An Introduction to R

R Workshop 4: World Bank Data

QS311 Section E: Baker University

November 5th, 2019

Contents

1 Introduction

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

Goal: To explore bioethics using health 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.

2 Lab Procedures

2.1 Obtaining the Variables of Interest

Setting up the data:
Loading World Bank data into R: We will construct a panel dataset from several World Bank indicators on global health. 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 immunizations indicator and call the resulting data frame immunizations, you would use the following code:

   ```
   immunizations <- wb(country = "countries_only", indicator = "SH.IMM.MEAS",
   startdate = 2000, enddate = 2017, removeNA = FALSE)
   ```

   The following are the additional indicator codes that you need. Call each data frame a different name. In parentheses are suggested names for each dataframe.

   - SP.POP.BRTH.MF (*birth_ratio*)
   - SP.UWT.TFRT (*contraception*)
   - SH.DTH.IMRT (*infant_deaths*)

- SH.DYN.AIDS.ZS (*hiv*)
- SH.MLR.INCD.P3 (*malaria*)
- SP.POP.TOTL (*pop*)
- NY.GDP.MKTP.KD.ZG (*gdp_growth*)
- NY.GDP.MKTP.CD (*gdp*)
- SI.SPR.PCAP (*cons_tot*)

2.2 Obtaining a Panel with All Variables

3. 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(contraception$iso3c, contraception$date, contraception
$value, birth_ratio$value, infant_deaths$value, hiv$value, immunizations
$value, malaria$value)
```

4. 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", "contraception", "ratio",
"infant_deaths", "hiv", "immunizations", "malaria")
```

5. 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")
```

6. You will want to create gdp per capita by dividing gdp by the population. Recall that you can create a new variable in the data frame in the following manner:

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

2.3 Variables

1. Primary Variables
   In this workshop we want to focus on global health and its bioethical implications. Following literature on global health and human capital across countries, we use World Bank indicators on unmet contraception needs, the ratio of male to female births, infant deaths, hiv prevalence, immunization against measles, and incidence of malaria.

2. Outcome Variables of Interest We will look at the impact of the global health variables on GDP per capita, GDP growth, and consumption.
   (a) *GDP per capita*: Gross Domestic Product per person living in the country.
   (b) *GDP growth*: growth rate (in percent) of GDP per capita.
   (c) *consumption*: average individual daily consumption converted into US dollars.

   These variables are described in Table 1.

Table 1. Summary of Variables

Variable	**Description**
Panel Variables	
country_code	The ISO3 Code for Country
year	Year, 2000 - 2017
Human Capital and Health	
hiv	Prevalence of HIV, total (% of population ages 15-49)
malaria	Incidence of malaria (per 1,000 population at risk)
immunizations	Immunization, measles (% of children ages 12-23 months)
Family Planning	
infant_deaths	Number of infant deaths
ratio	Sex ratio at birth (male births per female births)
contraception	Unmet need for contraception (% of married women ages 15-49)
Outcome variables	
gdp_growth	GDP growth (annual %)
gdppc	GDP per capita, PPP (current international $)
consumption	Survey mean consumption or income per capita, total population (2011 PPP $ per day)

2.4 Regressions

We want to understand the impact that family planning and human capital through health has on basic macroeconomic variables like GDP per capita, GDP per capita growth, and consumption. We can do this with boxplots, fitted lines, and scatterplots. It is also possible to look at these relationships with a regression.

A regression is a statistical concept with a very simple intuition. We make the simplifying assumption that there is a somewhat linear relationship between two variables. For instance, as the access to immunizations increases, people become healthier and work more, increasing GDP per capita. This relationship can be understood with a regression which fits a line to the data and gives you a coefficient estimate.

The coefficient is called beta, or β, and can be interpreted as the magnitude of the relationship between the dependent variable (e.g. GDP per capita) and independent variable (e.g. immunizations). So, for instance, we can run a regression in R looking at immunizations as our independent variable and GDP per capita as our dependent variable. Doing this obtains a $\beta = 443.4$. This means that each additional percentage of children ages 12-23 who are vaccinated against measles, is correlated with a corresponding 443.4 current international $ increase in GDP per capita.

The code for this in R looks like the following:

```
reg1 <- lm(final_panel$gdppc ~ final_panel$immunizations)
reg1
```

The **lm()** command stands for linear model. We are defining reg1, the name we are calling this regression, as the linear model which relates immunizations with GDP per capita. Typing reg1 into the command line or running it from your do file will show you the coefficients that are stored in the regression. This is where you can find your β.

3 Analyzing the Data

3.1 Questions

Please remember to submit in your homework all graphs generated in the questions.

1. Create scatterplots for each of the human capital and health variables using GDP per capita as your y-variable. Describe the correlation that you see for each of the three scatterplots.

2. Create scatterplots for each of the human capital and health variables using GDP growth as your y-variable. Describe the correlation that you see for each of the three scatterplots.

3. Obtain regression estimates for the impact of hiv, malaria, and immunizations on GDP per capita and GDP growth. Create a table with each coefficient and interpret the impact of these variables on human capital.

4. Create scatterplots for each of the family planning variables using consumption as your y-variable. Describe the correlation that you see for each of the three scatterplots.

5. Obtain regression estimates for the impact of infant deaths, the birth ratio, and lack of access to contraception on consumption. Interpret the meaning of these coefficients.

6. Create boxplots to look at the family planning variables and the human capital variables by both income and region. Do these add any insight that you did not already see in the scatterplots and regressions?

7. Knowing that the level of health and ability to engage in family planning has real consequences for the economic well-being of both the society and the individual, what can we say about bioethics and the concern for healthcare provision?

8. Does the heterogeneous distribution in access to healthcare provision have any moral implications from the perspectives of the theories we have discussed in class (mainly utilitarianism, cosmopolitanism, and deontological ethics)?