gdppc\_and\_inequality

Jon Minton

## Background

I’m interested in the extent to which inequality within country can explain differences between the observed life expectancy in a population, and the life expectancy predicted by the population’s GDP per capita.

Danny Dorling kindly supplied/sourced some data that should allow this area can be investigated

## Data and packages

library(here)

here() starts at C:/Users/Jon Minton/repos/gdppc\_and\_inequality

library(tidyverse)

── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──

✔ ggplot2 3.3.6 ✔ purrr 0.3.4  
✔ tibble 3.1.7 ✔ dplyr 1.0.9  
✔ tidyr 1.2.0 ✔ stringr 1.4.0  
✔ readr 2.1.2 ✔ forcats 0.5.1

── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
✖ dplyr::filter() masks stats::filter()  
✖ dplyr::lag() masks stats::lag()

dta\_gdp <- readxl::read\_excel(  
 "data-raw/Jon\_Data\_2.xlsx",   
 sheet = "GDP",   
 range = "C5:AG164"  
 ) %>%   
 rename(country = `Row Labels`) %>%   
 pivot\_longer(  
 -country,   
 names\_to = "year",   
 values\_to = "gdp\_pc"  
 )  
  
dta\_gini <- readxl::read\_excel(  
 "data-raw/Jon\_Data\_2.xlsx",   
 sheet = "GINI",  
 range = "E5:AZ164"  
 ) %>%   
 rename(country = `Row Labels`) %>%   
 pivot\_longer(  
 -country,   
 names\_to = "year",   
 values\_to = "gini"  
 ) # much to interpolate here   
  
dta\_e0 <- readxl::read\_excel(  
 "data-raw/Jon\_Data\_2.xlsx",  
 sheet = "Life Expectancy",   
 range = "G5:CE164"  
) %>%   
 select(-c("Notes", "Country code", "Type", "Parent code")) %>%   
 rename(country = `Region, subregion, country or area \*`) %>%   
 pivot\_longer(-country, names\_to = "year", values\_to = "e0")

The next tasks are:

1. ☒ Interpolate values for dta\_gini
2. ☒ Join the three datasets for common countries

## Interpolate gini

# test\_df1 <- tibble(  
# year = 1992:2010,  
# gini = c(rep(NA, 5), 20, rep(NA, 4), 31, rep(NA, 4), 28, rep(NA, 3))  
# )  
#   
# approx(test\_df1$year, test\_df1$gini, test\_df1$year)  
   
do\_interpolation <- function(df){  
 approx(df$year, df$gini, df$year)$y  
}  
# countgaps(tmp)  
  
dta\_gini\_interp <-   
 dta\_gini %>%   
 group\_by(country) %>%   
 arrange(year) %>%   
 nest() %>%   
 mutate(gini\_interp = map(data, possibly(do\_interpolation, otherwise = NULL))) %>%  
 filter(!is.null(gini\_interp)) %>%   
 mutate(data\_combined = map2(data, gini\_interp, ~ .x %>% mutate(gini\_interp = .y))) %>%   
 select(country, data\_combined) %>%   
 unnest(cols = c(data\_combined)) %>%   
 select(country, year, gini = gini\_interp) %>%   
 ungroup()  
   
dta\_gini\_interp

# A tibble: 7,473 × 3  
 country year gini  
 <chr> <chr> <dbl>  
 1 Albania 1967 NA  
 2 Albania 1969 NA  
 3 Albania 1971 NA  
 4 Albania 1974 NA  
 5 Albania 1975 NA  
 6 Albania 1978 NA  
 7 Albania 1979 NA  
 8 Albania 1980 NA  
 9 Albania 1981 NA  
10 Albania 1982 NA  
# … with 7,463 more rows

So we can now combine the three values

dta\_combined <-   
 dta\_gini\_interp %>%   
 inner\_join(dta\_e0) %>%   
 inner\_join(dta\_gdp) %>%   
 filter(complete.cases(.))

Joining, by = c("country", "year")  
Joining, by = c("country", "year")

dta\_combined

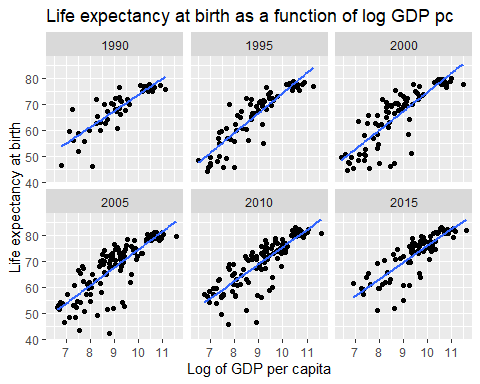
# A tibble: 2,663 × 5  
 country year gini e0 gdp\_pc  
 <chr> <chr> <dbl> <dbl> <dbl>  
 1 Albania 1996 27 74.6 4925.  
 2 Albania 1997 27.8 73.9 4415.  
 3 Albania 1998 28.6 75.0 4835.  
 4 Albania 1999 29.4 75.2 5493.  
 5 Albania 2000 30.1 75.4 5912.  
 6 Albania 2001 30.9 75.6 6462.  
 7 Albania 2002 31.7 75.9 6776.  
 8 Albania 2003 31.3 76.1 7178.  
 9 Albania 2004 31.0 76.4 7605.  
10 Albania 2005 30.6 76.6 8067.  
# … with 2,653 more rows

