Chapter 12
Targeted ANCOVA Estimator in RCTs

Daniel B. Rubin, Mark J. van der Laan

In many randomized experiments the primary goal is to estimate the average treatment effect, defined as the difference in expected responses between subjects assigned to a treatment group and subjects assigned to a control group. Linear regression is often recommended for use in RCTs as an attempt to increase precision when estimating an average treatment effect on a (nonbinary) outcome by exploiting baseline covariates. The coefficient in front of the treatment variable is then reported as the estimate of the average treatment effect, assuming that no interactions between treatment and covariates were included in the linear regression model.

In this setting, regression is actually not necessary but can lead to efficiency gains relative to the unadjusted estimator if the covariates are predictive of subject responses, and the consistency and asymptotic normality of the estimator of the average treatment effect does not depend on the linear model being correctly specified. However, we show that the usual least squares approach is a suboptimal way to fit a linear model in randomized experiments for the purpose of estimating the average treatment effect. A simple alternative linear regression fit utilizing TMLE guarantees that the average treatment effect estimator will be asymptotically efficient among a large class of popular methods. In addition, we argue and show that this TMLE often outperforms other proposed techniques if the sample size is small or moderate relative to the number of covariates, so that one can safely adjust for more predictors.

For a subject in the study, let W denote a vector of such baseline covariates. Let variable A be an indicator of treatment assignment, so that $A = 0$ signifies assignment to the control group, and $A = 1$ signifies assignment to the treatment group. Finally, let Y denote the primary outcome measurement that is taken on the subject at the end of the study. For our purposes it will not matter if Y is a continuous measurement, is restricted to some range, or even is a binary indicator.

It will be convenient to use counterfactuals, described in Chap. 2 as a consequence of the SCM. For a subject in the study, let Y_1 denote the response that the subject would have realized if he or she had been assigned to treatment. Likewise, let Y_0 be the response if the subject had been assigned to control. In reality each subject

M.J. van der Laan and S. Rose, *Targeted Learning: Causal Inference for Observational and Experimental Data*, Springer Series in Statistics, DOI 10.1007/978-1-4419-9782-1_12,

is assigned to only one group, either treatment or control, so one of these counterfactual outcomes will be missing. The observed response is $Y = AY_1 + (1 - A)Y_0$. The full data we would have liked to measure about a subject are $X = (W, Y_0, Y_1)$, while what we actually measure is $O = (W, A, Y)$. As this is an RCT, we assume the treatment assignment indicator A is independent of the full data (W, Y_0, Y_1).

For defining parameters we assume a superpopulation statistical model in which the study subjects are drawn with replacement from some larger population of subjects. That is, we assume the full data $(W_i, Y_{0,i}, Y_{1,i})$, $i = 1, ..., n$, are independent and identically distributed random triples. This assumption is mainly for simplicity. Freedman (2008a) shows how regression asymptotics can be analyzed in sequences of finite population statistical models, where the only randomness is that induced by the random assignment of subjects to treatment or control groups. The probability distribution of the full data structure (W, Y_0, Y_1) is unspecified. In addition, let $g_0(A \mid X)$ denote the probability distribution of treatment A, given X: by assumption, $g_0(1 \mid X) = g_0(1)$. Let P_0 denote the probability distribution of the observed data structure $(W, A, Y = Y_A)$.

The mean of the counterfactuals can be identified as a parameter of the probability distribution P_0:

$$\psi_0^{(1)} = E_0(Y \mid A = 1) = E_0(Y_1)$$

and

$$\psi_0^{(0)} = E_0(Y \mid A = 0) = E_0(Y_0),$$

the expected responses of subjects assigned to treatment or control. For quantifying the treatment effect, a parameter can then be defined as the contrast $\psi_0 = \psi_0^{(1)} - \psi_0^{(0)}$ between the mean responses among those assigned to the two arms (i.e., the average treatment effect). Note that we also have $\psi_0 = E_0(E_0(Y \mid A = 1, W) - E_0(Y \mid A = 0, W))$.

The statistical model for the probability distribution P_0 of $O = (W, A, Y)$ is identified by the nonparametric model for the full-data distribution, and that A is independent of the full-data structure. Thus the only testable assumption is that A is independent of W. This defines now the statistical model $\mathcal{M}$, and target parameter mapping $\Psi : \mathcal{M} \to \mathbb{R}$, $\Psi(P) = E_P(E_P(Y \mid A = 1, W) - E_P(Y \mid A = 0, W))$, and thereby the estimation problem.

The usual ANCOVA approach for using covariates to estimate this treatment effect ψ_0 on a continuous outcome is to use linear least squares to regress response Y on the treatment assignment indicator A and covariates W and then report the estimated coefficient in front of the treatment indicator.

In this chapter we propose an alternative to least squares fitting based on the TMLE algorithm such that:

- The treatment effect estimator generally becomes more efficient when $g(0) \neq 0.5$;

- The estimator tends to perform better with small or moderate samples than other common estimators with equivalent asymptotic efficiency, so more covariates can safely be used for the adjustment;
- The treatment effect estimator is the coefficient in front of A of a parsimonious fitted linear regression model. The technique should therefore be acceptable to nonstatistician investigators who are already familiar with interpreting regression coefficients in textbook linear models.

