Simple Approaches to Nonlinear Difference-in-Differences with Panel Data

Jeffrey M. Wooldridge
Department of Economics
Michigan State University

This version: January 1, 2023

Online Appendix

Simulations for Binary Responses

The potential outcomes, $Y_t(0)$ and $Y_t(1)$, are binary. There are $T = 6$ time periods with a common intervention time at $q = 4$. In generating the data, the difference in the means of the underlying latent variables, $E[Y_t^*(1)] - E[Y_t^*(0)]$, changes across $t = 4$, 5, and 6. The single (time-constant) covariate, X, is generated as the average, over the six time periods, of independent exponential random variables with unit means. Treatment status, D, follows a logit:

$$D = 1[-0.5 + (X - 1) + V > 0], \tag{o.1}$$

where V is independent of X with a logistic distribution. This mechanism is used for all simulations and implies $P(D = 1) \approx 0.382$. For the first simulation, the potential outcomes are generated as

$$Y_t(0) = 1[(X - 1)/2 - 2D + (X - 1) \cdot D/4 + U_t(0) > 0], t = 1, \ldots, T, \tag{o.2}$$

where the $U_t(0)$ follow logistic distributions with a common effect C and independent idiosyncratic shocks. This mechanism ensures the potential outcome $Y_t(0)$ is correlated with X as well as with treatment status, D, and that $U_t(0)$ has substantial serial correlation. In the treated state, $Y_t(1) = Y_t(0)$ for $t \leq 3$ to ensure no anticipation. Then

$$Y_t(1) = 1[0.5 + (X - 1) - 2D + 0.2 \cdot f5_t + 0.3 \cdot f6_t + U_t(1) > 0], t = 4, \ldots, T, \tag{o.3}$$

where $U_t(1)$ depends on the same heterogeneity as $U_t(0)$ with separate independent shocks and also has a logistic distribution. This formulation implies that the index inside the logit function has a time-varying treatment effect, although that effect does not vary with X. The treatment effects on the binary outcomes do vary with X because of the nonlinearity. The

parallel trends assumption, conditional on X, is in force for the linear index.

The data generating mechanism implies $T^{-1}\sum_{t=1}^{T} P[Y_t(0) = 1] \approx 0.316$ and $T^{-1}\sum_{t=1}^{T} P[Y_t(1) = 1] \approx 0.405$, and so there is reasonable balance in both potential outcomes. The ATTs – the target parameters – are obtained by averaging across the 5,000 Monte Carlo replications. The logit conditional mean is estimated using pooled logit and the linear mean is estimated using pooled OLS. The R-squared from the linear POLS estimation is about 0.205, which seems realistic given that D, year dummies, and the covariate are controlled for flexibly.

The method proposed by Callaway and Sant'Anna (2021), where the comparison group is the never treated group (the default in Stata), is also included in the simulation . The two tests for pre-trends are computed for the pooled logit and pooled OLS estimation. The findings are reported in Table A.1.

Table A.1: **Binary Response**, **Common Timing**, **Logistic Errors**

5,000 Replications	Sample ATT	Logit (Pooled Bernoulli)		Linear (Pooled OLS)		CS (2021)	
$N = 500$	Mean	Mean	SD	Mean	SD	Mean	SD
τ_4	0.060	0.060	0.021	0.061	0.035	0.061	0.042
τ_5	0.094	0.093	0.024	0.094	0.036	0.094	0.043
τ_6	0.113	0.112	0.026	0.113	0.038	0.113	0.044
Event Study Rejection Rate (2 df)	——	0.044		0.044		——	
Heterogeneous Trend Test (1 df)	——	0.046		0.042		——	

In this first scenario, the relative performance of the estimators is clear. Even with only 500 cross-sectional units, all three approaches show little bias when compared with the sample ATTs (which are averaged across the 5,000 simulations and should be close to the population ATTs). In terms of precision, the correctly specified logit is much preferred, having Monte Carlo standard deviations no more than 69% of the linear model SDs. Not surprisingly, the CS (2021) estimators, which use only the single period prior to the intervention as the control

period, are the least precise. Given that the logit model is correctly specified we hope that the parallel trends tests in the logit estimation have rejection frequencies close to the nominal size, 0.05, and that is the case. In principle, because the linear model is misspecified, the parallel trends tests could reject very often – but they also reject right around a 5% rate. In this scenario, that is not a bad thing because the linear model estimates show essentially no bias.

