Econometrics 6027
Lecture 7
Panel Data

Panel data sets collect information from many individuals at several different points in time. The same economic agents have been followed throughout the period of the sample. That is, with panel data we have:

$$y_{it} = \alpha + x'_{it}\beta + u_{it}$$

while for time series we had:

$$y_t = \alpha + x'_t\beta + u_t$$

Examples include:

- firm data
- patterns of individual behaviour over their life cycle
- macroeconomic trade data on multiple countries since World War II.

Panel data is advantageous compared to repeated cross-sections because:

- more data means more powerful regressions which yield increased precision of regression estimates
- the ability to control for individual fixed effects
- can capture dynamics by allowing for correlations across time without inducing aggregation bias
- can provide information to match groups for instrumental variable approaches
- can clarify the processes underlying the data: for example, whether if upon observing 50 % of women working, we have 50% of women working on average each period, or the same 50% of women working each period.

In panel data sets, the sample of individuals N is typically relatively large and the number of time periods T is generally short.

There are difficulties inherent in data sources with a longitudinal element. Data-gathering for panel data sets is expensive since you need to collect data from individuals over time. The sample is non-random. There may be attrition: though we need commitment from individuals over time, people may move or exit the sample for other reasons.

In determining the impact of regressors on the dependent variable, we have to take into account that particular individuals, or time periods, may have their own distinctive impact on the dependent variable. This is called individual

heterogeneity. As a result, we can see intercepts that are individual specific, or time specific. There are five different ways to model panel data linearly, as shown below: we will examine models marked with an (*).

$$y_{it} = \alpha + x'_{it}\beta + u_{it} \begin{cases} \alpha + x'_{it}\beta + u_{it}, & \text{1+K parameters (*)} \\ & \text{where the error may be correlated with the} \\ & \text{heterogeneity of the individual:} \\ & u_{it} = f_i + \epsilon_{it} \\ \alpha_i + x'_{it}\beta + u_{it} & \text{N+K parameters} \\ \alpha_t + x'_{it}\beta + u_{it} & \text{T+K parameters} \\ \alpha_i + \alpha_t + x'_{it}\beta + u_{it} & \text{N+T+K parameters (*)} \\ \alpha_i + \alpha_t + x'_{it}\beta_t + u_{it} & \text{T(1+K)+N parameters (*)} \end{cases}$$

Where there are K x's, N individuals, and T time periods. u_{it} is individual heterogeneity, which arises in panel data, but not in time series data. The last specification in the list is the most robust to misspecification, but the first one is the most efficient. However, if there is individual heterogeneity, the first specification won't be efficient.

Nonlinear panel models, and how to estimate dynamics in panels, are explored in a separate course. Examples include:

$$y_{it} = F(\alpha_1 + x'_i\beta) + u_i$$

$$y_t = \alpha_0 + \rho y_{t-1} + x\beta + u$$

The rest of this lecture will look at two categories of models that estimate linear panels. The first, fixed effects models, accounts for heterogeneity by introducing a "fixed" effect that acts like a parametric shift of the regression for particular individuals or time periods. The second, random effects models, assumes that individual effects on y are random draws distributed across the sample which are uncorrelated with the regressor.

1 Fixed Effects Models

Fixed effects models introduce individual and/or time-specific variables into the regression. They are formulated as variations of the following specification:

$$\begin{aligned} & y_{it} = f_i + f_t + x'_{it}\beta + u_{it} \\ \text{Assumptions: } & E(u_{it}|x_{it}) = 0 \\ & Var(u_{it}) = \sigma^2 \\ & E(u_{it}, u_{is}) = 0 \text{ where } t \neq s \\ & u_{it} \sim iid \end{aligned}$$

for $i = 1, .., N$ individuals over $t = 1, .., T$ time periods. The model could include an individual effect f_i (constant over time); a time effect f_t (constant over individuals); and marginal effects β for x_{it} (common across i and t).

