Prediction decomposition for causal analysis

Ofir Reich*

May 18, 2025

Abstract

There is rising interest in using Machine Learning (ML) model predictions as outcomes in causal analysis. However, these methods have faced challenges in finding the true treatment effects. It is also challenging to make choices about which prediction models to choose, since we are interested not only in the accuracy of the prediction but in its ability to produce the correct causal effect in the analysis. In this paper I propose a decomposition of the prediction into between-unit prediction, within-unit-across-time prediction, and counterfactual-treatment-effect prediction. I show that the third one is of interest, but only the first two can be estimated from non-experimental data. I propose that between-unit prediction accuracy is a better proxy for counterfactual-treatment-effect prediction than the overall prediction accuracy. This suggests a metric to measure whether the trained ML model is better or worse at producing outcomes which would reveal the true causal effect. Under some assumptions, it also enables constructing an unbiased estimate for the treatment effect. The method requires panel data, with at least two time periods. This allows making informed choices about the right model to generate the ML-predicted outcomes, and a criterion for telling whether it is not good enough. I illustrate the metric with simulations of synthetic data, and justify it with a theoretical framework.

1 Introduction

There is rising interest in using Machine Learning (ML) model predictions as outcomes in causal analysis. This could be in a Randomized Controlled Trial (RCT) or other kinds of analysis. The advantage of using ML predictions is that they are generally easier and cheaper to generate for large samples than individual data collection. One prominent example is research studying the effects of Unconditional Cash Transfers on various social outcomes (e.g. consumption or wealth), where the outcome is predicted using mobile phone Call Detail Records (Barriga-Cabanillas et al. [2025], Aiken et al. [2025]). Another example is studying the effects of agricultural interventions on yields, where yields are predicted using Remote Sensing data (Cole et al. [2025], Burke et al. [2021], Lobell et al. [2020]). And more (Ratledge et al. [2022]).

The general structure of these analyses is as follows:

*ofir.re@gmail.com

1. The actual outcome of interest is collected for a subsample of data.
2. A Machine Learning (ML) model is trained on this subsample of data for which both the actual label (the outcome of interest) and the broadly-available feature data (e.g. Call Detail Records, or Remote Sensing data) are known.
3. The ML model is used to predict the outcome of interest for each randomization unit, especially those outside the labeled subsample.
4. The causal analysis then proceeds as usual, using these ML-predicted outcomes as the outcome in the analysis. (This is not to be confused with using Machine Learning methods for the causal analysis itself when all data are known, which is an unrelated topic).

Intervention	**Outcome of interest**	**ML feature data**	**Unit of analysis**
Unconditional Cash Transfers	Consumption, wealth	Mobile phone Call Detail Records	Person
Agricultural extension	Yield	Remote sensing pictures	Plot

Table 1: *Examples of interventions studied using ML-predicted outcomes*

However, these methods have faced challenges in finding the true treatment effects (Barriga-Cabanillas et al. [2025], Aiken et al. [2025]). A true causal effect exists when using the actual outcome data, but not when using the ML-predicted outcome data (Aiken et al. [2025]). Or the distribution of the ML-predicted outcomes is compressed relative to the label distribution, biasing the estimated effect downwards (Ratledge et al. [2022]). This is naturally a problem for this approach, especially if it cannot be known in advance without collecting the ground-truth outcome data for the entire sample, which defeats the purpose of using ML-predicted outcomes. A related problem is choosing the best ML algorithm, the right features etc. - since we are interested not only in the accuracy of the prediction but in its ability to produce the right causal effect in the analysis.

In this paper I propose a metric to measure whether the trained ML model is better or worse at producing outcomes which would reveal the true causal effect. It requires panel data of the same form described above, for at least two time periods. This allows making informed choices about the right model to generate the ML-predicted outcomes, and a criterion for telling whether it is not good enough. I illustrate the metric with simulations of synthetic data, and justify it with a theoretical framework.

2 Intuition - fitting to between-unit differences

The problem with using ML models to predict outcomes which are later used in causal analysis is that ML models only "care" about prediction, whereas causal inference asks a

question about the counterfactual. More concretely, a ML-model could predict very accurately the observed outcome for an individual, but not necessarily the difference between the outcome with and without Treatment. The ML model fits (among other things) to the variation *between* units, whereas causal inference is interested in (counterfactual) variation *within* unit. So for example a Machine Learning model using Call Detail Records could learn to predict a person's consumption level by whether they live in a wealthier area, and get very good prediction results. But a Cash Transfer would not change the area of the person's house, so the model would predict a zero effect of Cash Transfers on consumption, even if such an effect exists. Another example is that a yield prediction model could use only the farmer's last-season yield to predict their current season's yield, have very good predictions but obviously will not be able to detect any treatment effect.

We want to know that our model does not only fit to between-unit variation, but also to the treatment effect, and at the very least to natural (not counterfactual) within-unit variation.

3 Basic case: 2 time periods

We shall start with the simple case where outcome data is collected for only two time periods, for a subsample of units. Suppose the setting is a RCT, and the units are people.

Suppose the actual outcome for a person i in time t with/without Treatment is modeled as:

$$\text{actualOutcome}_{i,t} = \alpha + \mu_i + \gamma \text{Treat}_{i,t} + \epsilon_{i,t} \tag{1}$$

where:

- α is a constant intercept
- μ_i are person fixed characteristics, with mean 0.
- $\text{Treat}_{i,t}$ is a random Treatment indicator
- $\epsilon_{i,t}$ is an independent error term
- t, the time period, takes values 1 and 2.

Note that α, μ_i and $\text{Treat}_{i,t}$ are pairwise uncorrelated. Likewise, $\epsilon_{i,1}$ and $\epsilon_{i,2}$ are uncorrelated, due to μ_i soaking up that correlation.

3.1 Pathological case: ML predictions fit only to between-person variation

Now suppose we trained a ML model only on untreated persons, and that the prediction fits only (and perfectly) to the person fixed characteristics.

$$\text{predictedOutcome}_{i,t} = \alpha + \mu_i \tag{2}$$

The prediction could be good when measured between people (say, on a test set without experimental variation), and even have good R-squared. But when using it to measure treatment effect, we would try to run the regression:

$$(\alpha + \mu_i =)\text{predictedOutcome}_{i,t} = \xi_0 + \xi_1 \text{Treat}_{i,t} + \nu_{i,t} \tag{3}$$

And since $\text{Treat}_{i,t}$ is independent of μ_i and so uncorrelated, we would estimate $\xi_1 = 0$, i.e. a treatment effect of 0.

A way to know that this is the case beforehand is to check if the ML-prediction explains within-person variation across time periods, even for non-treated persons (so $\text{Treat}_{i,t} = 0$).

In this pathological case we would find

$$\text{predictedOutcome}_{i,2} - \text{predictedOutcome}_{i,1} = \mu_i - \mu_i = 0 \tag{4}$$

$$\text{actualOutcome}_{i,2} - \text{actualOutcome}_{i,1} = (\epsilon_{i,2} - \epsilon_{i,1}) \neq 0 \tag{5}$$

So we would find that the predicted outcome explains precisely 0 percent of the within-person variation in outcome. This will be the basis of the metric we develop.

3.2 More realistic case: ML predictions fit partially to between-person variation

We shall now generalize to less pathological cases, where the ML predictions fit partially to between-person variation, partially to within-person between-period variation and partially to within-person between-counterfactuals variation.

As before

$$\text{actualOutcome}_{i,t} = \alpha + \mu_i + \gamma \text{Treat}_{i,t} + \epsilon_{i,t} \tag{6}$$

Again suppose we trained a ML model only on untreated persons. We now decompose the ML-predicted outcome (on the test set, still of untreated persons) into these same components, as follows:

$$\text{predictedOutcome}_{i,t} = \alpha + \eta_\mu \mu_i + \eta_T \gamma \text{Treat}_{i,t} + \eta_\epsilon \epsilon_{i,t} + \nu_{i,t} \tag{7}$$

where

- η_μ – Indicates how well the model fits to between-unit variation. A value of 1 means the model perfectly captures differences between individuals, while 0 means it ignores these differences entirely.
- η_ϵ – Measures how well the model fits to within-unit temporal variation. When $\eta_\epsilon = 1$, the model perfectly captures how individuals naturally change over time (unrelated to treatment), while $\eta_\epsilon = 0$ means the model misses this variation completely.
- η_T – Represents how well the model captures counterfactual treatment effects. An η_T value of 1 means the model perfectly predicts the causal impact of treatment, while 0 indicates the model is completely insensitive to treatment effects.
- $\nu_{i,t}$ is an independent error term, uncorrelated with the other elements.

- We assumed that the ML model gets the mean of the population, α, right.

Note that even though the ML model was not trained on Treated persons (so on a population without Treatment variation), it could still capture some or all of the counterfactual Treatment variation - for example a perfectly accurate predictor would capture all of it.

Now suppose that we use our ML predicted outcomes as the dependent variable in treatment effect regression, in the standard fashion.

$$\text{predictedOutcome}_{i,t} = \xi_0 + \xi_1 \text{Treat}_{i,t} + \delta_{i,t} \tag{8}$$

Using the decomposition of predicted outcomes above, and the fact that Treatment was randomly assigned and is uncorrelated with other elements, we can see that the estimated treatment effect would be:

$$\hat{\xi}_1 = \eta_T \cdot \gamma \tag{9}$$

and not the desired true treatment effect, γ.

Ideally, therefore, we want the model with $\eta_T = 1$, and we would find it using a population with experimental variation and check which model predicts it best. But usually this is not observable because the models are trained on non-Treated populations, and if we had a large enough sample to know which model best predicts the treatment effect, we would probably have a large enough sample to just estimate the treatment effect from that population directly. In lieu of estimating η_T, we conjecture that η_ϵ is a better proxy for η_T than the overall prediction accuracy, and therefore estimate η_ϵ. Before we derive our method, we shall see the problem with using prediction accuracy as the model selection criterion.

3.2.1 The problem with prediction accuracy

Normally ML models are chosen based on their prediction accuracy on the test set (a random holdout from the labeled training set). Considering our decomposition of prediction, we can now see why this is inadequate. If we estimated the overall prediction performance of the ML model on a non-Treated population (similar to its training set), using R-squared, we would have, using the equations above and the pairwise uncorrelatedness:

$$\text{Cov}[\text{predictedOutcome}_{i,t}, \text{actualOutcome}_{i,t}] = \eta_\mu \text{Var}[\mu_i] + \eta_\epsilon \text{Var}[\epsilon_{i,t}] \tag{10}$$

$$R^2 = \frac{\text{Cov}[\text{predictedOutcome}_{i,t}, \text{actualOutcome}_{i,t}]^2}{\text{Var}[\text{predictedOutcome}_{i,t}]\text{Var}[\text{actualOutcome}_{i,t}]} \tag{11}$$

$$= \frac{(\eta_\mu \text{Var}[\mu_i] + \eta_\epsilon \text{Var}[\epsilon_{i,t}])^2}{(\eta_\mu^2 \text{Var}[\mu_i] + \eta_\epsilon^2 \text{Var}[\epsilon_{i,t}] + \text{Var}[\nu_{i,t}])(\text{Var}[\mu_i] + \text{Var}[\epsilon_{i,t}])}. \tag{12}$$

This depends on the values of η, but also on the relative size of the variances, $\text{Var}[\mu_i]$, $\text{Var}[\epsilon_{i,t}]$, $\text{Var}[\nu_{i,t}]$. If $\text{Var}[\mu_i] \gg \text{Var}[\epsilon_{i,t}]$, then η_ϵ will not matter much for the R-squared. We shall see this in simulations later.

3.2.2 Estimating η_ϵ

We shall now define and calculate a few quantities that will help us estimate η_ϵ. Define the difference across time periods within person for any variable $X_{i,t}$ as

$$\Delta X_i = X_{i,t=2} - X_{i,t=1}. \tag{13}$$

Calculating this difference for actualOutcome we get:

$$\Delta\text{actualOutcome}_i = \gamma\Delta\text{Treat}_i + \Delta\epsilon_i \tag{14}$$

and similarly for $\Delta\text{predictedOutcome}_i$:

$$\begin{aligned}\Delta\text{predictedOutcome}_i &= \text{predictedOutcome}_{i,2} - \text{predictedOutcome}_{i,1} &(15)\\ &= \eta_T\gamma\Delta\text{Treat}_i + \eta_\epsilon\Delta\epsilon_i + \Delta\nu_i &(16)\end{aligned}$$

The left-hand-side quantities are observable for our subsample of persons with collected outcomes in both periods. We can see how well the actual outcome predicts the predicted outcome, in terms of linear regression.

$$\begin{aligned}\text{Cov}[\Delta\text{predictedOutcome}_i, \Delta\text{actualOutcome}_i] &= \eta_T\gamma^2\text{Var}[\Delta\text{Treat}_i] + \eta_\epsilon\text{Var}[\Delta\epsilon_i] &(17)\\ &= \eta_T\gamma^2\text{Var}[\Delta\text{Treat}_i] + \eta_\epsilon\text{Var}[\Delta\epsilon_i] &(18)\end{aligned}$$

But observe that

$$\begin{aligned}\text{Var}[\Delta\epsilon_i] &= \text{Var}[\epsilon_{i,2} - \epsilon_{i,1}] &(19)\\ &= \text{Var}[\epsilon_{i,2}] + \text{Var}[\epsilon_{i,1}] &(20)\\ &= 2\text{Var}[\epsilon_{i,t}] &(21)\end{aligned}$$

where the next-to-last equality is because errors are uncorrelated across periods, due to μ_i soaking up that correlation.

So finally we have

$$\text{Cov}[\Delta\text{predictedOutcome}_i, \Delta\text{actualOutcome}_i] = \eta_T\gamma^2\text{Var}[\Delta\text{Treat}_i] + \eta_\epsilon 2\text{Var}[\epsilon_{i,t}] \tag{22}$$

If we are working with a subsample of non-Treated persons, this simplifies even further.

$$\begin{aligned}\text{Cov}[\Delta\text{predictedOutcome}_i, \Delta\text{actualOutcome}_i] &= \eta_\epsilon 2\text{Var}[\epsilon_{i,t}] &(23)\\ \text{Var}[\Delta\text{actualOutcome}_i] &= 2\text{Var}[\epsilon_{i,t}] &(24)\\ & &(25)\end{aligned}$$

If we ran a linear regression of $\Delta\text{predictedOutcome}_i$ on $\Delta\text{actualOutcome}_i$, without an intercept:

$$\Delta\text{predictedOutcome}_i = \beta\Delta\text{actualOutcome}_i + \text{error}_i, \tag{26}$$

the estimated coefficient is

$$\hat{\beta} = \frac{\text{Cov}[\Delta\text{predictedOutcome}_i, \Delta\text{actualOutcome}_i]}{\text{Var}[\Delta\text{actualOutcome}_i]} = \frac{\eta_\epsilon 2\text{Var}[\epsilon_{i,t}]}{2\text{Var}[\epsilon_{i,t}]} = \eta_\epsilon \tag{27}$$

This is the parameter we are interested in estimating, in absence of a direct estimate of η_T.

This empirical diff-vs.-diff slope will be our estimate of η_ϵ, and the standard error from the regression will be our standard error for the estimate.

3.2.3 Estimating η_μ

In a non-Treated population, doing some algebra using the definitions and the uncorrelatedness of the different components, we have:

$$\text{Cov}[\text{predictedOutcome}_{i,t}, \text{actualOutcome}_{i,t}] = \eta_\mu \text{Var}[\mu_i] + \eta_\epsilon \text{Var}[\epsilon_{i,t}] \tag{28}$$

$$\text{Cov}[\Delta\text{predictedOutcome}_i, \Delta\text{actualOutcome}_i] = \eta_\epsilon 2\text{Var}[\epsilon_{i,t}] \tag{29}$$

$$\text{Var}[\text{actualOutcome}_{i,t}] = \text{Var}[\mu_i] + \text{Var}[\epsilon_{i,t}] \tag{30}$$

$$\text{Var}[\Delta\text{actualOutcome}_i] = 2\text{Var}[\epsilon_{i,t}] \tag{31}$$

We can perform some algebra to isolate η_μ, where for ease of notation we replace predictedOutcome with predOut and actualOutcome with actualOut:

$$\text{Var}[\text{actualOut}_{i,t}] - \frac{1}{2}\text{Var}[\Delta\text{actualOut}_i] = \text{Var}[\mu_i] \tag{32}$$

$$\text{Cov}[\text{predOut}_{i,t}, \text{actualOut}_{i,t}] - \frac{1}{2}\text{Cov}[\Delta\text{predOut}_i, \Delta\text{actualOut}_i] = \eta_\mu \text{Var}[\mu_i] \tag{33}$$

$$\Rightarrow \frac{\text{Cov}[\text{predOut}_{i,t}, \text{actualOut}_{i,t}] - \frac{1}{2}\text{Cov}[\Delta\text{predOut}_i, \Delta\text{actualOut}_i]}{\text{Var}[\text{actualOut}_{i,t}] - \frac{1}{2}\text{Var}[\Delta\text{actualOut}_i]} = \frac{\eta_\mu \text{Var}[\mu_i]}{\text{Var}[\mu_i]} = \eta_\mu \tag{34}$$

The left-hand side of the last equation will be our estimate for η_μ. Standard errors can be estimated using bootstrap.

3.2.4 Estimating η_T

This is done by estimating the Treatment Effect twice, with a standard treatment effect regression. Once using the actual outcome, and once using the predicted outcome. We obviously need experimental variation for this, which we did not need to estimate the previous two coefficients, which makes this approach infeasible in most practical applications.

The specification

$$\text{actualOutcome}_{i,t} = \alpha + \mu_i + \gamma \text{Treat}_{i,t} + \epsilon_{i,t} \tag{35}$$

(If is $\text{Treat}_{i,t}$ is fixed within unit, $\text{Treat}_{i,t} = \text{Treat}_i$, then the regression is run without the Fixed-Effects μ_i)

Gives us the estimate $\hat{\gamma}_{\text{actual}} \to \gamma$.

Running the same regression with predictedOutcome as the left-hand side gives us the estimate $\hat{\gamma}_{\text{predicted}} \to \eta_T \gamma$. So our estimate for η_T will be:

$$\hat{\eta}_T = \frac{\hat{\gamma}_{\text{predicted}}}{\hat{\gamma}_{\text{actual}}} \tag{36}$$

Here, too, standard errors can be estimated using bootstrap.

4 Simulations

4.1 Simulation process

We use the framework above to simulate synthetic data and run several different tests. All code used to generate these simulations and charts in the paper is publicly available on GitHub (Reich [2025]).

4.1.1 Simulating actual outcomes

Recall our notation:

$$\text{actualOutcome}_{i,t} = \alpha + \mu_i + \gamma \text{Treat}_{i,t} + \epsilon_{i,t} \tag{37}$$

We simulate actual outcomes for each person for two time periods. We selected the standard deviation of μ relative to the standard deviation of ϵ according to the results in a real setting which is not part of this paper (Cole et al. [2025]). In that setting the share of variance explained by person fixed effects was 0.92, so we used this value. Notice that this is a large share of the variance, but it comes from real data and is not necessarily atypical.

Specifically, we used values:

$$\alpha = 3200 \tag{38}$$
$$\text{SD}(\mu) = 1400 \tag{39}$$
$$\text{SD}(\epsilon) = 600 \tag{40}$$
$$\gamma = 200 \tag{41}$$
$$P(\text{Treat}_{i,t}) = P(\text{Treat}_i) = 0.5 \tag{42}$$

Treatment was fixed within person across time, so $\text{Treat}_{i,t} = \text{Treat}_i$.

We can then draw random numbers to obtain simulated values of μ_i, $\epsilon_{i,t}$, $\text{Treat}_{i,t}$ for each person in each time period, and add them up to calculate the actual outcome for each person and time period using the formula for $\text{actualOutcome}_{i,t}$ above. We used a Log-Normal distribution for $(\alpha + \mu_i)$ to simulate the fat-tailed distribution of between-person outcomes present in many real-world settings.

We simulate those outcomes once, and hold them fixed when simulating the predicted outcomes explained below.

4.1.2 Simulating predicted outcomes

Recall our decomposition for predicted outcomes:

$$\text{predictedOutcome}_{i,t} = \alpha + \eta_\mu \mu_i + \eta_T \gamma \text{Treat}_{i,t} + \eta_\epsilon \epsilon_{i,t} + \nu_{i,t} \tag{43}$$

We simulate the predicted outcomes as follows. We iterate over parameter values (between 0 and 1 in skips of 0.25) for each of η_μ, η_T, η_ϵ, and values (between 0 and 1,000 in skips of 250) for $\text{Var}[\nu]$. We used the full cartesian product of the parameter values, so each combination of values is obtained. So each simulation has a set of parameter values for $\eta_\mu, \eta_T, \eta_\epsilon, \text{Var}[\nu]$. For each simulation, we generate $\nu_{i,t}$ for each person and period, and calculate the predicted outcomes for each person and period using their values of μ_i, $\epsilon_{i,t}$, $\text{Treat}_{i,t}$, $\nu_{i,t}$, and the values of η_μ, η_T, η_ϵ by substituting in the formula for $\text{predictedOutcome}_{i,t}$ above.

4.1.3 Calculating stats

We calculate various stats:

- We run a linear regression of $\text{actualOutcome}_{i,t}$ on $\text{predictedOutcome}_{i,t}$ (without an intercept). We estimate:
  - "ML prediction R-squared": R-squared of this regression.
- We run a linear regression of $\Delta\text{actualOutcome}_i$ on $\Delta\text{predictedOutcome}_i$ (without an intercept).
$$\Delta\text{predictedOutcome}_i = \beta\Delta\text{actualOutcome}_i + \text{error}_i \tag{44}$$
  We estimate:
  - The slope, i.e. the coefficient estimated for $\Delta\text{predictedOutcome}_i$.
  - "Diff prediction R-squared": R-squared of this regression.
- Compression: ratio StD[predictedOutcome]/StD[actualOutcome].
- Share of variance in *actualOutcome* explained by person fixed-effects μ_i (calibrated to 0.92). Note that this is a parameter we have chosen for the simulations, but it comes from a real setting, and it holds in many real settings - the treatment effect is almost always small compared with between-unit variation. This is why RCTs often spend time and effort collecting pre-intervention outcomes and perform a difference-in-differences analysis, since it cancels precisely that variation between units.

We then run an ordinary treatment effects linear regression using those *predictedOutcomes* as the target variable:

$$\text{predictedOutcome}_{i,t} = \xi_0 + \xi_1\text{Treat}_{i,t} + \delta_{i,t} \tag{45}$$

We estimate the Treatment Effect, $\hat{\xi}_1$.

We also estimate the actual Treatment Effect (which is different than the expectation due to statistical noise in the simulation) using the actual outcome, which is unknown in real settings where actualOutcome was not collected for the entire sample:

$$\text{actualOutcome}_{i,t} = \gamma_0 + \gamma_1\text{Treat}_{i,t} + \delta_{i,t} \tag{46}$$

We calculate two other stats from this regression:

- "Scaled Treatment Effect": estimated slope divided by the actual Treatment Effect: $\hat{\xi}_1/\hat{\gamma}_1$.
- t-statistic for the slope coefficient.

4.2 Simulation results

We find the following.

4.2.1 Better ML-prediction does not guarantee more accurate treatment estimation.

There could be two different models, one of which would have vastly inferior R-squared in predicting actual outcome, but would be better at giving the true treatment effect. See Figure 1.

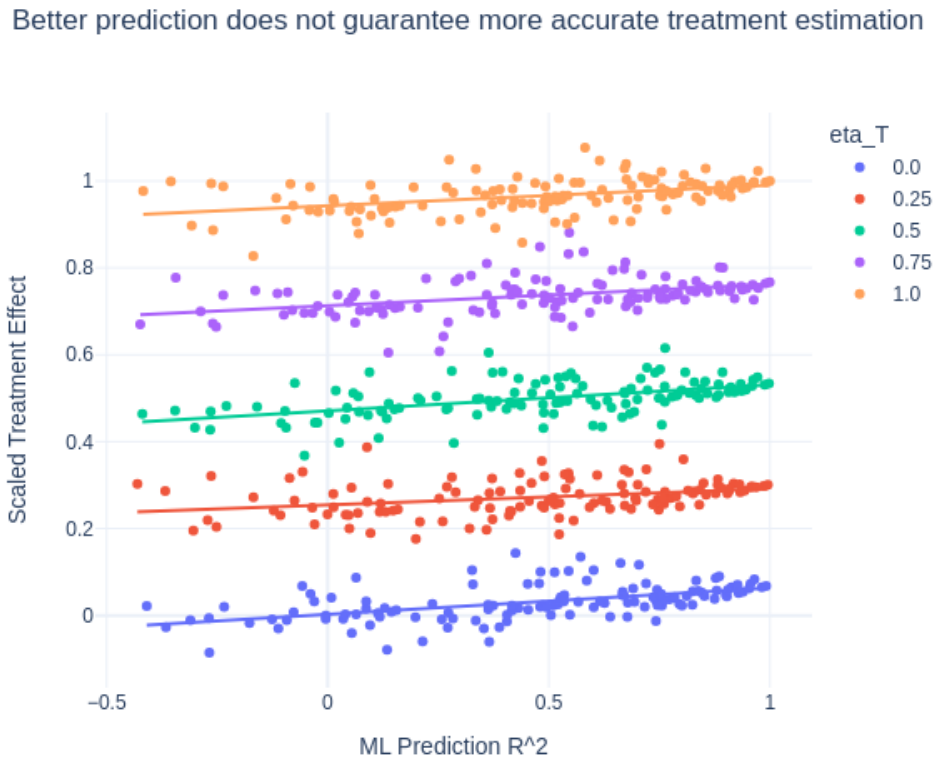

Figure 1: *Each dot is a specific simulation of actual outcomes and predicted outcomes. The x-axis marks the R-squared of the ML prediction, and the y-axis the Scaled Treatment Effect (where 1 is the correct effect). The color of the dot is by η_T. The trendline is for all points, using OLS. We can see the main determinant of the Scaled Treatment Effect is η_T, where the general prediction R-squared matters very little.*

4.2.2 Better ML-prediction is mostly determined by person attributes

In our setting, where $\text{StD}[\mu] \gg \text{StD}[\gamma\text{Treat}]$, better prediction is mostly about capturing the person fixed effects, and so η_T does not greatly affect the prediction accuracy. See Figures 2 and 3.

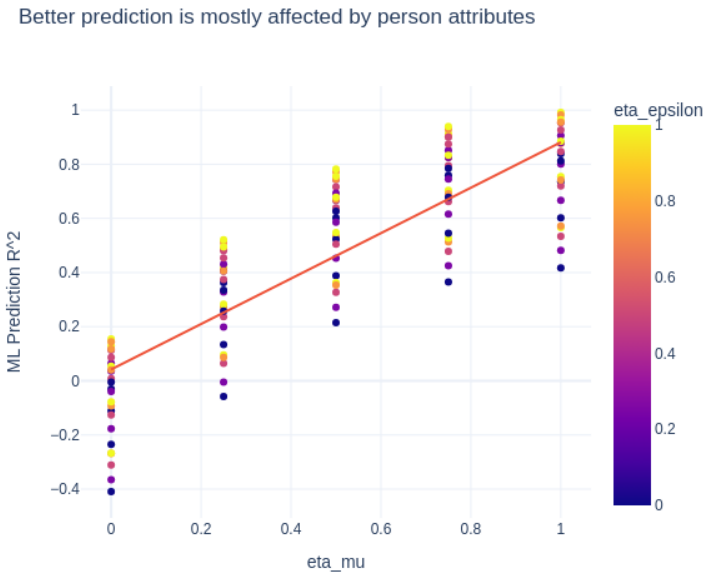

Figure 2: *Each dot is a simulation. X-axis is η_μ. Y-axis is ML prediction R-squared. Trendline is OLS. Higher η_μ strongly predicts higher prediction R-squared.*

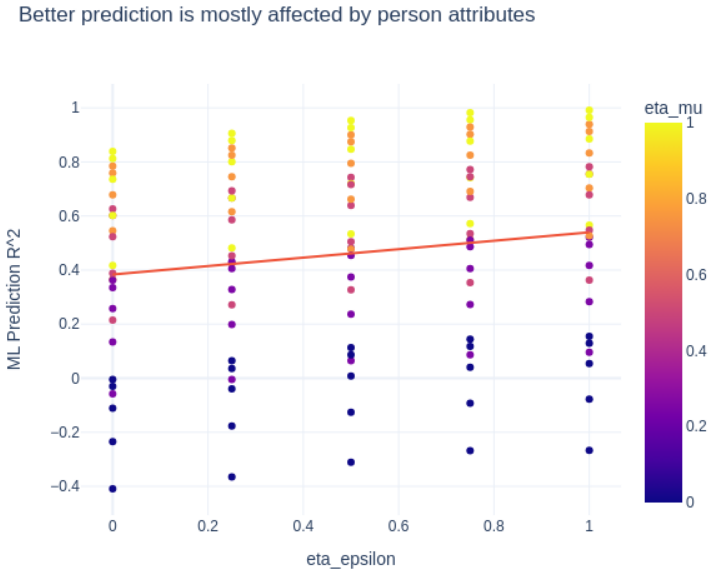

Figure 3: *Each dot is a simulation. X-axis is η_ϵ. Y-axis is ML prediction R-squared. Trendline is OLS. η_ϵ and prediction R-squared don't have a strong relationship.*

4.2.3 Better ML-prediction does not guarantee more statistical power for detecting treatment effect

See Figure 4.

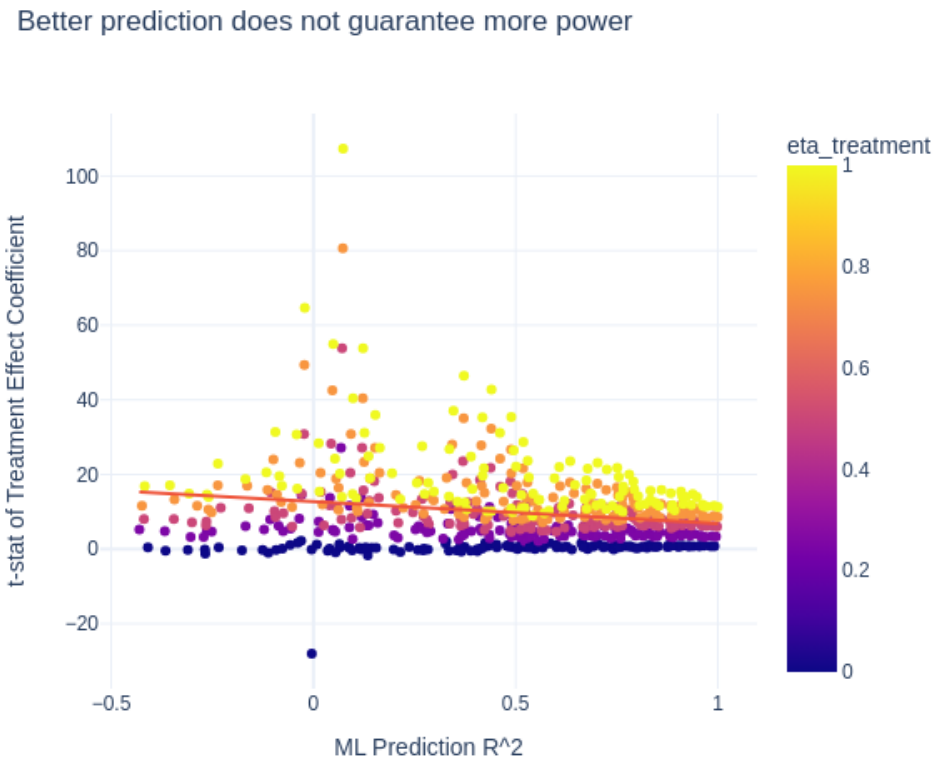

Figure 4: *Each dot is a simulation. X-axis is ML prediction R-squared. Y-axis is the t-statistic for the coefficient of Treat in the Treatment Effect regression. Trendline is OLS. The t-statistic is mostly determined by η_T, not by the prediction R-squared. A few very extreme outliers in t-statistic were discarded.*

4.2.4 Distribution compression of ML-Predicted outcomes is not predictive of treatment effect compression

This is for the same reason - the compression is mostly about η_μ where the correct estimated treatment effect is mostly affected by η_T. This is important since some papers (e.g. Ratledge et al. [2022]) have proposed to target the compression directly (and even artificially inflate the predictions) as a method for dealing with the treatment effect being compressed. This would only work if η_μ and η_T are similar, since artificial inflation of the prediction inflates both. See Figure 5.

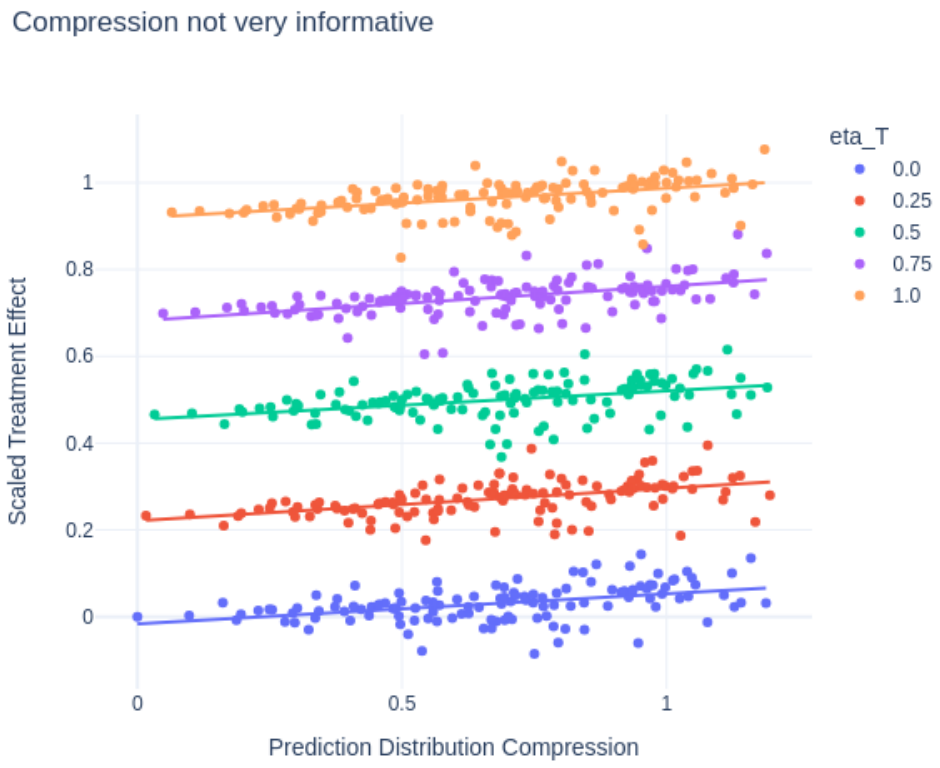

Figure 5: *Each dot is a simulation. X-axis is the compression ratio StD(predictedOutcome)/StD(actualOutcome). Y-axis is Scaled Treatment Effect. Trendline is OLS.*

4.2.5 Diff-vs-diff regression predicts the scaled treatment effect, when $\eta_T = \eta_\epsilon$

When we restrict ourselves to cases where $\eta_T = \eta_\epsilon$, meaning the prediction fits to within-person variation as well as it fits to counterfactual treatment variation, then our method of estimating η_ϵ using the diff-vs-diff regression predicts the Scaled Treatment Effect rather well. See Figure 6. This is to be expected, since the Scaled Treatment Effect is largely determined by η_T.

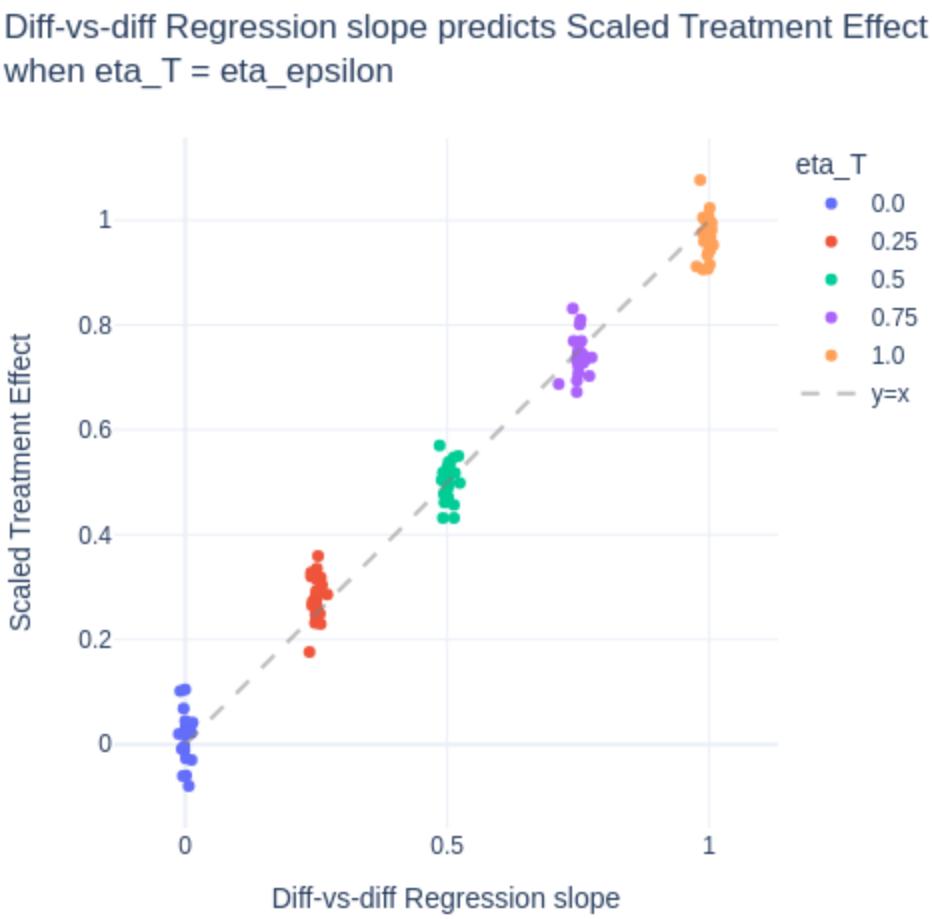

Figure 6: *Each dot is a simulation. X-axis is the diff-vs-diff regression slope (our estimate for η_ϵ). Y-axis is Scaled Treatment Effect. Color is by η_T. Dashed line is y=x. When we restrict ourselves to cases where $\eta_T = \eta_\epsilon$, meaning the prediction fits to within-person variation as well as it fits to counterfactual treatment variation, then our method of estimating η_ϵ using the diff-vs-diff regression predicts the Scaled Treatment Effect rather well.*

When our assumption that $\eta_T = \eta_\epsilon$ holds, we can correct the bias in the estimated Scaled Treatment Effect and arrive at an unbiased estimate of the Treatment Effect:

$$\text{UnbiasedTreatmentEffect} = \text{EstimatedTreatmentEffect}/\hat{\eta}_\epsilon \tag{47}$$

Importantly, this is only true if we assume $\eta_T = \eta_\epsilon$. If there is no correlation between η_T and η_ϵ, obviously our regression slope aimed at estimating η_ϵ is not helpful at predicting the Scaled Treatment Effect affected by η_T. See Figure 7.

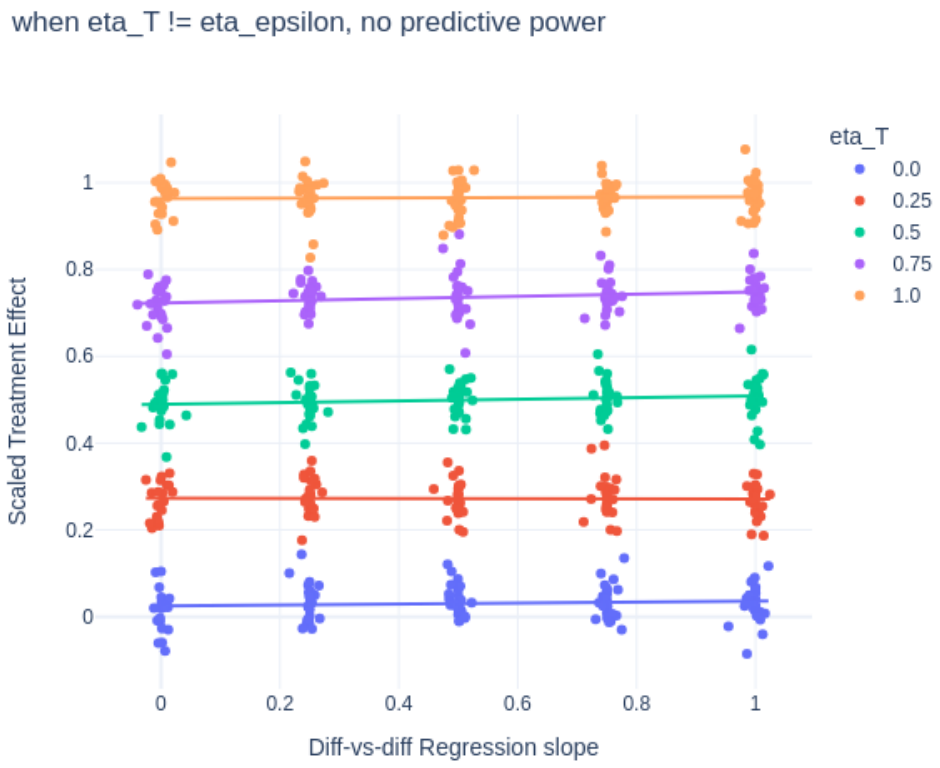

Figure 7: *Each dot is a simulation. X-axis is the diff-vs-diff regression slope (our estimate for η_ϵ). Y-axis is Scaled Treatment Effect. Trend lines are OLS. With no restriction that $\eta_T = \eta_\epsilon$, there is no relationship between our estimate and the Scaled Treatment Effect.*

It remains an empirical question what is the typical relationship between η_μ, η_T, η_ϵ. We conjecture that in many contexts η_T and η_ϵ will be related, at the very least more related than η_T and η_μ. But this remains to be studied in real settings, and it is possible that in practice some combination of η_μ, η_ϵ will be a better predictor of η_T than only η_ϵ.

5 Generalization to more than 2 time periods

The generalization to more than 2 time periods is relatively straightforward. The decomposition remains the same, except now t can take on more than 2 values, $t = 1..T$.

$$\text{actualOutcome}_{i,t} = \alpha + \mu_i + \gamma \text{Treat}_{i,t} + \epsilon_{i,t} \tag{48}$$

$$\text{predictedOutcome}_{i,t} = \alpha + \eta_\mu \mu_i + \eta_T \gamma \text{Treat}_{i,t} + \eta_\epsilon \epsilon_{i,t} + \nu_{i,t} \tag{49}$$

The generalization of the time-differences $\Delta X_i := X_{i,2} - X_{i,1}$ are centered variables,

denoted by tildes:

$$\tilde{X}_{i,t} := X_{i,t} - \bar{X}_i = X_{i,t} - \frac{1}{T}\sum_{s=1}^{T} X_{i,s} \tag{50}$$

5.1 Estimating η_ϵ

Our regression specification using differences was:

$$\Delta\text{predictedOutcome}_i = \beta\Delta\text{actualOutcome}_i + \text{error}_i \tag{51}$$

This is equivalent to the specification:

$$\text{predictedOutcome}_{i,t} = \alpha_i + \beta\text{actualOutcome}_{i,t} + \text{error}_{i,t} \tag{52}$$

This fixed-effects regression extends unchanged to multiple time periods, so this will be our specification, and our estimate for η_ϵ remains $\hat{\beta}$. Standard errors for this estimate come directly from the regression.

5.2 Estimating η_μ

This is very similar to the case of two time periods, except now we use not the deltas but the centered variables, $\tilde{X}_{i,t}$.

Here too we can perform some algebra to isolate η_μ, where for ease of notation we replace predictedOutcome with predOut and actualOutcome with actualOut:

$$\tilde{\text{actualOut}}_{i,t} = \text{actualOut}_{i,t} - \overline{\text{actualOut}}_i = \gamma(\text{Treat}_{i,t} - \overline{\text{Treat}}_i) + \epsilon_{i,t} - \bar{\epsilon}_i = \gamma\tilde{\text{Treat}}_{i,t} + \tilde{\epsilon}_{i,t} \tag{53}$$

$$\tilde{\text{predOut}}_{i,t} = \text{predOut}_{i,t} - \overline{\text{predOut}}_i = \eta_T\gamma\tilde{\text{Treat}}_{i,t} + \eta_\epsilon\tilde{\epsilon}_{i,t} + \tilde{\nu}_{i,t} \tag{54}$$

We'll now work out the Variance of $\tilde{\epsilon}_{i,t}$. First, we'll rewrite $\tilde{\epsilon}_{i,t}$:

$$\tilde{\epsilon}_{i,t} = \epsilon_{i,t} - \frac{1}{T}\sum_{s=1}^{T}\epsilon_{i,s} = \frac{T-1}{T}\epsilon_{i,t} - \frac{1}{T}\sum_{s\neq t}\epsilon_{i,s} \tag{55}$$

But since the errors in different time periods are assumed to be uncorrelated and have equal variance (if these assumptions do not hold, it requires a separate estimate which exceeds the bounds of this paper, but can perhaps be done using time-series methods):

$$\text{Var}[\tilde{\epsilon}_{i,t}] = \text{Var}[\epsilon_{i,t}]\left(\left(\frac{T-1}{T}\right)^2 + \left(\frac{1}{T}\right)^2(T-1)\right) = \text{Var}[\epsilon_{i,t}]\frac{T-1}{T} \tag{56}$$

In a non-treated population:

$$\text{Cov}[\tilde{\text{predOut}}_{i,t}, \tilde{\text{actualOut}}_{i,t}] = \eta_\epsilon\text{Var}[\tilde{\epsilon}_{i,t}] = \eta_\epsilon\text{Var}[\epsilon_{i,t}]\frac{T-1}{T} \tag{57}$$

$$\text{Cov}[\text{predOut}_{i,t}, \text{actualOut}_{i,t}] = \eta_\mu\text{Var}[\mu_i] + \eta_\epsilon\text{Var}[\epsilon_{i,t}] \tag{58}$$

$$\text{Var}[\text{actualOut}_{i,t}] = \text{Var}[\mu_i] + \text{Var}[\epsilon_{i,t}] \tag{59}$$

$$\text{Var}[\tilde{\text{actualOut}}_{i,t}] = \text{Var}[\tilde{\epsilon}_{i,t}] = \text{Var}[\epsilon_{i,t}]\frac{T-1}{T} \tag{60}$$

$$\text{Cov}[\text{predOut}_{i,t}, \text{actualOut}_{i,t}] - \frac{T}{T-1}\text{Cov}[\tilde{\text{predOut}}_{i,t}, \tilde{\text{actualOut}}_{i,t}] = \eta_\mu \text{Var}[\mu_i] \tag{61}$$

$$\text{Var}[\text{actualOut}_{i,t}] - \frac{T}{T-1}\text{Var}[\tilde{\text{actualOut}}_{i,t}] = \text{Var}[\mu_i] \tag{62}$$

$$\Rightarrow \frac{\text{Cov}[\text{predOut}_{i,t}, \text{actualOut}_{i,t}] - \frac{T}{T-1}\text{Cov}[\tilde{\text{predOut}}_{i,t}, \tilde{\text{actualOut}}_{i,t}]}{\text{Var}[\text{actualOut}_{i,t}] - \frac{T}{T-1}\text{Var}[\tilde{\text{actualOut}}_{i,t}]} = \frac{\eta_\mu \text{Var}[\mu_i]}{\text{Var}[\mu_i]} = \eta_\mu \tag{63}$$

And so the left hand side of this equation is our estimate for η_μ. Standard errors can be estimated using bootstrap, similar to the two-period case.

5.3 Estimating η_T

As in the two-period case, estimating η_T requires experimental variation. We run two regressions:

$$\text{actualOutcome}_{i,t} = \alpha + \mu_i + \gamma \text{Treat}_{i,t} + \epsilon_{i,t} \tag{64}$$

This gives us the estimate $\hat{\gamma}_{\text{actual}} \to \gamma$. Running the same regression with predictedOutcome as the dependent variable gives us $\hat{\gamma}_{\text{predicted}} \to \eta_T \gamma$. Our estimate for η_T is then:

$$\hat{\eta}_T = \frac{\hat{\gamma}_{\text{predicted}}}{\hat{\gamma}_{\text{actual}}} \tag{65}$$

Standard errors can similarly be estimated using bootstrap. This approach is identical to the two-period case, as it does not depend on the number of time periods.

6 Conclusion

In this paper, I have introduced a framework for decomposing ML predictions into three components: between-unit, within-unit-across-time, and counterfactual-treatment-effect. The first two components can be separately estimated with non-experimental panel data. The third one cannot be estimated absent experimental variation. Often measuring model performance using experimental variation is either impossible (because it is done before the intervention took place) or infeasible (as it would require a much bigger sample than estimating the treatment effect using collected actual outcomes directly, and so defeats the purpose of a lower cost or a larger sample). This decomposition is therefore useful when such predictions are used as outcomes in causal analysis. The key findings and contributions are:

1. I show that overall prediction accuracy is a poor proxy for a model's ability to detect treatment effects, as models that fit well to between-unit variation may completely miss treatment effects, especially when between-unit variation dominates.

2. I propose a metric based on within-unit variation across time (η_ϵ) that better predicts a model's ability to capture treatment effects (η_T) under certain assumptions. I conjecture that these assumptions hold in many practical settings.

3. I demonstrate through simulations that under the assumption $\eta_T = \eta_\epsilon$, we can correct for bias in estimated treatment effects, producing unbiased estimates.

4. The approach provides a practical way to evaluate ML models for causal analysis without requiring experimental data for the entire population.

The implications for practitioners are significant. When using ML-predicted outcomes for causal inference, researchers should not rely solely on overall prediction accuracy but should specifically evaluate the model's ability to capture within-unit variation over time. This criterion could also guide researchers in decisions about model training, for example including or excluding various features. This approach requires panel data with at least two time periods but provides valuable insight into which model is likely to perform better for causal analysis.

Future research should empirically validate the relationship between η_T and η_ϵ across different domains and data types. Additionally, this work suggests that researchers might benefit from training models specifically to predict *changes* rather than *levels* when the ultimate goal is causal inference.

References

Emily Aiken, Suzanne Bellue, Joshua E. Blumenstock, Dean Karlan, and Christopher Udry. Estimating impact with surveys versus digital traces: Evidence from randomized cash transfers in togo. *Journal of Development Economics*, 175:103477, June 2025. ISSN 0304-3878. doi: 10.1016/j.jdeveco.2025.103477. URL `http://dx.doi.org/10.1016/j.jdeveco.2025.103477`.

Oscar Barriga-Cabanillas, Joshua E. Blumenstock, Travis J. Lybbert, and Daniel S. Putman. Probing the limits of mobile phone metadata for poverty prediction and impact evaluation. *Journal of Development Economics*, 174:103462, May 2025. ISSN 0304-3878. doi: 10.1016/j.jdeveco.2025.103462. URL `http://dx.doi.org/10.1016/j.jdeveco.2025.103462`.

Marshall Burke, Anne Driscoll, David B Lobell, and Stefano Ermon. Using satellite imagery to understand and promote sustainable development. *Science*, 371(6535):eabe8628, 2021.

Shawn Cole, Jessica Goldberg, Tomoko Harigaya, and Jessica Zhu. The impact of digital agricultural extension service: Experimental evidence from rice farmers in india. Working Paper, 2025.

David B Lobell, George Azzari, Marshall Burke, Sydney Gourlay, Zhenong Jin, Talip Kilic, and Siobhan Murray. Eyes in the sky, boots on the ground: Assessing satellite-and ground-based approaches to crop yield measurement and analysis. *American Journal of Agricultural Economics*, 102(1):202–219, 2020.

Nathan Ratledge, Gabe Cadamuro, Brandon de la Cuesta, Matthieu Stigler, and Marshall Burke. Using machine learning to assess the livelihood impact of electricity access. *Nature*, 611(7936):491–495, November 2022. ISSN 1476-4687. doi: 10.1038/s41586-022-05322-8. URL http://dx.doi.org/10.1038/s41586-022-05322-8.

Ofir Reich. Prediction decomposition for causal analysis, 2025. URL https://github.com/ofir-reich/prediction_decomposition_for_causal_analysis.