Chapter 9
Marginal Structural Models

Michael Rosenblum

In many applications, one would like to estimate the effect of a treatment or exposure on various subpopulations. For example, one may be interested in these questions:

- What is the effect of an antidepressant medication on Hamilton Depression Rating Scale (HAM-D) score for those who enter a study with severe depression, and for those who enter with moderate depression?
- What is the effect of a cancer therapy for those who test positive for overexpression of a particular gene and for those who test negative for overexpression of that gene?
- What is the impact of low adherence to antiretroviral therapy on viral load for HIV-positive individuals who have just achieved viral suppression and for those who have maintained continuous viral suppression for 1 year?

In this chapter, we present a method for estimating the effect of a treatment or exposure in various subpopulations in an HIV treatment application. We first present an analysis in which there are only two subpopulations of interest. Then we present an analysis with 12 subpopulations of interest, where we use a marginal structural model as a working model. Marginal structural models, an important class of causal models and target parameters, were introduced by Robins (1998).

9.1 Impact of Missing Doses on Virologic Failure

For HIV-positive individuals taking antiretroviral medication, a danger in missing doses is that the HIV virus may increase replication. A measure of the amount of circulating virus is called "viral load." It is of interest to understand how different levels of missed doses (e.g., missing 20% of doses in a month or 40% of doses in a month) are related to the probability of subsequent increases in viral load. Furthermore, we'd like to understand how the impact of missed doses on viral load may

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_9,

differ depending on patient history of viral suppression. The aspect of patient history of viral suppression we focus on is the number of consecutive months in the past, starting just before the current month, that a subject has had viral load below 50 copies/ml (which we refer to as "duration of continuous suppression"). As an example, we'd like to understand the impact of low adherence to antiretroviral therapy on viral load for HIV-positive individuals who have just achieved viral suppression and for those who have maintained continuous viral suppression for 1 year. We describe a particular data analysis that aimed to answer this question, which is fully described in Rosenblum et al. (2009).

The population we consider is HIV-positive individuals in the Research in Access to Care for the Homeless (REACH) cohort; subjects in the study consist of a systematic, community-based sample of HIV-positive urban poor individuals in San Francisco (Moss et al. 2004). Adherence to antiretroviral therapy was assessed based on unannounced pill counts, as described in Bangsberg et al. (2001).

We consider four levels of percent adherence to therapy in a given month: 0–49%, 50–74%, 75–89%, and 90–100%. The outcome we consider is whether a patient's viral load is less than 50 copies/ml in a given month. We say a patient experiences virologic failure if her viral load is at least 50 copies/ml.

Three hundred and fifty-seven subjects were monitored monthly for medication adherence. Each subject who had a viral load of less than 50 copies/ml over 2 consecutive months (which is an indicator of successful suppression of the HIV virus) was included in the study; a total of 221 subjects met this criterion. For each included subject, we found the earliest occurrence of 2 consecutive months with viral load less than 50 copies/ml; we let "month 0" denote the first of these two consecutive months.

The goal is to produce estimates of the risk of virologic failure at the end of a given month, under each of the four adherence levels, controlling for variables measured prior to that month. We will get such estimates for each of the following 12 groups:

1 Risk of virologic failure at the end of month 2 among subjects who remained continuously suppressed through month 1;
2 Risk of virologic failure at the end of month 3 among subjects who remained continuously suppressed through month 2;

⋮

12 Risk of virologic failure at the end of month 13 among subjects who remained continuously suppressed through month 12.

We point out that all 221 subjects included in the study contribute data to the estimate in group 1 above (since the inclusion criterion described above requires that subjects be suppressed during month 1). Fewer subjects directly contribute data to the estimates in the latter groups. We also used a nonsaturated marginal structural

model in our analysis that "smoothed" estimates across the above 12 groups. In this case, data from each subject indirectly contributed to the estimates for each of the above groups. This is discussed further in Sect. 9.6.

Of special interest is to compare the relative risk of virologic failure between the highest adherence level (90–100%) and the lowest adherence level (0–49%) for each of the above 12 groups. We can then test for effect modification by comparing this relative risk across the 12 groups.

9.2 Data

Longitudinal data were collected on each subject in the REACH cohort. However, for clarity, we present a simplified data structure in which each subject contributes only a single time point of data. The extension to longitudinal data structures is described elsewhere (Rosenblum and van der Laan 2010a, Sect. 4.2). Longitudinal data structures are also discussed in Chaps. 24–26 of this book.

In our simplified data structure, each subject contributes a vector of data consisting of baseline variables (V, W) measured at the beginning of a month, percent adherence to antiretroviral medication during that month (A), and virologic failure at the end of the month (Y). The baseline variable V denotes duration of continuous viral suppression up to the current time point. The baseline variables W include the following potential confounders of the effect of adherence on virologic failure:

> prior adherence, prior duration of HAART, prior exposure to mono/dual nucleoside therapy, recent CD4+ T cell count (lagged 2 months), CD4+ T cell nadir (lagged 2 months), demographics (sex, ethnicity, age), years of education, past and current antiretroviral treatment characteristics, crack cocaine and alcohol use, calendar time, and homelessness (Rosenblum et al. 2009, p. 2).

