Appendix C Info-Metrics based Difference-in-Difference Estimator

Difference-in-differences (DiD) is a popular approach to estimate causal relationships when randomization is not feasible. In the basic set up, a group (the treatment group) is exposed to an intervention (treatment) at some period while another group (the control group) is never exposed to the treatment. Under the assumption that the two groups would have followed "parallel trends" in the absence of the treatment, the causal impact of the treatment can be estimated by comparing the difference in the outcome between the treatment and control groups before and after the intervention.

In practical work, employing the DiD methodology can pose several challenges:

1. Parallel trends assumption violation: if the two groups would not have followed parallel trends without the intervention, the DiD estimation is invalid. Violation of the parallel trends assumption could be due to:
    a. time-varying confounders: factors that change over time and affect the treatment and control group differently can bias the estimates.
    b. endogenous selection: assignment into the treatment group might be associated with unobserved factors affecting the outcome.
    c. heterogenous treatment effects: the impact of the intervention might vary across units or over time, and this variation might be correlated with membership to the control or treatment group.
2. Anticipation effects: If units in the treatment group anticipate the intervention and change their behavior beforehand, it can bias the results.
3. Spillover effects: The intervention might indirectly affect the control group, which would bias the results.
4. Staggered treatment timing: if the treatment is rolled out in several phases, the standard two-period DiD may not be appropriate.

This appendix develops an Info-Metrics (Generalized Maximum Entropy) based estimator for DiD that can correct for endogenous selection and heterogenous treatment effects, which violate the parallel trends assumption.

Theory

Consider the standard DiD model:

$$Y_{it} = \alpha + \beta D_i + \gamma T_t + \delta(D_i \times T_t) + \epsilon_{it}$$

where: Y_{it} is outcome Y for individual i at time t), D_i is a group indicator ($D_i = 1$ for the treatment group, $D_i = 0$ for the control group), T_t is a period indicator ($T_t = 1$ for the post-treatment period, $T_t = 0$ for the pre-treatment time period). $\alpha, \beta, \gamma, \delta$ are parameters (δ is the parameter of interest, the effect of the treatment and ϵ_{it} is an error term.

Under traditional DiD assumptions, this equation can be estimated consistently using Ordinary Least Squares (OLS). However, if the treatment effect varies with characteristics of the treatment unit or over time (*heterogenous treatment effects*), δ is no longer a consistent estimator of the treatment effect.

The traditional DiD model can be expanded so that:

$$Y_{it} = \alpha + \beta D_i + \gamma T_t + \delta(D_i \times T_t) + T_t \times Z_i' \Gamma + \sigma \epsilon_{it}$$

where: Z_i is a vector of covariates and Γ is a vector of parameters representing the effect of the covariates through the treatment channel. OLS will now again be a consistent estimator of the treatment effect.

However, in cases where treatment is endogenous (that is the probability of assignment to D_i is correlated with factors affecting the outcome Y_{it}), the above approach will lead to biased estimates of the treatment effect because the error terms are correlated with the treatment indicator. We can model the selection into treatment equation through a latent variable approach. I follow a method outlined by (Abadie 2005) that generalizes Heckman's (1997) correction model for selection. In this two-step process, the probability of being selected into treatment conditional on covariates, is modeled in step 1. In step 2, difference-in-differences is estimated, after weighting the observations by the probabilities obtained in step 1. Intuitively, units with certain characteristics are overrepresented in the untreated group. The weights from step 1, when applied to the regression, weights down the distribution of $Y(1) - Y(0)$ for those values of the covariates which are overrepresented and weighting up those values of the covariates that are underrepresented among the untreated (Abadie 2005).

I propose that both step 1 and step 2 in the above process be estimated using the semiparametric Generalized Maximum Entropy (GME) approach. In the first step, the probability of selection can be modeled using the generalized maximum entropy extension to the traditional logistic regression outlined in (Golan, Judge, and Perloff 1996). This method is known to have several advantages over traditional logistic regression: it is efficient for small samples, avoids strong parametric

assumptions, handles multicollinearity and is resilient to ill-conditioned data. Similarly, the difference-in-differences estimator in step 2 can be estimated using the "linear regression" maximum entropy estimator (Golan 2017). This estimator has similar advantages over traditional OLS.

