A Further Analysis of the Kernel Estimator

In this section, we analyze the asymptotic properties of the kernel odds ratio estimator. We work with the log of the odds ratio for an easier analysis of the asymptotic properties.

Theorem A.1. *Given a symmetric kernel function $K(u)$ such that $\int K(u)du = 1$, $\int uK(u)du = 0$, and λ is the bandwidth. Consider the kernel odds ratio estimator as defined in Section 4.1. Denote $f(x)$ as the density of $X, X \in \mathbb{R}^d$ and consider $\lambda = o(\frac{1}{N})^{\frac{1}{d+4}}$. Assume that $f(x)$ continuous, bounded-above, and second-order differentiable. Define $\boldsymbol{w} \triangleq (YZ, (1-Y)(1-Z), Y(1-Z), Z(1-Y))^\top$, $\Sigma(x) \triangleq \mathrm{Cov}(\boldsymbol{w}|x)$, $g(x) = (g^1(x), g^2(x), g^3(x), g^4(x))^\top \triangleq \mathbb{E}[\boldsymbol{w}|x]$, and $A(x) \triangleq (\frac{1}{g^1(x)}, \frac{1}{g^2(x)}, \frac{1}{g^3(x)}, \frac{1}{g^4(x)})^\top$. Then,*

$$(n_2\lambda^d)^{1/2}(\log \widehat{OR}(x) - \log OR(x)) \to \mathcal{N}\left(0, A^\top(x)\Sigma(x)A(x)f(x)^{-1}\int K^2(u)du\right),$$

in distribution as $N \to \infty$.

Proof. By definition,

$$\mathrm{OR}(x) = \frac{g^1(x)g^2(x)}{g^3(x)g^4(x)}. \tag{7}$$

Also by definition, the kernel estimator is

$$\widehat{\mathrm{OR}}(x) = \frac{\hat{g}^1(x)\hat{g}^2(x)}{\hat{g}^3(x)\hat{g}^4(x)}, \tag{8}$$

where $\hat{g}^j(x) = \frac{\sum_{i=1}^N K(\frac{x-X_i}{\lambda})W^j}{\sum_{i=1}^N K(\frac{x-X_i}{\lambda})}, i = 1\cdots 4$, N as the number of the samples, and $(W^1, W^2, W^3, W^4)^\top = \boldsymbol{w} = (YZ, (1-Y)(1-Z), Y(1-Z), Z(1-Y))^\top$. Then, $\log \widehat{\mathrm{OR}}(x)$ is simply $\log(\hat{g}^1(x)) + \log(\hat{g}^2(x)) - \log(\hat{g}^3(x)) - \log(\hat{g}^4(x))$.

We first study an asymptotic analysis for $g^j(x), j = 1\cdots 4$. Then we apply the Delta method to obtain the asymptotic consistency of $\log \widehat{\mathrm{OR}}(x)$. Denote that

$$\hat{g}^j(x) = \frac{\frac{1}{N\lambda^d}\sum_{i=1}^N K(\frac{x-X_i}{\lambda})W_i^j}{\frac{1}{N\lambda^d}\sum_{i=1}^N K(\frac{x-X_i}{\lambda})} = \frac{\hat{\tau}^j(x)}{\hat{f}(x)}.$$

Then we have

$$\begin{aligned}
\mathbb{E}[\hat{f}(x)] &= \mathbb{E}\left[\frac{1}{N\lambda^d}\sum_{i=1}^N K(\frac{x-X_i}{\lambda})\right] \\
&= \mathbb{E}\left[\frac{1}{\lambda^d}K(\frac{x-X_i}{\lambda})\right] \\
&= \int \frac{1}{\lambda^d}K(\frac{x-z}{\lambda})f(z)dz.
\end{aligned}$$

Making the change-of-variables formula for multivariate densities $u = \frac{x-z}{\lambda}$, $du = \lambda^{-d}dz$, then

$$\mathbb{E}[\hat{f}(x)] = \int K(u)f(x-\lambda u)du.$$

When $f(x)$ is continuous, bounded above and $\int K(u)du = 1$, the above converges to $f(x)$ as λ goes to 0 by dominated convergence theorem. To compute the bias, take the second order Taylor expansion of $f(x-\lambda u)$, we have:

$$f(x-\lambda u) = f(x) - \lambda\frac{\partial f(x)}{\partial x'}u + \frac{\lambda^2}{2}tr(\frac{\partial^2 f(x)}{\partial x \partial x'}uu') + o(\lambda^2)$$

Therefore, since the kernel is symmetric, the bias is $\mathcal{O}(\lambda^2)$.

Similarly, we calculate the variance of $\hat{f}(x)$:

$$\begin{aligned}
\text{Var}(\hat{f}(x)) &= \text{Var}\left(\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\lambda^d}K(\frac{x-X_i}{\lambda})\right) \\
&= \frac{1}{N}\text{Var}\left(\frac{1}{\lambda^d}K(\frac{x-X_i}{\lambda})\right) \\
&= \frac{1}{N}\mathbb{E}\left[\left(\frac{1}{\lambda^d}K(\frac{x-X_i}{\lambda})\right)^2\right] - \frac{1}{N}\left(\mathbb{E}\left[\frac{1}{\lambda^d}K(\frac{x-X_i}{\lambda})\right]\right)^2 \\
&= \frac{1}{N}\int\frac{1}{\lambda^{2d}}\left(K(\frac{x-z}{\lambda})\right)^2 f(z)dz - \frac{1}{N}\left(\mathbb{E}[\hat{f}(x)]\right)^2 \\
&= \frac{1}{N\lambda^d}\int\left(K(u)\right)^2 f(x-\lambda u)du - \frac{1}{N}\left(\mathbb{E}[\hat{f}(x)]\right)^2 \\
&= \frac{f(x)}{N\lambda^d}\int(K(u))^2du + o(\frac{1}{N\lambda^d}),
\end{aligned}$$

where we make the change of variable $u = \frac{x-z}{\lambda}$ again. Therefore, the variance of $\hat{f}(x)$ is $\mathcal{O}(\frac{1}{N\lambda^d})$. Recall its bias is $\mathcal{O}(\lambda^2)$ given that the kernel is symmetric, and the optimal bandwidth equates the rate of convergence of the squared bias and variance, i.e. $\mathcal{O}((\lambda^*)^4) = \mathcal{O}(\frac{1}{N(\lambda^*)^d})$. Therefore, the optimal bandwidth is $\lambda^* = \mathcal{O}(\frac{1}{N})^{\frac{1}{d+4}}$.

