Testing for Average Treatment Effects in Choice-Based Samples

Kentaro Akashi* **Tetsushi Horie†**

May 2024

Abstract

It is shown that in causal inference based on choice-based samples, the consistent estimation and t-tests of propensity scores and average treatment effects can be performed only from biased subsamples without external knowledge about the original random samples. Thus, program evaluation becomes more feasible.

Keywords: Choice-based samples, Cosslett's maximum likelihood estimator, Random subsample size, Inverse probability weighting.
JEL Classification: C12, C21.

1 Introduction

Due to the cost of data collection, program evaluation is often conducted with matched subsamples. The subsamples from which an econometrician arbitrarily selects treatment and control groups are called choice-based samples. Choice-based or endogenous sampling has the advantage of simple to implement because, unlike exogenous sampling with multivariate exogenous variables or propensity scores, the former method relies only on treatment and control groups with the endogenous dummy variable. Generally, however, the subsamples are not representative because they are drawn differently from the population ratio, which possibly leads to the selection bias in average treatment effects (ATE) estimation.

Although there has been much development in causal inference under random and choice-based samples with external knowledge, the following are the previous studies on robustness when only choice-based samples are available. Heckman and Todd (2009) suggested that even if the propensity scores has the bias, ATE can be identified from the propensity score matching based on its odds. For ATE on the treated (ATT), Kenndy et al. (2015) pointed out that ATT can be estimated using standard estimation methods even if propensity score is not identified. As a different approach, we present a method in which the propensity score is first estimated consistently, and the identification and consistent estimator of ATE are obtained by inverse probability weighting (IPW).

*Faculty of Economics, Gakushuin University, Tokyo, Japan.
†Policy Research Institute, Ministry of Finance, Tokyo, Japan.
Email address: kentaro.akashi@gakushuin.ac.jp
The authors appreciate the participants of the KIER-CAPS Workshop 2024 (Kyoto university) for their helpful comments.

Our goal is to present consistent estimators and t-tests for the propensity score and ATE when the original random sample size, n, is unknown and only choice-based samples are available. First, we identify the population ratio of the treatment group by applying the approach of Cosslett (1981) with the number of subsamples as random. Second, the IPW estimator is extended to choice-based samples. Third, we demonstrate that normalization and the standard error do not depend on n with regard to the ATE significance test. Therefore, it is easier to evaluate a program as only biased subsamples are needed without external knowledge.

The remainder of this paper is organized as follows. The next section provides an overview of choice-based sampling and presents the t-tests on the propensity score and the ATE. Section 3 describes the numerical experiments and the results and Section 4 is the conclusions. All proofs are summarized in the Appendix.

2 t-tests for propensity score and ATE

We consider the Rubin causal model:

$$y_{1i} = y_{1i}^* y_{2i} \ , \tag{2.1}$$

$$y_{0i} = y_{0i}^* (1 - y_{2i}) \ , \tag{2.2}$$

where y_{1i}^* and y_{0i}^* are unobservable and only one of y_{1i} or y_{i0} , $y_i = y_{1i} + y_{i0}$, is observed for each subject i. $y_{2i} = \mathbb{1}\{\boldsymbol{\beta}'\mathbf{x}_i + u_{2i} \geq 0\}$, where the indicator function $\mathbb{1}\{.\}$ takes the value of 1 if the argument is true; otherwise, it takes 0. $\mathbf{x}_i$ are the K-variate covariates that are independent of u_{2i} . The error term u_{2i} follows the standard normal distribution, i.e., the propensity score becomes the probit model.

$$\begin{aligned} p_i &= \Pr(y_{2i} = 1 | \mathbf{x}_i) \\ &= \Phi(\boldsymbol{\beta}'\mathbf{x}_i) \ , \end{aligned} \tag{2.3}$$

where Φ is the standard normal cumulative distribution function. The parameters of interest are $\boldsymbol{\beta}$, $\tau = \mathcal{E}[y_{1i}^* - y_{0i}^*]$ as the ATE, and $p = \mathcal{E}[p_i]$ which identifies τ .

2.1 Choice-based samples

For random samples ($i = 1, \ \cdots \ , \ n$) , let $n_1 = \sum_{i=1}^{n} y_{2i}$ be the number of the treatment group and $n_0 = \sum_{i=1}^{n}(1 - y_{2i})$ be the number of the control group. Then, the following holds:

$$\frac{n_1}{n} : \frac{n_0}{n} \simeq p : 1 - p \ , \tag{2.4}$$

where p is the proportion of the treatment group in the population,

$$p = \Pr(y_{2i} = 1) \ . \tag{2.5}$$

Choice-based samples ($j = 1, \cdots, m$) are subsamples ($m \leq n$) of random samples arbitrarily collected by an econometrician for each treatment and control group. Hence, the ratio r of the treatment group in the subsamples is known.

$$\frac{m_1}{m} : \frac{m_0}{m} = r : 1 - r \ , \tag{2.6}$$

where $m_1 = \sum_{j=1}^{m} y_{2j}$, $m_0 = \sum_{j=1}^{m}(1 - y_{2j})$, and $m = m_1 + m_0$. Hereafter, subscript j indicates that the sample of subject j is drawn as the choice-based sample.

The IPW estimator for ATE is widely used in program evaluation.

$$\tilde{\tau} = \frac{1}{m}\sum_{j=1}^{m} \frac{y_{1j}}{\tilde{p}_j} - \frac{y_{0j}}{1 - \tilde{p}_j} \ , \tag{2.7}$$

where $\tilde{p}_j = \Phi(\tilde{\boldsymbol{\beta}}' \mathbf{x}_j)$ and $\tilde{\boldsymbol{\beta}}$ is the probit maximum likelihood estimator (MLE). However, as the choice-based samples follow the biased distribution as $r : 1 - r \neq p : 1 - p$ in general, there may be a selection bias.

First, we consider the consistent estimation of $\boldsymbol{\beta}$ in the propensity score. The consistent estimation methods for dealing with choice-based samples are summarized in Amemiya (1985, Ch. 9). Given the true value of p , Manski and Lerman (1977) proposed the weighted MLE obtained by

$$\sum_{j=1}^{m} \frac{p}{r} y_{2j} \log p_j + \frac{1-p}{1-r}(1 - y_{2j}) \log(1 - p_j) \ . \tag{2.8}$$

Thus, a bias can arise in $\tilde{p}_j$ where p/r and $(1-p)/(1-r)$ are not weighted.

