Midterm Exam, Econ 613, Spring 2025

1 Return to Schooling (25 points)

A conceptual framework of the return to schooling can be described as the following

$$S_i = X_{1i}'\pi_x + Z_i\pi_z + \xi_i \tag{1}$$

$$Y_i = X_i'\alpha + \rho S_i + \eta_i \tag{2}$$

where S_i is the number of years schooling for individual i. It depends on a subset of X_i (which we denote as X_{1i}), and also importantly on an instrument z_i. Y_i is an individual's earnings - we are interested in the return to schooling parameter ρ. X_i is individual-specific observables, including experience, race, and regions of living.

1. How do we interpret the error terms ξ_i and η_i in both equations? What are the potential unobservables that are included in these error terms?
2. Suppose that X_i is plausibly exogenous concerning η_i, what is the assumption we need to have the OLS estimate of ρ consistent? Why this assumption is likely not to hold?
3. Schooling is a hard variable to measure in general. One will have to adjust for things like how many classes each student takes within a given year and potentially there are also significant quality differences in various schools. Suppose as an applied researcher, we only observe a noisy measure of true schooling

   $$S_i^m = S_i + e_i$$

   where e_i is the classical measurement error. Please show that if S_i is uncorrelated with η_i, we will have the OLS coefficient estimate of ρ biased towards zero.
4. Card (1993) proposed to use the distance to the nearest 4-year college as an instrument.In data we have variables $X_i = \{expr, race, region\}$, $z_i = \{dist\}$, $y_i = \{wage\}$, please write

down the STATA command that implements OLS and 2SLS regressions of wage equation. Explain what a robust standard error means and how to implement that in STATA.

5. What confounding factors might invalidate Card's empirical strategy? Can you suggest any remedies?

2 Return to Schooling Redux (20 points)

Ashenfelter and Krueger (1994) use another creative empirical strategy to estimate the return to schooling. They look at a sample of twins whose wages can be defined as

$$Y_{i1} = X_i^{'}\alpha_x + Z_{i1}^{'}\alpha_z + \rho S_{i1} + \omega_i + \eta_{i1} \tag{3}$$

$$Y_{i2} = X_i^{'}\alpha_x + Z_{i2}^{'}\alpha_z + \rho S_{i2} + \omega_i + \eta_{i2} \tag{4}$$

where 1 and 2 indicate the twin individual 1 and 2 of the same household i.

1. Explain again what could be in the household unobservable ω_i and whether it could be a confounder in the identification of return to schooling ρ.

2. Ashenfelter and Krueger (1994) proposed to use a "within household" differencing estimator, could you please define that estimator using the notation we had in the equations above? Explain what set of parameters we can/can't estimate.

3. If the schooling variables are again measured with errors, would the "within household" differencing work to estimate ρ? Explain in a simple formula and contrast that with the case in Card (1994) (where we use the level equation (2)).

4. In Ashenfelter and Krueger's data, there are self-reported education level S_{i1}^1, S_{i2}^2 as well as sibling reported education level of their twin brother/sister S_{i1}^2 and S_{i2}^1. They use a very creative IV strategy to overcome the measurement error problem. Could you describe it?

3 System OLS and FGLS (20 points)

Christensen and Greene (1978) investigate the returns to scale in the U.S. electric power generator industry. Their baseline model is a Translog cost function

$$\begin{aligned} lnC = \alpha_0 + \alpha_y lnY + 0.5\gamma_{yy}(lnY)^2 + \sum_i \alpha_i lnP_i \\ +0.5\sum_i\sum_j \gamma_{ij} lnP_i lnP_j + \sum_i \gamma_{yi} lnY lnP_i + u \end{aligned} \tag{5}$$

where Y is the total amount of electricity generated, and P_i is input prices of $i = L, K, F$ (labor, capital, and fuel) respectively. For the cost function to satisfy the Homogeneity of Degree one, we impose that $\sum_i \alpha_i = 1$, $\sum_i \gamma_{Yi} = 0$, and $\sum_i \gamma_{ij} = \sum_j \gamma_{ji} = \sum_i \sum_j \gamma_{ij} = 0$, as well as further symmetry $\gamma_{ij} = \gamma_{ji}$. We are interested in estimating the parameters of this cost function to evaluate returns to scale at the plant level.

