# To be, or not to be: Extending Difference-in-Differences to Binary Outcomes

Alonso M. Guerrero Castañeda

Duke University

April 5th, 2024

# Roadmap

# Introduction

- Difference-in-differences (DiD) is widely used in econometrics and causal inference to address confounding
- Traditional DiD assumes continuous outcomes and uses linear regression, which may introduce bias with binary outcome
- Objective: Propose alternative DiD estimator using logistic regression to handle binary outcomes
- Simulations: Compare different estimation methods and explore their relative performance

# Roadmap

Introduction

Literature Review

Methodology

Simulation

Conclusion

# Literature Review

- Resurgence in methodological research on DiD: challenges and extensions
  - Eg. Callaway & Sant'Anna (2021) incorporated multiple time periods $\rightarrow 4,200$ citations

- DiD with binary outcomes remains relatively underexplored

- Linear regression is dominant
  - Gomila (2021) even cautioned against using logit or probit models

- Lechner (2010) proposed 2 approaches to address nonlinear models in DiD:
  1. Nonlinear models with standard trend assumption (similar to Li & Li, 2019)
  2. Nonlinear models with a modified trend assumption (similar to Wooldridge, 2023).

- Our framework will be built upon Li and Li (2019)

# Roadmap

# Set-Up

- 2 periods: before ($t$) and after treatment ($t+1$)
- 2 groups: control ($G=0$) and treatment ($G=1$)
- $\theta_1 = \mathbb{E}[Y_{i,t+1}(1)|G_i = 1]$ and $\theta_0 = \mathbb{E}[Y_{i,t+1}(0)|G_i = 1]$.
- Quantity of interest: average treatment effect on the treated (ATT)

$$\tau = \theta_1 - \theta_0$$

# Estimating Assumptions

- Under SUTVA, $\theta_1$ is non-parametrically identified as $\theta_1 = \mathbb{E}[Y_{i,t+1} | G_i = 1]$
- $\theta_0$ is not directly observed. We need the **Parallel Trend Assumption:**

$$\mathbb{E}[Y_{i,t+1}(0) - Y_{i,t}(0) | X_i, G_i = 1] = \mathbb{E}[Y_{i,t+1}(0) - Y_{i,t}(0) | X_i, G_i = 0]$$

- Applying law of total expectations,

$$\theta_0 = \mathbb{E}[Y_{i,t} | G_i = 1] + \mathbb{E}_X[\mathbb{E}[Y_{i,t+1}(0) - Y_{i,t}(0) | X_i, G_i = 0] | G_i = 1]$$

First term (directly observed) can be consistently estimated with $\frac{\sum G_i Y_{i,t}}{\sum G_i}$. Second term requires a model specification

# Binary Outcome Modelling and ATT Estimation

- Asume Bernoulli models:

$$Y_{i,t}(0)|X_i, G_i = 0 \sim \text{Bernoulli}\left(\mu = \frac{1}{1+\exp(-X'\beta)}\right)$$

$$Y_{i,t+1}(0)|X_i, G_i = 0 \sim \text{Bernoulli}\left(\nu = \frac{1}{1+\exp(-X'\gamma)}\right)$$

- Logistic regression for estimating $\hat{\beta}$ and $\hat{\gamma}$. Thus,

$$\hat{\theta}_0^{reg} = \frac{\sum_i G_i Y_{i,t}}{\sum_i G_i} + \frac{\sum_i G_i\{\nu(X_i, \hat{\gamma}) - \mu(X_i, \hat{\beta})\}}{\sum_i G_i}$$

- If model is correctly specified, $\hat{\theta}_0^{reg}$ is consistent. If so, $\hat{\tau}^{reg} = \hat{\theta}_1 - \hat{\theta}_0^{reg}$ is also consistent

# Roadmap

# Simulating Covariates and Group Assignment

- Each simulation has $N = 1,000$ units
- Each unit has a binary covariate $X_1$ and a continuous covariate $X_2$:

$$X_1 \sim \text{Bernoulli}(0.25), \quad X_2 | X_1 \sim \text{Normal}(2 + 6X_1, 2^2)$$

- Group assignment $G_i$ follows Bernoulli distribution, with propensity score:

$$\text{logit}\{e(\mathbf{X})\} = -2 + X_1 - 0.2X_2 + 0.04X_2^2$$

# Simulating Binary Outcomes

- 200 replicates based on the models below (200 for each $k$)

$$\begin{aligned} Y_t(0)|\mathbf{X}, G = 0 &\sim \text{Bernoulli}\left(\mu_{00}(\mathbf{X})\right), \\ Y_t(0)|\mathbf{X}, G = 1 &\sim \text{Bernoulli}\left(\mu_{01}(\mathbf{X})\right), \\ Y_{t+1}(0)|\mathbf{X}, G = 0 &\sim \text{Bernoulli}\left(\nu_{00}(\mathbf{X})\right), \\ Y_{t+1}(1)|\mathbf{X}, G = 1 &\sim \text{Bernoulli}\left(\nu_{11}(\mathbf{X})\right), \end{aligned}$$

