Introduction to Statistics
lab 6

Johan A. Elkink
jos.elkink@ucd.ie

12 October 2015

Data

For this lab we will make use of the Irish National Election Study (INES) data from the 2007 general election. Download the file referred to as "Long datafile in Stata 10 file format" at `http://www.tcd.ie/ines/index.php?action=download`. The data is in a ZIP archive,[1] which you open by simply double clicking on the file and then copy the Stata file to a folder where you can open it from inside your statistical package as usual. Make sure you also download the "Data codebook".

Open the file and select only the survey from 2007 – the year of the survey is coded in the **ines** variable. Note that this variable also contains missings, so you might have to explicitly exclude those. At least in R, it should be: `ines <- ines[ines$ines == 2007 & !is.na(ines$ines), ]`, assuming that the file has been opened with `ines <- read.dta()`.[2] One way to find out whether there are many missings is by making a frequency table of the **ines** variable after selecting the subset of the data. In Stata and SPSS this will include a count of the missings, in R you will need to do: `table(ines$ines, exclude = NULL)`.

Standard error and confidence intervals of the mean

1. Produce a frequency table of **v0190**, which is the answer to "How probable is it that you will ever give your first preference vote to the following parties? – Labour Party".
2. Calculate the mean and the standard error of the mean for this variable:
   SPSS: `DESCRIPTIVES v0190 /STATISTICS = mean semean.` or
   `MEANS v0190 /CELLS = mean semean.`

[1] `https://en.wikipedia.org/wiki/Zip_(file_format)`
[2] The `is.na()` function checks for missing cases; the exclamation mark negates the finding. So combining the two means including all those that are not missing.

R: There is no function to do this, but it is easily calculated using the `mean()` and `sd()` functions from earlier labs. `table(is.na(ines$v0190))` will help identifying N.
Stata: `mean v0190`.

3. Calculate the mean and the standard error of the mean of **v0190** separately for those that are a member of a trade union and those that are not (**v0936**):
SPSS: `MEANS v0190 BY v0936 /CELLS = mean semean.`
R: For the mean, you can use: `tapply(ines$v0190, ines$v0936, mean, na.rm = TRUE)` – and similar for the standard deviation, which you can use to calculate the standard errors.[3]
Stata: `mean v0190, over(v0936)`.
4. Based on the calculated standard errors, manually calculate the confidence intervals of the means for union members and non-members.
5. If you consider the mean for union members, does it fall inside or outside the confidence interval of the mean for non-members? What do you think this might mean?
6. Repeat the above for Fianna Fáil (**v0187**).

Standard errors in regression

1. Regress support for Fine Gael (**v0188**) on the left-right self-placement of the respondent (**v0239**).
2. Interpret the regression coefficients. How do you express this in terms of left-right placement and support for Fine Gael?
3. Find the standard errors of the regression coefficients and manually calculate confidence intervals for each of the two coefficients.
4. For the intercept, does zero fall within the confidence interval? What do you think this might mean?
5. For the slope coefficient, does zero fall within the confidence interval? What do you think this might mean?
6. Produce a scatter plot with regression line to illustrate the finding.
7. Interpret the R^2 statistic.
8. Repeat the above for the Labour Party (**v0190**).

[3] The `tapply()` function is a tool to apply a function (in this case `mean()`) to a variable (here **v0190**), separately for the subcategories of another variable (here **v0936**). You can pass additional parameters to the function after the function name, like here we use `na.rm = TRUE` to instruct R to ignore missings.

Regression with dummy variables

Normally speaking, linear regression is for evaluating the relationship between two continuous variables. However, when the independent variable is a dummy variable – i.e. only has values 0 and 1 – then it is also valid to include this in a regression model as an explanatory variable.[4] The "slope coefficient" then does not indicate the slope of a line, but rather the difference in mean for two groups. E.g. if we have regression equation

$$y_i = \beta_1 + \beta_2 d_i + \varepsilon_i,$$

whereby Y is a continuous variable and D a dummy variable, then we have two estimates for Y:

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2$$

when $d_i = 1$ and

$$\hat{y}_i = \hat{\beta}_1$$

when $d_i = 0$, so just two values, two means of Y that differ by β_2.

1. Regress support for the Labour Party (**v0190**) on union membership (**v0936**). This should work without problems as any regression.[5]
2. Compare the estimate of the intercept (β_1 in the above equation) with the means calculated at the beginning of this lab exercise.
3. Compare the estimate of the slope coefficient (β_2) with those two means.
4. Consider the confidence interval of the slope coefficient. Is zero included in this interval? What might that mean?
5. Repeat for Fianna Fáil (**v0187**).

[4] It gets more complicated when it is the dependent variable.

[5] If you get error messages, try recoding **v0936** explicitly into a variable with only ones and zeros first.