



# **MOSTLY HARMLESS ECONOMETRICS**

**An Empiricist's Companion**

**Joshua D. Angrist and Jörn-Steffen Pischke**

## 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*

**T**he 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 instrumental variables, 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

$$E[Y_{0it}|A_i, X_{it}, t, D_{it}] = E[Y_{0it}|A_i, X_{it}, t],$$

where  $X_{it}$  is a vector of observed time-varying covariates and  $A_i$  is a vector of unobserved but fixed confounders that we'll call ability.

In other words, union status is as good as randomly assigned conditional on  $A_i$  and observed covariates, such as age, schooling, and region of residence.

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

$$E[Y_{0it}|A_i, X_{it}, t] = \alpha + \lambda_t + A_i' \gamma + X_{it}' \beta, \quad (5.1.1)$$

We also assume that the causal effect of union membership is additive and constant:

$$E[Y_{1it}|A_i, X_{it}, t] = E[Y_{0it}|A_i, X_{it}, t] + \rho.$$

Together with (5.1.1), this implies

$$E[Y_{it}|A_i, X_{it}, t, D_{it}] = \alpha + \lambda_t + \rho D_{it} + A_i' \gamma + X_{it}' \beta, \quad (5.1.2)$$

where  $\rho$  is the causal effect of interest. The set of assumptions leading to (5.1.2) is more restrictive than 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.<sup>1</sup>

<sup>1</sup>In some cases, we can allow heterogeneous treatment effects so that

$$E(Y_{1it} - Y_{0it}|A_i, X_{it}, t) = \rho_i.$$

See, for example, Wooldridge (2005), who discusses estimators for the average of  $\rho_i$ .

Equation (5.1.2) implies

$$Y_{it} = \alpha_i + \lambda_t + \rho D_{it} + X_{it}' \beta + \varepsilon_{it}, \quad (5.1.3)$$

where  $\varepsilon_{it} \equiv Y_{0it} - E[Y_{0it}|A_i, X_{it}, t]$  and

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

This is a fixed effects model. Given panel data (repeated observations on individuals), the causal effect of union status on wages can be estimated by treating  $\alpha_i$ , the fixed effect, as a parameter to be estimated. The year effect,  $\lambda_t$ , 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.<sup>2</sup>

It might seem that there are a lot of parameters to be estimated in the fixed effects model. For example, the Panel Survey of Income Dynamics, a widely used panel data set, includes data on 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,

$$\bar{Y}_i = \alpha_i + \bar{\lambda} + \rho \bar{D}_i + \bar{X}_i' \beta + \bar{\varepsilon}_i.$$

Subtracting this from (5.1.3) gives

$$Y_{it} - \bar{Y}_i = \lambda_t - \bar{\lambda} + \rho(D_{it} - \bar{D}_i) + (X_{it} - \bar{X}_i)' \beta + (\varepsilon_{it} - \bar{\varepsilon}_i), \quad (5.1.4)$$

<sup>2</sup>An alternative to the fixed effects specification is random effects (see, e.g., Wooldridge, 2006). The random effects model assumes that  $\alpha_i$  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 OLS standard errors. Random effects models can be estimated by GLS, which promises to be more efficient if the assumptions of the random effects model are satisfied. However, as in chapter 3, we prefer fixing OLS standard errors to GLS. GLS requires stronger assumptions than OLS, and the resulting asymptotic efficiency gain is likely to be modest, while finite-sample properties may be worse.

so deviations from means kills the unobserved individual effects.<sup>3</sup>

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

$$\Delta Y_{it} = \Delta \lambda_t + \rho \Delta D_{it} + \Delta X'_{it} \beta + \Delta \varepsilon_{it}, \quad (5.1.5)$$

where the  $\Delta$  prefix denotes the change from one year to the next. For example,  $\Delta Y_{it} = Y_{it} - Y_{it-1}$ . With two periods, differencing is algebraically the same as deviations from means, but not otherwise. Both should work, although with homoskedastic and serially uncorrelated  $\varepsilon_{it}$  and more than two periods, 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 freedom 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.<sup>4</sup>

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,

<sup>3</sup>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.

<sup>4</sup>The fixed effects are not estimated consistently in a panel where the number of periods  $T$  is fixed while  $N \rightarrow \infty$ . This is called the incidental parameters problem, a name that 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.

