Semi-parametric Label Shift: Statistics 992 Final Project

Evan Gorstein

December 2021

1 Intro

In this project, our goal is to estimate the severity of label/target shift under a parametric assumption. We assume that we observe n_s i.i.d. samples of $(Y, \mathbf{X})$ from the source population and observe n_t i.i.d. samples of only $\mathbf{X}$ from the target population, and we assume the two populations are linked by the so-called label/target shift assumption: $f_{\mathbf{X}|Y}(\mathbf{x} \mid y) = g_{\mathbf{X}|Y}(\mathbf{x} \mid y) =: f(\mathbf{x} \mid y)$. We adopt the notation $f_{Y|\mathbf{X}}(y \mid \mathbf{x})$ and $f_{\mathbf{X}}(\mathbf{x})$ for the source population, and $g_{Y|\mathbf{X}}(y \mid \mathbf{x})$ and $g_{\mathbf{X}}(\mathbf{x})$ for the target population. Note that $g_Y(y)$ is not estimable from the observed data. We impose the density ratio assumption

$$g_Y(y) = f_Y(y)\rho(y; \boldsymbol{\beta}).$$

In order for $g_Y(y)$ to be a valid density, $\rho(y; \boldsymbol{\beta})$ must be a positive function satisfying $E_1\{\rho(Y; \boldsymbol{\beta})\} = 1$, where subscripts 1 and 0 represent the source and target population, respectively. We use random variable S to denote whether it belongs to the source ($S = 1$) or the target ($S = 0$) population, and we define $\pi = \text{pr}(S = 1)$. Therefore, if the subject is from the source population, the corresponding likelihood is

$$f(\mathbf{x} \mid y) f_Y(y) \pi,$$

and, if the subject is from the target population, the corresponding likelihood is

$$\int f(\mathbf{x} \mid y) g_Y(y) \mathrm{d}y (1 - \pi) = \int f(\mathbf{x} \mid y) f_Y(y) \rho(y; \boldsymbol{\beta}) \mathrm{d}y (1 - \pi).$$

The likelihood for a single observation in the hypothesized combined population is therefore

$$L(\boldsymbol{\beta}, \pi, f_Y(\cdot), f(\cdot \mid \cdot); \mathbf{x}, y, s) = \{\pi f(\mathbf{x} \mid y) f_Y(y)\}^s \left\{(1 - \pi) \int f(\mathbf{x} \mid y) f_Y(y) \rho(y; \boldsymbol{\beta}) \mathrm{d}y\right\}^{1-s}.$$

Our parameter of interest is $\boldsymbol{\beta} \in \mathbb{R}^q$

2 The nuisance tagent space

Since we let $\pi = n_s/(n_s + n_t)$, we have only two (infinite dimensional) nuisances: $f_Y(y)$ and $f(\mathbf{x} \mid y)$. We begin by considering a typical parametric submodel for $f_Y(y)$: for some $\mathbf{h}(Y) \in \mathbb{R}^q$ satisfying $E_1(\mathbf{h}) = 0$,

$$\mathcal{P}_{\boldsymbol{\lambda}} = \{f_Y^{\boldsymbol{\lambda}}(y) = f_Y^0(y)[1 + \boldsymbol{\lambda}^{\mathrm{T}}\mathbf{h}(Y)] \mid \boldsymbol{\lambda} \in \mathbb{R}^q\},$$

where $f_Y^0(y)$ is the true density of Y in the source population. In order for these to be valid densities, we must also impose some condition on the boundedness of $\mathbf{h}$.

The log-likelihood of our model is

$$\log L(\boldsymbol{\beta}, \pi, f_Y(\cdot), f(\cdot \mid \cdot); \mathbf{x}, y, s) = s\log(\pi f(\mathbf{x} \mid y) f_Y(y)) + (1-s)\log\left\{(1-\pi)\int f(\mathbf{x} \mid y) f_Y(y)\rho(y; \boldsymbol{\beta})\mathrm{d}y\right\}$$

so the score function of this parametric submodel (with respect to λ, evaluated at the truth) is given by

