An Information Theoretic Extension to conditional differences-in-differences

The difference-in-differences approach (sometimes abbreviated DiD or diff-in-diff) to causal identification has a long history, perhaps tracing back to work of physician John Snow in identifying the cause of the 1854 London Cholera outbreak (Cunningham 2021). In economics, precursors to diff-in-diff methods began to appear in the literature in the in the 1950s (for example (Lester 1946) as recounted in (Halkiewicz 2024)). The modern diff-in-diff approach was popularized by Ashenfelter (1978), with further crucial work by Ashenfelter and Card (1984), Angrist (1990; 1991), Angrist and Kruger(1991; 1999), and Imbens (1995). This development was part of the "credibility revolution" taking place in economics in the 1990s, which saw a variety of methods for identifying causal effects (also known as identification strategies), such as regression discontinuity designs (RDD), instrumental variables (IV), natural experiments and randomized control trials (RCT), becoming popular in the field (Angrist and Pischke 2010).

Each identification strategy relies on some assumptions about the underlying data generation process (DGP) and causal mechanism to make a valid causal inference. The identification issue arises from the "missing counterfactual" problem: we cannot observe the same units both receive and not receive the treatment (the intervention or policy variable whose effects we are interested in studying). We try to fill in this missing counterfactual with some stand-in. In the case of diff-in-diff, we identify an alternative group that never receives the treatment during our period of study. We then assume that the groups would have followed "parallel trends" for the outcome variable (that is, the gap between the treated and untreated groups would have remained constant) in the absence of any intervention. Thus, any change in the gap between the outcome variables after treatment can be attributed to the intervention.

For a concrete example, consider a diff-in-diff study that attempts to identify the effect of a job training program on wages (for example, as in LaLonde, 1985). We observe the wages of those in the training program before and after they receive training. However, we cannot plausibly identify the impact of the training program on participants' wages based on this information alone, since we do not observe what their wages would have been had they not received training. After all, other variables (such as the business cycle) might have caused changes in wages, and if we simply compared the before and after training wages of this group, we would be mistakenly attributing the change in wages to job training, when another variable could have been the cause. We try to get around this missing counterfactual problem by identifying a group that never receives job training during the period of study. Let's say that before the job training program started, those not in the training program on average made $800 more a year than those in the training program (job training is often targeted to those who are unemployed, underemployed or low income). We now make the crucial assumption that this gap would have remained constant if not for the job training program. That is, although wages for both groups might have increased or declined during the period of study (e.g. due to the business cycle), the average gap in wages between the groups would have remained constant without the job training program. If we can plausibly make this claim, then any changes in the gap after training can be attributed to the program. For example, if after the program, the gap in wages between the two groups narrowed to $300, then the program can be attributed to an average $500 increase on the earnings of those who received it.

The diff-in-diff methodology requires the crucial and restrictive parallel trends assumption. This assumption is sometimes implausible. In the job training example, the comparison group might differ in terms of age, education, gender and racial composition from those receiving training. Let us assume that those in the job training program are on average, younger than those in the comparison group. If younger workers experienced higher wage growth throughout the economy during our period of study, then we could be falsely attributing some of the wage gains to the job training program when the gains are really driven by the difference in ages between the two groups. To make a valid causal inference in this situation, we might adopt the less restrictive "conditional parallel trends" assumption. That is, we could assume that the groups would have followed parallel trends conditional on some values of the covariates.

The literature suggests several ways to estimate this type of conditional diff-in-diff model. Firstly, if we have a good model of how the outcome variables would have evolved conditional on the covariates (for example, if we know how age affects wages), we can incorporate this into the two-way fixed-effects (TWFE) regression commonly used to estimate the classical diff-in-diff model (Baker et al. 2025). Other methods include matching and inverse-probability weighting (IPW) (Abadie 2005; Abadie and Imbens 2006), regression adjustment (J. J. Heckman, Ichimura, and Todd 1997) or doubly robust diff-in-diff (Sant'Anna and Zhao 2020), which incorporates both regression adjustment and inverse probability weighting.