TABLE 5.1.1  
Estimated effects of union status on wages

Survey	Cross Section Estimate	Fixed Effects Estimate
May CPS, 1974–75	.19	.09
National Longitudinal Survey of Young Men, 1970–78	.28	.19
Michigan PSID, 1970–79	.23	.14
QES, 1973–77	.14	.16

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

the table displays results from a fixed effects estimator and the corresponding cross section estimates. The cross section estimates are typically higher (ranging from .14 to .28) than the fixed effects estimates (ranging from .09 to .19). 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.

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 such as 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 differenced regressors in an equation like (5.1.4) or (5.1.5) than in the levels of the regressors. This fact may account for smaller fixed effects estimates.<sup>5</sup>

<sup>5</sup>See Griliches and Hausman (1986) for a more complete discussion of measurement error in panel data.

A variant on the measurement error problem in panel data 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 schooling. Surprisingly, the within-family estimates come out larger than OLS estimates. 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, neither is the difference in their schooling. Hence, small unobserved ability differences between 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 solution to measurement error is to use IV methods. Ashenfelter and Krueger (1994) use cross-sibling reports to construct instruments for schooling differences between 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 Krueger (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 or group level, such as state or cohort. For example, state policies regarding health care benefits for pregnant workers may change over time but are fixed across workers within states. The source of OVB when evaluating these policies must therefore be unobserved variables at the state and year level. In some cases, group-level omitted variables can be captured by group-level fixed effects, an approach that leads to the differences-in-differences (DD) identification strategy.

The DD idea was probably pioneered by physician John Snow (1855), who studied cholera epidemics in London in the mid-nineteenth century. Snow wanted to establish that cholera is transmitted by contaminated drinking water (as opposed to "bad air," the prevailing theory at the time). To show this, Snow compared changes in death rates from cholera in districts serviced by two water companies, the Southwark and Vauxhall Company and the Lambeth Company. In 1849 both companies obtained their water supply from the dirty Thames in central London. In 1852, however, the Lambeth Company moved its water works upriver to an area relatively free of sewage. Death rates in districts supplied by Lambeth fell sharply in comparison to the change in death rates in districts supplied by Southwark and Vauxhall.

To make matters more concrete, let us return to an example from economics. 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 labor 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.<sup>6</sup>

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 also 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 February-to-November change in employment in New Jersey to the change in employment in Pennsylvania over the same period.

DD is a version of fixed effects estimation using aggregate data. To see this, let  $Y_{1ist}$  be fast food employment at restaurant  $i$  in state  $s$  and period  $t$  if there is a high state minimum wage, and let  $Y_{0ist}$  be fast food employment at restaurant  $i$  in state  $s$  and period  $t$  if there is a low state minimum wage. These are potential outcomes; in practice, we only get to see one or the other. For example, we see  $Y_{1ist}$  in New Jersey in November 1992. The heart of the DD setup is an additive structure for potential outcomes in the no-treatment state. Specifically, we assume that

$$E[Y_{0ist}|s, t] = \gamma_s + \lambda_t, \quad (5.2.1)$$

<sup>6</sup>The DD idea was first used to study the effects of minimum wages by Obenauer and von der Nienburg (1915), writing for the U.S. Bureau of Labor Statistics.

where  $s$  denotes state (New Jersey or Pennsylvania) and  $t$  denotes period (February, before the minimum wage increase, or November, after the increase). This equation 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 section 5.1.

Let  $D_{st}$  be a dummy for high-minimum-wage states and periods. Assuming that  $E[Y_{1ist} - Y_{0ist}|s, t]$  is a constant, denoted  $\delta$ , observed employment,  $Y_{ist}$ , can be written:

$$Y_{ist} = \gamma_s + \lambda_t + \delta D_{st} + \varepsilon_{ist}, \quad (5.2.2)$$

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

$$\begin{aligned} E[Y_{ist}|s = PA, t = Nov] - E[Y_{ist}|s = PA, t = Feb] \\ = \lambda_{Nov} - \lambda_{Feb} \end{aligned}$$

and

$$\begin{aligned} E[Y_{ist}|s = NJ, t = Nov] - E[Y_{ist}|s = NJ, t = Feb] \\ = \lambda_{Nov} - \lambda_{Feb} + \delta. \end{aligned}$$

The population difference-in-differences,

$$\begin{aligned} \{E[Y_{ist}|s = NJ, t = Nov] - E[Y_{ist}|s = NJ, t = Feb]\} \\ - \{E[Y_{ist}|s = PA, t = Nov] - E[Y_{ist}|s = PA, t = Feb]\} = \delta, \end{aligned}$$