Further, we can consider $\hat{f}(x)$ as an average of a triangular array:

$$\hat{f}(x) = \frac{1}{N}\sum_{i=1}^{N} Z_{in},$$

with $Z_{in} = \frac{1}{\lambda^d}K(\frac{x-X_i}{\lambda})$. By the Lyapunov CLT theorem, when $\lambda \to 0$, $\mathcal{N}\lambda^d \to \infty$, we have

$$\frac{\hat{f}(x) - \mathbb{E}[\hat{f}(x)]}{\sqrt{\text{Var}(\hat{f}(x))}} \to N(0,1),$$

as $N \to \infty$.

Notice that

$$\frac{\hat{f}(x) - f(x)}{\sqrt{\text{Var}(\hat{f}(x))}} = \frac{\hat{f}(x) - \mathbb{E}[\hat{f}(x)]}{\sqrt{\text{Var}(\hat{f}(x))}} + \frac{\mathbb{E}[\hat{f}(x)] - f(x)}{\sqrt{\text{Var}(\hat{f}(x))}}.$$

Pick $\lambda = o(\frac{1}{N})^{\frac{1}{d+4}}$ ensures that and the bias term vanishes from the asymptotic distribution and is negligible relative to the variance. Therefore, we have

$$\frac{\hat{f}(x) - f(x)}{\sqrt{\text{Var}(\hat{f}(x))}} \to N(0,1),$$

in distribution as $N \to \infty$.

Next we analyze $\hat{\tau}^j(x), j = 1 \cdots 4$. By definition,

$$\begin{aligned}
\mathbb{E}[\hat{\tau}^j(x)] &= \mathbb{E}[\frac{1}{\lambda^d}K(\frac{x-X_i}{\lambda})W_i^j] \\
&= \mathbb{E}[\frac{1}{\lambda^d}K(\frac{x-X_i}{\lambda})g^j(x_i)] \\
&= \int\frac{1}{\lambda^d}K(\frac{x-X_i}{\lambda})g^j(z)f(z)dz
\end{aligned}$$

Similar to the derivation of $\mathbb{E}[\hat{f}(x)]$, we have

$$\begin{aligned}\mathbb{E}[\hat{\tau}^j(x)] &= g^j(x)f(x) + \mathcal{O}(\lambda^d) \\ &\to g^j(x)f(x),\end{aligned}$$

as $N \to \infty$, and

$$\begin{aligned}\mathrm{Var}(\hat{\tau}^j(x)) &= \mathrm{Var}\left(\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\lambda^d}K(\frac{x-X_i}{\lambda})W_i^j\right) \\ &= \frac{\mathbb{E}[(W^j)^2|x]}{N\lambda^d}\int K^2(u)du + o(\frac{1}{N\lambda^d}).\end{aligned}$$

As $N\lambda^d \to \infty$, $\begin{pmatrix}\hat{\tau}(x)\\ \hat{f}(x)\end{pmatrix}$ are jointly normal. Applying the Delta method, we have

$$(N\lambda^d)^{\frac{1}{2}}(\hat{g}^j(x) - g^j(x)) \to \mathcal{N}\left(0, \frac{\mathrm{Var}(W_i^j|x_i = x)}{f(x)}\int K^2(u)du\right).$$

Therefore, for $j = 1\cdots 4$, we have $\mathrm{Var}(\hat{g}^j(x)) = \frac{\mathrm{Var}(W_i^j|x_i=x)}{f(x)}\int K^2(u)du,$. Similarly, we have that $\mathrm{Cov}(\hat{g}^j(x), \hat{g}^k(x)) = \frac{\mathrm{Cov}(W_i^j, W_i^k|x_i=x)}{f(x)}\int K^2(u)du$ for all $j, k = 1\cdots 4, j \neq k$ as N goes to ∞. Thus,

$$(n_2\lambda^d)^{1/2}\left(\hat{g}(x) - g(x)\right) \to \mathcal{N}\left(0, \Sigma(x)f(x)^{-1}\int K^2(u)du\right), \tag{9}$$

Lastly, we use the Delta method again to analyze the asymptotic convergence of $\log\widehat{\mathrm{OR}}(x)$. Recall that $\log\widehat{\mathrm{OR}}(x) = \log(\hat{g}^1(x)) + \log(\hat{g}^2(x)) - \log(\hat{g}^3(x)) - \log(\hat{g}^4(x))$. By definition and Taylor expansion,

$$\begin{aligned}&\log\widehat{\mathrm{OR}}(x) - \log\mathrm{OR}(x) \\ &= \frac{1}{g^1(x)}(\hat{g}^1(x) - g^1(x)) + \frac{1}{g^2(x)}(\hat{g}^2(x) - g^2(x)) \\ &- \frac{1}{g^3(x)}(\hat{g}^3(x) - g^3(x)) - \frac{1}{g^4(x)}(\hat{g}^4(x) - g^4(x))\end{aligned}$$

