Inequality and Growth: *Regressions on Panel Data in R*

Jeffrey Yozwiak
SDGB 7847 Machine Learning for Statistics
Spring 2020

My Background

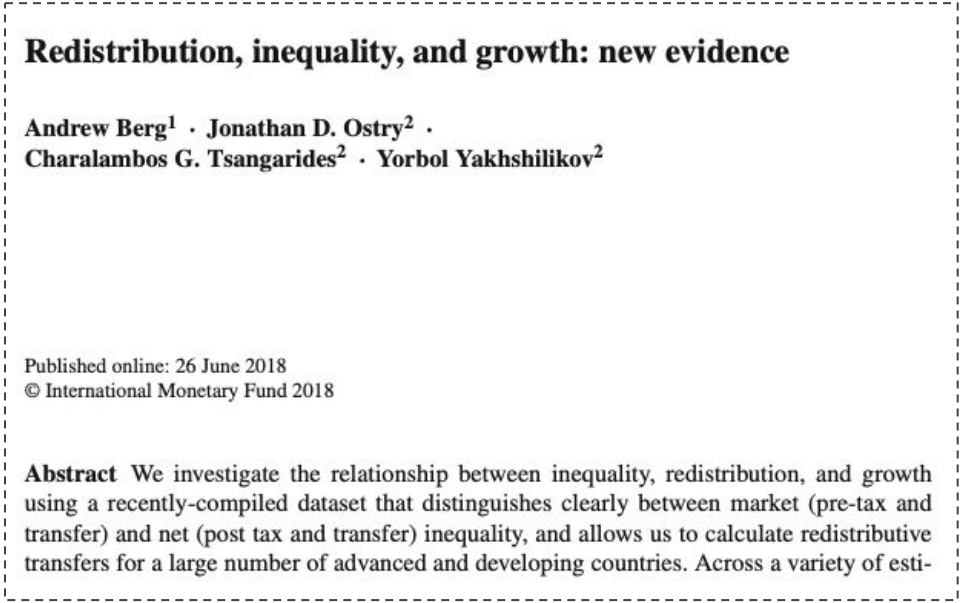
- M.A. in Economics, Spring 2020 (GSAS)
- Research interests:
  - Inequality and poverty in the US
  - Optimal tax theory/tax policy design
- Professional background: tech startups

Statistical Analysis in Economics

1. **Regressions** to establish causality. Interpretability is key.
2. **Feature elimination** informed by theory.
3. *Cross-country growth regressions* on **panel data** (country = *i*; year = *t*).

$$y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \ldots + u$$

Project Goals



Redistribution, inequality, and growth: new evidence

Andrew Berg[1] · Jonathan D. Ostry[2] · Charalambos G. Tsangarides[2] · Yorbol Yakhshilikov[2]

Published online: 26 June 2018

Abstract We investigate the relationship between inequality, redistribution, and growth using a recently-compiled dataset that distinguishes clearly between market (pre-tax and transfer) and net (post tax and transfer) inequality, and allows us to calculate redistributive transfers for a large number of advanced and developing countries. Across a variety of esti-

Goal: Reproduce Berg et al. (2018):

1. "Lower net inequality is robustly correlated with faster and more durable growth"

   ↑ inequality ⇒ ↓ growth

2. "Inequality seems to affect growth through human capital accumulation and fertility channels"

   ↑ inequality ⇒ ↓ education, health ⇒ ↓ growth

Raw Data

Observations: 2,079
Variables: 97
Sources: SWIID 3.1, PWT 7.1, WEO, Polity IV, Barro and Lee (2013), Lane and Milesi-Ferretti (2011)
NAs: 73,963 (37%)

Variable	Definition	Variable	Definition
wbcode	World Bank country code	lmfdebtl	Government debt
year	Year (every 5 years from 1960–2010)	p4polity2	10 = democracy to -10 = autocracy
logincome_pc	Income per capita (log transformed)	open	Openness to trade
gini_net	Gini after taxes and transfers (0 = equality to 1 = inequality)	lnpopgr	Pop. growth rate (log transformed)
gini_market	Gini before taxes and transfers	lfexp	Life expectancy
govexp	Government expenditure	ltoted	Avg. years of education
inv	Total investment	pvintv, pubinv, fert, chmort, admortm, admortf, primaryeducyears, secondaryeducyears, oecd_1975, yr_sch_f, yr_sch_pri_f, yr_sch_sec_f, yr_sch_ter_f, yr_sch_pri_m...	
lni	Ratio of investment to GDP (log trans.)		

