Problem Set 3
Econometrics I

Vikrant Vig

Due: January 14th, 2017

1. (Fixed Effects Estimator - 20 pts) For $T = 2$ consider the standard unobserved effects model

$$y_{it} = x'_{it}\beta + \delta_i + u_{it} \quad t = 1, 2$$

Let $\hat{\beta}_W$ and $\hat{\beta}_{FD}$ denote the within and first difference estimators, respectively. In other words:

$$\ddot{y}_{it} = \ddot{x}_{it}\hat{\beta}_W + \ddot{u}_{it}$$

where $\ddot{y}_i = y_{it} - \bar{y}_i$, the same for all the other variables. Note that this is the transformed version of the main regression, where variables are demeaned. $\hat{\beta}_W$ is the OLS coefficient of this regression. Also:

$$\Delta y_i = \Delta x_i\hat{\beta}_{FD} + \Delta u_i$$

where $\Delta y_i = y_{it} - y_{it-1}$, the same for the other variables. Note that this is the transformed version of the main regression, in which variables are time-differenced. $\hat{\beta}_{FD}$ is the OLS coefficient of this regression

   (a) Show that the within and first differences estimators are numerically equal

   (b) Show that the properly computed error variance estimates from the FE and FD methods are numerically identical.

2. (Measurement Error Bias - 15 pts) Suppose we observe a noisy version (x_i) of the true explanatory variable (x_i^*). The true model we want to estimate is

$$y_i = \beta_0 + \beta_1 x_i^* + e_i \tag{0.1}$$

but we only observe x_i, and not x_i^*. We know, however, that $x_i = x_i^* + \nu_i$, where ν_i is an iid error term with mean zero, variance σ_ν^2) and independent of x_i^*, i.e. $E(x_i^*\nu_i) = 0$. You estimate the regression of y_i on x_i and obtain the coefficient $\tilde{\beta}_1$ for the slope variable. Show that this coefficient is biased, i.e.: $\tilde{\beta}_1 = \left(\frac{\sigma^2_{x^*}}{\sigma^2_{x^*}+\sigma^2_\nu}\right)\beta_1$. Discuss the direction of this bias.

Hint: use the fact that the OLS estimator for the slope of a simple regression can be calculated as $\tilde{\beta}_1 = \frac{Cov(x_i,y_i)}{Var(x_i)}$

3. (15 pts) You have two regressors x_1 and x_2, and estimate a regression with all quadratic terms.

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i}^2 + \beta_4 x_{2i}^2 + \beta_5 x_{1i} x_{2i} + e_i \tag{0.2}$$

One of your advisor asks: Can we exclude the variable x_2 from this regression? How do you translate this question into a statistical test? When answering these questions, be specific, not general.

(a) What is the relevant null and alternative hypotheses?

(b) What is an appropriate test statistic and its distribution?

(c) What is the rule for acceptance/rejection of the null hypothesis?

4. (STATA only - 20 pts) This data set is from Schaller (1990). It contains investment (inv) and q variables from 164 firms over 35 years that Schaller uses in his paper. Include a constant term in the following regressions.

(a) Run a regression of inv on q. Discuss the results (please refer to the paper)

(b) Create dummy variables corresponding to each firm using the command "xi". Now using 'regress', run a regression of inv on q and the firm dummies. How has the coefficient on q changed?

(c) Now regress inv on the firm dummies alone and obtain the residual (call it e_inv) from the regression. Similarly, regress q on the firm dummies and obtain the residual (call it e_q) from this regression. finally, regress e_inv on e_q. Is the coefficient estimate on e_q the same as the coefficient estimate on q in part (b)? Why? What is the constant term estimate in this case? Why?

(d) How could you have used the standard error estimate on the q coefficient from part (b) to obtain the standard error estimate on the e_q coefficient in part (c) without running the regressions in part (c)?

(e) Regress inv on e_q. Is the coefficient estimate on e_q the same as in part (c)? Why?

5. (Computational - Panel Data - 30 pts) For this exercise, use the dataset PS3.xls. There are three variables (w: earnings, ed: education, and a:age) for three different years (0: 1990, 1:1991, and 2:1992). Note that this data-set is not in a panel format as in the question above. Either edit it in excel to transform it in a long panel, or use the function **reshape** in stata. You want to have a data with the following variables only: individual, year, w, ed, a. After that create variables corresponding to the individual means for log(wage), educ, exp, exp^2.

(a) For the 1990 portion of the data, regress log(wage) on constant, educ, exp, exp^2 (corresponding to the regression in the Problem Set 1). Now do the same OLS regression using all the data (1990-92), which is called the "pooled least squares" regression. Have the coefficients changed much?

(b) Discuss the validity of the assumption of homoskedastic errors in the regression above.

(c) How would you test if the coefficients estimates changed over the years using the pooled OLS from the regression in (a)?

(d) (Matlab Only) Now estimate a Fixed Effects Model as seen in the lecture notes. Create the matrix M_D and use it to obtain $\hat{\beta}_{FE}$. Should you include an intercept parameter here?

(e) Estimate the within estimator ($\hat{\beta}_W$) for the wage equation. How is $\hat{\beta}_W$ related to $\hat{\beta}_{FE}$ estimated above? *HINT: In this exercise, you need to demean the variables by each individual, i.e. subtract the time average of each variable by group. In Matlab see the function* ***grpstats(x,GROUP)*** *that creates a mean of x by group GROUP (****means_x****). To demean the variable x use* ***x−kron(means_x, ones(T,1))****, where T in this case equal 3 (years). In STATA see function* ***egen*** *combined with* ***by GROUP***

The following questions are **OPTIONAL**. 10 extra marks are going to be given if you do it right.

(f) (Matlab only) Obtain s_W^2 from the last regression. Also obtain the between coefficient estimates and s_B^2. Use both s_W^2 and s_B^2 to form $\hat{\theta}$, the weight needed for FGLS. Generate the transformed variables and perform an OLS regression of them (also known as the FGLS estimator).
Hint: The between estimator is calculated from a regression of the average by group of y ($\overline{y}_i$) on the average by group of x ($\overline{x}_i$) AND a constant. See the note on Random Effects for a formula to calculate s_W^2, s_B^2, $\hat{\theta}$ and the adjusted (y, x) variables, i.e. $\hat{\tilde{y}}_{it}$ and $\hat{\tilde{x}}_{it}$. Finally, note that you need to adjust the constant, as well.

(g) (Stata only) Check if the results match with Stata by using the command **xtgls** on the variables from the original wage equation.

(h) Are the FGLS results different than the FE ones? Perform a Hausman test. What do you conclude?