8/30/2004 NOTES

Logistics

Logistics, office hour times, the course website, relevant texts, and a list of topics to be covered are given on the course syllabus. This syllabus can be obtained from Mark van der Laan or Alan Hubbard. Course evaluation will be based on attendance, lecture notes (each student will have to transcribe one or more lectures), a midterm, and a final poster project. The final project will be an application of causal inference methodology from the class to real or simulated data, and students will be allowed to work in small groups.

The three big questions

Before a researcher comes to a statistician for help with data analysis, he or she should be able to answer the following questions.

(1) What is my data, and what is my population of interest?
(2) What is my model? (What assumptions can I make about the data generating distribution?)
(3) What is the parameter of interest? (What would I like to know about the population?)

A statistician's job is to then help the researcher estimate the parameter of interest. Causal inference problems fall into this three question framework, but the answers to (1), (2), and (3) are different from what one would find in more traditional fields of statistics, as we describe below.

1. What is the data in causal inference problems?

We will be interested in longitudinal data, where each subject is followed over time, and we define the following variables.

A(t). This denotes the treatment given to a subject at time t.
Y(t). This denotes some outcome of interest, measured at time t.
X(t). This include $Y(t)$, as well as time-dependent (and baseline) covariates measured on the subject.

Note that $A(t)$, $Y(t)$, $X(t)$ can be possibly multivariate. Define $\overline{A}$ as $(A(t) : t \geq 0)$, the process giving the value of $A(t)$ for each t, and define $\overline{Y}$ and $\overline{X}$ similarly. The observed data in causal inference problems is then n i.i.d. copies of $(\overline{A}, \overline{X}) \sim P_0$. If the reader is unfamiliar with the notation i.i.d., feel free to consult any introductory statistics text. Here P_0 is the (unknown) data generating distribution, which assigns probabilities to members of the population of interest.

As an example, consider an AIDS study. Suppose n patients are selected at random from an AIDS registry, and each is followed up for a period of time. Here P_0 is the probability distribution putting equal mass on each sample of n distinct subjects from the registry (this approximates i.i.d. sampling if the registry size is very large compared to n), and the population of interest consists of all members of the registry. We might have $A(t)$ represent the collection of medications being prescribed to the patient at time t, and the outcome $Y(t)$ might represent the viral load at time t or an indicator of whether the patient is still alive at time t. Here $X(t)$ could include baseline measurements such as the patient's sex, age, and income, as well as time-dependent covariates such as the patient's CD4 count at time t.

2. What is the model in causal inference problems?

Causal inference is the study of counterfactuals, which are the outcomes that would have been observed had the treatment somehow been different. Specifically, let $\overline{X}_{\overline{a}}$ be a counterfactual, and represent the process $\overline{X}$ that would have been observed had the treatment been set at $\overline{A} = \overline{a}$. When $\overline{A} = \overline{a}$, we refer to the observed $\overline{X}_{\overline{a}} = \overline{X}_{\overline{A}}$ as the factual. We will assume that $\overline{X}_{\overline{A}} = \overline{X}$, and this is called the consistency assumption. It states that the observed data is equal to what we would have observed in the counterfactual world had the treatment been set to the observed treatment. In the AIDS example, suppose $\overline{a}$ represents

no treatment being given. Then for a given patient, $\overline{X}_{\overline{a}}$ represents the covariate process that would have occured if, contrary to fact, no treatment had been given to that patient. The idea of counterfactuals raises philosophical issues that have been discussed at least since the time of David Hume, because in the real world each subject is only assigned one treatment process. In causal inference problems, our model will assume the existence of counterfactuals. If Θ represents the support of $\overline{A}$, then we refer to $\overline{X}^{FULL} = (\overline{X}_{\overline{a}} : \overline{a} \in \Theta)$ as the full data. For the estimation procedures discussed in this class to be effective, we must make additional assumptions on the distribution of $\overline{X}^{FULL}$ (the full data model), and the conditional distribution of $\bar{A}$, given $\overline{X}^{FULL}$, such as the sequential randomization or no unmeasured confounding assumptions, and these will be formally defined in subsequent lectures. The choice of models are typically heavily driven by the parameter of interest. Our general phylosophy is that one should make model assumptions on the parameter of interest, but try to minimize assumptions on the nuisance parameters.

Marginal structural models, models for direct and indirect effects, and history adjusted marginal structural models (three topics covered in this class) are just different models on (conditional) distributions of counterfactuals, describing how these conditional distributions change with a change in treatment regime $\bar{a}$.

3. What is the parameter of interest in causal inference?

In causal inference problems, parameters of interest are called causal parameters. Typically they are functions of the data generating distribution of $\overline{X}^{FULL}$, but in history adjusted marginal structural models they will be functions of conditional distributions of the counterfactual outcome $Y_{\bar{a}}$, given an observed past. Usually these parameters will be related to how the outcome process $\overline{Y}_{\overline{a}}$ varies with $\overline{a}$, and how this variation is modified by the covariates. For instance, in a marginal structural models our model might assume that $E[Y_{\overline{a}}(t)] = m(t, \overline{a}, \beta)$ for some known function $m(\cdot)$ and unknown Euclidean parameter β. In this case, β would be the causal parameter, and we would have to find a way to estimate it from the observed data. Note that because the observed data $(\overline{A}, \overline{X})$ is a strict subset of the full data $(\overline{A}, \overline{X}^{FULL})$, causal inference can be treated as a missing data problem. This will be heavily exploited in the following lectures. The general estimating function approach for censored/missing data structures as described in van der Laan, Robins (2002), and presented in this course, corresponds with first finding procedures (i.e., full data estimating functions) that can estimate the causal parameter from the full data, and then map these to procedures (i.e., observed data estimating functions) that estimate the causal parameter from the observed data.

Causal graphs, 9/1/2004 Notes.

In point treatment studies we are interested in studying the causal effect of treatment on outcome of interest. Let A denotes the treatment and Y the outcome variables respectively. Other covariates $X_1, X_2, \ldots, X_m$ are also collected from the patients in the study, requiring adjustment of the causal effect for the covariates. In some situations, the set of causal assumptions on the random variables is known before the study begins and can be represented in the form of a *causal graph*.

Before we proceed to give a formal definition of this concept, we introduce the notation using the causal graph in fig. 1 as an example. There are 5 variables in this hypothetical study - X_1, X_2, X_3, A and Y, which are represented in the graph as vertices. Directed edges represent causal dependence (of which the statistical conditional dependence is a special case), thus Y is causally dependent on A, X_1 and X_3; X_2 is dependent on X_1, etc. We define $PA(X)$ to be the set of all random variables Z in the graph which have a direct arrow going into the node X. In other words $PA(X)$ is the set of parent nodes of vertex X, i.e. all those vertices in the causal graph that have a directed edge that ends in X. For example $PA(Y) = \{A, X_1, X_3\}$ and $PA(X_1) = \emptyset$. Loops are not allowed, which is equivalent to requiring that the causal graph is DAG (directed acyclic graph).

Definition 1. A causal graph G for the set of R.V. $(X_1, X_2, \ldots, X_m, A, Y)$ in a point treatment study is a DAG defining a set of causal assumptions:

$$\begin{aligned} X_i &= f_i(PA(X_i), \epsilon_i), \; i = 1, \ldots, m, \\ A &= f_A(PA(A), \epsilon_A), \\ Y &= f_Y(PA(Y), \epsilon_Y), \end{aligned}$$

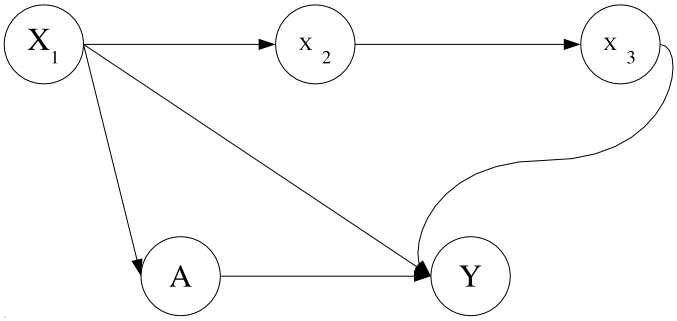

FIGURE 1. Example of a causal graph with three covariates.

for some deterministic functions f_i, $i = 1, \ldots, m$, f_A and f_Y, and ϵ_i, $i = 1, \ldots, m$, ϵ_A and ϵ_Y are random variables called *(exogenous) errors* satisfying the following assumptions $\epsilon_i \perp PA(X_i), \epsilon_A \perp PA(A), \epsilon_Y \perp PA(Y)$.