1. What will your concern of estimating equation (3) using OLS? Why this concern is not likely to be a problem for the U.S. electric power generator industry during the sample period?

2. Use Shephard's Lemma to show that

$$\frac{P_i X_i}{C} \equiv S_i = \alpha_i + \gamma_{Yi} lnY + \sum_j \gamma_{ij} lnP_j \tag{6}$$

where S_i is the cost expenditure share of input i.

3. Use the notations y_k and X_k, please briefly describe the procedure of implementing FGLS. (Hint: there will be two steps, and you will need to write down the estimator in terms of y_k, X_k, and $\hat{\Omega}$)

4. Define the Scale Economies (SCE) using the model we outlined above. What were the key findings of the paper in terms of SCE?

4 Linear GMM (20 points)

Hausman and Zona (1994) proposed to estimate lower-level demand for different brands of beer i, at city n and year t by the following demand function:

$$s_{int} = \alpha_{in} + \beta_i log(y_{Gnt}/P_{nt}) + \sum_{j=1}^{J} \gamma_{ij} logP_{jnt} + \epsilon_{int} \tag{7}$$

where y_{Gnt} is total category expenditure at city-year n, t, and P_{nt} is category specific index.

$$logP_{nt} = \sum_j \alpha_{jn} logP_{jnt} + 0.5 \sum_i \sum_j \gamma_{ij} logP_{int} logP_{jnt} \tag{8}$$

s_{int} is the expenditure share of brand i at city n and year t. The restrictions are $\sum_i \alpha_{in} = 1$, $\sum_i \beta_i = 0$, and $\sum_i \gamma_{ij} = \sum_j \gamma_{ij} = 0, \gamma_{ij} = \gamma_{ji}$

1. This demand system is proposed by Deaton and Muellbauer with the name of Almost Ideal Demand System. The starting point is an expenditure function

   $$logc(u, p_n) = \alpha_0 + \sum_i \alpha_i log(p_{in}) + 0.5 \sum_i \sum_j \gamma^*_{ij} log(p_{in}) log(p_{jn}) + u\beta_0 \Pi_i p_{in}^{\beta_i} \tag{9}$$

   Please show how $logP_n$ is derived from this demand system.

2. We are generally concerned about the endogeneity of own price $logP_{int}$ since local demand shock ϵ_{int} could be a result of brand-market specific advertising or marketing effort, that is partially reflected in the price. Hausman and Zona proposed IV strategy that utilizes a simple pricing model

   $$logP_{jnt} = \omega_{jn} + \delta_j logc_{jt} + \omega_{jnt} \tag{10}$$

   For the same brand j, there is a brand-year specific distribution cost c_{jt}. ω_{jnt} is brand-market-year specific cost shock. What are the potential IVs for $logP_{int}$ based on this model? What are the crucial assumptions on $\omega_{jnt}, n = 1, ..., N$ to validate these IVs?

3. We implement a linear GMM estimation brand by brand $i = 1, ..., J$, let $z_{int} = \{IVP_{int}; logP_{jnt}, j \neq i; log(y_{Gnt}/P_{nt})\}$, please write down the STATA command that implements the moment condition $E[z_{int}\epsilon_{int}] = 0$ for all brand i.

4. Once you estimate the model parameters, please explain how to calculate the demand elasticities $\frac{dlnq_{int}}{dlnp_{int}}$ and $\frac{dlnq_{int}}{dlnp_{jnt}}, j \neq i$.

5 Linear Panel Data Model (15 points)

Consider the linear panel data model

$$y_{it} = x_{it}\beta + v_{it}$$

$$v_{it} = c_i + u_{it}$$

1. Please define the “Random Effect” and “Fixed Effect” estimator of this model. Please contrast the assumptions needed for either of them to be Consistent. (Hint: Denote $\Omega = E(v_i v_i'), v_i = [v_{i1}, v_{i2}, ..., v_{iT}]$)

2. Further define the “First Difference” estimator. Can you come up with the *weakest* assumption possible for this estimator to be Consistent? Using the FD estimator, will you be able to test whether the u_{it} is serially correlated overtime, how?

3. Please define the well-known ”Hausman Test” based on the “Random Effect” vs “Fixed Effect” estimators. Which assumption we will hold true and which assumption are we testing for?