Practice Problems on Interpreting Linear Regression

1. Load the data set `clark-golder-2006.csv` available here.

2. Use the `glimpse()` function in the `dplyr` package to get a quick overview of the data. If you need more information, see the codebook here.

3. Clark and Golder's theory or model is more subtle than the explanation of the number of political parties we discussed in class. The authors suggest that it's social heterogeneity that really drives the number of political parties. After all, a society with more social groups needs more political parties to adequately represent those groups. However, they suggest that the electoral rules modify this relationship slightly. A simplified version of their theory suggests the following two hypotheses:

   Hypothesis 1 In single-member districts, just a few parties will emerge, regardless of the social heterogeneity.

   Hypothesis 2 In multimember districts, the number of parties that emerges depends on the social heterogeneity. If the society is heterogeneous, then many parties will emerge. If the society is homogenous, then only a few parties will emerge.

   Why is this? You might have heard of the "wasted vote" argument–if you vote for a third party in the U.S., you are "wasting" your vote. You hear this because single-member districts give elites and voters incentives to focus their resources and votes on the top two parties. We can crudely think of single-member districts as as countries with "non-permissive" electoral rules and multimember districts as as countries with "permissive" electoral rules. Non-permissive electoral rules give elites and voters an incentive to focus their resources and votes, while permissive electoral rules do not.

   To measure social heterogeneity, the authors use the effective number of ethnic groups (ENEG), which is exactly analogous to the effective number of parties, except applied to social groups.

   Now we'd like to see if the data seem to support this idea.

   (a) To do this, we'll break the data into two subsets. One subset should include only countries with an average district magnitude of one, and the other subset should include only countries with an average district magnitude of greater than one.

   (b) For the data set with single-member districts, create of scatterplot with ENEG along the x-axis and ENEP along the y-axis. What would Clark and Golder predict? Does that seem to be the case?

   (c) Now fit a linear model with ENEP as the outcome and ENEG as the explanatory variable. Are the coefficients consistent with Clark and Golder's prediction? Calculate the average ENEP for an ENEG of one. Calculate the average ENEP for an ENEG of four. If the ENEG increased from one to four, how much does the model say the average ENEP will change? Interpret the slope.

   (d) Now add the best fit line to the plot.

   (e) Repeat parts (b)-(d) using the data set with multimember districts.

4. Overall, do the data seem consistent with Clark and Golder's model? Does endogeneity seem plausible? Spuriousness?