Fixed Effects vs Random Effects Regression

The Econometric Model:

Consider the multiple linear regression model for individual i = 1,...,N who is observed at several time periods t = 1,...,T

$$y_{it} = \alpha + x'_{it}\beta + z'_i\gamma + c_i + u_{it}$$

where y_{it} is the dependent variable, x'_{it} is a K-dimensional row vector of time-varying explanatory variables and z'_i is a M-dimensional row vector of time-invariant explanatory variables excluding the constant, α is the intercept, β is a K-dimensional column vector of parameters, γ is a M-dimensional column vector of parameters, c_i is an individual-specific effect and u_{it} is an idiosyncratic error term.

The data generation process (dgp) is described by:

PL1: Linearity

$y_{it} = \alpha + x'_{it}\beta + z'_i\gamma + c_i + u_{it}$ where $E[u_{it}] = 0$ and $E[c_i] = 0$

The model is linear in parameters α, β, γ, effect c_i and error u_{it}.

PL2: Independence

$\{X_i, z_i, y_i\}_{i=1}^{N}$ i.i.d. (independent and identically distributed)

The observations are independent across individuals but not necessarily across time. This is guaranteed by random sampling of individuals.

PL3: Strict Exogeneity $E[u_{it}|X_i, z_i, c_i] = 0$ (mean independent)

The idiosyncratic error term u_{it} is assumed uncorrelated with the explanatory variables of all past, current and future time periods of the same individual. This is a strong assumption which e.g. rules out lagged dependent variables. PL3 also assumes that the idiosyncratic error is uncorrelated with the individual specific effect.

PL4: Error Variance

a) $V[u_i|X_i, z_i, c_i] = \sigma_u^2 I, \sigma_u^2 > 0$ and finite

(homoscedastic and no serial correlation)

b) $V[u_{it}|X_i, z_i, c_i] = \sigma_{u_{it}}^2 > 0$, finite and

$Cov[u_{it}, u_{is}|X_i, z_i, c_i] = 0 \; \forall \; s \neq t$ (no serial correlation)

c) $V[u_i|X_i, z_i, c_i] = \Omega_{u_i}(X_i, z_i)$ is positive definite and finite

The remaining assumptions are divided into two sets of assumptions: the random effects model and the fixed effects model.

The Random Effects Model:

Random effects analysis imposes more assumptions than those needed for pooled OLS: strict exogeneity in addition to orthogonality between c_i and $\boldsymbol{x_{it}}$, that is, $E(u_{it}|\boldsymbol{x_i}, c_i) = 0$, $t = 1, \dots, T$ and $E(c_i|\boldsymbol{x_i}) = E(c_i) = 0$ where $\boldsymbol{x_i} \equiv (\boldsymbol{x_{i1}}, \boldsymbol{x_{i2}}, \dots, \boldsymbol{x_{iT}})$. In a random effects model, the unobserved variables are assumed to be uncorrelated with (or, more strongly, statistically independent of) all the observed variables. That assumption will often be wrong but for the reasons given above (e.g. standard errors may be very high with fixed effects, RE lets you estimate effects for time-invariant variables), an RE model may still be desirable under some circumstances. RE models can de estimated via Generalized Least Squares (GLS).

In the random effects model, the individual-specific effect is a random variable that is uncorrelated with the explanatory variables.

RE1: Unrelated effects

$E[c_i|\mathrm{X}_i, z_i] = 0$

RE1 assumes that the individual-specific effect is a random variable that is uncorrelated with the explanatory variables of all past, current and future time periods of the same individual.

RE2: Effect Variance

a) $V[c_i|\mathrm{X}_i, z_i] = \sigma_c^2 < \infty$ (homoscedastic)
b) $V[c_i|\mathrm{X}_i, z_i] = \sigma_{c_i}^2(\mathrm{X}_i, z_i) < \infty$ (heteroscedastic)

RE2a assumes constant variance of the individual specific effect.

RE3: Identifiability

a) $rank\ (W) = K + M + 1 < NT\ and\ E\left[\acute{W}_\iota W_i\right] = Q_{WW}\ is\ p.d\ and\ finite.$ The typical element $\acute{w}_{\iota t} = [1\ \acute{x}_{\iota t}\acute{z}_\iota]$.
b) $rank\ (W) = K + M + 1 < NT\ and\ E\left[\acute{W}_\iota \Omega_{v_i}^{-1} W_i\right] = Q_{WOW}\ is\ p.d\ and\ finite.$ Ω_{v_i} is defined below.

