A Study on Program Evaluation via Difference-In-Difference (DID) Approach

Bonyoung Koo*　　Heungsun Park†

Abstract
Difference-In-Difference (DID) has been widely used as a program evaluation method in many disciplines such as econometrics, social studies, and education or social policies. However, when the lagged dependent variable is considered as a predictor, the impact size is often misinterpreted. This study illustrates how DID estimators of impact size are related to the conventional interaction effect models and time series models, and the appropriate estimation and testing methods are suggested.

Key Words: Quasi-experiment, Interaction Effect, Lagged Dependent Variable, Intervention Model.

1. Motivation

A program evaluation or intervention needs to be estimated or tested with before and after experiment data sets. As many social experiments cannot be controlled by complete randomization due to ethical reasons, control-treatment groups are influenced by selection bias (e.g. the unemployed who voluntarily participated in job-training programs are more likely to have better job opportunities regardless of the program participation).

So, the real impact size is the difference between control and treatment groups after the intervention minus the difference between control and treatment groups before the intervention, which is illustrated on Figure 1.

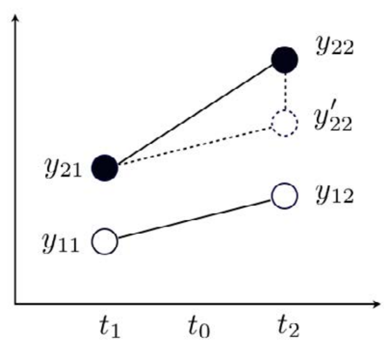

Figure 1: The Difference-In-Differences estimation of the training program effect between the treatment group (black circles) and the control group (white circles).

Figure 1 illustrates the program effect between the treatment group which is the program participant group (black circles) and the control group which is the program nonparticipant group (white circles) when the program intervention occurs at time point t_0. In this, the actual program effect size is not measured by the difference, $y_{22} - y_{12}$, but $y_{22} - y'_{22}$ where y'_{22} is the counterfactual point which would have happened to the treatment group if

*Assistant Manager, KB Insurance, Hapjeong-dong Mapo-gu, Seoul, South Korea
†Professor, Dept. of Statistics, Hankuk University of Foreign Studies, 81 Oedae-ro, Yongin-si, Gyeonggi-do, 17035, South Korea

there is no intervention. Therefore, the program effect size, $y_{22} - y'_{22}$, can be rewritten as the difference in two differences as follows:

$$\begin{aligned} y_{22} - y'_{22} &= (y_{22} - y_{12}) - (y_{21} - y_{11}) \\ &= (y_{22} - y_{21}) - (y_{12} - y_{11}) \end{aligned} \tag{1}$$

This Difference-In-Differences method was developed by Ashenfelter (1978), who served as Director of the Office of Evaluation in the U.S. Department of Labor and was interested whether a training program would increase earnings. In his study, he compared the earnings between the trainees and the non-trainees after the training. Since then, many researchers have used this method in different disciplines (Moffitt, 1991; Henchman and Robb, 1985). In this study, we are going to review the DID method from several statistical modeling point of views and clarify the relationship between the interaction term and the impact size, so that people may not be misled to interpret the coefficients.

1.1 Interaction Term in Factorial Designs

Figure 1 illustrates the DID estimate when we have only two data points. However, we could have several data points before and after intervention in program evaluation experiments, which leads us to the factorial design framework with an interaction term.

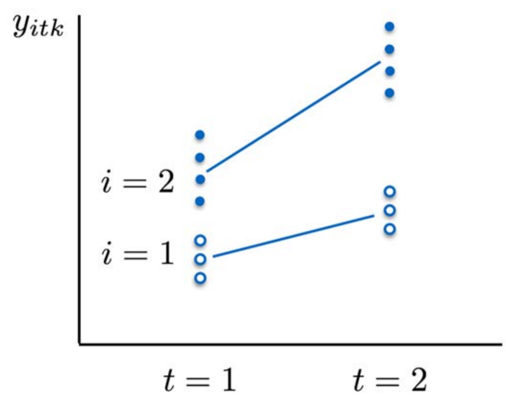

Figure 2: The Difference-In-Differences estimation in Factorial Design framework between the treatment group (black circles) and the control group (white circles).