Therefore, by Delta method and Eq equation 9, the asymptotic variance of $\log\widehat{\mathrm{OR}}(x)$ is:

$$A^\top(x)\Sigma(x)A(x)f(x)^{-1}\int K^2(u)du.$$

Therefore, we have

$$(n_2\lambda^d)^{1/2}(\log\widehat{\mathrm{OR}}(x) - \log\mathrm{OR}(x)) \to \mathcal{N}\left(0, A^\top(x)\Sigma(x)A(x)f(x)^{-1}\int K^2(u)du\right),$$

in distribution. This completes the proof. □

Let $\log\widehat{\mathrm{OR}}_1(x), \log\widehat{\mathrm{OR}}_2(x)$ denotes the two consistent estimators obtained from datasets $\mathcal{O}_1$ and $\mathcal{O}_2$. Let $\widehat{\tau}_2$ denote a consistent estimator of the true ATE τ that we obtain using dataset $\mathcal{O}_2$. Then

$$\begin{pmatrix} n_2^{1/2}(\widehat{\tau}_2 - \tau) \\ (n_2\lambda^d)^{1/2}\left(\log\widehat{\mathrm{OR}}_1(x) - \log\widehat{\mathrm{OR}}_2(x)\right)\end{pmatrix} \to \mathcal{N}\left\{0, \begin{pmatrix} v_2 & \Gamma^\top \\ \Gamma & V\end{pmatrix}\right\}, \tag{10}$$

for some V and Γ. If Eq. (10) holds exactly rather than asymptotically, by multivariate normal theory, we have the following the conditional distribution:

$$n_2^{1/2}(\widehat{\tau}_2 - \tau) \mid (n_2\lambda^d)^{1/2}\left(\log\widehat{\mathrm{OR}}_1(x) - \log\widehat{\mathrm{OR}}_2(x)\right)$$
$$\sim \mathcal{N}\left\{(n_2\lambda^d)^{1/2}\Gamma^\top V^{-1}(n_2\lambda^d)^{1/2}\left(\log\widehat{\mathrm{OR}}_1(x) - \log\widehat{\mathrm{OR}}_2(x)\right), v_2 - \Gamma^\top V^{-1}\Gamma\right\}.$$

Then, we apply the control variates method to build a new estimator of τ which has a lower variance than $\hat{\tau}_2$. The new bias-corrected estimator for ATE is as follows: $\widehat{\tau}_{\mathrm{CV}}(\beta) = \widehat{\tau}_2 - \beta\left(\log\widehat{\mathrm{OR}}_1(x) - \log\widehat{\mathrm{OR}}_2(x)\right)$.

Solving for the optimal β, we obtain the new estimator

$$\widehat{\tau}_{\mathrm{CV}} = \widehat{\tau}_2 - \sqrt{\lambda^d}\Gamma^\top V^{-1}\left(\log\widehat{\mathrm{OR}}_1(x) - \log\widehat{\mathrm{OR}}_2(x)\right), \tag{11}$$

where $V = \mathrm{Var}\left(\log\widehat{\mathrm{OR}}_1(x) - \log\widehat{\mathrm{OR}}_2(x)\right)^{-1}$, and $\Gamma = \mathrm{Cov}(\log\widehat{\mathrm{OR}}_1(x) - \log\widehat{\mathrm{OR}}_2(x), \widehat{\tau}_2)$, and $\lambda^* = o(\frac{1}{N})^{\frac{1}{d+4}}$.

Denote the asymptotic variance of $\widehat{\tau}_2$ as v_2. Under Assumption 2.1, if Equation (10) holds, then $\widehat{\tau}_{\mathrm{CV}}$ is consistent for τ, and we have:

$$n_2^{1/2}(\widehat{\tau}_{\mathrm{CV}} - \tau) \to \mathcal{N}(0, v_2 - \Gamma^\top V^{-1}\Gamma),$$

in distribution as $n_2 \to \infty$. Given a nonzero Γ, the asymptotic variance, $v_2 - \Gamma^\top V^{-1}\Gamma$, is smaller than v_2.

B Bootstrap Sampling Procedure

Our bootstrap sampling procedure is similar to the one in Yang & Ding (2020). For $b = 1, ..., B$, we construct bootstrap replicates for the estimators as follows:

Step 1. Sample n_2 units from $\mathcal{O}_2$ with replacement as $O_2^{*(b)}$, and sample n_1 units from $\mathcal{O}_1$ with replacement as $O_1^{*(b)}$.

Step 2. Compute the bootstrap replicate $\hat{\tau}_2^{(b)}$ using the dataset $O_2^{*(b)}$, and compute the bootstrap replicates $\hat{\psi}_2^{(b)}$, and $\hat{\psi}_1^{(b)}$ using the dataset $O_1^{*(b)}$.

Based on the bootstrap replicates, we estimate the sample covariance $\hat{\Gamma}$ and $\hat{V}$ by

$$\hat{\Gamma} = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\tau}_2^{(b)} - \hat{\tau}_2)(\hat{\psi}_2^{(b)} - \hat{\psi}_1^{(b)} - \hat{\psi}_2 + \hat{\psi}_1),$$
$$\hat{V} = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\psi}_2^{(b)} - \hat{\psi}_1^{(b)} - \hat{\psi}_2 + \hat{\psi}_1)(\hat{\psi}_2^{(b)} - \hat{\psi}_1^{(b)} - \hat{\psi}_2 + \hat{\psi}_1)^\top.$$

The bootstrap covariance estimates $\hat{\Gamma}$ and $\hat{V}$ are consistent if the estimators $\hat{\tau}_2$, $\hat{\psi}_1$, and $\hat{\psi}_2$ are regular asymptotically linear (RAL) estimators, as shown by Efron & Tibshirani (1986) and Shao & Tu (2012).

Definition B.1. *An estimator $\hat{\tau}$ for a statistic τ estimated from a dataset $\{Z_i, X_i, Y_i\}_{i=1}^n$ is RAL if it can be asymptotically approximated by a sum of IID random vectors with mean 0:*

$$\hat{\tau} - \tau \cong \frac{1}{n}\sum_{i=1}^{n}\phi(Z_i, X_i, Y_i)$$

$\phi(Z, X, Y)$ is also known as the influence function for $\hat{\tau}$.

A common example of a RAL estimator for the ATE is the regression imputation estimator, which we used in experiments.

