Linear Panel Data Models

Table of contents

1 Overview

In this handout we will see how to test **static linear** panel data regression models. We will review a number of estimators for these models, including a range of potential estimators:

- pooled OLS;
- between-group;
- feasible GLS;
- within-group;
- and first differences.

Further reading can be found in:

- Section 21 of Cameron and Trivedi (2005)
- Section 10.1-10.3 of Verbeek (2017)

2 Panel Data Regression Model

The basic, linear panel-data-regression model is given by,

$$Y_{it} = X_{it}'\beta + \alpha_i + \varepsilon_{it}$$

for $i = 1, .., n$ and $t = 1, ..., T$. By basic, we mean *static*. The contrast would be dynamic panel data models, which include lag(s) (and/or leads) of the outcome variable on the right-hand side. In dynamic models, the long-run (equilibrium) relationship between outcome and regressors differs from the short-run (or contemporaneous) relationship. This is the result of including a lagged dependent variable. This static model, can incorporate lags (and/or leads) of regressors; although, you may then need to consider carefully the exogeneity assumption if strict exogeneity does not apply.

For the purposes of this discussion, we will treat T as fixed. As n increases, T remains fixed, implying that the asymptotics concern only n.

If collect all T observations of unit i, we can describe them by the model,

$$Y_i = X_i\beta + \alpha_i\ell + \varepsilon_i$$

where Y_i is a $T \times 1$ random vector; X_i a $T \times k$ random matrix; and ℓ a $T \times 1$ vector of 1's.

This model has placed no restriction on the values of the outcome variable, Y_i, and regressors, X_i. In particular, the regressors may include time-varying as well as time-invariant variables.

We can extend this specification to include a linear or non-linear trend in time; for example, ϕt. However, the more common option is to include a very flexible time trend using fixed effects:

$$\delta_t = \sum_{j=1}^{T} \delta_j \mathbf{1}\{t = j\}$$

Time fixed-effects are essentially a dummy variable for each time-period. This flexible function - sometimes referred to as saturated - can approximate any functional form of the underlying time trend. Models that include both α_i and δ_t are referred to as **two-way fixed-effects** models.

2.1 Unobserved heterogeneity

The term α_i is particularly important. It represents an unobserved individual effect or **unobserved heterogeneity**. I strongly recommend you read the discussion on pages 285-286 of Wooldridge (2010). You should not refer to α_i as individual **fixed-effects**.

The language of fixed-effects comes from models where α_i is a unit-specific (population) parameter. As a parameter, it is by definition non-random. This would imply that the correlation between α_i and X_i is by definition 0. Here α_i is an unobserved random variable. Wooldridge (2010) refers to it as an unobserved effects model (UEM). The term random effects model is often used to describe UEM models where the individual effect is (mean) independent.

The α_i is unobserved, and therefore is a component of the **composite error term** $v_{it} = \alpha_i + \varepsilon_{it}$. The time-invariant component (α_i) of the error is *permanent*, while the time-varying component (ε_{it}) is *transient*; also referred to as the idiosyncratic error.

Given the model, we must consider all three components on the right-hand side - $\{X_i, \alpha_i, \varepsilon_{it}\}$ - when making assumptions regarding exogeneity. **We will make all assumptions conditional on α_i.** This is consistent with the idea that α_i is a permanent shock, and is therefore realized before $\varepsilon_i t$.

If we do not condition on α_i, then when we evaluate $E[Y_i|X_i]$, we also need to consider both $E[\varepsilon_i|X_i]$ and $E[\alpha_i|X_i]$; not just $E[\varepsilon_i|X_i, \alpha_i]$. Moreover, in general it will be that case that $E[\alpha_i|X_i] \neq E[\alpha_i]$ (i.e., not mean independent). For example, α_i may represent unobserved ability in a wage equation. This will correlated with regressors like education.

3 Exogeneity

Having assumed that the samples are independent across i, we can define mean independence of the transient error term component for unit i. There are three potential assumptions we can make:

1. Strict exogeneity:

$$E[\varepsilon_i|X_i, \alpha_i] = 0$$

or,

$$E[\varepsilon_{it}|X_{i1}, X_{i2}, ..., X_{iT}, \alpha_i] = 0 \qquad \forall\ t$$

2. Weak or sequential exogeneity:

$$E[\varepsilon_{it}|X_{i1}, X_{i2}, ..., X_{it}, \alpha_i] = 0 \qquad \forall\ t$$

Exogeneity with respect to the past sequence of regressors (or predetermined regressors).

3. Contemporaneous exogeneity:

$$E[\varepsilon_{it}|X_{it}, \alpha_i] = 0 \qquad \forall\ t$$

Exogeneity only with respect to the contemporaneous value of X_i.

3.1 Strict exogeneity

Strict exogeneity is a very strong assumption. It implies that X is uncorrelated with past, current, and future values of the transient error term (conditional on α_i): $E[X_{it}, \varepsilon_{is}|\alpha_i] = 0\ \forall\ t, s$. Crucially, X_{it} cannot respond to the history of idiosyncratic shocks $\varepsilon_{i1}, \varepsilon_{i2}, ..., \varepsilon_{it}$.

