Chapter 11
Robust Analysis of RCTs Using Generalized Linear Models

Michael Rosenblum

It is typical in RCTs for extensive information to be collected on subjects prior to randomization. For example, age, ethnicity, socioeconomic status, history of disease, and family history of disease may be recorded. Baseline information can be leveraged to obtain more precise estimates of treatment effects than the standard unadjusted estimator. This is often done by carrying out model-based analyses at the end of the trial, where baseline variables predictive of the primary study outcome are included in the model. As shown in Moore and van der Laan (2007), such analyses have potential to improve precision and, if carried out in the appropriate manner, give asymptotically unbiased, locally efficient estimates of the marginal treatment effect even when the model used is arbitrarily misspecified.

In this chapter, we extend the results of Moore and van der Laan (2007) for linear and logistic regression models to a wider range of generalized linear models. Our main result implies that a large class of generalized linear models, such as linear regression models (for continuous outcomes), logistic regression models (for binary outcomes), Poisson regression models (for count data), and gamma regression models (for positive-valued outcomes), can be used to produce estimators that are asymptotically unbiased even when the model is arbitrarily misspecified. The estimators that we show to have this property are TMLEs that use the fits of such generalized linear models as initial density estimators. The results hold when the canonical link function for the family of generalized linear models is used, which is commonly the default link function used.

In this chapter we describe (1) a class of model-based estimators particularly useful in analyzing RCT results and (2) a relatively simple and useful application of TMLE. We also show a new robustness property of Poisson regression models when used in RCTs; this result is the log-linear analog to the robustness to model misspecification of ANCOVA for linear models when used in RCTs.

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_11,

The TMLEs we present below are examples of parametric MLEs, double robust estimators (Scharfstein et al. 1999; Robins 2000; Robins and Rotnitzky 2001; van der Laan and Robins 2003), and estimators in Tsiatis (2006) and Zhang et al. (2008). The estimators we present are special cases of a class of estimators in the Comments to the Rejoinder to Scharfstein et al. (1999, p. 1141), for RCTs. Their arguments involving parametric generalized linear models with canonical link functions imply that the estimators given in this chapter are asymptotically unbiased under arbitrary model misspecification, and are locally efficient.

11.1 Summary of Main Result

We are interested in estimating the marginal treatment effect of a randomized treatment or intervention. We want to compare the average of the outcomes for the population of interest under the following two scenarios: (1) had everyone in the population received the treatment and (2) had everyone in the population received the control. These treatment-specific marginal means can be contrasted in many ways. For example, we could be interested in the risk difference, the relative risk, the log relative risk, log odds ratio, etc. Such marginal contrasts are often the goal of RCTs and enter into the decision-making process of the FDA in approving new drugs. We focus on estimating the marginal risk difference, but the methods can be modified to robustly estimate any of these other contrasts, as we describe in Sect. 11.3.1.

Robustness property. We show that a class of TMLEs is robust to misspecification of the working model.

We use the term *working model* to refer to a parametric statistical model that is used in computing an estimator, but that we don't assume to be correctly specified; that is, we don't assume it contains the true data-generating distribution. Here we use generalized linear models as working models.

The robustness property we demonstrate for a class of TMLEs is that these estimators are asymptotically unbiased and asymptotically normal, under arbitrary misspecification of the working model used, under mild regularity conditions. That is, even when the true data-generating distribution is not captured by a generalized linear model at all, the class of estimators we present will be consistent and asymptotically normal. If the generalized linear model used as working model is correctly specified, then the resulting estimator will in addition be efficient; that is, it will attain the semiparametric efficiency bound. These robustness properties require that data come from an RCT. Extensions to observational studies are discussed in Rosenblum and van der Laan (2010b).

Data, statistical model, and target parameter. For each subject i, we denote their baseline variables by W_i, treatment assignment by A_i, and outcome by Y_i. We assume that each triple (W_i, A_i, Y_i) is an independent draw from an unknown data-generating distribution P_0 on the random vector (W, A, Y). [We note that this often-made assumption is not guaranteed by randomization (Freedman 2008c). However, this, or a slightly weaker assumption, is often needed in order to prove even that the standard unadjusted estimator is asymptotically unbiased and asymptotically normal.] We also assume that A is binary, with $A = 1$ indicating assignment to the treatment arm and $A = 0$ indicating assignment to the control arm. Additionally, we assume A is independent of baseline variables W, which is guaranteed by randomization. The prerandomization variables W can take various values (e.g., they may be continuous, categorical, etc.). The outcome Y can also take various values. This defines the statistical model $\mathcal{M}$ for P_0.

