Spurious Significance of Treatment Effects in Overfitted Fixed Effect Models

Albrecht Ritschl[1]
LSE and CEPR

March 2009

1 Introduction

Evaluating subsample means across groups and time periods is common in panel studies that evaluate the treatment effects of training programs, labor market policies, currency unions etc. Comparison of means between treated and non-treated groups may occur along the time axis (fixed effects, FE), the cross section (pooled OLS, IV) or in a combination of the two (difference in differences, DiD, see Ashenfelter (1978), Ashenfelter and Card (1985)), depending on the choice of identifying assumptions about selectivity and common trends (see Heckman, Lalonde, and Smith (1999)). Despite their widespread use in evaluation studies, FE and DiD estimators have acquired a reputation for generating spuriously low standard errors on the estimated treatment effect. Bertrand, Duflo, and Mullainathan (2004) survey empirical applications of the DiD estimator, and find that much of this phenomenon can be attributed to autocorrelation. In a simulated dataset, they show that many standard methods for dealing with autocorrelation yield downward biased standard errors on the treatment effect coefficient.

The following note argues that spurious significance of treatment effects in panels may also occur in the absence of autocorrelation. This phenomenon arises in overfitted FE and DiD modeling of within-group comparisons. Overfitting in such models occurs if observation-specific individual fixed effects (IFE) are specified, although the comparison would be identified by group-specific fixed effects. In evaluation studies, identifying the average treatment effect on the treated through a within-group estimator would require a group fixed effect on the treated (FET), see e.g. Angrist and Pischke (2009). Yet specifying an overfitted regression with IFE instead may seem innocuous to the applied researcher, as the coefficient estimates on the treatment effect under both fixed effect specifications are identical. Moreover, standard software packages provide easy to use options for individual fixed effects, making an overfitted specification seem attractive. However, while the estimated treatment effect under IFE and FET is the same, its estimated standard error is not. Overfitting through IFE leads to spurious precision of the estimated treatment effect coefficient. The resulting bias is related to the reduction in the residual sum of squares induced by employing IFE instead of FET. Under ideal conditions where all

[1] Financial support from *Deutsche Forschungsgemeinschaft* under SFB649 “Economic Risk” at Humboldt University Berlin is gratefully acknowledged.

other regressors are uncorrelated to the treatment and the fixed effects, this relation is strictly proportional.

The rest of this note is structured as followed. The next section provides the setup. Section (3) presents the result. Section (4) concludes.

2 A Minimal and an Overfitted Setup

Consider a data panel with n observation units in the cross section and T time periods. In this panel, denote by $Y_{Tn\times 1}$ the dependent variable. Z is a matrix of characteristics of interest, as well as any time fixed effects, while X includes the regression constant and/or a suitably chosen matrix of either individual or group fixed effects. A policy treatment is applied to some observation units y_i during treatment period $\tau = \{s, \ldots, s+\tau\} \subset T$. Treatment during period $t \in \tau$ is indicated by a $(n \times 1)$-vector of dummy variables Δ_t, which are equal to one if unit i is under treatment at time t, and zero otherwise. $d = tr(\Delta) < n$ is the number of observation units i in the treatment group. Accordingly, $n-d$ is the number of observation units in the non-treated group.

A standard linear panel model of this treatment effect problem is:

$$Y = (XD)\beta + Z\gamma + v \tag{1}$$

where $v \sim N(0, \sigma_v^2)$ and where $D = (0 \ldots \Delta_s' \ldots \Delta_\tau' \ldots 0)'$ is a dummy vector capturing the policy treatment in τ periods. Fixed effects estimation of models like (1) is a popular (yet problematic) attempt to ensure the exogeneity of D with respect to the disturbance term v.

To focus on the essentials, consider an ideal regression in which any characteristics included in Z are orthogonal to the fixed effects X and the treatment dummy D. Define the detrended variable $y = M_z Y$ with $M_z = I - Z(Z'Z)^{-1}Z'$, where the influence on Y of any such characteristics, as well as any time fixed effects included in Z has been removed[2]. As $M_z XD = XD$ if $Z'XD = 0$, the model becomes a Least Squares Dummy Variables (LSDV) regression on the fixed effect terms and the treatment dummy only:

$$y = (XD)\beta + u \tag{2}$$

where X is a suitably chosen matrix of fixed effects, $D = (0 \ldots \Delta_1' \ldots \Delta_\tau' \ldots 0)'$ is a dummy vector capturing the policy treatment in τ periods, and $u \sim N(0, \sigma_u^2)$.