$$\begin{aligned}
\left.\frac{\partial \log L}{\partial \boldsymbol{\lambda}}\right|_{\boldsymbol{\lambda}=0} &= S\mathbf{h}(Y) + (1-S)\frac{\frac{\partial}{\partial \boldsymbol{\lambda}}\int f(\mathbf{X} \mid y) f_Y^0(y)[1 + \boldsymbol{\lambda}^{\mathrm{T}}\mathbf{h}(y)]\rho(y; \boldsymbol{\beta})\mathrm{d}y}{\int f(\mathbf{X} \mid y) f_Y^0(y)\rho(y; \boldsymbol{\beta})\mathrm{d}y} \\
&= S\mathbf{h}(Y) + (1-S)\frac{\int \mathbf{h}(y) f(\mathbf{X} \mid y) f_Y^0(y)\rho(y; \boldsymbol{\beta})\mathrm{d}y}{g_{\mathbf{X}}(\mathbf{X})} \\
&= S\mathbf{h}(Y) + (1-S)\int \mathbf{h}(y) g_{Y|\mathbf{X}}(y \mid \mathbf{X})\mathrm{d}y \\
&= S\mathbf{h}(Y) + (1-S)E_0(\mathbf{h}(Y) \mid \mathbf{X})
\end{aligned} \tag{1}$$

This suggests that the semi-parametric nuisance tangent space for $f_Y(y)$ is

$$\Lambda_1 = \{S\mathbf{h}(Y) + (1-S)E_0[\mathbf{h}(Y) \mid \mathbf{X}] \mid \mathbf{h}(Y) \in \mathbb{R}^q,\ E_1(\mathbf{h}) = 0\}$$

We consider a similar parametric submodel for $f(\mathbf{x} \mid y)$: for some $\mathbf{c}(\mathbf{X}, Y) \in \mathbb{R}^q$ satisfying $E[\mathbf{c}(\mathbf{X}, Y) \mid Y] = 0$,

$$\mathcal{P}_{\boldsymbol{\alpha}} = \{f^{\boldsymbol{\alpha}}(\mathbf{x} \mid y) = f^0(\mathbf{x} \mid y)[1 + \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{c}(\mathbf{x}, y)] \mid \boldsymbol{\alpha} \in \mathbb{R}^q\},$$

A derivation identical to (1) leads to

$$\frac{\partial \log L}{\partial \boldsymbol{\alpha}} = S\mathbf{c}(\mathbf{X}, Y) + (1-S)E_0(\mathbf{c}(\mathbf{X}, Y) \mid \mathbf{X}),$$

suggesting a semi-parametric nuisance tangent space for $f(\mathbf{x} \mid y)$

$$\Lambda_2 = \{S\mathbf{c}(\mathbf{X}, Y) + (1-S)E_0[\mathbf{c}(\mathbf{X}, Y) \mid \mathbf{X}] \mid \mathbf{c}(\mathbf{X}, Y) \in \mathbb{R}^q,\ E[\mathbf{c}(\mathbf{X}, Y) \mid Y] = 0\}$$

Unfortunately, these two spaces, Λ_1 and Λ_2 are not orthogonal to each other. Nonetheless, we can proceed to find the orthogonal complement to the overall nuisance tangent space, i.e. $\Lambda^\perp = \Lambda_1^\perp \cap \Lambda_2^\perp$.

3 The orthogonal complement to the nuisance tangent space

We begin by observing a fact relating expectations taken in the target population to expectations taken in the source population.

Lemma 1. *For any function* $\mathbf{b}(\mathbf{X}, Y)$*,*

$$E_0[\mathbf{b}(\mathbf{X}, Y)] = E_1[\rho(Y; \boldsymbol{\beta})\mathbf{b}(\mathbf{X}, Y)].$$

Proof. The proof is immediate:

$$\begin{aligned} E_0[\mathbf{b}(\mathbf{X}, Y)] &= \int\int \mathbf{b}(\mathbf{x}, y) f(\mathbf{x} \mid y) g_Y(y) \mathrm{d}x \, \mathrm{d}y \\ &= \int\int \mathbf{b}(\mathbf{x}, y) f(\mathbf{x} \mid y) f_Y(y) \rho(y; \boldsymbol{\beta}) \mathrm{d}x \, \mathrm{d}y \\ &= E_1[\mathbf{b}(\mathbf{X}, Y)\rho(Y; \boldsymbol{\beta})] \end{aligned}$$

□

Now let $\mathbf{m}(\mathbf{X}, Y, S) \in \Lambda^\perp$ and since S is binary and Y is not observed in the target population, we can write

$$\mathbf{m}(\mathbf{X}, Y, S) = \frac{S}{\pi}\mathbf{m}_1(X, Y) + \frac{1-S}{1-\pi}\mathbf{m}_0(X)$$

for some functions $\mathbf{m}_1$ and $\mathbf{m}_0$. Our goal is to uncover the relationship between $\mathbf{m}_0$ and $\mathbf{m}_1$ required for elements of $\Lambda^\perp$. We proceed in a sequence of steps.