Under strict exogeneity,

$$E[Y_{it}|X_i, \alpha_i] = E[Y_{it}|X_{it}, \alpha_i] = X'_{it}\beta + \alpha_i$$

The first equality implies that once you control for X_{it}, there is no additional partial effect of X_{is} ($\forall\ s \neq t$) on (the mean of) Y_{it}. This assumption relates to the assumed *static* nature of the model: the model includes not lags (or leads) of the dependent variable.

3.2 Weak exogeneity

Weak exogeneity, also referred to as sequential exogeneity, implies that the error term is uncorrelated with past and contemporaneous values of the regressors:

$$E[X_{is}\varepsilon_{it}] = 0 \qquad \forall\ s = 1, ..., t$$

In the static linear model, it also implies that,

$$E[Y_{it}|X_i, \alpha_i] = E[Y_{it}|X_{it}, \alpha_i] = X'_{it}\beta + \alpha_i$$

This structure can permit a lagged dependent variable amongst the regressors (assuming there is no serial correlation of the error term). However, the assumption remains violated if the regressors are endogenous.

3.3 Contemporaneous exogeneity

This assumption implies that the error term is only uncorrelated with regressors in the same time period,

$$E[X_{it}\varepsilon_{it}] = 0$$

Regardless, it still implies that,

$$E[Y_{it}|X_i, \alpha_i] = E[Y_{it}|X_{it}, \alpha_i] = X'_{it}\beta + \alpha_i$$

4 Static Linear Panel Model

We begin by describing the core assumptions underlying the static (or basic) panel data regression model (SLPM) assumptions. Many of these will be familiar to you

SLPM 1: The model is static and linear in parameters with a composite error term made up of a time-invariant and time-varying component:

$$Y_{it} = X'_{it}\beta + \alpha_i + \varepsilon_{it}$$

SLPM 2: Strict exogeneity: $E[\varepsilon_{it}|X_i, \alpha_i] = 0 \; \forall \; t$

SLPM 3: Conditional homoskedasticity and serial uncorrelatedness of the transient error term component.

$$Var(\varepsilon_i|X_i, \alpha_i) = \begin{bmatrix} \sigma_\varepsilon^2 & 0 & \cdots & 0 \\ 0 & \sigma_\varepsilon^2 & & \\ \vdots & & \ddots & \\ 0 & & & \sigma_\varepsilon^2 \end{bmatrix} = \sigma_\varepsilon^2 I_T$$

Together, assumptions 2 & 3 are referred to as a 'classical error term' structure. Of course, since the model is linear, we need a rank condition. Without placing any additional assumptions on the variation of X_{it}, we need to assume that:

SLPM 4: $rank(X) = k$

Next, we need to consider sampling.

SLPM 5: Independent sampling along the cross-sectional dimension (i).

SLPM 6: Balanced panel: we observe each i in all T time periods.

Under assumptions 1-6:

- $E[Y_i|X_i, \alpha_i] = X_i\beta + \alpha_i\ell$
- $Var(Y_i|X_i, \alpha_i) = \sigma_\varepsilon^2 I_T$

At this stage, the unanswered question is how to deal with the unobservables α_i in the equation. Generally, there are two approaches:

1. assume away the relationship between α_i and X_i;
2. or remove α_i from the equation prior to estimator.

Adopting approach (1), we will review the pooled OLS, between-group, and (feasible) Generalized Least Squares (GLS) estimators. Under approach (2), will review the within-group, fixed effects, and first-difference estimators.

5 Pooled OLS

We can 'assume away' the relationship between α_i and X_i. Specifically, we will assume conditional mean independence:

- **SLPM 7** $E[\alpha_i|X_i] = E[\alpha_i|X_{i1}, X_{i2}, ..., X_{iT}] = E[\alpha_i] = 0$

As in the CLRM, the assumption that the unconditional mean of α_i is zero is not binding given the inclusion of a constant among the regressors. This also implies uncorrelatedness: $E[X_i\alpha_i] = 0$. Under this assumptions, the model can be written as,

$$Y_{it} = X_{it}'\beta + \upsilon_{it}$$

where $E[\upsilon_i|X_i] = 0$. Alternatively, we can stack all nT observations together,

$$Y = X\beta + \upsilon$$

Second, we need to make an assumption regarding the variance of the unobserved heterogeneity:

- **SLPM 8** $Var(\alpha_i|X_i) = \sigma_\alpha^2$

Under these assumptions, the error term $\upsilon_i = \alpha_i \ell + \varepsilon_i$ has the variance,

$$Var(\upsilon_i|X_i) = E[\upsilon_i \upsilon_i'|X_i] = \sigma_\alpha^2 \ell\ell' + \sigma_\varepsilon^2 I_T = \Sigma$$

For all $s \neq t$ the $E[\upsilon_{it}, \upsilon_{is}] = \sigma_\alpha^2$. And the diagonal elements are given by $\sigma_\alpha^2 + \sigma_\varepsilon^2$.