Note that the causal graphs just assumes that nodes are certain functions of parents-nodes and exogenous error variables, but it assumes nothing about the functional form of these deterministic functions. However, in practice, if one aims to estimate these unknown deterministic functions, then, by the curse of dimensionality, one will parameterize each of these functions with a set of parameters such as coefficients of linear/logistic regression models. The causal graph can involve unmeasured variables: that is, A, Y or $X_1, \ldots, X_m$ can be subject to missingness or (right-)censoring. In this case one views the causal graph as a set of assumptions on the full-data random vector being the collection of nodes in the causal graph: thus, in this case the causal graph does not make any assumptions about the conditional distribution of censoring/missingness variables, given the full data $X^{FULL} \equiv (X_1, \ldots, X_m, A, Y)$. Specifying the whole DAG is usually hard in practice. Given such a DAG it is now possible to identify a *causal* effect of one node on another node in the graph from the distribution/density of the full-data $X^{FULL} = (X_1, \ldots, X_m), A, Y$. In addition, if the observed data is $O = \Phi(C, X^{FULL})$ for some known function Φ of a censoring variable C and X^{FULL} (this is the most general definition of a censored/missing data structure), and one assumes that the conditional distribution of C, given X^{FULL}, satisfies *coarsening at random* (see e.g., van der Laan, Robins, 2002, for literature overview and definitions), then one can often identify from the observed data distribution the full data distribution X^{FULL}, and thereby the wished causal effects. To understand what variables in the graph can be completely missing (i.e., they are not observed on any subject in the sample) while still having coarsening at random and (thereby) identification of the wished causal effect is an interesting problem, and area of research.

In class we we will present an alternative *counterfactual* approach exists that does not require full specification of a causal graph, but, does only require knowing which variables are pre-treatment, but does also need to assume that there are no unmeasured confounders.

Once we have the causal graph we can identify from the distribution of X^{FULL} the counterfactual distribution $P(Y_a = y)$ which is the marginal distr. of Y when treatment is set at level $A = a$. If the interest is only in the effect of A on Y, all covariates connected to A only through an undirected path that includes Y should be ignored. One important issue in the study of causal graphs is what's the minimal subset of covariates that is sufficient to identify the counterfactual distribution of Y_a; related to this is the question of what's the minimal set of confounders that needs to be stratified upon in a point treatment study.

Definition 2. A set of edges connecting two vertices A and Y is called a *back-door path* if $\exists X$ s.t. $X \in PA(A)$ and there is a undirected path between X and Y. A vertex in a back-door path is called a *collider* if the path edges incident with this vertex are incoming (i.e. having their direction towards the vertex). A *confounding* between A and Y is present if $\exists$a back-door path with no colliders connecting A and Y.

For example, there are two back-door paths present in fig. 1 - $A \to X_1 \to Y$ and $A \to X_2 \to X_3 \to Y$.

How to find the likelihood of the data given a causal graph? Using the chain rule for factoring the joint likelihood of discrete R.V. $(Z_1, Z_2, \ldots, Z_d)$:

$$P(Z_1, Z_2, \ldots, Z_d) = P(Z_1) \prod_{i=2}^{d} P(Z_i | Z_1, \ldots, Z_{i-1}), \tag{1}$$

together with the conditional dependence between variables implied by the causal graph $P(Z_i|Z_1, \ldots, Z_{i-1}) = P(Z_i|PA(Z_i))$, we obtain:

$$P(Z_1, Z_2, \ldots, Z_d) = \prod_{i=1}^{d} P(Z_i|PA(Z_i)).$$

When applying this to a single observation in a point-treatment study with discrete R.V. $(X_1, X_2, \ldots, X_m, A, Y)$ we obtain:

$$\begin{aligned} P(X_1 = x_1, \ldots, X_m = x_m, A = a, Y = y) &= \\ &= \prod_{j=1}^{m} P(X_j = x_j | PA(X_j)) P(A = a | PA(A)) P(Y = y | PA(Y)). \end{aligned} \tag{2}$$

Assuming that each of the functional relationships in definition 1 is parameterized as a (e.g., linear) regression in the parent nodes with known (or up till a finite dimensional parameter) conditional distribution of the error-term, given the parent-nodes, we can express the joint probability in 2 as a function of an unknown parameter vector (e.g., the collection of node-specific regression coefficients), which represents now a parametric model and for the density/likelihood of X^{FULL}.

One can estimate the unknown parameters with the maximum likelihood estimator obtained by maximizing the likelihood $\prod_{i=1}^{n} P(X^{FULL} = x_i^{FULL})$ of an observed sample x_i^{FULL}, $i = 1, \ldots, n$. Typically, this can be carried out with standard software (e.g. implementations of generalized linear regression).

Given the causal graph and its estimated functional relations, one can now define the distribution of Y_a as the distribution one obtains by fixing the treatment variable $A = a$ in the system of equations defined by the node-specific equations, and generating all nodes accordingly.

Example 1. The causal graph in this example is specified in fig. 2.

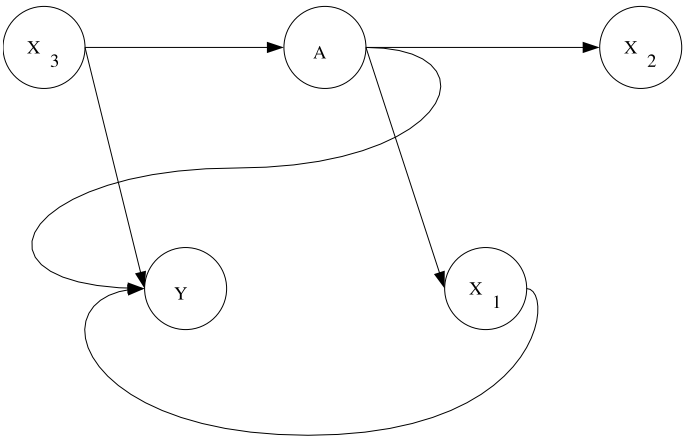

FIGURE 2. Another example of a causal graph with three covariates.

Assuming continuous distr. for all R.V in the causal graph, the density for a single observation can be written as follows:

$$f(X_1, X_2, X_3, A, Y) = f(X_3) f(A|X_3) f(X_2|A) f(Y|A, X_1, X_3) f(X_1|A). \tag{3}$$

Now let's find the counterfactual distribution for $A = a$:

(1) Erase $f(A|PA(A))$ from 3. This is equivalent to performing an incision to the graph in fig. 2 that removes both vertex A and the edges incident to A (fig. 3);
(2) Set $A = a$ in all functions where A belongs to the parents' set;
(3) Define $f_a(X_1, X_2, X_3, Y) = f(X_3) f(X_2|A = a) f(Y|A = a, X_1, X_3) f(X_1|A = a)$;
(4) Integrate out X_1, X_2 and X_3 in $f_a(X_1, X_2, X_3, Y)$ to get the counterfactual density $f_a(Y)$.

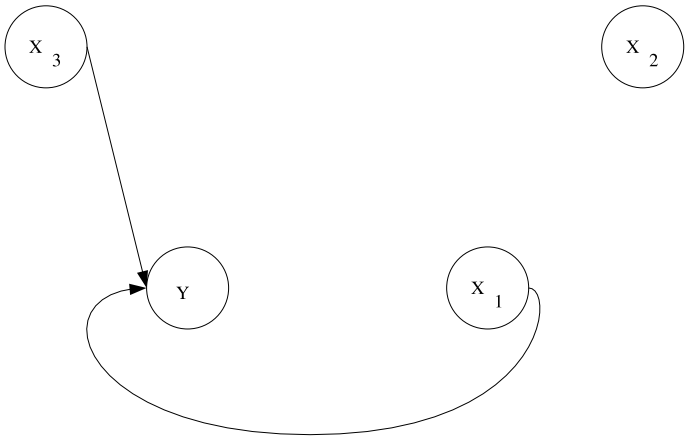

FIGURE 3. Graph from example 1 after incision of A.

Example 2. This is how Pearl defined the counterfactual probability distribution $f_a(Y)$. Hence, if A has two levels - treatment ($a = 1$) and control ($a = 0$), and the causal parameter (or effect) of interest is the treatment difference, one needs to calculate the quantity $Ef_1(Y) - Ef_0(Y)$.

The above method for doing causal inference can be summarized as follows. Using standard software we can do maximum likelihood estimation to fit the functional forms of each node in the causal graph, thus estimating the parameters in each functional relationships in def. 1. Subsequently, given the maximum likelihood estimator of the unknown parameters, we can identify corresponding estimates of the treatment specific disstribution of $(Y_a, X_{1a}, \ldots, X_{ma})$ by Monte-Carlo simulation for each choice of treatment level a.

If the functions in def. 1 are parameterized using flexible regression functions, one can employ data-adaptive model selection methods such as cross-validation or penalized likelihood methods. Such methods provide tools to decide data adaptively how flexible the parametrization should be. Clearly, if the number of parameters is larger than the sample size n, then the maximum likelihood estimator of the unknown parameter vector becomes too variable or ill defined, and thereby our estimate of the treatment specific distribution is too variable as well. That is, the size/dimension/complexity of the model needs to be data dependent. An important research area is the development of methods which data adaptively selects models for the purpose of estimating a particular parameter (such as in our case, the causal effect of treatment on the outcome Y) of interest.

To estimate the variability of the estimates of the causal parameters, (non-) parametric bootstrap is usually employed. This involves resampling repeatedly n observations from the actual sample (nonparametric bootstrap) or from a fit of the true probability distribution of the data (parametric bootstrap).

Is it possible to make a choice between different causal graphs based exclusively on the data from a study? Short answer is, "No". Consider a simple causal graph with 2 R.V. X_1, X_2, where the true causal graph and data generating distribution is: $X_1 \rightarrow X_2$, $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 = X_1 + c$. If we would choose between the only two possible causal graphs $X_1 \rightarrow X_2$ and $X_1 \leftarrow X_2$ the one with the largest corresponding fitted likelihood (with the second model assuming $X_1 = X_2 + c^{'}$, $X_2 \sim N(\mu_2, \sigma_2^2)$, and μ_2, σ_2 estimated from the data generated using the true data distribution) we'll get that approximately 50% of the time we pick the wrong causal graph. In general, if all nodes $(A, Y, X_1, \ldots, X_m)$ are discrete valued, then all possible causal graphs of X^{FULL} give the same corresponding fitted maximum likelihood estimate of the distribution of

X^{FULL}, *if* we use for all causal graphs the nonparametric model (that is, do not assume any parametric form for the functional relations). In this case, the causal-graph specific maximum likelihood is completely flat in the choice of causal graph. Consequently, any variability in the causal-graph specific maximum likelihood values across causal-graphs is NOT due to changes in the causal graphs, but it is due to the fact that different *parametrized* causal graphs result in different statistical models for X^{FULL}, and one might approximate the true distribution of X^{FULL} better than another. For example, if in truth $X_1 \rightarrow X_2$, and the conditional mean of X_1, given X_2 happens to be linear in X_2, then a wrong causal graph $X_2 \rightarrow X_1$ with corresponding linear normal regression assumption $X_1 \sim N(beta X_2, \sigma^2)$, will likely give a higher maximum likelihood value then a correct causal graph $X_1 \rightarrow X_2$ with corresponding assumption X_2 is exponential with $\lambda(X_1) = \beta X_1$.

Lecture of September 8, 2004

Assumptions of the counterfactual framework for causal inference

For each subject, the observed data consist of a file whose rows represent different time points t , $0 \leq t \leq \tau$, and whose columns contain a treatment process $A(t)$, an outcome process $Y(t)$, and a number of covariates $L = \{X_i(t) : i = 1, ..., p\}$, which may or may not be time-dependent. Let $X(t)$ denote $(Y(t), L(t))$. Recall the notation $\bar{A}(t_0) = \{A(t) : t \leq t_0\}$. The counterfactual framework for causal inference is based on the following assumptions:

(1) Temporal ordering assumption (TA)
The data are collected in the order $X(0), A(0), X(1), A(1), ...$, i.e. at a given time point t we first measure all the covariate processes $X(t)$ and then assign the treatment $A(t)$.

(2) Consistency assumption (CA)
Let $\mathcal{A}$ denote the collection of all possible treatment regimes. Then there exists, for each subject, a collection of treatment-specific processes $X^{Full} = \{\bar{X}_{\bar{a}} : \bar{a} \in \mathcal{A}\}$. The observed data for a given subject simply consist of that element of X^{Full} that corresponds to the treatment regime assigned to the subject: $O = (\bar{A}, \bar{X}_{\bar{A}})$. This assumption allows us to view causal inference as a missing-data problem: If we had access to X^{Full} for each subject, inference about causal parameters would be straightforward; the difficulties arise because we only observe one element of X^{Full} for each subject.

Example 1: Consider the case of a binary point treatment at baseline $A \in \{0, 1\}$, an outcome Y, and no other covariates. Then we assume that for each subject there exist Y_0 and Y_1, the outcomes we would observe if the treatment were 0 or 1, respectively. Thus $X^{Full} = \{Y_0, Y_1\}$. Depending on which treatment we observe for a given subject, we only have access to either Y_0 or Y_1.

This allows us to parameterize the data-generating distribution as

$$O = (\bar{A}, \bar{X}_{\bar{A}}) \sim P_{F_0, g_0}$$

where F_0 is the distribution of X^{Full} and $g_0(\cdot|X^{Full})$ is the conditional distribution of $\bar{A}$ given X^{Full}. In order to simulate such data for one subject in `R`, we could thus first generate a realization of X^{Full}, then a realization of the treatment process $\bar{A}$, and finally pick that file $X_{\bar{a}}$ that corresponds to the specific treatment process $\bar{a}$ we generated.

In example 1, we would do the following for each subject: We generate both Y_0 and Y_1, then the treatment assignment a, and then we pick Y_a.

Claim: If the treatment assignment in example 1 is at random, i.e. independent of $X^{Full} = (Y_0, Y_1)$, then the expectation of Y among subjects who actually received a given treatment a in your study equals the

expectation of Y among all subjects if everybody had received treatment a: $E[Y|A=a] = E[Y_a]$. This is very useful since we would like to make statements about the dependence of $E[Y_a]$ on a, but only have access to $E[Y|A=a]$.

Proof: $E[Y|A=a] = E[Y_A|A=a] = E[Y_a|A=a] = E[Y_a]$

where the first equality follows from the consistency assumption, and the third equality follows from the independence assumption $A \perp (Y_0, Y_1)$.

Example 2: Suppose in example 1 we also measured a baseline covariate W, and we knew that treatment assignment is only based on W, i.e. , conditional on W, treatment assignment is independent of X^{Full}: $A \perp (Y_0, Y_1)|W$. Then we have

$$E[Y|A=a,W] = E[Y_A|A=a,W] = E[Y_a|A=a,W] = E[Y_a|W]$$

In this case, we can thus make statements about causal effects of A on Y within strata of W. This corresponds to the usual case of adjusting for a baseline covariate by using multiple regression.

(3) Sequential randomization assumption (SRA)
At any time point t, treatment assignment is only dependent on the observed history of a given subject. It depends neither on future covariate or outcome measures, nor on unobserved counterfactual histories:

$$P(A(t)|\bar{A}(t-1), X^{Full}) = P(A(t)|\bar{A}(t-1), X_{\bar{A}}(t))$$

Note that this is an assumption on $g_0(\cdot|X^{Full})$, the conditional distribution of $\bar{A}$ given X^{Full}:

$$g_0(\bar{A}|X^{Full}) = \prod_{t=0}^{\tau} g_0(A(t)|\bar{A}(t-1), X^{Full}) = \prod_{t=0}^{\tau} g_0(A(t)|\bar{A}(t-1), X_{\bar{A}}(t))$$

We assume that, given the observed history of a subject, treatment assignment is independent of X^{Full}:

$$A(t) \perp X^{Full}|\bar{A}(t-1), X_{\bar{A}}(t)$$

Suppose there exists an unmeasured confounder U that is associated with the outcome Y as well as with treatment assignment. Then this conditional independence assumption will not hold. Thus the observed history must contain any variables that are associated with treatment assignment. The SRA informs our study design in that it requires us to measure any variables that are associated with treatment assignment.

Example 3: Sequentially randomized trials meet this assumption. For simplicity, assume that treatment is binary, $A(t) \in \{0, 1\}$. In order to decide if we assign a subject to treatment at any given time point t, we take into account the subject's history (side effects, drug resistance, viral mutations etc.). We use a known function of this history to obtain a probability p, $0 < p < 1$, with which we assign the subject to treatment. Note that treatment assignment is probabilistic rather than deterministic since the function does not return

values of zero or one.