First consider the generalized maximum entropy logistic regression for a binary outcome (selection into treatment). Assume that T units are observed, where each unit is either selected into treatment or not. Let

$$p_i = Prob(D_i = 1|x_i, \beta) = F(x_i'\beta)$$

be the probability of observing unit $i \in [1,2, \dots T]$ in the treatment state, conditional on covariates x_i (of dimension $1 \times K$) and unknown parameters β (of dimension $K \times 1$). If we have noisy data, we can write

$$y_i = F(\cdot) + e_i = p_i + e_i$$

where p_i denotes the unknown and unobservable probability of selection and e_i is a noise component for each observation contained in $[-1, 1]$. To proceed, we must reparametrize the error component:

$$e_i = \sum_h v_{ih} w_{ih}$$

where $\sum_h w_h = 1$ are an H-dimensional vector of weights and v_i are the H-dimensional support space. (Golan, Judge, and Perloff 1996) suggest $v_i = \left[-\frac{1}{\sqrt{T}}, \frac{0,1}{\sqrt{T}}\right]$ and thus $H = 3$.

We now maximize the entropy (of the error augmented) probability distribution:

$$\max_{p,w} \left(-\sum_i p_i \ln p_i - \sum_{ijh} w_{ijh} \ln w_{ijh} \right)$$

subject to the K moment constraints imposed by the data, where the k-th constraint is:

$$\sum_i y_i x_{ik} = \sum_i x_i p_i + \sum_{ih} x_i v_h w_{ih}$$

and the adding up constraints

$$\sum_h w_i = 1$$

$$\sum_{j \in [0,1]} p_{ij} = 1$$

We can solve the above using the method of Lagrange multipliers. In matrix-form:

$$\mathcal{L} = -\boldsymbol{p}\ln p - \boldsymbol{w}'\ln \boldsymbol{w} + \boldsymbol{\lambda}'(\boldsymbol{X}'\boldsymbol{p} + \boldsymbol{X}'\boldsymbol{V}\boldsymbol{w} - \boldsymbol{X}'\boldsymbol{y}) + \boldsymbol{\mu}'(1 - I_T\boldsymbol{p}) + \boldsymbol{\rho}'(1 - \mathbf{1}'w)$$

where $\boldsymbol{\lambda}.\boldsymbol{\mu},\boldsymbol{\rho}$ are Lagrange multipliers and $\beta = -\lambda$.

The first order conditions are:

$$\frac{\partial \mathcal{L}}{\partial p_i} = -\ln p_i - 1 - \sum_k \lambda_k x_{ik} - \mu_i = 0$$

$$\frac{\partial \mathcal{L}}{\partial w_{ih}} = -\ln w_{ih} - 1 - \sum_k \lambda_k x_{ik} v_h - \rho_i = 0$$

$$\frac{\partial \mathcal{L}}{\partial w_k} = \sum_i y_i x_{ik} - \sum_i x_{ik} p_i - \sum_k x_{ik} v_h w_{ih} = 0$$

$$\frac{\partial \mathcal{L}}{\partial u_i} = 1 - \sum_{j\in 0,1} p_j = 0$$

$$\frac{\partial \mathcal{L}}{\partial \rho_i} = 1 - \sum_h w_{ih} = 0$$

From which we obtain the estimated probability distributions:

$$\widehat{p_i} = \frac{\exp\left(-\sum \widehat{\lambda_k} x_{ik}\right)}{\sum_{j\in(0,1)} \exp\left(-\sum_k \widehat{\lambda_k} x_{ik}\right)}$$

and

$$w_{ih} = \frac{\exp\left(-\sum_k x_{ik}\widehat{\lambda_k} v_h\right)}{\sum_h \exp\left(-\sum_k x_{ik}\widehat{\lambda_k} v_h\right)}$$

Note that the above can be computed using the "dual unconstrained" method not discussed here (see (Golan 2017) for reference).

