Introduction to Statistics
homework 2

Johan A. Elkink
jos.elkink@ucd.ie

Due 26 October 2016, 5 pm

You will submit two files: one PDF file[1] *including all plots, tables and interpretations and one command file (SPSS Syntax file (.sps), or Stata do-file, or R-file) with all commands used to answer the exercise and no superfluous commands. Please send both files to* `jos.elkink@ucd.ie`.

(9%) of the grade is used for an overall evaluation of the presentation of your work (the PDF file) and (5%) of the grade for the evaluation of the clarity / presentation of your command file, including the use of comments, clear variable names, and whitespace.

1. For this homework, we will make use of the replication data for Hassan (2014). This article is still forthcoming, but a working paper has been published online and it is recommended to download and read this paper. The American Journal of Political Science, where this article will be published, requires authors to make replication data available through the Harvard DataVerse, which we have encountered in class. Access the data[2] and download `variable_names_replication.pdf` to have a list of what the different variables in the data set mean and the Stata version of `Shuffle_replication_data.tab`.

   (a) (5%) Open the data and make sure the command for doing so is in the command file.

   (b) (5%) Select only the observations from 1997 from this data and only those where the author has indicated that the case should be dropped in case average values are used, i.e. where **avg_drop_indicator** equals 1—see "Subsetting data" in Lab 3 for instructions.[3]

2. During the period covered by the data, Daniel arap Moi is president of Kenya, and he is a Kalenjin, an ethnic group in the Rift Valley province in Kenya. What we are interested in here is the appointment of district officers (DOs), regional officials, in different electoral districts in Kenya. The question is, does the president appoint officials of his own

[1] Word files will be sent back—note that newer versions of Word can easily save to PDF format.
[2] `https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/WPKTKJ`
[3] Note that you can use the ampersand symbol, as in: `year == 1997 & avg_drop_indicator == 1`.

ethnic group in regions where his electoral position is less secure? We therefore use as dependent variable **total_dos_9298_kal_p**, the percentage of DOs prior to the 1998 elections that are of the Kalenjin ethnic group. To measure the electoral position, we use the previous (i.e. 1992 elections) vote share of the president, **laggedvoteshare**.[4]

- (5%) Calculate variances and standard deviations for both variables.
- (5%) Calculate the covariance and correlation coefficient between those two variables.
- (5%) Regress the percentage of Kalenjin DOs on lagged vote share.
- (5%) Produce a publishable regression table.[5]
- (8%) What do you conclude about the relationship between lagged vote share and percentage of Kalenjin DOs? Do you find support for the expected relationship? Is there any impact of previous vote share on the likelihood of appointing Kalenjin DOs—and if so, how much? (approx. 200 words)

Ethno-linguistic fractionalization	0.174**
	(0.083)
Intercept	0.439***
	(0.033)
N	249
R^2	0.018
Adjusted R^2	0.014
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 1: OLS regression explaining Moi's (lagged) vote share by ethnic fractionalization.

3. Table 1 produces a regression table whereby the lagged vote share of President Moi is regressed on the level of ethnic fractionalization (**elf**). Since ethnic fractionalization is a relatively static variable, the fact that the vote variable is lagged is not so important—so read this as explaining vote share by ethnic diversity. The fractionalization index is a measure that is low when there are many, small ethnic groups and close to one when there is one large ethnic group.
   (a) (5%) Replicate the regression analysis.
   (b) (5%) Standardize both variables.
   (c) (3%) Repeat the regression analysis using the standardized variables.
   (d) (5%) Produce a regression table identical to Table 1, but with a column next to it including the regression results using standardized variables.

[4] "Lagged" means that it is observed in the previous time period.

[5] See `http://www.joselkink.net/wp-content/uploads/2013/01/POL5005_Spring_2013_note_regression_presentation_and_interpretation.pdf`, and, as an example, Table 1.

Model (1)		Model (2)	
Population	0.000000562	Log(Population)	0.436
Intercept	199.6	*Intercept*	-2.03
N	77		77
R^2	0.20		0.84

Table 2: Two regression tables. On the left with the number of seats in the lower house as dependent variable; on the right with the logarithm of the number of seats.

(e) (8%) Fully discuss what you would conclude about the relationship between ethnic fractionalization and vote share for the incumbent president, in particular fully interpreting the coefficients and R^2. (approx. 200 words)

4. For this question you will not need SPSS. We will look at the relationship between the number of seats in parliament and the size of the population. Only those parliaments classified as "lower house" in the table have been included.[6] We will investigate two models:

$$Seats_i = \beta_1 + \beta_2 Population_i \quad (1)$$

$$\log(Seats_i) = \beta_1 + \beta_2 \log(Population_i) \quad (2)$$

For Model (1), the regression coefficients can be found on the left-hand side in Table 2 and the scatter plot with regression line can be found in Figure 1. For Model (2), the regression coefficients can be found on the right-hand side in Table 2 and the regression line in Figure 2.

(a) (3%) Based on Model (1), what do you conclude about the relationship between a country's population and the size of the lower house?