This essay applies and expands on a weighting scheme suggested by Abadie (2005). Specifically, it uses the generalized maximum entropy logistic regression (GME Logit) as proposed by Golan

Judge and Perloff (1996) to obtain the probability of selection into treatment and then uses the weighting scheme developed by Abadie to obtain the treatment effect. I show that the information theory (IT) based estimation of the selection equation outperforms against a classical logistic regression (Logit or ML Logit) for estimating selection probabilities in certain situations. This is because the IT-based estimator is known to outperform ML Logit in all finite samples. The more complex and ill-behaved the data, the larger the advantage of IT Logit.

Researchers are often interested in studying the treatment effect for a subpopulation (that is, for specific values of a covariate). Abadie (2005) also develops a semiparametric method to estimate this treatment effect for subpopulations, using the weights discussed above. Abadie's method replaces the true (usually high dimensional) conditional average treatment effect with a parametric approximation. If the parametric approximation is a linear model (often the case in practice), Abadie's method becomes weighted least squares (WLS) with the weights obtained from estimating the selection equation in stage 1. This essay expands on Abadie's method in this area as well, replacing the WLS-based estimator he suggests with a (weighted) generalized maximum entropy-based estimator. I also show that this method outperforms against the WLS-based estimator under certain circumstances.

The essay proceeds as follows: section II reviews the literature on diff-in-diff and conditional diff-in-diff; section III develops the theory; section IV compares the GME-based and classical methods in a series of simulations; section V applies the method to a well-known dataset and section VI concludes.

Review of Literature

The history of canonical diff-in-diff was briefly discussed earlier. I do not recap it here but discuss developments since then. The diff-in-diff literature developed in two distinct waves (Cunningham 2023). The first wave, which started in the 1980s (and discussed earlier), introduced the term and saw the publication of several influential papers. It peaked around 2007. The second wave began in 2011 and was characterized by "an exponential growth in popularity" (Cunningham 2023) and consistent use of terminology (such as parallel-trends and event study) in many papers. This second wave continues today, with diff-in-diff becoming one of the most popular methods of causal identification.

The validity of what we now call the parallel trends assumption was under scrutiny very early in the history of the literature. For example, Ashenfelter (1978) discussed that participants in job training programs are often selected after they receive a temporary dip in their earnings or employment status. In such cases, wages could be expected to bounce back even in the absence of the training program. LaLonde (1986) showed that econometric methods such as diff-in-diff fail to match results obtained from a randomized control trial (RCT), and attributed this to mismatches in outcome trends between the treatment group and potential comparison groups (e.g. drawn from the current population survey, panel study of income dynamics or similar data sources for the general labor force) researchers might use. Heckman, et. al. (1998; 1997) proposed a method called regression adjustment that can recover the ATT when parallel trends are violated if the researcher correctly specifies the model for conditional outcome evolution. Recent research has progressed in three directions: (1) testing (conditional) parallel trends assumptions (e.g. (Roth 2022; Rambachan and Roth 2023)) (2) extending the method to staggered/multi-period treatment adoption (e.g. (Goodman-Bacon 2021; Sun and Abraham 2021; Callaway and Sant'Anna 2021)) or where the level/intensity of treatment can be varied continuously (e.g. (Callaway, Goodman-Bacon, and Sant'Anna 2024)) and relaxing the parallel trends assumption (e.g. (Abadie 2005; Athey and Imbens 2006)).

Theory

I begin by defining the diff-in-diff methodology in the canonical setting. The setup is described in many textbooks such as (Cunningham 2021) or (Angrist and Pischke 2009). The researcher selects a causal quantity of interest (target parameter) for study. Adapting the potential outcomes framework of Rubin (1976), let $Y_{i,t}(0,0)$ denote unit $i's$ potential outcome at time t if it remains untreated in both periods, and $Y_{i,t}(0,1)$ if it is untreated in period $t = 0$ and treated in period $t = 1$. Since all units are untreated in period $t = 0$, we can simplify notation to: $Y_{i,t}(0) = Y_{i,t}(0,0)$ and $Y_{i,t}(1) = Y_{i,t}(0,1)$. Ideally, we would like to observe $Y_{i,t}(0)$ and $Y_{i,t}(1)$ for the same units. Then the treatment effect for unit i would be $Y_{i,t}(1) - Y_{i,t}(0)$. In practice, we can only observe $Y_{i,t}(1)$ for the treated units and $Y_{j,t}(0)$ for the untreated units. If D_i is an indicator variable that takes a value $D_i = 1$ if unit i is treated in period $t = 1$, and $D_i = 0$ otherwise, then the observed values are:

$$Y_{i,t} = (1 - D_i)Y_{i,t}(0) + D_i Y_{i,t}(1)$$

(1)

The potential outcomes framework defines a treatment effect $Y_{i,t}(1) - Y_{i,t}(0)$ for every unit i. In practice, researchers are often interested in finding the average treatment effect on the treated:

$$ATT = \mathbb{E}\left[Y_{i,t}(1) - Y_{i,t}(0)\middle|D_i = 1\right] = \mathbb{E}\left[Y_{i,t}(1)\middle|D_i = 1\right] - \mathbb{E}[Y_{i,t}(0)|D_i = 1]$$

(2)

Without additional assumptions, the ATT is not identified since we are unable to observe $\mathbb{E}\left[Y_{i,t}(0)\middle|D = 1\right]$ (the expected potential outcome if unit i had not received treatment, given that it actually did receive treatment). If we are willing to make additional assumptions, we could potentially identify the ATT. Different research designs make different assumptions for identifying the missing counterfactual. In the case of diff-in-diff, the crucial assumption (known as the parallel trends assumption) is:

$$\mathbb{E}\left[Y_{i,t=1}(0)\middle|D_i = 1\right] - \mathbb{E}\left[Y_{i,t=0}(0)\middle|D_i = 1\right] = \mathbb{E}\left[Y_{i,t=1}(0)\middle|D_i = 0\right] - \mathbb{E}\left[Y_{i,t=0}(0)\middle|D_i = 0\right]$$

(3)

In words, the expected change in outcome between pre-treatment and post-treatment periods $(\mathbb{E}\left[Y_{i,t=1}(0) - Y_{i,t=0}(1)\right])$ is the same for treated and comparison groups. Under these assumptions, the ATT is:

$$ATT = \mathbb{E}\left[Y_{i,t=1}\middle|D_i = 1\right] - \left(\mathbb{E}\left[Y_{i,t=0}\middle|D_i = 1\right] + \left(\mathbb{E}\left[Y_{i,t=1}\middle|D_i = 0\right] - \mathbb{E}\left[Y_{i,t=0}\middle|D_i = 0\right]\right)\right)$$

(4)

where the term in parenthesis on the right-hand side is our counterfactual $\mathbb{E}\left[Y_{i,t}(0)\middle|D_i = 1\right]$. The above can be rewritten as:

$$ATT = \left(\left[Y_{i,t=1}\middle|D_i = 1\right] - \mathbb{E}\left[Y_{i,t=0}\middle|D_i = 1\right]\right) - \left(\mathbb{E}\left[Y_{i,t=1}\middle|D_i = 0\right] - \mathbb{E}\left[Y_{i,t=0}\middle|D_i = 0\right]\right)$$

(5)

where the first term in the parenthesis is the change in the average outcome for treated units between period $t = 0$ and $t = 1$ and the second term is the change in the average outcome for the untreated units during the same period. Thus, the average treatment effect (on the treated) has a natural interpretation as the "additional" average change in the outcome of the treated units on top of the change experienced by the untreated units during the period of study.

It is well-known (for a textbook treatment, see (Cunningham 2021)) that estimating this model, involving four means is the same as a two-way fixed effects (TWFE) regression:

$$Y_{i,t} = \theta_t + \eta_i + \beta D_{i,t} + e_{i,t}$$

(6)

where: θ_t and η_i are time and unit fixed effects respectively, $e_{i,t}$ are idiosyncratic shocks and β is the parameter of interest.