In practice, p is unknown. To obtain the feasible weighted MLE, Hsieh et al. (1985) assigned the sample average $\tilde{p}$ to p ,

$$\tilde{p} = \frac{1}{n}\sum_{i=1}^{n} y_{2i} \ . \tag{2.9}$$

This method consistently estimates propensity scores. However, we eventually need external knowledge about $(n,\ n_1)$ regarding the random sample for $\tilde{p}$ when estimating with the choice-based sample, m . As described in the previous study, there may be few datasets that cover all persons, including information necessary to determine whether a person is eligible for the program. Hence, n is assumed to be unavailable or unknown. Then, a feasible weighted MLE would be difficult to calculate.

Hence, we next consider another important estimation method described by Manski and MacFadden (1981). They proposed the following log-likelihood function, given the true value p :

$$l(\boldsymbol{\beta}) = \sum_{j=1}^{m} \log \frac{[\lambda_1 p_j]^{y_{2j}} [\lambda_0 (1 - p_j)]^{1 - y_{2j}}}{\lambda_1 p_j + \lambda_0 (1 - p_j)} \ , \tag{2.10}$$

where $\lambda_1 = r/p$ and $\lambda_0 = (1-r)/(1-p)$. Cosslett (1981) proposed the generalized choice-based sampling method, which for the binomial model is reduced to estimate λ_1

and λ_0 as unknown parameters. When normalizing with $\lambda_0 = 1$, $\lambda = \lambda_1/\lambda_0$ becomes the unknown parameter. Then the log-likelihood function of Cosslett (1981) becomes

$$l(\boldsymbol{\beta},\ \lambda) = \sum_{j=1}^{m} y_{2j} \log \lambda p_j + (1 - y_{2j}) \log(1 - p_j) - \log\left(\lambda p_j + (1 - p_j)\right) \ . \tag{2.11}$$

The MLE $\hat{\lambda}$ is the consistent estimator for the following odds ratio:

$$\hat{\lambda} \xrightarrow{p} \frac{1-p}{p} \frac{r}{1-r} \ . \tag{2.12}$$

Note that because r is known, by counting backwards, we can obtain the consistent estimator for p by only subsamples:

$$\hat{p} = \frac{r}{r + \hat{\lambda}(1-r)} \ . \tag{2.13}$$

Although p is needed to identify the ATE, $\hat{p}$ can replace the sample average $\tilde{p}$. Cosslett's MLE is essential in our approach, as $\hat{p}$ plays an important role in extending the IPW estimation discussed next.

2.2 Cosslett's MLE under a random subsample size

Importantly, the way subsamples are drawn affects the asymptotic distribution of the estimator. To characterize the choice-based sample design, we introduce sampling dummies $(d_{1i},\ d_{0i})$ and the sampling probabilities $(q_1,\ q_0)$. If sample i is included in subsample m in the treatment group, it is represented as $d_{1i} = 1$ and 0 otherwise. If the sample i is included in subsample m in the control group, it is represented as $d_{0i} = 1$ and 0 otherwise. We put $\mathbf{d}_i = (d_{1i},\ d_{i0})'$ and $\mathbf{x}_i^* = (\mathbf{x}_i',\ y_{1i}^*,\ y_{i0}^*)'$.

Assumption 1: $q_1 = \Pr(d_{1i} = 1 | y_{2i} = 1) > 0$, $q_0 = \Pr(d_{0i} = 1 | y_{2i} = 0) > 0$, *and* $\mathbf{d}_i \perp\!\!\!\perp \mathbf{x}_i^* \mid y_{2i}$.
Assumption 2: $\{\mathbf{d}_i,\ \mathbf{x}_i^*,\ u_{2i}\}_{i=1}^n$ *are independent and identically distributed.*

Assumption 1 implies choice-based or endogenous sampling in which sampling does not depend on the exogenous variables $\mathbf{x}_i$. That is, it depends only on the endogenous variable y_{2i} . Because there are the latent variables, including $(y_{1i}^*,\ y_{i0}^*)$, $\mathbf{x}_i^*$ is conditionally independent of $\mathbf{d}_i$.

In choice-based sampling, there are two ways to consider the numbers of subsamples $(m_1,\ m_0)$ as constant sequences or random variables. If $(m_1,\ m_0)$ are fixed, then conditioning $\sum_{i=1}^n d_{1i} y_{2i} = m_1$ and $\sum_{i=1}^n d_{0i}(1 - y_{2i}) = m_0$ means that the samples, $i = 1, \cdots, n$, are not mutually independent.

Assumption 3: $\sum_{i=1}^n d_{1i} y_{2i}$ *and* $\sum_{i=1}^n d_{0i}(1 - y_{2i})$ *are not conditioned.*

This study considers the case in which the random variables $(m_1,\ m_0)$ are not given. As

such, the results are derived from standard asymptotic theory for the sum of independent random variables. For instance, if the number of the treatment group is small and relatively rare, all of its samples may be collected. Hence, because $q_1 = 1$ or $m_1 = n_1$, it is suitable to let m_1 be random. When $(m_1,\ m_0)$ are random variables, the ratio, $r = m_1/(m_1 + m_0)$, is also a random variable. Therefore, we add the asymptotic evaluation of $\sqrt{n}(r - r_0)$ to Cosslett's MLE, where the probability limit r_0 of r is

$$r_0 = \frac{q_1 p}{q_1 p + q_0(1-p)} , \tag{2.14}$$

as shown in the Appendix.

Based on the invariance property, we directly estimate p as follows:

$$\begin{aligned} l(\boldsymbol{\beta},\ p) &= \sum_{j=1}^{m} y_{2j} \log \frac{r}{p} p_j + (1 - y_{2j}) \log \frac{1-r}{1-p}(1-p_j) - \log \left(\frac{r}{p} p_j + \frac{1-r}{1-p}(1-p_j) \right) \\ &= \sum_{j=1}^{m} l_j(\boldsymbol{\psi}) \quad (, say). \end{aligned} \tag{2.15}$$

Thus, with the maximization point as $\hat{\boldsymbol{\psi}} = (\hat{\boldsymbol{\beta}}_2',\ \hat{p})'$, the estimators of the propensity scores $\hat{p}_j = \Phi(\hat{\boldsymbol{\beta}}' \mathbf{x}_j)$ and $\hat{p}$ are simultaneously obtained. Notably, Cosslett's MLE includes the probit MLE as a special case, i.e., $\hat{p} = r$. Hence, even if we can assume random subsampling, applying Cosslett's MLE provides a more robust estimation.

