Homework 6

Due: Wed, Nov 19 lab

1. Someone asked in class (Michael?) about the relationship between the F-test and the t-test. Specifically, when we have only two factors – say an X and a Y – are the two the same? We decided, based on about 2 seconds of thinking, that they probably are. Are they in fact?

   (a) Using this data, evaluate if we reject with the same probability.

   ```
   set.seed(1414)
   x <- rnorm(100, 0, 1)
   y <- 1 * 2*x + rnorm(100, sd = 10)
   ```

   (b) Now, show why this is case (using the properties of the t-test and F-test).

2. Create an interactive model using the 1995-1997 world values survey data. Don't worry, you don't have to find it, John Fox has, and placed it on his data index.

   (a) Convert the poverty measure into a numeric variable with values {1,2,3} – this certainly the *wrong* model, but...

   (b) Estimate a model using one of the dummies

   (c) Test for the necessity of an interaction

   (d) Test for the significance of the interaction at a few different x-values. Do this both using the call `vcov(...)` and something built-in.

   (e) Make a table, and plot that *clearly* show the effect (or lack of effect) that you found

3. This is a legacy problem. The datasets for this problem posted at the addresses listed below - there are no hyperlinks on the front page. *Enjoy!*

   (a) Create a working measure of *cohesion*. Let's use this operational definition:

   $$C_i = \sum_{i=1}^{n} \frac{|Y_i - N_i|}{Y_i + N_i}$$

   where i indexes bills, Y_i is the number of Yes votes on bill i, N_i is the number of No votes on bill i. C_i then is the cohesion of the congress on a particular bill i. If the party is split 50-50, then this cohesion measure will be 0. If it is all Yes or No it will obtain the maximum value.

   There are lots of bills that are passed in congresses (at least those that are not the present US Congress), so we aren't typically interested in cohesion on *one particular* bill but rather the average cohesion across many bills. We also are usually more interested in intra-party cohesion, rather than cohesion of a congress as a whole.

   This suggests that we might *actually* be interested in

   $$C_j = \frac{1}{n} \sum_{i=1,j=1}^{n} C_{ij} = \frac{1}{n} \sum_{1=1}^{n} \frac{|Y_{ij} - N_{ij}|}{Y_{ij} + N_{ij}}$$

which is the average cohesion for party j across all votes i. Y_{ij} is the Yes votes cast by legislators on bill i in party j and N_{ij} are the No votes cast by legislators on bill i in party j.

(b) Analyze state party cohesion in Brazil, using roll-call votes from the Brazilian Congress, as well as supplemental data (discussed below). The data is in a "tall" file of roll-call votes with four variables. DEPNCODE is a variable that identifies each legislator. The fourth and fifth character of that variable are the two-letter state abbreviation of the legislator. DEPPID is the party of the legislator. VNUM is a unique identifier for each bill voted on. DEPVOTE is a variable recording the vote of each legislator, coded "1" for "yes", "2" for "no", and "9" or other values for other patterns, including abstentions.

(c) Explore the impact of ideology, electoral partisanship, and development on state-party cohesion. Identify which variables (if any) are significant, and report the F-test. Provide a regression table like you might find in an academic journal. Interpret the slopes and intercept.

(d) You will have to combine several datasets to do this:

i. The roll-call vote dataset used last week, but now with cohesion calculated at the STATE party level, not just party. The fourth and fifth digits of legislators' id variable is a two-letter abbreviation for state.

ii. An additional dataset, contains party abbreviations and ideological codes from left (-1) to right (1).

iii. The census dataset has demographics by municipality. The numeric variable MUN is a unique number for each municipality. The variable UF is a numeric code for STATE. The variable POP records the population of the municipality, the variable POPRURAL has the number of residents in rural areas, the variable HOUSEHOLDS has the number of households in a municipality, and the variable HHLT1YEAR has the number of head of households with less than one year of education.

iv. The dataset here has two-letter abbreviations for states "uf2" and the census code for states "uf".

v. The dataset here has vote totals for all candidates in *all* races in the 1994 elections in Brazil, BY MUNICIPALITY. The variable COD_MUNIC, is the municipality number. The variable SGL_UE is the two-letter state abbreviation. The variable NUM_VOTAVEL is the candidate or party receiving votes. The variable QTD_VOTOS is the number of votes received. The variable COD_CARGO is the office sought; you want Congress, which is coded #7.
Candidate numbers (NUM_VOTAVEL) between 1000 and 9999 are for individual candidates - "personal votes". Candidate numbers between 10 and 90 are for party lists - "party votes". Numbers between 90 and 99 are blank or null ballots. So in the first row of the dataset, in municipality 62855, in the state of SP (Sao Paulo), candidate 1313 received 117 votes.

vi. The dataset parcodes has party names and party numbers to allow you to join datasets using one or the other.

Some thoughts:

- Get rid of the "9" values in the votes. We don't know how to score those for right now.
- Ask me if you have questions.

- Make sure you are using the right case (upper or lower) in the file names. Many Brazilian parties have formed, split, and merged, so don't worry if the data look weird.

This sounds like a bummer. It is – but it is also basic data management and basic data analysis, the kinds of things you will be doing for the rest of your career. I recommend drawing some pictures and outlining your strategy, then breaking it down step-by-step. Just do one part at a time. Perhaps one variable at a time, Y, then X1, then X2, then X3, to keep things easy.

Process Recommendation

- Make a list of datasets and the unit of analysis in each dataset.
- Then make a list of the desired unit of analysis for the final dataset.
- Draw arrows showing what you've got to do to get from here to there. Merge and/or aggregate as appropriate.
- Then analyze the data.