The average treatment effect is a common summary of the effect of a treatment on a continuously varying outcome. The ATE is the difference in the outcome between the subpopulation receiving the treatment and a control subpopulation, [[]]

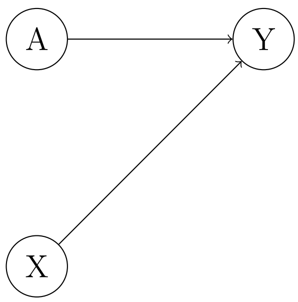
If the data consists only of the responses and the treatment indicators, the empirical ATE is efficient, in the sense that it achieves the lowest asymptotic variance among [[get class from tsiatis book]]. When covariates are available, further efficiency gains are attainable [[]]. Efficient estimators taking into account of covariates are described in [[]].

1. Method

1.1. Assume for the data the following model:

$$
\begin{aligned}
(Y_1, X_1, A_1), \ldots, (Y_n, X_n, A_n) &\overset{iid}{\sim} \mathcal{P} \\
A &\perp X \\
P(A = 1 \mid X) = P(A = 1) = 1 - P(A = 0) &= p
\end{aligned} \tag{1}
$$

for some law $\mathcal{P}$. In the context of a randomized trial with two arms, Y_i



is interpreted as the outcome of interest, X_i as a vector of covariates, and A_i as a binary indicator of treatment. The independence of A_i and X_i reflects an assumption of a random assignment mechanism. Beyond these restrictions on $\mathcal{P}$, certain regularity assumptions are also made on the [...].

The target of estimation is the average treatment effect,

$$\psi_0 = E(Y \mid A = 1) - E(Y \mid A = 0).$$

An estimator is obtained as the solution in ψ of

$$\sum_{i=1}^{n} U(Y_i, A_i; \psi) = 0,$$

where

$$U(Y, A; \psi) = (A - p)(Y - \psi A).$$

By a Taylor expansion, the influence function for $\hat{\psi}$ is

That is,

$$\sqrt{n}(\hat{\psi} - \psi_0) = [...]$$

A short calculation shows that the mean of the criterion function $U(Y_i, A_i; \psi)$ is 0 when $\psi = \psi_0$. The consistency and asymptotic normality of $\hat{\psi}$ then follow from the central limit theorem and [ref].
[1]

$$\begin{aligned}
E(U(Y, A; \psi_0)) &= E[(A - p)(Y - \psi_0 A)] \\
&= E[(A - p)(E(Y \mid A) - \psi_0 A)] \\
&= (E(Y \mid A = 1) - \psi_0 A)(1 - p)p + E(Y \mid A = 0)(-p)(1 - p) \\
&= p(1 - p)[E(Y \mid A = 1) - E(Y \mid A = 0) - \psi_0] = 0.
\end{aligned}$$

If the data consisted only of the outcomes Y_i and treatment indicators A_i, estimator $\hat{\psi}$ is semiparametric efficient, that is, its asymptotic variance $\operatorname{Var} U(Y_i, A_i; \psi_0)$ is not greater than the asymptotic variance of any other regular asymptotically linear estimator. [cite]

The availability of covariates provides an opportunity for further efficiency gains. [cite]. The semiparametric efficient influence function may be obtained by minimizing the variance of

over arbitrary measurable h, and is found to be [cite] [should we re-derive this result here?],

$$W(X, Y, A; \psi) = (A - p)[Y - (1 - p)E(Y \mid A = 1, X) - pE(Y \mid A = 0, X)] - p(1 - p)\psi$$

A drawback of estimators taking account of covariates is that the relationship of the covariates to the outcome must usually be modeled consistently. For example, $E(Y \mid A = 0, X)$ and $E(Y \mid A = 1, X)$ in [ref] must be modeled in order to obtain the semiparametric estimator. This modeling process introduces the possibility of bias by an analyst [cite]. One solution is to specify in advance of seeing the data the model to be used. A disadvantage of this solution is [what disadvantage? should any modeling decisions be made based on the data?] [Tsiatis 2008] presents a solution that allows an analyst to exercise fuller range of modeling expertise. Briefly, each regression $E(Y \mid A = 0, X)$ and $E(Y \mid A = 1, X)$ is modeled by separate teams