Once the probabilities are obtained, we can proceed to estimate the difference-in-differences model, weighting each observation y_i by:

$$\rho_{0i} = \frac{D_i - p_{ij}}{\widehat{p_{ij}} \cdot \left(1 - \widehat{p_{ij}}\right)}$$

The full generalized maximum entropy model of linear regression is not recapitulated here (see (Golan 2017) for reference) but briefly:

1. Reparametrize the β_ks and the error terms as $\beta_k = \sum_m z_{km} p_{km}$ and $\epsilon_i = \sum_j w_{ij} v_j$ where z_k is a $1 \times M$ vector (the discrete support space), p_k is a $M \times 1$ vector of weights, v_j is a $1 \times N$ vector (the discrete error support) and w_{ij} is a set of weights (that dd up to 1).
2. Maximize the entropy of $H_{p,w} = -\sum_i p_i \ln p_i - \sum_i w_i \ln w_i$ subject to the data constraint
$$\sum x_{ik} p_i \rho_0 + \sum v_i w_i \rho_0 - y_i \rho_0 = 0$$
and the adding up constraints.
3. Solve using the method of Lagrange multipliers.

Thus, both the first stage and second stages of the methodology described by Abadie can be implemented using the semiparametric generalized maximum entropy techniques (with estimators in both stages having desirable properties described earlier).

Simulation

I simulate the performance of this estimator using the following data generating process. For the baseline case, I create a dataset of $n = 500$ observations. Each observation i has observable covariates X_{i1} and X_{2i} and unobservable $u_i \sim \mathbb{N}(0,1)$. Selection into treatment is governed by the selection equation, simulating both selection on observable and selection on unobservables:

$$D_i = \begin{cases} 1 \; if \; Unif(0,1) < logit(3X_{1i} - 5X_{2i} - 2u_i) \\ 0, otherwise \end{cases}$$

We observe the outcomes twice: once in the pre period and once in the post period. The outcome process is governed by the equation

$$Y_{pre,i} = 1 + 0.3X_{1i} + 0.3X_{2i} + 0.2D_i + \epsilon_{pre,i}$$
$$Y_{post,i} = 3 + 0.3X_{1i} + 0.3X_{2i} + (1.5 + 0.5X_{1i})D_i + \epsilon_{post,i}$$

This equation induces some difference in the "pre" outcome between treatment and control groups. In addition, there is some heterogeneity in treatment outcomes based on the realized value of X_{1i}. The true Average Treatment Effect on the Treated (ATT) is 2.0.

I generate data using this process 5,000 times. Figure 10 shows the results of the simulation for the baseline case, compared against an uncorrected/naïve difference-in-differences and Abadie's matching estimator using logit for the first stage. Table 16 summarizes these results. Note that the uncorrected difference-on-differences is not a consistent estimator of the treatment effect under the circumstances of selection into treatment and heterogenous treatment effects (both causing violation of the parallel trends assumption). Abadie's method, using either logistic regression or

GME for the first stage perform better, with the GME method having much smaller variance (std. dev.) and mean squared error.

Figure 10 Comparison of GME matching estimator performance, in comparison to matching based on logit and an uncorrected difference-in-differences