Often the parallel trends assumption is not plausible. The treatment group might differ from the comparison group in systematic ways that affect the outcome (for example, workers participating in a job training program might be less educated or younger than the comparison group and would have experienced a different wage trend than older workers regardless of whether they participated in training or not). We can relax the parallel trends assumption from above with a conditional parallel-trends assumption. Let X_i be a vector of observed determinants of changes in $Y_{i,t}^0$. We now assume that the groups would have followed parallel trends *conditional* on the covariates:

$$\mathbb{E}\left[Y_{i,t=1}(0) - Y_{i,t=0}(0) \middle| X_i, D_i = 1\right] = \mathbb{E}\left[Y_{i,t=1}(0) - Y_{i,t=0}(0) \middle| X_i, D_i = 0\right]$$

(7)

There are several ways to estimate this conditional parallel-trends model. Firstly, the covariates can be incorporated into an augmented TWFE regression (Baker et al. 2025) either as:

$$Y_{i,t} = \theta_t + \eta_i + \beta D_{i,t} + X_{i,t}\beta_2 + e_{i,t}$$

(8)

or as

$$Y_{i,t} = \theta_t + \eta_i + \beta D_{i,t} + \left(\mathbf{1}\{t = 1\}X_{i,t=0}\right)\beta_3 + e_{i,t}$$

(9)

where: $X_{i,t}$ is the observed value of the covariate at time t, $\mathbf{1}\{t = 1\}$ is an indicator variable that takes a value of 1 in the post-treatment period and 0 otherwise, and β_2 and β_3 are vectors of coefficients associated with the covariates. Note that since time-invariant variables drop out in the TWFE regression, only the effect of changes in the levels of a covariate $\Delta X_{i,t} = \left(X_{i,t=1} - X_{i,t=0}\right)$ or differential trends related to baseline levels of the variable, interacted with a post-treatment dummy $\mathbf{1}\{t = 1\}X_{i,t=0}$ can be estimated with this type of model. In addition, researchers need to be careful about what variables are controlled for. For example, if the treatment affects the outcome by changing the value of $X_{i,t}$ (e.g. job training improving some measured skill), then controlling for this in the TWFE regression will bias our estimate of ATT.

Abadie (2005) provides another way to estimate the average treatment effect. If we are willing to make the mild assumptions that (1) at least some portion of the population receives treatment and (2) at every stratum/value of the covariate, some portion of the population remains untreated, that is:

$$\Pr(D_i = 1) > 0 \ \ and \ \ 0 < \ \Pr(D_i = 1|X_i) < 1$$

(10)

then Abadie shows that the ATT *conditional* on X_i is:

$$\mathbb{E}\left[Y_{i,t=1}(1) - E_{i,t}(0)\middle|X_i, D_i = 1\right] = \mathbb{E}\left[\rho_0 \cdot \left(Y(1) - Y(0)\right)\middle|X_i\right]$$

(11)

where:

$$\rho_0 = \frac{D_i - \Pr(D_i = 1|X_i)}{\Pr(D_i = 1|X_i)\,(1 - \Pr(D_i = 1|X_i)}$$

(12)

The (unconditional) ATT can be recovered as:

$$\mathbb{E}\left[Y_{i,t}(1) - Y_{i,t}(0)\middle|D_i = 1\right] = \int \mathbb{E}\left[\rho_0 \cdot \left(Y_{i,t}(1) - Y_{i,t}(0)\right)\middle|X_i\right] d\,\Pr(X_i|D_i = 1)$$

$$= \mathbb{E}\left[\frac{Y_{i,t}(1) - Y_{i,t}(0)}{\Pr(D_i = 1)} \cdot \frac{D_i - \Pr(D_i = 1|X_i)}{1 - \Pr(D_i = 1|X)}\right]$$

(13)

Abadie does not specify how $\Pr(D_i = 1|X_i)$ should be estimated. Commonly, researchers use ML Logit. However, other methods that estimate the selection probability could also be used. The GME approach for binary discrete choice (Golan, Judge, and Perloff 1996) is known to have several advantages over traditional ML Logit: it is efficient for small samples, avoids strong parametric assumptions, handles multicollinearity and is resilient to ill-conditioned data. The theory of GME Logit is recapped below (see Golan, Judge and Perloff 1996 for a more thorough development).

