Panel Econometrics

Notes

William Radaic Peron

EESP-FGV

April 19, 2021

Contents

1 Pooled OLS 3
1.1 Motivation . . . 3
1.2 Pooled Ordinary Least Squares Estimator . . . 3
1.2.1 Assumptions for identification . . . 4
1.2.2 Assumptions for inference . . . 4
1.3 Asymptotics . . . 4
1.4 Dynamic Completeness . . . 4
1.4.1 Lags . . . 4
1.4.2 Persistence . . . 5
1.5 Robust inference . . . 5
1.5.1 Asymptotic variance estimator . . . 5
1.5.2 Testing for serial correlation and heteroskedasticity . . . 6
1.6 Feasible GLS . . . 6
1.6.1 Consistency and asymptotic normality . . . 6

2 Fixed Effects Estimator 8
2.1 Unobserved effect . . . 8
2.2 Pooled OLS with unobserved effects . . . 9
2.3 Fixed effects with dummies . . . 9
2.4 Within (FE) estimator . . . 10
2.4.1 Transformation . . . 11
2.4.2 Consistency . . . 11
2.4.3 Inference . . . 12
2.5 First difference (FD) estimator . . . 13
2.5.1 Consistency . . . 14
2.5.2 Inference . . . 14
2.6 Clustering . . . 15
2.6.1 Moulton factor . . . 15
2.6.2 Clustering corrections . . . 16
2.7 Measurement error . . . 16

3 Differences-in-Differences Estimator 19

3.1 DD framework 19
3.2 Regression DD 21
3.2.1 Potential problems in DD design 22
3.3 Clustering and serial correlation for DD 22

4 Random Effects Estimator 25
4.1 Consistency 25
4.2 Inference 26
4.2.1 RE variance matrix 26
4.2.2 A robust variance matrix estimator for RE 28
4.2.3 General Feasible GLS 28
4.3 Testing for unobserved effects 28

5 Testing and comparing estimators 30
5.1 Comparing panel estimators 30
5.1.1 FE and FD for different assumption violations 30
5.1.2 Relationship between RE and FE 32
5.2 Hausman test 33

6 Dynamic models 36

7 Sample selection 37

Chapter 1

Pooled OLS

References: AW, 7.7, 7.4.

1.1 Motivation

In this course, we'll study econometric models and estimators for *panel* data.

Definition 1.1.1. *A panel is a data set that combines cross-sectional observations of the same units (individuals) over multiple periods of time.*

Panel data can offer multiple solutions for common issues in Econometrics, such as omitted variable bias (OVB), sample restrictions, dynamics and simultaneity. However, the temporal structure of such data usually requires further considerations to yield consistent and efficient estimates.

1.2 Pooled Ordinary Least Squares Estimator

Suppose a linear model of the form:

$$y_t = \mathbf{x}_t\beta + u_t, \qquad t = 1, 2, ..., T \tag{1.1}$$

Although this seems to restrict β to be fixed over time, varying parameters can be achieved through manipulations in the covariates (e.g., time dummies, trends).

> As a general rule, with large N and small T it is a good idea to allow for separate intercepts for each time period. Doing so allows for aggregate time effects that have the same influence on y_{it} for all i. (AW, p. 192)

The pooled OLS estimator is analogous to its cross-sectional counterpart, as will be now clear with its assumptions.

1.2.1 Assumptions for identification

Assumption 1.2.1. ***Contemporaneous exogeneity.***[1] $\mathbb{E}[\mathbf{x_t}'u_t] = 0, \forall t = 1, 2, ..., T.$

Assumption 1.2.2. ***Full rank.*** *$rank[\sum_{i=1}^{T} \mathbb{E}(\mathbf{x_t'x_t})] = K$ for k regressors.*

1.2.2 Assumptions for inference

A further assumption may be imposed for efficiency:

Assumption 1.2.3. ***Homoskedasticity.***

- $\mathbb{E}(u_t^2\mathbf{x_t}'x_t) = \sigma^2\mathbb{E}(\mathbf{x_t'x_t}), t = 1, 2, ..., T.$
- $\sigma^2 = \mathbb{E}(u_t^2), \forall t.$
- $\mathbb{E}(u_t u_s \mathbf{x_t'x_s}) = 0, t \neq s, (t, s) = 1, 2, ..., T.$

Intuitively, homoskedasticity means that the conditional variance does not depend on $\mathbf{x_t}$ and that the unconditional variance is time-invariant. This is important because it

> *essentially restricts the conditional covariances of the errors across different time periods to be zero.* the errors across different time periods to be zero. In fact, since $\mathbf{x}_t$ almost always contains a constant, POLS.3b [] requires at a minimum that $\mathbb{E}\left(u_t u_s\right) = 0, t \neq s$. Sufficient for POLS.3 b $[\mathbb{E}(u_t u_s \mathbf{x_t'x_s}) = 0, t \neq s]$ is $\mathbb{E}\left(u_t u_s \mid \mathbf{x_t}, \mathbf{x_s}\right) = 0, t \neq s, t, s = 1, \ldots, T$ (AW, p. 193)

1.3 Asymptotics

Theorem 1.3.1. ***Large-sample properties of Pooled OLS.*** *Under assumptions 1.2.1 (contemporaneous exogeneity) and 1.2.2 (full rank), $\hat{\beta}_{pOLS} \sim_A \mathcal{N}(\cdot)$. If 1.2.3 (homoskedasticity) also holds, then $Avar(\hat{\beta}_{pOLS}) = \sigma^2[\mathbb{E}(\mathbf{X_i'X_i})]^{-1}/N$. This implies that an appropriate estimator for $Avar(\hat{\beta}_{pOLS})$ is $\hat{\sigma}^2(\mathbf{X'X})^{-1}$, where $\hat{\sigma}^2$ is the usual (pooled) OLS variance estimator.*

1.4 Dynamic Completeness

1.4.1 Lags

When the model has dynamics, it is important to ensure that $\mathbb{E}(u_t u_s \mathbf{x_t'x_s}) = 0, t \neq s$ holds and that the model completely captures this aspect. Dynamic completeness can be stated as:

$$\mathbb{E}(y_t|\mathbf{x_t}, y_{t-1}, ..., y_1, \mathbf{x_1}) = \mathbb{E}(y_t|x_t) \tag{1.2}$$

In other words, the regressors (including lags of x and y) completely capture the dynamics of the model.

[1] Note that this assumption says nothing about the relationship between x_t, u_t for $s \neq t$. (AW, p. 192).

Note that this is equivalent to[2]:

$$\mathbb{E}(u_t|\mathbf{x_t}, u_{t-1}, \mathbf{x_{t-1}}, ..., u_1, \mathbf{x_1}) = 0 \tag{1.3}$$

Iterated expectations further implies that $\mathbb{E}(u_t u_s|\mathbf{x_t}, \mathbf{x_s}) = 0, s \neq t$. This means that dynamic completeness *implies contemporary exogeneity* (1.2.1) *and no nonzero error autocorrelations.*

> Often, once we start putting any lagged values of y_t into $\mathbf{x_t}$, then equation 1.3 is an intended assumption. But this generalization is not always true. [...] We may not care that serial correlation is still present in the error [...] [and] estimate the asymptotic variance of the pooled OLS estimator to be robust to serial correlation.
>
> In introductory econometrics, students are often warned that having serial correlation in a model with a lagged dependent variable causes the OLS estimators to be inconsistent. *While this statement is true in the context of a specific model of serial correlation, it is not true in general, and therefore it is very misleading.* Our analysis shows that, whatever is included in $\mathbf{x_t}$, pooled OLS provides consistent estimators of b whenever $\mathbb{E}(y_t|\mathbf{X_t}) = \mathbf{x_t}\beta$; it does not matter that the u_t might be serially correlated. (AW, p. 196)

1.4.2 Persistence

Also note that the asymptotic properties of the pooled OLS estimator, 1.3.1, do not impose restrictions on *time series persistence.* That is the case because our asymptotics refer to *fixed T, large N.* If T grows, we enter time series analysis, which requires further knowledge of the temporal dependence of the data – together with some assumptions, such as weak dependence[3].

If the processes $\{y_t, x_t\}$ are autoregressive in some way,

$$y_t = \mu + \theta y_{t-1} + u_t, \qquad \mathbb{E}(u_t|y_{t-1}, ..., y_0) = 0 \tag{1.4}$$

dynamic completeness (1.3) ensures consistency for fixed T, large N. However, as $T \to \infty$, it is necessary to assume $|\theta| < 1$ as a stability condition.

1.5 Robust inference

1.5.1 Asymptotic variance estimator

Because homoskedasticity (1.2.3) is too restrictive, we can obtain a roubst estimate of the asymptotic variance of $\hat{\beta}_{pOLS}$. The fully robust – to arbitrary heteroskedasticity and serial

[2] If the expectation conditional on X is equivalent to the conditional expectation given y, then the model is correctly specified.

[3] Weakly dependent processes are commonly, if somewhat misleadingly, referred to as "stationary" processes. (AW, p. 322)

correlation – $\widehat{Avar(\hat{\beta}_{pOLS})}$ can be stated as:

$$\text{Avar}(\hat{\boldsymbol{\beta}}) = \left(\sum_{i=1}^{N} \mathbf{X}_i'\mathbf{X}_i\right)^{-1} \left(\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T} \hat{u}_{it}\hat{u}_{is}\mathbf{x}_{it}'\mathbf{x}_{is}\right) \left(\sum_{i=1}^{N} \mathbf{X}_i'\mathbf{X}_i\right)^{-1} \tag{1.5}$$

1.5.2 Testing for serial correlation and heteroskedasticity

Serial correlation should not be present in the model if it is dynamically complete. To test for that, we can just model the error as an AR(1) process:

$$u_t = \rho_1 u_{t-1} + e_t,$$

and test for the null of no serial correlation – i.e., $\rho_1 = 0$. This can be done by regressing y_{it} on $\mathbf{x_{it}}, \mathbf{\hat{u}_{i,t-1}}$ and doing a standard $t-$test of $\hat{\rho}_1$.

The advantage of this approach compared to regressing $\mathbf{\hat{u}_{i,t}}$ on $\mathbf{\hat{u}_{i,t-1}}$ is that the above works regardless of strict exogeneity.

Heteroskedasticity can be tested as in the standard OLS framework. For example, regress $\mathbf{\hat{u}^2_{it}}$ on $1, \mathbf{h_{it}}$, where $\mathbf{h_{it}}$ is a vector of nonconstant functions of $\mathbf{x_{it}}$. The test statistic is NTR_c^2 of such regression, which is asymptotically χ^2_Q under the null. $\mathbf{h_{it}}$ usually includes $\mathbf{x_{it}}$, its squares and cross products.

1.6 Feasible GLS

When we don't have homoskedasticity, it seems reasonable to estimate the model with a FGLS approach. However, this procedure

> is not even guaranteed to produce consistent, let alone efficient, estimators under Assumptions 1.2.1 (contemporaneous exogeneity) and 1.2.2 (full rank).

Unless $\Omega = \mathbb{E}(\mathbf{u_t}\mathbf{u_t'})$ is diagonal, FGLS only produces consistent estimators with strict exogeneity.

