

POL 213: Research Methods III

Final Exam

Please answer question 1 and then choose to answer either question 2 or question 3.

1. Two data sets are available on the website. Both have the same set of variables but one is based on the NES 2000 survey and the second on the NES 2004 survey. Estimate the same models for both years (i.e. separate models, do not pool the data). The dependent variable in the data sets is `HispTherm` which is a 101-point feeling thermometer where higher scores indicate more favorable ratings for Hispanics and lower scores indicate less favorable ratings. The two independent variables are `Equal` and `Traits_Hispanic`. The first variable is 0 to 1 scale (not binary!) where higher scores indicate individuals with *less* equalitarian values and lower score indicate individuals with *higher* equalitarianism values. The second variable `Traits_Hispanic` is a scale ranging from .15 to 1 where higher scores indicate individuals who give very negative trait assessments of Hispanics and lower scores indicate individuals who give very positive trait assessments of Hispanics. The data are only for white, non-Latino respondents. For each data set set please do the following:

a. Estimate an OLS regression model and present results in tabular form. Give a succinct but precise interpretation of results. Interpretation need not be more than 1-2 paragraphs in length. In so-doing, discuss any differences you see between the 2000 and 2004 estimates. (20 points)

b. Next, estimate a generalized additive model using the R package `mgcv`. Present the GAM estimates in tabular form *as well as* graphical form. Specifically, plot the smoothing functions for each independent variable along with a contour plot from the model. For both models (2000 and 2004), fully interpret the results (most likely relying on the graphs). What if any, are the major differences between the GAMs and the OLS models? (25 points)

c. In either of the GAMS, would a linear term on the `Traits_Hispanic` variable fit the data better than a smoother? Why or why not. (5 points)

2. Using the R package `Matching`. Please use the data set `white_match.dta`, a Stata 10 data set for the input data for R. The data set has several variables. They are:

`new_inc_f`: Incumbent feeling thermometer on a 0-100 scale (100 most favorable).

`black_rep`: a binary indicator coded 1 if the incumbent MC is African-American; else 0.

`RESideology`: a seven-point scale where -1 denotes Extreme Conservative and 1 denotes Extreme Liberal.

`RESpid`: a seven-point scale where -1 denotes

Strong Republican and 1 denotes Strong Democrat.

tenure: denotes number of terms served in congress.

reseduc4: a four-point scale denoting education of respondent (4 denotes college degree).

age: respondent's age (in years)

redistrict: a binary indicator coded 1 if the cd was had been redistricted; 0 other wise.

cdcollege, cdblack, cdinc1k: percent college educated, percent black, and household income in the congressional district.

nom1: Poole-Rosenthal DW nominate scores (1st D)

Please note the following: these data are **only** based on white (non-Latino) respondents and the incumbents who are evaluated are **only** Democrat incumbents. Using these data, please answer the following questions/do the following exercises.

- a. Suppose a researcher believes whites “exposed” to an African-American MC will have lower incumbent evaluations than when compared to whites “exposed” to a non-African-American MC. The basis of this prediction might stem from some theory of racial prejudice or some theory of descriptive representation. Evaluate this expectation using a simple difference-in-means test. Quickly interpret the findings. (5 points)
- b. A researcher argues that the above scenario sounds like a treatment-control problem and that matching methods might be useful to evaluate the difference in incumbent ratings. Describe precisely why matching methods *may* be useful in this context. (5 points)
- c. Using the data given, estimate a matching model using the propensity scores method. For the propensity score model, use as covariates: `cdcollege` `cdblack` `cdinc1k` `RESideology` `RESpid` `redistrict` `tenure`. Report the logit estimates. Which covariates have the strongest relationship on predicting the treatment? What is the estimated treatment effect (just summarize the matching object, which will report the difference-in-means for treatment vs. control groups)? Is this difference significant? (15 points)
- d. Check for balance using not only the covariates included in the model but also the covariates: `nom1` `reseduc4` `age`. Which covariates seem to be balanced? Which do not? In general, why is balance a “good” thing? (10 points)
- e. Now estimate a matching model using a genetic algorithm. Consult Sekhon's *JSS* piece for a summary of the methodology. Below, I give the shell R script you can use to estimate the model (this may take a long time to converge):

```

X <- cbind(X1, X2, ... X3) #Vars used for matching; this gives the matrix we're matching on.

BalanceMatrix <- cbind(X1, X2, ... X3) #Vars used for matching

gen1 <- GenMatch(Tr=Tr, X=X, BalanceMatrix=BalanceMatrix,
  pop.size=1000, max.generations=10, wait.generations=1) #This will take several minutes to run!

mout <- Match(Y=Y, Tr=Tr, X=X, estimand="ATT", Weight.matrix=gen1) #Getting treatment effect

gm<-MatchBalance(treatment~ X1 + X2 + ... + XJ,
  match.out=mout, nboots=1000, data=INPUT DATA) #The Xs are the variables on which you're examining balance

```

f. Describe the basic differences in this approach compared to the propensity score method. Are the balance statistics appreciably different? If so, on which covariates? What is the treatment effect given by the genetic algorithm? Of the three models estimated of the causal effect, which would you choose to report and why? What possible problems do you see with the matching model, as stipulated? (15 points)

3. For this question, please access the data file `polviews.dta`, a **Stata** 10 dataset. This data set has the following covariates: **conservative** is a seven-point scale ranging from 1 (extremely liberal) to 7 (extremely conservative). Treat this as the response variable y in the models that follow (hold your nose; I'm going to ask you to run regression). The dependent variables are: **education** coded in years completed of schooling; **male** coded 1 if male, 0 if female; **white** coded 1 if white, 0 if non-white; **income** coded as intervals of income (higher scores indicate higher income); **religious** coded such that lower scores indicate greater religiosity; **partyid** coded on a scale where 1 denotes strong Democrat and 7 denotes strong Republican; **region** coded categorically to denote region of residence; **union** codes union membership; **age** codes age of the respondent. (Don't worry, you will not need to interpret all these variables).

a. Estimate an OLS model treating y as a function of: education, male, income, religious, partyid region, union, age. In a paragraph, summarize the major features of this model. (5 points)

b. In order for someone to be observed in part a, the respondent had to offer a response on y , the outcome variable. The variable `missingview` is binary coded 1 if the respondent offered a response and 0 if he/she did not. To assess selection effects, estimate a probit model treating the selection indicator as a function of: education, male, white, income, religious. In a paragraph, describe the major features of this model. (5 points)

c. Knowing the normal density is given by $[1/\sqrt{2\pi}] * \exp(-(x\hat{\beta})^2/2)$ and the normal cdf is given by the function `normal($x\hat{\beta}$)`, where `normal` is a **Stata** function and $x\hat{\beta}$ is the *linear prediction* from the probit model, generate the inverse Mills ratio and rerun the model in part a including the IMR as a predictor. Give a brief interpretation of the coefficient for the IMR. What is this model accounting for that the one in part a) did not? (Note that π in **Stata** can be accessed with command `_pi`.) (20 points)

d. Now, using **Stata's** `heckman` procedure (with the `twostep`) option, estimate the models in part b and part c. Compare the results across four models? Are your results from parts b and c similar to the results given in this model. Provide an interpretation of the education coefficient. Why might this coefficient change across these models? (20 points)