Assume that N units are observed, where each unit is either selected into treatment or not. Let

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

(14)

be the probability of observing unit $i \in [1,2, \dots N]$ 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

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

(15)

where $y_i \in [0,1]$ is the observed outcome, p_i denotes the unknown and unobservable probability of selection and e_i is an error or noise component for each observation contained in $[-1, 1]$. We reparametrize the error component as:

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

(16)

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

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

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

(17)

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

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

(18)

and the adding up constraint

$$\sum_h w_i = 1$$

(19)

(I omit the constraint the $\sum p_{i,j}$=1, since this is naturally satisfied in the case of binary choice). We can solve the above using the method of Lagrange multipliers. In matrix-form:

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

(20)

where $\boldsymbol{\lambda}.\,\boldsymbol{\mu}, \boldsymbol{\rho}$ are Lagrange multipliers and $\beta = -\lambda$. The first order conditions are:

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

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

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

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

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

(20)

From which we obtain the estimated probability distributions:

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

(21)

and

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

(22)

Note that the above can be computed using a "dual unconstrained" method that is usually less computationally intensive but not discussed here for brevity (see (Golan 2017) for details).

Simulation

To evaluate the performance of alternative estimators of the ATT, I conduct Monte Carlo simulations based on a stylized data generating process (DGP) designed to reflect common features of observational studies, including selection into treatment on observables and covariate-dependent potential outcomes.

I begin by describing the baseline simulation. The simulated dataset contains $n = 500$ observations and the simulation is replicated across 5,000 Monte Carlo runs. Each observation is characterized by two baseline covariates, X_1 and X_2, independently drawn from a standard normal distribution:

$$X_1 \sim \mathcal{N}(0,1)$$
$$X_2 \sim \mathcal{N}(0,1)$$

(23)

The potential outcome in the absence of treatment, denoted Y_0, is modeled as a linear function of the covariates, with additive normal noise:

$$Y_0 = 5 + 0.5X_1 + 0.3X_2 + \epsilon_0,$$
$$\epsilon_0 \sim \mathcal{N}(0,1)$$

(24)

This structure ensures that baseline outcomes are systematically related to observed characteristics. In addition to level differences, the untreated trend also and depends on the covariates:

$$\Delta_0 = 0.4X_1 - 0.1X_2$$

(25)

This trend is added to all individuals, regardless of treatment status, mimicking scenarios in which outcome trajectories differ across subgroups even in the absence of intervention. Treatment is assigned based on a logistic selection model that depends on the covariates:

$$\Pr(D = 1|X_1, X_2) = invlogit(0.5X_1 - 0.8X_2)$$

(26)

where $invlogit$ is $\Pr(D = 1|X1, X2) = \frac{\exp(0.5X_1 - 0.8X_2)}{\exp(0.5X_1 - 0.8X_2) + 1}$. Each unit is assigned to treatment with a probability equal to this score. This introduces selection on observables, such that the probability of receiving treatment is systematically related to covariates that also influence potential outcomes. Equations (25) and (26) are the crucial deviations from the unconditional parallel trends assumption that makes TWFE a biased estimator of the ATT. Since treatment into selection and the trend are both dependent on covariates X_1 and X_2, the treated and untreated groups no longer follow parallel trends.

The observed outcome Y_1 is constructed as the untreated outcome plus the untreated trend, a constant treatment effect for the treated, and additional noise:

$$Y_1 = Y_0 + \Delta_0 + 2 \cdot D + \epsilon_1,$$
$$\epsilon_1 \sim \mathcal{N}(0,1)$$

(27)

Thus, the ATT is set to a constant value of 2 across all treated units, independent of covariates. However, since both the baseline outcome and the untreated trend depend on X_1 and X_2, failure to account for these covariates can lead to biased estimates of the treatment effect.