## Analysis proper

First let’s look at relationship between e0 and gdp\_pc at five year intervals

dta\_combined %>%   
 filter(year %in% seq(1990, 2015, by = 5)) %>%   
 ggplot(aes(x = log(gdp\_pc), y = e0)) +   
 facet\_wrap(~year) +   
 geom\_point() +   
 stat\_smooth(method = "lm", se = FALSE) +   
 labs(x = "Log of GDP per capita", y = "Life expectancy at birth",  
 title = "Life expectancy at birth as a function of log GDP pc")

`geom\_smooth()` using formula 'y ~ x'



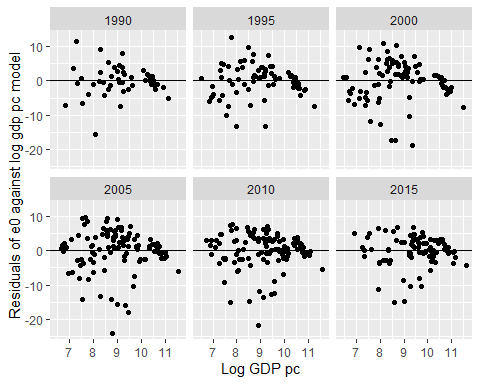
So what are the residuals?

mod\_gdp\_pc\_resid <-   
 dta\_combined %>%   
 filter(year %in% seq(1990, 2015, by = 5)) %>%   
 group\_by(year) %>%   
 nest() %>%   
 mutate(lm\_e0\_gdp = map(data, ~lm(e0 ~ log(gdp\_pc), data = .))) %>%   
 mutate(dta\_aug = map2(lm\_e0\_gdp, data, broom::augment)) %>%   
 select(year, dta\_aug) %>%   
 unnest(cols = dta\_aug)   
   
mod\_gdp\_pc\_resid

# A tibble: 537 × 11  
# Groups: year [6]  
 year country gini e0 gdp\_pc .fitted .resid .hat .sigma .cooksd  
 <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
 1 2000 Albania 30.1 75.4 5912. 65.0 10.4 0.0114 5.74 0.0188   
 2 2000 Algeria 32.9 70.5 8710. 67.8 2.66 0.0116 5.84 0.00124   
 3 2000 Angola 52 46.0 4728. 63.3 -17.3 0.0123 5.54 0.0557   
 4 2000 Argentina 51.1 73.9 18625. 73.4 0.507 0.0183 5.85 0.0000723  
 5 2000 Armenia 35.8 70.6 4048. 62.2 8.44 0.0133 5.78 0.0144   
 6 2000 Australia 33.4 79.6 38462. 78.8 0.872 0.0325 5.85 0.000390   
 7 2000 Austria 28.8 78.2 46551. 80.2 -1.93 0.0375 5.85 0.00222   
 8 2000 Azerbaijan 36.2 64.9 4063. 62.2 2.68 0.0133 5.84 0.00145   
 9 2000 Bangladesh 33.4 65.8 1938. 56.8 9.02 0.0229 5.77 0.0289   
10 2000 Belarus 31.2 69.2 8053. 67.2 1.93 0.0114 5.85 0.000646   
# … with 527 more rows, and 1 more variable: .std.resid <dbl>

What does the residual pattern look like ?

mod\_gdp\_pc\_resid %>%   
 ggplot(aes(log(gdp\_pc), .resid)) +   
 facet\_wrap(~year) +   
 geom\_point() +   
 geom\_hline(yintercept = 0) +   
 labs(x = "Log GDP pc", y = "Residuals of e0 against log gdp pc model")



So, for each year, which countries have the highest residuals?

mod\_gdp\_pc\_resid %>%   
 group\_by(year) %>%   
 arrange(desc(.resid), .by\_group = TRUE) %>%   
 top\_n(5)

Selecting by .std.resid

# A tibble: 30 × 11  
# Groups: year [6]  
 year country gini e0 gdp\_pc .fitted .resid .hat .sigma .cooksd  
 <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
 1 1990 China 32.2 68.0 1424. 56.5 11.5 0.0806 4.04 0.334   
 2 1990 Sri Lanka 32.4 71.9 3878. 62.7 9.18 0.0342 4.19 0.0814  
 3 1990 Costa Rica 45.3 76.6 9868. 68.6 8.04 0.0200 4.24 0.0354  
 4 1990 Tunisia 40.2 70.1 5432. 64.9 5.26 0.0258 4.33 0.0198  
 5 1990 Jamaica 41.1 72.3 8591. 67.7 4.58 0.0203 4.35 0.0117  
 6 1995 China 34.7 70.0 2391. 57.3 12.7 0.0254 4.44 0.0995  
 7 1995 Sri Lanka 35.4 72.3 4790. 62.6 9.62 0.0153 4.55 0.0337  
 8 1995 Tunisia 41.7 72.0 5931. 64.3 7.71 0.0140 4.60 0.0197  
 9 1995 Costa Rica 45.7 76.6 11295. 69.2 7.31 0.0151 4.61 0.0192  
10 1995 Honduras 55.5 67.1 4025. 61.3 5.84 0.0170 4.64 0.0138  
# … with 20 more rows, and 1 more variable: .std.resid <dbl>