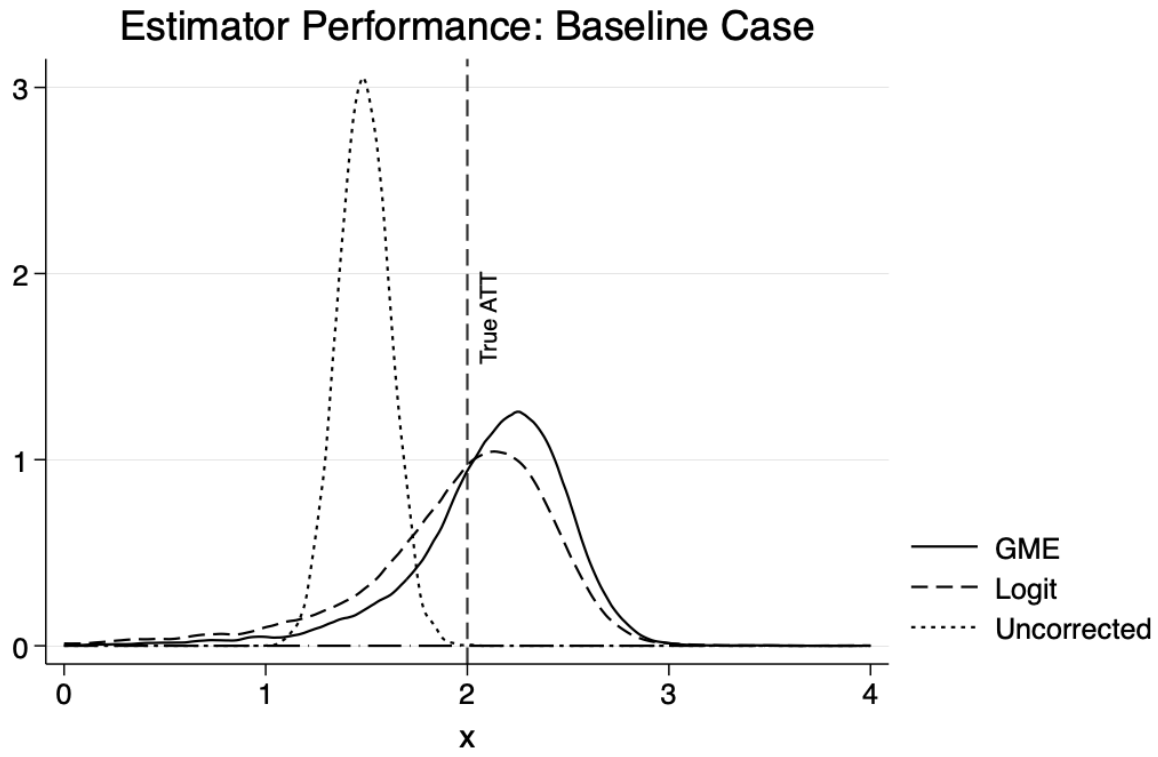

Table 16 Summary of Estimator Performance

	Baseline Performance		
	Bias	Std. Dev	MSE
Abadie (Logit)	-.1147202	.8791705	.7861014
Abadie (GME)	.0797839	.5604687	.3204907
Traditional DiD	-.5119547	.1295395	.2788781

To test the small sample performance, I reduce the size of the simulated dataset from $N = 500$ to $N = 25$. Figure 11 and Table 17 show the results. Now, the GME-based matching estimator clearly outperforms a logit-based first stage, with a lower bias, variance and MSE.