RE3 assumes that the regressors including a constant are not perfectly collinear, that all regressors (but the constant) have non-zero variance and too many extreme values.

Random Effects Model *(Greene's notation)***:**

The fixed effects model allows the unobserved individual effects to be correlated with the included variables. If the individual effects are strictly uncorrelated with the regressors, then it might be appropriate to model the individual specific constant terms as randomly distributed across cross-sectional units. This view would be appropriate if we believed that the sampled cross-sectional units are drawn from a large population.

The basic framework for this discussion is a regression model of the form

$$y_{it} = \acute{\boldsymbol{x}}_{\iota t}\beta + \acute{\boldsymbol{z}}_\iota \alpha + \varepsilon_{it}.$$

There are *K* regressors in $\boldsymbol{x}_{it}$, *not including a constant term*. The **heterogeneity,** or **individual effect** is $\boldsymbol{z}_i'\alpha$ where $\boldsymbol{z}_i$ contains a constant term and a set of individual or group specific variables, which may be observed, such as race, sex, location, and so on or unobserved, such as family specific characteristics, individual heterogeneity in skill or preferences, and so on, all of which are taken to be constant over time *t*. As it stands, this model is a classical regression model. If $\boldsymbol{z}_i$ is observed for all individuals, then the entire model can be treated as an ordinary linear model and fit by least squares. If $\boldsymbol{z}_i$ contains only a constant term, then the ***Pooled Regression*** ordinary least squares provides consistent and efficient estimates of the common α and the slope vector $\boldsymbol{\beta}$.

If the unobserved individual **heterogeneity**, however formulated, can be assumed to be uncorrelated with the included variables, then the model may be formulated as[1]

$$y_{it} = \acute{\boldsymbol{x}}_{\iota t}\boldsymbol{\beta} + E[\acute{\boldsymbol{z}}_\iota \boldsymbol{\alpha}] + \{\acute{\boldsymbol{z}}_\iota \boldsymbol{\alpha} - E[\acute{\boldsymbol{z}}_\iota \boldsymbol{\alpha}]\} + \varepsilon_{it}$$

$$y_{it} = \acute{\boldsymbol{x}}_{\iota t}\boldsymbol{\beta} + (\alpha + u_i) + \varepsilon_{it}$$

[1] There are *K* regressors in $\boldsymbol{x}_{it}$, not including s constant term. The **heterogeneity**, or **individual effect** is $\acute{\boldsymbol{z}}_i\boldsymbol{\alpha}$ where $\boldsymbol{z}_i$ contains a constant term and a set of individual or group specific variables, which may be observed or unobserved, all of which are taken to be constant over time *t*.

that is, as a linear regression model with a compound disturbance that may consistently, albeit inefficiently, estimated by least squares. This random effects approach specifies that u_i is a group specific random element, similar to ε_{it} except that for each group, there is but a single draw that enters the regression identically in each period. Again, the crucial distinction between these two cases is whether the unobserved individual effect embodies elements that are correlated with regressors in the model, not whether these effects are stochastic or not.

Consider the following model

$$y_{it} = \boldsymbol{x}'_{it}\boldsymbol{\beta} + (\alpha + u_i) + \varepsilon_{it}$$

Where there are K regressors including a constant and now the single constant term is the mean of the unobserved heterogeneity, $E[\boldsymbol{z}'_i\boldsymbol{\alpha}]$. The component u_i is the random heterogeneity specific to the ith observation and is constant through time; recall that, $u_i = \{\boldsymbol{z}'_i\boldsymbol{\alpha} - E[\boldsymbol{z}'_i\boldsymbol{\alpha}]\}$. For example, in an analysis of families, we can view u_i as the collection of factors, $\boldsymbol{z}'_i\boldsymbol{\alpha}$, not in the regression that are specific to that family.