We prepare the notation $l(\boldsymbol{\beta},\ p) = \sum_{i=1}^{n} l_i(\boldsymbol{\psi}; r) d_i$ with subscript i for the following assumptions, where

$$l_i(\boldsymbol{\psi}; r) = \log r_i^{d_{1i} y_{2i}} (1 - r_i)^{d_{0i}(1 - y_{2i})} , \tag{2.16}$$

$$r_i = \frac{r p^{-1} p_i}{r p^{-1} p_i + (1-r)(1-p)^{-1}(1-p_i)} \tag{2.17}$$

$$= \frac{q_1 p_i}{q_1 p_i + q_0(1-p_i)} + o_p(1) , \tag{2.18}$$

and $d_i = d_{1i} y_{2i} + d_{0i}(1 - y_{2i})$, in Equation (2.18), r is evaluated at r_0 . Thus $l(\boldsymbol{\beta},\ p)$ is asymptotically equivalent to the conditional log-likelihood function conditional on subject i being drawn, i.e., given $d_i = 1$ and $\mathbf{x}_i$.

Assumption 4: (i) *The parameter space of* $(\boldsymbol{\psi}',\ r_0)$ *is compact, and the true value of* $\boldsymbol{\psi}$ *is an interior point.* (ii) *Given* $d_i = 1$, $\boldsymbol{\psi} \neq \boldsymbol{\psi}^*$ *implies* $l_i(\boldsymbol{\psi}; r_0) \neq l_i(\boldsymbol{\psi}^*; r_0)$. (iii) *For some* $\epsilon > 0$, $\epsilon \leq p_i \leq 1 - \epsilon$ *w.p.1* .

These assumptions are similar to the ones made by Cosslett (1981). The identification condition of (ii) implicitly requires that $\mathbf{x}_i$ is not multicollinear to identify coefficients $\boldsymbol{\beta}$. Condition (ii) also requires that at least one covariate, x_i, exists and that its coefficient is non-zero. If there is only a constant term β_1 , then $p_i = \Phi(\beta_1) \propto p$, and p is unidentifiable from Equation (2.17). For instance, it is sufficient if x_i is a dummy variable and p_i takes two values, then p is identifiable. It is a mild condition that $\Pr(y_{2i} = 1 | \mathbf{x}_i)$

is not a constant; thus, p can be identified, even within biased subsamples. Condition (iii) is referred to as the strict overlap assumption in the causal inference literature, and it requires the index $\boldsymbol{\beta}'\mathbf{x}_i$ or $\mathbf{x}_i$ be bounded random variables. This condition applies to the IPW estimator described below, but it is also employed as the condition for the existence of moments related to the MLE.

The result of Cosslett (1981) is slightly changed to match the case of a random subsample size and parameterization to p :

Theorem 1: (i) *Under Assumptions 1-4, as* $n \to \infty$, $\hat{\boldsymbol{\psi}} \xrightarrow{p} \boldsymbol{\psi}$ *and*

$$\sqrt{m_1 + m_0}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}) \ , \tag{2.19}$$

where $\boldsymbol{\psi} = (\boldsymbol{\beta}', \ p)'$ *and*

$$\boldsymbol{\Omega} = \boldsymbol{\Psi}^{-1}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_{12} + \boldsymbol{\Sigma}_2)\boldsymbol{\Psi}^{-1} \ . \tag{2.20}$$

(ii) *Under the same assumptions and the null hypothesis* $H_0 : \ \psi_k = \psi_{0k}$,

$$t_k = \frac{\hat{\psi}_k - \psi_{0k}}{\hat{\sigma}_k} \ \xrightarrow{d} \ \mathcal{N}(0, 1) \ , \tag{2.21}$$

where $\hat{\sigma}_k$ *refers to the standard error of* $\hat{\psi}_k$ *for* $k = 1, \ \cdots \ , \ K + 1$.

For the asymptotic variance-covariance matrix of (i), definitions from $\boldsymbol{\Psi}$ to $\boldsymbol{\Sigma}_2$ are given in Equations (A.1) and (A.3) in the Appendix. Although $\boldsymbol{\Psi}$ and $\boldsymbol{\Sigma}_1$ correspond to the Hessian and the squares of the score function, respectively, the covariance $\boldsymbol{\Sigma}_{12}$ and variance $\boldsymbol{\Sigma}_2$ are added by Assumption 3.

With respect to the result of (ii), $\boldsymbol{\Omega}$ depends on the nuisance parameters $(q_1, \ q_0) \simeq (m_1/n_1, \ m_0/(n - n_1))$, however, the test statistic is also constructed without relying on $(n, \ n_1)$. The definition of $\hat{\sigma}_k^2$ is as follows:

$$\hat{\sigma}_k^2 = \mathbf{e}_k' \hat{\boldsymbol{\Psi}}^{-1} \hat{\boldsymbol{\Sigma}} \hat{\boldsymbol{\Psi}}^{-1} \mathbf{e}_k \ , \tag{2.22}$$

where $\mathbf{e}_k = (0, \ \cdots \ , \ 1, \ \cdots \ , \ 0)'$ is the vector with only the k-th element as 1 ,

$$\hat{\boldsymbol{\Psi}} \ = \ \frac{\partial^2 l(\hat{\boldsymbol{\psi}})}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}'} \ , \tag{2.23}$$

$$\hat{\boldsymbol{\Sigma}} \ = \ \sum_{j=1}^{m} \left(\frac{\partial l_j(\hat{\boldsymbol{\psi}})}{\partial \boldsymbol{\psi}} + \hat{\boldsymbol{\gamma}}_1 (y_{2j} - r) \right) \left(\frac{\partial l_j(\hat{\boldsymbol{\psi}})}{\partial \boldsymbol{\psi}} + \hat{\boldsymbol{\gamma}}_1 (y_{2j} - r) \right)' \ , \text{ and} \tag{2.24}$$

$$\hat{\boldsymbol{\gamma}}_1 \ = \ \frac{1}{m_1 + m_0} \frac{\partial^2 l(\hat{\boldsymbol{\psi}})}{\partial \boldsymbol{\psi} \partial r} \ . \tag{2.25}$$

The t-value of the M-estimator is expressed as

$$\frac{\sqrt{m}(\hat{\psi}_k - \psi_{0k})}{\sqrt{\mathbf{e}_k' (\frac{1}{m}\hat{\boldsymbol{\Psi}})^{-1} \frac{1}{m} \hat{\boldsymbol{\Sigma}} (\frac{1}{m}\hat{\boldsymbol{\Psi}})^{-1} \mathbf{e}_k}} = \frac{\hat{\psi}_k - \psi_{0k}}{\hat{\sigma}_k} \ , \tag{2.26}$$

where the normalization $1/m$ of $\hat{\boldsymbol{\gamma}}_1$ cannot be omitted on both sides of Equation (2.26). Meanwhile, standard error $\hat{\sigma}_k$ is computed without information on $(n,\ n_1)$, and the sampling probabilities $(q_1,\ q_0)$ can be unknown. Following the consistent estimation of $\boldsymbol{\beta}$ or the propensity score, we conclude that the asymptotic t-test is feasible given only the choice-based samples.