In other scenarios, the approaches based on the linear PT assumption can be badly biased. Table A.2 reports on a simulation where the potential outcomes are generated as

$$Y_t(0) = 1[0.4 \cdot f4_t + 0.5 \cdot f5_t + 0.6 \cdot f6_t + (X-1)/2 - 2D + U_t(0) > 0], t = 1, \ldots, T \quad \text{(o.4)}$$

$$Y_t(1) = 1[0.9 \cdot f4_t + 1.1 \cdot f5_t + 1.3 \cdot f6_t + (X-1)/2 - 2D + U_t(1) > 0], t = 4, \ldots, T \quad \text{(o.5)}$$

The ATTs in this case are time varying and increasing. Again, the index satisfies the conditional PT assumption. In this case the average response probabilities are slightly larger than before: $T^{-1}\sum_{t=1}^{T} P[Y_t(0) = 1] \approx 0.385$ and $T^{-1}\sum_{t=1}^{T} P[Y_t(1) = 1] \approx 0.459$.

The simulation findings are striking. The pooled logit estimation shows very little bias. (Simulation results for the probit estimator show a slight downward bias, but well within the acceptable range.) Not only are the pooled OLS and CS (2021) estimates badly biased, they are, on average, actually negative, whereas the SATTs are positive and practically large. To compound matters, the tests of parallel trends have no power for detecting the misspecification apparent in using linear methods. A topic for future research is to obtain diagnostics that would allow one to reject the misspecified linear model so that one is led to a nonlinear model.

Table A.2: **Binary Response**, **Common Timing**, **Logistic Errors**							
5,000 Replications	Sample ATT	Logit (Pooled Bernoulli)		Linear (Pooled OLS)		CS (2021)	
$N = 500$	Mean	Mean	SD	Mean	SD	Mean	SD
τ_4	0.078	0.078	0.027	–0.049	0.036	–0.049	0.043
τ_5	0.114	0.113	0.031	–0.037	0.038	–0.037	0.044
τ_6	0.158	0.157	0.036	–0.013	0.041	–0.012	0.046
Event Study Rejection Rate (2 df)	——	0.040		0.048		—	
Heterogeneous Trend Test (1 df)	——	0.050		0.042		—	

In a third scenario, the errors $U_t(0)$ and $U_t(1)$ are generated to have $Uniform(-2,2)$ distributions, where each depends on the common heterogeneity, C, as before. The potential outcomes are generated as in (o.2) and (o.3). Now, $T^{-1}\sum_{t=1}^{T} P[Y_t(0) = 1] \approx 0.320$ and $T^{-1}\sum_{t=1}^{T} P[Y_t(1) = 1] \approx 0.398$. When the support in the uniform distribution is sufficiently wide, the linear model is essentially correctly specified. Therefore, we might expect the the linear model to perform relatively well in this setting. The findings are reported in Table A.3.

Table A.3: **Binary Response**, **Common Timing**, **Uniform Errors**							
5,000 Replications	Sample ATT	Logit (Pooled Bernoulli)		Linear (Pooled OLS)		CS (2021)	
$N = 500$	Mean	Mean	SD	Mean	SD	Mean	SD
τ_4	0.110	0.110	0.025	0.111	0.038	0.111	0.044
τ_5	0.159	0.159	0.028	0.159	0.040	0.159	0.046
τ_6	0.183	0.183	0.030	0.183	0.041	0.183	0.047
Event Study Rejection Rate (2 df)	——	0.041		0.046		—	
Heterogeneous Trend Test (1 df)	——	0.047		0.048		—	

While all methods show little bias, the logit estimator is substantially more efficient than the linear model or CS estimates. Evidently, the ramp function is well approximated by the logistic function, leading to better precision (and essentially no bias). As before, the parallel trends tests reject right around 5% of the time even though, technically, both the linear and logit models are misspecified. And, again, this is a good outcome because the methods are doing

very well for uncovering the ATTs.

