Homework 2
PP 414 Applied Regression Analysis
Instructors: Black and Delgado
Due: November 5th
Fall 2018

Directions: Go to the IPUMS web site: https://www.ipums.org/. You will need to create an account. Then go to the IPUMS-USA and select your data from the 2016 American Community Survey (ACS). You will need to gather data from all women residing in the state of California between the ages of 20 and 40 (inclusive). Be sure to get the person weights. You will be estimating a binary choice model whether she has any of her own children in the household (please use the "nchild" variable). You will need to find information on her education level (please use the "educd" variable), whether she is currently enrolled in school, her age, race, Hispanic status, place of residence (please use the "met2013" variable), and for the next few assignments also gather information on her hours worked, weeks worked, and earnings (please use both "incwage" and "incearn"). Also for future assignments, be sure to pick up the data imputation codes.

For race, please divide the data into white, black, Asian, and other. Please include multi-racial people into other race category. Define Hispanic status as a binary variable. For the place of residence, use MSA from "met2013" or a dummy variable for not living in a MSA. For education variable, define a sequence of dummies that include, less than high school, some high school but no degree, GED, high school graduate, some college but no degree, associate's degree, bachelor's degree, master's degree, professional degree, and doctoral degree.

1. Run an OLS equation using whether the woman has a child or not against her age, education, race, Hispanic status, and place of residence. Recover the predicted values. Are they within the unit interval?

2. Run the same equation using a logit model rather than the OLS. Recover the predicted values. What is the correlation between the predicted values in question one and two?

3. Run a fully saturated OLS model. How many instances do you a predicted value of zero or one?

4. Run a logit model in which you use a second-order approximation. Recover the predicted values. What is the correlation among the predicted values from questions one, two, and four.

5. Use a five-fold cross validation to pick the best out-of-sample estimates for questions one, two, and four. How do the out-of-sample predictions compare to the within sample prediction you derive in question four?