# INFERENCE WITH DIFFERENCE-IN-DIFFERENCES AND OTHER PANEL DATA

Stephen G. Donald and Kevin Lang*

*Abstract*—We examine inference in panel data when the number of groups is small, as is typically the case for difference-in-differences estimation and when some variables are fixed within groups. In this case, standard asymptotics based on the number of groups going to infinity provide a poor approximation to the finite sample distribution. We show that in some cases the $t$-statistic is distributed as $t$ and propose simple two-step estimators for these cases. We apply our analysis to two well-known papers. We confirm our theoretical analysis with Monte Carlo simulations.

## I. Introduction

MANY policy analyses rely on panel data in which the dependent variable differs across individuals, but at least some explanatory variables, such as the policies being studied, are constant among all members of a group. For example, in the typical difference-in-differences model, we regress outcomes at the individual level (for example, employment in a firm in state $s$ in year $t$) on a policy that applies to all individuals in the group (for example, the minimum wage in state $s$ in year $t$). Moulton (1990) shows that in regression models with mixtures of individual and grouped data, the failure to account for the presence of common group errors can generate estimated standard errors that are biased downward dramatically.[1] The difference-in-differences estimator is a special case of this model.

Researchers use a number of standard techniques to adjust for common group effects:

- random-effects feasible GLS estimation, which under certain conditions is asymptotically efficient,
- correcting the standard errors using the error covariance matrix based on common group errors as in Moulton,
- correcting the standard errors using a robust covariance estimator according to a formula developed by Liang and Zeger (1986) and more commonly known as the Stata cluster command.

This paper makes two simple but, we believe important, points. First, when applied to variables that are constant within a group, the $t$-statistics generated using each of these techniques for correcting for common group errors are asymptotically normally distributed only as the number of groups goes to infinity.

Second, under standard restrictions, the efficient estimator can be implemented by a simple two-step procedure, and the resulting $t$-statistic may have, under restrictions on the distribution of the group-level error, an asymptotic $t$-distribution as the number of observations per group goes to infinity. In addition, under more restrictive assumptions, when the same procedure is used in finite samples, the $t$-statistics have a $t$-distribution.

Consequently, standard asymptotics cannot be applied when the number of groups is small, as in the case where we compare two states in two years, two cities over a small number of years, or self-employed workers and employees over a small number of years. In such cases, failing to take account of the group-error structure will not only generate underestimates of the standard errors as in Moulton, but applying the normal distribution to corrected $t$-statistics will dramatically overstate the significance of the statistics. Standard asymptotics should apply to comparisons across all fifty states, although other problems may arise in common time series/cross-section estimates based on states and using long panels (see Bertrand, Duflo, & Mullainathan, 2004).[2]

In the next section, we first present an intuitive argument and then formalize the conditions under which we can derive the distribution of the $t$-statistic when the number of groups is small. Readers who are not interested in the details can skip the later part of this section and proceed to the third section, where we discuss the common two-group/two-period case and also apply our approach to two influential papers: the Gruber and Poterba (1994) paper on health insurance and self-employment and Card's (1990) study of the Mariel boatlift. We show that analyzing the $t$-statistic, taking into account a possible group-error component, dramatically reduces our estimate of the precision of their results.

In the fourth section, we consider two other approaches to common group errors, the Moulton correction and the commonly applied Stata cluster correction. In the fifth section, we present Monte Carlo evidence regarding the distribution of the $t$-statistic using a variety of estimators. Our results indicate that the $t$-statistics produced by standard estimators have distributions that differ quite substantially from both the normal and the $t$-distributions. However, when the theory predicts that they should, the two-step estimators we

Received for publication October 29, 2004. Revision accepted for publication March 2, 2006.

* University of Texas at Austin; and Boston University and NBER, respectively.

This paper was written in part while Lang was visiting the Massachusetts Institute of Technology. We are grateful to them for their hospitality and to Josh Angrist, Eli Berman, George Borjas, David Card, Jon Gruber, Larry Katz, Alan Krueger, and participants in workshops at Boston University and MIT for helpful comments. Donald thanks the Sloan Foundation and the National Science Foundation (SES-0196372) for research support. Lang thanks the National Science Foundation for support under grant SES-0339149. The usual caveat applies.

[1] See also Kloek (1981), who considered the bias in standard errors, as well as the relationship between OLS and GLS in the case where only group-level regressors are present.

[2] For other recent work related to difference-in-differences methodology, see Abadie (2005) and Athey and Imbens (2005).

*The Review of Economics and Statistics*, May 2007, 89(2): 221–233

propose produce $t$-statistics with approximately a $t$-distribution with degrees of freedom equal to number of groups minus number of group-constant variables. Moreover, one of the two-step estimators we consider appears to be reasonably robust to the departures from the assumptions needed to guarantee that the $t$-statistic has a $t$-distribution.

## II. The Error Components Model with a Small Number of Groups

We begin with a standard time series/cross-section model of the form

$$Y_{is} = a + X_s\beta + Z_{is}\gamma + \alpha_{is} + \varepsilon_{is}, \tag{1}$$

where $\alpha_{is}$ is an error term that is correlated within group $s$, and $\varepsilon_{is}$ is an individual-specific term that is independent of the other errors. With a single cross section, $Y$ might be income, $X$ state laws, and $Z$ characteristics of individuals. In this case it would be natural to follow Moulton and assume that $\alpha$ is a state effect that does not vary among members in a group, that is that $\alpha_{is} = \Sigma_i\alpha_{is}/N_s \equiv \alpha_s$. We do not require that the error term take the Moulton structure, only that the $\sigma_\alpha^2$ (the variance of $\alpha_s$) depend only on the number of observations from group $s$ and that, as group size gets large, it converge in probability to some finite value.[3]

If the covariance matrix of the error term is known, GLS estimation of equation (1) is efficient. With some regularity conditions, feasible GLS is efficient if the covariance matrix can be estimated consistently. Depending on the structure of the covariance matrix, GLS can be computationally burdensome. Moreover, if the exact structure of the dependence is unknown, GLS estimation may be infeasible.

Estimating $\beta$ in two stages is often computationally simpler. In this case, we use OLS to estimate

$$y = Z\gamma + W\Gamma + \varepsilon, \tag{2}$$

where $W$ is a set of dummy variables indicating group membership. Note that

$$\Gamma = X\beta + \alpha. \tag{3}$$

We then can use the estimated $\hat{\Gamma}$ in GLS estimation of

$$\hat{\Gamma}_s = X_s\beta + \alpha_s + (\hat{\Gamma}_s - \Gamma_s), \tag{4}$$

where the error term has variance $\sigma_\alpha^2 I + \text{var}(\hat{\Gamma})$.

Amemiya (1978) shows that if the covariance matrices of $\alpha$ and $\varepsilon$ are known, then the two-step procedure and the GLS procedure applied directly to equation (1) are numerically identical. If instead feasible GLS is used, then provided the covariance terms are estimated in the same fashion, numerical equivalence continues to hold. More commonly, the two approaches lend themselves to different methods for obtaining consistent estimates of the covariance terms. If so, the equivalence is asymptotic rather than numeric.

Our contribution is twofold. First, we can see from equation (4) that if the number of groups is small, then it is not possible to rely on the consistency of estimates of $\sigma_\alpha^2$ to justify feasible GLS estimation of equation (4). However, if $\sigma_\alpha^2 I + \text{var}(\hat{\Gamma})$ is homoskedastic and diagonal, by the usual arguments, it is still possible to obtain an unbiased estimate of the variance of the error term in equation (4). Under normality assumptions on $\alpha_s$, the resulting $t$-statistic will have a $t$-distribution rather than a normal distribution. We explore circumstances under which the assumption of homoskedasticity is reasonable. In particular, the error term will be homoskedastic under at least two circumstances:

(i) if the number of observations per group is large, or
(ii) if there are no within-group varying characteristics and the number of observations is the same for all groups.

Second, when the error term in equation (4) is homoskedastic, by standard theorems, OLS estimation of equation (4) is efficient. Since OLS estimation of equation (4) is numerically equivalent to feasible GLS estimation of equation (1), we have full efficiency of estimation even when the number of groups is small.

We begin our formal treatment with the case where all variables are fixed within groups.

