Chapter 8
Direct Effects and Effect Among the Treated

Alan E. Hubbard, Nicholas P. Jewell, Mark J. van der Laan

Researchers are frequently interested in assessing the direct effect of one variable on an outcome of interest, where this effect is not mediated through a set of intermediate variables. In this chapter, we will examine direct effects in a gender salary equity study example. Such studies provide one measure of the equity of procedures used to set salaries and of decisions in promoting and advancing faculty based on performance measures. The goal is to assess whether gender, as determined at birth, has a direct effect on the salary at which a faculty member is hired, not mediated through intermediate performance of the subject up until the time the subject gets hired. If such a direct effect exists, then that means the salary was set in response not only to merit but also to the gender of the person, indicating a gender inequality issue.

We will start by defining the SCM and a natural direct effect of gender on salary controlling for the intermediate variables. In addition, we present the identifiability assumptions under which this causal quantity can be identified from the distribution of the observed data. We commit to an estimand and nonparametric statistical model, even though we accept that the estimand cannot be interpreted as the desired causal direct effect in the gender inequality study. Nevertheless, it represents an effect of gender that controls for the *measured* intermediate variables, and thereby represents a "best" approximation of the desired causal direct effect, given the restrictions of the observed data. We present TMLE for this estimand in a nonparametric statistical model and apply it to our gender equity data set.

8.1 Defining the Causal Direct Effect

The observed data structure is $O = (W, A, Y) \sim P_0$. Here, Y represents the salary for a specific year, A refers to gender with $A = 1$ for females and $A = 0$ for males, and W is the set of intermediate predictive factors available. In our data these intermediate

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_8,

factors include (1) the nature of the highest degree received, (2) years since receipt of highest degree, and (3) years since appointment at the institution.

We define the full data modeled by an SCM as $(X, U) \sim P_{X,U,0}$, where $X = (A, W, H, Y)$ are the endogenous nodes and U denotes the exogenous factors drawn from some distribution $P_{U,0}$. That is, given $U = u$, $X = (A, W, H, Y)$ is deterministically generated by a collection of functions, $f_A(u_A)$,$f_W(A, u_W)$, $f_H(A, W, u_H)$, $f_Y(A, W, H, u_Y)$. This SCM implies a counterfactual random variable $Y(a, w, h) = Y(a, w, h)(U)$ corresponding with intervening on the SCM by setting $A = a, W = w, H = h$, while keeping U random. H is a binary variable indicating whether or not a person is hired.

Because we only observe subjects if they have been hired, the actual observed data O follows the conditional distribution of (A, W, Y), given $H = 1$. We can now define a definition of the "gender" effect as a weighted average of the w-specific controlled direct effects of gender, where the weights are with respect to the probability distribution of the intermediate variables, given one is female ($A = 1$) and hired ($H = 1$). We will refer to this causal quantity as a generalized natural direct effect (NDE) parameter:

$$\Psi^F_{NDE}(P_{X,U,0}) = \sum_w E_{X,U,0}[Y(1, w, H = 1) - Y(0, w, H = 1)]Q_{W,0}(w), \qquad (8.1)$$

with $Q_{W,0}(w) \equiv P_0(W = w \mid A = 1, H = 1)$. As discussed in van der Laan and Petersen (2008), the so-called NDE of treatment A on outcome Y, controlling for intermediate variables W, can be presented as $E_0[\sum_w(Y(1, w) - Y(0, w))P_0(W = w \mid A = 0)]$. In our case the only difference is that we are averaging the counterfactual differences within strata w with weights $P_0(W = w \mid A = 1, H = 1)$, so that we should indeed use the same terminology.