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

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

TABLE 5.2.1  
Average employment in fast food restaurants before and after the  
New Jersey minimum wage increase

Variable	PA (i)	NJ (ii)	Difference, NJ – PA (iii)
1. FTE employment before, all available observations	23.33 (1.35)	20.44 (.51)	-2.89 (1.44)
2. FTE employment after, all available observations	21.17 (.94)	21.03 (.52)	-.14 (1.07)
3. Change in mean FTE employment	-2.16 (1.25)	.59 (.54)	2.76 (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 restaurants with data on employment. Employment at six closed restaurants is set to zero. Employment at four temporarily closed restaurants is treated as missing. Standard errors are reported in parentheses.

a positive difference-in-differences, the opposite of what we might expect if a higher minimum wage pushed 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 is meant to be captured by the state fixed effect, which plays the same role as the unobserved individual effect in (5.1.3).<sup>7</sup>

<sup>7</sup>The common trends assumption can be applied to transformed data, for example,

$$E[\ln y_{0ist} | s, t] = \gamma_s + \lambda_t.$$

Note, however, that common trends in logs rule out common trends in levels and vice versa. Athey and Imbens (2006) introduce a semiparametric DD estimator that allows for common trends after an unspecified transformation of the dependent variable. Poterba, Venti, and Wise (1995) and Meyer, Viscusi, and Durbin (1995) discuss DD-type models for quantiles.

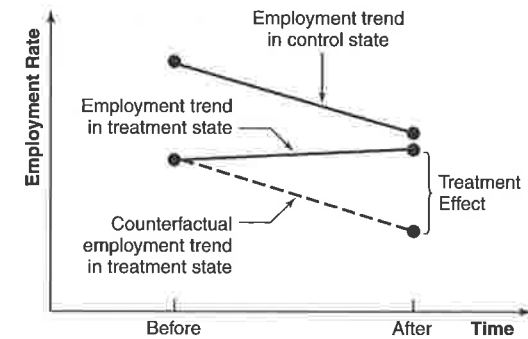


Figure 5.2.1 Causal effects in the DD model.

The common trends assumption can be investigated using data on multiple periods. In an update of their original minimum wage study, Card and Krueger (2000) obtained administrative payroll data for restaurants in New Jersey and a number of Pennsylvania counties. 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 the original Card and Krueger surveys were conducted, and the third vertical line indicates the October 1996 increase in the federal minimum wage to \$4.75, which affected Pennsylvania but not New Jersey. These data give us an opportunity to look at a new minimum wage experiment.

As in 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 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 increase in the federal minimum wage. So Pennsylvania may not provide a very good measure of counterfactual employment rates in New Jersey in the absence of a minimum wage change.

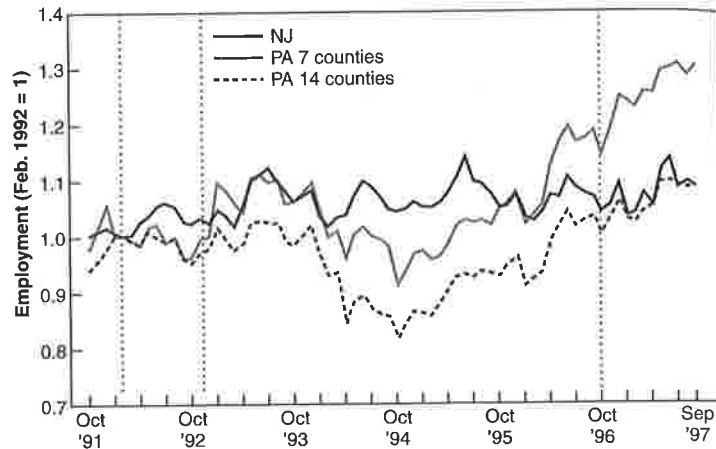


Figure 5.2.2 Employment in New Jersey and Pennsylvania fast food restaurants, October 1991 to September 1997 (from Card and Krueger 2000). Vertical lines indicate dates of the original Card and Krueger (1994) survey and the October 1996 federal minimum wage increase.

A more encouraging example comes from Pischke (2007), who looked 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 already had a fall start.