### A. *Only Within-Group-Constant Explanatory Variables*

We begin by treating the case where $X_s$ is a scalar and there are no within-group varying explanatory variables ($\gamma = 0$), so that

$$Y_{is} = a + X_s\beta + \alpha_{is} + \varepsilon_{is}. \tag{5}$$

This case provides much of the intuition for the more general case.[4]

Throughout we will assume that the $\varepsilon_{is}$ are independent of each other and of $\alpha$ for all $i$ and $s$. We further assume that $\alpha_{is}$ and $\alpha_{js}{}'$ are independent for $s \neq s'$, but do not assume that $\alpha_{is}$ and $\alpha_{js}$ are uncorrelated.

The two-step estimator in this case has a very simple interpretation. The first stage is equivalent to taking group means,

$$\hat{d}_s = \frac{\sum_{i=1}^{N_s} Y_{is}}{N_s}, \tag{6}$$

so that the second stage becomes

[3] Thus in a time series/cross-section context, we might have $\alpha_{t,s} = \rho\alpha_{t-1,s} + \mu_{t,s}$ with $0 < \rho < 1$, which would satisfy this requirement.

[4] Kloek (1981) provides some analysis of this model when the residual in the model is equicorrelated.

$$\hat{d}_s = \bar{Y}_s = a + X_s\beta + \frac{\sum_{i=1}^{N_s}\alpha_{is}}{N_s} + \frac{\sum_{i=1}^{N_s}\varepsilon_{is}}{N_s} \quad (7)$$

$$\equiv a + X_s\beta + \alpha_s + \varepsilon_s \quad (8)$$

$$\equiv a + X_s\beta + \eta_s, \quad (9)$$

which is just the "between-groups" estimator of $\beta$.

A few points follow immediately from the equivalence of GLS estimation of equations (5) and (9).

(i) $\beta$ can always be estimated efficiently by appropriate weighted least squares estimation of equation (9) if the weights are known or by feasible weighted least squares if they can be estimated consistently.

(ii) If either $\eta$ is homoskedastic or var($\eta_s$) is uncorrelated with $X_s$, then the efficient estimator is the unweighted between estimator. Note that homoskedasticity is a natural assumption either when all groups have the same number of observations ($N_s = N, \forall_S$) or when the number of observations in each group is large.

The latter point demonstrates that in many circumstances unweighted between-group estimation is the most efficient estimator[5] and that this efficient estimator can be achieved without knowledge of the exact covariance structure of $\alpha$, although as noted this does require that the variance of $\eta_s$ is constant across groups.

*Inference:* It should be apparent that our ability to perform inference on $\hat{\beta}$ depends primarily on $S$ and not on $N_s$. If the number of groups is large, then the standard theorems establish that when $\eta$ is homoskedastic, $\hat{\beta}_{ols}$ is normally distributed and the $t$-statistic follows the normal distribution. When $\eta$ is heteroskedastic, the same is true either for feasible GLS or for appropriately calculated standard errors.

In many cases it will be natural to treat $S$ as large and the error term as homoskedastic. For example, studies that use difference in laws across states and have large samples for all fifty states are likely to meet this requirement approximately. However, in many applications the number of groups is small. The well-known Card and Krueger (1994) minimum-wage study is a case in point in which there is a large number of observations per group but only four groups (New Jersey before and after the law and eastern Pennsylvania before and after the law). Other studies (Gruber & Poterba, 1994; Card, 1990; Eissa & Liebman, 1996) are based on a small number of group/year cells.

When the number of groups is small, in order to have a standard solution for the distribution of the $t$-statistic, we require that $\eta$ be i.i.d. normally distributed. Below, we present formal sufficient conditions for this requirement to be satisfied.

If the distribution of $\eta$ is i.i.d. normal, then it follows from standard theorems that the $t$-statistic for $\hat{\beta}$ has a $t$-distribution with $S - 2$ degrees of freedom. In effect, failure to recognize that the variance of the error term is estimated using very few observations can dramatically overstate the significance of findings.

Why does the distribution of $\hat{T}$ remain $t$ despite the large number of observations? The answer is quite intuitive. If we relied on published census data to estimate a relation based on the New England states, we would automatically assume that the resulting $t$-statistic had a $t$-distribution. Relying on the underlying individual data cannot help us if all of the information in the data is included in the mean.

Somewhat more formally, rewrite equation (9) as

$$\tilde{Y}_s = \tilde{X}_s\beta + \tilde{\eta}_s, \quad (10)$$

where ~ denotes a deviation from the mean. The usual $t$-statistic for hypotheses concerning $\beta$ is given by

$$\hat{T} = \frac{\hat{\beta} - \beta}{\hat{\sigma}_\eta(\sum_s \tilde{X}_s^2)^{1/2}}, \quad \hat{\sigma}_\eta^2 = \frac{1}{S-2}\sum_{s=1}^{S}(\tilde{Y}_s - \tilde{X}_s\hat{\beta})^2. \quad (11)$$

Given this fact, we can easily see that it will be reasonable to use a $t(S - 2)$ distribution for conducting inference whenever $\eta_s$ is exactly or is approximately a homoskedastic normal random variable.

**Finite Sample Result:** Here for the $\hat{T}$ statistic to have an exact $t(S - 2)$ distribution it is sufficient that

$$\eta_s = \frac{\sum_{i=1}^{N_s}\alpha_{is}}{N_s} + \frac{\sum_{i=1}^{N_s}\varepsilon_{is}}{N_s} \sim N(0, \sigma_\eta^2),$$

where it is important that $\sigma_\eta^2$ is constant across $s$. Although there may be a variety of conditions, the most obvious case is where $\alpha_{is} = \alpha_s \sim N(0, \sigma_\alpha^2)$ for all $i$, $\varepsilon_{is} \sim N(0, \sigma_\varepsilon^2)$ and $N_s = N$ for all $s$ so that $\eta_s \sim N(0, \sigma_\eta^2)$, where

$$\sigma_\eta^2 = \sigma_\alpha^2 + \frac{\sigma_\varepsilon^2}{N}. \quad (12)$$

This is the standard random-effects time series/cross-section model as well as the specification used by Moulton. This includes the possibility that there are no group-specific effects.

**Large Sample Result:** For the $\hat{T}$ statistic to have a distribution that is well approximated by $t(S - 2)$, it is sufficient that there be large $N_s$ and that

$$\eta_s \overset{A}{\sim} N(0, \sigma_\eta^2), \quad (13)$$

[5] This point is made in a somewhat different context by Dickens (1990).

which requires some form of asymptotic theory regarding $\eta_s$ with $N_s \to \infty$ but with the number of groups fixed. Here there are at least two interesting possibilities.

(i) For each $s$, $\alpha_{is} = \alpha_s \sim N(0, \sigma_\alpha^2)$ for all $i$ and $\varepsilon_{is}$ satisfy conditions for a Law of Large Numbers to imply that

$$p\lim_{N_s \to \infty} \frac{\sum_{i=1}^{N_s} \varepsilon_{is}}{N_s} = 0.$$

In this instance

$$p\lim_{N_s \to \infty} \eta_s = \alpha_s \sim N(0, \sigma_\alpha^2)$$

so that condition (13) is met. This does not require that $N_s$ be the same in all groups, but for the approximation to be valid we would need all groups to have large $N_s$ so that (14) is approximately true.

(ii) For each $s$, $\alpha_{is} = \alpha_s \sim N(0, \sigma_\alpha^2/N_s)$ for all $i$, $\varepsilon_{is}$ satisfy conditions for a Central Limit Theorem, and $N_s/N_t \to 1$ for any $s \neq t$ then,

$$\eta_s \overset{A}{\sim} N(0, (\sigma_\alpha^2 + \sigma_\varepsilon^2)/N_s),$$

where the variance asymptotically does not depend on $s$. This case is possible under a parameter sequence that keeps $\sigma_\alpha^2$ of the same (or smaller) order as $\sigma_\varepsilon^2/N_s$, as might be appropriate when the group effects are relatively small. Indeed, this includes the possibility that $\alpha_s = 0$ so that there are *no* group-specific effects. Also note that in this case we do not require normality for the $\varepsilon_{is}$, but for the approximation to be a good one we require that there be similar numbers of observations per group.