The null hypothesis of the t-test for the coefficient is usually $\mathrm{H}_0 : \beta_k = 0$, however, $\mathrm{H}_0 : p = 0$ is meaningless. On the other hand, the 95% confidence interval, $\hat{p} \pm 1.96\hat{\sigma}_{K+1}$, would be useful for predicting the population ratio from the choice-based samples. Moreover, using $t \propto \hat{p} - r$, we can consider the specification test for being random subsampling as $\mathrm{H}_0 :\ p = r_0$.

2.3 IPW for choice-based samples

This section describes the estimation and testing methods of the ATE under choice-based samples, which is our main interest. For the IPW estimator, it is insufficient to consistently estimate the propensity score $1/p_j$ as the inverse weight. Hence, we add $1/q_1$ as an inverse weight:

$$\begin{aligned} \frac{1}{q_1}\mathcal{E}\left[\frac{d_{1i}y_{1i}}{p_i}\right] &= \frac{1}{q_1}\mathcal{E}\left[\frac{\mathcal{E}[d_{1i}|y_{2i}=1]y_{1i}^* y_{2i}}{p_i}\right] \\ &= \mathcal{E}\left[\frac{\mathcal{E}[y_{1i}^*|\mathbf{x}_i]\mathcal{E}[y_{2i}|\mathbf{x}_i]}{p_i}\right] \\ &= \mathcal{E}[y_{1i}^*] \ . \end{aligned} \tag{2.27}$$

Similarly, $q_0^{-1}\mathcal{E}\left[d_{i0}y_{0i}(1-p_i)^{-1}\right] = \mathcal{E}[y_{0i}^*]$, where the first equality of (2.27) follows from Assumption 1, and the second equality is from the following assumption,

Assumption 5: $(y_{0i}^*,\ y_{1i}^*) \perp\!\!\!\perp y_{2i} \mid \mathbf{x}_i$.

This condition set by Rosenbaum and Rubin (1983) is called the ignorability assumption in the causal inference literature. Replacing the expectation with the sample average, it follows that

$$\frac{1}{q_1}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{d_{1i}y_{1i}}{p_i}\right) = \left(\frac{m_1}{np}\right)^{-1}\left(\frac{1}{n}\sum_{j=1}^{m}\frac{y_{1j}}{p_j}\right) + o_p(1) \ . \tag{2.28}$$

Thus, we construct the estimator as the sample analogue of (2.27):

$$\hat{\tau} = \frac{1}{m}\sum_{j=1}^{m}\frac{\hat{p}}{r}\frac{y_{1j}}{\hat{p}_j} - \frac{1-\hat{p}}{1-r}\frac{y_{0j}}{1-\hat{p}_j} \ . \tag{2.29}$$

The difference from $\tilde{\tau}$ is that Cosslett's MLE $\hat{\boldsymbol{\beta}}$ is employed to estimate the propensity score at the first stage. In the second stage, Cosslett's MLE $\hat{p}$ is employed and ($\hat{p}/r, (1-\hat{p})/(1-r)$) are also added as inverse weights. We may call $\hat{\tau}$ the choice-based IPW

(CIPW) estimator to distinguish it from $\tilde{\tau}$.

Assumption 6: *The second order moment of* $(y_{1i}^*,\ y_{0i}^*)$ *exists.*

Alongside the overlap assumption of Assumption 4 (iii), the variance of $\hat{\tau}$ is guaranteed to be finite, and the asymptotic normality holds.

Theorem 2: (i) *Under Assumptions 1-6, as* $n \to \infty$, $\hat{\tau} \xrightarrow{p} \mathcal{E}[y_{1i}^* - y_{0i}^*]$ *and*

$$\sqrt{m}(\hat{\tau} - \frac{n}{m}\mu_\tau) \xrightarrow{d} \mathcal{N}(0, \sigma^2) \ , \tag{2.30}$$

where $\mu_\tau = (q_1 p + q_0(1-p))\tau$.

(ii) *Under the same assumptions and the null hypothesis* $H_0 :\ \mathcal{E}[y_{1i}^* - y_{0i}^*] = 0$,

$$t_\tau \ = \ \frac{\sqrt{m}}{\hat{\sigma}}\hat{\tau} \ \xrightarrow{d} \mathcal{N}(0,1) \ , \tag{2.31}$$

where $\hat{\sigma}$ *stands for the standard error of* $\sqrt{m}\hat{\tau}$.

The definition of σ^2 in result (i) is given by (A.12) in the Appendix. Notably, the normalization of $\hat{\tau}$ depends on n; however, that of t_τ does not depend on n under $\mathrm{H}_0 : \tau = 0$. The estimate of the individual causal effect is expressed as

$$\hat{\tau}_j = \frac{\hat{p}}{r}\frac{y_{1j}}{\hat{p}_j} - \frac{1-\hat{p}}{1-r}\frac{y_{0j}}{1-\hat{p}_j} \ . \tag{2.32}$$

Then, regarding the definition of standard error in result (ii),

$$\hat{\sigma}^2 = \frac{1}{m}\sum_{j=1}^{m}\left(\hat{\tau}_j + \hat{\gamma}_2(y_{2j} - r) + \hat{\boldsymbol{\gamma}}_3'\frac{\partial l_j(\hat{\boldsymbol{\psi}})}{\partial \boldsymbol{\psi}}\right)^2 \ . \tag{2.33}$$