Under the causal graph assumption of no unblocked backdoor path from (A, W, H) to Y through the Us, or, equivalently, that (A, W, H) is independent of the counterfactuals $Y(a, w, h) = f_Y(a, w, h, U_Y)$, we can identify this causal quantity $\Psi^F_{NDE}(P_{X,U,0})$ from the probability distribution of the observed data structure O. Specifically, one can write parameter (8.1) as the following parameter mapping applied to the true observed data distribution P_0:

$$\begin{aligned}\Psi_{NDE}(P_0) &= E_0[E_0(Y \mid A = 1, W, H = 1) - E_0(Y \mid A = 0, W, H = 1) \mid A = 1, H = 1] \\ &= E_0[Y - E_0(Y \mid A = 0, W, H = 1) \mid A = 1, H = 1]. \qquad (8.2)\end{aligned}$$

Note that the estimand defines a statistical parameter mapping $\Psi_{NDE} : \mathcal{M} \to \mathbb{R}$, where $\mathcal{M}$ is the nonparametric statistical model. For notational convenience, henceforth we suppress the conditioning on $H = 1$ in all conditional distributions.

Standard practice for estimating the estimand in gender equity studies is to fit a standard parametric regression among the men for $E_0[Y \mid A = 0, W]$, use the resulting fit to predict the outcomes among the women, compute for each female the difference between the observed outcome and this predicted outcome, and average all these differences. This is indeed a particular method for estimation of estimand (8.2) obtained by substitution of a parametric regression fit for $E_0(Y \mid A = 0, W)$, and the empirical distribution for the conditional distribution of W, given $A = 1$. Since in truth we lack the knowledge that warrants a parametric model, we pose a nonparametric statistical model $\mathcal{M}$ for the probability distribution of $O = (W, A, Y)$. We now have to develop an estimator of the desired estimand in this nonparametric statistical model.

It is understood that the causal interpretation of the estimand as an actual natural direct causal effect is very questionable as relevant variables were not collected, such as merit, on the pathway from gender to the outcome. However, the causal model makes explicit the desired causal quantity of interest and provides a framework to understand what intermediate variables W need to be measured to make estimand (8.2) approach the causal direct effect one desires. For example, if one wishes to exclude certain intermediate variables due to other considerations, then the bias resulting from such steps could be studied analytically or through Monte Carlo simulations.

Conveniently, a causal effect among the treated defined in another SCM as $W = f_W(U_W)$, $A = f_A(W, U_A)$, $Y = f_Y(W, A, U_Y)$, under the randomization assumption that U_A is independent of U_Y, is identified by the same estimand (8.2) above (van der Laan 2010c). In that case, W are confounders of the treatment of interest, A, instead of being on the causal pathway.

As a consequence, the TMLE of the estimand for the effect among the treated in one SCM is identical to the TMLE of the estimand for the NDE in our SCM. Of course, the interpretation of the estimand is a function of the assumed SCM. So, though we present a TMLE of an average controlled direct effect among the treated in our causal model, we can also use this TMLE to estimate the causal effect among the treated for another causal model. This example nicely shows the distinct tasks of defining a causal parameter of interest in a causal model for the full data (that is, a function of the distribution of U, X) and developing estimators of the corresponding statistical estimand.

8.2 TMLE

We have defined a specific estimand $\Psi_{NDE}(P_0)$ as well as provided identifiability conditions under which one can interpret the parameter value as a type of weighted-average controlled direct effect. In this section we will now and then suppress the NDE in the notation of this estimand since no other estimands are considered. Suppressing the conditioning on hiring ($H = 1$), the estimand can be represented as

$$\Psi_{NDE}(P_0) = E_0[E_0(Y \mid A = 1, W) - E_0(Y \mid A = 0, W) \mid A = 1]. \tag{8.3}$$

We factorize P_0 in terms of the marginal distribution $Q_{W,0}$ of W, the conditional distribution g_0 of A, given W, and the conditional distribution $Q_{Y,0}$ of Y, given A, W.