B.1 Matching estimators

Another common class of ATE estimators is matching estimators. Matching estimators do not have smooth influence functions, so the direct bootstrap procedure above may not be consistent Abadie & Imbens (2008). However, Yang & Ding (2020) and Abadie & Imbens (2006) show that the bias of a matching estimator $\hat{\tau}$ can still be expressed in an asymptotically linear form:

$$\hat{\tau} - \tau \cong \frac{1}{n}\sum_{i=1}^{n} \phi_i$$

Using these linear terms, a slightly modified bootstrap procedure can be used, which Yang & Ding (2020) show to be consistent for both RAL estimators and matching estimators. This procedure uses a modified version of Step 2 which estimates the asymptotically linear terms. Let $\phi_i^{\tau_2}$ indicate the asymptotically linear term for estimator $\hat{\tau}_2$ (e.g. $\phi(Z_i, X_i, Y_i)$ for RAL $\hat{\tau}_2$), and let $\phi_i^{\psi_1}$, $\phi_i^{\psi_2}$ indicate the same for estimators $\hat{\psi}_2$, $\hat{\psi}_2$, respectively. Let $\hat{\phi}_i^{\tau_2}$, $\hat{\phi}_i^{\psi_1}$, $\hat{\phi}_i^{\psi_2}$ denote estimates for the population quantities.

Step 2 (modified for matching). Compute the bootstrap replicates using the dataset $O_2^{*(b)}$ as

$$\hat{\tau}_2^{(b)} - \hat{\tau}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} \hat{\phi}_i^{\tau_2}$$

and compute the bootstrap replicates using the dataset $O_1^{*(b)}$ as

$$\hat{\psi}_1^{(b)} - \hat{\psi}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} \hat{\phi}_i^{\psi_1}; \quad \hat{\psi}_2^{(b)} - \hat{\psi}_2 = \frac{1}{n_1}\sum_{i=1}^{n_1} \hat{\phi}_i^{\psi_2}.$$

Theorem B.1. *(Theorem 3 from Yang & Ding (2020)) If $\hat{\tau}_2$, $\hat{\psi}_1$, and $\hat{\psi}_2$ are RAL estimators or matching estimators, then under certain regularity conditions, the bootstrap estimates $\hat{\Gamma}, \hat{V}$ under this modified procedure are consistent for Γ, V.*

C Proofs and Further Analysis of the Odds Ratio

We provide proofs for the theorems and lemma presented in Sections 3 and 4, as well as further analysis of the odds ratio's effectiveness as a control variate.

C.1 Proofs from Section 3

Theorem 3.1. *Denote the asymptotic variance of $\widehat{\tau}_2$ as v_2. Under Assumption 2.1, if Equation (1) holds, then $\widehat{\tau}_{CV}$ is consistent for τ, and we have:*

$$n_2^{1/2}(\widehat{\tau}_{CV} - \tau) \to \mathcal{N}(0, v_2 - \Gamma^\top V^{-1}\Gamma),$$

in distribution as $n_2 \to \infty$. Given a nonzero Γ, the asymptotic variance, $v_2 - \Gamma^\top V^{-1}\Gamma$, is smaller than v_2.

Proof. The theorem statement follows directly from

$$n_2^{1/2}\begin{pmatrix} \widehat{\tau}_2 - \tau \\ \widehat{\psi}_2 - \widehat{\psi}_1 \end{pmatrix} \to \mathcal{N}\left\{0, \begin{pmatrix} v_2 & \Gamma^\top \\ \Gamma & V \end{pmatrix}\right\}.$$

By construction, $\widehat{\tau}_{\mathrm{CV}} = \widehat{\tau}_2 - \Gamma^\top V^{-1}(\widehat{\psi}_2 - \widehat{\psi}_1)$. To compute the asymptotic variance, notice that

$$
\begin{aligned}
&\mathrm{Var}(n_2^{1/2}(\widehat{\tau}_{\mathrm{CV}} - \tau)) \\
&= \mathrm{Var}(n_2^{1/2}(\widehat{\tau}_{\mathrm{CV}} - \tau)) \\
&= \mathrm{Var}(n_2^{1/2}(\widehat{\tau}_2 - \tau - \Gamma^\top V^{-1}(\widehat{\psi}_2 - \widehat{\psi}_1))) \\
&= \mathrm{Var}(n_2^{1/2}(\widehat{\tau}_2 - \tau)) + \Gamma^\top V^{-1}\mathrm{Var}(n_2^{1/2}(\widehat{\psi}_2 - \widehat{\psi}_1))V^{-1}\Gamma - 2\mathrm{Cov}(n_2^{1/2}(\widehat{\tau}_2 - \tau), n_2^{1/2}\Gamma^\top V^{-1}(\widehat{\psi}_2 - \widehat{\psi}_1)) \\
&= v_2 + \Gamma^\top V^{-1}\Gamma - 2\Gamma^\top V^{-1}\Gamma) \\
&= v_2 - \Gamma^\top V^{-1}\Gamma)
\end{aligned}
$$

Therefore, we have:

$$n_2^{1/2}(\widehat{\tau}_{\mathrm{CV}} - \tau) \to \mathcal{N}(0, v_2 - \Gamma^\top V^{-1}\Gamma),$$

in distribution as $n_2 \to \infty$, which completes the proof. □

C.2 Proofs from Section 4

Lemma 4.1. *If the selection S depends solely on Y (as in Figure 1), then the conditional odds ratio is transportable and given by:*

$$OR(x) = \frac{P(Y=1|S=1,Z=1,x)P(Y=0|S=1,Z=0,x)}{P(Y=0|S=1,Z=1,x)P(Y=1|S=1,Z=0,x)}.$$

Proof. By Bayes' theorem,

$$P(Y=y|Z=z,x) = \frac{P(Y=y|S=1,Z=z,x)P(S=1|Z=z,x)}{P(S=1|Y=y,Z=z,x)}.$$

Since S depends solely on Y, S is conditionally independent of X and Z given Y. Therefore, we can rewrite the equation above as

$$P(Y=y|Z=z,x) = \frac{P(Y=y|S=1,Z=z,x)P(S=1|Z=z,x)}{P(S=1|Y=y)}.$$

Note that here y does not need to be binary values, but can also take any categorical values, i.e. $y \in \{1, \cdots K\}$.

Substituting this into the odds ratio[1] Definition 4.1 (under Assumption 2.1),

$$
\begin{aligned}
\mathrm{OR}(x) &= \frac{P(Y=1|Z=1,x)P(Y=0|Z=0,x)}{P(Y=0|Z=1,x)P(Y=1|Z=0,x)} \\
&= \frac{\frac{P(Y=1|S=1,Z=1,x)P(S=1|Z=1,x)}{P(S=1|Y=1)}\frac{P(Y=0|S=1,Z=0,x)P(S=1|Z=0,x)}{P(S=1|Y=0)}}{\frac{P(Y=0|S=1,Z=1,x)P(S=1|Z=1,x)}{P(S=1|Y=0)}\frac{P(Y=1|S=1,Z=0,x)P(S=1|Z=0,x)}{P(S=1|Y=1)}} \\
&= \frac{P(Y=1|S=1,Z=1,x)P(Y=0|S=1,Z=0,x)}{P(Y=0|S=1,Z=1,x)P(Y=1|S=1,Z=0,x)}
\end{aligned}
$$

Therefore, the conditional odds ratio is transportable under selection bias given by S. □

Theorem 4.2. *If the selection S depends solely on Y (as in Figure 1) and $P(Y=1|Z=z,X=x)$ follows the logistic model in equation 4, then $P(Y=1|Z=z,X=x,S=1)$ also follows a logistic model, with the same coefficient β_1^x on Z as the logistic model for $P(Y=1|Z=z,X=x)$ for each covariate value x. Furthermore, the conditional odds ratio $OR(x) = e^{\beta_1^x}$.*

[1] In general, the odds ratio definition may be generalized to non-binary outcomes, for example, see Agresti (1980); Moser & Coombs (2004).

Proof. Given the assumed outcome model, we have

$$P(Y=1|Z=z,x)=\frac{e^{\beta_0^x+\beta_1^x z}}{1+e^{\beta_0^x+\beta_1^x z}}.$$

Let $p_1 = P(S=1|Y=1)$ and $p_0 = P(S=1|Y=0)$. Since the selection S depends solely on Y, S is conditionally independent of X and Z given Y.

The outcome model under selection bias is given by:

$$\begin{aligned}
&P(Y=1|Z=z,X=x,S=1)\\
&=\frac{P(Y=1|Z=z,X=x)P(S=1|Z=z,X=x,Y=1)}{P(S=1|Z=z,X=x)}\\
&=\frac{P(Y=1|Z=z,X=x)P(S=1|Z=z,X=x,Y=1)}{\sum_{y\in\{0,1\}}P(Y=y|Z=z,X=x,)P(S=1|Z=z,X=x,Y=y)}\\
&=\frac{P(Y=1|Z=z,X=x)p_1}{P(Y=1|Z=z,X=x)p_1+P(Y=0|Z=z,X=x)p_0}\\
&=\frac{\frac{e^{\beta_0^x+\beta_1^x z}}{1+e^{\beta_0^x+\beta_1^x z}}p_1}{\frac{e^{\beta_0^x+\beta_1^x z}}{1+e^{\beta_0^x+\beta_1^x z}}p_1+\frac{1}{1+e^{\beta_0^x+\beta_1^x z}}p_0}\\
&=\frac{e^{\beta_0^x+\beta_1^x z}p_1}{e^{\beta_0^x+\beta_1^x z}p_1+p_0}\\
&=\frac{e^{\beta_0^x+\beta_1^x z}p_1/p_0}{e^{\beta_0^x+\beta_1^x z}p_1/p_0+1}\\
&=\frac{e^{\delta+\beta_0^x+\beta_1^x z}}{1+e^{\delta+\beta_0^x+\beta_1^x z}}.
\end{aligned}$$

where $\delta = \log(p_1/p_0)$. Thus, on the selection biased dataset $\mathcal{O}_1$, the outcome model $P(Y=1|Z=z,X=x,S=1)$ also follows a logistic model, with the same coefficient β_1^x on Z as the logistic model for $P(Y=1|Z=z,X=x)$ for each covariate value x. Furthermore, a simple calculation shows that the conditional odds ratio is $\mathrm{OR}(x)=e^{\beta_1^x}$:

$$\begin{aligned}
\mathrm{OR}(x)&=\frac{P(Y=1|Z=1,x)P(Y=0|Z=0,x)}{P(Y=0|Z=1,x)P(Y=1|Z=0,x)}\\
&=\frac{\frac{e^{\beta_0^x+\beta_1^x}}{1+e^{\beta_0^x+\beta_1^x}}\frac{1}{1+e^{\beta_0^x}}}{\frac{1}{1+e^{\beta_0^x+\beta_1^x}}\frac{e^{\beta_0^x}}{1+e^{\beta_0^x}}}\\
&=\frac{e^{\beta_0^x+\beta_1^x}}{e^{\beta_0^x}}\\
&=e^{\beta_1^x}.
\end{aligned}$$

□

C.3 Analysis of Nonlinear Relationship Between ATE and OR

As discussed in Section 3, the variance reduction from adding control variates depends on the strength of the correlation between the control variates and the ATE estimator. Since we propose to use the odds ratio for selection biased datasets, here we examine the relationship between the ATE and the odds ratio. Specifically, we derive an explicit expression for the ATE using the marginal odds ratio OR assuming a binary covariate X and the following simple logistic outcome model:

$$P(Y=1|Z=z,X=x)=\frac{e^{\beta_0+\beta_1 z+\beta_2 x}}{1+e^{\beta_0+\beta_1 z+\beta_2 x}},$$

where the marginal odds ratio OR is defined as

$$\text{OR}=\frac{P(Y(1)=1)P(Y(0)=0)}{P(Y(1)=0)P(Y(0)=1)}.$$

Under this simple logistic outcome model, $\text{OR}=e^{\beta_1}$.