There may be other possibilities in the large sample case, although we refrain from giving more explicit conditions. For instance, provided that one can show that if $(\sum_i \alpha_{is})/N_s$ is approximately normal (with constant variance independent of $s$ asymptotically) and that $\varepsilon_{is}$ satisfy conditions for a Law of Large Numbers, then the condition in (13) will hold and using a $t(S-2)$ will provide a good approximation. There are a variety of possible assumptions one could use to obtain the approximate normality of $(\sum_i \alpha_{is})/N_s$ that relate to the dependence across observations within a group—the simplest case is given in (i) above, but it is apparent that the two-stage technique can accommodate inference even when the nature of the dependence within groups is unknown, provided that there is no correlation of errors across groups.[6] Also, as in case (ii), for this result to hold one would require $N_s$ to be (asymptotically) constant across groups.

It is worth noting that estimating the between-group estimator is a matter of convenience. When $N_s = N$, the between-group and OLS estimators are identical. If $N_s/N_t \to 1 \forall_s, t$, then OLS converges to the between-groups estimator. Therefore, it is possible to calculate a corrected standard error for the OLS estimate and generate a $t$-statistic that has a $t$-distribution. The between-groups estimator is, however, much more convenient.

[6] For more on this see Andrews (2005) and references therein.

### B. *Variables that Vary Within Group*

We consider now hybrid models in which some variables differ across observations within groups. For simplicity we ignore complications associated with nonconstant correlation of the group effect since they do not add to the analysis. As was seen in the discussion without within-group varying variables, all that is important is the variance of the mean group error.

Thus we analyze the standard Moulton model

$$Y_{is} = a + X_s\beta + Z_{is}\gamma + \alpha_s + \varepsilon_{is}. \tag{15}$$

We assume for simplicity that $Y_{is}$ and $X_s$ are scalars. The extension to the case of more than one group-varying or group variable is straightforward.

We further assume that

$$\alpha_s \sim N(0, \sigma_a^2) \quad \text{for all } s,$$

$$\varepsilon_{is} \sim N(0, \sigma_\varepsilon^2) \quad \text{for all } i, s,$$

and that these residuals are mutually independent for all $i$ and $s$.

As before, we know that GLS estimation of equation (15) is efficient and that if we can obtain consistent estimates of $\sigma_\alpha^2$ and $\sigma_\varepsilon^2$, feasible GLS estimation is asymptotically efficient. Finally, we know from Amemiya (1978) that if these variance terms are estimated in the same way, feasible GLS estimation is numerically identical to the following estimator—first use OLS to estimate the "within-group" estimate of $\gamma$

$$Y_{is} = d_s + Z_{is}\gamma + \varepsilon_{is}. \tag{16}$$

Then estimate $\beta$ by feasible GLS estimation of

$$\bar{Y}_s - \bar{Z}_s\bar{\gamma} = \hat{d}_s = a + \beta X_s + u_s, \tag{17}$$

where

$$\bar{Y}_s = \frac{\sum_{i=1}^{N_s} Y_{is}}{N_s}, \quad \bar{Z}_s = \frac{\sum_{i=1}^{N_s} Z_{is}}{N_s},$$

$\hat{d}_s$ is the estimate of $d_s$ in equation (16), and

$$\text{var}(u) = \sigma_\alpha^2 I + \Sigma_{\hat{d}},$$

where $\Sigma_{\hat{d}}$ is the covariance matrix of the fixed-effect parameter estimates.

Note that estimates of $\Sigma_{\hat{d}}$ can be obtained by selecting the covariance matrix corresponding to the fixed effects. $\sigma^2$ can be estimated by first estimating equation (17) by OLS and then using

$$\frac{\sum_{i=1}^{S} \hat{u}_i^2}{S-K} - \frac{1}{S}\sum_{i=1}^{S} \text{var}(\hat{d}_i) = \hat{\sigma}^a,$$

where $\text{var}(\hat{d}_i)$ is the variance of the $i$th fixed effect and $K$ is the number of explanatory variables in equation (17).

Finally, it is worth noting that since the groups are distinct, covariance of the $\hat{d}$'s arises only because $\gamma$ is estimated rather than known. If each $\hat{\gamma}_i$ is calculated from a separate sample, the covariances will be zero. Thus, the covariance problem can be avoided by estimating $\gamma$ for the $s$th state as

$$\hat{\gamma}(s) = (Z_s' M_s Z_s)^{-1} Z_s' M_s Y_s, \tag{18}$$

where

$$M_s = I_{N_s} - \iota_{N_s}(\iota_{N_s}'\iota_{N_s})^{-1}\iota_{N_s}',$$

and $t_{N_s}$, is a vector of 1s of length $N_s$.

Since $\hat{\gamma}_W$ involves the restriction that $\gamma$ is constant across states, it is more efficient but less robust.[7] Our focus, however, is not on estimation of $\gamma$ but of $\beta$.

When the number of groups is large, it is possible to estimate equation (17) by feasible GLS. If the number of groups is small, then we can still get efficient estimates and determine the distribution of the $t$-statistic if the error term in equation (17) is i.i.d. normal, except possibly for an error term common to all groups. The following propositions summarize conditions under which this condition holds and the resulting $t$-statistic has the $t$-distribution.

Let $\hat{T}_1$ be the $t$-statistic using $\hat{\gamma}_w$ and $\hat{T}_2$ be the $t$-statistic using $\hat{\gamma}(s)$.

**Proposition 1:** If the $\varepsilon_{is}$ are normally distributed,

(i) $\hat{T}_1 \sim t(S-2)$ when $N_s$ are identical for all $s$ and either (a) there are no $Z_{is}$ or (b) $\bar{Z}_s$ is constant across $s$.

(ii) $\hat{T}_2 \sim t(S-2)$ when $N_s$ are identical for all $s$ and either (a) there are no $Z_{is}$ or (b) $\bar{Z}_s'(Z_{is}' M_s Z_s)^{-1}\bar{Z}_s$ is constant across $s$.

We can also show that the statistics will have asymptotic $t(S-2)$ distributions under more general situations so that the $t(S-2)$ distribution can be used quite generally.

**Proposition 2:** If the $\varepsilon_{is}$ are not normally distributed and if $\sigma_\alpha^2$ is fixed, then $\hat{T}_j \overset{A}{\sim} t(S-2)$ (for $j = 1, 2$) when $N_s \to \infty$ for all $s$.

As in the case where there are no individual-specific covariates, one can show that the statistics $\hat{T}_j$ will be approximately $t(S-2)$ when the group-specific errors are small relative to the idiosyncratic errors.

The conditions that give rise to this are essentially asymptotic analogs of the conditions in proposition 1.

**Proposition 3:** If $\sigma_\alpha^2$ is small in the sense that $\sigma_\alpha^2 = O(N_s^{-1/2})$, then regardless of the distribution of $\varepsilon_{is}$

(i) $\hat{T}_1 \overset{A}{\sim} t(S-2)$ when $N_s$ are asymptotically identical[8] for all $s$ and either (a) there are no $Z_{is}$ or (b) $p\lim \bar{Z}_s$ is constant across $s$.

(ii) $\hat{T}_2 \overset{A}{\sim} t(S-2)$ when $N_s$ are asymptotically identical for all $s$ and either (a) there are no $Z_{is}$ or (b) $p\lim \bar{Z}_s'(Z_s' M_s Z_s/N_s)^{-1}\bar{Z}_s$ is constant across s.

These situations can be stated with reference to the residuals in the second-stage equation:

$$\bar{Y}_s - \bar{Z}_s'\hat{\gamma} = a + \beta X_s + \alpha_s + \bar{\varepsilon}_s + \bar{Z}_s'(\gamma - \hat{\gamma}). \tag{19}$$

For the distribution of the $t$-statistic to be exactly $t$, we require that the error term be normally distributed and i.i.d. except possibly for a common component for all observations. When all groups have the same sample size, we can readily check whether

$\bar{Z}_s$ is identical across groups, in the case where we use the within estimator to obtain $\hat{\gamma}_w$: or

$\text{var}(\hat{\gamma}(s))$ is identical for all groups, in the case where $\gamma$ is estimated separately for each group.