The leading term corresponds to the sample variance of $\hat{\tau}_j$. The second term is the variation caused by the random subsample size, and the third term relates to the use of estimates $\hat{\boldsymbol{\psi}}$, where $(\hat{\gamma}_2,\ \hat{\boldsymbol{\gamma}}_3') = (\hat{\delta}_r + \hat{\boldsymbol{\gamma}}_1'\hat{\boldsymbol{\gamma}}_3,\ -(\hat{\boldsymbol{\delta}}_\beta',\ \hat{\delta}_p)(m^{-1}\hat{\boldsymbol{\Psi}})^{-1})$,

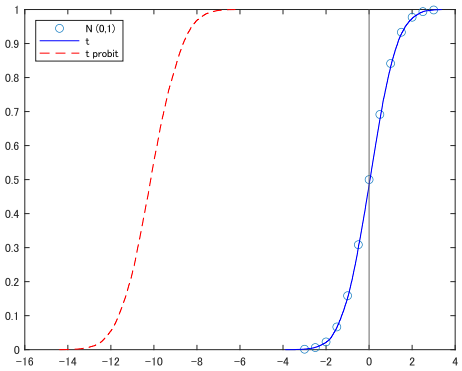
$$\hat{\delta}_r \ = \ \frac{1}{m}\sum_{j=1}^{m} -\frac{\hat{p}}{r^2}\frac{y_{1j}}{\hat{p}_j} - \frac{1-\hat{p}}{(1-r)^2}\frac{y_{0j}}{1-\hat{p}_j} \ , \tag{2.34}$$

$$\hat{\boldsymbol{\delta}}_\beta \ = \ \frac{1}{m}\sum_{j=1}^{m}\left(-\frac{\hat{p}}{r}\frac{y_{1j}}{\hat{p}_j^2} - \frac{1-\hat{p}}{1-r}\frac{y_{0j}}{(1-\hat{p}_j)^2}\right)\phi(\hat{\boldsymbol{\beta}}'\mathbf{x}_j)\mathbf{x}_j \ , \tag{2.35}$$

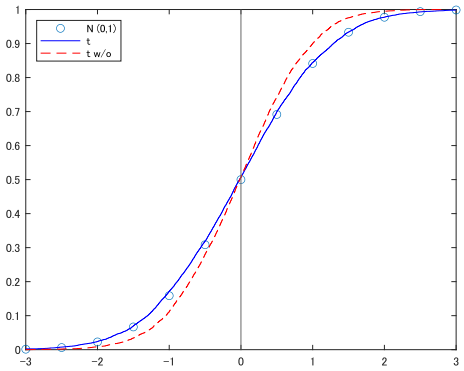
$$\hat{\delta}_p \ = \ \frac{1}{m}\sum_{j=1}^{m}\frac{1}{r}\frac{y_{1j}}{\hat{p}_j} + \frac{1}{1-r}\frac{y_{0j}}{1-\hat{p}_j} \ , \tag{2.36}$$

and ϕ represents the standard normal density function. The t-statistic t_τ depends only on the observable normalizer m . Therefore, the consistent estimation and significance test for the ATE can be performed without external knowledge of $(n,\ n_1)$ from the random sample.

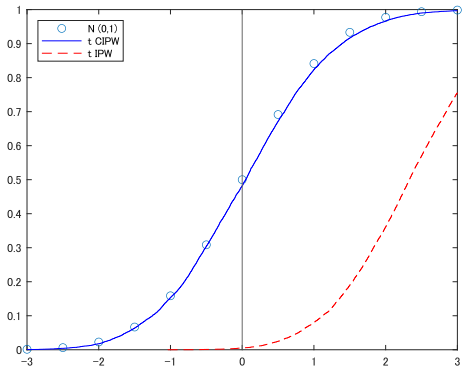
3 Monte Carlo experiments



(a) β



(b) p



(c) τ

Fig. 1: Empirical cumulative distribution functions of t-statistics

The finite sample properties of a simple simulation are explained forthwith. The exogenous variable is generated by $x_i \sim \mathcal{B}e(\alpha^*,\ \beta^*) - 0.5$, and the true value is $\beta = 5$. Then, the support of βx_i becomes $[-2.5,\ 2.5]$, where $\epsilon = \Phi(-2.5) = 0.006$. In an application scenario, p can be considered small. When the beta distribution is $(\alpha^*, \beta^*) = (2, 4.19)$, the right tail is longer and p is 0.25. When $(\alpha^*, \beta^*) = (2, 12.57)$, p is 0.05. As the true value p cannot be obtained analytically, it is approximated by $p = N^{-1} \sum_{i=1}^{N} \Phi(\beta x_i)$ for $N = 10^4 \times n$. Then the effect on the asymptotic distribution is $O_p(\sqrt{n/N})$, which is negligible. These true values are commonly used with 10000 iterations. Given $(n_1,\ n_0)$, subsamples $(m_1,\ m_0)$ are drawn using uniform random numbers, $d_{1i} = \mathbb{1}\{u_{1i} \leq q_1\}$ and $d_{0i} = \mathbb{1}\{u_{0i} \leq q_0\}$. On average, the set-

ting is $r \simeq 0.5$ for 10000 iterations. When $(p,\ q_1,\ q_0) = (0.25,\ 0.80,\ 0.267)$ and $n = 1000$, it follows that $(m_1,\ m_0) \simeq (200,\ 200)$ on average. If $n = 4000$, we obtain $(m_1,\ m_0) \simeq (800,\ 800)$. When $(p,\ q_1,\ q_0) = (0.05,\ 1.00,\ 0.053)$ and $n = 4000$, it holds that $(m_1,\ m_0) \simeq (200,\ 200)$ on average. If $n = 16000$, we obtain $(m_1,\ m_0) \simeq (800,\ 800)$.

The latent variables are $y_{1i}^* = \beta_1^* x_i + u_{1i}^*$ and $y_{0i}^* = \beta_0^* x_i + u_{0i}^*$, where $(u_{1i}^*,\ u_{0i}^*)$ both follow the standard normal distribution. Under $\mathrm{H}_0 : \tau = 0$, the bias of the IPW estimator is approximated by

$$\mathcal{E}[\tilde{\tau}] - 0 \simeq q_1 \beta_1^* \mathcal{E}\left[x_i \frac{p_i}{\tilde{p}_i}\right] - q_0 \beta_1^* \mathcal{E}\left[x_i \frac{1-p_i}{1-\tilde{p}_i}\right] . \tag{3.37}$$

If $q_1 = q_0$ or $p = r_0$, then choice-based sampling is reduced to random sampling. Under $q_1 \neq q_0$, there are two sources of bias: $q_1 \neq q_0$ and $p_i \neq \tilde{p}_i$. When $\beta_1^* = \beta_0^* = 0$, there is no bias by chance; hence, we set $\beta_1^* = 1.5$.

Figure 1 illustrates the empirical cumulative distribution functions of the t-statistic for $p = 0.25$ and $m_1 \simeq 800$. Figure 1-(a) shows that for coefficient β of the propensity scores, the asymptotic t-distribution based on $\tilde{\beta}$ is shifted from the reference distribution because the probit MLE $\tilde{\beta}$ is biased. Meanwhile, the asymptotic t distribution based on Cosslett's MLE $\hat{\beta}$ is well approximated by $\mathcal{N}(0,\ 1)$. Figure 1-(b) is for the population ratio p . Because the probit MLE does not estimate p, we compare t_{K+1} and t_{K+1} without $\mathbf{\Sigma}_{12}$ and $\mathbf{\Sigma}_2$ in Theorem 1 (t w/o). It can be seen that these additional terms make the standardization more precise. Figure 1-(c) illustrates the significance test for ATE τ . The t distribution of the IPW estimator deviates from the reference distribution, and that of the CIPW estimator t_τ is well approximated.

Table 1 list the bias and standard deviation (SD) of the IPW estimators and the actual size, $\Pr(|t_\tau| \geq 1.96)$, of the corresponding t-test statistics under $\mathrm{H}_0 : \tau = 0$. From the table we can see that the CIPW estimator has less bias than the IPW estimator. Although the nominal size is 0.05, size distortions occur for the t-statistics of the IPW estimator for both $p = 0.25$ and $p = 0.05$. As m increases, the size based on the CIPW estimator approaches the nominal size and the convergence in distribution is confirmed. Conversely, as m increases, the size based on the IPW estimator increases due to the inconsistency, i.e., the Type I error is not controlled.

Table 1: Finite sample properties of IPW estimators and t-statistics

		IPW			CIPW	
$r \simeq 0.5,\ \tau = 0$	Bias	SD	Size	Bias	SD	Size
$p = 0.25,\ m_1 \simeq 200$	0.144	0.113	0.200	-0.0001	0.173	0.064
800	0.145	0.057	0.656	0.0010	0.084	0.059
$p = 0.05,\ m_1 \simeq 200$	0.138	0.102	0.272	-0.0004	0.131	0.053
800	0.137	0.050	0.768	0.0001	0.064	0.050

4 Conclusions

The Cosslett's MLE includes the probit MLE as a special case, which provides the robust estimation for the propensity score in causal inference. This study applied Cosslett's MLE to identify the treatment group proportion of the population based only on choice-based samples. By adding the estimated ratios as inverse weights, we extended the IPW estimator to a choice-based samples version. The consistent estimation and significant test for ATE can be performed without external information on the original random sample. Thus, program evaluation would be made easier even when external information is difficult to obtain.

Appendix

Proof of Theorem 1: **(i)** r is expressed as $n^{-1}\sum_{i=1}^{n} d_{1i}y_{2i}/(n^{-1}\sum_{i=1}^{n} d_{1i}y_{2i}+n^{-1}\sum_{i=1}^{n} d_{0i}(1-y_{2i}))$, and the law of large numbers holds because $\mathcal{E}[(d_{1i}y_{2i})^2] < \infty$. Then, $r \xrightarrow{p} r_0$ as $\mathcal{E}[d_{1i}y_{2i}] = q_1 p$. The term involving only r in $n^{-1}l(\boldsymbol{\psi})$ converges in probability to $\mathcal{E}[d_{1i}y_{2i}]\log r_0 = \mathcal{E}[d_{1i}y_{2i}\log r_0]$. Then, using Assumptions 4 (i) and (iii), $n^{-1}\sum_{i=1}^{n} l_i(\boldsymbol{\psi};r)d_i \xrightarrow{p} \mathcal{E}[l_i(\boldsymbol{\psi};r_0)d_i] < \infty$ uniformly in $\boldsymbol{\psi}$. Because $n^{-1}l(\boldsymbol{\psi})$ is the conditional log-likelihood function when $r = r_0$, $\mathcal{E}[l_i(\boldsymbol{\psi}^*;r_0)|d_i = 1,\ \mathbf{x}_i] \leq \mathcal{E}[l_i(\boldsymbol{\psi};r_0)|d_i = 1,\ \mathbf{x}_i]$ by Jensen's inequality. Moreover, the inequality strictly holds due to Assumption 4 (ii). Thus, $\mathcal{E}[\mathcal{E}[l_i(\boldsymbol{\psi}^*;r_0)|d_i = 1,\ \mathbf{x}_i]d_i] < \mathcal{E}[\mathcal{E}[l_i(\boldsymbol{\psi};r_0)|d_i = 1,\ \mathbf{x}_i]d_i]$, because $\mathcal{E}[d_i|\mathbf{x}_i] = q_1 p_i + q_0(1-p_i) > 0$ due to Assumption 1. As the maximum point is unique at the true value, from the arguments of the consistency for the M-estimator, it follows that $\hat{\boldsymbol{\psi}} \xrightarrow{p} \boldsymbol{\psi}$.

From the Taylor expansion with respect to $\boldsymbol{\psi}$, $\mathbf{0} = m^{-\frac{1}{2}}\sum_{j=1}^{m} \partial l_j(\boldsymbol{\psi})/\partial\boldsymbol{\psi} + \boldsymbol{\Psi}\sqrt{m}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}) + o_p(1)$, where $\Pr(d_i = 1) = q_1 p + q_0(1-p) > 0$ and

$$\boldsymbol{\Psi} = \frac{1}{\Pr(d_i = 1)}\mathcal{E}\left[d_i \frac{\partial^2 l_i(\boldsymbol{\psi};r_0)}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}'}\right] . \tag{A.1}$$

