Panel Data Estimation

Econometrics

J. Eduardo Vera-Valdés

eduardo@math.aau.dk

Department of Mathematical Sciences
Aalborg University
Denmark

AALBORG UNIVERSITY
DENMARK

Outline

J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark

Outline

J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark

Introduction

A panel data set contains repeated observations over the same units collected over a number of periods.

Examples of panel data are census of population, price of a set of stocks through time, economic growth of countries.

Today we will study an estimation method that allows us to model this type of data.

J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark

Introduction

Advantages of Panel Data Estimation:

- Controlling for individual heterogeneity.
- They give more variability, less collinearity, more efficiency.
- Are better able to study the dynamics of adjustment.
- Allows us to construct and test more complicated behavioural models.

J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark

Introduction
Error-components model

The model for Panel Data looks like

$$Y_{i,t} = \alpha + X_{i,t}\beta + U_{i,t},$$

where $i = 1, \cdots, N$ denote households, firms, countries, etc., and $t = 1, \cdots, T$ denotes time.

Because we repeatedly observe the same units, it is usually no longer appropriate to assume that different observations are independent.

The **error-components model** was developed to model this dependency.

The idea is to decompose the error in two (or three) separate shocks, each assumed to be independent of the others.

J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark

The error-components model looks like

$$U_{i,t} = \mu_i + V_{i,t},$$

where μ_i are individual specific components and $V_{i,t}$ are remainder effects.

Furthermore, we assume that μ_i are independent across i and that $V_{i,t}$ are independent across all i and t, orthogonal to $X_{i,t}$.

We think of μ_i as containing individual-specific information which may be observed (race, sex, location), or unobserved (skill, preferences); all of which are taken to be constant over time t.

J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark

Depending on the characteristics of the μ_i, different models can be considered.

Pooled regression: If $E[\mu_i|X_{i,t}] = 0$, OLS provides consistent estimates of α and β.

Fixed effects: If $E[\mu_i|X_{i,t}] \neq 0$, OLS is biased and inconsistent due to omitted variable. We add a "fixed" term that accounts for this.

Random effects: If $E[\mu_i|X_{i,t}] = 0$, we can obtain efficient estimates by adding a "random" term that accounts for the autocorrelation.

J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences Aalborg University Denmark

Outline

J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark

Pooled Regression

The assumptions for pooled regression are

$$Y_{i,t} = \alpha + X_{i,t}\beta + U_{i,t},$$

where

$$E[U_{i,t}|X_{i,t}] = 0 \quad \text{and} \quad Var[U_{i,t}|X_{i,t}] = \sigma_U^2.$$

We know that OLS is unbiased and consistent in this specification.

Moreover, if $E[U_{i,t}U_{j,s}] = 0 \; \forall i, j, t, s, \; i \neq j, \; t \neq s$; then OLS is efficient.

Nonetheless, this assumption is hard to argue in practice.

J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark

Pooled Regression

Given that we are measuring the same individuals, it is more realistic to assume

$$E[U_{i,t}U_{i,s}] \neq 0,$$

so that the error shows autocorrelation.

Hence, we can use an autocorrelation-robust matrix for inference, or model the random component explicitly.

Outline

J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark

First difference

Using the error-components form, we can write the model like

$$Y_{i,t} = \alpha + X_{i,t}\beta + \mu_i + V_{i,t}.$$

Given that μ_i does not change in time, one way to deal with them is to eliminate them by taking first differences.

Thus, we can estimate β by running OLS on the model

$$\Delta Y = \Delta X\beta + \Delta V.$$

The OLS estimator on this equation is called **first difference (FD)** estimator. It is quite common in models for "treatment effects".

First differencing introduces autocorrelation of known form.

Panel Data Estimation

J. Eduardo Vera-Valdés

eduardo@math.aau.

Department of Mathematical Sciences Aalborg University Denmark

Outline

J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark

Fixed Effects

Least square dummy variables

J. Eduardo Vera-Valdés
eduardo@math.aau.

If $E[\mu_i | X_{i,t}] \neq 0$, we can write the model like

$$Y_{i,t} = \alpha_i + X_{i,t}\beta + V_{i,t},$$

where $\alpha_i = \alpha + \mu_i$, so that the remaining error term is unrelated to the regressors.

Note that the model implies a different constant for each individual.

It can be written as

$$Y = X\beta + D\eta + V,$$

where $D = \iota_N \otimes I_T$, is a matrix of dummy variables indicating individual.

The OLS estimator on this equation is called **fixed effects (FE)** or **least-squares dummy variables (LSDV)** model.

