Political Analysis Using R

Regression Analysis II

Hilary 2016

Authors: Roosmarijn de Geus, Niels Goet, and Julia de Romémont

In this final lab session, we'll revisit some of the key things we have learned this term. We'll again be investigating the relationship between representation of women in politics (i.e. the degree to which they hold political office) and a number of other variables.

1. Inspecting variables and computing summary statistics

This time, we'll consider the percentage of women that hold cabinet positions (see `womens_cabinet_representation_1995` and `womens_cabinet_representation_2008` in the Lijphart dataset).

1. Download the Lijphart dataset:

```
data <- read.csv("http://andy.egge.rs/data/L.csv")
```

2. Inspect the two cabinet representation variables that Lijphart uses.

Again, rename the variables in the interest of simplicity:

```
names(data)[57] <- "women_cabinet1995"
names(data)[58] <- "women_cabinet2008"
```

Look at the values of the variables. Are there any missing values? For which countries? (`Hint`: You can simply look at the variable in the upper-left pane of `RStudio`, but it might be easier to look at the two key columns in isolation using a command like:

```
View(data[ , c('country', 'women_cabinet1995',
          'women_cabinet2008')])
```

The `is.na()` command may also be helpful, together with subsetting the country variable to see which countries have missing data:

```
data$country[is.na(data$women_cabinet1995)]
```

- What country has the highest value of representation in 1995? And how about 2008?
- What is the average value of these variables?
- Plot a histogram for both variables.
- Rename the variable `eiu_democracy_index_2006_2010` into `democracy_index`. Now check for missing values for the variable `democracy_index` that we used in our multivariate regression analysis in lab 3. Are there any missing values? For which countries?
- What is the average value of the environmental index? (HINT: Don't forget to specify `na.rm = TRUE` to tell `R` to ignore missing values).

2. Means difference test

The mean for represenation of women in the cabinet is higher in 2008 than in 1995. What if we want to know whether this difference is statistically significant? Use the two-sample t-test to check whether this is the case:

```
t.test(data$women_cabinet1995,data$women_cabinet2008)
```

- What is the null hypothesis for this test? And, what is the alternative hypothesis?
- What p-value do you get, and what does this tell you?

3. Correlation

Now let's find out whether the independent variables that we used in lab 3 are also associated with representation of women in the cabinet. We'll first use a simple correlation to see whether this is the case.

Let's first explore the relationship between `eiu_democracy_index_2006_2010`[1] (the variable that we found had a statistically significant effect on women representation in parliament in lab 3) and `womens_cabinet_representation_2008` visually. Create a scatterplot in which `women_cabinet_2008` is on the vertical axis and `democracy_index` is on the horizontal axis.

- Does the correlation between the two variables look positive or negative?
- Calculate the correlation between the two variables. Use the following syntax:

```
cor.test(first_variable,second_variable,use="complete")
```

- What is the size of the correlation coefficient? What does this tell you?
- Is the correlation statistically signficant? At what confidence level?

4. Bivariate regression

Use the `lm()` command to regress the representation of women in cabinet for 2008 (dependent variable, i.e. Y) on the EUI Index of Democracy (independent variable, i.e. X).

- What is the estimate of the intercept in this regression? What does this tell you?
- What is the estimate of the coefficient on the independent variable (consensus democracy)? What does this tell you?
- What is the standard error of the coefficient on the independent variable? What does this tell you? HINT: You can get regression coefficients simply by entering:

```
lm(dependent_variable ~ independent_variable)
```

but the easiest way to see detailed regression results is by wrapping that command in the `summary()` command, as follows:

[1] Recall from lab 3 that The EIU's Index of Democracy 'is based on five categories: electoral process and pluralism; civil liberties; the functioning of government; political participation; and political culture. Countries are placed within one of four types of regimes: full democracies; flawed democracies; hybrid regimes; and authoritarian regimes.' See: *https* : *//graphics.eiu.com/PDF/Democracy_Index*$_2$*010_web.pdf*

```
summary(lm(dependent_variable ~ independent_variable))
```

- What is the p-value associated with the coefficient on the independent variable? What does this tell you?

5. Dummy variables and multivariate regression

What if we do not only want to know whether there is a statistically significant difference in representation between the sample of 1995 and 2008, but also in a particular set of countries *vis-à-vis* another group of states? In that case, we can create a dummy variable that codes which countries belong to what group. Here, we are going to check whether being in Europe or not affects representation of women in the cabinet in 2008.

1. First, create a variable that takes the value `TRUE` for countries in Europe and `FALSE` otherwise:

```
data$europe = data$country %in%  c('AUT', 'BEL', 'DEN', 'FIN', 'FRA', 'GER'
                                   ,'ICE', 'IRE',  'ITA', 'LUX', 'MAL',
                                   'NET', 'NOR', 'POR','SPA', 'SWE', 'SWI',
                                   'UK')
```

- Have a look at the variable you have just created (`data$europe`). Through the code above, you have basically asked `R` to check all countries (`data$ecountry`) against the list of countries in parentheses, and attribute `TRUE` whenever this was the case; and `FALSE` otherwise.

2. Use this dummy variable to test whether there is a difference in means between countries in Europe and outside Europe:

```
t.test(data$women_cabinet2008[!data$europe],
       data$women_cabinet2008[data$europe])
```

- Is there a statistically significant difference in representation of women in the cabinet between European and non-European countries? How can you tell?

3. Calculate the mean of representation in European countries (HINT: have a look at the previous code. In R, "`!`" is used to indicate a negative (i.e. "does not equal"))

4. Rerun the regression, this time controlling for whether a country is in Europe or not. (HINT: Recall from lab 3 that you can simply use "`+`" to add an extra independent variable).

- What is the estimate of the intercept? What does this tell you?
- What is the estimate of the coefficient on the EUI democracy index? What does this tell you?
- Is the relationship between representation of women in the cabinet and the EUI democracy index statistically significant when we control for whether a country is in Europe?
- If we want to understand the effect of the EUI democracy index on representation of women in the cabinet, is it appropriate to control for whether a country is in Europe?

5. Sometimes, we do want to conduct regression analysis on a subsample of our data. For example, perhaps we wish to know what the effect of the EUI democracy index on representation of women is in non-European countries only. We can do this through the following syntax:

```r
lm(women_cabinet2008 ~ democracy_index,
   data = data[!data$europe,])
```

NOTE: Here, you have not *controlled* for the effect of whether a country is in Europe or not; rather, you have run the regression for countries outside Europe only. In other words: you have conducted the analysis on a *subset* of the sample.

- What is the estimate of the intercept? What does this tell you?
- What is the estimate of the coefficient on the EUI democracy index? What does this tell you?
- Is the relationship between the EUI democracy index and representation of women in the cabinet statistically significant when we focus only on non-European countries?

6. How about the situation in Europe? Repeat exercise 5 above, but now focus on European states.

Q-Step survey

In order to know what you thought of these lab sessions, it would be great if you could fill out the following survey: https://www.surveymonkey.co.uk/r/QS1_HT16

Thank you so much!

Homework

None (yay!).

Appendix

1. Online material

Have a look at the following courses to improve your regression analysis skills in R:

- www.datacamp.com, A Hands-on Introduction to Statistics with `R` (Course 6: Multiple Regression).

- Swirl: have a look at the course `Statistical_Inference`. Use the code below to access this course (skip the first command if you have already installed the package):

```r
install.packages("swirl")

#load the swirl package
library(swirl)

#install the courses
install_from_swirl("Statistical_Inference", mirror = "bitbucket")

#enter your name and choose a course
swirl()
```