Figure 5.2.3 plots the likelihood of grade repetition for the 1962–73 cohorts of second graders in Bavaria and affected states (there are no repetition data for 1963–65). Repetition rates in Bavaria were reasonably flat from 1966 on at around 2.5 percent. Repetition rates are higher in the short-school-year (SSY) states, at around 4–4.5 percent 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

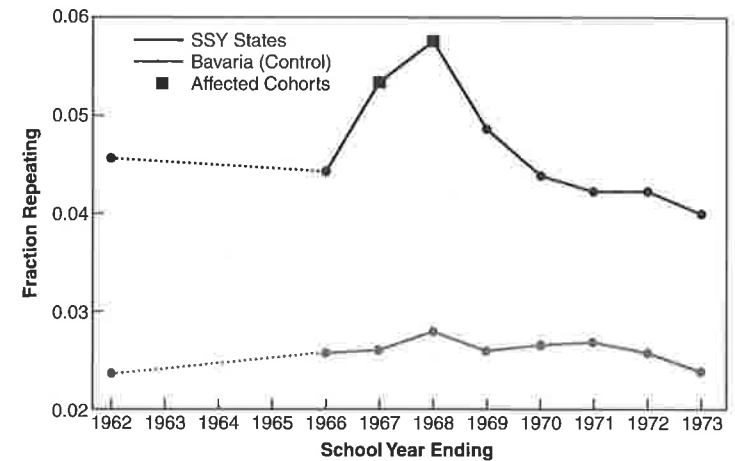


Figure 5.2.3 Average grade repetition rates in second grade for treatment and control schools in Germany (from Pischke, 2007). The data span a period before and after a change in term length for students outside Bavaria (SSY states).

in these states, a bit more so for the second cohort than for 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.

### 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

$$y_{ist} = \alpha + \gamma NJ_s + \lambda d_t + \delta(NJ_s \cdot d_t) + \varepsilon_{ist} \quad (5.2.3)$$

is the same as (5.2.2) where  $NJ_s \cdot d_t = D_{st}$ . 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  $E(y_{ist}|s, t)$  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

$$\begin{aligned}\alpha &= E[y_{ist}|s = PA, t = Feb] = \gamma_{PA} + \lambda_{Feb} \\ \gamma &= E[y_{ist}|s = NJ, t = Feb] - E[y_{ist}|s = PA, t = Feb] \\ &= \gamma_{NJ} - \gamma_{PA} \\ \lambda &= E[y_{ist}|s = PA, t = Nov] - E[y_{ist}|s = PA, t = Feb] \\ &= \lambda_{Nov} - \lambda_{Feb} \\ \delta &= \{E[y_{ist}|s = NJ, t = Nov] - E[y_{ist}|s = NJ, t = Feb]\} \\ &\quad - \{E[y_{ist}|s = PA, t = Nov] - E[y_{ist}|s = PA, t = Feb]\}.\end{aligned}$$

The regression formulation of the DD 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 setup. We might, for example, add additional control states and pretreatment 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 the study of policies other than those that can be described by 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 wage of \$4.25 an hour 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

$$y_{ist} = \gamma_s + \lambda_t + \delta(\text{FA}_s \cdot d_t) + \varepsilon_{ist}, \quad (5.2.4)$$

where the variable  $\text{FA}_s$  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 in 1990, when the federal minimum wage increased from \$3.35 to \$3.80. The  $\text{FA}_s$  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 1990. 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 microdata (provided the group-level estimates are weighted by cell size). Note that  $\text{FA}_s \cdot d_t$  is an interaction term, like  $\text{NJ}_s \cdot d_t$  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:

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

where  $\Delta \bar{y}_s$  is the change in average teen employment in state  $s$  and  $\Delta \bar{\varepsilon}_s$  is the error term in the differenced equation.<sup>8</sup>

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  $\text{FA}_s$  (fraction of affected teens) variable is a good predictor of the wage changes induced by an increase in the federal minimum. Employment, on the other

<sup>8</sup>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 wage rates.

TABLE 5.2.2  
Regression DD estimates of minimum wage effects on teens,  
1989 to 1990

Explanatory Variable	Change in Mean Log Wage		Change in Teen Employment-Population Ratio	
	(1)	(2)	(3)	(4)
1. Fraction of affected teens ( $FA_s$ )	.15 (.03)	.14 (.04)	.02 (.03)	-.01 (.03)
2. Change in overall emp./pop. ratio	—	.46 (.60)	—	1.24 (.60)
3. $R^2$	.30	.31	.01	.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 1990 CPS. Regressions are weighted by the CPS sample size for each state.

hand, seems largely unrelated to  $FA_s$ , 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.