(4) Experimental treatment assignment assumption (ETA)

Let $\mathcal{A}^*(t|\bar{A}(t-1))$ denote the set of all marginally possible treatment assignments at time point t, i.e. the set of all $a^*(t)$ such $(\bar{A}(t-1), a^*(t))$ is a possible treatment regime according to our data generating mechanism. Then we require that

$$P(A(t) = a(t)|\bar{A}(t-1), \bar{X}(t)) > 0 \qquad \forall\, a^* \in \mathcal{A}^*(t|\bar{A}(t-1))$$

Thus we do not allow treatment assignment rules that, based on a subject's history, give zero probability to certain treatment options that we would consider otherwise reasonable.

Example 4: When we show in example 2 that $E[Y|A = a, W = w] = E[Y_a|W = w]$, we need to assume that $P(A = a, W = w) > 0$. Otherwise, the conditional expectation $E(Y \mid A = a, W)$ on the left is undefined. Since we wont have any data of the form $(Y, A = a, W = w)$, estimation (even when it would be defined) $E[Y|A = a, W = w]$ is not non-parametrically possible. If we estimate $E(Y \mid A, W)$ according to a parametric regression model, then this fit will give us an estimate of $E(Y_{a*}|W) = E(Y \mid A = a*, W)$ for all (a^*, W) for which $P(A = a^*, W)$ has positive probability. We can extrapolate this regression model to the data point $(A = a, W = w)$ and hope that it will give us an approximation of $E[Y_a|W]$.

The G-computation formula

We can factorize the likelihood of the observed data as

$$P(O) = P\left(X(0)\right) P\left(A(0)|X(0)\right) P\left(X(1)|X(0), A(1)\right) ... P\left(A(\tau)|\bar{X}(\tau), \bar{A}(\tau-1)\right) =$$
$$\prod_t P\left(X(t)|\bar{X}(t-1), \bar{A}(t-1)\right) \prod_t P\left(A(t)|\bar{A}(t-1), \bar{X}(t)\right)$$

Under SRA, the second product term gives the likelihood of the observed treatment process conditional on the full data:

$$\prod_t P\left(A(t)|\bar{A}(t-1), \bar{X}(t)\right) = g_0\left(\bar{A}|X^{Full}\right)$$

If we fix $\bar{A} = \bar{a}$ for some $\bar{a}$, we can rewrite the first product term as

$$\prod_t P\left(X(t)|\bar{X}(t-1), \bar{A}(t-1) = \bar{a}(t-1)\right) = \prod_t P\left(X_{\bar{a}}(t)|\bar{X}_{\bar{a}}(t-1), \bar{A}(t-1) = \bar{a}(t-1)\right) =$$
$$\prod_t P\left(X_{\bar{a}}(t)|\bar{X}_{\bar{a}}(t-1)\right) = P(X_{\bar{a}} = \bar{x}) = P\left[(X_{\bar{a}}(0), X_{\bar{a}}(1), ..., X_{\bar{a}}(\tau)) = (x(0), x(1), ..., x(\tau))\right]$$

where the second equality follows from SRA $(X_a(t) \perp \bar{A}(t-1))$. Note that we rely on ETA for the conditional probabilities to be defined. The equality

$$P(X_{\bar{a}} = \bar{x}) = \prod_t P\left(X(t)|\bar{X}(t-1), \bar{A}(t-1) = \bar{a}(t-1)\right)$$

is known as the G-computation formula for longitudinal data.

September 13, 2004 Notes

This lecture deals with the G-computation formula when there is censored data. Modifications to the uncensored data formula involve minor bookkeeping, and a clever new definition of counterfactuals.

What is the data? Let T denote a random time of death, and C denote the time at which a patient was censored, so that the patient is followed until time $\min(T, C)$. Note that if $T < C$ then we know the patient's time of death, while if $C < T$ then we only know the patient lived at least until time C. Let $A_1(t)$ denote the treatment given at time t, $A_2(t) = I(C \leq t)$ be an indicator of whether the patient has been censored by time t, and $A(t) = (A_1(t), A_2(t))$. Let $L(t)$ represent covariate values measured at time t, $Y(t)$ represent an outcome process evaluated at time t, and $X(t) = (L(t), Y(t))$. Let $\overline{X} = \{X(t) : t \leq \min(T, C)\}$ and $\overline{A} = \{A(t) : t \leq \min(T-1, C)\}$. The observed data is then:
$O = ((X(0), A(0)), (X(1), A(1)), ..., (X(\min(T-1, C)), A(\min(T-1, C))), X(\min(T, C)) = (\overline{A}, \overline{X})$.

What are the assumptions? Temporal ordering assumption: At each time point t where observations are made, $X(t)$ is observed before $A(t)$. This is exactly as in the uncensored case.

Consistency assumption: Assume the existence of counterfactuals $\overline{X}_{\overline{a}}$, which represents the process that would have been observed had we set $\overline{A} = \overline{a}$, where $a(t) = (a_1(t), a_2(t))$. The consistency assumption states that $\overline{X} = \overline{X}_{\overline{A}}$, exactly as in the uncensored case.

Sequential Randomization Assumption (SRA): If $\overline{X}^{Full}$ denotes the set of all possible counterfactuals $\overline{X}_{\overline{a}}$, then the SRA assumption is that $g(A(t)|\overline{A}(t-1), \overline{X}^{Full}) = g(A(t)|\overline{A}(t-1), \overline{X}(t))$.
From SRA: $g(\overline{A}|\overline{X}^{Full}) = \prod_{t=0}^{\min(T-1,C)} g(A_1(t)|A_2(t), \overline{A}(t-1), \overline{X}(t)) \prod_{t=0}^{\min(T-1,C)} g(A_2(t)|\overline{A}(t-1), \overline{X}(t))$.
Here the first product is called the treatment mechanism, and the second the censoring mechanism.

Experimental Treatment Assumption (ETA): Exactly as in the uncensored case, this is the assumption that $g(A(t) = a(t)|\overline{A}(t-1), \overline{X}(t)) > 0$ for all $\overline{a}$ of interest.

What would we like to know, and how can we find it? Denote the counterfactual distribution by $P(\overline{X}_{\overline{a}}) = P(\overline{X}_{\overline{a}_1, \overline{a}_2})$. Then we would like to known the distribution of $\overline{X}_{\overline{a}_1, 0}$, or what would have happened under treatment regime $\overline{a_1}$ had we set $C = \infty$, so that there was no censoring. This trick of incorporating the censoring into the treatment variable $\overline{A}$ allows us to write our parameter of interest as part of the counterfactual distribution, and we can then treat our problem as a missing data problem.

From SRA, we can factor $P(O) = \prod_{t=0}^{\min(T,C)} P(X(t)|\overline{X}(t-1), \overline{A}(t-1)) g(\overline{A}|\overline{X}^{Full})$,
and $P(\overline{X}_{\overline{a}_1, \overline{a}_2}) = \prod_{t=1}^{\min(T,C)} P(X(t)|\overline{X}(t-1), \overline{A}_1(t-1) = \overline{a}_1(t-1), \overline{A}_2(t-1) = \overline{a}_2(t-1))$.
Thus, $P(\overline{X}_{\overline{a}_1, 0}) = \prod_{t=1}^{\min(T,C)} P(X(t)|\overline{X}(t-1), \overline{A}_1(t-1) = \overline{a}_1(t-1), C \geq t)$.
To then estimate this quantity, we must fit models for each $X(t)|[\overline{X}(t-1), \overline{A}_1(t-1) = \overline{a}_1(t-1), C \geq t]$, and these can be fitted from the observed data. This technique is referred to as the G-computation method for censored data.

Data Reduction. In order to fit the G-computation formula for $P(\overline{X}_{\overline{a}_1,0})$, it is often helpful in practice to reduce the dimension of the covariate process $L(\cdot)$. There is a vast literature on dimensionality reduction, and many common methods such as PCA, factor analysis, ICA, and principal curves.
Specifically, it is useful to extract $L_1^*(t)$ and $L_2^*(t)$ from the fits of the conditional distribution $g(\cdot|A_2(t), \overline{A}(t-1), \overline{X}(t))$ of $A_1(t)$, and the conditional distribution $g(\cdot|\overline{A}(t-1), \overline{X}(t))$ of $A_2(t)$, which are predictors of $A_1(t)$ and $A_2(t)$. The suggested procedure is then to carry out the previous analysis, while replacing $L(t)$ with $L^*(t) \equiv (L_1^*(t), L_2^*(t))$.