The $t(S-2)$ can be justified as an approximation based on large $N_s$ asymptotics under more general conditions. This occurs because the error term in equation (19) converges to the homoskedastic normal error $\alpha_s$ as $N_s \to \infty$ because of the consistency of $\hat{\gamma}$ and the fact that $\bar{\varepsilon}_s \overset{P}{\to} 0$ by the usual Law of Large Numbers. When $\sigma_\alpha^2$ is small, as in proposition 3, the approximation can be justified because $\bar{\varepsilon}_s$ and $\bar{Z}_s'(\gamma - \hat{\gamma})$ are approximately normally distributed by the Central Limit Theorem so that the error term in equation (19) is approximately normal.

[7] The greater robustness of the estimator using $\hat{\gamma}(s)$ gives rise to an additional issue. If $\hat{\gamma}$ varies across groups, then generally $\hat{d}_s$ and thus $\hat{\beta}$ will not be invariant to linear reparameterization of the $Z$'s. $\hat{\gamma}$ may vary across groups because of (a) sampling variation, (b) differences in $\gamma$ across groups that are related to the policy intervention being studied, or (c) differences in $\gamma$ across groups that are unrelated to the policy intervention being studied. The first problem should vanish as the number of observations per group becomes large. In the second case, it is not meaningful to discuss a single treatment effect that is constant across groups. This is analogous to the case where the return to schooling differs between whites and blacks. One cannot then estimate a single black-white wage differential for all schooling groups. In the last case, we can think of $\gamma(s)$ as a random variable. As the number of groups gets large, $\hat{\beta}$ will be independent of the parameterization, but since we are interested in the case where the number of groups is small, invariance is a problem that can be avoided by relying on $\hat{\gamma}_W$.

[8] We say that the $N_s$ are asymptotically identical provided that $p\lim N_s/N_t = 1$ for all $s$ and $t$.

While the theory above and the Monte Carlo evidence below suggest that using one of the two-stage estimators will generally be preferable to using OLS, there are two caveats that must be recognized. First, when $\sigma_\alpha^2 = 0$, OLS is the preferred estimator of equation (15) and the $t$-statistic has the conventional distribution. If one knows that $\sigma_\alpha^2 = 0$, OLS is the preferred estimator. What proposition 3 tells us is that if we proceed under the mistaken belief that $\sigma_\alpha^2 > 0$, two-stage estimation will still produce a statistic with a $t$-distribution.

Second, the distinction between two-stage estimation and one-stage estimation is really one of convenience. Whenever $\hat{T}_1$ has a $t$-distribution, the same statistic can be produced by relying on the OLS coefficients. For example, in the case where there are no group-varying covariates and each group has the same number of observations, the OLS coefficient is identical to the between-groups (two-step) estimator. However, it is more convenient to estimate the standard error of the estimate using between-group estimation. Similarly, it is easy to show that when $\bar{Z}_s$ is identical across groups and all groups have the same sample size, OLS produces the same $\hat{\beta}$ as in the case where we use the within estimator to obtain $\hat{\gamma}_w$ and estimate $\beta$ in two steps. Again, however, estimation of the correct standard error is much easier using the two-step estimator.

## III. Examples

In this section, we first review the two-by-two case, which features prominently in the literature. The main feature of this case is that we cannot calculate the standard error of the estimate and thus must exercise considerable caution in drawing conclusions. We then review two prominent papers that provide at least some difference-in-differences estimates in which there are no covariates that vary within group. The first case, Gruber and Poterba (1994), shows that accounting properly for error components can dramatically reduce the implied precision of the estimates in some specifications but that the estimate remains precise in at least one specification. In the second case we reexamine Card's (1990) Mariel boatlift study and suggest that the data cannot exclude large effects of the migration on blacks in Miami. This is consistent with the results of Angrist and Krueger's (1999) finding of a large impact of the "Mariel boatlift that didn't happen."

### A. *The Two-by-Two Case*

In the canonical difference-in-differences model, mean outcomes are calculated for groups A (the treatment group) and B (the control group) in each of periods 0 (the pretreatment period) and 1 (the posttreatment period). A standard table shows each of these means, plus the difference between groups A and B in each period and the difference between the pre- and posttreatment outcomes for each group. Finally, the difference between either pair of differences is the classic difference-in-differences estimate. Classic and recent examples that include tables in this form are Card and Krueger's (1994) study of the minimum wage; Eissa and Leibman's (1996) study of the effect of the earned-income tax credit; Meyer, Viscusi, and Durbin's (1995) study of workers compensation; Imbens, Rubin, and Sacerdote's (2001) study of lottery winners and labor supply, Eberts, Hollenbeck, and Stone's (2002) study of merit pay for teachers; and Finkelstein's (2002) study of tax subsidies and health insurance provision. Each of these studies provides additional analysis, but in each case, the two-by-two analysis is an important component of the study.

In a well-developed two-by-two case, the authors make a compelling case that other than the treatment, there is no reason to expect the outcome variable to evolve differently for the treatment and control groups.

The statistical model is

$$y_{igt} = \alpha_{gt} + bT_{gt} + \varepsilon_{igt}, \tag{20}$$

where $y$ is the outcome being measured for individual $i$ in group $g$ in year $t$, $T$ is a dummy variable for the treatment group in the posttreatment period, $\alpha$ is a group/year error which may be correlated over time or across groups, and $\varepsilon$ is an i.i.d. error term.

Without loss of generality, we can subtract the first period from the second period and rewrite the equation as

$$\Delta y_{ig} = \Delta a_g + b\Delta T_g + \varepsilon_{ig1} - \varepsilon_{ig0}. \tag{21}$$

Now $\Delta T$ equals 1 for the treatment group and 0 for the control group. Therefore, we can replace $\Delta T$ with $A$, a dummy variable for group A. And we can rewrite $\Delta a_g = c + \bar{\alpha}A$, where $\bar{\alpha} = \Delta\alpha_A - \Delta\alpha_B$. Letting $\varepsilon_{ig1} - \varepsilon_{ig0} = \varepsilon_{ig}$ we have

$$\Delta y_{ig} = c + (\bar{\alpha} + b)A_g + \varepsilon_{ig}. \tag{22}$$

The weighted least squares (where weights are chosen to make the sample sizes identical) coefficient on $A$ is the difference-in-differences estimator and is numerically identical to taking the difference between the change in the outcome for the treatment and control groups. It is an unbiased estimate of $\bar{\alpha} + b$. Since $E(\bar{\alpha}) = 0$, it is also an unbiased estimate of $b$. However, it is not consistent. No matter how many observations from either the control or treatment groups we add to the sample, the coefficient will not converge in probability to $b$.

The variance reported by econometric packages includes the sampling variance but not that part of the variance due to the common error. Thus if there are any shocks that are correlated within year/group cells, the reported $t$-statistic will be too high. We will tend to find an effect of the treatment even if none exists.

Unfortunately, if there are common errors, the two-by-two model has zero degrees of freedom. Therefore, it is not possible to determine the significance of any estimate solely

from within-sample information. It may be possible to use information from outside the sample to get a plausible estimate of the magnitude of common within-group errors, but even in this case, we will not know the sampling distribution of the resulting statistic. Thus, analysis of the two-by-two case requires extreme caution.

### B. *Card (1990)*

Card examines the impact of the mass migration of Cubans to Miami during the Mariel boatlift. He compares, among other outcomes, unemployment rates for whites, blacks, and Hispanics in Miami with unemployment rates of these groups in four comparison cities (Atlanta, Houston, Los Angeles, and Tampa–St. Petersburg). Surprisingly, he finds little evidence that the mass migration significantly affected the Miami labor market. For example, from 1979 to 1981 black unemployment in Miami increased by 1.3 percentage points, compared with 2.6 percentage points in the comparison communities. Angrist and Krueger (1999) replicate Card's study for a Cuban boatlift that was anticipated but did not occur. They find that "the Mariel boatlift that didn't happen" had a large adverse effect on unemployment in Miami. Their analysis cast doubt on the power of Card's original finding.

Our analysis helps to explain why Card found no effect and why it is possible to find a large effect of a nonexistent event. To understand this, we need to examine the true confidence interval around Card's estimates. Because Card provides seven years of data for both Miami and the comparison cities, we can, with auxiliary assumptions, calculate the variance of his estimate.

