DAVID SALAZAR

PANEL DATA

LONELINESS

Contents

Hopefully, a warranted effort.

Introduction

A research design that collects information of the same units repeatedly over time is called a panel. The main motivation for collecting panel data is an interest in the analysis of change; more specifically, an interest in the analysis of change at the (individual) level of units. Hence, besides answering questions on individual change, panel data can also be used to answer typical cross-sectional questions about level and trend. In other words: panel data allow us to address all the research questions that we are used to analyze with cross-sectional data and some additional questions that cross-sectional data cannot deal with; among them the question of individual change

Controlling for Omitted Variable Bias

The fact that we have access to repeated measurements of the same units allows us to control at least for their unknown characteristics that are constant over time. Units are used as their own controls, a technique known from experimental research as the pre-test post-test design. The underlying idea is the following: if a variable X influences the variable of interest Y , then a change of X at some time point t should result in a different value of Y at t + 1 than the value of Y at t - 1. **Since this design compares identical units measured at t - 1 and t + 1, it also controls for all their characteristics that do not change in between.**

Measurements Order

At the ideal setting, we would manipulate X for some people and leave it alone for others. Then, we would observe the change in Y resulting from an exogenous variation to conclude what would be the effect of altering X (assuming non treated are a good counterfactual to the treated population. However, in cross-sectional data, all measurements are observed at once. Panel Data allows us to check the timing of the changes.

Attrition

The main problem arises when sample members permanently drop out of the panel. This process is called panel attrition or panel mortality. Problem? If the non-response is selective, then the available information provides a biased picture of the population. Quite generally, representing a population over time is a much more complex issue than representing a population at a given point in time. Hence, besides the problem of panel attrition the panel design also suffers from significant changes of the population itself.

Conclusion

In sum, a panel has the ability to answer research questions that the cross- sectional design cannot address. However, it has selectivity problems due to panel attrition and population change. Hence, to exploit the unique features of the panel design, much effort must be invested to minimize these problems of representation.

Overview

Measurements Over Time are Not Independent

Panel Data introduces the problem that, conditional on the parameters, not all observations are independent. Although we can assume independence between cross-section units, we cannot assume independence between the different time measurements within the units. In sum, panel data usually do not include independent information. We have certainly more information than in a cross-section of n units, but not as much as in a cross-section of $N = n \times T$ units. Why? The errors are the part of the variation in Y that is not explained by X; if we do not account for dependence within observations for a given unit, this dependence will go to the error and exhibit a similarity. Thus, for a given unit, a serial correlation between its errors.

Ignoring these statistical dependencies is potentially dangerous when applying regression models, because traditional estimation procedures assume independent observations. Thus, they will estimate standard errors that tend to be too low. As a consequence, test statistics will be too high and correspondingly, p-values too low, **such that significance tests will lead to erroneous conclusions. Furthermore, although parameter estimates may be unbiased, they could be estimated more efficiently, if the statistical dependencies were explicitly modeled.**

Summarizing Variables

Although we can compute simple measures and see how they change through time, we are not using all the power of Panel Data. We are not using the repeated observation: what is variation within units? Between units?

Dependent variables and regressors can potentially vary over both time and individuals. Variation over time or a given individual is called within variation, and variation across individuals is called between variation. This distinction is important because estimators differ in their use of within and between variation. In particular, in

the FE model the coefficient of a regressor with little within variation will be imprecisely estimated and will be not identified if there is no within variation at all.

For continuous variables, we can divide (heuristically) the overall heterogeneity in two. One, resulting from the heterogeneity between units, and the other is the variance within the units. If both are similar, we could conclude: wage heterogeneity across different individuals does not differ very much from wage heterogeneity across different years (taking into account the average level of wages for each individual). If between heterogeneity is much larger than within heterogeneity, within estimation will lead to considerable efficiency loss as there isn't much variation to leverage.