Two-sample problem, or one i.i.d. sample? Suppose there are n subjects in the study, and a randomly selected subgroup of m of them are assigned to treatment, with the random selection not depending on covariates. Here n and m are fixed. Let $g = m/n$ represent the proportion assigned treatment. In many studies $m = n/2$, so g will be $1/2$.

In our statistical formulation above we assume that A is a Bernoulli random variable. If in truth A is Bernoulli with probability (say) 0.5, then by chance the treatment group will not be of the same size as the control group. However, by design, the study often arranges the treatment and control groups to be of the same size, showing that it is not completely accurate to state that the sample is an i.i.d. sample from (W, A, Y) with $P_0(A = 1) = 0.5$.

This suggests another description of the data-generating distribution of the actual observed data structure. Suppose that $O = (W, A, Y) \sim P_0$ is the random variable in which A is random with $P_0(A = 1) = 0.5$, and the causal effect is identified by $\Psi(P_0)$ defined above, but that our observed data consist of a sample of n observations from (W, Y), conditional on $A = 1$, and a sample of n observations from (W, Y), given $A = 0$. That is, we took a "biased" "case-control" sample from P_0, where a case is defined as "$A = 1$." (This type of sampling has been referred to in other literature as a particular type of cohort sampling, and we note that the "case-control" terminology we use here is not typically applied to sampling conditional on A.) The results for semiparametric estimation based on case-control data (van der Laan 2008a; Rose and van der Laan 2008) state that, without loss of consistency or efficiency, we can apply an estimator developed for an i.i.d. sample from P_0, *but* we have to assign weights $q_0 = P_0(A = 1)$ to the observations with $A = 1$ and $1 - q_0$ for the observations with $A = 0$ in the pooled case-control sample. We also refer the interested reader to Chaps. 13 and 14.

As a consequence, the case-control-weighted TMLE is identical to the non-weighted TMLE. Therefore one can simply apply the estimators developed under i.i.d. sampling from P_0 and act as if the two-sample problem was an i.i.d. sample from O. Indeed, in this article, we proceed under our posed statistical model and suffice with the remark that all our estimators and results apply by letting $g_n(1)$ play the role of this set g.

12.1 Previously Proposed Estimators

In this section we review the strengths and weaknesses of common methods for estimating the treatment effect previously defined.

Unadjusted estimation. The simplest approach is to ignore baseline covariates altogether. Then, the obvious estimator of the expected response in the treatment group is the empirical mean of responses for subjects assigned to treatment, and analogously for the control subjects. Estimators for $\psi_0^{(0)}$, $\psi_0^{(1)}$, and ψ_0 become

$$\psi_n^{(0)} = \frac{1}{n}\sum_{i=1}^{n} \frac{(1-A_i)}{g_n(0)} Y_i,$$

$$\psi_n^{(1)} = \frac{1}{n}\sum_{i=1}^{n} \frac{A_i}{g_n(1)} Y_i,$$

and $\psi_n = \psi_n^{(1)} - \psi_n^{(0)}$.

The unadjusted estimator of treatment effect is consistent, and it is also asymptotically normal, in that $\sqrt{n}(\psi_n - \psi_0)$ will converge in law to an $N(0, \sigma^2)$ distribution (e.g., Yang and Tsiatis 2001). The influence curve of this estimator is given by $IC(Q, g_0) = h_{g_0}(A)(Y - \bar{Q}(A))$, where $\bar{Q}(A) = E_0(Y \mid A)$ and $h_{g_0}(A) = (2A-1)/g_0(A)$, so that $\sigma^2 = P_0 IC(Q, g_0)^2$ is the variance of this influence curve.

The asymptotic variance σ^2 of this limiting normal distribution can be used to gauge the precision of this estimator, and compare it to other estimators. Unfortunately, it is known that the unadjusted estimator can be much less efficient than other techniques when covariates are predictive of the response. This is because only one of the two counterfactual responses can be measured for a subject, while the covariates may contain information about what the missing response would have been.

ANCOVA. As noted earlier, the most popular way to adjust for covariates is to use linear least squares in fitting the regression model:

$$Y = \alpha + \psi_0 A + \gamma^\top W + \text{error}.$$

The least squares fit of ψ_0 then estimates the average treatment effect defined earlier. Let $\bar{Q}_I$ denote the limit of the linear regression ANCOVA estimator of $\bar{Q}_0(A, W) = E_0(Y \mid A, W)$. This ANCOVA estimator of ψ_0 will also generally be asymptotically normal, even if $E_0(Y \mid A, W)$ is not actually linear in (A, W), or if the errors are not homoscedastic or exogenous (Yang and Tsiatis 2001; Leon et al. 2003; Tsiatis et al. 2008). See the previous chapter for a more general presentation and proof of this result for RCTs based on the observation that the TMLE of ψ_0 that takes as initial estimator a maximum likelihood estimator according to a (possibly misspecified) generalized linear regression model will result in no update in the TMLE step. As a consequence, the ANCOVA estimator is a TMLE targeting $(E_0(Y_0), E_0(Y_1))$, corresponding with squared error loss and linear fluctuation

