Democracy and Redistribution

Gov 50 Section 5

Introduction

A long-standing debate in the social sciences is: **Do democracies redistribute more to low-income citizens than autocracies?** In other words, do democracies tend to collect more taxes from the rich than from the poor, and also redistribute more of those taxes to the poor as social welfare? The conventional wisdom is that democracies redistribute more to the poor, because politicians are democratically elected and thus are more attentive to mass welfare. However, the following paper finds **statistical** evidence against this conventional wisdom:

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

Ross argued that previous studies had paid insufficient attention to the problem of **missing data**. Information about particular countries and variables is often missing, and this absence of data is often **not random**. For example, autocratic countries are less likely to report their statistics to international institutions like the World Bank, so researchers often lack complete data from autocracies. Yet, starting in the 1990s, countries have become better at both collecting and reporting data on different indicators such as economic growth or infant mortality. As such, the patterns of missing data depend on both regime type and time trends, and analyzing data without considering these factors might bias our conclusions.

Let's delve into Ross' data to understand better the issue of missing not at random. Below you will find a dictionary with the main variables in two datasets we will analyze:

World Bank: `wb.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 (current US$).
`infant_death`.	Infant mortality (deaths per 1000 children under 5).
`population_density`	Population density (per sq. km).
`budget_edu`	Education spending as percentage of government budget.
`budget_health`	Public health spending as percentage of government budget.

Polity IV: `polity.csv`

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

Question 1: Wide to Long Data

Before we analyze these two datasets, you'll notice that the `polity.csv` data is in the wide format, rather than the long format as shown by the table above. In the wide format, the rows correspond to countries, and the values across the columns in each row are the Polity scores for a country across all years. For data analysis, we usually want to use data in long format, such that we have a single column for each variable.

Now, load the two datasets and save them as `wb` and `polity`, respectively; then, use `pivot_longer()` to convert the `polity` dataset into long format with three columns: `year`, `country_code`, and `polity`. Be sure to save this back as the `polity` dataset.

Answer 1

```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
# Load the data
polity <- read_csv("data/polity.csv")
```

```
## Rows: 217 Columns: 195
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (195): year, AFG, ALB, ALG, ANG, ARG, ARM, AUL, AUS, AZE, BAD, BAH, BAV,...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
wb <- read_csv("data/wb.csv")
```

```
## Warning: One or more parsing issues, call `problems()` on your data frame for details,
## e.g.:
##   dat <- vroom(...)
##   problems(dat)
```

```
## Rows: 6422 Columns: 9
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): country_name, country_code
## dbl (6): year, gdp_growth, gdp_per_capita, population_density, budget_edu, b...
## lgl (1): infant_death
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
# Pivot
## Note: Identify the columns to pivot into long format using `!year`
```

```
polity <- pivot_longer(polity,
                       cols = !year,
                       names_to = "country_code",
                       values_to = "polity")
```

Question 2: Joining

Now that we have two long datasets, let's join them together. Two of the most common types of joins are `left_join()`, which keeps all rows in the "left-hand" data regardless of whether there is a match in the data being joined (the "right-hand" data), and `inner_join()`, which keeps only the rows that are present in both datasets.

Now, choose either `left_join()` or `inner_join()` to perform the joining. Before attempting either method, decide which type you think you should use and why, supposing that the `wb` data is the left-hand-side dataset. Keep in mind that all subsequent analysis here will look at the relationship between `polity` scores and various measures in `wb`.

Then, perform your chosen join, store the result in a new object called `wb_polity` and inspect the resulting data. How many rows does it have? Do you notice any missing data? Do you think you chose the right join?

Answer 2

```
# inner_join
wb_polity <- inner_join(wb, polity, by = c("country_code", "year"))

nrow(wb_polity)
```

```
## [1] 3002
```

Using `inner_join()` is the right choice because we want to compare `polity` scores to factors in `wb`. If we used `left_join()` with `wb` on the left, we would end up with a lot of rows from `wb` that do not have a `polity` score and thus will be useless. In this case, we only want the exact overlapping rows, which is accomplished via `inner_join()`. There is still a lot of missing data, but that was present before we joined the two datasets – the introduction explains that there is a lot of missing data from various countries, so this is not surprising.

