Estimation of Average Treatment Effects for Massively Unbalanced Binary Outcomes

Jinyong Hahn[1] Xueyuan Liu Geert Ridder

UCLA UCLA USC

May 31, 2022

[1] Corresponding Author. hahn@econ.ucla.edu

Abstract

The MLE of the ATE in the logit model for binary outcomes may have a significant second order bias if the event has a low probability, the case we focus on in this paper. We derive the second order bias of the logit ATE estimator. We also propose bias-corrected estimators of the ATE. We also propose a variation on the logit model with parameters that are elasticities. Finally, we propose a computational trick that avoids numerical instability in the case of estimation for rare events.

Keywords: Rare events, Logit, Second Order Bias, Average Treatment Effects, Constant Elasticity

Word Count: 11,111

1 Introduction

We examine the properties of the maximum likelihood estimator (MLE) for logit models when the outcome is the occurrence or not of a low probability event. King and Zeng (2001) considered logistic regression for rare events data and focused on correcting the bias of estimators of the regression coefficients and event probabilities. They were motivated by the fact that in political science data the binary dependent variable takes the value one (for "events", such as wars, coups, presidential vetoes, the decision of citizens to run for political office, or infection by an uncommon disease) much less frequently than the value zero (for "nonevents"). The analysis of rare events is relevant for economics because some of the big data sets are collected from online sources where the number of events (such as "clicks" and "purchases") is much smaller than the number of nonevents.

King and Zeng (2001) considered various statistical problems with rare event data, including sample selection problems and finite sample biases. They primarily discussed issues related to the predicted event probabilities. In this paper, we focus on the finite sample bias in the estimator of the average treatment effects (ATE). We derive the higher order bias of the logit MLE and the implied bias of the estimator of the ATE, and analyze the finite sample properties of the bias corrected estimator by Monte Carlo simulations.

It may seem unnatural to use a higher order expansion to derive the finite sample bias in the case of rare events. Indeed, Wang (2020) proposed an intuitive asymptotic approximation where the intercept term of the logit model diverges to negative infinity as a function of the sample size, so that the implied probability of the event converges to zero. We show that his asymptotic approximation is equivalent to the usual first-order asymptotic approximation where the sample size grows to infinity and the parameters are fixed. Therefore, Wang's (2020) results have implications for the efficiency of various sampling methods, but do not shed light on the finite sample behavior of logit MLE if events are rare.

In Section 2, we derive the higher order bias of the logit MLE as well as of the related estimator of the ATE. In Section 3, we provide intuition for the bias of the logit MLE exploiting the invariance property of the MLE. In Section 4, we argue why the higher order approach should be preferred over Wang's (2020) intuitive asymptotics. In Section 5, we present the simulation evidence. In Section 6, we propose a new binary response model with constant elasticity, and in Section 7, we develop a trick to reduce the numerical instability in the calculation of the MLE for to rare events.

2 Higher Order Bias of the Logit MLE and Related ATE Estimator

In this section, we consider the logit model where the binary dependent variable $y_i = 1$ with probability equal to $\Lambda\left(x_i'\theta_0\right)$, where $\Lambda\left(t\right) \equiv \exp\left(t\right)/\left(1+\exp\left(t\right)\right)$ denotes the cumulative distribution function (CDF) of the logistic distribution. We first derive the second order bias of the logit MLE as discussed in Rilstone, Srivastava, and Ullah (1996). We show that the second order bias thus calculated is larger if events are rare.

The idea underlying the second order expansion of the generic MLE is straightforward. Suppose that the density of the random vector z_i is given by $f\left(z;\theta_0\right)$, and the MLE $\widehat{\theta}$ maximizes the joint log likelihood $\sum_{i=1}^{n}\log f\left(z_i;\theta\right)$. The first order condition is $\frac{1}{n}\sum_{i=1}^{n} v\left(z_i;\widehat{\theta}\right) = 0$, where $v(z;\theta) = \partial \log f(z;\theta)/\partial\theta$. By manipulating this first order condition, we can derive the implication that

$$\widehat{\theta} - \theta_0 = \frac{1}{\sqrt{n}}\theta^{\epsilon}(0) + \frac{1}{2}\left(\frac{1}{\sqrt{n}}\right)^2 \theta^{\epsilon\epsilon}(0) + o_p\left(\frac{1}{n}\right),$$

where $\theta^{\epsilon}(0) = O_p(1)$ and $\theta^{\epsilon\epsilon}(0) = O_p(1)$. The $\theta^{\epsilon}(0)$ has mean zero,[1] and the second order bias is defined as the expectation of the second term on the RHS above, i.e., $\frac{1}{2n}E[\theta^{\epsilon\epsilon}(0)]$. In Appendix A.3, we show how the second order bias can be estimated; letting B denote a consistent estimator of $\frac{1}{2}E[\theta^{\epsilon\epsilon}(0)]$, the bias corrected estimator $\widetilde{\theta}$ is $\widetilde{\theta} = \widehat{\theta} - B/n$. Our bias formula looks somewhat different from the one presented in King and Zeng (2001), which in turn is based on McCullagh and Nelder (1989). It can be shown that they are in fact identical. See Appendix B.

We now present the main result of this section. We consider the treatment effect model where we have $y_i = 1$ with probability $\Lambda\left(x_i'\theta_{0,(1)}\right)$ under treatment ($D = 1$), and with probability $\Lambda\left(x_i'\theta_{0,(0)}\right)$ under control ($D = 0$). We can estimate the $\theta_{0,(1)}$ and $\theta_{0,(0)}$ for the two sub-samples with $D = 1$ and 0. [2] The average treatment effect (ATE) is equal to $E\left[\Lambda\left(x_i'\theta_{0,(1)}\right)\right] - E\left[\Lambda\left(x_i'\theta_{0,(0)}\right)\right]$, which can be estimated by the natural estimator

$$\frac{1}{n}\sum_{i=1}^{n}\Lambda\left(x_i'\widehat{\theta}_{(1)}\right) - \frac{1}{n}\sum_{i=1}^{n}\Lambda\left(x_i'\widehat{\theta}_{(0)}\right),$$

where $\widehat{\theta}_{(1)}$ and $\widehat{\theta}_{(0)}$ denote the MLE of $\theta_{0,(1)}$ and $\theta_{0,(0)}$ using the treated and control subsamples. When the outcome 1 is a rare event under treatment and/or control, the bias of the natural estimator above can be corrected, where the bias correction has to reflect the nonlinearity of

[1] Not surprisingly, the $\theta^{\epsilon}(0)$ converges in distribution to a normal distribution with mean zero and variance equal to the inverse of the information matrix.

[2] We assume that the treatment assignment is unconfounded

Λ as well as the finite sample bias of the MLE. In particular, the bias correction follows from the second order expansion

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\Lambda\left(x_i'\widehat{\theta}_{(1)}\right)-\frac{1}{n}\sum_{i=1}^{n}\Lambda\left(x_i'\theta_{0,(1)}\right)\\
&=\frac{1}{n}\sum_{i=1}^{n}\Lambda\left(x_i'\theta_{0,(1)}\right)\left(1-\Lambda\left(x_i'\theta_{0,(1)}\right)\right)x_i'\left(\widehat{\theta}_{(1)}-\theta_{0,(1)}\right)\\
&+\frac{1}{2}\frac{1}{n}\sum_{i=1}^{n}\Lambda\left(x_i'\theta_{0,(1)}\right)\left(1-\Lambda\left(x_i'\theta_{0,(1)}\right)\right)\left(1-2\Lambda\left(x_i'\theta_{0,(1)}\right)\right)\left(x_i'\left(\widehat{\theta}_{(1)}-\theta_{0,(1)}\right)\right)^2\\
&+o_p\left(n^{-1}\right).
\end{aligned}
$$

The bias of the ATE estimator derives from the two terms on the right. The (expectation of the) first term is proportional to the second order bias of $\widehat{\theta}_{(1)}$. The second term on the right contributes through the expectation of $\left(x_i'\left(\widehat{\theta}_{(1)}-\theta_{0,(1)}\right)\right)^2$ and is non-zero due to the curvature of Λ. See Appendix C for details on implementing the bias correction, if this expansion is applied to both $\frac{1}{n}\sum_{i=1}^{n}\Lambda\left(x_i'\widehat{\theta}_{(1)}\right)$ and $\frac{1}{n}\sum_{i=1}^{n}\Lambda\left(x_i'\widehat{\theta}_{(0)}\right)$.

Remark 1 *The second order bias of the ATE estimator can be shown to be zero if the treatment is randomly assigned. See Appendix E for a proof. By random assignment, we mean that D is independent of x so that the propensity score $\Pr\left(D_i=1\middle|\, x_i\right)$ is constant in x. Under random assignment, the distribution of x is identical across the two subsamples $D=1$ and $D=0$. Therefore, the intuition to be discussed in Remark 2 applies, and the second order bias is zero. Under unconfounded treatment assignment we should expect some amount of second order bias. Its magnitude is an empirical matter.*

3 Intuition

In this section, we provide the intuition underlying the second order bias. Using a model without regressors, we explain that the bias is due to the inherent nonlinearity of the logit model. We show this using the invariance property of the MLE. The invariance is also used to explain the bias of the predicted probabilities in models without regressors.

We consider the binary response model where $y = 1$ with probability equal to $\Lambda(\theta) = \exp(\theta) / (1 + \exp(\theta))$. The second order bias of the logit MLE, presented in Appendix A.3, simplifies to

$$-\frac{1}{2n}\frac{(1-2\Lambda)}{\Lambda(1-\Lambda)} \tag{1}$$

with $\Lambda = \Lambda(\theta_0)$. Note that if $\Lambda \approx 0$, the bias is negative. Also note that

$$\lim_{\Lambda \to 0}\left(-\frac{(1-2\Lambda)}{\Lambda(1-\Lambda)}\right) = -\infty$$

so the bias is larger if we are dealing with rare events. In order to understand the bias, we note that the first order condition can be rewritten as $\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \Lambda\left(\widehat{\theta}\right)\right) = 0$. In other words, the logit MLE solves

$$\Lambda\left(\widehat{\theta}\right) = \overline{y} \tag{2}$$

and hence

$$\widehat{\theta} = \Lambda^{-1}(\overline{y}) = \ln\frac{\overline{y}}{1-\overline{y}}.$$

By the CLT, we have $\sqrt{n}(\overline{y} - \Lambda) \to N(0, \Lambda(1-\Lambda))$, so we can write without loss of generality that

$$\overline{y} = \Lambda + \frac{1}{\sqrt{n}}\left(\sqrt{\Lambda(1-\Lambda)}Z + o_p(1)\right)$$

where $Z \sim N(0,1)$. It follows that

$$\begin{aligned}
\widehat{\theta} &= \ln\left(\frac{\Lambda + \frac{1}{\sqrt{n}}\left(\sqrt{\Lambda(1-\Lambda)}Z + o_p(1)\right)}{1-\Lambda-\frac{1}{\sqrt{n}}\left(\sqrt{\Lambda(1-\Lambda)}Z + o_p(1)\right)}\right) \\
&= \ln\left(\frac{\Lambda}{1-\Lambda}\right) + \frac{n^{-1/2}}{\sqrt{\Lambda(1-\Lambda)}}Z - n^{-1}\frac{1}{2}\frac{1-2\Lambda}{\Lambda(1-\Lambda)}Z^2 + o_p\left(n^{-1}\right) \\
&= \theta_0 + \frac{n^{-1/2}}{\sqrt{\Lambda(1-\Lambda)}}Z - n^{-1}\frac{1}{2}\frac{1-2\Lambda}{\Lambda(1-\Lambda)}Z^2 + o_p\left(n^{-1}\right)
\end{aligned}$$

It follows that the second order bias of $\widehat{\theta}$ is

$$-n^{-1}\frac{1}{2}\frac{1-2\Lambda}{\Lambda(1-\Lambda)}E\left[Z^2\right] = -n^{-1}\frac{1}{2}\frac{1-2\Lambda}{\Lambda(1-\Lambda)} \tag{3}$$

confirming the second order bias calculation in (1).

Remark 2 *The first order condition of the MLE can be a convenient tool to understand the (lack of) bias of the average of the predicted probabilities in a logit model with regressors. For this purpose, we note that the probability $E[\Lambda(x'\theta_0)]$ of $y = 1$ can be estimated by $\frac{1}{n}\sum_{i=1}^{n}\Lambda\left(x_i'\widehat{\theta}\right)$. The same second-order expansion as in Section 2 gives*

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\Lambda\left(x_i'\widehat{\theta}\right) - \frac{1}{n}\sum_{i=1}^{n}\Lambda\left(x_i'\theta_0\right) &= \frac{1}{n}\sum_{i=1}^{n}\Lambda\left(x_i'\theta_0\right)\left(1-\Lambda\left(x_i'\theta_0\right)\right)x_i'\left(\widehat{\theta}-\theta_0\right) \\
&\quad + \frac{1}{2}\frac{1}{n}\sum_{i=1}^{n}\Lambda\left(x_i'\theta_0\right)\left(1-\Lambda\left(x_i'\theta_0\right)\right)\left(1-2\Lambda\left(x'\theta_0\right)\right)\left(x_i'\left(\widehat{\theta}-\theta_0\right)\right)^2 \\
&\quad + o_p\left(n^{-1}\right)
\end{aligned}$$

We will assume that the first component of x_i is an intercept term. If so, we recall by the first order condition that

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \Lambda\left(x_i'\widehat{\theta}\right)\right) = 0$$

so the bias should be zero. Straightforward algebra shows that the second order bias of the $\frac{1}{n}\sum_{i=1}^{n}\Lambda\left(x_i'\widehat{\theta}\right)$ is indeed zero, confirming the intuition. See Appendix D.

Note that the ATE estimator is the difference of the average predicted probabilities for the treated and the controls. This estimator is biased because the average is over the full sample and not over the subsamples of the treated and the controls. Under random assignment the distribution of X is the same in the subsamples, so that the ATE estimator is unbiased, even if events are rare.

4 Comparison with Wang (2020)

We now compare our higher order asymptotics with Wang's (2020) asymptotics where the probability of $y = 1$ is assumed to converge to zero as a function of the sample size. Using the same model without regressors that was discussed in the previous section, we examine Wang's (2020) asymptotic analysis, and argue that his asymptotics is identical to the traditional fixed parameter asymptotics for all practical purposes.