Under these assumptions, the OLS estimator,

$$\begin{aligned}\hat{\beta}^{OLS} =& (X'X)^{-1}X'Y \\ =& \beta + (X'X)^{-1}X'\upsilon \\ =& \beta + (\sum_i X_i'X_i)^{-1} \sum_i X_i'\upsilon_i\end{aligned}$$

is both unbiased and consistent. This is because,

$$p\lim \frac{1}{n} \sum_i X_i'\upsilon_i = \sum_{t=1}^{T} p\lim \frac{1}{n} \sum_{i=1}^{n} X_{it}\upsilon_{it} = \sum_{t=1}^{T} E[X_{it}\upsilon_{it}] = 0$$

The asymptotic distribution is given by,

$$\begin{aligned}\sqrt{n}(\hat{\beta}^{OLS} - \beta) =& \left(\frac{1}{n}\sum_i X_i'X_i\right)^{-1} \frac{1}{\sqrt{n}} \sum_i X_i'\upsilon_i \\ &\to_d N(0, V^{-1}\Omega V^{-1})\end{aligned}$$

where,

- $V = E[X_i'X_i]$
- $\Omega = E[X_i'\Sigma X_i]$

We say that the approximate distribution of the pooled OLS estimator is given by,

$$\hat{\beta}^{OLS} \overset{a}{\sim} N\left(\beta, (\sum_i X_i'X_i)^{-1} \sum_i X_i'\Sigma X_i (\sum_i X_i'X_i)^{-1}\right)$$

Note, the variance is not homoskedastic. You must therefore estimate heteroskedastic (or clustered) standard errors. The usual homoskedastic estimator for the variance will be biased and inconsistent. Unobserved heterogeneity will result in a serial correlation across the error terms that is not accountant for by the standard estimator.

For this reason, **pooled OLS is NOT efficient**.

6 Between Group

An alternative to pooled OLS, is to collapse the multiple observations of unit i into a single cross-section aggregate. This transforms the model into,

$$\bar{Y}_i = \bar{X}_i'\beta + \bar{\upsilon}_i$$

where $\bar{\upsilon}_i = \alpha_i + \bar{\varepsilon}_i$. The variance of this error term is,

$$E[\bar{\upsilon}_i^2|X_i] = \sigma_\alpha^2 + \frac{\sigma_\varepsilon^2}{T}$$

The OLS estimator for β is given by,

$$\hat{\beta}^{BG} = (\bar{X}'\bar{X})^{-1}\bar{X}'\bar{Y} = (\sum_i \bar{X}_i\bar{X}_i')^{-1}\sum_i \bar{X}_i\bar{Y}_i$$

Since the variance term is now homoskedastic, the standard homoskedastic variance estimator will unbiased and consistent. This approach removes the problem of serially correlated error terms across repeated observations of i in pooled OLS by collapsing all observations to a single observation. However, it also reduces the information in the data and is therefore less efficient.

7 Generalized Least Squares

The efficient solution is to account for the error term structure in the estimation using Generalized Least Squares. The structure of the composite error-term variance-covariance matrix is,

$$\Sigma = \begin{bmatrix} \sigma_\alpha^2 + \sigma_\varepsilon^2 & \sigma_\alpha^2 & \cdots & \sigma_\alpha^2 \\ \sigma_\alpha^2 & \sigma_\alpha^2 + \sigma_\varepsilon^2 & & \\ \vdots & & \ddots & \\ \sigma_\alpha^2 & & & \sigma_\alpha^2 + \sigma_\varepsilon^2 \end{bmatrix}$$

The $nT \times nT$ matrix, $E[\upsilon\upsilon'|X]$, is a block-diagonal matrix in which the off-diagonal values are $E[\upsilon_{it}\upsilon_{js}|X] = \sigma_\alpha^2$ only for $i = j$ and $s \neq t$; and zero otherwise. We can describe this matrix using a Kronecker-product operator:

$$E[\upsilon\upsilon'|X] = \begin{bmatrix} \Sigma & 0 & \cdots & 0 \\ 0 & \Sigma & & \\ \vdots & & \ddots & \\ 0 & & & \Sigma \end{bmatrix} = I_n \otimes \Sigma$$

In this $nT \times nT$ matrix, eacg off-diagonal element is a $T \times T$ matrix of 0's, representing the cross-unit (i) covariance terms.

The Generalized Least Squares solution solves the following least squares problem:

$$\hat{\beta}^{GLS} = \arg\min_{b} \quad (Y - Xb)'(I_n \otimes \Sigma^{-1})(Y - Xb)$$

which can be written as,

$$\hat{\beta}^{GLS} = \arg\min_{b} \quad (Y^{+} - X^{+}b)'(Y^{+} - X^{+}b)$$

where $[Y^{+}, X^{+}] = [\Sigma^{-1/2}Y, \Sigma^{-1/2}X]$. In this instance, the net result of this linear transform of the model is give by,

$$\underbrace{Y_{it} - \theta\bar{Y_i}}_{Y_{it}^{+}} = \underbrace{(X_{it} - \theta\bar{X}_i)'}_{X_{it}^{+\prime}}\beta + \upsilon_{it}^{+}$$

where,

$$\theta = 1 - \frac{\sigma_{\varepsilon}}{\sqrt{T\sigma_{\alpha}^2 + \sigma_{\varepsilon}^2}}$$

Consider the transformed error term ν,

$$\upsilon_{it}^{+} = \upsilon_{it} - \theta\bar{\upsilon}_i = (1 - \theta)\alpha_i + \varepsilon_{it} - \frac{\theta}{T}\sum_{t}\varepsilon_{it}$$