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

$$E[y_{0ist} | s, t, X_{st}] = \gamma_s + \lambda_t + X'_{st} \beta.$$

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 that Card (1992) analyzes state averages instead of individual data. He might have used a pooled multiyear sample of microdata from the CPS to estimate an

equation like

$$y_{ist} = \gamma_s + \lambda_t + \delta(FA_s \cdot d_t) + X'_{ist} \beta + \varepsilon_{ist}, \quad (5.2.5)$$

where  $X_{ist}$  can include individual level characteristics such as race as well as 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 microdata on dependent variables with group-level regressors, however. The key issue is how best to adjust standard errors for possible group-level random effects, as we discuss in chapter 8.

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 epigraph 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 dummies for future policy changes should not matter in an equation like

$$y_{ist} = \gamma_s + \lambda_t + \sum_{\tau=0}^m \delta_{-\tau} D_{s,t-\tau} + \sum_{\tau=1}^q \delta_{+\tau} D_{s,t+\tau} + X'_{ist} \beta + \varepsilon_{ist}, \quad (5.2.6)$$

where the sums on the right-hand side allow for  $m$  lags ( $\delta_{-1}, \delta_{-2}, \dots, \delta_{-m}$ ) or posttreatment effects and  $q$  leads ( $\delta_{+1}, \delta_{+2}, \dots, \delta_{+q}$ ) 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. In the U.S., 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 are employed by someone else besides the firm for which they are executing tasks. As a result, firms 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 four 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 list of controls. In other words, we estimate

$$Y_{ist} = \gamma_{0s} + \gamma_{1s}t + \lambda_t + \delta D_{st} + X'_{ist}\beta + \varepsilon_{ist}, \quad (5.2.7)$$

where  $\gamma_{0s}$  is a state-specific intercept, as before, and  $\gamma_{1s}$  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

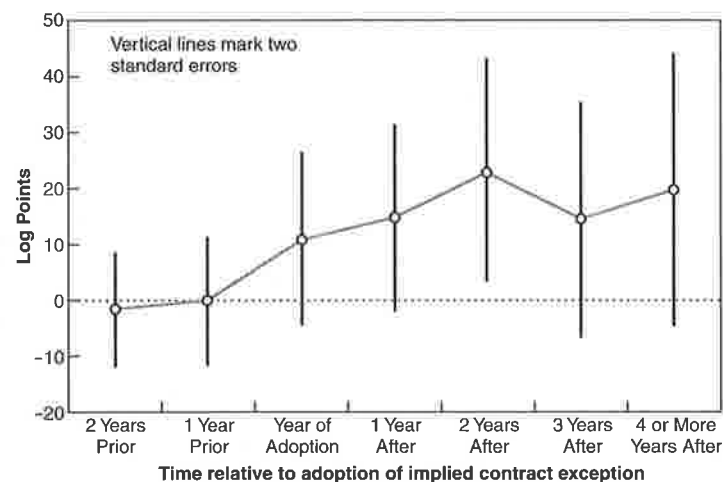


Figure 5.2.4 The estimated impact of implied-contract exceptions to the employment-at-will doctrine on the use of temporary workers (from Autor, 2003). The dependent variable is the log of state temporary help employment in 1979–1995. Estimates are from a model that allows for effects before, during, and after exceptions were adopted.

otherwise. Note, however, that we need at least three periods to estimate a model with state-specific trends. Moreover, in practice, three 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 pretreatment data establish a clear trend that can be extrapolated into the posttreatment period.

In a study of the effects 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

TABLE 5.2.3  
Estimated effects of labor regulation on the performance of firms  
in Indian states

	(1)	(2)	(3)	(4)
Labor regulation (lagged)	-.186 (.064)	-.185 (.051)	-.104 (.039)	.0002 (.020)
Log development expenditure per capita		.240 (.128)	.184 (.119)	.241 (.106)
Log installed electricity capacity per capita		.089 (.061)	.082 (.054)	.023 (.033)
Log state population		.720 (.96)	0.310 (1.192)	-1.419 (2.326)
Congress majority			-.0009 (.01)	.020 (.010)
Hard left majority			-.050 (.017)	-.007 (.009)
Janata majority			.008 (.026)	-.020 (.033)
Regional majority			.006 (.009)	.026 (.023)
State-specific trends	No	No	No	Yes
Adjusted $R^2$	.93	.93	.94	.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 1958-92. There are 552 observations.

the estimates in columns 2 and 3 add time-varying state-specific covariates, such as 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 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 increased in states where output was 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 setup *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 studied the effect of changes in state abortion laws on teen pregnancy using variation by state and year of birth. Regardless of the group labels, 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 treatment and control groups changes as a result of treatment. Going back to a design based on state and 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 are 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 is easily addressed with the IV methods discussed in chapter 4 (state of birth or previous residence can be used to construct instruments for current location).