The data are $y_1, \ldots, y_n$ which are IID Bernoulli variables such that $y_i = 1$ with probability p_n. We assume that $p_n \propto n^{-\delta}$ with $0 \leq \delta < 1$, and consider the normalized sum

$$\sum_{i=1}^{n} \frac{y_i - p_n}{\sqrt{np_n\left(1 - p_n\right)}}.$$

In Appendix F we show that the Lyapunov condition is satisfied[3] and

$$\frac{\sqrt{n}\left(\overline{y} - p_n\right)}{\sqrt{p_n\left(1 - p_n\right)}} = \sum_{i=1}^{n} \frac{y_i - p_n}{\sqrt{np_n\left(1 - p_n\right)}} \rightarrow N\left(0, 1\right), \tag{4}$$

so we can write

$$\overline{y} = p_n + \frac{\sqrt{p_n\left(1 - p_n\right)}}{\sqrt{n}} Z_n,$$

[3] Appendix F also notes that $\delta = 1$ (so that $p_n \propto n^{-1}$) is not compatible with asymptotic normality, and that this rate is not appropriate for logit models.

where $Z_n = O_p(1)$ is such that $E[Z_n] = 0$ and $\operatorname{Var}(Z_n) = 1$. In Appendix G we derive the expansion

$$\sqrt{np_n(1-p_n)}\left(\widehat{\theta} - \theta_0\right) = Z_n - \frac{1}{2}\frac{1-2p_n}{\sqrt{np_n(1-p_n)}}Z_n^2 + O_p\left(n^{-(1-\delta)}\right) \tag{5}$$

The expansion (5) implies that the higher order bias can be calculated as

$$E\left[-\frac{1}{2}\frac{1-2p_n}{np_n(1-p_n)}Z_n^2\right] = -\frac{1}{2}\frac{1-2p_n}{np_n(1-p_n)}.$$

If we compare this expression with the higher order bias (3) based on the fixed probability Λ, equating Λ and p_n, we see that the higher order bias under the asymptotics where $p_n \to 0$ is identical to the higher order bias under the asymptotics where p_n is fixed at Λ. In other words, Wang's asymptotics gives the same higher order bias as the fixed parameter asymptotics.

This analysis raises the question about the relevance of Wang's (2020) asymptotic framework for our purpose. It is helpful to make an explicit link between p_n and the logit model, which we will do by writing $p_n = \Lambda(\theta_n)$. The sum of y_i over the entire sample is binom$(n, \Lambda(\theta_n))$. Using Wang's notation, we have $n_1 \sim$binom$(n, \Lambda(\theta_n))$. His equation (2) means that he is considering $p_n \propto n^{-\delta}$ with $0 \le \delta < 1$, as we do here.[4]

His Theorem 1 boils down to $\sqrt{n_1}\left(\widehat{\theta} - \theta\right) \to N(0,1)$ in the model with only an intercept. Using his equation (3), this is equivalent to the approximation $\sqrt{n\Lambda(\theta_n)}\left(\widehat{\theta} - \theta\right) \approx N(0,1)$ or

$$\widehat{\theta} \approx N\left(\theta, \frac{1}{n\Lambda(\theta_n)}\right) = N\left(\theta, \frac{1}{np_n}\right)$$

[4] His equation (2) implies that $\Lambda(\theta_n) \to 0$, $n\Lambda(\theta_n) \to \infty$. Note that $\Lambda(\theta_n) \to 0$ means that $1/(1+\exp(\theta_n)) \to 1$, which in turn means that $\exp(\theta_n) \to 0$, or $\theta_n \to -\infty$. Also note that $n\Lambda(\theta_n) \to \infty$ rules out the Poisson approximation, because if $\Lambda(\theta_n) \propto n^{-1}$, we cannot have $n\Lambda(\theta_n) \to \infty$. On the other hand, $n\Lambda(\theta_n) \to \infty$ is satisfied as long as $\Lambda(\theta_n) \propto n^{-\delta}$ with $0 \le \delta < 1$.

If we ignore the higher order term involving Z_n^2, our (5) along with (4) implies $\sqrt{np_n\left(1-p_n\right)}\left(\widehat{\theta}-\theta\right)\to N\left(0,1\right)$ or

$$\widehat{\theta}\approx N\left(\theta,\frac{1}{np_n\left(1-p_n\right)}\right),$$

which is the same approximation that we obtain when $p_n=\Lambda$ and does not vary as a function of the sample size. When $p_n\approx 0$, the difference between the two standard errors is small

$$\frac{1/\left(np_n\right)}{1/\left(np_n\left(1-p_n\right)\right)}=1-p_n\approx 1.$$

Therefore, Wang's (2020) asymptotic framework leads to the same first order asymptotic approximation as our (5) that was derived using the classical first order asymptotics with fixed parameters. A straightforward calculation suggests that even with regressors, Wang's (2020) asymptotic framework does not offer a substantively different approximation than the classical first order asymptotic approximation. In particular Wang is silent on the second order bias.

5 A New Binary Response Model for Rare Events

In this paper, the primary object of interest is the ATE if the outcome is the occurrence of a rare event. On the other hand, in some applications it is more useful to have an estimate of the elasticity of the rare event probability with respect to the independent variables. Consider the logit model where $y=1$ with probability $\Lambda\left(\alpha+x\beta\right)$, where x is a scalar that is measured on the log scale. The elasticity of the event probability with respect to x is equal to

$$\frac{\Lambda'\left(\alpha+x\beta\right)\beta}{\Lambda\left(\alpha+x\beta\right)}=\left(1-\Lambda\left(\alpha+x\beta\right)\right)\beta,$$

where we used $\partial\Lambda(t)/\partial t = \Lambda(t)(1-\Lambda(t))$. Note that $1-\Lambda(\alpha+x\beta)\approx 1$ if the event is rare, so the logit model exhibits near constant elasticity for rare events. If the elasticity is approximated by β, one may be interested in correcting the bias of the MLE of β. The simulation results in Tables 2, 3, and 4 show, that the bias in the MLE for β is modest but may be important depending on the application.[5] Therefore, one can use the bias corrected estimator of the slope coefficient as the bias corrected estimator of the elasticity in the rare event case.

Given that the β is only an approximate elasticity, the case for bias correction may not be so compelling for logit models. Instead one may choose to work with a model that has a constant elasticity. Let $P(x)$ denote the probability that $y=1$ as a function of x. A model that has a constant elasticity should satisfy

$$\frac{P'(x)}{P(x)} = \text{ constant}$$

assuming that x is measured in logs. This is equivalent to

$$\frac{d\ln P(x)}{dx} = \beta$$

for some β, so by integration we obtain $\ln P(x) = \alpha + \beta x$, or

$$P(x) = \exp(\alpha+\beta x)$$

as a model of constant elasticity.[6] Because $\exp(\alpha+\beta x)\approx \exp(\alpha+\beta x)/(1+\exp(\alpha+\beta x))$

[5] In Table 2, for the $\alpha_0 = -2.5$ and $n = 500$ combination, the bias of $\hat{\beta}$ is 0.0168, where the true value of β is 1. So, the MLE overestimates the elasticity by 1.68 %, which is reduced to 0.41 % by the bias corrected estimator.

[6] When $\alpha\approx -\infty$ and the support of x is bounded, we can guarantee that $\exp(\alpha+\beta x) < 1$. If the support

when $\exp(\alpha + \beta x) \approx 0$,[7] one can argue that this new model is an approximation of the logit model but with the convenient feature of a constant elasticity.

6 Artificial Censoring - Overcoming the Numerical Instability

In our Monte Carlo simulations, we encountered numerical stability problems with the computation of the MLE for extremely small values of α. The problem is that the optimization algorithm does not converge for such values of α that may be visited during the search for the MLE. As a consequence, we could not calculate the MLE for these data sets. Given that the log likelihood of the logit model is globally concave, in theory this sort of problem should not happen. The problem is that the Hessian of the log-likelihood is close to singular for rare events. We offer a simple solution, which seems to resolve the problem.

Consider the logit model where $y = 1$ with probability $\Lambda(\alpha + x\beta)$, where x is a scalar. We would normally maximize the log-likelihood

$$L(\alpha, \beta) = \sum_{i=1}^{n} \left(y_i \log \Lambda(\alpha + x_i\beta) + (1 - y_i) \log\left(1 - \Lambda(\alpha + x_i\beta)\right)\right).$$

of x is not bounded, but if we are sure that $\alpha + \beta x < 0$ for most values of x, we may want to adopt a parameterization

$$P(x) = \Psi(\alpha + \beta x)$$

where $\Psi(t) = \frac{1}{2}\exp(t)$ if $t < 0$, and $\Psi(t) = 1 - \frac{1}{2}\exp(-t)$ if $t > 0$.

[7] It is in the sense that

$$\frac{t}{1+t} = t + O\left(t^2\right).$$

It is straightforward to see that

$$\begin{aligned}
\frac{\partial^2 L(\alpha,\beta)}{\partial\alpha^2} &= -\sum_{i=1}^{n} \Lambda(\alpha+x_i\beta)(1-\Lambda(\alpha+x_i\beta)) \\
\frac{\partial^2 L(\alpha,\beta)}{\partial\alpha\partial\beta} &= -\sum_{i=1}^{n} \Lambda(\alpha+x_i\beta)(1-\Lambda(\alpha+x_i\beta))\,x_i \\
\frac{\partial^2 L(\alpha,\beta)}{\partial\beta^2} &= -\sum_{i=1}^{n} \Lambda(\alpha+x_i\beta)(1-\Lambda(\alpha+x_i\beta))\,x_i^2
\end{aligned}$$

so the second derivative is in theory strictly negative definite. On the other hand, if $\alpha \approx -\infty$, we have $\Lambda(\alpha + x_i\beta) \approx 0$ and as a consequence, the second derivative matrix is close to zero, and therefore close to singular. Because the step length of the Newwton-Raphson algorithm depends on the inverse of the Hessian, the algorithm may not converge.

In order to overcome this problem, we consider artificial censoring of the $y = 0$ outcomes. The rare event logit model is unchanged and x is always observed. However we censor observations with $y = 0$ with probability π. So there are three possible outcomes, $y = 1$ (and observed), with probability $\Lambda(\alpha + x\beta)$, $y = 0$ and observed with probability $(1-\pi)(1-\Lambda(\alpha+x\beta))$, and $y = 0$ and not observed with probability $\pi(1-\Lambda(\alpha+x\beta))$.

Therefore, the probability that the econometrician observes outcome $y = 1$ conditional on x, is

$$\begin{aligned}
\frac{\Lambda(\alpha+x\beta)}{\Lambda(\alpha+x\beta)+(1-\pi)(1-\Lambda(\alpha+x\beta))} &= \frac{\frac{\exp(\alpha+x\beta)}{1+\exp(\alpha+x\beta)}}{\frac{\exp(\alpha+x\beta)}{1+\exp(\alpha+x\beta)}+\exp(\delta)\frac{1}{1+\exp(\alpha+x\beta)}} \\
&= \frac{\exp(\alpha+x\beta)}{\exp(\alpha+x\beta)+\exp(\delta)} \\
&= \frac{\exp(\alpha^*+x\beta)}{\exp(\alpha^*+x\beta)+1} \\
&= \Lambda(\alpha^*+x\beta),
\end{aligned}$$

and the probability the econometrician observes $y = 0$ given x is

$$\begin{aligned}\frac{(1-\pi)(1-\Lambda(\alpha+x\beta))}{\Lambda(\alpha+x\beta)+(1-\pi)(1-\Lambda(\alpha+x\beta))} &= \frac{\exp(\delta)\frac{1}{1+\exp(\alpha+x\beta)}}{\frac{\exp(\alpha+x\beta)}{1+\exp(\alpha+x\beta)}+\exp(\delta)\frac{1}{1+\exp(\alpha+x\beta)}} \\ &= \frac{\exp(\delta)}{\exp(\alpha+x\beta)+\exp(\delta)} \\ &= \frac{1}{\exp(\alpha^*+x\beta)+1} \\ &= 1-\Lambda(\alpha^*+x\beta),\end{aligned}$$

where $\Lambda(\alpha+x\beta)+(1-\pi)(1-\Lambda(\alpha+x\beta))$ is the probability that the outcome is observed, and

$$\exp(\delta) \equiv 1-\pi, \quad \alpha^* \equiv \alpha-\delta.$$

Note that $\delta < 0$. If we choose δ (i.e., π) such that α^* is not close to $-\infty$, the Hessian for the censored sample is not (close to) singular. In general π can be chosen so that the expected number of observed 0 outcomes is about equal to the number of observed 1 outcomes.

Because δ is chosen by the econometrician, from an estimate α^* for the artificially censored sample we can back out α.

In order to see whether this trick is useful, we drew one sample of size $n = 50,000$ and another sample of size $n = 100,000$ such that $\alpha = -10$, $\beta_1 = \cdots = \beta_9 = 1$, and all the nine independent variables are independent and $N(0,1)$. We used $\beta_1 = \cdots = \beta_9 = 0$ as the starting value of the Newton-Raphson algorithm. As for the starting value for α, we chose the true value of α (i.e., -10) for the full sample, while we used $\alpha^* = -10-\delta$ for the subsample after the random censoring. In our sample of $n = 50,000$, the number of 1's was 116 and the missing probability is set to $\pi = 1-0.002$, implying that $\delta = \ln(0.002) = -6.21$.) Convergence results for Newton-Raphson method are shown in Table 1.

7 Monte Carlo Simulations

7.1 Bias-corrected MLE

We examined the performance of MLE estimators and bias-corrected MLE estimators in a sampling experiment. We consider a logit model in which $y_i = 1$ with probability $p_i = e^{\alpha+x_i\beta} \big/ \left(1 + e^{\alpha+x_i\beta}\right)$. For simplicity, we assume that x_i is a scalar random variable with the uniform distribution U[0,1]. We let $\hat{\alpha}$ and $\hat{\beta}$ denote the MLE estimators of α and β, and $\tilde{\alpha}$ and $\tilde{\beta}$ the bias corrected estimators, using the formula discussed in Section 2.

Tables 2, 3, and 4 present the mean biases of these estimators for various combinations of parameters, based on Monte Carlo simulations with 5000 runs. Note that the αs were chosen of so that events are rare. Consistent with the theory, the bias corrected estimators remove most of the bias.

The bias of the MLE increases if the probability of the event decreases, and the bias decreases with the sample size.The mean bias of the bias corrected MLE does not follow the same pattern which confirms that the bias correction removes the systematic bias of the MLE.

7.2 Bias-corrected ATE Estimator

In this section, we report properties of bias corrected estimators of the average treatment effect (ATE). The treatment assignment is assumed to be unconfounded. In our simulations, we generate the conditioning variable x separately for the treated and the controls. The distribution of x is allowed to be different for these two sub-samples. When they are equal, the treatment assignment is independent of x so that the propensity score is constant, and the treatment is randomly assigned. When the distribution of x is different for the treated and controls, the treatment is not randomly assigned.

