the SWIID estimates are *estimates*
- Taking the standard errors into account is crucial to making well-grounded, defensible comparisons
- Fortunately, this is easier to do than it used to be . . .

CCTs in Brazil

- Has inequality declined in Brazil since the introduction of national conditional cash transfers in 2001?



CCTs in Brazil

```
. mi estimate: reg gini_net year if country=="Brazil" & year>=2001

Multiple-imputation estimates                     Imputations     =        100
Linear regression                                 Number of obs   =         11
                                                  Average RVI     =     3.7471
                                                  Largest FMI     =     0.8895
                                                  Complete DF     =          9
DF adjustment:   Small sample                     DF:     min     =       1.28
                                                          avg     =       1.28
                                                          max     =       1.29
Model F test:       Equal FMI                     F(   1,    1.3) =      61.93
Within VCE type:          OLS                     Prob > F        =     0.0483
```

gini_net	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
year	-.5262775	.0668735	-7.87	0.048	-1.038384	-.0141711
_cons	1104.367	134.0916	8.24	0.046	78.99686	2129.737

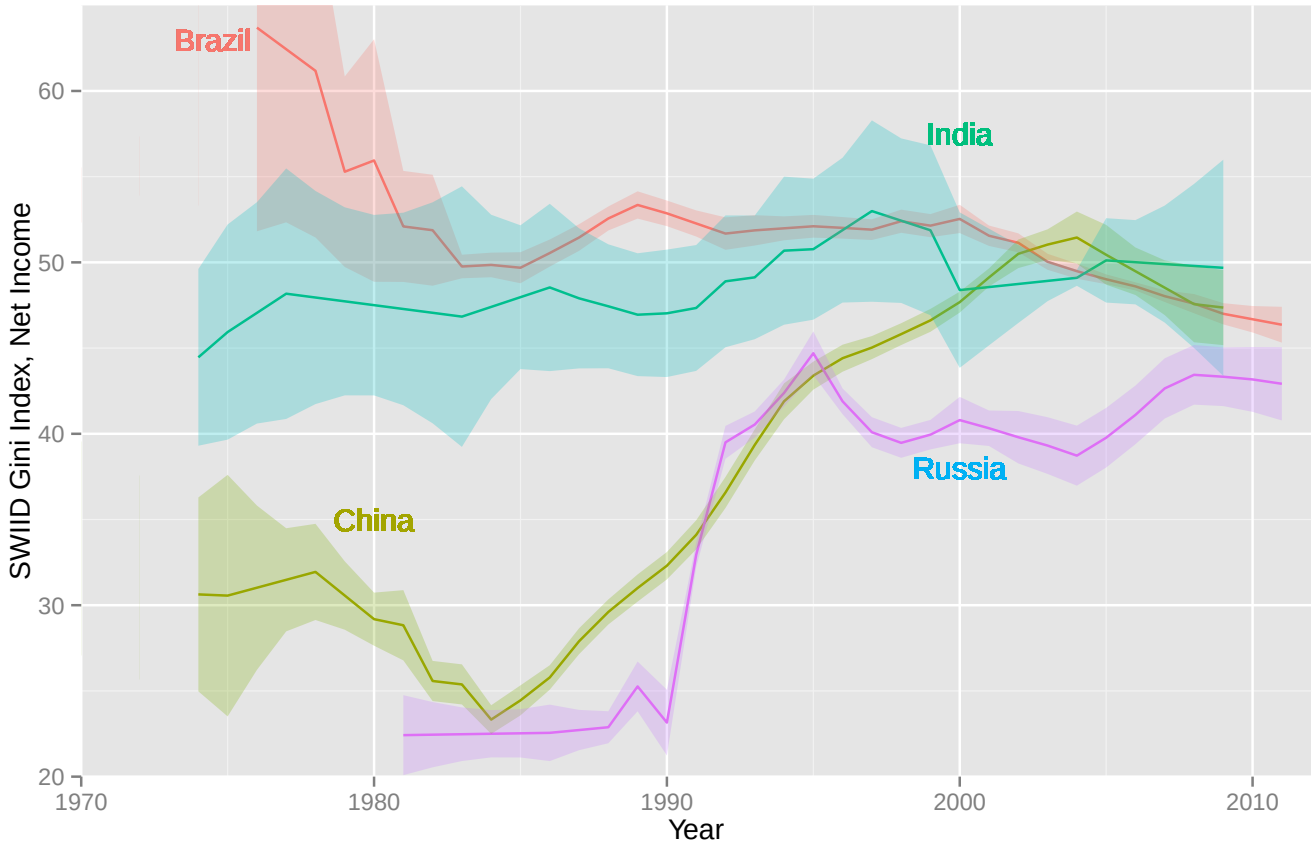
CCTs in Brazil

- Has the *trend* in inequality changed in Brazil since 2001?

```
. use "/Users/fredsolt/Downloads/SWIIDv4_pre-release-1/SWIIDv4_0_100pre.dta", clear

. gen no_cct=(country=="Brazil" & year<2001)

. xi: mi estimate: reg gini_net i.no_cct*year if (country=="Brazil" & year>= 1985)
i.no_cct          _Ino_cct_0-1        (naturally coded; _Ino_cct_0 omitted)
i.no_cct*year     _Ino_Xyear_#        (coded as above)

Multiple-imputation estimates                     Imputations      =        100
Linear regression                                 Number of obs    =         27
                                                  Average RVI      =     0.9517
                                                  Largest FMI      =     0.5545
                                                  Complete DF      =         23
DF adjustment:   Small sample                     DF:     min      =      10.04
                                                          avg      =      10.62
                                                          max      =      11.21
Model F test:       Equal FMI                     F(   3,   17.1)  =      37.89
Within VCE type:          OLS                     Prob > F         =     0.0000

------------------------------------------------------------------------------
    gini_net |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  _Ino_cct_1 |  -1210.569   218.0292    -5.55   0.000    -1696.08   -725.0585
        year |  -.5262775   .0897359    -5.86   0.000    -.7233491   -.3292059
_Ino_Xyear_1 |   .6056418   .1089116     5.56   0.000     .3630885    .8481951
       _cons |   1104.367   179.9681     6.14   0.000     709.1566    1499.577
------------------------------------------------------------------------------
```

Post-Communism

- Since 1989, has inequality been increasing more rapidly in China than in Russia?



Post-Communism

- Since 1989, has inequality been increasing more rapidly in China than in Russia?

```
. gen china = (country=="China")

. drop if year<1989
(1595 observations deleted)

