Ec142, Spring 2018

Professor Bryan Graham

Problem Set 1

Due: February 16th, 2017 (note this is the final day to add/drop classes for the Spring 2018 semester)

Problem sets are due at 5PM in the GSIs mailbox. You may work in groups, but each student should turn in their own write-up (including a printout of a narrated/commented and executed Jupyter Notebook). Please also e-mail a copy of any such notebooks to the GSI (if applicable). Please use markdown boxes within your Jupyter notebook for narrative answers to the questions that appear below.

1 Using inverse probability weighting (IPW) to evaluate an innovative secondary education program

Read the paper "Improving middle school quality in poor countries: evidence from the Honduran *Sistema de Aprendizaje Tutorial*" (McEwan et al., 2015). You should also review your lectures notes and read Holland (1986), Efron & Hastie (2016, Chapter 8) and Hirano & Imbens (2001).

Overview of dataset

This problem set uses the dataset `math_10.out`; available on the course GitHub page. The dataset includes information on 713 youth that are part the larger dataset analyzed by McEwan et al. (2015). These students resided in 40 of the 59 matched SAT-CEB pairs described in the article. The following variables are included:

`c10_zmath` – math test score (standardized) at follow-up in the Fall of 2010

`wgt10_math` – test score weight (equals 1/2 if student took "in home" version of test, 1 otherwise). See McEwan et al. (2015, p. 118) for more information.

`sat` - dummy variable indicating whether student resides in a SAT village or a CEB village.

`constant` – a column with a 1 in every row

`c08_zlang` - language test score at baseline in the Fall of 2008

`c08_zmath` - math test score at baseline in the Fall of 2008

`feeder_school` – feeder_score or "village" code `m_id_pairs` – code for 40 matched SAT-CEB village pairs

The dataset is a tab delimited text file. You can uses the `pandas.read_csv()` function to read it into your Notebook as a dataframe. The following snippet of code loads the dataset as a pandas dataframe and computes questions 1 and 2 below.

```
# Load core data science libraries
import numpy as np
import scipy as sp
import pandas as pd

# Import StatsModels library
import statsmodels.api as sm

# Location of math_10.out file
data  = '/Users/bgraham/Dropbox/Research/SAT/created_data/'

# Read in tab delimited dataset into a pandas dataframe
col_dtypes = {'c10_zmath' : float, 'wgt10_math' : float, 'sat' : int, \
        'constant' : int, 'c08_zlang' : float, 'c08_zmath' : float, \
        'feeder_school' : int, 'm_id_pairs' : int}
df = pd.read_csv(data + 'math_10.out', dtype = col_dtypes, \
        na_values='', engine='c', sep = '\t', encoding = 'utf-8')

# Construct a list of all matched SAT-CEB village pairs in the dataset
included_pairs = sorted(df['m_id_pairs'].unique())

# Form dummies for included matched SAT/CEB pairs
pair_dums = pd.get_dummies(df['m_id_pairs'].astype('category'), prefix='mp')

# Concatenate matched pair dummies onto dataframe
df = pd.concat([df, pair_dums], axis=1)

# Construct outcome vector, design matrix and
# test instrument inverse weights
Y = df['c08_zlang']                # Outcome
test_wgt = 1./df['wgt10_math'] # Test instrument dummies
X = df[['constant','sat']]         # Design matrix
```

```
X = pd.concat([X, df.loc[:, 'mp_' + str(included_pairs[0]) : \
        'mp_'+ str(included_pairs[-2])]], axis = 1)
# NOTE: omit last matched pair to avoid "dummy variable trap"

# Compute weighted least squares fit
# NOTE: cluster-robust standard errors
wls = sm.WLS(Y,X,weights=test_wgt).fit(cov_type='cluster', \
        cov_kwds={'groups': df['feeder_school']},use_t=True)
wls.summary()
```

Analysis

1. Create a dummy variable for each of the 40 matched SAT-CEB pairs. You can use the `pandas.get_dummies()` function to do this (see code snippet above).

2. Compute the weighted least squares (WLS) fit of `c08_zmath` onto a constant, `sat`, and the matched SAT-CEB pair dummies (you will need to exclude one dummy to avoid the "dummy variable trap"). Weight by the inverse of the `wgt10_math` weights included in the dataset (see code snippet above). Report cluster-robust standard errors (see code snippet above). Interpret the coefficient on `sat` in light of the research design described by McEwan et al. (2015). What is accomplished by weighting by the inverse of the `wgt10_math` weights?

3. Compute the WLS fit of `c10_zmath` onto a constant, `sat`, and the matched SAT-CEB pair dummies (you will need to excluded on dummy to avoid the "dummy variable trap"). Weight by the inverse of the `wgt10_math` weights. Report cluster-robust standard errors. Interpret the coefficient on `sat` in light of the research design described by McEwan et al. (2015).

4. Additionally control for `c08_zmath` and `c08_zmath` in the WLS fit computed immediately above. Interpret the coefficient on `sat`.

5. Compute the logistic regression fit of `sat` onto a constant, the matched SAT-CEB pair dummies and the two baseline test scores (i.e., `c08_zmath` and `c08_zmath`). Compute the fitted propensity score values. Is the overlap condition satisfied? Why? Present evidence for your answer.

6. Compute the IPW weights for average treatment effect (ATE) estimation as described in lecture and also Hirano & Imbens (2001). Multiply these weights by the `test_wgt`

constructed in problem 2 above. Compute the weighted least squares fit of `c08_zmath` onto a constant and `sat` using these weights. Report cluster-robust standard errors. What is accomplished by using these weights. Interpret the coefficient on `sat.`

References

Efron, B. & Hastie, T. (2016). *Computer Age Statistical Inference*. Cambridge: Cambridge University Press.

Hirano, K. & Imbens, G. W. (2001). Estimation of causal effects using propensity score weighting: an application to data on right heart catheterization. *Health Services and Outcomes Research Methodology*, 2(3-4), 259 – 278.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396), 945 – 960.

McEwan, P. J., Murphy-Graham, E., Irribarra, D. T., Aguilar, C., & Rápalo, R. (2015). Improving middle school quality in poor countries: evidence from the honduras sistema de aprendizaje tutorial. *Educational Evaluation and Policy Analysis*, 37(1), 113 – 137.