1. Since $\mathbf{m}$ is an element of the Hilbert space, it has mean 0, so that

$$\begin{aligned} 0 &= E\left[\frac{S}{\pi}\mathbf{m}_1(\mathbf{X}, Y) + \frac{1-S}{1-\pi}\mathbf{m}_0(\mathbf{X})\right] \\ &= E_1[\mathbf{m}_1(\mathbf{X}, Y)] + E_0[\mathbf{m}_0(\mathbf{X})] \\ &= E_1[\mathbf{m}_1(\mathbf{X}, Y) + \rho(Y; \boldsymbol{\beta})\mathbf{m}_0(\mathbf{X})] \quad \text{by Lemma 1} \end{aligned}$$

2. Next, since $\mathbf{m}$ is an element of $\Lambda_1^\perp$, we have for every $\mathbf{h}(Y)$ such that $E_1(\mathbf{h}) = 0$,

$$\begin{aligned} 0 &= E\left[\frac{S}{\pi}\mathbf{m}_1(\mathbf{X}, Y)^{\mathrm{T}}\mathbf{h}(Y) + \frac{1-S}{1-\pi}\mathbf{m}_0(\mathbf{X})^{\mathrm{T}}\mathbf{h}(Y)\right] \\ &= E_1[\mathbf{m}_1(\mathbf{X}, Y)^{\mathrm{T}}\mathbf{h}(Y)] + E_0[\mathbf{m}_0(\mathbf{X})^{\mathrm{T}}\mathbf{h}(Y)] \\ &= E_1[(\mathbf{m}_1(\mathbf{X}, Y) + \rho(Y;\boldsymbol{\beta})\mathbf{m}_0(\mathbf{X}))^{\mathrm{T}}\mathbf{h}(Y)] \quad \text{by Lemma 1} \end{aligned}$$

3. Lastly, since m is an element of $\Lambda_2^\perp$, we have for every $c(\mathbf{X}, Y)$ such that $E[\mathbf{c}(\mathbf{X}, Y) \mid Y] = 0$,

$$\begin{aligned} 0 &= E\left[\frac{S}{\pi}\mathbf{m}_1(\mathbf{X}, Y)^{\mathrm{T}}\mathbf{c}(\mathbf{X}, Y) + \frac{1-S}{1-\pi}\mathbf{m}_0(\mathbf{X})^{\mathrm{T}}\mathbf{c}(\mathbf{X}, Y)\right] \\ &= E_1[\mathbf{m}_1(\mathbf{X}, Y)^{\mathrm{T}}\mathbf{c}(\mathbf{X}, Y)] + E_0[\mathbf{m}_0(\mathbf{X})^{\mathrm{T}}\mathbf{c}(\mathbf{X}, Y)] \\ &= E_1[(\mathbf{m}_1(\mathbf{X}, Y) + \rho(Y;\boldsymbol{\beta})\mathbf{m}_0(\mathbf{X}))^{\mathrm{T}}\mathbf{c}(\mathbf{X}, Y)] \quad \text{by Lemma 1} \end{aligned}$$

We then have for any $\mathbf{b}(\mathbf{X}, Y) \in \mathbb{R}^q$ such that $E_1[\mathbf{b}] = 0$,

$$\begin{aligned} E_1[(\mathbf{m}_1(\mathbf{X}, Y) + \rho(Y;\boldsymbol{\beta})\mathbf{m}_0(\mathbf{X}))^{\mathrm{T}}\mathbf{b}(\mathbf{X}, Y)] &= E_1\left\{(\mathbf{m}_1(\mathbf{X}, Y) + \rho(Y;\boldsymbol{\beta})\mathbf{m}_0(\mathbf{X}))^{\mathrm{T}}E[\mathbf{b} \mid Y]\right\} \\ &+ E_1\left\{(\mathbf{m}_1(\mathbf{X}, Y) + \rho(Y;\boldsymbol{\beta})\mathbf{m}_0(\mathbf{X}))^{\mathrm{T}}(\mathbf{b}(\mathbf{X}, Y) - E[\mathbf{b} \mid Y])\right\} = 0 + 0 = 0 \end{aligned}$$

by steps 2 and 3.

