Causal Inference with Observational Data

4. Panel Data and Difference-in-Differences

Marc F. Bellemare

May 2018

Introduction

Panel data are data for which you have more than one observation per unit of observation.

For example, you might have several months of data on a sample of firms, or several years of data on a sample of countries, or you might have several individuals per household for a sample of households, or several plots per household for a sample of households.

From the latter two examples, you can see why equating the expression "panel" with "longitudinal" or "time-series cross-sectional" is mistaken; though there is often a time dimension that is involved with panel data, one can have a panel data that is also cross-sectional.

Introduction

Panel data are also a good example of how economics is not immune to fads and fashions.

In the late 1990s, all the cutting-edge studies used panel data, and it was difficult to publish in top journals without them.

Nowadays, panel data are just another tool among many used by applied microeconomists. This helps put into perspectives current fads and fashions.

Fixed vs. Random Effects

Suppose we have data on a number of individuals $i \in \{1, ..., N\}$ over time $t \in \{1, ..., T\}$. The first thing to note is that in most applied micro applications, $N > T$. With $T > N$ (say, if you have hourly price data on a sample of stocks), the methods involved tend to differ a bit from what we will be discussing in this class.

As always, we are interested in estimating the causal impact of the variable of interest D on some outcome y while controlling for a vector of covariates x. Thus, our equation of interest is such that

$$y_{it} = \alpha + \beta x_{it} + \gamma D_{it} + \epsilon_{it}. \tag{1}$$

Fixed vs. Random Effects

As it stands, except for the subscripts, the problem is identical to what we had earlier, and the previous equation does not take into account the panel structure of the data at all.

Indeed, witness how we could simply re-label each individual-time period observation it as j and we would revert back to the usual cross-sectional equation. In other words, without anything more, the equation above is a pooled cross-section, wherein we throw all observations together indiscriminately.

Fixed vs. Random Effects

Obviously, there is a better way, which involves taking into account the panel structure of the data.

An agricultural economist named Yair Mundlak was interested in estimating production functions for a sample of farms over time, and he noticed that an important source of bias was farmer managerial ability.

Fixed vs. Random Effects

In a paper published in 1961 in the *Journal of Farm Economics* (which would eventually be renamed the *American Journal of Agricultural Economics*), Mundlak wrote down the following specification:

$$y_{it} = \alpha + \beta x_{it} + \gamma D_{it} + \delta_i + \epsilon_{it}, \tag{2}$$

which is almost identical to the previous equation except that it now includes the vector δ of dummy variables, which are such that $\delta_i = 1$ for all observations i and $\delta_i = 0$ for all observations $-i$.

Fixed vs. Random Effects

For Mundlak, this allowed controlling for the owner of each farm i's managerial ability for all observations $i1$ to iT, since managerial ability presumably does not change over time within a given farm.

In other words, if the presence in the error term of farmer managerial ability means that $Cov(D, \epsilon) \neq 0$, then taking managerial ability (and other things which remain constant across all observations common to i) out of the error term would lessen $Cov(D, \epsilon)$ and get it closer to zero, if not make it so that $Cov(D, \epsilon) = 0$.

Fixed vs. Random Effects

Mundlak's approach is commonly referred to as the least-squares dummy variables approach—you create one dummy variable for each unit i in your sample and estimate it—but it is also known as the *fixed effects* (FE) estimator, because the vector δ controls for unit-specific fixed effects. Note that equation 2 can also be estimated by estimating

$$y_{it} - \overline{y}_i = \beta(x_{it} - \overline{x}_i) + \gamma(D_{it} - \overline{D}_i) + (\epsilon_{it} - \overline{\epsilon}_i), \qquad (3)$$

which we refer to as the *within* estimator, because it exploits the variation within each unit of observation to generate estimates of β and γ.

Fixed vs. Random Effects

The within estimator can be particularly useful in situations where you would rather not have to contend with a high number of dummy variables.

In Bellemare, Barrett, and Just (2013), for example, we wanted to estimate a system of seven equations by seemingly unrelated regression while incorporating household fixed effects, with about 1600 households observed on average five times each.

To simplify the problem, we centered the data—that is, we subtracted the within-observation mean of each variable—which allowed incorporating household fixed effects without having to worry about computational power.

Fixed vs. Random Effects

An approach that is sometimes touted as an alternative to the fixed effects estimator is the *random effects* (RE) estimator, which is such that

$$y_{it} = \alpha + \beta x_{it} + \gamma D_{it} + \nu_i + \epsilon_{it}, \tag{4}$$

where instead of being a parameter to be estimated, ν_i is a component of the error term which varies for each unit of observation.

