Political Analysis Using R

Regression Analysis I

Hilary 2016

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

In the previous lab sessions, we learned how to load data, how to visualise it, and how to conduct some very simple tests (e.g. the correlation test). In this lab, we take the plunge and start with proper inferential statistics. In other words: we are going to formulate a hypothesis and use regression analysis to *test* the hypothesised relationship between our response variable (i.e. representation of women in parliament) and our independent variable(s).

Our research question for this lab is: *What determinants affect the level of underrepresentation of women in politics?*

For most of this lab, we will be exploring the link between representation and socio-economic performance. We will primarily use `womens_parl_representation_2010` from the Lijphart dataset as an indicator of representation of women in politics, and `hdi_2010` as a measure of socio-economic performance.

Note: the Appendix to this lab sheet contains definitions and advice for accurately representing your results in a regression table.

1. Association between categorical variables

Before we start with regression analysis, however, we'll briefly look into two important tests: the **means difference test** and the **test of independence** for categorical variables.

1.1 Means difference test

Let's assume you have two groups of observations, and you want to check whether there is a statistically significant difference in representation of women in parliament between these two groups. This all sounds a bit abstract, but consider the following scenario: you have data on the number of female MPs in a number of countries in 1990 (`women1990`), and the same data for 2010 (`womens_parl_representation_2010`). You now want to know whether there is a difference that is statistically different from zero between these two years. In other words, has representation of women increased (or decreased), and is this increase (or decrease) statistically significant? Let's try and find out.

First, load the Lijphart dataset into R:

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

To make things easier, rename the variables for women representation in parliaments:

```
names(data)[55] <- "women1990"
names(data)[56] <- "women2010"
```

Now, inspect (`women1990` and `women2010`) and calculate the **mean** and the **standard deviation** for both variables.

How do we know whether there is a statistically significant difference in the mean of these two groups (1990 and 2010)? In this case, we can use the means difference test (the t-test):

```r
t.test(data$women1990,data$women2010)
```

Question 1.1: What are the null and the alternative hypotheses for the t-test?

Null hypothesis:____
Alternative hypothesis:____

Question 1.2: Can we reject the null hypothesis in this case, and what does that mean for your interpretation of the test results?

Answer:____
____

1.2 Test of independence

Let's now consider the test of independence between two categorical variables - the Chi-squared (χ^2) test of independence. In the previous lab session, we looked at the relationship between two continuous variables (i.e. variables that can take any value within a certain bound). Now, we are going to consider the association between two categorical variables, i.e. variables that can only have a finite set of values delimited by an upper and a lower bound. In order to do so, we will first create a *dummy variable* for representation of women in parliament that takes two values: `1` for high levels of representation, and `0` for low levels of representation. We use the median value as a way to split representation into two groups:

Calculate the median of `women2010`, and store it in an object:

```r
representation_median <- median(data$women2010)
```

Create the dummy variable, using the following two steps:

1. Create a new variable that copies the original representation variable, and store it in the dataset:

```r
dummyrepresentation <- data$women2010
```

2. Give all observations that are higher than the median a value of `1`; and those that are lower than the median a value of `0`:

```r
dummyrepresentation[dummyrepresentation < representation_median] <- 0
dummyrepresentation[dummyrepresentation > representation_median] <- 1
```

Now do the same for `hdi_2010`, and store the result in an object called "`dummyhdi`":

```r
hdi_median <- median(data$hdi_2010)
dummyhdi <- data$hdi_2010
dummyhdi[dummyhdi < hdi_median] <- 0
dummyhdi[dummyhdi > hdi_median] <- 1
```

Create a contingency table for these two categorical variables using the following commands in R:

```
representation.hdi <- table(dummyrepresentation, dummyhdi)
representation.hdi
prop.table(representation.hdi)
```

