Jeff Grogger
Harris School

Winter 2022
University of Chicago

PP 420: Problem Set 3
(due Tuesday, Feb 22 by 11:59pm via Canvas)

A. Analytical problems

1. Let the true model be given by

$$y = \beta_2 x_2 + \beta_3 x_3 + \gamma_2 x_2^2 + \gamma_3 x_3^2 + \gamma_4 x_2 x_3 + \varepsilon \qquad (1)$$

where the data are centered and $E(\varepsilon|X) = 0, X = (x_2, x_3)$. Suppose you estimate the regression

$$y = \beta_2 x_2 + \beta_3 x_3 + u. \qquad (2)$$

Show that, if the joint distribution of *X* is symmetric, then the OLS estimates of β_j from (2) are consistent for $E[\frac{\partial y}{\partial x_j}]$, j=2,3. *Hint:* In a multivariate symmetric distribution, all third-order central moments are zero.

2. Consider the one-way fixed effects model

$$y_{it} = x_{it}'\beta + \alpha_i + \varepsilon_{it} \qquad i=1,\ldots,N;\ t=1,\ldots,T. \qquad (1)$$

First-differencing (1) and subtracting yields

$$y_{it} - y_{it-1} = (x_{it} - x_{i,t-1})\beta + \varepsilon_{it} - \varepsilon_{it-1}. \qquad (2)$$

Applying OLS to (2) yields the first-difference estimator, denoted b_{FD}.
(a) Provide conditions sufficient to establish the consistency of b_{FD}.
(b) Derive $V(b_{FD})$.
(c) Show that $b_{FE} = b_{FD}$, that is, that the two estimates are numerically equivalent, when T=2.

3. Consider the one-way fixed-effects model

$$y_{it} = x_{it}'\beta + \alpha_i + \varepsilon_{it} \qquad i=1,\ldots,N;\ t=1,\ldots,T. \qquad (1)$$

Let b_{FE} denote the estimate of β obtained by applying OLS to (1), that is, by regressing y_{it} on x_{it}' and a set of unit dummies. Define $\bar{x}_i. = T^{-1}\sum_{t=1}^{T} x_{it}$ and consider the estimating equation

$$y_{it} = x_{it}'\beta + \bar{x}_i.'\,\pi + v_{it}. \qquad (2)$$

Let $\hat{\beta}$ denote the estimate of β obtained by applying OLS to (2). Show that $b_{FE} = \hat{\beta}$. *Hint:* Use the partitioned inverse rule. Explain why there might be a practical advantage to basing estimation on (2) rather than (1).

B. Computational Problems

(1) Using the pp420_data.dta data set, regress earnings on education, age, and age squared.

(2) This model assumes that the value of an additional year's schooling is the same, regardless how much education the worker has. An alternative specification assumes that education has a discrete effect at particular thresholds: having 12 years of schooling vs. having less than 12, and having more than 12 years of schooling vs. having 12. Formulate and estimate such a model (check that your sample size is the same as in (1)). What is the value of having a high school diploma vs. not having one (you get a diploma at the end of the 12th year of school). What is the value of having more than a diploma? Is this number statistically different from zero? Between this model and the model from (1), which provides a better fit to the data? Explain.

(3) A problem with the model from (2) is that it assumes that the value of an additional year's education is zero below the high school graduation threshold. Formulate and estimate a model that allows you to test this implicit assumption. Does it provide a better fit to the data than the model from (2)? What is the value of having 11 years of schooling, versus having only 9 years of schooling? What is the value of having 12 years, versus having only 11?

(4) Now add race/ethnicity dummies to your model from (3). Do they belong in the regression? How do they affect the estimates of the education and age coefficients? What does this say about the correlation between the race/ethnicity dummies and the education and age variables? Can you verify this in the data?

(d) Referring again to the model from 2(b), provide an informal check for the presence of heteroskedasticity. Based on this check, do you think it is important to provide a formal test?