Many people think of FE and RE estimators as interchangeable. When I was first taught about this, I was told that you should use FE if you have all of the units of observations in a population (e.g., all ten Canadian provinces), but you should use RE if you have a random sample of observations from a population (e.g., a random sample of survey respondents).

Fixed vs. Random Effects

Fortunately, thinking has evolved since then, and we know better. The rule of thumb is this: One should favor RE if and only if one can credibly make the claim that the variable of interest D in equation 1 is uncorrelated with the error term, i.e., if $Cov(D, \epsilon) = 0$. With $Cov(D, \epsilon) \neq 0$, then fixed effects are superior.

But when do we ever have $Cov(D, \epsilon) = 0$? The answer is that it pretty much never happens with observational (i.e., nonexperimental) data, but that the random effects estimator is okay to use with experimental data.

Fixed vs. Random Effects

In Bellemare, Lee, and Just (2018), for example, we run lab and lab-in-the-field experiments wherein we randomize people into certain or uncertain price treatments and, conditional on being randomized into the uncertain price treatment, we randomize them into four different uncertain price treatments of differing variances, and we look at their production decisions.

In that case, we estimate RE specifications, because our RHS variables of interest are assigned at random, and thus uncorrelated with unobservable factors such as individual preference for risk.

Fixed vs. Random Effects

Even in that case, we only do so after running a Hausman test to discriminate between the two.

This does not mean, however, that you should use the Hausman test to argue that you should use the RE estimator in an observational setting.

Indeed, recall that the null in the Hausman test is one of exogeneity (here, exogeneity of D), and that failure to reject the null would lead you to use the RE estimator. A rejection of the null is a much stronger result, and it would lead you to use the FE estimator.

Fixed vs. Random Effects

So there are four situations:

1. Experimental data and a failure to reject the null: Use the RE estimator.
2. Experimental data and a rejection of the null: Use the FE estimator.
3. Observational data and a failure to reject the null: Here, a failure to reject is not very convincing; use the FE estimator.
4. Observational data and a rejection of the null: Use the FE estimator.

Fixed vs. Random Effects

With that being said, you almost never see people using the RE estimator with observational data in applied microeconomics, so unless a reviewer makes it a necessary condition for publication (and even then, you should plead with the editor that that reviewer is wrong), you should not rely on the RE estimator unless you have experimental data.

Fixed vs. Random Effects

One other issue is that measurement error is more of a problem with FE than with other approaches.

That is, there is more noise in panel data, and so FE estimates can be significantly smaller than pooled OLS estimates. Angrist and Pischke (2009) suggest using an IV to correct serious measurement error issues.

SUTVA Violations

One issue that often crops up more with panel data than with other kinds of data and which threatens causal identification is the potential violation of the stable unit treatment value assumption (SUTVA), which roughly says that one observation's outcomes are unaffected by another observation's treatment assignment—that is, there are no spillovers.

Unfortunately, SUTVA violations are extremely common with panel data.

SUTVA Violations

Here are some examples:

- If you have a cross-section of households, each with several plots, and you are interested in productivity on each plot, then the SUTVA is almost surely violated by the fact that any resource expended on one plot by a household is not expended on another plot owned by the same household, and so there are necessarily spillovers between plots because of substitution.
- If you have longitudinal data on individuals and you are interested in those individuals' consumption behavior, then intertemporal substitution clearly causes a violation of the SUTVA.

SUTVA Violations

- In Bellemare and Nguyen (2018), we look at all 50 US states as well the District of Columbia for the period 2004-2013. We are interested in whether increases in the number of farmers markets per capita translate into increases in the number of outbreaks and cases of food-borne illness per capita. It is possible, however, that a resident of Hudson, WI shops at the St. Paul, MN farmers market. It is also possible that I shop at the farmers market in late December 2016 and get sick from the foods I purchased there only in early January 2017. If both those stories are frequent enough, identification is compromised. As a solution, we control for the number of farmers markets per capita in neighboring states, but that is not perfect.

SUTVA Violations

Any SUTVA violation compromises causal identification, and almost no application will be entirely free from those.

The solution, as always, is to be honest about the limitations of your approach, and to discuss potential SUTVA violations along with the other sources of statistical endogeneity.

Bellemare and Nguyen (2018)

In this paper, we are interested in the relationship between food-borne illness and farmers markets, so we assembled a data set of the number of all outbreaks and cases of food-borne illness on the one hand and of the number of farmers markets on the other hand.

For food-borne illness, we look at both aggregate numbers of outbreaks and cases, but also at specific illnesses (e.g., norovirus, Campylobacter, Salmonella, etc.)

We have data for all 50 US states plus the District of Columbia for eight years (2004, 2006, and 2008-2013), for a total of 408 observations.