Figure 11 Small Sample Performance of Abadie's Estimator

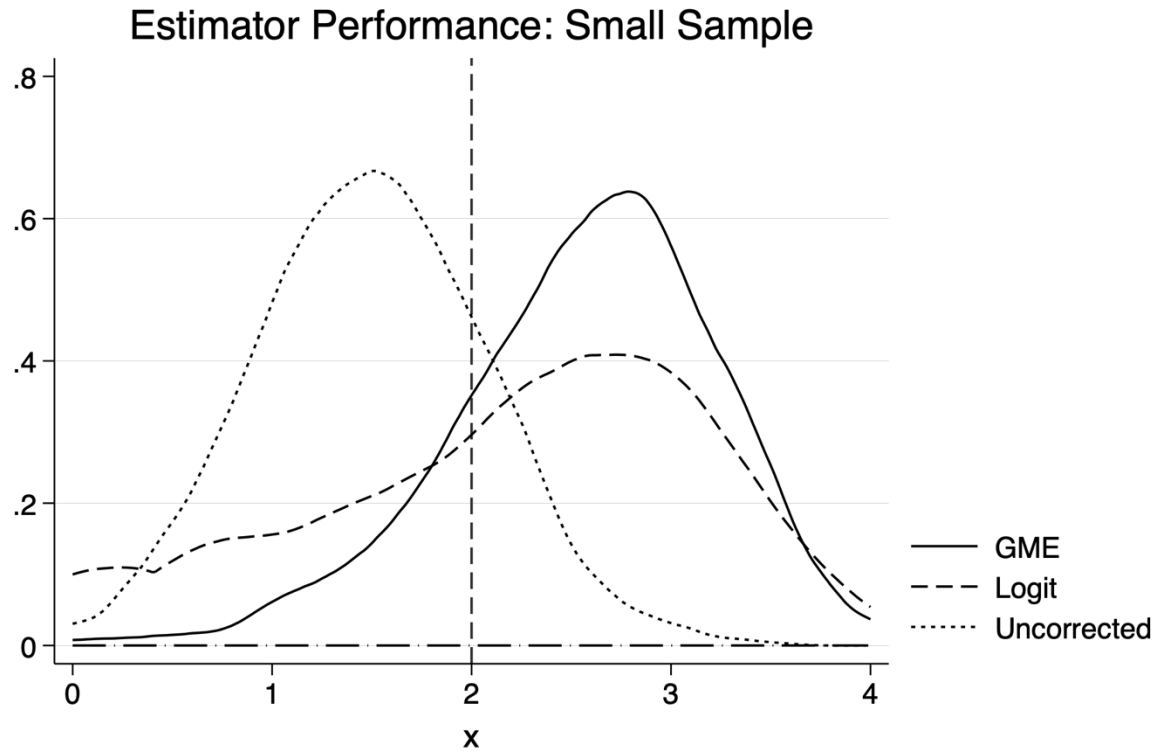

Table 17 Small Sample Performance of Abadie's Estimator

	Small Sample Performance		
	Bias	Std Dev	MSE
Abadie (Logit)	.1169992	1.279985	1.652051
Abadie (GME)	.564807	.683304	.7859113
Traditional DiD	-.5137796	.5946682	.6175997

To test the impact of an ill-conditioned design matrix, I modify the data-generating process as follows:

$$X_{3,i} = X_{1i} + X_{2i} + u_i$$
$$u_i \sim \mathbb{N}(0, 0.001)$$
$$Y_{pre,i} = 1 + 0.3X_{1,i} + 0.3X_{2,i} + 0.2D_i + X_{3,i} + \epsilon_{pre,i}$$
$$Y_{post,i} = 1 + 0.3X_{1,i} + 0.3X_{2,i} + (1.5 + 0.5X_{1i})D_i + X_{3,i} + \epsilon_{post,i}$$

That is, the design matrix now has an additional column X_3 that is nearly a linear combination of X_1 and X_2. Figure 12 and Table 18 shows how the estimators perform under these circumstances. The performance is very similar to the baseline case, with the GME based method still outperforming a logit first stage and the traditional difference-in-differences estimator.

Figure 12 Performnce of Abadie's estimator when the design matrix is ill-conditioned