1.1 Pooled OLS (POLS)

Pooled Ordinary Least Squares is the simplest approach to estimation. Individual effects f_i are fixed and common across economic agents such that $f_i = f$ for all $i = 1, .., N$. If the assumption $f_i = f$ holds, OLS produces consistent and efficient estimates of f and β.

$$y = f + x_{it}\beta + u_{it}$$

We estimate, and eliminate fixed effects, by differencing against population means $\overline{x}$ and $\overline{y}$. Define:

$$\overline{y} = \frac{1}{NT}\sum_i \sum_t y_{it}$$
$$\overline{x} = \frac{1}{NT}\sum_i \sum_t x_{it}$$

and

$$\widetilde{x}_{it} = x_{it} - \overline{x}$$
$$\widetilde{y}_{it} = y_{it} - \overline{y}$$

Then the Pooled OLS estimator is:

$$\widehat{f} = \overline{y} - \widehat{\beta}\overline{x}$$
$$\widehat{\beta}_{POLS} = \frac{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\widetilde{x}_{it}\widetilde{y}_{it}}{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\widetilde{x}_{it}^2}$$
$$= \frac{\frac{1}{NT}\sum_i\sum_t(y_{it} - \overline{y})(x_{it} - \overline{x})}{\frac{1}{NT}\sum_i\sum_t(x_t - \overline{x})^2}$$

The variance of the POLS estimator is:

$$var\widehat{\beta} = var\Big(\frac{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\widetilde{x}_{it}\widetilde{y}_{it}}{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\widetilde{x}_{it}^2}\Big)$$

but

$$\overline{y} = \frac{1}{NT}\sum_{i}\sum_{t} y_{it} = \alpha + \beta\overline{x} + \overline{u}$$

so

$$\begin{aligned}\widetilde{y_{it}} = y_{it} - \overline{y} &= \alpha + \beta x_t + u_{it} - \alpha - \beta\overline{x} - \overline{u} \\ &= \beta\widetilde{x_{it}} + \widetilde{u_{it}}\end{aligned}$$

Thus,

$$var\widehat{\beta} = var\Big(\frac{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\widetilde{x}_{it}(\beta\widetilde{x_{it}} + \widetilde{u_{it}})}{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\widetilde{x}_{it}^2}\Big)$$

$$\begin{aligned}var\widehat{\beta}_{POLS} &= var\Big(\frac{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\widetilde{x}_{it}^2\beta + \widetilde{x}_{it}\widetilde{u_{it}})}{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\widetilde{x}_{it}^2}\Big) \\ &= var\Big(\beta + \frac{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\widetilde{x}_{it}\widetilde{u_{it}})}{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\widetilde{x}_{it}^2}\Big)\end{aligned}$$

since $var(\beta) = 0$

$$\begin{aligned}&= \Big(\frac{\sum\sum x_{it}^2 var(\widetilde{u_{it}})}{(\sum\sum \widetilde{x_{it}}^2)^2}\Big) \\ &= \Big(\frac{var(\widetilde{u_{it}})}{\sum\sum \widetilde{x_{it}}^2}\Big) \\ &= \Big(\frac{\sigma^2}{\sum\sum \widetilde{x_{it}}^2}\Big)\end{aligned}$$

if x and y are not correlated. If they are correlated,

$$\begin{aligned}var(\widetilde{u_{it}}) &= var(x) + var(y) - 2\mathrm{cov}(x,y) \\ &= \sigma^2 + \frac{\sigma^2}{NT} - 2\sigma^2 NT \\ &= \sigma^2\Big(\frac{NT-1}{NT}\Big)\end{aligned}$$

thus,

$$var(\hat{\beta}_{POLS}) = \frac{\sigma^2(1-\frac{1}{NT})}{\sum\sum \widetilde{x_{it}}^2}$$

1.2 Least Squares Dummy Variables

With Least Squares Dummy Variables (LSDV), fixed effects are specific to individuals and represented by a dummy variable f_i for each individual.

$$y_{it} = f_i + x'_{it}\beta + u_{it}$$

To estimate, define a series of individual dummy variables d_i whose coefficients, f_i, represent individual fixed effects:

$$y_{it} = f_1 d_{1t} + f_2 d_{2t} + ... + f_N d_{Nt} + x'_{it}\beta + u_{it}$$

Estimate by standard OLS yielding $\widehat{\beta}_{LSDV}$.