Bellemare and Nguyen (2018)

For identification, we rely on state fixed effects first (FEs) and foremost with year fixed effects, but we also look at

1. State FEs and a linear trend,
2. State FEs and state-specific linear trends, and
3. Region-year FEs.

In an earlier version, we even had specifications that had state FEs and an IV (average minimum temperature) for the number of farmers markets.

Bellemare and Nguyen (2018)

Table 2. Ordinary Least Squares Estimation Results for the Total Number of Reported Outbreaks of Food-Borne Illness per Million

Variables	(1) *Total*	(2) *Norovirus*	(3) *Salmonella*	(4) *E. coli*	(5) *C. perfringens*	(6) *Campylobacter*	(7) *Scombroid*	(8) *Staph*
Dependent Variable: Reported Outbreaks per Million								
Farmers markets per million	0.089**†	0.026**	0.008	−0.002	0.005	0.009***§	0.013	0.004
	(0.034)	(0.012)	(0.006)	(0.003)	(0.004)	(0.003)	(0.011)	(0.003)
Multistate outbreaks not recorded	1.102	0.279	−0.806**	−0.617***	0.118	0.032	0.317	0.175*
	(0.867)	(0.348)	(0.313)	(0.120)	(0.073)	(0.054)	(0.249)	(0.104)
GDP per million	0.031	−0.088**	0.021	0.004	0.002	−0.004	−0.004	0.003*
	(0.055)	(0.034)	(0.024)	(0.004)	(0.006)	(0.010)	(0.004)	(0.002)
Proportion college graduates	−0.371	−0.027	−0.076	−0.020	−0.038*	−0.011	−0.081	−0.028
	(0.337)	(0.136)	(0.077)	(0.018)	(0.022)	(0.020)	(0.075)	(0.040)
Restaurants per million	−0.004	−0.003	−0.001	−0.001*	0.000	0.000	0.001	0.000
	(0.005)	(0.003)	(0.001)	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)
Constant	18.503*	11.788*	3.498	2.206***	0.881	−0.311	−0.331	−0.014
	(10.053)	(6.629)	(2.177)	(0.749)	(0.703)	(0.789)	(0.440)	(0.379)
Observations	408	408	408	408	408	408	408	408
State fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R-squared	0.095	0.206	0.282	0.187	0.027	0.061	0.140	0.114

Standard errors clustered at the state level in parentheses.

*** $p < .01$, ** $p < .05$, * $p < .1$.

§ $p < .05$, † $p < .1$ adjusting for multiple comparisons across columns (1) to (8) using the Dunn–Šidák correction (Šidák 1967).

Bellemare and Nguyen (2018)

Table 3. Ordinary Least Squares Estimation Results for the Total Number of Reported Cases of Food-Borne Illness per Million

Variables	(1) *Total*	(2) *Norovirus*	(3) *Salmonella*	(4) *E. coli*	(5) *C. perfringens*	(6) *Campylobacter*	(7) *Scombroid*	(8) *Staph*
Dependent Variable: Reported Cases per Million								
Farmers markets Per million	1.164**	0.774**	−0.076	−0.013	0.213	0.051*	0.032	0.094
	(0.515)	(0.337)	(0.131)	(0.019)	(0.194)	(0.029)	(0.024)	(0.081)
Multistate outbreaks not recorded	24.351	21.335**	−12.762	−1.468**	5.085	1.438	0.775	2.733
	(15.367)	(9.502)	(9.238)	(0.639)	(6.947)	(2.096)	(0.516)	(1.863)
GDP per million	−2.668**	−3.116**	−0.301	0.049	0.315	0.298	−0.011	0.042
	(1.320)	(1.352)	(0.438)	(0.068)	(0.361)	(0.376)	(0.011)	(0.044)
Proportion college graduates	−3.705	1.733	−1.617	0.045	−1.571	−0.409	−0.181	−0.456
	(6.205)	(5.340)	(2.596)	(0.133)	(0.939)	(0.415)	(0.138)	(0.711)
Restaurants per million	−0.107	−0.064	0.016	−0.003	0.003	−0.005	0.003	−0.001
	(0.105)	(0.090)	(0.022)	(0.002)	(0.021)	(0.008)	(0.002)	(0.002)
Constant	477.006**	226.242	50.514	3.705	25.194	5.848	−0.551	11.262
	(228.132)	(182.882)	(55.438)	(4.231)	(39.175)	(11.923)	(1.061)	(16.437)
Observations	408	408	408	408	408	408	408	408
State fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R-squared	0.145	0.186	0.056	0.033	0.020	0.034	0.102	0.036

Standard errors clustered at the state level in parentheses.
*** $p < .01$, ** $p < .05$, * $p < .1$.

