Econ 742. Introductory Econometrics (2)

Chapter ?. Panel Data Models

Fixed effect model:

\[ y_{it} = \alpha_i + x_{it}\beta + \epsilon_{it}, \quad i = 1, \cdots, n; t = 1, \cdots, T, \]

where \( \epsilon_{it} \) are i.i.d. \((0, \sigma^2)\).

This model can be estimated by OLS. It is often called LSDV – least squares with dummy variables model.

\[ \text{OLS}: \min_{\alpha_i, \beta} Q(\alpha_1, \cdots, \alpha_n, \beta) = \sum_{i=1}^{n} \sum_{t=1}^{T} (y_{it} - \alpha_i - x_{it}\beta)^2. \]

The first order conditions give

\[ \sum_{t=1}^{T} (y_{it} - \hat{\alpha}_i - x_{it}\hat{\beta}) = 0, \quad i = 1, \cdots, n; \]
\[ \sum_{i=1}^{n} \sum_{t=1}^{T} (y_{it} - \hat{\alpha}_i - x_{it}\hat{\beta})x'_{it} = 0, \]

which implies that

\[ \hat{\alpha}_i = \bar{y}_i - \bar{x}_i\hat{\beta}, \quad \text{and} \quad \hat{\beta} = \left( \sum_{i=1}^{n} \sum_{t=1}^{T} (x_{it} - \bar{x}_i)'(x_{it} - \bar{x}_i) \right)^{-1} \sum_{i=1}^{n} \sum_{t=1}^{T} (x_{it} - \bar{x}_i)'(y_{it} - \bar{y}_i). \]

In production studies, \( y \) is output and \( x \) is input. The \( \alpha_i \) is assumed to measure the efficiency – such as managerial input specific to the \( i \)th firm.

Error Component Model (or Variance Component Model):

In this model, the intercept term \( \alpha_i \) are treated as random variables. More specifically,

\[ y_{it} = x_{it}\beta + u_{it}, \quad u_{it} = \alpha_i + \epsilon_{it}, \]

where \( \alpha_i \sim \text{i.i.d.}(0, \sigma^2_\alpha) \), \( \epsilon_{it} \sim \text{i.i.d.}(0, \sigma^2) \), and \( \alpha_i \) and \( \epsilon_{it} \) are independent.

Note: In this model, in addition to that \( \epsilon_{it} \) are uncorrelated (or conditional on \( x \)) with \( x \), the \( \alpha_i \) are also uncorrelated (conditional on \( x \)) with \( x \).

Let \( u_i = (\alpha_i + \epsilon_{i1}, \cdots, \alpha_i + \epsilon_{in})' \) for \( i = 1, \cdots, n \) and \( u = (u_1', \cdots, u_n')' \). Let \( l_T \) be a \( T \)-dimensional vector of ones and \( \epsilon_i = (\epsilon_{i1}, \cdots, \epsilon_{iT})' \). Then

\[ E(u_i u_i') = E((\alpha_i l_T + \epsilon_i)(\alpha_i l_T + \epsilon_i)') = \sigma^2_a l_T l_T' + \sigma^2 I_T = \sigma^2_A, \]

where \( \sigma^2_u = \sigma^2_a + \sigma_c^2 \), and

\[ A = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix} = (1 - \rho)I_T + \rho l_T l_T' \]

with \( \rho = \sigma^2_a / (\sigma^2 + \sigma^2_c) \). The variance of the \( nT \times nT \) disturbance vector \( u \) is a block diagonal matrix \( V = E(uu') = \sigma^2_u (I_n \otimes A) \), and its inverse is \( V^{-1} = \frac{1}{\sigma^2_u} (I_n \otimes A^{-1}) \). As \( A \) has a specific pattern, its inverse is

\[ A^{-1} = \lambda_1 l_T l_T' + \lambda_2 I_T, \]

where

\[ \lambda_1 = -\frac{\rho}{(1 - \rho)(1 - \rho + T)} \quad \lambda_2 = \frac{1}{1 - \rho}. \]
Let $x_i = (x_{i1}, \ldots, x_{iT})'$ and $y_i = (y_{i1}, \ldots, y_{iT})'$. This model can be estimated by the GLS (assume $\rho$ is known or estimated)

$$
\hat{\beta}_G = (\sum_{i=1}^{n} x_i' A^{-1} x_i)^{-1} \sum_{i=1}^{n} x_i' A^{-1} y_i.
$$

