Food for Thought: Evidence from a Randomized Controlled Trial

October 13, 2024

In this note, we examine the results of a randomized placebo-controlled trial conducted on a sample of 2,500 patients, aimed at assessing the impact of a food additive on health outcomes. We perform causal discovery and causal inference using several frequentist estimators under different assumptions. Our analysis indicates that the point estimate of the population average treatment effect (PATE) is 6.333, with a standard error of 0.009. If the analyzed sample is representative of the entire population and we consider 6.333 to be an unbiased estimate, we can treat this value as the estimand of the PATE. In this case, the PATE has no variance, and its standard error becomes zero.

Table of contents

1 Description of the Experiment and the Data

The experiment[1] was conducted on a sample of $n = 2500$ patients, who were randomly selected from a population of unknown size, N. Of these patients, $n_1 = 507$ were assigned to receive a food additive ($D = 1$), while the remaining $n_0 = 1993$ received a placebo, placing them in the control group ($D = 1$). The outcome variable is a post-treatment health measure (Y), and another key variable is the pre-treatment number of steps recorded for each patient (X).

The results of the exploratory data analysis are summarized as follows:

1. **Censoring in Post-Treatment Health Outcome (Y)**

   The post-treatment health outcome (Y) is subject to both left and right censoring (Figure 1). This feature must be considered during the causal inference stage. In the presence of censoring, estimates of the average treatment effect (ATE) based on standard estimators like OLS may be biased toward zero, as their statistical properties depend on the assumption of symmetric noise distribution. To control for the magnitude of this bias, we compare ATE estimates from censored regressions (CR) to those obtained via OLS.

2. **Bias Due to Heteroscedasticity**

   Even after accounting for censoring, CR parameter estimates may still be biased if heteroscedasticity in the noise is ignored during estimation. In our RCT data, one potential source of heteroscedasticity is the inequality in the standard deviations of the outcome between the treatment and control groups (Table 2). According to the simulation experiments by Brown and Moffitt (1983), this could introduce a negative bias in ATE estimates if the assumption of homoscedasticity is incorrectly applied.

3. **Randomization Check**

   The distribution of the pre-treatment number of steps (X) suggests that the treatment assignment (D) is not determined by patient characteristics. We further verify this by comparing the sample means between the treatment and control groups (Table 3) and running a linear regression of D on X (Table 4). This confirms that the RCT protocol was correctly followed, and patients were randomly assigned to the treatment group. Therefore, omitting X during the estimation of ATE should not lead to omitted variable bias. To ensure robustness, we perform alternative estimates of ATE using exact

[1] The replication code is available on GitHub and Google Colab. Additional materials (figures and tables) can be found in the Appendix A of this note.

matching[2], where only those observations with identical values of X in both groups are retained.

4. **Interaction Between Pre- and Post-Treatment Variables**

   An examination of the relationship between the post-treatment outcome (Y) and the pre-treatment covariate (X) reveals that the treatment tends to alter the slope of the dependence between the two variables (Figure 3). Specifically, a negative association before treatment becomes a positive correlation after treatment. This suggests that the treatment's impact on the health outcome depends on the patient's initial characteristics. As a result, we control for both X and an interaction term ($X * D$), because including these terms always improves the precision of ATE estimates (Lin 2013).

5. **Non-linearity in the Relationship Between X and Y**

   The relationship between X and Y appears deterministic, indicating that a model-free ATE estimator (e.g., "difference-in-means") would yield similar results to more sophisticated estimators, such as "Double Machine Learning", which can model the relationship between X and Y using advanced prediction methods like random forests or neural networks. While partial[3] disregard for this nonlinearity may not introduce bias in ATE estimates (since D is independent of X), it could reduce the estimator's efficiency. When choosing between unbiasedness and efficiency, we prioritize unbiasedness[4]. Ignoring nonlinearity might also introduce heteroscedasticity in the error term, so we apply Eicker-Huber-White robust standard errors wherever possible to ensure valid statistical inference.

6. **Causal Discovery Using the Peter-Clark Algorithm**

   To uncover causal relationships between variables, we apply the Peter-Clark algorithm, which assumes no unobserved latent confounders in the observational data. By constructing a directed acyclic graph using this method (Figure 4), we find that Y is independently influenced by both D and X, but not the other way around. This is crucial for identifying the ATE, as reverse causality would be a major roadblock for the study that required an instrumental variable to estimate the causal impact of an independent variable on the outcome.

[2] Of course, this results in dropping observations that cannot be matched with treated units. In our case, if strict matching is applied, the sample size will decrease from n to $2n_1$, which in turn reduces the precision of the ATE estimates compared to the scenario without matching.

[3] In our regressions, we include polynomials up to the fourth order.

[4] A first-best solution would involve combining Double ML with censored regression estimation, and accounting for conditional heteroscedasticity of the noise during the estimation procedure. However, we rely on suboptimal approaches in this analysis and leave the optimal method for future work.

2 Estimation of the Population Average Treatment Effect and Its Precision

We estimate the causal effect of the food additive using the average treatment effect (ATE):

$$\text{ATE} = \mathbb{E}[Y_1 - Y_0],$$

where Y_1 represents the patients' outcome under treatment, and Y_0 is the outcome under control. Due to the fundamental problem of causal inference, we can observe only one of these outcomes for each individual, which means we cannot calculate the individual treatment effect for a patient ($\text{TE} = Y_1 - Y_0$). Alternatively, we can estimate the association between treated and untreated patients, known as the average predictive effect (APE):

$$\text{APE} = \mathbb{E}[Y_1|D = 1] - \mathbb{E}[Y_0|D = 0],$$

and interpret it as an ATE estimate because, as we have seen earlier, treated and untreated patients differ only in terms of the treatment itself and not in their initial characteristics (X). This implies that there is no selection bias ($\mathbb{E}[Y_0|D = 1] = \mathbb{E}[Y_0|D = 0]$), which could potentially contaminate the APE, making it differ from the ATE.

Our empirical strategy for obtaining an estimate of the population average treatment effect (PATE) is as follows. First, we leverage the fact that the distribution of Y is susceptible to censoring. As a result, using sample averages for a difference-in-means estimator (and, equivalently, an OLS estimator) will lead to downward bias, as these methods assume a symmetric i.i.d. distribution of outcomes. Second, we assume that censoring arises from exogenous factors unrelated to the patient's treatment status (i.e., no sample selection). To take both of these circumstances into account, we employ an old-fashioned (Type I) censored regression model originally proposed by Tobin (1958).

In this model, the post-treatment health outcome $Y_{n\times 1}$ is subject to bottom (a) and top (b) censoring

$$Y = \begin{cases} a & y^* \leq a \\ y^* & a < y^* < b \\ b & y^* \geq b \end{cases}$$

while a latent variable y^* is influenced by the treatment dummy D and centered covariates W (such as $\text{E}(W) = 0$)

$$
\begin{aligned}
y^* &= \vec{1}\alpha + D\beta + W\gamma + \varepsilon \\
\mathrm{E}(y^*|a < y^* < b|D, W) &= \vec{1}\alpha + D\beta + W\gamma,
\end{aligned}
\tag{1}
$$

where $W = (X, XD, X^2, X^2D, X^3, X^3D)_{n\times 2}$ and $\varepsilon \rightarrow$ i.i.d.$\mathcal{N}(0, \sigma^2 I_{n\times n})$ in large samples due to the Central Limit Theorem (CLT). To obtain sample estimates of the parameters, we employ a Maximum Likelihood estimator, which provides a consistent estimate of β according to Amemiya (1973).

To mitigate a potential downward bias in ATE estimates due to heteroscedasticity (Brown and Moffitt 1983), we estimate an alternative variant of the model, relaxing the homoscedasticity assumption and allowing the standard deviation of the noise to depend on observable regressors:

$$
\log(\sigma_i) = z_i'\delta + \epsilon_i, \tag{2}
$$

where $z_i' = (1, d_i, x_i, x_i^2, x_i^3, x_i^4)$ and ϵ_i is an i.i.d. innovation with zero mean and unit variance.

In addition to censored regressions, we also perform a simple linear regression based on the specification given in Equation 1. This helps us measure the degree of downward bias that occurs when using the OLS estimator.

As discussed earlier, the population[5] average treatment effect estimate (PATE) in the models is derived as follows:

$$
\widehat{\mathrm{PATE}} = \widehat{\mathrm{SATE}} = \mathbb{E}[\widehat{Y|D=1}] - \mathbb{E}[\widehat{Y|D=0}]
$$

For a linear regression, this difference in conditional expectations corresponds to $\hat{\beta}$, while in censored regression, the effect[6] depends on the distribution function:

$$
\begin{aligned}
\mathbb{E}[\widehat{Y|D=1}] - \mathbb{E}[\widehat{Y|D=0}] = \hat{\beta}\Bigg[&\Phi\left(\frac{b - \vec{1}\hat{\alpha} - W\hat{\gamma}}{\hat{\sigma}}\right) - \Phi\left(\frac{a - \vec{1}\hat{\alpha} - W\hat{\gamma}}{\hat{\sigma}}\right)\Bigg] \\
&+ \left(\vec{1}\hat{\alpha} + W\hat{\gamma} - a + \hat{\beta}\right)\Phi\left(\frac{a - \vec{1}\hat{\alpha} - W\hat{\gamma}}{\hat{\sigma}}\frac{\hat{\beta}}{\hat{\sigma}}\right) \\
&- \left(\vec{1}\hat{\alpha} + W\hat{\gamma} - b + \hat{\beta}\right)\Phi\left(\frac{b - \vec{1}\hat{\alpha} - W\hat{\gamma}}{\hat{\sigma}}\frac{\hat{\beta}}{\hat{\sigma}}\right)
\end{aligned}
\tag{3}
$$

[5] We assume that, due to the RCT protocol, the estimate of the sample average treatment effect ($\widehat{\mathrm{SATE}}$) is an unbiased estimator of the true SATE. Additionally, we consider SATE to be an unbiased estimator of the population average treatment effect PATE, provided the sample is representative of the population.

[6] The complete derivation of this estimate is provided in Appendix B of this note.

For the precision of these estimates, we calculate the standard error of $\widehat{\text{PATE}}$. In censored regressions, we use the Delta method, while for OLS estimates, we apply Eicker-Huber-White robust standard errors. To capture the full picture of the estimates' uncertainty, we construct 99% confidence intervals using the estimated standard errors.

Now, let's review the point estimates of the PATE (Table 1).

The results are presented in two panels. Panel A utilizes the entire sample of 2,500 patients, while Panel B ("Exact Matching") uses a subsample where each treated patient is matched with a "twin" based on their pre-treatment characteristics.

The first two columns show estimates obtained from linear regression models: one that includes pre-treatment covariates ("OLS w/ X") and one that does not ("OLS"). In columns 3 and 4, we see results from censored regressions ("CR" and "CR w/ X"), which employ a maximum likelihood (ML) estimator of the $\widehat{\text{PATE}}$ under the assumption of homoscedasticity. The final column provides estimates from a censored regression that controls for conditional heteroscedasticity, using the model specified in Equation 2 ("CR w/ het. err.").

The main findings can be summarized as follows:

1. As expected, OLS estimates of PATE are lower than those from censored regressions. The downward bias is more pronounced when controls are excluded from the model (5.936/5.941 in the pairwise linear regression vs. 5.997/5.998 in the censored regression).
2. Surprisingly, adding controls to the models slightly reduces the PATE estimates. This effect is more noticeable in censored regressions (5.997/5.998 vs. 5.947/5.950), possibly due to accounting for the non-linearity in the conditional average treatment effect.
3. Including interaction terms improves the precision of the estimates, consistent with theoretical predictions. This effect, seen in both linear and censored regression models, leads to a decrease in standard errors.
4. The most substantial bias occurs when conditional heteroscedasticity is ignored in censored regression models. Relaxing this assumption leads to a sharp increase in the PATE point estimate, from 5.997/5.947 to 6.333 (and from 5.998/5.950 to 6.315 with Exact Matching).

Based on these results, we conclude that the causal effect of the food additive on health outcomes is 6.333, with a standard error of 0.009. If we assume that (1) the RCT protocol is correct, and (2) our sample is representative of the population, then this estimate serves as an unbiased estimate of the population average treatment effect (PATE). We can infer that, on average, administering the food additive to the entire population would increase the health indicator by 6.333. The 99% confidence interval for the causal effect ranges from 6.316 to 6.351.

Table 1: Estimates of the Population Average Treatment Effect

	OLS	OLS w/ X	CR	CR w/ X	CR w/ het. err.
Panel A: No Matching					
PATE	5.936***	5.931***	5.997***	5.947***	6.333***
	s.e. = 0.022	s.e. = 0.003	s.e. = 0.023	s.e. = 0.004	s.e. = 0.009
	[5.893, 5.978]	[5.925, 5.938]	[5.951, 6.043]	[5.939, 5.956]	[6.316, 6.351]
Num.Obs.	2500	2500	2500	2500	2500
RMSE	0.65	0.14	0.66	0.14	0.82
Panel B: Exact Matching					
PATE	5.941***	5.941***	5.998***	5.950***	6.315***
	s.e. = 0.036	s.e. = 0.005	s.e. = 0.037	s.e. = 0.005	s.e. = 0.017
	[5.871, 6.010]	[5.931, 5.950]	[5.925, 6.070]	[5.940, 5.961]	[6.282, 6.349]
Num.Obs.	1014	1014	1014	1014	1014
RMSE	0.56	0.08	0.57	0.08	0.70

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001

3 Treating the Sample as a Population

If we assume that the collected sample represents the entire population, this simplifies the task of estimating the PATE, as there is no need to make assumptions about the sample's representativeness. In this case, any unbiased estimator will directly provide the estimand of the PATE.

As shown in Table 1, the most refined estimate of the PATE is 6.333. If we accept this as an unbiased estimate of the PATE, we conclude that the estimand of the PATE is 6.333. Additionally, since this "true" PATE represents the entire population, it has no variance, meaning the precision (i.e., the standard error) of this value would be zero.

4 Conclusion

In this note, we examine the results of a randomized placebo-controlled trial conducted on a sample of 2,500 patients, aimed at assessing the impact of a food additive on health outcomes. We perform causal discovery and causal inference using several frequentist estimators under different assumptions.

Our analysis indicates that the point estimate of the population average treatment effect (PATE) is 6.333, with a standard error of 0.009.

If the analyzed sample is representative of the entire population and we consider 6.333 to be an unbiased estimate, we can treat this value as the estimand of the PATE. In this case, the PATE has no variance, and its standard error becomes zero.

References

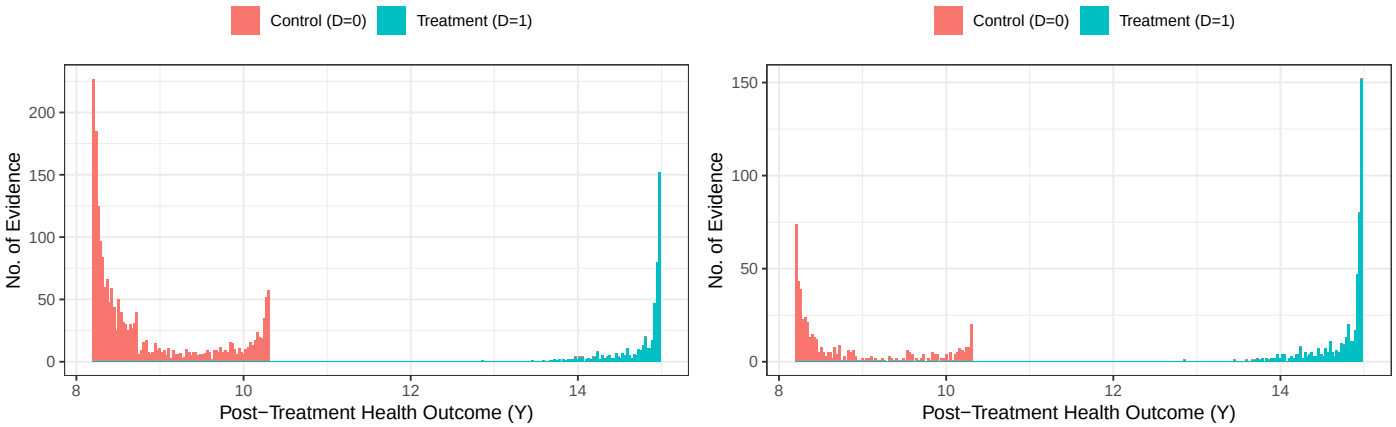
Amemiya, Takeshi. 1973. "Regression Analysis When the Dependent Variable Is Truncated Normal." *Econometrica* 41 (6): 997–1016. http://www.jstor.org/stable/1914031.

Brown, Charles, and Robert Moffitt. 1983. "The Effect of Ignoring Heteroscedasticity on Estimates of the Tobit Model." *NBER Working Papers*. https://doi.org/10.3386/t0027.

Lin, Winston. 2013. "Agnostic notes on regression adjustments to experimental data: Re-examining Freedman's critique." *The Annals of Applied Statistics* 7 (1): 295–318. https://doi.org/10.1214/12-AOAS583.

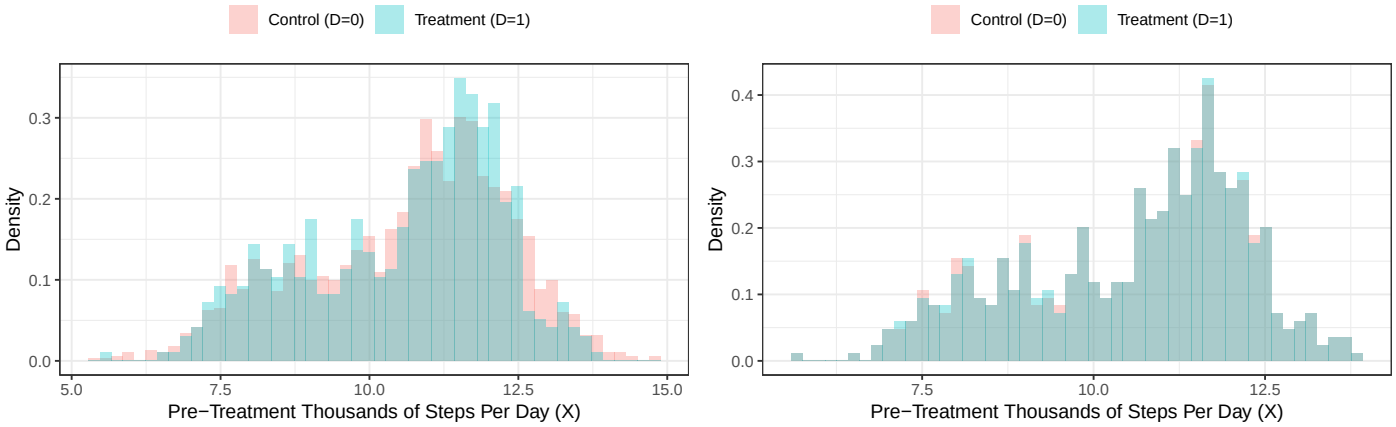
Tobin, James. 1958. "Estimation of Relationships for Limited Dependent Variables." *Econometrica* 26 (1): 24–36. http://www.jstor.org/stable/1907382.

Appendix A



(a) No Matching (b) Exact Matching

Figure 1: Distribution of Post-Treatment Health Outcome (Y)



(a) No Matching (b) Exact Matching

Figure 2: Distribution of Pre-Treatment Thousands of Steps Per Day (X)

Table 2: Descriptive Statistics of Post-Treatment Health Outcome (Y)

(a) No Matching

D	n	Mean	Std
0	1993	8.82	0.71
1	507	14.75	0.33

(b) Exact Matching

D	n	Mean	Std
0	507	8.81	0.73
1	507	14.75	0.33

Table 3: Descriptive Statistics of Pre-Treatment Thousands of Steps Per Day (X)

(a) No Matching

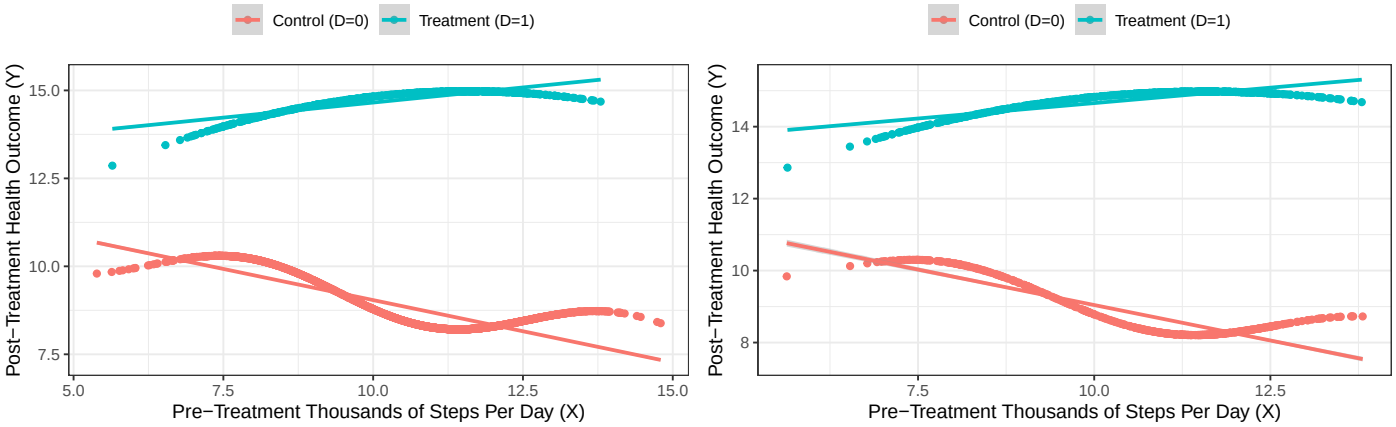
D	n	Mean	Std
0	1993	10.63	1.71
1	507	10.58	1.64

(b) Exact Matching

D	n	Mean	Std
0	507	10.58	1.64
1	507	10.58	1.64

Table 4: Results from a Pairwise Linear Regression of Treatment Status (D) on Pre-Treatment Patients' Characteristics (X)

	Panel A: No Matching	Panel B: Exact Matching
	OLS	OLS
(Intercept)	0.230	0.501
	[0.102, 0.359]	[0.235, 0.767]
	s.e. = 0.050	s.e. = 0.103
	t = 4.617	t = 4.854
	p = <0.001	p = <0.001
No. of steps (X)	-0.003	0.000
	[-0.015, 0.009]	[-0.025, 0.025]
	s.e. = 0.005	s.e. = 0.010
	t = -0.561	t = -0.008
	p = 0.575	p = 0.994
Num.Obs.	2500	1014
R2 Adj.	0.000	-0.001
F	0.315	0.000
Std.Errors	HC3	HC3



(a) No Matching

(b) Exact Matching

Figure 3: Scatter Plot of Pre-Treatment Thousands of Steps Per Day (X) and Post-Treatment Health Outcome (Y)

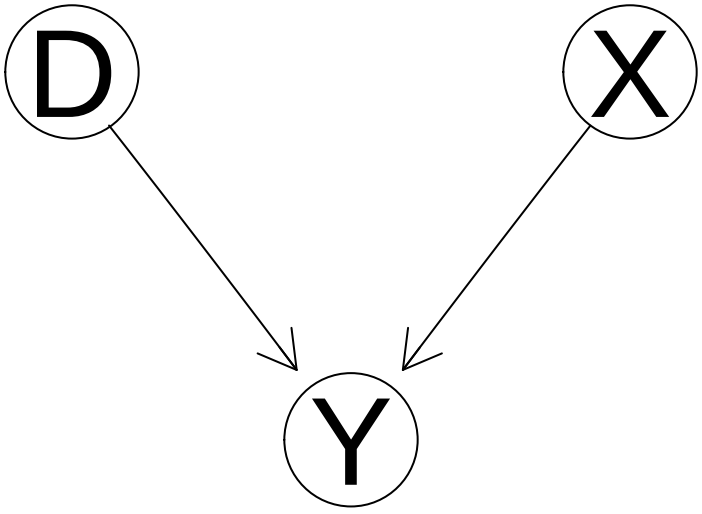


Figure 4: Directed Acyclic Graph Based on Peter-Clark Algorithm

Appendix B

Firstly, we derive a conditional expectation for the outcome Y:

$$\begin{aligned}\mathbb{E}[Y|D,W] = a\mathrm{P}(y^* \leq a|D,W) \\ + P(a < y^* < b|D,W)\mathrm{E}(y^*|a < y^* < b|D,W) \\ + b\mathrm{P}(y^* \geq b|D,W)\end{aligned}$$

According to Equation 1,

$$\begin{aligned}\mathbb{E}[Y|D,W] = a\mathrm{P}(\varepsilon \leq a - \vec{1}\alpha - D\beta - W\gamma) \\ + P(a < \vec{1}\alpha + D\beta + W\gamma + \varepsilon < b|D,W)\mathrm{E}(y^*|a < y^* < b|D,W) \\ + b\mathrm{P}(\varepsilon \geq b - \vec{1}\alpha - D\beta - W\gamma|D,W)\end{aligned}$$

and because of $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_{n\times n})$,

$$\begin{aligned}\mathbb{E}[Y|D,W] = a\Phi\left(\frac{a - \vec{1}\alpha + D\beta + W\gamma}{\sigma}\right) \\ + \mathrm{E}(y^*|a < y^* < b|D,W)\left[\Phi\left(\frac{b - \vec{1}\alpha + D\beta + W\gamma}{\sigma}\right) - \Phi\left(\frac{a - \vec{1}\alpha + D\beta + W\gamma}{\sigma}\right)\right] \\ + b\left[1 - \Phi\left(\frac{b - \vec{1}\alpha + D\beta + W\gamma}{\sigma}\right)\right]\end{aligned}$$

We can rewrite the last equation like this:

$$\begin{aligned}\mathbb{E}[Y|D,W] = \Phi\left(\frac{b - \vec{1}\alpha - D\beta - W\gamma}{\sigma}\right)[\mathrm{E}(y^*|a < y^* < b|D,W) - b] \\ - \Phi\left(\frac{a - \vec{1}\alpha - D\beta - W\gamma}{\sigma}\right)[\mathrm{E}(y^*|a < y^* < b|D,W) - a] \\ + b\end{aligned}$$

Secondly, we calculate the expectation outcome Y for the treated (using Equation 1)

$$\begin{aligned}\widehat{\mathbb{E}[Y|D=1]} = &\Phi\left(\frac{b-\vec{1}\hat{\alpha}-\hat{\beta}-W\hat{\gamma}}{\hat{\sigma}}\right)\left(\vec{1}\hat{\alpha}+\hat{\beta}+W\hat{\gamma}-b\right)\\ &-\Phi\left(\frac{a-\vec{1}\hat{\alpha}-\hat{\beta}-W\hat{\gamma}}{\hat{\sigma}}\right)\left(\vec{1}\hat{\alpha}+\hat{\beta}+W\hat{\gamma}-a\right)\\ &+b\end{aligned} \tag{4}$$

and for the untreated

$$\begin{aligned}\widehat{\mathbb{E}[Y|D=0]} = &\Phi\left(\frac{b-\vec{1}\hat{\alpha}-W\hat{\gamma}}{\hat{\sigma}}\right)\left(\vec{1}\hat{\alpha}+W\hat{\gamma}-b\right)\\ &-\Phi\left(\frac{a-\vec{1}\hat{\alpha}-W\hat{\gamma}}{\hat{\sigma}}\right)\left(\vec{1}\hat{\alpha}+W\hat{\gamma}-a\right)\\ &+b\end{aligned}$$

Thirdly, we re-express the Equation 4:

$$\begin{aligned}\widehat{\mathbb{E}[Y|D=1]} = &\Phi\left(\frac{b-\vec{1}\hat{\alpha}-W\hat{\gamma}}{\hat{\sigma}}-\frac{\hat{\beta}}{\hat{\sigma}}\right)\left(\vec{1}\hat{\alpha}+W\hat{\gamma}-b\right)\\ &+\Phi\left(\frac{b-\vec{1}\hat{\alpha}-W\hat{\gamma}}{\hat{\sigma}}-\frac{\hat{\beta}}{\hat{\sigma}}\right)\hat{\beta}\\ &-\Phi\left(\frac{a-\vec{1}\hat{\alpha}-W\hat{\gamma}}{\hat{\sigma}}-\frac{\hat{\beta}}{\hat{\sigma}}\right)\left(\vec{1}\hat{\alpha}+W\hat{\gamma}-a\right)\\ &-\Phi\left(\frac{a-\vec{1}\hat{\alpha}-W\hat{\gamma}}{\hat{\sigma}}-\frac{\hat{\beta}}{\hat{\sigma}}\right)\hat{\beta}\\ &+b\end{aligned}$$

$$
\begin{aligned}
\widehat{\mathbb{E}[Y|D=1]} = &\left[\Phi\left(\frac{b-\vec{1}\hat{\alpha}-W\hat{\gamma}}{\hat{\sigma}}\right)-\Phi\left(\frac{b-\vec{1}\hat{\alpha}-W\hat{\gamma}}{\hat{\sigma}}\frac{\hat{\beta}}{\hat{\sigma}}\right)\right]\left(\vec{1}\hat{\alpha}+W\hat{\gamma}-b\right)\\
&+\left[\Phi\left(\frac{b-\vec{1}\hat{\alpha}-W\hat{\gamma}}{\hat{\sigma}}\right)-\Phi\left(\frac{b-\vec{1}\hat{\alpha}-W\hat{\gamma}}{\hat{\sigma}}\frac{\hat{\beta}}{\hat{\sigma}}\right)\right]\hat{\beta}\\
&-\left[\Phi\left(\frac{a-\vec{1}\hat{\alpha}-W\hat{\gamma}}{\hat{\sigma}}\right)-\Phi\left(\frac{a-\vec{1}\hat{\alpha}-W\hat{\gamma}}{\hat{\sigma}}\frac{\hat{\beta}}{\hat{\sigma}}\right)\right]\left(\vec{1}\hat{\alpha}+W\hat{\gamma}-a\right)\\
&-\left[\Phi\left(\frac{a-\vec{1}\hat{\alpha}-W\hat{\gamma}}{\hat{\sigma}}\right)-\Phi\left(\frac{a-\vec{1}\hat{\alpha}-W\hat{\gamma}}{\hat{\sigma}}\frac{\hat{\beta}}{\hat{\sigma}}\right)\right]\hat{\beta}\\
&+b
\end{aligned}
$$

$$
\begin{aligned}
\widehat{\mathbb{E}[Y|D=1]} = &\Phi\left(\frac{b-\vec{1}\hat{\alpha}-W\hat{\gamma}}{\hat{\sigma}}\right)\left(\vec{1}\hat{\alpha}+W\hat{\gamma}-b\right)-\Phi\left(\frac{b-\vec{1}\hat{\alpha}-W\hat{\gamma}}{\hat{\sigma}}\frac{\hat{\beta}}{\hat{\sigma}}\right)\left(\vec{1}\hat{\alpha}+W\hat{\gamma}-b\right)\\
&+\left[\Phi\left(\frac{b-\vec{1}\hat{\alpha}-W\hat{\gamma}}{\hat{\sigma}}\right)-\Phi\left(\frac{b-\vec{1}\hat{\alpha}-W\hat{\gamma}}{\hat{\sigma}}\frac{\hat{\beta}}{\hat{\sigma}}\right)\right]\hat{\beta}\\
&-\Phi\left(\frac{a-\vec{1}\hat{\alpha}-W\hat{\gamma}}{\hat{\sigma}}\right)\left(\vec{1}\hat{\alpha}+W\hat{\gamma}-a\right)+\Phi\left(\frac{a-\vec{1}\hat{\alpha}-W\hat{\gamma}}{\hat{\sigma}}\frac{\hat{\beta}}{\hat{\sigma}}\right)\left(\vec{1}\hat{\alpha}+W\hat{\gamma}-a\right)\\
&-\left[\Phi\left(\frac{a-\vec{1}\hat{\alpha}-W\hat{\gamma}}{\hat{\sigma}}\right)-\Phi\left(\frac{a-\vec{1}\hat{\alpha}-W\hat{\gamma}}{\hat{\sigma}}\frac{\hat{\beta}}{\hat{\sigma}}\right)\right]\hat{\beta}\\
&+b
\end{aligned}
$$

Then, we express $\widehat{\text{PATE}}$ as the difference between $\widehat{\mathbb{E}[Y|D=1]}$ and $\widehat{\mathbb{E}[Y|D=0]}$:

$$
\begin{aligned}
\widehat{\mathbb{E}[Y|D=1]} = &\Phi\left(\frac{b-\vec{1}\hat{\alpha}-W\hat{\gamma}}{\hat{\sigma}}\right)\left(\vec{1}\hat{\alpha}+W\hat{\gamma}-b\right)-\Phi\left(\frac{b-\vec{1}\hat{\alpha}-W\hat{\gamma}}{\hat{\sigma}}\frac{\hat{\beta}}{\hat{\sigma}}\right)\left(\vec{1}\hat{\alpha}+W\hat{\gamma}-b\right)\\
&+\left[\Phi\left(\frac{b-\vec{1}\hat{\alpha}-W\hat{\gamma}}{\hat{\sigma}}\right)-\Phi\left(\frac{b-\vec{1}\hat{\alpha}-W\hat{\gamma}}{\hat{\sigma}}\frac{\hat{\beta}}{\hat{\sigma}}\right)\right]\hat{\beta}\\
&-\Phi\left(\frac{a-\vec{1}\hat{\alpha}-W\hat{\gamma}}{\hat{\sigma}}\right)\left(\vec{1}\hat{\alpha}+W\hat{\gamma}-a\right)+\Phi\left(\frac{a-\vec{1}\hat{\alpha}-W\hat{\gamma}}{\hat{\sigma}}\frac{\hat{\beta}}{\hat{\sigma}}\right)\left(\vec{1}\hat{\alpha}+W\hat{\gamma}-a\right)\\
&-\left[\Phi\left(\frac{a-\vec{1}\hat{\alpha}-W\hat{\gamma}}{\hat{\sigma}}\right)-\Phi\left(\frac{a-\vec{1}\hat{\alpha}-W\hat{\gamma}}{\hat{\sigma}}\frac{\hat{\beta}}{\hat{\sigma}}\right)\right]\hat{\beta}\\
&+b
\end{aligned}
$$

$$
\begin{aligned}
\widehat{\mathbb{E}[Y|D=0]} = &\Phi\left(\frac{b-\vec{1}\hat{\alpha}-W\hat{\gamma}}{\hat{\sigma}}\right)\left(\vec{1}\hat{\alpha}+W\hat{\gamma}-b\right)\\
&-\Phi\left(\frac{a-\vec{1}\hat{\alpha}-W\hat{\gamma}}{\hat{\sigma}}\right)\left(\vec{1}\hat{\alpha}+W\hat{\gamma}-a\right)\\
&+b
\end{aligned}
$$

$$
\begin{aligned}
\widehat{\mathrm{PATE}} = &-\Phi\left(\frac{b-\vec{1}\hat{\alpha}-W\hat{\gamma}}{\hat{\sigma}}\frac{\hat{\beta}}{\hat{\sigma}}\right)\left(\vec{1}\hat{\alpha}+W\hat{\gamma}-b\right)\\
&+\Phi\left(\frac{b-\vec{1}\hat{\alpha}-W\hat{\gamma}}{\hat{\sigma}}\right)\hat{\beta}-\Phi\left(\frac{b-\vec{1}\hat{\alpha}-W\hat{\gamma}}{\hat{\sigma}}\frac{\hat{\beta}}{\hat{\sigma}}\right)\hat{\beta}\\
&+\Phi\left(\frac{a-\vec{1}\hat{\alpha}-W\hat{\gamma}}{\hat{\sigma}}\frac{\hat{\beta}}{\hat{\sigma}}\right)\left(\vec{1}\hat{\alpha}+W\hat{\gamma}-a\right)\\
&-\Phi\left(\frac{a-\vec{1}\hat{\alpha}-W\hat{\gamma}}{\hat{\sigma}}\right)\hat{\beta}+\Phi\left(\frac{a-\vec{1}\hat{\alpha}-W\hat{\gamma}}{\hat{\sigma}}\frac{\hat{\beta}}{\hat{\sigma}}\right)\hat{\beta}
\end{aligned}
$$

And finally, we get the formula:

$$
\begin{aligned}
\widehat{\mathrm{PATE}} = &\hat{\beta}\left[\Phi\left(\frac{b-\vec{1}\hat{\alpha}-W\hat{\gamma}}{\hat{\sigma}}\right)-\Phi\left(\frac{a-\vec{1}\hat{\alpha}-W\hat{\gamma}}{\hat{\sigma}}\right)\right]\\
&+\left(\vec{1}\hat{\alpha}+W\hat{\gamma}-a+\hat{\beta}\right)\Phi\left(\frac{a-\vec{1}\hat{\alpha}-W\hat{\gamma}}{\hat{\sigma}}\frac{\hat{\beta}}{\hat{\sigma}}\right)\\
&-\left(\vec{1}\hat{\alpha}+W\hat{\gamma}-b+\hat{\beta}\right)\Phi\left(\frac{b-\vec{1}\hat{\alpha}-W\hat{\gamma}}{\hat{\sigma}}\frac{\hat{\beta}}{\hat{\sigma}}\right)
\end{aligned}
$$