Percent adherence A has four levels: $\mathcal{A} = \{0, 1, 2, 3\}$, representing adherence in a given month at 0–49%, 50–74%, 75–89%, and 90–100%, respectively. Y is a binary-valued indicator of virologic failure. Duration of past continuous suppression V takes levels $\mathcal{V} = \{0, 1, 2, \ldots, 11\}$. We denote this vector of data for each subject i by (V_i, W_i, A_i, Y_i). We assume that each subject's data vector is an independent draw from an unknown distribution P_0 of a random vector (V, W, A, Y).

9.3 Statistical Model

We assume a nonparametric statistical model for P_0; that is, we put no restrictions on the true data-generating distribution except that it can be represented as a density with respect to a known dominating measure. Since each distribution we consider has a corresponding density, with a slight abuse of notation, we sometimes refer to distributions such as P_0 as densities. The likelihood of the data at a candidate probability distribution P can be written

$$\prod_{i=1}^{n} P(Y_i, A_i, V_i, W_i) = \prod_{i=1}^{n} P_Y(Y_i \mid A_i, V_i, W_i) P_A(A_i \mid V_i, W_i) P_{V,W}(V_i, W_i).$$

9.4 Parameter of Interest

We are interested in the impact of percent adherence to antiretroviral therapy during a given month on virologic failure at the end of that month. We would furthermore like to know how this impact of adherence varies depending on duration of continuous viral suppression prior to that month.

Let Y_a denote the potential outcome that would have been observed had adherence been at level $a \in \mathcal{A}$. We'd like to learn the probability that $Y_a = 1$, within strata of duration of continuous, past suppression V, that is

$$P(Y_a = 1 \mid V = v), a \in \mathcal{A}, v \in \mathcal{V}. \tag{9.1}$$

We also would like to express the above display as a mapping from the distribution of the observed data (since for each subject three of the four potential outcomes $\{Y_a\}_{a \in \mathcal{A}}$ are unobserved). We make the following assumptions, described in Chap. 2, which we use to connect the potential outcomes to the observed data:

- Time-ordering assumption: W, V precede A, which precedes Y;
- Consistency assumption: For all $a \in \mathcal{A}$, $Y = Y_a$ on the event $A = a$;
- Randomization assumption (no unmeasured confounders): $\{Y_a\}_{a \in \mathcal{A}} \perp\!\!\!\perp A \mid W, V$; and
- Positivity assumption: $P(A = a \mid W = w, V = v) > 0$ for all $a \in \mathcal{A}$ and all (w, v) in the support of P_0.

Under these assumptions, we can equate function (9.1) of the potential outcomes we are interested in with a mapping from the distribution of the observed data, as follows:

$$P(Y_a = 1 \mid V = v) = E_{W|V=v} P(Y = 1 \mid A = a, V = v, W), a \in \mathcal{A}, v \in \mathcal{V},$$

where $E_{W|V=v}$ is expectation with respect to the distribution of baseline variables W given $V = v$.

We define our parameter of interest $\Psi(P)$ to be the mapping from the observed data distribution given on the right-hand side of the previous display:

$$\Psi(P)(a, v) = E_{W|V=v} P(Y = 1 \mid A = a, V = v, W). \tag{9.2}$$

If A and V each had only a couple levels, we could estimate $\Psi(P_0)(a, v)$ (where P_0 is the true, unknown data-generating distribution) directly for each value of a and v. As a stepping stone to the more complex case, we give such an estimator below in Sect. 9.5. Then, in Sect. 9.6, we handle the case described in Sect. 9.2 where there are 48 levels of (A, V) (that come from four possible values for A and 12 for V), and

where we define a different parameter of interest using a marginal structural model as a working model.

9.5 Effect Modification: Simplified Case

We consider the case where both A and V are binary valued. The goal is to estimate (9.2) for all four of the possible combinations of a, v. This is a special case of the more general situation we consider in Sect. 9.6. Here we show an estimator for $\Psi(P_0)(0,0)$; the estimators for the parameter at the other values of a and v are similar. We note that it is also possible to construct an estimator of the vector of parameter values $(\Psi(P_0)(0,0), \Psi(P_0)(0,1), \Psi(P_0)(1,0), \Psi(P_0)(1,1))$ using a single iteration of the targeted maximum likelihood algorithm, but for clarity of exposition we do not present the details here, and instead focus on estimating just $\Psi(P_0)(0,0)$.

9.5.1 Obtaining Q_n^0, an Initial Estimate of Q_0

Parameter (9.2) depends on the data-generating distribution P only through the conditional distribution of Y given (A, V, W), and the marginal distribution of (V, W). We let $Q = (P(Y \mid A, V, W), P(V, W))$ denote these relevant parts of the density P. We let Q_0 denote these relevant parts of the true density P_0. There are many ways to construct an initial estimator Q_n^0 of Q_0. For example, one could fit a parametric statistical model. Here, for simplicity, we fit a parametric statistical model for the conditional distribution of virologic failure Y given (A, V, W), and use the empirical distribution for the baseline variables (V, W). We assume that for at least one subject i in our sample, $V_i = 0$.

