Introduction to Social Data (SOC1004 / POL1008)
Take-home exam

August 2017

Please answer all the questions for both exercises in the exam. Your answer to each question must include: a) R code, b) R output with the results (and graphs where necessary), c) a short (one or two paragraphs) interpretation of the results. Each question is worth 10 points.

Please submit your assignments on eBart. The submission deadline is 15 August, 2pm.

You can use any sources to work on the assignment (the Internet, your notes, textbook, etc.), but you cannot consult with other students or use other students' work. Your exam must be the result of your individual work.

The following two exercises are included as additional exercises in K.Imai, "A first course in quantitative social science".

1 Indiscriminate Violence and Insurgency (60 points)[1]

In this exercise, we analyse the relationship between indiscriminate violence and insurgent attacks using data about Russian artillery fire in Chechnya from 2000 to 2005. Some believe that indiscriminate violence increases insurgent attacks by creating more cooperative relationships between citizens and insurgents. Others believe that indiscriminate violence can be effective in suppressing insurgents' activities. Table 1 contains the names and descriptions of variables in the data file `chechen.csv`.

1.1. How many villages were and were not attacked by Russians? Give the overall breakdown. (10 points)

[1] This exercise is based on: Lyall, J. (2009). "Does Indiscriminate Violence Incite Insurgent Attacks? Evidence from Chechnya." Journal of Conflict Resolution 53 (3): 331-362.

Table 1: The Russian Artillery Fire Data

Name	Description
`village`	the name of Chechnya village
`groznyy`	a variable indicating whether a village is in Groznyy (1) or not (0)
`fire`	whether Russians struck a village with artillery fire (1) or not (0)
`deaths`	estimated number of individuals killed during Russian artillery fire
`preattack`	the number of insurgent attacks before Russian artillery fire
`postattack`	the number of insurgent attacks after Russian artillery fire

1.2. Did Russian artillery result in a greater number of deaths in Groznyy compared to the villages outside of Groznyy? Conduct the comparison in terms of the mean and median. (10 points)

1.3. Compare the average number of insurgent attacks after Russian fire for villages hit by artillery fire and those that were not hit. Also, compare the quartiles. Would you conclude that indiscriminate violence reduces insurgent attacks? Why or why not? (10 points)

1.4. What is the difference in the average number of insurgent attacks before Russian artillery fire between the villages that were hit by artillery fire and those that were not? What does this tell you about the validity of comparison behind the identification strategy used for the previous question? (10 points)

1.5. Create a new variable called `diffattack` by calculating the difference in the number of insurgent attacks before and after the Russian artillery fire. Among the villages shelled by Russians, did the number of insurgent attacks increase after the artillery fire? Give a substantive interpretation of the result. (10 points)

1.6. Compute the mean difference in the diffattack variable between villages shelled and villages not shelled. Does this analysis support the claim that indiscriminate violence reduces insurgency attacks? Is the validity of this analysis improved over the analyses you conducted in the previous questions? Why or why not? Specifically, explain what additional factor this analysis addresses when compared to the analyses conducted in the previous questions. (10 points)

2 Oil, Democracy, and Development (40 points)[2]

Researchers have theorized that natural resources may have an inhibiting effect on the democratization process. Although there are multiple explanations as to why this might be the case, one hypothesis posits that governments in countries with large natural resource endowments (like oil) are able to fund their operations without taxing civilians. Since representation (and other democratic institutions) are a compromise offered by governments in exchange for tax revenue, resource-rich countries do not need to make this trade. In this exercise, we will not investigate causal effects of oil on democracy. Instead, we examine whether the association between oil and democracy is consistent with the aforementioned hypothesis. The data set is in the csv file `resources.csv`. Table 2 presents the names and descriptions of variables in this data set.

Table 2: Oil, Democracy, and Development Data

Name	Description
`cty_name`	country name
`year`	year
`logGDPcp`	logged GDP per capita
`regime`	a measure of a country's level of democracy: - 10 (authoritarian) to 10 (democratic)
`oil`	amount of oil exports as a percentage of the country's GDP
`metal`	amount of non-fuel mineral exports as a percentage of the country's GDP
`illit`	the percentage of the population that is illiterate
`life`	the life expectancy in the country

2.1. Use scatterplots to examine the bivariate relationship between logged GDP per capita and life expectancy as well as between logged GDP per capita and illiteracy. Be sure to add informative axis labels. Also, compute the correlation separately for each bivariate relationship. Briefly

[2] This exercise is based on: Ross, M.L. (2001). "Does Oil Hinder Democracy?" World Politics 53 (3): 325-361.

comment on the results. To remove missing data when applying the `cor` function, set `use` argument to `"complete.obs"`. (10 points)

2.2. We focus on the following subset of the variables: `regime`, `oil`, `logGDPcp`, and `illit`. Remove observations that have missing values in any of these variables. Using the `scale()` function, scale these variables so that each variable has a mean of zero and a standard deviation of one. Fit the k-means clustering algorithm with two clusters. How many observations are assigned to each cluster? What are the characteristics of each cluster? (10 points)

2.3. Using the clusters obtained above, modify the scatterplot between logged GDP per capita and illiteracy rate in the following manner. Use different colors for the clusters so that we can easily tell the cluster membership of each observation. In addition, make the size of each circle proportional to the `oil` variable so that oil-rich countries stand out. Briefly comment on the results. (10 points)

2.4. Repeat the previous two questions but this time with three clusters instead of two. How are the results different? Which clustering model would you prefer and why? (10 points)