We first assume that the difference between the annual unemployment rates in Miami and the comparison cities is subject to an i.i.d. shock. This allows for a common year shock which may be persistent but assumes that any shocks that are idiosyncratic to a city are not persistent. Given this assumption, we use the data reported by Card to regress the difference between the unemployment rate for blacks in Miami and the control cities on a dummy for the period after 1980 on all years except 1980. The resulting coefficient is 1.4 with a standard error of 4.0. Under the assumption that the error terms are homoskedastic and normal, and given that we have four degrees of freedom, the confidence interval is from −9.7 to 12.1, effectively including very large positive and negative impacts on blacks.

In sum, while the data certainly provide no support for the view that the Mariel immigration dramatically increased unemployment among blacks in Miami, they do not provide much evidence against this view either. In this case, the difference-in-differences approach lacks power.

### C. *Gruber and Poterba (1994)*

Gruber and Poterba analyze a change in the tax law that they anticipate would increase the purchase of health insurance by self-employed individuals but not by individuals who work for someone else. They find that, comparing the period before the tax change with the period after, purchase of health insurance grew more rapidly among the self-employed than among other employed workers. The difference-in-differences estimator therefore indicates that the tax law did increase the purchase of health insurance among the self-employed.

TABLE 1.—AGGREGATE INSURANCE RATES: EMPLOYED AND SELF-EMPLOYED WORKERS (1982–1989)

| | Self-Employed | Employed | Difference | Difference-in-Differences (relative to previous year) |
|---|---|---|---|---|
| 1982 | 68.9 | 88.6 | −19.7 | — |
| 1983 | 72.0 | 88.9 | −16.9 | 2.8 |
| 1984 | 68.9 | 88.1 | −19.2 | −2.3 |
| 1985 | 68.6 | 88.0 | −19.4 | −0.2 |
| 1986 | 70.1 | 88.0 | −17.9 | 1.5 |
| 1987 | 76.1 | 86.8 | −10.7 | 7.2 |
| 1988 | 73.2 | 86.1 | −12.9 | −2.2 |
| 1989 | 73.5 | 84.5 | −11.0 | 1.9 |

Source: Gruber and Poterba (1994).

In the simplest version of the Gruber/Poterba model, the authors compare the fraction of the self-employed who had health insurance in 1985–1986 and 1988–1989 with the fraction of employed (not self-employed) workers with health insurance in these years.

Gruber and Poterba report a difference-in-differences estimate of 6.7 with a standard error of 0.8, indicating that the effect of the change in tax law is quite precisely measured. To examine the importance of common group/year effects, we begin by examining "the tax changes that didn't happen." We reproduce Gruber and Poterba's annual data in table 1. The fourth column shows the annual difference-in-differences estimates.

There are a number of points to make about the difference-in-differences. First, the 1986–1987 change stands out. One would be hard-pressed to look at the fourth column of table 1 and find no evidence supporting an impact of the tax law change on the insurance rate among the self-employed.

More significantly from the perspective of this paper, it would be easy to look at other years and, using difference-in-differences, draw strong and possibly erroneous conclusions about the impact of other policies in those years. Based on information in Gruber and Poterba, the part of the standard error of the difference-in-differences estimator that is due to sampling error is approximately 1.1.[9] Relying on this standard error, for three of the six years in which no major policy change occurred, there appears to be a statistically significant change in the relative purchase of health insurance by the self-employed. For two of the remaining three years, the change falls short of conventional significance levels but remains sufficiently large relative to its "standard error" to provide support for the hypothesis of a policy effect.

[9] The standard error reported in Gruber/Poterba is 0.8. Their difference-in-differences estimator pools data from two years. The standard error for individual years is therefore approximately $\sqrt{2} \times 0.8$, or 1.1.

We can get a more accurate estimate of the standard error of the difference-in-differences estimator if we are willing to make auxiliary assumptions about the distribution of the group/year errors. First we assume that the difference between the employment rates for the two groups is i.i.d. normal. We therefore regress the differences in column 3 on a dummy variable for the period after 1986. This effectively treats 1982–1986 as one group (but with different random year errors) and 1987–1989 as a second group. The results from this estimation are strikingly similar to those obtained by Gruber and Poterba. The coefficient is 7.1 with a standard error of 0.9. Because there are only six degrees of freedom in the second-stage regression, the confidence interval is obtained by multiplying the standard error by 2.45 instead of the more common 1.96. Still, the cost of the group-error structure is largely offset by the increased sample size from using all nine years of data.

If instead we assume that the difference in the differences between the employment rates is i.i.d., in other words that the differences are a possibly correlated, random walk and normal, we can estimate the regression using differences. We therefore regress the difference-in-differences on a dummy variable for 1986–1987. This produces a difference-in-differences estimate of 6.9 and a standard error of 2.3. Using OLS and a $t$-distribution for the parameter estimate, the estimated impact of the policy change has a confidence interval ranging from 0.9 to 13.0.

Finally, since the difference-in-differences approach is predicated on the assumption that the expected difference-in-differences in nonexperimental years is 0, we reestimate the equation without a constant term. The resulting coefficient is 7.2 with a confidence interval from 2.3 to 12.1.

While the evidence against the hypothesis of no policy effect remains statistically significant, our confidence in its magnitude is diminished dramatically by taking into account random year effects in two of our three specifications. Nevertheless, the results demonstrate that it is possible to obtain precise coefficients in at least in some specifications.

## IV. OLS and Variance Adjustment: Moulton and Cluster

It is interesting to consider the properties of two commonly used approaches to statistical inference in the context of the model described earlier. To make these procedures easier to follow, we consider the special case where there are no within-group varying variables and the covariance across observations in the group is constant, as represented by the model with a random group effect,

$$Y_{is} = a + X_s\beta + \alpha_s + \varepsilon_{is}. \tag{23}$$

Moulton (1986) suggested that one adjust the standard errors for OLS for the fact that the errors are correlated within the groups because of the common group effect. Under the assumption that all residuals are homoskedastic, this correlation is given by $\rho = \sigma_\alpha^2/(\sigma_\alpha^2 + \sigma_\varepsilon^2)$ while the variance of the residual is $\sigma^2 = \sigma_\alpha^2 + \sigma_\varepsilon^2$. Then using the notation $X$ for the $(\sum_{s=1}^{S} N_s \times 2)$ regressor matrix in the above model (consisting of column of 1s and a column of $X_s$), then as shown in Moulton (1986) the variance of OLS has the form

$$V = (X'X)^{-1}X'\Sigma X(X'X)^{-1}, \tag{24}$$

where $\Sigma$ is block diagonal with blocks given by $\sigma^2(1-\rho)I_{N_s} + \sigma^2\rho\iota_{N_s}\iota_{N_s}'$. Moulton's suggestion is to estimate $\sigma^2$ and $\rho$ and to use the estimated version of the matrix $V$ to compute standard errors. Given the definition of $\rho$ and $\sigma^2$, this can be done as follows:

$$\hat{\sigma}^2 = \frac{\sum_s\sum_i(Y_{is} - \hat{a} - X_s\hat{\beta})^2}{N^*} = \frac{\sum_s\sum_i e_{is}^2}{N^*}, \tag{25}$$

where $N^* = \sum_s N_s$ and

$$(S-2)\hat{\sigma}^2\hat{\rho} = \sum_s\sum_i\sum_{j\neq i}\frac{(Y_{is} - \hat{a} - X_s\hat{\beta})(Y_{js} - \hat{a} - X_s\hat{\beta})}{N_s(N_s-1)} = \sum_s(\bar{Y}_s - \hat{a} - X_s\hat{\beta})^2 - \frac{\sum_s\sum_i(Y_{is} - \hat{a} - X_s\hat{\beta})(Y_{is} - \bar{Y}_s)}{(N_s-1)} \tag{26}$$

where the latter is an estimator of the within-group covariance of the residuals ($\sigma_\alpha^2$) that has been pooled across the groups.[10] Also note that we have included a degrees-of-freedom adjustment in the denominator of $\hat{\sigma}^2\hat{\rho}$.

The "cluster" adjusted standard error (as performed in programs such as Stata) is aimed at dealing with the within-group correlation structure but does not impose homogeneity of the variances. It takes the form

$$\hat{V} = \frac{S}{S-1}\frac{\sum_{s=1}^{S} N_s - 1}{\sum_{s=1}^{S} N_s - 2}(X'X)^{-1}\Big(\sum_{s=1}^{S} u_s u_s'\Big)(X'X)^{-1} \tag{27}$$

$$u_s = \sum_{i=1}^{N_s} e_{is}\begin{pmatrix}1\\ X_s\end{pmatrix}, \tag{28}$$

where $e_{is}$ is the OLS residual. This procedure allows for general within-group covariance and heteroskedasticity.