By step 1, $\mathbf{d}(\mathbf{X}, Y) \coloneqq \mathbf{m}_1(\mathbf{X}, Y) + \rho(Y;\boldsymbol{\beta})\mathbf{m}_0(\mathbf{X})$ itself satisfies $E_1[\mathbf{d}] = 0$, from which it follows that $E_1[\mathbf{d}^{\mathrm{T}}\mathbf{d}] = 0 \iff \mathbf{d} = 0$.

To recap, we have shown that every element $\mathbf{m}(\mathbf{X}, Y, S) \in \Lambda^\perp$ is of the form

$$\mathbf{m}(\mathbf{X}, Y, S) = \frac{S}{\pi}\mathbf{m}_1(\mathbf{X}, Y) + \frac{1-S}{1-\pi}\mathbf{m}_0(X), \tag{2}$$

for functions $\mathbf{m}_1$ and $\mathbf{m}_0$ satisfying

$$\mathbf{m}_1(\mathbf{X}, Y) + \rho(\mathbf{Y};\boldsymbol{\beta})\mathbf{m}_0(\mathbf{X}) = 0 \tag{3}$$

Substituting (3) into (2), we have

$$\mathbf{m}(\mathbf{X}, Y, S) = -\frac{S}{\pi}\rho(Y;\boldsymbol{\beta})\mathbf{m}_0(\mathbf{X}) + \frac{1-S}{1-\pi}\mathbf{m}_0(\mathbf{X})$$

So we obtain

$$\Lambda^{\perp} = \left\{ \left[\frac{1-S}{1-\pi} - \frac{S}{\pi}\rho(Y;\boldsymbol{\beta}) \right] \mathbf{m}_0(\mathbf{X}) \text{ for all } \mathbf{m}_0(\mathbf{X}) \in \mathbb{R}^q \right\}$$

Note that this leads directly to estimating equations

$$\frac{1}{1-\pi} \sum_{i\,:\,S_i=0} \mathbf{m}_0(\mathbf{X}_i) - \frac{1}{\pi} \sum_{i\,:\,S_i=1} \rho(Y_i;\boldsymbol{\beta}) \mathbf{m}_0(\mathbf{X}_i) = 0 \tag{4}$$

which we can solve for $\boldsymbol{\beta}$ to obtain an estimator $\hat{\boldsymbol{\beta}}$. This estimating equation will be used in our simulation study in Section 5. Before we get there, we first study whether we can find the element of the nuisance tangent space that gives rise to the most efficient estimator for $\boldsymbol{\beta}$.

4 The efficient score function $\mathbf{S}_{\text{eff}}$

The efficient score function for estimating $\boldsymbol{\beta}$ is obtained by projecting the score function with respect to $\boldsymbol{\beta}$ onto $\Lambda^{\perp}$. The score function with respect to $\boldsymbol{\beta}$ is

$$\begin{aligned}
\mathbf{S}_{\boldsymbol{\beta}} = \frac{\partial \log L}{\partial \boldsymbol{\beta}} &= \frac{(1-S)\int f(\mathbf{X} \mid y) f_Y(y) \frac{\partial \rho}{\partial \boldsymbol{\beta}}(y;\boldsymbol{\beta}) \mathrm{d}y}{\int f(\mathbf{X} \mid y) f_Y(y) \rho(y;\boldsymbol{\beta}) \mathrm{d}y} \\
&= (1-S) \frac{\int f(\mathbf{X} \mid y) f_Y(y) \rho(y;\boldsymbol{\beta}) \frac{\partial \log \rho}{\partial \boldsymbol{\beta}}(y;\boldsymbol{\beta}) \mathrm{d}y}{g_{\mathbf{X}}(\mathbf{X})} \\
&= (1-S) \int \frac{f(\mathbf{X} \mid y) g_Y(y)}{g_{\mathbf{X}}(\mathbf{X})} \frac{\partial \log \rho}{\partial \boldsymbol{\beta}}(y;\boldsymbol{\beta}) \mathrm{d}y \\
&= (1-S) E_0 \left[\frac{\partial \log \rho(Y;\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \mid \mathbf{X} \right]
\end{aligned}$$