We assume further that

$$E(\varepsilon_{it}|\mathbf{X}) = E(u_i|\mathbf{X}) = 0\ ,$$

$$E(\varepsilon_{it}^2|\mathbf{X}) = \sigma_\varepsilon^2,$$

$$E(u_i^2|\mathbf{X}) = \sigma_u^2,$$

$$E(\varepsilon_{it}u_j|\mathbf{X}) = 0 \text{ for all } i, t, \text{and } j,$$

$$E(\varepsilon_{it}\varepsilon_{js}|\mathbf{X}) = 0 \text{ if } t \neq s \text{ or } i \neq j,$$

$$E(u_iu_j|\mathbf{X}) = 0 \text{ if } i \neq j.$$

It is useful to view the formulation of the model in blocks of T observations for groups $i, \boldsymbol{y}_i, \mathbf{X}_i, u_i\mathbf{i}$, and $\boldsymbol{\varepsilon}_i$. For these T observations, let,

$$\eta_{it} = \varepsilon_{it} + u_i$$

and

$$\boldsymbol{\eta}_i = [\eta_{i1}, \eta_{i2}, \dots, \eta_{iT}]'.$$

In view of this form of $\boldsymbol{\eta}_{it}$, we have what is often called an “error components model.” For this model,

$$E[\eta_{it}^2|\mathbf{X}] = \sigma_\varepsilon^2 + \sigma_u^2,$$

$$E[\eta_{it}\eta_{is}|\mathbf{X}] = \sigma_u^2, \quad t \neq s$$

$$E\left[\eta_{it}\eta_{js}\middle|\mathbf{X}\right] = 0, \quad for\ all\ t\ and\ s\ if\ i \neq j.$$

Do there exist individual-specific Random effects? *(Greene's notation)*

Breusch and Pagan (1980) have devised a Lagrange multiplier test for the random effects model based on the OLS residuals. For

$$H_0: \sigma_u^2 = 0 \quad (or\ Corr[\eta_{it}, \eta_{is}] = 0),$$

$$H_1: \sigma_u^2 \neq 0,$$

The test statistic is

$$LM = \frac{nT}{2(T-1)}\left[\frac{\sum_{i=1}^{n}[\sum_{t=1}^{T} e_{it}]^2}{\sum_{i=1}^{n}\sum_{t=1}^{T} e_{it}^2} - 1\right]^2 = \frac{nT}{2(T-1)}\left[\frac{\sum_{i=1}^{n} T\bar{e}_{i.}{}^2}{\sum_{i=1}^{n}\sum_{t=1}^{T} e_{it}^2} - 1\right]^2.$$

Under the null hypothesis, LM is distributed as chi-squared with one degree of freedom.

The Fixed Effects model:

Fixed effects models control for, partial out, the effects of time-invariant variables with time invariant effects. This is true whether the variable is explicitly measured or not.

If there are omitted variables, and these variables are ***correlated*** with the variables in the model, then fixed effects models may provide a means for controlling for omitted variable bias. In a fixed effects model, subjects serve as their own controls. The idea/hope is that whatever effects the omitted variables have on the subject at a time, they will have the same effect at a later time; hence their effects will be constant, or "*fixed*". However, in order for this to be true, the effects must have time-invariant values (the value of the variable does not change across time) with time-invariant effects (the variable has the same effect across time)[2].

The random effects approach to estimating $\boldsymbol{\beta}$ effectively puts c_i into the error term under the assumption that c_i is orthogonal to $\boldsymbol{x_{it}}$, and then accounts for the implied serial correlation in the composite error $v_{it} = c_i + u_{it}$ using GLS analysis. In many applications the whole point of using panel data is to allow for c_i to be arbitrarily correlated with x_{it}. A fixed effects analysis achieves this purpose explicitly.

[2] If either of these assumptions is violated, we need to have explicit measurements of the variables in question and include them in our models. In case of time-varying effects, we can include the interaction of the time-invariant variable with time.

In the fixed effects model, the individual-specific effect is a random variable that is allowed to be correlated with the explanatory variables.

FE1: Related effects

FE1 explicitly states that the absence of the unrelatedness assumption in *FE1*.