(b) (5%) If the population size of country A is one million people larger than the population size of country B, how much would you expect their size of their lower houses to differ? Which one will be larger? Include the reasoning and/or calculation.

(c) (3%) Both visually and based on Table 2, how well do you think this linear regression summarizes the data? Explain why you think so.

(d) (3%) For Model (2), both variables are transformed before performing the linear regression. Both visually and based on the information in the table, how does this affect the extent to which the model describes the data? Explain why you think so.

(e) (3%) Based on both regressions, what would you conclude about the Irish and Indian cases: Are they typical or unusual compared to other countries? Explain why you think so.

(f) (5%) Based on Model (2), if country C has a population of 30 million, how many seats would you expect the lower house in parliament to have? Include the reasoning and/or calculation.

[6] The data is taken from `http://en.wikipedia.org/wiki/List_of_legislatures_by_country`.

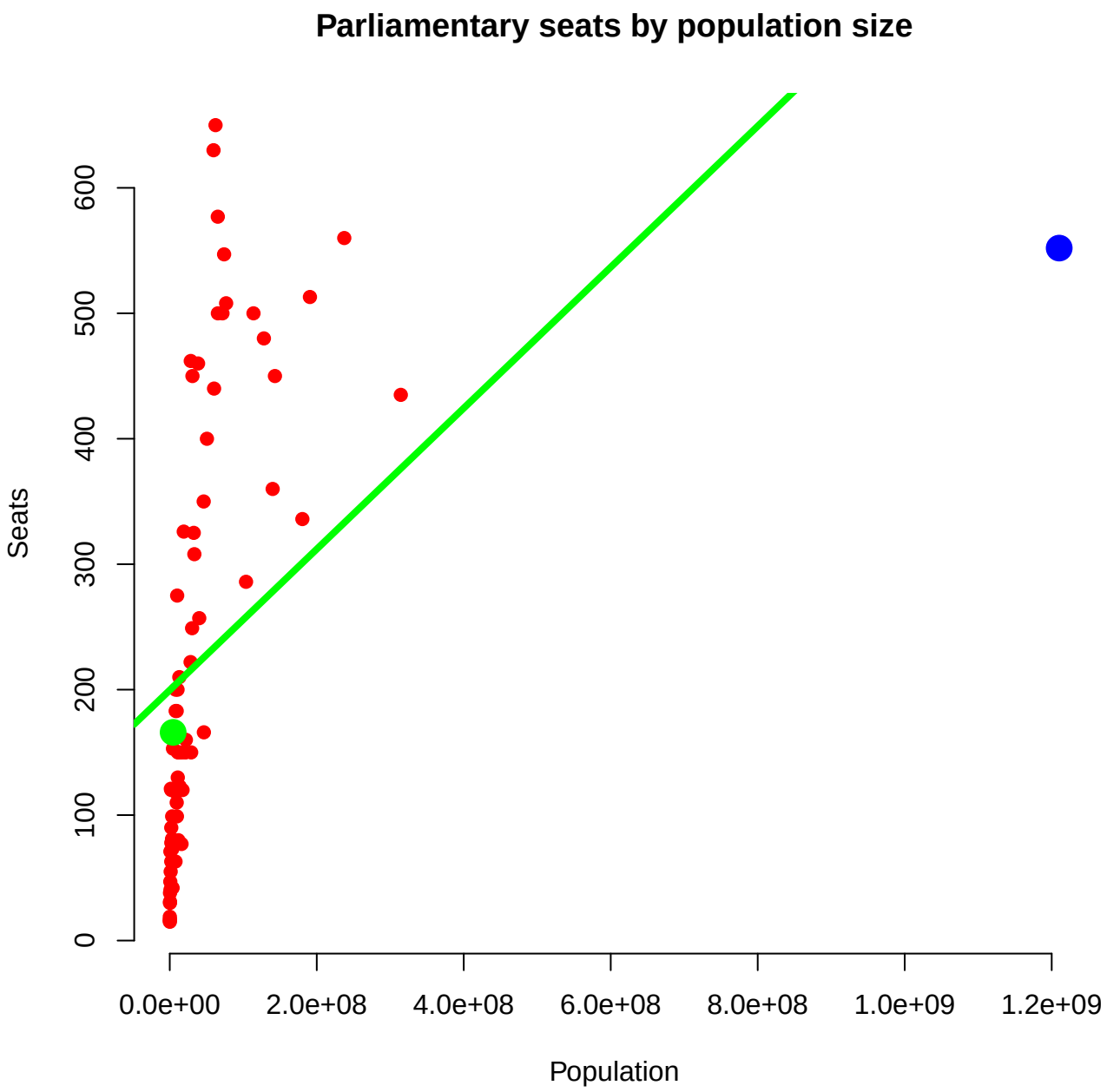

Figure 1: Seats by population and linear regression line. Green dot represents Ireland and blue dot represents India.

