PS1 - Econometrics 1.2

Bruno Neves and Matheus Pimentel

2025-11-30

Problem 1)

a)

$$\mathbb{E}[Y_1|X,Z] = \beta X + \mathbb{E}[\Lambda|X,Z] + \mathbb{E}[\eta_1|X,Z] \implies \mathbb{E}[Y_1|X,Z] = \beta X + \gamma_0 + \gamma_1(X+Z) + \alpha$$

$$\mathbb{E}[\mathbb{E}[Y_1|X,Z]|X] = \mathbb{E}[Y_1|X] = \beta X + \gamma_0 + \gamma_1(X + \mathbb{E}[Z|X]) + \alpha = (\beta+\gamma_1)X + (\gamma_0+\alpha) = aX + b,$$

calling $a = (\beta + \gamma_1)$ and $b = (\gamma_0 + \alpha)$. Then, as many pairs of (β, γ_1) could give the same $\mathbb{E}[Y_1|X]$, it is not possible to identify β.

An exclusion restriction that we could impose is that the conditional mean of Λ does not depend on X, making $\mathbb{E}[\Lambda|X,Z] = \gamma_0 + \gamma_1 Z$:

$$\mathbb{E}[Y_1|X] = \beta X + (\gamma_0 + \alpha).$$

b)

$$\mathbb{E}[Y_1 - Y_2|X,Z] = \beta(X-Z) + \mathbb{E}[\eta_1 - \eta_2|X,Z] \implies \mathbb{E}[Y_1 - Y_2|X,Z] = \beta(X-Z).$$

Thus, to identify β we can do like in OLS:

$$\mathbb{E}[(Y_1-Y_2)(X-Z)] = \mathbb{E}[\mathbb{E}[(Y_1-Y_2)(X-Z)|X,Z]]$$

$$\mathbb{E}[(Y_1-Y_2)(X-Z)] = \beta\mathbb{E}[(X-Z)^2]$$

$$\beta = \frac{\mathbb{E}[(Y_1-Y_2)(X-Z)]}{\mathbb{E}[(X-Z)^2]}.$$

Problem 2

a)

$$\mathbb{E}[Y|Z] = \alpha + \beta(\mathbb{E}[X-\eta|z]) + \mathbb{E}[\varepsilon|Z] \implies \mathbb{E}[Y|Z] = \alpha + \beta\mathbb{E}[X|z] + \mathbb{E}[\varepsilon|Z].$$

For us to identify β we need to impose that $\mathbb{E}[\varepsilon|Z]$ do not vary with z. Therefore,

$$\mathbb{E}[\varepsilon|Z=z] = e, \forall z.$$

We did this because with e being a constant we can make $c = \alpha + e$ and identify β directly.

b)

$$Cov(Z,Y) = Cov(Z, \alpha + \beta X^* + \varepsilon) = \beta Cov(Z, X-\eta) + Cov(Z,\varepsilon),$$

and given the statement and what we imposed, respectively, $\mathbb{E}[\eta|Z] = 0 = \mathbb{E}[\varepsilon|Z]$,

$$Cov(Z,Y) = \beta Cov(Z,X) \implies \beta = \frac{Cov(Z,Y)}{Cov(Z,X)}$$

Changes in Z shift X^*, and, therefore, in X, making Y shift proportionally to some β. As η, α and ε are constant to variations of Z, we have that, in the end, the part of Y and X that co-moves with Z reflects only the causal channel $X^* \to Y$.

Problem 3

1)

The paper works with a two–equation linear model for schooling and log wages:

$$S_i = X_i\gamma + v_i$$

$$y_i = X_i\alpha + S_i\beta + u_i.$$

where S_i is years of schooling, y_i log wages, X_i observed covariates and β is the structural "return to schooling". The admissible structures are defined by linearity, mean independence of errors from observables, and an exclusion restriction: college proximity can enter the schooling equation but is excluded from the wage equation in the baseline specification. Then, under these restrictions the model delivers a constant (or average) marginal return β to be identified.

2)

The key structural feature is that schooling is an endogenous choice, so S_i is correlated with the wage disturbance u_i through unobserved ability, measurement error in schooling, and heterogeneous returns. These channels imply that OLS estimates of β are biased. Another central feature is the presence of a cost shifter, the geographic proximity to a four–year college, that affects the schooling decision but is assumed not to shift the wage equation directly. Heterogeneity by parental education (college proximity matters more for low–background youths) is also part of the structure and is later exploited for over–identification.

3)

Identification of β relies on an instrumental–variables strategy using college proximity as an excluded instrument for schooling. The main assumptions are:

1. Relevance: living near a college shifts schooling choices;
2. Exogeneity: proximity affects wages only through schooling (no correlation with unobserved ability, school quality, or permanent wage premia);
3. IV-based identification strategy: any direct wage effect of proximity does not vary with family background, so interactions of proximity with low–background indicators can be used as additional instruments.

4)

The model is taken to the data using the NLSYM cohort: y_i is log hourly wages from 1976 and 1978, S_i is completed years of schooling, X_i includes experience, race, region and family–background controls, and the instrument Z_i is an indicator for a nearby accredited four–year college. In some specifications, its also interactions with parental–education dummies. First–stage regressions of S_i on Z_i and X_i recover the effect of proximity on schooling. Reduced–form regressions of y_i on Z_i and X_i recover the effect on wages. Under the IV assumptions, the structural parameter is identified as

$$\beta^{IV} = \frac{Cov(y_i, Z_i)}{Cov(S_i, Z_i)},$$

that is, the ratio of reduced–form effects of proximity on wages and schooling, or equivalently the Wald difference in mean wages divided by the difference in mean schooling between "near–college" and "far–from–college" groups.