An Introduction to R
R Workshop 2: World Bank Data
QS311 Section E: Baker University
October 22th, 2019

Contents

1 Introduction

Exploratory Data Analysis with One and Two Variables
R Workshop in World Bank Poverty Data
QS311 Section E: Baker University

Goal: To explore poverty data from the World Bank in R.
In this workshop we will focus on creating and loading World Bank data sets into R as well as basic exploratory commands.

1. To access World Bank data we need to use the API (Application Programming Interface). R has a package for the World Bank API called *wbstats*. Recall: if you use a computer without the API, you must install the package with *install.packages("wbstats")*. Once you have installed the package, remember to run the library command to call it into your session on R with the *library(wbstats)* command.

   (a) Last time we looked at aid effectiveness, one of 21 topics the World Bank uses to categorize its data. For this workshop we want to focus on global poverty, a different topic with its own set of indicators. The World Bank has several different variables which help provide insight into global poverty including income by quintile, the gini coefficient (a measure of inequality), and various measures of consumption.

2 Lab Procedures

2.1 Obtaining the Variables of Interest

1. First, we need to get the indicators of interest into R. We will use the same command as we did with aid effectiveness changing the indicator codes. For instance, for the gini coefficient, we can use the following code:

   ```
   gini <- wb(country = "countries_only", indicator =
   "SI.POV.GINI", startdate = 2000, enddate = 2017, removeNA = FALSE)}.
   ```

   (a) The "*countries_only*" command means that all countries will be included and no aggregate data will be added.

   (b) The indicator code is World Bank specific for their data on the gini coefficient.

   (c) The removeNA=FALSE command means that we will have a balanced panel, rather than a dataset only with non-missing datapoints included.

2. Run this command for each of the additional indicators of interest for this lab:

   - SI.DST.FRST.20 (*income*20)
   - SI.DST.02ND.20 (*income*40)

- SI.DST.03RD.20 (*income*60)
- SI.DST.04TH.20 (*income*80)
- SI.DST.05TH.20 (*income*100)
- SI.SPR.PC40 (*cons*40)
- SI.SPR.PCAP (*cons_tot*)

2.2 Obtaining a Panel with All Variables

We now want to take each of these variables and construct a panel dataset where all of the variables are in a single data frame. We can take advantage of the fact that the World Bank data is organized alphabetically and by year and that it has the same length. This means that we can directly concatenate the columns we want into a new data frame using the *data.frame* command.

- Note: this is a shortcut from what we did last time with the merge command. Merge is typically more helpful because it works for data frames that are different lengths as well.
- The structure of the command will be something like:

```
panel <- data.frame(gini$iso3c, gini$date, gini$value,
income20$value, income40$value, income60$value, income80$value,
income100$value, cons40$value, $cons_tot$value).
```

- Here, the name before the $ is the data frame. If you have chosen different names for your data frames when loading in the World Bank variables, you will need to change your code accordingly. The name after the $ is the column inside the data frame. You can check to make sure that you have the right name for the column by clicking on the data frame in R and looking at the column title.

We now want to change the names to something descriptive and aesthetically sensible. We can do this with the names() command.

- ```
  names(panel) <- c("country_code", "year", "gini",
  "income20","income40", "income60", "income80", "income100", "cons40",
   "cons_tot")
  ```
- The part of the command after the assignment code combines the names as titles in the data frame. This happens by putting them in order and and covering every column.

Factor Variable to Numeric Variable
There are many different types of variable categories: string, numeric, factor, long, float, etc. Your year variable may be a factor variable and you will want to have a numeric year variable for some plots.
  ```

1. Create a numeric year variable with the following code:

```
panel$year_num <- as.numeric(levels(panel$year))
```

You may also want to discretize the year so that you have five year increments. This can be accomplished with the following code:

1. `panel$five_year <- cut(panel$year_num, c(1999,2005,2010,2015,2018))`

   - This structure is something we have seen before. First, you are adding a new variable to the dataframe called $five\_year$.
   - Second, you are dividing the variable $year\_num$ into discrete increments between 1999 and 2005, 2005 and 2010, and so forth.

2. You will want to merge in the WB classifications dataset again. Recall that you did this in the previous R workshop. Import the WB classifications data into the R data environment. Then use the merge command to combine the panel dataframe and the WB classifications dataframe. The code should look like the following:

```
final_panel <- merge(panel, wb_classifications, by.x = "country_code",
by.y = "Code")
```

2.3 Variables

The variables which result are:

Variable	**Description**
$country\_code$	The ISO3 Code for Country
year	Year
$income20$	Share of income of the bottom quintile
$income40$	Share of income of the second quintile
$income60$	Share of income of the third quintile
$income80$	Share of income of the fourth quintile
$income100$	Share of income of the top quintile
$cons40$	Mean consumption, bottom 40% (2011 PPP $ per day)
$cons\_tot$	Mean consumption, population (2011 PPP $ per day)
$year\_num$	A number, 1 through 18, corresponding with the years
$five\_year$	Discretized five year gaps
$name$	Country Name
$region$	Region of the World
$income$	The level of income group of a country

3 Analyzing the Data

3.1 Questions

Please answer the following questions and submit them on Moodle for Homework Assignment 5 (Due: October 29).

1. First, obtain distribution information on each of the primary variables (each income quintile, the gini coefficient, and the two consumption variables). What can you say about the skewness?

2. Create a fitted line for each of the income quintiles. First create them on separate graphs. Then, put each fitted line on a single graph. What is striking about this graph?

3. Compare the fitted lines to the scatter plots for each of the income quintiles. What is apparent in the scatter plot that is obscured in the fitted line?

4. Create boxplots over five-year periods and compare these to the fitted lines. Do these provide any useful context for the change in share of income by quintile over time? What about the dispersion of income share by quintile?

5. Look at the fitted lines for income by region. What variation do you see? How does this inform the reading from the textbook and the Singer (1972) article on poverty?

6. Compare the fitted lines you obtained for income quintiles with fitted lines for consumption. Do you see any different patterns? Regional variation? Use this to critique and discuss the textbook reading and the Singer (1972) article.