The GLS estimator is equivalent to the OLS approach to

$$
y_{it} - c\bar{y}_i = (x_{it} - c\bar{x}_i)\beta + (u_{it} - c\bar{u}_i), \quad i = 1, \ldots, n; \quad t = 1, \ldots, T,
$$

where $c = 1 - [(1 - \rho)/(1 - \rho + \rho T)]^{1/2}$. This is so, because

$$
z' A^{-1} z = z' (\lambda_1 t_T l_T' + \lambda_2 l_T) z = \lambda_1 \sum_{t=1}^{T} z_t^2 + \lambda_2 \sum_{t=1}^{T} z_t^2 = \lambda_2 \sum_{t=1}^{T} z_t^2 - \frac{\rho T^2}{1 - \rho + \rho T}(\bar{z})^2,
$$

and

$$
z' A^{-1} y = z' (\lambda_1 t_T l_T' + \lambda_2 l_T) y = \lambda_1 \sum_{t=1}^{T} z_t' (\sum_{t=1}^{T} y_t) + \lambda_2 \sum_{t=1}^{T} z_t' y_t = \lambda_2 \sum_{t=1}^{T} z_t' y_t - \frac{\rho T^2}{1 - \rho + \rho T} \bar{z} \bar{y}.
$$

Define $\bar{z} = z - c\bar{z} l_T$ and $\bar{y} = y - c\bar{y} l_T$ (quasi-deviation),

$$
\bar{z}' \bar{z} = z' z + \rho \bar{z}' \bar{z} T - 2c\bar{z}(T\bar{z}) = z' z - (2cT - \rho T) \bar{z}^2,
$$

and

$$
\bar{z}' \bar{y} = z' y + \rho \bar{z}' \bar{y} T - 2c\bar{z}' \bar{y} = z' y - (2cT - \rho T) \bar{z}' \bar{y}.
$$

So, choose $c$ such that $2cT - \rho T = \rho T^2/(1 - \rho + \rho T)$, or, equivalently solve $c^2 - 2c + \rho T/(1 - \rho + \rho T) = 0$ gives $c = 1 + \frac{1}{2} (1 - \rho)/(1 - \rho + \rho T)$ or $c = 1 - \frac{1}{2} (1 - \rho)/(1 - \rho + \rho T)$.

The GLS has another interpretation. The equation $y_{it} = x_{it} \beta + \alpha_i + \epsilon_{it}$ implies that

$$
y_{it} - \bar{y}_i = (x_{it} - \bar{x}_i)\beta + (\epsilon_{it} - \bar{\epsilon}_i), \quad i = 1, \ldots, n; \quad t = 1, \ldots, T, \quad (1)
$$

and

$$
\bar{y}_i = \bar{x}_i \beta + (\alpha_i + \bar{\epsilon}_i), \quad i = 1, \ldots, n. \quad (2)
$$

The first equation (1) can be estimated by OLS (‘within estimator’)– fixed effects:

$$
\hat{\beta}_w = \left[ \sum_{i=1}^{n} \sum_{t=1}^{T} (x_{it} - \bar{x}_i)' (x_{it} - \bar{x}_i) \right]^{-1} \sum_{i=1}^{n} \sum_{t=1}^{T} (x_{it} - \bar{x}_i)' (y_{it} - \bar{y}_i),
$$

and its variance matrix is

$$
\text{var}(\hat{\beta}_w) = \sigma^2 \left[ \sum_{i=1}^{n} \sum_{t=1}^{T} (x_{it} - \bar{x}_i)' (x_{it} - \bar{x}_i) \right]^{-1}.
$$

The second equation (2) can also be estimated by OLS (‘between estimator’):

$$
\hat{\beta}_b = \left( \sum_{i=1}^{n} \bar{x}_i \bar{y}_i \right)^{-1} \sum_{i=1}^{n} \bar{x}_i \bar{y}_i,
$$

and its variance matrix is

$$
\text{var}(\hat{\beta}_b) = (\sigma^2 + \sigma^2 \bar{x}_i \bar{y}_i)^{-1}.
$$
because $\text{var}(\bar{u}_i) = \sigma_E^2\sigma_P^2/T$. The within and between estimators are uncorrelated. It can be shown that the GLS estimator is a weighted average of the between and within estimators (weights are the shares of inverse of variances):

$$\hat{\beta}_G = \frac{1}{\sigma_E^2} \sum_{i=1}^{n} \sum_{t=1}^{T} (x_{it} - \bar{x}_i)'(x_{it} - \bar{x}_i) + \frac{1}{\sigma_E^2 + \sigma_P^2/T} \sum_{i=1}^{n} x_{it}'x_{it}^{-1}$$

$$\cdot \{\frac{1}{\sigma_E^2} \sum_{i=1}^{n} \sum_{t=1}^{T} (x_{it} - \bar{x}_i)'(x_{it} - \bar{x}_i)\hat{\beta}_w + \frac{1}{\sigma_E^2 + \sigma_P^2/T} \sum_{i=1}^{n} x_{it}'\hat{\beta}_b\}.$$  