1.2.1 F-test: Pooled vs. Fixed LSDV

If we don't have individual heterogeneity, we don't need to include the dummies in our regression, so the Pooled OLS is the appropriate model. We test this via an F-test where the null is POLS.

$$\begin{aligned} H_0 &: POLS : f_1 = f_2 = ... = f_N \\ H_1 &: LSDV : f_i \neq f_j \\ \mathbf{R}^2_R &: Y_{it} = f + X_{it}\beta + u_{it} \\ \mathbf{R}^2_U &: Y_{it} = f_1 D_{1t} + f_2 D_{2t} + ...f_N D_{NT} + x_{it}\beta + u_{it} \end{aligned}$$

F-test:

$$F = \frac{\mathbf{R}^2_U - \mathbf{R}^2_R}{1 - \mathbf{R}^2_U} * \frac{NT - N - K}{N - 1}$$

or

$$F = \frac{(\mathbf{R}^2_U - \mathbf{R}^2_R)/(N-1)}{(1 - \mathbf{R}^2_U)/(NT - N)}$$

Which is distributed $F_{N-1,NT-N-k}$ under the null of equality of f_i. If $F_{test} > F_{N-1,NT-(N+K)}$ reject Pooled OLS.

1.2.2 The Two-Way Fixed Effects Model

This model is like LSDV, but you also include fixed effects γ_t that are specific to each time period.

$$y_{it} = f_i + \gamma_t + x'_{it}\beta + u_{it}$$

To estimate, define a series of time dummy variables $z_{sit} = 1(s = t)$ in addition to our individual dummy variables variables $d_{git} = 1(g = i)$ to give:

$$y_{it} = f_1 d_{1it} + f_2 d_{2it} + ... + f_N d_{Nit} + g_2 z_{2it} + ... + g_T z_{Tit} + x'_{it}\beta + u_{it}.$$

We need only $t-1$ time dummies. We have to drop one of the time dummies to prevent a situation of perfect collinearity among the dummy variables (note that $\sum_{s=1}^{T} z_{sit} = \sum_{g=1}^{N} d_{git}$). The problem with this model is that since we need to estimate a large number of parameters, we lose degrees of freedom (NT-N-T-K).

1.3 Within-Group Estimator

The Fixed Effects Within Group Estimator has individual-specific fixed effects, as with LSDV, but they are eliminated by differencing against the individual average. Define:

$$\begin{aligned}
\overline{y}_i &= \frac{1}{T}\sum_{t=1}^{T} y_{it} \\
&= f_i + \beta\overline{x_i} + \overline{u_i} \\
\overline{x}_i &= \frac{1}{T}\sum_{t=1}^{T} x_{it} \\
\overline{u}_i &= \frac{1}{T}\sum_{t=1}^{T} u_{it}
\end{aligned}$$

and

$$\begin{aligned}
x^*_{it} &= x_{it} - \overline{x}_i \\
y^*_{it} &= y_{it} - \overline{y}_i \\
u^*_{it} &= u_{it} - \overline{u}_i
\end{aligned}$$

We have:

$$\begin{aligned}
y^*_{it} &= y_{it} - \overline{y}_i = (f_i - f_i) + (x_{it} - \overline{x}_i)'\beta + (u_{it} - \overline{u}_i) \\
&= x^*_{it}\beta + u^*_{it}.
\end{aligned}$$

Hence, the Within-Groups estimator is:

$$\widehat{\beta}_{WG} = \frac{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} x^*_{it}y^*_{it}}{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} x^{*2}_{it}}$$

1.3.1 Variance of Within Groups estimator

$$\begin{aligned}
S_{xx} &= \sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it} - \overline{x})^2, \\
S^w_{xx} &= \sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it} - \overline{x}_i)^2 \\
S^b_{xx} &= \sum_{i=1}^{N} T(\overline{x}_i - \overline{x})^2
\end{aligned}$$

You can show that

$$S_{xx} = S^w_{xx} + S^b_{xx}.$$

Given that $var(u_{it}^*) = \left(\frac{T-1}{T}\right) var(u_{it})$, we have

$$var(\widehat{\beta}^{WG}) = \frac{var\left(u_{it}^*\right)}{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}^{*2}} = \frac{\left(\frac{T-1}{T}\right) var(u_{it})}{S_{xx} - S_{xx}^b} > var\widehat{\beta}.$$

1.3.2 Analysis of the Within-Groups estimator

- β_{WG} should be equal to β_{LSDV}, though they are estimated differently. β_{LSDV} will have a larger variance
- A major flaw of this within-groups model is that it eliminates (ie fails to estimate) any regressors that are constant over time. If there are such time-invariant regressors, we can either use the two-way model (and lose degrees of freedom) or use the following procedure (so long as there's no correlation between x_i and u_i).