So the top five (‘punching above weight’) by year are:

* 1990
  + China
  + Sri Lanka
  + Costa Rica
  + Tunisa
  + Jamaica
* 1995
  + China
  + Sri Lanka
  + Tunisia
  + Costa Rica
  + Honduras
* 2000
  + China
  + Albania
  + Tajikistan
  + Bangladesh
  + Armenia
* 2005
  + Solomon Islands
  + China
  + Tajikistan
  + Albania
  + Bangladesh
* 2010
  + Solomon Islands
  + Tajikistan
  + Albania
  + Nicaragua
  + Bangladesh
* 2015
  + Cabo Verde
  + Albania
  + Tajikistan
  + Bangladesh
  + Honduras

Conversely the bottom five (‘punching below weight’)

mod\_gdp\_pc\_resid %>%   
 group\_by(year) %>%   
 arrange(desc(.resid), .by\_group = TRUE) %>%   
 top\_n(-5)

Selecting by .std.resid

# A tibble: 30 × 11  
# Groups: year [6]  
 year country gini e0 gdp\_pc .fitted .resid .hat .sigma .cooksd  
 <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
 1 1990 Luxembourg 26.8 75.6 67858. 80.7 -5.06 0.0795 4.34 0.0633  
 2 1990 Madagascar 45.6 51.7 1853. 58.1 -6.42 0.0653 4.30 0.0810  
 3 1990 Uganda 43.4 46.4 908. 53.6 -7.21 0.112 4.26 0.194   
 4 1990 Botswana 58.3 60.5 9074. 68.1 -7.53 0.0201 4.26 0.0313  
 5 1990 Nigeria 43.2 46.0 3260. 61.7 -15.6 0.0399 3.74 0.278   
 6 1995 Luxembourg 28 76.7 76972. 84.0 -7.30 0.0634 4.61 0.0887  
 7 1995 Guinea-Bissau 41.8 49.0 2088. 56.3 -7.30 0.0284 4.61 0.0369  
 8 1995 Zambia 49.7 45.6 1909. 55.6 -10.0 0.0306 4.53 0.0750  
 9 1995 Botswana 61.7 55.1 9935. 68.3 -13.2 0.0143 4.42 0.0588  
10 1995 Nigeria 50.2 45.5 2902. 58.8 -13.3 0.0217 4.41 0.0923  
# … with 20 more rows, and 1 more variable: .std.resid <dbl>

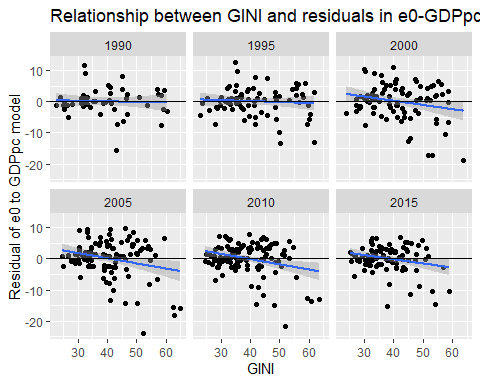
* 1990
  + Luxembourg
  + Madagascar
  + Uganda
  + Botswana
  + Nigeral
* 1995
  + Luxembourg
  + Guinea-Bissau
  + Zambia
  + Botswana
  + Nigeria
* 2000
  + Zambia
  + Nigeria
  + Eswatini
  + Angola
  + Botswana
* 2005
  + Lesotho
  + Namibia
  + South Afria
  + Botswana
* 2010
  + South Africa
  + Namibia
  + Nigeria
  + Lesotho
  + Eswatini
* 2015
  + Botswana
  + Namibia
  + Lesotho
  + Eswatini
  + Nigeria

## Model of residuals

Now the final part: To what extent are these residuals explained by inequalities within country?

mod\_gdp\_pc\_resid %>%   
 ggplot(aes(gini, .resid)) +   
 facet\_wrap(~year) +   
 geom\_point() +   
 stat\_smooth(method = "lm") +   
 geom\_hline(yintercept = 0) +   
 labs(  
 x = "GINI", y= "Residual of e0 to GDPpc model",   
 title = "Relationship between GINI and residuals in e0-GDPpc model"  
 )

`geom\_smooth()` using formula 'y ~ x'



So, there is a relationship in the expected direction, which on the one hand is fairly modest, but on the other appears to have become more prominent in the 21st century than in the 20th century