In other words, FGLS is an option if $\mathbb{E}(u_t u_s \mathbf{x_t'}\mathbf{x_s}) = 0, t \neq s$ does not hold – i.e., if the model presents correlation between errors and regressors across time.[4]

1.6.1 Consistency and asymptotic normality

(F)GLS is consistent under fairly weak assumptions. However, strict exogeneity is required. It is now presented in Kronecker product form.

Assumption 1.6.1. ***Strict exogeneity.*** $\mathbb{E}(\mathbf{X_i} \otimes \mathbf{u_i}) = \mathbf{0}$.

[4] Note that $\mathbb{E}(u_t u_s \mathbf{x_t'}\mathbf{x_s}) = 0, t \neq s$ is a less restrictive form of dynamic completeness, since it only refers to linear dependence.

A sufficient condition for 1.6.1 is the usual $\mathbb{E}(\mathbf{u_i}|\mathbf{x_i}) = 0, \forall i$.

Now, define $\Omega := \mathbb{E}(\mathbf{u_i}\mathbf{u_i'})$ as the unconditional variance-covariance matrix of $\mathbf{u_i}$. Under 1.6.1, it can be stated that the (F)GLS estimator is consistent – i.e., the following condition holds:

$$\mathbb{E}(\mathbf{X_i'}\mathbf{\Omega^{-1}}\mathbf{X_i}) = 0. \tag{1.6}$$

This motivates the following assumption:

Assumption 1.6.2. ***Positive definite matrix.*** *Ω is positive definite and $\mathbb{E}(\mathbf{X_i'}\mathbf{\Omega^{-1}}\mathbf{X_i})$ is nonsingular.*

> The usual motivation for the GLS estimator is to transform a system of equations where the error has a nonscalar variance-covariance matrix into a system where the error vector has a scalar variance-covariance matrix. (AW, p. 174)

Multiplying the model by $\Omega^{-1/2}$ transforms the variance-covariance matrix to a scalar.

$$\mathbf{\Omega}^{-1/2}\mathbf{y}_i = \left(\mathbf{\Omega}^{-1/2}\mathbf{X}_i\right)\boldsymbol{\beta} + \mathbf{\Omega}^{-1/2}\mathbf{u}_i, \quad \text{or} \quad \mathbf{y}_i^* = \mathbf{X}_i^*\boldsymbol{\beta} + \mathbf{u}_i^* \tag{1.7}$$

Then, the GLS estimator, denoted by β^*, is,

$$\boldsymbol{\beta}^* := \left(\sum_{i=1}^{N} \mathbf{x}_i^*\mathbf{X}_i^*\right)^{-1}\left(\sum_{i=1}^{N} \mathbf{X}_i^*\mathbf{y}_i^*\right) = \left(\sum_{i=1}^{N} \mathbf{X}_i'\mathbf{\Omega}^{-1}\mathbf{X}_i\right)^{-1}\left(\sum_{i=1}^{N} \mathbf{X}_i'\mathbf{\Omega}^{-1}\mathbf{y}_i\right), \tag{1.8}$$

which is consistent with the above assumptions (strict exogeneity and matrix conditions). We now define $\mathbf{A} := \mathbb{E}\left(\mathbf{X}_i'\mathbf{\Omega}^{-1}\mathbf{X}_i\right)$, $\mathbf{A}$ is nonsingular.

(W)LLN yields:

$$\boldsymbol{\beta}^* = \boldsymbol{\beta} + \left(N^{-1}\sum_{i=1}^{N} \mathbf{X}_i'\mathbf{\Omega}^{-1}\mathbf{X}_i\right)^{-1}\left(N^{-1}\sum_{i=1}^{N} \mathbf{X}_i'\mathbf{\Omega}^{-1}\mathbf{u}_i\right) = \beta, \text{ as } N \to \infty. \tag{1.9}$$

The first part of the second expression converges asymptotically to $\mathbf{A}$, and the second part converges to zero.

GLS has an asymptotically normal distribution.

$$\sqrt{N}\left(\boldsymbol{\beta}^* - \boldsymbol{\beta}\right) = \left(N^{-1}\sum_{i=1}^{N} \mathbf{X}_i'\mathbf{\Omega}^{-1}\mathbf{X}_i\right)^{-1}\left(N^{-1/2}\sum_{i=1}^{N} \mathbf{X}_i'\mathbf{\Omega}^{-1}\mathbf{u}_i\right) \tag{1.10}$$

CLT ensures that $\left(N^{-1/2}\sum_{i=1}^{N} \mathbf{X}_i'\mathbf{\Omega}^{-1}\mathbf{u}_i\right) \sim_A \mathcal{N}(\mathbf{0}, \mathbf{B}), \mathbf{B} := \mathbb{E}\left(\mathbf{X}_i'\mathbf{\Omega}^{-1}\mathbf{u}_i\mathbf{u}_i'\mathbf{\Omega}^{-1}\mathbf{X}_i\right)$.

Then,

$$\sqrt{N}\left(\boldsymbol{\beta}^* - \boldsymbol{\beta}\right) \sim_A \mathcal{N}\left(\mathbf{0}, \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}\right), \tag{1.11}$$

and $Avar(\hat{\boldsymbol{\beta}}) = \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}/N$.

The *feasible* GLS estimator has the same properties under analogous assumptions, substituting Ω by $\widehat{\Omega}$, estimated in the first step by pooled OLS.

Chapter 2

Fixed Effects Estimator

References: AP, 5.1; D, 14.1, 14.2, AW, 10.1-10.3.

Panel data allows us to address for OVB in multiple clever ways. Fixed effects is the most powerful of them.

2.1 Unobserved effect

Formally, define c as an unobserved random variable in our model that contains variables y and $\mathbf{x}$. If c is completely exogenous to the covariates, then it is just another unobserved factor that affects y not systematically – and can be included in the general error term without much concern, since it doesn't cause bias. However, if $Cov(c, x_j) \neq 0$ for some j, then leaving c in the error term can obviously yield biased estimates.

We can correct for this in a number of already known ways: for example, with proxies or instrumental variables. With panel data, we have a possibly much better way.

Assumption 2.1.1. ***Unobserved effect.*** *c is time-invariant for each observation, and is denoted by c_i.*

From this, we can write our unobserved effects model (UEM) as:

$$y_{it} = \mathbf{x_{it}}\beta + c_i + u_{it}, \qquad t = 1, 2, ..., T. \tag{2.1}$$

where c_i is individual heterogeneity and u_{it} are idiosyncratic disturbances.

The assumptions we impose on the unobserved effect c_i is what distinguish the fixed and random effects estimators. In simple terms, the *random effects framework* is synonymous with $Cov(x_{it}, c_i) = 0, \forall t$, and in the *fixed effects framework* we allow for correlation between the unobserved effect c_i and the covariates.[1]

With models that contain unobserved effects, we always impose strict exogeneity.

[1] Note that c_i is *always* treated as a random variable.

Assumption 2.1.2. ***Strict exogeneity for unobserved effects.***

$$\mathbb{E}\left(y_{it} \mid \mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iT}, c_i\right) = \mathbb{E}\left(y_{it} \mid \mathbf{x}_{it}, c_i\right) = \mathbf{x}_{it}\boldsymbol{\beta} + c_i.$$

The main difference between FE and RE frameworks is the exogeneity of the unobserved effects *conditional on the regressors.* FE allows for correlation, while RE does not.

As usual, we also impose a full rank assumption – which will restrain in a number of important ways our FE estimates.

Assumption 2.1.3. ***Full rank for unobserved effects.*** *rank*$[\sum_{i=1}^{T} \mathbb{E}(\mathbf{x_t'}\mathbf{x_t})] = K$ *for* k *regressors.*

> in any panel data application we should initially focus on two questions: (1) Is the unobserved effect, c_i, uncorrelated with x_{it} for all t? (2) Is the strict exogeneity assumption (conditional on c_i) reasonable? (AW, p. 289)

2.2 Pooled OLS with unobserved effects

There's a way to obtain consistent pooled OLS estimates even with unobserved effects. Consider the model

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + v_{it}, \quad t = 1, 2, \ldots, T, \qquad v_{it} := c_i + u_{it}, \forall t \tag{2.2}$$

Our error term is now *composite*, and encompasses (i) unobserved time-invariant individual effects; and (ii) idiosyncratic disturbances.

Pooled OLS will be consistent if

$$\mathbb{E}(\mathbf{x_{it}'}v_{it}) = 0, \forall t, \tag{2.3}$$

which means that we are assuming both $\mathbb{E}(\mathbf{x_{it}'}u_{it}) = 0$ and $\mathbb{E}(\mathbf{x_{it}'}c_i) = 0$. The second assumption is more restrictive, which requires that the model be correctly specified for strict exogeneity.

2.3 Fixed effects with dummies

We now consider the case when c_i is potentially correlated with X_{it}. One possible first approach is to treat the unobserved individual effects as *parameters* to be estimated. Since Frisch-Waugh-Lovell ensures that adding variables to the regression partials out its effects, by including individual dummies we can essentially remove c_i from the error term and avoid inconsistency.

Consider the model

$$y_{it} = \mathbf{x}_{it} + c_i D_i + u_{it}, \quad t = 1, 2, ..., T, \tag{2.4}$$

where D_i is a dummy variable for observation i.

It is worth noting that, as $N \to \infty$, while $\hat{\beta} \to \beta$, the same isn't true for c_i. That happens because c_i grows at the same rate as N, which implies that the number of observations for each unobserved effect stays constant.

> Each $\hat{c}_i$ is an unbiased estimator of c_i when the c_i are treated as parameters, at least if we maintain [exogeneity conditional on c_i, x_i] and [full rank]. [...] The $\hat{c}_i$ give practical examples of estimators that are unbiased but not consistent.

Note that, in OLS regression, $\hat{c}_i$ can be estimated by:

$$\hat{c}_i = \bar{y_i} - \bar{\mathbf{x}}_\mathbf{i}\hat{\beta}_{FE}, \quad i = 1, 2, ..., N. \tag{2.5}$$

This makes sense because $\hat{c}_i$ is effectively the intercept for cross section unit i. It is, thus, sometimes useful to estimate and plot $\hat{c}_i$ and its moments (mean, median, quantiles) to get a sense of the heterogeneity of the sample.

We can consistently estimate σ_c^2 (variance of unobserved effects) if we assume that the error process, $\{u_{it}\}$ is serially uncorrelated with constant variance. First, it is possible to estimate $\sigma_v^2 = \sigma_c^2 + \sigma_u^2$. Since $v_{it} = y_{it} - x_{it}\beta$,

$$\hat{\sigma}_v^2 = (NT - K)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(y_{it} - \mathbf{x}_{it}\hat{\boldsymbol{\beta}}_{FE} \right)^2 \tag{2.6}$$

is an asymptotically consistent estimator for σ_v^2. Rewriting $\hat{\sigma}_c^2 = \hat{\sigma}_v^2 - \hat{\sigma}_u^2$ yields the sample variance of c_i. $\hat{\sigma}_u^2$ is calculated with

$$\hat{\sigma}_u^2 = \frac{\text{SSR}}{[N(T-1) - K]} \tag{2.7}$$

Note that $\hat{\sigma}_v^2$ correctly corrects for the number of degrees of freedom in the model, as K includes N dummies. This is not given in within (FE) or first difference (FD) approaches, and requires correction, as will be shown.