This is because from $(m/n,\ r,\ \hat{\boldsymbol{\psi}}')' \xrightarrow{p} (q_1 p + q_0(1-p),\ r_0,\ \boldsymbol{\psi}')'$, we have $(n/m)n^{-1}\sum_{i=1}^{n} d_i\partial^2 l_i(\hat{\boldsymbol{\psi}} + o_p(1);r)/\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}' = (n/m)\mathcal{E}[d_i l_i(\boldsymbol{\psi};r_0)/\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}'] + o_p(1)$. The true value is a local maximum point, hence, $\partial^2\mathcal{E}[l_i(\boldsymbol{\psi};r_0)d_i]/\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}'$ is negative-definite. Under the Assumption 4, all elements of $l_i(\boldsymbol{\psi};r_0)/\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}'$ are bounded; thus, differentiation and integration are interchangeable. Then, as shown in Akashi and Horie (2022), $\boldsymbol{\Psi}$ is also negative-definite or invertible. Therefore, $\sqrt{m}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}) = -\boldsymbol{\Psi}^{-1}m^{-\frac{1}{2}}\sum_{j=1}^{m} \partial l_j(\boldsymbol{\psi})/\partial\boldsymbol{\psi} + o_p(1)$. By expanding at r_0 , since the numerator of r is also the sum of independent random variables,

$$\begin{aligned}
\frac{1}{\sqrt{m}}\sum_{j=1}^{m}\frac{\partial l_j(\boldsymbol{\psi})}{\partial\boldsymbol{\psi}} &= \frac{1}{\sqrt{m}}\sum_{i=1}^{n} d_i\frac{\partial l_i(\boldsymbol{\psi};r_0)}{\partial\boldsymbol{\psi}} + \frac{1}{m}\sum_{i=1}^{n} d_i\frac{\partial^2 l_i(\boldsymbol{\psi};r_0)}{\partial\boldsymbol{\psi}\partial r}\sqrt{m}(r - r_0) + o_p(1) \\
&= \sqrt{\frac{n}{m}}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{s}_i + o_p(1) ,
\end{aligned} \tag{A.2}$$

where $\mathbf{s}_i = d_i \partial l_i(\boldsymbol{\psi}; r_0)/\partial \boldsymbol{\psi} + \boldsymbol{\gamma}_1 (d_{1i} y_{2i} - d_i r_0)$ and $\boldsymbol{\gamma}_1 = \Pr(d_i = 1)^{-1} \mathcal{E}[d_i \partial^2 l_i(\boldsymbol{\psi}; r_0)/\partial \boldsymbol{\psi} \partial r]$. The conditional log-likelihood function implies that $\mathcal{E}[\partial l_i(\boldsymbol{\psi}; r_0)/\partial \boldsymbol{\psi} | d_i = 1,\ \mathbf{x}_i] = \mathbf{0}$. Hence, it holds that $\mathcal{E}[\mathbf{s}_i] = \mathbf{0}$ because $\mathcal{E}[d_i r_0] = q_1 p$. From Assumptions 2, and 3, and the Lindeberg-Lévy central limit theorem, $(n/m)^{\frac{1}{2}} n^{-\frac{1}{2}} \sum_{i=1}^{n} \mathbf{s}_i \xrightarrow{d} \mathcal{N}(\mathbf{0},\ \boldsymbol{\Sigma})$ as $n \to \infty$, where $\boldsymbol{\Sigma} = \Pr(d_i = 1)^{-1} \mathcal{E}[\mathbf{s}_i \mathbf{s}_i']$. Therefore, for the notations of the asymptotic covariance matrix of Theorem 1, we have

$$
\begin{aligned}
\boldsymbol{\Sigma}_1 &= \frac{1}{p_d} \mathcal{E}\left[d_i \frac{\partial l_i(\boldsymbol{\psi}; r_0)}{\partial \boldsymbol{\psi}} \frac{\partial l_i(\boldsymbol{\psi}; r_0)}{\partial \boldsymbol{\psi}'} \right] , \quad \boldsymbol{\Sigma}_2 = \frac{\mathcal{E}[(d_{1i} y_{2i} - d_i r_0)^2]}{p_d} \boldsymbol{\gamma}_1 \boldsymbol{\gamma}_1' , \text{ and} \\
\boldsymbol{\Sigma}_{12} &= \frac{1}{p_d} \mathcal{E}\left[d_{1i} y_{2i} \left(\frac{\partial l_i(\boldsymbol{\psi}; r_0)}{\partial \boldsymbol{\psi}} \boldsymbol{\gamma}_1' + \boldsymbol{\gamma}_1 \frac{\partial l_i(\boldsymbol{\psi}; r_0)}{\partial \boldsymbol{\psi}'} \right) \right] , \qquad \text{(A.3)}
\end{aligned}
$$

where $p_d = \Pr(d_i = 1)$, $d_i (d_{1i} y_{2i}) = d_{1i} y_{2i}$, and $\mathcal{E}[(d_{1i} y_{2i} - d_i r_0)^2] = q_1 p (1 - r_0)$. Thus, we obtain the representation that $\boldsymbol{\Omega} = \boldsymbol{\Psi}^{-1} (\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_{12} + \boldsymbol{\Sigma}_2) \boldsymbol{\Psi}^{-1}$. ■

(ii) From the above arguments, $m^{-1} \hat{\boldsymbol{\Psi}} \xrightarrow{p} \boldsymbol{\Psi}$, $\hat{\boldsymbol{\gamma}}_1 \xrightarrow{p} \boldsymbol{\gamma}_1$, and $m^{-1} \hat{\boldsymbol{\Sigma}} \xrightarrow{p} \boldsymbol{\Sigma}$. By Equation (2.26), the desired result is obtained. ■

Proof of Theorem 2: (i) For the first term $\hat{\tau}_1$ of $\hat{\tau}$, from $\hat{p} - \tilde{p} \xrightarrow{p} 0$ and $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$, we have $\hat{\tau}_1 = (\hat{p}/m_1) \sum_{j=1}^{m} y_{1j}/\hat{p}_j = (n_1/m_1) n^{-1} \sum_{i=1}^{n} d_{1i} y_{1i}^* y_{2i}/p_i + o_p(1)$. By the law of large numbers and Assumption 1,

$$
\hat{\tau}_1 = \frac{n_1}{m_1} \mathcal{E}\left[\frac{\mathcal{E}[d_{1i} | y_{2i},\ \mathbf{x}_i^*] y_{1i}^* y_{2i}}{p_i} \right] + o_p(1) = \frac{n_1}{m_1} q_1 \mathcal{E}[y_{1i}^*] + o_p(1) . \qquad \text{(A.4)}
$$

$m_1/n_1 = q_1 + o_p(1)$; hence, $\hat{\tau}_1 \xrightarrow{p} \mathcal{E}[y_{1i}^*]$. Similarly, for the second term, $\hat{\tau}_0 = (1 - \hat{p})/m_0 \sum_{j=1}^{m} y_{0j}/(1 - \hat{p}_j) \xrightarrow{p} \mathcal{E}[y_{0i}^*]$. Therefore, $\hat{\tau} = \hat{\tau}_1 - \hat{\tau}_0 \xrightarrow{p} \tau$.