Total sample size is n; n_1 is the number of $D_i = 1$, the treated, and n_0 is the number of $D_i = 0$, the controls. The number of replications is 1000. We let $\widehat{ATE_1}$ and $\widehat{ATE_2}$ denote the estimator of the ATE based on the MLE $\left(\widehat{\theta}_{(1)}, \widehat{\theta}_{(0)}\right)$, and on the biased corrected MLE, $\left(\widetilde{\theta}_{(1)}, \widetilde{\theta}_{(0)}\right)$, respectively, i.e.,

$$\begin{aligned}
\widehat{ATE_1} &\equiv \frac{1}{n}\sum_{i=1}^{n}\Lambda\left(x_i'\widehat{\theta}_{(1)}\right) - \frac{1}{n}\sum_{i=1}^{n}\Lambda\left(x_i'\widehat{\theta}_{(0)}\right), \qquad (6)\\
\widehat{ATE_2} &\equiv \frac{1}{n}\sum_{i=1}^{n}\Lambda\left(x_i'\widetilde{\theta}_{(1)}\right) - \frac{1}{n}\sum_{i=1}^{n}\Lambda\left(x_i'\widetilde{\theta}_{(0)}\right).
\end{aligned}$$

Let $\widetilde{ATE_1}$ and $\widetilde{ATE_2}$ denote the bias corrected versions of $\widehat{ATE_1}$ and $\widehat{ATE_2}$. The higher order bias of $\widehat{ATE_1}$ can be removed by using (32) in Appendix C. Likewise, the higher order bias of $\widehat{ATE_2}$ can be removed, noting that the first two terms in (32) can be ignored because $\left(\widetilde{\theta}_{(1)}, \widetilde{\theta}_{(0)}\right)$ already are bias-corrected.

In Tables 5 - 11, we report the mean bias of these estimators. The mean biases over the Monte Carlo replications are calculated as the averages of $\widehat{ATE_1} - ATE_0$, $\widehat{ATE_2} - ATE_0$, $\widetilde{ATE_1} - ATE_0$, $\widetilde{ATE_2} - ATE_0$, where

$$ATE_0 \equiv \frac{1}{n}\sum_{i=1}^{n}\Lambda\left(x_i'\theta_{0,(1)}\right) - \frac{1}{n}\sum_{i=1}^{n}\Lambda\left(x_i'\theta_{0,(0)}\right).$$

We will discuss the Monte Carlo results separately for the “random assignment” case, where the distribution of x is identical across the treatment and control subsamples, and for the “nonrandom assignment”, where the distribution of x is different across the treatment and control subsamples.

We adopt a different notation for the parameters of the model:

$$\theta_{0,(0)} = (\alpha_0 \ \beta_0)' \quad \theta_{0,(1)} = (\alpha_1 \ \beta_1)'$$

with α_0, α_1 the intercepts and β_0, β_1 the slope coefficients in the logit outcome models for the treated and controls.

7.2.1 Bias-corrected ATE - Constant Propensity Score

As was discussed in Remark 1, the second order bias of the ATE is zero when the propensity score is constant. In order to verify this result, we will first consider the case that the distributions of x are identical over the $D = 1$ and $D = 0$ subsamples. In Tables 5 - 7, we evaluate the performance of various estimators of the ATE under random assignment. Overall, we see that the original ATE estimator $\widehat{ATE_1}$ is largely free of bias, consistent with Remark 1. All estimators are unbiased even in the rare event case and if the treatment assignment is unbalanced. The bias also does not depend on the sample size and the event probability.

In Table 5, $\beta_1 = 2$ and $\alpha_1 = \alpha_0 + 1$ for $D_i = 1$. For $D_i = 0$, we let $\beta_0 = 1$, and we consider the different values of α_0 listed in Table 5. In Table 6, the treatment effect is larger, $\beta_1 = 4$. The rest of the DGP is unchanged. The conclusions are the same as for Table 5. In Table 7, we let $\beta_1 = \beta_0 = 1$, and we consider different values of $\alpha_1 = \alpha_0$ listed in Table 7. For each parameter value the ATE is 0. The rest of the DGP is unchanged. The conclusions are the same as for Table 5.

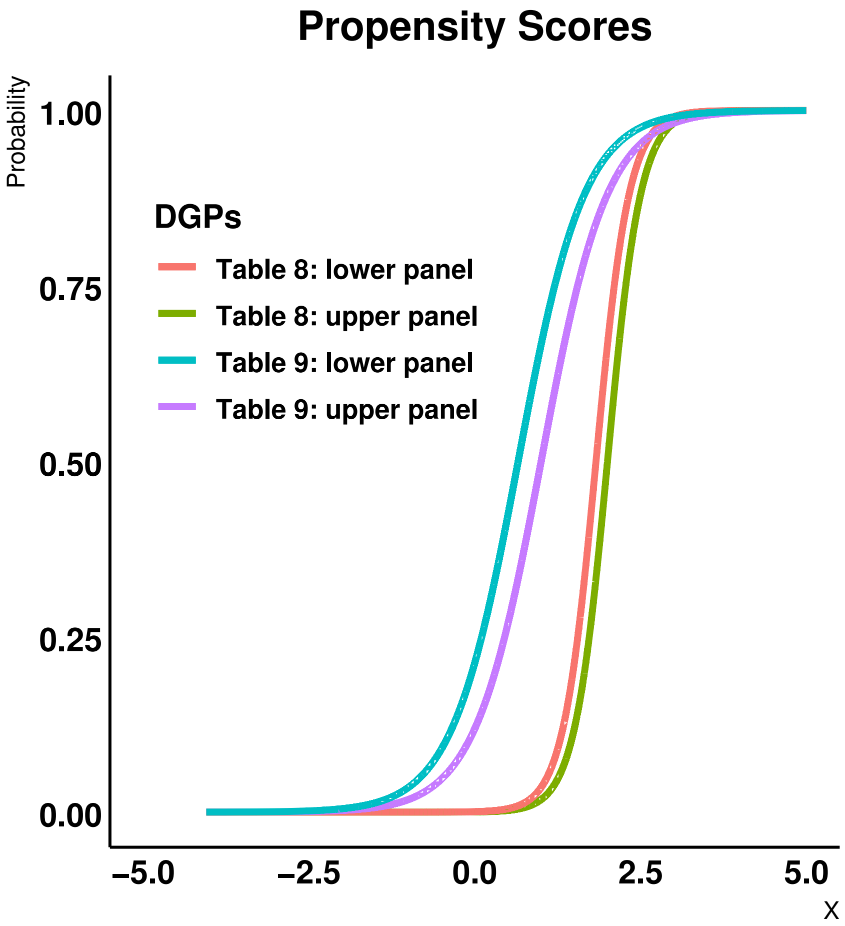
7.2.2 Bias-corrected ATE - General Propensity Score

As was discussed in Remark 1, the second order bias of the ATE is in general not equal to zero when the propensity score is not constant, and therefore, the distribution of x is different for the treated and the controls. In Tables 8 - 11, we consider such "nonrandom assignment". The discussion in Remark 1 suggests that the bias is large if there is a stark difference between the distribution of x in the two subsamples. Table 8 is meant to represent such a situation: for $D_i = 1$, $X \sim N[4, 1]$, and for $D_i = 0$, $X \sim N[0, 1]$. Figure 1 shows the implied propensity scores for the DGPs considered in Table 8 and Table 9. We see a modest bias for the selected parameter values. In any case, the bias corrected estimator removes most of the bias. The estimator based on the bias-corrected MLE is as biased as that based on the non-corrected MLE. The bias induced by the curvature of Λ is relatively large.

Based on Remark 1, we can speculate that if the difference of the distributions is not as stark, we expect the ATE to have a smaller bias. In order to verify this conjecture, we consider in Table 9 the case where the X is distributed as $N[2, 1]$ and $N[0, 1]$ in the treated and control subsamples, respectively. In general, the bias in $\widehat{ATE_1}$ is a lot smaller than in Table 8.

The bias formula (18) shows that it is an average of (16) and (17) weighted by the density of X. If the θs are identical across the two subsamples, then the counterparts of (16) and (17) are also identical across the two subsamples. Therefore, the bias of $\widehat{\theta}_{(1)}$ and $\widehat{\theta}_{(0)}$ may be similar, with the difference only arising from the possible difference of "weights". Therefore, one may think that the biases of the two terms on the right of (6) may almost cancel each other out when the θs are similar. In order to examine this conjecture, we considered cases

Figure 1: Propensity Scores



where the θs are identical across the two subsamples in Tables 10 and 11. In Table 10, we consider nonrandom assignment, for $D_i = 1$: $X \sim N[4, 1]$; for $D_i = 0$: $X \sim N[0, 1]$. For both $D_i = 1$ and $D_i = 0$: $\beta_0 = \beta_1 = 1$, $\alpha_0 = \alpha_1$ is listed in the table. In Table 11, we consider nonrandom assignment and same parameters of interests, for $D_i = 1$: $X \sim N[2, 1]$; for $D_i = 0$: $X \sim N[0, 1]$. For both $D_i = 1$ and $D_i = 0$: $\beta_0 = \beta_1 = 1$, $\alpha_0 = \alpha_1$ is listed in the table. We conclude that the intuition that the biases cancel is not correct.

8 Conclusion

If the treatment assignment is unconfounded, then the MLE of the ATE in the logit model for binary outcomes is biased both because the MLE of the parameters of the logit model are biased and the curvature of the logit model contributes to the bias of the ATE estimator. The bias is larger if the event has a low probability, the case we focus on in this paper.

The logit ATE estimator is unbiased if the treatment assignment is random. This is obvious if there are no covariates, but it is also true with covariates.

We derive the second order bias of the logit ATE estimator. We also propose bias-corrected estimators of the ATE.

Simulation experiments show that it is not sufficient to bias correct the MLE of the logit parameters, that the bias of the logit ATE estimator is moderately large, and that the bias-corrected estimators remove most of the bias.

Finally we propose a variation on the logit model with parameters that are elasticities. We also propose a computational trick that avoids numerical instability in the case of estimation for rare events.

References

[1] Rilstone, P., V.K. Srivastava, and A. Ullah (1996): "The Second-Order Bias and Mean Squared Error of Nonlinear Estimators", Journal of Econometrics 75, 369 - 395.

[2] Hahn, J., and W. Newey (2004): "Jackknife and Analytical Bias Reduction for Nonlinear Panel Models", Econometrica 72, 1295-1319.

[3] King, G. and L. Zeng (2001): "Logistic regression in rare events data", Political analysis 9, 137–163.

[4] McCullagh, P., and J. A. Nelder (1989): Generalized Linear Models, 2nd ed. New York: Chapman and Hall.

[5] Wang, H.(2020): "Logistic regression for massive data with rare events," in H. D. III and A. Singh,eds., Proceedings of the 37th International Conference on Machine Learning, vol. 119 of Proceedings of Machine Learning Research, 9829–9836.

Appendix

A Second Order Bias of Logit MLE

We derive the second order bias of the logit MLE. For notational simplicity, we will omit the i subscript whenever obvious.

A.1 Second Order Expansion of Generic MLE

The MLE $\widehat{\theta}$ maximizes the joint log likelihood $\sum_{i=1}^{n} \log f\left(z_i; c\right)$, and satisfies the first order condition $0 = \frac{1}{n}\sum_{i=1}^{n} v\left(z_i; \widehat{\theta}\right)$. Let F denote the collection of (marginal) distribution functions of z. Let $\hat{F}$ denote the empirical distribution function. Define $F\left(\epsilon\right) \equiv F + \epsilon\sqrt{n}\left(\hat{F} - F\right)$ for $\epsilon \in \left[0, n^{-1/2}\right]$. For each fixed ϵ, let $\theta\left(\epsilon\right)$ be the solution to the estimating equation

$$0 = \int v\left[\cdot; \theta\left(\epsilon\right)\right] dF\left(\epsilon\right). \tag{7}$$

Note that (7) is equivalent to $0 = \frac{1}{n}\sum_{i=1}^{n} v\left(z_i; \theta\left(n^{-1/2}\right)\right)$ when evaluated at $\epsilon = n^{-1/2}$, so we can see that $\theta\left(n^{-1/2}\right) = \widehat{\theta}$. By a Taylor series expansion, we have

$$\widehat{\theta} - \theta_0 = \theta\left(\frac{1}{\sqrt{n}}\right) - \theta\left(0\right) = \frac{1}{\sqrt{n}}\theta^{\epsilon}\left(0\right) + \frac{1}{2}\left(\frac{1}{\sqrt{n}}\right)^2 \theta^{\epsilon\epsilon}\left(0\right) + \frac{1}{6}\left(\frac{1}{\sqrt{n}}\right)^3 \theta^{\epsilon\epsilon\epsilon}\left(\tilde{\epsilon}\right), \tag{8}$$

where $\theta^{\epsilon}\left(\epsilon\right) \equiv d\theta\left(\epsilon\right)/d\epsilon$, $\theta^{\epsilon\epsilon}\left(\epsilon\right) \equiv d^2\theta\left(\epsilon\right)/d\epsilon^2$, ..., and $\tilde{\epsilon}$ is somewhere in between 0 and $n^{-1/2}$.

Let

$$h\left(\cdot, \epsilon\right) \equiv v\left[\cdot; \theta\left(\epsilon\right)\right] \tag{9}$$

The first order condition may be written as

$$0 = \int h\left(\cdot, \epsilon\right) dF\left(\epsilon\right) \tag{10}$$

Differentiating repeatedly with respect to ϵ, we obtain

$$0 = \int \frac{dh(\cdot,\epsilon)}{d\epsilon} dF(\epsilon) + \int h(\cdot,\epsilon)\, d\Delta_n \tag{11}$$

$$0 = \int \frac{d^2h(\cdot,\epsilon)}{d\epsilon^2} dF(\epsilon) + 2\int \frac{dh(\cdot,\epsilon)}{d\epsilon} d\Delta_n \tag{12}$$

$$0 = \int \frac{d^3h(\cdot,\epsilon)}{d\epsilon^3} dF(\epsilon) + 3\int \frac{d^2h(\cdot,\epsilon)}{d\epsilon^2} d\Delta_n \tag{13}$$

where $\Delta_n \equiv \sqrt{n}\left(\hat{F} - F\right)$. We will ignore the third order term, which can be justified under the type of regularity conditions discussed in Hahn and Newey (2004), and find the analytic expression for the second order bias.