The dummy variable approach makes a point very clear in regards to the fixed effects framework: regressors that are time-invariant to all observations, such as age, race etc., *cannot be included in the model.* In the dummy approach, it effectively yields *multicolinearity,* as we're controlling twice for some time-invariant individual characteristics in this case.

2.4 Within (FE) estimator

We can achieve the same result of eliminating the unobserved effects c_i with a transformation.

2.4.1 Transformation

Consider the model

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + c_i + u_{it}, \quad t = 1, \ldots, T \tag{2.8}$$

Compute averages for all processes *over time, for each observation*:

$$\bar{y}_i = \overline{\mathbf{x}}_i\boldsymbol{\beta} + c_i + \bar{u}_i \tag{2.9}$$

Now, subtract 2.9 from 2.8:

$$y_{it} - \bar{y}_i = \left(\mathbf{x}_{it} - \overline{\mathbf{x}}_i\right)\boldsymbol{\beta} + u_{it} - \bar{u}_i \tag{2.10}$$

From this, we redefine the variables as:

$$\ddot{y}_{it} := y_{it} - \bar{y}_i, \ddot{\mathbf{x}}_{it} := \mathbf{x}_{it} - \overline{\mathbf{x}}_i, \text{ and } \ddot{u}_{it} := u_{it} - \bar{u}_i \tag{2.11}$$

The *within* transformation then yields:

$$\ddot{y}_{it} = \ddot{\mathbf{x}}_{it}\boldsymbol{\beta} + \ddot{u}_{it}, \quad t = 1, 2, \ldots, T \tag{2.12}$$

Note that time demeaning for each individual purges the unobserved effect, c_i, from the model.

2.4.2 Consistency

As per usual, we need assumptions for consistency.

Assumption 2.4.1. ***Strict exogeneity for FE.*** $\mathbb{E}(u_{it}|\mathbf{x_i}, c_i) = 0, \quad t = 1, 2, ..., T.$

Assumption 2.4.2. ***Full rank for FE.*** $rank\left(\sum_{t=1}^{T}\mathbb{E}\left(\ddot{\mathbf{x}}_{it}'\ddot{\mathbf{x}}_{it}\right)\right) = rank\left[\mathbb{E}\left(\ddot{\mathbf{X}}_i'\ddot{\mathbf{X}}_i\right)\right] = K.$

It follows, then, that estimating this equation can be done with pooled OLS. More formally, it is necessary that $\mathbb{E}(\ddot{x}_{it}'\ddot{u}_{it}) = 0, \forall t$ (strict exogeneity) holds.

Strict exogeneity is required because time demeaning implies that every observation for a unit will include a fraction of observations in other periods. If there's only contemporaneous exogeneity (1.2.1), for example, then the possible cross correlations will cause bias.[2]

Again, it is worth pointing out that the full rank condition (2.4.2) rules out time-constant variables as regressors. The within transformation will remove, aside from c_i, all of these variables (as time demeaning them will be equal to every observation). This implies that $\mathbb{E}(\ddot{x}_{it}'\ddot{u}_{it})$ will be a matrix with columns equal to vectors of zeroes – which violates full rank, as this means linear dependence for columns.

Considering all of these remarks, then, the *within estimator* is given by

$$\hat{\boldsymbol{\beta}}_{FE} = \left(\sum_{i=1}^{N}\ddot{\mathbf{x}}_i'\mathbf{X}_i\right)^{-1}\left(\sum_{i=1}^{N}\ddot{\mathbf{x}}_i'\mathbf{y}_i\right) = \left(\sum_{i=1}^{N}\sum_{t=1}^{T}\ddot{\mathbf{x}}_{ii}'\mathbf{x}_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\ddot{\mathbf{x}}_{it}'\ddot{y}_{it}\right). \tag{2.13}$$

[2] More info on AW, p. 303.

2.4.3 Inference

Initially, we'll consider the usual condition for efficiency, i.e., homoskedasticity.

Assumption 2.4.3. ***Homoskedasticity for FE.*** $\mathbb{E}(\mathbf{u_i u_i'}|\mathbf{x_i}, c_i) = \sigma_u^2 \mathbf{I}_T$.

Along with strict exogeneity (2.4.1), this means that the unconditional variance matrix can be decompose as $v_i = c_i j_T + u_i$, where j_T is a matrix of individual dummies.[3] However, the conditional variance is different.

Efficiency isn't trivial with this assumption, because it is necessary that $\{\ddot{u}_{it}\}$ exhibits no serial correlation and is homoskedastic. The variance of $\ddot{u}_{it}$ can be computed as:

$$\begin{aligned}\mathbb{E}\left(\ddot{u}_{it}^2\right) &= \mathbb{E}\left[(u_{it} - \bar{u}_i)^2\right] = \mathbb{E}\left(u_{it}^2\right) + \mathbb{E}\left(\bar{u}_i^2\right) - 2\mathbb{E}\left(u_{it}\bar{u}_i\right) \\ &= \sigma_u^2 + \sigma_u^2/T - 2\sigma_u^2/T = \sigma_u^2(1 - 1/T),\end{aligned} \tag{2.14}$$

which demonstrates homoskedasticity. However, for $t \neq s$, we have:

$$\begin{aligned}\mathbb{E}\left(\ddot{u}_{it}\ddot{u}_{is}\right) &= \mathbb{E}\left[\left(u_{it} - \bar{u}_i\right)\left(u_{is} - \bar{u}_i\right)\right] = \mathbb{E}\left(u_{it}u_{is}\right) - \mathbb{E}\left(u_{it}\bar{u}_i\right) - \mathbb{E}\left(u_{is}\bar{u}_i\right) + \mathbb{E}\left(\bar{u}_i^2\right) \\ &= 0 - \sigma_u^2/T - \sigma_u^2/T + \sigma_u^2/T = -\sigma_u^2/T < 0,\end{aligned} \tag{2.15}$$

which clearly *shows negative autocorrelation* for the process $\{\ddot{u}_{it}\}$. Combining both results, we have

$$\operatorname{Corr}\left(\ddot{u}_{it}, \ddot{u}_{is}\right) = -1/(T-1), \tag{2.16}$$

which converges to zero asymptotically.

To find the asymptotic distribution of the FE estimator, write

$$\sqrt{N}\left(\hat{\boldsymbol{\beta}}_{FE} - \boldsymbol{\beta}\right) = \left(N^{-1}\sum_{i=1}^{N} \ddot{\mathbf{X}}_i'\ddot{\mathbf{X}}_i\right)^{-1}\left(N^{-1/2}\sum_{i=1}^{N}\ddot{\mathbf{X}}_i'\mathbf{u}_i\right) \tag{2.17}$$

where we have used the important fact that $\ddot{\mathbf{X}}_i'\mathbf{u}_i = \mathbf{X}_i'\mathbf{Q}_T\mathbf{u}_i = \mathbf{X}_i'\mathbf{u}_i$.[4] (AW, p. 305)

It follows, then, that

$$\sqrt{N}\left(\hat{\boldsymbol{\beta}}_{FE} - \boldsymbol{\beta}\right) \sim_A \mathcal{N}\left(\mathbf{0}, \sigma_u^2\left[\mathbb{E}\left(\ddot{\mathbf{X}}_i'\ddot{\mathbf{X}}_i\right)\right]^{-1}\right) \tag{2.18}$$

and

$$Avar\left(\hat{\boldsymbol{\beta}}_{FE}\right) = \sigma_u^2\left[\mathbb{E}\left(\ddot{\mathbf{X}}_i'\ddot{\mathbf{X}}_i\right)\right]^{-1}/N \tag{2.19}$$

This can be easily estimated with its sample analogue,

$$\operatorname{Avar}\left(\hat{\boldsymbol{\beta}}_{FE}\right) = \hat{\sigma}_u^2\left(\sum_{i=1}^{N}\ddot{\mathbf{X}}_i'\ddot{\mathbf{X}}_i\right)^{-1} = \hat{\sigma}_u^2\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\ddot{\mathbf{x}}_{it}'\ddot{\mathbf{x}}_{it}\right)^{-1} \tag{2.20}$$

The asymptotic standard errors of the FE estimates are obtained as the square roots of the diagonal elements of the [above] matrix (AW, p. 306)

[3] This is analogous to the RE vcov matrix.

[4] $\mathbf{Q_T}$ is idempotent and symmetric.

There's one subtlety with FE inference, notably regarding the degrees of freedom correction for sample variance. To see this, note that

$$\sum_{i=1}^{T} \mathbb{E}(\ddot{u}_i t^2) = (T-1)\sigma_u^2 \implies [N(T-1)]^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} \mathbb{E}\left(\ddot{u}_{it}^2\right) = \sigma_u^2, \tag{2.21}$$

where the implication follows from summing over all individuals to yield the general variance for u.

The FE residuals can be written as

$$\hat{\boldsymbol{u}}_{it} = \ddot{y}_{it} - \ddot{\mathbf{x}}_{it}\hat{\boldsymbol{\beta}}_{FE}, \quad t = 1, 2, \ldots, T; i = 1, 2, \ldots, N.$$

Then, a consistent estimator for σ_u^2 is

$$\hat{\sigma}_u^2 = \frac{\text{SSR}}{[N(T-1)-K]}, \tag{2.22}$$

where SST is the total sum of squared residuals *for the (within) transformed residuals.* Note that there's a correction for N degrees of freedom – i.e., for the number of unobserved effects removed from the model with the within transformation. In the dummy framework, this is not necessary, as the dummies imply that K actually encompasses all individual unobserved effects as parameters.

2.5 First difference (FD) estimator

Another possible transformation to remove the unobserved effects c_i from the model is *taking first differences* for all variables.

Lagging the usual model yields:

$$\Delta y_{it} = \Delta \mathbf{x}_{it}\boldsymbol{\beta} + \Delta u_{it}, \quad t = 2, 3, \ldots, T \tag{2.23}$$

The FD estimator is, then, the pooled OLS regresion of

$$\Delta y_{it} \text{ on } \Delta \mathbf{x_{it}}, \quad t = 2, 3, \ldots, T, \quad i = 1, 2, \ldots, N.$$

Note that we now have $T-1$ periods, because we taking first differences implies losing the first period.

> Di¤erences for observation numbers $1, T+1, 2T+1; 3T+1, ..., and(N-1)T+1$ should be set to missing. These observations correspond to the first time period for every cross section unit in the original data set; by definition, there is no first difference for the $t = 1$ observations.

The FD transformation makes it clear, again, that all regressors *must be time-varying.* If an x_j is constant for all t, then Δx_j is zero and must be removed from the model.