We note that $\psi_{NDE,0}$ depends on P_0 through $\bar{Q}_0(A,W) = E_0(Y \mid A,W)$, $Q_{W,0}$, and g_0. The TMLE presented below will yield a targeted estimator $\bar{Q}_n^*$ and g_n^* of $\bar{Q}_0$ and g_0, and the empirical distribution $Q_{W,n}$ of the marginal distribution $Q_{W,0}$ of W. Let $Q_0 = (Q_{W,0}, \bar{Q}_0)$, so that we can also present the estimand as $\Psi_{NDE}(Q_0, g_0)$. The TMLE of $\Psi_{NDE}(Q_0, g_0)$ is a substitution estimator $\Psi(Q_n^*, g_n^*)$ obtained by plugging in this estimator (Q_n^*, g_n^*). Due to the particular form of g_n^*, we show below that this substitution estimator corresponds with using the empirical distribution for the conditional distribution of W, given $A = 1$, so that the TMLE of the estimand $\psi_{NDE,0}$ can also be represented as

$$\Psi_{NDE}(Q_n^*, g_n^*) = \frac{1}{\sum_{i=1}^n I(A_i = 1)} \sum_{i=1}^n I(A_i = 1) * [\bar{Q}_n^*(1, W_i) - \bar{Q}_n^*(0, W_i)]. \tag{8.4}$$

To develop the TMLE we first need an initial estimator of the outcome regression $\bar{Q}_0$, and the treatment mechanism g_0, while estimating the marginal distribution of W with the empirical probability distribution of $W_1, \ldots, W_n$. We can estimate both $\bar{Q}_0$ and g_0 with loss-based super learning using the appropriate loss function for $\bar{Q}_0$ and log-likelihood loss function for g_0, respectively. If Y is binary, then we would use the log-likelihood loss function $L(\bar{Q})(O) = Y \log \bar{Q}(A,W) + (1-Y)\log(1 - \bar{Q}(A,W))$ for $\bar{Q}_0$. This same loss function can also be used if $Y \in [0,1]$, as shown in Chap. 7. If Y is continuous and bounded between a and b so that $P_0(a < Y < b) = 1$, we can either use the squared error loss function or this quasi-log-likelihood loss function applied to a linearly transformed $Y^* = (Y-a)/(b-a)$. Let $\bar{Q}_n^0, g_n^0$ be the resulting super learner fits of $\bar{Q}_0$ and g_0, respectively. This provides us with our initial estimator (Q_n^0, g_n^0) of (Q_0, g_0).

To determine the parametric submodel through the initial estimator that can be used to encode the fluctuations in the TMLE algorithm, we need to know the efficient influence curve of the target parameter $\Psi_{NDE} : \mathcal{M} \to \mathbb{R}$. This statistical target parameter was studied in van der Laan (2010c) in the context of a causal model for the causal effect among the treated, and the efficient influence curve at a $P \in \mathcal{M}$ was derived as (see also Appendix A)

$$\begin{aligned} D^*(P) &= \left(\frac{I(A=1)}{P(A=1)} - \frac{I(A=0)g(1 \mid W)}{P(A=1)g(0 \mid W)} \right) [Y - \bar{Q}(A,W)] \\ &\quad + \frac{I(A=1)}{P(A=1)} [\bar{Q}(1,W) - \bar{Q}(0,W) - \Psi(P)], \end{aligned} \tag{8.5}$$

where $\bar{Q} = \bar{Q}(P)$ and $g = g(P)$ are the conditional mean and probability distribution, respecively, under P. The first component, $D_Y^*(P)$, is a score of the conditional distribution of Y, given A, W, and the second component is a score $D_{A,W}^*(P)$ of the joint distribution of (A, W). The latter component of the efficient influence curve $D^*(P)$ can be orthogonally decomposed as $D_{A,W}^*(P) = D_W^*(P) + D_A^*(P)$, where

$$D^*_W(P) = \frac{g(1 \mid W)}{P(A=1)}(\bar{Q}(1,W) - \bar{Q}(0,W) - \Psi(P)) \text{ and}$$
$$D^*_A(P) = \frac{I(A=1) - g(1 \mid W)}{P(A=1)}(\bar{Q}(1,W) - \bar{Q}(0,W) - \Psi(P))$$

are scores for the marginal distribution of W and the conditional distribution of A, given W. One can represent (8.5) as a function of the parameter of interest, ψ, Q, and g, or $D^*(Q, g, \psi)$. The resulting estimating function is double robust, or, formally,

