Barcelona School of Economics

Master´s Degree in Economics and Finance
PhD Track

"Estimating Causal Effects in the Absence of Treatment Observability"

Author: Joon Sup Park
Supervisor: Davide Debortoli

Date: July 22, 2023

ABSTRACT IN ENGLISH

In this paper, I propose a novel method to estimate causal effects when the treatments are not observable. Treatment observability is not affordable in many instances of noncompliance settings, where the treatments actually taken by the subjects may be different from the treatments they were assigned to. The standard procedure is to adhere to the intention-to-treat (ITT) principle and simply settle at estimating causal effects on the level of assigned treatments. In contrast, I propose a statistical algorithm that recovers the actual treatments from the outcome and pretreatment variables by iterative applications of regression and clustering algorithm, and use it to estimate causal effects on the level of actual treatments.

ABSTRACT IN CATALAN/ SPANISH

En este artículo, propongo un método novedoso para estimar los efectos causales cuando los tratamientos no son observables. La observabilidad del tratamiento no es asequible en muchos casos de entornos de incumplimiento, donde los tratamientos que realmente toman los sujetos pueden ser diferentes de los tratamientos a los que fueron asignados. El procedimiento estándar es adherirse al principio de intención de tratar (ITT) y simplemente conformarse con estimar los efectos causales en el nivel de los tratamientos asignados. Por el contrario, propongo un algoritmo estadístico que recupera los tratamientos reales de las variables de resultado y pretratamiento mediante aplicaciones iterativas de regresión y algoritmo de agrupamiento, y lo utiliza para estimar los efectos causales en el nivel de los tratamientos reales.

KEYWORDS IN ENGLISH (3)**:** Causal Inference, Noncompliance, Machine Learning

KEYWORDS IN CATALAN/ SPANISH (3)**:** Inferencia Causal, Incumplimiento, Machine Learning

MASTER PROJECT

Estimation of Causal Effects in the Absence of Treatment Observability

Author: Joon Sup Park

Program: Ph.D. Track

Advisor: Valeria Gargiulo

Academic Year: 2022-2023

Contents

1. Introduction **3**

2. Methodology **4**

2.1. Assumptions 5

2.2. Method 5

3. Results **7**

4. Discussions **9**

Appendix **11**

A. References 11

B. Figures 12

Abstract

In this paper, I develop a method to estimate causal effects when the treatment variable is unobserved. Treatment unobservability is a common problem in noncompliance settings, where the actual treatment individuals take may be different from the treatment they were assigned to. The standard procedure is to adhere to the intention-to-treat (ITT) principle and simply settle at estimating the causal effect on the level of assigned treatments. However, such estimates are misplaced if the object of interest is not the causal effect of the intervention itself. This is especially the case if there is a systematic way that individuals with different characteristics defy or comply with the treatment assignment. Given fairly standard assumptions, I propose a method that directly estimates the causal effect on the level of actual treatments by recovering the actual treatments from the information provided in the outcome and confounders variables. In the process, I present a heuristic test that tells us if the method delivered a reliable estimate of the actual treatments and the average treatment effect (ATE). Then I conduct a simulation study to assess the performance and sensitivity of the method over different parameter values. Limitations of the method will be discussed with considerations of further research questions.

1. Introduction

Causal inference is about estimating the counterfactual effect of treatment on outcome. Here, the term "*counterfactual*" is what distinguishes it from usual statistical inference. Ideally, we want to compare the outcomes of the *same unit* under different treatments, but every unit receives only one treatment at a time with only one outcome as a result, making within-unit comparison difficult. Thus, the problem of missing data lies at the heart of causal inference, and how we deal with it through different imputation techniques defines its methods. Despite the difference, they all aim at a reasonable comparison between the treated and untreated outcomes by pooling the information across different units while

controlling for their pre-treatment characteristics.

It is, then, not a surprise that most methods in causal inference presume the observability of treatment variable. The task already involves imputation, and further missingness in data could be critical to conduct causal inference. More often than not, however, we cannot afford to observe the treatment each unit receives. Take the case of non-compliance, where the actual treatment a unit receives may be different from the treatment it is assigned. Health agency might prescribe a medicine to patients but they may covertly not take it, worrying that it may have some unknown side effects. Thus, treatment unobservability may be far from limited to pathological cases and it could be a real issue in many problems we want to address.

This motivates the goal of this paper. In this paper, I propose a method to systematically recuperate the treatment variable from the information provided by the outcome variable and the pre-treatment variables. Section 2 will elucidate the method and present a heuristic that tells us if this method is applicable. Simulation results will be presented in Section 3 that enable us to assess the performance and sensitivity of the method. Section 4 discusses the advantage and shortcomings of the method, future research questions, and comparison with two existing methods also adequate for recovering the unobserved treatment variable.

2. Methodology

Broadly speaking, there are two approaches in causal inference: 1)a parametric approach based on outcome modeling and 2)a non-parametric approach based on propensity score. Outcome modeling assumes that we have good prior knowledge in the functional form how outcome depends on treatment and pre-treatment characteristics, i.e., in $Y_i = f(X_i, Z_i) + \epsilon_i$. Propensity score, on the other hand, is a probability that a unit i receives treatment conditional on its pre-treatment characteristics, i.e., $\Pr(Z_i = 1 \mid X_i)$. In both cases, conducting causal inference critically relies on the knowledge of treatment variable.

The method I propose is based on outcome modeling. It leverages the functional form $Y_i = f(X_i, Z_i) + \epsilon_i$ as well as the observability of outcome Y_i's and pre-treatment

characteristics X_i's to recover our knowledge of treatment Z_i's.

2.1. Assumptions

Let us assume the following outcome model as data generating process: for every unit $i \in \{1, \cdots, n\}$,

$$Y_i = f(X_i, Z_i) + \epsilon_i = X_i^T \beta + Z_i \delta + \epsilon_i, \qquad \epsilon_i \sim_{iid} N(0, \sigma_y^2)$$

where treatment $Z_i \in \{0,1\}$ is unobserved. For identifiability, we make assumptions that the ATE δ is positive and that we have knowledge of the functional form $f(X_i, Z_i)$. Also, we retain the two standard assumptions in causal inference:

$$\text{Assumption I (Ignorability): } \{Y_i(0), Y_i(1)\} \perp Z_i \mid X_i, \ \forall i$$

$$\text{Assumption II (Overlap): } 0 < \Pr(Z_i \mid X_i) < 1, \ \forall i$$

Importantly, the ignorability assumption implies that there is no unmeasured confounders: all variables that affect the potential outcomes $Y_i(0)$ and $Y_i(1)$ are summarized in the treatment Z_i and the measured confounders X_i. The overlap assumption, on the other hand, implies that every unit i has a positive probability of receiving treatment $Z_i = 1$ and no-treatment $Z_i = 0$.

2.2. Method

Now we present the method. For traceability, first rewrite the outcome model in matrix form:

$$Y = X\beta + Z\delta + \epsilon, \qquad \epsilon \sim N(0, \sigma_y^2 I_n)$$

Step 1: Regress Y on X alone to obtain our first estimate of β,

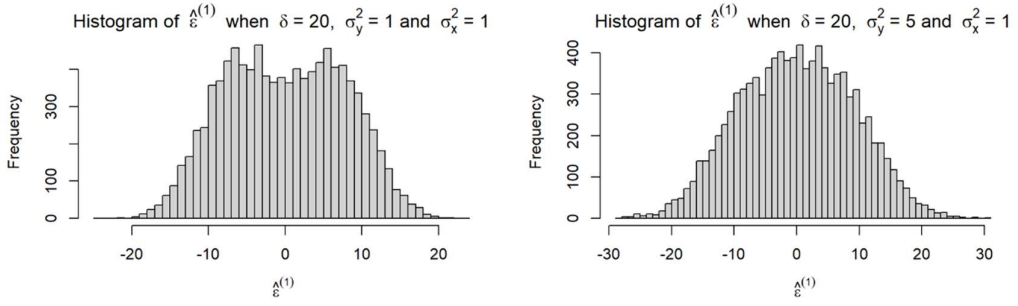
$$\hat{\beta}^{(1)} = (X^T X)^{-1} X^T y = \beta + (X^T X)^{-1} X^T Z\delta + (X^T X)^{-1} X^T \epsilon$$

which is biased because of Z. Now, obtain the residual

$$\hat{\epsilon}^{(1)} \equiv y - X\hat{\beta}^{(1)} = (I - X(X^T X)^{-1} X^T)\epsilon + (I - X(X^T X)^{-1} X^T) Z\delta$$

We use the bias in the residual $\hat{\epsilon}^{(1)}$ stemming from $(I - X(X^T X)^{-1} X^T) Z\delta$ to

reconstruct our estimate of Z. Since $Z_i = 0$ or $Z_i = 1$ for every unit i, $\hat{\epsilon}^{(1)}$ be distributed around 2 centers. This suggests that clustering $\hat{\epsilon}^{(1)}$ into 2 groups may be feasible. It depends on 3 terms: ϵ, $X(X^TX)^{-1}X^T$, and δ. Thus, the larger the variance of ϵ relative to δ, the more difficult it would be to successfully cluster $\hat{\epsilon}^{(1)}$. The following histograms of $\hat{\epsilon}^{(1)}$ with different noise-to-signal ratios illustrate cases where clustering seems promising and where it seems not:



Note that the histogram of $\hat{\epsilon}^{(1)}$ under the lower noise-to-signal ratio, as represented by σ_y^2/δ, suggests a better potential for clustering: its shape exhibits the mixture of two distributions.

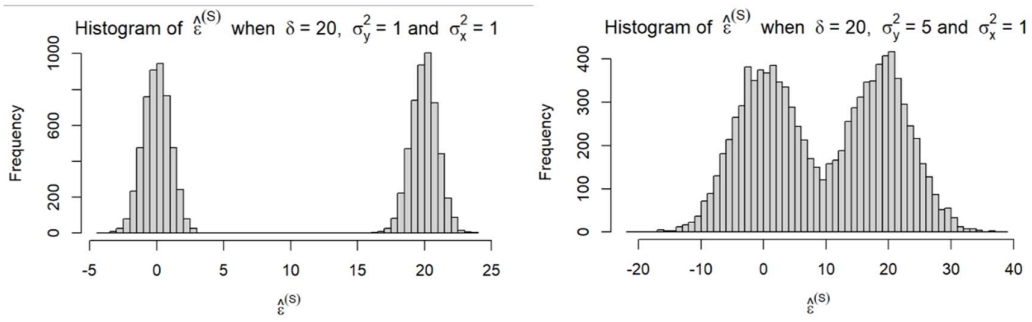
Step 2: Cluster $\hat{\epsilon}^{(1)}$ into 2 groups. We can use various probabilistic or non-probabilistic clustering algorithms for it. As an illustrative example, I used a simplest clustering algorithm, K-means, for this paper. Once we have two clusters, label the unit i's in the group with the larger center with $\hat{Z}_i^{(1)} = 1$, and label the units in the group with the smaller center with $\hat{Z}^{(1)} = 0$. This is our first estimate of $Z, \hat{Z}^{(1)}$.

Step 3: Regress Y on X and $\hat{Z}^{(1)}$ to obtain our second estimate of β, $\hat{\beta}^{(2)}$, and our first estimate of the ATE δ, $\hat{\delta}^{(1)}$. This completes the first cycle of this method.

Step 4: Now, if the histogram of $\hat{\epsilon}^{(1)}$ suggested good potential for clustering, then our estimate $\hat{Z}^{(1)}$ would provide good information about Z. Then our estimate $\hat{\beta}^{(2)}$ obtained from regressing Y on X and $\hat{Z}^{(1)}$ will be a better estimate of β than $\hat{\beta}^{(1)}$ that is obtained from regressing Y on X alone. This, in turn, implies that our second "residual" $\hat{\epsilon}^{(2)} = Y - X\hat{\beta}^{(2)}$ will provide more precise information about Z than $\hat{\epsilon}^{(1)} = Y - X\hat{\beta}^{(1)}$. Thus, clustering $\hat{\epsilon}^{(2)}$ into 2 groups will result in finer labeling and a

better estimate of Z, $\hat{Z}^{(2)}$. Finally, regressing Y on X and $\hat{Z}^{(2)}$ will result in better estimates of β, $\hat{\beta}^{(3)}$, and of δ, $\hat{\delta}^{(2)}$.

Step 5: Iterate Step 4 until $\left\|\hat{\beta}^{(S+1)} - \hat{\beta}^{(S)}\right\| < t$ for some small threshold value t, and obtain our final estimates of Z and ATE, $\hat{Z}^{(S)}$ and $\hat{\delta}^{(S)}$. Note that if $\left\|\hat{\beta}^{(S+1)} - \beta\right\| \approx 0$, then $\hat{\epsilon}^{(S)} = y - X\hat{\beta}^{(S+1)} \approx Z\delta + \epsilon$ and the histogram of $\hat{\epsilon}^{(S)}$ will have 2 centers 0 and δ. The following histograms of $\hat{\epsilon}^{(S)}$ illustrates the results of the iteration to the cases with two noise-to-signal ratios presented above:



Note that the histogram of $\hat{\epsilon}^{(S)}$ under the lower noise-to-signal ratio, σ_y^2/δ, exhibits a better separation of $\hat{\epsilon}^{(S)}$. But the histogram of $\hat{\epsilon}^{(S)}$ with the higher noise-to-signal ratio also centers around 0 and $\delta = 20$, suggesting that $\hat{\beta}^{(S+1)}$ converged to β quite well.

3. Results

I conducted a simulation study to measure how the performance of the proposed method changes along the two sources of noise-to-signal ratio: the variance of error terms σ_y^2 and the variance of the pre-treatment characteristics σ_x^2. For simulation, I used the following data generating process:

$$X_i = \left(X_{i,1}, X_{i,2}, \cdots, X_{i,16}\right) \sim_{iid} Multivariate\ Normal(0, \sigma_x^2 I_{16}),$$
$$\forall i \in \{1, \cdots, n = 10000\}$$

$$Z_i \sim Bernoulli(\pi_i), \qquad where\ \pi_i = \frac{\exp\{X_i\theta\}}{1 + \exp\{X_i\theta\}}, \qquad \forall i \in \{1, \cdots, n = 10000\}$$

$where\ \theta$

$= (-1, 0.5, -0.25, -0.1, -1, 0.5, -0.25, -0.1, -1, 0.5, -0.25, -0.1, -1, 0.5, -0.25, -0.1)$

$$Y_i = 210 + X_i^T\beta + Z_i\delta + \epsilon_i, \qquad where\ \epsilon \sim_{iid} N(0, \sigma_y^2), \qquad \forall i \in \{1, \cdots, n = 10000\}$$

