Democracy and Redistribution

A long-standing debate in the social sciences is whether democracies redistribute more to the poor than autocracies. Research on this topic is challenging, however, due to the prevalence of missing data. Information about particular countries (and variables) are often missing, and this absence of data is not random. For example, autocratic countries are less likely to report their data to international institutions like the World Bank. Also, starting in the 1990s, countries have become better at both collecting and reporting data on different indicators such as economic growth or infant mortality. So if we just analyze data without taking these factors into account, we might bias our results. This exercise is based on the following prominent paper:

Ross, Michael (2006), "Is Democracy Good for the Poor", *American Journal of Political Science*, Vol. 50, No. 4, pp. 860 - 874.

Prior to Ross' work, the prevailing belief was that democracies redistributed to the poor more than autocracies. Ross's work challenged this belief. Specifically, Ross argued that previous studies had paid insufficient attention to differences between countries and time trends. Further, Ross argued that their analysis did not address the problem of missing data.

Below you will find a dictionary with the main variables in two datasets we analyze:

World Bank: `world_bank.csv`

Name	Description
`country_name`	Country name.
`country_code`	Country abbreviation.
`year`	Year.
`gdp_growth`	GDP growth rate (percentage).
`gdp_per_capita`	GDP per capita (2000 US$).
`inf_mort`	Infant mortality (deaths per 1000 children under 5).
`pop_density`	Population density (per sq. km).

Polity IV: `polity.csv`

Name	Description
`country`	Country name.
`ccode`	Country abbreviation.
`year`	Year.
`polity`	Polity Score. Ranges from -10 (most autocratic) to 10 (most democratic)

Question 1

Read in the two files, `world_bank.csv` and `polity.csv`. Discuss what the observations in each dataset are and report the number of observations in each. Additionally, calculate the proportion of missing data for each and every variable in the World Bank data. Which variables seem to be missing the most data? You can use a loop to answer this question or you can do it some other way. You may find that the function `is.na()` may be helpful. The function can take as an input a dataframe as well. Remember that `R` takes `TRUE` or `FALSE` statement as a binary variable. Specifically, a `TRUE` is equal to `1` and a `FALSE` is a `0`.

Question 2

Let's clean the data and prep it for merging. First, subset the `polity` data so it contains only years from 1970 to 2015 (make sure to include both 1970 and 2015). Second, let's keep only the columns we need: `scode`, `year`, `polity`. Let's rename the column `scode` to `country_code`, so we can merge this dataset with the World Bank dataset. Finally, merge the two datasets using both `country_code` and `year`. You may want to use the function `merge()`. How many observations are in this new merged dataset?

Question 3

Now we are going to investigate the pattern of missing data. In a linear regression in `R`, if any of the variables used are missing for an observation, that row will be deleted and not included in the analysis. This is a major problem in previous analyses, as pointed out by Ross (2006). As a basis for future questions, create a new column variable in the merged dataset called `missing` which has a value of `1` if any of the variables in your merged dataset are missing, and `0` otherwise. **Hint:** Using the function `ifelse()`, and `apply(x, 1, anyNA)` may be helpful. The function `apply()` will go row by row in dataset `x`, check if there are any NA's (hence `anyNA`) in any of the columns, and output `TRUE` if any of the entries are missing, and `FALSE` otherwise. These types of columns are usually called indicators: they indicate the presence of missing data.

Question 4

Let's visualize the pattern of missingness across time. Calculate the proportion of rows with missing data by year. Then plot that in a graph, with the years as the x-axis, and the proportion of missingness as the y-axis. Make sure to include informative titles and labels. How has the pattern of missing data evolved over time?

Question 5

Let's compare the polity scores of country-year observations with missing data and those without. Make a boxplot graph of the polity scores in the group of observations with missing data, and in the group without missing data. Using the `formula` argument of the `boxplot()` function may help you in placing these two boxplots in the same plot. Do countries with and without missing data differ in their polity scores?

Question 6

In his study, Ross analyzes whether a democracy is better for the poor by looking at infant mortality. The intuition is that countries that do more for the poor help decrease infant mortality by improving access to public healthcare, among other policies. First we will run a regression without taking into account between-country difference and time trends. In other words, we will not address the problems that Ross identifies in his study. In order to run this regression, we first use the log transformation of population density, infant mortality, and GDP per capita to address their skewness. Use the function `log()` to do so. With your merged data set, regress logged infant mortality on the following predictors: polity score, gdp growth, logged gdp per capita, logged population density. State the null hypothesis regarding the polity score. Interpret your point estimates, standard error, and p-value on the polity variable. **Hint:** For a model of the form $\ln(Y) = \beta_0 + \beta_1 X + \beta_2 Z + \epsilon$, where Y is our outcome, X is our covariate of interest, Z is all other covariates, and ϵ is the error, we need to undo the logarithm in order to correctly interpret β_1. If we change X by one unit (an increase in polity score of 1), this model expects our Y to change by $100 \times (\exp(\beta_1) - 1)$ percent. **Bonus question:** try to explain how this result is obtained).

Question 7

Let's rerun the statistical model presented by Ross (2006), slightly modified. Regress logged infant mortality on the following predictors: polity score, gdp growth, logged gdp per capita, and logged population density. Following Ross, we now add fixed effects by country and year. That is, we add a dummy variable for each country and a dummy variable for each year in order to adjust for any factors that are specific to each country and year. Interpret the results again and compare them to question 5. What are the differences between

the results of this model and those of the model in the previous question? Give a substantive interpretation. Finally, how do you think that the pattern of missing data may affect the results obtained in this question?