$$P_0 D^*(Q, g, \psi_0) = 0 \text{ if } Q = Q_0 \text{ or } g = g_0,$$

where $Pf \equiv \int f(o)dP(o)$. Since the TMLE Q^*_n, g^*_n solves the efficient influence curve equation, $P_n D^*(Q^*_n, g^*_n, \Psi(Q^*_n, g^*_n)) = 0$, this double robustness implies that the TMLE $\Psi(Q^*_n, g^*_n)$ is consistent for ψ_0 if either Q^*_n is consistent for Q_0 or g^*_n is consistent for g_0. Apparently, even though the estimand depends on both Q_0 and g_0, we still obtain a consistent estimator if either Q_0 or g_0 is consistently estimated!

The next step in defining the TMLE is to select loss functions $L_1(\bar{Q})$ and $L_2(g) = -\log g$ for $\bar{Q}_0$ and g_0, respectively, and construct parametric submodels $\{\bar{Q}^0_n(\epsilon_1) : \epsilon_1\}$ and $\{g^0_n(\epsilon_2) : \epsilon_2\}$ so that the two "scores"

$$\frac{d}{d\epsilon_1} L_1(\bar{Q}^0_n(\epsilon_1)) \text{ and}$$

$$\frac{d}{d\epsilon_2} L_2(g^0_n(\epsilon_2))$$

at $\epsilon_1 = \epsilon_2 = 0$ span the efficient influence curve $D^*(Q^0_n, g^0_n)$ at the initial estimator. If we use the squared error loss function $L_1(\bar{Q})(O) = (Y - \bar{Q}(A,W))^2$, then we select the linear fluctuation working model $\bar{Q}^0_n(\epsilon_1)(A,W) = \bar{Q}^0_n(A,W) + \epsilon_1 C_1(g^0_n)(A,W)$, where

$$C_1(g)(A,W) = I(A=1) - \frac{I(A=0)g(1 \mid W)}{g(0 \mid W)}.$$

If Y is binary or continuous in $[0,1]$ and we select the quasi-log-likelihood loss function for $L_1(\bar{Q})$, then we select the logisitic fluctuation working model logit $\bar{Q}^0_n(\epsilon_1) =$ logit $\bar{Q}^0_n + \epsilon_1 C_1(g^0_n)$. In the context of theoretical or practical violations of the positivity assumption, $P_0(g_0(0 \mid W) > 0) = 1$, as required for a bounded variance of the efficient influence curve, and Y being continuous, we strongly recommend the quasi-log-likelihood loss function $L_1(\bar{Q})$ for $\bar{Q}_0$, since the resulting TMLE will then fully respect the global bounds $[a,b]$ of the statistical model. As parametric submodel through g^0_n, we select $\text{logit}(g^0_n(\epsilon_2)(1 \mid W)) = \text{logit}(g^0_n(1 \mid W)) + \epsilon_2 C_2(g^0_n, Q^0_n)(W)$, where

$$C_2(Q,g)(W) = \bar{Q}(1,W) - \bar{Q}(0,W) - \Psi(Q,g).$$

Finally, we select the log-likelihood loss function $L(Q_W) = -\log Q_W$ for the probability distribution $Q_{W,0}$, and, as a parametric submodel through $Q_{W,n}$, we select $\{Q_{W,n}(\epsilon_3) = (1 + \epsilon_3 D^*_W(Q^0_n, g^0_n))Q_{W,n} : \epsilon_3\}$, where

$$D^*_W(Q^0_n, g^0_n)(W) = g^0_n(1 \mid W)\left\{\bar{Q}^0_n(1, W) - \bar{Q}^0_n(0, W) - \Psi(Q^0_n)\right\}.$$