The model equation can be written as

$$y_{itk} = \mu + \tau_i + \delta_t + (\tau\delta)_{it} + \epsilon_{itk} \tag{2}$$

or, equivalently, if the corresponding indicator functions used,

$$y_{itk} = \beta_0 + \beta_1 I_T(t) + \beta_2 I_G(i) + \beta_{12} I_T(t) \times I_G(i) + \epsilon_{itk} \tag{3}$$

where $I_T(t)$ is an indicator function whose value is 1 for $t = 2$ or 0 for $t = 1$, and $I_G(i)$ is an indicator function whose value is 1 for $i = 2$ or 0 for $i = 1$.

The DID estimator in this example is the interaction term, β_{12}, because

$$\begin{aligned} (E(y_{22k}) - E(y_{12k})) - (E(y_{21k}) - E(y_{11k})) &= (\beta_2 + \beta_{12}) - (\beta_2) \\ &= \beta_{12} \end{aligned} \tag{4}$$

so we could estimate the impact size as β_{12} and do the significance t-test for it using

$$t_0 = \frac{\hat{\beta}_{12}}{SE(\hat{\beta}_{12})} \sim t_{n-4}, \qquad \text{if} \quad H_0 : \beta_{12} = 0 \quad \text{is true.} \tag{5}$$

1.2 Interaction Term in Random Effect Model

The previous factorial framework ignores the correlation within a subject. With considering each subject factor as a random component, we could get around of this correlation problem. In this, the interaction coefficient can be assessed in random effect models with the same analogy.

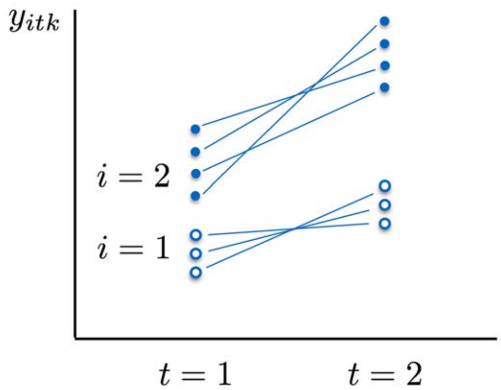

Figure 3: The Difference-In-Differences estimation in random effect model framework between the treatment group (black circles) and the control group (white circles).

The model equation is

$$\begin{aligned} y_{itk} &= \beta_0 + \beta_1 I_T(t) + \beta_2 I_G(i) + \beta_{12} I_T(t) \times I_G(i) + \rho_{ik} + \epsilon_{itk}, \\ \rho_{ik} &\sim \text{i.i.d. } N(0, \sigma_\rho^2), \\ \epsilon_{itk} &\sim \text{i.i.d. } N(0, \sigma^2), \end{aligned} \tag{6}$$

and the DID impact size can be written as

$$\begin{aligned} (E(y_{22k}) - E(y_{12k})) - (E(y_{21k}) - E(y_{11k})) &= (\beta_2 + \beta_{12}) - (\beta_2) \\ &= \beta_{12} \end{aligned} \tag{7}$$

which is estimated by the popular GLS-REML algorithm in a linear mixed model.

2. Time Series with Shifts

The previous examples demonstrate that the intervention impact size in program evaluations can be measured by the interaction term in the corresponding models. However, it is no longer true in more complicated models, specially when the lagged dependent variable is considered as a predictor.

One of the examples is a time series model with shifts. There could be sequences of observations before and after the program occurred, in which time series models might be applied with different shifts.

The autoregressive model of order 1, AR(1), could be fitted with different shifts when there is assumed to have an intervention, and the model equation is

$$y_{it} = \beta_0 + \beta_1 I_T(t) + \beta_2 I_G(i) + \beta_{12} I_T(t) \times I_G(i) + \phi y_{it-1} + \epsilon_{it}. \tag{8}$$