[10] Moulton (1986) actually uses $\hat{\sigma}^2$ and $\hat{\rho}$ obtained from the random-effects MLE of the model. The formulas we have used are similar, although they use the OLS residual rather than the residual from the random-effects MLE.

It is interesting to examine the properties of inference procedures for these approaches in the case where we have a fixed small number of groups and where there are large (but possibly different) numbers of observations within each group.[11] It is possible to show that the OLS estimator under these conditions satisfies

$$\begin{pmatrix} \hat{a} \\ \hat{\beta} \end{pmatrix} \xrightarrow{P} \begin{pmatrix} a \\ \beta \end{pmatrix} + (X_*'PX_*)^{-1}X_*'P\alpha_*, \tag{29}$$

where $X_*$ is the $S \times 2$ matrix consisting of a column of 1s and a column of the different group constant variables $X_s$, and where $P$ is a diagonal matrix consisting of $p(s) = p \lim N_s/N^*$. Thus the OLS estimator behaves under these asymptotics like a weighted least squares estimator in the model,

$$p\lim \bar{Y}_s = a + X_s\beta + \alpha_s. \tag{30}$$

By contrast, the two-step estimator discussed above behaves (asymptotically) like OLS in this model and will be more efficient when $\alpha_s$ is homoskedastic. It is also possible to show that the $t$-statistic based on the "Moulton" standard error will satisfy

$$t \xrightarrow{d} \frac{\Lambda'\alpha}{\sqrt{\tilde{\sigma}_\alpha^2 \Lambda'\Lambda}},$$

$$\Lambda' = w'(X_*'PX_*)^{-1}X_*'P,\ w' = (0,1)$$

$$\tilde{\sigma}_\alpha^2 = \frac{\alpha'(1 - X_*(X_*'X_*)^{-1}X_*')\alpha}{S-2}.$$

There are a few things to note about this distribution. First, assuming normality of $\alpha$, it is possible to show that although one can write the limiting random variable as a ratio of a $N(0, 1)$ and a $\chi^2(S-2)$, the two random variables will not be independent unless $(I - X_*(X_*'X_*)^{-1}X_*')\Lambda = 0$. This condition will occur only when the $p(s)$ are all identical. Thus in the case where the number of observations per group are similar a $t(S-2)$ distribution will provide a good approximation to the $t$-statistic using the Moulton correction. It is also worth noting that the scale adjustment suggested for Moulton is crucial for this result—the scale adjustment results in the estimate of $\sigma_\alpha^2$ being asymptotically unbiased as well.[12]

For the approach based on the cluster correction one can show that

$$t \xrightarrow{d} \frac{\Lambda'\alpha}{\sqrt{\tilde{V}_{22}}},$$

where $\tilde{V}_{22}$ *is the (2, 2) element of*

$$\tilde{V} = \frac{S}{S-1}(X_*'PX_*)^{-1}X_*'PAPX_*(X_*'PX_*)^{-1}$$

$$A = diag\{a_s^2\}$$

$$a_s = \alpha_s + (1, X_s)(X_*'PX_*)^{-1}X_*'P\alpha,$$

which is the (stochastic) limit of the cluster variance estimator. It is interesting to note that $\tilde{V}$ is a scaled version of the Eicker-White variance-covariance matrix in the weighted least squares regression (30). This interpretation leads to some conclusions. First, the only real justification for the use of these adjusted standard errors is asymptotic in the sense that $S$ must be large. When $S$ is small the distribution will generally be unknown, and there do not appear to be cases where the $t$-distribution would be a good approximation. Moreover, it is well-known that there can be substantial small sample downward bias in the Eicker-White standard errors, with evidence suggesting that they lead to overrejection of true null hypotheses (see MacKinnon & White, 1985). This suggests that the cluster approach may be quite unreliable except in the case where there are many groups.

## V. Monte Carlo Evidence

In this section we provide Monte Carlo estimates of the distribution of the $t$-statistic for a variety of estimators used with panel data.

Our first set of experiments addresses the four most common estimators applied to grouped data—ordinary least squares with conventional standard errors, OLS with Eicker-White heteroskedasticity robust standard errors for grouped data,[13] random-effects estimation,[14] and two-step estimation in which the first stage is fixed-effects estimation. We address two-step estimation in which the first stage is estimated separately for each group in a later experiment.

For the first set of experiments, we assume that the underlying model is

$$y_{is} = \alpha_s + \varepsilon_{is},$$

with the error terms independent normals and $100\sigma_\alpha^2 = \sigma_\varepsilon^2 = 1$. We estimate models of the form

$$y_{is} = X_sB + Z_{is}\Gamma + \alpha_s + \varepsilon_{is}$$

so that the true parameter values are 0. $X$ takes on the values 1, 2, 3, and 4 with equal numbers from each group. $Z$ is distributed uniform on the unit interval for each of the four groups. When $B$ is estimated using the two-step estimator, the second stage takes the form

$$\hat{d}_s = a + X_sB + \tilde{\alpha}_s.$$

[11] We note in passing that in contrast to the estimator discussed above, test statistics based on OLS and using either the "Moulton" or "cluster" standard errors will not have exact $t$-distributions under normality.

[12] When the groups have different numbers of observations but normality of $\alpha_s$, it is assumed it should be possible to simulate the distribution.

[13] These are obtained using the cluster option in Stata.

[14] These are obtained using the xtreg command in Stata.

For each case, we simulated 50,000 estimates. We also experimented with including and excluding $Z$ from the equation.

Within this set of estimates, we experiment with two sample sizes: 250 observations per group (1,000 total) and 2,500 observations per group (10,000 total). We note that for the larger sample, we would expect the large $N$, fixed variances, asymptotics to apply so that the asymptotic distribution of the $t$-statistic for the two-step estimator would be the $t$-distribution. For the smaller sample, the asymptotics with $\sigma_\alpha^2$ proportional to $\sqrt{N}$ should apply. Since in this set of experiments $E(\bar{Z}_s) = 0.5$, $\forall_s$, the $t$-statistic should have a $t$-distribution with two degrees of freedom in all cases.

Tables 2 and 3 report the results of these Monte Carlo experiments.

The first thing to notice from the tables is that they support the theoretical predictions regarding the distribution of the $t$-statistic when two-step estimation is used. The 0.01, 0.05, and 0.1 critical values of the $t$-distribution with two degrees of freedom are 9.92, 4.30, and 2.92. In all twelve cases, the Monte Carlo estimates are close to these critical values, and the true critical values lie within the confidence intervals.

The results for ordinary least squares with conventional standard errors confirm Moulton's findings. Even with a relatively modest number of observations and a low covariance across observations within a group, using the normal distribution to determine the significance of the $t$-statistic is badly biased. In our experiments with 250 observations per group, approximately 30% of the estimates obtained with OLS have conventional $t$-statistics that would be deemed significant at the 0.05 level. With 2,500 observations per group, this fraction rises to 70%.