Rewrite (11) as

$$0 = \left(\int v^{\theta}\left[\cdot;\theta(\epsilon)\right] dF(\epsilon)\right)\theta^{\epsilon}(\epsilon) + \int v\left[\cdot;\theta(\epsilon)\right] d\Delta_n,$$

where

$$v^{\theta} \equiv \frac{\partial v\left[\cdot;\theta(\epsilon)\right]}{\partial\theta'}.$$

Evaluating it at $\epsilon = 0$, and noting that $E\left[v_i\right] = 0$, we obtain

$$0 = \left(\int v^{\theta}\left[\cdot;\theta(0)\right] dF\right)\theta^{\epsilon}(0) + \int v\left[\cdot;\theta(0)\right]\Delta_n,$$

so

$$\theta^{\epsilon}(0) = -\left(E\left[v^{\theta}\right]\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} v_i\right) \tag{14}$$

Recall that $\dim(\theta) = K$, and write

$$v\left[\cdot;\theta(\epsilon)\right] = \begin{bmatrix} v_1\left[\cdot;\theta(\epsilon)\right] \\ \vdots \\ v_K\left[\cdot;\theta(\epsilon)\right] \end{bmatrix}$$

We can then write

$$\frac{d^2 h\left(\cdot,\epsilon\right)}{d\epsilon^2} = \begin{bmatrix} d^2 h_1\left(\cdot,\epsilon\right)/\,d\epsilon^2 \\ \vdots \\ d^2 h_K\left(\cdot,\epsilon\right)/\,d\epsilon^2 \end{bmatrix} = \begin{bmatrix} \theta^\epsilon\left(\epsilon\right)' \frac{\partial^2 v_1[\cdot;\theta(\epsilon)]}{\partial\theta\partial\theta'}\theta^\epsilon\left(\epsilon\right) + \frac{\partial v_1[\cdot;\theta(\epsilon)]}{\partial\theta'}\theta^{\epsilon\epsilon}\left(\epsilon\right) \\ \vdots \\ \theta^\epsilon\left(\epsilon\right)' \frac{\partial^2 v_K[\cdot;\theta(\epsilon)]}{\partial\theta\partial\theta'}\theta^\epsilon\left(\epsilon\right) + \frac{\partial v_K[\cdot;\theta(\epsilon)]}{\partial\theta'}\theta^{\epsilon\epsilon}\left(\epsilon\right) \end{bmatrix}$$

so we can rewrite (12) as

$$0 = \begin{bmatrix} \theta^\epsilon\left(\epsilon\right)'\left(\int \frac{\partial^2 v_1[\cdot;\theta(\epsilon)]}{\partial\theta\partial\theta'}dF\left(\epsilon\right)\right)\theta^\epsilon\left(\epsilon\right) \\ \vdots \\ \theta^\epsilon\left(\epsilon\right)'\left(\int \frac{\partial^2 v_K[\cdot;\theta(\epsilon)]}{\partial\theta\partial\theta'}dF\left(\epsilon\right)\right)\theta^\epsilon\left(\epsilon\right) \end{bmatrix} + \begin{bmatrix} \int \frac{\partial v_1[\cdot;\theta(\epsilon)]}{\partial\theta'}dF\left(\epsilon\right) \\ \vdots \\ \int \frac{\partial v_K[\cdot;\theta(\epsilon)]}{\partial\theta'}dF\left(\epsilon\right) \end{bmatrix}\theta^{\epsilon\epsilon}\left(\epsilon\right) + 2\begin{bmatrix} \int \frac{\partial v_1[\cdot;\theta(\epsilon)]}{\partial\theta'}d\Delta_n \\ \vdots \\ \int \frac{\partial v_K[\cdot;\theta(\epsilon)]}{\partial\theta'}d\Delta_n \end{bmatrix}\theta^\epsilon\left(\epsilon\right)$$

Evaluating it at $\epsilon = 0$, we obtain

$$0 = \begin{bmatrix} \theta^\epsilon\left(0\right)'\left(E\left[\frac{\partial^2 v_1}{\partial\theta\partial\theta'}\right]\right)\theta^\epsilon\left(0\right) \\ \vdots \\ \theta^\epsilon\left(0\right)'\left(E\left[\frac{\partial^2 v_K}{\partial\theta\partial\theta'}\right]\right)\theta^\epsilon\left(0\right) \end{bmatrix} + \begin{bmatrix} E\left[\frac{\partial v_1}{\partial\theta'}\right] \\ \vdots \\ E\left[\frac{\partial v_K}{\partial\theta'}\right] \end{bmatrix}\theta^{\epsilon\epsilon}\left(0\right) + 2\begin{bmatrix} \int \frac{\partial v_1[\cdot;\theta(\epsilon)]}{\partial\theta'}d\Delta_n \\ \vdots \\ \int \frac{\partial v_K[\cdot;\theta(\epsilon)]}{\partial\theta'}d\Delta_n \end{bmatrix}\theta^\epsilon\left(0\right)$$

so

$$\frac{1}{2}\theta^{\epsilon\epsilon}\left(0\right) = -\frac{1}{2}\left(E\left[v^\theta\right]\right)^{-1}\begin{bmatrix} \theta^\epsilon\left(0\right)'\left(E\left[\frac{\partial^2 v_1}{\partial\theta\partial\theta'}\right]\right)\theta^\epsilon\left(0\right) \\ \vdots \\ \theta^\epsilon\left(0\right)'\left(E\left[\frac{\partial^2 v_K}{\partial\theta\partial\theta'}\right]\right)\theta^\epsilon\left(0\right) \end{bmatrix} - \left(E\left[v^\theta\right]\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(v_i^\theta - E\left[v_i^\theta\right]\right)\right)\theta^\epsilon\left(0\right)$$

A.2 Second order bias

We now calculate $E\left[\frac{1}{2}\theta^{\epsilon\epsilon}\left(0\right)\right]$. We first compute the expectation of

$$E\left[v^\theta\right]\left(\frac{1}{2}\theta^{\epsilon\epsilon}\left(0\right)\right) = -\frac{1}{2}\begin{bmatrix} \theta^\epsilon\left(0\right)'\left(E\left[\frac{\partial^2 v_1}{\partial\theta\partial\theta'}\right]\right)\theta^\epsilon\left(0\right) \\ \vdots \\ \theta^\epsilon\left(0\right)'\left(E\left[\frac{\partial^2 v_K}{\partial\theta\partial\theta'}\right]\right)\theta^\epsilon\left(0\right) \end{bmatrix} - \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(v_i^\theta - E\left[v_i^\theta\right]\right)\right)\theta^\epsilon\left(0\right)$$

The kth component of the first term has an expectation equal to

$$\begin{aligned}
-\frac{1}{2}E\left[\theta^{\epsilon}(0)'\left(E\left[\frac{\partial^2 v_k}{\partial\theta\partial\theta'}\right]\right)\theta^{\epsilon}(0)\right] &= -\frac{1}{2}E\left[\operatorname{trace}\left(\theta^{\epsilon}(0)'\left(E\left[\frac{\partial^2 v_k}{\partial\theta\partial\theta'}\right]\right)\theta^{\epsilon}(0)\right)\right] \\
&= -\frac{1}{2}E\left[\operatorname{trace}\left(E\left[\frac{\partial^2 v_k}{\partial\theta\partial\theta'}\right]\theta^{\epsilon}(0)\,\theta^{\epsilon}(0)'\right)\right] \\
&= -\frac{1}{2}\operatorname{trace}\left(E\left[\frac{\partial^2 v_k}{\partial\theta\partial\theta'}\right]E\left[\theta^{\epsilon}(0)\,\theta^{\epsilon}(0)'\right]\right)
\end{aligned}$$

Because

$$\begin{aligned}
E\left[\theta^{\epsilon}(0)\,\theta^{\epsilon}(0)'\right] &= E\left[\left(-\left(E\left[v^{\theta}\right]\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}v_i\right)\right)\left(-\left(E\left[v^{\theta}\right]\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}v_i\right)\right)'\right] \\
&= \left(E\left[v^{\theta}\right]\right)^{-1}E\left[vv'\right]\left(E\left[v^{\theta}\right]\right)^{-1} \\
&= -\left(E\left[v^{\theta}\right]\right)^{-1}
\end{aligned}$$

where we used the information equality, we can write

$$-\frac{1}{2}E\left[\theta^{\epsilon}(0)'\left(E\left[\frac{\partial^2 v_k}{\partial\theta\partial\theta'}\right]\right)\theta^{\epsilon}(0)\right] = \frac{1}{2}\operatorname{trace}\left(E\left[\frac{\partial^2 v_k}{\partial\theta\partial\theta'}\right]\left(E\left[v^{\theta}\right]\right)^{-1}\right).$$

The kth component of the second term has an expectation equal to

$$\begin{aligned}
&-E\left[\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(v_{k,i}^{\theta}-E\left[v_{k,i}^{\theta}\right]\right)\right)\theta^{\epsilon}(0)\right] \\
&= -E\left[\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(v_{k,i}^{\theta}-E\left[v_{k,i}^{\theta}\right]\right)\right)\left(-\left(E\left[v^{\theta}\right]\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}v_i\right)\right)\right] \\
&= E\left[\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{\partial v_{k,i}}{\partial\theta'}-E\left[\frac{\partial v_{k,i}}{\partial\theta'}\right]\right)\right)\left(\left(E\left[v^{\theta}\right]\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}v_i\right)\right)\right] \\
&= E\left[\operatorname{trace}\left\{\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{\partial v_{k,i}}{\partial\theta'}-E\left[\frac{\partial v_{k,i}}{\partial\theta'}\right]\right)\right)\left(\left(E\left[v^{\theta}\right]\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}v_i\right)\right)\right\}\right] \\
&= E\left[\operatorname{trace}\left\{\left(\left(E\left[v^{\theta}\right]\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}v_i\right)\right)\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{\partial v_{k,i}}{\partial\theta'}-E\left[\frac{\partial v_{k,i}}{\partial\theta'}\right]\right)\right)\right\}\right] \\
&= \operatorname{trace}\left(\left(E\left[v^{\theta}\right]\right)^{-1}E\left[v_i\left(\frac{\partial v_{k,i}}{\partial\theta'}-E\left[\frac{\partial v_{k,i}}{\partial\theta'}\right]\right)\right]\right) \\
&= \operatorname{trace}\left(E\left[v\left(\frac{\partial v_k}{\partial\theta'}-E\left[\frac{\partial v_k}{\partial\theta'}\right]\right)\right]\left(E\left[v^{\theta}\right]\right)^{-1}\right).
\end{aligned}$$

Therefore, the kth component of $E\left[v^{\theta}\right] E\left[\left(\frac{1}{2}\theta^{\epsilon\epsilon}(0)\right)\right]$ is equal to

$$\begin{aligned}
&\frac{1}{2}\operatorname{trace}\left(E\left[\frac{\partial^2 v_k}{\partial\theta\partial\theta'}\right]\left(E\left[v^{\theta}\right]\right)^{-1}\right)+\operatorname{trace}\left(E\left[v\left(\frac{\partial v_k}{\partial\theta'}-E\left[\frac{\partial v_k}{\partial\theta'}\right]\right)\right]\left(E\left[v^{\theta}\right]\right)^{-1}\right)\\
&=\operatorname{trace}\left(\left(\frac{1}{2}E\left[\frac{\partial^2 v_k}{\partial\theta\partial\theta'}\right]+E\left[v\left(\frac{\partial v_k}{\partial\theta'}-E\left[\frac{\partial v_k}{\partial\theta'}\right]\right)\right]\right)\left(E\left[v^{\theta}\right]\right)^{-1}\right) \qquad (15)
\end{aligned}$$

A.3 Getting back to Logit model

We now consider the general logit model. We assume that y is equal to 1 with probability equal to

$$\Lambda\left(x'\theta\right)=\frac{\exp\left(x'\theta\right)}{1+\exp\left(x'\theta\right)},$$

where x is the vector of regressors that may include the intercept term. The log likelihood is then given by

$$\log f = y\log\Lambda\left(x'\theta\right)+(1-y)\log\left(1-\Lambda\left(x'\theta\right)\right)$$

Using that $\frac{d}{dt}\Lambda(t)=\Lambda(t)\left(1-\Lambda(t)\right)$, we obtain

$$\begin{aligned}
v &= \frac{\partial\log f}{\partial\theta}=\left(\frac{y}{\Lambda}-\frac{1-y}{1-\Lambda}\right)\Lambda\left(1-\Lambda\right)x=\left(y-\Lambda\right)x\\
v^{\theta} &= \frac{\partial^2\log f}{\partial\theta\partial\theta'}=-\Lambda\left(1-\Lambda\right)xx'\\
\frac{\partial v_k}{\partial\theta'} &= \frac{\partial\left(\left(y-\Lambda\right)x_k\right)}{\partial\theta'}=-\Lambda\left(1-\Lambda\right)x_k x' \qquad (16)\\
\frac{\partial^2 v_k}{\partial\theta\partial\theta'} &= \frac{\partial^2\left(\left(y-\Lambda\right)x_k\right)}{\partial\theta\partial\theta'}=-\Lambda\left(1-\Lambda\right)\left(1-2\Lambda\right)x_k xx' \qquad (17)
\end{aligned}$$

Note that

$$E\left[vv_k^{\theta}\right]=E\left[\left(\left(y-\Lambda\right)x\right)\left(-\Lambda\left(1-\Lambda\right)x_k x'\right)\right]=0$$

because the conditional expectation of y given x is equal to Λ. It follows that $E\left[v\left(\frac{\partial v_k}{\partial\theta'}-E\left[\frac{\partial v_k}{\partial\theta'}\right]\right)\right]=0$ and the second order bias (15) simplifies; for the logit model, the kth component of

$E\left[v^{\theta}\right] E\left[\left(\frac{1}{2}\theta^{\epsilon\epsilon}(0)\right)\right]$ is equal to

$$\frac{1}{2}\operatorname{trace}\left(E\left[\frac{\partial^2 v_k}{\partial\theta\partial\theta'}\right]\left(E\left[v^{\theta}\right]\right)^{-1}\right). \tag{18}$$

Therefore, the bias correction would take the form of

1. Calculate the MLE $\widehat{\theta}$

2. Let
$$\hat{\Lambda}_i = \frac{\exp\left(x_i'\widehat{\theta}\right)}{1+\exp\left(x_i'\widehat{\theta}\right)},$$

3. Let
$$A \equiv \widehat{E}\left[v^{\theta}\right] = -\frac{1}{n}\sum_{i=1}^{n}\hat{\Lambda}_i\left(1-\hat{\Lambda}_i\right)x_i x_i' \tag{19}$$