Under the baseline assumptions (large samples, well-behaved data), GME Logit and ML logit models of selection into treatment should behave similarly. To test whether GME Logit outperforms ML Logit under certain circumstances, I test several cases where GME is known to outperform. Firstly, GME is known to outperform ML logit for all finite samples. These differences are likely to be highlighted when sample sizes are very small. I modify the above (baseline) simulation, so that the dataset is of size $n = 25$ observations. Next, I test how the two estimators perform when the probability of receiving treatment is very small (this is done by shifting the selection equation to:

$$\Pr(D = 1|X_1, X_2) = invlogit(-5 + 0.4X_1 - 0.1X_2)$$

(28)

which causes only ~1% of observations to be selected into treatment). GME logit is also known to outperform ML logit when covariates are highly collinear. I simulate this by drawing X_1 and X_2

from a multivariate normal distribution with a high correlation $\rho(X_1, X_2) = 0.99$. The tables below summarize the simulation settings.

Table 1: Baseline simulation settings. All simulations have 5,00 runs

Observations (N)	$n = 500$
Covariate Distribution (X_1, X_2)	$X_1 \sim \mathcal{N}(0,1)$ $X_2 \sim \mathcal{N}(0,1)$
Heterogenous Trend (Δ_0)	$\Delta_0 = 0.4X_1 - 0.1X_2$
Selection equation	$\Pr(D = 1 \vert X_1, X_2)$ $= invlogit(0.5X_1 - 0.8X_2)$
Errors	$\epsilon_0 \sim \mathcal{N}(0,1)$ $\epsilon_1 \mathcal{N}(0,1)$

Table 2: Modifications to the baseline case for specific scenarios.
All simulations have 5,000 Monte Carlo runs

Case: Small sample size	$N = 25$
Case: Rare selection	$\Pr(D = 1 \vert X_1, X_2) = invlogit(-5 + 0.5X_1 - 0.8X_2)$
Case: Highly collinear X_1, X_2	$X_1, X_2 \sim \mathcal{N}(0, \begin{pmatrix} 1 & 0.99 \\ 0.99 & 1 \end{pmatrix}$

For the estimation of the GME Logit model, I use a support space of 3 with $v_{ih} = \left[\frac{-1}{\sqrt{n}}, \frac{0,1}{\sqrt{n}}\right]$ as suggested by Golan, Judge and Perloff (1996). I do not use any prior distributions for GME Logit. Note that unlike the GME Linear Model (discussed later), GME Logit does not require us to reparametrize the coefficients $(\hat{\beta})$ and the error is naturally bounded between 0 and 1.

First Stage

I begin my analysis of the simulation results by first examining how successful ML Logit and GME Logit are at recovering the selection mechanism. I examine two aspects of the methods' performance: (1) how well each predicts treatment assignment (treated vs. untreated) and (2) how accurately each recovers the true model parameters. To assess (1), I report the number of misclassifications (following (Golan, Judge, and Perloff 1996)) as well as the Area Under the Receiver Operating Curve (AUC-ROC). The AUC-ROC is a widely used metric that summarizes a classifier's ability to rank positive cases above negative ones across all possible threshold values. A value of 0.5 corresponds to guessing at random, while a value of 1 indicates perfect classification. For reporting misses, I set the classifier threshold to 0.5 (that is, for each method, when the predicted probability is > 0.5, the observation is set to be in the treated group). For assessing (2), I report the Mean Squared Error (MSE) for each method as in Golan Judge and Perloff (1996).

Table 3: Model Performance for Stage 1 over 5,000 Monte Carlo simulation runs. Each run has 500 observations, except for the small sample case, which has 50 observations. The worst, average and best misclassification rates over the 5,000 runs are reported (misclassification is defined as predicted incorrectly to receive treatment when the unit did not receive treatment or predicted incorrectly to not receive treatment, when the unit did in-fact receive treatment). The average AUC and the MSE (calculated as bias^2-variance) is also reported.

