# Empirical Methods in Applied Economics Lecture Notes

Jörn-Steffen Pischke
LSE

October 2005

## 1 Differences-in-differences

### 1.1 Basics

The key strategy in regression was to estimate causal effects by controlling for confounding factors. A key variable in such a strategy is frequently the outcome of interest in a period before the treatment took place. Differences-in-differences is a strategy to model the role of pre-treatment outcomes in a particular fashion.

For example, say you are interested in the effect of the minimum wage on employment. A number of studies have exploited changes in minimum wages at the state level, and we will use the example of Card and Krueger (1994) here, who studied the increase in the minimum wage in New Jersey from 4.25 to 5.05. This change took effect on April 1, 1992. Card and Krueger collected data on employment at fast food restaurants in New Jersey in February and in November 1992. They also collected similar data on restaurants in eastern Pennsylvania, the neighboring state, for the same period. The minimum wage in Pennsylvania remained at 4.25 throughout this period.

Formally, the assumptions underlying differences-in-differences estimation are as follows. Let

$$\begin{aligned} y_1 &= \text{fast food employment for high minimum wage} \\ y_0 &= \text{fast food employment for low minimum wage} \end{aligned}$$

be the counterfactual outcomes. Recall that conditioning means that we are willing to assume that $E(y_0|D,X) = E(y_0|X)$. Here, we are assuming a particular functional form for $E(y_0|X)$, namely

$$E(y_0|X) = E(y_0|s,t) = \gamma_s + \lambda_t$$

where $s$ denotes the state (New Jersey or Pennsylvania) and $t$ denotes the period (February, before the minimum wage increase or November, after the increase). This says that in the absence of a minimum wage change employment is given by state effect, and a time effect, which is the same in both states. The treatment, a higher minimum wage, changes the employment level conditional on $s$ and $t$:

$$E(y_1|s,t) = E(y_0|s,t) + \beta = \gamma_s + \lambda_t + \beta$$

So we can write observed employment in restaurant $i$ as

$$y_i = \gamma_s + \lambda_t + \beta D_{st} + \varepsilon_i \tag{1}$$

where $D_{st}$ is a dummy for the treatment, a high minimum wage, which was in place in New Jersey in November.

Notice that

$$E(y_i|s = PA, t = Nov) - E(y_i|s = PA, t = Feb) = \lambda_{Nov} - \lambda_{Feb}$$

and

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

Hence, the difference-in-difference

$$[E(y_i|s = PA, t = Nov) - E(y_i|s = PA, t = Feb)]$$

$$-[E(y_i|s = NJ, t = Nov) - E(y_i|s = NJ, t = Feb)] = \beta$$

estimates the treatment effect $\beta$.

Table 3 in Card and Krueger (1994), rows 1. to 3. and columns (i) to (iii) display estimates of employment in the four cells (PA versus NJ, before versus after), as well as the state differences, the changes over time, and the difference-in-difference. Employment in PA restaurants is somewhat higher than in NJ in Februrary and falls by November. Employment in NJ, in contrast, increases slightly. This results in a positive estimate for the difference-in-difference. This is the opposite result from what we might expect if restaurants were moving up their labor demand curve as the minimum wage increases.

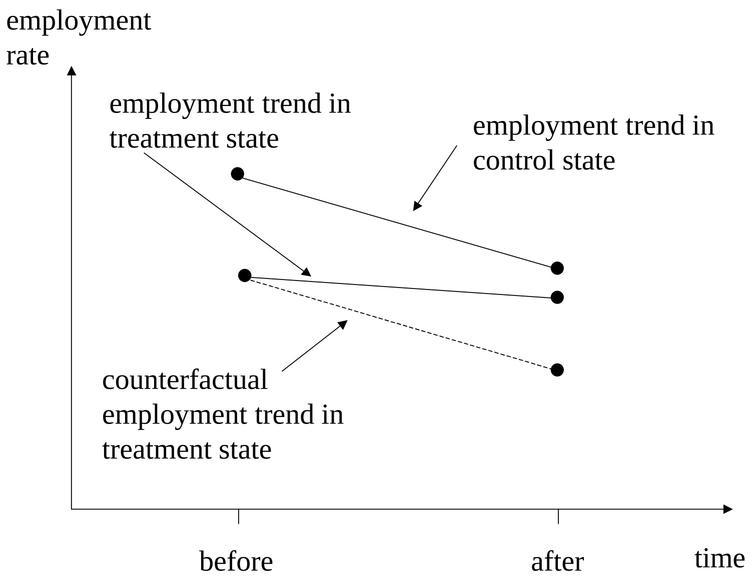


Figure 1: Identification in the difference-in-difference model

What is the key identifying assumption of the difference-in-difference estimator? The assumption is that employment trends would have been the same in both states in the absence of the treatment. Hence, the employment trend in the treatment state has the same slope as in the control state, but is displaced to account for the different employment levels before the treatment, as in figure 1.

Depending on the context, there may be various forms of this identifying assumption, which are reasonable. Card and Krueger (1994) assume that it is the levels of employment which evolve in the same way in PA and NJ. If employment levels were somewhat different ex ante, an equally reasonable assumption might be that the log of employment evolves in the same way absent minimum wage changes, or

$$\log y_i = \gamma_s + \lambda_t + \beta D_{ts} + \varepsilon_i.$$

This specification is different from (1), and involves a different assumption about the counterfactual trends. If one assumption is true, the other one must be necessarily false. Since the assumption is about an unobserved counterfactual, it is not testable with the data we have examined so far.

## 1.2 Group Specific Trends

Much of the recent discussion of differences-in-differences models has been about ascertaining, whether the underlying assumption of equal trends in the absence of treatment is a reasonable one. One possible way to look at this issue is if there are data available on multiple periods. For a later update of their study, Card and Krueger (2000) obtained time series of administrative payroll data for restaurants in New Jersey and Pennsylvania. This data are plotted in Figure 2 in their paper. The vertical lines indicate the dates when their original surveys were conducted. The administrative data also show a slight decline in employment from February to November 1992 in Pennsylvania, and little change in New Jersey. However, the data also reveal a large amount of ups and downs in employment in the two states. The employment trends in periods when the minimum wage was constant are often not the same in the two states. In particular, employment in New Jersey and Pennsylvania was rather similar at the end of 1991. Relative employment in Pennsylvania declined over the next three years (at least using the larger set of 14 PA counties), with much of this trend occuring at periods unrelated to the 1992 minimum wage change. Hence, eastern Pennsylvania restaurants may not be a perfect control group for New Jersey restaurants, because employment trends differ somewhat in periods with no treatment.

A more positive example is the paper by Hastings (2004). She studies the effect of the competitive environment in the retail gasoline market on gasoline prices. She uses the takeover of a large number of previously independent Thrifty gas stations in southern California in September 1997 by ARCO, a large, vertically integrated gasoline retailer. Gas stations belonging to a vertically integrated retailer typically sell gasoline at a higher price than independent stations. The hypothesis is that the presence of more independent gas stations in a local market increases cometition and therefore lowers the market price of competitors as well. Hastings investigates this hypothesis by looking at the prices of other gas stations before and after the ARCO purchase of the Thrifty stations. The treatment group in her setup are gas stations which are located near a Thrifty station, while the control group are gas stations with no Thrifty station nearby.

Figures 1a and 1b in her paper tell the story. These figures plot gasoline prices for Thrifty competitors and other stations during 1997. Prices move in parallel throughout the period, except between June and October, the period of the ARCO purchase. Prices at Thrifty competitors increase by more during this period than at comparison stations. The graphs are highly compelling that the comparison stations provide a good control group for the Thrifty competitors. Being able to produce pictures like these should be the goal of any differences-in-differences analysis.

## 1.3 Difference-in-differences in Regression Format, Multiple Contrasts, and Robustness

It is easy to see that (1) is a regression equation. If there are only two groups and two periods, then

$$\begin{aligned} y_i &= \gamma_s + \lambda_t + \beta D_{st} + \varepsilon_i \\ &= \alpha + \gamma 1(s = NJ) + \lambda 1(t = Nov) + \beta 1(s = NJ) \cdot 1(t = Nov) + \varepsilon_i \end{aligned}$$

where $1(\cdot)$ is the indicator function. Taking conditional expectations for different states and periods, and subtracting easily yields

$$\begin{aligned} \alpha &= E(y_i|s = PA, t = Feb) \\ \gamma &= E(y_i|s = NJ, t = Feb) - E(y_i|s = PA, t = Feb) \\ \lambda &= E(y_i|s = PA, t = Nov) - E(y_i|s = PA, t = Feb) \\ \beta &= [E(y_i|s = PA, t = Nov) - E(y_i|s = PA, t = Feb)] \\ &\quad - [E(y_i|s = NJ, t = Nov) - E(y_i|s = NJ, t = Feb)] \end{aligned}$$

The regression formulation of the difference-in-difference model is useful for multiple reasons. First of all, it is a convenient way of estimating the difference-in-difference, and obtaining standard errors and t-statistics. Second, it is easy to incorporate additional states or periods in the analysis now. For example, instead of just comparing the impact of the change in the minimum wage in New Jersey in a particular period, we may want to look at the impact comparing many state pairs, or comparing different periods. In this case, the formulation of the model would simply be

$$y_{st} = \gamma_s + \lambda_t + \beta D_{st} + \varepsilon_{st}$$

where $s$ and $t$ may now take on more than two values, and $y_{ts}$ is employment in state $s$ at time $t$. $D_{st}$ now indicates whether state $s$ has raised the minimum wage by date $t$.

This immediately suggests a third advantage of the regression formulation. In some cases, like in the minimum wage example, the treatment may not be binary but continuous. Different states could have different levels of the minimum wage, or the same nominal minimum wage may have a different impact depending on the distribution of wages in the state. The regression formulation would now be

$$y_{st} = \gamma_s + \lambda_t + \beta M_{st} + \varepsilon_{st} \tag{2}$$

where the variable $M_{st}$ is a measure of the "bite" of the minimum wage in state $s$ at time $t$. Despite the continuous nature of the treatment, this formulation still retains the basic features of the differences-in-differences model.

An example of the model in (2) is the paper by Card (1992). He studies the effect of the federal increase in the minimum wage in April 1990 using all the US states. The federal minimum wage was \$3.35 before the increase, and was raised to 3.80. Some states already had state minimum wages of \$3.80 or higher at the time of the federal increase. Moreover, the same increase will have more of an effect in a low wage state, where many workers are subject to the minimum, than in a high wage state. Card's measure of the impact of the increase of the minimum wage is the fraction of workers, who are paid less than \$3.80 just before the increase of the minimum wage, something he calls the "fraction of affected workers."

Since there are still only two time periods in the Card (1992) setup, before and after the minimum wage increase, (2) can be differenced over time to obtain

$$\begin{aligned} \Delta y_{st} &= \lambda_t - \lambda_{t-1} + \beta M_{st} + \Delta \varepsilon_{st} \\ &= \lambda + \beta M_{st} + \Delta \varepsilon_{st}. \end{aligned}$$

The difference in the time effect simply becomes a constant term, so that this is a standard bivariate relationship for the outcome and the treatment variable. Figures 4 and 5 plot the change in the wage and in employment against the fraction of workers affected by the minimum wage change. Figure 4 reveals that wages increased more in states where the minimum wage increase had more bite. On the other hand, figure 5 shows that there is no relationship with employment growth. Table 3 in the paper displays these results in regression format in columns (1) and (4).

Given the multiple contrasts from using 51 state level changes, it is now possible to introduce controls for other state specific factors. Since $M_{st}$ is one such variable, it is obvious that this cannot be done non-parametrically. Nevertheless, it is possible to use parametric controls for trends at the state level as in

$$\Delta y_{st} = \lambda + \beta M_{st} + X_{st}\delta + \Delta \varepsilon_{st}.$$

Card examines the wages and employment outcomes of teenagers, a group typically strongly affected by the minimum wage. He uses adult employment trends as a control (the $X_{st}$). These trends are supposed to pick up differences in the business cycle in the various states. These results are shown in columns (2), (3) and (5), (6) in his table. There is little change in the minimum wage coefficients, which is a comforting result.

Return to our model in levels. Frequently, there will be multiple periods as well as multiple treatment and control groups. In addition, outcomes are often measured at the individual level, while treatment takes place at the state level, e.g. because of a policy change. The model is now

$$y_{ist} = \gamma_s + \lambda_t + \beta M_{st} + X_{ist}\delta + \varepsilon_{ist}.$$

Variables in $X_{ist}$ could now be individual level variables or time varying variables at the state level. Including individual level variables may not only help to control for confouning trends, but may also reduce the variance of $\varepsilon_{ist}$, which may reduce the standard errors of the estimate of $\beta$.

In a model with multiple treatment groups (states) and multiple periods, it becomes more difficult to provide a simple visual inspection for the evolution of state specific trends in the periods when there is no treatment, as in Card and Krueger (2000) and Hastings (2004). Of course, identical counterfactual trends in treatment and control states is still the identifying assumption. One way to test this assumption is to allow for leads and lags of the treatment. In order to see how this works, return to the model with a binary treatment. Let $k$ be the time at which the treatment is being switched on in state $s$. Then our model is

$$y_{ist} = \gamma_s + \lambda_t + \sum_{j=-m}^{q} \beta_j D_{st}(t = k + j) + X_{ist}\delta + \varepsilon_{ist}.$$

Instead of a single treatment effect, we have now also included $m$ "leads" and $q$ "lags" of the treatment effect. $\beta_j$ is the coefficient on the $j$th lead or lag. A test of the differences in differences assumption is $\beta_j = 0 \; \forall j < 0$, i.e. the coefficients on all leads of the treatment should be zero. Moreover, the $\beta_j, j \geq 0$ may not be identical. For example, the effect of the treatment could accumulate over time, so that $\beta_j$ increases in $j$.