The new dataset `wb_polity` contains 3002 rows, meaning there are 3002 overlapping rows between `wb` and `polity`.

Question 3: Democracy and Public Health

Now that we have joined the two datasets together, we can examine the relationship between variables in them. We want to know whether and how the degree of democracy (measured by the Polity score `polity`) is related to a measure of government redistribution: Public Health Spending as % of Government Budget. We will examine their relationship in two steps:

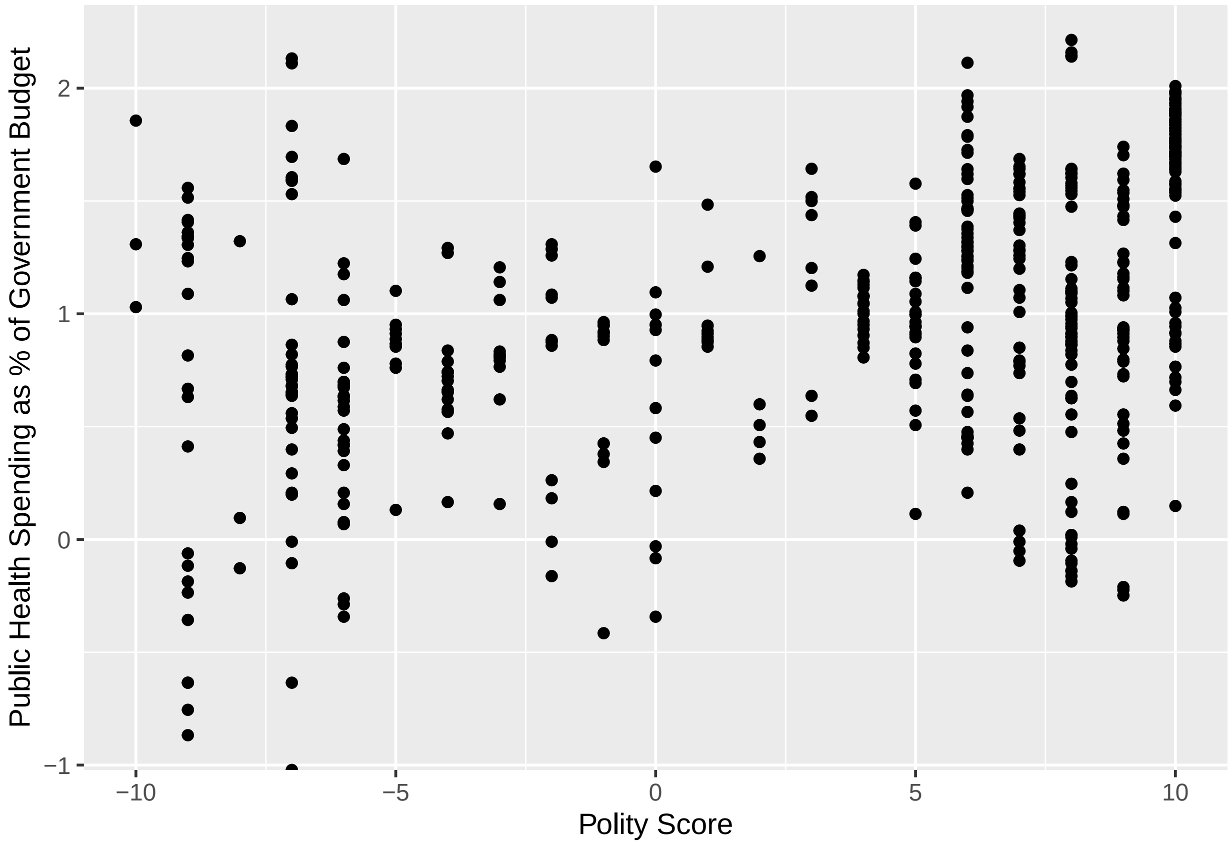
First, create a scatter plot via `ggplot()` and `geom_point()` with Polity score on the x-axis and the public health spending measure (look at the variables in the Introduction to see what column this is in) on the y-axis. Can you tell what their relationship is just by looking at the scatter plot?

Second, let's calculate the correlation between the two variables using `cor()`. Note that the `use` argument of `cor()` will need to be changed so that the correlation is only using observations where both variables are not missing (i.e., no NAs). Interpret what the correlation coefficient suggests for the relationship between the degree of democracy and public health spending.

