PS C236A/ Stat C239A
Problem Set 7
Due: November 2, 2009

Problem 1 In this homework, you will analyze data used in:

> Chris Blattman and Jeannie Annan (2009). "The Consequences of Child Soldiering." *The Review of Economics and Statistics*

The data is from a panel survey of male youth in war-afflicted regions of Uganda. Blattman and Annan are interested in estimating the impact of abduction by the rebel group "Lord's Resistance Army" on economic and psychological outcomes.

They describe abduction as follows:

> Abduction was large-scale and seemingly indiscriminate; 60,000 to 80,000 youth are estimated to have been abducted and more than a quarter of males currently aged 14 to 30 in our study region were abducted for at least two weeks. Most were abducted after 1996 and from one of the Acholi districts of Gulu, Kitgum, and Pader.
>
> Youth were typically taken by roving groups of 10 to 20 rebels during night raids on rural homes. Adolescent males appear to have been the most pliable, reliable and effective forced recruits, and so were disproportionately targeted by the LRA. Youth under age 11 and over 24 tended to be avoided and had a high probability of immediate release. Lengths of abduction ranged from a day to ten years, averaging 8.9 months in our sample. Youth who failed to escape were trained as fighters and, after a few months, received a gun. Two thirds of abductees were forced to perpetrate a crime or violence. A third eventually became fighters, and a fifth were forced to murder soldiers, civilians, or even family members in order to bind them to the group, to reduce their fear of killing, and to discourage disobedience.

As Blattman and Annan note, the most important factors determining selection into abduction (the "treatment") are location and age. In addition to these covariates, the dataset includes a number of variables related to the socioeconomic status of the respondents' families. The outcome variables are years of education, log wages, and an additive index of psychological distress using 19 self-reported symptoms of depression and anxiety. The variables in the dataset are listed in the codebook on the course website.

a. Select a set of covariates to condition on. Be sure to consider if any higher order terms and interactions are appropriate. Using these variables, perform Mahnolobis distance matching on a propensity score and "orthogonalized" covariates, with ATT as your estimand. Report your balance statistics, preferably using a plot.

b. Now using the same set of covariates (propensity score and "orthogonalized" covariates), use `GenMatch` to generate weights that optimize balance. Use the default setting for the loss function, but feel free to adjust other parameters of the function. Present balance before and after matching.

c. Now find balance with `GenMatch` using your own loss function. Explain the logic behind your choice of loss function. You may want to prioritize important selection variables, such as age. Present balance statistics after matching.

d. Create a loss function that ensures that GenMatch will not return a matched data set with worse balance on any variable in `BalanceMatrix` than the balance obtained by pscore matching—as judged by eQQ-plots and difference of means. Do this so that this property holds by design—i.e., it holds regardless of the dataset used. In order to make this happen, you will have to both write a custom loss function (you may alter the one created in question "c" or write a new one), and provide GenMatch with `starting.values` so that it starts with the pscore matched dataset. Report your balance statistics after using this loss function.

e. Pick the matching method that produced, in your judgement, the best balance. Estimate treatment effects and report them. How do they differ from the reported estimates?