Consider estimation of the risk difference $E_{P_0}(Y \mid A = 1) - E_{P_0}(Y \mid A = 0)$. In an RCT, this target parameter identifies the additive causal effect as defined by the SCM. Due to the independence between A and W assumed by the model $\mathcal{M}$, this statistical parameter is equivalent to the parameter $\Psi : \mathcal{M} \to \mathbb{R}$ defined by $\Psi(P) = E_P[E_P(Y \mid A = 1, W) - E_P(Y \mid A = 0, W)]$. There are n total subjects.

The choice of generalized linear (working) model that defines the estimators below will in part depend on the possible values taken by the outcome variable Y. For example, if Y is a count variable (nonnegative integer), then a Poisson regression model may (but won't necessarily) be appropriate.

Class of estimators with robustness property. We refer to the following estimator as the unadjusted estimator of the risk difference:

$$\frac{1}{N_A}\sum_{i=1}^{n} Y_i A_i - \frac{1}{N_{A^c}}\sum_{i=1}^{n} Y_i(1 - A_i),$$

where $N_A = \sum_{i=1}^{n} A_i$ and $N_{A^c} = n - N_A$. This is the difference in sample averages between the treatment and control arms. No baseline variables are used.

The class of TMLEs that we will show to have the robustness property defined above is constructed as follows. First, fit a generalized linear model m, resulting in an estimate $\bar{Q}_n(A, W) = \mu(A, W, \beta_n)$ for $E_{P_0}(Y \mid A, W)$. Then compute the following estimator based on this model fit:

$$\psi_n = \frac{1}{n}\sum_{i=1}^{n} \mu(1, W_i, \beta_n) - \frac{1}{n}\sum_{i=1}^{n} \mu(0, W_i, \beta_n). \tag{11.1}$$

This can be thought of as the difference in the average of the predicted outcomes based on a model fit of m, using the baseline variables W, had all subjects been assigned to the treatment arm vs. had all subjects been assigned to the control arm. We show below that such estimators result from applying the TMLE when using the model fit of m as initial density estimator.

When the generalized linear model is one of the models we describe below, estimator (11.1) will be asymptotically unbiased, even when the model is arbitrarily misspecified, under the mild regularity conditions given below. It will be asymptotically unbiased regardless of whether the true data-generating distribution $P_0(Y \mid A, W)$ is an element of model m. It will be asymptotically unbiased even when $P_0(Y \mid A, W)$ is not in an exponential family at all.

The well-known robustness of analysis of covariance (ANCOVA) to model misspecification in RCTs [as described, e.g., by Yang and Tsiatis (2001) and Leon et al. (2003)] is a special case of the results here and in Moore and van der Laan (2007). This follows since ANCOVA is equivalent to estimator (11.1) when the generalized linear model used is from the normal family with identity link function, and only an intercept and main terms are included in the linear part.

Similarly, ANCOVA II has been shown to be robust to model misspecification (Yang and Tsiatis 2001; Leon et al. 2003); this too is a special case of estimator (11.1), using the same generalized linear model as just described for ANCOVA, except also including an interaction term in the linear part. The definition of the ANCOVA II estimator in Yang and Tsiatis (2001) and Leon et al. (2003) involves ordinary least squares regression of the centered outcome on an intercept, the centered treatment indicator, the centered baseline variable, and a corresponding interaction term involving centered variables; the estimated coefficient of the centered treatment indicator is the ANCOVA II estimator. This estimator is identical to estimator (11.1) when the generalized linear model used is from the normal family with identity link function, and only an intercept, main terms A and W, and an interaction term $A \times W$ are included in the linear part.

11.2 The Generalized Linear (Working) Models

We briefly summarize several facts we use below concerning generalized linear models. More information can be found in, for example, McCullagh and Nelder (1989). We then give examples of generalized linear models that can be used as working models in constructing the estimators (11.1) that are robust to model misspecification in RCTs.

Generalized linear models are a special class of parametric models for the conditional distribution of an outcome (or sequence of outcomes) Y conditional on predictor variables. Here, we use study arm assignment A and baseline variables W as the predictor variables. We only consider generalized linear models with canonical link functions below. Generalized linear models with canonical link functions relate the density p of the outcome Y to the predictors through a linear part η and functions b and c that depend on the generalized linear model family as follows:

$$P(Y \mid A, W) = \exp(Y\eta - b(\eta(A, W)) + c(Y, \phi)),$$

where $\eta(A,W) = \sum_{i=1}^{k} \beta_i h_i(A,W)$, for some functions $h_i(A,W)$, and where ϕ is a dispersion parameter. We require that $h_1(A,W) = 1$ (which gives an intercept term) and that $h_2(A,W) = A$ (which gives a main term A) in the linear part. (Actually, it suffices that these two terms are in the linear span of the terms $h_i(A,W)$.) The link function f_{link} relates the conditional mean $E(Y \mid A,W)$ under the model to the linear part η as follows:

$$f_{\text{link}}[E(Y \mid A,W)] = \eta(A,W).$$

A canonical link function f_{link} satisfies $\dot{b}(\eta) = f_{\text{link}}^{-1}(\eta)$, where $\dot{b}$ is the first derivative of b. When the above model is fit, the coefficients β, ϕ are estimated by maximum likelihood estimation. We denote the conditional mean $E(Y \mid A,W)$ corresponding to the model fit by $\mu(A,W,\beta_n)$ (where, for simplicity, we suppress the fit of the dispersion parameter ϕ_n in our notation). For ease of exposition, we sometimes refer to the generalized linear model as the parameterization "$\beta \to \mu(A,W,\beta)$," even though this only represents a model for the conditional mean of Y given A, W (and not the full conditional distribution of Y given A, W implied by the generalized linear model). We will also use the notation μ_β for the function $(A,W) \to \mu(A,W,\beta)$.

The results in this chapter hold for generalized linear models from the following families: normal, binomial, Poisson, gamma, and inverse normal. We provide below a list of example generalized linear models with canonical links that can be used as working models to construct the TMLEs (11.1). Each example corresponds to a particular choice of the functions b, c, and h_i, $1 \le i \le k$, in the definition above.

Examples of Working Models Satisfying Given Requirements

1. **Least squares regression:** For Y continuous, the normal model assuming $E(Y \mid A,W)$ has the form $\mu_1(A,W,\beta) = \beta_0 + \beta_1 A + \beta_2 W + \beta_3 AW + \beta_4 W^2$.
2. **Logistic regression:** For Y binary and $\text{logit}(x) = \log(x/(1-x))$, the following model for $P(Y = 1 \mid A,W)$: $\mu_2(A,W,\beta) = \text{logit}^{-1}(\beta_0 + \beta_1 A + \beta_2 W)$.
3. **Poisson regression:** For Y a "count" (that is, Y a nonnegative integer), the Poisson (log-linear) model with mean of Y given A, W of the form $\mu_3(A,W,\beta) = \exp(\beta_0 + \beta_1 A + \beta_2 W)$.
4. **Gamma regression:** For Y positive, real valued, the gamma model with mean of Y given A, W modeled by
   $$\mu_4(A,W,\beta) = 1/(\beta_0 + \beta_1(1+A) + \beta_2 \exp(W) + \beta_3 \exp(AW)),$$
   where all coefficients β_j are assumed to be positive and bounded away from 0 by some $\delta > 0$.
5. **Inverse normal regression:** For Y positive, real valued, the inverse normal model with mean of Y given A, W modeled by
   $$\mu_5(A,W,\beta) = 1/\sqrt{\beta_0 + \beta_1 \exp(A) + \beta_2 \exp(W)},$$
   where all coefficients β_j are assumed to be positive and bounded away from 0 by some $\delta > 0$.

The additional restrictions in the gamma and inverse normal regression examples above are needed to ensure the corresponding $\mu(A,W \mid \beta)$ is bounded. We make two

assumptions that, along with the assumptions above, guarantee that the maximum likelihood estimator for the generalized linear model is well defined and converges to a value β^* as sample size goes to infinity. The first assumption guarantees the design matrix will have full rank, with probability 1, as long as sample size is greater than the number of terms in the linear part of the model. The second assumption guarantees convergence of the maximum likelihood estimator to a value β^*.

Assumption 1: If a set of constants c_j satisfies $\sum_j c_j h_j(A, W) = 0$ with probability 1, then $c_j = 0$ for all j.

Assumption 2: There exists a maximizer β^* of the expected log-likelihood

$$E_{P_0}\left[Y\eta - b(\eta) + c(Y,\phi)\right] = E_{P_0}\left[Y\sum_j \beta_j h_j(A,W) - b\left(\sum_j \beta_j h_j(A,W)\right) + c(Y,\phi)\right],$$

and the maximum likelihood estimator β_n of the generalized linear model converges in probability to β^*. In addition, we assume each component of β^* has absolute value smaller than some prespecified bound M.

These two assumptions imply there is a unique maximizer β^* of the expected log-likelihood give above, which follows by concavity of the expected log-likelihood for these generalized linear model families when using canonical links.

11.3 TMLE Using Generalized Linear Model in Initial Estimator

We are interested in estimating the marginal effect of assignment to treatment vs. control in an RCT. We make no assumptions on the unknown, true data-generating distribution P_0, except the following two assumptions. (1) Study arm assignment A is independent of baseline variables and takes values 1 with probability $g_0(1)$ and 0 with probability $g_0(0)$, which is enforced by design in an RCT. (2) P_0 has a smooth density with respect to some dominating measure. The likelihood of the data at a candidate density P can be written

$$\begin{aligned}\prod_{i=1}^n P(Y_i, A_i, W_i) &= \prod_{i=1}^n P_Y(Y_i \mid A_i, W_i) P_A(A_i \mid W_i) P_W(W_i) \\ &= \prod_{i=1}^n P_Y(Y_i \mid A_i, W_i) P_W(W_i) g_0(A_i).\end{aligned}$$

The second equality follows by the first assumption above.