An example of this approach is the paper by Autor (2003). He investigates the effect of employment protection on outsourcing by firms. To this end, he relates the employment of temporary help workers in a state to indicators whether the state courts had a adopted more stringent exceptions to the employment at will doctrine. Figure 3 in his paper plots the $\beta_j$

coefficients. These coefficients are zero in the two years before the courts adopted the new rule, increasing in the first few years after the adoption, and then flat. This indicates that the differences-in-differences strategy seems successful in this context.

An alternative way to probe the robustness of the differences-in-differences identification is to include state specific parametric time trends among the regressors in $X_{ist}$. Of course, this is only possible with multiple periods again. This is done, for example, in the paper by Besley and Burgess (2004). They examine the effect of labor regulation on the performance of firms in Indian states. Different states change the regulatory regime at different times, giving rise to a differences-in-differences design. Table IV in their paper shows the key results. Column (1) shows that labor regulation leads to lower output per capita. In columns (2) and (3) they include state specific-time varying regressors like development expenditures per capita. This is a similar strategy to using the adult employment rate in Card (1992) above. This affects the estimates little. However, when they include linear state specific trends in column (4) the coefficient on labor regulation drops to zero. This suggests that the introduction of additional labor regulation correlates with other trends in state level output, and it is not possible to disentangle the causal effect of the regulation from these underlying trends. Effectively, after including a parametric trend, the identification hinges on there being a sharp change in the outcome at the date of the treatment, as in Autor (2003). If the treatment effect grows gradually, this may be much more difficult to pick up with state specific trends.

Controlling for state specific trends only works well, when there is a sufficient sample period available before the treatment starts. This is particularly true when there is a dynamic response to the treatment, so that looking at trends after the treatment is not particularly informative. Including state specific trends with the best intention can actually backfire, as Wolfers (2003) illustrates. He discusses the impact of unilateral divorce laws on divorce in the US. Before the 1970s, a divorce was only possible if both spouses agreed. In the 1970s, states introduced unilateral divorce laws, which allow a divorce if one spouse wants the divorce. An influential paper by Friedberg (1998) estimated the effect of the introduction of unilateral divorce on divorce rates, and found a sizeable and lasting effect.

Wolfers (2003) reanalyzes the data, and points that much of the result hinges on Friedberg's treatment of state specific trends. Figure 5 in his paper illustrates the problem for one state: California. Friedberg's sample starts only one year before California introduced unilateral divorce. Her estimate of the California specific trend therefore relies almost completely on the post-law trend in the state. Wolfers, using a sample going back to the late 1950s demonstrates that California's pre-exisiting trend was very different. Extrapolating this pre-existing trend results in a very different estimate of the divorce effect. As is true in many applications, this suggests both the power of large samples, and of plotting the data.

## 1.4 Picking a Good Control Group

In the discussion above, we have labeled the two dimensions $s$ and $t$ in the differences-in-differences setup "states" and "time." While there are many applications were the treatment or policy is time varying at the regional level, the identification strategy is not at all limited to these dimensions. $s$ and $t$ can be any two dimensions, so that treatment only takes place for particular combinations of $s$ and $t$. While contrasts simply across $s$ or $t$ may not plausibly identify the treatment effect, this may be more likely for the differences-in-differences estimator. Insteat of states, $s$ may denote different demographic groups, some of which are affected by a policy and others are not. For example, unemployment benefits may be changed differentially for various age groups. Anti-discrimination or job protection legislation may not apply to firms below a particular size cutoff but could be extended to additional firms. Welfare benefits may only be payable to low income families with a single parent, but not other demographic groups. The excluded groups may or may not be appropriate comparison groups. It is often the main challenge for the researcher to identify a particularly appropriate comparison group, which satifies the necessary identifying assumption, that the treated groups $s$ would behave similarly as $t$ is varied as the untreated groups.

One potential pitfall in defining treatment and control groups in a differences-in-differences setup is that $s$ or $t$ maybe directly affected by the treatment. For example, if $s$ is a state as before, we may be concerned that the policy induces some inter-state migration. Hence, the population resident in the treatment state before and after the policy becomes effective may not be identical. Say state $s = 1$ lowers wefare benefits, and this leads some poor families to move to another state $s = 0$, which forms the control group. We are interested in estimating how the lower benefits affect the fraction of the population on welfare. Also suppose, that in the absence of the policy change, welfare receipt would not have changed over time. In this case, the change of welfare receipt over time in the control state does not provide a valid counterfactual anymore: in the absence of the policy, welfare receipt in state $s = 1$ would have been unchanged. Instead, welfare receipt in state

$s = 0$ increases because of the induced welfare migration. Hence, we are overestimating the effect of the policy.

Sometimes this problem can be overcome if we know where an individual starts out. Say we know the state of residence in the period before treatment, or we know the individual's state of birth. This is immutable, i.e. cannot be affected by the treatment itself. If we assign individuals to the treatment or control group on the basis of this immutable characteristic, we can circumvent the problem outlined above. This introduces a new problem, however, that the new dimension, say state of birth, is not really the correct delineation for the treatment. I.e. some individuals "born" (or previously residing) in the treatment state move, and we would now assign them to the treatment group, even so they are not affected by the policy after their move. However, this divergence of treatment group assignment and actual treatment can easily be addressed by using instrumental variables methods, as we will discuss later.

The difference-in-difference design allows a comparison over time in the treatment group, controlling for concurrent time trends by using a control group. If there were no trends, no control group would be necessary, and a simple before-and-after design using the treatment group would be sufficient. In some circumstances, this might be quite plausible. Alternatively, it may be reasonable to assume that the underlying trend is constant over time in the treatment group, so that it is sufficient to extrapolate from different pre-treatment periods within the treatment group only.

An example of this approach can be found in Davis and Weinstein (2002). They study the growth of cities, in order to test models in economic geography. Their natural experiment is the decline in population caused by the bombing of Japanese cities in the second world war, and they are interested in how growth in these cities compares to the growth that would have taken place in the absence of bombing. Using a variety of pre-treatment years, it seems plausible that city growth is roughly exponential with a city specific growth rate. This means that extrapolating the growth of the log of population from before the war should be a valid counterfactual for the no-bombing case. Figure 2 in their paper shows the actual growth paths for Hiroshima and Nagasaki, compared to their pre-war trends. Growth in both cities is faster after the war, so that city size converges back to the pre-war growth path.

Alternatively, rather than using single differences, the treatment assignment rule may sometimes suggest a triple or higher order differences setup for the estimation. An example, is the extension of Medicaid coverage in the U.S., studied by Yelowitz (1992). Medicaid, health insurance for the poor, was traditionally tied to eligibility for AFDC, the cash welfare program. In the late 1980s, various states introduced extensions of Medicaid coverage for families with earnings high enough so they would not qualify for AFDC anymore. These extensions happened at different times for different states. This would give rise to a classical differences-in-differences design. However, different states introduced these extensions for children in different age groups. Hence, the age of the youngest child is a third dimension along which the treatment varies. Hence, Yelowitz analyzes employment effects of these extensions using the model

$$y_{iast} = \gamma_{st} + \lambda_{at} + \theta_{as} + \beta D_{ast} + X_{iast}\delta + \varepsilon_{iast}.$$

There are now three dimensions, state ($s$), time ($t$), and age of the youngest child ($a$). This allows the researcher to control non-parametrically for state specific shocks $\gamma_{st}$, i.e. each time period receives a separate dummy variable in each state. In order to only exploit the triple differences, it is also necessary to include interactions of age and time effects $\lambda_{at}$, and age and state effects $\theta_{as}$. Sometimes it may not be possible to identify the effect of the treatment with such a rich set of controls, and some of the second level interactions may have to be excluded. However, when the full set of controls is feasible, triple differences may allow for a more credible analysis.

An important challenge in evaluation design is to find a good control group. We have seen that a good control group in a differences-in-differences analysis should have similar pre-treatment trends to the treatment group, say, as in the example of Hastings (2004). Sometimes the choice of treatment group is obvious but sometimes it is not. For example, Abadie and Gareazabal (2003) try to identify the cost in terms of lost output of terrorism in the Basque country region of Spain. Basque terrorism in the 1970s was largely confined to the Basque region itself, although other Spanish regions were affected to a minor degree. Abadie and Gareazabal therefore compare growth in the Basque region to other regions in Spain. However, no single other region in Spain is a good comparison group, since the Basque country has relatively more manufacturing than the rest of the country. Abadie and Gareazabal therefore devise a method of constructing a counterfactual for the Basque region, using a weighted average of all other Spanish regions. The weights are chosen so as to mimic the pre-terrorism growth trends of the Basque country as closely as possible.[1]

Figure 1 in their paper plots GDP per capita for the Basque region as well as for their synthetic counterfactual region. Not surprisingly, the

[1] For a detailed description of the method of constructing weights, see Abadie and Gareazabal (2003).

synthetic control region tracks growth in the Basque country well in the 1960s. There is a slowdown in Basque growth in the 1970s, and output tracks the counterfactual again fairly well in the latter part of the sample when terrorist activity subsides somewhat.

## 1.5 Fixed Effects

Suppose the model of interest is again

$$y_{ist} = \gamma_s + \lambda_t + \beta D_{st} + X_{ist}\delta + \varepsilon_{ist}$$

which is one of the models we analyzed above. One of the great advantages of this setup is that the treatment happened at a well defined level of aggregation, $s$. In order to estimate this model, all we need is to sample from the population in the relevant groups $s$ in various periods $t$. The samples do not have to include the same individuals over time. This allows the use of repeated cross-section samples, which are often large and available over long periods of time. A typical example are labor force surveys, like the U.S. Current Population Survey of the UK Labor Force Survey.

However, sometimes there is no natural unit $s$ where treatment is assigned. Instead, some individuals get treated at a particular point in time, and others do not. The treatment itself may not be randomly assigned. Individuals with certain characteristics may be much more likely to be treated than others, i.e. $E(y_0|D_{it} = 1) \neq E(y_0|D_{it} = 0)$. Nevertheless, the same principle as in differences-in-differences may still apply: if we know outcomes for the individual before the treatment, and we observe other, untreated individuals, who experience the same trends over time, we can still estimate the effect of the treatment. The only additional requirement now is that we will actually need data on the same individuals over time.

Our model now simply becomes

$$y_{it} = \alpha_i + \lambda_t + \beta D_{it} + X_{it}\delta + \varepsilon_{it} \tag{3}$$

where we have simply replaced the state effect $\gamma_s$ with an individual effect $\alpha_i$. As long as $E(y_0|\alpha_i, X_{it}, D_{it}) = E(y_0|\alpha_i, X_{it})$ holds, this model will allow us to identify the treatment effect. The identifying assumption says that counterfactual outcomes in the absence of treatment are independent of treatment, conditional on an individual effect $\alpha_i$ and covariates $X_{it}$. Alternatively, it says that treatment is only determined by the fixed individual effect $\alpha_i$ and covariates $X_{it}$. The model in equation (3) can be estimated as a *fixed effects model*, i.e. treating $\alpha_i$ as a parameter to be estimated.

In practice, with many individuals it is typically not feasible to estimate the individual dummy variables in equation (3) directly. Instead, the model is either estimated by differencing out the fixed effect

$$\Delta y_{it} = \Delta \lambda_t + \beta \Delta D_{it} + \Delta X_{it} \delta + \Delta \varepsilon_{it}$$

or by taking deviations from means. Note that taking means across individuals yields

$$\overline{y}_i = \alpha_i + \overline{\lambda} + \beta \overline{D}_i + \overline{X}_i \delta + \overline{\varepsilon}_i$$

so that

$$y_{it} - \overline{y}_i = \lambda_t - \overline{\lambda} + \beta \left( D_{it} - \overline{D}_i \right) + \left( X_{it} - \overline{X}_i \right) \delta + \varepsilon_{it} - \overline{\varepsilon}_i,$$

which also sweeps out the individual effect. The two estimators are identical if there are only two periods. For more periods, if $\varepsilon_{it}$ is iid, then the difference estimator introduces serial correlation in the new error $\Delta \varepsilon_{it}$. This would have to be accounted for in calculating the covariance matrix. Regression packages will typically implement the deviations from means estimator, with an appropriate adjustment for the degrees of freedoms lost in estimating the $N$ individual level means. This estimator will often be referred to as the within estimator, and the procedure as absorbing the $\alpha_i$s. In Stata, this estimator is implemented in the commands `xtreg` and `areg`.

In looking at model (3), this could also be considered a *random effects model*

$$\begin{aligned} y_{it} &= \lambda_t + \beta D_{it} + X_{it} \delta + u_{it} \\ u_{it} &= \alpha_i + \varepsilon_{it} \end{aligned}$$

where $\alpha_i$ is now considered a random disturbance, not a parameter to be estimated. Of course, we were originally worried about the fact that selection, i.e. $D_{it}$, depends on $\alpha_i$. Hence, simply treating $\alpha_i$ as part of the error term and estimating the model, for example, by GLS does not solve the problem, since the standard random effects estimator still assumes that the error term $u_{it}$ is uncorrelated with the regressors. This is because the random effects model still uses the cross sectional variation, and hence will be contaminated by the correlation between $D_{it}$ and $\alpha_i$.

However, the distinction between fixed and random effects is not really critical for the consistent estimation of $\beta$. Instead, important is whether we treat $\alpha_i$ as a correlated or uncorrelated random effect. I.e. even in the random effects model it is possible to acknowledge the correlation between

$D_{it}$ and $\alpha_i$ by modeling the relationship. See Chamberlain (1994) for details on how this can be done.

A classic example for the fixed effects model is the estimation of the union wage differential. Let $y_{it}$ equal log earnings, and $D_{it}$ the individual's union status. We are interested in the causal effect of union membership (or coverage by a union contract) on earnings. However, we are concerned that union firms may hire different types of workers. For example, union firms may have workers who are more productive on average. As long as the productivity difference is confined to the individual effect $\alpha_i$, the fixed effects model (3) will allow consistent estimation of the union wage effect $\beta$.