By Assumption 2.1, we have

$$\begin{aligned}\mathbb{E}[Y(1)]&=\int \mathbb{E}[Y|Z=1,X=x]P(X=x)\mathrm{d}x\\&=\int \frac{e^{\beta_0+\beta_1+\beta_2 x}}{e^{\beta_0+\beta_1+\beta_2 x+1}}P(X=x)\mathrm{d}x\\&=\frac{e^{\beta_0+\beta_1+\beta_2}}{e^{\beta_0+\beta_1+\beta_2+1}}P(X=1)+\frac{e^{\beta_0+\beta_1}}{e^{\beta_0+\beta_1+1}}(1-P(X=1))\end{aligned}$$

Similarly,

$$\begin{aligned}\mathbb{E}[Y(0)]&=\int \mathbb{E}[Y|Z=0,X=x]P(X=x)\mathrm{d}x\\&=\int \frac{e^{\beta_0+\beta_2 x}}{e^{\beta_0+\beta_2 x+1}}P(X=x)\mathrm{d}x\\&=\frac{e^{\beta_0+\beta_2}}{e^{\beta_0+\beta_2+1}}P(X=1)+\frac{e^{\beta_0}}{e^{\beta_0+1}}(1-P(X=1))\end{aligned}$$

Therefore, by some algebra we obtain that

$$\begin{aligned}\tau&=\mathbb{E}[Y(1)]-\mathbb{E}[Y(0)]\\&=\frac{\gamma ab\psi}{ab\psi+1}-\frac{\gamma a\psi}{a\psi+1}+\frac{a\psi-a}{a^2\psi-a\psi+a+1}-C,\end{aligned}$$

where $\psi=\text{OR}$, $\gamma=P(X=1)$, $a=e^{\beta_0}$, $b=e^{\beta_2}$, with a constant term $C=\frac{a}{a+1}-\frac{ab}{ab+1}$.

D Additional Experimental Details and Results for Simulation Study

This section provides additional experimental details and results for the simulation study. All code for running the simulation study is provided with the supplementary materials.

D.1 Data generation

We generate the dataset $\mathcal{O}_2$ by sampling n_2 samples using the following data-generating process. Let $X\in\mathbb{R}^2$ have two components X_1, X_2, which are i.i.d. Bernoulli($p=0.5$). Given X, the treatment assignment Z is distributed as $P(Z=1|X=x)=\frac{e^{a_0+a_1^T x}}{1+e^{a_0+a_1^T x}}$.

As specific parameters, we set $a_1=[-1,1]$, and $a_0=-E[a_1^T X]$, which implies that $P(Z=1)=0.5$. Setting $a_1=[0,0]$ would correspond to a randomized study, whereas we set $a_1=[-1,1]$ to simulate an observational study with confounding. The potential outcomes are distributed as

$$P(Y(0)=1|x)=\frac{e^{b_{0,0}+b_{0,1}^T x}}{1+e^{b_{0,0}+b_{0,1}^T x}},\quad P(Y(1)=1|x)=\frac{e^{b_{1,0}+b_{1,1}^T x}}{1+e^{b_{1,0}+b_{1,1}^T x}}. \tag{12}$$

Eq. equation 12 builds in ignorability in Assumption 2.1, which is also equivalent to generating the outcome Y from

$$P(Y=1|Z=z,x) = \frac{e^{\beta_0+\beta_1 z+\beta_2^T x+\beta_3^T xz}}{1+e^{\beta_0+\beta_1 z+\beta_2^T x+\beta_3^T xz}},$$

where $\beta_0 = b_{0,0}$, $\beta_1 = b_{1,0} - b_{0,1}$, $\beta_2 = b_{0,1}$, and $\beta_3 = b_{1,1} - b_{0,1}$. When $\beta_3 = 0$ and $b_{1,1} = b_{0,1}$, then there is no interaction term between X and Z and the conditional odds ratio is simply e^{β_1}. We set $b_{0,1} = [-1, 1]$ and $b_{1,1} = [1, -1]$ ($\beta_3 \neq 0$) so that the conditional odds ratio varies as a function of x. As done by Zhang (2009), the intercept terms are determined by $b_{0,0} = -0.5 - E[b_{0,1}^T X]$ and $b_{1,0} = 0.5 - E[b_{1,1}^T X]$.

D.2 Finite sample estimator for Correa et al. (2019)

For our simulation experiments with low dimensional discrete X, we compute a finite sample estimator $\widehat{\tau}_{[\text{CTB'19}]}$ based on Lemma 2, Lemma 3, and Algorithm 3 from Correa et al. (2019), but replacing the true probabilities with finite sample estimates. Our finite sample estimator is given by

$$\widehat{\tau}_{[\text{CTB'19}]} = \sum_{x\in\mathcal{X}} \left(\frac{\hat{P}(Y=1,Z=1,X=x)}{\hat{P}(Z=1,X=x)} - \frac{\hat{P}(Y=1,Z=1,X=x)}{\hat{P}(Z=1,X=x)} \right) \hat{P}(X=x),$$

where $\hat{P}(Y=y,Z=z,X=x) = \hat{P}(X=x,Z=z|Y=y,S=1)\hat{P}(Y=y)$, since $X, Z \perp\!\!\!\perp S|Y$.

The first term $\hat{P}(X=x,Z=z|Y=y,S=1)$ comes from $\mathcal{O}_1$, and is given by

$$\hat{P}(X=x,Z=z|Y=y,S=1) = \frac{\sum_{i=1}^{n_1} \mathbb{1}(X_i=x,Z_i=z,Y_i=y,S=1)}{\sum_{i=1}^{n_1} \mathbb{1}(Y_i=y,S=1)}.$$

The second term comes from $\mathcal{O}_2$, and is given by

$$\hat{P}(Y=y) = \frac{1}{n_2}\sum_{j=1}^{n_2} \mathbb{1}(Y_j=y).$$

The other terms in $\widehat{\tau}_{[\text{CTB'19}]}$ come from applying the law of total probability to $\hat{P}(Y=y,Z=z,X=x)$, where

$$\hat{P}(Z=z,X=x) = \sum_{y\in\{0,1\}} \hat{P}(Y=y,Z=z,X=x),$$

$$\hat{P}(X=x) = \sum_{y\in\{0,1\}}\sum_{z\in\{0,1\}} \hat{P}(Y=y,Z=z,X=x).$$

We note that this method is not so straightforward when X is high dimensional or continuous.

D.2.1 Simple logistic outcome model without interaction between X and Z

In addition to the data generation setting described in Section 5, we also include results with a simpler data generation setting without interaction between X and Z. We set $b_{1,1} = b_{0,1} = (-1, 1)^\top$, which implies that $\beta_3 = 0$ in Section D.1. In this simpler model, the conditional odds ratio is constant in X and is given by e^{β_1}.