11.3.1 Parameter as a Mapping from the Distribution of the Data

The target parameter is the risk difference. In an RCT, where study arm assignment A is independent of baseline variables W, we have

$$\Psi(P_0) = E_{P_0}(Y \mid A = 1) - E_{P_0}(Y \mid A = 0) \tag{11.2}$$
$$= E_{P_0}\left[E_{P_0}(Y \mid A = 1, W) - E_{P_0}(Y \mid A = 0, W)\right]. \tag{11.3}$$

We note that parameter (11.2) is a function of the data-generating distribution P_0 only through the conditional mean $\bar{Q}_0(A, W) = E_{P_0}(Y \mid A, W)$ and the marginal distribution $Q_{W,0}$ of W. We denote this relevant part of the data-generating distribution by $Q_0 = (\bar{Q}_0, Q_{W,0})$, and we also denote the target parameter by $\Psi(Q_0)$. As in previous chapters, $D^*(Q_0, g_0)$ denotes the efficient influence curve of the risk difference $\Psi : \mathcal{M} \to \mathbb{R}$ at P_0, which will also be defined below. Let $D_1^*(Q_0, g_0)$ and $D_0^*(Q_0, g_0)$ denote the efficient influence curves of the statistical parameters $\psi_0^{(1)} = E_0[E_0(Y \mid A = 1, W)]$ and $\psi_0^{(0)} = E_0[E_0(Y \mid A = 0, W)]$, respectively.

The adjusted estimator (11.1) is the substitution estimator of (11.3) at the generalized linear model fit for $E_{P_0}(Y \mid A, W)$, and using the empirical distribution for the marginal distribution of the baseline variables. Denoting these choices for $E_{P_0}(Y \mid A, W)$ and $Q_{W,0}$ by Q_n, we then have the estimator ψ_n defined in (11.1) equals the substitution estimator $\Psi(Q_n)$.

Note that we can also use Q_n to obtain the substitution estimators $\psi_n^{(1)} = \Psi^{(1)}(Q_n)$ and $\psi_n^{(0)} = \Psi^{(0)}(Q_n)$ of the two treatment-specific means. We have $\psi_n = \psi_n^{(1)} - \psi_n^{(0)}$.

Normally the TMLE involves computing a substitution estimator at a density that is an updated version (via iteratively fitting suitably chosen parametric models) of the initial density estimator. As we show below, for the target parameter ψ_0 and model we consider in this chapter, the updating step of the TMLE always leaves the initial density estimator unchanged. This is due to the initial density estimator already being a maximum likelihood estimator for a parametric working model that happens to have a score that equals the efficient influence curve (at the maximum likelihood estimator) for our parameter of interest ψ_0: that is, $P_n D^*(Q_n, g_0) = 0$, and, also $P_n D^*(Q_n, g_n) = 0$ if $g_n(1) = \sum_i A_i/n$ is the empirical proportion. In fact, we also have $P_n D_1^*(Q_n, g_0) = P_n D_0^*(Q_n, g_0) = 0$ so that $\psi_n^{(1)}$ and $\psi_n^{(0)}$ are also TMLEs of $\psi_0^{(1)}$ and $\psi_0^{(0)}$. In other words, Q_n equals the estimate of the relevant part of the data-generating distribution obtained by the TMLE that targets the parameter $(\psi_0^{(0)}, \psi_0^{(1)})$. This property of generalized linear models with canonical link functions was, to the best of our knowledge, first noted in the Comments to the Rejoinder to Scharfstein et al. (1999, Sect. 3.2.3, p. 1141). This property can be used to show the consistency, asymptotic normality, and local efficiency of the simple estimator (11.1).

A TMLE of any smooth functions of treatment-specific means $[E_{P_0}(Y \mid A = 0), E_{P_0}(Y \mid A = 1)]$ is obtained by substitution of Q_n. For example, for a count variable Y, we might be interested in the marginal log rate ratio

$$\log\left[E_{P_0}(Y \mid A = 1)/E_{P_0}(Y \mid A = 0)\right], \tag{11.4}$$

which we could estimate by the substitution estimator

$$\log\left[\frac{1}{n}\sum_{i=1}^{n}\mu(1,W_i,\beta_n)\Big/\left\{\frac{1}{n}\sum_{i=1}^{n}\mu(0,W_i,\beta_n)\right\}\right]. \tag{11.5}$$

Below we focus on estimating the risk difference (11.2), but analogous arguments apply to other smooth functions of the treatment-specific means.