**Estimation of $\sigma_\alpha^2$ and $\sigma_e^2$:**

Because $u_{it} = \alpha_i + \epsilon_{it}$, it implies that $\bar{u}_i = \alpha_i + \bar{\epsilon}_i$ and $\bar{u} = \bar{\alpha} + \bar{\epsilon}$. As $E(\sum_{t=1}^{T}(\epsilon_{it} - \bar{\epsilon}_i)^2) = (T-1)\sigma_\epsilon^2$, $E(\sum_{i=1}^{n}\sum_{t=1}^{T}(\epsilon_{it} - \bar{\epsilon})^2) = n(T-1)\sigma_\epsilon^2$. Because $\sum_{i=1}^{n}\sum_{t=1}^{T}(\epsilon_{it} - \bar{\epsilon})^2 = \sum_{i=1}^{n}\sum_{t=1}^{T}(\epsilon_{it} - \bar{\epsilon}_i)^2$, $\sigma_\epsilon^2$ can be estimated as

$$\hat{\sigma}_\epsilon^2 = \frac{1}{n(T-1)} \sum_{i=1}^{n} \sum_{t=1}^{T} (\bar{u}_it - \bar{u}_i)^2$$

with $\bar{u}_it$ being the estimated OLS residual. Furthermore, because

$$\sum_{i=1}^{n} (\bar{u}_i - \bar{u})^2 = \sum_{i=1}^{n} [(\alpha_i - \bar{\alpha})^2 + (\bar{\epsilon}_i - \bar{\epsilon})^2 + 2(\alpha_i - \bar{\alpha})(\bar{\epsilon}_i - \bar{\epsilon})],$$

$E(\sum_{i=1}^{n}(\alpha_i - \bar{\alpha})^2)$, $E(\sum_{i=1}^{n}(\bar{\epsilon}_i - \bar{\epsilon})^2)$, $E(\sum_{i=1}^{n}(\epsilon_{it} - \bar{\epsilon})^2)$, and $E(\alpha_i - \bar{\alpha})(\bar{\epsilon}_i - \bar{\epsilon}) = 0$, it follows that

$$E(\sum_{i=1}^{n}(\bar{u}_i - \bar{u})^2) = (n-1)\sigma_\alpha^2 + (n-1)\sigma_e^2/T,$$

and $\sigma_\alpha^2$ can be estimated by

$$\hat{\sigma}_\alpha^2 = \frac{1}{n-1} \sum_{i=1}^{n} \bar{u}_i^2 - \hat{\sigma}_\epsilon^2/T.$$  

**Hausman’s specification test:**

Suppose that there exists two estimators $\hat{\theta}$ and $\hat{\theta}_0$ in $R^k$ with the following properties:

1. $\hat{\theta}$ is asymptotically efficient under $H_0$ but is inconsistent under $H_1$,
2. $\hat{\theta}_0$ is asymptotically inefficient under $H_0$ but is consistent under both $H_0$ and $H_1$. The Hausman test statistic is based on the comparison of $\hat{\theta}$ and $\hat{\theta}_0$. Let $V$ be a consistent estimate of the asymptotic covariance matrix of $\hat{\theta} - \hat{\theta}_0$. If $V$ is nonsingular,

$$(\hat{\theta} - \hat{\theta}_0)'V^{-1}(\hat{\theta} - \hat{\theta}_0) \overset{D}{\rightarrow} \chi^2(k).$$

Under the null hypothesis, the asymptotic efficiency of $\hat{\theta}$ implies that $V(\hat{\theta} - \hat{\theta}_0) = V(\hat{\theta}) - V(\hat{\theta}_0)$. The asymptotic efficiency property indeed implies $\text{cov}(\hat{\theta}, \hat{\theta}_0) = V(\hat{\theta})$. This can be shown as a contradictory argument. Suppose that this is not so, $\text{cov}(\hat{\theta}, \hat{\theta}_0) - V(\hat{\theta}) \neq 0$. Consider an estimator defined as

$$\theta^* = \hat{\theta} + (V(\hat{\theta}) - \text{cov}(\hat{\theta}, \hat{\theta}))\text{[V(\hat{\theta} - \hat{\theta})]}^{-1}(\hat{\theta} - \hat{\theta}).$$

It follows that

$$V(\theta^*) = V(\hat{\theta}) + [V(\hat{\theta}) - \text{cov}(\hat{\theta}, \hat{\theta})][V(\hat{\theta} - \hat{\theta})]^{-1}\text{[V(\hat{\theta} - \hat{\theta})]}' + [\text{cov}(\hat{\theta}, \hat{\theta}) - V(\hat{\theta})][V(\hat{\theta} - \hat{\theta})]^{-1}[\text{cov}(\hat{\theta}, \hat{\theta}) - V(\hat{\theta})]'$$

$$+ [\text{cov}(\hat{\theta}, \hat{\theta}) - V(\hat{\theta})][V(\hat{\theta} - \hat{\theta})]^{-1}[\text{cov}(\hat{\theta}, \hat{\theta}) - V(\hat{\theta})]'> V(\hat{\theta}),$$

which is a contradiction because $\hat{\theta}$ is asymptotically efficient. (Note: the above construction is motivated by regressing $\hat{\theta}$ on $\hat{\theta}_0$ and $\theta^*$ is the LS residual.)

The Hausman test statistic can be rewritten as $(\hat{\theta} - \hat{\theta}_0)'[V(\hat{\theta}) - V(\hat{\theta})]^{-1}(\hat{\theta} - \hat{\theta}_0)$.