Freeman (1984) analyzes this case. Table 6 displays estimates for the union wage effects from four data sets, using both the cross section and the fixed effects estimator (the fixed effects estimates are the ones for the group labeled (*NU - UN)/2*). The cross section estimates are typically higher (in the range of 0.15 to 0.25) than the fixed effects estimates (in the range 0.10 to 0.20). This may indicate positive selection of union workers.

One problem with fixed effects estimation is that it tends to accentuate measurement error problems. Freeman presents an example in table 1. In the example there are 100 workers, 75 unionized and 25 not unionized, the true union wage premium is 30 percent, and there is no selection among union workers. Suppose we use the cross section only, and there are two union members and two non-union members each who are misclassified. This would result in an estimated union wage effect of 0.268. The measurement error leads to some attenuation.

Now consider a two period panel. Of the 75 union members 10 leave the union after the first period, and 10 non-union members join. In each period, there are still two of each type of workers misclassified. The fixed effects estimator only uses the wage variation for the workers who join or leave the union. There are 12 observed joiners and 12 observed leavers, but only 9 of each group represent a true transition. Because relatively few workers change union status, a much larger fraction of the transitions is now due to measurement error. The estimated union wage differential among the leavers and joiners is 0.225, substantially below the cross-sectional estimate. Hence, measurement error may well be responsible for the lower fixed effects estimate of the union wage effect.

The same problem arises with a continuous regressor $M_{it}$. Suppose instead of $M_{it}$ we observe the mismeasured variable $\widetilde{M_{it}} = M_{it} + v_{it}$, where $v_{it}$ is a classical measurement error, uncorrelated with any other variable. Consider the model

$$y_{it} = \alpha_i + \beta M_{it} + \varepsilon_{it}.$$

In the bivariate cross-sectional regression

$$\widehat{\beta}_{CS} = \frac{cov(y_{it}, \widetilde{M}_{it})}{var(\widetilde{M}_{it})}$$

and hence

$$\text{plim}\widehat{\beta}_{CS} = \frac{cov(\alpha_i + \beta M_{it} + \varepsilon_{it}, \widetilde{M}_{it})}{var(\widetilde{M}_{it})} = \beta \frac{var(M_{it})}{var(M_{it}) + var(v_{it})}$$

assuming that $\alpha_i$ is uncorrelated with the regressor $M_{it}$. The estimate of $\widehat{\beta}$ is attenuated by the factor $var(M_{it})/[var(M_{it}) + var(v_{it})]$.

Now consider the differeced estimator of the same model

$$\Delta y_{it} = \beta \Delta M_{it} + \Delta \varepsilon_{it}.$$

Hence

$$\widehat{\beta}_{FE} = \frac{cov(\Delta y_{it}, \Delta \widetilde{M}_{it})}{var(\Delta \widetilde{M}_{it})}$$

and

$$\text{plim}\widehat{\beta}_{FE} = \frac{cov(\beta \Delta M_{it} + \Delta \varepsilon_{it}, \Delta M_{it} + \Delta v_{it})}{var(\Delta M_{it} + \Delta v_{it})} = \beta \frac{var(\Delta M_{it})}{var(\Delta M_{it}) + var(\Delta v_{it})}.$$

Let $\rho_x$ denote the autocorrelation coefficient in variable $x_{it}$. Then we have

$$\begin{aligned} \text{plim}\widehat{\beta}_{FE} &= \beta \frac{var(M_{it})(1 - \rho_M)}{var(M_{it})(1 - \rho_M) + var(v_{it})(1 - \rho_v)} \\ &= \beta \frac{var(M_{it})}{var(M_{it}) + var(v_{it}) \frac{(1-\rho_v)}{(1-\rho_M)}}. \end{aligned}$$

In many applications, $\rho_M$ will tend to be high, because many economic variables are rather persistent. In the other hand, $\rho_v$ may be small. Measurement error is likely to have a large transitory component. This implies $(1 - \rho_v)/(1 - \rho_M)$ will be larger than one, and hence the attenuation bias in the fixed effects estimator will be larger than in the cross-section estimator. If the difference in $\rho_M$ and $\rho_v$ is large, the difference in the bias can also be large. For example, if $\rho_M = 0.9$ and $\rho_v = 0$, then the variance component $var(v_{it})$ gets a weight 10 times as large as in the cross-section case.

The insights from the analysis of measurement error are more general. Taking out fixed effects may remove a lot of the variance in the treatment effect $D_{it}$ or $M_{it}$. Our assumption is that this variance is harmful in our exercise because (part of) it is correlated with the individual fixed effect. However, as we have seen in the measurement error case, it is quite possible that taking out fixed effects removes both "good" and "bad" variation. This is particularly troubling if the fixed effects strategy is imperfect, and some "bad" variation is left in the fixed effects estimates. Because much of the "good" variation has been filtered out, the consequences of the "bad" variation also get accentuated. Hence, it easy to throw out the baby with the bathwater.

An example for this type of concern is related to twin based estimates of returns to schooling. Ashenfelter and Krueger (1994) and Ashenfelter and Rouse (1998) present estimates of the returns to schooling among twins, controlling for twin pair fixed effects. Hence, these estimates compare the difference in earnings across twins to differences in schooling. The idea that ability is related to either genetics, family background or school environment, which are all captured by the twin fixed effect. Any remaining difference in schooling should therefore be unrelated to ability.

But how do differences in schooling come about between individuals who otherwise so much alike? Bound and Solon (1999) point out that there are small differences between twins, with first borns typically having higher birthweight but also higher IQ scores. While these twin differences are not large, neither is the difference in schooling. Hence, a small amount of remaining selection among twins in terms of their schooling attainment could be responsible for a large amount of bias in the resulting estimates. The challenge of good evaluation work is always to remove "bad" variation and to leave as much "good" variation intact as possible.

### 1.6 Selection on Past Outcomes

A further question arises when the researcher has panel data available. One assumption is that selection depends on the fixed effect $E(y_0|\alpha_i, X_{it}, D_{it}) = E(y_0|\alpha_i, X_{it})$. But an alternative assumption would be $E(y_0|y_{it-h}, X_{it}, D_{it}) = E(y_0|y_{it-h}, X_{it})$, i.e. that selection depends on some value of the lagged outcome. In this case, the correct model to estimate would be

$$y_{it} = \alpha + \theta y_{it-h} + \lambda_t + \beta D_{it} + X_{it}\delta + \varepsilon_{it}. \qquad (4)$$

But now assume that model (3) is correct and the researcher estimates (4) instead. Ignoring other covariates, i.e. the $\lambda_t$ and $X_{it}\delta$ terms, we have

$$\widehat{\beta} = \frac{var(y_{it-h})cov(y_{it}, D_{it}) - cov(y_{it-h}, D_{it})cov(y_{it}, y_{it-h})}{var(D_{it})var(y_{it-h}) - cov(y_{it-h}, D_{it})^2}.$$

If $D_{it-h} = 0$ and $E(D_{it}) = p$, then we get

$$\begin{aligned}
\text{plim}\widehat{\beta} &= \frac{(\sigma_\alpha^2 + \sigma_\varepsilon^2)\left[cov(\alpha_i, D_{it}) + \beta p(1-p)\right] - cov(\alpha_i, D_{it})(\sigma_\alpha^2 + \sigma_\varepsilon^2)}{p(1-p)(\sigma_\alpha^2 + \sigma_\varepsilon^2) - cov(\alpha_i, D_{it})^2} \\
&= \frac{(\sigma_\alpha^2 + \sigma_\varepsilon^2)\left[cov(\alpha_i, D_{it}) + \beta p(1-p) - cov(\alpha_i, D_{it})\right]}{p(1-p)(\sigma_\alpha^2 + \sigma_\varepsilon^2) - cov(\alpha_i, D_{it})^2} \\
&= \beta \frac{p(1-p)(\sigma_\alpha^2 + \sigma_\varepsilon^2)}{p(1-p)(\sigma_\alpha^2 + \sigma_\varepsilon^2) - cov(\alpha_i, D_{it})^2}.
\end{aligned}$$

Similarly, the treatment effect will not be consistently estimated if (4) is correct and the researcher estimates (3) instead. Ignoring the other covariates again

$$\begin{aligned}
\widehat{\beta} &= \frac{cov(\Delta y_{it}, \Delta D_{it})}{var(\Delta D_{it})} \\
\text{plim}\widehat{\beta} &= \beta + \frac{\theta cov(\Delta y_{it-1}, \Delta D_{it})}{var(\Delta D_{it})}.
\end{aligned}$$

The second term will not be equal to zero because selection (and hence $D_{it}$) depends on $y_{it-1}$.

One possible solution to this would be to be agnostic about the correct model and estimate

$$y_{it} = \alpha_i + \theta y_{it-h} + \lambda_t + \beta D_{it} + X_{it}\delta + \varepsilon_{it}$$

which allows for both a fixed effect and a lagged dependent variable. One problem with this model is that the standard fixed effects estimators are no longer consistent with a once lagged dependent variable. Set $h = 1$, and consider, for example, differencing of the equation, which yields

$$\Delta y_{it} = \theta \Delta y_{it-1} + \Delta \lambda_t + \beta \Delta D_{it} + \Delta X_{it}\delta + \Delta \varepsilon_{it}.$$

This still removes the fixed effect but the $\varepsilon_{it-1}$, which is part of the error term, is correlated with $y_{it-1}$. Hence, this equation cannot be estimated consistently directly. This problem was first pointed out by Nickell (1981). There are various potential solutions to this problem, typically involving instrumenting $\Delta y_{it-1}$ with further lags of the dependent or independent variables, which are uncorrelated with $\Delta \varepsilon_{it}$.

An example where treatment assignment may depend on past realizations of $y_{it}$ is in active labor market programs. For example, workers may be selected for a government training program because their earnings are below a certain threshold. Ashenfelter and Card (1985) analyze the earnings of workers who participated in the CETA (Comprehensive Education and Training Act) training program in 1976-77. Data on trainees are from their social security records. Data on a control group are drawn from individuals from the March 1976 Current Population Survey, who were eligible for the program but did not participate. Social security earnigns data for the control group are also available. In their table 1, Ashenfelter and Card find that trainees look worse in terms of pre-training earnings than controls.

Assume that training for trainees takes place in year $k$. Hence we define

$$D_{it} = 1 \text{ if } t > k \text{ and } i \text{ is a trainee}$$

so that we identify the treatment with the period when the individual is trained.

Suppose the selection rule into training is

$$D_{ik+1} = 1 \text{ if } \alpha_i < \overline{y}.$$

Standard differences-in-differences or fixed effects estimates produce a valid estimate of $\beta$ in this case.

$$E(y_{ik+1} - y_{ik-j} | D_{ik+1} = 1) - E(y_{ik+1} - y_{ik-j} | D_{ik+1} = 0) = \beta$$

Since selection does not depend on $\varepsilon_{it}$ this is true irrespective of the process for $\varepsilon_{it}$.

Now suppose instead that the selection rule is

$$\begin{aligned} D_{ik+1} &= 1 \text{ if } y_{ik-h} < \overline{y} \\ &\Rightarrow \alpha_i + \lambda_{k-h} + \varepsilon_{ik-h} < \overline{y} \end{aligned}$$

i.e. selection is based on actual earnings $h$ years before the training. Differences-in-differences produces now

$$\begin{aligned} E(y_{ik+1} - y_{ik-j} | D_{ik+1} &= 1) - E(y_{ik+1} - y_{ik-j} | D_{ik+1} = 0) \\ &= \beta + E(\varepsilon_{ik+1} - \varepsilon_{ik-j} | D_{ik+1} = 1) - E(\varepsilon_{ik+1} - \varepsilon_{ik-j} | D_{ik+1} = 0). \end{aligned}$$

The differences-in-differences estimator will yield an estimate of $\beta$ on if

- $j > h$,
- $\varepsilon_{it}$ is serially uncorrelated.

This implies that the estimates using different pre-treatment years should be the same, as long as the above two assumptions are satisfied. However, Ashenfelter and Card (1985) find that using different pre-training years produces very different estimates of the training effect. This is worrisome, and may suggest that the assumption of uncorrelated earnings is not satisfied. In fact, they note that earnings for the trainees in 1975, just before the training, are particularly low. Hence, higher earnings for the trainees after training may simply reflect mean reversion: the fact that earnings are rebounding from a temporary drop in 1975. The temporarily low earnings in right before the advent of training are often referred to as "Ashenfelter's dip."

Ashenfelter and Card implement a solution to this problem first suggested by James Heckman: to take symmetric differences-in-differences around the year of selection. This estimator should produce consistent estimates of the training effect, as long as $\varepsilon_{it}$ is covariance stationary. The intution for this simple: covariance stationarity implies that the rebound in earnings from a temporary low going forward by $h$ years is just as strong as is the rebound going back $h$ years. Without a training effect, earnings at these two points in time are expected to be equal.