We note that the scores with respect to the loss functions $L_1(\bar{Q})$, $L_2(g)$ and $L(Q_W)$ generated by these three submodels are $D^*_Y(Q^0_n, g^0_n) \equiv C_1(g^0_n)(Y - \bar{Q}^0_n)$, $D^*_A(Q^0_n, g^0_n) \equiv C_2(g^0_n, Q^0_n)(A - g^0_n(1 \mid W))$, and $D^*_W(Q^0_n, g^0_n)$, respectively, and the sum of these three scores equals $D^*(Q^0_n, g^0_n)$ up till the scalar $P_0(A = 1)$. Thus, if we define the single loss function $L(Q^0_n, g^0_n) = L_1(\bar{Q}^0_n) + L(Q_{W,n}) + L_2(g^0_n)$, then

$$\frac{d}{d\epsilon} L(Q^0_n(\epsilon), g^0_n(\epsilon))$$

at $\epsilon = 0$ spans the efficient influence curve $D^*(Q^0_n, g^0_n)$. That is, we successfully carried out the second step in defining the TMLE involving the selection of a loss function and corresponding parametric submodel through the initial estimator whose score spans the efficient influence curve at the initial estimator. The TMLE algorithm is now defined.

The maximum likelihood estimator of $\epsilon_n = (\epsilon_{1,n}, \epsilon_{2,n}, \epsilon_{3,n})$, according to the working parametric submodel through (Q^0_n, g^0_n), defines an updated fit $(Q^1_n = Q^0_n(\epsilon_{1,n}, \epsilon_{3,n})$, $g^1_n = g^0_n(\epsilon_{2,n}))$. Since we selected $Q_{W,n}$ to be the nonparametric maximum likelihood estimator of $Q_{W,0}$, we have that $\epsilon_{3,n} = 0$, so that the empirical distribution of W is not updated. This TMLE updating of $\bar{Q}^0_n$ and g^0_n is iterated until convergence, and we denote the final fit with $(Q^*_n = (\bar{Q}^*_n, Q_{W,n}), g^*_n)$. The TMLE of ψ_0 is the corresponding substitution estimator $\Psi(Q^*_n, g^*_n)$.

Due to the fact that $P_n D^*_{A,W}(Q^*_n, g^*_n) = 0$, it follows immediately that the TMLE equals the substitution estimator (8.4) obtained by plugging in $\bar{Q}^*_n$ for $\bar{Q}_0$, and plugging in the empirical probability distribution for the conditional distribution of W, given $A = 1$. Suppose that we replace in the TMLE algorithm the parametric submodel $\bar{Q}^0_n(\epsilon_1)$ with

$$\bar{Q}^0_n(\epsilon_{11}, \epsilon_{12}) = \bar{Q}^0_n + \epsilon_{11} C_{11} + \epsilon_{12} C_{12}(g^0_n),$$

where $C_{11}(A, W) = I(A = 1)$, and $C_{12}(g)(A, W) = I(A = 0)g(1 \mid W)/g(0 \mid W)$, and thus fit $\epsilon_1 = (\epsilon_{11}, \epsilon_{12})$ accordingly. Then the TMLE $\bar{Q}^*_n$ also solves the equation $\sum_{i=1}^n I(A_i = 1)(Y_i - \bar{Q}^*_n(1, W_i))$, so that we obtain the following representation of the TMLE:

$$\Psi(Q^*_n, g^*_n) = \frac{1}{\sum_{i=1}^n I(A_i = 1)} \sum_{i=1}^n I(A_i = 1)\left\{Y_i - \bar{Q}^*_n(0, W_i)\right\}. \tag{8.6}$$

That is, the difference between this TMLE and the standard practice in assessing gender inequality is only in the choice of estimator of $\bar{Q}_0$: the TMLE uses a targeted data-adaptive estimator, while standard practice would use a maximum likelihood estimator according to a parametric model.

Regarding implementation of this TMLE, we make the following remark. Technically, in each TMLE update step for $\bar{Q}_0$ one treats the most recent updated estimate of $\bar{Q}_0$ as offset, and one computes the appropriate maximum likelihood estimate of

ϵ_1 according to the parametric working model $\bar{Q}_n^k(\epsilon_1) = \bar{Q}_n^{k-1} + \epsilon_1 C_1(g_n)$. So if it takes K iterations until convergence, then the final fit can be represented as

$$\bar{Q}_n^* = \bar{Q}_n^0 + \sum_{k=1}^{K} \epsilon_{1n}^k C_1(g_n^{k-1}),$$

where g_n^k represents the estimated g after the kth iteration of estimation (and convergence would imply that $\epsilon_{1n}^K \approx 0$). Similarly, this applies to the final fit g_n^*. From a programming standpoint, this means that all one needs to save from the estimation procedure is the sequence, $(\epsilon_{1n}^k, \epsilon_{2n}^k), k = 1, \ldots, K$, as well as the initial fits, $(\bar{Q}_n^0, g_n^0)$. Finally, statistical inference can be based on the sample variance of the estimated efficient influence curve $D^*(Q_0, g_0)$, or the bootstrap.