		Misclassification Rate				
		Worst case	Average	Best Case	AUC	MSE
Baseline	GME	40.8%	33.1%	26.2%	0.73	0.02
	ML Logit	40.8%	33.1%	26.2%	0.73	0.02
Small Sample	GME	60.0%	28.0%	5.0%	0.77	0.53
	ML Logit	60.0%	28.0%	5.0%	0.77	2.7
Rare Selection	GME	3.0%	1.0%	0.4%	0.78	0.30
	ML Logit	3.0%	1.0%	0.2%	0.78	0.60
Highly Collinear	GME	35.4%	43.4%	50.6%	0.59	0.80
	ML Logit	35.4%	43.4%	50.6%	0.59	0.81

The results for stage 1 are similar to those obtained by Golan, Judge and Perloff (1996) in their exploration of this method. In the baseline case, GME Logit performs as well as ML Logit at classification, and has similar MSE. In addition to the metrics reported by them, I also include the AUC-ROC. These are also comparable between GME Logit and ML Logit.

Comparing GME Logit's performance with ML Logit for deviations from the baseline case, I note that both methods perform equally well at classification (whether looking at misclassification rates or AUC-ROC). Although the GME Logit has much lower MSE than ML Logit for small samples and rare selection, these might not translate over to gains in the second stage of estimation. Only the estimate of the probability of selection into treatment (i.e. the propensity score) enters the second stage of estimation, and here both methods give similar results.

Second Stage

Recall that in my data generating process, the true treatment effect is "hard-coded" to a value of 2.0 (that is, a unit that receives treatment will have a 2.0 higher value for the outcome $Y_{i,t=1}$ if it receives treatment than it would otherwise). I now examine the ability of the second stage estimation (which simply involves replacing the weighted population expectations in equation 13 with their sample counterparts) to recover this treatment effect. In addition to the results obtained with weights derived from GME Logit and ML Logit first steps, I also show the results for the unweighted (or equal weighted) estimator. This estimator is biased and not consistent with the true ATT, but is shown to give researchers an idea of what to expect when unconditional parallel trends assumption is violated in the underlying data but the analysis does not account for the violation.

Table 4: Second Stage regression results. See above for specification of each case (leftmost column). The true ATT is 2.0.

	MSE			Bias		
	ME 1st -stage	ML 1st -stage	Unweighted	ME 1st -stage	ML 1st -stage	Unweighted
Baseline	0.02	0.02	0.08	-0.07	-0.07	0.24
Small Sample	4.4	4.4	0.46	-0.01	-0.01	0.47
Rare Selection	0.51	0.55	0.18	0.00	-0.01	-0.01
Highly Collinear	0.01	0.01	0.03	-0.08	-0.08	-0.16

In general, ML Logit and GME Logit perform equally well (when measured by bias or MSE) in recovering the true treatment effect. This is even the case when the first stage GME estimator has lower MSE (e.g. for rare selection into treatment or small sample sizes). In the second stage, the parameter estimates of the selection equation are not explicitly used. Only the propensity (likelihood of selection into treatment conditional on the covariates) enters the second stage weighting scheme. Golan, Judge and Perloff (1996) show that the GME estimator performs equally well with ML Logit when comparing misclassification rates, but has lower variance and lower MSE. Thus, the results of the second stage are not unexpected. However, researchers might still prefer to use the GME Logit for the first stage if estimating the parameters of the selection mechanism is important, since GME Logit has lower variance than ML Logit.

Extension: Sub-population treatment effects

So far, our treatment effect has been constant (set to 2.0 in the previous examples) regardless of the value of the covariates. Thus, although the baseline values ($Y_0 = 5 + 0.5X_1 + 0.3X_2 + \epsilon_0$), trend ($\Delta_0 = 0.4X_1 - 0.1X_2$) and selection into treatment ($\Pr(D = 1|X_1, X_2) = invlogit(0.5X_1 - 0.8X_2$) all depend on the covariates, the treatment effect for the entire population remains constant regardless of the values of X_1 and X_2. This might not be an accurate way to model how treatment works. Going back to the job training example, we might imagine that younger and more skilled workers could benefit more from training. If X_1 and X_2 are workers' ages and a proxy for skill, then the true treatment effect would be $\tau = g(X_1, X_2)$. Abadie's weighting method above would still recover the population-wide ATT, even if the treatment effect is dependent on covariates. However, researchers might be interested in studying the treatment effect for a substratum of the population (for example, for younger workers, for males vs. females, etc.). Abadie (2005) also suggests a way to estimate these subpopulation treatment effects, which is outlined below.