We fit a logistic regression model for $P_0(Y \mid A, V, W)$ such as

$$P(Y = 1 \mid A, V, W) = \operatorname{expit}(\alpha_0 + \alpha_1 A + \alpha_2 V + \alpha_3 W).$$

Denote the model fit by $\bar{Q}_n(Y = 1 \mid A, V, W)$. There are no constraints on what model could be used, e.g., interaction terms could have been included as well. For the initial estimator of $P_0(V, W)$, we use the empirical distribution, which we denote by $Q_{V,W,n}$. Our initial estimator Q_n^0 is defined as the pair $\left(\bar{Q}_n(Y = 1 \mid A, V, W), Q_{V,W,n}(V, W)\right)$. Below, in constructing the fluctuation, we will use the substitution estimator at the initial density estimate Q_n^0:

$$\begin{aligned}\Psi(Q_n^0)(0,0) &= E_{Q_n^0}\left[\bar{Q}_n(Y = 1 \mid A = 0, V = 0, W) \mid V = 0\right] \\ &= \frac{1}{\sum_{i=1}^n I(V_i = 0)} \sum_{i=1}^n I(V_i = 0)\bar{Q}_n(Y = 1 \mid A = 0, V = 0, W_i), \quad (9.3)\end{aligned}$$

where $I(S)$ is the indicator function taking value 1 when S is true and 0 otherwise. The second equality follows since in Q_n^0 we set the distribution of (V, W) to be the empirical distribution, so that the corresponding expectation conditional on $V = 0$ is obtained by averaging over all data points with $V_i = 0$.

9.5.2 Calculating the Optimal Fluctuation

To compute the optimal fluctuation, we first need the efficient influence curve for the parameter $\Psi(P)(0,0)$ in the nonparametric model. This can be derived using the methods in Appendix A. The efficient influence curve is (up to a normalizing constant)

$$D_{0,0}(Y,A,V,W) = I(A=0,V=0)\left(\frac{Y - P(Y=1 \mid A=0,V=0,W)}{P(A=0 \mid V=0,W)}\right) + I(V=0)[P(Y=1 \mid A=0,V=0,W) - \Psi(P)(0,0)]. \quad (9.4)$$

We now construct a parametric model $\{P(\epsilon) : \epsilon\}$ that (1) contains the initial estimator Q_n^0 at $\epsilon = 0$ and (2) has a score at $\epsilon = 0$ whose linear span contains the efficient influence curve at Q_n^0. To do this, we first define the clever covariate $H_1^*(A,V,W)$ for fluctuation of the outcome-regression, and function $H_2^*(V,W)$ for fluctuation of the distribution of (V,W):

$$H_1^*(A,V,W) = \frac{I(A=0,V=0)}{g_n(A=0 \mid V=0,W)}$$

and

$$H_2^*(V,W) = I(V=0)[\bar{Q}_n(Y=1 \mid A=0,V=0,W) - \Psi(Q_n^0)(0,0)],$$

where $\Psi(Q_n^0)(0,0)$ is defined in (9.3) and $g_n(A \mid V,W)$, for example, is defined based on fitting a logistic regression model.

Let $\epsilon = (\epsilon_1, \epsilon_2)$. Define the parametric model $\{P(\epsilon) : \epsilon\}$:

$$\begin{aligned} P(\epsilon)(Y=1 \mid A,V,W) &= \operatorname{expit}\left(\epsilon_1 H_1^*(A,V,W) + \operatorname{logit}\left(\bar{Q}_n(Y=1 \mid A,V,W)\right)\right), \quad (9.5)\\ P(\epsilon)(A \mid V,W) &= g_n(A \mid V,W),\\ P(\epsilon)(V,W) &= s_{\epsilon_2} \exp(\epsilon_2 H_2^*(V,W)) Q_{V,W,n}(V,W), \end{aligned}$$

where the constant $s_{\epsilon_2} = 1/[\frac{1}{n}\sum_{i=1}^n \exp(\epsilon_2 H_2^*(V_i,W_i))]$ is chosen such that $P(\epsilon)(V,W)$ integrates to 1 for each ϵ. It is straightforward to verify that conditions (1) and (2) above are satisfied for the parametric model $\{P(\epsilon) : \epsilon\}$.

9.5.3 Obtaining Q_n^*, a Targeted Estimate of Q_0

We fit the above parametric model using maximum likelihood estimation to get estimates $\epsilon_n = (\epsilon_{1,n}, \epsilon_{2,n})$ of (ϵ_1, ϵ_2). We give arguments below to show that the maximum likelihood estimate $\epsilon_n = (\epsilon_{1,n}, \epsilon_{2,n})$ can be obtained simply as follows: to obtain $\epsilon_{1,n}$, fit the logistic regression model (9.5), which has a single term (H_1^*) and offset equal to $\text{logit}\left(\bar{Q}_n(Y = 1 \mid A, V, W)\right)$; we show below that $\epsilon_{2,n}$ must equal 0.

First, since the only term involved in the likelihood that depends on ϵ_1 is the term in (9.5), the ϵ_1 component of the maximum likelihood estimator is obtained by fitting logistic regression model (9.5). We now show that $\epsilon_{2,n} = 0$. The derivative of the log-likelihood with respect to ϵ_2 is zero at $\epsilon_2 = 0$; also, the second derivative of the log-likelihood with respect to ϵ_2 is everywhere strictly negative as long as the values $H_2^*(V_i, W_i)$, for i in $\{1, \ldots, n\}$, are not all equal. (If these values were all equal, the model $P(\epsilon)(V, W) = Q_{V,W,n}$ for all ϵ.) Therefore, the maximum likelihood estimator $\epsilon_{1,n}, \epsilon_{2,n}$ must have $\epsilon_{2,n} = 0$. This means $P(\epsilon_n)(V, W)$ equals the initial density estimator $Q_{V,W,n}$, which was chosen to be the empirical distribution of (V, W).

It is not necessary here to iterate the above steps, since a second iteration [involving fitting a parametric model as above, but now with clever covariates defined in terms of the density $P(\epsilon_n)$ instead of the initial density estimate Q_n^0] would lead to no update of the current density estimate $P(\epsilon_n)$. This follows since the covariate $H_1^*(A, V, W)$ only depends on g_n, which is not updated in the above model fitting; the covariate $H_2^*(V, W)$ does not lead to any update of the density as argued in the previous paragraph. Thus, a single iteration of the above step suffices for convergence. Our final estimator for the relevant part Q_0 of the density of the data-generating distribution is

$$Q_n^* = P(\epsilon_n) = \left(P(\epsilon_{1,n})(Y = 1 \mid A, V, W), Q_{V,W,n}\right). \tag{9.6}$$

9.5.4 Estimation of Parameter

Lastly, we compute the substitution estimator $\Psi(Q_n^*)(0, 0)$:

$$\psi_n(0, 0) = \frac{1}{\sum_{i=1}^n I(V_i = 0)} \sum_{i=1}^n I(V_i = 0) Q_n^*(Y = 1 \mid A = 0, V = 0, W_i). \tag{9.7}$$

This estimator was obtained by evaluating parameter (9.2) at the final density estimate Q_n^* defined in (9.6). This involved first taking the estimated conditional distribution $Q_n^*(Y = 1 \mid A = 0, V = 0, W)$ and computing its average given V, again according to Q_n^*. Since in Q_n^* we set the distribution of (V, W) to be the empirical distribution, this is obtained by averaging $Q_n^*(Y = 1 \mid A = 0, V = 0, W_i)$ over all data points with $V_i = 0$, as in (9.7). In summary, the above estimator involved obtaining initial estimators for $P_0(Y \mid A, V, W)$ and $P_0(A \mid V, W)$, then fitting a logistic regression involving clever covariates constructed from these initial estimators, and

finally averaging this logistic regression fit over the empirical distribution of baseline variables as in (9.7).

A class of estimators that is a special case of the above class of estimators was given in a previous paper (Scharfstein et al. 1999, p. 1141). To the best of our knowledge, they were the first to include the inverse of the propensity score as a covariate in a parametric regression-based estimator for the parameter considered in this section. The estimators there are parametric regression estimators that include the inverse propensity score (H_1^*) as a term in the regression model; these estimators were shown to be double robust and locally efficient.

We next consider the case where V can take 12 values, rather than just 2 values as in this subsection. In that case, estimator (9.7) will not perform well, since for some values of V there may be very few data points contributing to the summation. We will use a marginal structural model to smooth across values of V.

9.6 Effect Modification: Marginal Structural Models

We consider the case introduced in Sect. 9.2, where adherence A can take four possible values, and the number of months of continuous viral suppression V can take 12 values. Here, instead of trying to estimate $\Psi(a, v)$ defined in (9.2) for all 48 possible combinations of (a, v), we will define a different parameter Ψ'. This involves a working statistical model (i.e., marginal structural model as working model) for $\Psi(a, v)$, which can be thought of as smoothing over (a, v). This approach of using marginal structural models as working models to define target parameters is presented in detail Neugebauer and van der Laan (2007) for general longitudinal data structures and multiple time point treatments. The following presentation in this section closely follows that in Rosenblum and van der Laan (2010a, Sect. 4.1).

Marginal Structural Models

For a given treatment level a and duration of past suppression v, the TMLE above for the parameter $\psi_0(a, v)$ defined in (9.2) involves the clever covariate:

$$\frac{I(A = a, V = v)}{g_n(a \mid v, W)}.$$

As a consequence, this estimator may become unstable if there are few subjects in the sample with $A = a$ and $V = v$. In particular, the variance of the estimator will depend on the number of subjects in the category defined by $A = a$ and $V = v$. We present two possible approaches for dealing with this, both of which involve smoothing over the different values of a and v.

The first approach is to assume a statistical model for the parameter $\psi_0(a, v)$ such as:

$$\text{logit}\, \psi_0(a, v) = \beta_0(a, v),$$

indexed by a Euclidean parameter β_0 of lower dimension than $\{\psi_0(a,v), a \in \mathcal{A}, v \in \mathcal{V}\}$. Such a model allows one to focus on estimating the parameter β_0, and the TMLE of β_0 will smooth across all the observations. However, this requires making a model assumption, and if this model assumption is incorrect (i.e., if there is model misspecification, which may be difficult to rule out), then β_0 (and thereby ψ_0) is not defined.