We next show the asymptotic normality. By expanding at $(r_0,\ \boldsymbol{\beta}',\ p)$, under the assumptions,

$$
\sqrt{m}(\hat{\tau} - \frac{n}{m} p_d \tau) = e_\tau + \delta_r \sqrt{m}(r - r_0) + \boldsymbol{\delta}_\beta' \sqrt{m}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \delta_p \sqrt{m}(\hat{p} - p) + o_p(1) , \qquad \text{(A.5)}
$$

where $\mathcal{E}[\sqrt{m} e_\tau] = 0$ and the law of large numbers leads to

$$
\begin{aligned}
e_\tau &= \frac{1}{\sqrt{m}} \sum_{i=1}^{n} \frac{p}{r_0} \frac{d_{1i} y_{1i}}{p_i} - \frac{1 - p}{1 - r_0} \frac{d_{0i} y_{0i}}{1 - p_i} - p_d \tau , \qquad \text{(A.6)} \\
\delta_r &= \frac{1}{p_d} \mathcal{E}\left[-\frac{p}{r_0^2} \frac{d_{1i} y_{1i}}{p_i} - \frac{1 - p}{(1 - r_0)^2} \frac{d_{0i} y_{0i}}{1 - p_i} \right] , \qquad \text{(A.7)} \\
\boldsymbol{\delta}_\beta &= \frac{1}{p_d} \mathcal{E}\left[\left(-\frac{p}{r_0} \frac{d_{1i} y_{1i}}{p_i^2} - \frac{1 - p}{1 - r_0} \frac{d_{0i} y_{0i}}{(1 - p_i)^2} \right) \phi(\boldsymbol{\beta}' \mathbf{x}_i) \mathbf{x}_i \right] , \text{ and} \qquad \text{(A.8)} \\
\delta_p &= \frac{1}{p_d} \mathcal{E}\left[\frac{1}{r_0} \frac{d_{1i} y_{1i}}{p_i} + \frac{1}{1 - r_0} \frac{d_{0i} y_{0i}}{1 - p_i} \right] . \qquad \text{(A.9)}
\end{aligned}
$$

Moreover, expressing $r - r_0$ and $\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}$ as asymptotically linear forms, we have

$$\begin{aligned}\sqrt{m}(\hat{\tau} - \frac{n}{m}p_d\tau) &= e_\tau + \frac{\delta_r}{\sqrt{m}}\sum_{i=1}^{n}(d_{1i}y_{2i} - d_i r_0) - (\boldsymbol{\delta}'_\beta,\ \delta_p)\boldsymbol{\Psi}^{-1}\frac{1}{\sqrt{m}}\sum_{i=1}^{n}\mathbf{s}_i + o_p(1) \\ &= \sqrt{\frac{n}{m}}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}e_i + o_p(1)\ , \end{aligned} \tag{A.10}$$

where

$$e_i = \left(\frac{p}{r_0}\frac{d_{1i}y_{2i}}{p_i} - \frac{1-p}{1-r_0}\frac{d_{0i}y_{0i}}{1-p_i} - p_d\tau\right) + \gamma_2(d_{1i}y_{1i} - d_i r_0) + d_i\boldsymbol{\gamma}'_3\frac{\partial l_i(\boldsymbol{\psi}; r_0)}{\partial\boldsymbol{\psi}}\ , \tag{A.11}$$

$\boldsymbol{\gamma}'_3 = -(\boldsymbol{\delta}'_\beta,\ \delta_p)\boldsymbol{\Psi}^{-1}$, and $\gamma_2 = \delta_r + \boldsymbol{\gamma}'_3\boldsymbol{\gamma}_1$. Because $\mathbf{s}_i$ contains $(d_{1i}y_{2i} - d_i r_0)$, it is collected as the second term of e_i . It follows that $\mathcal{E}[e_i] = 0$ and $\mathcal{E}[e_i^2] < \infty$ under the same assumptions. From the Lindeberg-Lévy theorem, $\sqrt{m}(\hat{\tau} - nm^{-1}p_d\tau) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$. Then, the definition of σ^2 is given by

$$\sigma^2 = \frac{\mathcal{E}[e_i^2]}{p_d}\ . \tag{A.12}$$

Thus, the desired result is obtained. ■

(ii) From the results of Theorem 1 and $(\hat{\delta}_r,\ \hat{\boldsymbol{\delta}}'_\beta,\ \hat{\delta}_p) \xrightarrow{p} (\delta_r,\ \boldsymbol{\delta}'_\beta,\ \delta_p)$, we have $\hat{\sigma}^2 \xrightarrow{p} \sigma^2$. Under $\mathrm{H}_0 : \tau = 0$, it holds that $\mu_\tau = 0$. Thus, we conclude the asymptotic normality of t_τ . ■

References

[1] Akashi, K. and T. Horie (2022), "Note on the Uniqueness of the Maximum Likelihood Estimator for a Heckman's Simultaneous Equations Model," *Econometrics and Statistics*, forthcoming.

[2] Amemiya, T. (1985), *Advanced Econometrics*, Harvard University Press.

[3] Cosslett, S. R. (1981), "Maximum Likelihood Estimator for Choice-Based Samples," *Econometrica*, Vol. 49(5), 1289-1316.

[4] Manski, C. F. and D. McFadden (1981), "Alternative Estimators and Sample Designs for Discrete Choice Analysis," *Structural Analysis of Discrete Data with Econometric Applications*, 2-49. Cambridge, MA: MIT Press.

[5] Heckman, J. J. and P. E. Todd (2009), "A Note on Adapting Propensity Score Matching and Selection Models to Choice Based Samples," *Econometrics Journal*, Vol. 12(1), 230-234.

[6] Hsieh, D. A., C. F. Manski, and D. McFadden (1985), "Estimation of Response Probabilities from Augmented Retrospective Observations," *Journal of the American Statistical Association*, Vol. 80(391), 651-662.

[7] Kennedy, E. H., A. Sjölander, and D. S. Small (2015), "Semiparametric Causal Inference in Matched Cohort Studies," *Biometrika*, Vol. 102(3), 739-746.

[8] Manski, C. F. and S. R. Lerman (1977), "The Estimation of Choice Probabilities from Choice Based Samples," *Econometrica*, Vol. 45(8), 1977-1988.

[9] Rosenbaum, P. R. and D. B. Rubin (1983), " The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika* Vol. 70(1), 41-55.