11.3.2 Obtaining Q_n^0, an Initial Estimate of Q_0

We set the initial estimate $\bar{Q}_n(A,W)$ of $E_{P_0}(Y \mid A,W)$ to be the fit $\mu(A,W,\beta_n)$ of the generalized linear model we are using. This is the maximum likelihood estimate based on using the generalized linear model as a parametric working model. We set the initial density estimate $Q_{W,n}$ of the marginal density $Q_{W,0}$ to be the empirical distribution of $W_1,\ldots,W_n$. Summarizing, we have our initial estimate $Q_n^0 = (\bar{Q}_n, Q_{W,n})$, which also implies an initial density estimate of P_0, according to the generalized linear model, given by $P_n^0 = P_{Q_n^0}$.

11.3.3 Loss Function for TMLE Step

One possible loss function to use is minus the log-likelihood $L(P)(O) = -\log P_Y(Y \mid A,W)P_A(A \mid W)P_W(W)$. Here, however, we use a loss function that only depends on the part Q_0 of the data-generating distribution that is relevant to the parameter of interest. We use $L(Q)(O) = -\log P_{\bar{Q}}(Y \mid A,W) - \log Q_W(W)$ as the loss function for Q_0, where $P_{\bar{Q}}(Y \mid A,W)$ is defined as

$$P_{\bar{Q}}(Y \mid A,W) = \exp(Y f_{\text{link}}(\bar{Q}(A,W)) - b(f_{\text{link}}(\bar{Q}(A,W))) + c(Y)),$$

which is the conditional distribution of Y, given (A,W), implied by the conditional mean function $\bar{Q}$ and the generalized linear working model defined in Section 11.2 (where we omit the dispersion parameter ϕ). This is a valid loss function, i.e., for any of the generalized linear working models we allow, the expected value of the loss function is minimized at the true Q_0.

11.3.4 Calculating the Optimal Fluctuation/Submodel

We now determine a parametric model $\{P_n^0(\epsilon) = P_{Q_n^0(\epsilon)} : \epsilon\}$ that (1) equals the initial density at $\epsilon = 0$ and (2) has a score at $\epsilon = 0$ whose linear span contains the efficient influence function at the initial density estimate $P_n^0 = P_{Q_n^0}$. Given the definition

of the loss function $L(Q)$ above, this is equivalent to stating that we determine a submodel $\{Q_n^0(\epsilon) : \epsilon\}$ so that $Q_n^0 = Q_n^0(0)$, and $d/d\epsilon L(Q_n^0(\epsilon))$ at $\epsilon = 0$ equals the efficient influence curve at $P_n^0 = P_{Q_n^0}$.

The efficient influence function for $\Psi(P)$ defined in (11.2) is

$$D^*(P) = H_{g_0}^*(A)(Y - E_P(Y \mid A, W)) + E_P(Y \mid A = 1, W) - E_P(Y \mid A = 0, W) - \Psi(P),$$

where $H_{g_0}^*(A) = (2A - 1)/g_0(A)$. Note that the function $D^*(P)$ only depends on the relevant part $Q = Q(P)$ of the joint density P since g_0 is known. We sometimes write $D(P) = D^*(Q, g_0)$ as $D(Q)$ below. Let $\epsilon = (\epsilon_1, \epsilon_2, \epsilon_3)$. Define the scalar covariates: $H_1^* = 1$, $H_2^*(A) = A$, and $H_3^*(P)(W) = E_P(Y \mid A = 1, W) - E_P(Y \mid A = 1) - [E_P(Y \mid A = 0, W) - E_P(Y \mid A = 0)]$. Let $H_3^*(Q_n^0)$ be the covariate at the initial estimator Q_n^0. The two covariates (H_1^*, H_2^*) can be replaced by $H_1^*(A) = A/g_0(A)$ and $H_2^*(A) = (1 - A)/g_0(A)$, which are the clever covariates that define the TMLE that targets $(\psi_0^{(0)}, \psi_0^{(1)})$ in general, as in Chaps. 4 and 5. This follows since the linear span of $(1, A)$ is identical to the linear span of $(A/g_0(A), (1-A)/g_0(A))$, so that these different choices do not affect the TMLE.

Recall that f_{link} is the canonical link function for the generalized linear model family used in the initial density estimator. For example, if the generalized linear model family is the Poisson family, then we have $f_{\text{link}} = \log$. We define the following function, which will be used to construct the parametric fluctuation:

$$\eta(\epsilon, A, W) = \epsilon_1 H_1^* + \epsilon_2 H_2^*(A) + f_{\text{link}}\left(\bar{Q}_n(A, W)\right).$$