Formally, this can be seen as follows. Note, if $(x_1, x_2) \sim$ jointly normal then

$$E(x_1|x_2) = E(x_1) + \frac{cov(x_1, x_2)}{var(x_2)}\left[x_2 - E(x_2)\right].$$

Assume for the moment that $\beta = 0$. Using the result above we have

$$E(y_{ik+1}|y_{ik-h} < \overline{y}) = E(y_{ik+1}) + \frac{cov(y_{ik+1}, y_{ik-h})}{var(y_{ik-h})}\left[E(y_{ik-h}|y_{ik-h} < \overline{y}) - E(y_{ik-h})\right]$$

and similarly, $h+1$ years before training

$$\begin{aligned} E(y_{ik-h-(h+1)}|y_{ik-h} &< \overline{y}) = E(y_{ik-h-(h+1)}) \\ &+ \frac{cov(y_{ik-h-(h+1)}, y_{ik-h})}{var(y_{ik-h})}\left[E(y_{ik-h}|y_{ik-h} < \overline{y}) - E(y_{ik-h})\right]. \end{aligned}$$

Now by covariance stationarity of $y_t$, $cov(y_{ik+1}, y_{ik-h}) = cov(y_{ik-h-(h+1)}, y_{ik-h})$. This implies that

$$E(y_{ik+1}|y_{ik-h} < \overline{y}) - E(y_{ik-h-(h+1)}|y_{ik-h} < \overline{y}) = E(y_{ik+1}) - E(y_{ik-h-(h+1)})$$

and hence the symmetric difference-in-difference is independent of the selection into training. Returning to the case where $\beta \neq 0$ again, we therfore have:

$E(y_{ik+1} - y_{ik-h-(h+1)}|D_{ik+1} = 1) - E(y_{ik+1} - y_{ik-h-(h+1)}|D_{ik+1} = 0) = \beta.$

For the CETA trainees, who finish with training in 1976, there are two post-training years, 1977 and 1978, so we can construct two such estimates. Both estimates should be similar, and this in fact provides an overidentification test for the underlying model. If the year of selection is 1975, then the two comparisons would be 1977 with 1973 or 1978 with 1972. Both of these yield an estimate of the training effect of around -\$800. If the year of selection is 1976 instead, the two comparisons would be 1977 with 1975 or 1978 with 1974. The estimates in these cases are about \$400 and \$0 (Ashenfelter and Card, 1985, table 2). Hence, if we do not know the year of selection, we cannot ascertain what the correct estimate is and whether the model fits the data.

## 1.7 Inference in Panel Data and Differences-in-Differences Models

The discussion so far has concentrated on identification of the effect of interest. Obviously, this always should be the main concern. However, there are also a number of important inference issues which arise in the use of panel data and differences-in-differences models. All these problems have to do with correlation of the errors across the units of observation. Start by considering the simple model

$$y_{it} = \alpha + \beta x_t + \varepsilon_{it} \tag{5}$$

where the outcome is observed at the individual level but the regressor of interest, $x_t$, varies only at a higher level of aggregation. If $x_t$ is as good as randomly assigned, then the OLS estimator is unbiased and consistent but OLS standard errors will not be consistent if the error term has group structure

$$\varepsilon_{it} = v_t + \eta_{it}.$$

This problem of correlation in the errors is, of course, well known in econometrics. Moulton (1986), however, pointed out how important it can be in the grouped regressor case.

In order to analyze the problem, let

$$y_t = \begin{bmatrix} y_{1t} \\ y_{2t} \\ \vdots \\ y_{n_t t} \end{bmatrix} \quad \varepsilon_t = \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \\ \vdots \\ \varepsilon_{n_t t} \end{bmatrix}$$

and

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{bmatrix} \quad x = \begin{bmatrix} \iota_1 x_1 \\ \iota_2 x_2 \\ \vdots \\ \iota_T x_T \end{bmatrix} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_T \end{bmatrix}$$

where $\iota_t$ is a column vector of $n_t$ ones. Notice that

$$\begin{aligned} E(\varepsilon\varepsilon') &= \sigma_\varepsilon^2 G = \sigma_\varepsilon^2 \begin{bmatrix} G_1 & 0 & \cdots & 0 \\ 0 & G_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & G_T \end{bmatrix} \\ G_t &= \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & & \vdots \\ \vdots & & \ddots & \rho \\ \rho & \cdots & \rho & 1 \end{bmatrix} = (1-\rho)I + \rho\iota_t\iota_t' \\ \rho &= \frac{\sigma_v^2}{\sigma_v^2 + \sigma_\eta^2}. \end{aligned}$$

Now

$$\begin{aligned} X'X &= \sum_t n_t x_t x_t' \\ X'GX &= \sum_t x_t \iota_t' G_t \iota_t x_t'. \end{aligned}$$

But

$$\begin{aligned} x_t \iota_t' G_t \iota_t x_t' &= x_t \iota_t' \begin{bmatrix} 1 + (n_t - 1)\rho \\ 1 + (n_t - 1)\rho \\ \cdots \\ 1 + (n_t - 1)\rho \end{bmatrix} x_t' \\ &= n_t \left[1 + (n_t - 1)\rho\right] x_t x_t'. \end{aligned}$$

Denote $\tau_t = 1 + (n_t - 1)\rho$, so we get

$$\begin{aligned} x_t \iota_t' G_t \iota_t x_t' &= n_t \tau_t x_t x_t' \\ X'GX &= \sum_t n_t \tau_t x_t x_t'. \end{aligned}$$

With this at hand, we can compute the covariance matrix of the OLS estimator, which is

$$\begin{aligned} var(\widehat{\beta}_{OLS}) &= \sigma_{\varepsilon}^{2}\left(X'X\right)^{-1}X'GX\left(X'X\right)^{-1} \\ &= \sigma_{\varepsilon}^{2}\left(\sum_{t} n_t x_t x_t'\right)^{-1}\sum_{t} n_t \tau_t x_t x_t'\left(\sum_{t} n_t x_t x_t'\right)^{-1}. \end{aligned}$$

We want to compare this with the standard OLS covariance estimator

$$var^{*}(\widehat{\beta}_{OLS}) = \sigma_{\varepsilon}^{2}\left(\sum_{t} n_t x_t x_t'\right)^{-1}.$$

In the case of equal group sizes, i.e. $n_t = n$, we have $\tau = 1 + (n-1)\rho$ and

$$\frac{var(\widehat{\beta}_{OLS})}{var^{*}(\widehat{\beta}_{OLS})} = \tau = 1 + (n-1)\rho. \tag{6}$$

Notice that the OLS standard error formula will be worse if $n$ is large and if $\rho$ is large. To see the intuition, consider the case where $\rho \to 1$. In this case, all the errors within a group are the same. This is just like taking a data set and making $n$ identical copies. The covariance matrix of the replicated data set is going to be $1/n$ times the original covariance matrix, although no information has been added. Hence, $\tau = n$.

In order to see how this problem is related to the group structure in the regressor $x$, consider the generalization of (6):

$$\begin{aligned} \frac{var(\widehat{\beta}_{OLS})}{var^{*}(\widehat{\beta}_{OLS})} &= 1 + \left[\frac{var(n_t)}{\overline{n}} + \overline{n} - 1\right]\rho_x \rho \\ \rho_x &= \frac{\sum_t \sum_{i \neq k}\left(x_{it} - \overline{x}\right)\left(x_{kt} - \overline{x}\right)}{var(x_{it})\sum_t n_t(n_t - 1)}. \end{aligned}$$

$\rho_x$ is the within group correlation of $x_{it}$. What the formula says is that the bias in the OLS formula is much worse when $\rho_x$ is large but vanishes when $\rho_x = 0$. If the $x_{it}$'s are uncorrelated within groups, the error structure does not matter for the estimation of the standard errors.

The magnitude of the problem can be assessed by returning to the case $\rho_x = 1$ and $n_t = n$:

| | $\rho$ | | |
|---|---|---|---|
| $n$ | 0.05 | 0.20 | 0.50 |
| 10 | 1.20 | 1.67 | 2.35 |
| 50 | 1.86 | 3.29 | 5.05 |
| 500 | 5.09 | 10.04 | 15.83 |

It is easy to see that either a moderate $\rho$ or a moderate $n$ is enough to lead to seriously misleading inference from OLS standard errors. With large micro data sets and a limited number of groups, and hence very large $n$, even a very small $\rho$ is sufficient for misleading inference.

There are various solutions to this problem:

1. Obtain an estimate of $\rho$ and calculate the standard errors using the correct formula given by Moulton. The only non-standard part is the estimate of the intraclass correlation $\rho_x$, but this can typically be implemented with the programming tools in standard regression packages.

2. Clustered standard errors: A non-parametric correction for the standard errors is given by the following estimate of the covariance matrix

$$\begin{aligned} var(\widehat{\beta}_{OLS}) &= \left(\sum_t n_t x_t x_t'\right)^{-1} \sum_t n_t x_t \iota_t' \widehat{G}_t \iota_t x_t' \left(\sum_t n_t x_t x_t'\right)^{-1} \\ \widehat{G}_t &= \begin{bmatrix} \widehat{\varepsilon}_{1t}^2 & \widehat{\varepsilon}_{1t}\widehat{\varepsilon}_{2t} & \cdots & \widehat{\varepsilon}_{1t}\widehat{\varepsilon}_{n_t t} \\ \widehat{\varepsilon}_{1t}\widehat{\varepsilon}_{2t} & \widehat{\varepsilon}_{2t}^2 & & \vdots \\ \vdots & & \ddots & \widehat{\varepsilon}_{(n_t-1)t}\widehat{\varepsilon}_{n_t t} \\ \widehat{\varepsilon}_{1t}\widehat{\varepsilon}_{n_t t} & \cdots & \widehat{\varepsilon}_{(n_t-1)t}\widehat{\varepsilon}_{n_t t} & \widehat{\varepsilon}_{n_t t}^2 \end{bmatrix}. \end{aligned}$$

   This is implemented in Stata as the cluster option, and works well with a reasonable number of groups (as few as 10 in many applications).

3. Aggregation to the group level: Calculate $\overline{y}_t$ first and then run a weigthed least squares regression

$$\overline{y}_t = \alpha + \beta x_t + \overline{\varepsilon}_t$$

   with the number of observations in the group as weights (or the inverse of the sampling variance of $\overline{y}_t$). The error term at this aggregated level is $\overline{\varepsilon}_t = v_t + \overline{\eta}_t$, and the error component $v_t$ is therefore considered in the usual second step standard errors. If there are other micro level regressors in the model, as in

$$y_{it} = \lambda_t + \beta M_t + X_{it}\delta + \varepsilon_{it},$$

   we can do the aggregation by running the regression

$$y_{it} = \sum_\tau \widetilde{y}_\tau 1(t = \tau) + X_{it}\delta + \varepsilon_{it}.$$

The coefficients on these dummies are our time means, purged of the effect of the individual level variables in $X_{it}$.

4. Block bootstrap: Bootstrapping means to draw random samples from the empirical distribution of the data. Since the best representation of the empirical distribution of the data is the data itself, this means in practice for a sample of size $n$, to draw another sample of size $n$ with replacement from the original data set. This can be done many times, and the estimate is computed for all the bootstrap samples. The standard error of the estimate is the standard deviation of the estimates across all the bootstrap samples. In block bootstrapping, the bootstrap draws will not be a tuple $\{y_{it}, x_{it}\}$, but instead a whole block as defined by the groups $t$ is drawn together. Hence, any correlation across the errors within the block will be kept intact with the block bootstrap sampling, and should therefore be reflected in the standard error estimate.

5. Estimate a random effects GLS or ML model of equation (5). This has not been particularly popular in the recent applied microeconomics literature compared to adjusting the OLS standard errors.

Now suppose that there are only two groups, i.e. the regressor of interest is a dummy variable again:

$$y_{it} = \alpha + \beta D_t + v_t + \eta_{it}. \tag{7}$$

The Moulton problem does not arise in this case, because OLS fits the regression line perfectly through the two points defined by the dummy variable. However, it really fits the line through the two points defined by both $D_t$ and $v_t$. In practice this means that the estimate of $\beta$ will be unbiased but not consistent, as pointed out by Donald and Lang (2001). In every new sample, there will be a new draw of $v_t$. So the regression line will be somewhat off, and the estimate will not exactly equal $\beta$. However, on average, there will be no bias: sometimes $\beta$ will be overestimated, sometimes underestimated. Now suppose we take a single sample and let $n \to \infty$, while $T = 2$ remains constant. The bias that exists in any particular sample will not go to zero, because $v_t$ is just as imporant in the big sample as in the small sample. Only the sampling variation due to $\eta_{it}$ will vanish, not the sampling variation due to $v_t$.

This problem also arises in the standard 2x2 difference-in-difference model if there is a state-time specific component to the error term

$$y_{ist} = \gamma_s + \lambda_t + \beta D_{st} + v_{st} + \eta_{ist}.$$

Because the model is saturated (the maximum number of dummies of $s$ and $t$ have been included) this is really equivalent to the model (7) if there are two states and two periods. By simply taking $n \to \infty$, the error component $v_{st}$ does not vanish. Moreover, there is really no way to get consistent standard errors which acknowledge this problem because $\beta D_{st} + v_{st}$ is completely collinear. So no separate estimate of $\beta$ and $v_{st}$ possible. This means that 2x2 difference-in-differences, like the original New Jersey-Pennsylvania comparison of the employment effects of the minimum wage are not really very informative.

The solution is to have either multiple time periods on two states, as in the Card and Krueger (2000) reanalysis of the New Jersey-Pennsylvania experiment with a longer time series of payroll data , or multiple contrasts for two time periods, like in Card (1992). If $v_{st}$ is iid, adjusting the standard errors is relatively straightforward. Donald and Lang (2001) give formulas for consistent standard errors based on estimation $\sigma^2_{vst}$. The alternatives are similarly to the ones discussed above, i.e. to cluster the standard errors by $s * t$, i.e. at the state-time level or to aggregate the data to the state-time level.[2]

Bertrand, Duflo, and Mullainathan (2004) point out a further problem. Many of the economic outcome variables of interest tend to be correlated over time. This means, $v_{st}$ is most likely serially correlated. Hence, the solutions which treat $v_{st}$ as iid are not sufficient. They investigate a variety of remedies, like clustering at the state level, block bootstrap methods at the state level, ignoring the time series information by aggregating the data into two periods, or parametric modeling of the serial correlation. They conclude that most of the non-parametric methods perform well but only when there is a sufficiently large number of states available.

The conclusion from this discussion is that correlated errors are likely to be a problem in many applications. This means that adjusting the standard errors for this correlation is important. Unfortunately, consistent estimators for the covariance matrix are not available when the number of groups is small. Really the only way out is to have a design that allows for many contrasts.

[2] See Conley and Taber (2004) for another approach to this problem.