2-step procedure to recover time-invariant regressors from a within group regression:

1. Run your within-group OLS regression and save your residuals $\hat{u_{it}}^*$
2. Run a second regression of $\hat{u_{it}}^*$ on your regressors, including the time-invariant ones. The coefficients on the time-invariant variables in this second regression are your estimates.

1.4 Example: A Simple Dynamic Linear Model with fixed effects

Suppose we observe N individuals for T periods. Consider the dynamic linear model with fixed effects,

$$y_{it} = \rho y_{i,t-1} + f_i + \varepsilon_{it} \tag{1}$$

where $E(\varepsilon_{it}|y_{i1},...,y_{it}) = 0$, $E(\varepsilon_{it}^2|y_{i1},...,y_{it}) = \sigma^2$, $E\left(\varepsilon_{it}\varepsilon_{is}|y_{i1},...,y_{i,t-1}\right) = 0$, for $s \neq t$, $s = 1,...,T$, and $i = 1,...N$.
Consider the following likelihood function

$$L^i(\rho, f_i) = \sum_t L^{it}(\rho, f_i),$$

where f_i is a nuisance parameter that controls for heterogeneity. To estimate the parameter of interest ρ, we need that f_i to be orthogonal to ρ. If this is the case, we consider elimination of nuisance parameters by integration.

Cox and Reid (1987) and Jeffreys (1961) use this concept and refer to it as 'orthogonality'. We prefer the term information-orthogonality to distinguish it from the other orthogonality concepts and to stress that it is defined in terms of

the properties of the information matrix. See Tibshirani and Wasserman (1994) and Woutersen (2000) for an overview of orthogonality concepts.

Given that in most of the data f_i is not orthogonal to ρ, we use a reparametrization of the fixed effect given by Lancaster (2002) . The information-orthogonal parametrization is conditioned on y_{i0},is explicit and is given by

$$f_i = (1-\rho)y_{i0} + \lambda_i e^{-b(\rho)} \text{ where } b(\rho) = \frac{1}{T}\sum_{t=1}^{T}\frac{T-t}{t}\rho^t \text{ and } \frac{1}{N}\sum_{i=1}^{N}|y_{i0}|^2 = O(1).$$

The model (**??**) becomes

$$y_{it} = \rho y_{i,t-1} + (1-\rho)y_{i0} + \lambda_i e^{-b(\rho)} + \varepsilon_{it},$$

where the fixed effect λ_i is a nuisance parameter that controls for heterogeneity and that is now orthogonal to ρ. Now, let the log likelihood contribution of the t^{th} period of individual i be denoted by L^{it}. Summing over the contribution of individual i yields the log likelihood contribution,

$$L^i(\rho, \lambda_i) = \sum_t L^{it}(\rho, \lambda_i).$$

To estimate the parameter of interest ρ, we consider elimination of nuisance parameters λ_i by integration. We denote the logarithm of the integrated likelihood contribution by $L^{i,I}$, i.e.

$$L^{i,I}(\rho) = \ln \int e^{L^i} d\lambda_i.$$

Summing over i yields the logarithm of the integrated likelihood,

$$L^I(\rho) = \sum_i L^{i,I}(\rho) = \sum_i \ln \int e^{L^i} d\lambda_i. \tag{2}$$

Analogue to quasi-maximum likelihood estimators, normality of the error terms is assumed in order to derive the integrated likelihood estimator. The estimator depends only on the first two moments of y_{it} and is given by Lancaster (2002). Integrating the likelihood contribution of individual i with respect to λ gives $e^{L^{i,I}}$, where:

$$e^{L^{i,I}} \propto \frac{1}{\sigma^{T-1}} e^{b(\rho) - \frac{1}{2\sigma^2}\sum_t (y_{it}-\rho y_{i,t-1})^2 + \frac{T}{2}(\overline{y_{it}-\rho y_{i,t-1}})^2}$$

or the total integrated likelihood

$$L^I = -N(T-1)\ln(\sigma) + b(\rho) - \frac{1}{2\sigma^2}\sum_t (y_{it}-\rho y_{i,t-1})^2 + \frac{T}{2}(\overline{y_{it}-\rho y_{i,t-1}})^2.$$

After integrating out the fixed effects, the mode of the integrated likelihood can be used as an estimator[1]. Let $\hat{\rho}_I$ be the *integrated likelihood estimator.* Then,

$$\hat{\rho}_I = \arg\max_{\rho} L^I(\rho).$$

[1] As $N \to \infty$, using the marginal posteriors is asymptotically equivalent. Considering the mode of the posterior, however, simplifies the algebra.