Here η will be the linear part (including offset) of a generalized linear model. The offset guarantees that at $\epsilon = 0$ we have $\eta(\epsilon, A, W)$ equals $f_{\text{link}}(\bar{Q}_n(A, W))$, which is the linear part corresponding to the initial density estimator $P_n^0 = P_{Q_n^0}$. This can also be stated as a submodel $\{\bar{Q}_n(\epsilon) : \epsilon\}$ defined as

$$f_{\text{link}}\left(\bar{Q}_n(\epsilon)\right) = f_{\text{link}}\left(\bar{Q}_n(A, W)\right) + \epsilon_1 H_1^* + \epsilon_2 H_2^*(A).$$

Define the parametric model $\{P_n^0(\epsilon) : \epsilon\}$:

$$\begin{aligned} P_n^0(\epsilon)(Y \mid A, W) &= \exp(Y\eta_n^0(\epsilon, A, W) - b(\eta_n^0(\epsilon, A, W)) + c(Y, \phi)), \\ P_n^0(\epsilon)(A \mid W) &= g_0(A), \\ P_n^0(\epsilon)(W) &= Q_{W,n}(\epsilon)(W) = s_{\epsilon_3} \exp(\epsilon_3 H_3^*(Q_n^0)(W)) Q_{W,n}(W), \end{aligned}$$

where the constant $s_{\epsilon_3} = 1/[\frac{1}{n}\sum_{i=1}^n \exp(\epsilon_3 H_3^*(Q_n^0)(W_i))]$ is chosen such that $P_n^0(\epsilon)(W)$ integrates to 1 for each ϵ. This also implies a corresponding submodel $\{Q_n^0(\epsilon) = (\bar{Q}_n(\epsilon), Q_{W,n}(\epsilon)) : \epsilon\}$ that equals Q_n^0 at $\epsilon = 0$, so that condition (1) above is satisfied. It is straightforward to verify that condition (2) above is satisfied for the parametric model $\{P_n^0(\epsilon) : \epsilon\}$, or, equivalently, $d/d\epsilon L(Q_n^0(\epsilon))$ at $\epsilon = 0$ equals $D^*(Q_n^0)$.

11.3.5 Obtaining Q_n^*, a Targeted Estimate of Q_0

Consider the maximum likelihood estimator $\epsilon_n = \arg\max P_n \log P_n^0(\epsilon)$ for the parametric model $\{P_n^0(\epsilon) = P_{Q_n^0(\epsilon)} : \epsilon\}$. Note that we also have $\epsilon_n = \arg\min_\epsilon P_n L(Q_n^0(\epsilon))$. The TMLE of P_0 is defined by $P_n^* = P_n^0(\epsilon_n)$. Let $Q_n^* = Q_n^0(\epsilon_n)$, so that $P_n^* = P_{Q_n^*}$. We have $\epsilon_n = (\epsilon_{1,n}, \epsilon_{2,n}, \epsilon_{3,n}) = (0, 0, 0)$. The components $\epsilon_{1,n}, \epsilon_{2,n}$ of the maximum likelihood estimator are found by fitting the generalized linear model $P_n^0(\epsilon)(Y \mid A, W)$. Since the initial density estimator was assumed to have an intercept term and main term A, and was itself fit by maximum likelihood estimation, we must have $(\epsilon_{1,n}, \epsilon_{2,n}) = (0, 0)$. Since $Q_{W,n}$ is the empirical distribution and thereby also a nonparametric maximum likelihood estimator, it follows that $\epsilon_3 = 0$. Thus, the targeted estimate $P_n^* = P_{Q_n^*}$ is identical to the initial density estimate P_n^0, and $Q_n^* = Q_n^0$.

11.3.6 Estimation of Marginal Treatment Effect

The TMLE is the substitution estimator of the risk difference (11.2) evaluated at the targeted density $P_n^* = P_{Q_n^*}$. This density, as described above, consists of the maximum likelihood estimate of the generalized linear model, and the empirical distribution of the baseline variables W. The substitution estimator $\Psi(P_n^*) = \Psi(Q_n^*)$ is given by estimator (11.1). Since Q_n^* is the final density estimate for the TMLE that targets both $(\psi_0^{(0)}, \psi_0^{(1)})$, we also have that Q_n^* maps into the TMLEs of the two treatment specific means $(\psi_0^{(0)}, \psi_0^{(1)})$.

11.4 Main Theorem

The following theorem about the performance of estimator (11.1) under possible misspecification of the working model is a special case of the theorem proved in Rosenblum and van der Laan (2010b).