Card and Krueger 1994: Table 3

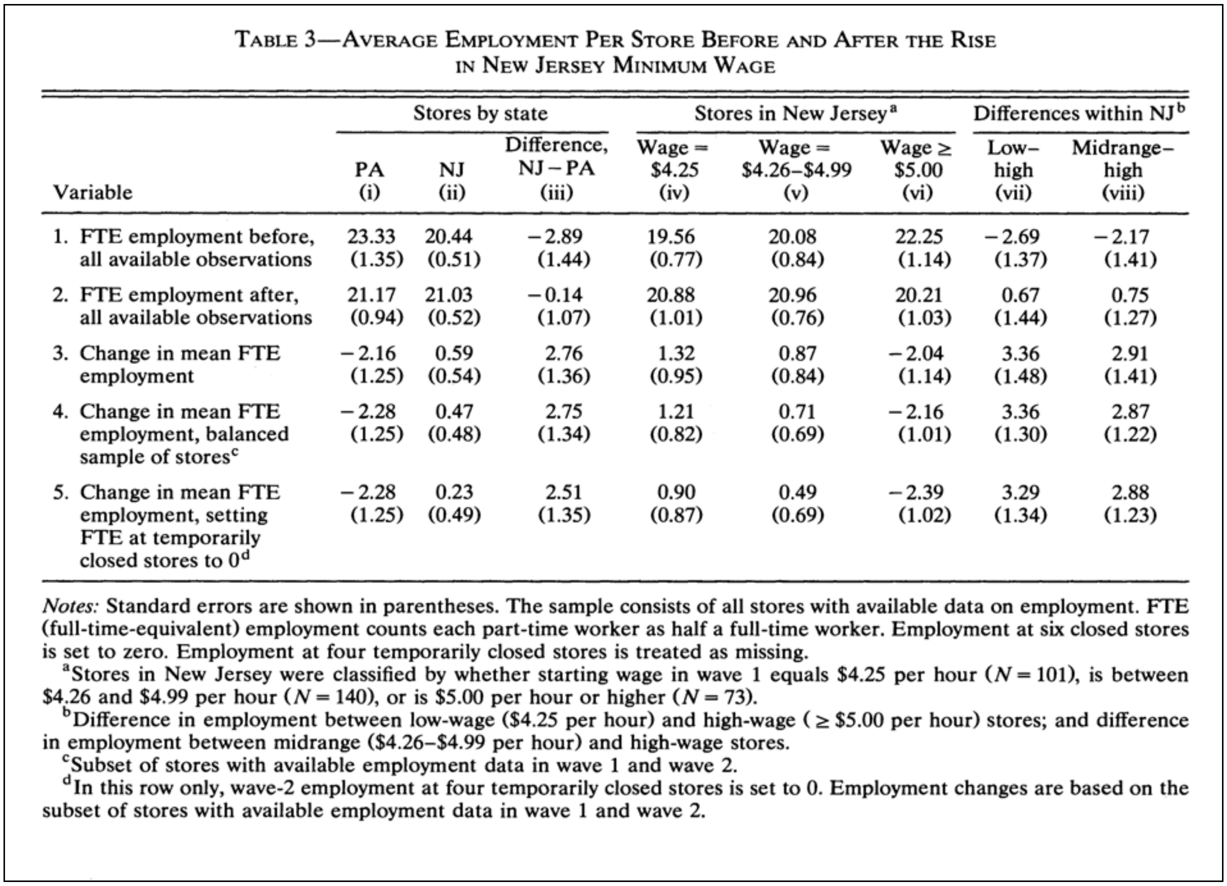


TABLE 3—AVERAGE EMPLOYMENT PER STORE BEFORE AND AFTER THE RISE IN NEW JERSEY MINIMUM WAGE

| | Stores by state | | | Stores in New Jersey[a] | | | Differences within NJ[b] | |
|---|---|---|---|---|---|---|---|---|
| Variable | PA<br>(i) | NJ<br>(ii) | Difference, NJ − PA<br>(iii) | Wage = $4.25<br>(iv) | Wage = $4.26–$4.99<br>(v) | Wage ≥ $5.00<br>(vi) | Low–high<br>(vii) | Midrange–high<br>(viii) |
| 1. FTE employment before, all available observations | 23.33<br>(1.35) | 20.44<br>(0.51) | −2.89<br>(1.44) | 19.56<br>(0.77) | 20.08<br>(0.84) | 22.25<br>(1.14) | −2.69<br>(1.37) | −2.17<br>(1.41) |
| 2. FTE employment after, all available observations | 21.17<br>(0.94) | 21.03<br>(0.52) | −0.14<br>(1.07) | 20.88<br>(1.01) | 20.96<br>(0.76) | 20.21<br>(1.03) | 0.67<br>(1.44) | 0.75<br>(1.27) |
| 3. Change in mean FTE employment | −2.16<br>(1.25) | 0.59<br>(0.54) | 2.76<br>(1.36) | 1.32<br>(0.95) | 0.87<br>(0.84) | −2.04<br>(1.14) | 3.36<br>(1.48) | 2.91<br>(1.41) |
| 4. Change in mean FTE employment, balanced sample of stores[c] | −2.28<br>(1.25) | 0.47<br>(0.48) | 2.75<br>(1.34) | 1.21<br>(0.82) | 0.71<br>(0.69) | −2.16<br>(1.01) | 3.36<br>(1.30) | 2.87<br>(1.22) |
| 5. Change in mean FTE employment, setting FTE at temporarily closed stores to 0[d] | −2.28<br>(1.25) | 0.23<br>(0.49) | 2.51<br>(1.35) | 0.90<br>(0.87) | 0.49<br>(0.69) | −2.39<br>(1.02) | 3.29<br>(1.34) | 2.88<br>(1.23) |

*Notes:* Standard errors are shown in parentheses. The sample consists of all stores with available data on employment. FTE (full-time-equivalent) employment counts each part-time worker as half a full-time worker. Employment at six closed stores is set to zero. Employment at four temporarily closed stores is treated as missing.

[a] Stores in New Jersey were classified by whether starting wage in wave 1 equals $4.25 per hour ($N = 101$), is between $4.26 and $4.99 per hour ($N = 140$), or is $5.00 per hour or higher ($N = 73$).

[b] Difference in employment between low-wage ($4.25 per hour) and high-wage (≥ $5.00 per hour) stores; and difference in employment between midrange ($4.26–$4.99 per hour) and high-wage stores.

[c] Subset of stores with available employment data in wave 1 and wave 2.

[d] In this row only, wave-2 employment at four temporarily closed stores is set to 0. Employment changes are based on the subset of stores with available employment data in wave 1 and wave 2.

Card and Krueger 2000: Figure 2

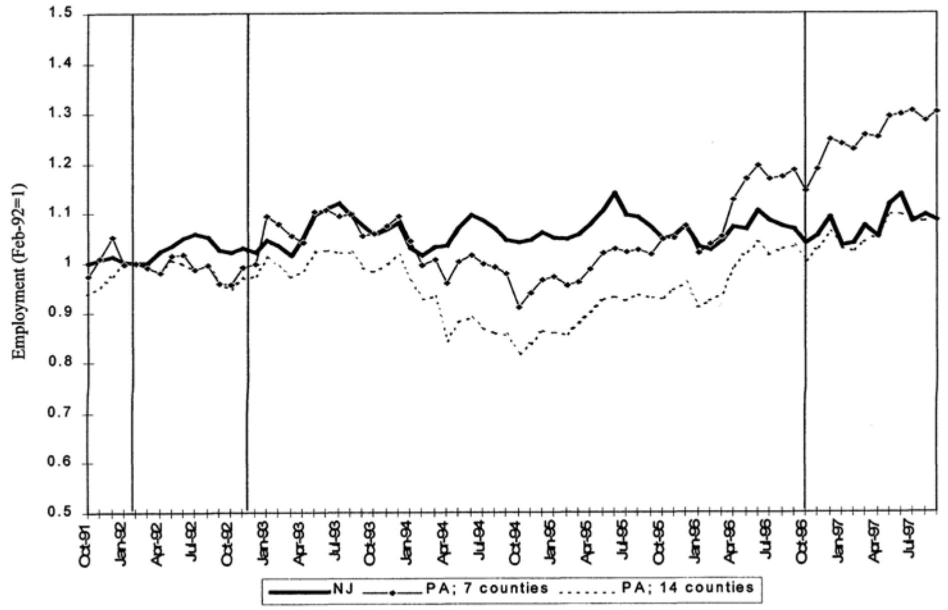
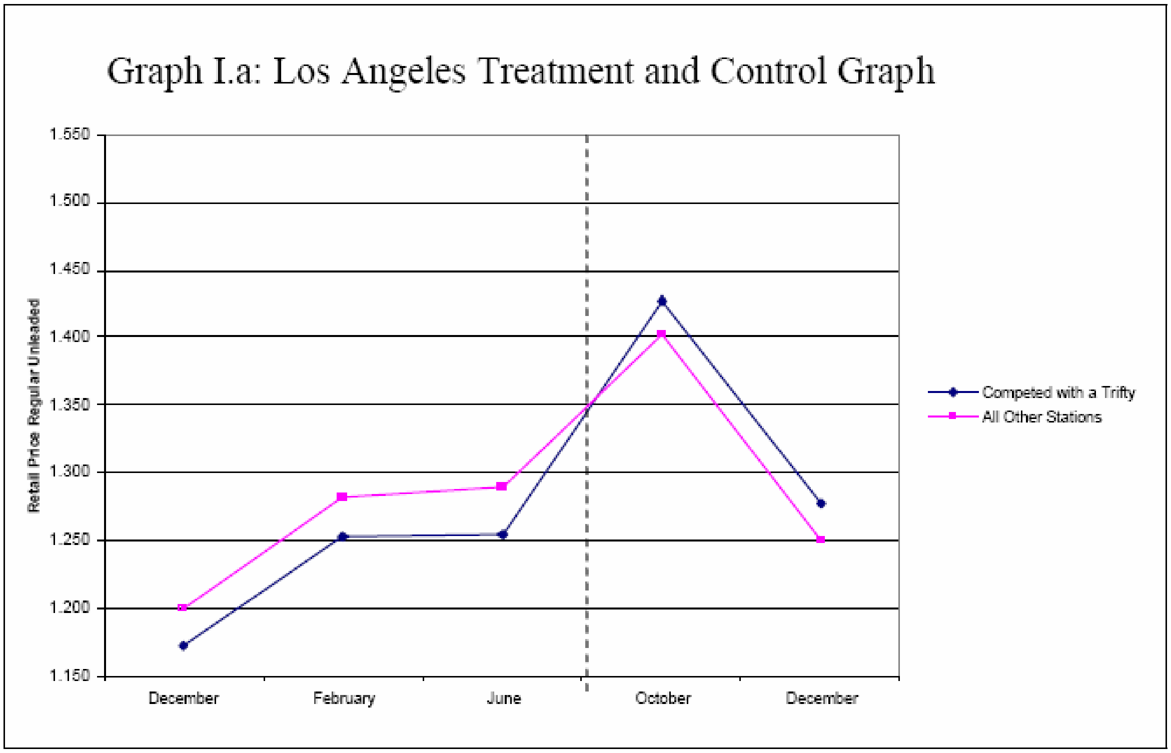


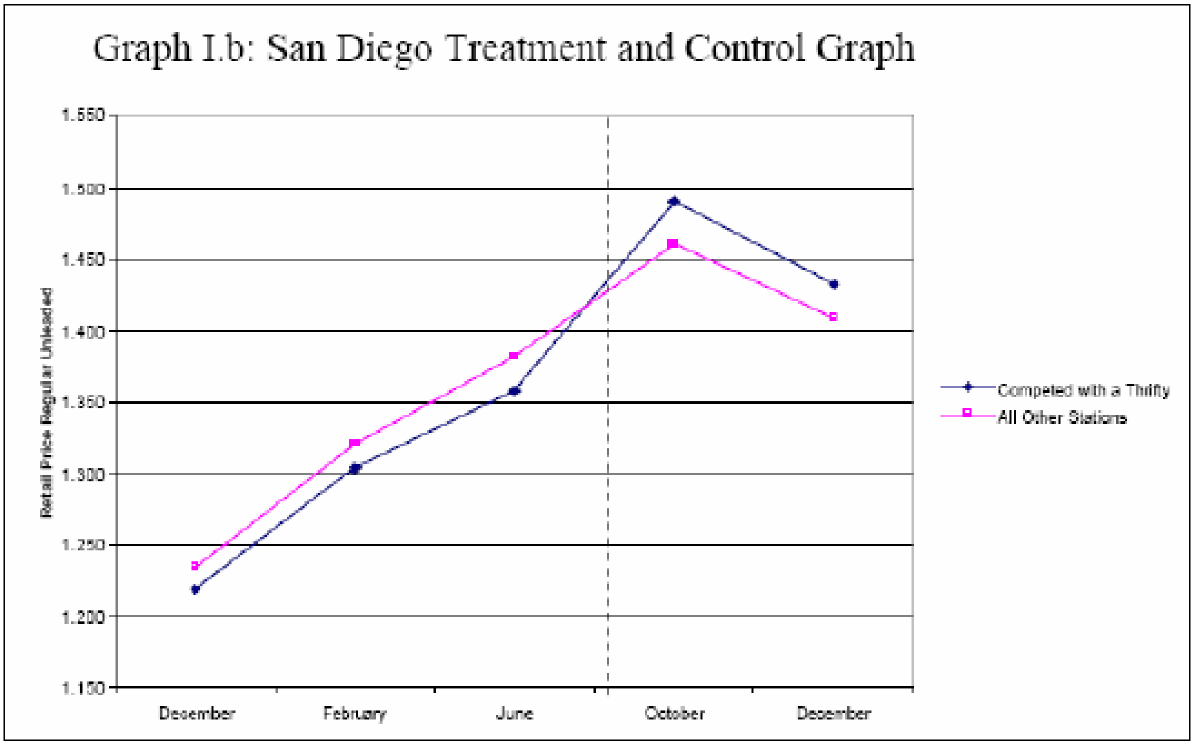
FIGURE 2. EMPLOYMENT IN NEW JERSEY AND PENNSYLVANIA FAST-FOOD RESTAURANTS, OCTOBER 1991 TO SEPTEMBER 1997

*Note:* Vertical lines indicate dates of original Card-Krueger survey and the October 1996 federal minimum-wage increase.
*Source:* Authors' calculations based on BLS ES-202 data.