2.5.1 Consistency

Assumption 2.5.1. ***Strict exogeneity for FD.*** $\mathbb{E}(u_{it}|\mathbf{x_i}, c_i) = 0, \quad t = 1, 2, ..., T.$

Assumption 2.5.2. ***Full rank for FD.*** rank $\left(\sum_{t=2}^{T} \mathbb{E}\left(\Delta\mathbf{x}'_{it}\Delta\mathbf{x}_{it}\right)\right) = K$

Strict exogeneity here (2.5.1) implies not only that FD is consistent:

$$\mathbb{E}(\Delta u'_{it}\Delta x_{it}) = 0, \quad t = 2, 3, ..., T. \tag{2.24}$$

but also *unbiased*, because the following condition holds:

$$\mathbb{E}\left(\Delta u_{it} \mid \Delta\mathbf{x}_{i2}, \Delta\mathbf{x}_{i3}, \ldots, \Delta\mathbf{x}_{iT}\right) = 0, \quad t = 2, 3, \ldots, T. \tag{2.25}$$

2.5.2 Inference

The key aspect of FD when compared to FE is that it is not *efficient.* The FE estimator assumes (i) homoskedasticity (2.4.3); and (ii) *no serial correlation* for $\{u_{it}\}$. Assuming that $\{u_{it}\}$ is serially uncorrelated may be too strong, and a possible alternative is assuming that $\{\Delta u_{it}\}$ is serially uncorrelated.

Assumption 2.5.3. ***Homoskedasticity for FE.***

$$\mathbb{E}\left(\mathbf{e}_i\mathbf{e}'_i \mid \mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}, c_i\right) = \sigma_e^2\mathbf{I}_{T-1},$$

where $\mathbf{e}_i$ *is the* $(T-1)\times 1$ *vector containing* $e_{it}, t = 2, \ldots, T,$ *and* $e_{it} := \Delta u_{it}$*.*

This assumption represents an extreme opposite from FE, since it implies that $\{u_{it}\}$ is a random walk, and has, therefore, substantial serial dependence.

Efficiency depends, then, on the serial correlation assumption that we take as being valid. If $\{u_{it}\}$ is assumed not to be serially correlated, then FE is efficient. FD is efficient if 2.5.3 holds, i.e., if $\{u_{it}\}$ is a random walk.

The assumption of no serial correlation of $\{e_{it}\}$ can be tested with a simple regression:

$$\hat{e}_{it} = \hat{\rho}_1\hat{e}_{i,t-1} + \text{error}_{it}, \quad t = 3, 4, \ldots, T; i = 1, 2, \ldots, N, \tag{2.26}$$

where we compute the usual $t-$test for $\hat{\rho}_1$.

Usually, though, that's not the case, and it implies the need for adjustments.

> If the idiosyncratic errors $\{u_{it} : t = 1, 2, \ldots, T\}$ are uncorrelated to begin with, $\{e_{it} : t = 1, 2, \ldots, T\}$ will be autocorrelated. In fact, under 2.4.3 (Homoskedasticity for FE) it is easily shown that $Corr(e_{it}, e_{it-1}) = -0.5$. (AW, p. 320)

AP highlight this point:

> the differenced standart errors should be adjusted for the fact that the differenced residuals are serially correlated. (AP, p. 224)

We now present a robust variance estimator for FD.

$$\text{Avar}\left(\hat{\beta}_{FD}\right) = \left(\Delta\mathbf{X}'\Delta\mathbf{X}\right)^{-1}\left(\sum_{i=1}^{N}\Delta\mathbf{X}'_i\hat{\mathbf{e}}\hat{\mathbf{e}}'_i\Delta\mathbf{X}_i\right)\left(\Delta\mathbf{X}'\Delta\mathbf{X}\right)^{-1}. \tag{2.27}$$

2.6 Clustering

It is possible that the observations come from different groups, with some form of dependence between them. In this case, since we don't have a random sample at the individual level, but instead at the group level, our standard errors will need corrections. That is the motivation behind clustering.

2.6.1 Moulton factor

Consider the example from the STAR experiment (AP, 8.2):

$$y_{ig} = \beta_0 + \beta_1 x_g + e_{ig}. \tag{2.28}$$

Note that x_g varies only at the group level g. It is possible that individuals in the same group exhibit some form of correlation:

$$\mathbb{E}[e_{ig}e_{jg}] = \rho_e \sigma_e^2 > 0. \tag{2.29}$$

ρ_e is the residual intraclass correlation, σ_e^2 is the residual variance.

We can assume that the error term is composed by:

$$e_{ig} = v_g + \eta_{ig}, \tag{2.30}$$

that is, a group-specific term added to an idiosyncratic disturbance.

Given this structure,

$$\rho_e = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_\eta^2} \tag{2.31}$$

Let $V_c(\hat{\beta}_1)$ be the usual OLS variance, and $V(\hat{\beta}_1)$ be the correct variance given the error structure. We divide both terms, which yields:

$$\frac{V\left(\hat{\beta}_1\right)}{V_c\left(\hat{\beta}_1\right)} = 1 + (n-1)\rho_e, \tag{2.32}$$

where n denotes the size of the groups and the regressors are fixed at the group level.

Taking the square root of this expression yields the *Moulton factor* for this special case.

$$\sqrt{\frac{V\left(\hat{\beta}_1\right)}{V_c\left(\hat{\beta}_1\right)}} = \sqrt{1 + (n-1)\rho_e} \tag{2.33}$$

More generally, we allow the regressors x_{ig} to vary at the individual level and for groups of different sizes. Then, the Moulton factor is the square root of:

$$\frac{V\left(\hat{\beta}_1\right)}{V_c\left(\hat{\beta}_1\right)} = 1 + \left[\frac{V\left(n_g\right)}{\bar{n}} + \bar{n} - 1\right]\rho_x \rho_e, \tag{2.34}$$

where $\bar{n}$ is the average group size and ρ_x is the intraclass correlation of x_ig:

$$\rho_x = \frac{\sum_g \sum_j \sum_{i \neq j} \left(x_{ig} - \bar{x}\right)\left(x_{ig} - \bar{x}\right)}{V\left(x_{ig}\right) \sum_g n_g \left(n_g - 1\right)}. \tag{2.35}$$

2.6.2 Clustering corrections

There are some possible solutions for the Moulton problem:

1. **Parametric.** Fix conventional standard errors with 2.34 and the intraclass correlation coefficients.

2. **Clustered standard errors.**

$$\hat{\Omega}_{cl} = (X'X)^{-1}\left(\sum_g X_g \hat{\Psi}_g X_g\right)(X'X)^{-1}, \tag{2.36}$$

   where

$$\begin{aligned} \hat{\Psi}_g &= a\hat{e}_g\hat{e}_g' \\ &= a\begin{bmatrix} \hat{e}_{1g}^2 & \hat{e}_{1g}\hat{e}_{2g} & \cdots & \hat{e}_{1g}\hat{e}_{ng}g \\ \hat{e}_{1g}\hat{e}_{2g} & \hat{e}_{2g}^2 & \cdots & \vdots \\ \vdots & \vdots & & \hat{e}_{n_g-1,g}\hat{e}_{ng}g \\ \hat{e}_{1g}\hat{e}_{n_g g} & \cdots & \hat{e}_{n_g-1,g}\hat{e}_{n_g g} & \hat{e}_{n_g g}^2 \end{bmatrix}. \end{aligned} \tag{2.37}$$

   X_g is the matrix of regressors for group g and a is the degrees of freedom correction.

   The more general Ω_{cl}, then, is *block-diagonal*:

$$\text{Var}(e) = \begin{pmatrix} \Sigma_1 & & & & 0 \\ & \ddots & & & \\ & & \Sigma_g & & \\ & & & \ddots & \\ 0 & & & & \Sigma_G \end{pmatrix} \tag{2.38}$$ [5]

3. **Use group averages.** Our new model is

$$\bar{y_g} = \beta_0 + \beta_1 x_g + \bar{e}_g. \tag{2.39}$$

4. **Block bootstrap.** This is basically resampling while maintaining the intraclass dependence structure.

5. **GLS or MLE.**

2.7 Measurement error

The fixed effects approach can be very susceptible to measurement error. This comes especially from variables that usually vary litte over time, such as marital status or institutions. Time variations observed for such variables can reflect mostly noise. However, since FE depends on these variations, be it through time demeaning or first differencing, this noise becomes substantial in the estimating procedures – which leads to large attenuation biases.

[5] http://fmwww.bc.edu/repec/usug2007/crse04.pdf

> It is widely believed in econometrics that the di¤erencing and FE transformations exacerbate measurement error bias (even though they eliminate heterogeneity bias). However, it is important to know that this conclusion rests on the classical errors-in-variables (CEV) model under strict exogeneity, as well as on other assumptions. (AW, p. 365)

Consider the model

$$y_{it} = \beta x_{it}^* + c_i + u_{it} \tag{2.40}$$

where we assume strict exogeneity

$$\mathbb{E}\left(u_{it} \mid \mathbf{x}_i^*, \mathbf{x}_i, c_i\right) = 0, \quad t = 1, 2, \ldots, T. \tag{2.41}$$

x^* is the unobserved true variable, and x is the observed variable. The measurement error is defined as

$$r_{it} = x_{it} - x_{it}^* \tag{2.42}$$

We now assume that r_{it} is uncorrelated with x_{it}^* – CEV. Then, the asymptotic results are:

$$\begin{aligned}\text{plim}_{N\to\infty}\,\hat{\beta}_{POLS} &= \beta + \frac{\text{Cov}\left(x_{it}, c_i + u_{it} - \beta r_{it}\right)}{\text{Var}\left(x_{it}\right)} \\ &= \beta + \frac{\text{Cov}\left(x_{it}, c_i\right) - \beta\sigma_r^2}{\text{Var}\left(x_{it}\right)}\end{aligned} \tag{2.43}$$

This shows that the pooled OLS estimator has two sources of bias: (i) unobserved effect c_i; (ii) measurement error (negative).

> If x_{it} and c_i are positively correlated and $\beta > 0$, the two sources of bias tend to cancel each other. (AW, p. 366)

Consider an example with $T = 2$ and where CEV holds. If we take first differences,

$$\begin{aligned}\text{plim}_{N\to\infty}\,\hat{\beta}_{FD} &= \beta + \frac{\text{Cov}\left(\Delta x_{it}, \Delta u_{it} - \beta\Delta r_{it}\right)}{\text{Var}\left(\Delta x_{it}\right)} = \beta - \beta\frac{\text{Cov}\left(\Delta x_{it}, \Delta r_{it}\right)}{\text{Var}\left(\Delta x_{it}\right)} \\ &= \beta - 2\beta\frac{\left[\sigma_r^2 - \text{Cov}\left(r_{it}, r_{i,t-1}\right)\right]}{\text{Var}\left(\Delta x_{it}\right)} \\ &= \beta\left(1 - \frac{\sigma_r^2\left(1-\rho_r\right)}{\sigma_{x^*}^2\left(1-\rho_{x^*}\right) + \sigma_r^2\left(1-\rho_r\right)}\right)\end{aligned} \tag{2.44}$$

> As the autocorrelation of x_{it}^* increases relative to that in r_{it}, the measurement error bias in $\hat{\beta}_{FD}$ increases. In fact, as $\rho_{x^*} \to 1$, the measurement error bias approaches $-\beta$.

It is possible to achieve consistent estimates even under measurement error with some assumptions.

$$y_{it} = \mathbf{z}_{it}\gamma + \delta w_{it}^* + c_i + u_{it}, \quad t = 1, 2, \ldots, T, \tag{2.45}$$

where w_{it}^* is measured with error, $r_{it} = w_{it} - w_{it}*$.

