Problem Set 3

Due Thurs Nov 10

1.) We'll plot some ROC curves using the spam example from `lec03--regression.nb.html`. You can use the R function for plotting ROC curves, uploaded to Blackboard.

a.) Estimate the logit specification from the lecture notes on the full data set (you can just copy the code from the lecture). Plot an ROC curve for classifying spam vs not-spam from this logit regression.

b.) Plot an out-of-sample ROC curve by estimating the same logit specification on 4/5 of the data (sampled randomly), and evaluating classification error on the remaining 1/5 of the data.

2.) The data set `dw_data.csv` is taken from Dehejia and Wahba[1]. It contains data on individuals who participated in a job training program (for whom `treated` equals `TRUE`) and "control" individuals from the PSID (for whom `treated` equals `FALSE`). Additional variables are: age, education, Black (1 if black, 0 otherwise), Hispanic (1 if Hispanic, 0 otherwise), married (1 if maried, 0 otherwise), RE74 and RE75 (earnings in 1974 and 1975), UE74 and UE75 (indicator for unemployment in 1974 and 1975), and RE78 (earnings in 1978). The training program was done before 1978, so the last varible is post-training earnings.

a.) Estimate a linear model with constant treatment effects by regressing RE78 on the treatment indicator and the remaining variables. Report the estimate of the average treatment effect from this regression (using the assumptions of selection-on-observables and correct specification of this regression).

b.) Add interactions of the treatment indicator with the remaining variables to the regression in (a). Compute an estimate of the average treatment effect from this regression (using the assumptions of selection-on-observables and correct specification of this regression).

c.) The data set `dw_experimental_data.csv` includes both the treated observations in the `dw_data.csv` data set and a set of true control individuals, such that treatment was randomized among this population. Parts (a) and (b) can thus be considered an attempt to recover the outcome of this experiment using observational data where a true control group is not available.

Using the data set `dw_experimental_data.csv`, estimate the average treatment effect by taking the difference in means between treated and control groups. Is the answer similar to (a) and (b)?

[1] Dehejia, Rajeev H., and Sadek Wahba. "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs." Journal of the American Statistical Association 94, no. 448 (December 1, 1999): 1053–62. https://doi.org/10.1080/01621459.1999.10473858.

3.) We'll run a placebo test for the RD estimator using the Lee data.

a.) Run a local linear regression estimator (with uniform kernel) as described in the lecture notes. Use $h = 5$ as the bandwidth.

b.) Form placebo estimates β_b^* at cutoffs $c_1, \ldots, c_B$ given by $-80, -79, \ldots, -7, -6, -5$ and $5, 6, 7, \ldots, 79, 80$. Plot a histogram of the placebo estimates β_b^* along with a vertical line at the estimate computed in part (a).