Under individual fixed effects, X consists of T stacked $(n \times n)$ identity matrices:

$$X^I = \begin{pmatrix} I_{n\times n} \\ \vdots \\ I_{n\times n} \end{pmatrix}_{Tn\times n}$$

Under the alternative assumption of a group fixed effect on the treated, matrix X takes the form:

[2] Time fixed effects would be orthogonal to X. Their inclusion in Z makes the FE and DiD estimators in y identical.

$$X^G = \begin{pmatrix} \mathbf{1} & \Delta \\ \vdots & \vdots \\ \mathbf{1} & \Delta \end{pmatrix}_{Tn\times 2}$$

Note that the column dimension of X^G is 2 as opposed to n in X^I. LSDV estimation of (2) under the two different fixed effect specifications yields:

$$\begin{aligned} \hat{\beta}^I &= ([X^I D]'[X^I D])^{-1}[X^I D]'y \\ \hat{\Omega}_{\hat{\beta},I} &= \hat{\sigma}^2_{u,I}\,([X^I D]'[X^I D])^{-1} \end{aligned} \tag{3}$$

$$\begin{aligned} \hat{\beta}^G &= ([X^G D]'[X^G D])^{-1}[X^G D]'y \\ \hat{\Omega}_{\hat{\beta},G} &= \hat{\sigma}^2_{u,G}\,([X^G D]'[X^G D])^{-1} \end{aligned} \tag{4}$$

Let b^I be the $n+1$th (i.e., last) element of $\hat{\beta}^I$, and b^G be the 3rd (i.e., last) element of $\hat{\beta}^G$. b^I and b^G are the coefficients on the treatment dummy under Individual Fixed Effects (I) and the Group Fixed Effect on the Treated (G), respectively. Likewise, let $\hat{\sigma}^2(b^I) = \hat{\Omega}_{\hat{\beta},I,(n+1,n+1)}$ and $\hat{\sigma}^2(b^G) = \hat{\Omega}_{\hat{\beta},G,(3,3)}$ be the estimated variances of these coefficients, with $S^I_{n+1,n+1} = [X^I D]'[X^I D])^{-1}_{n+1,n+1}$ and $S^G_{3,3} = [X^G D]'[X^G D])^{-1}_{3,3}$ as the pertaining elements of the matrix inverses in (3) and (4), respectively.

3 Spurious Significance under Overfitting

Consider a treatment effect model as in eq. (2), in which the endogenous variable has been detrended from any time effects, and in which any further characteristics are orthogonal to the fixed effects and treatment dummy, and have been eliminated as well. Estimation under the alternatives of Individual Fixed Effects (IFE) and Fixed Effects on the Treated (FET) as in (3) yields identical coefficient estimates on the treatment effects. However, the estimated variances on these coefficients in (4) differ, owing to the presence of unnecessary dummy variables in the IFE specification that artificially increases the fit of the regression. This is expressed in the following

Proposition 1. *In a treatment effect model as in eq. (2), the estimated variance of the treatment effect coefficient is downward biased under IFE relative to FET. The bias is equal to the ratio of the estimated residual variances under IFE and FET:* $\frac{\hat{\sigma}^2(b^I)}{\hat{\sigma}^2(b^G)} = \frac{\hat{\sigma}^2_{u,I}}{\hat{\sigma}^2_{u,G}}$.

