Chapter 5

Parallel Worlds: Fixed Effects, Differences-in-differences, and Panel Data

> The first thing to realize about parallel universes . . . is that they are not parallel.
>
> Douglas Adams, *Mostly Harmless* (1995)

The key to causal inference in chapter 3 is control for observed confounding factors. If important confounders are unobserved, we might try to get at causal effects using IV as discussed in Chapter 4. Good instruments are hard to find, however, so we'd like to have other tools to deal with unobserved confounders. This chapter considers a variation on the control theme: strategies that use data with a time or cohort dimension to control for unobserved-but-fixed omitted variables. These strategies punt on comparisons in levels, while requiring the counterfactual *trend* behavior of treatment and control groups to be the same. We also discuss the idea of controlling for lagged dependent variables, another strategy that exploits timing.

5.1 Individual Fixed Effects

One of the oldest questions in Labor Economics is the connection between union membership and wages. Do workers whose wages are set by collective bargaining earn more because of this, or would they earn more anyway? (Perhaps because they are more experienced or skilled). To set this question up, let Y_{it} equal the (log) earnings of worker i at time t and let D_{it} denote his union status. The observed Y_{it} is either Y_{0it} or Y_{1it}, depending on union status. Suppose further that

$$E(\text{Y}_{0it}|A_i, \text{X}_{it}, t, \text{D}_{it}) = E(\text{Y}_{0it}|A_i, \text{X}_{it}, t),$$

i.e. union status is as good as randomly assigned conditional on unobserved worker ability, A_i, and other observed covariates X_{it}, like age and schooling.

The key to fixed-effects estimation is the assumption that the unobserved A_i appears without a time subscript in a linear model for $E(\mathrm{Y}_{0it}|A_i,\mathrm{X}_{it},t)$:

$$E(\mathrm{Y}_{0it}|A_i,\mathrm{X}_{it},t) = \alpha + \lambda_t + A_i'\gamma + \mathrm{X}_{it}\delta, \tag{5.1.1}$$

Finally, we assume that the causal effect of union membership is additive and constant:

$$E(\mathrm{Y}_{1it}|A_i,\mathrm{X}_{it},t) = E(\mathrm{Y}_{0it}|A_i,\mathrm{X}_{it},t) + \rho.$$

This implies

$$E(\mathrm{Y}_{it}|A_i,\mathrm{X}_{it},t,\mathrm{D}_{it}) = \alpha + \lambda_t + \rho\mathrm{D}_{it} + A_i'\gamma + \mathrm{X}_{it}\delta, \tag{5.1.2}$$

where ρ is the causal effect of interest. The set of assumptions leading to (5.1.2) is more restrictive those we used to motivate regression in Chapter 3; we need the linear, additive functional form to make headway on the problem of *unobserved* confounders using panel data with no instruments.[1]

Equation (5.1.2) implies

$$\mathrm{Y}_{it} = \alpha_i + \lambda_t + \rho\mathrm{D}_{it} + \mathrm{X}_{it}\delta + \varepsilon_{it}. \tag{5.1.3}$$

where

$$\alpha_i \equiv \alpha + A_i'\gamma.$$

This is a *fixed-effects model.* Given panel data, i.e., repeated observations on individuals, the causal effect of union status on wages can be estimated by treating α_i, the fixed effect, as a parameter to be estimated. The *year effect*, λ_t, is also treated as a parameter to be estimated. The unobserved individual effects are coefficients on dummies for each individual while the year effects are coefficients on time dummies.[2]

It might seem like there are an awful lot of parameters to be estimated in the fixed effects model. For

[1] In some cases, we can allow heterogeneous treatment effects so that

$$E(\mathrm{Y}_{1it} - \mathrm{Y}_{0it}|A_i,\mathrm{X}_{it},t) = \rho_i.$$

See, e.g., Wooldridge (2005), who discusses estimators for the average of ρ_i.

[2] An alternative to the fixed-effects specification is "random effects" (See, e.g., Wooldridge, 2006). The random-effects model assumes that α_i is uncorrelated with the regressors. Because the omitted variable in a random-effects model is uncorrelated with included regressors there is no bias from ignoring it - in effect, it becomes part of the residual. The most important consequence of random effects is that the residuals for a given person are correlated across periods. Chapter 8 discusses the implications of this for standard errors. An alternative approach is GLS, which promises to be more efficient if the assumptions of the random-effects model are satisfied (linear CEF, homoskedasticity). We prefer OLS/fix-the-standard-errors to GLS under random-effects assumptions. As discussed in Section 3.4.1, GLS requires stronger assumptions than those we are comfortable with and the resulting efficiency gain is likely to be modest.

example, the Panel Survey of Income Dynamics, a widely-used panel data set, includes about 5,000 working-age men observed for about 20 years. So there are roughly 5,000 fixed effects. In practice, however, this doesn't matter. Treating the individual effects as parameters to be estimated is algebraically the same as estimation in deviations from means. In other words, first we calculate the individual averages

$$\overline{\mathrm{Y}}_i = \alpha_i + \overline{\lambda} + \rho\overline{\mathrm{D}}_i + \overline{\mathrm{X}}_i\delta + \overline{\varepsilon}_i.$$

Subtracting this from (5.1.3) gives

$$\mathrm{Y}_{it} - \overline{\mathrm{Y}}_i = \lambda_t - \overline{\lambda} + \rho\left(\mathrm{D}_{it} - \overline{\mathrm{D}}_i\right) + \left(\mathrm{X}_{it} - \overline{\mathrm{X}}_i\right)\delta + \left(\varepsilon_{it} - \overline{\varepsilon}_i\right), \tag{5.1.4}$$

so deviations from means kills the unobserved individual effects.[3]

An alternative to deviations from means is differencing. In other words, we estimate,

$$\Delta\mathrm{Y}_{it} = \Delta\lambda_t + \rho\Delta\mathrm{D}_{it} + \Delta\mathrm{X}_{it}\delta + \Delta\varepsilon_{it}, \tag{5.1.5}$$

where the Δ prefix denotes the change from one year to the next. For example, $\Delta\mathrm{Y}_{it} = \mathrm{Y}_{it} - \mathrm{Y}_{it-1}$. With two periods, differencing is algebraically the same as deviations from means, but not otherwise. Both should work, although with homoskedastic and serially uncorrelated ε_{it} deviations from means is more efficient. You might find differencing more convenient if you have to do it by hand, though the differenced standard errors should be adjusted for the fact that the differenced residuals are serially correlated.

Some regression packages automate the deviations-from-means estimator, with an appropriate standard-error adjustment for the degrees of freedoms lost in estimating N individual means. This is all that's needed to get the standard errors right with a homoskedastic, serially uncorrelated residual. The deviations-from-means estimator has many names, including the "within estimator" and "analysis of covariance". Estimation in deviations-from-means form is also called *absorbing* the fixed effects.[4]

Freeman (1984) uses four data sets to estimate union wage effects under the assumption that selection into union status is based on unobserved-but-fixed individual characteristics. Table 5.1.1 displays some of his estimates. For each data set, the table displays results from a fixed-effects estimator and the corresponding cross-section estimates. The cross section estimates are typically higher (ranging from .15-.25) than the

[3] Why is deviations from means the same as estimating each fixed effect in (5.1.3)? Because, by the regression anatomy formula, (3.1.3), any set of multivariate regression coefficients can be estimated in two steps. To get the multivariate coefficient on one set of variables, first regress them on all the other included variables, then regress the original dependent variable on the residuals from this first step. The residuals from a regression on a full set of person-dummies in a person-year panel are deviations from person means.