2 Random Effects Models

The fixed effects model is appropriate when differences between individual agents may reasonably be viewed simply as parametric shifts in the regression function itself. This might be considered reasonable if the cross-sectional used in estimation represents a broadly exhaustive sample of the population of economic agents, as might be the case in a study which covers a full sample of countries, or in a study of the performance of firms in a particular industry, where the sample of firms represents a broadly complete coverage of those within the industry.

If, on the other hand, the cross-section is drawn from a larger population (so that the sample of cross-sectional agents may not reasonably be considered exhaustive) then it may be more appropriate to view the individual-specific terms in the sample as randomly distributed effects across the full cross-section of agents. Defining the fixed effect as τ_i, we have $\tau_i = \alpha + f_i$, where f_i has a zero (unconditional) mean. This would suggest a random effects specification of the following form:

$$y_{it} = \alpha + x'_{it}\beta + f_i + e_{it}$$

where the error term has two components: $u_{it} = f_i + e_{it}$. Here, f_i represents an individual disturbance which is fixed over time. The individual heterogeneity is in the error term, which means that we have fewer parameters to estimate and we don't eliminate any time-invariant errors. The following assumptions relate to the random components in the model:

$$\begin{aligned}
&(1) E\left(e_{it}|f_i\right) = 0 \\
&(2) E\left(e_{it}^2|f_i\right) = \sigma_e^2 \\
&(3) E\left(f_i|x_{it}\right) = 0 \text{ for all } i,t \\
&(4) E\left(f_i^2|x_{it}\right) = \sigma_f^2 \\
&(5) E\left(e_{it}f_j\right) = 0 \text{ for all } i,t,j \\
&(6) E\left(e_{it}u_{js}\right) = 0 \text{ for } i \neq j \text{ or } t \neq s \\
&(7) E\left(f_i f_j\right) = 0 \text{ for } i \neq j
\end{aligned}$$

Among these assumptions, number (3) is perhaps the most restrictive. It states that the expectation of f_i conditional on the value of x_{it} is assumed to be zero for the simple random effects model. If the individual fixed effects are correlated with the explanatory variables, assumption (3) will be violated, the random effects model is unusable, and we must use the fixed effects model since otherwise the estimates will be biased.

On the other hand, if the individual heterogeneity is random, then assumption (3) holds and we can use random effects models. This is preferable since RE models are more robust to mis-specification, retain time-invariant variables, and retain the maximum degrees of freedom. We will investigate a test for this assumption later in the lecture.

2.1 The Generalized Least Squares (GLS) estimator

To estimate the linear random effects model (sometimes called the variance components or random components model) requires a Generalized Least Squares approach to deal with heteroskedasticity in the errors. To see this, consider the characteristics of the combined error term $u_{it} = f_i + e_{it}$. It is certainly true that $E(u_{it}) = 0$. However,

$$\begin{aligned} E(u_{it}^2) &= \sigma_e^2 + \sigma_f^2 \text{ for all } i, t \\ E(u_{it}u_{is}) &= \sigma_f^2 \text{ for all } t \neq s \text{; time difference} \\ E(u_{it}u_{jt}) &= 0 \text{ for } i \neq j \text{; individual difference.} \end{aligned}$$

So, if we collect the T disturbances for individual i in a vector of the form $u_i = (u_{i1}, u_{i2}, ..., u_{iT})'$, we have that

$$E(u_i u_i') = \Omega = \begin{pmatrix} (\sigma_e^2 + \sigma_f^2) & \sigma_f^2 & \sigma_f^2 & \dots & \sigma_f^2 \\ \sigma_f^2 & (\sigma_e^2 + \sigma_f^2) & \sigma_f^2 & \dots & \sigma_f^2 \\ \sigma_f^2 & \sigma_f^2 & (\sigma_e^2 + \sigma_f^2) & \dots & \sigma_f^2 \\ \vdots & \dots & \dots & \ddots & \vdots \\ \sigma_f^2 & \sigma_f^2 & \sigma_f^2 & \dots & (\sigma_e^2 + \sigma_f^2) \end{pmatrix}$$

