Network and Spatial Analyses

19. February 2020

Lecture:

Topics in Economic Geography II.

- Economic growth and development
- FDI/MNEs
- Why networks?

Seminar:

Econometrics reminder

Course Github page: https://github.com/bokae/anet_course

Gauss-Markov assumptions

I. The population process should be linear

The dependent variable Y is assumed to be a linear function of the explanatory variables X in the model.

$$y_i = \alpha + \beta_1 x_1 + u_i$$

II. Set of sample data is a random sample

Every observation in the population is equally likely to be picked. Implicitly means that all of our data points come from the same sample.

$$\{X, Y\} \longrightarrow \{x_i, y_i\}$$

$$P\{x_i, y_i\} = P\{x_j, y_j\}$$

Gauss-Markov assumptions

III. Zero conditional mean of errors

If I know the value of X that does not help to know whether they will be over or under the population average regression line!

$$E(u_i|X) = E(u_i|x_1, \dots x_n) = 0$$

This assumption is violated when the variables are stochastic or are endogenous!

$$E(x_j \cdot u_i) = \begin{bmatrix} E(x_{j1} \cdot u_i) \\ E(x_{j2} \cdot u_i) \\ \dots \\ E(x_{j1k} \cdot u_i) \end{bmatrix} = 0$$

Gauss-Markov assumptions

IV. No collinearity

Some explanatory variables are linearly dependent.

$$y_i = \alpha + \beta_1 x_1 + \beta_2 x_2 + u_i$$

$$x_1 \neq \delta_0 + \delta_1 x_2 \qquad corr(x_1, x_2) \neq 0$$

The sample data matrix X should have a full column rank:

$$rank(\mathbf{X}) = k$$

Gauss-Markov assumptions

IV. No collinearity

Some explanatory variables are linearly dependent.

$$y_i = \alpha + \beta_1 x_1 + \beta_2 x_2 + u_i$$

$$x_1 \neq \delta_0 + \delta_1 x_2 \qquad corr(x_1, x_2) \neq 0$$

The sample data matrix X should have a full column rank:

$$rank(\mathbf{X}) = k$$

Perfect collinearity

$$price_i = \alpha + \beta_1 sqm_i + \beta_2 sqf_i + u_i$$

$$sqf = 10.76 sqm$$

Non- perfect multicollinearity

$$price_i = \alpha + \beta_1 sqm_i + \beta_2 rooms_i + u_i$$

Gauss-Markov assumptions

V. Homoscedastic errors

Heteroscedasticity occurs when the error term is correlated with the level of the explanatory variable.

$$var(u_i|x_i) = \sigma^2 \qquad var(u_i) = \sigma^2$$

The error term *u* does not vary systematically with *x*.

$$E(uu^T|X) = Var(u|X) = \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix} = \sigma^2\mathbf{I}$$

Gauss-Markov assumptions

VI. No serial correlation

Autocorrelation occurs when a given observation is more likely to be above the regression line if adjacent observations also are above the fitted regression line

$$cov(u_i, u_j) = 0$$

If I know *i* that does NOT help me to predict the error of *j*.

$$E(uu^T|X) = Var(u|X) = \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix} = \sigma^2 \mathbf{I}$$

Gauss-Markov assumptions

I. The population process should be linear

I. Set of sample data is a random sample (from the same population!)

I. Zero conditional mean of errors

I. No collinearity

I. Homoscedastic errors

I. No serial correlation

Non-normal distribution of the DV - Discrete Choice Models

I. Linear probability model

The biggest advantage of the LP model is its simplicity.

$$P(y=1|x) = \alpha + \beta_1 x_i + e_i$$

Disadvantages:
i) Heteroscedasticity

$$var(e_i|x_i) = E(e_i^2|x_i) = \sum_j P(y_i = y_j)e_i^2$$

$$= P(y_i = 0|x_i)(-\beta x_i)^2 + P(y_i = 1|x_i)(1 - \beta x_i)^2$$

$$= (1 - \beta x_i)\beta x_i(\beta x_i + 1 - \beta x_i)$$

ii) Non-normality of errors

iii) LP model allows predicted probabilities outside of the normal [0,1] range.

Non-normal distribution of the DV - Discrete Choice Models

II. Logit and Probit models

$$P(Uni = 1|pwage) = F(\alpha + \beta|pwage)$$

$$F(-\,\text{inf}) = 0$$
$$F(+\,\text{inf}) = 1$$

Logit Model

$$F(z) = \frac{e^z}{1+e^z} = L(z)$$

$$Z \rightarrow -\infty \qquad F \rightarrow 0$$

$$Z \rightarrow +\infty \qquad F \rightarrow \frac{e^z}{e^z} = 1$$

Probit Model

$$F(z) = \int_{-\infty}^{z} \phi(u)du$$

Integrate of the density function $\Phi(z)$

$$\Phi(-\infty) = 0 \qquad \Phi(+\infty) = 1$$

Non-normal distribution of the DV - Count Data

III. Poisson and Negative Binomial models

Discrete: $P(y|\lambda) = \frac{\lambda^y e^{-\lambda}}{y!}$; $y = 0, 1, 2, 3...$

Count:
-Events occur rate uniformly random both in time and space.
-Events must be independent.