[4] The fixed effects are not estimated consistently in a panel where the number of periods T is fixed while $N \to \infty$. This is called the "incidental parameters problem," a name which reflects the fact that the number of parameters grows with the sample size. Nevertheless, other parameters in the fixed effects model - the ones we care about - are consistently estimated.

fixed effects estimates (ranging from .10-.20). This may indicate positive selection bias in the cross-section estimates, though selection bias is not the only explanation for the lower fixed-effects estimates.

Table 5.1.1: Estimated effects of union status on log wages

Survey	Cross section estimate	Fixed effects estimate
May CPS, 1974-75	0.19	0.09
National Longitudinal Survey of Young Men, 1970-78	0.28	0.19
Michigan PSID, 1970-79	0.23	0.14
QES, 1973-77	0.14	0.16

Notes: Adapted from Freeman (1984). The table reports cross-section and panel estimates of the union relative wage effect. The estimates were calculated using the surveys listed at left. The cross-section estimates include controls for demographic and human capital variables.

Although they control for a certain type of omitted variable, fixed-effects estimates are notoriously susceptible to attenuation bias from measurement error. On one hand, economic variables like union status tend to be persistent (a worker who is a union member this year is most likely a union member next year). On the other hand, measurement error often changes from year-to-year (union status may be misreported or miscoded this year but not next year). Therefore, while union status may be misreported or miscoded for only a few workers in any single year, the observed year-to-year changes in union status may be mostly noise. In other words, there is more measurement error in the regressors in an equation like (5.1.5) or (5.1.4) than in the levels of the regressors. This fact may account for smaller fixed-effects estimates.[5]

A variant on the measurement-error problem arises from that fact that the differencing and deviations-from-means estimators used to control for fixed effects typically remove both good and bad variation. In other words, these transformations may kill some of the omitted-variables-bias bathwater, but they also remove much of the useful information in the baby - the variable of interest. An example is the use of twins to estimate the causal effect of schooling on wages. Although there is no time dimension to this problem, the basic idea is the same as the union problem discussed above: twins have similar but largely unobserved family and genetic backgrounds. We can therefore control for their common family background by including a family fixed effect in samples of pairs of twins.

Ashenfelter and Krueger (1994) and Ashenfelter and Rouse (1998) estimate the returns to schooling using samples of twins, controlling for family fixed effects. Because there are two twins from each family, this is the same as regressing differences in earnings within twin-pairs on differences in their schooling. Surprisingly, the with-family estimates come our larger than OLS. But how do differences in schooling come about between individuals who are otherwise so much alike? Bound and Solon (1999) point out that there are small differences between twins, with first-borns typically having higher birth weight and higher IQ scores (here differences in birth timing are measured in minutes). While these within-twin differences are not large,

[5] See Griliches and Hausman (1986) for a more complete analysis of measurement error in panel data.

neither is the difference in their schooling. Hence, a small amount of unobserved ability differences among twins could be responsible for substantial bias in the resulting estimates.

What should be done about measurement error and related problems in models with fixed effects? A possible fix-up for measurement error is instrumental variables. Ashenfelter and Krueger (1994) use cross-sibling reports to construct instruments for schooling differences across twins. For example, they use each twin's report of his brother's schooling as an instrument for self-reports. A second approach is to bring in external information on the extent of measurement error and adjust naive estimates accordingly. In a study of union wage effects, Card (1996) uses external information from a separate validation survey to adjust panel-data estimates for measurement error in reported union status. But data from multiple reports and repeated measures of the sort used by Ashenfelter and Rouse (1994) and Card (1996) are unusual. At a minimum, therefore, it's important to avoid overly strong claims when interpreting fixed-effects estimates (never bad advice for an applied econometrician in any case).

5.2 Differences-in-differences: Pre and Post, Treatment and Control

The fixed effects strategy requires panel data, that is, repeated observations on the same individuals (or firms or whatever the unit of observation might be). Often, however, the regressor of interest varies only at a more aggregate level such as state or cohort. For example, state policies regarding health care benefits for pregnant workers or minimum wages change across states but not within states. The source of omitted variables bias when evaluating these policies must therefore be unobserved variables at the state and year level.

To make this concrete, suppose we are interested in the effect of the minimum wage on employment, a classic question in Labor Economics. In a competitive labor market, increases in the minimum wage move us up a downward-sloping demand curve. Higher minimums therefore reduce employment, perhaps hurting the very workers minimum-wage policies were designed to help. Card and Krueger (1994) use a dramatic change in the New Jersey state minimum wage to see if this is true.

On April 1, 1992, New Jersey raised the state minimum from $4.25 to $5.05. Card and Krueger collected data on employment at fast food restaurants in New Jersey in February 1992 and again in November 1992. These restaurants (Burger King, Wendy's, and so on) are big minimum-wage employers. Card and Krueger collected data from the same type of restaurants in eastern Pennsylvania, just across the Delaware river. The minimum wage in Pennsylvania stayed at $4.25 throughout this period. They used their data set to compute differences-in-differences (DD) estimates of the effects of the New Jersey minimum wage increase. That is, they compared the change in employment in New Jersey to the change in employment in Pennsylvania around the time New Jersey raised its minimum.

DD is a version of fixed-effects estimation using aggregate data.[6] To see this, let

$$\begin{aligned} \mathrm{Y}_{1ist} &= \text{fast food employment at restaurant } i \text{ and period } t \\ &\quad \text{if there is a high state minimum wage} \\ \mathrm{Y}_{0ist} &= \text{fast food employment at restaurant } i \text{ and period } t \\ &\quad \text{if there is a low state minimum wage} \end{aligned}$$

These are potential outcomes - in practice, we only get to see one or the other. Fort example, we see Y_{1ist} in New Jersey in November of 1992. The heart of the DD setup is an additive structure for potential outcomes in the no-treatment state. Specifically, we assume that

$$E(\mathrm{Y}_{0ist}|s,t) = \gamma_s + \lambda_t \tag{5.2.1}$$

where s denotes state (New Jersey or Pennsylvania) and t denotes period (February, before the minimum wage increase or November, after the increase). This equations says that in the absence of a minimum wage change, employment is determined by the sum of a time-invariant state effect and a year effect that is common across states. The additive state effect plays the role of the unobserved individual effect in the previous subsection.