A modification of the two-by-two DD setup with possibly improved control groups uses higher-order contrasts to draw causal inferences. An example is the extension of Medicaid coverage in the United States, studied by Yelowitz (1995). Eligibility for Medicaid, the massive U.S. health insurance program for the poor, was once tied to eligibility for Aid for Families with Dependent Children (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 of publicly provided health insurance for children affected, among other things, mothers' labor force participation and earnings.

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

$$y_{iast} = \gamma_{st} + \lambda_{at} + \theta_{as} + \delta D_{ast} + X'_{iast} \beta + \varepsilon_{iast},$$

where  $s$  indexes states,  $t$  indexes time, and  $a$  is the age of the youngest child in a family. This model provides full nonparametric control for state-specific time effects that are

common across age groups ( $\gamma_{st}$ ), time-varying age effects ( $\lambda_{at}$ ), and state-specific age effects ( $\theta_{as}$ ). The regressor of interest,  $D_{ast}$ , indicates families with children in affected age groups in states and periods where Medicaid 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 DD 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

$$E[y_{0it} | \alpha_i, X_{it}, D_{it}] = E[y_{0it} | \alpha_i, X_{it}], \quad (5.3.1)$$

where  $\alpha_i$  is an unobserved personal characteristic that determines, along with covariates,  $X_{it}$ , whether individual  $i$  gets training. To be concrete,  $\alpha_i$  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  $E(y_{0it} | \alpha_i, X_{it})$ , 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's 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, such as 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 preprogram dip. Past earnings is a time-varying confounding variable that cannot be subsumed in a time-invariant omitted variable like  $\alpha_i$ .

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

$$E[Y_{0it}|Y_{it-b}, X_{it}, D_{it}] = E[Y_{0it}|Y_{it-b}, X_{it}]. \quad (5.3.2)$$

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

$$Y_{it} = \alpha + \theta Y_{it-b} + \lambda_t + \delta D_{it} + X'_{it}\beta + \varepsilon_{it}, \quad (5.3.3)$$

where the causal effect of training is  $\delta$ . To make this more general,  $Y_{it-b}$  can be a vector including lagged earnings for multiple periods.<sup>9</sup>

Applied researchers using panel data are often faced with the challenge of choosing between fixed effects and lagged dependent variables models, that is, 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

$$E[Y_{0it}|\alpha_i, Y_{it-b}, X_{it}, D_{it}] = E[Y_{0it}|\alpha_i, Y_{it-b}, X_{it}], \quad (5.3.4)$$

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

$$Y_{it} = \alpha_i + \theta Y_{it-b} + \lambda_t + \delta D_{it} + X'_{it}\beta + \varepsilon_{it}. \quad (5.3.5)$$

<sup>9</sup>Abadie, Diamond, and Hainmueller (2007) develop a semiparametric version of the lagged dependent variables model, more flexible than the traditional regression setup. As in 5.3.2, the key assumption in this model is independence of treatment status and potential outcomes conditional on lagged earnings. The Abadie, Diamond, and Hainmueller approach works for microdata and for data with a group structure. The Dehejia and Wahba (1999) matching strategy also uses lagged dependent variables.

Unfortunately, the conditions for consistent estimation of  $\delta$  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

$$\Delta Y_{it} = \theta \Delta Y_{it-1} + \Delta \lambda_t + \delta \Delta D_{it} + \Delta X'_{it}\beta + \Delta \varepsilon_{it}. \quad (5.3.6)$$

The problem here is that the differenced residual,  $\Delta \varepsilon_{it}$ , is necessarily correlated with the lagged dependent variable,  $\Delta Y_{it-1}$ , because both are a function of  $\varepsilon_{it-1}$ . 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  $\Delta Y_{it-1}$  in (5.3.6).<sup>10</sup> But this requires that  $Y_{it-2}$  be uncorrelated with the differenced residuals,  $\Delta \varepsilon_{it}$ . 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 also likely to be correlated with  $\Delta \varepsilon_{it}$ . If  $\varepsilon_{it}$  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, so we get 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 one and get the other as a special case if need be.

