

ECON7360 Causal Inference for Microeconometrics

Homework 2

Rigissa Megalokonomou

October 12, 2020

Instruction

Use STATA to conduct all empirical analysis. When you are asked to explain or discuss something, your response should be concise (no more than four sentences). To facilitate the grading work, please clearly label all your answers. You should upload your work (a PDF or MS Word report collecting all your answers) via the “Turnitin” submission link (in the “Homework 2 Submission” folder under “Assessment”) by **11:59 am on the new updated due date November 1, 2020**. It is not necessary to submit your STATA script (if you want, please include your code in an “Appendix” attached to the end of your report). You are encouraged to work on this problem set in groups; that is, you can discuss how to answer these questions with your group members. However, **this is not a group assignment, which means that you must answer all the questions in your own words and submit your report separately**. The marking system will check the similarity, and UQ’s student integrity and misconduct policies on plagiarism apply.

Question: POLS, Instrumental Variable and Panel Data Regressions

Consider the following linear panel data model

$$y_{it} = \beta_0 + \beta_1 x_{it,1} + \beta_2 x_{it,2} + \beta_3 x_{it,3} + \beta_4 x_{it,4} + u_{it}, \quad (1)$$

where $x_i = (x_{it,1}, x_{it,2}, x_{it,3}, x_{it,4})$ are explanatory variables, $\beta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)$ are unknown parameters of interest, and u_{it} is unobservable error. Here, β_1 is of our primary interest, and we include $(x_{it,2}, x_{it,3}, x_{it,4})$ as control variables. As usual, $i = 1, \dots, N$ refers to individuals (id, cross-sectional units) and $t = 1, \dots, T$ refers to time periods. Use the data file **SIMUDATA.dta** to answer the following questions. Unless otherwise specified, use 5% as the significance level for all the tests below.

- (6 points) Declare the data to be a panel via specifying the individual identifier (**id**) and time identifier (**t**). Which regressor(s) are *not* time-varying?¹ What are N and T ? Is the dataset of long form or wide form?
- (4 points) Use (pooled) OLS to estimate model (1), compute heteroskedasticity-robust SE, and report estimation results.²

¹Hint: You can use the **egen** command along with the **by** option to compute the standard deviation of each regressor for each i . Which regressor(s) have zero variation over time?

²Hint: Estimation results can be summarized either in equation form or by one or more regression tables. There are numerous illustrating examples for these two forms in Wooldridge’s (2013) textbook. For a standard equation form, see, e.g., Example 4.2. Section 4.6 provides guidelines for reporting estimation results in table(s). Instead of typing tables in Word/Excel manually, you can use the “**estout**” package to conveniently create nice-looking regression tables in STATA.

- (c) (10 points) It is well-known that the standard errors (SE) of panel data estimation need to be adjusted to control for likely correlation of the error u_{it} over time for given i (clustering on i), i.e., $C(u_{it}, u_{is}|x_i) \neq 0$ for $t \neq s$. Re-estimate model (1) using OLS and calculate cluster-robust SE. Compare the estimation results with those obtained in Part (b). Comment on your findings. If $C(u_{it}, u_{is}|x_i) \neq 0$ is true, is OLS still BLUE?
- (d) (6 points) One of your friends argues that the OLS estimator may suffer from omitted variable bias (OVB) as the error u_{it} may contain factor(s) correlated with $x_{it,1}$. If this were true, which assumption of linear regression would not be valid, and what could be wrong with using OLS?
- (e) (6 points) This friend suggests that you should use TSLS rather than OLS. In particular, he proposes two instrumental variables (IV), $z_{it,1}$ and $z_{it,2}$, for $x_{it,1}$. What conditions must hold for $z_{it,1}$ and $z_{it,2}$ to be valid IV?
- (f) (10 points) Estimate model (1) using TSLS with $z_{it,1}$ and $z_{it,2}$ as IV and report estimation results. As in Part (c), you should compute and report cluster-robust SE. Compare the TSLS estimates with the OLS estimates obtained in Part (c), and comment on your findings. Assuming both $z_{it,1}$ and $z_{it,2}$ are valid IV, do you think $x_{it,1}$ is an endogenous regressor? Explain.³
- (g) (10 points) Write the expression for the first-stage regression of the TSLS. How to assess the strength of $(z_{it,1}, z_{it,2})$ as IV? Do you think the TSLS estimation in Part (f) has weak IV problem?
- (h) (16 points) To capture potential omitted time effects, consider the following model

$$y_{it} = \beta_0 + \beta_1 x_{it,1} + \beta_2 x_{it,2} + \beta_3 x_{it,3} + \beta_4 x_{it,4} + \sum_{s=2}^T \gamma_t d_{s,t} + v_{it}, \quad (2)$$

where $d_{s,t}$ are time dummies ($d_{s,t} = 1$ if $s = t$, and 0 otherwise). Note that the sample includes data from $t = 1$ to $t = T$, but model (2) includes only dummies for $t = 2$ to $t = T$. Why? Estimate model (2) using OLS and TSLS (again with $z_{it,1}$ and $z_{it,2}$ as IV) and test if time effects are significant, i.e., at least one γ_t are not zero. With time effects controlled, do you think $x_{it,1}$ is still an endogenous regressor?⁴

- (i) (10 points) Suppose $v_{it} = \alpha_i + e_{it}$ and $Cov(e_{it}, x_{it,k}) = 0$ holds for all $k = 1, \dots, 4$. Re-write model (2) as

$$y_{it} = \beta_0 + \beta_1 x_{it,1} + \beta_2 x_{it,2} + \beta_3 x_{it,3} + \beta_4 x_{it,4} + \sum_{s=2}^T \gamma_t d_{s,t} + \alpha_i + e_{it}. \quad (3)$$

Treat α_i as random effects (RE). Use an RE estimator to estimate model (3) and report estimation results. To account for possible serial correlation in e_{it} , compute the cluster-robust SE. Compare the RE estimates with the TSLS estimates obtained in Part (h). Comment on your findings.

- (j) (12 points) Now treat α_i in model (3) as fixed effects (FE). Use an FE estimator to estimate model (3) and report estimation results. Justify the fact that the FE estimator cannot estimate all slope coefficients. Compare the FE estimates with the TSLS estimates obtained in Part (h).⁵ Comment on your findings.

³Hint: Compare OLS and TSLS estimates for model (1).

⁴Hint: Again, compare the OLS and TSLS estimates.

⁵Hint: In addition to comparing estimated coefficient, take a look at the SE and use your findings in Part (g).

- (k) (10 points) Perform the Hausman test to compare FE and RE estimators. What is the null hypothesis of the Hausman test. Which model do you think is more appropriate, FE or RE? Note that to implement the Hausman test with heteroskedasticity- or cluster-robust SE, you need to use the user-contributed command `xtoverid` after an RE estimation. To install `xtoverid` to your STATA, type “`ssc install xtoverid`”.

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder