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 additional assumptions about the causal mechanisms to make a valid 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 (i.e. the gap in wages between program participants and non-participants would have remained constant in the absence of the training program). 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. Other methods include matching, inverse-probability weighting (IPW) or doubly robust diff-in-diff (which incorporates both a model for treatment assignment and model for outcome evolution conditional on covariates).

This paper applies and expands on a weighting scheme suggested by Abadie (2005). Specifically, it uses the generalized maximum entropy logistic regression (GME Logit) 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 information-theory (IT) based estimation of the selection equation outperforms the classical logistic regression (Logit) for estimating selection probabilities in certain situations. This is because the IT-based estimator is known to outperform maximum likelihood (ML, aka classical) Logit in all finite samples. The more complex and ill-behaved the data, the larger is 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 also develops a way to estimate this treatment effect for subpopulations, using the weights discussed above. This paper expands on Abadie’s method in this area as well, replacing the least-squares-based estimator he suggests with a generalized maximum entropy-based estimator. I also show that this method outperforms against the least-squares-based estimator under certain circumstances.

The paper proceeds as follows: section II reviews the literature on diff-in-diff, conditional diff-in-diff and generalized maximum entropy methods; 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 on job training first analyzed by LaLonde (1986) 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 an RCT, and attributed this to mismatches in outcome trends between the treatment group and potential comparison groups researchers might use (e.g. drawn from the current population survey, panel study of income dynamics or similar data sources for the general labor force). 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 difference-in-differences 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 denote unit potential outcome at time if it remains untreated in both periods, and if it is untreated in period and treated in period . Since all units are untreated in period , we can simplify notation to: and . Ideally, we would like to observe and for the same units. Then the treatment effect for unit would be . In practice, we can only observe for the treated units and for the untreated units. If is an indicator variable that takes a value if unit is treated in period , and otherwise, then the observed values are:

The potential outcomes framework defines a treatment effect for every unit . In practice, researchers are often interested in finding the average treatment effect on the treated (ATT):

Note that without additional assumptions, the ATT is not identified since we are unable to observe (the expected potential outcome if unit 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 difference-in-differences, the crucial assumption (known as the parallel trends assumption, summarized briefly earlier) is:

In words, the expected change in outcome between pre-treatment and post-treatment periods () is the same for treated and comparison groups. Under these assumptions, the ATT is:

Where the term in parenthesis on the right-hand side is our counterfactual . The above can be rewritten as:

Where the first term in the parenthesis is the change in the average outcome for treated units between period and 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 that estimating this model, involving four means is the same as a two-way fixed effects (TWFE) regression:

where: and are time and unit fixed effects and is the parameter of interest and is an idiosyncratic error term.

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 be a vector of observed determinants of changes in . We now assume that the groups would have followed parallel trends *conditional* on the covariates:

The literature offers several ways to estimate this conditional parallel-trends model. Firstly, the covariates can be incorporated into an augmented TWFE regression either as:

or as

where: is the observed value of the covariate at time , is an indicator variable that takes a value of 1 in the post-treatment period and 0 otherwise. Note that because time-invariant variables drop out in the TWFE regression, only the effect of changes in the levels of a covariate or differential trends related to baseline levels of the variable, interacted with a post-treatment dummy 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 (e.g. job training increasing earnings by 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 (together called the strong overlap assumption), some portion of the population remains untreated, that is:

then Abadie shows that the average treatment effect on the treated conditional on is:

where:

The average treatment effect can be recovered as:

Abadie does not specify how should be estimated. We need a model for the probability of selection . Commonly, researchers use logit. However, other methods that estimate the selection probability could also be used. The generalized maximum entropy approach for discrete choice (Golan, Judge, and Perloff 1996) (of which gme logit is a special case, with only two choices) is known to have several advantages over traditional 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 units are observed, where each unit is either selected into treatment or not. Let

be the probability of observing unit in the treatment state, conditional on covariates (of dimension and unknown parameters (of dimension ). If we have noisy data, we can write

where denotes the unknown and unobservable probability of selection and is a noise component for each observation contained in . We can reparametrize the error component as:

where are an H-dimensional vector of weights and are the H-dimensional support space. (Golan, Judge, and Perloff 1996) suggest and thus .

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

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

and the adding up constraints

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

where are Lagrange multipliers and .

The first order conditions are:

From which we obtain the estimated probability distributions:

and

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 a Monte Carlo simulation. The Monte Carlso simulations are based on my 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 observations and is replicated across 5,000 Monte Carlo runs. Each observation is characterized by two baseline covariates, ​ and ​, independently drawn from a standard normal distribution:

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

This structure ensures that baseline outcomes are systematically related to observed characteristics. In addition to level differences, the untreated trend (i.e., the evolution of outcomes in the absence of treatment) is also heterogeneous and depends on the covariates:

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:

where is the logistic cumulative density function (CDF). Each unit is then 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. The observed outcome ​ is constructed as the untreated outcome plus the untreated trend, a constant treatment effect for the treated, and additional noise:

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 ​ and ​, failure to account for these covariates can lead to biased estimates of the treatment effect.

Under the baseline assumptions, GME logit-based and classical logit-based models of selection into treatment should behave similarly. To test whether GME logit outperforms classical logit under certain circumstances, I test several cases. Firstly, GME is known to outperform classical 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 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

which causes only ~1% of observations to be selected into treatment). GME logit is also known to outperform classical logit when covariates are highly collinear. I simulate this by drawing and from a multivariate normal distribution with . The tables below summarize the simulation settings.

|  |  |  |
| --- | --- | --- |
| Table 1: Baseline simulation settings. All simulations have 5,00 runs | | |
| Observations (n) |  |  |
| Covariate Distribution ( |  |  |
| Treatment Heterogeneity ( |  |  |
| Selection equation |  |  |
| Errors |  |  |

|  |  |
| --- | --- |
| Table 2: Modifications to the baseline case for specific scenarios. All simulations have 5,000 runs | |
| Case: Small sample size |  |
| Case: Rare selection |  |
| Case: Highly collinear |  |

For the estimation of the GME logit model, I use a support space of 3 with as suggested by Golan, Judge and Perloff (1996). I do not use any prior distributions for GME logit. Note that GME does not require re-parametrization of (as is the case for GME linear regression).

Table 3 shows the Monte Carlo simulation results with the above settings:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 3: Simulation Results | | | | | | | | | |
|  | GME | | | Logit | | | Uncorrected | | |
|  | Bias | Variance | MSE | Bias | Variance | MSE | Bias | Variance | MSE |
| Baseline Case | -0.071 | 0.104 | 0.016 | -0.071 | 0.011 | 0.016 | 0.244 | 0.017 | 0.077 |
| Small Sample | -1.062 | 0.196 | 1.323 | -1.313 | 0.206 | 1.929 | -0.136 | 0.255 | 0.274 |
| Highly Collinear | -0.039 | 0.004 | 0.006 | -0.039 | 0.004 | 0.006 | -0.105 | 0.008 | 0.019 |
| Rare Selection | -0.238 | 57.384 | 57.440 | -0.236 | 57.794 | 57.850 | -0.031 | 0.274 | 0.275 |
| Noisy case | 0.000 | 0.009 | 0.009 | 0.000 | 0.009 | 0.009 | 0.280 | 0.018 | 0.096 |
| misspecified (case 1) | 0.042 | 0.009 | 0.011 | 0.042 | 0.009 | 0.011 | 0.119 | 0.119 | 0.032 |
| misspecified (case 2) | 0.098 | 0.010 | 0.019 | 0.099 | 0.010 | 0.019 | 0.369 | 0.018 | 0.154 |
| misspecified (case 3) | -0.620 | 286.708 | 287.093 | -0.742 | 287.014 | 287.565 | 0.224 | 0.467 | 0.517 |
| misspecified (case 4) | 0.020 | 0.207 | 0.207 | -0.002 | 0.224 | 0.224 | 0.262 | 0.384 | 0.453 |
| random params | -0.045 | 0.010 | 0.012 | -0.045 | 0.010 | 0.012 | 0.277 | 0.027 | 0.103 |

The GME-based approach outperforms traditional logit in small-sample cases (as seen in MSE for rows small sample, misspeficied case 4) and performs as well as classical logit in all other cases. Note that this improvement in the MSE is usually driven by smaller variance of the estimator, while bias remains comparable.

# 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. 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 3: 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]](#footnote-1) | Without Age | With Age |
| Actual Controls | $886 (476) | $798 (472) | $847 (560) | $856 (558) |
| CPS-1 | -$8,870 (562) | -$4,416 (557) | $1,714 (452) | $195 (441) |

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 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 4: 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) |

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.

1. The exogenous variables used in the regression adjusted equations are age, age squared, years of schooling, high school dropout status and race [↑](#footnote-ref-1)