Let D_{st} be a dummy for high-minimum-wage states, where states are index by s and observed in period t. Assuming that $E(\mathrm{Y}_{1ist} - \mathrm{Y}_{0ist}|s,t)$ is a constant, denoted β, we have:

$$\mathrm{Y}_{ist} = \gamma_s + \lambda_t + \beta \mathrm{D}_{st} + \varepsilon_{ist} \tag{5.2.2}$$

where $E(\varepsilon_{ist}|s,t) = 0$. From here, we get

$$E[\mathrm{Y}_{ist}|s = PA, t = Nov] - E(\mathrm{Y}_{ist}|s = PA, t = Feb) = \lambda_{Nov} - \lambda_{Feb}$$

and

$$E(\mathrm{Y}_{ist}|s = NJ, t = Nov) - E(\mathrm{Y}_{ist}|s = NJ, t = Feb) = \lambda_{Nov} - \lambda_{Feb} + \beta.$$

The population difference-in-differences,

$$[E(\mathrm{Y}_{ist}|s = PA, t = Nov) - E(\mathrm{Y}_{ist}|s = PA, t = Feb)]$$

$$- [E(\mathrm{Y}_{ist}|s = NJ, t = Nov) - E(\mathrm{Y}_{ist}|s = NJ, t = Feb)] = \beta,$$

[6] The DD idea is at least as old as IV. Kennan (1995) references a 1915 BLS report using DD to study the employment effects of the minimum wage (Obenauer and von der Nienburg, 1915).

is the causal effect of interest. This is easily estimated using the sample analog of the population means.

Table 5.2.1: Average employment per store before and after the New Jersey minimum wage increase

Variable		PA (i)	NJ (ii)	Difference, NJ-PA (iii)
1.	FTE employment before,	23.33	20.44	-2.89
	all available observations	(1.35)	(0.51)	(1.44)
2.	FTE employment after,	21.17	21.03	-0.14
	all available observations	(0.94)	(0.52)	(1.07)
3.	Change in mean FTE	-2.16	0.59	2.76
	employment	(1.25)	(0.54)	(1.36)

Notes: Adapted from Card and Krueger (1994), Table 3. The table reports average full-time equivalent (FTE) employment at restaurants in Pennsylvania and New Jersey before and after a minimum wage increase in New Jersey. The sample consists of all stores with data on employment. Employment at six closed stores is set to zero. Employment at four temporarily closed stores is treated as missing. Standard errors are reported in parentheses

Table 5.2.1 (based on Table 3 in Card and Krueger, 1994) shows average employment at fast food restaurants in New Jersey and Pennsylvania before and after the change in the New Jersey minimum wage. There are four cells in the first two rows and columns, while the margins show state differences in each period, the changes over time in each state, and the difference-in-differences. Employment in Pennsylvania restaurants is somewhat higher than in New Jersey in February but falls by November. Employment in New Jersey, in contrast, increases slightly. These two changes produce a positive difference-in-differences, the opposite of what we might expect if a higher minimum wage pushes businesses up the labor demand curve.

How convincing is this evidence against the standard labor-demand story? The key identifying assumption here is that employment *trends* would be the same in both states in the absence of treatment. 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 in captured by the state fixed effect, which plays the same role as the unobserved individual effect in (5.1.3).[7]

The common trends assumption can be investigated using data on multiple periods. In an update of their

[7]The common trends assumption can be applied to transformed data, for example,

$$E(\log Y_{0ist}|s,t) = \gamma_s + \lambda_t.$$

Note, however, that if there is a common trend in logs, there will not be one in levels and vice versa. Athey and Imbens (2006) introduce a semi-parametric DD estimator that allows for common trends after an unknown transformation, which they propose to use the data to estimate. Poterba, Venti and Wise (1995) and Meyer, Viscusi, and Durbin (1995) discuss DD-type models for quantiles.

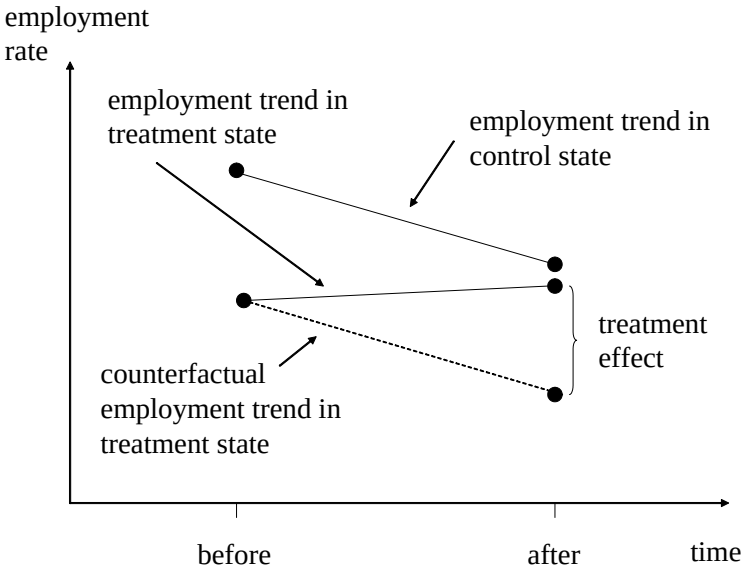

Figure 5.2.1: Causal effects in the differences-in-differences model

original minimum wage study, Card and Krueger (2000) obtained administrative payroll data for restaurants in New Jersey and Pennsylvania for a number of years. These data are shown here in Figure 5.2.2, similar to Figure 2 in their follow-up study. The vertical lines indicate the dates when their original surveys were conducted, and the third vertical line denotes the increase in the federal minimum wage to $4.75 in October 1996, which affected Pennsylvania but not New Jersey. These data give us an opportunity to look at a new minimum wage "experiment".

Like the original Card and Krueger survey, the administrative data show a slight decline in employment from February to November 1992 in Pennsylvania, and little change in New Jersey over the same period. However, the data also reveal fairly substantial year-to-year employment variation in other periods. These swings often seem to differ substantially in the two states. In particular, while employment levels in New Jersey and Pennsylvania were similar at the end of 1991, employment in Pennsylvania fell relative to employment in New Jersey over the next three years (especially in the 14-county group), mostly before the 1996 change in Federal minimum. So Pennsylvania may not provide a very good measure of counterfactual employment rates in New Jersey in the absence of a policy change, and vice versa.

A more encouraging example comes from Pischke (2007), who looks at the effect of school term length on student performance using variation generated by a sharp policy change in Germany. Until the 1960s, children in all German states except Bavaria started school in the Spring. Beginning in the 1966-67 school year, the Spring-starters moved to start school in the Fall. The transition to a Fall start required two short school years for affected cohorts, 24 weeks long instead of 37. Students in these cohorts effectively had their time in school compressed relative to cohorts on either side and relative to students in Bavaria, which

