The University of Texas at Austin
ECO 348K (Advanced Econometrics)
Prof. Jason Abrevaya
Fall 2016

PROBLEM SET #6 (due by Tuesday, November 22nd, 9:30am at the BRB front desk)

1. (Wooldridge 17.C3) Use the data in **fringe.dta** for this question. This dataset is for a 1977 sample of workers, where the main outcome of interest for this question is *pension*, which is the monetary value of an employee's pension.

   (a) For what percentage of the workers in the sample is *pension* equal to zero? What is the range of *pension* for workers with non-zero pension benefits? Why is a Tobit model appropriate for modeling *pension*?

   (b) Estimate a Tobit model explaining *pension* in terms of *exper*, *age*, *tenure*, *educ*, *depends*, *married*, *white*, and *male*. Do whites and males have statistically significant higher expected pension benefits?

   (c) Use the results from part (b) to estimate the difference in expected pension benefits for a white male and a nonwhite female, both of whom are 35 years old, are single with no dependents, have 16 years of education, and have 10 years of experience.

   (d) Add *union* to the Tobit model and comment on its significance.

   (e) Apply the Tobit model from part (d) but with *peratio*, the pension-earnings ratio, as the dependent variable. (Notice that *peratio* is a fraction between zero and one, but, though it often takes on the value zero, it never gets close to one. Thus, a Tobit model is fine as an approximation.) Does gender or race have an effect on the pension-earnings ratio?

2. (Adapted from Wooldridge 17.5) Let *patents* be the number of patents applied for by a firm during a given year. Assume that the conditional expectation of *patents* given *sales* and *RD* is

$$E(patents|sales, RD) = \exp\left[\beta_0 + \beta_1 \ln(sales) + \beta_2 RD + \beta_3 RD^2\right],$$

   where *sales* is annual firm sales and *RD* is total spending on research and development over the past 10 years.

   (a) How would you estimate this model? (This is easy— one sentence.)

   (b) What is the interpretation of β_1?

   (c) Find the partial effect of *RD* on $E(patents|sales, RD)$.

   (d) (Extra— not in Wooldridge) Suppose you want to estimate the conditional-variance model

$$Var(patents|sales, RD) = \exp\left[\gamma_0 + \gamma_1 \ln(sales) + \gamma_2 RD + \gamma_3 RD^2\right].$$

If you already had estimates $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$ of the parameters from the conditional-expectation model, how would you estimate the conditional-variance model? (Hint: Think about estimated residuals.)

3. (Based upon a question from Wooldridge's graduate textbook) A common setup for program evaluation with two periods of panel data is the following. Let y_{it} denote the outcome of interest for unit i in period t. At $t = 1$, no one is in the program. Let $prog_{it}$ be a binary indicator equal to one if unit i is in the program in period t; by the program design, $prog_{i1} = 0$ for all i. At $t = 2$, some units are in the control group ($prog_{i2} = 0$) and others are in the experimental ("treatment", $prog_{i2} = 1$) group. A fixed effects model without additional covariates is

$$y_{it} = \theta_1 + \theta_2 d2_t + \delta_1 prog_{it} + a_i + u_{it}, \quad E(u_{it}|prog_{i2}, a_i) = 0,$$

where $d2_t$ is a dummy variable equal to one if $t = 2$ and zero if $t = 1$, and a_i is the unobserved effect.

   (a) Explain why including $d2_t$ is important in these contexts. If particular, what problems might be caused by leaving it out?

   (b) Why is it important to include a_i in the equation?

   (c) Using the first differencing method, show that $\hat{\theta}_2 = \overline{\Delta y}_{control}$ and $\hat{\delta}_1 = \overline{\Delta y}_{treat} - \overline{\Delta y}_{control}$, where $\overline{\Delta y}_{control}$ is the average change in y over the two periods for the group with $prog_{i2} = 0$, and $\overline{\Delta y}_{treat}$ is the average change in y for the group where $prog_{i2} = 1$. (This formula shows that $\hat{\delta}_1$, the difference-in-differences (DD) estimator, arises out of an unobserved effects panel data model.)