E.g.: y measures the number of birth per hour on a given day in a given hospital.

$$E[y] = \sum_y \frac{\lambda^y e^{-\lambda}}{y!} = \lambda e^{-\lambda} x e^{\lambda} = \lambda$$

$$Var[y] = \lambda$$

If the variance is higher than the mean we should use NB!

$$X \sim NB(N, P) \quad \binom{N}{K} P^K (1-P)^{N-K}$$

- N large
- P small

$$X \sim Po(NP)$$

Non-normal distribution of the DV - Count Data

IV. Zero-Inflation models

$$Y_i = \begin{cases} 0, & \text{with probability} P_i + (1-P_i)e^{-\lambda i} \\ y_i & \text{with probability} (1-P_i)\frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i} \end{cases}$$

If there is an excessive number of zeros we have to deal with the possibility of false zeros!
We need a combination of a probability function and a count process based on a negative binomial distribution.

$$P(Y_i = 0) = P_i + (1-P_i)(1+k\lambda_i)^{-1/k}$$

We need to run two regressions:
Logit (or probit): $P(Y_i = 0) = \gamma_0 + \gamma_1 z_i$

Negative Binomial Regression: $\log(Y_i = y_i|\hat{P}_i) = \alpha + \beta_1 x_1 + e_i$

Non-normal distribution of the DV - Quantile Regression

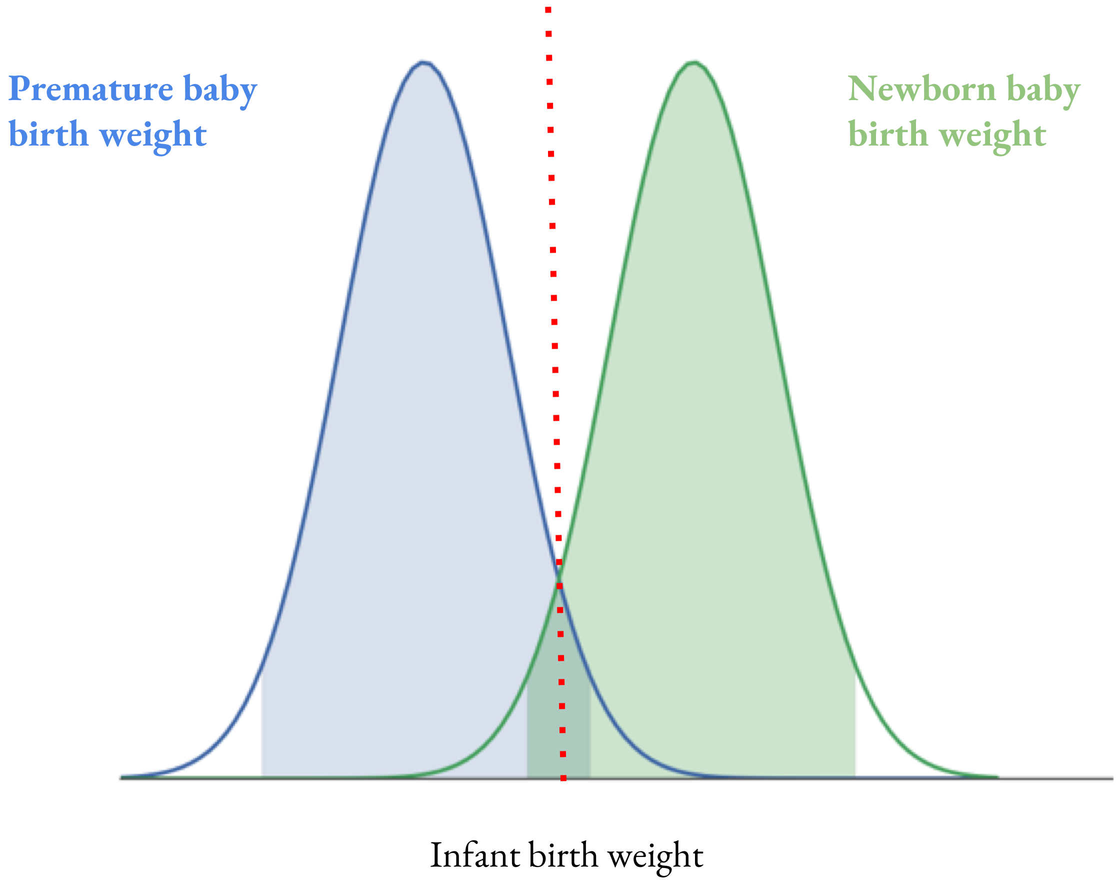
The LS estimation minimizes the sum of the squared errors:

$$\sum_i^N = (y_i - \hat{y}_i)^2$$

while, the quantile regression minimizes a weighted sum of the errors, where τ is the quantile level:

$$\tau \sum_{y_i > \hat{\beta}'_\tau X_i} |y_i - \hat{\beta}'_\tau X_i| + (1 - \tau) \sum_{y_i < \hat{\beta}'_\tau X_i} |y_i - \hat{\beta}'_\tau X_i|$$

Non-normal distribution of the DV - Quantile Regression



Selection bias - sampling problem

$$E[\delta_i] = E[Y_{i(\omega+1)} - Y_{i(\omega)}]$$

Comparing the averages of the two groups does not represent the effect difference of $(\omega + 1) - \omega$

$$\Delta\mu \neq \text{Average Causal Effect}$$

,because there is a selection bias!

In OLS: $E[\delta_i | X_i] \quad (\omega + 1) \Leftarrow X_i \Rightarrow \omega$

Selection bias - sampling problem

$$E[\delta_i] = E[Y_{i(\omega+1)} - Y_{i(\omega)}]$$

Comparing the averages of the two groups does not represent the effect difference of $(\omega + 1) - \omega$

$$\Delta\mu \neq \text{Average Causal Effect}$$

,because there is a selection bias!

$$J_i = \begin{cases} 1 \\ 0 \end{cases} \quad E[\delta_i] = E[x_{i1} - x_{i0}] \neq E[x_i | J_i = 1] - E[x_i | J_i = 0]$$

Solutions:

i) Heckman correction

ii) [Propensity Score] Matching

Endogeneity - Instrumental Variable

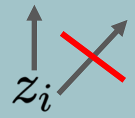
$$y_i = \alpha + \beta x_i + e_i$$

OLS doesn't give the true population parameter, because the chance in *y* consists two effects, the change in *x* and the error:

$$\hat{\beta}_{OLS} = \frac{\Delta y}{\Delta x} = \frac{\Delta y_x + \Delta y_e}{\Delta x} = \beta + \frac{\Delta y_e}{\Delta x}$$

We need a third variable, which not associated with the error term:

$$y_i = \alpha + \beta x_i + e_i$$



$$\begin{aligned} cov(z_i, y_i) &= cov(z_i, \alpha + \beta x_i + e_i) \\ &= cov(z_i, \alpha) + \beta(z_i, x_i) + cov(z_i, e_i) \end{aligned}$$

Endogeneity - Instrumental Variable

$$\begin{aligned} cov(z_i, y_i) &= cov(z_i, \alpha + \beta x_i + e_i) \\ &= cov(z_i, \alpha) + \beta(z_i, x_i) + cov(z_i, e_i) \end{aligned}$$

To derive the IV estimator for a bivariate model: $\hat{\beta}_{IV} = \frac{cov(z_i, y_i)}{cov(z_i, x_i)}$

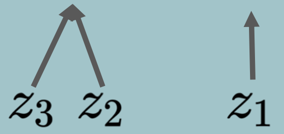
Assumption to be made to have the IV estimator:

i) $cov(z_i, e_i) = 0$

ii) $cov(z_i, x_i) \neq 0$

Endogeneity - 2 Stage-Least Square

$$y = \alpha + \beta_1 x + \beta_2 z_1 + e$$



$$cov(z_2, e) = 0$$

$$cov(z_3, e) = 0$$

First-stage to estimate x with the exogenous variables z:

$$\hat{x} = \hat{\delta}_0 + \hat{\delta}_1 z_i + \hat{\delta}_2 z_2 + \hat{\delta}_3 z_3$$

Second-stage we include the estimated values of x:

$$y = \alpha + \beta_1 \hat{x} + \beta_2 z_1 + e$$

Hence, we get a more efficient estimation on population parameter: $\hat{\beta}_{2SLS} > \hat{\beta}_{IV} > \hat{\beta}_{OLS}$

“If applied econometrics was easy, theorists would do it. But it’s not as hard as the dense pages of Econometrica might lead you to believe. Carefully applied to coherent causal questions, regression and 2SLS almost always make sense. Your standard errors probably won’t be quite right, but they rarely are. Avoid embarrassment by being your own best skeptic - and, especially, Don’t Panic!”

- Joshua D. Angrist (MIT)