Data Cleaning

Results:

Full dataset:
Observations: 1,078
Variables: 14

OECD dataset:
Observations: 231
Variables: 13

1. Drop superfluous variables (e.g., nonoecd_1975).
2. Many variables measure roughly the same concept (e.g., health, education, etc.). ⇒ Prefer the variables with fewer NAs.
    - *Example:* yr_sch-, primaryeducyears/ secondaryeducyears, and ltoted all measure education. However, yr_sch- has 37% NAs vs. 13–17% NAs for the others. ⇒ Drop yr_sch-.
3. Replace NAs with the mean value for that country (Edureka). (Thank you, Fred Viole!)
4. Create a dataset of just OECD countries.
5. Drop collinear variables.
6. Drop any observations that still have NAs.
7. PCA fails (Statalist).

Implementation

- `plm` (Croissant and Millo 2008) and `pmdyplr` (Huntington-Klein 2020) libraries.
- Regressions:

Data	Formula
Full	$\texttt{logincome\_pc}_{i,t} \sim \texttt{gini\_net}_{i,t} + X_{i,t}$
Full	$\texttt{logincome\_pc}_{i,t} \sim \texttt{gini\_market}_{i,t} + X_{i,t}$
OECD	$\texttt{logincome\_pc}_{i,t} \sim \texttt{gini\_net}_{i,t} + X_{i,t}$
OECD	$\texttt{logincome\_pc}_{i,t} \sim \texttt{gini\_market}_{i,t} + X_{i,t}$

Results: Full Dataset

```
> summary(model_plm_full_giniNet)
Oneway (individual) effect Within Model

Call:
plm(formula = logincome_pc ~ gini_net + govexp + inv + lni +
    lmfdebtl + p4polity2 + open + lnpopgr + lfexp + ltoted, data
= data_reg_full)

Balanced Panel: n = 98, T = 11, N = 1078

Residuals:
     Min.   1st Qu.    Median   3rd Qu.      Max.
-1.309544 -0.153136  0.013491  0.157904  1.574041

Coefficients:
             Estimate  Std. Error t-value  Pr(>|t|)
gini_net   6.0957e-03  2.9933e-03  2.0365 0.0419757 *
govexp     3.9237e-07  2.4365e-07  1.6104 0.1076322
inv        5.9467e-03  1.8487e-03  3.2167 0.0013396 **
lni       -4.8112e-02  6.5041e-02 -0.7397 0.4596469
lmfdebtl   9.9123e-05  2.5109e-04  0.3948 0.6930987
p4polity2 -2.5171e-03  2.3119e-03 -1.0887 0.2765450
open       4.9845e-03  5.4131e-04  9.2082 < 2.2e-16 ***
lnpopgr   -2.3009e-01  9.0771e-02 -2.5349 0.0114045 *
lfexp      3.0843e-02  2.0954e-03 14.7196 < 2.2e-16 ***
ltoted     1.6255e-01  4.5294e-02  3.5889 0.0003487 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    149.39
Residual Sum of Squares: 80.85
R-Squared:      0.45881
Adj. R-Squared: 0.39911
F-statistic: 82.2347 on 10 and 970 DF, p-value: < 2.22e-16
```

```
> summary(model_plm_full_giniMarket)
Oneway (individual) effect Within Model

Call:
plm(formula = logincome_pc ~ gini_market + govexp + inv + lni +
    lmfdebtl + p4polity2 + open + lnpopgr + lfexp + ltoted, data
= data_reg_full)

Balanced Panel: n = 98, T = 11, N = 1078

Residuals:
     Min.   1st Qu.    Median   3rd Qu.      Max.
-1.307783 -0.155921  0.013247  0.162623  1.572907

Coefficients:
               Estimate  Std. Error t-value  Pr(>|t|)
gini_market  4.6881e-03  2.3202e-03  2.0205 0.0436057 *
govexp       3.8045e-07  2.4373e-07  1.5610 0.1188605
inv          6.0124e-03  1.8483e-03  3.2530 0.0011812 **
lni         -6.0783e-02  6.5299e-02 -0.9309 0.3521605
lmfdebtl     9.7201e-05  2.5116e-04  0.3870 0.6988357
p4polity2   -2.3552e-03  2.3092e-03 -1.0199 0.3080265
open         4.8708e-03  5.4175e-04  8.9909 < 2.2e-16 ***
lnpopgr     -2.3325e-01  9.0807e-02 -2.5687 0.0103567 *
lfexp        3.0928e-02  2.0941e-03 14.7692 < 2.2e-16 ***
ltoted       1.8668e-01  4.8366e-02  3.8596 0.0001211 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    149.39
Residual Sum of Squares: 80.855
R-Squared:      0.45877
Adj. R-Squared: 0.39907
F-statistic: 82.2228 on 10 and 970 DF, p-value: < 2.22e-16
```