The serial correlation of this error term is 0. Consider, for $t \neq s$

$$\begin{aligned}
E[v_{it}^{+}v_{is}^{+}|X_i] =& E[((1-\theta)\alpha_i + \varepsilon_{it} - \frac{\theta}{T}\sum_{t'}\varepsilon_{it'})((1-\theta)\alpha_i + \varepsilon_{is} - \frac{\theta}{T}\sum_{t'}\varepsilon_{it'})|X_i] \\
=& (1-\theta)^2\sigma_\alpha^2 - 2\frac{\theta}{T}\sigma_\varepsilon^2 + \frac{\theta^2}{T^2}\sum_{t'}\sigma_\varepsilon^2 \\
=& \frac{\sigma_\varepsilon^2\sigma_\alpha^2}{T\sigma_\alpha^2+\sigma_\varepsilon^2} + \frac{\theta(\theta-2)}{T}\sigma_\varepsilon^2 \\
=& \frac{\sigma_\varepsilon^2\sigma_\alpha^2}{T\sigma_\alpha^2+\sigma_\varepsilon^2} - \frac{\sigma_\varepsilon^2}{T}\left(1 - \frac{\sigma_\varepsilon}{\sqrt{T\sigma_\alpha^2+\sigma_\varepsilon^2}}\right)\left(1 + \frac{\sigma_\varepsilon}{\sqrt{T\sigma_\alpha^2+\sigma_\varepsilon^2}}\right) \\
=& \frac{\sigma_\varepsilon^2\sigma_\alpha^2}{T\sigma_\alpha^2+\sigma_\varepsilon^2} - \frac{\sigma_\varepsilon^2}{T}\left(1 - \frac{\sigma_\varepsilon^2}{T\sigma_\alpha^2+\sigma_\varepsilon^2}\right) \\
=& \frac{\sigma_\varepsilon^2\sigma_\alpha^2}{T\sigma_\alpha^2+\sigma_\varepsilon^2} - \frac{\sigma_\varepsilon^2\sigma_\alpha^2}{T\sigma_\alpha^2+\sigma_\varepsilon^2} \\
=& 0
\end{aligned}$$

The GLS estimator is then given by,

$$\hat{\beta}^{GLS} = [\sum_i X_i^{+'}X_i^{+}]^{-1}\sum_i X_i^{+'}Y_i^{+}$$

You can show that $\hat{\beta}^{GLS}$ is a weighted average of $\hat{\beta}^{BG}$ and $\hat{\beta}^{WG}$ (see below). Also, take note of the fact that as $T \to \infty$, $\theta \to 1$. Thus, $\hat{\beta}_{GLS} \to \hat{\beta}^{WG}$ as $T \to \infty$. In addition, if if $\sigma_\alpha^2 = 0$, then $\theta = 0$ and $\hat{\beta}^{GLS} = \hat{\beta}^{OLS}$, the pooled OLS estimator.

However, this estimator is NOT feasible. This is because we do not observe $\{\sigma_\alpha^2, \sigma_\varepsilon^2\}$.

7.1 Feasible GLS

A feasible version of the GLS estimator is given by the following steps:

1. Estimate σ_ε^2 using the WG estimator (see below).
2. Use the pooled OLS or BG estimator then estimate σ_α^2, using the value of σ_ε^2 from step 1. For example, the RSS from pooled OLS (divided by $nT-k$) is a consistent estimator for $\sigma_\varepsilon^2 + \sigma_\alpha^2$. Similar, the RSS from BG estimator (divided by $n-k$) is a consistent estimator for $\sigma_\varepsilon^2/T + \sigma_\alpha^2$.
3. Using the estimated $\{\hat{\sigma}_\alpha^2, \hat{\sigma}_\varepsilon^2\}$, compute the transformed model (using $\hat{\theta}$) and estimate using $\hat{\beta}^{FGLS}$ using OLS.

In Stata, this estimator is referred to as the random effects estimtor within the `xtreg` package: `xtreg , re`.

8 Within Group

The second approach to dealing with unobserved heterogeneity is to transform the model in such a way that α_i is eliminated. Having done so, we do not need to make any assumption regarding $E[\alpha_i|X_i]$.

Here we will exploit the fact that α_i is time-invariant. We begin by computing the unit-level average of the model. For the left-hand side,

$$\bar{Y}_i = \frac{1}{T}\sum_{t=1}^{T} Y_{it}$$

and the right-hand side,

$$\frac{1}{T}\sum_{t=1}^{T}\left(X_{it}'\beta + \alpha_i + \varepsilon_{it}\right) = \bar{X}_i'\beta + \alpha_i + \bar{\varepsilon}_i$$

Next, subtract this from each value, to create a demeaned expression

$$\underbrace{Y_{it} - \bar{Y}_i}_{\tilde{Y}_{it}} = \underbrace{(X_{it} - \bar{X}_i)'\beta}_{\tilde{X}_{it}'\beta} + \underbrace{\alpha_i - \alpha_i}_{=0} + \underbrace{\varepsilon_{it} - \bar{\varepsilon}_i}_{\tilde{\varepsilon}_{it}}$$

The permanent error-term component drops out precisely because it is time-invariant. The transformed model is given by,

$$\tilde{Y}_{it} = \tilde{X}_{it}'\beta + \tilde{\varepsilon}_{it}$$

This model can be estimated by OLS. The solution is given by,

$$\hat{\beta}^{WG} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{Y} = (\sum_{i=1}^{n}\tilde{X}_i'\tilde{X}_i)^{-1}\sum_{i=1}^{n}\tilde{X}_i'\tilde{Y}_i$$

where $\tilde{X}_i$ is a $T \times k$ matrix. **Note, this is different to the standard cross-section expression.**