$\bar{Q}_n^0(\epsilon) = \bar{Q}_n^0 + \epsilon_1 A/g_n(1) + \epsilon_2(1-A)/g_n(0)$. Note that the targeting step in the TMLE corresponds with adding $\epsilon(1, A)$ and that $\epsilon_n = 0$ since $(1, A)$ is already included in the working linear regression model. Invoking the known asymptotics of the TMLE, it follows that it is asymptotically linear with influence curve

$$IC(Q, g_0)(O) = D^*(Q_I, g_0)(O) - C(A),$$

where

$$D^*(Q, g_0) = h_{g_0}(A)(Y - \bar{Q}(A, W)) + \bar{Q}(1, W) - \bar{Q}(0, W) - \psi_0$$

is the efficient influence curve of Ψ at (Q, g_0), and C is a correction term, due to g_n being an estimator of g_0, defined as $C(A) = E_0(D^*(Q, g_0)(O) \mid A)$. It is easy to verify that, if Q is such that $\Psi^{(1)}(Q) = \psi_0^{(1)}$, $\Psi^{(0)}(Q) = \psi_0^{(0)}$, then it follows that $C = 0$. This holds for the limit Q of a TMLE that targets both $E_0(Y_1)$ and $E_0(Y_0)$ such as this ANCOVA estimator. As a consequence, the ANCOVA estimator is asymptotically linear with influence curve $D^*(Q, g_0)$.

The only real additional assumption for the asymptotic linearity of this estimator is that the distributions of W and Y do not have overly heavy tails. The linear model can thus be viewed as a working model, used in an intermediate step to estimate the average treatment effect.

However, Freedman (2008a) shows that unless $g_0(1) = 0.5$, this ANCOVA estimator can be less efficient than the unadjusted estimator, in terms of asymptotic variance. It might also be biased in finite samples under model misspecification, although the $n^{-1/2}$-scale asymptotic normality result suggests that this bias will quickly become negligible relative to the variance.

The properties of this method are not fully understood when the number of covariates is large relative to the sample size, as the asymptotic approximations may begin to break down. Therefore, the usual recommendation is to simply adjust for an a priori specified handful of covariates that are considered to be the most important predictors.

ANCOVA II. A simple extension of ANCOVA is to add interaction terms and fit the model

$$Y = \theta_1 + \theta_2 A + \theta_3^T W + \theta_4^T (A \times W) + \text{error}$$

with linear least squares. The estimate of the treatment effect is no longer a coefficient fit, but the value obtained when using the fitted model $\bar{Q}_n$ of $\bar{Q}_0$ to impute missing counterfactuals, i.e.,

$$\psi_n = \Psi(Q_n) = \frac{1}{n} \sum_{i=1}^{n} \{\bar{Q}_n(1, W_i) - \bar{Q}_n(0, W_i)\}.$$

Again, as remarked above, this estimator of ψ_0 equals the same TMLE mentioned above, but now using as initial estimator the least squares regression fit of this parametric working model.

By the same arguments as above, this estimator is asymptotically linear with influence curve $D^*(Q_{II}, g_0)$, where Q_{II} denotes the limit of this linear regression ANCOVA II estimator of $\bar{Q}_0$. The ANCOVA II estimator is also asymptotically normal under the same minimal conditions we have discussed, and its asymptotic variance is guaranteed to be at least as small as the ANCOVA estimator and unadjusted estimator, while under many data-generating distributions it is asymptotically more efficient (Yang and Tsiatis 2001).

In fact, the ANCOVA II approach possesses an optimality property (Tsiatis et al. 2008). To appreciate this optimality property, one must know about the following alternative representation of $D^*(Q, g_0)$ for Q satisfying $\Psi(Q) = \psi_0$:

$$D^*(Q, g_0)(O) = h_{g_0}(A)(Y - f(Q)(W)) - \psi_0 \equiv D_{f(Q)}(\psi_0)(O), \tag{12.1}$$

where

$$f(Q)(W) \equiv g_0(1)\bar{Q}(0, W) + g_0(0)\bar{Q}(1, W)$$

and $h_{g_0}(A_i) = A_i/g_0(1) - (1 - A_i)/g_0(0)$. This representation defines a class of estimators $\psi_n(f) = 1/n \sum_i h_{g_0}(A_i)(Y_i - f(W_i))$ as solutions of the estimating equation $P_n D_f(\psi) = 0$, indexed by a choice f. These estimators are consistent and asymptotically linear with influence curve $h_{g_0}(A)(Y - f(W))$. Suppose two estimators are asymptotically equivalent if their difference is of order $o(n^{-1/2})$ in probability, under the distribution P_0 governing (W, A, Y). The asymptotic variance of the ANCOVA II estimator is no larger than that of any regular asymptotically normal estimator that is asymptotically equivalent to one of the form

$$\psi_n(f) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{A_i}{g_0} - \frac{1 - A_i}{1 - g_0} \right) (Y_i - f(W_i)) \tag{12.2}$$

for $f(W) = \eta^\top(1, W)$ linear in the components of W. The unadjusted, ANCOVA, and ANCOVA II estimators are all asymptotically equivalent to estimators in this class corresponding with functions $f(W)$ that are linear in W.