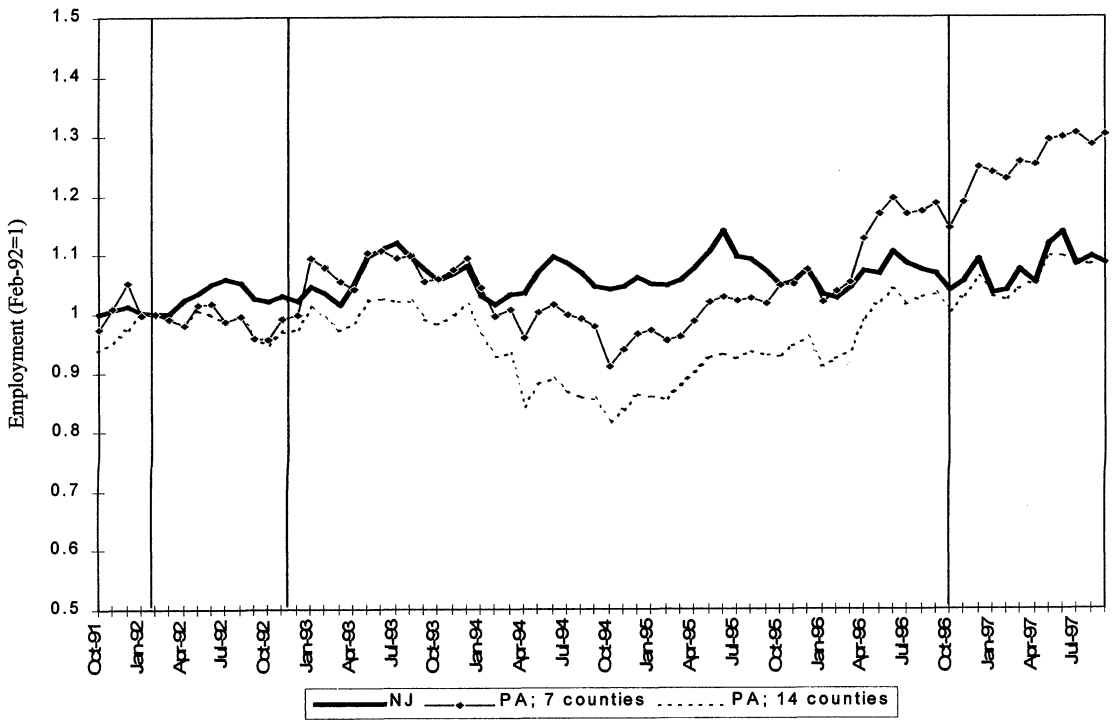

Figure 5.2.2: Employment in New Jersey and Pennsylvania fast-food restaurants, October 1991 to September 1997 (from Card and Krueger 2000). Vertical lines indicate dates of the original Card and Krueger (1994) survey and the October 1996 federal minimum-wage increase.

already had a Fall start.

Figure 5.2.3 plots the likelihood of grade repetition for the 1962-73 cohorts of 2nd graders in Bavaria and affected states (there are no repetition data for 1963-65). Repetition rates in Bavaria were reasonably flat from 1966 onwards at around 2.5%. Repetition rates are higher in the short-school-year states, at around 4 - 4.5% in 1962 and 1966, before the change in term length. But repetition rates jump up by about a percentage point for the two affected cohorts in these states, a bit more so for the second cohort than the first, before falling back to the baseline level. This graph provides strong visual evidence of treatment and control states with a common underlying trend, and a treatment effect that induces a sharp but transitory deviation from this trend. A shorter school year seems to have increased repetition rates for affected cohorts.

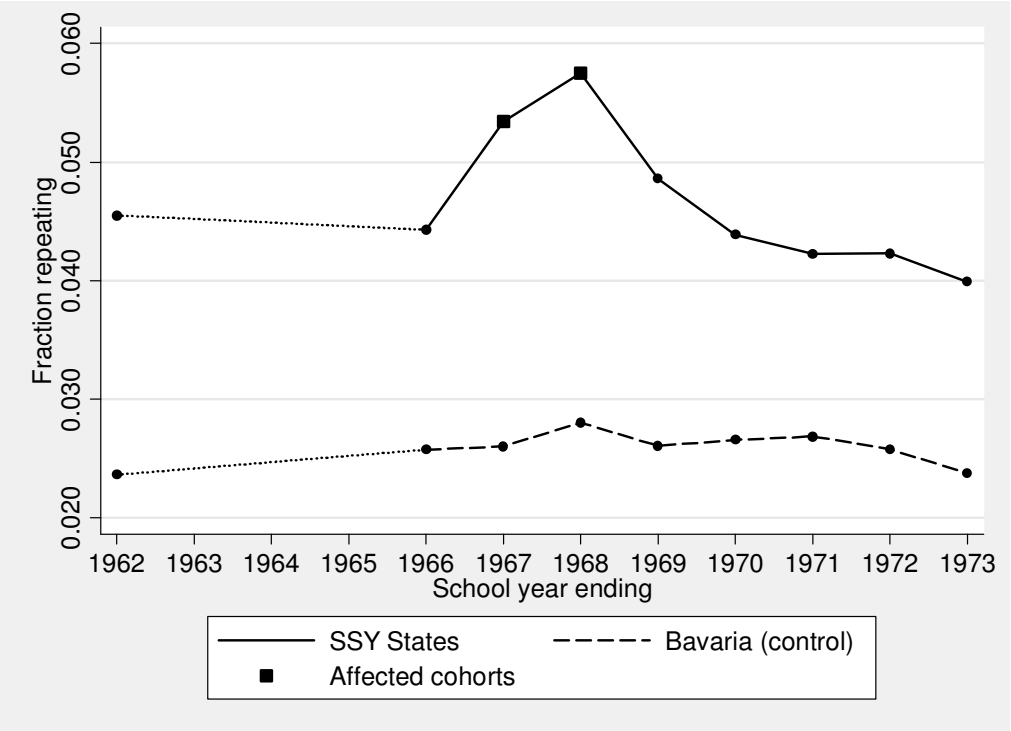

Figure 5.2.3: Average rates of grade repetition in second grade for treatment and control schools in Germany (from Pischke 2007). The data span a period before and after a change in term length for students outside of Bavaria.

5.2.1 Regression DD

As with the fixed effects model, we can use regression to estimate equations like (5.2.2). Let NJ_s be a dummy for restaurants in New Jersey and d_t be a time-dummy that switches on for observations obtained

in November (i.e., after the minimum wage change). Then

$$\text{Y}_{ist} = \alpha + \gamma NJ_s + \lambda d_t + \beta(NJ_s \cdot d_t) + \varepsilon_{ist} \tag{5.2.3}$$

is the same as (5.2.2) where $NJ_s \cdot d_t = \text{D}_{st}$. In the language of Section 3.1.4, this model includes two main effects for state and year and an interaction term that marks observations from New Jersey in November. This is a saturated model since the conditional mean function $E(\text{Y}_{ist}|s,t)$ takes on four possible values and there are four parameters. The link between the parameters in the regression equation, (5.2.3), and those in the DD model for the conditional mean function, (5.2.2), is

$$\begin{aligned}
\alpha &= E(\text{Y}_{ist}|s = PA, t = Feb) = \gamma_{PA} + \lambda_{Feb} \\
\gamma &= E(\text{Y}_{ist}|s = NJ, t = Feb) - E(\text{Y}_{ist}|s = PA, t = Feb) = \gamma_{NJ} - \gamma_{PA} \\
\lambda &= E(\text{Y}_{ist}|s = PA, t = Nov) - E(\text{Y}_{ist}|s = PA, t = Feb) = \lambda_{Nov} - \lambda_{Feb} \\
\beta &= \{E(\text{Y}_{ist}|s = NJ, t = Nov) - E(\text{Y}_{ist}|s = NJ, t = Feb)\} \\
&\quad -\{E(\text{Y}_{ist}|s = PA, t = Nov) - E(\text{Y}_{ist}|s = PA, t = Feb)\}.
\end{aligned}$$

The regression formulation of the difference-in-difference model offers a convenient way to construct DD estimates and standard errors. It's also easy to add additional states or periods to the regression set-up. We might for example, add additional control states and pre-treatment periods to the New Jersey/Pennsylvania sample. The resulting generalization of (5.2.3) includes a dummy for each state and period but is otherwise unchanged.

