An Introduction to R
R Workshop 1: World Bank Data
QS311 Section E: Baker University
October 15th, 2019

Contents

1 Introduction

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

Goal: To explore global governance 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.

This workshop makes use of the *ggplot* command in R using a variety of different geoms. In order to obtain commands in R, you must first use the library in which they are stored. For *ggplot*, the corresponding library is *ggplot2*. Thus, you must run *library(ggplot2)* before you can do any plotting.

2 Lab Procedures

2.1 Obtaining the Variables of Interest

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*. To use it, 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.

   (a) World Bank data is separated into categories of country, indicator, and year. We can call into R this data using the *wb* command available through the wbstats package specifying which countries, indicators, and years we want.

   (b) One example might be to look at the topic category Aid Effectiveness. There, variables include net migration and net overseas development assistance. Both of these are important topics in the global governance literature which have many ethical considerations.

Calling in the Indicators

1. First, we need to get the indicators of interest into R. For instance, to obtain GDP, use the following code:

```
gdp <- wb(country = "countries_only", indicator = "NY.GDP.MKTP.CD",
 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 GDP.

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

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

   - SP.POP.TOTL (*pop*)
   - DT.ODA.ODAT.PC.ZS (*net_oda*)
   - NY.GDP.MKTP.KD.ZG (*gdp_growth*)
   - SM.POP.NETM (*net_migration*).

2.2 Obtaining a Panel with All Variables

We now want to take each of these variables and construct a panel dataset where the four of them are merged together. To do so, you can choose the columns from each dataframe that you want. For us, we want our new dataframe to have the country code, year, and each of the value columns that represent the five different variables. To do this, we can use the $data.frame$ command in R.

The following code will allow you to create a panel with the five variables of interest:

1.
```
panel $<-$ data.frame(gdp$iso3c, gdp$date, gdp
$value, pop$value, net_oda$value, gdp_growth$value,
net_mig$value)
```

   - First, we are calling this new dataframe panel.
   - Second, the data.frame command means we are constructing a new dataframe in R using data from existing dataframes.
   - Inside the dataframe command, the item before the dollar sign is the dataframe and the item after the dollar sign is the column of interest.

2. We now want to name each of the variables as their names were changed when the data frame command was applied. We can do this with the following code:

```
 names(panel) <- c("country_code", "year", "gdp","pop",
"net_oda", "gdp_growth", "net_mig")
```

   - We now have a dataframe with the variables that we want. However, we are missing some interesting discrete variables which could help visualize the data. In particular, we are missing a regional designation for each country and an income group designation for each country.
   - Use the World Bank Classifications dataset $wb\_classifications$ provided on Moodle and import it to the R environment.

- The data.frame command will not work because the length of the panel dataframe and the wb classifications dataframe is not the same. To combine these, we need to merge the two dataframes and R gives us a merge command which can facilitate this. The code looks 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
gdp	GDP in 2010 USD
pop	Population
gdp_growth	GDP Growth
net_mig	Net Migration
net_oda	Net Overseas Development Assistance
name	Country Name
region	Region of the World
income	The level of income group of a country

3 Analyzing the Data

R has numerous commands which allow you to manipulate the data in ways which can help you better understand it and analyze it. Below are some commands which may be helpful to think about the questions which follow.

- You will want to know many of the different geoms that accompany the ggplot command: *geom_boxplot*, *geom_point*, *geom_smooth*, and *geom_hist*. For instance, to create a boxplot using the final panel created above, you might use something like the following:

```
ggplot(data = final_panel)+
     geom_boxplot(aes(Region, net_mig))
```

- You may also want to create variables based on existing variables in the dataframe. For instance, you may want GDP per capita and net migration per capita. To do this, you must create a new variable in R and divide the *gdp* or *net_mig* variables respectively by pop. For instance:

```
final_panel$gdp_pc <- final_panel$gdp/final_panel$pop.
```

- When you have variables with particularly large numbers it is often useful to take the logarithm and transform the data down to smaller numbers. This can be done with the following code:

```
final_panel$ln_pop <- log(final_panel$pop)
```

- Summary statistics help you think about the variables you are using. There are a variety of ways to obtain summary variables in R. In this assignment, we will make use of the describe command in the Hmisc package. First run the library(Hmisc) command and remember that you may have to do install.packages("Hmisc") first if the library command does not work. Then, you can run code like to following to obtain summary statistics:

```
describe(final_panel$gdp_growth)
```

  This will give you the number of observations, the number of observations that are missing (reported as NA in World Bank data), the mean, and different percentiles.

Data Analysis Tip: The default histogram counts the number of instances in each bin. If you want to get the fraction of the overall percentage represented by each bin, you should use $stat\_bin(aes(y = ..count../sum(..count..)))$ instead of $geom\_histogram()$.

3.1 Questions

Please answer the following questions and submit them on Moodle for Homework Assignment 4 (Due: October 22).

1. First, obtain distribution information on each of the primary variables (net oda, net mig, ln pop, gdp growth, and gdp per capita). What can you say about the location of each of the variables? (Relation of the mean and the median? Skewness?)
   (a) For question 1, please submit a graph for each primary variable.
   (b) Include the code that you used to obtain each of those graphs.
   (c) Then, describe the location of each variable.
2. Obtain summary statistics on the five main variables (net oda, net mig, ln pop, gdp growth, and gdp per capita). Use the describe command from the Hmisc library and report in a table in your submission the number of observations, the number of missing observations and the mean for each variable.
3. Explore the relationship between *net_oda* and *gdp_growth*.
   (a) What does the scatter plot indicate about this relationship? Put *net_oda* on the x-axis and *gdp_growth* on the y-axis.
   (b) Put a fitted line over the scatterplot. Does this indicate further investigation might be warranted? If so, why? If not, why not?

4. If more overseas development aid lifts growth rates, it is important to think about the relationship between gdp per capita and growth rates.
   (a) Use a scatterplot and fitted line to see the relationship between gdp per capita and the growth rate of gdp?
   (b) Use the information about the relationship between *net_oda* and *gdp_growth* and the relationship *gdp_growth* and gdp per capita to comment on the William Easterly article about the efficacy of overseas development assistance.

5. Find another indicator from the Aid Effectiveness topic in the World Bank data and add it to the panel dataframe. **This Link Takes You to World Bank Indicators**. You can also find the indicators in the Aid Effectiveness excel spreadsheet provided on Moodle.

6. Obtain a histogram of the new variable and compute summary statistics like above.

7. Create one informative plot for the variable and include it with a description in your homework submission.