Difference-in-Differences

A related approach to the FE estimator, one that has become more popular in recent years but which relies on different assumptions, is the differences-in-differences (DID) estimator.

Technically, Angrist and Pischke (2009) note that "[differences-in-differences] is a version of fixed effects estimation using aggregate data."

The DID estimator lends itself best to situations where you have a dichotomous treatment D whose rollout is staggered across different units, and you observe a large enough amount of observations both pre- and post-treatment.

Difference-in-Differences

Let's consider the simplest case. Suppose we observe two time periods $T \in \{0, 1\}$, which we will refer to as pre- and post-some treatment, and a unit i is either treated or not, so that $D_i \in \{0, 1\}$. Suppose also that you have a vector of controls which change within each unit over time x_{it}. The DID estimator is thus

$$y_{it} = \alpha + \beta_D D_i + \beta_T T_t + \gamma D \times T + \beta_x x_{it} + \epsilon_{it}. \tag{5}$$

Difference-in-Differences

The important things to note here are (i) the value of D does not change for a unit over time, i.e., whether a unit is assigned to treatment or control at all over the time period you consider remains constant over that time period; (ii) the variable T accounts for the passage of time, and (iii) the interaction term $D \times T$ is what we refer to as the DID term, and it captures the difference between treatment and control as well as the difference pre- and post-treatment.

In other words, the DID term accounts *both for a within- and between-unit effect of treatment*.

Difference-in-Differences

Specifically, we have that

1. $\alpha = E(y|D = 0, T = 0)$,
2. $\beta_D = E(y|D = 1, T = 0)$,
3. $\beta_T = E(y|D = 0, T = 1)$, and
4. $\gamma = E(y|D = 1, T = 1) - E(y|D = 0, T = 1) - E(y|D = 1, T = 0) - E(y|D = 0, T = 0)$.

Difference-in-Differences

Examples of DID abound in the applied microeconomics literature, but the most famous illustration of the DID estimator is probably Card and Krueger's study of the minimum wage effects in the fast-food industry in New Jersey, in which the authors compare the effects on unemployment of an increase in the minimum wage (the treatment) in both NJ and PA, and for which the authors conclude that the treatment led to more unemployment.

Difference-in-Differences

The original idea for the DID estimator is due to the work of John Snow, who discovered the causes of cholera in London in the 19th century, a story that told by Freedman (1991) in an article titled "Statistical Models and Shoe Leather," which illustrates that sometimes, you do not even need a regression to make the case that a given relationship is causal.

Obviously, the DID estimator does not give you identification for free, and a number of things must hold for it to generate credible causal estimates.

Difference-in-Differences

In no particular order,

- *Parallel trends*. It has to be the case that treated and untreated units follow parallel trends. This can sometimes be shown to hold if you have enough pre-treatment data for treated and control units of observation. In Bellemare, Fajardo-Gonzalez, and Gitter (2015), for example, we show figure 3.1 to show that consumption expenditure trends were roughly the same for all household categories before the quinoa price spike of 2012-2013, which is our treatment variable.