Since empirical economists have become increasingly aware of Moulton's critique and since robust. standard errors accounting for group errors are available for some statistical packages, it has become common for researchers to present these robust standard errors in lieu of conventional standard errors. As can be seen from tables 2 and 3, the distribution of the $t$-statistic using Eicker-White standard errors has the distinct advantage of being unaffected by the number of observations within each group. However, applying the normal distribution can again give rise to highly misleading inference. The $t$-statistic exceeds 1.96 in about one-third of cases. In fact, the $t$-statistic divided by about 1.565 seems to have a $t$-distribution with two degrees of freedom. We have not yet established why the Eicker-White $t$-statistic should have this distribution and hence cannot say whether it should generalize to other cases.[15]

TABLE 2.—MONTE CARLO ESTIMATION
DISTRIBUTION OF ABSOLUTE VALUE OF $T$-STATISTICS
(FOUR GROUPS, 250 OBSERVATIONS PER GROUP)

| | 99th Percentile | 95th Percentile | 90th Percentile | % > 1.645 | % > 1.96 |
|---|---|---|---|---|---|
| OLS (conventional standard errors) | | | | | |
| No $Z$ | 4.83 | 3.65 | 3.06 | 37.8 | 29.5 |
| $Z$ | 4.83 | 3.65 | 3.06 | 37.8 | 29.4 |
| OLS (Eicker-White standard errors) | | | | | |
| No $Z$ | 15.29 | 6.68 | 4.57 | 39.4 | 32.9 |
| $Z$ | 14.92 | 6.68 | 4.56 | 39.3 | 32.8 |
| Feasible GLS (random effects) | | | | | |
| No $Z$ | 4.03 | 2.82 | 2.22 | 19.0 | 13.5 |
| $Z$ | 4.30 | 3.05 | 2.44 | 22.6 | 16.7 |
| Two-Step | | | | | |
| No $Z$ | 9.75 | 4.29 | 2.93 | 24.0 | 18.8 |
| $Z$ | 9.74 | 4.29 | 2.93 | 24.0 | 18.8 |

TABLE 3.—MONTE CARLO ESTIMATION
DISTRIBUTION OF ABSOLUTE VALUE OF $T$-STATISTICS
(FOUR GROUPS, 2,500 OBSERVATIONS PER GROUP)

| | 99th Percentile | 95th Percentile | 90th Percentile | % > 1.645 | % > 1.96 |
|---|---|---|---|---|---|
| OLS (conventional standard errors) | | | | | |
| No $Z$ | 13.01 | 9.93 | 8.40 | 74.5 | 69.8 |
| $Z$ | 13.01 | 9.93 | 8.40 | 74.5 | 69.9 |
| OLS (Eicker-White standard errors) | | | | | |
| No $Z$ | 15.56 | 6.74 | 4.58 | 39.6 | 33.0 |
| $Z$ | 14.58 | 6.74 | 4.58 | 39.6 | 33.0 |
| Feasible GLS (random effects) | | | | | |
| No $Z$ | 7.34 | 4.00 | 2.82 | 23.8 | 18.6 |
| $Z$ | 9.74 | 6.07 | 4.30 | 32.7 | 27.4 |
| Two-Step | | | | | |
| No $Z$ | 9.72 | 4.28 | 2.92 | 24.1 | 18.9 |
| $Z$ | 9.75 | 4.28 | 2.92 | 24.1 | 18.9 |

Finally, random-effects estimation is asymptotically efficient as both $S$ and $N$ go to infinity. Therefore, it is natural to estimate such models by feasible GLS. When the random-effects estimator is used, the distribution of the $t$-statistic depends not only on sample size, but also on whether $Z$ is included in the equation. Depending on whether $Z$ is included and on the sample size, the bias from assuming that the $t$-statistic is normally distributed ranges from modest (14% significant at the 0.05 level) to quite large (27% significant at the 0.05 level).

Having established that with two-step estimation the $t$-statistic has the $t$-distribution in those cases where the theory predicts that it should, we now turn to cases where the theory does not predict that the distribution will be $t$. We modify the true model so that

$$y_{is} = Z_{is}\Gamma + \alpha_s + \varepsilon_{is}.$$

We generate

$$Z_{is} = \bar{Z}_s + z_{is},$$

[15] Bell and McCaffrey (2002) propose an adjustment to the clustered standard error which seems to work well even when group size is small and there are twenty groups.

where $z_{is}$ is uniform (0,1). We consider the two cases. In the uncorrelated case. $(X, \bar{Z}) \in \{(0, 1), (1, 3), (2, 2), (3, 0)\}$. In this case, the relation between $X$ and $\bar{Z}$ has been chosen so that they are uncorrelated. In the correlated case, $X = \bar{Z}$.

We again perform 50,000 replications for each case we consider. For each scenario, we assume four groups and allow both 250 and 2,500 observations per group. Finally, we perform our two-step estimation in two ways. In the first, we follow convention and estimate fixed group effects in the first stage and regress the fixed effects on $X$ in the second stage. In the other approach, we estimate the constant and coefficient on $Z_i$ separately for each group and regress the constants from these equations on $X$.

Table 4 reports the results of this estimation. When there are only 250 observations per group, then the theoretical analysis suggests that there is no reason to assume that the $t$-distribution will be a good approximation to the distribution of the $t$-statistic. This is confirmed by the first part of the table, which shows that the centiles of the simulated distributions depart quite significantly from the critical values of the $t$-distribution with two degrees of freedom. Thus the 0.05 critical value is 4.30. Except in the case where $\Gamma$ is allowed to vary across groups in the first-stage estimation, the 95th percentile of the distributions are quite far from this number. Perhaps even more disturbing, the distribution is greatly affected by the correlation between $X_s$ and $\bar{Z}_s$, at least in the case where $\Gamma$ is constrained to be identical across groups.

We anticipated that with 2,500 observations per group, the large $N$ asymptotics would apply and that, therefore, the $t$-statistic would have a $t$-distribution. Apparently convergence to the $t$-distribution is slower than we anticipated.

TABLE 4.—MONTE CARLO ESTIMATION
DISTRIBUTION OF ABSOLUTE VALUE OF $t$-STATISTICS
(FOUR GROUPS, $Z$ ENTERING THE STRUCTURAL EQUATION)

| | 99th Percentile | 95th Percentile | 90th Percentile | % > 1.645 | % > 1.96 |
|---|---|---|---|---|---|
| 250 Observations/Group | | | | | |
| $X_s$ and $\mu_z(s)$ uncorrelated | | | | | |
| $\gamma$ constant across groups | 6.55 | 2.81 | 1.88 | 12.4 | 9.3 |
| $\gamma$ varies across groups | 8.10 | 3.69 | 2.56 | 21.2 | 16.1 |
| $X_s = \mu_z(s)$ | | | | | |
| $\gamma$ constant across groups | 22.78 | 9.98 | 6.76 | 55.0 | 48.5 |
| $\gamma$ varies across groups | 9.09 | 4.15 | 2.94 | 27.4 | 20.7 |
| 2,500 Observations/Group | | | | | |
| $X_s$ and $\mu_z(s)$ uncorrelated | | | | | |
| $\gamma$ constant across groups | 8.46 | 3.81 | 2.59 | 20.4 | 15.7 |
| $\gamma$ varies across groups | 8.84 | 3.86 | 2.66 | 21.7 | 16.6 |
| $X_s = \mu_z(s)$ | | | | | |
| $\gamma$ constant across groups | 12.98 | 5.35 | 3.63 | 31.8 | 25.7 |
| $\gamma$ varies across groups | 9.83 | 4.33 | 2.99 | 26.3 | 20.4 |

Again, except in the case where $\Gamma$ is allowed to vary across groups in the first-stage estimation, the 95th percentile of the distributions are quite far from 4.30, although in each case, it is closer than when there are only 250 observations per group.

We note, however, that in no case does the 95th percentile of the $t$-distribution significantly exceed 4.30 when $\Gamma$ varies across groups. Moreover, the distribution is considerably less sensitive to the precise specification of the experiment when $\Gamma$ varies than when it is constant across groups. It appears that estimating the first stage separately for each group and using the $t$-distribution to assess the significance of the second-stage coefficients is a conservative strategy.

Table 4 does not address the performance of the other estimators evaluated in tables 2 and 3. The poor performance of the random-effects estimator and uncorrected OLS even under the best of circumstances makes it pointless to examine them in the presence of other explanatory variables. However, it appears from tables 2 and 3 that the $t$-statistic from OLS estimation and Eicker-White standard errors might provide, with suitable modification, a reliable basis for inference. Unfortunately, this test statistic turns out to be very sensitive to the inclusion of $Z$ with a nonzero coefficient. Indeed, in the experiment where $\bar{Z}$ and $X$ are forced to be independent, the lowest Eicker-White statistic (in absolute value) that we obtained in 50,000 draws was 7.56. Clearly, the Eicker-White standard errors do not provide a reliable basis for inference.

In our final Monte Carlo exercise, we draw on the example from Gruber and Poterba to assess the power of the two-step estimator. We assume that there are eight years of data with 50,000 observations in each year. The first six years are the preexperiment years and the last two the postexperiment years. In each year, 4,000 of the observations are from members of the group that will be subject to the experimental treatment (the self-employed). For the nonexperimental group (employees), each observation has a probability of 0.87 of success (being insured). For the potential treatment group, the probability of success is 0.71 plus a normally distributed year error. The year error has a standard deviation of 0.009, corresponding to our estimate of the variance of the year effect in Gruber and Poterba.[16]

We provide three sets of estimates. In the first two cases, we estimate linear probability models in which we regress whether the individual is insured on a constant, a dummy for whether the individual is self-employed, a dummy for being in the experimental period, and an interaction term between being self-employed and being in the experimental period. The first estimator relies on conventional standard errors, while the second uses Eicker-White standard errors clustered by year/group. The third set of estimates uses our two-step estimator.