---

[1]

$$\begin{aligned}
0 = \sum_i U(X_i, Y_i, A_i; \hat{\psi}) &= \sum_i U(X_i, Y_i, A_i; \psi_0) + (\hat{\psi} - \psi_0) \sum_i U'(X_i, Y_i, A_i; \psi_*) \\
n^{1/2}(\hat{\psi} - \psi_0) &= -(n^{-1} \sum_i U'(X_i, Y_i, A_i; \psi_*))^{-1} \times n^{-1/2} \sum_i U(X_i, Y_i, A_i; \psi_0) \\
&\rightsquigarrow -E(U'(X, Y, A; \psi_*))^{-1} \times \mathcal{N}(0, \sigma^2)
\end{aligned}$$

where use the lemma in concluding that the normal distribution in the last line has mean zero. Implying root-n consistency.

or individuals, each with access only to data corresponding to their arm. The resulting regression estimates may then be combined in [[]] to obtain $\hat{\psi}$.

1.2. We propose another approach, eliminating the regressions on treatment in []. Specifically, by weighting the response

$$Y = \tilde{Y}\frac{p^A(1-p)^{1-A}}{(1-p)^A p^{1-A}} = \tilde{Y}\left(\frac{p}{1-p}\right)^{2A-1}.$$

the terms may be combined as

The criterion function [ref] may be rewritten as

In case $p = P(A = 1) = P(A = 0) = 1/2$,

$$W(X, Y, A; \psi) = (A - 1/2)(Y - E(Y \mid X)) - \psi/4. \tag{2}$$

An analyst given the weighted data may then model the regression on covariates. The estimate may then be substituting in [] to obtain the semi-parametric estimate of the ATE.

This approach avoids the need for a separate analyst for each arm. Additionally, a single regression estimate will often be more precise than two combined regression estimates. [toy example?]

1.3. In some fields it is common to regress out a variable of interest Y on covariates, studying instead the residuals $Y' = Y - E(Y \mid X)$. The average treatment effect of Y' is obtained as the slope in the model $E(Y') = \beta_0 + \beta_1 A$. [Example from epi literature] This estimator is typically inefficient.

Define

$$\tilde{Y} = Y - E(Y \mid X)$$

and consider the regression

$$E(\tilde{Y} \mid A) = \beta_0 + \beta_1 A.$$

Then $\beta_0 + \beta_1 = E(\tilde{Y} \mid A = 1) = E(Y \mid A = 1) - E(Y)$ and $\beta_0 = E(\tilde{Y} \mid A = 0) = E(Y \mid A = 0) - E(Y) = E(Y \mid A = 0, X) - pE(Y \mid A = 1, X) - (1 - p)E(Y \mid A = 0, X) = p(E(Y \mid A = 0, X) - E(Y \mid A = 1, X))$, so

$$\begin{aligned}
\beta_0 &= -p\psi_0 \\
\beta_1 &= \psi_0.
\end{aligned}$$

The influence function of (β_0, β_1) is obtained as:

$$
\begin{aligned}
0 &= \sum_{i=1}^{n} \begin{pmatrix} 1 \\ A_i \end{pmatrix} (\tilde{Y}_i - \hat{\beta}_0 - A_i \hat{\beta}_1) \\
&= \sum_{i=1}^{n} \left\{ \begin{pmatrix} 1 \\ A_i \end{pmatrix} (\tilde{Y}_i - \beta_0 - A_i \beta_1) + \begin{pmatrix} -1 & -A_i \\ -A_i & -A_i \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 - \beta_0 \\ \hat{\beta}_1 - \beta_1 \end{pmatrix} \right\} \\
n^{1/2} \begin{pmatrix} \hat{\beta}_0 - \beta_0 \\ \hat{\beta}_1 - \beta_1 \end{pmatrix} &= \left(\frac{1}{n} \sum_i \begin{pmatrix} 1 & A_i \\ A_i & A_i \end{pmatrix} \right)^{-1} n^{-1/2} \sum_i \begin{pmatrix} 1 \\ A_i \end{pmatrix} (\tilde{Y}_i - \beta_0 - A_i \beta_1) \\
&= \begin{pmatrix} 1 & p \\ p & p \end{pmatrix}^{-1} n^{-1/2} \sum_i \begin{pmatrix} 1 \\ A_i \end{pmatrix} (\tilde{Y}_i - \beta_0 - A_i \beta_1) + o_P(1) \\
n^{1/2} (\hat{\beta}_1 - \beta_1) &= n^{-1/2} \sum_i \begin{pmatrix} \frac{-1}{1-p} & \frac{1}{p(1-p)} \end{pmatrix} \begin{pmatrix} 1 \\ A_i \end{pmatrix} (\tilde{Y}_i - \beta_0 - A_i \beta_1) + o_P(1) \\
&= \frac{n^{-1/2}}{p(1-p)} \sum_i (A_i - p)(\tilde{Y}_i - \beta_0 - A_i \beta_1) + o_P(1) \\
&= \frac{n^{-1/2}}{p(1-p)} \sum_i (A_i - p)(\tilde{Y}_i - (A_i - p)\psi_0) + o_P(1) \\
&= \frac{n^{-1/2}}{p(1-p)} \sum_i \left\{ (A_i - p)(Y_i - E(Y \mid X_i)) - p^2 \left(\frac{1-p}{p} \right)^{2A_i} \psi_0 \right\} + o_P(1).
\end{aligned}
$$

In case $p = 1/2$,

$$
n^{1/2} (\hat{\beta}_1 - \beta_1) = 4 n^{-1/2} \sum_i \{ (A_i - 1/2)(Y_i - E(Y \mid X_i)) - \psi_0/4 \} + o_P(1).
$$

By comparison with (2), we find that when $p = 1/2$, the augmented estimator $\hat{\psi}$ is asymptotically equivalent to $\hat{\beta}_1$.

Since the estimator [[]] is efficient, the estimator [[]] is inefficient when $p \neq 1/2$, i.e., its asympototic variance is larger. Specifically, the asymptotic variance of [[]] is

and the asymptotic variance of [[]] is

.

The difference is therefore

$$
(1 - 2p)^2 \operatorname{Var}(E(Y \mid A = 1, X) - E(Y \mid A = 0, X)),
$$

which is 0 if and only $p = 1/2$ or the stratified ATEs $E(Y \mid A = 1, X) - E(Y \mid A = 0, X)$ are constant.

2. Simulation

As an illustration, suppose the response follows a linear model,

$$Y = \alpha A + \beta^T X + \gamma^T AX + \epsilon.$$

The fixed parameters γ represent interaction between A and X. The errors ϵ are assumed to have mean zero and the variables A, X, ϵ are assumed mutually independent. In this case, after a short calculation, the difference [[]] evaluates to

$$(1-2p)^2 \gamma^T \operatorname{Var}(X)\gamma.$$

When there are no interactions, $\gamma = 0$, the difference vanishes, since in this case the stratified ATEs $E(Y \mid A = 1, X) - E(Y \mid A = 0, X)$ are constant. By another short calculation, the asymptotic relative efficiency of $\hat{\beta}_1$ to $\hat{\psi}$ is

$$1 + \frac{(1-2p)^2 \gamma^T \operatorname{Var}(X)\gamma}{\frac{\operatorname{Var}(\epsilon)}{p(1-p)} + \gamma^T \operatorname{Var}(X)\gamma}.$$

See Fig. [[]].