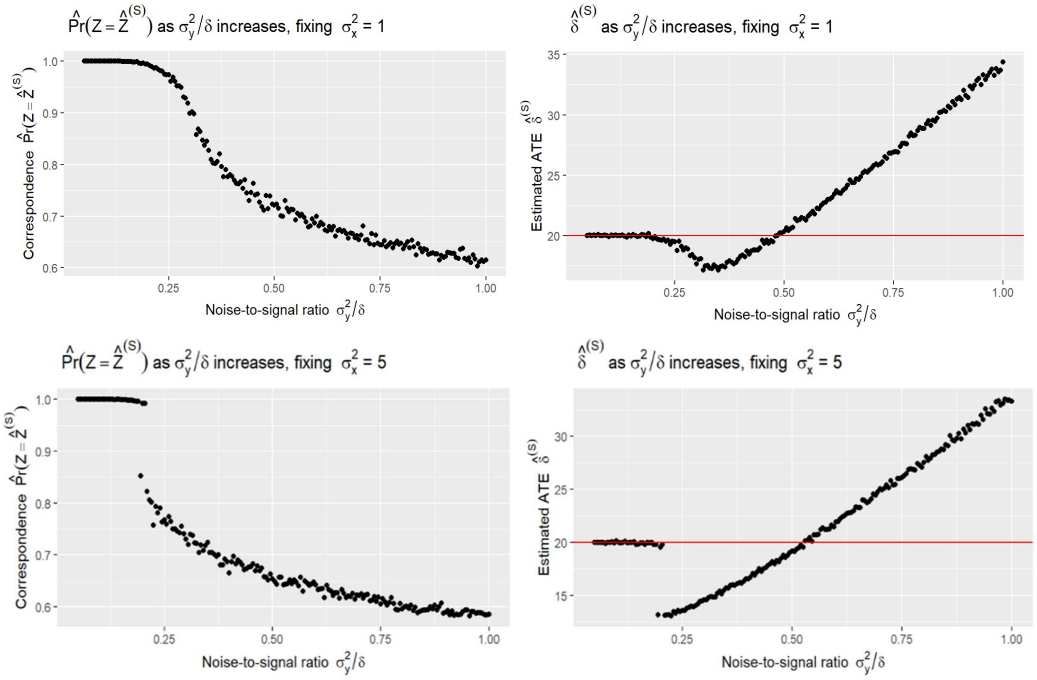
$where\ \beta$

$= (27.4, 13.7, -10, 20, 27.4, 13.7, -10, 20, 27.4, 13.7, -10, 20, 27.4, 13.7, -10, 20)\ and\ \delta$

$= 20$

I considered two measures for the performance of the method: $\widehat{Pr}(\{\hat{Z}^{(S)} = Z\})$ and $\hat{\delta}^{(S)}$. The first refers to the proportion of the final estimate $\hat{Z}^{(S)}$ that equals the true Z. The second lets us assess how close the final estimate $\hat{\delta}^{(S)}$ is to the true δ.

For the first set of simulations, I fixed σ_x^2 and increased σ_y^2 from 1 to 20 by 0.1. In terms of the noise-to-signal ratio σ_y^2/δ, it is an increase from 0.05 to 1. The plots are as follows:



The method performed surprisingly well, exhibiting $\widehat{Pr}(\{\hat{Z}^{(S)} = Z\})$ nearly 1 and $\delta^{(S)}$ very close to $\delta = 20$ until the noise-to-signal ratio σ_y^2/δ reaches around 0.1875, which

translates to σ_y^2 around 3.75. Past this point, the performance drops down increasingly in σ_y^2/δ. Interestingly, setting the higher value for σ_x^2 adversely affected the performance. Both $\widehat{Pr}\left(\{\hat{Z}^{(S)} = Z\}\right)$ and $\delta^{(S)}$ exhibits a point of discontinuity around the noise-to-signal ratio σ_y^2/δ of 0.1875.

For the second set of simulations, I fixed σ_y^2 and increased σ_x^2 from 1 to 20 by 0.1. In terms of the ratio σ_x^2/δ, it is an increase from 0.05 to 1. The plots are contained in the Appendix B1. They show that, for $\sigma_y^2 = 1$, the performance of the method is consistently good for any value of the ratio σ_x^2/δ, exhibiting $\widehat{Pr}\left(\{\hat{Z}^{(S)} = Z\}\right)$ nearly 1 and $\delta^{(S)}$ very close to $\delta = 20$ throughout. For $\sigma_y^2 = 3.75$, however, the performance plunges with discontinuity around the ratio σ_x^2/δ of 0.35. This shows that increasing σ_x^2 may adversely affect the performance of the method when the noise-to-signal ratio σ_y^2/δ is sufficiently large.

4. Discussion

The simulation results confirm the expectation that successful clustering of $\hat{\epsilon}^{(S)}$ depends on how large δ is relative to σ_y^2 and σ_x^2. Furthermore, the histograms of $\hat{\epsilon}^{(S)}$ presented in Section 2.2 and Appendix B2 suggest that the shape of the histogram of $\hat{\epsilon}^{(S)}$ can serve as a heuristic that tells us if our estimate of Z, $\hat{Z}^{(S)}$, is credible and if the noise-to-signal ratio was sufficiently small for our method to be applicable.

The method was robust to increasing sample size up to $n = 100{,}000$ at least and the dimensions of covariates up to $p = 40$ at least. The plots that summarize the simulation results are presented in Appendix B3.

Uncertainty quantification of the estimate $\hat{Z}^{(S)}$ and $\hat{\delta}^{(S)}$ is tricky due to the iterative nature of our method. For low values of noise-to-signal ratio, specifically, usual bootstrapping may not give us the variance estimate as $\widehat{Pr}\left(\{\hat{Z}^{(S)} = Z\}\right)$ would be consistently close to 1. Perturbation approach is another option to consider, but further research is required for the full exposition of this question.

On the other hand, application of this method to non-compliance settings is straightforward. There, we would have an additional variable, "assigned treatment" W, as distinguished from "actual treatment" Z. But the assigned treatment W_i would affect outcome Y_i only through its effect on the actual treatment Z_i, implying that W_i does not enter in the outcome model. Since our method only relies on the outcome model to estimate Z, no change is needed to use it for recuperation of the actual treatment variable Z.

Now we compare our method to two other methods that can be used for recuperating the treatment variable Z: EM algorithm and Gibbs sampling. Given the outcome model we assumed, EM algorithm for two-component mixture of normals with regression mean structure will do the job. The EM algorithm had the value of $\widehat{Pr}\left(\left\{\hat{Z}^{(1)} = Z\right\}\right)$ comparable to the first round of demeaning followed by clustering, and sometimes better depending on the initialization. However, the value of $\widehat{Pr}\left(\left\{\hat{Z}^{(S)} = Z\right\}\right)$ after the full rounds of the proposed method was consistently and considerably higher than that of the EM algorithm, especially when the noise-to-signal ratio was low so that it was nearly 1. Given that the estimated ATE deviates considerably from the true ATE even when the value of $\widehat{Pr}\left(\left\{\hat{Z}^{(S)} = Z\right\}\right)$ is around 0.85, this difference may not be negligible. Over the full Bayesian implementation of Gibbs sampling, on the other hand, our method has computational advantage that increases in sample size. Finally, since our method is distance-based, it can be flexibly adapted to non-spherical error terms by changing the metrics.

Appendix

A. Reference

Daskalakis, C. Tzamos, C. & Zampetakis M.. (2017) Ten steps of em suffice for mixtures of two gaussians. In Conference on Learning Theory, pages 704–710.

Kwon, J., Qian, W., Caramanis, C., Chen, Y. & Davis, D.. (2019). Global Convergence of the EM Algorithm for Mixtures of Two Component Linear Regression. *Proceedings of the Thirty-Second Conference on Learning Theory*, in *Proceedings of Machine Learning Research*, 99:2055-2110.

Mantz, A., (2016) A Gibbs sampler for multivariate linear regression, *Monthly Notices of the Royal Astronomical Society*, Volume 457, Issue 2, 1279–1288,

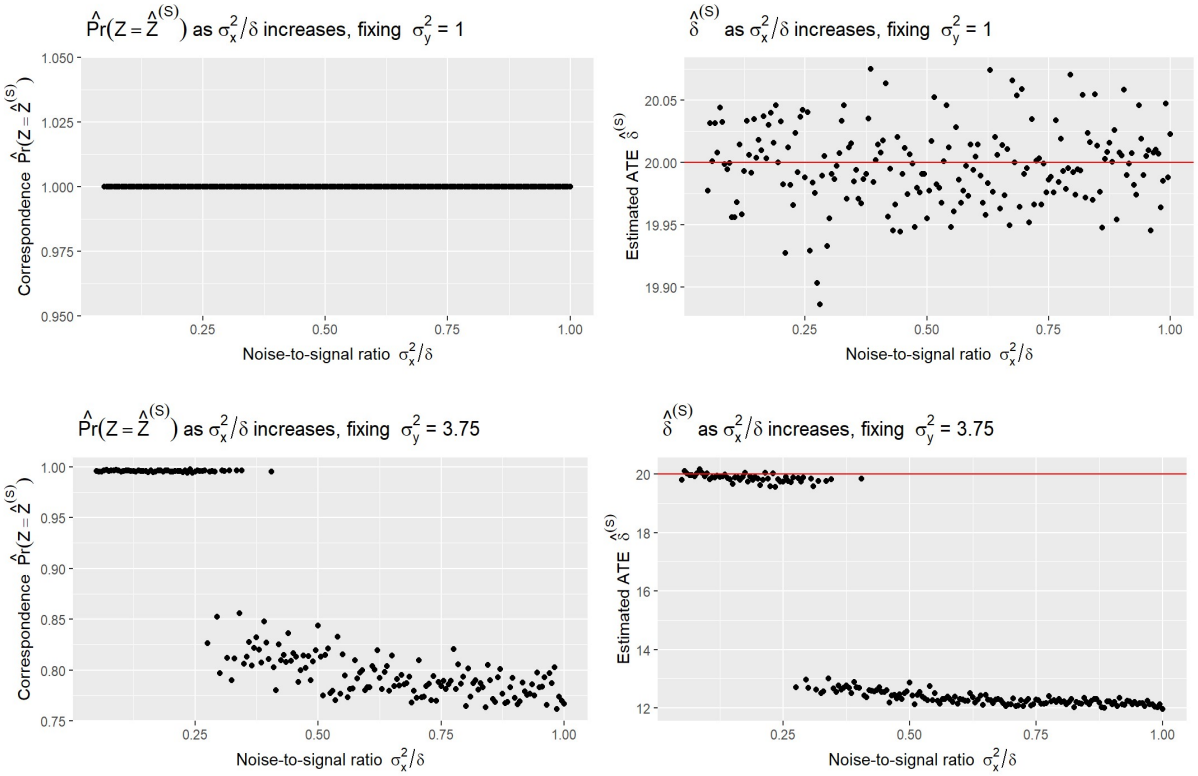
Rubin, D. B. (1973). The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*, 185-203.

Rubin D. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*; 74:318-324.

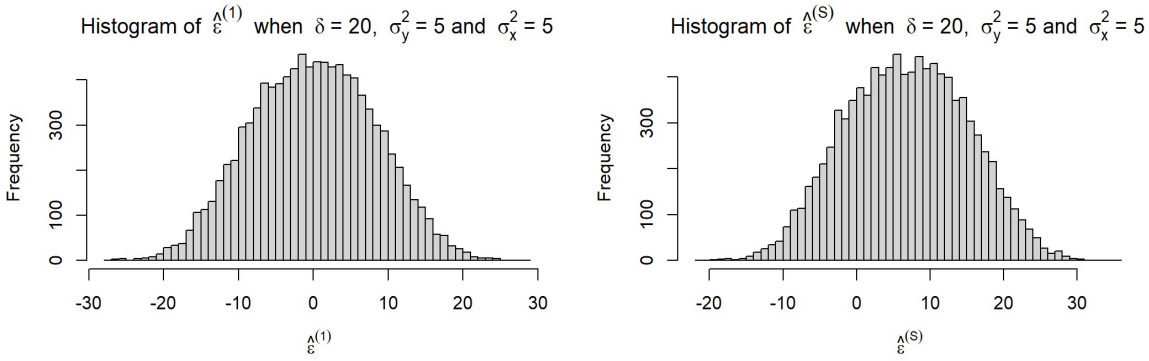
Rosenbaum P, Rubin D. (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika*; 70(1):41-55.

B. Figures

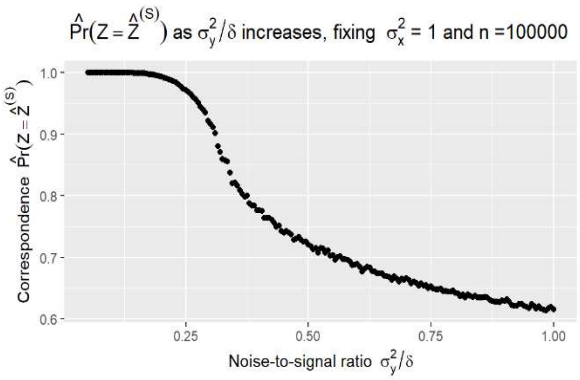
B1.

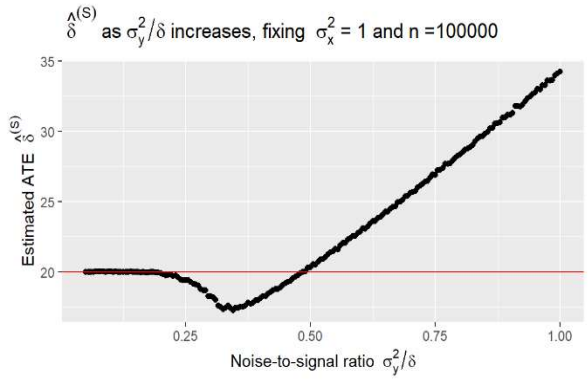


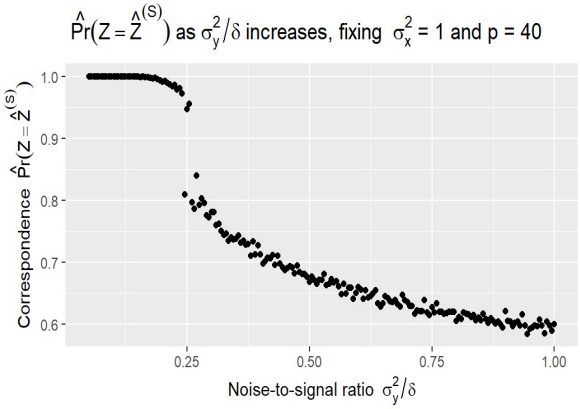
B2.

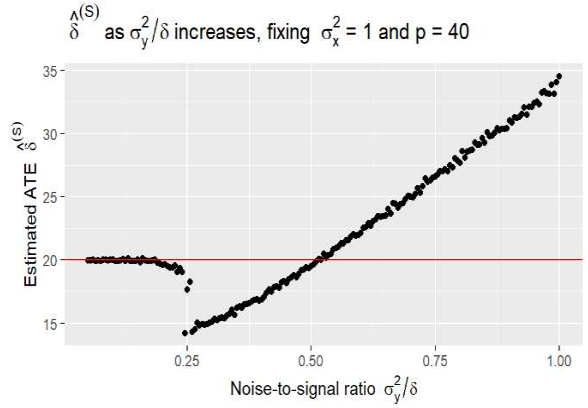


B3.


$\hat{Pr}(Z = \hat{Z}^{(S)})$ as σ_y^2/δ increases, fixing σ_x^2 = 1 and n =100000
Correspondence $\hat{Pr}(Z = \hat{Z}^{(S)})$
1.0
0.9
0.8
0.7
0.6
0.25
0.50
0.75
1.00
Noise-to-signal ratio σ_y^2/δ


$\hat{\delta}^{(S)}$ as σ_y^2/δ increases, fixing σ_x^2 = 1 and n =100000
Estimated ATE $\hat{\delta}^{(S)}$
35
30
25
20
0.25
0.50
0.75
1.00
Noise-to-signal ratio σ_y^2/δ


$\hat{Pr}(Z = \hat{Z}^{(S)})$ as σ_y^2/δ increases, fixing σ_x^2 = 1 and p = 40
Correspondence $\hat{Pr}(Z = \hat{Z}^{(S)})$
1.0
0.9
0.8
0.7
0.6
0.25
0.50
0.75
1.00
Noise-to-signal ratio σ_y^2/δ


$\hat{\delta}^{(S)}$ as σ_y^2/δ increases, fixing σ_x^2 = 1 and p = 40
Estimated ATE $\hat{\delta}^{(S)}$
35
30
25
20
15
0.25
0.50
0.75
1.00
Noise-to-signal ratio σ_y^2/δ