Analysis:

- gini- is significant but positive (expected negative).
- gini_net vs. gini_market doesn’t matter.
- Drivers of growth: investment, openness, and human capital.
- R^2 = 46% so there’s a lot of variation in logincome_pc that these features *don’t* explain.

Results: OECD Dataset

```
> summary(model_plm_oecd_giniNet)
Oneway (individual) effect Within Model

Call:
plm(formula = logincome_pc ~ gini_net + govexp + inv + lni +
    p4polity2 + open + lnpopgr + lfexp + ltoted, data = data_reg
_oecd)

Balanced Panel: n = 21, T = 11, N = 231

Residuals:
     Min.   1st Qu.    Median   3rd Qu.      Max.
-0.697918 -0.079813  0.023697  0.096447  0.721968

Coefficients:
             Estimate  Std. Error t-value  Pr(>|t|)
gini_net  -2.8035e-03  5.4200e-03 -0.5173  0.605549
govexp     2.1431e-06  2.8429e-06  0.7538  0.451833
inv        3.9486e-02  7.8035e-03  5.0600 9.436e-07 ***
lni       -1.0250e+00  2.1851e-01 -4.6909 5.019e-06 ***
p4polity2  1.2663e-02  4.7133e-03  2.6867  0.007820 **
open       3.8136e-03  9.2019e-04  4.1444 5.019e-05 ***
lnpopgr    2.5457e-01  2.4630e-01  1.0336  0.302585
lfexp      9.4577e-02  5.6991e-03 16.5949 < 2.2e-16 ***
ltoted    -3.1402e-01  1.1721e-01 -2.6791  0.007993 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    38.136
Residual Sum of Squares: 6.2894
R-Squared:      0.83508
Adj. R-Squared: 0.81129
F-statistic: 113.086 on 9 and 201 DF, p-value: < 2.22e-16
```

```
> summary(model_plm_oecd_giniMarket)
Oneway (individual) effect Within Model

Call:
plm(formula = logincome_pc ~ gini_market + govexp + inv + lni +
    p4polity2 + open + lnpopgr + lfexp + ltoted, data = data_reg
_oecd)

Balanced Panel: n = 21, T = 11, N = 231

Residuals:
     Min.   1st Qu.    Median   3rd Qu.      Max.
-0.697755 -0.078250  0.023201  0.094924  0.725561

Coefficients:
              Estimate  Std. Error t-value  Pr(>|t|)
gini_market -4.4643e-03  3.3217e-03 -1.3440  0.180473
govexp       2.4291e-06  2.8407e-06  0.8551  0.393501
inv          3.9678e-02  7.7743e-03  5.1038 7.692e-07 ***
lni         -1.0216e+00  2.1722e-01 -4.7030 4.759e-06 ***
p4polity2    1.2061e-02  4.6449e-03  2.5966  0.010111 *
open         3.9948e-03  9.2684e-04  4.3101 2.554e-05 ***
lnpopgr      2.5980e-01  2.4507e-01  1.0601  0.290381
lfexp        9.4915e-02  5.6838e-03 16.6991 < 2.2e-16 ***
ltoted      -3.2729e-01  1.1585e-01 -2.8250  0.005204 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    38.136
Residual Sum of Squares: 6.2416
R-Squared:      0.83633
Adj. R-Squared: 0.81272
F-statistic: 114.121 on 9 and 201 DF, p-value: < 2.22e-16
```

