An Introduction to R
R Workshop 5: World Bank Data
QS311 Section E: Baker University
November 11th, 2019

Contents

1 Introduction **2**

2 Lab Procedures **2**
2.1 Obtaining the Variables of Interest 2
2.2 Obtaining a Panel with All Variables 3
2.3 Variables 3
2.4 Mapping 4

3 Analyzing the Data **5**
3.1 Questions 5

1 Introduction

Mapping One Variable
R Workshop in World Bank Global Environmental Data
QS311 Section E: Baker University

Goal: To explore global environmental 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 some of the basic mapping capabilities in R.

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.

2 Lab Procedures

2.1 Obtaining the Variables of Interest

Loading World Bank data into R: We will construct a panel dataset from several World Bank indicators on the global environment. Once we analyze these indicators we will think about the bioethical implications of the information contained in this data.

Steps:

1. Remember to run the library for wbstats: **library("wbstats")**.
2. Now, read in the data for each indicator code. For example, to get the CO2 emissions indicator and call the resulting data frame *co2_emissions*, you would use the following code (notice that the date is different from before):

```
co2_emissions <- wb(country = "countries_only", indicator =
 "EN.ATM.CO2E.PP.GD", startdate = 1960, enddate = 2017,
removeNA = FALSE)
```

   The following are the indicator codes that you need. Call each data frame a different name. In parentheses are the names that I used for each data frame in my code.

   - EN.ATM.CO2E.PP.GD (*co2_emissions*)
   - EG.USE.COMM.CL.ZS (*alternative_energy*)
   - AG.LND.PRCP.MM (*ave_precipitation*)

- EN.POP.EL5M.ZS ($pop\_below5m$)
- AG.YLD.CREL.KG ($cereal\_yield$)
- AG.LND.FRST.ZS ($forest\_area$)

2.2 Obtaining a Panel with All Variables

1. Create a panel comprising each of these indicators. Remember that we can combine individual columns from different data frames into a single data frame with the variables we want. The following code enables this:

```
panel <- data.frame(co2_emissions$iso3c, co2_emissions$date,
co2_emissions$value, alternative_energy$value, ave_precipitation$value,
pop_below5m$value, cereal_yield$value, forest_area$value)
```

2. As a result of using the **data.frame** command, the variable names at the top of each column are a concatenation of the original variable name and the original data frame. We want to rename these. The following shows the names that I chose.

```
names(panel) <- c("country_code", "year", "co2_emissions",
"alternative_energy", "ave_precipitation", "pop_below5m",
"cereal_yield", "forest_area")
```

3. The final step requires you to load in the $wb\_classifications$ data. Remember that you have used this in previous R Workshops for the region and income variables. You need to import the excel file so that the data is in your R environment. Then, based on a variable that identifies rows in both the *panel* and the $wb\_classifications$ data frame, you should merge the two. The code I used looks like the following:

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

2.3 Variables

The full dataset will have many variables based on the merge with the WB classifications data and, later, with the world data. However, the primary variables of interest are listed below:

Table 1. Summary of Variables

Variable	Description
pop_below5m	Population living in areas where elevation is below 5 meters (% of total population)
co2_emissions	CO2 emissions (kg per PPP $ of GDP)
alternative_energy	Alternative and nuclear energy (% of total energy use)
ave_precipitation	Average precipitation in depth (mm per year)
cereal_yield	Cereal yield (kg per hectare)
forest_area	Forest area (% of land area)
name	Country Name
region	Region of the World
income	The level of income group of a country

2.4 Mapping

It is common to map environmental data and R makes it possible to map data using centroids. Centroids are latitude and longitude coordinates of the center of a particular area. In our case, we will use centroids for the countries of the world. In order to do this, we will need to install several packages that are required to map in R. You can install multiple packages at once in the following manner:

```
install.packages(c("cowplot", "googleway", "ggplot2", "ggrepel", "rgeos",
           "ggspatial", "libwgeom", "sf", "rnaturalearth", "rnaturalearthdata"))
```

Remember to read in the libraries once these packages are installed. In addition to reading in the libraries, you should set the theme of the graphs. The *theme_bw*() is common for graphing in R.

```
library("ggplot2")
theme_set(theme_bw())
library("sf")
library(rgeos)
library("rnaturalearth")
library("rnaturalearthdata")
```

In order to map, R needs a certain set of data on geography. We will be using country borders and that data requires latitude and longitude coordinates for the centroid of each country. The command that obtains these data points is *ne_countries* and the full code is below.

```
world <- ne_countries(scale = "medium", returnclass = "sf")
```

In order to make use of the mapping capability, you need to combine the *world* dataframe with the *final_panel* dataframe. This can be accomplished with another **merge** command.

```
merge <- merge(world, final_panel, by.x = "adm0_a3", by.y = "country_code")
```

Below is an example which uses one of the variables of interest for the environment under examination today.

```
ggplot(subset(final_panel, year %in% c("2016"))) +
  geom_sf(aes(fill = cereal_yield)) +
  scale_fill_viridis_c(option = "plasma", na.value = "grey50")
```

3 Analyzing the Data

3.1 Questions

1. Create maps for each of the six primary variables in three time periods. This requires you to subset the data on year. Recall that in the poverty code data you needed to subset based on region. The idea here is similar so you can make use of that code if you do not remember the format to subset.

   For cereal and forest area compare 2000 and 2016.
   For alternative energy, co2 emissions, and pop below 5m compare 2000 and 2010.
   For average precipitation, compare 2014 with 1997.

   What change over time do you see in these maps?

2. Run regressions for each of these variables against time to see what kind of time trend exists. Report the results for the primary coefficient in a table and compare them with the maps that you created. Can you see the increase or decrease in the variables?

3. Generate boxplots for each of the variables in each of the time periods that you examined. (again, you will need to subset the data on year). Describe the geographic variation that you see? Do particular regions seem to drive the results for the regressions?

4. Find one additional variable of interest to you from the Climate Change topic excel file located on Moodle. Go through the steps for questions 1-3 with this variable.