[16] It is easier to assign the error to one group rather than to apportion it between the two groups. Since we care about the variance of the difference in the group errors, it is irrelevant how we assign the group error.

TABLE 5.—MONTE CARLO ESTIMATION
DISTRIBUTION OF ABSOLUTE VALUE OF *T*-STATISTICS
(MIMICKING GRUBER & POTERBA, 1994)

| | 99th Percentile | 95th Percentile | 90th Percentile | % > 1.645 | % > 1.96 |
|---|---|---|---|---|---|
| | | No True Treatment Effect | | | |
| Linear probability model | 5.17 | 3.98 | 3.35 | 41.8 | 33.5 |
| Linear probability model (clustered standard errors) | 5.51 | 3.61 | 2.83 | 28.2 | 21.5 |
| Two-stage estimation | 3.73 | 2.45 | 1.95 | 15.2 | 9.9 |
| | | Treatment Effect = 0.07 | | | |
| Linear probability model | 19.67 | 18.32 | 17.59 | 100.0 | 100.0 |
| Linear probability model (clustered standard errors) | 30.05 | 22.09 | 18.82 | 100.0 | 100.0 |
| Two-stage estimation | 20.17 | 14.64 | 12.48 | 100.0 | 100.0 |

We check the distribution and significance of the standard errors for two cases. In the first case, the true experimental effect is zero. In the second case, it is 0.07, our estimate of the effect size in Gruber and Poterba. The results in table 5 are based on 60,500 simulations.[17] As would be expected with group sizes of 46,000 and 4,000, the large $N$ results are quite accurate. In the case where the true effect size is zero, using the two-step technique, the 90th, 95th, and 99th percentiles of the $t$-statistic are very close to their predicted values of 1.94, 2.45, and 3.71. In contrast, consistent with previous results, OLS rejects way too frequently. The $t$-statistic exceeds 1.96 in just about one-third of cases. Using the Stata cluster command reduces the number of rejections, but the $t$-statistic still exceeds 1.96 in over one-fifth of the simulations.

When the true effect size is 0.07, the two-step estimator is nevertheless powerful. The null hypothesis of no effect can be rejected at the 0.05 level in 60,449 out of 60,500 simulations and at the 0.1 level in every case.

## VI. Discussion and Conclusion

This paper makes the following basic points regarding difference-in-differences estimation, or more generally estimation with grouped data:

(i) Moulton's critique of estimation with grouped data applies to difference-in-differences estimation.
(ii) When the number of groups is small, $t$-statistics obtained using standard methods (OLS, OLS with Eicker-White standard errors, feasible GLS estimation of the random-effects model, two-stage estimation) are not normally distributed.
(iii) If the number of members of each group is large, two-step estimation is efficient and $t$-statistics from two-step estimation have $t$-distributions if the underlying common group errors are normally distributed.
(iv) If the number of members of each group is small, only under special circumstances is two-step estimation efficient and do $t$-statistics from two-step estimation have $t$-distributions. The criticism of inference using other estimators still applies.

While it will not normally be feasible to check whether the underlying distribution of common group errors is normally distributed, it is relatively straightforward to verify the remaining conditions under which the $t$-statistic from the two-step estimators is distributed $t$. When sample sizes for each group are similar, it will frequently be the case that the standard errors of first-stage coefficients estimated separately for each sample will also be similar. In other cases, groups will have similar distributions of individual-specific variables.

Our reanalysis of two studies suggests that taking these considerations into account is important. As a practical matter, when there are variables that vary within group, the two-step estimator in which the first stage is estimated separately for each group seems to us to be promising. It is robust both in theory and in our simulations. In addition, it is easily used with conventional software packages. However, as noted in footnote 7, in many situations it will not be invariant to the parameterization of within-group varying variables. In such cases constraining $\gamma$ to be equal across groups will generally be necessary. There are other settings, however, in which the estimate will be invariant. For example, if we are interested in the effect of a law on the black-white wage differential, the estimate of this differential in each year is invariant to linear reparameterizations of the remaining variables. In such settings, estimation allowing the remaining parameters to vary across groups is likely to be the most desirable estimator.

[17] We inadvertently exceeded our intended 50,000 simulations and saw no reason to throw away the extra simulations. Needless to say, the results for the first 50,000 simulations are similar.

## REFERENCES

Abadie, Alberto, "Semiparametric Difference in Difference Estimators," *Review of Economic Studies* 72 (January 2005), 1–19.
Amemiya, Takeshi, "A Note on a Random Coefficients Model," *International Economic Review* 19 (October 1978), 793–796.
Andrews, Donald W. K., "Cross-Section Regression with Common Shocks," *Econometrica* 73 (September 2005), 1551–1585.
Angrist, Joshua D, and Alan B. Krueger, "Empirical Strategies in Labor Economics" (pp. 1277–1366), in Orley E. Ashenfelter and David Card

(Eds.), *Handbook of Labor Economics, Volume 3A* (Amsterdam; New York, and Oxford: Elsevier Science, North-Holland, 1999).

Athey, Susan, and Guido W. Imbens, "Identification and Inference in Nonlinear Difference-in-Differences Models," *Econometrica* (2005).

Bell, Robert M., and Daniel F. McCaffrey, "Bias Reduction in Standard Errors for Linear Regression with Multi-Stage Samples," *Survey Methodology* 28 (December 2002), 169–181.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan, "How Much Should We Trust Differences-in-Differences Estimates?" *Quarterly Journal of Economics* 119 (February 2004), 249–276.

Card, David, "The Impact of the Mariel Boatlift on the Miami Labor Market," *Industrial and Labor Relations Review* 43 (January 1990), 245–257.

Card, David, and Alan B. Krueger, "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania," *American Economic Review* 84 (September 1994), 772–793.

Dickens, William T., "Error Components in Grouped Data: Is It Ever Worth Weighting?" this REVIEW 72 (May 1990), 328–333.

Eberts, Randall, Kevin Hollenbeck, and Joe Stone, "Teacher Performance Incentives and Student Outcomes," *Journal of Human Resources* 37 (Fall 2002), 913–927.

Eissa, Nada, and Jeffrey B. Liebman, "Labor Supply Response to the Earned Income Tax Credit," *Quarterly Journal of Economics* 111 (May 1996), 605–637.

Finkelstein, Amy, "The Effect of Tax Subsidies to Employer-Provided Supplementary Health Insurance: Evidence from Canada," *Journal of Public Economics* 84 (June 2002), 305–339.

Gruber, Jonathan, and Jems Poterba, "Tax Incentives and the Decision to Purchase Health Insurance: Evidence from the Self-Employed," *Quarterly Journal of Economics* 109 (August 1994), 701–734.

Imbens, Guido W., Donald B. Rubin, and Bruce I. Sacerdote, "Estimating the Effect of Unearned Income on Labor Earnings, Savings, and Consumption: Evidence from a Survey of Lottery Players," *American Economic Review* 91 (September 2001), 778–794.

Kloek, Teun, "OLS Estimation in a Model Where a Microvariable Is Explained by Aggregates and Contemporaneous Disturbances are Equicorrelated," *Econometrica* 49 (January 1981), 205–207.

Liang, Kung-yee, and Scott L. Zeger, "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika* 73 (March 1986), 13–22.

MacKinnon, James G., and Halbert White, "Some Heteroskedasticity-Consistent Covariance Matrix Estimators with Improved Finite Sample Properties," *Journal of Econometrics* 29 (September 1985), 305–325.

Meyer, Bruce D., W. Kip Viscusi, and David L. Durbin, "Workers' Compensation and Injury Duration: Evidence from a Natural Experiment," *American Economic Review* 85 (June 1995), 322–340.

Moulton, Brent R., "Random Group Effects and the Precision of Regression Estimates," *Journal of Econometrics* 32 (August 1986), 385–397.

——— "An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Units," this REVIEW 72 (May 1990), 334–338.