FE2: Effect Variance

FE2 explicitly states the absence of the assumption in *RE2.*

FE3: Identifiability

$rank(\ddot{X}) = K < NT \ and \ E(\acute{\ddot{x}}_i \ddot{x}_i)$ is p.d. and finite where the typical element

$$\ddot{x}_{it} = x_{it} - \bar{x}_i \ \ and \ \bar{x}_i = 1/T \sum_t x_{it}$$

FE3 assumes that the time-varying explanatory variables are not perfectly collinear, that they have non-zero within-variance (i.e variation over time for a given individual) and not too many extreme values. Hence, x_{it} cannot include a constant or any time-invariant variables. Note that only the parameters β but neither α nor γ are identifiable in the fixed effects model.

Fixed Effects Model Estimation:

Firm effects:

Model: $$\boldsymbol{y_{it} = x'_{it}\beta + \mu + \alpha_i + \varepsilon_{it}}$$

Within effect model: $$\boldsymbol{y_{it} - \bar{y}_{i.} = (x_{it} - \bar{x}_{i.})\acute{}\beta + \varepsilon_{it} - \bar{\varepsilon}_{i.}}$$

Between effect model: $$\boldsymbol{\bar{y}_{i.} = \bar{x}_{i.}\beta + \alpha + \bar{\varepsilon}_{i.}}$$

Constant: $$\boldsymbol{\hat{\mu} = m = \bar{y}_{i.} - \bar{x}'_{i.}b}$$

Dummy coefficients: $$\boldsymbol{\hat{\alpha}_i = a_i = (y_{it} - \bar{y}_{i.}) - (x_{it} - \bar{x}_{i.})\acute{}b}$$

Time Effects:

Model: $$\boldsymbol{y_{it} = x'_{it}\beta + \mu + \gamma_t + \varepsilon_{it}}$$

Within effect model: $$\boldsymbol{y_{it} - \bar{y}_{.t} = (x_{it} - \bar{x}_{.t})\acute{}\beta + \varepsilon_{it} - \bar{\varepsilon}_{.t}}$$

Between effect model: $\bar{y}_{.t} = \bar{x}_{.t}\beta + \gamma + \bar{\varepsilon}_t$

Constant: $\hat{\mu} = m = \bar{y}_{.t} - \bar{x}'_{.t}b$

Dummy coefficients: $\hat{\gamma}_t = c_t = (y_{it} - \bar{y}_{.t}) - (x_{it} - \bar{x}_{.t})'b$

Both (Firm & Time) effects:

Model: $y_{it} = x'_{it}\beta + \mu + \alpha_i + \gamma_t + \varepsilon_{it}$

$$\sum_i \alpha_i = \sum_t \gamma_t = 0$$

Within effect model: $y_{*it} = y_{it} - \bar{y}_{i.} - \bar{y}_{.t} + \bar{\bar{y}}$

$$x_{*it} = x_{it} - \bar{x}_{i.} - \bar{x}_{.t} + \bar{\bar{x}}$$

$$\bar{y}_{.t} = \frac{1}{n}\sum_{i=1}^{n} y_{it}$$

$$\bar{\bar{y}} = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} y_{it}$$

Overall constant: $\hat{\mu} = m = \bar{\bar{y}} - \bar{\bar{x}}'b$

Dummy coefficients: $\hat{\alpha}_i = a_i = (\bar{y}_{i.} - \bar{\bar{y}}) - (\bar{x}_{i.} - \bar{\bar{x}})'b$

Dummy coefficients: $\hat{\gamma}_{.t} = c_t = (\bar{y}_{.t} - \bar{\bar{y}}) - (\bar{x}_{.t} - \bar{\bar{x}})'b$

CAUTION: If the subjects change a little, or not at all, across time, a fixed effects model may not work very well or even at all. There needs to be within-subject variability in the variables if we are to use subjects as their own controls. If there is little variability within subjects then the standard errors from fixed effects models may be too large to tolerate. Conversely, random effects models will often have smaller standard errors. But, the trade-off is that their coefficients are more likely to be biased.

- If you think there are no omitted variables – or if you believe that the omitted variables are ***<u>uncorrelated</u>*** with the explanatory variables that are in the model – then a random effects model is probably best. It will produce unbiased estimates of the coefficients, use all the data available, and produce the smallest standard errors. More likely, however, is that omitted variables will produce at least some bias in the estimates.

Which model (Random Effects vs. Fixed Effects) fits the data Better?