4. Let
$$C_k \equiv \widehat{E}\left[\frac{\partial^2 v_k}{\partial\theta\partial\theta'}\right] = -\frac{1}{n}\sum_{i=1}^{n}\hat{\Lambda}_i\left(1-\hat{\Lambda}_i\right)\left(1-2\hat{\Lambda}_i\right)x_{i,k}x_i x_i' \tag{20}$$

5. Let
$$T_k = \frac{1}{2}\operatorname{trace}\left(C_k A^{-1}\right) = \frac{1}{2}\operatorname{trace}\left(\widehat{E}\left[\frac{\partial^2 v_k}{\partial\theta\partial\theta'}\right]\left(\widehat{E}\left[v^{\theta}\right]\right)^{-1}\right) \tag{21}$$

6. Let
$$B = (A)^{-1}\begin{bmatrix} T_1 \\ \vdots \\ T_k \end{bmatrix} = \left(\widehat{E}\left[v^{\theta}\right]\right)^{-1}\begin{bmatrix} T_1 \\ \vdots \\ T_k \end{bmatrix} \tag{22}$$

7. Let $\widetilde{\theta} = \widehat{\theta} - B/n$

B Comparison with King and Zeng's (2001) Bias Formula

In terms of comparison with King and Zeng's (11), we note that their bias formula can be rewritten $\left(\frac{1}{n}X'WX\right)^{-1}\left(\frac{1}{n}X'W\xi\right)$, where W in our context is equal to a diagonal matrix with diagonal elements equal to $\hat{\Lambda}_i\left(1-\hat{\Lambda}_i\right)$, $\xi_i = \frac{Q_{ii}}{2}\left(2\hat{\Lambda}_i - 1\right)$, and $Q = X\left(X'WX\right)^{-1}X'$. Note first that

$$\frac{1}{n}X'WX = \frac{1}{n}\sum_{i=1}^{n} x_i\hat{\Lambda}_i\left(1-\hat{\Lambda}_i\right)x_i' = -A, \tag{23}$$

where A is from (19). Also note that $Q = \frac{1}{n}X\left(\frac{1}{n}X'WX\right)^{-1}X' = -\frac{1}{n}XA^{-1}X'$, so we have $Q_{ii} = -\frac{1}{n}x_i'Ax_i = -\frac{1}{n}\operatorname{trace}\left(x_ix_i'A\right)$. It follows that the k-th component of $\frac{1}{n}X'W\xi$ is equal to

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} x_{i,k}\hat{\Lambda}_i\left(1-\hat{\Lambda}_i\right)\xi_i \\
&= \frac{1}{n}\sum_{i=1}^{n} x_{i,k}\hat{\Lambda}_i\left(1-\hat{\Lambda}_i\right)\frac{1}{2}\left(-\frac{1}{n}\operatorname{trace}\left(Ax_ix_i'\right)\right)\left(2\hat{\Lambda}_i-1\right) \\
&= \frac{1}{2n}\frac{1}{n}\sum_{i=1}^{n}\hat{\Lambda}_i\left(1-\hat{\Lambda}_i\right)\left(1-2\hat{\Lambda}_i\right)\operatorname{trace}\left(x_ix_i'A\right) \\
&= \frac{1}{2n}\operatorname{trace}\left(\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\Lambda}_i\left(1-\hat{\Lambda}_i\right)\left(1-2\hat{\Lambda}_i\right)x_ix_i'\right)A\right) \\
&= -\frac{1}{2n}\operatorname{trace}\left(C_kA\right) = -\frac{1}{n}T_k,
\end{aligned}$$

where C_k and T_k are from (20) and (21). Therefore, we can understand that

$$\frac{1}{n}X'W\xi = -\frac{1}{n}\begin{bmatrix} T_1 \\ \vdots \\ T_k \end{bmatrix} \tag{24}$$

Combining (23) and (24), we obtain

$$\left(\frac{1}{n}X'WX\right)^{-1}\left(\frac{1}{n}X'W\xi\right) = (-A)^{-1}\left(-\frac{1}{n}\begin{bmatrix} T_1 \\ \vdots \\ T_k \end{bmatrix}\right) = B$$

for B in (22).

C Details of Estimating the Bias of the ATE

The second order expansion indicates that the second order bias through $\widehat{\theta}_{(1)}$ is equal to

$$\begin{aligned}
&\frac{1}{2n_1}\left(\frac{1}{n}\sum_{i=1}^{n}\Lambda\left(x_i'\theta_{0,(1)}\right)\left(1-\Lambda\left(x_i'\theta_{0,(1)}\right)\right)x_i'\right)E_{(1)}\left[\theta^{\epsilon\epsilon}(0)\right] \\
&-\frac{1}{2n_1}\operatorname{trace}\left\{\left(\frac{1}{n}\sum_{i=1}^{n}\Lambda\left(x_i'\theta_{0,(1)}\right)\left(1-\Lambda\left(x_i'\theta_{0,(1)}\right)\right)\left(1-2\Lambda\left(x_i'\theta_{0,(1)}\right)\right)x_ix_i'\right)\left(E_{(1)}\left[v^{\theta}\right]\right)^{-1}\right\}
\end{aligned}$$

where $E_{(1)}\left[\theta^{\epsilon\epsilon}(0)\right]$ and $E_{(1)}\left[v^{\theta}\right]$ denote the counterparts of $E\left[\theta^{\epsilon\epsilon}(0)\right]$ and $E\left[v^{\theta}\right]$ for $\widehat{\theta}_{(1)}$. Because of the similar reasoning applied to $\widehat{\theta}_{(0)}$, we can conclude that the second order bias of the ATE is given by

$$\begin{aligned}
&\frac{1}{2n_1}\left(\frac{1}{n}\sum_{i=1}^{n}\Lambda\left(x_i'\theta_{0,(1)}\right)\left(1-\Lambda\left(x_i'\theta_{0,(1)}\right)\right)x_i'\right)E_{(1)}\left[\theta^{\epsilon\epsilon}(0)\right] \\
&-\frac{1}{2n_0}\left(\frac{1}{n}\sum_{i=1}^{n}\Lambda\left(x_i'\theta_{0,(0)}\right)\left(1-\Lambda\left(x_i'\theta_{0,(0)}\right)\right)x_i'\right)E_{(0)}\left[\theta^{\epsilon\epsilon}(0)\right] \\
&-\frac{1}{2n_1}\operatorname{trace}\left\{\left(\frac{1}{n}\sum_{i=1}^{n}\Lambda\left(x_i'\theta_{0,(1)}\right)\left(1-\Lambda\left(x_i'\theta_{0,(1)}\right)\right)\left(1-2\Lambda\left(x_i'\theta_{0,(1)}\right)\right)x_ix_i'\right)\left(E_{(1)}\left[v^{\theta}\right]\right)^{-1}\right\} \\
&+\frac{1}{2n_0}\operatorname{trace}\left\{\left(\frac{1}{n}\sum_{i=1}^{n}\Lambda\left(x_i'\theta_{0,(0)}\right)\left(1-\Lambda\left(x_i'\theta_{0,(0)}\right)\right)\left(1-2\Lambda\left(x_i'\theta_{0,(0)}\right)\right)x_ix_i'\right)\left(E_{(0)}\left[v^{\theta}\right]\right)^{-1}\right\}
\end{aligned} \tag{25}$$

which can be estimated in the standard way.

1. Suppose that there are n_1 observations such that $D = 1$. We estimate $\theta_{0,(1)}$ by MLE $\widehat{\theta}_{(1)}$ from this sample. Our preceding discussion implies that the counterparts of $\widehat{E}\left[\theta^{\epsilon\epsilon}(0)\right]$ and $\widehat{E}\left[v^{\theta}\right]$, which we will denote as $\widehat{E}_{(1)}\left[\theta^{\epsilon\epsilon}(0)\right]$ and $\widehat{E}_{(1)}\left[v^{\theta}\right]$ can be characterized by the following steps:

   (a) Calculate the MLE $\widehat{\theta}_{(1)}$

   (b) Let
$$\hat{\Lambda}_{i,(1)} = \frac{\exp\left(x_i'\widehat{\theta}_{(1)}\right)}{1+\exp\left(x_i'\widehat{\theta}_{(1)}\right)}. \tag{26}$$

   (c) Let
$$\widehat{E}_{(1)}\left[v^{\theta}\right] \equiv A_{(1)} = -\frac{1}{n_1}\sum_{D=1}\hat{\Lambda}_{i,(1)}\left(1-\hat{\Lambda}_{i,(1)}\right)x_i x_i'. \tag{27}$$

   (d) Let
$$C_{k,(1)} \equiv \widehat{E}_{(1)}\left[\frac{\partial^2 v_k}{\partial\theta\partial\theta'}\right] = -\frac{1}{n_1}\sum_{D=1}\hat{\Lambda}_{i,(1)}\left(1-\hat{\Lambda}_{i,(1)}\right)\left(1-2\hat{\Lambda}_{i,(1)}\right)x_{i,k}x_i x_i'.$$

   (e) Let
$$T_{k,(1)} = \frac{1}{2}\operatorname{trace}\left(C_{k,(1)}A_{(1)}^{-1}\right).$$

   (f) Let
$$\widehat{E}_{(1)}\left[\theta^{\epsilon\epsilon}(0)\right] \equiv \left(A_{(1)}\right)^{-1}\begin{bmatrix} T_{1,(1)} \\ \vdots \\ T_{k,(1)} \end{bmatrix}. \tag{28}$$

2. Likewise, we calculate the $\widehat{E}_{(0)}\left[\theta^{\epsilon\epsilon}(0)\right]$ and $\widehat{E}_{(0)}\left[v^{\theta}\right]$:

   (a) Calculate the MLE $\widehat{\theta}_{(0)}$

(b) Let

$$\hat{\Lambda}_{i,(0)} \equiv \frac{\exp\left(x_i'\widehat{\theta}_{(0)}\right)}{1+\exp\left(x_i'\widehat{\theta}_{(0)}\right)}. \tag{29}$$

(c) Let

$$\widehat{E}_{(0)}\left[v^{\theta}\right] \equiv A_{(0)} = -\frac{1}{n_0}\sum_{D=0}\hat{\Lambda}_{i,(0)}\left(1-\hat{\Lambda}_{i,(0)}\right)x_i x_i'. \tag{30}$$

(d) Let

$$C_{k,(0)} \equiv \widehat{E}_{(0)}\left[\frac{\partial^2 v_k}{\partial\theta\partial\theta'}\right] = -\frac{1}{n_0}\sum_{D=0}\hat{\Lambda}_{i,(0)}\left(1-\hat{\Lambda}_{i,(0)}\right)\left(1-2\hat{\Lambda}_{i,(0)}\right)x_{i,k}x_i x_i'.$$

(e) Let

$$T_{k,(0)} = \frac{1}{2}\operatorname{trace}\left(C_{k,(0)}A_{(0)}^{-1}\right).$$

(f) Let

$$\widehat{E}_{(0)}\left[\theta^{\epsilon\epsilon}(0)\right] \equiv \left(A_{(0)}\right)^{-1}\begin{bmatrix} T_{1,(0)} \\ \vdots \\ T_{k,(0)} \end{bmatrix}. \tag{31}$$

3. The second order bias is computed to be

$$\begin{aligned}\frac{B_{ATE}}{n} \equiv & \frac{1}{2n_1}\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\Lambda}_{i,(1)}\left(1-\hat{\Lambda}_{i,(1)}\right)x_i'\right)\widehat{E}_{(1)}\left[\theta^{\epsilon\epsilon}(0)\right] \\ & -\frac{1}{2n_0}\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\Lambda}_{i,(0)}\left(1-\hat{\Lambda}_{i,(0)}\right)x_i'\right)\widehat{E}_{(0)}\left[\theta^{\epsilon\epsilon}(0)\right] \\ & -\frac{1}{2n_1}\operatorname{trace}\left\{\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\Lambda}_{i,(1)}\left(1-\hat{\Lambda}_{i,(1)}\right)\left(1-2\hat{\Lambda}_{i,(1)}\right)x_i x_i'\right)\left(\widehat{E}_{(1)}\left[v^{\theta}\right]\right)^{-1}\right\} \\ & +\frac{1}{2n_0}\operatorname{trace}\left\{\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\Lambda}_{i,(0)}\left(1-\hat{\Lambda}_{i,(0)}\right)\left(1-2\hat{\Lambda}_{i,(0)}\right)x_i x_i'\right)\left(\widehat{E}_{(0)}\left[v^{\theta}\right]\right)^{-1}\right\}\end{aligned} \tag{32}$$

using (26), (27), (28), (29), (30), and (31) computed in previous steps.