However, the ANCOVA II estimator is based on a less parsimonious model than the usual ANCOVA. It essentially fits separate linear regressions in the treatment and control arms and thus can be less stable with small or moderate samples due to loss of degrees of freedom. An analyst using the ANCOVA II method might therefore adjust for fewer covariates than someone using the regular ANCOVA technique, and thus make less use of potentially informative predictors.

Interestingly, in the special case that $g_0(1) = 0.5$, it happens that $f(Q_I) = f(Q_{II})$, so that the ANCOVA estimator and the ANCOVA II estimator have identical influence curves ($D^*(Q_I, g_0) = D^*(Q_{II}, g_0)$) (Yang and Tsiatis 2001; Leon et al. 2003; Tsiatis et al. 2008). Apparently the simple ANCOVA estimator achieves the same asymptotic efficiency as the more data-adaptive ANCOVA II estimator, and should thus be favored in small samples.

Koch estimator. Koch et al. (1998) introduced the treatment effect estimator defined by

$$\psi_n = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{A_i}{g_n(1)} - \frac{1-A_i}{1-g_n(1)}\right)Y_i + \frac{n}{m(n-m)}V^\top U^{-1}\sum_{i=1}^{n}(A_i - g_n(1))W_i.$$

Here $m = \sum_{i=1}^{n} A_i$, $V = V^{(0)}/(n-m) + V^{(1)}/m$ and $U = U^{(0)}/(n-m) + U^{(1)}/m$. Matrices $U^{(0)}$ and $U^{(1)}$ are unbiased sample estimates of covariance matrices of W in the control and treatment groups. Vectors $V^{(0)}$ and $V^{(1)}$ are likewise unbiased sample estimates of covariances between elements of W and the response Y in the control and treatment groups.

This estimator is also asymptotically normal, with the same asymptotic variance as the ANCOVA II estimator (Tsiatis et al. 2008), although it is motivated from a different perspective. Hence, it also has the optimality property of being asymptotically efficient among estimators asymptotically equivalent to those in (12.2).

This estimator appears to not be consistent with fitting a regression model for the response on both covariates and treatment assignment, i.e., it is not a substitution estimator. Additionally, like the ANCOVA II method, it requires estimating covariances between the outcome and each element of the covariate vector within each of the two treatment arms. If the sample size is small relative to the number of covariates, this might lead to more instability than the unadjusted or standard ANCOVA estimators.

Leon estimator. Leon et al. (2003) discuss a general class of estimators asymptotically equivalent to those of the form of (12.2), but with f a linear combination of given basis functions of W, and they mention using quadratic or cross product terms. We consider the reduction of their method when $f(W)$ must be a linear combination of the elements of W, along with an intercept. The estimator is defined by

$$\psi_n = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{A_i}{g_n(1)} - \frac{1-A_i}{1-g_n(1)}\right)Y_i - n(S_1/m^2 + S_2/(n-m)^2)^\top S_3^{-1} S_4.$$

To explain the definition, let $F = [1, W]^\top$ be the addition of constant 1 to a subject's covariate vector. Here,

$$S_3 = \sum_{i=1}^{n} F_i F_i^\top,$$

$$S_4 = \sum_{i=1}^{n}(A_i - g_n(1))F_i,$$

$$S_1 = \sum_{i=1}^{n} A_i(Y_i - \bar{Y}_{1,n})F_i,$$

and

$$S_2 = \sum_{i=1}^{n}(1-A_i)(Y_i - \bar{Y}_{0,n})F_i,$$

where $\bar{Y}_{1,n}$ and $\bar{Y}_{0,n}$ are the empirical means of response Y in the treatment and control groups.

This estimator is asymptotically equivalent to the ANCOVA II and Koch estimators. Like these two estimators, in both treatment arms it requires computing covariances between the outcome and each element of the covariate vector. This is not a computational issue, but it causes instability in small samples. Also, the method does not seem to correspond with fitting a simple linear regression model for the outcome on treatment assignments and covariates, and is thus not a substitution estimator.

12.2 Targeted ANCOVA

We now introduce a new treatment effect estimator. Recall the linear model

$$Y = \alpha + \psi_0 A + \gamma^\top W + \text{error}$$

used in the ANCOVA approach. Rather than fit coefficients with linear least squares, we proceed as follows. First, we let δ_n and γ_n minimize:

$$\sum_{i=1}^{n} \left(\frac{A_i}{g_n(1)} - \frac{1 - A_i}{1 - g_n(1)} \right)^2 |Y_i - \delta - \gamma^\top W_i|^2.$$