Hastings 2004: Figure 1a



Hastings 2004: Figure 1b



Card 1992: Figure 4

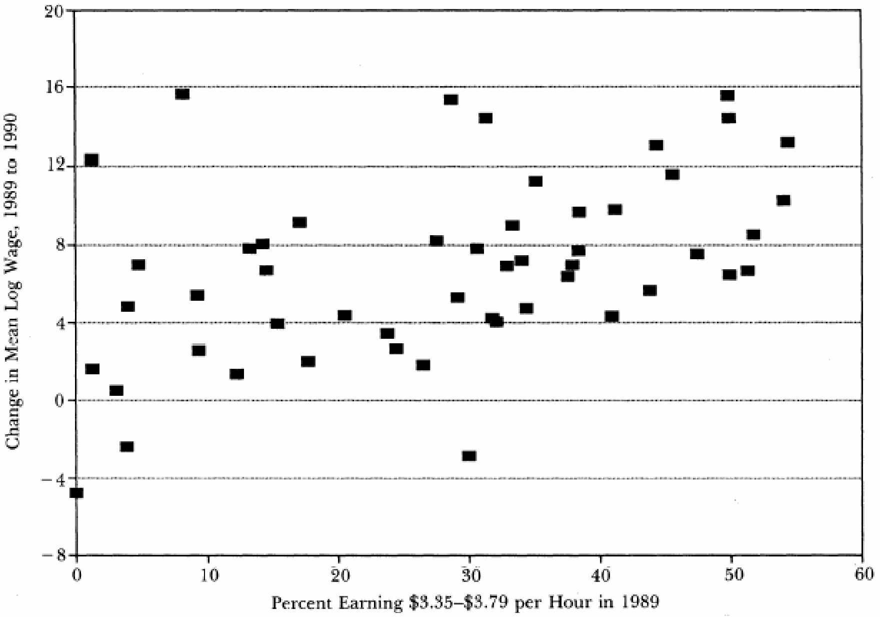


*Figure 4.* Change in Mean Log Wage of Teenage Workers Versus Percent Earning $3.35–$3.79 per Hour in 1989.

Card 1992: Figure 5

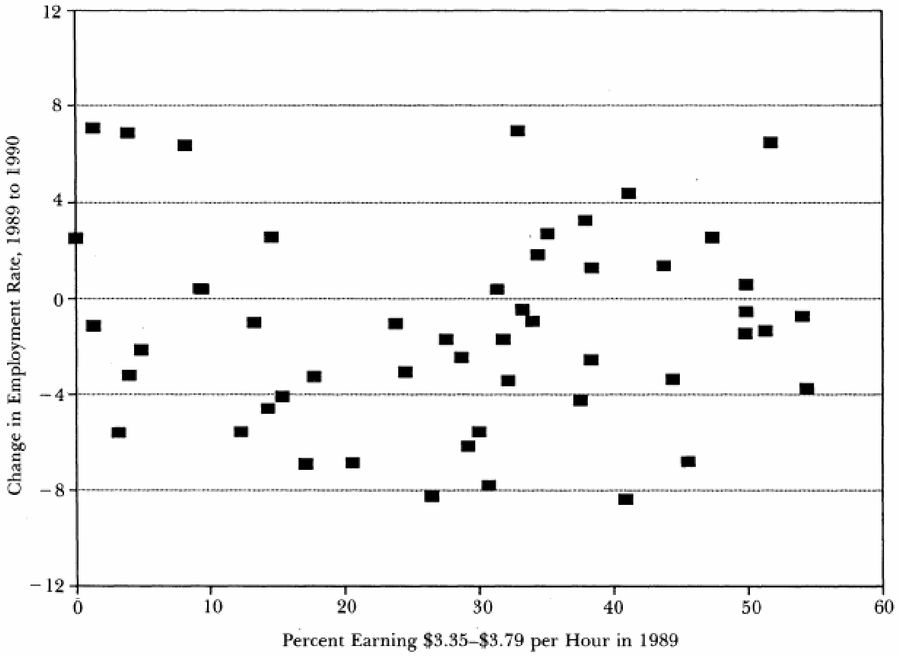


*Figure 5.* Change in Teenage Employment Rates Versus Percent Earning $3.35–$3.79 per Hour in 1989.

Card 1992: Table 3

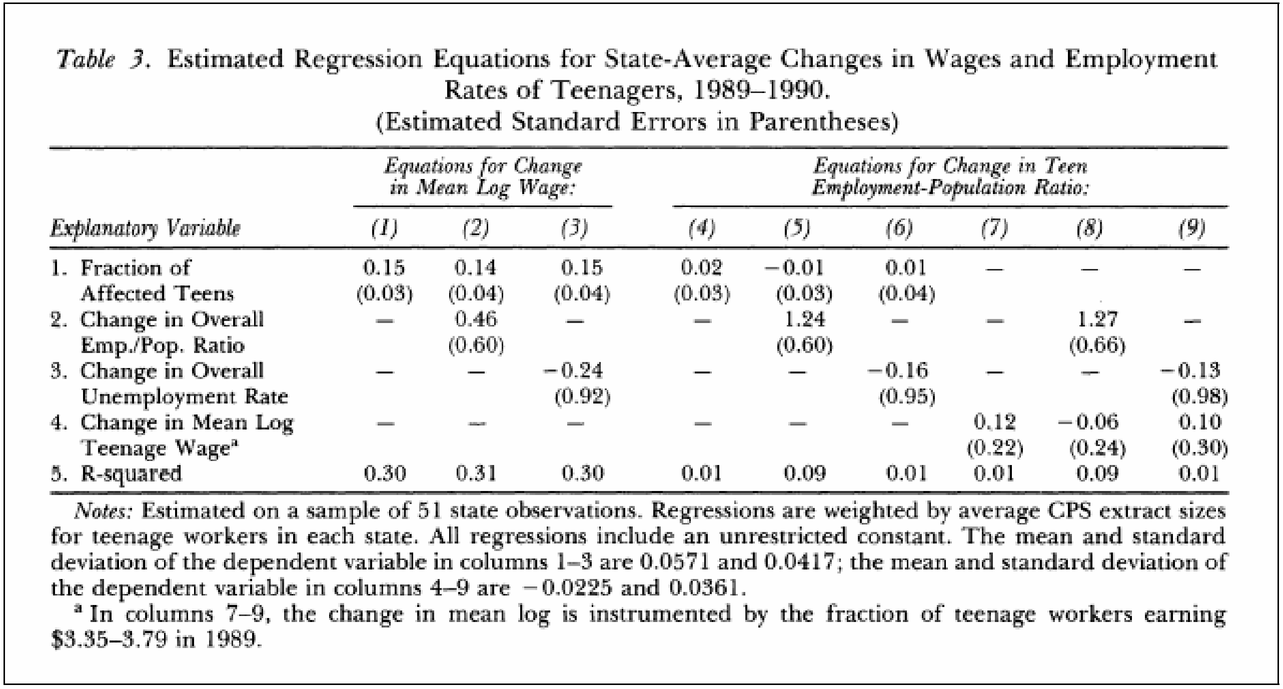


*Table 3.* Estimated Regression Equations for State-Average Changes in Wages and Employment Rates of Teenagers, 1989–1990.
(Estimated Standard Errors in Parentheses)

| | *Equations for Change in Mean Log Wage:* | | | *Equations for Change in Teen Employment-Population Ratio:* | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| *Explanatory Variable* | *(1)* | *(2)* | *(3)* | *(4)* | *(5)* | *(6)* | *(7)* | *(8)* | *(9)* |
| 1. Fraction of Affected Teens | 0.15<br>(0.03) | 0.14<br>(0.04) | 0.15<br>(0.04) | 0.02<br>(0.03) | −0.01<br>(0.03) | 0.01<br>(0.04) | — | — | — |
| 2. Change in Overall Emp./Pop. Ratio | — | 0.46<br>(0.60) | — | — | 1.24<br>(0.60) | — | — | 1.27<br>(0.66) | — |
| 3. Change in Overall Unemployment Rate | — | — | −0.24<br>(0.92) | — | — | −0.16<br>(0.95) | — | — | −0.13<br>(0.98) |
| 4. Change in Mean Log Teenage Wage[a] | — | — | — | — | — | — | 0.12<br>(0.22) | −0.06<br>(0.24) | 0.10<br>(0.30) |
| 5. R-squared | 0.30 | 0.31 | 0.30 | 0.01 | 0.09 | 0.01 | 0.01 | 0.09 | 0.01 |

*Notes:* Estimated on a sample of 51 state observations. Regressions are weighted by average CPS extract sizes for teenage workers in each state. All regressions include an unrestricted constant. The mean and standard deviation of the dependent variable in columns 1–3 are 0.0571 and 0.0417; the mean and standard deviation of the dependent variable in columns 4–9 are −0.0225 and 0.0361.

[a] In columns 7–9, the change in mean log is instrumented by the fraction of teenage workers earning $3.35–3.79 in 1989.

Autor 2003: Figure 3

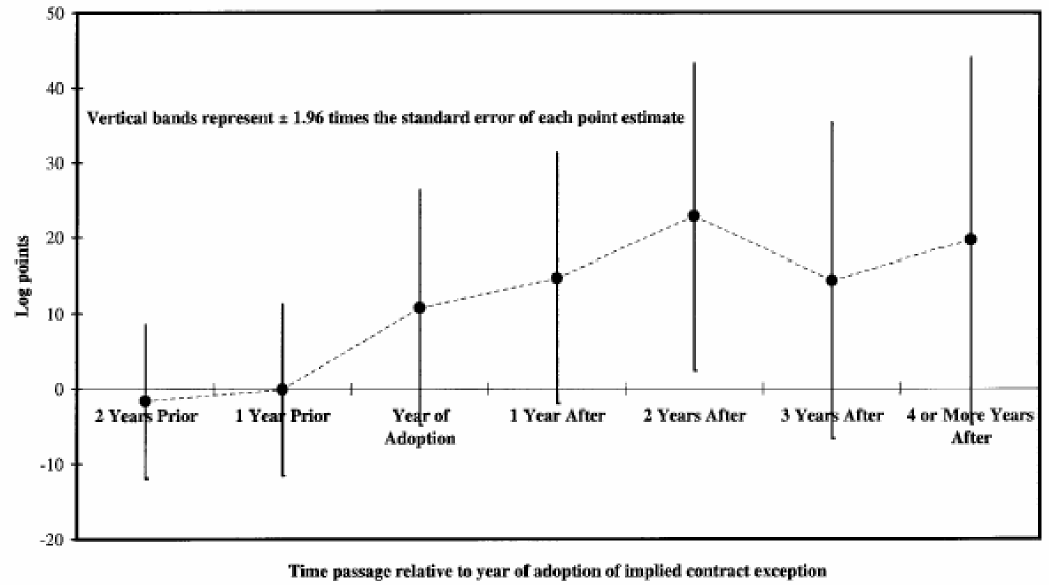


Fig. 3.—Estimated impact of implied contract exception on log state temporary help supply industry employment for years before, during, and after adoption, 1979–95.