Difference-in-Differences

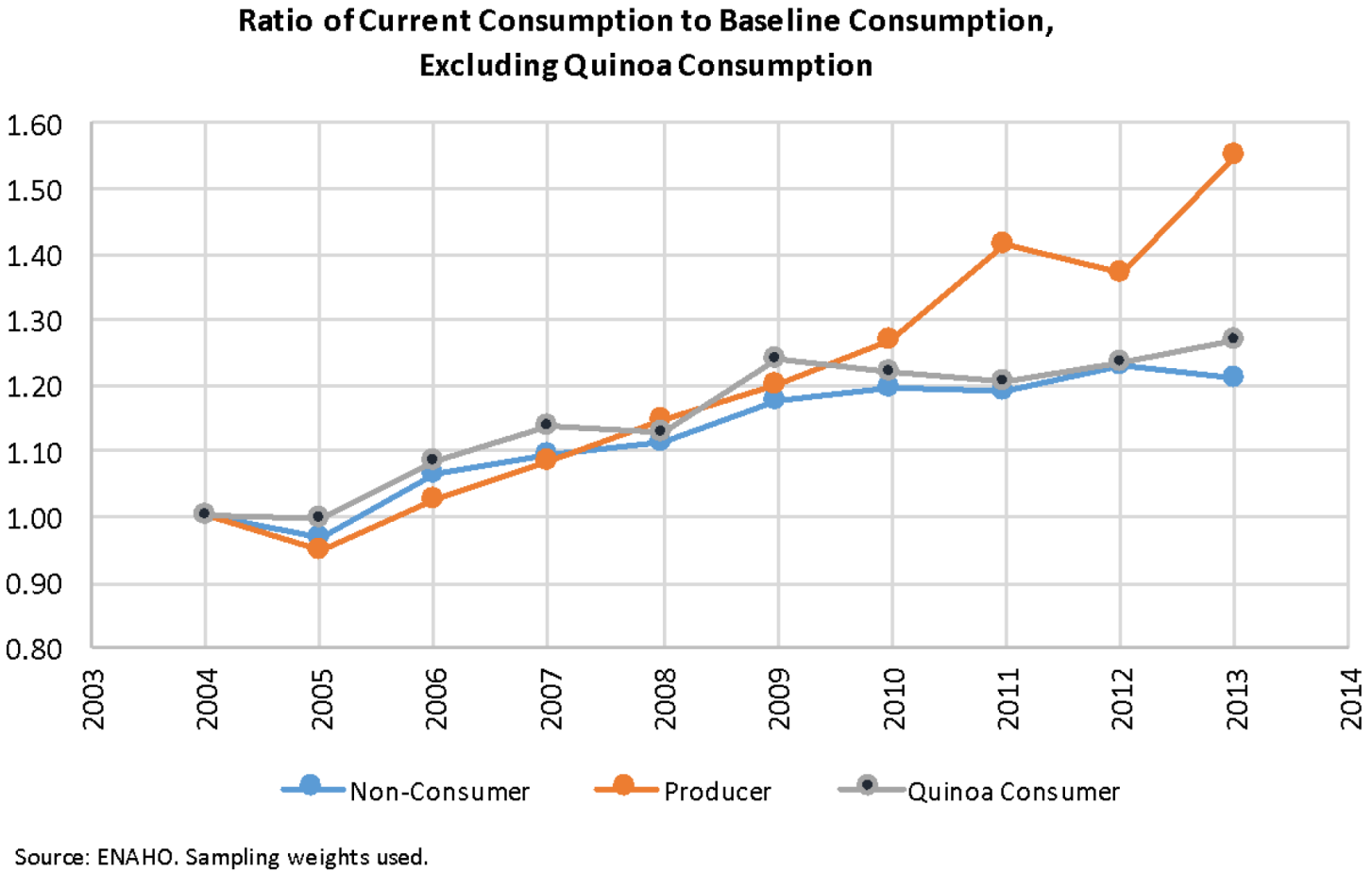

Figure: Parallel Trends in Consumption Expenditures in Peru. (Source: Bellemare, Fajardo-Gonzalez, and Gitter, 2015).

Difference-in-Differences

- *No Ashenfelter dips*. Though the expression is relatively uncommon, an Ashenfelter dip occurs when treatment units might respond by decreasing their value of the dependent variable pre-treatment in anticipation of receiving the treatment. For example, if the treatment consists of receiving SNAP benefits, people might respond the month before they start receiving those benefits by cutting back on their expenditures on fresh fruits and vegetables (FFV), which would make your DID estimate of the effect of SNAP benefits on FFV consumption to be biased upward, and to be much too high relative to the true effect of treatment.

Difference-in-Differences

- *No autocorrelation*. Straight standard errors for the DID estimator are relatively untrustworthy due to the presence of autocorrelation (see Bertrand et al., 2004). One thing that is pretty much necessary nowadays when using either the FE or DID estimators is the clustering of standard errors at the level of the unit of observation observed over time.

Repeated Cross Sections

Suppose you have data that consists of repeated cross-sections. To take an example from my own work (Bellemare, Fajardo-Gonzalez, and Gitter, 2015), suppose you have 10 years worth of a nationally representative household survey, but the data are not longitudinal.

That is, for each year, whoever was in charge of collecting the data collected them on a brand new sample of households. Obviously, because the data are not longitudinal, the usual panel data tricks (e.g., household fixed effects) are not available.

So what can you do if you want to get closer to credible identification?

Repeated Cross Sections

Enter pseudo-panel methods (Deaton, 1985), which are a set of very useful tools that I did not get to hear about in grad school.

To keep with the 10 of a nationally representative household survey example, suppose you have data on a random sample of households i in village v in periods $t \in \{1, ..., 10\}$, and suppose you are interested in the effect of some treatment D_{ivt} on some outcome while controlling for a vector of other factors x_{ivt}.

Repeated Cross Sections

In other words, the treatment and the outcome both vary at the household level, but because you have a repeated cross-section rather than longitudinal data, it is not possible to estimate an equation of the form

$$y_{ivt} = \alpha + \beta x_{ivt} + \gamma D_{ivt} + \delta_i + \tau t + \epsilon_{ivt}. \tag{6}$$

where δ is a household fixed effect and τ is a linear trend to account for the passage of time.

This is because you have as many household-village-year observations as you have observations in the entire data set, and so you cannot identify γ.