Analysis:

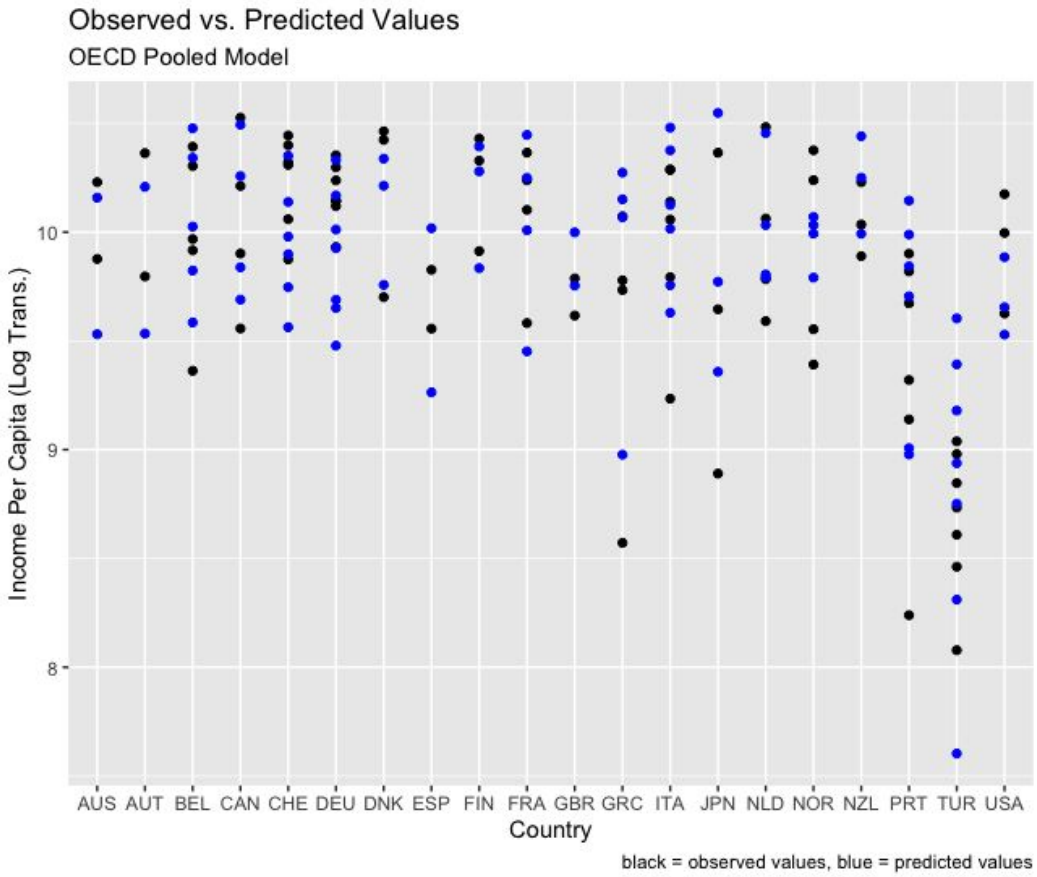
- gini- is not significant.
- Drivers of growth in OECD countries: investment, openness, democracy, and human capital.
- F-stat = 114: stronger evidence of joint significance.
- R^2 = 84%: explains much more of the variation in logincome_pc.
- More complete, more accurate data?