A second advantage of regression-DD is that it facilitates empirical work with regressors other than switched-on/switched-off dummy variables. Instead of New Jersey and Pennsylvania in 1992, for example, we might look at all state minimum wages in the United States. Some of these are a little higher than the federal minimum (which covers everyone regardless of where they live), some are a lot higher, and some are the same. The minimum wage is therefore a variable with differing "treatment intensity" across states and over time. Moreover, in addition to statutory variation in state minima, the local importance of a minimum wage varies with average state wage levels. For example, the early-1990s Federal minimum of $4.25 was probably irrelevant in Connecticut - with high average wages - but a big deal in Mississippi.

Card (1992) exploits regional variation in the impact of the federal minimum wage. His approach is motivated by an equation like

$$\text{Y}_{ist} = \gamma_s + \lambda_t + \beta(\text{FA}_s \cdot d_t) + \varepsilon_{ist} \tag{5.2.4}$$

where the variable FA_s is a measure of the fraction of teenagers likely to be affected by a minimum wage increase in each state and d_t is a dummy for observations after 1990, when the federal minimum increased from $3.35 to $3.80. The FA_s variable measures the baseline (pre-increase) proportion of each state's teen

labor force earning less than $3.80.

As in the New Jersey/Pennsylvania study, Card (1992) works with data from two periods, before and after, in this case 1989 and 1992. But this study uses 51 states (including the District of Columbia), for a total of 102 state-year observations. Since there are no individual-level covariates in (5.2.4), this is the same as estimation with micro data (provided the group-level estimates are weighted by cell size). Note that $\text{FA}_s \cdot d_t$ is an interaction term, like $NJ_s \cdot d_t$ in (5.2.3), though here the interaction term takes on a distinct value for each observation in the data set. Finally, because Card (1992) analyzes data for only two periods, the reported estimates are from an equation in first-differences:

$$\Delta\bar{\text{Y}}_s = \lambda^* + \beta\text{FA}_s + \Delta\bar{\varepsilon}_s,$$

where $\Delta\bar{\text{Y}}_s$ is the change in average teen employment in state s and $\Delta\bar{\varepsilon}_s$ is the error term in the differenced equation.[8]

Table 5.2.2, based on Table 3 in Card (1992), shows that wages increased more in states where the minimum wage increase is likely to have had more bite (see the estimate of .15 in column 1). This is an important step in Card's analysis - it verifies the notion that the *fraction affected* variable is a good predictor of the wage changes induced by an increase in the federal minimum. Employment, on the other hand, seems largely unrelated to *fraction affected*, as can be seen in column 3. Thus, the results in Card (1992) are in line with the results from the New Jersey/Pennsylvania study.

Table 5.2.2: Regression-DD estimates of minimum wage effects on teens, 1989 to 1992

	Equations for Change in Mean Log Wage:		Equations for change in Teen Employment-Population Ratio:	
Explanatory Variable	(1)	(2)	(3)	(4)
1. Fraction of Affected Teens	0.15 (0.03)	.14 (0.04)	0.02 (0.03)	-.01 (0.03)
2. Change in Overall Emp./Pop. Ratio	–	0.46 (0.60)	–	1.24 (0.60)
3. R-squared	0.30	0.31	0.01	0.09

Notes: Adapted from Card (1992). The table reports estimates from a regression of the change in average teen employment by state on the fraction of teens affected by a change in the federal minimum wage in each state. Data are from the 1989 and 1992 CPS. Regressions are weighted by the CPS sample size by state and year.

Card's (1992) analysis illustrates a further advantage of regression-DD: it's easy to add additional covariates in this framework. For example, we might like to control for adult employment as a source of omitted

[8] Card weights estimates of (5.2.4) by the sample size used to construct averages for each state. Other specifications in the spirit of (5.2.4) put a normalized function of state and federal minimum wages on the right hand side instead of $\text{FA}_s \cdot d_t$. See, for example, Neumark and Wascher (1992), who work with the difference between state and federal minima, adjusted for minimum-wage coverage provisions, and normalized by state average hourly wages.

state-specific trends. In other words, we can model counterfactual employment in the absence of a change in the minimum wage as

$$E[\mathrm{Y}_{0ist}|s,t,\mathrm{X}_{st}] = \gamma_s + \lambda_t + \mathrm{X}'_{st}\delta.$$

where X_{st} is a vector of state-and-time-varying covariates, including adult employment (though this may not be kosher if adult employment also responds to the minimum wage change, in which case it's *bad control*; see Section 3.2.3). As it turns out, the addition of an adult employment control has little effect on Card's estimates, as can be seen in columns 2 and 4 in Table 5.2.2.

It's worth emphasizing the fact that Card (1992) analyzes state averages instead of individual data. He might have used a pooled multi-year sample of micro data from the CPS to estimate an equation like

$$\mathrm{Y}_{ist} = \gamma_s + \lambda_t + \beta(\mathrm{FA}_s \cdot d_t) + \mathrm{X}'_{ist}\delta + \varepsilon_{ist}, \tag{5.2.5}$$

where X_{ist} can include individual level characteristics such as race. The covariate vector might also include time-varying variables measured at the state level. Only the latter are likely to be a source of omitted variables bias, but individual-level controls can increase precision, a point we noted in Section 2.3. Inference is a little more complicated in a framework that combines of micro data on dependent variables with group-level regressors, however. The key issue is how best to adjust for possible group-level random effects, as we discuss in Chapter 8, below.

When the sample includes many years, the regression-DD model lends itself to a test for causality in the spirit of Granger (1969). The Granger idea is to see whether causes happen before consequences and not vice versa (though as we know from the epigram at the beginning of Chapter 4, this alone is not sufficient for causal inference). Suppose the policy variable of interest, D_{st}, changes at different times in different states. In this context, Granger causality testing means a check on whether, conditional on state and year effects, past D_{st} predicts Y_{ist} while future D_{st} does not. If D_{st} causes Y_{ist} but not vice versa, then leads should not matter in an equation like:

$$\mathrm{Y}_{ist} = \gamma_s + \lambda_t + \sum_{\tau=0}^{m} \beta_{-\tau}\mathrm{D}_{s,t-\tau} + \sum_{\tau=1}^{q} \beta_{+\tau}\mathrm{D}_{s,t+\tau}\mathrm{X}_{ist}\delta + \varepsilon_{ist}, \tag{5.2.6}$$

where the sums on the right-hand side allow for m lags $(\beta_{-1}, \beta_{-2}, ..., \beta_{-m})$ or post-treatment effects and q leads $(\beta_{+1}, \beta_{+1}, ..., \beta_{+q})$ or anticipatory effects. The pattern of lagged effects is usually of substantive interest as well. We might, for example, believe that causal effects should grow or fade as time passes.