The assumptions for consistency are:

Assumption 2.7.1. ***Strict exogeneity for measurement error.***

$$\mathbf{E}\left(u_{it} \mid \mathbf{z}_i, \mathbf{w}_i^*, \mathbf{w}_i, c_i\right) = 0, \quad t = 1, 2, \ldots, T$$

Assumption 2.7.2. ***CEV.***

$$\mathbb{E}\left(r_{it} \mid \mathbf{z}_i, \mathbf{w}_i^*, c_i\right) = 0, \quad t = 1, 2, \ldots, T$$

Taking first differences yields:

$$\Delta y_{it} = \Delta \mathbf{z}_{it}\gamma + \delta \Delta w_{it} + \Delta u_{it} - \delta \Delta r_{it} \tag{2.46}$$

Now, CEV implies that Δr_{it} is correlated with Δw_{it}. We need an instrument for Δw_{it}. A natural choice is another measure of Δw_{it}^*, h_{it}, whose measurement error is orthogonal to the measurement error in $w_{is}, \forall t, s$[6].

[6] More info on AW, 11.5.

Chapter 3

Differences-in-Differences Estimator

Differences-in-Differences (DD) is another framework for data spanning multiple periods. Most importantly, it doesn't rely on *panel* data, which requires observations for the same individuals over time. Often, it uses data at an aggregate level for its parameters of interest.

> DD is a version of fixed effects estimation using aggregate data. (AP, p. 228)

3.1 DD framework

DD will be presented with an example from Card and Krueger (1994) (AP, 5.2).

Let y_{1ist} be fast food employment at restaurant i, state s and time t *if there's a high state minimum vage* (1); y_{0ist} *if there's a low state minimum vage* (0). Therefore, 0 or 1 denote treatment status.

We'd ideally like to observe y_{1ist} and y_{0ist} for all states. However, that's obviously not possible, since the counterfactual for treated states won't be observed in these states.

The main assumption in DD is its additive effects structure.

Assumption 3.1.1. ***Additive effects for DD.*** *$\mathbb{E}[y_{0ist}|s,t] = \gamma_s + \lambda_t$, where γ_s is a state-specific, time-invariant effect and λ_t is a time-specific, state-invariant effect.*

Note that γ_s is a state-level fixed effect.

Now, we define the dummy D_{st} for high minimum wage laws on state s at time t. We also assume that $\mathbb{E}(y_{1st} - y_{0st}|s,t)$ is a constant, denoted by δ. Then,

$$y_{ist} = \gamma_s + \lambda_t + \delta D_{st} + \varepsilon_{ist} \tag{3.1}$$

is our linear model.

As per usual, we assume strict exogeneity of the disturbances.

Assumption 3.1.2. ***Strict exogeneity for DD.*** $\mathbb{E}(\varepsilon_{ist}|s,t) = 0, \forall s,t.$

We first consider the time difference for both states:

$$\begin{aligned} &\mathbb{E}\left[\mathrm{Y}_{ist} \mid s = PA, t = \text{Nov}\right] - \mathbb{E}\left[\mathrm{Y}_{ist} \mid s = PA, t = Feb\right] \\ &\quad = \lambda_{Nov} - \lambda_{Feb} \end{aligned} \tag{3.2}$$

and

$$\begin{aligned} &\mathbb{E}\left[\mathrm{Y}_{ist} \mid s = NJ, t = \text{Nov}\right] - \mathbb{E}\left[\mathrm{Y}_{ist} \mid s = NJ, t = Feb\right] \\ &\quad = \lambda_{Nov} - \lambda_{Feb} + \delta \end{aligned} \tag{3.3}$$

Now, we also take differences *across states* (treatment and control):

$$\begin{aligned} &\{\mathbb{E}\left[\mathrm{Y}_{ist} \mid s = NJ, t = Nov\right] - \mathbb{E}\left[\mathrm{Y}_{ist} \mid s = NJ, t = Feb\right]\} \\ &\quad - \{\mathbb{E}\left[\mathrm{Y}_{ist} \mid s = PA, t = Nov\right] - \mathbb{E}\left[\mathrm{Y}_{ist} \mid s = PA, t = Feb\right]\} = \delta \end{aligned} \tag{3.4}$$

which yields the causal effect of interest.

The key identifying assumption for causality here is the idea of **parallel trends** between control and treatment groups. This means that the control group will provide a way to calculate an estimate of the counterfactual of the treatment state by removing the time trend and considering the state-specific unobserved effects.

> Treatment induces a deviation from this common trend, as illustrated in figure 5.2.1. Although the treatment and control states can differ, this difference is meant to be captured by the state fixed effect, which plays the same role as the unobserved individual effect in [FE]. (AP, p. 230)

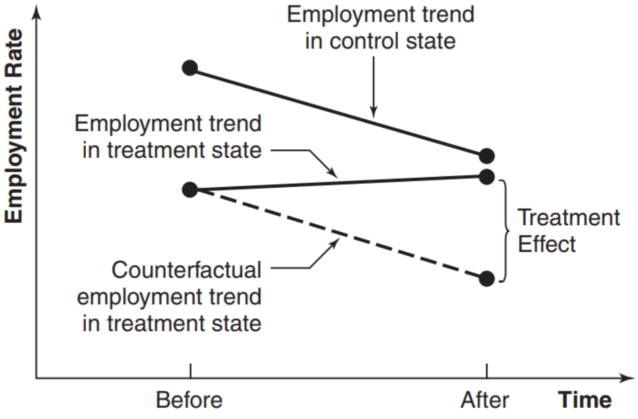
Employment trend in control state
Employment trend in treatment state
Treatment Effect
Counterfactual employment trend in treatment state
Employment Rate
Before
After
Time

Figure 5.2.1 Causal effects in the DD model.

This assumption can be investigated with data on multiple periods. Card and Krueger update their study with time series data on employment, which indicates that Pennsylvania wasn't very fit for a control, due to state-specific shocks and overall variations that are not related systematically to New Jersey (treatment).

3.2 Regression DD

It is also possible to use DD in a regression framework, also controlling for relevant variables that could help make a stronger case for causality. For example, consider

$$\mathrm{Y}_{ist} = \alpha + \gamma NJ_s + \lambda d_t + \delta \left(NJ_s \cdot d_t \right) + \varepsilon_{ist} \tag{3.5}$$

where NJ_s is a dummy for restaurants in New Jersey. The interaction term now has the same role as D_{st} in the previous model.

This model is saturated, which means that the conditional mean takes four possible values:

$$\begin{aligned}
\alpha &= E\left[\mathrm{Y}_{ist} \mid s = PA, t = Feb\right] = \gamma_{PA} + \lambda_{Feb} \\
\gamma &= E\left[\mathrm{Y}_{ist} \mid s = NJ, t = Feb\right] - E\left[\mathrm{Y}_{ist} \mid s = PA, t = \mathrm{Feb}\right] \\
&= \gamma_{NJ} - \gamma_{PA} \\
\lambda &= E\left[\mathrm{Y}_{ist} \mid s = PA, t = Nov\right] - E\left[\mathrm{Y}_{ist} \mid s = PA, t = Feb\right] \\
&= \lambda_{Nov} - \lambda_{Feb} \\
\delta &= \left\{E\left[\mathrm{Y}_{ist} \mid s = NJ, t = Nov\right] - E\left[\mathrm{Y}_{ist} \mid s = NJ, t = Feb\right]\right\} \\
&- \left\{E\left[\mathrm{Y}_{ist} \mid s = PA, t = Nov\right] - E\left[\mathrm{Y}_{ist} \mid s = PA, t = Feb\right]\right\}.
\end{aligned} \tag{3.6}$$

Regression DD helps to conveniently estimate standard errors, allow for multiple controls and treatments, and treatments that are not binary – i.e., that do not involve dummies, but instead some variation. For example:

$$\mathrm{Y}_{ist} = \gamma_s + \lambda_t + \delta \left(\mathrm{FA}_s \cdot d_t \right) + \varepsilon_{ist} \tag{3.7}$$

where FA_s measures the fraction of teenagers likely to be affected by a minimum wage increase in each state (Card, 1992; AP, p. 235).

It is also possible to add controls to the regression[1]:

$$\mathrm{Y}_{ist} = \gamma_s + \lambda_t + \delta \left(\mathrm{FA}_s \cdot d_t \right) + \mathrm{X}'_{ist}\beta + \varepsilon_{ist}. \tag{3.8}$$

Another useful addition in the model for robustness is *state-specific time trends*, which allow for states to follow (linearly) different trends:

$$\mathrm{Y}_{ist} = \gamma_{0s} + \gamma_{1s} t + \lambda_t + \delta \mathrm{D}_{st} + \mathrm{X}'_{ist}\beta + \varepsilon_{ist} \tag{3.9}$$

Here, γ_{0s} represents different intercepts for each state. Note that this requires at least 3 periods for estimation – and is still a low number for adequate inference.

[1] AP, p. 237 discusses Granger causality tests, that are particularly suited for regression DD – as it is possible to check the relevance of anticipatory or lagged effects.

3.2.1 Potential problems in DD design

DD always, be it implicitly of explicitly, takes the form of a control-treament comparison. This implies that it is necessary to correctly consider difficulties regarding the composition and comparison of these groups.

The composition of groups may change after treatment. For example, studying the effects of welfare programs for income maintenance may be specially difficult because, when such a program is implemented in a state, poor people in neighboring states that aren't attached to his/her land might move to the treated state. This would cause a downward bias for income maintenance programs in regards to employment.

A possible way to fix this problem is to use the initial location (for example) as an instrument for the actual location in the model.[2]

3.3 Clustering and serial correlation for DD

Consider the standard DD model with additive state and time effects.

$$\mathrm{Y}_{ist} = \gamma_s + \lambda_t + \delta \mathrm{D}_{st} + \varepsilon_{ist} \tag{3.10}$$

We can consider that ε_{ist} is the sum of a state-year shock, v_{st}, and an idiosyncratic disturbance, η_{ist}. In this case, v_{st} could reflect some regional shock like a business cycle.

Rewriting the model yields

$$\mathrm{Y}_{ist} = \gamma_s + \lambda_t + \delta \mathrm{D}_{st} + v_{st} + \eta_{ist} \tag{3.11}$$

As per usual, we assume $\mathbb{E}(v_{st}) = 0, \mathbb{E}(\eta_{ist}|s,t) = 0$.

> As with the Moulton problem, state- and time-specific random effects generate a clustering problem that affects statistical inference. But that might be the least of our problems in this case. (AP, p. 317)

Aside from the clustering problem, we also may have an inconsistent estimate of the causal effect. The empirical For Card and Krueger (1994), the DD estimator is:

$$\hat{\delta}_{CK} = \left(\overline{\mathrm{Y}}_{s=NJ,t=Nov} - \overline{\mathrm{Y}}_{s=NJ,t=Feb}\right) - \left(\overline{\mathrm{Y}}_{s=PA,t=Nov} - \overline{\mathrm{Y}}_{s=PA,t=Feb}\right) \tag{3.12}$$

This is unbiased, since $\mathbb{E}(v_{st}) = \mathbb{E}(\eta_{ist}) = 0$. However, consider the asymptotic argument for increasing group size:

$$\operatorname{plim} \hat{\delta}_{CK} = \delta + \left\{(v_{s=NJ,t=Nov} - v_{s=NJ,t=Feb}) - (v_{s=PA,t=Nov} - v_{s=PA,t=Feb})\right\}.$$

> Averaging larger and larger samples within New Jersey and Pennsylvania in a pair of periods does nothing to eliminate the regional shocks specific to a

[2] More on AP, p. 242-3.