This solution assumes that the $k \times k$ matrix $\sum_{i=1}^{n}\tilde{X}_i'\tilde{X}_i$ is invertible. We must therefore make a modified rank assumption,

- **SLPM 4'**: $rank(\tilde{X}) = k$

Given that $\tilde{X}$ is the within-group demeaned vale of X, this implies the regressors must all be time-varying. In addition, if the model includes a time trend (including time-FEs), the included variables cannot vary uniformly with time. An example of this is age. If all units' age values increase by the same amount each period, then

$$a\tilde{g}e_{it} = age_{it} - a\bar{g}e_i = t - \bar{t}$$

This is because an individuals age can be expressed as a time-invariant value of their year of birth yob_i plus a linear time-trend that has a unit-specific intercept. The demeaned value of age is perfectly colinear with a demeaned linear time-trend. It would also be perfectly colinear with a higher-order polynomial time-trend or year FEs.

$$\tilde{X}_i = \begin{bmatrix} \tilde{X}_{i11} & \tilde{X}_{i12} & \cdots & \tilde{X}_{i1k} \\ \tilde{X}_{i21} & \tilde{X}_{i22} & & \\ \vdots & & \ddots & \\ \tilde{X}_{iT1} & & & \tilde{X}_{iTk} \end{bmatrix} = \begin{bmatrix} \tilde{X}'_{i1} \\ \tilde{X}'_{i2} \\ \vdots \\ \tilde{X}'_{iT} \end{bmatrix}$$

$\tilde{X}'_i\tilde{X}_i$ is therefore a $k \times k$ matrix, which can be expressed as,

$$\tilde{X}'_i\tilde{X}_i = \sum_{t=1}^{T} \tilde{X}_{it}\tilde{X}'_{it} = \begin{bmatrix} \sum_t \tilde{X}^2_{it1} & \sum_t \tilde{X}_{it1}\tilde{X}_{it2} & \cdots & \sum_t \tilde{X}_{it1}\tilde{X}_{itk} \\ \sum_t \tilde{X}_{it2}\tilde{X}_{it1} & \sum_t \tilde{X}^2_{it2} & & \\ \vdots & & \ddots & \\ \sum_t \tilde{X}_{itk}\tilde{X}_{it1} & & & \sum_t \tilde{X}^2_{itk} \end{bmatrix}$$

Finally, we can express $\tilde{X}'\tilde{X}$ as,

$$\tilde{X}'\tilde{X} = \sum_{i=1}^{n} \tilde{X}'_i\tilde{X}_i = \begin{bmatrix} \sum_i (\sum_t \tilde{X}^2_{it1}) & \sum_i (\sum_t \tilde{X}_{it1}\tilde{X}_{it2}) & \cdots & \sum_i (\sum_t \tilde{X}_{it1}\tilde{X}_{itk}) \\ \sum_i (\sum_t \tilde{X}_{it2}\tilde{X}_{it1}) & \sum_i (\sum_t \tilde{X}^2_{it2}) & & \\ \vdots & & \ddots & \\ \sum_i (\sum_t \tilde{X}_{itk}\tilde{X}_{it1}) & & & \sum_i (\sum_t \tilde{X}^2_{itk}) \end{bmatrix}$$

We can use the same method to describe $\sum_{i=1}^{n} \tilde{X}'_i\tilde{Y}_i$, a $k \times 1$ vector. Substituting in the definition of $\tilde{Y}_i$ from the transformed model, we get that,

$$\hat{\beta}^{WG} = \beta + (\sum_{i=1}^{n} \tilde{X}'_i\tilde{X}_i)^{-1} \sum_{i=1}^{n} \tilde{X}'_i\tilde{\varepsilon}_i$$

8.1 Conditional variance

Given the demeaning of the model, the error term is no longer uncorrelated across t for the same i:

$$\begin{aligned}
E[\tilde{\varepsilon}_{is}\tilde{\varepsilon}_{it}|X] =& E[(\varepsilon_{is}-\bar{\varepsilon}_i)(\varepsilon_{it}-\bar{\varepsilon}_i)|X] \\
=& E\left[\left(\varepsilon_{is}-\frac{1}{T}\sum_{s'}\varepsilon_{is'}\right)\left(\varepsilon_{it}-\frac{1}{T}\sum_{s'}\varepsilon_{it'}\right)\middle|X_i\right] \\
=& \begin{cases} \sigma_\varepsilon^2(1-1/T) & \text{for} \quad s=t \\ -\sigma_\varepsilon^2/T & \text{for} \quad s\neq t \end{cases}
\end{aligned}$$

The off-diagonal elements for the same i are the same for all time-periods. Together, this gives us the $T \times T$ matrix,

$$\begin{aligned}
\Sigma =& E[\tilde{\varepsilon}_i\tilde{\varepsilon}_i'|X_i] \\
=& \begin{bmatrix} \sigma_\varepsilon^2(1-1/T) & -\sigma_\varepsilon^2/T & \cdots & -\sigma_\varepsilon^2/T \\ -\sigma_\varepsilon^2/T & \sigma_\varepsilon^2(1-1/T) & & \\ \vdots & & \ddots & \\ -\sigma_\varepsilon^2/T & & & \sigma_\varepsilon^2(1-1/T) \end{bmatrix} \\
=& \sigma_\varepsilon^2 M_\ell
\end{aligned}$$

where $M_\ell = I_T - \frac{\ell\ell'}{T}$ for the $T \times 1$ vector of ones ℓ. This follows from the fact that $\tilde{\varepsilon}_i = M_\ell \varepsilon_i$. Which implies that,

$$\begin{aligned}
E[\tilde{\varepsilon}_i\tilde{\varepsilon}_i'|X_i] =& E[M_\ell\varepsilon_i\varepsilon_i'M_\ell'|X_i] \\
=& M_\ell E[\varepsilon_i\varepsilon_i'|X_i]M_\ell \\
=& M_\ell\sigma_\varepsilon^2 I_T M_\ell \\
=& \sigma_\varepsilon^2 M_\ell
\end{aligned}$$