So we'd like to find the unique functions $\mathbf{S}_{\text{eff}}$ and $\mathbf{a}$, elements of $\Lambda^{\perp}$ and Λ respectively, such that

$$(1-S) E_0 \left[\frac{\partial \log \rho(Y;\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \mid \mathbf{X} \right] = S_{\boldsymbol{\beta}} = \mathbf{S}_{\text{eff}} + \mathbf{a} \tag{5}$$

Recalling the forms that we have derived for arbitrary elements of Λ and $\Lambda^{\perp}$, we have

$$\mathbf{S}_{\text{eff}} = \left[\frac{1-S}{1-\pi} - \frac{S}{\pi}\rho(Y;\boldsymbol{\beta}) \right] \mathbf{m}_*(\mathbf{X})$$

for some $\mathbf{m}_*(\mathbf{X}) \in \mathbb{R}^q$ and

$$a = S[\mathbf{h}(Y) + \mathbf{c}(\mathbf{X}, Y)] + (1-S) E_0(\mathbf{h}(Y) + \mathbf{c}(\mathbf{X}, Y) \mid \mathbf{X})$$

for some $\mathbf{h}(Y), \mathbf{c}(\mathbf{X}, Y) \in \mathbb{R}^q$ satisfying $E_1(\mathbf{h}) = E(\mathbf{c}|Y) = 0$.

Theorem 1. *The choice of*

$$\mathbf{m}_*(\mathbf{X}) = \frac{E_0\left(\frac{\partial \log \rho(Y;\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \mid \mathbf{X}\right)}{\frac{E_0(\rho(Y;\boldsymbol{\beta})|\mathbf{X})}{\pi} + \frac{1}{1-\pi}}$$

yields the efficient score function.

Proof. We have to suppose that there exist $\mathbf{h}(Y), \mathbf{c}(\mathbf{X}, Y) \in \mathbb{R}^q$ satisfying $E_1(\mathbf{h}) = E(\mathbf{c}|Y) = 0$ and such that

$$\mathbf{h}(Y) + \mathbf{c}(\mathbf{X}, Y) = \pi^{-1}\rho(\mathbf{Y};\boldsymbol{\beta})\mathbf{m}_*(\mathbf{X})$$

We then have

$$\begin{aligned}
\mathbf{S}_{\text{eff}} + \mathbf{a} &= S[\mathbf{h}(Y) + \mathbf{c}(\mathbf{X}, Y) - \pi^{-1}\rho(\mathbf{Y};\boldsymbol{\beta})\mathbf{m}_*(\mathbf{X})] \\
&\quad +(1-S)[E_0(\mathbf{h}(Y) + \mathbf{c}(\mathbf{X}, Y) \mid \mathbf{X}) + (1-\pi)^{-1}\mathbf{m}_*(\mathbf{X})] \\
&= (1-S)[E_0(\pi^{-1}\rho(\mathbf{Y};\boldsymbol{\beta})\mathbf{m}_*(\mathbf{X}) \mid \mathbf{X}) + (1-\pi)^{-1}\mathbf{m}_*(\mathbf{X})] \\
&= (1-S)[E_0(\pi^{-1}\rho(\mathbf{Y};\boldsymbol{\beta}) \mid \mathbf{X}) + (1-\pi)^{-1}]\mathbf{m}_*(\mathbf{X}) \\
&= (1-S)E_0\left[\frac{\partial \log \rho(Y;\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \mid \mathbf{X}\right] \\
&= \mathbf{S}_{\boldsymbol{\beta}}
\end{aligned}$$

□

5 Simulation Study

In this section, we estimate $\boldsymbol{\beta}$ in a very simple setting. We will look at the estimator corresponding to a very simple element of the orthogonal complement of the nuisance tangent space: $\left[\frac{1-S}{1-\pi} - \frac{S}{\pi}\rho(Y;\boldsymbol{\beta})\right]\mathbf{X}$.

In our simple setting, we generate $Y \sim \text{Exp}(\lambda)$ and a single covariate $X \mid Y_i \sim N(Y_i, 1)$, and we then generate population membership as $S \mid Y \sim \text{Bern}(1, \text{expit}(\beta Y))$. It turns out that in this setting, the density ratio is given as

$$\rho(y;\beta) = \frac{g(y;\beta)}{f(y;\beta)} = \frac{\pi}{1-\pi}e^{-\beta y}, \tag{6}$$

where the dimension of our parameter of interest β is just 1.

