Quant I - Problem Set V
OLS

Kenneth Benoit

Assigned: Wednesday, December 1st, 2010
Due: Wednesday, December 15th, 2010

1. Economic development and homicide in some former Soviet states

   Table 1 gives information on economic development (GDP per capita in 1000 US$ at PPP) and homicide rates (homicides per 100,000 inhabitants) for a number of former Soviet Union countries in 2004.

Table 1: Economic development and homicide rates in 2004

Country	GDP	Homicide rate
Belarus	6.970	10.3
Moldova	1.729	8.4
Russia	9.902	21.9
Ukraine	6.394	8
Armenia	4.101	2.3
Azerbaijan	4.153	2.6
Georgia	2.844	6.5
Kazakhstan	7.44	13.9
Kyrgyzstan	1.935	8.3
Tajikistan	1.202	2.1
Turkmenistan	4.315	4
Uzbekistan	1.869	3.7

See also notes in text. *Source*: UNESCO TransMONEE database, 2006 edition.

   (a) Enter the data as three vectors into R (you can use three digit-abbreviations for the countries). (1 point)

   (b) We want to regress homicide rates (dependent variable) on GDP (independent variable). For doing so, calculate the OLS estimators of slope and intercept for this example. Please do this with R code that reflects the single steps of the calculation. (5 points)

(c) Verify your results in R with the built-in regression command. (2 points)

(d) Create a scatterplot of GDP and the homicide rate (make sure the homicide rate is on the vertical axis!). Choose the range of the axes appropriately and make sure that the intercept is visible. Add labels to the axes. Substitute your variable names in the command `text(xvariable, yvariable, countryvariable, cex=0.8, pos=4)` in order to add labels to the data points. (3 points)

(e) Add the regression line to the plot and include vertical and horizontal dashed lines at the respective means of the variables. What do the dashed lines illustrate? (4 points)

(f) Is there a negative association between economic development and homicide rates in this sample? Discuss in one sentence. (1 point)

2. Age and ideology in Germany

Load the dataset `GER02ageideol.RData`. As usual, the data come in form of a dataframe called `d`. After `load`ing the dataset, you can see the variable names by typing `names(d)` and access them using the `d$...` notation. The data come from a German election survey from 2002, and the variables are *age* in years and *leftright*, a self-placement on a left-right scale where 1 means most leftist and 11 most rightist.

(a) Calculate the mean of the age variable. (1 point)

(b) Center the age variable by subtracting its sample mean from every value. (1 point)

(c) Regress the left-right self-placement (we assume that this can be treated as a measure at the interval-level) on age using R's built-in regression command. (2 points)

(d) Interpret the estimates (intercept and slope) - make sure you make precise (numeric) statements with regard to what the numbers tell you. (You may e.g. use a difference of 10 years to discuss the coefficient of the age variable if you find that more convenient.) Why was it useful to center the independent variable? (5 points)

(e) Which coefficients are statistically significant and what does this mean? (4 points)

(f) Interpret the R-squared figure, include a precise numeric statement. What does this tell you about the association between age and ideological leaning (in Germany in 2002)? (3 points)

(g) According to this (very) simple model, what is the expected average self-placement of somebody who is 20 years older than the person of average age? (2 points)

(h) *Extra credit:* What is the expected average self-placement of somebody who is 18 years old? (*optional +2 points*)

Total: 34 points (*+2 optional*)