Figure 5 shows that adding control variates still improves the variance of the ATE estimator under this simpler outcome model without interaction between X and Z. The bias for the simple logistic outcome model without interaction between X and Z is given in Figure 6.

E Additional Experimental Details and Results for Real Data Case Studies

This section provides additional experimental details and results for the real data case studies. All code for generating data and running the experiments is provided with the supplementary materials.

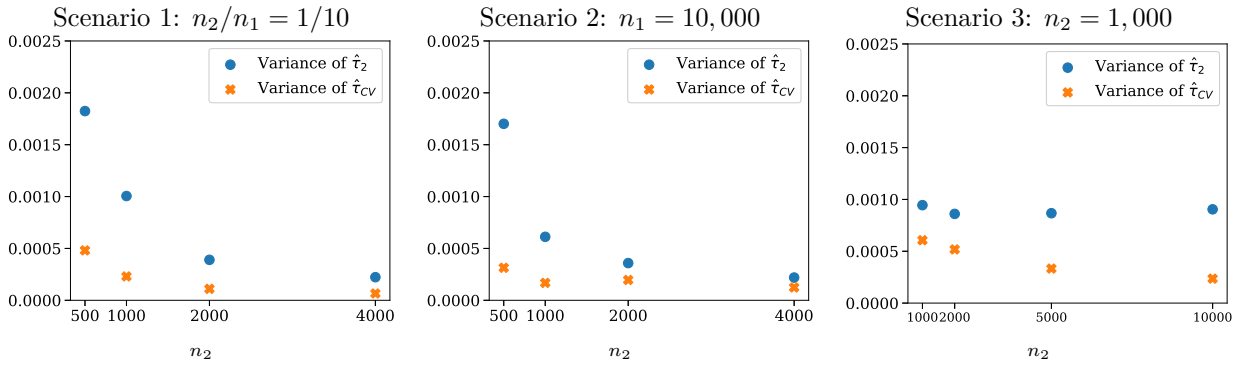

Figure 5: Simple logistic model ($\beta_3 = 0$): Comparisons of variance for $\widehat{\tau}_2$ and $\widehat{\tau}_{\text{CV}}$ over 100 bootstrap replicates. *Lower is better.*

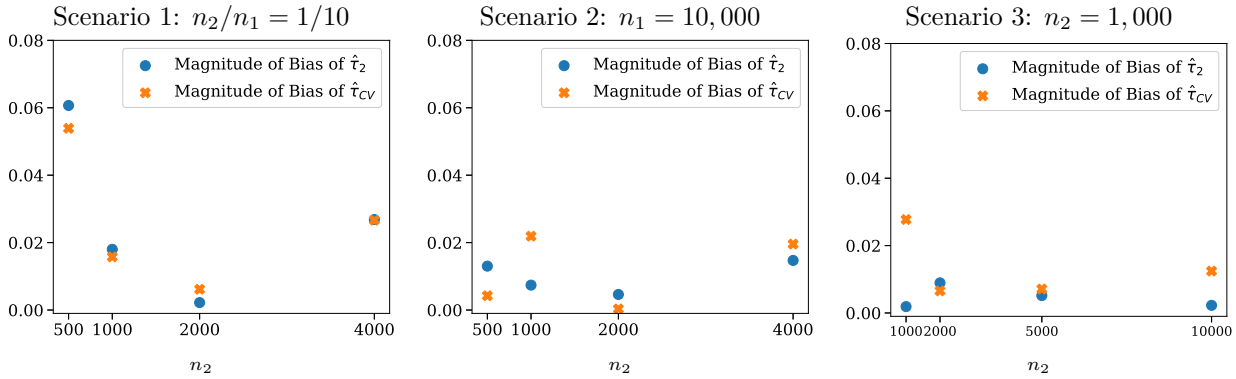

Figure 6: Simple logistic model ($\beta_3 = 0$): Comparisons of bias for $\widehat{\tau}_2$ and $\widehat{\tau}_{\text{CV}}$ over 100 bootstrap replicates. The magnitude of bias reported is the absolute value of the difference between the average value of the estimator over the boostrap replicates and the true ATE. *Lower is better.*

E.1 Data generation for case study 1: flu shot encouragement

We provide a detailed breakdown for generating the selection biased dataset $\mathcal{O}_1$ for Case Study 1 on flu shot encouragement below. Code for this is also included with the supplementary materials.

1. Fit a logistic regression model on the original dataset $\mathcal{O}_2$ with inputs X, Z and outcome Y according to
$$P(Y = 1|Z = z, X = x) = g(\beta, z, x) = \frac{e^{\beta_0+\beta_1 z+\beta_2 x+\beta_3 xz}}{1 + e^{\beta_0+\beta_1 z+\beta_2 x+\beta_3 xz}}. \tag{13}$$
This results in estimated parameters $\hat{\beta}$.

2. Fit a logistic regression model to estimate the propensity score according to
$$P(Z = 1|X = x) = h(a, x) = \frac{e^{a_0+a_1^T x}}{1 + e^{a_0+a_1^T x}},$$
This results in estimated parameters $\hat{a}$.

3. Sample covariates $\{X_i\}_1^N$ with replacement from $\mathcal{O}_2$.

4. Generate Z_i according to the estimated propensity score distribution, $P(Z_i = 1|X = X_i) = h(\hat{a}, X_i)$.

5. Generate Y_i according to the estimated outcome distribution, $P(Y_i = 1|Z = Z_i, X = X_i) = g(\hat{\beta}, Z_i, X_i)$.

6. Apply selection bias on the outcome according to the distribution $P(S_i = 1|Y_i = 1) = 0.9$ to generate the final dataset $\mathcal{O}_1$.

E.2 Data generation for case study 2: spam email detection

For case study 2, we directly apply the code provided by the Atlantic Causal Inference Conference (ACIC) Data Challenge to generate both $\mathcal{O}_1$ and $\mathcal{O}_2$. The ACIC data generation code is publicly available at https://sites.google.com/view/acic2019datachallenge/data-challenge. The ACIC data generation code modifies existing real datasets in a variety of ways to generate datasets for the ACIC data challenge with known true ATEs. Specifically, we use ACIC's "modification 1" of the Spambase spam email detection dataset from UCI, which applies a logistic outcome model very close to the logistic regression models fitted to the actual data. For convenience, we include the exact script for "modification 1" in our supplementary materials.