D Zero Second Order Bias of the Average Predicted Probability

Similar calculation as in the previous section indicates that the second order bias of is $\frac{1}{n}\sum_{i=1}^{n}\Lambda\left(x_i'\widehat{\theta}\right)$ can be estimated by

$$\frac{1}{2n}\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\Lambda}_i\left(1-\hat{\Lambda}_i\right)x_i'\right)\widehat{E}\left[\theta^{\epsilon\epsilon}\left(0\right)\right]$$
$$-\frac{1}{2n}\operatorname{trace}\left\{\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\Lambda}_i\left(1-\hat{\Lambda}_i\right)\left(1-2\hat{\Lambda}_i\right)x_ix_i'\right)\left(\widehat{E}\left[v^{\theta}\right]\right)^{-1}\right\}$$

Recalling that the first component of x_i is an intercept term, i.e., 1, we may rewrite it as

$$\frac{1}{2n}\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\Lambda}_i\left(1-\hat{\Lambda}_i\right)x_{i,1}x_i'\right)\widehat{E}\left[\theta^{\epsilon\epsilon}\left(0\right)\right]$$
$$-\frac{1}{2n}\operatorname{trace}\left\{\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\Lambda}_i\left(1-\hat{\Lambda}_i\right)\left(1-2\hat{\Lambda}_i\right)x_{i,1}x_ix_i'\right)\left(\widehat{E}\left[v^{\theta}\right]\right)^{-1}\right\} \tag{33}$$

Recalling that $\widehat{E}\left[v^{\theta}\right]=-\frac{1}{n}\sum_{i=1}^{n}\hat{\Lambda}_i\left(1-\hat{\Lambda}_i\right)x_{i,1}x_i'$, we can understand the first term above as

$$-\frac{1}{2n}\left(\text{1st row of }\widehat{E}\left[v^{\theta}\right]\right)\widehat{E}\left[\theta^{\epsilon\epsilon}\left(0\right)\right]$$

In Appendix A, we saw that the kth component of $\widehat{E}\left[v^{\theta}\right]\widehat{E}\left[\theta^{\epsilon\epsilon}\left(0\right)\right]$ is

$$-\operatorname{trace}\left\{\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\Lambda}_i\left(1-\hat{\Lambda}_i\right)\left(1-2\hat{\Lambda}_i\right)x_{i,k}x_ix_i'\right)\left(\widehat{E}\left[v^{\theta}\right]\right)^{-1}\right\}$$

so we can see that (33) can be rewritten

$$-\frac{1}{2n}\left(\text{1st row of }\widehat{E}\left[v^{\theta}\right]\right)\widehat{E}\left[\theta^{\epsilon\epsilon}\left(0\right)\right]+\frac{1}{2n}\left(\text{1st component of }\widehat{E}\left[v^{\theta}\right]\widehat{E}\left[\theta^{\epsilon\epsilon}\left(0\right)\right]\right)=0$$

E Zero Second Order Bias of the ATE Under Random Assignment

In view of (25) in Appendix C, it suffices to prove that

$$\begin{aligned}&\frac{n}{2n_1}\left(\frac{1}{n}\sum_{i=1}^{n}\Lambda\left(x_i'\theta_{0,(1)}\right)\left(1-\Lambda\left(x_i'\theta_{0,(1)}\right)\right)x_i'\right)E_{(1)}\left[\theta^{\epsilon\epsilon}\left(0\right)\right]\\&-\frac{n}{2n_1}\operatorname{trace}\left\{\left(\frac{1}{n}\sum_{i=1}^{n}\Lambda\left(x_i'\theta_{0,(1)}\right)\left(1-\Lambda\left(x_i'\theta_{0,(1)}\right)\right)\left(1-2\Lambda\left(x_i'\theta_{0,(1)}\right)\right)x_ix_i'\right)\left(E_{(1)}\left[v^{\theta}\right]\right)^{-1}\right\}\\&=o_p\left(1\right),\end{aligned}\tag{34}$$

and

$$\begin{aligned}&\frac{1}{2n_0}\left(\frac{1}{n}\sum_{i=1}^{n}\Lambda\left(x_i'\theta_{0,(0)}\right)\left(1-\Lambda\left(x_i'\theta_{0,(0)}\right)\right)x_i'\right)E_{(0)}\left[\theta^{\epsilon\epsilon}\left(0\right)\right]\\&-\frac{1}{2n_0}\operatorname{trace}\left\{\left(\frac{1}{n}\sum_{i=1}^{n}\Lambda\left(x_i'\theta_{0,(0)}\right)\left(1-\Lambda\left(x_i'\theta_{0,(0)}\right)\right)\left(1-2\Lambda\left(x_i'\theta_{0,(0)}\right)\right)x_ix_i'\right)\left(E_{(0)}\left[v^{\theta}\right]\right)^{-1}\right\}\\&=o_p\left(1\right),\end{aligned}$$

under random assignment. We will only prove the former, because the latter can be established similarly.

Write

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\Lambda\left(x_i'\theta_{0,(1)}\right)\left(1-\Lambda\left(x_i'\theta_{0,(1)}\right)\right)x_i'=&\frac{n_1}{n}\frac{1}{n_1}\sum_{D_i=1}\Lambda\left(x_i'\theta_{0,(1)}\right)\left(1-\Lambda\left(x_i'\theta_{0,(1)}\right)\right)x_i'\\&+\frac{n_0}{n}\frac{1}{n_0}\sum_{D_i=0}\Lambda\left(x_i'\theta_{0,(1)}\right)\left(1-\Lambda\left(x_i'\theta_{0,(1)}\right)\right)x_i'\end{aligned}$$

Under the random assignment, the x_i has the identical distribution so

$$\frac{1}{n_1}\sum_{D_i=1}\Lambda\left(x_i'\theta_{0,(1)}\right)\left(1-\Lambda\left(x_i'\theta_{0,(1)}\right)\right)x_i'=\frac{1}{n_0}\sum_{D_i=0}\Lambda\left(x_i'\theta_{0,(1)}\right)\left(1-\Lambda\left(x_i'\theta_{0,(1)}\right)\right)x_i'+o_p\left(1\right),$$

and hence, we can write

$$
\begin{aligned}
&\left(\frac{1}{n}\sum_{i=1}^{n}\Lambda\left(x_i'\theta_{0,(1)}\right)\left(1-\Lambda\left(x_i'\theta_{0,(1)}\right)\right)x_i'\right)E_{(1)}\left[\theta^{\epsilon\epsilon}\left(0\right)\right]\\
&=\left(\frac{1}{n_1}\sum_{D_i=1}\Lambda\left(x_i'\theta_{0,(1)}\right)\left(1-\Lambda\left(x_i'\theta_{0,(1)}\right)\right)x_i'\right)E_{(1)}\left[\theta^{\epsilon\epsilon}\left(0\right)\right]+o_p\left(1\right)\\
&=\left(\text{1st row of }-E_{(1)}\left[v^{\theta}\right]\right)E_{(1)}\left[\theta^{\epsilon\epsilon}\left(0\right)\right]+o_p\left(1\right). \qquad (35)
\end{aligned}
$$

Likewise, we can write

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\Lambda\left(x_i'\theta_{0,(1)}\right)\left(1-\Lambda\left(x_i'\theta_{0,(1)}\right)\right)\left(1-2\Lambda\left(x_i'\theta_{0,(1)}\right)\right)x_ix_i'\\
&=\frac{1}{n_1}\sum_{D_i=1}\Lambda\left(x_i'\theta_{0,(1)}\right)\left(1-\Lambda\left(x_i'\theta_{0,(1)}\right)\right)\left(1-2\Lambda\left(x_i'\theta_{0,(1)}\right)\right)x_ix_i'+o_p\left(1\right)\\
&=-E_{(1)}\left[\frac{\partial^2 v_1}{\partial\theta\partial\theta'}\right]+o_p\left(1\right)
\end{aligned}
$$

under random assignment, so

$$
\begin{aligned}
&\operatorname{trace}\left\{\left(\frac{1}{n}\sum_{i=1}^{n}\Lambda\left(x_i'\theta_{0,(1)}\right)\left(1-\Lambda\left(x_i'\theta_{0,(1)}\right)\right)\left(1-2\Lambda\left(x_i'\theta_{0,(1)}\right)\right)x_ix_i'\right)\left(E_{(1)}\left[v^{\theta}\right]\right)^{-1}\right\}\\
&=-E_{(1)}\left[\frac{\partial^2 v_1}{\partial\theta\partial\theta'}\right]\left(E_{(1)}\left[v^{\theta}\right]\right)^{-1}+o_p\left(1\right)\\
&=-\text{1st component of }\widehat{E}\left[v^{\theta}\right]\widehat{E}\left[\theta^{\epsilon\epsilon}\left(0\right)\right]+o_p\left(1\right). \qquad (36)
\end{aligned}
$$

Combining (35) and (36), we conclude that (34) under random assignment as long as $n_1, n_0 \to \infty$ at the same rate.

F Lyapunov Condition

Note that

$$\begin{aligned} E\left|\frac{y_i - p_n}{\sqrt{np_n(1-p_n)}}\right|^3 &= \left(\frac{1-p_n}{\sqrt{np_n(1-p_n)}}\right)^3 p_n + \left(\frac{p_n}{\sqrt{np_n(1-p_n)}}\right)^3 (1-p_n) \\ &= \frac{1}{n\sqrt{n}} \frac{2p_n^2 - 2p_n + 1}{\sqrt{p_n(1-p_n)}} \end{aligned}$$

so

$$\sum_{i=1}^{n} E\left|\frac{y_i - p_n}{\sqrt{np_n(1-p_n)}}\right|^3 = \frac{1}{\sqrt{n}} \frac{2p_n^2 - 2p_n + 1}{\sqrt{p_n(1-p_n)}} \to 0$$

if $p_n \propto n^{-\delta}$ with $0 \leq \delta < 1$.

If $p_n \propto \frac{1}{n}$, we can see that

$$\lim_{n\to\infty} \frac{1}{\sqrt{n}} \frac{2p_n^2 - 2p_n + 1}{\sqrt{p_n(1-p_n)}} = 1$$

so Lyapunov condition is violated. This is consistent with the fact that we should expect Poisson approximation not normal approximation when p_n is very small. For concreteness, we will assume that $p_n = \frac{\lambda}{n}$ and analyze

$$\begin{aligned} \widehat{\theta} - \theta &= \ln \frac{\overline{y}}{1-\overline{y}} - \ln \frac{p_n}{1-p_n} = \ln \frac{\overline{y}}{p_n} - \ln \frac{1-\overline{y}}{1-p_n} \\ &= \ln \frac{n\overline{y}}{\lambda} - \ln \frac{1-\overline{y}}{1-\frac{\lambda}{n}} \end{aligned}$$

As for the second term, we can use $n\overline{y} = O_p(1)$ to conclude that

$$\ln \frac{1-\overline{y}}{1-\frac{\lambda}{n}} = \ln \frac{1 - O_p\left(\frac{1}{n}\right)}{1-\frac{\lambda}{n}} = o_p(1)$$

while $\ln \frac{n\overline{y}}{\lambda}$ is asymptotically equivalent to log of Poisson(λ) divided by its own mean. Problem is that we need to confront the fact that a Poisson distribution has a positive probability of being equal to 0, at which point the ln is not well-defined.

G Derivation of (5)

We analyze

$$\widehat{\theta} - \theta = \ln \frac{\overline{y}}{1 - \overline{y}} - \ln \frac{p_n}{1 - p_n} = \ln \frac{\overline{y}}{p_n} - \ln \frac{1 - \overline{y}}{1 - p_n}.$$

As for the second term, we have

$$\begin{aligned}
\ln \frac{1 - \overline{y}}{1 - p_n} &= \ln \frac{1 - \left(p_n + \frac{\sqrt{p_n (1 - p_n)}}{\sqrt{n}} Z_n \right)}{1 - p_n} \\
&= \ln \left(1 - \frac{\sqrt{p_n}}{\sqrt{n \left(1 - p_n\right)}} Z_n \right) \\
&= - \frac{\sqrt{p_n}}{\sqrt{n \left(1 - p_n\right)}} Z_n - \frac{1}{2} \left(\frac{\sqrt{p_n}}{\sqrt{n \left(1 - p_n\right)}} Z_n \right)^2 + O_p \left(n^{-3(1+\delta)/2} \right) \\
&= - \frac{\sqrt{p_n}}{\sqrt{n \left(1 - p_n\right)}} Z - \frac{1}{2} \left(\frac{\sqrt{p_n}}{\sqrt{n \left(1 - p_n\right)}} \right)^2 Z_n^2 + O_p \left(n^{-3(1+\delta)/2} \right)
\end{aligned}$$

noting that $\frac{\sqrt{p_n}}{\sqrt{n(1-p_n)}} = O\left(n^{-(1+\delta)/2}\right)$. As for the first term, we have

$$\begin{aligned}
\ln \frac{\overline{y}}{p_n} &= \ln \left(1 + \frac{\sqrt{1 - p_n}}{\sqrt{n p_n}} Z_n \right) \\
&= \frac{\sqrt{1 - p_n}}{\sqrt{n p_n}} Z_n - \frac{1}{2} \left(\frac{\sqrt{1 - p_n}}{\sqrt{n p_n}} Z_n \right)^2 + O \left(n^{-3(1-\delta)/2} \right) \\
&= \frac{\sqrt{1 - p_n}}{\sqrt{n p_n}} Z_n - \frac{1}{2} \left(\frac{\sqrt{1 - p_n}}{\sqrt{n p_n}} \right)^2 Z_n^2 + O \left(n^{-3(1-\delta)/2} \right)
\end{aligned}$$

noting that $\frac{\sqrt{1-p_n}}{\sqrt{n p_n}} = O\left(n^{-(1-\delta)/2}\right)$. To conclude, we have

$$\begin{aligned}
\widehat{\theta} - \theta &= \ln \frac{\overline{y}}{1 - \overline{y}} - \ln \frac{p_n}{1 - p_n} = \ln \frac{\overline{y}}{p_n} - \ln \frac{1 - \overline{y}}{1 - p_n} \\
&= \frac{\sqrt{1 - p_n}}{\sqrt{n p_n}} Z_n - \frac{1}{2} \left(\frac{\sqrt{1 - p_n}}{\sqrt{n p_n}} \right)^2 Z_n^2 + O_p \left(n^{-3(1-\delta)/2} \right) \\
&+ \frac{\sqrt{p_n}}{\sqrt{n \left(1 - p_n\right)}} Z_n + \frac{1}{2} \left(\frac{\sqrt{p_n}}{\sqrt{n \left(1 - p_n\right)}} \right)^2 Z_n^2 + O_p \left(n^{-3(1+\delta)/2} \right) \\
&= \frac{1}{\sqrt{n p_n \left(1 - p_n\right)}} Z_n - \frac{1}{2} \frac{1 - 2 p_n}{n p_n \left(1 - p_n\right)} Z_n^2 + O_p \left(n^{-3(1-\delta)/2} \right),
\end{aligned}$$

and using $\sqrt{np_n\left(1-p_n\right)} = O_p\left(n^{-(1-\delta)/2}\right)$, we conclude that

$$\sqrt{np_n\left(1-p_n\right)}\left(\widehat{\theta}-\theta\right) = Z_n - \frac{1}{2}\frac{1-2p_n}{\sqrt{np_n\left(1-p_n\right)}}Z_n^2 + O_p\left(n^{-(1-\delta)}\right).$$

H Tables

Table 1: Non-convergence and Artificial Censoring

Initial guess	α	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9
n=50000										
	Full Sample									
$\alpha = -10$	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	Random Censoring									
$\alpha = -10 - (-6.21)$	-9.81	0.79	1.06	0.97	0.57	0.78	1.00	0.91	0.76	0.89
n=100000										
	Full Sample									
$\alpha = -10$	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	Random Censoring									
$\alpha = -10 - (-6.91)$	-10.4	0.95	1.44	1.19	1.02	1.21	0.99	1.24	1.41	0.77

Table 2: Mean Bias of MLE; $\beta_0 = 1$

	$\alpha_0 = -2.5$		$\alpha_0 = -3$		$\alpha_0 = -3.5$		$\alpha_0 = -4$	
	p[8] $= 12.25\%$		$p = 7.83\%$		$p = 4.91\%$		$p = 3.04\%$	
	Mean bias of α							
	$\hat{\alpha}$	$\tilde{\alpha}$	$\hat{\alpha}$	$\tilde{\alpha}$	$\hat{\alpha}$	$\tilde{\alpha}$	$\hat{\alpha}$	$\tilde{\alpha}$
n=500	-0.0249	-0.0033	-0.0357	-0.0005	-0.0703	-0.0109	-0.1164	-0.0130
n=750	-0.0179	-0.0036	-0.0243	-0.0013	-0.0435	-0.0053	-0.0784	-0.0135
n=1000	-0.0060	0.0046	-0.0143	0.0028	-0.0297	-0.0016	-0.0499	-0.0028
n=5000	-0.0021	0.0000	-0.0055	-0.0021	-0.0076	-0.0021	-0.0115	-0.0025
	Mean bias of β							
	$\hat{\beta}$	$\tilde{\beta}$	$\hat{\beta}$	$\tilde{\beta}$	$\hat{\beta}$	$\tilde{\beta}$	$\hat{\beta}$	$\tilde{\beta}$
n=500	0.0168	0.0041	0.0183	-0.0006	0.0391	0.0088	0.0630	0.0112
n=750	0.0152	0.0068	0.0165	0.0042	0.0319	0.0124	0.0564	0.0239
n=1000	-0.0007	-0.0068	0.0070	-0.0021	0.0189	0.0046	0.0285	0.0054
n=5000	-0.0002	-0.0014	0.0030	0.0012	0.0036	0.0008	0.0036	-0.0007

[8] p denotes the probability of y=1 for each parameter combination.