One may also wish to employ a C-TMLE (Chaps. 19–21 and 23) approach, which would choose among a sequence of candidate estimates of g_0 in a targeted fashion. Remarkably, the efficient influence curve also satisfies collaborative double robustness results so that the C-TMLE can indeed be utilized to build a targeted regression estimator of g_0.

8.3 Simulation

To demonstrate the double robustness of the TMLE, we perform a simple simulation. Specifically, we have a binary $A \in \{0, 1\}$ with $P_0(A = 1) = 0.5$, a simple binary W where $P_0(W = 1|A = 1) = 0.27$ and $P(W = 1|A = 0) = 0.12$, and Y is normally distributed with a conditional mean given by $50000 + 4000W - 1000A + 3000A \times W$, and constant variance 100. The NDE is given by $\Psi_{NDE}(Q_0, g_0) = -192$. We simulate from this same data-generating distribution for sample sizes of $10^i, i = 2, \ldots, 6$, and for each data set we use a correctly specified model for the conditional probability distribution g_0 of A, given W, but a misspecified parametric regression model $E_0(Y \mid A, W) = \beta_0 A + \beta_1 A$ for $\bar{Q}_0$. Thus, the MLE of the estimand will be substantially biased, but the TMLE should eliminate that bias because of the correct model used for g_0. The results (Fig. 8.1) confirm the double robustness property of the TMLE, and thus its value for rectifying bias that can result from nontargeted initial estimates of $\bar{Q}_0$.

8.4 Data Analysis

We use data on 9-month faculty salaries at two schools at the University of California, Irvine, for the academic year 2007–08. There were 579 male and 269 female faculty members. The W variables were as follows: Ph.D. degree, years of UC service (in any position), years since earning highest degree, and department. From a causal direct effect point of view, one should use other measures of performance as

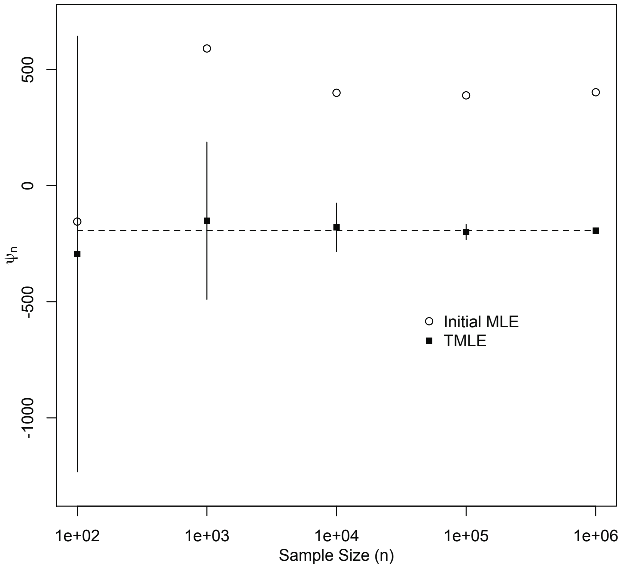

Fig. 8.1 Simulation results for the NDE. *Vertical bars* represent 95% confidence intervals based on standard errors derived from the influence curve (8.5); the *dotted line* is the true value, $\psi_{NDE}(Q_0, g_0)$

well, but there is great controversy in defining academic merit (which itself might be subject to biases, including gender biases), and thus only the most noncontroversial variables associated with salary are typically used.