Repeated Cross Sections

With a large enough data set, one thing you can do to get out of this bind and get more credible identification is to use pseudo-panel methods.

Here, rather than treating the household as the unit of observation, you can simply treat the village as the unit of observation, and take the within-village mean of each variable over all households. Ultimately, you would estimate

$$\overline{y}_{vt} = \alpha + \beta \overline{x}_{vt} + \gamma \overline{D}_{vt} + \delta_v + \tau t + \overline{\epsilon}_{vt}, \tag{7}$$

where a bar denotes a within-village average.

Repeated Cross Sections

What are the assumptions that need to be satisfied for pseudo-panel methods to work?

First, at whatever level you choose as your unit of observation (here, the village level), the sample needs to be random. This is necessary because if you want to be able to compare a village today with the same village tomorrow, it has to be the case that the households sampled from that village today and tomorrow are matched their observable and unobservable characteristics.

The way to make sure that this holds is to have a random sample, i.e., a sample where respondents do not choose to answer the survey on the basis of some unobservable characteristic.

Repeated Cross Sections

Second, you also need to account for the passage of time. Even if the first condition holds and the households in a given village are randomly selected in each time period, and thus each village-level average is comparable with the previous and the next one, something might change over time that makes them incomparable.

For robustness, you can do this with a linear trend, year fixed effects, village-specific linear trends, and so on.

Repeated Cross Sections

Another important thing to keep in mind with pseudo-panel methods is the trade-off between sample size and measurement error.

In Bellemare, Fajardo-Gonzalez, and Gitter (2015), we are lucky to have three administrative levels which we could treat as our unit of observation: (i) 1,840 districts nested in (ii) 195 provinces nested in (iii) 25 departments. Since we have 10 years worth of data, we could then estimate everything at each level, respectively with in theory 18,400 district-year observations, 1,950 province-year observations, and 250 department-year observations.

Repeated Cross Sections

For robustness, we estimate all of our specification at each of those levels, but the trade-off is that the more households go into making an average (e.g., there are more households sampled in a department than in a province, and in a province than in a district), the more precise that average will be, and so the less measurement error there is.

But the more households go into making an average, the smaller the pseudo-panel sample size, too: There are fewer departments than there are provinces, and there are fewer provinces than there are districts. This is nothing new under the sun—the trade-off between sample size and precision is part and parcel of statistics—but it is useful to keep it in mind nevertheless.

Bellemare, Chua, Santamaria, and Vu (2018)

In this paper, we were interested in looking at the impact of a reduction in tenurial insecurity (y) on the investment behavior of Vietnamese "landowners" of annual crop plots (D) after the passage of the Land Law of 2013 (T).

In Vietnam, all plots are owned by the state. In 1993, landowners of perennial (annual) crop plots were given 50 (20) years of usufruct rights. In 2013, in a largely unanticipated change, the Vietnamese government gave all landowners an extra 50 years of usufruct rights.

Bellemare, Chua, Santamaria, and Vu (2018)

So we rely a difference-in-differences design, given that we find support for the notion that both types of landowners were following parallel trends before the passage of the Land Law of 2013.

We look at four types of investment, but one of those (investment in aquaculture) serves as placebo, since we don't expect people to change their investment in aquaculture behavior as a consequence of a change in the degree of tenurial insecurity for landowners of annual crops.

We also control for endogenous switching of the type of crop grown (i.e., from annual to perennial).

Bellemare, Chua, Santamaria, and Vu (2018)

Table 5: Effects of the Land Law on investment for owned, non-restricted plots

Outcomes	Model 1	Model 2	Model 3	Model 4
Irrigation	0.16***	0.18***	0.30***	0.16*
	(0.04)	(0.04)	(0.06)	(0.08)
	[8273]	[8273]	[8273]	[8273]
Infrastructure	0.00	0.00	0.00	0.01
	(0.01)	(0.01)	(0.01)	(0.01)
	[8273]	[8273]	[8273]	[8273]
Tree-planting	-0.04	-0.07	-0.04	-0.11
	(0.04)	(0.04)	(0.07)	(0.08)
	[6122]	[6122]	[6122]	[6122]
Aquaculture	0.00	0.00	0.00	0.01
	(0.01)	(0.01)	(0.01)	(0.01)
	[8262]	[8262]	[8262]	[8262]
Year FE	No	Yes	Yes	No
Plot FE	No	No	Yes	Yes
Province-year FE	No	No	No	Yes

Note: Clustered standard errors in parenthesis. Number of observations in bracket. Each cell corresponds to an individual OLS regression in which the outcome variable is specified in every row. Four specifications are analyzed. Model (1) regress the corresponding outcome to a dummy of whether the plot is annual or perennial, a time dummy that takes the value of 1 in year 2016, and the interaction of the latter two dummies. Model (2) includes year fixed effects. Model (3) adds plot fixed effects. Model (4) add province-year fixed effects. All four specifications control for plot area. *** p¡0.01, ** p¡0.05, * p¡0.1

Figure: Results not accounting for endogenous switching.