This is a weighted linear least squares regression of the response on the covariates, with an intercept, weighting subjects in the treatment group by $g_n(1)^{-2}$ and subjects in the control group by $(1 - g_n(1))^{-2}$. We have thus fitted γ. Next, let α_n and ψ_n minimize the sum of squares:

$$\sum_{i=1}^{n} |Y_i - \gamma_n^\top W_i - \alpha - \psi_0 A_i|^2.$$

That is, we regress response Y on an intercept and A, using the initial weighted least squares regression as offset. The targeted ANCOVA estimate of the treatment effect is ψ_n^*. Let $\bar{Q}_n^*$ be the targeted ANCOVA regression fit. Note that $\psi_n^* = \Psi(Q_n^*)$. One can also estimate expected responses in the treatment and control arms by using the fitted regression model $\bar{Q}_n^*$ to impute missing counterfactuals and obtain $\psi_n^{(0)} = \Psi^{(0)}(Q_n^*) = n^{-1} \sum_{i=1}^{n} (\alpha_n + \gamma_n^\top W_i)$ and $\psi_n^{(1)} = \Psi^{(1)}(Q_n^*) = \psi_n^{(0)} + \psi_n^*$.

This estimator equals the TMLE using the squared error loss and linear regression submodel $\bar{Q}_n^0(\epsilon) = \bar{Q}_n^0 + \epsilon(1, A)$, with the additional feature that the initial estimator $\bar{Q}_n^0$ is a *weighted* least squares estimator according to a linear regression model of Y in W. Specifically,

- The initial estimator $\bar{Q}_n^0$ of $\bar{Q}_0(A, W) = E_0(Y \mid A, W)$ is targeted by minimizing a weighted least squares criterion, and $Q_{W,0}$ is estimated with the empirical distribution $Q_{W,n}$. This defines the initial estimator Q_n^0 of $Q_0 = (Q_{W,0}, \bar{Q}_0)$. The

weighted least squares loss function $L_{g_n}(\bar{Q})(O) = h^2_{g_n}(A)(Y - \bar{Q}(A,W))^2$ is still a valid loss function for $\bar{Q}_0$ but is tailored to correspond with minimizing the asymptotic variance of the resulting TMLE.

- For the TMLE step, we use a squared error loss function $L(\bar{Q})(O) = (Y - \bar{Q}(A,W))^2$ for $\bar{Q}_0$ and a log-likelihood loss function $L(Q_W) = -\log Q_W$ for $Q_{W,0}$. This results in a loss function $L(Q) = L(\bar{Q}) + L(Q_W)$ for $Q_0 = (Q_{W,0}, \bar{Q}_0)$ for the TMLE step.
- For the TMLE step, we use the linear regression submodel $\bar{Q}^0_n(\epsilon_2) = \bar{Q}^0_n + \epsilon_2 H^*$, where $H^*(A,W) = (1,A)$, or, equivalently, $H^*(A,W) = (A/g_n(1), (1-A)/g_n(0))$, is chosen so that the score $d/d\epsilon_2 L(\bar{Q}^0_n(\epsilon_2))$ at $\epsilon_2 = 0$ spans the component $D^*_Y(\bar{Q}^0_n, g_n)(O) = (2A-1)/g_n(A)(Y - \bar{Q}^0_n(W))$ of the efficient influence curve $D^*(Q^0_n, g_n) = D^*_Y(\bar{Q}^0_n, g_n) + D^*_W(Q^0_n)$. The empirical distribution of W is separately fluctuated with a submodel $Q_{W,n}(\epsilon_1) = (1 + \epsilon_1 D^*_W(Q^0_n))Q_{W,n}$ with score $D^*_W(Q^0_n)(O) = \bar{Q}^0_n(1,W) - \bar{Q}^0_n(0,W) - \Psi(Q^0_n)$. Since $Q_{W,n}$ is a nonparametric maximum likelihood estimator, the maximum likelihood estimators of ϵ_1 in the TMLE algorithm equals zero. The score of $L(Q(\epsilon_1, \epsilon_2))$ at $\epsilon_1 = \epsilon_2 = 0$ spans the efficient influence curve $D^*(Q^0_n, g_n)$.

12.2.1 Asymptotic Optimality

What are the statistical properties of this estimator $\Psi(Q^*_n)$ of the additive causal effect of treatment ψ_0? Since the targeted ANCOVA estimator is a TMLE targeting $(E_0(Y_1), E_0(Y_0))$, we can refer to asymptotic linearity theorems established for such estimators. Since $\bar{Q}^*_n$ is a simple linear regression estimator, all the empirical process conditions and convergence rate conditions are trivially met. As a consequence, $\psi^{(0)}_n$, $\psi^{(1)}_n$, and ψ^*_n will all be consistent and asymptotically normal estimators of the respective target parameters under minimal conditions, such as W and Y not having overly heavy tails.

Moreover, we claim that the asymptotic variance of treatment effect estimator $\psi^*_n = \Psi(Q^*_n)$ will equal that of the ANCOVA II estimator, Koch estimator, and Leon estimator. Thus, we achieve the optimality bound discussed for estimators of the form (12.2) with linear $f(W)$ and are guaranteed to be at least as asymptotically efficient as the unadjusted and standard ANCOVA techniques.