For categorical variables, we can explore they change as how tabulations of the percentages change: what percentage of people did vote in the first wave and also voted in the second? What percentage didn't in neither of them.

Explaining the dependent over Time and Units

Now that we have an idea of what our dependent variable looks like, we can turn to the question of how to explain its distribution over time and between units. Why do some individuals earn higher wages or have higher probabilities of union membership? Why do wages increase more steeply for some individuals? Why is union membership rather stable for some individuals, but a more transient phenomenon for others? These questions relate to the level and change of the dependent variable.

Variables change between units and within units. The variation between units is given by both time varying variables and time-constant variables; Variation within units is given by time varying variables. As a proxy to explain either type of variation, we can use time to proxy for the effect of any other variable that is relevant for the analysis and changes over time in a predictable fashion (tendency).

Unobserved heterogenity

Quid: one group pre-test post-test design. It focuses only on the treatment group and takes measures of the outcome variable before and after the treatment. The reasoning behind this design is the following: Even if the members of the treatment group have specific characteristics, if the treatment is effective there should be a significant difference between the post-test and the pre-test measurement. In other words, by looking at differences within the treatment group, the (time-constant) selective characteristics of their members are con-

trolled for.

Why are panel data useful to control for unobserved heterogeneity?? The answer is: Because panel data include repeated observations for each unit of analysis, and this allows us to base our estimations on the within-unit variation, which is unaffected by time-constant characteristics of the units (both the known and the unknown ones). Each unit, so to speak, is used as its own control.

Continuous Variables

With Panel Data, both the consistency and efficiency of traditional models is at stake. The consistency rests on an orthogonality assumption, as usual; the efficiency is at stake due to very possible fact that repeated observations for a given unit must be related through time. Thus, there's a serial correlation in the errors. We can try to fix it with robust standard errors, but that wouldn't be an efficient solution, as we wouldn't be exploding the benefits of Panel Data. Why? Robust standard errors controls for any kind of correlation in the data, while in this case we know the specific cause of the serial correlation. Thus:

1. Random Effects: t allows us to compute more efficient estimates of the parameters and their stan- dard errors than in traditional models even with robust standard errors. This method is called random effects (RE) estimation
2. Fixed Effects: It allows us to control for omitted variables bias at the unit level, when the unob- served variables at the unit level correlate with the variables in the model, which is not possible with cross-section data. However, if the exogeneity assumption is true (the omitted variable bias does not exist), it is less efficient than RE estimation.

Pooled Data

Treat $n \times T$ as independent observations. Thus, the model would be:

$$y_{it} = x_{it}\beta + c_i + e_{it}$$

The usual OLS assumptions thus transform into the following:

- $E(e_{it}|x_{it}, c_i) = E(e_{it})\ \forall t = 1, \ldots, T$
- $E(c_i|x_{it}) = E(c_i)\ \forall t = 1, \ldots, T$

Thus, the $\hat{\beta}$ estimator would be consistent and unbiased. However, what we would be estimating would be something of the following form:

$$y_{it} = x_{it}\beta + v_{it}$$

The errors v_{it} wouldn't be homoscedastic. Why? Given a unit, our predictions for that unit will fail at each period to catch the effect of c_i; thus, the errors within units will be systematically correlated. Thus, OLS estimation would fail to accurately estimate the standard errors. We could solve it by estimating the standard errors of $\hat{\beta}$ with robust standard errors. However, this method controls for any kind of serial correlation, but we know the specific type of serial correlation. Using this information, we can do better: Random Effects.