I- Hausman's Specification Test for the random Effects Model: *(Greene's notation)*

The specification test designed by Hausman (1978) is used to test for orthogonality of the random effects and the regressors. The test is based on the idea that under the hypothesis of no correlation, both OLS in the LSDV model and GLS are consistent, but OLS is inefficient, where under the alternative, OLS is consistent, but GLS is not. Therefore, under the null hypothesis, the two estimates should not differ systematically, and a test can be based on the difference. The other essential ingredient for the test is the convariance matrix of the difference vector, $\left[\mathbf{b}, \hat{\boldsymbol{\beta}}\right]$:

$$\mathbf{Var}\left[\mathbf{b}-\hat{\boldsymbol{\beta}}\right] = \mathbf{Var}(\mathbf{b}) + \mathbf{Var}\left(\hat{\boldsymbol{\beta}}\right) - \mathbf{Cov}\left[\mathbf{b}-\hat{\boldsymbol{\beta}}\right] - \mathbf{Cov}\left[\mathbf{b}-\hat{\boldsymbol{\beta}}\right]$$

Hausman's essential result is that the convariance of an efficient estimator with its difference from an inefficient estimator is zero, which implies that

$$\mathbf{Cov}\left[\left(\mathbf{b}-\hat{\boldsymbol{\beta}}\right), \hat{\boldsymbol{\beta}}\right] = \mathbf{Cov}\left[\mathbf{b}.\hat{\boldsymbol{\beta}}\right] - \mathbf{Var}\left(\hat{\boldsymbol{\beta}}\right) = \mathbf{0}$$

or that

$$\mathbf{Cov}\left[\mathbf{b}.\hat{\boldsymbol{\beta}}\right] = \mathbf{Var}\left(\hat{\boldsymbol{\beta}}\right)$$

Inserting this result in the first equation produces the required covariance matrix for the test,

$$\mathbf{Var}\left[\mathbf{b}-\hat{\boldsymbol{\beta}}\right] = \mathbf{Var}[\mathbf{b}] - \mathbf{Var}\left[\hat{\boldsymbol{\beta}}\right] = \boldsymbol{\Psi}.$$

The chi-squared test is based on the Wald criterion:

$$W = \chi^2[K-1] = [\boldsymbol{b}-\hat{\boldsymbol{\beta}}]' \hat{\boldsymbol{\Psi}}^{-\mathbf{1}}\left[\boldsymbol{b}-\hat{\boldsymbol{\beta}}\right].$$

For $\hat{\Psi}$, we use the estimated convariance matrices of the slope estimator in the LSDV model and the estimated convariance matrix in the random effects model, excluding the constant term. Under the null hypothesis, *W* has a limiting chi-squared distribution with *K*-1 degrees of freedom.

II- Auxiliary Regression (Wooldridge 2010, p. 332, eq. 10.88, Mundlak, 1978)

The random effects model can be consistently estimated by both the RE estimator or the FE estimator. We would prefer the RE estimator if we can be sure that the individual-specific effect is an unrelated effect (RE1). This is usually tested by a (Durbin-Wu-) Hausman test. However, the Hausman test is only valid under homoscedasticity and cannot include time fixed effects. The unrelatedness assumption (RE1) is better tested by running an auxiliary regression (Wooldridge 2010, p. 332, eq. 10.88, Mundlak, 1978):

$$y_{it} = \alpha + \acute{x}_{it}\beta + \acute{z}_i\gamma + \acute{\bar{x}}_i\lambda + \delta_t + u_{it}$$

where $\bar{x}_i = 1/T \sum_t x_{it}$ are the time averages of all time-varying regressors. Include time fixed δ_t if they are included in the RE and FE estimation. A joint Wald-test on $H_0 : \lambda = 0$ tests RE1. Use cluster-robust standard errors to allow for heteroscedasticity and serial correlation.

Note: Assumption RE1 is an extremely strong assumption and the FE estimator is almost always much more convincing than the RE estimator. Not rejecting RE1 does not mean accepting it. Interest in the effects of a time-invariant variable is no sufficient reason to use the RE estimator.

References:

- Greene, W., ***Econometric Analysis***, 5th edition, 2003. Printice Hall.
- https://www3.nd.edu/~rwilliam/stats3/Panel04-FixedVsRandom.pdf
- https://www.schmidheiny.name/teaching/panel.pdf