- Mean Functions ($k$ takes values from $-9$ to 3):

$$\begin{aligned} \mu_{00}(\mathbf{X}) &= \text{expit}\left(k + 0.5X_1 + 0.05X_2\right) \\ \mu_{01}(\mathbf{X}) &= \text{expit}\left(k - 0.5 + 0.5X_1 + 0.05X_2\right) \\ \nu_{00}(\mathbf{X}) &= \text{expit}\left(k - 0.5 + 0.5X_1 + 0.05X_2\right) \\ \nu_{11}(\mathbf{X}) &= \text{expit}\left(k + 0.5X_1 + 0.05X_2\right) \end{aligned}$$

# Outcome Approaches to Estimation: Linear Regression

- *Linear regression approach*: standard additive fixed effects regression model

$$Y_{it} = \alpha + \gamma G_i + \delta_T + \tau G_i \mathbb{1}\{T = t + 1\} + \beta \mathbf{X_i} + \epsilon_{iT}, \quad \epsilon_{iT} | G_i, T, X_i \sim N(0, \sigma^2).$$

  $\hat{\tau}^{lm}$ and its 95% CI are directly estimated by running the linear regression based on the model above

# Outcome Approaches to Estimation: Logistic Regression

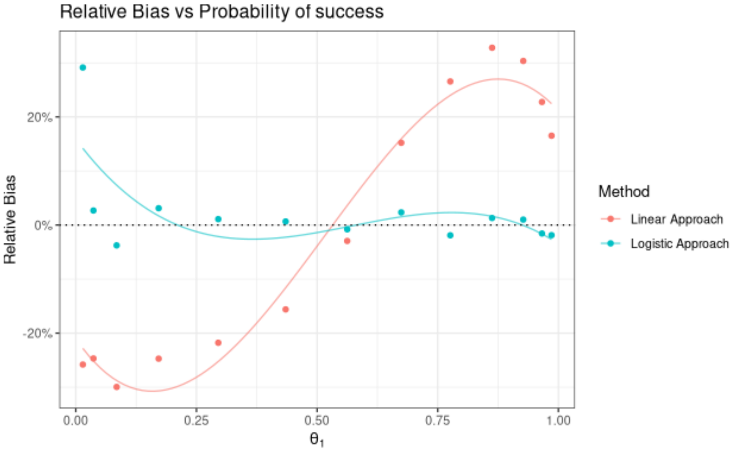
- *Logistic regression*: proposed estimator with correctly specified mean functions $\mu(\mathbf{X})$ and $\nu(\mathbf{X})$, each estimated using logistic regression

$$\hat{\theta}_0^{reg} = \frac{\sum_i G_i Y_{i,t}}{\sum_i G_i} + \frac{\sum_i G_i \{\nu(X_i, \hat{\gamma}) - \mu(X_i, \hat{\beta})\}}{\sum_i G_i}$$

$\hat{\theta}_1$ is directly observed. Thus, $\hat{\tau}^{reg} = \hat{\theta}_1 - \hat{\theta}_0^{reg}$. Nonparametric bootstrap to obtain 95% CI

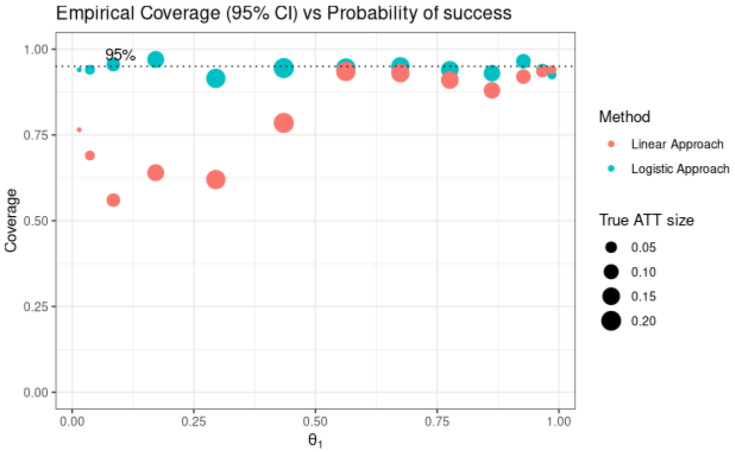
# Results: Relative Bias

- Logistic: unbiased, except for extreme probabilities
- Linear: under-estimation and over-estimation are highest around 10% and 90%, respectively. Unbiased around 55%



# Results: Coverage

- Logistic: coverage is fairly stable for the logistic approach (between 90% and 100%)
- Linear: coverage is unstable. It is competitive around 55%



# Roadmap

# Conclusion

- We proposed a DiD extension to handle binary outcomes
- Logistic regression approach offers robust alternative to linear regression
- Importance of considering appropriate estimation techniques when applying DiD to settings with binary outcomes
- Future research: double-robust estimation approach for binary outcomes