Mini-project 1

S681

Upload your draft project through the Assignments tab on Canvas by 11:59 pm, Wednesday 13th February. Final submissions will be due at a later date to be announced.

Work in groups of 3 or 4 (4 is recommended.) Form groups using the "Mini-project 1 groups" tab on Canvas.

Income inequality

Income inequality in the U.S. and in other countries is a major research topic in the social sciences. A well-regarded collection of data is maintained by Branko Milanovic of the Stone Center on Socio-Economic Inequality. We will primarily deal with two data sets that we'll call **WPID** and **Ventile**.

WPID is the Lakner-Milanovic World Panel Income Distribution data set (Stata file `LM_WPID_web_2.dta`.) Potentially useful variables include:

- `country`: country name
- `contcod`: 3-letter country code
- `bin_year`: the year for which incomes are estimated (1988 to 2008.) For technical reasons, use this rather than the `year` variable.
- `group`: the income decile group the estimate is for, where a decile is 10% of the population. "1" means the 10% of individuals in the country with the lowest income, while "10" means the 10% of the individuals in the country with the highest income.
- `RRinc`: the per capita income of that decile in that country, in 2005 US dollars.
- `RRmean`: the mean per capita income of the whole country, in 2005 US dollars.

Ventile (Stata file `ventile_2011_for_release_LCU.dta`) contains more recent data (circa 2011.) Some important variables are:

- `contcod`: 3-letter country code
- `ventile`: the income ventile group the estimate is for, where a ventile is 5% of the population. "1" means the 5% of individuals in the country with the lowest income, while "20" means the 5% of the individuals in the country with the highest income.

- `ventile_income`: the per capita income of that ventile in that country. Important: This number is **NOT** adjusted for currency or inflation, so it is not directly comparable to the incomes in the WPID data set (or to other countries in the Ventile data set.)

One further data set, the World Income Distribution (**WYD**, Stata file `wyd_88_05_for_release.dta`) has been uploaded for your interest/reference, but you are not required to use this.

Full descriptions of the data sets can be found on Milanovic's website.

Questions to answer

1. **How has per capita income for each decile in the U.S. changed since 1988?**

   Use the WPID data set to answer this. Give both graphical and numerical results, and describe in words both the differences in income level between the deciles and the differences in changes in income for each decile.

2. **How does the present distribution of income, relative to a country's mean, differ between selected countries?**

   Use the Ventile data set for the most recent data. For this question, compare the United States (USA), the United Kingdom (GBR), Germany (DEU), and TWO other countries of your choice using the most recent available data. We know that (for example) relatively low-income people in the U.S. have higher incomes than relatively high-income people in Ghana, so you'll need to find some way of removing this effect. Find a clear way or ways to graphically display your results, and briefly write about each country in turn, explain how its distribution differs (or doesn't differ) from the others.

3. **Is the percentage of income earned by the top 5 percent in a country related to mean income in that country? What about the percentage of income earned by the bottom 5 percent?** If so, what's the relationship, and does the relationship have a simple explanation, such as regional differences? Are there any outliers that require special explanation?

   Again, the Ventile data set has the most recent data.

Write a PDF report of no more than six pages, including graphs, addressing these questions. The body of the report should *not* include code — it should be readable to someone who has never used R (generally you should not just copy-paste output.) Additional graphs for model checking can be placed in an appendix that does not count toward the six page limit and which probably no one will read.

Notes and hints

- There is not necessarily any single objectively right answer to any question in the project. However, some answers are better than others.
- Whenever you present something that isn't just raw data, carefully explain what it is and how it's calculated.
- Think about whether it's more appropriate to describe differences in absolute terms or in percentage terms (or both.)
- You'll be penalized for any P-value that appears in your report.

What to submit

- A PDF or other file containing your report.
- A .Rmd or other file containing your code.
- Any other supplementary files required to reproduce your work.

Grading

- Question 1: 5 points
- Question 2: 5 points
- Question 3: 10 points
- Communication: 10 points. Full credit for presentation requires a readable, informative, comprehensive, clearly labeled set of graphs, and a comprehensible write-up with few glaring spelling and grammatical errors that makes the main points of the analysis clear.