Theorem 1. *Let $\mu(A, W, \beta)$ be the generalized linear regression model for $E_0(Y \mid A, W)$ implied by a generalized linear model from the normal, binomial, Poisson, gamma, or inverse Gaussian family, with canonical link function, in which the linear part contains the treatment variable A as a main term and also contains an intercept (and possibly contains other terms as well). Under the assumptions in the previous sections, estimator (11.1) is an asymptotically consistent and asymptotically linear estimator of the risk difference ψ_0 defined in (11.2), under arbitrary misspecification of the working model $\mu(A, W, \beta)$. Its influence curve is given by $D^*(Q^*, g_0)$, where $Q^* = (\mu_{\beta^*}, Q_{W,0})$, and $\bar{Q}^* = \mu_{\beta^*}$ denotes the limit of $\bar{Q}_n = \mu_{\beta_n}$. It is locally efficient, meaning that if the working model is correctly specified, then its influence curve $D^*(Q^*, g_0) = D^*(Q_0, g_0)$ is the efficient influence curve, so that the asymptotic vari-*

ance of estimator (11.1) achieves the semiparametric efficiency bound.

Since we showed that μ_{β_n} is also the TMLE that targets both treatment-specific means, it follows that the same theorem applies to any parameter defined as a function of $(E_0(Y \mid A = 0), E_0(Y \mid A = 1))$. See Rosenblum and van der Laan (2010b) for the more general version of Theorem 1.

11.5 Special Robustness of Poisson Model with Only Main Terms

Consider the Poisson regression model given as the third example in Sect. 11.2:

$$\mu_3(A, W, \beta) = \exp(\beta_0 + \beta_1 A + \beta_2 W). \tag{11.6}$$

Denote the maximum likelihood estimator for coefficient β_1 at sample size n by $\beta_{1,n}$. This can be found by simply fitting the above Poisson regression model using standard statistical software. It follows from the generalization of Theorem 1 that $\beta_{1,n}$ is an asymptotically consistent and linear estimator for the marginal log rate ratio (11.4), under arbitrary misspecification of the working model (11.6). To see this, note that the generalization of Theorem 1 discussed above implies that

$$\log\left\{\frac{1}{n}\sum_{i=1}^{n}\mu_3(1, W_i, \beta_n)\right\} \Big/ \left\{\frac{1}{n}\sum_{i=1}^{n}\mu_3(0, W_i, \beta_n)\right\}$$

is an asymptotically consistent and asymptotically linear estimator for the marginal log rate ratio (11.4). But in the special case here, where the Poisson model has only main terms, we have that the above display simplifies to coefficient estimate $\beta_{1,n}$. This can be seen from the following chain of equalities:

$$\begin{aligned}
&\log\left\{\frac{1}{n}\sum_{i=1}^{n}\mu_3(1, W_i, \beta_n)\right\} \Big/ \left\{\frac{1}{n}\sum_{i=1}^{n}\mu_3(0, W_i, \beta_n)\right\} \\
&= \log\left\{\frac{1}{n}\sum_{i=1}^{n}\exp(\beta_{0,n} + \beta_{1,n} + \beta_{2,n} W_i)\right\} \Big/ \left\{\frac{1}{n}\sum_{i=1}^{n}\exp(\beta_{0,n} + \beta_{2,n} W_i)\right\} \\
&= \log\exp(\beta_{1,n})\left\{\frac{1}{n}\sum_{i=1}^{n}\exp(\beta_{0,n} + \beta_{2,n} W_i)\right\} \Big/ \left\{\frac{1}{n}\sum_{i=1}^{n}\exp(\beta_{0,n} + \beta_{2,n} W_i)\right\} \\
&= \beta_{1,n},
\end{aligned}$$

where $\beta_n = (\beta_{0,n}, \beta_{1,n}, \beta_{2,n})$ is the maximum likelihood estimator for β at sample size n. Thus, $\beta_{1,n}$ is an asymptotically unbiased estimator for the marginal log rate ratio (11.4). This is the log-linear analog of the ANCOVA estimator (discussed briefly in Sect. 11.1) being robust to arbitrary model misspecification. This result for the above Poisson model was shown by Gail (1986) under stronger model assumptions than used here.

11.6 Standard Errors and Confidence Intervals

The asymptotic variance σ^2 of the estimator ψ_n defined in (11.1) can be estimated based on its influence curve evaluated at the limit Q^* of Q_n^*. Let β^* be the probability limit of β_n, where β_n is the maximum likelihood estimator for the generalized linear model being used. Then $\bar{Q}^* = \mu_{\beta^*}$. The variance of $\sqrt{n}(\psi_n - \psi_0)$ converges to $\sigma^2 = E_{P_0}\{D(Q^*, g_0)(O)\}^2$. We can estimate σ^2 by replacing E_{P_0} with the empirical mean E_{P_n} and substituting $D(Q_n^*, g_0)$ for $D(Q^*, g_0)$ to get

$$\begin{aligned}\sigma_n^2 &= E_{P_n}\left(D(Q_n^*, g_0)(O)\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\left\{H_{g_0}(A_i)(Y_i - \mu(A_i, W_i, \beta_n)) + \mu(1, W_i, \beta_n) - \mu(0, W_i, \beta_n) - \psi_n\right\}^2.\end{aligned}$$

The standard error of ψ_n can then be approximated by $\sigma_n/\sqrt{n}$ and 95% confidence intervals can be constructed as $(\psi_n - 1.96\sigma_n/\sqrt{n}, \psi_n + 1.96\sigma_n/\sqrt{n})$, which has coverage probability that converges to 95% as sample size tends to infinity.