Department of Mathematical Sciences
Aalborg University
Denmark

By the FWL theorem, the fixed effects estimator is given by

$$\hat{\beta}_{FE} = (X^T M_D X)^{-1} X^T M_D Y,$$

where M_D is the maker of residuals of D.

Note that

$$M_D X = X_{i,t} - \bar{X}_i,$$

where $\bar{X}_i$ is the group mean.

Thus, the estimator is also called **within-groups estimator**.

Panel Data Estimation
J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark

Fixed Effects

Fixed effects is consistent and unbiased.

It is BLUE as long as $V_{i,t}$ is the standard classical disturbance with mean 0 and variance covariance matrix $\sigma_\epsilon^2 I_{NT}$.

Nonetheless, fixed effects cannot estimate the effect of any time-invariant variable like sex, race, religion, schooling, or union participation.

If $T = 2$, then FE is numerically the same as the first difference estimator.

J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark

Outline

Department of Mathematical Sciences
Aalborg University
Denmark

Random Effects

There are too many parameters in the fixed effects model (α_i), which implies a loss of degrees of freedom.

We can avoid this if the μ_i can be assumed random.

Thus, it is possible to improve on the efficiency of the fixed-effects estimator.

Note that we then require that $E[\mu_i | X_{i,t}] = 0$, which guarantees that OLS is consistent but may be inefficient.

J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark

Random Effects

Recall that

$$U_{i,t} = \mu_i + V_{i,t},$$

the assumptions imposed for random effects are given by

$$E[V_{i,t}V_{j,s}] = 0 \ \forall i \neq j, t \neq s, \quad Var[V_{i,t}] = \sigma_V^2,$$

and,

$$E[\mu_i \mu_j] = 0 \ \forall i \neq j, \quad Var[\mu_i] = \sigma_\mu^2,$$

moreover

$$E[V_{i,t}\mu_i] = 0 \ \forall i, t.$$

J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark

Random Effects

Thus, note that the covariance of the error term in the model

$$Y_{i,t} = X_{i,t}\beta + U_{i,t},$$

is given by

$$\Omega = I_N \otimes \Sigma,$$

with

$$\Sigma = \sigma_V^2 I_T + \sigma_\mu^2 \iota_T \iota_T^T.$$

We can use this to estimate the model by GLS if we know σ_V^2 and σ_μ^2.

Alternatively, we can use FGLS if we find a way to estimate them.

First, recall that the error term in the specification for FE,

$$Y = X\beta + D\eta + V,$$

is just V.

This suggests that we can use FE to obtain an estimate for σ_V^2,

$$\hat{\sigma}_V^2 = \frac{\sum_{i=1}^{TN} \hat{V}_{i,t}^2}{NT - N - K},$$

where $\hat{V}_{i,t}$ are the residuals from FE.

J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark

Random Effects

Variance estimation

Now, to obtain an estimate of σ_μ^2, recall that the original model is

$$Y_{i,t} = \alpha + X_{i,t}\beta + U_{i,t} = \alpha + X_{i,t}\beta + \mu_i + V_{i,t},$$

so that the residual of pooled regression contains information on both.

This suggests that we can use pooled regression to obtain

$$s_{PR}^2 = \frac{\sum_{i=1}^{TN} \hat{U}_{i,t}^2}{NT - K - 1} \approx \sigma_\mu^2 + \sigma_V^2.$$

where $\hat{U}_{i,t}$ are the residuals from pooled regression.

Thus,

$$\hat{\sigma}_\mu^2 = s_{PR}^2 - \hat{\sigma}_V^2.$$

J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark

Random Effects

If the estimated variance turns out to be negative we could try a different estimate with more degrees of freedom, or (perhaps better) we should re-evaluate the model.

Under the assumptions, RE estimator is more efficient than either pooled regression estimator or the FE estimator.

Nonetheless, if the assumptions on the individual error term are not satisfied, then both pooled and RE are inconsistent.

Outline

Fixed effects vs pooled regression

We can test the suitability of fixed effects by testing the joint significance of the dummy variables.

Thus, we perform a F-test of the unrestricted model given by FE against the restricted one given by pooled regression.

The statistic is given by

$$F = \frac{(R^2_{FE} - R^2_{PR})/(N-1)}{(1-R^2_{FE})/(NT-N-K)},$$

which follows an F distribution.

Fixed effects vs random effects

To decide between using RE and FE, we can use a Hausman test.

Under all the assumptions, both RE and FE are consistent but FE is inefficient, while only FE is consistent under the null.

