
1. (25 points) Suppose you work for the department of agriculture and have been charged with the task of determining the effect of irrigation on corn yields. You have cross-sectional data on 1000 different farms containing crop yield in bushels per acre (Y), and a dummy variable indicating whether the land is irrigated or not (IRR). Keep in mind that farmers decide whether to irrigate based upon the profitability of such an investment. If a plot of land generally receives a lot of rain, irrigation is not likely to be profitable.

- a. Suppose you estimate a simple OLS model of $Y = a + b \cdot IRR$. Is the estimated effect of IRR likely to be biased upward or downward? Why?
- b. Explain how you could eliminate the bias in the above model using cross-sectional data. Describe any additional variables that would be useful in your estimation process, and outline the steps you would take to properly estimate the effect of irrigation on crop yields.
- c. Describe how you could calculate a test statistic for the hypothesis that the OLS estimates are biased/
- d. If you had panel data on crop yields (i.e. multiple years of data on the 1000 farms), what approach could you use to eliminate the bias in the OLS estimates? Describe the process and explain how it eliminates the source of the bias in the OLS estimates.

2. (30 points) In a recent article¹, the effectiveness of seat belts and airbags in saving lives was measured. Quoting from the study, "Measurement of seat belt and air bag effectiveness is complicated by the fact that systematic data are collected only for crashes in which a fatality occurs. These data suffer from sample selection since seat belt and air bag usage influences survival rates which in turn determine whether a crash is included in the sample."

In this paper, the dependent variable of interest is a dummy variable indicating whether a person dies in an automobile accident. The relevant control variable is a dummy indicating whether a seat belt is used. That is,

$$F_{ij} = b_0 + b_1 \cdot SB_{ij} + b_2 \cdot Carsize_{ij} + e_{ij}$$

where $F_{ij}=1$ if person i is fatally injured in accident j , SB_{ij} is a dummy that equals one if the person is wearing a seatbelt, $carsize_{ij}$ is the size of the car (measured in tons) that person i is an occupant in for accident j .

F_{ij} is observed only if at least one person dies in accident j .

Presumably the error term e would pick up unobservables that would be important to the fatality, such as the speed of the car(s) involved in the accident.

- a. Is there likely to be positive or negative sample selection bias in the example? Explain.
- b. Is the sample selection likely to cause the estimated effect of seatbelt use to be over- or under-estimated? Why?
- c. Assume that heavier cars are safer, is the sample selection likely to cause the estimated effect of carsize to be over- or under-estimated? Why?

¹ Steve Levitt and Jack Porter. "Sample Selection in the Estimation of Air Bag and Seat Belt Effectiveness." NBER Website. Friday, March 25, 2005. <http://www.nber.org/papers/w7210>.

As noted above, an accident shows up in the sample only if at least one person is killed. Thus, for example, if there is a 3 car accident, there is information on seatbelt use and fatality on all three cars only if at least one person is killed in one of the 3 cars. The authors use the multi-car accidents to get around the sample selection problem. For any given multi-car accident with a fatality, they estimate the above regression for the other cars involved in the accident. (e.g. if there is a three-car accident and a person is killed in car A, the data for car B and car C is used. If there is a three car-accident and a person is killed in both car A and car B, the data for car B and C is used once, and the data for car A and C is used once).

d. How does the “multi-car” accident approach fix the sample selection problem? Hint: Remember that the sample selection bias is the result of the sample selection problem causing unobservables (like car speed) for the observed sample to be correlated with the explanatory variable (e.g. seat belt use). How does the multi-car approach eliminate this correlation?

```

. xtreg empstock part_dir assets yeardum2-yeardum10, re
Random-effects GLS regression           Number of obs   = 333506
Group variable (i): einpn              Number of groups = 77839
Random effects u_i ~ Gaussian          Wald chi2(11)   = 1483.24
corr(u_i, X) = 0 (assumed)             Prob > chi2     = 0.0000

empstock   Coef.           Std. Err.      z    P>z   [95% Conf. Interval]
part_direct  -.0171184      .0004972     -34.43  0.000   -.0180929   -.016144
assets       2.35e-11      1.62e-12     14.51  0.000   2.04e-11    2.67e-11
_cons       .0434864              .000835     52.08  0.000   .0418498    .045123

sigma_u .16656507
sigma_e .07025808
rho     .84895412      (fraction of variance due to u_i)

```

- a. Based on the information above, explain which of the three models is most appropriate for examining the effect of participant direction on employer stock holdings. Be sure to provide the basis for your answer (i.e. what test statistics or other information in the above results support this conclusion?)

- b. Compared to the OLS and random effects estimates, the fixed effect model finds a much smaller effect of participant direction on the share of assets invested in employer stock. What could explain this result? Provide an econometric explanation that is linked to investment behavior.

- c. Suppose that you want to estimate the effect of unionism on the share of assets invested in employer stock. Explain whether it would be possible to estimate the effect of unionism in the
 - i. fixed-effects model?
 - ii. random effects model?

4. (20 points) Several recent articles have investigated the importance of “peer effects” on student performance. Many educators have argued that better peers lead to increased student performance. A standard way of analyzing this issue is to regress some measure of a student’s performance on measures of the performance of their peers. For example,

$$S_i = a_0 + a_1 \text{Peerscore}_i + X_i b + e_i$$

Where S_i is student i ’s score on some standard test, Peerscore_i represents the average score of student i ’s peers (e.g. classmates) on the test, and X_i is a vector of other controls (e.g. family income, parental education, etc.).

One study investigated the issue of peer effects on subsequent performance in medical school.³ Quoting from the abstract “We estimate positive peer effects that disappear when school-specific fixed effects are added to control for ... peer group.” For example, one regression was specified as:

$$\text{Boardscore}_i = b_0 + b_1 * \text{Admission_score}_i + b_2 * (\text{peer_admission_score}_i) + e_{it}$$

where Boardscore_i is student i ’s score on the board exam (given at the end of medical school), Admission_score_i is student i ’s own test score at admission, $\text{peer_admission_score}_i$ is the average score of student i ’s admission scores. The coefficient b_2 was significant and positive when OLS was used to estimate the model. After “school fixed effects” are included (i.e. controls for the medical student that the student attends), b_2 becomes smaller and statistically insignificant.

a. Describe how the fixed effects model could be used using a “deviations from means” approach. Define your notation carefully!

b. Recall that that the OLS model controls for the individual’s own performance on the admission test. Why does the OLS model find a positive effect of peer admission scores on own performance whereas the fixed effects model does not? Provide both an econometric explanation and the type of behavior on the part of **medical school admissions personnel** that could lead to this result.

³ Peter Arcidiacono and Sean Nicholson. "Peer Effects in Medical School." NBER Website. Saturday, March 26, 2005. <<http://www.nber.org/papers/w9025>>.