

HOW MUCH SHOULD WE TRUST DIFFERENCES-IN-DIFFERENCES ESTIMATES?*

MARIANNE BERTRAND, ESTHER DUFLO, SENDHIL MULLAINATHAN

The main motivation of the paper is analyzing the efficiency of D-i-D method used in the literature. D-i-D method becomes popular among the scholar who tries to estimate an effect of a treatment over time on a treatment group. The appeal of the method stems on the fact that it is *easy to use* and it *alleviates the endogeneity* problem. However the D-i-D estimation suffers from *biasness problem* and *standard error problem*. In this paper, Bertrand et al focus on the **standard error problem** rather than the biasness problem. The standard errors to construct the confidence interval of the coefficients are obtained by OLS estimation but it is efficient under very restrictive assumptions such that the changes in dependent variable over time would have been same in both treatment and control groups if there were no treatment dummy. The reason that the OLS standard errors are inefficient is that the OLS regression generally suffers from severe **serial correlation**. That is the error term this period is highly correlated with error term last period. Serial correlation is especially important in D-i-D context because first these estimation generally rely on long time series; second the dependent variables used in the literature is generally highly serially correlated and finally the dummy variable changes little over time. Because of inefficiency problem in standard errors, authors concludes that most of the research fall into the mistake of **over rejecting** (over estimation of t-stats) of null hypothesis of no effect of treatment.

He surveyed 92 published papers in six journals from 1990 to 2000 and found 65 of them with severe serial correlation problem and only 5 of these papers use some correction for the problem. To show the over rejection of serially correlated data, the author used Current Population Survey and introduced an imaginary dummy for treatment. The OLS estimation gives the coefficient of the treatment as well as the standard error which in return is used for the t-statistics. Then he repeated this methodology over a large number of times. He concluded that if “OLS gives a consistent standard error, we should expect to reject null hypothesis of no effect ($\beta=0$) roughly 5 % of the time when a threshold of 1.96 t-stats is used.” However the OLS rejected the null

hypothesis of no effect, stunningly, 67.5 percent of the time. This over rejection is due to the failure of OLS to account for correlation within states. The authors then used arbitrary correlation of error terms at the state-year level as well as aggregating the data into state level and estimate 21 years of panel data, but none of these methods yields expected standard errors. In addition the authors also proved that OLS estimates are consistent, if the error terms are not serially correlated. In order to show this the authors pick the years from 1979-1999 randomly which assures that dummy variable is repeatedly turned on and off and its value in one year tells one nothing about the next year's value. Even using the AR(1) correction, the authors find 37% rejection rate of hypothesis of no impact of dummy.

The solutions that the authors are offering can be classified into five: *parametric* and block bootstrap, ignoring time series info, empirical var-cov matrix and finally arbitrary var-cov matrix. First parametric solution would be to specify an autocorrelation structure for error term then estimate its parameters and by using these estimated parameters find the standard errors. However this parametric solution does little help to the problem. The authors believe that the failure of this parametric correction is in part due to the downward bias in the estimator of the autocorrelation coefficient. Second, block bootstrap method presents a major improvement over the parametric techniques, however the method performs less well when the number of states declines, as well as the power of this test also declines fast. In the third method, time series info is ignored by averaging the before and after intervention and constructing a panel of length two. This simple procedure gives very good rejection rates for null hypothesis, 5.3%. The downside of the procedure is its power is low and diminishes fast with sample size. If one assumes that the autocorrelation process is the same across all the states and that there is no cross sectional heteroscedasticity, empirical var-cov matrix can be used to estimate consistent standard errors as number of groups goes to infinity. However this method performs poorly on small sample sizes. Finally, arbitrary var-cov matrix method relaxes the assumption of no cross sectional heteroscedasticity. The authors suggest using White-like formula to calculate the standard errors. The estimator is consistent for fixed panel length as the number of states goes to infinity. However again the rejection rates increase as sample size decreases.

Personally after reading all the methods to account for auto correlation problem, I prefer to use ignoring time series information. It seems easy to implement such that dividing the data into before and after intervention and averaging the states to construct panel data of length two seems to offer no difficulty. However then I believe we may loose some important information that time series data reveal such as the dynamics of the pattern etc.