Model: Summary and Performance

```
> summary(model_plm_full_train_pooled)
Pooling Model

Call:
plm(formula = logincome_pc ~ gini_net + inv + open + lnpopgr +
    lfexp + ltoted, data = data_reg_full_train, model = "pooling
")

Unbalanced Panel: n = 98, T = 3-11, N = 755

Residuals:
    Min.  1st Qu.   Median  3rd Qu.     Max.
-2.51464 -0.41835  0.03203  0.46537  2.51214

Coefficients:
              Estimate  Std. Error t-value  Pr(>|t|)
(Intercept)  5.15310338  0.47025970 10.9580 < 2.2e-16 ***
gini_net     0.00023597  0.00289273  0.0816 0.9350092
inv          0.01170909  0.00262657  4.4579 9.543e-06 ***
open         0.00028528  0.00063546  0.4489 0.6536082
lnpopgr     -1.01892131  0.19641375 -5.1876 2.745e-07 ***
lfexp        0.06988113  0.00378439 18.4656 < 2.2e-16 ***
ltoted       0.24264790  0.07148001  3.3946 0.0007235 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    1233.8
Residual Sum of Squares: 340.91
R-Squared:      0.72368
Adj. R-Squared: 0.72146
F-statistic: 326.501 on 6 and 748 DF, p-value: < 2.22e-16
```

```
> summary(model_plm_oecd_train_pooled)
Pooling Model

Call:
plm(formula = logincome_pc ~ inv + lni + p4polity2 + open + lfex
p +
    ltoted, data = data_reg_oecd_train, model = "pooling")

Unbalanced Panel: n = 21, T = 4-11, N = 162

Residuals:
      Min.    1st Qu.     Median    3rd Qu.       Max.
-0.6022010 -0.1698074  0.0015389  0.1590369  0.6656394

Coefficients:
              Estimate  Std. Error t-value  Pr(>|t|)
(Intercept)  5.45373133  0.70949139  7.6868 1.616e-12 ***
inv          0.03402271  0.01204384  2.8249 0.0053520 **
lni         -0.99010401  0.29615186 -3.3432 0.0010383 **
p4polity2    0.02263520  0.00653562  3.4634 0.0006898 ***
open         0.00211020  0.00077192  2.7337 0.0069917 **
lfexp        0.08507361  0.00614246 13.8501 < 2.2e-16 ***
ltoted       0.04920239  0.09668027  0.5089 0.6115330
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    44.659
Residual Sum of Squares: 9.2507
R-Squared:      0.79286
Adj. R-Squared: 0.78484
F-statistic: 98.8801 on 6 and 155 DF, p-value: < 2.22e-16
```

- *Process:*
  - Feature elimination: regress only on significant variables.
  - Pooled specification adds an intercept (necessary for `predict`).
- *Results:*
  - Full dataset: MSE = 0.392
  - OECD dataset: MSE = 0.070 (!)



OECD Model: Observed vs. Predicted Values for logincome_pc

Conclusions

1. Failed to reproduce ↑ inequality ⇒ ↓ growth. We found ↑ inequality ⇒ ↑ growth (full dataset only).
2. Reproduced ↑ health, ↑ education ⇒ ↑ growth.
3. Evidence for "Washington Consensus:" ↑ democracy, ↑ openness ⇒ ↑ growth.
4. Model performed better for OECD dataset than full dataset. Better data?
5. Why did we fail to replicate Berg et al.'s primary result (↑ inequality ⇒ ↓ growth)? Possibilities:
    a. Different regression specification; Berg et al. used sGMM.
    b. Replacing NAs changed the data too much.
        - Replacing NAs assumes that values for a country are normally distributed over time. In retrospect, this is probably a bad assumption for many variables (e.g., life expectancy tends to improve over time).