Proof. It suffices to show that $S^I_{n+1,n+1} = S^G_{3,3}$, i.e. the last elements on the main diagonal of the inverted product sum matrices in eqs. (3) and (4) are identical. By elementary operations, $X^{I'}X^I = T \cdot I_{Tn\times Tn}$. Hence, under Individual Fixed Effects:

$$[X^I D]'[X^I D] = \begin{pmatrix} T \cdot I_{Tn\times Tn} & \tau \cdot \Delta \\ \tau \cdot \Delta' & \tau \cdot d \end{pmatrix}$$

where, as defined further above, $d = tr(\Delta) = tr(\Delta'\Delta)$. Inverting this partitioned matrix, we find for the (n+1,n+1)-element of the inverse:

$$([X^I D]'[X^I D])^{-1}_{n+1,n+1} = (\tau d - \tau\Delta' \cdot \frac{1}{T}\Delta\tau)^{-1} = \frac{T}{\tau d(T-\tau)} \tag{5}$$

Under Fixed Effects on the Treated, the product sum matrix becomes:

$$X^{G'}X^{G} = T \cdot \begin{pmatrix} n & d \\ d & d \end{pmatrix}$$

Hence,

$$[X^{G}D]'[X^{G}D] = \begin{pmatrix} T \cdot n & T \cdot d & \tau \cdot d \\ T \cdot d & T \cdot d & \tau \cdot d \\ \tau \cdot d & \tau \cdot d & \tau \cdot d \end{pmatrix}$$

Inverting this partitioned matrix, we find for the element (3,3) of the inverse:

$$([X^{G}D]'[X^{G}D])^{-1}_{3,3} = \left[\tau d - \tau (d \quad d)(X^{G'}X^{G})^{-1}\tau \begin{pmatrix} d \\ d \end{pmatrix}\right]^{-1} \tag{6}$$

Using

$$(X^{G'}X^{G})^{-1} = \frac{1}{Td(n-d)} \begin{pmatrix} d & -d \\ -d & n \end{pmatrix}$$

this becomes:

$$\begin{aligned} ([X^{G}D]'[X^{G}D])^{-1}_{3,3} &= \left[\tau d - \tfrac{1}{Td(n-d)}\,\tau (d \quad d) \begin{pmatrix} d & -d \\ -d & n \end{pmatrix} \tau \begin{pmatrix} d \\ d \end{pmatrix}\right]^{-1} \\ &= \left[\tau d - \tfrac{1}{Td(n-d)}\,\tau (d \quad d) \begin{pmatrix} 0 \\ \tau \cdot d(n-d) \end{pmatrix}\right]^{-1} \qquad (7) \\ &= \left[\tau d - \tfrac{\tau^2 d}{T}\right]^{-1} = \frac{T}{\tau d (T - \tau)} \end{aligned}$$

(7) is equal to (5), which completes the proof.

□

In applied work, the possible correlation of additional characteristics Z with XD means the above relation no longer obtains exactly. Unless, however, this correlation amounts to near-collinearity, its effect is small relative to the overfitting effect described in the proposition.

4 Conclusion

Applications of fixed effect and difference in differences estimators sometimes employ individual, observation-unit specific fixed effects when group-specific fixed effects would suffice for identification. This note has examined the properties of difference in differences estimators of treatment effects under two different fixed effects specifications. It shows that overfitting under individual, observation-unit specific fixed effects generates lower standard errors on the treatment effect coefficient than estimation under a minimal specification with group specific effects. Depending on the correlation with other regressors, this bias grows at or near the relative decrease of the residual sum of squares as the number of overfitted fixed effects increases. In large samples, which are frequent in evaluation studies, this overfitting bias may lead to substantial underestimation of the standard errors on treatment effect coefficients, and hence to substantial false positives.

References

Angrist, J., and S. Pischke (2009): *Mostly Harmless Econometrics.* Princeton University Press.

Ashenfelter, O. (1978): “Estimating the Effect of Training Programs on Earnings,” *Review of Economics and Statistics*, 60, 47–57.

Ashenfelter, O., and D. Card (1985): “Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs,” *Review of Economics and Statistics*, 64, 648–660.

Bertrand, M., E. Duflo, and S. Mullainathan (2004): “How Much Should We Trust the Differences-in-Differences Estimator?,” *Quarterly Journal of Economics*, 119, 249–275.

Heckman, J., R. Lalonde, and J. Smith (1999): “The Economics and Econometrics of Active Labor Market Programs,” in *Handbook of Labor Economics*, ed. by O. Ashenfelter, and D. Card, pp. 1865–2033. Elsevier.