This is shown as follows. The TMLE $Q^*_n = (Q_{W,n}, \bar{Q}^*_n)$ solves the efficient influence curve estimating equation $P_n D^*(Q^*_n, g_n, \psi^*_n) = 0$, where we denoted the efficient influence curve $D^*(Q, g_0)$ as an estimating function $D^*(Q, g_0, \Psi(Q))$ in ψ. By (12.1) this can also be represented as $P_n D_{f(Q^*_n)}(\psi^*_n) = 0$, and thereby

$$\psi^*_n = \frac{1}{n}\sum_i h_{g_n}(A_i)(Y_i - f(Q^*_n)(W_i)).$$

Recall that $\bar{Q}^*_n = \bar{Q}^0_n + \epsilon_n(1,A)$, so that $f(Q^*_n) = f(Q^0_n) + c$ for some constant c. However, this constant c cancels out since $P_n h_{g_n} c = P_n h_{g_n} = 0$. Thus, it follows that

$$\psi_n^* = \frac{1}{n}\sum_i h_{g_n}(A_i)(Y_i - f(Q_n^0)).$$

Standard analysis now shows that ψ_n^* is asymptotically linear with influence curve

$$IC(O) = D_{f(Q)}(\psi_0)(O) - E_0(D_{f(Q)}(\psi_0)(O) \mid A),$$

where $D_{f(Q)}(\psi_0)$ can also be represented as $D^*(Q, g_0, \psi_0)$, Q denotes the limit of Q_n^0, and the additional projection term is due to the estimation of g_0 with g_n. However, by definition of $\bar{Q}_n^0$ as the linear regression that minimizes the empirical variance of $D_f^*(\psi_0)$ over all linear functions $f(W) = \eta^\top(1, W)$, we know that

$$\text{var } D_{f(Q)}(\psi_0) = \arg\min_\eta \text{var}\{h_{g_0}(A)(Y - \eta^\top(1, W)) - \psi_0\}.$$

The additional term in IC can only reduce the variance, which proves that the variance of IC is smaller than or equal to the variance of all the influence curves $D_f^*(\psi_0)$ with $f(W) = \eta^\top(1, W)$ for some η. Since the additional term only changes the intercept in f, it also follows that the additional term will not affect the variance relative to the already optimal $f(Q)$. Thus, $IC = D_{f(Q)}^*(\psi_0) = D^*(Q, g_0, \psi_0)$.

12.2.2 Targeted ANCOVA Is a Substitution Estimator

A byproduct of our method is a fit of a parsimonious linear model. Hence, targeted ANCOVA could be easier to use than the Koch or Leon methods for nonstatistician investigators who are already used to fitting parametric linear regression models. If the linear ANCOVA model is a good approximation to the unknown data-generating distribution, then, like linear least squares, our fit should accurately approximate the regression function $(A, W) \to E_0(Y \mid A, W)$. To see this, note that by first fitting the coefficient vector γ of covariates W and then fixing it in the next step, we are merely implementing forward stagewise modeling with weights in one of the stages, which is a well-known regression technique (Hastie et al. 2001, Sect. 10.3). Since A and W are independent, the separation of the two stages does not harm the fit of $\bar{Q}_0$.

12.2.3 Small and Moderate Sample Performance

While we have noted that our targeted ANCOVA technique will perform similarly in large samples to the ANCOVA II, Koch, and Leon methods, we claim that asymptotics will often kick in more quickly for the targeted ANCOVA estimator, so we could safely adjust for more predictors. Simulations in the following sections will be used to investigate this issue in more depth, but for now we give an explanation for our confidence.

Table 12.1 Summary of ANCOVA methods and their properties

	Unadjusted	ANCOVA	ANCOVA II	Koch	Leon	Targeted ANCOVA
Meets asymptotic bound of (12.2)			×	×	×	×
Parametric regression model		×	×			×
Doesn't estimate $2p$ parameters	×	×				×

Suppose that the covariate vector is p-dimensional. The ANCOVA II estimator reduces to performing two linear regressions of the response on the covariate vector, one in each arm. Hence, fits in each model are based on fewer observations than with ordinary ANCOVA. Similarly, the Koch and Leon estimators both involve estimating the $2p$ quantities corresponding to how each of the p baseline predictors covaries with the response in each treatment arm. The optimal linear $f(W)$ in (12.2) involves a mixture of two p-dimensional vectors: the vector of covariances between predictors and the response in the treatment arm, and likewise for controls (Tsiatis et al. 2008, Eq. 11). While earlier methods attaining the efficiency bound fit both vectors, we try to directly fit the optimal linear $f(W)$ in (12.2) and implicitly just estimate the relevant mixture $f(Q_0)(W) = g_0(1)\bar{Q}_0(0, W)+g_0(0)\bar{Q}_0(1, W)$ according to a linear working model. A related way of viewing targeted ANCOVA is that relative to ANCOVA II and other techniques, we improve finite sample performance by sacrificing asymptotic efficiency for our separate estimators of expected responses $\psi_0^{(0)}$ and $\psi_0^{(1)}$ in the two arms, as the treatment effect contrast will usually be of primary importance. There is subjectivity in these statements, just as there are many ways to represent estimators and how many parameters they fit in intermediate steps. Still, Table 12.1 seems to summarize the added value of our method.