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 plausible alternative models. Fixed effects and lagged dependent variables estimates also have a useful bracketing property. The appendix to this

<sup>10</sup>See Holtz-Eakin, Newey, and Rosen (1988), Arellano and Bond (1991), and Blundell and Bond (1998) for details and examples.

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, such as (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 of interest (given some assumptions about the nature of selection bias). Guryan (2004) illustrates this sort of reasoning in a study estimating the effects of court-ordered busing on black students' high school graduation rates.

#### 5.4 Appendix: More on Fixed Effects and Lagged Dependent Variables

To simplify, we ignore covariates, intercepts, 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,  $\delta$ , is positive. Suppose first that treatment (training status) is correlated with an unobserved individual effect,  $\alpha_i$ , uncorrelated with lagged outcome residuals,  $\varepsilon_{it-1}$ , and that outcomes can be described by

$$Y_{it} = \alpha_i + \delta D_{it} + \varepsilon_{it}, \quad (5.4.1)$$

where  $\varepsilon_{it}$  is serially uncorrelated, and also uncorrelated with  $\alpha_i$  and  $D_{it}$ . We also have

$$Y_{it-1} = \alpha_i + \varepsilon_{it-1},$$

where  $\alpha_i$  and  $\varepsilon_{it-1}$  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  $\frac{\text{Cov}(Y_{it}, \tilde{D}_{it})}{V(\tilde{D}_{it})}$ , where  $\tilde{D}_{it} = D_{it} - \gamma Y_{it-1}$  is the residual from a regression of  $D_{it}$  on  $Y_{it-1}$ .

Now substitute  $\alpha_i = Y_{it-1} - \varepsilon_{it-1}$  in (5.4.1) to get

$$Y_{it} = Y_{it-1} + \delta D_{it} + \varepsilon_{it} - \varepsilon_{it-1}.$$

From here, we get

$$\begin{aligned} \frac{\text{Cov}(Y_{it}, \tilde{D}_{it})}{V(\tilde{D}_{it})} &= \delta - \frac{\text{Cov}(\varepsilon_{it-1}, \tilde{D}_{it})}{V(\tilde{D}_{it})} \\ &= \delta - \frac{\text{Cov}(\varepsilon_{it-1}, D_{it} - \gamma Y_{it-1})}{V(\tilde{D}_{it})} = \delta + \frac{\gamma \sigma_\varepsilon^2}{V(\tilde{D}_{it})}, \end{aligned}$$

where  $\sigma_\varepsilon^2$  is the variance of  $\varepsilon_{it-1}$ . Since trainees have low  $Y_{it-1}$ ,  $\gamma < 0$ , and the resulting estimate of  $\delta$  is too small.

Suppose instead that treatment is determined by low  $Y_{it-1}$ . Causal effects can be estimated using a simplified version of (5.3.3), say

$$Y_{it} = \alpha + \theta Y_{it-1} + \delta D_{it} + \varepsilon_{it}, \quad (5.4.2)$$

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

$$\frac{\text{Cov}(Y_{it} - Y_{it-1}, D_{it} - D_{it-1})}{V(D_{it} - D_{it-1})} = \frac{\text{Cov}(Y_{it} - Y_{it-1}, D_{it})}{V(D_{it})}. \quad (5.4.3)$$

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

$$Y_{it} - Y_{it-1} = \alpha + (\theta - 1)Y_{it-1} + \delta D_{it} + \varepsilon_{it}.$$

Substituting this in (5.4.3), the inappropriately differenced model yields

$$\frac{\text{Cov}(Y_{it} - Y_{it-1}, D_{it})}{V(D_{it})} = \delta + (\theta - 1) \left[ \frac{\text{Cov}(Y_{it-1}, D_{it})}{V(D_{it})} \right].$$

In general, we think  $\theta$  is a positive number less than one, otherwise  $Y_{it}$  is nonstationary (i.e., an explosive time series process). Therefore, since trainees have low  $Y_{it-1}$ , the estimate of  $\delta$  in first differences is too big. Note that in this simple model, differencing turns out to be ok in the unlikely event  $\theta = 1$  in (5.4.2), but that is not true in general.