Table 3: Mean Bias of MLE; $\beta_0 = 1.5$

	$\alpha_0 = -2.5$		$\alpha_0 = -3$		$\alpha_0 = -3.5$		$\alpha_0 = -4$	
	$p = 15.62\%$		$p = 10.19\%$		$p = 6.48\%$		$p = 4.05\%$	
				Mean bias of α				
	$\hat{\alpha}$	$\tilde{\alpha}$	$\hat{\alpha}$	$\tilde{\alpha}$	$\hat{\alpha}$	$\tilde{\alpha}$	$\hat{\alpha}$	$\tilde{\alpha}$
n=500	-0.0243	-0.0043	-0.0345	-0.0028	-0.0617	-0.0098	-0.1049	-0.0169
n=750	-0.0155	-0.0023	-0.0224	-0.0017	-0.0374	-0.0037	-0.0674	-0.0116
n=1000	-0.0058	0.0018	-0.0128	0.0027	-0.0252	-0.0004	-0.0395	0.0013
n=5000	-0.0022	-0.0003	-0.0045	-0.0014	-0.0075	-0.0027	-0.0105	-0.0027
				Mean bias of β				
	$\hat{\beta}$	$\tilde{\beta}$	$\hat{\beta}$	$\tilde{\beta}$	$\hat{\beta}$	$\tilde{\beta}$	$\hat{\beta}$	$\tilde{\beta}$
n=500	0.0190	0.0031	0.0270	0.0042	0.0444	0.0092	0.0724	0.0146
n=750	0.0168	0.0063	0.0193	0.0043	0.0278	0.0052	0.0585	0.0219
n=1000	0.0040	-0.0060	0.0076	-0.0034	0.0190	0.0023	0.0274	0.0011
n=5000	0.0008	-0.0007	0.0025	0.0037	0.0042	0.0009	0.0060	0.0010

Table 4: Mean Bias of MLE; $\beta_0 = 2$

	$\alpha_0 = -2.5$		$\alpha_0 = -3$		$\alpha_0 = -3.5$		$\alpha_0 = -4$	
	$p = 19.76\%$		$p = 13.23\%$		$p = 8.58\%$		$p = 5.43\%$	
				Mean bias of α				
	$\hat{\alpha}$	$\tilde{\alpha}$	$\hat{\alpha}$	$\tilde{\alpha}$	$\hat{\alpha}$	$\tilde{\alpha}$	$\hat{\alpha}$	$\tilde{\alpha}$
n=500	-0.0240	-0.0054	-0.0309	-0.0023	-0.0532	-0.0079	-0.0888	-0.0142
n=750	-0.0136	-0.0013	-0.0195	-0.0007	-0.0322	-0.0026	-0.0535	-0.0056
n=1000	-0.0062	0.0030	-0.0120	0.0020	-0.0194	0.0025	-0.0342	0.0010
n=5000	-0.0027	-0.0010	-0.0044	-0.0016	-0.0064	-0.0021	-0.0087	-0.0019
				Mean bias of β				
	$\hat{\beta}$	$\tilde{\beta}$	$\hat{\beta}$	$\tilde{\beta}$	$\hat{\beta}$	$\tilde{\beta}$	$\hat{\beta}$	$\tilde{\beta}$
n=500	0.0232	0.0048	0.0270	0.0021	0.0487	0.0118	0.0720	0.0140
n=750	0.0170	0.0049	0.0192	0.0028	0.0306	0.0066	0.0447	0.0077
n=1000	0.0044	-0.0046	0.0084	-0.0038	0.0146	-0.0031	0.0280	0.0009
n=5000	0.0023	0.0006	0.0035	0.0010	0.0046	0.0011	0.0058	0.0005

Table 5: Mean Bias of ATE Estimators; Random Assignment

$\beta_0 = 1$	$\alpha_0 = -2.5$		$\alpha_0 = -3$		$\alpha_0 = -3.5$		$\alpha_0 = -4$	
$\beta_1 = 2$	$\alpha_1 = -1.5$		$\alpha_1 = -2$		$\alpha_1 = -2.5$		$\alpha_1 = -3$	
$x \mid D = 1 \sim N(0,1)$	p_1[9] $= 28.61\%$		$p_1 = 22.45\%$		$p_1 = 17.17\%$		$p_1 = 12.86\%$	
$x \mid D = 0 \sim N(0,1)$	p_0[10] $= 10.45\%$		$p_0 = 6.89\%$		$p_0 = 4.41\%$		$p_0 = 2.76\%$	
	$\frac{n_1}{n} = \frac{1}{2}$							
	$\widehat{ATE}_1$	$\widetilde{ATE}_1$	$\widehat{ATE}_1$	$\widetilde{ATE}_1$	$\widehat{ATE}_1$	$\widetilde{ATE}_1$	$\widehat{ATE}_1$	$\widetilde{ATE}_1$
n=1000	0.0005	0.0005	0.0000	0.0000	0.0005	0.0005	0.0003	0.0003
n=1500	-0.0001	-0.0001	0.0004	0.0004	0.0001	0.0001	-0.0004	-0.0004
n=2000	0.0004	0.0004	0.0003	0.0003	0.0003	0.0003	0.0000	0.0000
	$\widehat{ATE}_2$	$\widetilde{ATE}_2$	$\widehat{ATE}_2$	$\widetilde{ATE}_2$	$\widehat{ATE}_2$	$\widetilde{ATE}_2$	$\widehat{ATE}_2$	$\widetilde{ATE}_2$
n=1000	-0.0003	0.0006	-0.0009	0.0004	-0.0005	0.0012	-0.0007	0.0007
n=1500	-0.0007	0.0000	-0.0002	0.0004	-0.0005	0.0000	-0.0011	-0.0004
n=2000	0.0000	0.0004	-0.0001	0.0003	-0.0002	0.0002	-0.0004	0.0002
	$\frac{n_1}{n} = \frac{2}{3}$							
	$\widehat{ATE}_1$	$\widetilde{ATE}_1$	$\widehat{ATE}_1$	$\widetilde{ATE}_1$	$\widehat{ATE}_1$	$\widetilde{ATE}_1$	$\widehat{ATE}_1$	$\widetilde{ATE}_1$
n=1500	0.0000	0.0000	-0.0007	-0.0007	-0.0006	-0.0006	-0.0008	-0.0008
n=2250	0.0000	0.0000	0.0000	0.0000	0.0008	0.0008	0.0005	0.0005
n=3000	-0.0006	-0.0006	-0.0002	-0.0002	0.0002	0.0002	0.0003	0.0003

[9] p_1 denotes the probability of y=1 for $D_i = 1$, i.e., the treated sub-sample.

[10] p_0 denotes the probability of y=1 for $D_i = 0$, i.e., the control sub-sample.

	$\widehat{ATE_2}$	$\widetilde{ATE_2}$	$\widehat{ATE_2}$	$\widetilde{ATE_2}$	$\widehat{ATE_2}$	$\widetilde{ATE_2}$	$\widehat{ATE_2}$	$\widetilde{ATE_2}$
n=1500	-0.0011	0.0000	-0.0019	-0.0008	-0.0019	-0.0003	-0.0021	-0.0006
n=2250	-0.0006	0.0002	-0.0007	0.0000	0.0000	0.0009	-0.0004	0.0008
n=3000	-0.0011	-0.0006	-0.0008	-0.0003	-0.0004	0.0002	-0.0004	0.0003

Table 6: Mean Bias of ATE Estimators; Random Assignment

$\beta_0 = 1$	$\alpha_0 = -2.5$	$\alpha_0 = -3$	$\alpha_0 = -3.5$	$\alpha_0 = -4$
$\beta_1 = 4$	$\alpha_1 = -1.5$	$\alpha_1 = -2$	$\alpha_1 = -2.5$	$\alpha_1 = -3$
$x\|D = 1 \sim N(0,1)$	$p_1 = 36.57\%$	$p_1 = 32.39\%$	$p_1 = 28.40\%$	$p_1 = 24.68\%$
$x\|D = 0 \sim N(0,1)$	$p_0 = 10.45\%$	$p_0 = 6.89\%$	$p_0 = 4.41\%$	$p_0 = 2.76\%$

	$\frac{n_1}{n} = \frac{1}{2}$							
	$\widehat{ATE_1}$	$\widetilde{ATE_1}$	$\widehat{ATE_1}$	$\widetilde{ATE_1}$	$\widehat{ATE_1}$	$\widetilde{ATE_1}$	$\widehat{ATE_1}$	$\widetilde{ATE_1}$
n=1000	0.0004	0.0004	0.0002	0.0002	0.0012	0.0012	0.0008	0.0008
n=1500	-0.0003	-0.0003	0.0000	0.0000	0.0007	0.0007	0.0005	0.0005
n=2000	0.0008	0.0008	0.0004	0.0004	-0.0002	-0.0002	0.0003	0.0004
	$\widehat{ATE_2}$	$\widetilde{ATE_2}$	$\widehat{ATE_2}$	$\widetilde{ATE_2}$	$\widehat{ATE_2}$	$\widetilde{ATE_2}$	$\widehat{ATE_2}$	$\widetilde{ATE_2}$
n=1000	-0.0007	0.0005	-0.0010	0.0006	-0.0002	0.0019	-0.0006	0.0013
n=1500	-0.0011	-0.0003	-0.0009	0.0000	-0.0002	0.0007	-0.0005	0.0005
n=2000	0.0002	0.0007	-0.0002	0.0004	-0.0008	-0.0002	-0.0003	0.0004
	$\frac{n_1}{n} = \frac{2}{3}$							
	$\widehat{ATE_1}$	$\widetilde{ATE_1}$	$\widehat{ATE_1}$	$\widetilde{ATE_1}$	$\widehat{ATE_1}$	$\widetilde{ATE_1}$	$\widehat{ATE_1}$	$\widetilde{ATE_1}$
n=1500	0.0002	0.0002	-0.0005	-0.0005	-0.0004	-0.0004	-0.0002	-0.0002
n=2250	0.0002	0.0002	0.0000	0.0000	0.0004	0.0004	0.0005	0.0005
n=3000	-0.0003	-0.0003	-0.0005	-0.0005	-0.0002	-0.0002	-0.0002	-0.0002
	$\widehat{ATE_2}$	$\widetilde{ATE_2}$	$\widehat{ATE_2}$	$\widetilde{ATE_2}$	$\widehat{ATE_2}$	$\widetilde{ATE_2}$	$\widehat{ATE_2}$	$\widetilde{ATE_2}$

n=1500	-0.0010	0.0004	-0.0018	-0.0006	-0.0019	-0.0001	-0.0019	0.0000
n=2250	-0.0005	0.0004	-0.0010	0.0000	-0.0006	0.0005	-0.0006	0.0008
n=3000	-0.0009	-0.0003	-0.0011	-0.0005	-0.0010	-0.0003	-0.0010	-0.0001

Table 7: Mean Bias of ATE Estimators; Random Assignment

$\beta_0 = 1$	$\alpha_0 = -2.5$		$\alpha_0 = -3$		$\alpha_0 = -3.5$		$\alpha_0 = -4$	
$\beta_1 = 1$	$\alpha_1 = -1.5$		$\alpha_1 = -2$		$\alpha_1 = -2.5$		$\alpha_1 = -3$	
$x \mid D = 0,1 \sim N(0,1)$	$p_1 = p_0 = 10.45\%$		$p_1 = p_0 = 6.89\%$		$p_1 = p_0 = 4.41\%$		$p_1 = p_0 = 2.76\%$	
	$\frac{n_1}{n} = \frac{1}{2}$							
	$\widehat{ATE}_1$	$\widetilde{ATE}_1$	$\widehat{ATE}_1$	$\widetilde{ATE}_1$	$\widehat{ATE}_1$	$\widetilde{ATE}_1$	$\widehat{ATE}_1$	$\widetilde{ATE}_1$
n=1000	0.0006	0.0006	0.0007	0.0007	0.0006	0.0006	0.0004	0.0004
n=1500	-0.0009	-0.0009	-0.0003	-0.0003	0.0000	0.0000	0.0000	0.0000
n=2000	-0.0003	-0.0003	-0.0001	-0.0001	0.0000	0.0000	0.0000	0.0000
	$\widehat{ATE}_2$	$\widetilde{ATE}_2$	$\widehat{ATE}_2$	$\widetilde{ATE}_2$	$\widehat{ATE}_2$	$\widetilde{ATE}_2$	$\widehat{ATE}_2$	$\widetilde{ATE}_2$
n=1000	0.0006	0.0008	0.0007	0.0012	0.0006	0.0016	0.0004	0.0006
n=1500	-0.0009	-0.0009	-0.0003	-0.0002	0.0000	0.0000	0.0000	0.0001
n=2000	-0.0003	-0.0004	-0.0001	-0.0002	0.0000	0.0000	0.0000	0.0000
	$\frac{n_1}{n} = \frac{2}{3}$							
	$\widehat{ATE}_1$	$\widetilde{ATE}_1$	$\widehat{ATE}_1$	$\widetilde{ATE}_1$	$\widehat{ATE}_1$	$\widetilde{ATE}_1$	$\widehat{ATE}_1$	$\widetilde{ATE}_1$
n=1500	-0.0005	-0.0005	-0.0005	-0.0005	-0.0002	-0.0002	-0.0003	-0.0003
n=2250	0.0003	0.0003	0.0000	0.0000	0.0004	0.0004	0.0003	0.0003
n=3000	0.0000	0.0000	-0.0004	-0.0004	0.0000	0.0000	0.0001	0.0001
	$\widehat{ATE}_2$	$\widetilde{ATE}_2$	$\widehat{ATE}_2$	$\widetilde{ATE}_2$	$\widehat{ATE}_2$	$\widetilde{ATE}_2$	$\widehat{ATE}_2$	$\widetilde{ATE}_2$
n=1500	-0.0011	-0.0004	-0.0013	-0.0007	-0.0010	-0.0011	-0.0012	-0.0002

n=2250	0.0000	0.0005	-0.0004	0.0002	-0.0001	0.0005	-0.0003	0.0007
n=3000	-0.0003	0.0000	-0.0007	-0.0004	-0.0004	0.0000	-0.0003	0.0002