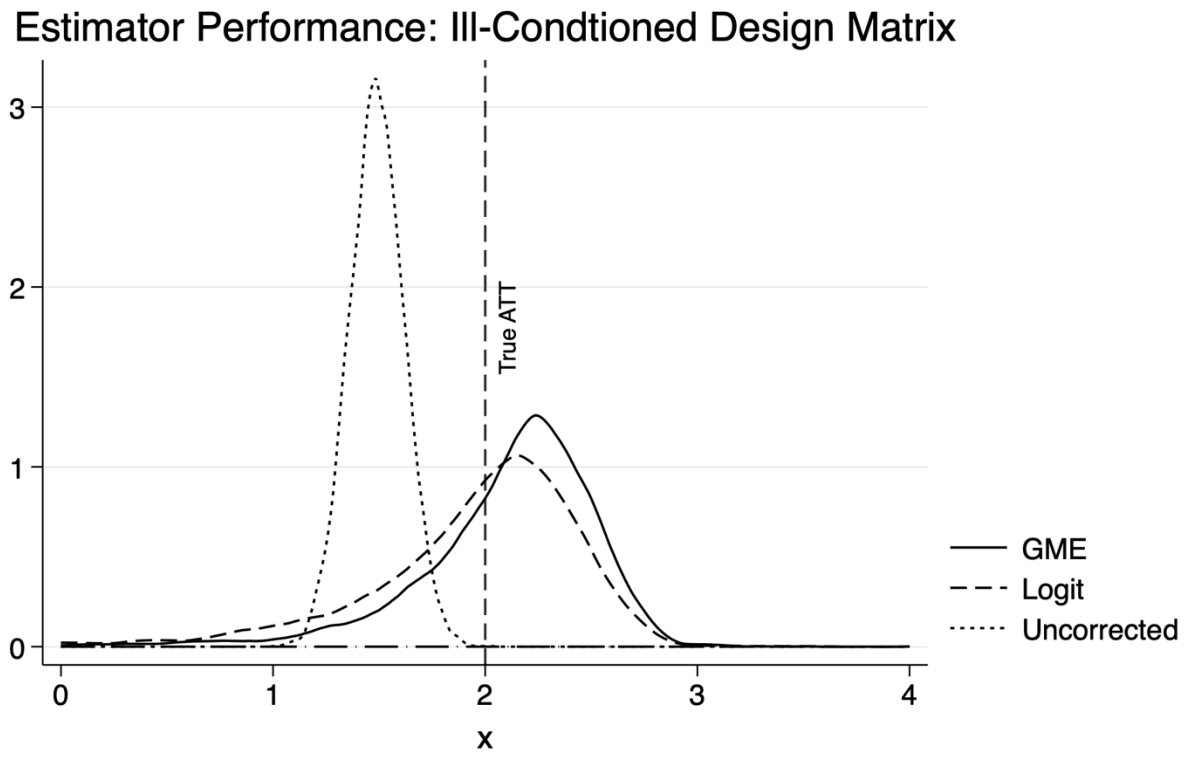

Table 18 Ill-Performance of Abadie's estimator when the design matrix is ill-conditioned

	Ill-Conditioned Data Performance		
	Bias	Std. Dev	MSE
Abadie(Logit)	-.1448158	2.065859	4.288745
Abadie(GME)	.0646625	1.160147	1.350122
Traditional DiD	-.5113141	.1286859	.2780022

Finally, I test the case when there is some heteroskedasticity/correlation between covariates and the error. I modify the data generating process as follows:

$$\epsilon_{pre,i} = \mathbb{N}(0.5\, X_{1i}, 1)$$

$$\epsilon_{post,i} = \mathbb{N}(0.5\, X_{1i}, 1)$$

The results are shown in Figure 13 and Table 19. The GME-based first-stage estimator clearly outperforms the Logit-based one, in terms of bias, variance and mean squared error. In summary, the GME-based first stage estimator outperforms a logit-based method in all cases simulated, with the outperformance especially pronounced in the case of small samples and heteroskedasticity.