The second approach, which we take here, is to define our target parameter as a summary measure of the parameters $\{\psi_0(a,v) : a, v\}$. For example, for a given adherence level a, one could define our target parameter as the minimizer (β_0, β_1) of the expectation (with respect to the true data-generating distribution) of the squared residuals $(\psi_0(a,V) - \beta_0 - \beta_1 V)^2$. In this case $\beta_0 + \beta_1 V$ represents the least squares projection of the true treatment-specific mean at level a as a function of V onto a linear trend.

The choice of working statistical model, such as the linear statistical model $\beta_0 + \beta_1 V$, defines the target parameter of interest, but it does not represent a statistical assumption.

The parameter $\Psi(P)$ is now well defined for any probability distribution P, including the true distribution P_0. One could also define a whole collection of such summary measures as target parameters, thereby allowing the investigation of a whole collection of features of the true response curve $\psi_0(a,v)$ as a function of a and v.

Define the working model m as follows:

$$m(a, v, \Psi') = \text{expit}(\Psi^{(0)'} + \Psi^{(1)'} a_1 + \Psi^{(2)'} a_2 + \Psi^{(3)'} a_3 + \Psi^{(4)'} v),$$

where a_1, a_2, a_3 are indicator variables for the first three (out of four total) adherence levels defined in Sect. 9.2. The parameter we will estimate throughout this section, in terms of the potential outcomes Y_a, is

$$\Psi'_0 = \arg\max_{\Psi'} \sum_{a \in \mathcal{A}} E_{P_0} h(a,V) \log \left[m(a,V,\Psi')^{Y_a} (1 - m(a,V,\Psi'))^{1-Y_a} \right], \tag{9.8}$$

for some bounded, measurable weight function $h(a,V) \geq 0$ that we specify. When the model m is correctly specified, this can be interpreted as the maximizer of a weighted log-likelihood, in terms of the potential outcomes Y_a. When the model m is misspecified, the parameter is still well defined.

We assume there is a unique maximizer Ψ' to the expression on the right-hand side of (9.8). In this case, the parameter Ψ'_0 is the unique solution of

$$\sum_{a \in \mathcal{A}} E_{P_0} h(a,V)(Y_a - m(a,V,\Psi'))(1, a_1, a_2, a_3, V)' = 0. \tag{9.9}$$

Under the assumptions in Sect. 9.4 linking potential outcomes to a mapping of the observed data, we have that Ψ' is also the unique solution to

$$\sum_{a\in\mathcal{A}} E_{P_0} h(a,V)(P_0(Y=1\mid A=a,V,W)-m(a,V,\Psi'))(1,a_1,a_2,a_3,V)'=0. \quad (9.10)$$

This last display involves a mapping from the distribution of the observed data, and no reference to potential outcomes Y_a; we will use it below in constructing the TMLE for Ψ'.

9.6.1 Obtaining Q_n^0, an Initial Estimate of Q_0

Just as in Sect. 9.5, parameter (9.8), which is the solution to (9.10), depends on the data-generating distribution only through the conditional distribution of Y given A, V, W and the marginal distribution of (V, W). We let $Q = (P(Y \mid A, V, W), P(V, W))$ denote those relevant parts of the density P, and let Q_0 denote those relevant parts of the density at the true data-generating distribution P_0. There are many ways to construct an initial estimator Q_n^0 of Q_0. Just as in Sect. 9.5, for simplicity, here we fit a single logistic regression model to obtain an estimator for the first component of Q_0 and use the empirical distribution as estimator for the second component of Q_0. The resulting initial estimator Q_n^0 is denoted by $\left(\bar{Q}_n(Y=1\mid A,V,W), Q_{V,W,n}(V,W)\right)$. We fit a multinomial logistic regression model for $P_0(A \mid V, W)$, which we denote by g_n. Below we will use the substitution estimator at the initial density estimate Q_n^0, denoted by $\Psi'(Q_n^0)$, which satisfies [by property (9.10) above]

$$\sum_{a\in\mathcal{A}}\sum_{i=1}^{n} h(a,V_i)(\bar{Q}_n^0(Y=1\mid A=a,V_i,W_i)-m(a,V_i,\Psi'(Q_n^0)))(1,a_1,a_2,a_3,V_i)'=0. \quad (9.11)$$

We assume there is a unique solution $\Psi'(Q_n^0)$ to the above display.

9.6.2 Calculating the Optimal Fluctuation

To compute the optimal fluctuation, we need the efficient influence curve for the parameter Ψ' in the nonparametric model. The efficient influence curve is (up to a normalizing matrix) given by

$$D^*(P)(Y,A,V,W)=\left[\frac{h(A,V)(Y-P(Y=1\mid A,V,W))}{P(A\mid V,W)}(1,A_1,A_2,A_3,V)'\right.$$
$$\left.+\sum_{a\in\mathcal{A}} h(a,V)\left(P(Y=1\mid A=a,V,W)-m(a,V,\Psi')\right)(1,a_1,a_2,a_3,V)'\right],$$

where A_1, A_2, A_3 are indicator variables of adherence levels $A = 1, A = 2$, and $A = 3$, respectively. Note that the above efficient influence function reduces to its counterpart (9.4) for the simpler case in Sect. 9.5, for the special case where the weight function $h(a, v)$ is the indicator that $a = 0, v = 0$ and the working model m has only an intercept term.