Besley & Burgess 2004: Table 4

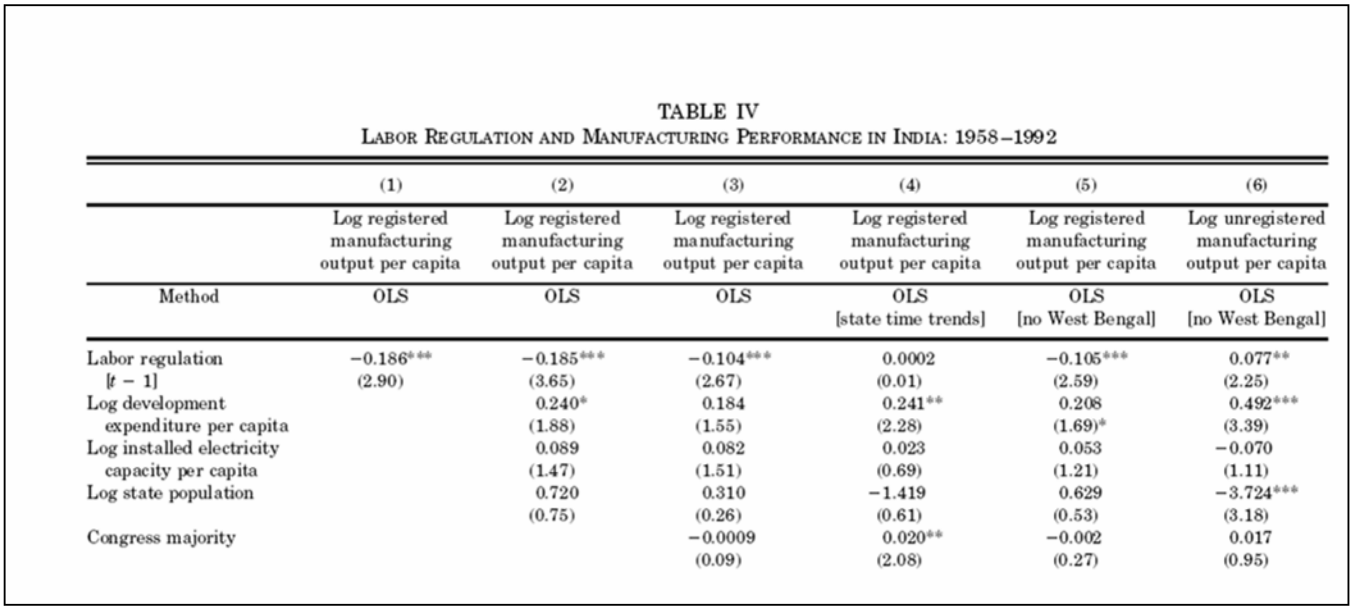
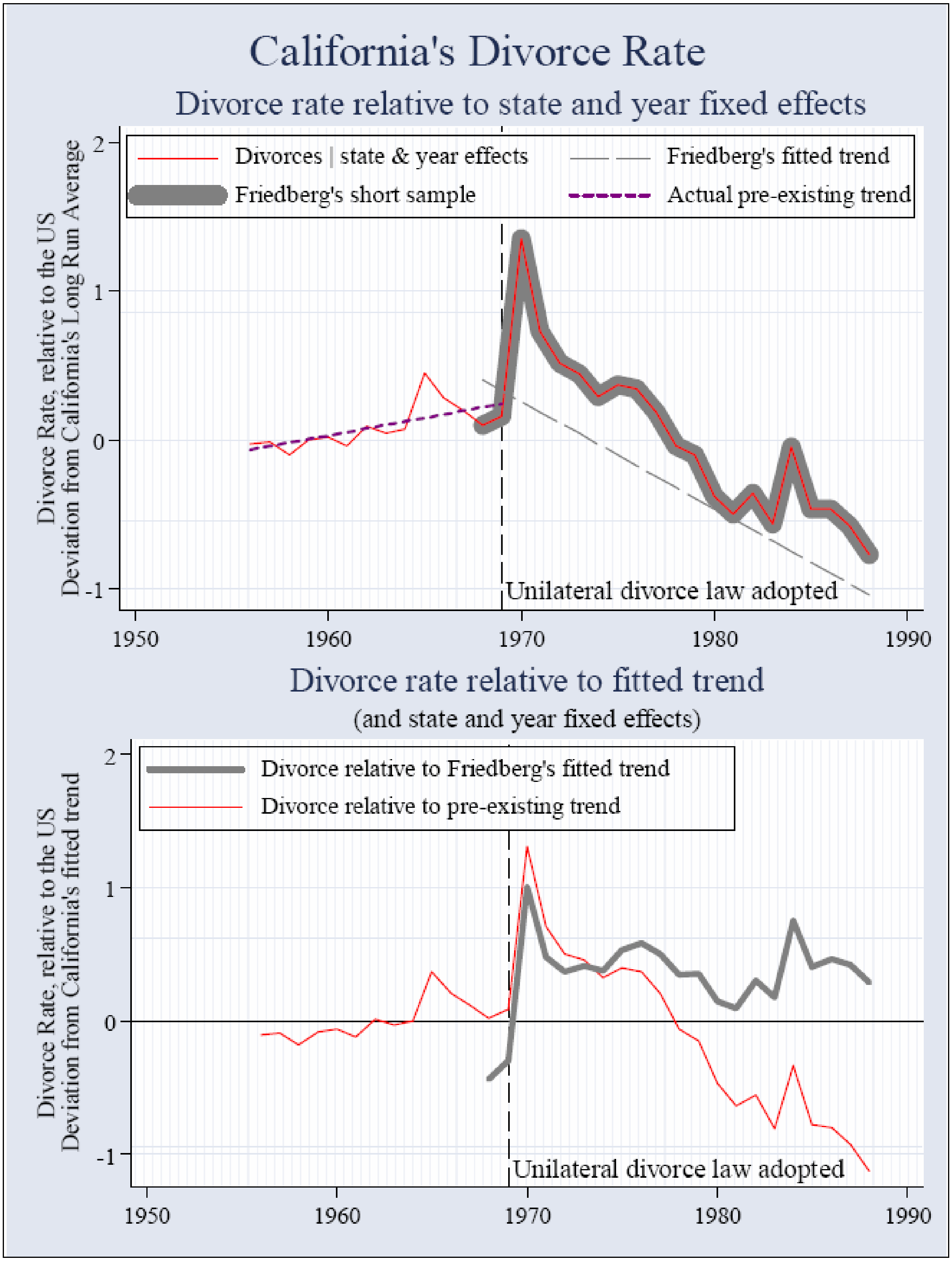


TABLE IV
LABOR REGULATION AND MANUFACTURING PERFORMANCE IN INDIA: 1958–1992

|  | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
|  | Log registered manufacturing output per capita | Log registered manufacturing output per capita | Log registered manufacturing output per capita | Log registered manufacturing output per capita | Log registered manufacturing output per capita | Log unregistered manufacturing output per capita |
| Method | OLS | OLS | OLS | OLS<br>[state time trends] | OLS<br>[no West Bengal] | OLS<br>[no West Bengal] |
| Labor regulation | −0.186*** | −0.185*** | −0.104*** | 0.0002 | −0.105*** | 0.077** |
| [t − 1] | (2.90) | (3.65) | (2.67) | (0.01) | (2.59) | (2.25) |
| Log development |  | 0.240* | 0.184 | 0.241** | 0.208 | 0.492*** |
| expenditure per capita |  | (1.88) | (1.55) | (2.28) | (1.69)* | (3.39) |
| Log installed electricity |  | 0.089 | 0.082 | 0.023 | 0.053 | −0.070 |
| capacity per capita |  | (1.47) | (1.51) | (0.69) | (1.21) | (1.11) |
| Log state population |  | 0.720 | 0.310 | −1.419 | 0.629 | −3.724*** |
|  |  | (0.75) | (0.26) | (0.61) | (0.53) | (3.18) |
| Congress majority |  |  | −0.0009 | 0.020** | −0.002 | 0.017 |
|  |  |  | (0.09) | (2.08) | (0.27) | (0.95) |

Wolfers 2003: Figure 5



California's Divorce Rate

Divorce rate relative to state and year fixed effects

Divorce rate relative to fitted trend

(and state and year fixed effects)

Davis & Weinstein 2002: Figure 2

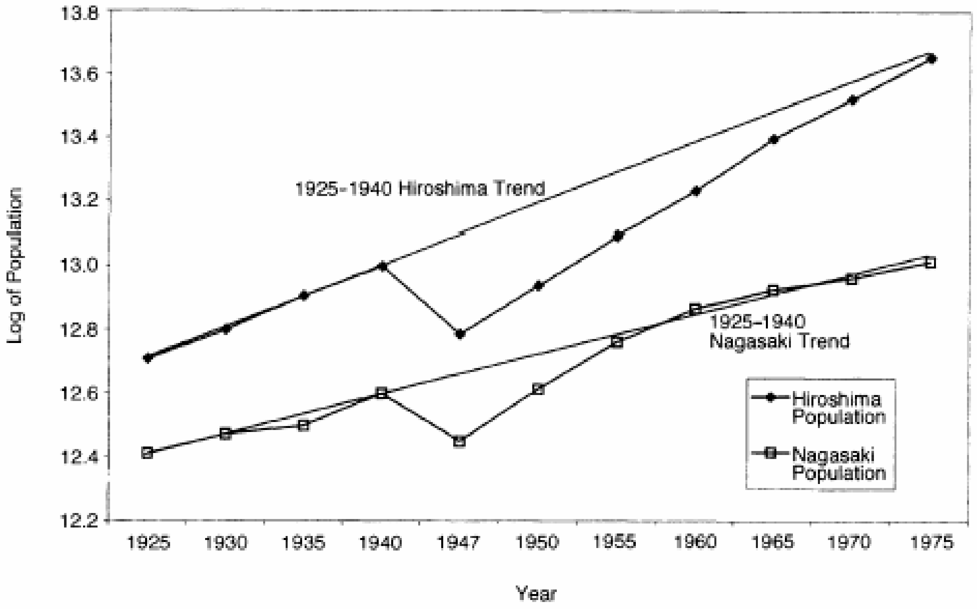


FIGURE 2. POPULATION GROWTH

Abadie & Garbeazabal 2003: Figure 1

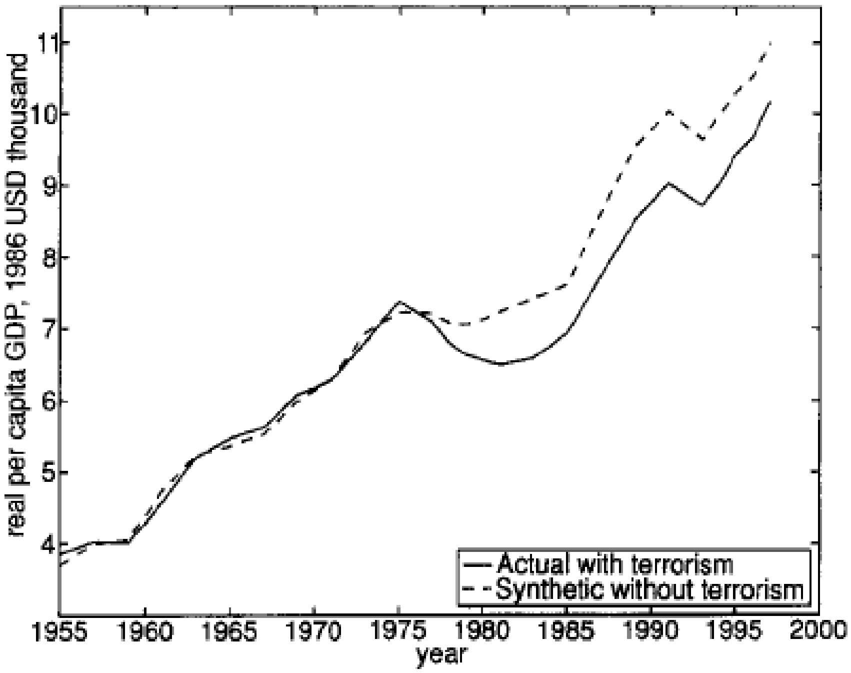


FIGURE 1. PER CAPITA GDP FOR THE BASQUE COUNTRY

Freeman 1984: Table 6

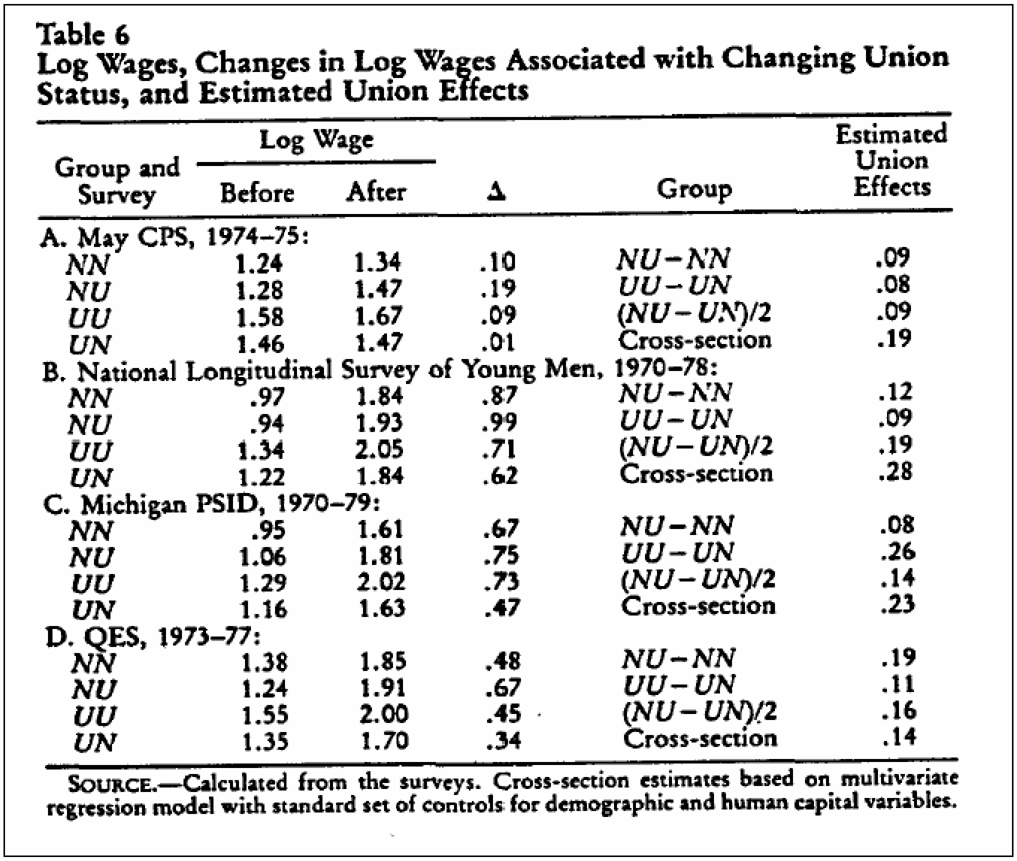


Table 6
Log Wages, Changes in Log Wages Associated with Changing Union Status, and Estimated Union Effects

| Group and Survey | Log Wage | | Δ | Group | Estimated Union Effects |
|---|---|---|---|---|---|
| | Before | After | | | |
| A. May CPS, 1974–75: | | | | | |
| *NN* | 1.24 | 1.34 | .10 | *NU–NN* | .09 |
| *NU* | 1.28 | 1.47 | .19 | *UU–UN* | .08 |
| *UU* | 1.58 | 1.67 | .09 | (*NU–UN*)/2 | .09 |
| *UN* | 1.46 | 1.47 | .01 | Cross-section | .19 |
| B. National Longitudinal Survey of Young Men, 1970–78: | | | | | |
| *NN* | .97 | 1.84 | .87 | *NU–NN* | .12 |
| *NU* | .94 | 1.93 | .99 | *UU–UN* | .09 |
| *UU* | 1.34 | 2.05 | .71 | (*NU–UN*)/2 | .19 |
| *UN* | 1.22 | 1.84 | .62 | Cross-section | .28 |
| C. Michigan PSID, 1970–79: | | | | | |
| *NN* | .95 | 1.61 | .67 | *NU–NN* | .08 |
| *NU* | 1.06 | 1.81 | .75 | *UU–UN* | .26 |
| *UU* | 1.29 | 2.02 | .73 | (*NU–UN*)/2 | .14 |
| *UN* | 1.16 | 1.63 | .47 | Cross-section | .23 |
| D. QES, 1973–77: | | | | | |
| *NN* | 1.38 | 1.85 | .48 | *NU–NN* | .19 |
| *NU* | 1.24 | 1.91 | .67 | *UU–UN* | .11 |
| *UU* | 1.55 | 2.00 | .45 | (*NU–UN*)/2 | .16 |
| *UN* | 1.35 | 1.70 | .34 | Cross-section | .14 |

SOURCE.—Calculated from the surveys. Cross-section estimates based on multivariate regression model with standard set of controls for demographic and human capital variables.