8.4.1 Unadjusted Mean Difference in Salary

We first carried out a standard least squares regression analysis in which we only have gender (females are $A = 1$) in the regression. Assuming our SCM, and again suppressing the dependence on hiring, H, the coefficient on A is an estimate of the total additive effect due to gender: $\psi^F_{0,Total} \equiv E_0[Y_1 - Y_0]$, where now $Y(a)$ represents the outcome when, possibly contrary to fact, $A = a$, or, using our system of SCM above, $Y(a) = f_Y(a, f_W(a, U_W), U_Y)$. Under our SCM, and the assumption that U_A is independent of U_Y, this is identifiable as a mapping of P_0 as $\psi_{0,Total} = E_0(Y \mid A = 1) - E_0(Y \mid A = 0)$. The corresponding plug-in estimator is simply the difference in average salary among the women ($A = 1$) vs. the men ($A = 0$). This analysis resulted in an estimate $\psi_{n,Total}$ of −\$13,853 and a 95% confidence interval given by (−\$18,510, −\$9,197). Thus, the salary among women is on average \$13,000 less than the average salary among men, and this difference is statistically significant. These results are also displayed in Table 8.1.

8.4.2 Adjusted Mean Difference in Salary: $\Psi_{NDE}(P_0)$

As is evident from representation (8.6) of the TMLE, the only distinction between the TMLE and the standard approach is the choice of estimator of $\bar{Q}_0(1, W)$, i.e., the conditional mean of the salaries among men as a function of the intermediate variables W. This regression function is then used to predict the salary for the females, and one takes the average of the prediction errors among all the females in the sample. The standard approach uses a simple multivariate linear regression to fit $\bar{Q}_0(1, W)$ among the males ($A = 1$). This main term linear regression includes as main terms the indicator of having earned a Ph.D., years of service, years since highest degree, and nine dummy variables representing the ten departments. To obtain robust standard errors, we used a nonparametric bootstrap. This analysis resulted in an estimate of the estimand given by −\$2,235, with a 95% confidence interval (−\$5,228, \$756).

We then used a TMLE that acknowledges that the model for both $\bar{Q}_0$ and g_0 are unknown (and thus uses a data-adaptive estimator) but targets these adaptive estimators to optimize the estimate of $\psi_{NDE,0}$. However, we must start with initial estimates $\bar{Q}_n^0$ of $\bar{Q}_0$ and g_n^0 of g_0. For $\bar{Q}_n^0$, we used a super learner with three linear regression candidate estimators (one with all covariates entered linearly, one with only the covariates besides the dummy variables indicating the department, and one with only gender), polymars, and DSA. The latter two algorithms are machine learning procedures, and for both we enforced the inclusion of gender in the model fit. As initial estimate g_n^0 of the conditional distribution of gender given W, we used simple multivariate linear logistic regression with all covariates entered as main terms. In order to estimate the variance of the TMLE, we used both the empirical variance of the estimated influence curve (8.5), plugging in the final fits, $(\bar{Q}_n^*, g_n^*, \Psi(Q_n^*, g_n^*))$, as well as a more conservative cross-validated empirical variance of the estimated influence curve using a 10-fold cross-validation, where the training sample was used to estimate the influence curve, while the validation sample was used to estimate the variance of this influence curve (van der Laan and Gruber 2010).

In Table 8.1, we display estimates of the unadjusted effect, the naive direct effect, and the TMLE direct effect. Neither estimate of the direct effect of gender is statistically significantly different from the null value. It is of interest to note that the TMLE had a much smaller estimate of the variability: the estimate of the variance of the TMLE was 30% smaller than the estimate of the variance of the naive estimate. (We note that the more conservative CV-based variance of the TMLE was approximately 1400, still reflecting a significant gain in variance.) Thus, although the statistical model for P_0 for the TMLE was much larger than the main term linear regression model of the naive approach, and the TMLE carried out a subsequent tar-

Table 8.1 Analysis results for University of California, Irvine salary gender equity study estimates for the average salary difference between genders, as well as estimates of ψ_{NDE} using both the naive estimator and the TMLE