Censored Longitudinal Data and Causality: Notes for 9/15/04
G-computation by simulation: Alan Hubbard

Point Treatment Study To illustrate the implementation of G-computation by simulation, we first rely on a simple point treatment study. In this study, the data consist of three binary variables: the outcome ($Y = (0, 1)$), the treatment ($A = (0, 1)$), and a covariate ($L = (0, 1)$) (see Figure 1).

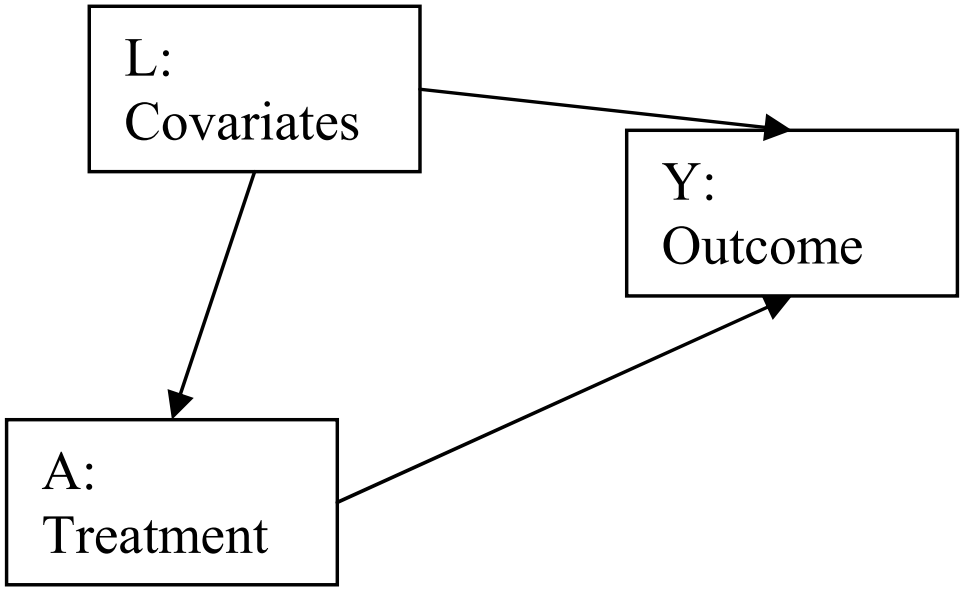

FIGURE 4

The full data can be written as the counterfactual outcomes under the two possible treatments $A = (0, 1)$, and the covariate L: $X^{full} = (Y_0, Y_1, L)$. We are interested in the expectation of the counterfactual outcome under a given treatment ($A = a$): $E[Y_a]$.
We assume Sequential Randomization (SRA): $A \perp Y_0, Y_1 \mid L$
The Likelihood of the observed data can be written as:
$P(Y \mid A, L)P(A \mid L)P(L)$
Under the SRA, the likelihood of the observed data can be rewritten in terms of the full data,
$P(Y \mid A, L)P(A \mid L)P(L) = P(Y \mid A, L)P(A \mid X^{full})P(L)$.
In the point treatment setting, we have the G-computation formula:
$P(X_a = (y, l)) = P(Y = y \mid A = a, L = l)P(L = l)$, where $X_a = (L, Y_a)$.
(Note that $P(X_a = x)$ gives us $E[Y_a]$).
Implementing the G-computation formula in this setting requires two models:

- $P(L)$. We use the empirical distribution to estimate this.
- $P(Y \mid A = a, L) \equiv P(a, L)$. We will model this using logistic regression: $logit(P(a, L)) = \beta_0 + \beta_1 a + \beta_2 L$ (Note that, due to the simplicity of the data, we don't really need a model here. We can just estimate Y empirically among subpopulations defined by a and L).

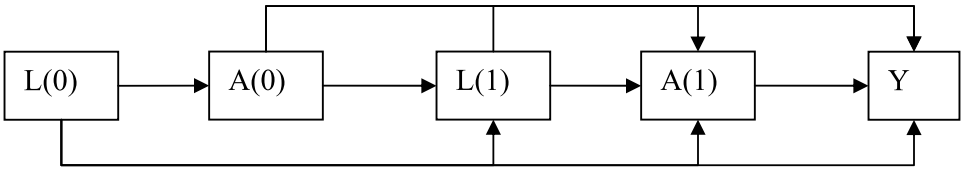

FIGURE 5

To complete one simulation of Y_0, we first draw L from the empirical distribution. We then use our model, $logit(P(a, L)) = \beta_0 + \beta_1 a + \beta_2 L$ to estimate $P(Y \mid A = a, L) \equiv P(a, L)$, setting $a = 0$ and L equal to the value drawn from the empirical distribution. $\hat{P}(Y \mid A = 0, L) \equiv P(0, L)$ gives us an estimate of the probability that $Y_0 = 1$ for this simulation; from this we get a simulated value of Y_0. If we repeat this many times, we get an estimate of the distribution of Y_0 in the population. (We note that,for this simple example, $\hat{P}(0, L)$ has only two possible values, corresponding to $L = (0, 1)$). We repeat the same process, setting $a = 1$, to get an estimate of the distribution of Y_1 in the population. Note that we are estimating the marginal distributions of Y_0 and Y_1, not their joint distribution..

In the above analysis, we were interested in $E[Y_a]$. What if we are interested in the counterfactual outcome, given a baseline covariate, for example, $E[Y_a \mid L = 1]$? We could then just do the same Monte Carlo simulation process, and evaluate the average simulated outcome $E[Y_a]$ among the population with $L = 1$.

G-computation simulation in a setting with two time points We now consider G-computation in a setting where both treatment and covariate are measured at two time points (see Figure 2). Once again, we assume that treatment at each time point, covariate at each time point, and outcome are all binary. Treatment over the course of the study is now represented as $\bar{A} = (A(0), A(1))$. Similarly, the covariate process over the course of the study is $\bar{L} = (L(0), L(1))$. We now have four counterfactuals of interest: $Y_{\bar{a}}$, where $\bar{a}$ varies over the possible combinations of $(0, 1)$. In other words, our counterfactuals of interest are: $Y_{0,0}, Y_{0,1}, Y_{1,0}, Y_{1,1}$, corresponding to whether treatment was taken or not at each time point in the study.

We can write the likelihood of the observed data as:
$P(L(0))P(A(0) \mid L(0))P(L(1) \mid A(0), L(0))P(A(1) \mid (L(0), A(0), L(1))P(Y \mid L(0), A(0), L(1), A(1))$
The G-comp formula here is: $P(X_{\bar{a}} = x) = P(L(0) = l(0))P(L(1) = l(1) \mid L(0) = l(0), A(0) = a(0))P(Y = y \mid L(0) = l(0), A(0) = a(0), L(1) = l(1), A(1) = a(1))$
We now simulate our counterfactuals of interest as above, using the G-comp formula. First, draw $L(0)$ from the empirical distribution. Then set $A(0)$ and generate $L(1)$. Then use the $L(0)$ you drew and the $L(1)$ you generated previously to generate Y, setting A(0) and A(1). Do this many (10,000?) times for each counterfactual treatment history of interest. This gives you a distribution of the outcome under each counterfactual treatment history.

G-computation in the presence of censoring Consider the same data structure as proposed above, with both treatment and covariate measured at two time points. Now assume that some individuals are censored (i.e. we do not get to observe their outcomes). We further assume (for the purposes of this example, only) that all censoring occurs after treatment at the second time point is assigned ($A(1)$), and before outcome Y is measured (see Figure 3). Define $C = 1$ if a subject is censored, $C = 0$ if a subject is uncensored. We now refine our definition of our counterfactuals of interest to reflect the presence of censoring: $Y_{\bar{a},c=0}$ refers to the counterfactual outcome under treatment history $\bar{a}$ and no censoring $c = 0$. Note: if we are going to talk about counterfactual outcomes Y in the presence of censoring and use the same framework for censoring that we used for treatment, we need to define $Y_{\bar{a},c=1}$, or in other words, we need to define the outcome even if the outcome is not measured due to censoring. For more on this, see the notes from the previous lecture on truncated data, where we defined $Y_{\bar{a},c=1}$ as the last observed value for the covariate representing the outcome.