> given location and period. With only two states and years, we have no way to distinguish the differences-in-differences generated by a policy change from the difference-in-dfferences due to the fact that, say, the New Jersey economy was holding steady in 1992 while Pennsylvania was experiencing a cyclical downturn. *The presence of vst amounts to a failure of the common trends assumption discussed in section 5.2.* (AP, p. 318)

A possible solution to this problem is to analyze multiple groups over multiple time periods, *hoping that v_{st} average out to zero.* Note that it is important, here, to have *multiple groups,* not the group size.

Inference in DD models relies heavily on the behavior of v_{st}. If we consider it to be independent across groups and over time, then we go back to the Moulton problem discussed in FE. However, in most cases it is unreasonable to assume that $\{v_{st}\}$ isn't serially correlated.

> Almost certainly, for example, regional shocks are highly serially correlated: if things are bad in Pennsylvania in one month, they are likely to be about as bad in the next. [...] Any research design with a group structure where the group means are correlated can be said to have the serial correlation problem. (AP, p. 318)

This means that, aside from the correction needed because of the correlation within groups that the presence v_{st} implies, it is necessary to correct for the serial dependence of $\{v_{st}\}$. The simplest way to tackle this issue is to cluster one level higher – for example, clustering for state instead of state and year.

The main issue with this approach is that it reduces – often substantially – the number of groups in our sample, which may hinder inference.

> With few clusters, we tend to underestimate either the serial correlation in a random shock like v_{st} or the intraclass correlation, ρ_e, in the Moulton problem. *The relevant dimension* for counting clusters in the Moulton problem *is the number of groups, G.* (AP, p. 319)

What can we do when the cluster count is low?

1. **Get more clusters by collecting more data.** This is obvioulsy the first-best solution, but not always feasible.

2. **Bias correction of clustered standard errors.** This means, in essence, inflating standard errors hoping to reduce bias, as in small samples (i.e., low G), $\mathbb{E}(\hat{e}_g \hat{e}_g') \neq \mathbb{E}(e_g e_g')$.[3]

3. **Recognizing that the fundamental unit of observation is a cluster and not an individual unit within clusters.** This widens the confidence intervals for

[3] More on AP, p. 320.

testing and will help avoding inference mistakes.

4. **Using group means instead of individual data.** The level of aggregation is the level at which you'd like to cluster [...]. For serial correlation, this is the state, but state averages cannot be used to estimate a model with a full set of state effects. Also, since treatment status varies within states, averaging up to the state level averages the regressor of interest as well, changing the rules of the game in a way we may not like (the estimator becomes IV using group dummies as instruments). The group means approach is therefore out of bounds for the serial correlation problem. (AP, p. 321)

5. **Block boostrap.**

6. **Parametric corrections.** For the Moulton problem, this amounts to use of the Moulton factor. With serial correlation, this means correcting your standard errors for first-order serial correlation at the group level. (AP, p. 322)

Chapter 4

Random Effects Estimator

The random effects framework is an alternative way of dealing with the unobserved effect c_i. As with pooled OLS, RE considers c_i as part of the error term.

> In fact, random effects analysis imposes more assumptions than those needed for pooled OLS: *strict exogeneity*[1] in addition to orthogonality between c_i and x_{it}. (AW, p. 292)

4.1 Consistency

Assumption 4.1.1. ***Strict exogeneity for RE.***

1. $\mathbb{E}(u_{it}|\mathbf{x_i}, c_i) = 0, t = 1, ..., T$;
2. $\mathbb{E}(c_i|\mathbf{x_i}) = \mathbb{E}(c_i) = 0$.

Note that this is a very strong assumption: it requires *strict* exogeneity for all regressors (1) *and* for the unobserved effects (2) for all periods.

For consistency, $\mathbb{E}\left(\mathbf{x}'_{it} c_i\right) = \mathbf{0}, \quad t = 1, 2, \ldots, T$ is a sufficient condition instead of (2); however, it doesn't afford much more generality, so (2) will be used (AW).

The advantage of imposing a more restrictive assumption than pooled OLS is the ability to exploit the serial correlation in the composite error term,$v_{it} = c_i + u_{it}$, in a GLS framework. But GLS requires strict exogeneity for consistency.

Under 4.1.1, the model can be rewritten as

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + v_{it}, \tag{4.1}$$

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + v_{it} \tag{4.2}$$

[1] Remember that pooled OLS needs only contemporaneous exogeneity for consistency.

Writing the model for all time periods yields

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{v}_i \tag{4.3}$$

From this, $\mathbf{v_i}$ can be decomposed as $\mathbf{v_i} = c_i\mathbf{j_T} + \mathbf{u_i}$. We now define the unconditional variance matrix of $\mathbf{v_i}$ as

$$\boldsymbol{\Omega} := \mathbb{E}(\mathbf{v_i}\mathbf{v_i'}),$$

a $T \times T$ positive definite matrix.

This motivates our usual second assumption, i.e., the GLS full rank assumption.

Assumption 4.1.2. ***GLS full rank for RE.*** $\operatorname{rank}\mathbb{E}\left(\mathbf{X}_i'\boldsymbol{\Omega}^{-1}\mathbf{X}_i\right) = K$.

Under 4.1.1 and 4.1.2, (F)GLS is consistent. We could use a general unrestricted form for Ω; however, this doesn't exploit the particular variance structure of the RE framework.

4.2 Inference

4.2.1 RE variance matrix

First, assume that $\mathbb{E}(u_{it}^2) = \sigma_u^2, \quad t = 1, 2, ..., T$ – i.e., constant unconditional varianec across time.

The second assumption is that the idiosyncratic disturbances are not serially correlated:

$$\mathbb{E}(u_{it}u_{is}) = 0, \forall t \neq s.$$

Aside from the formal consistency assumptions, these two conditions, central to the RE estimator, imply that

$$\mathbb{E}\left(v_{it}^2\right) = \mathbb{E}\left(c_i^2\right) + 2\mathbb{E}\left(c_i u_{it}\right) + \mathbb{E}\left(u_{it}^2\right) = \sigma_c^2 + \sigma_u^2 \tag{4.4}$$

where $2\mathbb{E}\left(c_i u_{it}\right) = 0$ from strict exogeneity, 4.1.1. Furthermore,

$$\mathbb{E}\left(v_{it}v_{is}\right) = \mathbb{E}\left[\left(c_i + u_{it}\right)\left(c_i + u_{is}\right)\right] = \mathbb{E}\left(c_i^2\right) = \sigma_c^2. \tag{4.5}$$

These results can be used to build the entire $\boldsymbol{\Omega}$ variance-covariance matrix for $\mathbf{v_i}$:

$$\boldsymbol{\Omega} = \mathrm{E}\left(\mathbf{v}_i\mathbf{v}_i'\right) = \begin{pmatrix} \sigma_c^2 + \sigma_u^2 & \sigma_c^2 & \cdots & \sigma_c^2 \\ \sigma_c^2 & \sigma_c^2 + \sigma_u^2 & \cdots & \vdots \\ \vdots & & \ddots & \sigma_c^2 \\ \sigma_c^2 & & & \sigma_c^2 + \sigma_u^2 \end{pmatrix}, \tag{4.6}$$

which can be alternatively written as $\boldsymbol{\Omega} = \sigma_u^2\mathbf{I}_T + \sigma_c^2\mathbf{j}_T\mathbf{j}_T'$.

When $\boldsymbol{\Omega}$ has this form, we say that it has the *random effects structure.* The advantage of such $\boldsymbol{\Omega}$ is that it only requires two parameters for estimation, namely, $\sigma_c^2, sigma_u^2$. This greatly improves inference via degrees of freedom.

We also know the serial correlation of the composite errors:

$$\operatorname{Corr}\left(v_{is}, v_{it}\right) = \sigma_c^2 / \left(\sigma_c^2 + \sigma_u^2\right) \geq 0, s \neq t. \tag{4.7}$$

Note that this doesn't tend to zero as t, s get far apart.

> Unlike standard models for serial correlation in time series settings, the random effects assumption implies strong persistence in the unobservables over time, due, of course, to the presence of c_i. (AW, p. 294)

We sum up all of the assumptions required for the RE variance matrix in the following assumption.

Assumption 4.2.1. ***Random effects variance structure.***

1. $\mathbb{E}\left(\mathbf{u}_i \mathbf{u}_i' \mid \mathbf{x}_i, c_i\right) = \sigma_u^2 \mathbf{I}_T$;
2. $\mathbb{E}\left(c_i^2 \mid \mathbf{x}_i\right) = \sigma_c^2$.

Note that (1) implies the unconditional variance of σ_u^2 and (2) implies no serial correlation of the idiosyncratic disturbances.

For FGLS, we define $\sigma_v^2 = \sigma_c^2 + \sigma_u^2$. Suppose that we have consistent estimators for σ_c^2, σ_u^2 (as will be the case). Then,

$$\hat{\boldsymbol{\Omega}} \equiv \hat{\boldsymbol{\sigma}}_u^2 \mathbf{I}_T + \hat{\sigma}_c^2 \mathbf{j}_T \mathbf{j}_T' \tag{4.8}$$

is a positive definite $T \times T$ matrix.

From this, we can define the *random effects estimator:*

$$\hat{\boldsymbol{\beta}}_{RE} = \left(\sum_{i=1}^N \mathbf{X}_i' \hat{\boldsymbol{\Omega}}^{-1} \mathbf{X}_i\right)^{-1} \left(\sum_{i=1}^N \mathbf{X}_i' \hat{\boldsymbol{\Omega}}^{-1} \mathbf{y}_i\right). \tag{4.9}$$

It is important to highlight (again) that $\hat{\beta}_{RE}$ *is consistent regardless of the form of* Ω – which implies that 4.2.1 isn't necessary for consistency.

We can estimate $\hat{\sigma}_c^2, \hat{\sigma}_u^2$ by first estimating $\hat{\sigma}_v^2$.

$$\hat{\sigma}_v^2 = \frac{1}{(NT - K)} \sum_{i=1}^N \sum_{t=1}^T \check{v}_{it}^2 \tag{4.10}$$

is a consistent estimator for $\hat{\sigma}_v^2$, where $\check{v}_{it}$ denotes the pooled OLS residuals.