Autor (2003) implements the Granger test in an investigation of the effect of employment protection on firms' use of temporary help. Employment protection is a type of labor law - promulgated by state legislatures or, more typically, through common law as made by state courts - that makes it harder to fire workers. As a rule, U.S. labor law allows "employment at will," which means that workers can be fired for

just cause or no cause, at the employer's whim. But some state courts have allowed a number of exceptions to the employment-at-will doctrine, leading to lawsuits for "unjust dismissal". Autor is interested in whether fear of employee lawsuits makes firms more likely to use temporary workers for tasks for which they would otherwise have increased their workforce. Temporary workers work for someone else besides the firm for which they are executing tasks. As a result, the firm using them cannot be sued for unjust dismissal when they let temporary workers go.

Autor's empirical strategy relates the employment of temporary workers in a state to dummy variables indicating state court rulings that allow exceptions to the employment-at-will doctrine. His regression-DD model includes both leads and lags, as in equation (5.2.6). The estimated leads and lags, running from two years ahead to 4 years behind, are plotted in Figure 5.2.4, a reproduction of Figure 3 from Autor (2003). The estimates show no effects in the two years before the courts adopted an exception, with sharply increasing effects on temporary employment in the first few years after the adoption, which then appear to flatten out with a permanently higher rate of temporary employment in affected states. This pattern seems consistent with a causal interpretation of Autor's results.

An alternative check on the DD identification strategy adds state-specific time trends to the regressors in X_{ist}. In other words, we estimate

$$\mathrm{Y}_{ist} = \gamma_{0s} + \gamma_{1st} + \lambda_t + \beta \mathrm{D}_{st} + \mathrm{X}'_{ist}\delta + \varepsilon_{ist}, \tag{5.2.7}$$

where γ_{0s} is a state-specific intercept as before and γ_{1s} is a state-specific trend coefficient multiplying the time-trend variable, t. This allows treatment and control states to follow different trends in a limited but potentially revealing way. It's heartening to find that the estimated effects of interest are unchanged by the inclusion of these trends, and discouraging otherwise. Note, however, that we need at least 3 periods to estimate a model with state-specific trends. Moreover, in practice, 3 periods is typically inadequate to pin down both the trends and the treatment effect. As a rule, DD estimation with state-specific trends is likely to be more robust and convincing when the pre-treatment data establish a clear trend that can be extrapolated into the post-treatment period.

In a study of the effect of labor regulation on businesses in Indian states, Besley and Burgess (2004)use state trends as a robustness check. Different states change regulatory regimes at different times, giving rise to a DD research design. As in Card (1992), the unit of observation in Besley and Burgess (2004) is a state-year average. Table 5.2.3 (based on Table IV in their paper) reproduces the key results.

The estimates in column 1, from a regression-DD model without state-specific trends, suggest that labor regulation leads to lower output per capita. The models used to construct the estimates in columns 2 and 3 add time-varying state-specific covariates like government expenditure per capita and state population. This is in the spirit of Card's (1992) addition of state-level adult employment rates as a control in the minimum

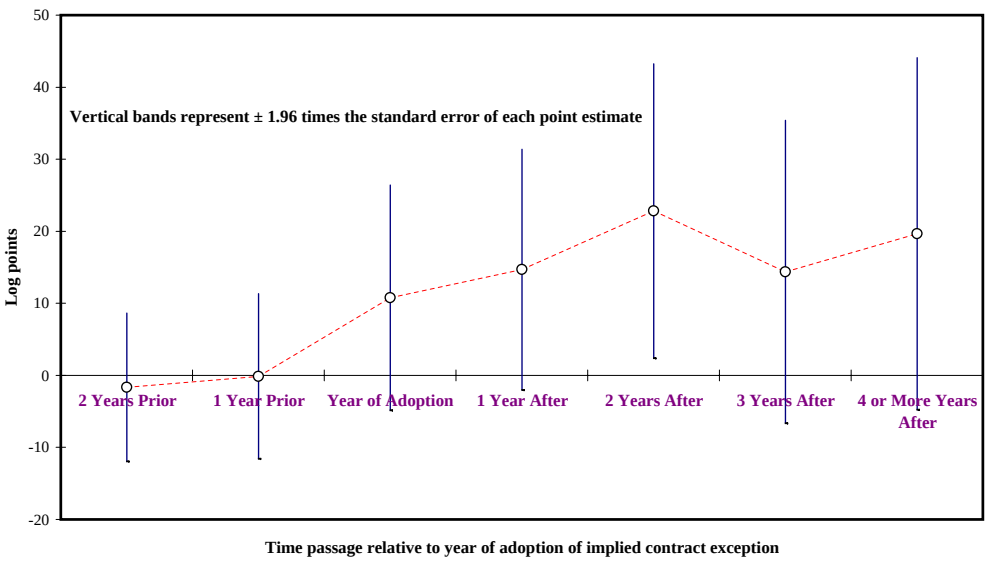

Figure 5.2.4: Estimated impact of state courts' adoption of an implied-contract exception to the employment-at-will doctrine on use of temporary workers (from Autor 2003). The dependent variable is the log of state temporary help employment in 1979 - 1995. Estimates are from a model that allows for effects before, during, and after adoption.

Table 5.2.3: Effect of labor regulation on the performance of firms in Indian states

	(1)	(2)	(3)	(4)
Labor regulation (lagged)	-0.186	-0.185	-0.104	0.0002
	(.0641)	(.0507)	(.039)	(.02)
Log development		0.240	0.184	0.241
expenditure per capita		(.1277)	(.1187)	(.1057)
Log installed electricity		0.089	0.082	0.023
capacity per capita		(.0605)	(.0543)	(.0333)
Log state population		0.720	0.310	-1.419
		(.96)	(1.1923)	(2.3262)
Congress majority			-0.0009	0.020
			(.01)	(.0096)
Hard left majority			-0.050	-0.007
			(.0168)	(.0091)
Janata majority			0.008	-0.020
			(.0235)	(.0333)
Regional majority			0.006	0.026
			(.0086)	(.0234)
State-specific trends	NO	NO	NO	YES
Adjusted R-squared	0.93	0.93	0.94	0.95

Notes: Adapted from Besley and Burgess (2004), Table IV. The table reports regression-DD estimates of the effects of labor regulation on productivity. The dependent variable is log manufacturing output per capita. All models include state and year effects. Robust standard errors clustered at the state level are reported in parentheses. State amendments to the Industrial Disputes Act are coded 1=pro-worker, 0 = neutral, -1 = pro-employer and then cumulated over the period to generate the labor regulation measure. Log of installed electrical capacity is measured in kilowatts, and log development expenditure is real per capita state spending on social and economic services. Congress, hard left, Janata, and regional majority are counts of the number of years for which these political groupings held a majority of the seats in the state legislatures. The data are for the sixteen main states for the period 1958-1992. There are 552 observations.

wage study. The addition of controls affects the Besley and Burgess estimates little. But the addition of state-specific trends kills the labor-regulation effect, as can be seen in column 4. Apparently, labor regulation in India increases in states where output is declining anyway. Control for this trend therefore drives the estimated regulation effect to zero.

Picking Controls

We've labeled the two dimensions in the DD set-up "states" and "time" because this is the archetypical DD example in applied econometrics. But the DD idea is much more general. Instead of states, the subscript s might denote demographic groups, some of which are affected by a policy and others are not. For example, Kugler, Jimeno, and Hernanz (2005) look at the effects of age-specific employment protection

policies in Spain. Likewise, instead of time, we might group data by cohort or other types of characteristics. An example is Angrist and Evans (1999), who study the effect of changes in state abortion laws on teen pregnancy using variation by state and year of birth. Implicitly, however, DD designs always set up an implicit treatment-control comparison. The question of whether this comparison is a good one deserves careful consideration.