Figure 13 Estimator performance when covariates are correlated wth errors

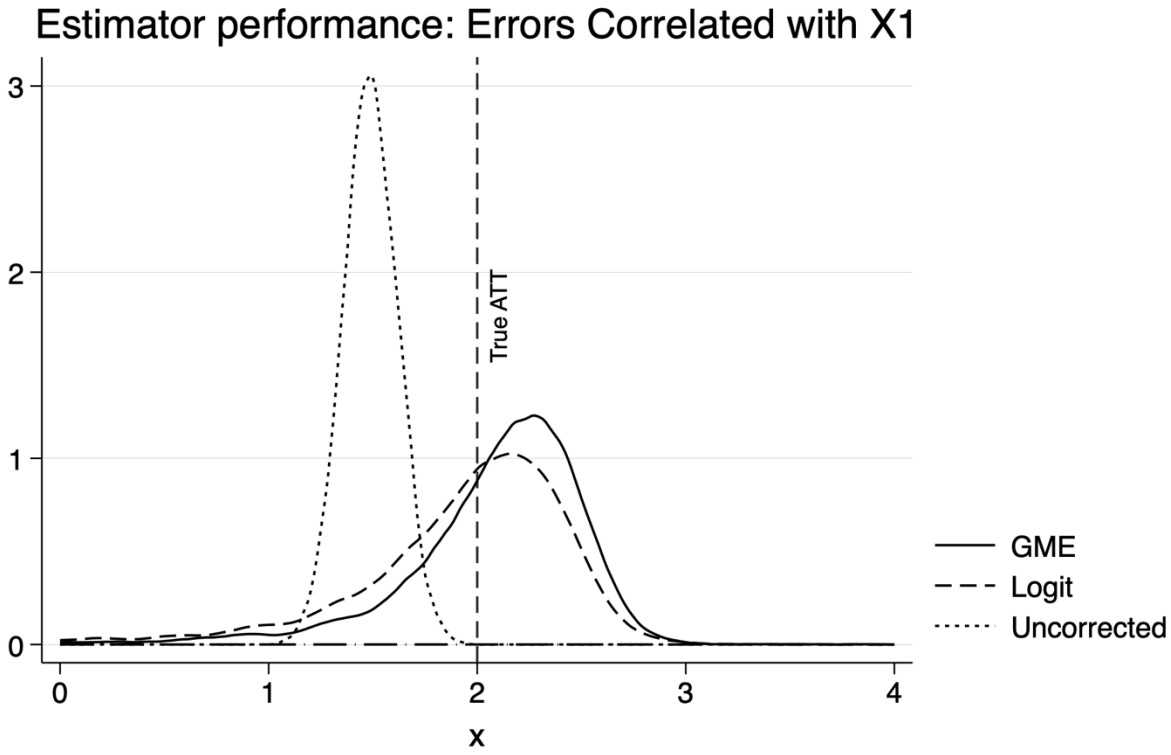

Table 19 Estimator Performance when covariates are correlated with errors

	Heteroskedasticity Performance		
	Bias	Std. Dev.	MSE
Abadie (Logit)	-.1560549	1.163898	1.379011
Abadie (GME)	.052799	.6806262	.4660398
Traditional DiD	-.512848	.1297478	.2798475

Real World Data

I now apply this method to a real-world dataset. I use the data provided by LaLonde (1986) for his evaluation of the impact of the National Supported Work (NSW) demonstration. The data is readily available in many software packages as well as through the NBER website. The NSW program randomly assigned individuals to a treatment or control group. Those in the treatment group received "supported work" where participants were employed at construction, service or similar industries in a supportive but performance-oriented work environment. LaLonde examines the data for male and female participants separately. For the male group, the outcome of interest is the 1978 real earnings due to participation in the program (or earnings growth from baseline 1975 wages in a difference-in-differences context). Since we have data from a control group, the difference in mean 1978 earnings between the experimental and control group is an unbiased estimator of the treatment effect. LaLonde asks what if researchers used an alternative control group to analyze the data instead of the actual control group. He constructs several sets of control groups: one based on the Current Population Survey (CPS), one based on the Panel Study of Income Dynamics (PSID), with additional datasets further sub-setting these to match demographic and pre-treatment earnings characteristics to more closely match that of program participants. A part of LaLonde's results is reproduced here for reference.

Table 20 LaLonde's Analysis of NSW program effects on male participants (abbreviated)

Comparison Group	Treatment Less Comparison Group Earnings		Difference-in-Differences: Difference in Earnings Growth 1975-1978 Less Comparison Group	
	Unadjusted	Adjusted[4]	Without Age	With Age
Controls	$886	$798	$847	$856
	(476)	(472)	(560)	(558)
CPS-1	-$8,870	-$4,416	$1,714	$195
	(562)	(557)	(452)	(441)

[4] The exogenous variables used in the regression adjusted equations are age, age squared, years of schooling, high school dropout status and race

Note that naïve difference in 1975 earnings between the CPS comparison group and the treatment group recovers a large negative impact of participating in the program. Using the experimental control group as our reference, we know participation in the program is associated with a ~$886 increase in earnings. The large negative impact obtained when using the CPS as the comparison group is likely due to differences in demographic characteristics between the treatment group and the CPS population. Column 2 attempts to adjust for some demographic characteristics such as educational attainment, race and age. However, the estimate of the program's effect continues to be negative, likely due to uncontrolled (and unobservable) differences between the two groups. Columns 3 and 4 show the results of a Difference-in-Differences type estimation (using 1975 earnings as a baseline). Column 4 controls for age (and age squared) in the DiD estimate. The CPS group shows wildly different impact of program participation when age is controlled for (which is inconsistent with our findings for the experimental control group).

I now apply Abadie's matching method, using the GME based first stage described earlier. I obtain standard errors through bootstrap. Table 21 shows the results. Note that the estimate of the program's effect is now much closer to one obtained from using the experimental control group. The standard error of the estimate is also similar.

Table 21 LaLonde's Analysis of CPS population with matching

Comparison Group	Difference-in-Differences: Difference in Earnings Growth 1975-1978 Less Comparison Group
	Without Age
Controls	$847
	(560)
CPS-1 with Abadie's	$575
method	(562)

In conclusion, through simulation I show that Abadie's matching method (especially when using GME first stage to estimate the selection equation) performs well in recovering the impact of treatment when the narrow parallel trends assumption is violated (due to selection or heterogenous treatment effects. However, the parallel trends assumption is still maintained after conditioning on treatment and covariates). The GME-based method performs especially well when there is heteroskedasticity or small sample sizes. In addition, I apply the method to a well-known dataset

(once where we have an experimental control group to verify our result). I show the Difference-in-Differences with matching estimator performs much better in estimating the treatment effect than traditional DiD.