E.3 Further details on the estimators

We provide further details on the estimators for the ATE and odds ratio used in the real data experiments.

Logistic model with interaction: We model the outcome Y using a logistic model with an interaction term between X and Z in Eq. equation 5, repeated here for convenience:

$$P(Y = 1|Z = z, X = x) = g(\beta, z, x) = \frac{e^{\beta_0+\beta_1 z+\beta_2 x+\beta_3 xz}}{1 + e^{\beta_0+\beta_1 z+\beta_2 x+\beta_3 xz}}.$$

To estimate the ATE from the dataset without selection bias $\mathcal{O}_2$, we perform logistic regression of Y on X, Z, and the interaction term XZ, to produce estimates $\hat{\beta}_{\mathcal{O}_2}$. The ATE estimator is then given by

$$\widehat{\tau}_2 = g(\hat{\beta}_{\mathcal{O}_2}, 1, x) - g(\hat{\beta}_{\mathcal{O}_2}, 0, x).$$

We then estimate the conditional odds ratio from $\mathcal{O}_2$ as $\widehat{\mathrm{OR}}_2(x) = e^{\hat{\beta}_{1,\mathcal{O}_2}+\hat{\beta}_{3,\mathcal{O}_2}}$.

We similarly estimate the conditional odds ratio from $\mathcal{O}_1$ by following the same logistic regression procedure as above, resulting in $\widehat{\mathrm{OR}}_1(x) = e^{\hat{\beta}_{1,\mathcal{O}_1}+\hat{\beta}_{3,\mathcal{O}_1}}$.

Neural network: For a more general outcome model, we model the outcome Y using a logistic model with varying coefficients in Eq. equation 4, where the functions $\beta_0^x = f_0(x; \theta)$, $\beta_1^x = f_1(x; \theta)$ make up the two-dimensional output of a single neural network with parameters θ:

$$P(Y = 1|Z = z, X = x) = g(\theta, z, x) = \frac{e^{f_0(x;\theta)+f_1(x;\theta)z}}{1 + e^{f_0(x;\theta)+f_1(x;\theta)z}}.$$

The optimization objective for the neural network is the logistic loss on the final outcome prediction, $g(\theta, z, x)$. We optimize the neural network using using ADAM with a default learning rate of 0.001 for $1,000$ epochs with batch size n_2. We choose the neural network architecture using five-fold cross validation over $\mathcal{O}_2$. Specifically, we search over $\{4, 8\}$ hidden layers, and hidden layer sizes of $\{4, 8, 16, 32\}$. We use the TensorFlow framework and include all code in the supplementary materials.

Once an architecture has been chosen by the method above, we estimate the ATE from $\mathcal{O}_2$ by optimizing the neural network parameters θ over the dataset $\mathcal{O}_2$ to obtain an estimate $\hat{\theta}_{\mathcal{O}_2}$, and the ATE estimator is given by

$$\widehat{\tau}_2 = g(\hat{\theta}_{\mathcal{O}_2}, 1, x) - g(\hat{\theta}_{\mathcal{O}_2}, 0, x).$$

We then estimate the conditional odds ratio from $\mathcal{O}_2$ as $\widehat{\mathrm{OR}}_2(x) = e^{f_1(x;\hat{\theta}_{\mathcal{O}_2})}$.

Finally, using the same architecture as chosen above, we estimate the conditional odds ratio from $\mathcal{O}_1$ by optimizing the neural network parameters θ over the dataset $\mathcal{O}_1$ to obtain an estimate $\hat{\theta}_{\mathcal{O}_1}$, resulting in $\widehat{\mathrm{OR}}_1(x) = e^{f_1(x;\hat{\theta}_{\mathcal{O}_1})}$.

E.4 Bias results for real-data case studies

We report the bias for the ATE estimators with and without control variates, $\widehat{\tau}_2, \widehat{\tau}_{CV}$ for the real data experiments. The bias is calculated over $B = 300$ bootstrap replicates, and is defined as the difference between the average value of the ATE estimator over $B = 300$ bootstrap replicates and the true ATE.

Tables 3 and 4 show the bias of $\widehat{\tau}_2, \widehat{\tau}_{\text{CV}}$ for each of the different estimation methods for case study 1 and case study 2, respectively. While there was not much difference in bias between $\widehat{\tau}_2$ and $\widehat{\tau}_{\text{CV}}$ in the simulation study, we often observe higher bias for $\widehat{\tau}_{\text{CV}}$ in finite samples on the real data. This bias is more pronounced in case study 2, which may be due to a high dimensional continuous covariate vector X.

Table 3: Biases for Case study 1: flu shot encouragement data with $n_1 = 10,000$ and $n_2 = 2,861$. The bias reported is the difference between the average value of the estimator over the boostrap replicates and the true ATE.

Model type	Bias $\widehat{\tau}_2$	Bias $\widehat{\tau}_{\text{CV}}$
Logistic	4.283×10^{-4}	6.431×10^{-4}
Kernel	5.384×10^{-4}	-4.464×10^{-3}

Table 4: Biases for Case Study 2: spam email detection data with $n_1 = 30,000$. We provide results for a smaller validation dataset $n_2 = 3,000$ on the left, and results for a larger validation dataset $n_2 = 10,000$ on the right. The bias reported is the difference between the average value of the estimator over the boostrap replicates and the true ATE.

	$n_2 = 3,000$		$n_2 = 10,000$	
Model type	Bias $\widehat{\tau}_2$	Bias $\widehat{\tau}_{\text{CV}}$	Bias $\widehat{\tau}_2$	Bias $\widehat{\tau}_{\text{CV}}$
Logistic	6.785×10^{-4}	2.132×10^{-3}	7.230×10^{-3}	3.085×10^{-3}
Kernel	-4.250×10^{-3}	-6.390×10^{-3}	-2.827×10^{-4}	-1.378×10^{-3}
Neural Net	-1.095×10^{-2}	-1.129×10^{-2}	3.427×10^{-4}	1.361×10^{-3}