. xi: mi estimate: reg gini_net i.china*year if country=="China" | country=="Russian Federati
> on"
i.china           _Ichina_0-1         (naturally coded; _Ichina_0 omitted)
i.china*year      _IchiXyear_#        (coded as above)

Multiple-imputation estimates                     Imputations     =        100
Linear regression                                 Number of obs   =         44
                                                  Average RVI     =     0.1204
                                                  Largest FMI     =     0.1322
                                                  Complete DF     =         40
DF adjustment:   Small sample                     DF:     min     =      33.23
                                                          avg     =      33.26
                                                          max     =      33.29
Model F test:       Equal FMI                     F(   3,   37.5) =      21.92
Within VCE type:          OLS                     Prob > F        =     0.0000

------------------------------------------------------------------------------
    gini_net |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   _Ichina_1 |   -809.946   394.6244    -2.05   0.048    -1612.548   -7.343584
        year |   .4740573   .1298514     3.65   0.001     .2099419    .7381727
_IchiXyear_1 |   .4079588   .1974129     2.07   0.047     .0064452    .8094723
       _cons |  -908.7984   259.6369    -3.50   0.001    -1436.884   -380.7126
------------------------------------------------------------------------------
```

The Standardized World Income Inequality Database

- The SWIID maximizes comparability for the widest possible sample
- It multiply-imputes missing data in the LIS, using as much information as possible from the same country in proximate years
- Residual incomparability is represented as uncertainty
- Incorporating the standard errors is therefore crucial . . .
- . . . and is now relatively easy

The Standardized World Income Inequality Database

- The SWIID maximizes comparability for the widest possible sample
- It multiply-imputes missing data in the LIS, using as much information as possible from the same country in proximate years
- Residual incomparability is represented as uncertainty
- Incorporating the standard errors is therefore crucial . . .
- . . . and is now relatively easy

The Standardized World Income Inequality Database

- The SWIID maximizes comparability for the widest possible sample
- It multiply-imputes missing data in the LIS, using as much information as possible from the same country in proximate years
- Residual incomparability is represented as uncertainty
- Incorporating the standard errors is therefore crucial . . .
- . . . and is now relatively easy

The Standardized World Income Inequality Database

- The SWIID maximizes comparability for the widest possible sample
- It multiply-imputes missing data in the LIS, using as much information as possible from the same country in proximate years
- Residual incomparability is represented as uncertainty
- Incorporating the standard errors is therefore crucial . . .
- . . . and is now relatively easy

The Standardized World Income Inequality Database

- The SWIID maximizes comparability for the widest possible sample
- It multiply-imputes missing data in the LIS, using as much information as possible from the same country in proximate years
- Residual incomparability is represented as uncertainty
- Incorporating the standard errors is therefore crucial . . .
- . . . and is now relatively easy