Ec142, Spring 2018

Professor Bryan Graham

Problem Set 4

Due: May 11th, 2018

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 Jupyter Notebook to the GSI (if applicable).

NOTE: Because this problem set is being made available late, I have decided to make this homework optional. Specifically, if you do not do this assignment, your homework grade will be determined by the two best homework grades out of the three homeworks assigned so far. If you do complete this assignment, your homework grade will be computed using the two best homework grades out of the four homework assignments overall (i.e., including this one). To provide some incentive to do this assignment irrespective of your homework performance so far, I will add 3 bonus points to the homework aggregate to any student completing this assignment. In principle students could then get a maximum homework grade of 23 out of 20.

1 Discrete time hazard analysis: computation/illustration

The file **marriage.out** contains age at marriage information for 616 female participants in the study described in McEwan et al. (2015). The dataset in organized in the "person-period" format described in lecture and also in Chapter 10 of Singer and Willet (2003) book listed on the course syllabus. The following variables are included in the dataset:

- PID – an individual respondent identification number
- AGE – respondents age
- Y – binary indicator equal to one if the respondent married for the first time at the current age and zero otherwise
- SAT – binary indicator equal to one if the respondent lives in a SAT village and zero otherwise (see McEwan et al. (2015) for more information on the SAT program)
- MATH – standardized baseline math test score
- LANG – standardized baseline language arts test score
- VILLAGE_PAIR – identification number of matched SAT-CEB village pairs (you will not use this in the problem set, but again see the paper if you are interested)

Missing values in the dataset are coded as "-999".

1. After loading the data as a Pandas dataframe called "PersonYear" type into your notebook

```
print PersonYear[0:32]
```

   This prints out the rows of the "person-period" dataset associated with the first five respondents. Describe the marriage histories, including a discussion of censoring (if any), of these four women. [1 to 2 paragraphs].

2. Create a dummy variable for each unique age value in the dataset and add these to your dataframe. You can do this with the following line of code

```
PersonYear = pd.concat([PersonYear, \
                        pd.get_dummies(PersonYear['AGE'], \
                                       prefix = 'AGE')], axis=1)
```

   This will create dummies "AGE_12", "AGE_13",..., "AGE_28". Due to the study design, there are very few girls followed beyond age 22. Create a dummy variable for all rows corresponding to ages 23 and above and call it "AGE_23+".

3. Compute the discrete-time baseline hazard using logistic regression as described in lecture and Singer and Willet (2003, Chapter 11). You may use the StatsModels "Logit" command which has a syntax almost identical to that of the "OLS" command. An accessible introduction to logistic regression analysis in Python can be found online at `http://blog.yhat.com/posts/logistic-regression-python-rodeo.html`. In your model include the dummies "AGE_12", "AGE_13",.... "AGE_22" and "AGE_23+". Do not include a constant in your model. Why? [2 -3 sentences] Plot the baseline hazard for ages 12 to 22. Remember the coefficients on the dummy variables correspond to the logit of the baseline hazard so you will need to transform them prior to plotting.

4. Plot the Kaplan-Meier survival function estimate based on your analysis in #3.

5. Plot 95 percent point-wise confidence intervals around your estimated survival function.

6. Add the covariates MATH, LANG, their interaction as well as their squares. Comment on your results [2 to 4 sentences].

7. Add the SAT dummy variable to the model introduced in #6 above. Interpret the estimated coefficient on SAT in light of the discussion of the SAT program provided in McEwan et al. (2015). How does exposure to SAT influence (or not influence) age at first marriage [1 to 2 paragraphs].

References

[1] McEwan, P. J., Murphy-Graham, E., Torres Irribarra, D., Aguilar, C., & Rápalo, R. (2015). Improving middle school quality in poor countries: Evidence from the Honduran Sistema de Aprendizaje Tutorial. *Educational Evaluation and Policy Analysis*, 37(1), 113-137.