Answer 3

```
# Scatter plot
ggplot(data=wb_polity,
       aes(x = polity, y = log(budget_health))) +
  geom_point() +
  labs(x = "Polity Score", y = "Public Health Spending as % of Government Budget")
```

```
## Warning: Removed 2435 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



In the scatter plot, there seems to be a positive relationship between the degree of democracy and public health spending, but the relationship does not appear to be strong.

```
# Calculate the correlation
cor(wb_polity$budget_health, wb_polity$polity, use = "pairwise")
```

```
## [1] 0.4432674
```

```
# or
wb_polity |> summarize(cor = cor(budget_health, polity, use = "pairwise"))
```

```
## # A tibble: 1 x 1
##     cor
##   <dbl>
## 1 0.443
```

There is a positive correlation between the degree of democracy and public health spending. Thus, it seems that democratic governments tend to spend relatively more on public health than autocratic ones; put it differently, more democratic governments spend relatively more on public health.

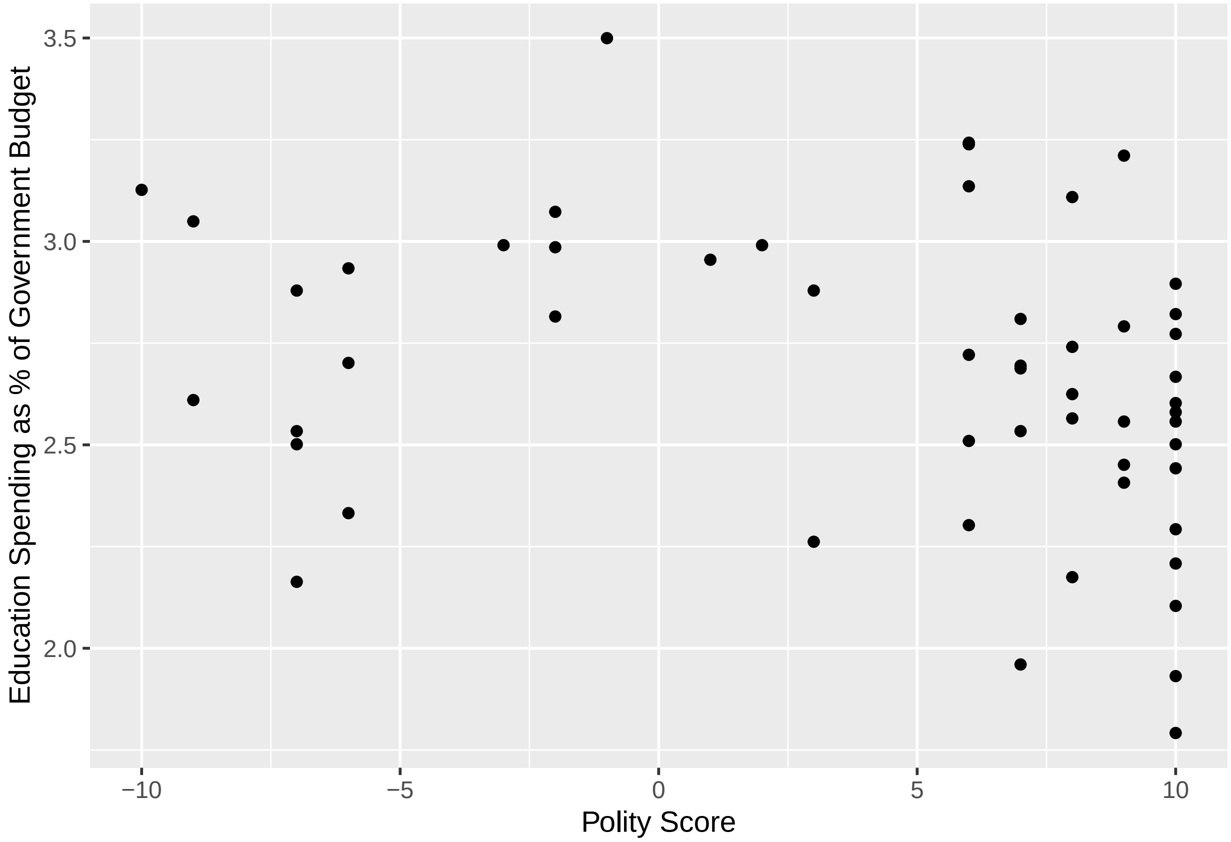
Question 4: Democracy and Education

Now, let's consider the other measure of redistribution in the dataset: Education Spending as % of Government Budget. Identify which column in the joined dataset contains this information, and then repeat the previous two steps for this variable and interpret your results.

Answer 4