This is the covariance matrix for one individual. Its dimensions are T by T. Compile these into one big covariance matrix for the full panel of observations in the whole sample. The covariance matrix of the NT vector of disturbances $u = (u_1, u_2, ..., u_N)'$ may be derived as

$$\begin{aligned} P_{(NTxNT)} = E(uu') &= \begin{pmatrix} \Omega & 0 & 0 & \dots & 0 \\ 0 & \Omega & 0 & \dots & 0 \\ 0 & 0 & \Omega & \dots & 0 \\ \vdots & \dots & \dots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \Omega \end{pmatrix} \\ &= I_N \otimes \Omega \end{aligned}$$

where I_N is the identity matrix of dimension N and $\otimes$ represents the Kronecker product. This is the covariance matrix for the population: its dimensions are NT by NT.

As we can see, the covariance matrix P is heteroskedastic. To estimate random effects, we must thus use a Generalized Least Squares method to weigh the data to remove heteroskedasticity. We weigh the data using the square root of the covariance matrix:

$$\begin{aligned} P^{1/2}Y_{it} &= P^{-1/2}\beta X_{it} - P^{-1/2}u_{it} \\ Y^* &= X^*\beta + u^* \end{aligned}$$

The estimator is:

$$\begin{aligned}\beta_{RE} &= \hat{\beta_{GLS}} \\ &= (X'P^{-1}X)^{-1}X'PY\end{aligned}$$

The variance of this will be an identity matrix I_{NTxNT} without any heteroskedasticity:

$$\begin{aligned}E(u^*u^{*\prime}) &= E(P^{-1/2}uu'P^{-1/2}) \\ &= P^{-1/2}E(uu')P^{-1/2} \\ &= P^{-1/2}PP^{-1/2} \\ &= P^{-1/2}P^{1/2}P^{1/2}P^{-1/2} \\ &= I_{NT}\end{aligned}$$

This can also be expressed as follows. Recall:

$$\Omega = \begin{pmatrix} (\sigma_e^2+\sigma_f^2) & \sigma_f^2 & \sigma_f^2 & \dots & \sigma_f^2 \\ \sigma_f^2 & (\sigma_e^2+\sigma_f^2) & \sigma_f^2 & \dots & \sigma_f^2 \\ \sigma_f^2 & \sigma_f^2 & (\sigma_e^2+\sigma_f^2) & \dots & \sigma_f^2 \\ \vdots & \dots & \dots & \ddots & \vdots \\ \sigma_f^2 & \sigma_f^2 & \sigma_f^2 & \dots & (\sigma_e^2+\sigma_f^2) \end{pmatrix}$$

and define:

$$\theta = 1 - \frac{\sigma_\epsilon}{T(\sigma_\epsilon^2+\sigma_f^2)^{1/2}}$$

such that

$$\Omega^{-1/2} = I_T - \frac{\theta}{T}$$

Next, recall:

$$\begin{aligned} P &= \begin{pmatrix} \Omega & 0 & 0 & \dots & 0 \\ 0 & \Omega & 0 & \dots & 0 \\ 0 & 0 & \Omega & \dots & 0 \\ \vdots & \dots & \dots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \Omega \end{pmatrix} \\ &= I_N \otimes \Omega \\ P^{-1/2} &= I_N \otimes \Omega^{-1/2} \end{aligned}$$

substituting for $\Omega^{-1/2}$, we have a new definition of our projection matrix:

$$P^{-1/2} = I_N \otimes (I_T - \frac{\theta}{T})$$

since we know

$$Y^* = P^{-1/2}\beta X_{it} - P^{-1/2}u_{it}$$

we can substitute in our new definition of P:

$$Y^* = (I_N \otimes (I_T - \frac{\theta}{T}))\beta X_{it} - (I_N \otimes (I_T - \frac{\theta}{T}))u_{it}$$

so that:

$$y^* = \Omega^{-1/2} y_i = \begin{pmatrix} Y_{i1} & -\theta\overline{Y_i} \\ Y_{i2} & -\theta\overline{Y_i} \\ \vdots & \\ Y_{it} & -\theta\overline{Y_i} \end{pmatrix}$$