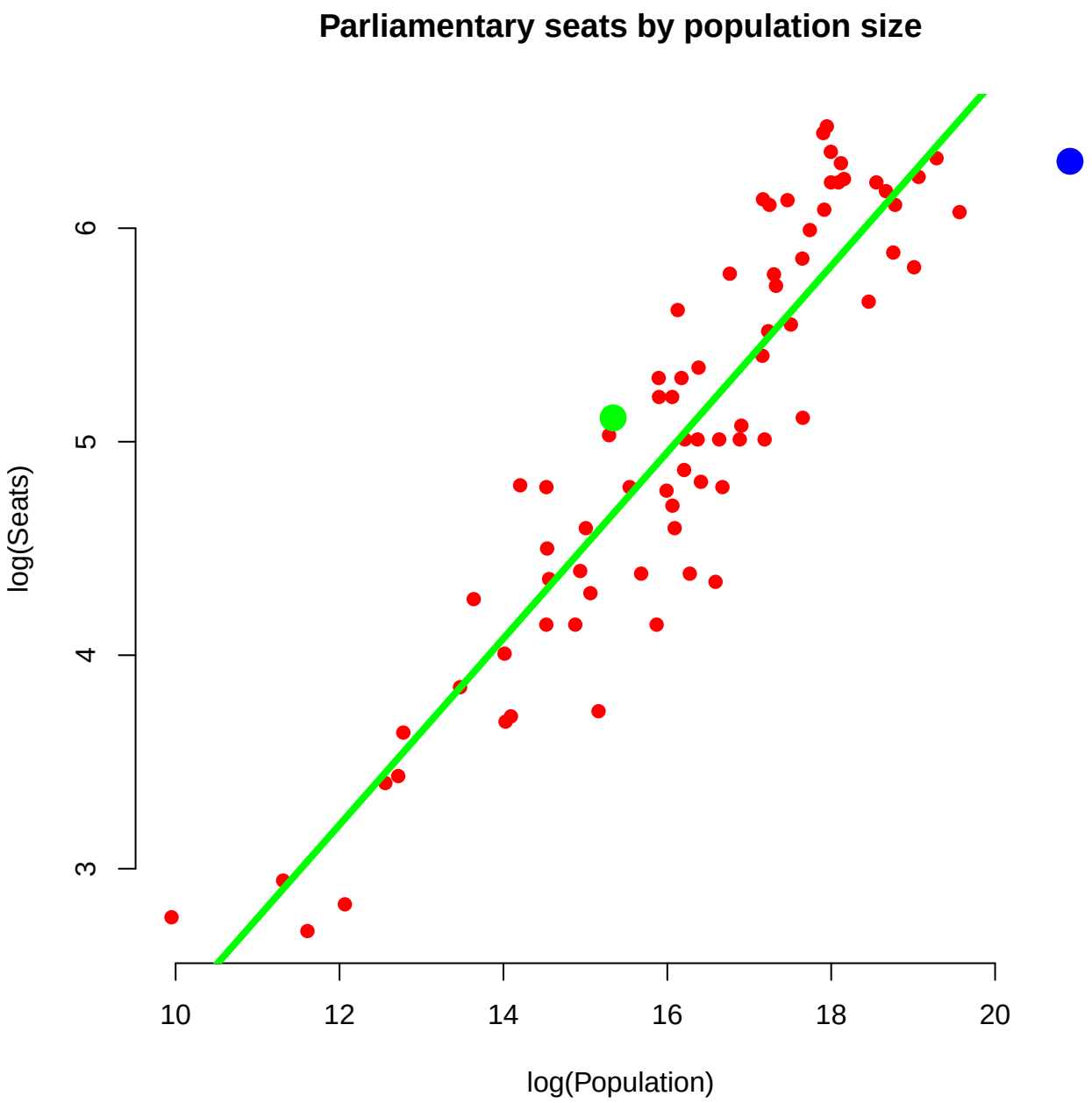

Figure 2: Seats (logged) by population (logged) and linear regression line. Green dot represents Ireland and blue dot represents India.

References

Hassan, Mai. 2014. "The Strategic Shuffle: Ethnic Geography, the Internal Security Apparatus, and Elections in Kenya." revised version forthcoming in *American Journal of Political Science*.
URL: *http://dosen.narotama.ac.id/wp-content/uploads/2014/11/The-Strategic-Shuffle-Ethnic-Geography-the-Internal-Security-Apparatus-and-Elections-in-Kenya.pdf*

Grade conversion scheme

Homeworks	UCD	MDP	Homeworks	UCD	MDP
97-100%	A+	78.33	74-76%	C-	51.67
94-96%	A	75.00	71-73%	D+	48.33
91-93%	A-	71.67	68-70%	D	45.00
88-90%	B+	68.33	65-67%	D-	41.67
85-87%	B	65.00	54-64%	E+	38.33
83-84%	B-	61.67	44-53%	E	35.00
80-82%	C+	58.33	33-43%	E-	31.67
77-79%	C	55.00	0-32%	F	25.00

Note that the percentage scores will be translated to UCD grades before entering on the system. Overall module grade will be calculated by the system based on the UCD grades. For MDP students, grades will then be translated to TCD marks. Note that TCD marks are *not* percentages and will therefore reflect the above scale. Thus, a 95% score on all your homeworks will generate an A grade on the UCD system, and a 75 mark on the TCD system.