We now construct a parametric model $\{P(\epsilon) : \epsilon\}$ that (1) contains the initial estimator Q_n^0 at $\epsilon = 0$ and (2) has a score at $\epsilon = 0$ whose linear span contains the efficient influence function at Q_n^0. To do this, we first define the clever covariates $H_1^*(A, V, W)$ and $H_2^*(V, W)$:

$$H_1^*(A, V, W) = \frac{h(A, V)}{g_n(A \mid V, W)}(1, A_1, A_2, A_3, V)'$$

and

$$H_2^*(V, W) = \sum_{a \in \mathcal{A}} h(a, V) \left(\bar{Q}_n(Y = 1 \mid A = a, V, W) - m(a, V, \Psi'(Q_n^0)) \right) (1, a_1, a_2, a_3, V)'.$$

Here H_1^* and H_2^* are each column vectors with five components.

Let $\epsilon = (\epsilon_1, \epsilon_2)$, where ϵ_1 and ϵ_2 are each row vectors with five components (so as to have the same length as H_1^* and H_2^*, respectively). Define the parametric model $\{P(\epsilon) : \epsilon\}$:

$$\begin{aligned} P(\epsilon)(Y = 1 \mid A, V, W) &= \text{expit}\left(\epsilon_1 H_1^*(A, V, W) + \text{logit}\left(\bar{Q}_n(Y = 1 \mid A, V, W)\right)\right), \quad (9.12) \\ P(\epsilon)(A \mid V, W) &= g_n(A \mid V, W), \\ P(\epsilon)(V, W) &= s_{\epsilon_2} \exp(\epsilon_2 H_2^*(V, W)) Q_{V,W,n}(V, W), \end{aligned}$$

where the constant $s_{\epsilon_2} = 1/[\frac{1}{n}\sum_{i=1}^n \exp(\epsilon_2 H_2^*(V_i, W_i))]$ is chosen such that $P(\epsilon)(V, W)$ integrates to 1 for each ϵ. It is straightforward to verify that conditions (1) and (2) above are satisfied for the parametric model $\{P(\epsilon) : \epsilon\}$.

9.6.3 Obtaining Q_n^*, a Targeted Estimate of Q_0

We fit the above parametric model using maximum likelihood estimation to get the estimate $\epsilon_n = (\epsilon_{1,n}, \epsilon_{2,n})$ of (ϵ_1, ϵ_2). One can show (using slight extensions of the arguments in Sect. 9.5.3) that the maximum likelihood estimate $\epsilon_n = (\epsilon_{1,n}, \epsilon_{2,n})$ can be obtained by fitting the logistic regression model (9.12), which has five terms (one for each component of H_1^*) and offset equal to $\text{logit}\left(\bar{Q}_n(Y = 1 \mid A, V, W)\right)$, to obtain $\epsilon_{1,n}$; arguments as in Sect. 9.5.3 can be used to show $\epsilon_{2,n}$ must equal 0 and that no iteration is necessary, since convergence occurs in a single step. Our final estimator for the relevant part Q_0 of the density of the observed data is

$$Q_n^* = P(\epsilon_n) = \left(P(\epsilon_{1,n})(Y = 1 \mid A, V, W), Q_{V,W,n}\right). \quad (9.13)$$

9.6.4 Estimation of Parameter

We compute the substitution estimator $\Psi'(Q_n^*)$, which by property (9.10) solves

$$\sum_{a \in \mathcal{A}} \sum_{i=1}^{n} h(a, V_i)(\bar{Q}_n^*(Y = 1 \mid A = a, V_i, W_i) - m(a, V_i, \Psi'(Q_n^*))(1, a_1, a_2, a_3, V_i)' = 0. \tag{9.14}$$

We assume there is a unique solution to the above equation. The solution $\Psi'(Q_n^*)$ to the above equation can be computed using iteratively reweighted least squares, where the set of outcomes is $\bar{Q}_n^*(Y = 1 \mid A = a, V_i, W_i)$ for each $a \in \mathcal{A}$ and each subject i, which are regressed on the working model $m(a, V_i, \Psi')$ using weights $h(a, V_i)/[m(a, V_i, \Psi')(1 - m(a, V_i, \Psi'))]$.

This iteratively reweighted least squares solution can be implemented in the statistical programming language R with the generalized linear statistical model (glm) function. This involves first constructing a new data set where there are four rows for each subject, one for each possible level of adherence $a \in \mathcal{A}$. For subject i and adherence level $a \in \mathcal{A}$, the following entries make up the corresponding row of this new data set:

1. $\bar{Q}_n^*(Y = 1 \mid A = a, V_i, W_i)$ (which is the "outcome" in the new data set);
2. a (the adherence level under consideration; note that this is not the subject's observed adherence level);
3. V_i (the number of continuous months of past viral suppression);
4. $h(a, V_i)$ (the weight).

