**Education 265: Problem Set 5**

**IMPORTANT: I have marked the section that you should do on your own.**

**Due:** Friday, February 5, 2020 at 1:00 p.m. in the class Canvas folder (Assignments tab 🡪 Problem Set 5). Turn in

* Your answer sheet as a Word document (titled “**LastName\_PS5**”)
* Type your answers in **bold directly** under the questions
* Your do-file (titled “**LastName\_PS5\_Code.do**”).

**Objective**

1. We will practice estimating dichotomous outcome variables using a Linear Probability Model.
2. Combine what you know about interactions with Linear Probability Models.

**Set up your do-file.** Start by setting the path directory (use globals or just the cd command) so that any data files are saved to the right folder. The dataset to use for this problem set is “***for\_ps5.dta***”.

**Dichotomous outcome variables**. We will run models that investigate the relation between spring of kindergarten externalizing behaviors, gender, and school-entry math achievement.

1. We will add a set of controls that includes the following,

* standardized fall of kindergarten reading achievement (‘zread1’);
* dummies for Blacks, Hispanics, Asians, and Other ethnicity (omit Whites as the comparison group);
* dummy for mothers with a high school diploma, vocational school, or some college; and dummy for mothers that graduated from college or attended graduate school (omit mothers with less than a high school diploma as the comparison group);
* and the Fryer and Levitt variables, namely: age in month, birthweight in ounces, female, number of children’s books, ***number of children’s books squared***, mother over 30 at first birth, mother a teenager at first birth (mother age 19-29 at first birth omitted as reference group), continuous measure of socioeconomic status, mother received WIC benefits.
* Note: we have already cleaned these variables, so you only need to find the variables with look up, inspect that the variables to make sure they look okay, and put them into a global
* Create a global list for the variables (“controls”), write just 1 line of code to quickly inspect that nothing looks weird, then paste the code below.

**global controls "zread1 dblack dhispanic dasian dother dhighcol dcolplus ageinmonth bw\_oz books books2 dmage\_30plus dmage\_teen wksesl momwic"**

**codebook $controls**

1. Create an analysis sample (“sample\_ps5”) that is equal to 1 if students are non-missing on fall and spring of kindergarten measures of math achievement (‘zmath1’ & ‘zmath2’), non-missing on our outcome variable “t2extern”, and non-missing on all our control variables. (*Hint:* *analysis sample should include 13,721 observations).* Paste your code for creating the analysis sample to your answer sheet.

*All of the following analyses should be restricted to the analysis sample.*

**gen sample\_ps5 = 0**

**replace sample\_ps5 = 1 if !mi(zmath1, zmath2, t2extern, dblack, dhighcol, ageinmonth, bw\_oz, dfemale, books, dmage\_30plus, wksesl, momwic, zread1)**

**tab sample\_ps5**

1. Create a dummy variable called “highext” (short for high in externalizing behavior). Set it equal to 1 if a student was above the mean in teacher-rating of spring kindergarten externalizing (“t2extern”). Set it to 0 if the student had valid data on the measure, but was *at or below* the mean. (*Hint: “highext” should have 19,188 valid cases, 8,260 of which should be coded as 1).*

Paste your code for creating “highext” and the analysis sample to your answer sheet.

**egen ext\_mean = mean(t2extern)**

**gen highext = .**

**replace highext = 1 if (t2extern > ext\_mean & t2extern != .)**

**replace highext = 0 if t2extern <= ext\_mean**

**tab highext**

1. Run a linear probability model (LPM) in which you predict ‘highext’ using fall of kindergarten math score, gender, and our set of controls. Report and interpret the coefficient produced by zmath1.

**Coefficient: -.037 (.006) at p-value < .001**

**A 1 standard deviation increase in fall of kindergarten math score is associated with a 3.7 percentage point decrease in probability of getting a higher that mean score in spring kindergarten externalizing, while holding control variables constant.**

1. Run the model from Q4 again and predict the probabilities for each student. (hint: use the *predict* command). Make a histogram for the distribution of the predicted probability and paste it below. What do you notice about the distribution?



**There are values below zero in our predictions which is in contrast with the nature of probability range that is [0,1] that is one of the main issues with LPM models.**

Hint: add the following options to the histogram code to make easier to read (and use these kinds of options to make graphs look better later)

percent ///

title("probability of high externalizing behavior at school-entry") ///

subtitle("predicted by fall-K math score and controls, using LPM") ///

xtitle("probability") ///

ytitle("percent of observations") ///

note("Data come from ECLS-K", span) ///

name(lpm, replace)

Do the following on your own.

**Dichotomous outcome predicted using interactions**. Still predicting the probability of being rated high on externalizing problem in spring-k using standardized math score in fall-k and our set of controls, we now examine whether the association differs between females and males.

1. Run the LPM model (i.e., Q4) on females only, report and interpret the coefficient of zmath1. (Hint: *you should take dfemale out of the global control list since we are restricting our sample based on gender now*).

**Coefficient: -.043 (.01) at p-value < .001**

**For female students, a 1 standard deviation increase in fall of kindergarten math score is associated with a 4.2 percentage point decrease in probability of being rated high on externalizing problem in spring-k, while holding control variables constant.**

1. Run the LPM model (i.e., Q4) on males only, report and interpret the coefficient of zmath1, how does it compare to the coefficient we got in Q11?

**Coefficient: -.034 (.008) at p-value < .001**

**For male students, a 1 standard deviation increase in fall of kindergarten math score is associated with a 3.4 percentage point decrease in probability of being rated high on externalizing problem in spring-k, while holding control variables constant.**

1. Now run the LPM model (i.e., Q4) predicting the probability of being rated high on externalizing problem in spring-k using standardized math score in fall-k, gender, our full set of controls and female interacted with math score and all the controls (run a fully saturated model). Report and interpret the one coefficient on the interaction between math and female (be specific with that this tells you).

**Coefficient: -.009 (.013) at p-value = 0.469**

**1 SD increase in fall of kindergarten math score is associated with .9 percentage points more decrease for female students compared with males in probability of being rated high on externalizing problem in spring-k, while holding control variables constant.**

**Also, the p-value suggests that this coefficient is statistically insignificant.**

1. Time you took to finish this problem set: \_\_\_**3**\_\_\_ (sorry, this question got left off last week by accident).