12.3 Standard Error Estimation

Yang and Tsiatis (2001), Leon et al. (2003), and Tsiatis et al. (2008) give semiparametric representations of asymptotic variances for estimators in this problem, and this framework can be applied to our method. Alternatively, we use that our targeted ANCOVA estimator $\psi_n^* = \Psi(Q_n^*)$ is a TMLE that solves the efficient influence curve estimating equation $0 = P_n D^*(Q_n^*, g_0, \psi_n^*) = 0$, so that inference can proceed accordingly. Specifically, the TMLE $\Psi(Q_n^*)$ is asymptotically linear with an influence curve given by

$$D^*(Q^*, g_0)(O) = \frac{2A-1}{g_0(A)}(Y - \bar{Q}^*(A,W)) + \bar{Q}^*(1,W) - \bar{Q}^*(0,W) - \psi_0,$$

where Q^* denotes the limit of the TMLE Q_n^*.

Let α_n, ψ_n^*, and γ_n denote the fitted coefficients from targeted ANCOVA, with α, ψ_0, and γ the corresponding large sample limits. We have that $\bar{Q}_n^*(A,W) = \alpha_n + \gamma_n W + \psi_n^* A$, which converges to $\bar{Q}^*(A,W) = \alpha + \gamma + \psi_0 A$. The variance of $\psi_n = \Psi(Q_n^*)$ can thus be estimated as

$$\sigma^2_{\psi_n,n} = \frac{1}{n^2}\sum_{i=1}^{n} D^*(Q_n^*, g_0)(O_i)^2.$$

The same variance formulas can be applied to estimate the variance of the TMLE of E_0Y_1 and E_0Y_0. These formulas correspond with the formulas in the above referenced articles.

Confidence intervals and test statistics can now be constructed based on the normal approximation of $\psi_n^* - \psi_0$. Although the consistency and asymptotic normality properties of the standard ANCOVA estimator of the treatment effect do not depend on the linear model's being correctly specified, the usual nominal variance formulas produced by the software can be incorrect (Freedman 2008a). The above standard error estimators for targeted ANCOVA here do not depend on any parametric modeling assumptions since they are based on the influence curve of the estimator in the posed semiparametric model that only assumed the randomization assumption.

12.4 Simulations

We investigated the performance of targeted ANCOVA through simulating four data-generating distributions studied in Yang and Tsiatis (2001). For each distribution we evaluated estimators in simulated experiments with sample sizes $n = 20$, $n = 50$, and $n = 100$. In addition to redoing the original Yang and Tsiatis simulations, in which each subject had a single covariate W, we also considered scenarios with more covariates. For each subject, three additional covariates were generated from the same marginal distribution originally used, but independently of all the subject's other measurements. This was to simulate clinical trial settings where substantial baseline information is available yet "most covariates are not strongly related to the outcome" (Pocock et al. 2002).

For each of the four data-generating distributions, each of the three sample sizes, and both choices of including one covariate or four covariates, we ran 100,000 Monte Carlo simulations of randomized experiments. The true treatment effect in all cases was $\psi_0 = 1/2$, and the Monte Carlo replications allowed us to estimate the root mean squared errors (RMSEs) of different estimators. Although the results that follow do not present estimates of simulation error, the number of replications was chosen to be large enough so that this simulation error can be ignored.

Table 12.2 Simulation 1: RMSE

Extra covariates	n	Unadjusted	ANCOVA	ANCOVA II	Koch	Leon	Targeted ANCOVA
No	20	0.69	0.59	0.60	0.59	0.59	0.56
	50	0.43	0.37	0.38	0.37	0.37	0.37
	100	0.31	0.26	0.26	0.26	0.26	0.26
Yes	20	0.69	0.65	0.69	0.65	0.62	0.53
	50	0.44	0.39	0.39	0.39	0.38	0.36
	100	0.31	0.27	0.27	0.27	0.27	0.26

Table 12.3 Simulation 2: RMSE

Extra covariates	n	Unadjusted	ANCOVA	ANCOVA II	Koch	Leon	Targeted ANCOVA
No	20	0.46	0.47	0.47	0.47	0.47	0.45
	50	0.29	0.29	0.29	0.29	0.29	0.29
	100	0.21	0.21	0.21	0.21	0.21	0.20
Yes	20	0.46	0.52	0.56	0.52	0.49	0.43
	50	0.29	0.30	0.30	0.30	0.30	0.28
	100	0.21	0.21	0.21	0.21	0.21	0.20

Table 12.4 Simulation 3: RMSE

Extra covariates	n	Unadjusted	ANCOVA	ANCOVA II	Koch	Leon	Targeted ANCOVA
No	50	0.51	0.42	0.43	0.43	0.43	0.42
	100	0.36	0.30	0.30	0.30	0.30	0.30
Yes	50	0.51	0.44	0.47	0.45	0.45	0.42
	100	0.36	0.30	0.31	0.31	0.31	0.30