Thus,

$$E[\tilde{\varepsilon}\tilde{\varepsilon}'|X] = I_n \otimes \Sigma = \sigma_\varepsilon^2 I_n \otimes M_\ell$$

where the $rank(I_n \otimes M_\ell) = nT - n$. This follows from the fact that each M_ℓ has $rank(M_\ell) = T-1$. The matrix $M_{I_n \otimes M_\ell}$ is a $nT \times nT$ matrix: a block diagonal matrix of n M_ℓ matrices. With this we can now solve for the conditional variance of the WG estimator.

$$\begin{aligned}
Var(\hat{\beta}^{WG}|X) =& (\tilde{X}'\tilde{X})^{-1}\tilde{X}'E[\tilde{\varepsilon}\tilde{\varepsilon}'|X]\tilde{X}(\tilde{X}'\tilde{X})^{-1} \\
=& \sigma_\varepsilon^2(\tilde{X}'\tilde{X})^{-1}\tilde{X}'(I_n \otimes M_\ell)\tilde{X}(\tilde{X}'\tilde{X})^{-1} \\
=& \sigma_\varepsilon^2(\tilde{X}'\tilde{X})^{-1}
\end{aligned}$$

The final line follows from the fact that $\tilde{X} = (I_n \otimes M_\ell)X$ and $I_n \otimes M_\ell$ is, itself, an idempotent projection matrix.

An unbiased and consistent estimator for σ_ε^2 is given by,

$$\hat{\sigma}_\varepsilon^2 = \frac{RSS}{dof} = \frac{\sum_{i=1}^n \sum_{t=1}^T \left(Y_{it} - \bar{Y}_i - (X_{it} - \bar{X}_i)'\hat{\beta}^{WG}\right)^2}{nT - n - k}$$

The residual degrees of freedom is equal to $nT - n - k$, not $nT - k$. While there are nT observations and k parameters, we must deduct the n unit-level means computed. Another way to see this is to use the above projection matrix decomposition. For the purpose of this decomposition, denote $I_n \otimes M_\ell = M_{n\ell}$ an idempotent (orthogonal) projection matrix. Then,

$$\begin{aligned} RSS =& \hat{\varepsilon}'\hat{\varepsilon} \\ =& \tilde{Y}'(I_{nT} - \tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}')\tilde{Y} \\ =& Y'M_{n\ell}(I_{nT} - M_{n\ell}X(X'M_{n\ell}X)^{-1}X'M_{n\ell})M_{n\ell}Y \\ =& Y' \underbrace{(M_{n\ell} - M_{n\ell}X(X'M_{n\ell}X)^{-1}X'M_{n\ell})}_{M_+} Y \end{aligned}$$

This new matrix M_+ is also an idempotent projection matrix, with rank $rank(M_+) = rank(M_{n\ell}) - \min\{rank(M_{n\ell}), rank(X)\} = nT - n - k$. As we saw with the CLRM, this term with have a χ^2 distribution with dof equal to the rank of the projection matrix.

8.2 Consistency and asymptotic distribution

Under assumptions SLPM 1-6, $\hat{\beta}^{WG}$ is consistent,

$$\hat{\beta}^{WG} \to_p \beta \qquad \text{as} \quad n \to \infty$$

and asymptotically normal,

$$\hat{\beta}^{WG} \overset{a}{\sim} N(\beta, \sigma_\varepsilon^2 (\sum_i \tilde{X}_i' \tilde{X}_i)^{-1})$$

Both results require that $E[\tilde{X}_i, \tilde{\varepsilon}_i] = 0$. This is maintained by strict exogeneity (CLPM 4), which states that in addition to $E[X_{is}, \varepsilon_{it}] = 0 \; \forall \; s, t$,

$$E[\bar{X}_i, \bar{\varepsilon}_i] = 0$$

This would not be true under the weak exogeneity assumption.

8.3 Fixed Effects

In Stata's `xtreg` package, the WG estimator is referred to as the Fixed Effects (FE) estimator. This not not mean that α_i is non-random. It simply means that the WG estimator is equivalent to the OLS estimator for a (individual/unit) fixed effects model. Consider the model,

$$Y_{it} = \sum_{j=1}^{n} \phi_j \mathbf{1}\{i = j\} + X_{it}'\beta + \upsilon_{it}$$

This model includes a dummy variable for each unit. For each unit, only one dummy variable can be =1, the dummy variable with parameter ϕ_i. Thus, the expression $\sum_{j=1}^{n} \phi_j \mathbf{1}\{i = j\} = \phi_i$ for any i. The model can therefore be written as,

$$Y_{it} = \phi_i + X_{it}'\beta + \upsilon_{it}$$

This looks very similar to our SLPM, but with the distinguishing feature that ϕ_i is taken as a non-random population parameter. This is sometimes referred as a unit-specific constant.