One potential pitfall in this context arises when the composition of the implicit treatment and control groups changes as a result of treatment. Going back to a design based on state/time comparisons, suppose we're interested in the effects of the generosity of public assistance on labor supply. Historically, U.S. states have offered widely-varying welfare payments to poor unmarried mothers. Labor economists have long been interested in the effects of such income maintenance policies - how much of an increase in living standards they facilitate, and whether they make work less attractive (see, e.g., Meyer and Rosenbaum, 2001, for a recent study). A concern here, emphasized in a review of research on welfare by Moffitt (1992), is that poor people who would in any case have weak labor force attachment might move to states with more generous welfare benefits. In a DD research design, this sort of program-induced migration tends to make generous welfare programs look worse for labor supply than they really are.

Migration problems can usually be fixed if we know where an individual starts out. Say we know state of residence in the period before treatment, or state of birth. State of birth or previous state of residence are unchanged by the treatment but still highly correlated with current state of residence. The problem of migration is therefore eliminated in comparisons using these dimensions instead of state of residence. This introduces a new problem, however, which is that individuals who do move are incorrectly located. In practice, however, this problem is easily addressed with the IV methods discussed in chapter 4 (state of birth or previous residence is used to construct instruments for current location).

A modification of the two-by-two DD set-up uses higher-order contrasts to draw causal inferences. An example is the extension of Medicaid coverage in the U.S. studied by Yelowitz (1995). Eligibility for Medicaid, the massive U.S. health insurance program for the poor, was once tied to eligibility for AFDC, a large cash welfare program. At various times in the 1980s, however, some states extended Medicaid coverage to children in families ineligible for AFDC. Yelowitz was interested in how this expansion affected, among other things, mothers' labor force participation and earnings.

In addition to state and time, children's age provides a third dimension in which Medicaid policy varies. Yelowitz exploits this variation by estimating

$$\mathrm{Y}_{iast} = \gamma_{st} + \lambda_{at} + \theta_{as} + \beta \mathrm{D}_{ast} + \mathrm{X}_{iast}\delta + \varepsilon_{iast},$$

where s index states, t indexes time, and a is the age of the youngest child in a family. This model provides full non-parametric control for state-specific time effects that are common across age groups (γ_{st}), time-varying

age effects (λ_{at}), and state-specific age effects (θ_{as}). The regressor of interest, D_{ast}, indicates children in affected age groups in states and periods where coverage is provided. This triple-differences model may generate a more convincing set of results than a traditional DD analysis that exploits differences by state and time alone.

5.3 Fixed Effects versus Lagged Dependent Variables

Fixed effects and differences-in-differences estimators are based on the presumption of time-invariant (or group-invariant) omitted variables. Suppose, for example, we are interested in the effects of participation in a subsidized training program, as in the Dehejia and Wahba (1999) and Lalonde (1986) studies discussed in section (3.3.3). The key identifying assumption motivating fixed effects estimation in this case is

$$E(\mathrm{Y}_{0it}|\alpha_i, \mathrm{X}_{it}, \mathrm{D}_{it}) = E(\mathrm{Y}_{0it}|\alpha_i, \mathrm{X}_{it}), \tag{5.3.1}$$

where α_i is an unobserved personal characteristic that determines, along with covariates, X_{it}, whether individual i gets training. To be concrete, α_i might be a measure of vocational skills, though a strike against the fixed-effects setup is the fact that the exact nature of the unobserved variables typically remains somewhat mysterious. In any case, coupled with a linear model for $E(\mathrm{Y}_{0it}|\alpha_i,\mathrm{X}_{it})$, assumption (5.3.1) leads to simple estimation strategies involving differences or deviations from means.

For many causal questions, the notion that the most important omitted variables are time-invariant doesn't seem plausible. The evaluation of training programs is a case in point. It seems likely that people looking to improve their labor market options by participating in a government-sponsored training program have suffered some kind of setback. Many training programs explicitly target people who have suffered a recent setback, e.g., men who recently lost their jobs. Consistent with this, Ashenfelter (1978) and Ashenfelter and Card (1985) find that training participants typically have earnings histories that exhibit a pre-program dip. Past earnings is a time-varying confounder that cannot be subsumed in a time-invariant variable like α_i.

The distinctive earnings histories of trainees motivates an estimation strategy that controls for past earnings directly and dispenses with the fixed effects. To be precise, instead of (5.3.1), we might base causal inference on the conditional independence assumption,

$$E(\mathrm{Y}_{0it}|\mathrm{Y}_{it-h}, \mathrm{X}_{it}, \mathrm{D}_{it}) = E(\mathrm{Y}_{0it}|\mathrm{Y}_{it-h}, \mathrm{X}_{it}). \tag{5.3.2}$$

This is like saying that what makes trainees special is their earnings h periods ago. We can then use panel data to estimate

$$\mathrm{Y}_{it} = \alpha + \theta \mathrm{Y}_{it-h} + \lambda_t + \beta \mathrm{D}_{it} + \mathrm{X}_{it}\delta + \varepsilon_{it}, \tag{5.3.3}$$

where the causal effect of training is β. To make this more general, Y_{it-h} can be a vector including lagged earnings for multiple periods.[9]

Applied researchers using panel data are often faced with the challenge of choosing between fixed-effects and lagged-dependent variables models, i.e., between causal inferences based on (5.3.1) and (5.3.2). One solution to this dilemma is to work with a model that includes both lagged dependent variables and unobserved individual effects. In other words, identification might be based on a weaker conditional independence assumption:

$$E(\mathrm{Y}_{0it}|a_i, \mathrm{Y}_{it-h}, \mathrm{X}_{it}, \mathrm{D}_{it}) = E(\mathrm{Y}_{0it}|\alpha_i, \mathrm{Y}_{it-h}, \mathrm{X}_{it}), \tag{5.3.4}$$

which requires conditioning on both α_i and Y_{it-h}. We can then try to estimate causal effects using a specification like

$$\mathrm{Y}_{it} = \alpha_i + \theta \mathrm{Y}_{it-h} + \lambda_t + \beta \mathrm{D}_{it} + \mathrm{X}_{it}\delta + \upsilon_{it}. \tag{5.3.5}$$

Unfortunately, the conditions for consistent estimation of β in equation (5.3.5) are much more demanding than those required with fixed effects or lagged dependent variables alone. This can be seen in a simple example where the lagged dependent variable is Y_{it-1}. We kill the fixed effect by differencing, which produces

$$\Delta \mathrm{Y}_{it} = \theta \Delta \mathrm{Y}_{it-1} + \Delta\lambda_t + \beta \Delta \mathrm{D}_{it} + \Delta \mathrm{X}_{it}\delta + \Delta\upsilon_{it}. \tag{5.3.6}$$

The problem here is that the differenced residual, $\Delta\upsilon_{it}$, is necessarily correlated with the lagged dependent variable, $\Delta \mathrm{Y}_{it-1}$, because both are a function of υ_{it-1}. Consequently, OLS estimates of (5.3.6) are not consistent for the parameters in (5.3.5), a problem first noted by Nickell (1981). This problem can be solved, though the solution requires strong assumptions. The easiest solution is to use Y_{it-2} as an instrument for $\Delta \mathrm{Y}_{it-1}$ in (5.3.6).[10] But this requires that Y_{it-2} be uncorrelated with the differenced residuals, $\Delta\upsilon_{it}$. This seems unlikely since residuals are the part of earnings left over after accounting for covariates. Most people's earnings are highly correlated from one year to the next, so that past earnings are an excellent predictor of future earnings and earnings growth . If υ_{it} is serially correlated, there may be no consistent estimator for (5.3.6). (Note also that the IV strategy using Y_{it-2} as an instrument requires at least three periods to obtain data for t, $t-1$, and $t-2$).