Freeman 1984: Table 1

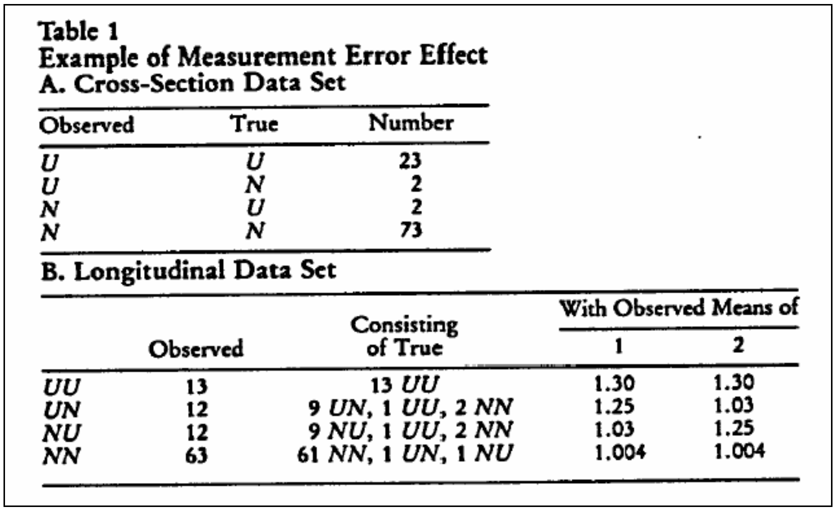


Table 1
Example of Measurement Error Effect
A. Cross-Section Data Set

| Observed | True | Number |
|---|---|---|
| *U* | *U* | 23 |
| *U* | *N* | 2 |
| *N* | *U* | 2 |
| *N* | *N* | 73 |

B. Longitudinal Data Set

| | Observed | Consisting of True | With Observed Means of | |
|---|---|---|---|---|
| | | | 1 | 2 |
| *UU* | 13 | 13 *UU* | 1.30 | 1.30 |
| *UN* | 12 | 9 *UN*, 1 *UU*, 2 *NN* | 1.25 | 1.03 |
| *NU* | 12 | 9 *NU*, 1 *UU*, 2 *NN* | 1.03 | 1.25 |
| *NN* | 63 | 61 *NN*, 1 *UN*, 1 *NU* | 1.004 | 1.004 |

Ashenfelter & Card 1985: Table 1

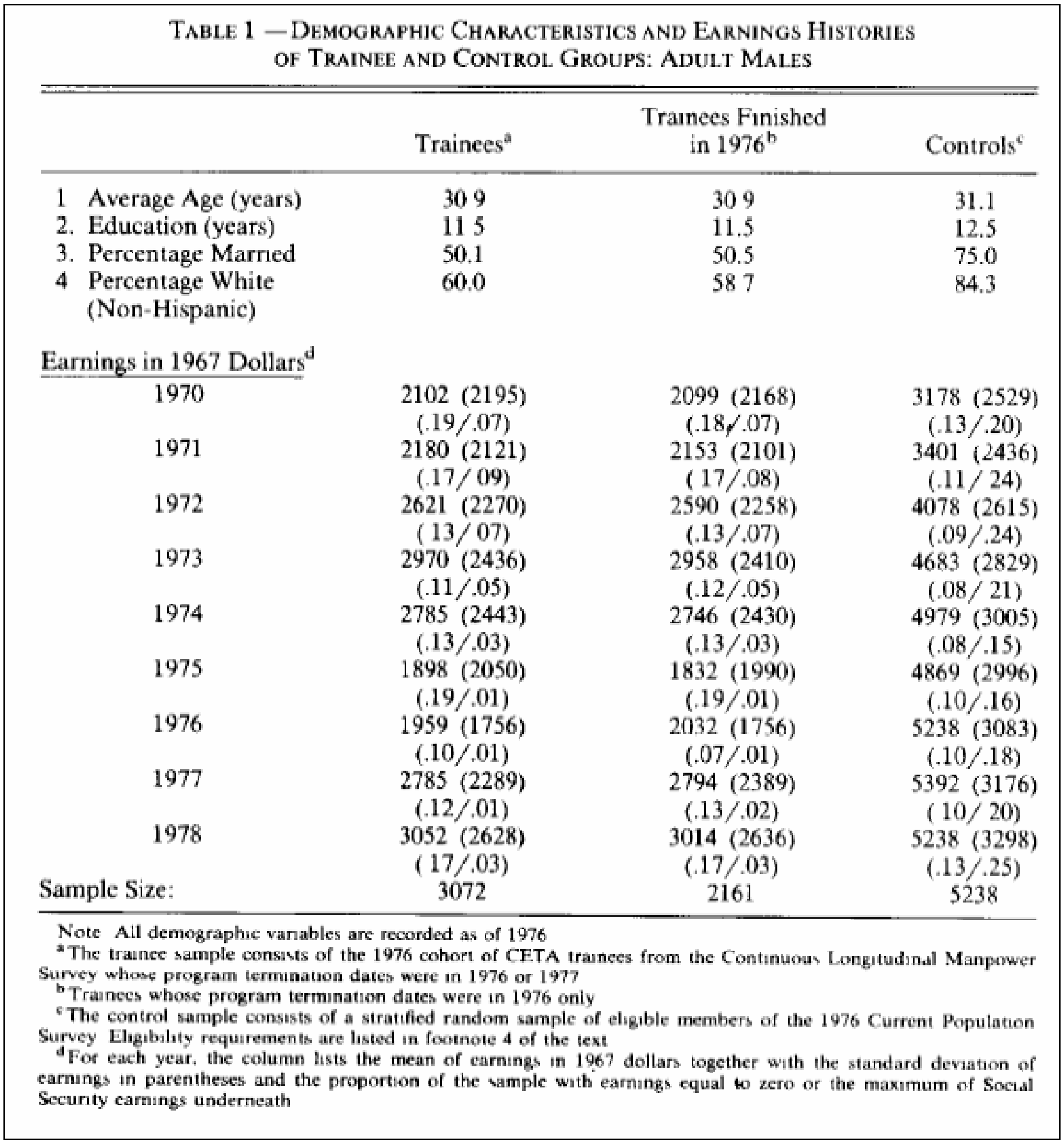


TABLE 1 — DEMOGRAPHIC CHARACTERISTICS AND EARNINGS HISTORIES OF TRAINEE AND CONTROL GROUPS: ADULT MALES

| | Trainees[a] | Trainees Finished in 1976[b] | Controls[c] |
|---|---|---|---|
| 1. Average Age (years) | 30.9 | 30.9 | 31.1 |
| 2. Education (years) | 11.5 | 11.5 | 12.5 |
| 3. Percentage Married | 50.1 | 50.5 | 75.0 |
| 4. Percentage White (Non-Hispanic) | 60.0 | 58.7 | 84.3 |
| Earnings in 1967 Dollars[d] | | | |
| 1970 | 2102 (2195)<br>(.19/.07) | 2099 (2168)<br>(.18/.07) | 3178 (2529)<br>(.13/.20) |
| 1971 | 2180 (2121)<br>(.17/.09) | 2153 (2101)<br>(.17/.08) | 3401 (2436)<br>(.11/.24) |
| 1972 | 2621 (2270)<br>(.13/.07) | 2590 (2258)<br>(.13/.07) | 4078 (2615)<br>(.09/.24) |
| 1973 | 2970 (2436)<br>(.11/.05) | 2958 (2410)<br>(.12/.05) | 4683 (2829)<br>(.08/.21) |
| 1974 | 2785 (2443)<br>(.13/.03) | 2746 (2430)<br>(.13/.03) | 4979 (3005)<br>(.08/.15) |
| 1975 | 1898 (2050)<br>(.19/.01) | 1832 (1990)<br>(.19/.01) | 4869 (2996)<br>(.10/.16) |
| 1976 | 1959 (1756)<br>(.10/.01) | 2032 (1756)<br>(.07/.01) | 5238 (3083)<br>(.10/.18) |
| 1977 | 2785 (2289)<br>(.12/.01) | 2794 (2389)<br>(.13/.02) | 5392 (3176)<br>(.10/.20) |
| 1978 | 3052 (2628)<br>(.17/.03) | 3014 (2636)<br>(.17/.03) | 5238 (3298)<br>(.13/.25) |
| Sample Size: | 3072 | 2161 | 5238 |

Note: All demographic variables are recorded as of 1976.

[a] The trainee sample consists of the 1976 cohort of CETA trainees from the Continuous Longitudinal Manpower Survey whose program termination dates were in 1976 or 1977.

[b] Trainees whose program termination dates were in 1976 only.

[c] The control sample consists of a stratified random sample of eligible members of the 1976 Current Population Survey. Eligibility requirements are listed in footnote 4 of the text.

[d] For each year, the column lists the mean of earnings in 1967 dollars together with the standard deviation of earnings in parentheses and the proportion of the sample with earnings equal to zero or the maximum of Social Security earnings underneath.

Ashenfelter and Card 1985: Table 2

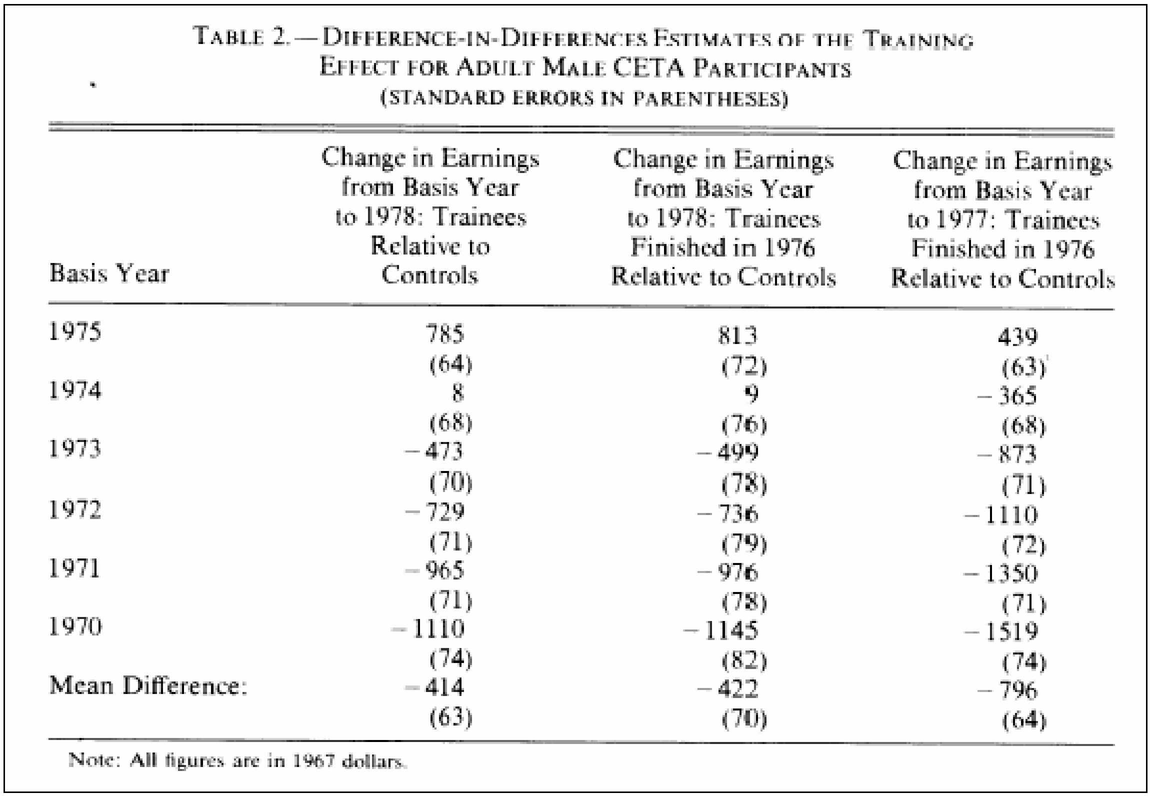


TABLE 2.—DIFFERENCE-IN-DIFFERENCES ESTIMATES OF THE TRAINING EFFECT FOR ADULT MALE CETA PARTICIPANTS (STANDARD ERRORS IN PARENTHESES)

| Basis Year | Change in Earnings from Basis Year to 1978: Trainees Relative to Controls | Change in Earnings from Basis Year to 1978: Trainees Finished in 1976 Relative to Controls | Change in Earnings from Basis Year to 1977: Trainees Finished in 1976 Relative to Controls |
|---|---|---|---|
| 1975 | 785 | 813 | 439 |
| | (64) | (72) | (63) |
| 1974 | 8 | 9 | −365 |
| | (68) | (76) | (68) |
| 1973 | −473 | −499 | −873 |
| | (70) | (78) | (71) |
| 1972 | −729 | −736 | −1110 |
| | (71) | (79) | (72) |
| 1971 | −965 | −976 | −1350 |
| | (71) | (78) | (71) |
| 1970 | −1110 | −1145 | −1519 |
| | (74) | (82) | (74) |
| Mean Difference: | −414 | −422 | −796 |
| | (63) | (70) | (64) |

Note: All figures are in 1967 dollars.