Data generated in this way for three choices of β and a total sample size of $n_s + n_t = N = 10000$ are shown below. Setting $\beta = 0$ corresponds with the situation in which there is no label shift: the distribution of Y (and hence the X distribution) is identical in the target

and source population. Setting $\beta = 1$ and $\beta = 3$ produces label shift, i.e. differences in the target and source distributions, more pronounced in the case where $\beta = 3$, which we will call severe label shift.

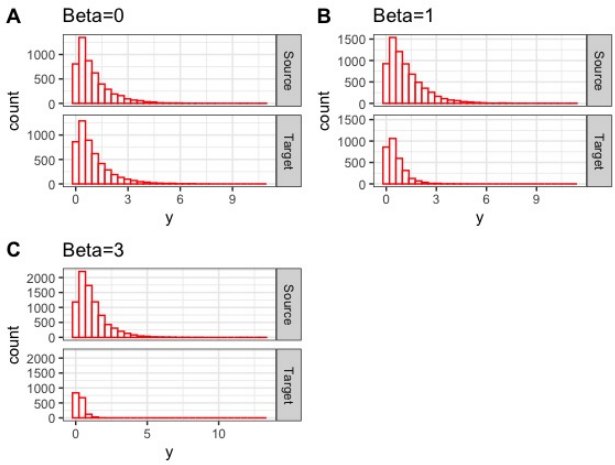

Figure 1: Distribution of Y in simulated data for **A:** $\beta = 0$, **B:** $\beta = 1$, **C:** $\beta = 3$

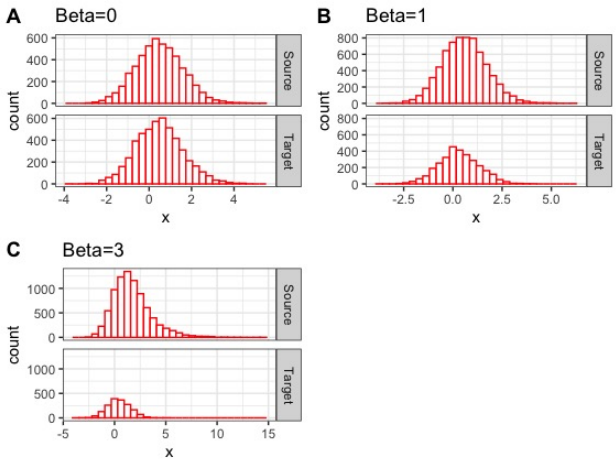

Figure 2: Distribution of X in simulated data for **A:** $\beta = 0$, **B:** $\beta = 1$, **C:** $\beta = 3$

Because of the way we generate the data (where population membership S is a logistic function of Y with coefficient β), we get that $n_t > n_s$ whenever $\beta > 0$, i.e. in the presence of leftward label shift. The larger we choose β, i.e. the more intense we make the shift, the smaller the sample from the source population becomes and the larger the sample from the target population.

Recall that our estimator is always obtained by solving equation (4), and in this simple study, we choose $\mathbf{m}_0(\mathbf{X}) = \mathbf{X}$.

For each β, we generate R=500 data sets and solve equation (4) to calculate our estimate of β for each one. To make the calculations go a bit faster, each data set is of size $n_{\text{source}} + n_{\text{target}} = N = 1000$, instead of N=10,000.

The empirical mean, median, and standard deviation of our estimators for the three different settings of β are provided in Table 1. In addition, histograms for

beta	mean	median	sd
0	1.23	0.00	27.38
1	1.03	0.99	0.27
3	379.18	2.91	3783.79

Table 1: Summary statistics calculated on simulated estimates

We see from the table that except for in the case $\beta = 1$, our estimator is biased. However, this is simply due to occasional extreme values of the estimator, which skew its distribution to the right. The possibility of estimating such extreme values may be due to computational inaccuracies in calculating the solution to (4).

However, the median estimate across the simulation is quite accurate (so the estimator is median-unbiased), and if we remove the outlier estimates, we get a nice, more-or-less

symmetric distribution of the estimates. Figure 3 shows histograms of the estimates once the extreme outliers are removed.

Figure 3: Distribution of estimates of with outliers removed