Bellemare, Chua, Santamaria, and Vu (2018)

Table 6: Effects of the Land Law on investment for owned, non-restricted, non-switching plots

Outcomes	Model 1	Model 2	Model 3	Model 4
Irrigation	0.16***	0.19***	0.32***	0.27*
	(0.05)	(0.04)	(0.06)	(0.11)
	[7400]	[7400]	[7400]	[7400]
Infrastructure	0.00	0.00	0.00	0.00
	(0.01)	(0.01)	(0.01)	(0.00)
	[7400]	[7400]	[7400]	[7400]
Tree-planting	0.00	-0.03	0.01	-0.10
	(0.05)	(0.05)	(0.08)	(0.09)
	[5389]	[5389]	[5389]	[5389]
Aquaculture	0.00	0.00	-0.01	-0.01
	(0.01)	(0.01)	(0.01)	(0.01)
	[7390]	[7390]	[7390]	[7390]
Year FE	No	Yes	Yes	No
Plot FE	No	No	Yes	Yes
Province-year FE	No	No	No	Yes

Note: Clustered standard errors in parenthesis. Number of observations in bracket. Each cell corresponds to an individual OLS regression in which the outcome variable is specified in every row. Four specifications are analyzed. Model (1) regress the corresponding outcome to a dummy of whether the plot is annual or perennial, a time dummy that takes the value of 1 in year 2016, and the interaction of the latter two dummies. Model (2) includes year fixed effects. Model (3) adds plot fixed effects. Model (4) add province-year fixed effects. All four specifications control for plot area. *** p¡0.01, ** p¡0.05, * p¡0.1

Figure: Results accounting for endogenous switching.

Bellemare, Chua, Santamaria, and Vu (2018)

Table 7: Long term effects of the Land Law on investment for owned, non-restricted plots

Outcomes	Model 1	Model 2	Model 3	Model 4
Irrigation	0.15**	0.15**	0.49***	0.49**
	(0.05)	(0.05)	(0.11)	(0.17)
	[3590]	[3590]	[3590]	[3590]
Infrastructure	-0.01	-0.01	-0.02*	0.00
	(0.01)	(0.01)	(0.01)	(0.00)
	[3590]	[3590]	[3590]	[3590]
Tree-planting	-0.11*	-0.11*	0.03	-0.07
	(0.05)	(0.05)	(0.11)	(0.14)
	[3590]	[3590]	[3590]	[3590]
Aquaculture	0.00	0.00	-0.02	-0.06
	(0.01)	(0.01)	(0.02)	(0.05)
	[3590]	[3590]	[3590]	[3590]
Year FE	No	Yes	Yes	No
Plot FE	No	No	Yes	Yes
Province-year FE	No	No	No	Yes

Note: Clustered standard errors in parenthesis. Number of observations in bracket. Each cell corresponds to an individual OLS regression in which the outcome variable is specified in every row. Four specifications are analyzed. Model (1) regress the corresponding outcome to a dummy of whether the plot is annual or perennial, a time dummy that takes the value of 1 in year 2016, and the interaction of the latter two dummies. Model (2) includes year fixed effects. Model (3) adds plot fixed effects. Model (4) add province-year fixed effects. All four specifications control for plot area. *** p¡0.01, ** p¡0.05, * p¡0.1

Figure: Results for long-term effects.

Bellemare, Fajardo-Gonzalez, and Gitter (2018)

Table 4a. Pseudo-Panel Regression of Total Household Consumption on the Proportion of Households Who Consume Quinoa Treating Districts as Units of Observation, 2004-2014.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Dependent Variable: Mean Log of Total Household Consumption (in 2004 PEN)							
% of Quinoa Consumers x Log of International Price of Quinoa	0.059***	0.041***	0.037***	0.040***	0.040***	0.033***	0.036***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.005)
Constant	8.842***	8.652***	8.654***	-62.951***	-60.810***	8.731***	8.255***
	(0.006)	(0.009)	(0.004)	(2.138)	(2.292)	(0.011)	(0.021)
Observations	10,774	10,774	10,774	10,774	10,774	7,814	2,678
R-squared	0.050	0.196	0.426	0.266	0.231	0.186	0.203
District Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Linear Trend	No	Yes	No	No	Yes	Yes	Yes
Year Fixed Effects	No	No	No	No	No	No	No
District-Specific Linear Trends	No	No	Yes	No	No	No	No
Province-Specific Linear Trends	No	No	No	Yes	No	No	No
Province-Year Fixed Effects	No	No	No	No	No	No	No
Department-Specific Linear Trends	No	No	No	No	Yes	No	No
Quinoa consuming only	No	No	Yes	No	No	Yes	No
Quinoa producing only	No	No	No	No	No	No	Yes

Note: *, **, and *** denote statistical significance at the 10, 5, and 1 percent levels, respectively.