Given the difficulties that arise when trying to estimate (5.3.6), we might ask whether the distinction between fixed effects and lagged dependent variables matters. The answer, unfortunately, is yes. The fixed-effects and lagged dependent variables models are not nested, which means we cannot hope to estimate

[9] Abadie, Diamond, and Hainmueller (2007) develop a semiparametric version of the lagged-dependent variables model, more flexible than the traditional regression setup. As with our regression setup, the key assumption in this model is conditional independence of potential outcomes conditional on lagged earnings, i.e., assumption (5.3.2).

[10] See Holtz-Eakin, Newey and Rosen (1988), Arellano and Bond (1991), Blundell and Bond (1998) for details and examples.

one and get the other as a special case if need be. Only the more general and harder-to-identify model, (5.3.5), nests both fixed effects and lagged dependent variables.[11].

So what's an applied guy to do? One answer, as always, is to check the robustness of your findings using alternative identifying assumptions. That means that you would like to find broadly similar results using both models. Fixed effects and lagged dependent variables estimates also have a useful bracketing property. The appendix to this chapter shows that if (5.3.2) is correct, but you mistakenly use fixed effects, estimates of a positive treatment effect will tend to be too big. On the other hand, if (5.3.1) is correct and you mistakenly estimate an equation with lagged outcomes like (5.3.3), estimates of a positive treatment effect will tend to be too small. You can therefore think of fixed effects and lagged dependent variables as bounding the causal effect you are after. Guryan (2004) illustrates this sort of reasoning in a study estimating the effects of court-ordered busing on Black high school graduation rates.

5.4 Appendix: More on fixed effects and lagged dependent variables

To simplify, we ignore covariates and year effects and assume there are only two periods, with treatment equal to zero for everyone in the first period (the punch line is the same in a more general setup). The causal effect of interest, β, is positive. Suppose first that treatment is correlated with an unobserved individual effect, a_i, and that outcomes can be described by

$$\mathrm{Y}_{it} = a_i + \beta \mathrm{D}_{it} + \varepsilon_{it}. \tag{5.4.1}$$

where ε_{it} is serially uncorrelated, and uncorrelated with a_i and D_{it}. We also have

$$\mathrm{Y}_{it-1} = a_i + \varepsilon_{it-1},$$

where a_i and ε_{it-1} are uncorrelated. You mistakenly estimate the effect of D_{it} in a model that controls for Y_{it-1} but ignores fixed effects. The resulting estimator has probability limit $\frac{Cov(\mathrm{Y}_{it},\tilde{\mathrm{D}}_{it})}{V(\mathrm{D}_{it})}$, where $\tilde{\mathrm{D}}_{it} = \mathrm{D}_{it} - \gamma \mathrm{Y}_{it-1}$ is the residual from a regression of D_{it} on Y_{it-1}.

[11] In particular, setting $\theta = 1$ in (5.3.3) does not produce the fixed-effects model as a special case of the lagged dependent variables model. Instead we get

$$\Delta \mathrm{Y}_{it} = \alpha + \lambda_t + \beta \mathrm{D}_{it} + \mathrm{X}_{it}\delta + \varepsilon_{it}$$

i.e., a differenced dependent variable with regressors in levels. This is not the model with first differences on both the right and left side needed to kill the fixed effect.

Now substitute $a_i = \mathrm{Y}_{it-1} - \varepsilon_{it-1}$ in (5.4.1) to get

$$\mathrm{Y}_{it} = \mathrm{Y}_{it-1} + \beta \mathrm{D}_{it} + \varepsilon_{it} - \varepsilon_{it-1}.$$

From here, we get

$$\frac{Cov(\mathrm{Y}_{it}, \tilde{\mathrm{D}}_{it})}{V(\tilde{\mathrm{D}}_{it})} = \beta - \frac{Cov(\varepsilon_{it-1}, \tilde{\mathrm{D}}_{it})}{V(\tilde{\mathrm{D}}_{it})} = \beta - \frac{Cov(\varepsilon_{it-1}, \mathrm{D}_{it} - \gamma \mathrm{Y}_{it-1})}{V(\tilde{\mathrm{D}}_{it})} = \beta + \frac{\gamma \sigma_{\varepsilon}^2}{V(\tilde{\mathrm{D}}_{it})}.$$

where σ_{ε}^2 is the variance of ε_{it-1}. Since trainees have low Y_{it-1}, $\gamma < 0$ and the resulting estimate of β is too small.

Suppose instead that treatment is determined by low Y_{it-1}. The correct specification is a simplified version of (5.3.3), say

$$\mathrm{Y}_{it} = \alpha + \theta \mathrm{Y}_{it-1} + \beta \mathrm{D}_{it} + \varepsilon_{it}, \tag{5.4.2}$$

where ε_{it} is serially uncorrelated. You mistakenly estimate a first-differenced equation in an effort to kill fixed effects. This ignores lagged dependent variables. In this simple example, where $\mathrm{D}_{it-1} = 0$ for everyone, the first-differenced estimator has probability limit

$$\frac{Cov(\mathrm{Y}_{it} - \mathrm{Y}_{it-1}, \mathrm{D}_{it} - \mathrm{D}_{it-1})}{V(\mathrm{D}_{it} - \mathrm{D}_{it-1})} = \frac{Cov(\mathrm{Y}_{it} - \mathrm{Y}_{it-1}, \mathrm{D}_{it})}{V(\mathrm{D}_{it})}. \tag{5.4.3}$$

Subtracting Y_{it-1} from both sides of (5.4.2), we have

$$\mathrm{Y}_{it} - \mathrm{Y}_{it-1} = \alpha + (\theta - 1)\mathrm{Y}_{it-1} + \beta \mathrm{D}_{it} + \varepsilon_{it}.$$

Substituting this in (4.2.2), the inappropriately differenced model yields

$$\frac{Cov(\mathrm{Y}_{it} - \mathrm{Y}_{it-1}, \mathrm{D}_{it})}{V(\mathrm{D}_{it})} = \beta + (\theta - 1)\left[\frac{Cov(\mathrm{Y}_{it-1}, \mathrm{D}_{it})}{V(\mathrm{D}_{it})}\right].$$

In general, we think θ is a number between zero and one, otherwise Y_{it} is non-stationary (i.e., an explosive time series process). Therefore, since trainees have low $\mathrm{Y}_{it-1,}$ the estimate of β in first differences is too big.