One regresses the first column (the new "outcome") on the model $m(a, V_i, \Psi')$ using the glm function with family binomial and logistic link function and using weights $h(a, V_i)$ (from the fourth column of the new data set). Even though the new "outcome" is not binary valued but lies in the interval [0, 1], the glm function computes the desired iteratively reweighted least squares solution, as long as the algorithm converges. It is shown in Rosenblum and van der Laan (2010a) that if this algorithm converges to a value Ψ_n', then this is the unique solution to (9.14).

We now summarize the steps in constructing the TMLE for parameter (9.8). First, we obtained the initial estimators of the conditional densities $P_0(Y = 1 \mid A, V, W)$ and $P_0(A \mid V, W)$. Next, we fit a logistic regression model for Y, with terms H_1^* and offset both depending on the initial density estimators and the formula for the efficient influence function for the parameter. Lastly, we used iterated reweighted least squares to solve Eq. (9.14), yielding the final estimate Ψ_n'.

An important special case of the class of TMLEs given above was previously given in the Rejoinder to Comments in Scharfstein et al. (1999), on p. 1142. To the best of our knowledge, their class of parametric regression-based estimators for the parameter defined by (9.10) is the first to include the covariate H_1^*. Their class of parametric regression-based estimators is double robust and locally efficient.

9.7 Constructing Confidence Intervals

We constructed separate 95% confidence intervals for $m(a, v, \Psi')$, for each $a \in \mathcal{A}, v \in \mathcal{V}$, using the nonparametric bootstrap bias-corrected and accelerated (BCa) method (Efron 1987), with 10,000 iterations. The entire procedure, including refitting the initial regressions, was iterated for each bootstrap replicate. Note that these are not simultaneous 95% confidence intervals. It is also possible to use the nonparametric bootstrap to construct simultaneous confidence intervals. In addition, the asymptotic multivariate normal distribution or the bootstrap distribution of the estimator of $(m(a, v, \Psi') : a, v)$ can be used to carry out multiple testing procedures, controlling a user-supplied type I error rate such as the familywise error rate using methods in Dudoit and van der Laan (2008).

9.8 Results

We now apply the method from Sects. 9.6 and 9.7 to the data from the REACH cohort described in Sect. 9.1. The analysis in Sect. 9.6 was implemented using a working model $m(a, V, \Psi')$ with main terms and interaction terms. Unlike the simplified description above, in which each subject contributed a single time point of data, in the actual analysis subjects contributed multiple time points of data. Overall there were 1201 patient-months of data used in the analysis. The parameter of interest involved a generalization of (9.8) to this setting, as described in Rosenblum and van der Laan (2010a).

For the initial density estimator $Q_n^0 = \left(\bar{Q}_n(Y = 1 \mid A, V, W), Q_{V,W,n}(V, W)\right)$, we let $\bar{Q}_n$ be the fit of a logistic regression model, which included the following terms:

- Intercept;
- Indicator variables for the first three levels of adherence (A_1, A_2, A_3);
- Duration of continuous suppression (V);
- Interactions of (A_1, A_2, A_3) and V;
- Main terms for each confounder variable from Sect. 9.2.

We let $Q_{V,W,n}(V, W)$ be the empirical distribution of (V, W). We fit a multinomial logistic regression model for adherence level A given V, W, which included as terms: intercept, duration of continuous suppression (V), V^2, and main terms for each confounder variable from Sect. 9.2. We denote this by g_n. Similarly, we fit a multinomial logistic regression model for adherence level A given just V, which included as terms intercept, duration of continuous suppression (V), and V^2, which we denote by h_n.

We chose the weight function $h(a, V)$ in the definition of parameter (9.8) to be an approximation to $g_0(a \mid V)$. The motivation behind such a choice was to help stabilize the inverse weights in the clever covariate $H_1^*(A, V, W)$ defined in Sect. 9.6.2. The approximation to $g_0(a \mid V)$ that we use in defining $h(a, V)$ is the limit in probability, as $n \to \infty$, of h_n. Having thus defined $h(a, V)$, we still need to be able to