The statistic is given by

$$H = (\hat{\beta}_{FE} - \hat{\beta}_{RE})^T [Var(\hat{\beta}_{FE}) - Var(\hat{\beta}_{RE})]^{-1} (\hat{\beta}_{FE} - \hat{\beta}_{RE}),$$

which follows a chi-square distribution with degrees of freedom equal to the number of parameters.

Panel Data Estimation
J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark

Random effects vs pooled regression

To test on whether we should estimate RE or pooled regression.

The null hypothesis of the **Breusch-Pagan** test is

$$H_0 : \sigma_\mu^2 = 0,$$

so that there is no need to run RE.

The test statistic is

$$LM = \frac{NT}{2(T-1)} \left[\frac{\sum_{i=1}^{N} \left(\sum_{t=1}^{T} \hat{U}_{i,t} \right)^2}{\sum_{i=1}^{N} \sum_{t=1}^{T} \hat{U}_{i,t}^2} - 1 \right]^2,$$

which follows a chi-squared with one degree of freedom.

Outline

J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark

Dynamic Panel Data

Many economic issues are dynamic by nature:

$$Y_{i,t} = Y_{i,t-1}\delta + X_{i,t}\beta + U_{i,t}.$$

They are called **dynamic** due to the presence of lags of the dependant variable as regressors. They pose estimation problems.

The one-way error component model for dynamic panel data is given by

$$U_{i,t} = \mu_i + V_{i,t},$$

with i, t as before.

J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark

By definition, $Y_{i,t}$ is correlated with μ_i, which implies that $Y_{i,t-1}$ is correlated with μ_i. Thus,

$$E[\mu_i | Y_{i,t-1}] \neq 0.$$

FE subtracts the group mean, thus, the dynamic term is correlated with the error term.

Analogously, random effects pseudo-demeans the data, thus, the dynamic regressor is correlated with the error term.

J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark

Dynamic Panel Data
IV-estimation

Note that,

$$E[Y_{i,t-2}\Delta V_{i,t}] = 0,$$

so that $Y_{i,t-2}$ is uncorrelated to the error term.

Moreover,

$$E[Y_{i,t-2}\Delta Y_{i,t-1}] \neq 0,$$

so that $Y_{i,t-2}$ is correlated to the dynamic regressor.

Thus, FD eliminates the individual component using only one lag, which leaves previous lags to use as possible instruments.

In particular, $Y_{i,t-2}$ can be used as an instrument for $\Delta Y_{i,t-1}$. This is Anderson and Hsiao's **first instrument**.

J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark

Dynamic Panel Data

IV-estimation

Alternatively, Anderson and Hsiao suggest a **second instrument**, $\Delta Y_{i,t-2}$.

$$E[\Delta Y_{i,t-2} \Delta V_{i,t}] = 0,$$

so that $\Delta Y_{i,t-2}$ is uncorrelated to the error term.

$$E[\Delta Y_{i,t-2} \Delta Y_{i,t-1}] \neq 0,$$

so that $\Delta Y_{i,t-2}$ is correlated to the dynamic regressor.

The second instrument loses one more observation than the first instrument.

J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark

Hence, to estimate a dynamic panel model, the instruments for individual i are given by,

$$W_i = [Y_{i,t-2}, \ \Delta X_{i,t}],$$

using the first instrument.

Or by

$$W_i = [\Delta Y_{i,t-2}, \ \Delta X_{i,t}],$$

for the second.

We then construct the matrix of instruments, $W = [W_i^T]^T$. And compute the IV estimator.

Panel Data Estimation
J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark

Anderson and Hsiao's IV estimator is consistent when either $N \to \infty$ or $T \to \infty$.

In practice, IV using the first instrument has been shown to have smaller variance and it is thus recommended.

Nonetheless, note that the IV estimation does not control for the correlation in the error term introduced by taking first differences.

Moreover, additional instruments can be obtained if we use more lags.

Outline

J. Eduardo Vera-Valdés
eduardo@math.aau.

Department of Mathematical Sciences
Aalborg University
Denmark

Summing Up

- We have studied panel data methods that allows us to extract more information from the data.
- If the individual errors are not related to the regressors, pooled regression is consistent and unbiased but it is inefficient.
- If the errors are related to the regressors, we can control for them by adding "fixed" terms. The estimation is thus consistent.
- If the individual errors are not related to the regressors, we can gain efficiency by controlling for the "random" individual terms.
- Under dynamic specifications, pooled regression, fixed effects, and random effects are biased. Use FD and IV.