Then, using the assumptions above (no serial correlation for u_{it}, u_{is}) and the $T(T-1)/2$ nonredundant error products, we estimate $\hat{\sigma}_c^2$.

$$\begin{aligned}\mathrm{E}\left(\sum_{t=1}^{T-1} \sum_{s=t+1}^T v_{it} v_{is}\right) &= \sum_{t=1}^{T-1} \sum_{s=t+1}^T \mathrm{E}\left(v_{it} v_{is}\right) = \sum_{t=1}^{T-1} \sum_{s=t+1}^T \sigma_c^2 = \sigma_c^2 \sum_{t=1}^{T-1} (T - t) \\ &= \sigma_c^2((T-1) + (T-2) + \cdots + 2 + 1) = \sigma_c^2 T(T-1)/2,\end{aligned} \tag{4.11}$$

The sample equivalent, then, is:

$$\hat{\sigma}_c^2 = \frac{1}{[NT(T-1)/2 - K]} \sum_{i=1}^{N} \sum_{t=1}^{T-1} \sum_{s=t+1}^{T} \check{v}_{it}\check{v}_{is} \tag{4.12}$$

where we again use $\check{v}_{it}$, the pooled OLS residuals, and correct for the degrees of freedom.

Now that we have $\hat{\sigma}_c^2$ and $\hat{\sigma}_v^2$, we can calculate $\hat{\sigma}_u^2 = \hat{\sigma}_v^2 - \hat{\sigma}_c^2$.

> As a practical matter, equation $[\hat{\sigma}_c^2]$ is not guaranteed to be positive, although it is in the vast majority of applications. A negative value for $\hat{\sigma}_c^2$ is indicative of negative serial correlation in u_{it}, probably a substantial amount, which means that Assumption RE.3a [RE variance structure, conditional homoskedasticity] is violated. (AW, p. 296)

4.2.2 A robust variance matrix estimator for RE

If assumption 4.2.1 fails, we can simply use

$$Avar(\hat{\beta}) = \left(\sum_{i=1}^{N} \mathbf{X}_i'\hat{\mathbf{\Omega}}^{-1}\mathbf{X}_i\right)^{-1} \left(\sum_{i=1}^{N} \mathbf{X}_i'\hat{\mathbf{\Omega}}^{-1}\hat{\mathbf{v}}_i\hat{\mathbf{v}}_i'\hat{\mathbf{\Omega}}^{-1}\mathbf{X}_i\right) \left(\sum_{i=1}^{N} \mathbf{X}_i'\hat{\mathbf{\Omega}}^{-1}\mathbf{X}_i\right)^{-1} \tag{4.13}$$

and use the usual robust Wald statistics $W = (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})' \left(\mathbf{R}\hat{\mathbf{V}}\mathbf{R}'\right)^{-1} (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})$.

4.2.3 General Feasible GLS

If $\{u_{it}\}$ are generally heteroskedastic and serially correlated, a general estimator for Ω is

$$\hat{\mathbf{\Omega}} = N^{-1} \sum_{i=1}^{N} \check{\mathbf{v}}_i\check{\mathbf{v}}_i' \tag{4.14}$$

where $\check{\mathbf{v}}_i$ is the pooled OLS residuals. This is an asymptotically efficient variance estimator.

This is also more general than the RE framework. It isn't used very much for a couple of reasons. First, the idiosyncratic disturbances $\{u_{it}\}$ are usually considered axiomatically not serially correlated; any serial correlation would come from c_i. Furthermore, this general strategy requires estimation of many more parameters, which can yield poor finite sample properties. $\hat{\Omega}$ here requires the estimation of $T(T+1)/2$ parameters, while the RE matrix requires only estimation of $\hat{\sigma_c^2}$ and $\hat{\sigma_u^2}$.

4.3 Testing for unobserved effects

If all of the RE assumptions hold but the true model does not contain c_i, then pooled OLS is efficient and consistent. This can be tested with the null $H_0 : \sigma_c^2 = 0$, using a simple AR(1) model. This is valid because, under the null, v_{it} is serially uncorrelated.

Breusch and Pagan also developed a test based on a Lagrange Multiplier. The test statistic is[2]

$$\frac{\sum_{i=1}^{N}\sum_{t=1}^{T-1}\sum_{s=1+1}^{T}\hat{v}_{it}\hat{v}_{is}}{\left[\sum_{i=1}^{N}\left(\sum_{t=1}^{T-1}\sum_{s=t+1}^{T}\hat{v}_{it}\hat{v}_{is}\right)^{2}\right]^{1/2}} \tag{4.15}$$

[2] More on AW, p. 300.

Chapter 5

Testing and comparing estimators

In this chapter, we develop a methodology that guides the selection of the correct estimator for a given model and shows the relations between different estimators.

5.1 Comparing panel estimators

5.1.1 FE and FD for different assumption violations

First, when $T = 2$, FE and FD are *identical* in all respects (as time demeaning is equivalent to first differencing.) When $T > 2$, our choice depends on the assumptions regarding $\{u_{it}\}$, specifically its serial dependence. FD is more efficient when $\{u_{it}\}$ follows a random walk; FE takes the opposite stance, assuming that $\{u_{it}\}$ has no serial correlation.[1]

However, if strict exogeneity does not hold, and u_{it} is correlated with $\mathbf{x}_{is}$, for some t, s, FE and FD usually yield *different probability limits.* In fact, any of the usual endogeneity problems, like OVB (for time-varying variables), simultaneity, measurement error, etc. will cause the two probability limits to differ.

If x_{it}, u_{it} are correlated, we then no longer have *contemporaneous exgoeneity* – necessary even for pooled OLS (1.2.1) –, which yields inconsistent estimators. If u_{it} is correlated with $\mathbf{x}_{is}$, $t \neq s$, the estimates will also be inconsistent. When u_{it} is correlated with lags of $\mathbf{x}_{it}$, a possible way to solve this is by adding lags as regressors. The same can be said for correlation between u_{it} and future values of $\mathbf{x}_{it}$, although adding future regressors does not yield usual interpretations in an economic model.

We now attempt to analyze the possible biases of FE and FD under some endogeneity problems.

First, assume that contemporaneous exogeneity still holds, i.e.,

$$\mathbb{E}(x'_{it}u_{it}) = 0, \forall t \tag{5.1}$$

[1] This section seems not that important for our purposes.

Under 2.4.1,

$$\text{plim}_{N\to\infty}\left(\hat{\boldsymbol{\beta}}_{FE}\right) = \boldsymbol{\beta} + \left[T^{-1}\sum_{t=1}^{T}\mathrm{E}\left(\ddot{\mathbf{x}}_{ii}'\ddot{\mathbf{x}}_{it}\right)\right]^{-1}\left[T^{-1}\sum_{t=1}^{T}\mathrm{E}\left(\ddot{\mathbf{x}}_{it}'u_{it}\right)\right]^{-1} \tag{5.2}$$

Under contemporaneous exogeneity,

$$\mathrm{E}\left(\ddot{\mathbf{x}}_{it}'u_{it}\right) = \mathrm{E}\left[\left(\mathbf{x}_{it}-\overline{\mathbf{x}}_i\right)'u_{it}\right)\right] = -\mathrm{E}\left(\overline{\mathbf{x}}_i'u_{it}\right) \tag{5.3}$$

and so

$$T^{-1}\sum_{t=1}^{T}\mathrm{E}\left(\ddot{\mathbf{x}}_{it}'u_{it}\right) = -T^{-1}\sum_{t=1}^{T}\mathrm{E}\left(\overline{\mathbf{x}}_i'u_{it}\right) = -\mathrm{E}\left(\overline{\mathbf{x}}_i'\bar{u}_i\right) \tag{5.4}$$

If $\{\mathbf{x_it}, u_{it}\}$ is stable and weakly dependent, then both of these average moments are bounded.

The probability limit of $\hat{\beta}_{FE}$ is[2]

$$\text{plim}_{N\to\infty}\left(\hat{\boldsymbol{\beta}}_{FE}\right) = \boldsymbol{\beta} + O(1)\cdot O\left(T^{-1}\right) = \boldsymbol{\beta} + O\left(T^{-1}\right) \equiv \boldsymbol{\beta} + \mathbf{r}_{FE}(T), \tag{5.5}$$

where $\mathbf{r}_{FE}(T) = O(T^{-1})$. This means, Intuitively, that the inconsistency in the FE estimator can be small if T is large.

For the FD estimator, the general plim is:

$$\text{plim}_{N\to\infty}\left(\hat{\boldsymbol{\beta}}_{FD}\right) = \boldsymbol{\beta} + \left[(T-1)^{-1}\sum_{t=2}^{T}\mathrm{E}\left(\Delta\mathbf{x}_{it}'\Delta\mathbf{x}_{it}\right)\right]^{-1}\left[(T-1)^{-1}\sum_{t=2}^{T}\mathrm{E}\left(\Delta\mathbf{x}_{it}'\Delta u_{it}\right)\right]^{-1} \tag{5.6}$$

Under weak dependence of $\{\mathbf{x_it}\}$, the first average of the above equation is bounded. Under contemporaneous exogeneity,

$$\mathrm{E}\left(\Delta\mathbf{x}_{it}'\Delta u_{it}\right) = -\left[\mathrm{E}\left(\mathbf{x}_{it}'u_{i,t-1}\right) + \mathrm{E}\left(\mathbf{x}_{i,t-1}'u_{it}\right)\right], \tag{5.7}$$

which is usually $\neq 0$ – even as T grows[3].

> The previous analysis shows that under contemporaneous exogeneity and weak dependence of the regressors and idiosyncratic errors, the FE estimator has an advantage over the FD estimator when T is large. (AW, p. 324)

This analysis favors FE *when contemporaneous exogeneity holds but strict exogeneity fails*, even if $\{y_{it}\}$ and some elements of $\{\mathbf{x}_{it}\}$ have unit roots. There's a catch, though: it depends on the critical assumption that $\{u_{it}\}$ is I(0) – i.e., $y_{it}, \mathbf{x}_{it}$ must be *cointegrated.* If that's not the case, then FE is no longer superior to FD in regards to inconsistency, and FE's inconsistency will grow as T gets large. FD, by contrast, removes any unit roots, which rules out the possibility of spurious regressions.

[2] This is much more detailed on AW, p. 322-3.

[3] Again, more on AW, p. 323. This is probably beyond the scope of the course.

5.1.2 Relationship between RE and FE

IF X varies little over time, FE and FD may lead to imprecise estimates, forcing us to use RE to learn anything about our parameters. If RE is appropriate, then it has much smaller variance than FE.

We can rewrite Ω under RE using the fact that $\mathbf{j_T j_T}' = T$:

$$\begin{aligned}\boldsymbol{\Omega} &= \sigma_u^2\mathbf{I}_T + \sigma_c^2\mathbf{j}_T\mathbf{j}_T' = \sigma_u^2\mathbf{I}_T + T\sigma_c^2\mathbf{j}_T\left(\mathbf{j}_T'\mathbf{j}_T\right)^{-1}\mathbf{j}_T' \\ &= \sigma_u^2\mathbf{I}_T + T\sigma_c^2\mathbf{P}_T = \left(\sigma_u^2 + T\sigma_c^2\right)\left(\mathbf{P}_T + \eta\mathbf{Q}_T\right)\end{aligned} \tag{5.8}$$

where $\mathbf{P}_T := \mathbf{I}_T - \mathbf{Q}_T = \mathbf{j}_T\left(\mathbf{j}_T'\mathbf{j}_T\right)^{-1}\mathbf{j}_T'$[4] and $\eta := \sigma_u^2/\left(\sigma_u^2 + T\sigma_c^2\right)$. Now, define $\mathbf{S}_T := \mathbf{P}_T + \eta\mathbf{Q}_T$.