However, the DID estimator for the impact size is not β_{12}, the interaction coefficient any more as expected before, but

$$\frac{\beta_{12}}{1-\phi} \tag{9}$$

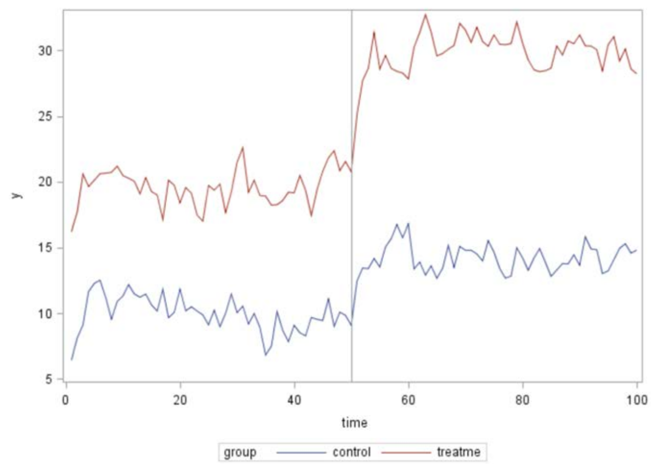

Figure 4: The Difference-In-Differences estimation in an AR(1) time series framework between the treatment group (black circles) and the control group (white circles).

because DID impact size is

$$(E(y_{2t}) - E(y_{1t})) - (E(y_{2t-1}) - E(y_{1t-1})) = \frac{\beta_{12}}{1-\phi}. \tag{10}$$

Therefore, the relevant significance test for the impact size should be based on $\frac{\beta_{12}}{1-\phi}$ not the interaction coefficient, β_{12}.

3. Time Series with Trends

Time series data are more than often accompanied with time trends. In this, we could use the differencing technique, $Z_t = Y_{t-1} - Y_t$, in order to eliminate the time trend. For the

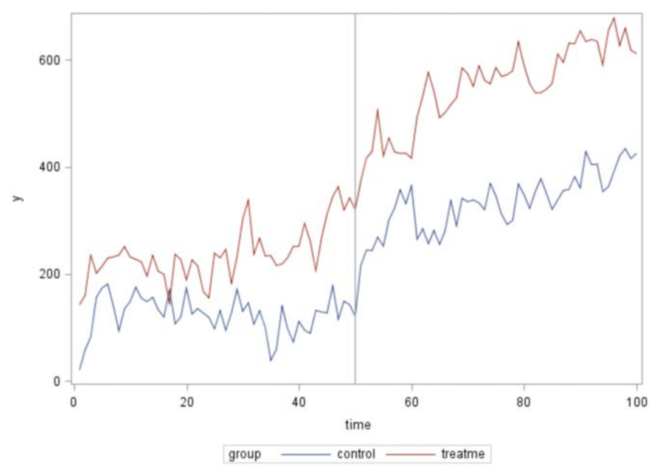

Figure 5: The Difference-In-Differences estimation in an ARIMA(1,1,0) time series framework between the treatment group (black circles) and the control group (white circles).

control group,

$$\begin{aligned} y_{it} &= \beta_0 + \beta_1 I_T(t) + \beta_2 I_G(i) + \beta_{12} I_T(t) \times \\ &\quad + \{\alpha_{i1} I(t < t_0) + \alpha_{i2} I(t \geq t_0)\}t + \phi y_{it-1} + \epsilon_{it} \end{aligned} \tag{11}$$

which means, if $z_{it} = y_{it} - y_{it-1}$,

$$z_{it} = \{\alpha_{i1} I(t < t_0) + \alpha_{i2} I(t \geq t_0)\} + \phi z_{it-1} + \epsilon_{it} \tag{12}$$

therefore, the DID impact size is

$$\frac{1}{1-\phi}\{(\alpha_{22} - \alpha_{12}) - (\alpha_{21} - \alpha_{11})\} \tag{13}$$

not the interaction term β_{12}.

4. Conclusion

Difference-In-Difference (DID) estimations are discussed and the possible estimation and testing methods are illustrated. The interaction term is not always representing the impact size as we presume. The proper alternative functionals were presented.

REFERENCES

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

Hechman, J.J. and Robb, R. (1985), “Alternative Methods for Evaluating the Impact of Interventions”, *Journal of Econometrics*, 30, 239-267.

Moffitt, R. (1991) “Program Evaluation with Nonexperimental Data”, *Evaluation Review*, Vol 15, No 3, 291-314.