2.2 The Breusch-Pagan (Lagrange Multiplier) test: Pooled vs Random

The Breusch-Pagan LM statistic provides a test of the random effects (RE) model against the pooled OLS model. The specific hypothesis under investigation is the following:

$$H_0 : POLS : \sigma_f^2 = 0$$
$$H_1 : RE : \sigma_f^2 \neq 0$$

Under the null ($\sigma_f^2 = 0$), $P = \sigma_f^2 I_{NT}$, so that the random effects model reduces to a pooled OLS regression. The test of this hypothesis, based on OLS residuals $\widehat{u}_{it}$ from the pooled regression model, requires use of the LM statistic

$$LM = \frac{NT}{2(T-1)} \left[\frac{\sum_{i=1}^{N} (\sum_{i=1}^{T} \widehat{u}_{it})^2}{\sum_{i=1}^{N} \sum_{i=1}^{T} \widehat{u}_{it}^2} - 1 \right]^2 \sim \chi_1^2$$

Under the null, this statistic should be distributed as a χ_1^2 with one degree of freedom because we are testing one parameter (σ_f^2).

2.3 The Hausman (Wald) Test: Random vs Fixed

We may be interested in comparing directly the random effects estimator $\widehat{\beta}_{RE}$ with the fixed effects estimator $\widehat{\beta}_{FE}$. As noted above, in the presence of correlation between the regressors x_{it} and individual effects f_i the GLS estimator $\widehat{\beta}_{RE}$ is inconsistent, whilst the OLS estimate $\widehat{\beta}_{FE}$ is consistent, so we should use a fixed effects model. If on the other hand $E = (f_i | x_{it}) = 0$, then the RE estimator is consistent and efficient whilst the OLS estimator is consistent but inefficient, so we should use a random effects model.

This motivated Hausman (1978) to construct a test based on the difference between $\widehat{\beta}_{RE}$ and $\widehat{\beta}_{FE}$. He noted that in the presence of correlation between the regressors x_i and individual effects f_i, $\hat{\beta}_{RE}$ is inconsistent and $\hat{\beta}_{FE}$ is consistent, so the β's will be different. His test evaluates whether the βs are the same. If the null hypothesis is rejected, then we conclude that the β's are significantly

different, that there is correlation between the regressors and the individual effects, and so we should use a fixed effects model such as LSDV.

$$H_0 : RE : cov(f_i, x_i) = 0$$
$$H_1 : FE : cov(f_i, x_i) \neq 0$$

Hausman noted that under the null the variance of the difference $\widehat{\beta}_{RE} - \widehat{\beta}_{FE}$ may be derived as:

$$var(\widehat{\beta}_{RE} - \widehat{\beta}_{FE}) = var(\widehat{\beta}_{RE}) + var(\widehat{\beta}_{FE}) - 2cov(\widehat{\beta}_{RE}, \widehat{\beta}_{FE})$$

But the covariance is (adding and subtracting β_{FE}):

$$\begin{aligned} cov(\widehat{\beta}_{RE}, \widehat{\beta}_{FE}) &= cov(\widehat{\beta}_{RE} - \widehat{\beta}_{FE} + \widehat{\beta}_{FE}, \widehat{\beta}_{FE}) \\ &= cov(\widehat{\beta}_{RE} - \widehat{\beta}_{FE}, \widehat{\beta}_{FE}) + cov(\widehat{\beta}_{FE}, \widehat{\beta}_{FE}) \\ &= cov(\widehat{\beta}_{RE} - \widehat{\beta}_{FE}, \widehat{\beta}_{FE}) + var(\widehat{\beta}_{FE}) \end{aligned}$$

but under null, $\widehat{\beta}_{RE} - \widehat{\beta}_{FE} = 0$

Thus,

$$cov(\widehat{\beta}_{RE} - \widehat{\beta}_{FE}, \widehat{\beta}_{FE}) = var(\widehat{\beta}_{FE})$$

And so

$$\begin{aligned} var(\widehat{\beta}_{RE} - \widehat{\beta}_{FE}) &= var(\widehat{\beta}_{RE}) + var(\widehat{\beta}_{FE}) - 2(var(\widehat{\beta}_{FE})) \\ &= var(\widehat{\beta}_{RE}) - var(\widehat{\beta}_{FE}) \end{aligned}$$

The Hausman test is then:

$$W = (\hat{\beta}_{RE} - \hat{\beta}_{FE})' Var(\hat{\beta}_{RE} - \hat{\beta}_{FE})^{-1} (\hat{\beta}_{RE} - \hat{\beta}_{FE}) \sim \chi^2_K$$

which is distributed as a chi-squared with K degrees of freedom under the null, K being the number of regressors in x_{it}.