Is the association statistically significant? The test to determine the significance of the association is called the **Test of Independence**. The key idea behind independence tests is Observed and Expected Values. Whereas Observed values are more or less self–explanatory, the Expected Values relate to the scores we would get under the Null hypothesis. The expected value for each cell in a two-way table equals to: $\left(\frac{RowTotal \times ColumnTotal}{n}\right)$, where n is the total number of observations included in the table. To test for independence we examine the χ^2 statistic which is given by the following formula:

$$\chi^2 = \sum_{i=1}^{n} \frac{(Observed_i - Expected_i)^2}{Expected_i}$$

where i indexes the cells in the table. Thus, χ^2 is the sum of the difference between expected (under the null of no association) and observed frequencies in each cell. A χ^2 statistic helps determine if the associations are due to pure luck or if there is a systematic pattern in our sample. The test can be ran using the following code:

```
chisq.test(representation.hdi)
```

> QUESTION 1.3: Is the association between the two categorical variables statistically significant? Why (not)?
>
> Answer:____
>
> ____

2. Regression analysis with two variables

The χ^2-test tells us something about the association between two dichotomous variables, just as the correlation test tells us something about the association between two continuous variables. Now, we'll start with the basics of regression analysis - one of the most important tools in testing the existence of a *relationship* between variables. Before we start, take a minute to review the key terms in the Appendix (section 1).

Let us first explore the relationship between the Human Development Index[1] (`hdi_2010`) and women representation in parliament (`women2010`). We do this by first formulating our hypotheses:

> QUESTION 2.1: Formulate a null hypothesis and an alternative hypothesis for the relationship between HDI and representation of women in parliament.
>
> Null hypothesis:____
> Alternative hypothesis:____

[1] According to the UNDP, the Human Development Index (HDI) is 'a summary measure of average achievement in key dimensions of human development: a long and healthy life, being knowledgeable and have a decent standard of living. The HDI is the geometric mean of normalized indices for each of the three dimensions.' See: http://hdr.undp.org/en/content/human-development-index-hdi

Subsequently, we estimate a bivariate regression model. In R, the general formula for running a regression is: `lm(dependent_variable ∼ independent_variable)`. So let's apply this formula to our variables of interest:

```
model1 <- lm(data$women2010 ~ data$hdi_2010)
```

Note that we have stored the estimates in the object called "model1". If we want to see details of the model, including the p-value, degrees of freedom and the R_2, we call the `summary()` command on the object we have just created:

```
summary(model1)
```

Question 2.2: Is there a statistically significant relationship between the HDI and representation of women in parliament?

Answer:____

____

Question 2.3: How would you interpret the coefficient for `hdi_2010`?

Answer:____

____

Question 2.4: What does the adjusted R-squared tell us about model fit?

Answer:____

____

Optional exercise

There are different ways of measuring representation of women in politics. Find two such measures in the Lijphart dataset and answer questions 2.1-2.4 for these variables.

3. Multivariate regression analysis

Bivariate regression only tells us whether a relationship exists between two variables of interest. Usually, we want to know whether our predictor (the variable that we have theorised will have an effect on the phenomenon that we are investigating) is the most important one. In other words: we want to *control* for the effect of other variables, and check whether the relationship holds after we have added these determinants to the model. This allows us to make stronger claims about the validity of our theory, and to show that other explanations are incorrect (or do not explain as much as our own theory).

In section 2, we formulated a hypothesis about the effect of HDI on the representation of women in parliament. We are now going to estimate a *multivariate* regression model to see whether the effect still holds when we control for other determinants. In R, we can add extra variables by simply using the "`+`" sign:

```
lm(dependent_variable ∼ independent_variable1 + independent_variable2 +
independent_variable3).
```

Here, we are going to control for the impact of two variables: i) the Economist Intelligence Unit (EIU) democracy index[2] (`eiu_democracy_index_2006_2010`); ii) the proportionality of the electoral system, for which we use the effective number of parties as an indicator (`eff_num_parl_parties_1945_2010`).

Both variables have been theorised to have an effect on the representation of women in legislative assemblies:
- A number of contributions theorise that the quality of democracy positively affects the number of women who enter into politics (Inglehart, Norris, and Welzel 2004).
- Several authors have found that proportional representation systems with party lists tend to have more female MPs than plurality systems (Duverger 1955; Lakeman 1994). The theory behind this claim is that in PR systems, parties seek to widen their appeal by adding women to the list. The risk of doing so is perceived to be less when the female candidate is part of a group, rather than the only candidate.