In the final simulation for the binary response case, the potential outcomes are generated as in (o.4) and (o.5) with the errors having $Uniform(-2,2)$ distributions. The summary statistics in this simulation do not seem unusual: $T^{-1}\sum_{t=1}^{T} P[Y_t(0) = 1] \approx 0.375$, $T^{-1}\sum_{t=1}^{T} P[Y_t(1) = 1] \approx 0.450$, and the R-squared from the pooled LPM estimation is about 0.207, on average. Because D is generated as before, $P(D = 1) \approx 0.382$. And yet, as shown in Table A.4, the logit model now fares considerably worse than the linear model and CS approach, exhibiting a severe upward bias in each of the three ATTs. Both the linear regression and CS estimates have slight downward biases.

Table A.4: Binary Response, Common Timing, Uniform Errors

5,000 Replications	Sample ATT	Logit (Pooled Bernoulli)		Linear (Pooled OLS)		CS (2021)	
$N = 500$	Mean	Mean	SD	Mean	SD	Mean	SD
τ_4	0.125	0.197	0.031	0.111	0.041	0.111	0.047
τ_5	0.150	0.243	0.033	0.135	0.043	0.135	0.048
τ_6	0.174	0.287	0.035	0.160	0.044	0.160	0.048
Event Study Rejection Rate (2 df)	——	0.040		0.049		——	
Heterogeneous Trend Test (1 df)	——	0.051		0.047		——	

Simulations are necessarily special and not always realistic. For example, a uniform distribution for the composite error term, where the response probability can reach zero and one, does not seem particularly plausible. The purpose of these simulations is to show that one can obtain very different estimates depending on functional form. With at least one continuous covariate one can at least explore goodness-of-fit as a possible way of choosing among different conditional mean models. Unfortunately, the parallel trends tests do not provide guidance in the simulation settings reported in Tables A.2 and A.4.

In some sense, simulations over many different scenarios can be viewed as exploring the

limits of the bounds on the treatment effects of the kind derived in Athey and Imbens (2006) when $T = 2$. In empirical practice, one can try a linear analysis along with a sensible nonlinear model, such as logit, and hopefully draw robust conclusions. Of course, one can replace logit with probit or even complementary log-log in a robustness check.

A.2. Simulations for Nonnegative Responses

In the first case, $Y_t(\infty)$ has a Poisson distribution conditional on unobserved heterogeneity. In particular,

$$Y_t^*(\infty) = 2 + 0.2 \cdot f2_t + 0.3 \cdot f3_t + 0.4 \cdot f4_t + 0.5 \cdot f5_t + 0.6 \cdot f6_t + X/5 - (D_4 + D_5 + D_6) + C$$

$$Y_t(\infty) \sim Poisson(\exp(Y_t^*(\infty)))$$

where X is generated the average over time of six independent exponential random variables with unit means and $C \sim Normal(0,1)$. The distribution of $Y_t^*(\infty)$ conditional only on the cohort indicators is a mixture of a Poisson and lognormal random variable. The outcomes in the treated states are generated as

$$\begin{aligned} Y_t(4) &\sim Poisson(\exp(Y_t^*(\infty) + (X - 1/5 + 0.4 \cdot f4_t + 0.8 \cdot f5_t + f6_t)), t \geq 4 \\ Y_t(5) &\sim Poisson(\exp(Y_t^*(\infty) + (X - 1/5 + 0.6 \cdot f5_t + f6_t)), t \geq 5 \\ Y_t(6) &\sim Poisson(\exp(Y_t^*(\infty) + (X - 1/5 + 0.4 \cdot f6_t)), t = 6 \end{aligned}$$

The treatment cohorts are generated using an ordered probit, resulting in the following (approximate) shares:

$$P(D_\infty = 1) \approx 0.357,\ P(D_4 = 1) \approx 0.291$$
$$P(D_5 = 1) \approx 0.225,\ P(D_6 = 1) \approx 0.127$$

The population R-squared in the pooled OLS estimation is about 0.126 and $T^{-1}\sum_{t=1}^{T} P(Y_t(\infty) = 0) \approx 0.058$ (so zero is not a dominant outcome). The results are reported

in Table A.5, where the sample ATTs are averaged across the 5,000 replications and 500 observations.