3. Other estimands

As above, the full data is $Y_i^* = (Y_i^*(0), Y_i^*(1)), i = 1, \ldots, n$, the observed data is $(Y_i, A_i, X_i), i = 1, \ldots, n$, and we assume

$$\begin{aligned} &Y = AY^*(1) + (1-A)Y^*(0), \\ &P(A = 1) = p \in (0, 1) \\ &A \perp X, A \perp Y^*. \end{aligned}$$

Besides the mean treatment difference $E(Y \mid A = 1) - E(Y \mid A = 0)$ discussed above, we consider other estimands:

(1) $\psi_0 = \log \frac{E(Y^*(1))}{E(Y^*(0))} = \log \frac{E(Y|A=1)}{E(Y|A=0)}$
(2) the slope in the model

$$\operatorname{logit}(AE(Y^*(1)) + (1-A)E(Y^*(0))) = \operatorname{logit}(P(Y = 1 \mid A)) = \psi_0 + \psi_1 A,$$

for a binary-valued response Y

In each case, we obtain the efficient augmented influence function following the approach of [[Tsiatis ch. 13]]:

(1) obtain a full-data influence function $\phi^F(Y^*)$
(2) obtain an observed data influence function $\phi(Y, A, X)$ corresponding to ϕ^F under the mapping $\phi \mapsto E(\phi \mid Y^*)$
(3) compute the efficient augmentation term

$$\begin{aligned} h^*(Y, A, X) &= (A - p)(E(\phi \mid A = 1, X) - E(\phi \mid A = 0, X)) \\ &= E(\phi \mid A, X) - E(\phi \mid X) \end{aligned}$$

We then eliminate regressions on treatment level, i.e., the terms $E(Y \mid A = 1, X)$ and $E(Y \mid A = 0, X)$.

3.1. $\log \frac{E(Y|A=1)}{E(Y|A=0)}$**.** The problem is to estimate

$$\psi_0 = \log \frac{E(Y^*(1))}{E(Y^*(0))} = \log \frac{E(Y \mid A=1)}{E(Y \mid A=0)}.$$

A full-data estimator is given by the solution to

$$\sum_i (Y_i^*(1) - e^{\psi_0} Y_i^*(0)) = 0,$$

with influence function

$$\begin{aligned}
\phi^F(Y, A, X; \psi) &= (e^{\psi} E(Y^*(0)))^{-1}(Y^*(1) - e^{\psi} Y^*(0)) \\
&= (E(Y^*(1)))^{-1}(Y^*(1) - e^{\psi} Y^*(0)).
\end{aligned}$$

An influence functions ϕ of the observed data satisfies

$$\begin{aligned}
E(Y_i^*(1))\phi(Y, A, X) &= \left(\frac{A}{p} - e^{\psi_0}\frac{1-A}{1-p}\right) Y + h(Y, A, X) \\
&= (A-p)\left(\frac{A}{(A-p)p} - e^{\psi_0}\frac{1-A}{(A-p)(1-p)}\right) Y + h(Y, A, X) \\
&= \frac{A-p}{p(1-p)}(A + e^{\psi_0}(1-A))Y + h(Y, A, X) \\
&= \frac{A-p}{p(1-p)} e^{(1-A)\psi_0} Y + h(Y, A, X),
\end{aligned}$$

where h satisfies $E(h(Y, A, X) \mid Y^*) = 0$. The minimizing h is given by subtracting out

$$\begin{aligned}
h^*(Y, A, X) &= (A-p)[E(\frac{A-p}{p(1-p)} e^{(1-A)\psi_0} Y \mid A = 1, X) - E(\frac{A-p}{p(1-p)} e^{(1-A)\psi_0} Y \mid A = 0, X)] \\
&= (A-p)[\frac{1}{p} E(Y \mid A = 1, X) + \frac{e^{\psi_0}}{1-p} E(Y \mid A = 0, X)].
\end{aligned}$$

The efficient influence function is therefore

$$\begin{aligned}
\phi^*(Y, A, X) &= (E(Y^*(1)))^{-1}\phi(Y, A, X) - h^*(Y, A, X) \\
&= (E(Y^*(1)))^{-1}(A-p)\left(\frac{e^{(1-A)\psi_0}}{p(1-p)} Y - \frac{1}{p} E(Y \mid A = 1, X) - \frac{e^{\psi_0}}{1-p} E(Y \mid A = 0, X)\right)
\end{aligned}$$