Including a dummy variable each i has the same effect as demeaning the model prior to estimation (as with WG). This approach is referred to as the Least Squares Dummy Variable (LSDV) method.

$$\hat{\beta}^{WG} = \hat{\beta}^{LSDV}$$

This equivalence can be shown using Frisch Waugh Lovell Theorem (or partitioned regression). The LSDV estimator can be computed in two steps. First, regress each regressor and outcome on a setting of unit level dummies.

$$X_k = \sum_{j=1}^{n} \phi_j \mathbf{1}\{i = j\} + \xi$$

Each $\mathbf{1}\{i = j\}$ corresponds to a dummy variable where T values are =1 (for unit i), and the remainder 0. Next, use the residuals in the main equation. The residual from the regression is given by,

$$(I_n \otimes M_\ell)X_k = \tilde{X}_k$$

Thus, we regression is given by,

$$(I_n \otimes M_\ell)Y = \tilde{Y} = (I_n \otimes M_\ell)X\beta + (I_n \otimes M_\ell)\varepsilon = \tilde{X}\beta + \tilde{\varepsilon}$$

Employing the LSDV approach, one can estimated the unit-FE as,

$$\hat{\phi}_i = \bar{Y}_i - \bar{X}_i'\hat{\beta}^{LSDV}$$

While this estimator is unbiased, $E[\hat{\phi}_i|X_i] = \phi$, it is NOT consistent for fixed T. It is only consistent if $T \to \infty$ as $n \to \infty$.

The equivalene of these approaches also explains why the degrees of freedom in the residual is $nT - n - k$. By including n dummy variables, the number of parameters we need to estimate is $n + k$.

9 First Difference

As with the WG estimator, he first-difference (FD) estimator removes α_i from the model through differencing. However, this time the transformation is just a single difference:

$$\underbrace{Y_{it} - Y_{it-1}}_{\Delta Y_{it}} = \underbrace{(X_{it} - X_{it-1})'\beta}_{\Delta X_{it}'\beta} + \underbrace{\alpha_i - \alpha_i}_{=0} + \underbrace{\varepsilon_{it} - \varepsilon_{it-1}}_{\Delta \varepsilon_{it}}$$

As a result, the estimation sample will include 1 less period: $t = 2, ..., T$ for each i. In addition, we must make a modified rank assumption,

- **SLPM 4"**: $rank(\Delta X) = k$

This transformation can be describe using the linear transformation (i.e. matrix) D:

$$D = \begin{bmatrix} -1 & 1 & 0 & 0 & \cdots & 0 \\ 0 & -1 & 1 & 0 & \cdots & 0 \\ 0 & 0 & -1 & 1 & & \\ \vdots & \vdots & & \ddots & \ddots & \\ 0 & 0 & & & -1 & 1 \end{bmatrix}$$

D is a $(T-1) \times T$ matrix. When applied to Y_i or X_i, it reduces the number of observations by 1. Note, we assume here that the data is sorted by t. Thus, DX_i is a $(T-1) \times k$ matrix of first-differences. The model can therefore be expressed as,

$$DY_i = DX_i\beta + D\varepsilon_i$$

Under SLPM 1-6, $\hat{\beta}^{FD}$ is consistent and asympotitcally normal. However, we need to account for the error term structure. This is because,

$$E[\Delta\varepsilon_{is}\Delta\varepsilon_{it}|X_i] = \begin{cases} E[(\varepsilon_{is} - \varepsilon_{is-1})(\varepsilon_{it} - \varepsilon_{it-1})|X_i] = 2\sigma_\varepsilon^2 & \text{for } s = t \\ E[(\varepsilon_{is} - \varepsilon_{is-1})(\varepsilon_{it} - \varepsilon_{it-1})|X_i] = -\sigma_\varepsilon^2 & \text{for } s = t-1 \\ 0 \quad \text{otherwise} \end{cases}$$

This is a MA(1) error term structure, in which the first-order correlation is non-zero. Using the linear transformation D, we can express this as,

$$E[D\varepsilon_i(D\varepsilon_i)'|X_i] = \sigma_\varepsilon^2 DD'$$

9.1 OLS vs GLS

The OLS estimator is given by,

$$\hat{\beta}^{FD} = \left(\sum_{i=1}^{n}(DX_i)'DX_i\right)^{-1}\sum_{i=1}^{n}(DX_i)'DY_i$$

For $T = 2$, you can show that,

$$\hat{\beta}^{FD} = \hat{\beta}^{WG}$$

As scene above, the error term is not homoskedastic, which makes this estimator inefficient. The efficient GLS estimator is given by,

$$\hat{\beta}^{GLS} = \arg\min_b \sum_{i=1}^{n}(DY_i - DX_i b)'(\sigma_\varepsilon^2 DD')^{-1}(DY_i - DX_i b)$$

Since the scalar σ_ε^2 can be ignored, the GLS solution is given by,

$$\hat{\beta}^{GLS} = \left(\sum_{i=1}^{n} X_i' D'(DD')^{-1}DX_i\right)^{-1}\sum_{i=1}^{n} X_i' D'(DD')^{-1}DY_i$$

It turns out that,

$$D'(DD')^{-1}D = M_\ell$$

This result holds for $T \geq 2$. The GLS estimator for first differences is equivalent to the Within-Group estimator.