Table A.5: Count Outcome, Staggered Intervention							
5,000 Replications	Sample ATT	Exponential (Pooled Poisson)		Linear (Pooled OLS)		CS (2021)	
$N = 500$	Mean	Mean	SD	Mean	SD	Mean	SD
τ_{44}	4.07	4.06	0.59	2.64	0.80	2.72	0.89
τ_{45}	11.13	11.13	1.34	8.20	1.71	8.32	1.70
τ_{46}	17.25	17.25	2.01	11.86	2.59	12.80	2.48
τ_{55}	8.12	8.12	1.20	5.72	1.55	6.62	1.55
τ_{56}	18.43	18.43	2.51	13.33	3.12	15.26	2.94
τ_{66}	6.42	6.41	1.35	1.64	2.27	4.67	2.07
Event Study Rejection Rate (9 df)	—	0.099		0.995		—	
Heterogeneous Trend Test (3 df)	—	0.069		0.999		—	

The clear-cut winner in this simulation is the exponential mean function (which is correctly specified) estimated using pooled Poisson regression (where the Poisson distribution is misspecified). The pooled QMLE is essentially unbiased for each ATT, whereas the pooled OLS estimators have severe downward biases. The CS estimators show less bias but are still very different, on average, from the sample ATTs. Moreover, the precision of the Poisson regression estimates is much better than either POLS or CS.

The outcomes of the parallel trends tests are promising. The event-study-type test rejects in the linear case 99.5% of the time heterogeneous trends has a 99.9% rejection rate. Therefore, one would conclude that the PT assumption is violated in the linear model. Because the observed Y_{it} is a count variable, the hope is that one would turn to pooled Poisson regression with an exponential mean. The PT tests for the exponential model reject only 9.9% and 6.9% of the time, with the latter test using the three heterogeneous linear trends having particularly good size with $N = 500$ and three treated cohorts.

In the second simulation, I generate $Y_{it}(\infty)$ as a corner solution outcome: $P(Y_{it}(\infty) = 0) > 0$ with $Y_{it}(\infty)$ continuous over strictly positive values. Unobserved heterogeneity is allowed by setting

$$Y_t^*(\infty) = 0.2 + X/5 - (D_4 + D_5 + D_6) + C,$$

where $C \sim Normal(0, 1)$. The corner solution outcome is generated with idiosyncratic variation over time as

$$Y_t(\infty) = R_t(\infty) \cdot \exp(Y_t^*(\infty)),$$

where the $R_t(\infty)$ are independent $Poisson(1)$ random variables. In other words, $Y_t(\infty)$ is the product of a Poisson random variable with unit mean and a random variable with a lognormal distribution conditional on X. The other cohort potential outcomes, with no anticipation imposed, are generated to have their own multiplicative idiosyncratic shocks as

$$\begin{aligned}
Y_t(4) &= R_t(4) \cdot \exp(Y_t^*(\infty) + (X - 1)/5 + 1.2 \cdot f4_t + 1.6f5_t + f6_t), \; t \geq 4 \\
Y_t(5) &= R_t(5) \cdot \exp(Y_t^*(\infty) + (X - 1)/5 + 1.2 \cdot f5_t + 1.8 \cdot f6_t), \; t \geq 5 \\
Y_t(6) &= R_t(6) \cdot \exp(Y_t^*(\infty) + (X - 1)/5 + f6_t), \; t = 6
\end{aligned}$$

where the $R_t(g)$ are mutually independent and distributed as $Poisson(1)$. The R-squared from the POLS regression is about 0.083. The proportion of zero outcomes in the population is fairly large, $T^{-1} \sum_{t=1}^{T} P(Y_{it}(\infty) = 0) \approx 0.368$. The simulation results are shown in Table A.6.

Table A.6: Corner Solution Outcome, Staggered Intervention							
5,000 Replications	Sample ATT	Exponential (Pooled Poisson)		Linear (Pooled OLS)		CS (2021)	
$N = 500$	Mean	Mean	SD	Mean	SD	Mean	SD
τ_{44}	2.47	2.47	0.60	2.48	0.63	2.47	0.84
τ_{45}	4.19	4.18	0.90	4.19	0.94	4.18	1.08
τ_{46}	1.84	1.82	0.49	1.83	0.64	1.83	0.77
τ_{55}	2.76	2.75	0.77	2.76	0.81	2.75	1.12
τ_{56}	5.90	5.88	1.40	5.89	1.49	5.90	1.61
τ_{66}	2.45	2.41	0.97	2.44	1.17	2.46	1.72
Event Study Rejection Rate (9 df)	—	0.157		0.040		—	
Heterogeneous Trend Test (3 df)	—	0.087		0.047		—	