We will add these variables one by one:

```
model2 <- lm(data$women2010 ~ data$hdi_2010 +
  data$eiu_democracy_index_2006_2010)
summary(model2)
```

Question 3.1: What happens to the impact of HDI when we add the EUI's index of democracy? How would you interpret the coefficient and p-values of these two variables?

Answer:____

____

```
model3 <- lm(data$women2010 ~ data$hdi_2010 +
               data$eiu_democracy_index_2006_2010 + data$eff_num_parl_parties_1945_2010)
summary(model3)
```

Question 3.2: What happens to the coefficients and the p-values now? How about the model fit?

Answer:____

____

Optional exercise

What other variables do you think may have an impact on representation of women in parliament? Why? Estimate a multivariate regression model with these variables and interpret the results.

Homework

Send your answers to the following two questions (including `R` code used) to your lab instructor by noon on Friday of week 7:

- Create a scatterplot that shows representation of women in parliament on the y-axis (dependent variable) and an appropriate independent variable on the x-axis. Add a regression line.

- Run a regression with these two variables and make a regression table that includes all necessary information (see section 2 for an example of how to present regression results).

[2] 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*

Appendix

1. Presenting regression results effectively

Table 1 below shows the regression results for the binary and multivariate models that you estimated in secions 2-3 of this lab sheet. There are a number of conventions when you present your regression results. Researchers usually report the following information:

- The estimates of the coefficients
- The standard errors (S.E.)
- The intercept
- The (adjusted) r-squared (a measure of model fit)
- The number of observations (HINT: to find the number of observations, simply call the `nobs()` command on the object in which you stored the regression model. For example, to get the number of observations for the first model that we estimated in section 2, use `nobs(model1)`)
- Statistical significance: The level of statistical significance of the effect of an independent variable is indicated by placing asterisks next to the coefficient.

Table 1: Regression Results

	Model 1	Model 2	Model 3
Variables	Estimate (S.E.)	Estimate (S.E.)	Estimate (S.E.)
HDI 2010	49.65** (20.29)	14.64 (24.18)	14.86 (23.89)
EUI Democracy Index		5.94** (2.80)	5.21* (2.82)
Eff. number of parties			2.12 (1.60)
Intercept	−16.20 (16.90)	-36.37* (19.66)	-37.35* (19.44)
N	36	34	34
Adjusted R^2	0.125	0.199	0.218

Notes: *p<0.1; **p<0.05; ***p<0.01; ****p<0.001.

2. Key terms in regression analysis

Term	Description
Ordinary Least Squares Regression	Finding the “line of best fit” between the dependent variable “y” (or: response variable) and the independent variable (i.e. the line that minimises the size of the residuals).
Residuals	Distances between the regression line and the observations.
Coefficient	A one-unit increase in the value of the independent variable leads to an increase of the size of the coefficient in the dependent variable.
Standard Error	Estimate of the standard deviation of the sample mean (based on the population mean).
Constant/Intercept	Expected mean value of the dependent variable when all independent variables’s are zero.
P-value	Probability of observing a value larger than the t- value here, given the t-distribution with d.f. (n-1) (general: the probability of obtaining an effect that is at least as extreme as the estimate in your sample data, assuming the null hypothesis is true. I.e. A P-value evaluates how well the sample data support the idea that the null hypothesis is true.
T-value	The t-value is the coefficient divided by the standard error. To find the p-value associated with a coefficient, compare the t-value with the critical value on the t-distribution with degrees of freedom (n-1) (fortunately, `R` does this for you automatically!).

3. 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 2: Students’ T-Test and Course 5: Correlation and Regression).

- Swirl: have a look at the courses `Data_Analysis` and `Regression_Models`. 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("Data Analysis", mirror = "bitbucket")
install_from_swirl("Statistical Inference", mirror = "bitbucket")
install_from_swirl("Regression Models", mirror = "bitbucket")

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