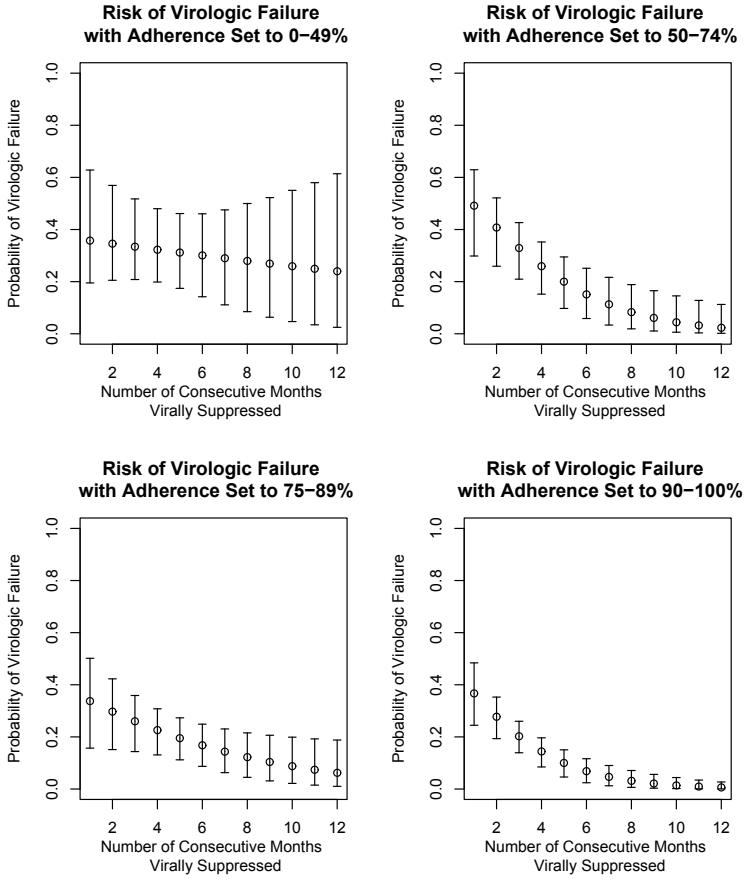

Fig. 9.1 "Estimates and 95% Confidence Intervals for the Risk of Virologic Failure, at Four Ranges of Adherence, Given Duration of Continuous Viral Suppression." This figure and caption are reproduced from Rosenblum et al. (2009)

compute it, or at least approximate it, based on the data we have to work with. In implementing the targeted maximum likelihood algorithm, as in Sect. 9.6, we substitute h_n for h in the definition of $H_1^*(A, V, W)$. We point out that super learning could have been used to construct density estimators in this problem.

The resulting estimate Ψ_n' corresponds to the following working model fit:

$$m(a, v, \Psi_n') = \text{expit}[-0.5 - 0.04A_1 + 0.6A_2 - 0.13A_3 \\ - 0.4V + 0.36A_1V + 0.07A_2V + 2.3A_3V].$$

We show a plot of this function of a, v, along with 95% confidence intervals computed with the nonparametric bootstrap, in Fig. 9.1. The null hypothesis of no effect

modification on the relative risk scale, comparing lowest to highest adherence levels was tested.

$$H_0 : \frac{m(0, v, \Psi')}{m(3, v, \Psi')} = \frac{m(0, v', \Psi')}{m(3, v', \Psi')} \text{ for all } v, v' \in \mathcal{V}.$$

The test involved computing substitution estimates of the relative risks

$$RR(v) = \frac{m(0, v, \Psi')}{m(3, v, \Psi')}$$

for all $v \in \mathcal{V}$ (where the final estimator Ψ'_n was substituted for Ψ'), and then regressing $\log RR(V)$ on the model $\alpha_0 + \alpha_1 V$. The hypothesis of no effect modification on the relative risk scale is rejected if the confidence interval (based on the nonparametric bootstrap) for the coefficient α_1 excludes 0. This hypothesis was rejected at a p-value of 0.001. This hypothesis-testing procedure relies on the consistency and asymptotic normality of the estimator Ψ'_n. To be interpretable in terms of causal relative risks, we additionally need the assumptions given above relating potential outcomes to observed data, and also the assumption that the working model m is a correctly specified marginal structural model.

9.9 Discussion

Under the assumptions given above, and under weak regularity conditions, the TMLEs from Sects. 9.5 and 9.6 are doubly robust, locally efficient. In contrast to this, standard propensity score methods, regression-based methods, and inverse probability weighted methods are generally not doubly robust. Since our goal was to look at effect modification by number of months of continuous suppression (V), we are, by design, comparing effects across different subpopulations. These subpopulations are the 12 groups listed in Sect. 9.1. Observed differences do not, therefore, point to any causal mechanism. However, the results of our analysis are relevant in predicting the impact of missed doses for patients, based on the number of months of continuous viral suppression. Due to the relatively small sample size of the study, and the set of assumptions required by the analysis (which are common to many such analyses), any conclusions from this single study should be made with care.

9.10 Notes and Further Reading

We focused on an application in HIV treatment that can be found in Rosenblum et al. (2009). The content of this chapter is based on work previously published in Rosenblum and van der Laan (2010a). The seminal paper promoting the use of marginal structural working models to define a parameter is Neugebauer and van der Laan (2007). As previously mentioned, a special case of the class of TMLEs presented in

this chapter for the parameter of a marginal structural model was previously given in the Rejoinder to Comments in Scharfstein et al. (1999), on p. 1142. A description of the estimator of Scharfstein et al. (1999) and its relationship to the estimators in this chapter is given in Appendix 2 of Rosenblum and van der Laan (2010a).

We started this chapter with several motivating examples for the estimation of effects in subpopulations. Kirsch et al. (2008) found that there was a difference between the effect of an antidepressant medication on HAM-D score for those entering a study with severe depression vs. less severe depression. An example of the effect of a cancer therapy differing among those with and without overexpression of a particular gene is given in Baselga (2001).