4. (Adapted from Wooldridge 13.C8) Use the data in **vote2.dta** for this question. Note that each observation has information for two time periods (1988 and 1990). This data considers House of Representatives elections for those two years, where only winners from 1988 who also run in 1990 are included in the sample. These 1988 winners are known as "incumbents". A fixed effects model explaining the share (a number between 0 and 100) of incumbent's vote in terms of expenditures by both candidates is

$$vote_{it} = \beta_1 + \beta_2 \ln(inexp_{it}) + \beta_3 \ln(chexp_{it}) + \beta_4 incshr_{it} + \delta_1 d90_t + a_i + u_{it},$$

where $incshr_{it}$ is the incumbent's share of total campaign spending (in percentage form, a number between 0 and 100), $inexp_{it}$ is the incumbent's expenditures, and $chexp_{it}$ is the challenger's expenditures. The unobserved effect a_i contains characteristics of the incumbent, such as "quality", as well as things about the district that are constant. The incumbent's gender and party are constant over time, so these are subsumed in a_i.

   (a) Difference the model across the two years, and estimate the FD model using OLS. Which variables are significant at a 5% level?

   (b) Test for the joint significance of β_2 and β_3, and report the p-value.

(c) Re-estimate the FD model after dropping the ln($inexp$) and ln($chexp$) variables. Interpret the estimated slope on $incshr$. For example, if the incumbent's share of spending increases by 10 percentage points, how is this predicted to affect the incumbent's share of the vote?

(d) If you thought that error disturbances might be correlated for elections within the same state (see the *state* variable), how would you correct the standard errors from part (c)? Re-run the regression using your suggested correction. Does it make a difference?

(e) Re-do part (c), but now use only the observations in which the challenger is the same in 1988 and 1990. (These observations are indicated by $rptchall = 1$.) Note that this regression allows us to also control for unobserved (fixed) characteristics of the challenger within a_i. Do your results change at all?

5. Use the data **married_bmi_sample.dta**, which has data on 7,055 married couples from the Community Tracking Study. There are 14,110 total observations, with household identifer *hhid* and the spouse identity indicated by the *male* indicator. (So **xtset hhid male** will set up the panel dataset appropriately.)

   (a) Run pooled OLS with robust (but not clustered) standard errors. Use *bmi* as the outcome variable, and use *male*, *educ*, *age*, *agesq*, *smoke*, $logfaminc$, and *withkid* as RHS variables.

   (b) Determine whether or not the residuals for husband and wife are correlated. To do this, form estimated residuals in the usual way. Then, use the **gen** command to generate a variable containing the spouse's residual (for example, if you have estimated residuals in **uhat**, you would do **gen uhat_wife = uhat[_n-1] if hhid==hhid[_n-1]**). Find the correlation for between spouse's residuals. Is it statistically significant? Is this what you expected?

   (c) Re-run the pooled OLS regression from part (a) using appropriately clustered standard errors. Interpret the $logfaminc$ slope estimate. Interpret the *withkid* slope estimate.

   (d) Now set up a fixed-effects model using the same RHS variables but which now has a "family effect" a_i. Estimate this fixed-effects model. Which parameters are you no longer able to estimate? How do your results compare to what was found in part (c)?

   (e) Now consider the binary outcome *obese* (defines as 1 if $bmi \geq 30$ and 0 if $bmi < 30$) and the following fixed-effects version of the LPM

   $$obese_{it} = x_{it}\beta + a_i + u_{it},$$

   with a_i being the family effect. Estimate this model with **xtreg** and the appropriate options.

      i. Why do you think the *educ* estimate becomes statistically insignificant whereas the *smoke* estimate remains statistically significant? (Hint: Think about the variation being used to identify the parameters in the FE regression, which here is equivalent to the FD regression.)

ii. What is the interpretation of the *smoke* slope estimate? In thinking about the effect of smoking on obesity, do you prefer this estimate or the one from part (a)?

iii. Explain why you can not consistently estimate the obesity probability for a specific individual based upon the fixed-effects LPM.

iv. Explain why you can consistently estimate the difference between husband's obesity probability and his wife's obesity probability.