10 Wu-Hausman Test

The Wu-Hausman test is used to test the exogeneity assumption underlying a particular estimator. You need two estimators: $\{\hat{\beta}_1, \hat{\beta}_2\}$ such that,

- Under $H_0 : \beta_1 = \beta_2$
- $\hat{\beta}_1$ is consistent
- $\hat{\beta}_2$ is consistent
- $Var(\hat{\beta}_1|X) < Var(\hat{\beta}_2|X)$: the former is more efficient
- Under $H_1 : \beta_1 \neq \beta_2$
- $\hat{\beta}_1$ is inconsistent
- $\hat{\beta}_2$ is consistent

The test statistic is given by,

$$\text{Stat} = (\hat{\beta}_2 - \hat{\beta}_1)'(Var(\hat{\beta}_2 - \hat{\beta}_1|X))^{-1}(\hat{\beta}_2 - \hat{\beta}_1)$$

Under H_0, this statistic converges in distribution to χ_k, where k is the number of regressors. The inner matrix is the inverse of the variance-covariance matrix.

$$\begin{aligned} Var(\hat{\beta}_2 - \hat{\beta}_1|X) =& Var(\hat{\beta}_2|X) + Var(\hat{\beta}_1|X) - 2Cov(\hat{\beta}_2, \hat{\beta}_1|X) \\ =& Var(\hat{\beta}_2|X) - Var(\hat{\beta}_1|X) \end{aligned}$$

Line 2 follows from line 1, because of a result demonstrated by Hausman (1978):

$$0 = Cov(\hat{\beta}_1, \hat{\beta}_1 - \hat{\beta}_2) = Var(\hat{\beta}_1) - Cov(\hat{\beta}_1, \hat{\beta}_2)$$

This result only holds in cases where the variances of the respectively estimators can be ranked: i.e. one estimator is more efficient than the other. As a result, the $Var(\hat{\beta}_2 - \hat{\beta}_1|X)$ matrix is positive definite.

This test can be applied in this setting to test the null: $H_0 : E[X_{it}\alpha_i] = 0$ (i.e. uncorrelatedness). This condition must hold for the (F)GLS estimator to be consistent. If H_0 is false, then we know that the WG estimator is consistent, but that it is less efficient. To test this hypothesis we use the coefficients from the FGLS and WG estimators. Of course, this means that you can only test for restrictions on time-varying regressors (a restriction of WG estimator).

10.1 Mundlack correction

The Mundlack correction provides a way of including time-invariant variables in a WG estimator, and therefore also in a Hausman test. Consider, the within-group transformation gives us the model,

$$\tilde{Y}_i = \tilde{X}_i\beta + \tilde{\varepsilon}_i$$

which can be written as,

$$M_\ell Y_i = M_\ell X_i\beta + M_\ell\varepsilon_i$$

given that M_ℓ is the orthogonal projection matrix that demeans each variable. It turns out that the OLS estimator for β in the above expression is equivalent to the OLS estimator from the equation,

$$Y_i = X_i\beta + \bar{X}_i\ell\gamma + \varepsilon_i$$

Where $\bar{X}_i\ell = P_\ell X_i$, the (individual-specific) mean of each variable. Demeaning each variable is equivalent to controlling for the mean.

We can demonstrate this result using partitioned regression. In this new equation, there are two sets of regressors: X_i and $\bar{X}_i\ell$ (the unit-specific mean of each X). We know that the OLS estimator for β can be arrived at by first regressing X_i on $\bar{X}_i\ell$,

$$X_i = \bar{X}_i\ell\eta + \nu_i$$

and then regressing the residual from this equation on Y. The residual from this expression is given by the orthogonal projection of $\bar{X}_i\ell = P_\ell X_i$. In this case, tt is sufficient to consider the orthogonal projection for unit i,

$$\begin{aligned}
&(I_T - P_\ell X_i(X_i'P_\ell P_\ell X_i)^{-1}X_i'P_\ell)X_i \\
=&X_i - P_\ell X_i(X_i'P_\ell X_i)^{-1}X_i'P_\ell X_i \\
=&(I_T - P_\ell)X_i \\
=&M_\ell X_i
\end{aligned}$$

Thus, partitioned regression says that the Mundlack correction will get the same OLS estimator as the regression of,

$$Y_i = M_\ell X_i\beta + \epsilon_i$$

which is equivalent, in terms of its OLS projection to,

$$M_\ell Y_i = M_\ell X_i \beta + M_\ell \varepsilon_i$$

Why does this matter? It means that we can estimate a model of the form,

$$Y_{it} = X_{it}'\beta + W_i'\psi + \bar{X}_i\gamma + \epsilon_{it}$$

that includes time invariant regressors W_i, and gives us the same OLS estimates for β as $\widehat{\beta}^{WG}$. Since this model, with time invariant regressors, can also be estimated using FGLS (under the assumption $E[\alpha_i|X_i] = 0$), we can conduct a Hausman test that compares the estimates of both β *and* ψ.

References

Cameron, A Colin, and Pravin K Trivedi. 2005. *Microeconometrics: Methods and Applications*. Cambridge university press.

Verbeek, Marno. 2017. *A Guide to Modern Econometrics*. John Wiley & Sons.

Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data*. MIT press.