The G-computation formula can now be written: $P(X_{\bar{a},c} = x) = P(L(0) = l(0))P(L(1) = l(1) \mid L(0) = l(0), A(0) = a(0))P(Y = y \mid L(0) = l(0), A(0) = a(0), L(1) = l(1), A(1) = a(1), C = c)$
Note that we assume that censoring only occurs after $A(1)$. To implement this G-comp simulation, we do

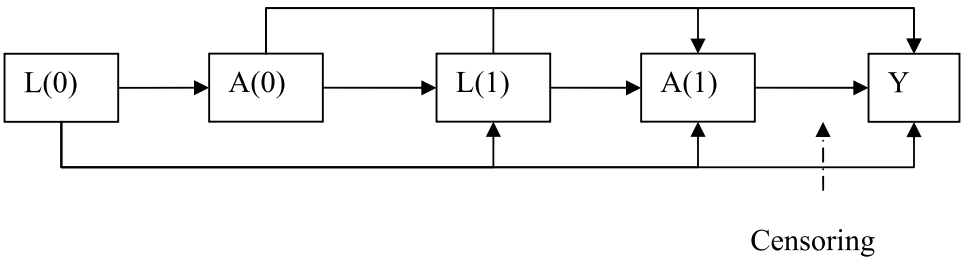

FIGURE 6. Figure 3

exactly the same as in the two time point setting, above, only now we only estimate outcome among people who were uncensored.

Mark van der Laan

Some comments on Alan's lecture

- Start by thinking how you would do everything empirically. In the examples given above, we had few time points, binary variables, and low dimensional L. So really, in all these examples, we probably don't need any models to do G-comp. Just estimate everything empirically. For example, in the two time point case, we have only four possible combinations of $L(0), A(0)$. So, to estimate $L(1)$, all we need to do is take the average $L(1)$ among each of the possible subpopulations: $(L(0) = 0, A(0) = 0), (L(0) = 0, A(0) = 1), (L(0) = 1, A(0) = 0), (L(0) = 1, A(0) = 1)$. However, as we have more covariates, continuous covariates, and covariates and treatment history measured over more time points, we will begin to run out of data to do things empirically. This is where models enter in- to help us deal with the "curse of dimensionality". On this topic, also see discussion of data-reduction methods in the previous lecture.
- To implement G-comp simulation, we need two data matrices: The first is the observed data. We use this to get estimates for our models. The second is the simulated data. We use this to record the results of our simulation and to estimate the distribution of our counterfactual outcomes of interest. A summary of the implementation of G-comp by simulation:
  (1) Write down the likelihood
  (2) Sample from the Likelihood and L(0)
  (3) Get 10,000 simulated counterfactual outcomes (in our example, Y_0, Y_1)
  (4) This gives us a data set with each line of data consisting of the counterfactual treatment (a), the covariates, and the simulated outcome. In this data, just do a regression of outcome on treatment (possibly conditional on covariates, if you are interested in that), or fit some other model, depending on what you are interested in.

Using traditional methods for causal inference in the point treatment setting Say we are interested in $E[Y_a \mid V] = \beta_0 + \beta_1 a + \beta_2 aV + \beta_3 V, V \subset L$, (ie V is a subset of L). We can write: $E[Y_a \mid L] = E[Y \mid A = a, L]$. Also, $E[Y_a \mid V] = E[E[Y_a \mid L] \mid V] = E[E[Y \mid A = a, L] \mid V]$. Call $E[Y \mid A = a, L] \equiv P(a, L)$. We can estimate $P(a, L)$ using a normal regression model. This regression model gives us a predicted outcome for every subject under each treatment of interest that is a function of that subject's value L. We now have a data set in which each subject contributes one line of data for each treatment of interest (e.g. in the case of a binary treatment, each subject contributes two lines of data). Each line of data contains the subject's covariate values (L), a treatment ($a = 0$ or $a = 1$), and the predicted outcome given the subject's covariate values and that treatment ($P(a, L)$). This is just a repeated measures data set with two outcomes measured for each subject, corresponding to the two treatments. We can pool the whole dataset and do a simple regression of Y on a and V.

Introduction to dynamic treatment regimens Take the observed data to be a collection of treatment history and covariates over time, $O = (\bar{A}, \bar{X})$. A dynamic treatment regime is a collection of j-specific decision rules, d_j, where d_j is a function of $\bar{X}(j)$ which assigns a treatment at time $j = 0, ...$ We assume the consistency assumption (CA): For all (or a set containing all static treatment regimes and the dynamic treatment regimes the user is interested in) dynamic treatment regimens $\bar{d} = (d_j : j = 0, ...)$, we assume the existence of $\bar{X}_{\bar{d}}$, $X^{full} = (\bar{X}_{\bar{d}} : \bar{d} \in D)$. Under the CA, the observed data can be represented as $O = (\bar{A}, \bar{X}_{\bar{A}})$.

We further assume the SRA: $A(j) \perp X^{full} \mid \bar{X}(j), \bar{A}(j)$ and a d-specific experimental treatment assignment assumption (ETA), which, informally, states that we need support in our data for this particular dynamic treatment regime d. Formally, this ETA assumption is defined as: for any $\bar{l}(j)$ with $P(\bar{A}(j-1) = \bar{d}_j(\bar{l}(j-1)), \bar{L}(j) = \bar{l}(j)) > 0$, we have $P(\bar{A}(j) = \bar{d}_j(\bar{l}(j)), \bar{L}(j) = \bar{l}(j)) > 0$. We can then write the G-comp formula for dynamic treatment regimes. this is the same as the previous G-comp formula, only now we set A using the first action of $d(x)$:
$P(\bar{X}_{\bar{d}} = \bar{x}) = \prod_{j=0} P(X(j) = x(j) \mid \bar{X}(j-1) = \bar{x}(j-1), \bar{A}(j-1) = d_0(X(0)), d_1(\bar{X}(1)), ..., d_{j-1}(\bar{X}(j-1)))$

More to come on dynamic treatment regimes...

Estimation of G-Computation Formula for Dynamic Treatment Regimes in Point Treatment Studies, Keith Betts, September 20, 2004

Define the data structure as $O = (W, A, Y)$,
$W \longrightarrow A \longrightarrow Y$ (time ordering)
The treatment regime is a function of W (where W are the baseline covariates)

A possible question of interest is if everyone in the population is assigned a rule that given their past what treatment should be, what would be the mean outcome?

Example:
Cohort of children with asthma
Treatment is use of steriod spray; Outcome is lung function(FEV) in two years.
Based on history get assigned treatment, follow, at end observe outcomes.

Necessary Assumptions:

Sequential Randomization Assumption(SRA):
We can identify from the data what the treatment was (i.e. no unmeasured confounders)

Dynamic treatment regime means that the treatment is dependent on the past (covariates)

Consistancy Assumption:
For a given treatment rule $W \mapsto d(W)$ (function that maps history at baseline into treatment). Let $X_d = (W, Y_d)$, the treatment specific counterfactual, be the data we would have seen on the subject if they would have followed rule d.
Let $X^{Full} = (X_d : d) = (W, (Y_d : d \in \mathcal{D}))$
where $\mathcal{D} = \{w \mapsto d_{\delta_1, \delta_2}(w) : \delta_1, \delta_2\}$
Assumption is $O = (A, X_a) = (W, A, Y_A)$

How to simulate?
Simulate baseline covariates.
Simulate dynamic treatment specific outcomes.
1. $Y_{d_{\delta_1,\delta_2}} = Y_0 + \beta_0\delta_1 + \beta_1\delta_2 + \beta_2\delta_1\delta_2$ (Draw Y_0 from N(0,1)
2. $Y_0|W$
1 and 2 gives us $X^{Full} = (W, Y_d : d \in \mathcal{D})$
Then we simulate $A|W \Rightarrow O = (A, W, Y_A)$
To generate counterfactuals:
treat as missing data, apply G-Computation formula to observed data.

$P(Y_d = y, W = w) = P(W = w)P(Y = y|A = d(w), W = w)$
Can check if close to the truth.