Then, $\mathbf{S}_T^{-1} = \mathbf{P}_T + (1/\eta)\mathbf{Q}_T$ and $\mathbf{S}_T^{-1/2} = \mathbf{P}_T + (1/\sqrt{\eta})\mathbf{Q}_T$ ($\mathbf{S_T}, \mathbf{S_T^{-1/2}}$ are symmetric, since $\mathbf{P_T}, \mathbf{Q_T}$ are symmetric.)

$\mathbf{S}_T^{-1/2} = (1-\lambda)^{-1}\left[\mathbf{I}_T - \lambda\mathbf{P}_T\right]$, where $\lambda = 1 - \sqrt{\eta}$.

We now write $\boldsymbol{\Omega}^{-1/2}$ as

$$\boldsymbol{\Omega}^{-1/2} = \left(\sigma_u^2 + T\sigma_c^2\right)^{-1/2}(1-\lambda)^{-1}\left[\mathbf{I}_T - \lambda\mathbf{P}_T\right] = (1/\sigma_u)\left[\mathbf{I}_T - \lambda\mathbf{P}_T\right], \tag{5.9}$$

where $\lambda = 1 - \left[\sigma_u^2/\left(\sigma_u^2 + T\sigma_c^2\right)\right]^{1/2}$.

If we know λ, we can effectively get the RE estimator by estimating the transformed equation

$$\mathbf{C}_T\mathbf{y}_i = \mathbf{C}_T\mathbf{X}_i\boldsymbol{\beta} + \mathbf{C_T v_i} \tag{5.10}$$

by system OLS, where $\mathbf{C}_T \equiv \left[\mathbf{I}_T - \lambda\mathbf{P}_T\right]$. The transformed equation is: $\breve{\mathbf{y}}_i = \breve{\mathbf{X}}_i\boldsymbol{\beta} + \breve{\mathbf{v}}_i$

The variance matrix of $\breve{\mathbf{v}}_i$ is $\mathrm{E}\left(\breve{\mathbf{v}}_i\breve{\mathbf{v}}_i'\right) = \mathbf{C}_T\boldsymbol{\Omega}\mathbf{C}_T = \sigma_u^2\mathbf{I}_T$. This is ideal for OLS (homoskedasticity).

Note that $\breve{\mathbf{y}}_i$ is $y_{it} - \lambda\bar{y}_i$, and similarly for $\breve{\mathbf{X}}_i$. We can then interpret this as simply estimating pooled OLS for:

$$y_{it} - \lambda\bar{y}_i = \left(\mathbf{x}_{it} - \lambda\overline{\mathbf{x}}_i\right)\boldsymbol{\beta} + \left(v_{it} - \lambda\bar{v}_i\right) \tag{5.11}$$

The feasible RE estimator replaces λ for $\hat{\lambda}$, using variance estimates from the above pooled OLS regression.

The RE estimator can be written as

$$\hat{\boldsymbol{\beta}}_{RE} = \left(\sum_{i=1}^{N}\sum_{t=1}^{T}\breve{\mathbf{x}}_{it}'\breve{\mathbf{x}}_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\breve{\mathbf{x}}_{it}'\breve{y}_{it}\right) \tag{5.12}$$

[4] P_T is the projection matrix.

> Equation [above] shows that the RE estimator is obtained by a *quasi-time demeaning*: rather than removing the time average from the explanatory and dependent variables at each t, RE estimation removes a *fraction of the time average.* If $\hat{\lambda}$ is close to unity, the RE and FE estimates tend to be close. (AW, p. 327)

$$\hat{\lambda} = 1 - \left\{1 / \left[1 + T\left(\hat{\sigma}_c^2 / \hat{\sigma}_u^2\right)\right]\right\}^{1/2} \tag{5.13}$$

When $T\left(\hat{\sigma}_c^2/\hat{\sigma}_u^2\right)$ is large, the second term in $\hat{\lambda}$ is small, so it is close to unity – and RE is close to FE. $\hat{\lambda} \to 1$ as $T \to \infty, \hat{\sigma}_c^2/\hat{\sigma}_u^2 \to \infty$.

> As λ approaches unity, the precision of the RE estimator approaches that of the FE estimator, and the effects of time-constant explanatory variables become harder to estimate. (AW, p. 327)

This shows that the inconsistency of RE can be small when σ_c^2 is large relative to σ_u^2 or if T is large.

In essence:

1. **Pooled OLS.** $\lambda = 0$.
2. **FE.** $\lambda = 1$.
3. **RE.** $\lambda = 1 - \left\{1/\left[1 + T\left(\hat{\sigma}_c^2/\hat{\sigma}_u^2\right)\right]\right\}^{1/2}, 0 < \lambda < 1$.

Between these estimators, we're varying how much is being time demeaned.

5.2 Hausman test

Hausman developed a test to systematically guide the decision between RE or FE frameworks. It is based on the difference between RE and FE estimates – since FE is consistent and RE isn't, when c_i is correlated with the regressors. This assumes that x_{it} is strictly exogenous with respect to u_{it}.

Some caveats:

- Strict exogeneity with respect to u_{it} holds under the null and the alternative. If it doesn't, then both estimates would be biased.
- The test is usually implemented with 4.2.1 (variance structure of RE) being valid under the null. It implies that RE is more efficient than FE and simplifies computation of the test statistic – but it does not yield a way to *test* this assumption. Failure of this assumption causes the Hausman test to have a non-standard limiting distribution. (AW, p. 329)
- We can only compare coefficients on *time-varying* regressors, because of the inherent restrictions that FE imposes.

Consider the model

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + c_i + u_{it} = \mathbf{z}_i\gamma + \mathbf{w}_{it}\boldsymbol{\delta} + c_i + u_{it}, \tag{5.14}$$

where $\mathbf{z_i}$ is a vector of intercept and time-invariant regressors, and $\mathbf{w_{it}}$ is a vector of time-varying covariates.

> A key component of the traditional Hausman test is showing that the asymptotic variance of the FE estimator is never smaller, and is usually strictly larger, than the asymptotic variance of the RE estimator. (AW, p. 330)

$$\text{Avar}\left(\hat{\boldsymbol{\delta}}_{FE}\right) = \sigma_u^2\left[\text{E}\left(\ddot{\mathbf{W}}_i'\ddot{\mathbf{W}}_i\right)\right]^{-1}/N \tag{5.15}$$

Let $\breve{\mathbf{w}}_{it} = \mathbf{w}_{it} - \lambda\overline{\mathbf{w}}_i$ be the quasi-time demeaned time-varying covariates. To get Avar$(\hat{\delta}_{RE})$, we need the residuals from the pooled regression: $\breve{\mathbf{w}}_{it}$ on $(1-\lambda)\mathbf{z}_i$. These residuals are called $\tilde{\mathbf{w}}_{it}$. Then,

$$\text{Avar}\left(\hat{\boldsymbol{\delta}}_{RE}\right) = \sigma_u^2\left[\text{E}\left(\tilde{\mathbf{W}}_i'\tilde{\mathbf{W}}_i\right)\right]^{-1}/N \tag{5.16}$$

$\text{Avar}\left(\hat{\boldsymbol{\delta}}_{FE}\right) - \text{Avar}\left(\hat{\boldsymbol{\delta}}_{RE}\right)$ is positive definite because $\text{E}\left(\tilde{\mathbf{W}}_i'\tilde{\mathbf{W}}_i\right) - \text{E}\left(\ddot{\mathbf{W}}_i'\ddot{\mathbf{W}}_i\right)$ is positive definite[5].

$$\begin{aligned}\text{E}\left(\tilde{\mathbf{W}}_i'\tilde{\mathbf{W}}_i\right) - \text{E}\left(\ddot{\mathbf{W}}_i'\ddot{\mathbf{W}}_i\right) =& (1-\lambda)^2\text{E}\left[\left(\overline{\mathbf{w}}_i - \overline{\mathbf{w}}_i^*\right)'\left(\overline{\mathbf{w}}_i - \overline{\mathbf{w}}_i^*\right)\right] \\ &+ (1-\lambda)\sum_{t=1}^{T}\ddot{\mathbf{w}}_{it}'\left(\overline{\mathbf{w}}_i - \overline{\mathbf{w}}_i^*\right) + (1-\lambda)\sum_{t=1}^{T}\left(\overline{\mathbf{w}}_i - \overline{\mathbf{w}}_i^*\right)'\ddot{\mathbf{w}}_{it} \\ =& (1-\lambda)^2\text{E}\left[\left(\overline{\mathbf{w}}_i - \overline{\mathbf{w}}_i^*\right)'\left(\overline{\mathbf{w}}_i - \overline{\mathbf{w}}_i^*\right)\right],\end{aligned} \tag{5.17}$$

where the last equality follows from $\sum_{t=1}^{T}\ddot{\mathbf{w}}_{it} = \mathbf{0}, \forall i$. For $\lambda < 1$, this is positive definite.

The **Hausman statistic** is defined as:

$$H = \left(\hat{\boldsymbol{\delta}}_{FE} - \hat{\boldsymbol{\delta}}_{RE}\right)'\left[\text{Avar}\left(\hat{\boldsymbol{\delta}}_{FE}\right) - \text{Avar}\left(\hat{\boldsymbol{\delta}}_{RE}\right)\right]^{-1}\left(\hat{\boldsymbol{\delta}}_{FE} - \hat{\boldsymbol{\delta}}_{RE}\right) \tag{5.18}$$

which has a χ_M^2 distribution under the null.[6]

> It is best to use the same estimate of σ_u^2 (based on either FE or RE) in both places [Avar $\left(\hat{\boldsymbol{\delta}}_{FE}\right)$ and Avar $\left(\hat{\boldsymbol{\delta}}_{RE}\right)$]. (AW, p. 331)

If we want to test for a single coefficient, we can apply a standard $t-$test: $\left(\hat{\delta}_{FE} - \hat{\delta}_{RE}\right)/\left\{\left[\text{se}\left(\hat{\delta}_{FE}\right)\right]^2 - \left[\text{se}\left(\hat{\delta}_{RE}\right)\right]^2\right\}^{1/2}$

> First, if there are no time-constant variables (except an overall intercept) in the RE estimation, the null hypothesis is $Cov(\bar{\mathbf{w}}_i, c_i) = 0$, which means we are *really testing whether the time-average of the w_{it} is correlated with the*

[5] More info on AW, p. 330.

[6] This is expected, because it is effectively a sum of standard normal distributions.

> *unobserved effect.* [...] with a rich set of controls in z_i, it is possible for $\bar{\mathbf{w}}_i - \bar{\mathbf{w}}_i^*$ to be uncorrelated with ci even though $\bar{\mathbf{w}}_i$ is correlated with c_i. (AW, p. 331)

The Hausman test can be written in regression form as:

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + \overline{\mathbf{w}}_i\xi + a_i + u_{it}, \tag{5.19}$$

where $c_i = \psi + \overline{\mathbf{w}}_i\xi + a_i$. In this case, the null is $H_0 : \xi = 0$.[7]

[7] More on AW, p. 332-3.

Chapter 6

Dynamic models

AP, 5.3-4.

Chapter 7

Sample selection

AW, 17.7 (new ed.)