Consider X_k to be a deterministic function of the underlying covariates X (for example, X_k maybe a subset of X, or may contain indicator variables for X). Let $\mathcal{G} = \{g(X_k;\ \theta)|\theta \in \Theta \subset \mathbb{R}^k\}$ be a class of approximating functions, square integrable with respect to $\Pr(X_k|D = 1)$. Then a least squares approximation from $\mathcal{G}$ to $\mathbb{E}[Y_1(1) - Y_0(1)|X_k D = 1]$ is given by $g(X_k; \theta_0)$, where:

$$\theta_0 = argmin_{\theta \in \Theta} \mathbb{E}[\{E[Y_1(1) - Y_0(1)|X_k, D = 1]\}^2|D = 1]$$

(29)

For example, if $\mathcal{G} = \{X_k'\theta|\theta \in \Theta \subset \mathbb{R}^k\}$ then θ_0 defines a linear LS approximation to $\mathbb{E}[Y_1(1) - Y_0(1)|X_k D = 1]$. Abadie shows that the θ_0 can be recovered as:

$$\theta_0 = argmin_{\theta \in \Theta} \mathbb{E}\left[\Pr(D = 1|X) \cdot \left\{\rho_0 \cdot \left(Y_1 - Y(0)\right) - g(X_k; \theta)\right\}^2\right]$$

(30)

In the case when $\mathcal{G} = \{X_k'\theta|\theta \in \Theta \subset \mathbb{R}^k\}$, this simply becomes weighted LS (with weights obtained from the first stage).

It is known (see (Golan, Judge, and Miller 1996; Golan 2007) that the generalized maximum entropy linear model (GME Linear) is superior to LS for finite samples and ill-behaved data. Thus, it is natural to extend the weighted LS proposed of Abadie to a weighted GME approach, to see how it performs. To test this, I replace the constant treatment effect $\tau = 2.0$ with

$$\tau = 2.0 + 0.5X_1 - 0.3X_2$$

Since the treatment effect is now dependent on X_1 and X_2 and units with higher values of X_1 and lower values of X_2 are more likely to be selected into treatment, our ATT is now slightly modified. The true ATT is now $ATT \approx 2.34$.

For the weighted LS estimator, I use the predicted probabilities from the first stage as analytic weights (this is the same as multiplying the covariate matrix and outcome vector by $\sqrt{}(\Pr(D = 1|X)$. I use the GME Logit recovered probabilities (rather than the ML Logit probabilities) as both methods perform similarly, and researchers might prefer to use GME Logit for reasons discussed previously.

For the GME Linear estimation, transforming each observation by the square-root of the selection probability is not appropriate since the maximum entropy objective function is not of the square-loss variety. I use an unweighted GME Linear, but I rescale the variables ($y_i' = \frac{y_i}{sd(y_i)}; x_{n,i}' = \frac{x_{n,i}' - \bar{x}}{sd(x_n')}; where\ y_i = \rho_0 Y_i$) so that the GME Linear regression is around the same scale as the weighted LS. GME estimation also requires us to specify a support matrix for the βs and errors. I use empirical 6σ for the errors and the following support matrix for the coefficients

$$z = \begin{bmatrix} -20 & -10 & 0 & 10 & 20 \\ -20 & -10 & 0 & 10 & 20 \\ -20 & -10 & 0 & 10 & 20 \end{bmatrix}$$

(31)

Table 4 shows the results

… discussion of results to follow

Application to 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. As described by LaLonde, he 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 4: Treatment effects for male NWS participants as reported in LaLonde (1986) using the actual control group, and CPS-1 group drawn from the current population survey with similar characteristics as the treatment group

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

Note that naïve difference in 1978 earnings between the treatment group and CPS comparison group recovers a large negative impact of participating in the program. Using the experimental

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

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 for (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 4 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 5: Treatment effects using Abadie's weighting scheme (with GME first step)

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

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.