Standard errors clustered at the district level are shown in parentheses. Each household is weighted according to the sampling weight it was given in the ENAHO. In addition to being expressed in constant (i.e., 2004) terms, all prices are deflated using departmental-level deflators.

Bellemare, Fajardo-Gonzalez, and Gitter (2018)

Table 5a. Pseudo-Panel Regression of Total Household Consumption on the Proportion of Households Who Produce Quinoa Treating Districts as Units of Observation, 2004-2014.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Dependent Variable: Mean Log of Total Household Consumption (in 2004 PEN)							
% of Quinoa Producers x Log of International Price of Quinoa	0.021***	0.013**	0.015***	0.014***	0.013**	0.007	0.006
	(0.006)	(0.005)	(0.006)	(0.005)	(0.005)	(0.006)	(0.007)
Constant	8.938***	8.703***	8.695***	-67.769***	-65.428***	8.794***	8.343***
	(0.003)	(0.008)	(0.003)	(2.151)	(2.310)	(0.010)	(0.020)
Observations	10,774	10,774	10,774	10,774	10,774	7,814	2,678
R-squared	0.003	0.174	0.412	0.246	0.210	0.169	0.180
District Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Linear Trend	No	Yes	No	No	Yes	Yes	Yes
Year Fixed Effects	No	No	No	No	No	No	No
District-Specific Linear Trends	No	No	Yes	No	No	No	No
Province-Specific Linear Trends	No	No	No	Yes	No	No	No
Province-Year Fixed Effects	No	No	No	No	No	No	No
Department-Specific Linear Trends	No	No	No	No	Yes	No	No
Quinoa Producing Only	No	No	No	No	No	No	Yes
Quinoa Consuming Only	No	No	No	No	No	Yes	No

Note: *, **, and *** denote statistical significance at the 10, 5, and 1 percent levels, respectively.

Standard errors clustered at the district level are shown in parentheses. Each household is weighted according to the sampling weight it was given in the ENAHO. In addition to being expressed in constant (i.e., 2004) terms, all prices are deflated using departmental-level deflators.

Summary

- With observational data, you almost always should rely on FE instead of RE. With experimental data, you almost always should rely on RE instead of FE. Only in the latter case can you credibly use the result of a Hausman test to guide your choice of estimator.

Summary

- Panel data are not a panacea for endogeneity issues. They certainly help, but it pays to ascertain how they help in the face of our usual three ruffians of endogeneity—unobserved heterogeneity, reverse causality, and measurement error.
- More so than in the absence of panel data, it helps to think about potential violations of the SUTVA. There almost always are some spillover effects that can compromise identification. Sometimes, you can gauge whether they do.

Summary

- Whatever estimator you settle on, it helps to show both the results of a pooled OLS (i.e., a naive specification that ignores the panel structure) as well as your FE results.

Summary

- If your data follow units over time, in addition to unit FEs, it helps to look at (i) a linear time trend, (ii) time-period FEs, (iii) unit-specific linear trends and, if possible, (iv) higher-level units-year FEs. For example, in a state panel, you'd have specifications that control for (i) state FEs with a linear time trend, (ii) state FEs with year FEs, (ii) state FEs as well as state-specific trends, and (iv) census region-year FEs in addition to state FEs. Autor (2003) is a good example of what to do with panel data in terms of robustness checks.

Summary

- As in so many other cases, it is best to stick to linear estimators when you have panel data. ML-based estimators do especially badly with FEs because of the incidental parameter problem. So instead of relying on a probit with FEs, use an LPM to incorporate your FEs. In Bellemare, Novak, and Steinmetz (2015), we offer a discussion of why LPM should be preferred.
- With aggregate data, you should use the DID estimator.

Summary

- The DID estimator is most clearly understood with a dichotomous treatment and few time periods.
- The DID relies on three important assumptions (i) parallel trends, (ii) no Ashenfelter dips, and (iii) no autocorrelation.

Summary

- Regarding autocorrelation, the thing to do nowadays is to cluster at the level of the unit of observation. Clustering is more powerful than correcting for heteroskedasticity or autocorrelation, or HAC-style covariances in that clustering accounts for arbitrary within-unit correlation.
- Repeated cross sections can sometimes lend themselves to pseudo-panel techniques, wherein a village, a cohort, etc. is the unit of observation.