Randomization Assumption:
$A|(W, Y_d : d) \sim A|W$
Y's are random variables; The distributions for Y_0, Y_1, and Y_d are different
Y_0 is the random variable observed if no one receives treatment
Y_1 is the random variable observed if everyone given static treatment
Now:
Theorem: Distribution of Y_d:
$P(Y_d = y, W = w) = P(W = w)P(Y = y|A = d(w), W = w)$

Given a rule d:

δ_1	δ_2	$\hat{EY_{\delta_1,\delta_2}}$
1	20	96
1	30	95
1	50	91
1	80	90
2	20	97
2	30	.
⋮	⋮	⋮

For every combination of δ_1, δ_2 do a G-compuation to get $\hat{EY}_{\delta_1,\delta_2}$.
Can see which combination of δ_1, δ_2 is the "best".
Can bootstrap.
Can Assume model.
e.g. $\hat{EY}_{\delta_1,\delta_2} = \beta_0 + \beta_1\delta_1 + \beta_2\delta_2 + \beta_3\delta_1\delta_2$
$lm(\hat{EY}_{\delta_1,\delta_2} = \delta_1 + \delta_2 + \delta_1\delta_2)$

Alternatively could have modeled:
$\hat{EY_{\delta_1,\delta_2}}|W = \beta_0 + \beta_1\delta_1$
This is the analog of Marginal Structural Models, except for dynamic treatment regimes.

Inference for the Causal Effects of Dynamic Treatment Regimes in Point Treatment Studies 9/22/04

Definitions

In a point treatment study, the observed data for each of n subjects is $O = (W, A, Y)$, where W is a vector of baseline covariates, A is the treatment received, and Y is the outcome. It is assumed that W is measured before A is measured before Y.

Let $\mathcal{A}$ be the set of possible treatments in a study and $\mathcal{W}$ the set of possible values for W. Then a **dynamic treatment regime** is a function $d : \mathcal{W} \to \mathcal{A}$. That is, a dynamic treatment regime is a rule or algorithm for assigning treatment based on baseline values. We use $\mathcal{D}$ to denote the set of dynamic treatment regimes of interest for a particular study.

This is in contrast to the **static treatment regimes** we have previously considered, in which a given treatment is assigned to all study participants, irrespective of their baseline values; e.g., assigning everyone to a study's control treatment is a static regime. In previous lectures, when we were trying to estimate $E[Y_1 - Y_0]$, we were estimating the average difference in the outcome from the static regime that assigns everyone to the active treatment with the outcome from the static regime that assigns everyone to the control treatment. Each treatment in $\mathcal{A}$ can thus be associated with the static regime that assigns that treatment to everyone, so $\mathcal{A}$ can also be used to denote the set of static treatment regimes. Note that "officially" a static treatment regime is also a (degenerate) dynamic regime, since it is a constant function from $\mathcal{W}$ to $\mathcal{A}$.

The parameters of interest that we wish to estimate, which reflect the usefulness of the dynamic regimes under consideration, are

$$(E[Y_d|V] : d \in \mathcal{D}, V \subseteq W).$$

We will develop G-computation estimators for these parameters below.

Example Consider an observational study of short-term psychotherapy ($A = 1$) versus antidepressants ($A = 0$) for the treatment of depression, based on archival records from a large outpatient psychiatry clinic. The observed outcome Y is depression score (0-66, higher scores indicate greater depression) after three months of treatment. The two baseline covariates are pretreatment depression score W_1 (taking values between 14-66, since one needs a score of at least 14 to qualify for the study) and W_2, an indicator of sex (female=1, male=0). The researcher believes that psychotherapy is the treatment of choice for mildly or moderately depressed individuals, while antidepressant therapy is preferable for severely depressed individuals. The aim of the study is to determine the optimal cutoff score (with regard to treatment assignment) for differentiating severe from nonsevere depression. Thus, the dynamic regimes of interest are of the form

$$d_\theta(w_1, w_2) = I(w_1 < \theta)$$

where $I()$ is an indicator function. That is, an individual with a pretreatment depression score below the θ cutoff is assigned psychotherapy, while individuals at or above the cutoff receive antidepressants. For the sake of simplicity, we will consider only two nonstatic regimes, d_{25} and d_{35}. Thus, $\mathcal{D} = \{d_{25}, d_{35}\}$. The parameters of interest are hence $E[Y_{d_{25}}]$ and $E[Y_{d_{35}}]$ ($V = \phi$ here). If the researcher was additionally interested in the cutoffs evaluated separately for men and women, then additional parameters of interest would include $E[Y_{d_{25}}|W_2]$ and $E[Y_{d_{35}}|W_2]$.

The data for subjects 1 through 5 are given below. Y_0 denotes the counterfactual outcome for the static treatment regime of assigning antidepressants to everyone, Y_1 denotes the counterfactual outcome for the static treatment regime of assigning psychotherapy to everyone, and $Y_{d_{25}}/Y_{d_{35}}$ denote the counterfactual outcomes for the nonstatic regimes with cutoffs 25/35, respectively.

Subject	W_1	W_2	A	Y	Y_0	Y_1	$Y_{d_{25}}$	$Y_{d_{35}}$
1	20	1	1	4	?	4	4	4
2	30	0	1	11	?	11	?	11
3	40	1	1	29	?	29	?	?
4	40	0	0	21	21	?	21	21
5	30	1	0	13	13	?	13	?

Several notes are in order. First, since W_1 can take 53 different values and W_2 is a binary variable, $W = (W_1, W_2)$ can take 106 different values, and hence there are 2^{106} dynamic treatment regimes that could be examined with this data set. Any realistic study of these data will thus examine only a very small proportion of the dynamic regimes that could be examined. Second, we are performing a post hoc examination of the dynamic regimes of interest. We are not assuming that the psychiatrists who treated the patients in the study actually employed any of these regimes when making treatment decisions with their patients. Third, for some of the subjects, we observe multiple counterfactual outcomes. For example, for subjects 1 and 4, we observe one static counterfactual and both nonstatic counterfactuals. As discussed previously, this cannot happen when only static regimes are studied.

4. Assumptions

In order to develop the G-computation estimators, the following assumptions are needed:

Consistency Assumption (CA) $O = (W, A, Y_A)$

Randomization Assumption (RA) $A \perp (Y_d : d \in \mathcal{D})|W$

That is, assigned treatment is assumed to be independent of the set of counterfactual outcomes, within strata of W. In point treatment studies, this is equivalent to $A \perp (Y_a : a \in \mathcal{A})|W$.

Example, continued Alas, RA does not hold in our example study. A competent psychiatrist will ask patients about previous depressive episodes and the treatments received for them, and this information will both influence treatment choice and predict counterfactual outcome. For example, if a patient had received a successful course of psychotherapy for a previous bout of depression, then he/she is more likely to be assigned psychotherapy again and is more likely to again have a good outcome with it. For RA to hold

here, treatment history (and no doubt a bunch more information) would have to be included as a baseline covariate. For the sake of simplicity, we will ignore this assumption violation.

5. G-computation Estimators

proposition Let $W = (V, U)$. Then

$$E[Y|V = v_0] = E[E[Y|W]|V = v_0]$$

proof

$$\begin{aligned}
E[Y|V = v_0] &= \int y f_{Y|V}(y|v_0)dy \\
&= \int y \frac{f(y, v_0)}{f(v_0)} dy \\
&= \int y \left[\int \frac{f(y, v_0, u)}{f(v_0)} du \right] dy \\
&= \int y \int \frac{f(y, w_0)}{f(w_0)} \frac{f(w_0)}{f(v_0)} du dy \qquad \text{where } w_0 = (v_0, u) \\
&= \int \left[\int y f_{Y|W}(y|w_0)dy \right] f_{W|V}(w_0|v_0)du \\
&= E\left[E[Y|W]|V = v_0\right]
\end{aligned}$$

We can now derive the **G-computation formula** for parameter $E[Y_d|V]$

$$\begin{aligned}
E[Y_d|V] &= E[E[Y_d|W]|V] \qquad \text{by the proposition} \\
&= E[E[Y_d|A = d(W), W]|V] \qquad \text{by RA} \\
&= E[E[Y|A = d(W), W]|V] \qquad \text{by CA}
\end{aligned}$$

Hence, the parameter of interest is expressible as a function of the observed data O, assuming CA and RA hold.

Example, continued To see how the G-computation formula can be converted into an estimator of the parameter of interest, let's consider how we might estimate $E[Y_{d_{25}}|W_2 = 1]$.

step 1 To estimate the inner expectation $E[Y|A = d(W), W]$ from the G-computation formula, first regress Y on A and W. Suppose this yields the regression equation

$$\hat{E}[Y|A, W_1, W_2] = \hat{\alpha}_0 + \hat{\alpha}_1 A + \hat{\alpha}_2 W_1 + \hat{\alpha}_3 W_2 + \hat{\alpha}_4 A W_1$$

step 2 Then, for each subject, plug the values for $A = d_{25}(W)$, W_1, and W_2 into the regression equation to compute $\hat{Y}_{d_{25}} \equiv \hat{E}[Y|A = d_{25}(W), W_1, W_2]$. We now have an estimate of the inner expectation.

step 3 Compute the mean of $\hat{Y}_{d_{25}}$ in the female subsample (this is an estimate of the outer expectation from the G-computation formula). This gives $\hat{E}[Y_{d_{25}}|W_2 = 1]$.