Table 12.5 Simulation 4: RMSE

Extra covariates	n	Unadjusted	ANCOVA	ANCOVA II	Koch	Leon	Targeted ANCOVA
No	20	0.99	0.81	0.84	0.81	0.80	0.77
	50	0.63	0.53	0.55	0.53	0.53	0.52
	100	0.45	0.39	0.39	0.39	0.39	0.38
Yes	20	0.99	0.90	0.96	0.90	0.85	0.72
	50	0.63	0.55	0.56	0.55	0.55	0.51
	100	0.45	0.39	0.40	0.39	0.39	0.38

Simulation 1. For the initial simulation the covariate followed a standard normal distribution. Half of the subjects were assigned to treatment, meaning that $g_0 = 1/2$. Responses were generated through

$$Y = (-1/4 + \psi_0 A) + (\beta_1 + \beta_2 A)W + (\beta_3 + \beta_4 A)(W^2 - \text{var}(W)) + U,$$

where $(\beta_1, \beta_2, \beta_3, \beta_4) = (1/2, 3/5, 2/5, 3/10)$. The error U followed a standard normal distribution, independently of covariates. RMSEs are shown in Table 12.2. Targeted ANCOVA appeared slightly more accurate than other methods, particularly as the sample size got smaller or the number of baseline covariates got larger. The unadjusted estimator was noticeably less efficient than all covariate-adjusted estimators, as the covariate was strongly predictive of the response for this artificial data-generating distribution.

Simulation 2. In the second simulation we took $(\beta_1, \beta_2, \beta_3, \beta_4) = (1/10, 1/10, 1/10, 1/10)$, making the baseline covariate less predictive of the response. Recall that in settings where three extra covariates were added, these were unrelated to the outcome. RMSEs are reported in Table 12.3. For this distribution the unadjusted estimator performed more favorably. All methods were mostly similar, but it was notable that targeted ANCOVA had the best performance across each of these six independent Monte Carlo settings.

Simulation 3. The third simulation was identical to the first, except that the proportion of subjects assigned treatment was $g_0 = 3/10$ instead of $1/2$. Following Yang and Tsiatis (2001), we don't report results for the $n = 20$ sample size as too few subjects were assigned treatment to make meaningful generalizations. RMSE results are shown in Table 12.4. The unadjusted estimator appeared slightly worse than the others. Adjusted estimators were all similar, with targeted ANCOVA consistently having slightly smaller RMSEs than its competitors over these four settings.

Simulation 4. The final simulation was identical to simulation 1, except covariates W and error U were drawn from t-distributions with seven degrees of freedom instead of standard normal distributions. Table 12.5 shows RMSEs. Once more, we found that targeted ANCOVA performed best, particularly when the sample size became small or when the three extra covariates were added. An unusual feature of the simulations was that, unlike the other estimators, targeted ANCOVA occasionally seemed to perform better with the three extra covariates than without them, even though they were unrelated to the treatment or response.

12.5 Discussion

We have introduced a new alternative to least squares for fitting linear models in RCTs. Our estimator of the average treatment effect generally increases the asymp-

totic efficiency of the usual ANCOVA approach and still produces a regression fit. It may also often outperform estimators with similar asymptotic efficiency if the sample size is moderate relative to the number of covariates, so that one could safely adjust for more predictors. Our work is a special case of a TMLE, which is based on the idea that fitting a regression model isn't always an end in itself. Rather, it is an intermediate step in estimating the target parameter of interest, which in our case is the average causal effect. In such circumstances it may be suboptimal to use standard parametric maximum likelihood or least squares for model fitting of the initial estimator in the TMLE and advantageous to keep the final desired estimand in mind while targeting the initial regression fit accordingly. This approach has beneficial applications in a variety of problems, including RCTs. In this chapter we focused on using a parametric regression working model. Instead, the initial estimator $\bar{Q}_n^0$ in the targeted ANCOVA could be replaced by a super learner based on the weighted squared error loss function $L_{g_n}(\bar{Q}) = (Y - f(\bar{Q}))^2 h_{g_n}^2$. We also refer the interested reader to Appendix A.19.

To summarize, in the TMLE one has the option to select a loss function for the initial estimator, separate from the loss function selected for the targeting step in the TMLE. For example, this loss function can be selected so that minimizing its empirical risk over candidate TMLEs $Q_n^*(Q^0)$ indexed by different initial estimators Q^0 (or values such as regression functions) corresponds with minimizing the variance of the efficient influence curve $D^*(Q_n^*(Q^0), g_0)$ at these TMLEs over a working model for Q^0. As a result of such a procedure, the variance of the influence curve of the TMLE will be the variance of $D^*(Q^*, g_0)$, where Q^* has been tailored to minimize the variance of these influence curves over a specified set of candidate functions for Q^*. In this chapter's example, this loss function was the weighted least squares loss $L_{g_n}(\bar{Q}) = (Y - f(\bar{Q}))^2 h_{g_n}^2$ and the working model was linear regression functions in W for $\bar{Q}$. To conclude, we showed that TMLE can accommodate the incorporation of additional targeting of the initial estimator through empirical efficiency maximization (Rubin and van der Laan 2008), so that additional optimality properties, such as having an influence curve that is more optimal than a user-supplied class of influence curves, can be guaranteed.

Disclaimer

This work concerns only the views of the authors and does not necessarily represent the position of the Food and Drug Administration.