Table 8: Mean Bias of ATE Estimators; Non-random Assignment

$\beta_0 = 1$	$\alpha_0 = -2.5$		$\alpha_0 = -3$		$\alpha_0 = -3.5$		$\alpha_0 = -4$	
$\beta_1 = 1$	$\alpha_1 = -3$		$\alpha_1 = -3$		$\alpha_1 = -3$		$\alpha_1 = -3$	
$x \mid D = 1 \sim N(4,1)$	$p_1 = 69.60\%$							
$x \mid D = 0 \sim N(0,1)$	$p_0 = 10.45\%$		$p_0 = 6.89\%$		$p_0 = 4.41\%$		$p_0 = 2.76\%$	
	$\frac{n_1}{n} = \frac{1}{2}$							
	$\widehat{ATE}_1$	$\widetilde{ATE}_1$	$\widehat{ATE}_1$	$\widetilde{ATE}_1$	$\widehat{ATE}_1$	$\widetilde{ATE}_1$	$\widehat{ATE}_1$	$\widetilde{ATE}_1$
n=1000	0.0049	-0.0010	0.0061	-0.0010	0.0068	-0.0004	0.0033	-0.0014
n=1500	0.0037	-0.0002	0.0064	0.0015	0.0079	0.0028	0.0085	0.0051
n=2000	0.0022	-0.0007	0.0027	-0.0010	0.0038	-0.0003	0.0009	-0.0024
	$\widehat{ATE}_2$	$\widetilde{ATE}_2$	$\widehat{ATE}_2$	$\widetilde{ATE}_2$	$\widehat{ATE}_2$	$\widetilde{ATE}_2$	$\widehat{ATE}_2$	$\widetilde{ATE}_2$
n=1000	0.0061	-0.0012	0.0073	-0.0019	0.0073	-0.0019	0.0025	-0.0030
n=1500	0.0045	-0.0004	0.0072	0.0014	0.0083	0.0027	0.0079	0.0046
n=2000	0.0029	-0.0006	0.0033	-0.0010	0.0040	-0.0006	0.0005	-0.0032
	$\frac{n_1}{n} = \frac{2}{3}$							
	$\widehat{ATE}_1$	$\widetilde{ATE}_1$	$\widehat{ATE}_1$	$\widetilde{ATE}_1$	$\widehat{ATE}_1$	$\widetilde{ATE}_1$	$\widehat{ATE}_1$	$\widetilde{ATE}_1$
n=1500	0.0056	-0.0008	0.0062	-0.0017	0.0072	-0.0006	0.0047	0.0003
n=2250	0.0053	0.0010	0.0090	0.0034	0.0102	0.0041	0.0091	0.0051
n=3000	0.0037	0.0005	0.0067	0.0024	0.0088	0.0041	0.0068	0.0033
	$\widehat{ATE}_2$	$\widetilde{ATE}_2$	$\widehat{ATE}_2$	$\widetilde{ATE}_2$	$\widehat{ATE}_2$	$\widetilde{ATE}_2$	$\widehat{ATE}_2$	$\widetilde{ATE}_2$

n=1500	0.0071	-0.0011	0.0077	-0.0014	0.0080	-0.0018	0.0040	-0.0013
n=2250	0.0064	0.0004	0.0100	0.0033	0.0107	0.0038	0.0083	0.0040
n=3000	0.0046	0.0004	0.0075	0.0025	0.0091	0.0041	0.0062	0.0027

Table 9: Mean Bias of ATE Estimators; Non-random Assignment

$\beta_0 = 1$	$\alpha_0 = -2$		$\alpha_0 = -3$		$\alpha_0 = -3.5$		$\alpha_0 = -4$	
$\beta_1 = 1$	$\alpha_1 = -2$		$\alpha_1 = -2$		$\alpha_1 = -2$		$\alpha_1 = -2$	
$x \mid D = 1 \sim N(2, 1)$	$p_1 = 50.05\%$							
$x \mid D = 0 \sim N(0, 1)$	$p_0 = 10.45\%$		$p_0 = 6.89\%$		$p_0 = 4.41\%$		$p_0 = 2.76\%$	
	$\frac{n_1}{n} = \frac{1}{2}$							
	$\widehat{ATE_1}$	$\widetilde{ATE_1}$	$\widehat{ATE_1}$	$\widetilde{ATE_1}$	$\widehat{ATE_1}$	$\widetilde{ATE_1}$	$\widehat{ATE_1}$	$\widetilde{ATE_1}$
n=1000	0.0002	-0.0002	0.0000	0.0000	-0.0003	0.0005	-0.0023	-0.0004
n=1500	0.0006	0.0003	0.0014	0.0014	0.0014	0.0020	0.0007	0.0021
n=2000	0.0000	-0.0002	-0.0003	-0.0003	-0.0004	0.0000	-0.0014	-0.0005
	$\widehat{ATE_2}$	$\widetilde{ATE_2}$	$\widehat{ATE_2}$	$\widetilde{ATE_2}$	$\widehat{ATE_2}$	$\widetilde{ATE_2}$	$\widehat{ATE_2}$	$\widetilde{ATE_2}$
n=1000	0.0006	-0.0002	-0.0003	0.0002	-0.0013	0.0008	-0.0041	-0.0007
n=1500	0.0009	0.0003	0.0012	0.0014	0.0007	0.0019	-0.0005	0.0023
n=2000	0.0002	-0.0002	-0.0004	-0.0003	-0.0008	0.0002	-0.0023	0.0000
	$\frac{n_1}{n} = \frac{2}{3}$							
	$\widehat{ATE_1}$	$\widetilde{ATE_1}$	$\widehat{ATE_1}$	$\widetilde{ATE_1}$	$\widehat{ATE_1}$	$\widetilde{ATE_1}$	$\widehat{ATE_1}$	$\widetilde{ATE_1}$
n=1500	-0.0008	-0.0010	-0.0023	-0.0018	-0.0030	-0.0016	-0.0045	-0.0017
n=2250	0.0010	0.0089	0.0017	0.0020	0.0013	0.0023	-0.0002	0.0018
n=3000	0.0000	0.0000	0.0006	0.0008	0.0005	0.0013	-0.0008	0.0007
	$\widehat{ATE_2}$	$\widetilde{ATE_2}$	$\widehat{ATE_2}$	$\widetilde{ATE_2}$	$\widehat{ATE_2}$	$\widetilde{ATE_2}$	$\widehat{ATE_2}$	$\widetilde{ATE_2}$

n=1500	-0.0005	-0.0010	-0.0027	-0.0018	-0.0044	-0.0014	-0.0069	-0.0012
n=2250	0.0012	0.0009	0.0014	0.0020	0.0003	0.0022	-0.0020	0.0017
n=3000	0.0002	0.0000	0.0004	0.0008	-0.0002	0.0013	-0.0021	0.0013

Table 10: Mean Bias of ATE Estimators; Non-random Assignment

$\beta_0 = 1$	$\alpha_0 = -2.5$		$\alpha_0 = -3$		$\alpha_0 = -3.5$		$\alpha_0 = -4$	
$\beta_1 = 1$	$\alpha_1 = -2.5$		$\alpha_1 = -3$		$\alpha_1 = -3.5$		$\alpha_1 = -4$	
$x\|D = 1 \sim N(4,1)$	$p_1 = 77.81\%$		$p_1 = 69.60\%$		$p_1 = 60.19\%$		$p_1 = 50.05\%$	
$x\|D = 0 \sim N(0,1)$	$p_0 = 10.45\%$		$p_0 = 6.89\%$		$p_0 = 4.41\%$		$p_0 = 2.76\%$	
	$\frac{n_1}{n} = \frac{1}{2}$							
	$\widehat{ATE_1}$	$\widetilde{ATE_1}$	$\widehat{ATE_1}$	$\widetilde{ATE_1}$	$\widehat{ATE_1}$	$\widetilde{ATE_1}$	$\widehat{ATE_1}$	$\widetilde{ATE_1}$
n=1000	0.0044	-0.0021	0.0061	-0.0010	0.0062	-0.0005	0.0027	-0.0013
n=1500	0.0041	-0.0003	0.0064	0.0015	0.0083	0.0034	0.0090	0.0061
n=2000	0.0025	-0.0008	0.0027	-0.0010	0.0040	0.0002	0.0010	-0.0020
	$\widehat{ATE_2}$	$\widetilde{ATE_2}$	$\widehat{ATE_2}$	$\widetilde{ATE_2}$	$\widehat{ATE_2}$	$\widetilde{ATE_2}$	$\widehat{ATE_2}$	$\widetilde{ATE_2}$
n=1000	0.0058	-0.0023	0.0073	-0.0019	0.0066	-0.0019	0.0018	-0.0028
n=1500	0.0050	-0.0005	0.0072	0.0014	0.0086	0.0033	0.0083	0.0056
n=2000	0.0033	-0.0007	0.0033	-0.0010	0.0043	0.0000	0.0006	-0.0027
	$\frac{n_1}{n} = \frac{2}{3}$							
	$\widehat{ATE_1}$	$\widetilde{ATE_1}$	$\widehat{ATE_1}$	$\widetilde{ATE_1}$	$\widehat{ATE_1}$	$\widetilde{ATE_1}$	$\widehat{ATE_1}$	$\widetilde{ATE_1}$
n=1500	0.0047	-0.0019	0.0062	-0.0017	0.0073	-0.0004	0.0042	0.0000
n=2250	0.0051	0.0006	0.0090	0.0034	0.0104	0.0045	0.0092	0.0054
n=3000	0.0037	0.0003	0.0067	0.0024	0.0085	0.0039	0.0066	0.0032
	$\widehat{ATE_2}$	$\widetilde{ATE_2}$	$\widehat{ATE_2}$	$\widetilde{ATE_2}$	$\widehat{ATE_2}$	$\widetilde{ATE_2}$	$\widehat{ATE_2}$	$\widetilde{ATE_2}$

n=1500	0.0062	-0.0023	0.0077	-0.0014	0.0081	-0.0016	0.0036	-0.0015
n=2250	0.0062	0.0000	0.0100	0.0033	0.0109	0.0041	0.0084	0.0043
n=3000	0.0045	0.0002	0.0075	0.0025	0.0089	0.0038	0.0060	0.0026

Table 11: Mean Bias of ATE Estimators; Non-random Assignment

$\beta_0 = 1$	$\alpha_0 = -2.5$		$\alpha_0 = -3$		$\alpha_0 = -3.5$		$\alpha_0 = -4$	
$\beta_1 = 1$	$\alpha_1 = -2.5$		$\alpha_1 = -3$		$\alpha_1 = -3.5$		$\alpha_1 = -4$	
$x\|D = 1 \sim N(2,1)$	$p_1 = 39.93\%$		$p_1 = 30.38\%$		$p_1 = 22.13\%$		$p_1 = 15.50\%$	
$x\|D = 0 \sim N(0,1)$	$p_0 = 10.45\%$		$p_0 = 6.89\%$		$p_0 = 4.41\%$		$p_0 = 2.76\%$	
	$\frac{n_1}{n} = \frac{1}{2}$							
	$\widehat{ATE}_1$	$\widetilde{ATE}_1$	$\widehat{ATE}_1$	$\widetilde{ATE}_1$	$\widehat{ATE}_1$	$\widetilde{ATE}_1$	$\widehat{ATE}_1$	$\widetilde{ATE}_1$
n=1000	-0.0001	-0.0005	-0.0006	-0.0005	-0.0009	0.0000	-0.0026	-0.0006
n=1500	0.0003	0.0000	0.0010	0.0010	0.0008	0.0014	-0.0002	0.0012
n=2000	0.0001	-0.0001	-0.0002	-0.0002	-0.0006	-0.0001	-0.0017	-0.0007
	$\widehat{ATE}_2$	$\widetilde{ATE}_2$	$\widehat{ATE}_2$	$\widetilde{ATE}_2$	$\widehat{ATE}_2$	$\widetilde{ATE}_2$	$\widehat{ATE}_2$	$\widetilde{ATE}_2$
n=1000	0.0003	-0.0004	-0.0007	-0.0003	-0.0017	0.0003	-0.0042	-0.0010
n=1500	0.0006	0.0000	0.0009	0.0010	0.0002	0.0014	-0.0013	0.0014
n=2000	0.0003	0.0000	-0.0002	-0.0001	-0.0010	0.0000	-0.0025	-0.0003
	$\frac{n_1}{n} = \frac{2}{3}$							
	$\widehat{ATE}_1$	$\widetilde{ATE}_1$	$\widehat{ATE}_1$	$\widetilde{ATE}_1$	$\widehat{ATE}_1$	$\widetilde{ATE}_1$	$\widehat{ATE}_1$	$\widetilde{ATE}_1$
n=1500	-0.0008	-0.0010	-0.0024	-0.0020	-0.0035	-0.0020	-0.0046	-0.0017
n=2250	0.0009	0.0008	0.0014	0.0017	0.0012	0.0022	-0.0004	0.0016
n=3000	-0.0002	-0.0003	0.0005	0.0007	0.0003	0.0010	-0.0007	0.0007
	$\widehat{ATE}_2$	$\widetilde{ATE}_2$	$\widehat{ATE}_2$	$\widetilde{ATE}_2$	$\widehat{ATE}_2$	$\widetilde{ATE}_2$	$\widehat{ATE}_2$	$\widetilde{ATE}_2$

n=1500	-0.0005	-0.0010	-0.0027	-0.0019	-0.0046	-0.0019	-0.0068	-0.0013
n=2250	0.0012	0.0008	0.0011	0.0017	0.0003	0.0021	-0.0020	0.0015
n=3000	0.0000	-0.0003	0.0003	0.0007	-0.0003	0.0011	-0.0019	0.0014