Let $\hat{\psi}_n$ be a consistent estimator of ψ_0. Under the transformation

$$Y = \frac{p^{2A}(1-p)^{2(1-A)}}{e^{(1-A)\hat{\psi}_n}} \tilde{Y}$$

the efficient influence function may be rewritten

$$\phi^*(Y, A, X) = (E(Y^*(1)))^{-1}(A-p)\left(\frac{e^{(1-A)\psi}}{p(1-p)} Y - E(\tilde{Y} \mid X)\right) + o_P(1).$$

In case $p = 1/2$, $\tilde{Y} = 4e^{(1-A)\hat{\psi}_n}Y$, and

(3)
$$\phi^*(Y, A, X) = 2(E(Y^*(1)))^{-1}(2A-1)[e^{(1-A)\psi}Y - E(e^{(1-A)\hat{\psi}_n}Y \mid X)] + o_P(1).$$

3.1.1. *Two-step regression.* Let

$$Z = Y - e^{(A-1)\psi_0}[E(e^{(1-A)\psi_0}Y \mid X) - E(Y \mid A = 1)]$$

and consider the log-linear regression model

$$\log(E(Z \mid A)) = \beta_0 + \beta_1 A. \tag{4}$$

Then

$$E(e^{(1-A)\psi_0}Y) = (1/2)[e^{\psi_0}E(Y \mid A = 0) + E(Y \mid A = 1)] = E(Y \mid A = 1)$$

implies

$$\begin{aligned} E(Z \mid A = 1) &= E(Y \mid A = 1) - E[E(e^{(1-A)\psi_0}Y \mid X) - E(Y \mid A = 1) \mid A = 1] \\ &= E(Y \mid A = 1) - E(e^{(1-A)\psi_0}Y) + E(Y \mid A = 1) \\ &= E(Y \mid A = 1), \end{aligned}$$

and similarly

$$\begin{aligned} E(Z \mid A = 0) &= E(Y \mid A = 0) - E[E(e^{(1-A)\psi_0}Y \mid X) - E(Y \mid A = 1) \mid A = 0] \\ &= E(Y \mid A = 0). \end{aligned}$$

Therefore, under model (4),

$$\begin{aligned} \beta_0 &= \log(E(Z \mid A = 0)) = \log(E(Y \mid A = 0)), \\ \beta_0 + \beta_1 &= \log(E(Z \mid A = 1)) = \log(E(Y \mid A = 1)), \\ \beta_1 &= \frac{\log(E(Y \mid A = 1))}{\log(E(Y \mid A = 0))} = \psi_0. \end{aligned}$$

An estimator $(\hat{\beta}_0, \hat{\beta}_1)$ under the log-linear regression model (4) is given by the estimating equations

$$0 = \sum_{i=1}^{n} \begin{pmatrix} 1 \\ A_i \end{pmatrix} (Z_i - e^{\hat{\beta}_0 + \hat{\beta}_1}).$$

The influence function of $\beta_1 = \psi_0$ is computed to be

$$
\begin{aligned}
\phi_{\beta_1}(Y, A, X; \beta) &= 2(2A-1)(e^{-\beta_0-\beta_1 A} Z - 1) \\
&= 2(2A-1)\left(\frac{Z}{E(Z \mid A)} - 1\right) \\
&= 2(2A-1)\left(\frac{e^{(1-A)\beta_1} Z}{E(Z \mid A=1)} - 1\right) \\
&= 2(2A-1)\frac{e^{(1-A)\beta_1} Y - E(e^{(1-A)\psi_0} Y \mid X) + E(Y \mid A=1) - E(Z \mid A=1)}{E(Z \mid A=1)} \\
&= 2(2A-1)\frac{e^{(1-A)\beta_1} Y - E(e^{(1-A)\psi_0} Y \mid X)}{E(Y \mid A=1)} \\
&= 2(E(Y^*(1)))^{-1}(2A-1)[e^{(1-A)\psi_0} Y - E(e^{(1-A)\psi_0} Y \mid X)].
\end{aligned}
$$

This influence function is the same as the influence function (3), so the estimator $\hat{\beta}_1$ under the log-linear regression is asymptotically equivalent to the efficient estimator $\hat{\psi}$.