```
# Scatter plot
ggplot(data=wb_polity,
       aes(x = polity, y = log(budget_edu))) +
  geom_point() +
  labs(x = "Polity Score", y = "Education Spending as % of Government Budget")
```

```
## Warning: Removed 2948 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



In the scatter plot, there seems to be a negative relationship between the degree of democracy and education spending, but it is not very apparent.

```
# Calculate the correlation
cor(wb_polity$budget_edu, wb_polity$polity, use = "pairwise")
```

```
## [1] -0.270358
```

```
# or
wb_polity |> summarize(cor = cor(budget_edu, polity, use = "pairwise"))
```

```
## # A tibble: 1 x 1
##      cor
##    <dbl>
## 1 -0.270
```

There is a negative correlation between the degree of democracy and education spending. Thus, it seems that democratic governments tend to spend relatively less on education than autocratic ones; put it differently, proportional spending on education tends to decrease as governments are measured as more democratic.

Question 5: Z-Score Function

Next, let's write a function called `my_z()` that will take a column of our data and return the standardized version of the column. Recall the formula for the Z-score for some variable x:

$$Z = \frac{x - \text{mean}(x)}{\text{sd}(x)}$$

You can also use the following template for the general composition of a user-defined function. You get to decide how many arguments (also called inputs), so you may not need the same number as are in the template. Make sure you include the `return(...)` statement to tell your function what you want to output.

```
my_function <- function(arg1, arg2, arg3, arg4, ...) {
  result <- ...do something...
  return(result)
}
```

Once your create your function `my_z()`, use it to transform the `polity` variable and store the output in a new column called `polity_std`. (Hint: you can either use `mutate()` to create a new variable and `my_z()` as the function inside of that `mutate` call, or you can assign the output of your `my_z()` function to a new column in your dataset directly by doing `data$column <-`) Check to see that the standardized variable has mean 0 and standard deviation 1.

Answer 5

```
my_z <- function(x) {
  z <- (x - mean(x, na.rm = TRUE)) / sd(x, na.rm = TRUE)
  return(z)
}

wb_polity$polity_std <- my_z(wb_polity$polity)

# or in tidyverse

wb_polity <- wb_polity %>%
  mutate(polity_std = my_z(polity))

# check whether the standardized variable has mean 0 and standard deviation 1.
mean(wb_polity$polity_std, na.rm = TRUE)
```

```
## [1] 2.434164e-17
```

```
sd(wb_polity$polity_std, na.rm = TRUE)
```

```
## [1] 1
```

Question 6: Missing Data

Let's consider the issue of non-random missing data now. First, how serious is the issue of missing data in the dataset? Calculate the proportion of observations (country-year) that are missing in each of the health and education spending variables.

Second, by calculating the correlations using `cor()`, examine whether democracies or autocracies are more likely to have missing data in those two variables. To do this, create two new columns (`missing_edu` and `missing_health`) that are equal to 1 if the observation is missing that variable and 0 if it is not. Then use `cor()` to look at the correlation between those variables and `polity`.

Answer 6

```
sum(is.na(wb_polity$budget_edu))/nrow(wb_polity)
```

```
## [1] 0.9810127
```

```
sum(is.na(wb_polity$budget_health))/nrow(wb_polity)
```

```
## [1] 0.7974684
```

```
wb_polity <- wb_polity |>
  mutate(missing_edu = is.na(budget_edu),
         missing_health = is.na(budget_health))

cor(wb_polity$missing_edu, wb_polity$polity, use = "pairwise")
```

```
## [1] -0.07225149
```

```
cor(wb_polity$missing_health, wb_polity$polity, use = "pairwise")
```

```
## [1] -0.2375382
```

The negative correlations suggest that democracies are less likely to have missing data in both the two variables.