Parameter	Approach	Estimate	SE	95% CI
ψ_{Unadj}		−\$13,853	2,372	(−\$18,510, −\$9,197)
ψ_{NDE}	Naive	−\$2,235	1,526	(−\$5,228, \$756)
ψ_{NDE}	TMLE	−\$2,212	1,336	(−\$4,830, \$406)

geted bias reduction as well, the variability of the TMLE was still lower. Regarding data analysis, the bottom line is what appeared to be very strong evidence of salary inequity based on a simple difference in average salaries; the naive and TMLE direct effects of gender, controlling for some intermediate variables, show that these data do not support the presence of salary inequality, but it can be stated that the estimated salary gap is around \$2,200, and, with a 95% confidence level, the salary gap is at most \$4,800.

8.5 Discussion

The road map for assessing a causal direct effect (1) states explicitly the causal model and causal direct effect ψ_0^F of interest, (2) states explicitly the identifiability assumptions required to identify the causal direct effect from the observed data distribution, P_0, thereby defining a statistical target parameter mapping, $\Psi(P_0)$, (3) defines the statistical model $\mathcal{M}$ for P_0 based on what is truly known about P_0, (4) estimates the required components of P_0 respecting the statistical model $\mathcal{M}$ using loss-based machine learning, (5) targets the estimation of these components of P_0 for optimizing estimation of $\Psi(P_0)$, where this definition of optimal is based on efficiency theory for estimators. This salary equity example serves as a useful exercise in demonstrating how current ad hoc approaches are more a function of traditional practice than a rigorous methodology used to derive as much information about a scientific question as possible from the data at hand. As discussed, when the covariates W are confounders of a treatment variable A of interest, the *average treatment effect among the treated* (Heckman et al. 1997) is identified by the identical statistical estimand that is the focus of this chapter. Thus, this discussion also applies to situations where the estimand is the same, but the interpretation is different based on a different causal model. We note that much has been written about the average treatment effect among the treated, being a common parameter of interest in fields such as economics and political science. The approach emphasized here has significant relevance to current practice in several disciplines.

8.6 Notes and Further Reading

A popular class of methods for estimation of the treatment effect among the treated is semiparametric matching methods, including propensity score matching algorithms (Rosenbaum and Rubin 1983), and the more recently developed genetic matching algorithms (Sekhon 2006). However, these methods are somewhat inefficient, i.e., the stratification of the sample in clusters of matched treated and nontreated subjects is only informed by the data on (W, A), ignoring the outcome. Attempts to rectify this (Hansen 2008) have failed to provide useful solutions. One can think of such an estimator as an NPMLE that stratifies on W according to an unsupervised method. Two disadvantages of such a method are the lack of smoothing in estimation of $\bar{Q}_0$, and the lack of targeting of the stratification strategy with respect to the target parameter. For example, a treated person might be considered similar to a nontreated person based on similarity in variables that are nonpredictive of the outcome. (C-)TMLE resolves both of these issues.

There is considerable literature on pay equity studies in the academy. In particular, the American Association of University Professors (AAUP) has published a guidebook on implementation of pay equity studies (Haignere 2002). The naive approach presented for ψ_{NDE} is in fact the recommended AAUP approach. Interpretation of these studies is complicated by the absence of adequate measures of the quality and quantity of an individual's performance in terms of research, teaching, and public and professional service despite the fact that it is these very attributes that are presumably the predominant factors in the salary reward system in academia. Without such factors available, gender comparisons are usually adjusted solely by various demographic factors, largely reflecting the "academic age" of individuals. The two principal variables of this type that are associated with current salaries are (1) the number of years since receipt of the highest degree (usually the Ph.D.) and (2) the number of years since appointment to the current institution. The latter variable is important in capturing the influence of market forces in determining salary, almost always at play at the time an individual is hired.

In this chapter, we postulated the existence of an underlying counterfactual salary for an individual whose gender was different. However, this chapter was not focused on questions of the existence of such counterfactuals, and other ontological issues, but on defining sensible parameters of the data-generating distribution that aim to address gender equity. For debates on defining the "causal" effect of a variable such as gender, see, for instance, Holland (1988).