Random Effects

Random Effects is a special kind of estimation with FGLS, where we use the information of the provenance of the serial correlation. To specify this estimation, and its assumptions, it's useful to express the model in stacked form thus:

$$y_{it} = x_{it}\beta + c_i + e_{it}$$

- $Y_i = (y_{i1}, \ldots, y_{iT})'$
- $X_i = (x_{i1}, \ldots, x_{iT})'$
- $e_i = (e_{i1}, \ldots, e_{iT})'$
- $V_i = (v_{i1}, \ldots, v_{iT})'$

$$Y_i = X_i\beta + c_i + e_i$$
$$Y_i = X_i\beta + V_i$$

Under the following assumptions, FGLS is unbiased, consistent and efficient:

- $E(e_{it}|X_i, c_i) = E(e_{it})\ \forall t = 1, \ldots, T$
- $E(c_i|X_i) = E(c_i)$
- $E(e_i e_i'|X_i) = \sigma_h^2 I$ Idiosyncratic errors as homoscedastic.

There, by estimating the following with FGLS:

$$Y_i = X_i\beta + V_i$$

$$\hat{\beta_{RE}} = (\sum_i^N X_i'\hat{\Omega^{-1}}X_i)^{-1}(\sum_i^N X_i'\hat{\Omega^{-1}}Y_i)$$

Where:

$$\Omega = E(V_i V_i')$$

Thus, we need to estimate Ω. To do so, note that Ω is a matrix where each component in the diagonal is:

$$E(v_{it} v_{is}) = E(c_i^2) = \sigma_c^2$$

Off the diagonal:

$$E(v_{it}^2) = \sigma_c^2 + \sigma_e^2$$

Note that both values require and estimation of v_{it}; estimation that can be obtained from Pooled OLS of the model.

Fixed Effects

In many applications, the main benefit of using Panel Data is the correction of endogenity at the individual level. That is, correlation between our variables of interest and unobserved time constant heterogeneity in the individuals. Then, we allow $E(c_i|x_i)$ to be any arbitrary correlation.

However, we'd still assume the following:

$$y_{it} = x_{it}\beta + c_i + e_{it}$$

- $E(e_{it}|X_i, c_i) = E(e_{it})\ \forall t = 1, \ldots, T$

As we suggested, this robustness comes at a price: without further assumptions, we cannot include time-constant factors in x_{it}. The reason is simple: if c_i can be arbitrarily correlated with each element of x_{it}, there is no way to distinguish the effects of time-constant observables from the time-constant unobservable c_i. When analyzing individuals, factors such as gender or race cannot be included in $x_i t$.

The idea for estimating β under the former assumption is to transform the equations to eliminate the unobserved effect c_i. When at least two time periods are available, there are several transformations that accomplish this purpose.

Within Transformation

We transform our structural equation into the following:

$$\ddot{y_{it}} = \ddot{x_{it}}\beta + \ddot{c_i} + \ddot{e_{it}}$$

Where the dotted variables are time-demeaned variables. Note $\ddot{c}_i = 0$. Thus, this estimating equation can be consistently and efficiently estimated using Pooled OLS, given our former strict exogeneity assumption. In strictu sensu, this assumption is too much, we need:

$$E(\ddot{x_{it}}'\ddot{e_{it}}) = 0 \; \forall t, s = 1, \ldots, T$$

First Differencing Transformation

Transform the model by subtracting its lag (e.g., $\Delta y_{it} = y_{it} - y_{i,t-1}$):

$$\Delta y_{it} = \Delta x_{it}\beta + \Delta c_i + \Delta e_{it}$$

Note that $\Delta c_i = 0$. Thus, the former estimating equation can be consistently estimated by Pooled OLS, as long as the following assumption holds:

$$E(\Delta x_{it}' \Delta e_{it}) = 0 \; \forall t = 1, \ldots, T$$

That is, e_{it} cannot be correlated with the variables in the former period, nor the actual period neither the future period. The most problematic of these three types is that current errors aren't correlated with future variables, that is: $E(u_{it}' x_{is}) = 0, \; s > t$. Assumption that can be dropped using First Differencing -IV.

Assumption that holds under the fixed effects assumption.