For parameters other than the risk difference, it is just as easy to compute the asymptotic variance σ^2 of the corresponding TMLE defined above. This follows since, in general, the efficient influence curve of a parameter $f(\psi_0^{(0)}, \psi_0^{(1)})$ for some real-valued function f is given by $d/d\psi_0^{(0)} f(\psi_0^{(0)}, \psi_0^{(1)}) D_0^* + d/d\psi_0^{(1)} f(\psi_0^{(0)}, \psi_0^{(1)}) D_1^*$, where D_j^* is the efficient influence curve of $\psi_0^{(j)}$, $j = 0, 1$. The influence curve of $f(\psi_n^{(0)}, \psi_n^{(1)})$ is determined with the delta method accordingly as a linear combination of the influence curves $D_j^*(Q^*, g_0)$ of $\psi_n^{(j)}$, $j = 0, 1$. For example, the asymptotic variance σ^2 of estimator (11.5) of the marginal log rate ratio can be derived from its influence curve (Rosenblum and van der Laan 2010b, Sect. 4) and can be estimated by

$$\begin{aligned}\sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n}\Bigg(&-\frac{1}{\psi_n^{(0)}}\left\{\frac{1 - A_i}{g_0(A_i)}(Y_i - \mu(0, W_i, \beta_n)) + \mu(0, W_i, \beta_n) - \psi_n^{(0)}\right\} \\ &+ \frac{1}{\psi_n^{(1)}}\left\{\frac{A_i}{g_0(A_i)}(Y_i - \mu(1, W_i, \beta_n)) + \mu(1, W_i, \beta_n) - \psi_n^{(1)}\right\}\Bigg)^2,\end{aligned}$$

where $\psi_n^{(0)} = \frac{1}{n}\sum_{i=1}^{n}\mu(0, W_i, \beta_n)$ and $\psi_n^{(1)} = \frac{1}{n}\sum_{i=1}^{n}\mu(1, W_i, \beta_n)$. Inference for $f(\psi_0^{(0)}, \psi_0^{(1)})$ based on $f(\psi_n^{(0)}, \psi_n^{(1)})$ proceeds as above.

11.7 Discussion

We showed an application of the targeted maximum likelihood algorithm for estimating marginal treatment effects in RCTs. These estimators use generalized linear models as working models. The resulting estimators are simple to compute, and

have the robustness property that they are asymptotically unbiased and asymptotically normal even under arbitrary misspecification of the working model used. If the working model is correctly specified, then these estimators are also asymptotically efficient. For the linear normal error regression model and $g_0(0) = g_0(1) = 0.5$, the TMLE is always at least as efficient, asymptotically, as the unadjusted estimator. In general, the TMLEs considered here may have more precision than the unadjusted estimator for misspecified working models, but there is no guarantee of such an improvement, and it is possible that the TMLE given here could be less efficient than the unadjusted estimator. In the next chapter, we will show how TMLE based on generalized linear regression models can be constructed to provide such guaranteed improvement.

11.8 Notes and Further Reading

The material in this chapter is based on Rosenblum and van der Laan (2010b). Proofs of these results are given in that paper. Previous work related to estimators in RCTs (and in general in observational studies with known probabilities of treatment) that are robust to model misspecification include, for example, Robins (1994), Robins et al. (1995), Scharfstein et al. (1999), van der Laan and Robins (2003), Leon et al. (2003), Tan (2006), Tsiatis (2006), Moore and van der Laan (2007), Zhang et al. (2008), Rubin and van der Laan (2008), Freedman (2008a,b), and Rosenblum and van der Laan (2009a).

As noted in the introduction, the estimators (11.1) are special cases of the class of parametric regression-based estimators in the Comments to the Rejoinder to Scharfstein et al. (1999, Sect. 3.2.3, p. 1141). Scharfstein et al. (1999, Sect. 3.2.3, p. 1141) construct simple, parametric regression-based estimators of the risk difference. These estimators are double robust and locally efficient. Some of these estimators involve generalized linear models with canonical link functions, in which certain simple functions of the inverse of the propensity score are included as terms in the linear part of the model. These estimators take a special form for RCTs; in this case, including the additional terms of Scharfstein et al. (1999) is equivalent to including a treatment variable and an intercept. It follows that their estimator is equal to ours in the special case of estimating the risk difference in an RCT. Their arguments imply that this estimator is consistent under arbitrary model misspecification, and locally efficient. Also, the class of estimators we give is not identical but asymptotically equivalent to the class of estimators given in Tsiatis (2006, Sect. 5.4, p. 132).