Hausman Test

Because the key consideration in choosing between an RE and an FE approach is whether c_i and x_{it} are correlated, it is important to have a method for testing this assumption.

First, strict exogeneity, ($E(e_{it}|X_i, c_i) = E(e_{it}) \; \forall t = 1, \ldots, T$), is maintained under the null and the alternative. Then, we specify under the null that the estimation technique that is both consistent and efficient under $cov(x_{it}, c_i) = 0 \; \forall t = 1, \ldots, T$. Under the alternative, the other technique, that is always consistent, is specified. A deviation between both techniques is interpreted as a evidence against the null, as both techniques should yield similar results if $cov(x_{it}, c_i) = 0 \; \forall t = 1, \ldots, T$. The efficiency is used to compute the test statistic, as the asymptotic variance of the given method should always be smaller under the null (and thus the order is specified such that the difference yields a positive number).

Instrumental Variables in Panel Data

Sometimes, the assumption of strict exogeneity conditional on unobserved heterogeneity fails. Instead, we assume we have available instrumental variables (IVs) that are uncorrelated with the idiosyncratic errors in all time periods. Depending on whether these instruments are also uncorrelated with the unobserved effect, we are led to random effects or fixed effects IV methods. That is, as always instruments are assumed to be uncorrelated with idiosyncratic errors, but not necessarily with unobserved time constant effects. Conditional on this latter assumption, we use either estimation technique.

Sequential Exogeneity

Under this new framework, we allow any arbitrary correlation between the idiosyncratic errors and future values of the variables. That is:

$$E(e_{it}|x_{it}, x_{i,t-1}, \ldots, x_{i1}) = E(e_{it})$$

Which is equivalent to:

$$E(y_{it}|x_{it}, x_{i,t-1}, \ldots, x_{i1}) = E(e_{it}|x_{it})$$

Then, under the usual transformations in the fixed effect framework, under sequential exogeneity, the Pooled OLS assumptions won't hold. However, there's a solution: first differencing using lagged values of the dependent variables as instruments. That is, in the first difference model:

$$\Delta y_{it} = \Delta x_{it}\beta + \Delta c_i + \Delta e_{it}$$

We instrument Δx_{it} with $x_{i,t-1}$ (and other higher order lags) or $\Delta x_{i,t-1}$. Such that the necessary First Difference assumption ($E(\Delta x'_{it}\Delta e_{it}) = 0 \; \forall t = 1, \ldots, T$), with the instruments, does hold under sequential exogeneity.

Autoregressive Models in Panel Data

Under sequential exogeneity, a similar problem happens with autoregressive models: the usual transformations under the fixed effects framework will yield models that, under the usual assumptions, cannot be consistently estimated. Thus, instead of using past values of the x, we use past values of the dependent variables (that under sequential exogeneity are exogenous w.r.t relevant errors) as instruments.

Under sequential exogeneity, the necessary condition for consistently using Pooled OLS on the transformed models won't hold. Thus, instrumental variable estimation with first differencing.

Categorical Variables

With Panel Data, both the consistency and efficiency of traditional models is at stake. The consistency rests on an orthogonality assumption, as usual; the efficiency is at stake due to very possible fact that repeated observations for a given unit must be related through time. Thus, there's a serial correlation in the errors. We can try to fix it with robust standard errors, but that wouldn't be an efficient solution, as we wouldn't be exploding the benefits of Panel Data. Why? Robust standard errors controls for any kind of correlation in the data, while in this case we know the specific cause of the serial correlation. Thus:

1. Random Effects: t allows us to compute more efficient estimates of the parameters and their stan- dard errors than in traditional models even with robust standard errors. This method is called random effects (RE) estimation

2. Fixed Effects: It allows us to control for omitted variables bias at the unit level, when the unob- served variables at the unit level correlate with the variables in the model, which is not possible with cross-section data. However, if the exogeneity assumption is true (the omitted variable bias does not exist), it is less efficient than RE estimation.