Suppose W_2 had been a continuous variable like age (say, with range 18-65) instead of a binary variable like sex. Then step 3 above might not work well, since there likely wouldn't be enough subjects at any given age to usefully compute an empirical mean. Instead, we could assume a model $m(age|\beta)$ such as, say,

$$E[Y_{d_{25}}|age] = m(age|\beta) \equiv \beta_0 + \beta_1 age + \beta_2 age^2$$

and then regress $\hat{Y}_{d_{25}}$ on age to estimate the β parameters. Then $\hat{E}[Y_{d_{25}}|age] = m(age|\hat{\beta})$.

In fact, since we're interested in evaluating dynamic regimes d_θ for multiple values of θ, rather than formulating a separate model $m_\theta(age|\beta)$ for each θ, it would be more useful to formulate a single model $m(age, \theta|\beta)$ that includes θ as an independent variable; e.g.,

$$E[Y_{d_\theta}|age] = m(age, \theta|\beta) \equiv \beta_0 + \beta_1 age + \beta_2 age^2 + \beta_3 \theta + \beta_4 \theta age$$

To fit this model, we would first complete step 2 for each value of θ, so that for each subject we compute the vector $(\hat{Y}_{d_\theta} : all\theta)$. Then we would regress $\hat{Y}_{d_\theta}$ (all θ) on age and θ to estimate the β parameters. In our example, where we're just interested in two θ values (i.e., 25 and 35), this repeated-measures regression would

be based on $2n$ data points, two per subject. Note that because the regression is not based on independent observations (since $\hat{Y}_{d_{25}}$ and $\hat{Y}_{d_{35}}$ may be correlated), we might want to use generalized least squares to fit the model. Finally, set $\hat{E}[Y_{d_\theta}|age] = m(age, \theta|\hat{\beta})$.

As a final note, the bootstrap can be used to get standard errors and confidence intervals for the parameter estimates.

Estimation of direct and indirect causal effects (Kasper Daniel Hansen)

We focus on the point treatment case. The reader is referred to "Estimation of Direct and indirect causal effects in longitudinal studies" by Mark J. van der Laan and Maya L. Petersen, available from http://www.bepress.com/ucbbiostat/paper155.

We are looking at an observed data structure $O = (W, A, Z, Y)$ with the time (causal) ordering $W \mapsto A \mapsto Z \mapsto Y$. As usual W is the baseline covariates, A is the treatment, Z is the intermediate variable and Y is the outcome.

We are interested in the direct effect of the treatment on the outcome, which (intuitively) is the effect of A which is *not* mediated through the intermediate variable Z.

In our approach we will start by treating Z as a response (we will construct counterfactuals for it) as well as a treatment variable (we will index the counterfactual outcomes of the response by Z). This will allow us to define the direct effect of A (on Y).

We assume existence of counterfactuals

- $(Y_{az}, a \in \mathcal{A}, z \in \mathcal{Z})$ (where $\mathcal{A}$ is the set of possible values for A and $\mathcal{Z}$ is the set of possible values for Z.
- $(Z_a, a \in \mathcal{A})$.

The full data structure is thus $X^{\text{full}} = (W, (Z_a, a \in \mathcal{A}), (Y_{az}, a \in \mathcal{A}, z \in \mathcal{Z}))$.

We are now able to define the parameter of interest: the direct effect of A (on Y) is defined to be

$$\text{DF}(a) = E(Y_{aZ_0}) - E(Y_{0Z_0})$$

Z_0 in the definition is a random variable - the counterfactual outcome had A been assigned to 0. Note that we here define the direct effect as a difference between two expectations – this specific form (as opposed to eg. an odds ratio or similar) will be important for the calculations below.

As usual we need a couple of assumptions. They are (more or less) equal to the standard assumptions of causal inference with one extra assumption.

CA (consistency assumption): We assume the existence of counterfactuals (described above) such that the observed data are $O = (W, A, Z_A, Y_{AZ_A})$.

RA (randomization assumption): We assume

$$(A, Z) \perp (Y_{az}, a \in \mathcal{A}, z \in \mathcal{Z}) \,|\, W$$

and

$$A \perp (Z_a, a \in \mathcal{A}) \,|\, W$$

The last (and new) assumption – in lack of better terminology we will call it the direct effect assumption – has a couple of variants in the literature. In addition a new one saw light during the lecture.

DEF (direct effect assumption) (lecture version):

$$E(Y_{az} - Y_{0z} \,|\, Z_0 = z, W) = E(Y_{az} - Y_{0z} \,|\, W), \text{ for all } z$$

DEF (article version) (the article referred to in the beginning of the section)

$$Y_{az} - Y_{0z} \perp Z_0 \,|\, W, \text{ for all } a, z$$

DEF (Robins' version) : $Y_{az} - Y_{0z}$ is a random variable *not depending* on z, *only* on a

DEF (Pearl's version) :

$$Y_{az} \perp Z_0 \,|\, W \text{for all } a, z$$

A couple of comments on the different assumptions

- It is clear that the article version as well as Pearl's version imply the lecture version. Apparently the lecture version is the weakest of the four.

- Pearl's assumption is very strong: See Robins, Greenland's article for relevant discussion.
- If the following model for the counterfactual hold

$$EY_{az} = \beta_0 + \beta_1 a + \beta_2 z + \beta_3 az$$

then

$$E(Y_{az} - Y_{0z}) = \beta_1 a + \beta_3 az$$

If $\beta_3 \neq 0$ then Robin's assumption fails (since if the expectation depends on z the random variables do as well). On the other hand if $\beta_3 = 0$ then Robin's assumption *might* hold.

We now derive a formula used in the estimation of the direct effect.

Theorem We have the following relation

$$\mathrm{DF}(a) = E_w \int_{\mathcal{Z}} E(Y \mid A = a, Z = z, W) - E(Y \mid A = 0, Z = z, W)\, dP_{Z_0|W}(z)$$

Proof We have

$$\begin{aligned} DF(a) &= E(Y_{aZ_0} - Y_{0Z_0}) \\ &= E_W(E\big(Y_{aZ_0} - Y_{0Z_0} \mid W\big)) \\ &= E_W(E_{Z_0|W}\big(E\Big(Y_{aZ_0} - Y_{0Z_0} \mid Z_0, W\Big)\big)) \\ &= E_W \int_{\mathcal{Z}} E(Y_{aZ_0} - Y_{0Z_0} \mid Z_0 = z, W)\, dP_{Z_0|W}(z) \end{aligned}$$

Nothing much has happened so far, we have used the principle of successive conditioning, the last equality is (depending on one's framework) a case of definition. Now, in the last expression we are conditioning on the event $(Z_0 = z)$. Using this event we may write

$$= E_W \int_{\mathcal{Z}} E(Y_{az} - Y_{0z} \mid Z_0 = z, W)\, dP_{Z_0|W}(z)$$

Now we use the DEF assumption (lecture version) to trivially get

$$= E_W \int_{\mathcal{Z}} E(Y_{az} - Y_{0z} \mid W)\, dP_{Z_0|W}(z)$$

Finally we use the standard causal result that $E(Y_{az}) = E(Y \mid A = a, Z = z)$ (this requires the consistency and randomization assumptions) to get

$$= \quad E_W \int_{\mathcal{Z}} E(Y \mid A = a, Z = z, W) - E(Y \mid A = 0, Z = z, W) \mid W)\, dP_{Z_0|W}(z)$$

and the proof is complete.

Example 1: Let us assume that we have a model

$$E(Y \mid A = a, Z = z, W = w) = \alpha_0 + \alpha_1 a + \alpha_2 z + \alpha_3 w$$

In that case

$$E(Y \mid A = a, Z = z, W) - E(Y \mid A = 0, Z = z, W) = \alpha_1 a$$

is the integrand of the integral in the above theorem. This integrand now needs to be integrated according to the theorem. As the integrand only depends on a, it may be moved outside the integral and we get

$$\mathrm{DF}(a) = \alpha_1 a$$

Example 2 : We assume a more elaborate model than in example 1:

$$E(Y \mid A = a, Z = z, W = w) = \beta_0 + \beta_1 a + \beta_2 z + \beta_3 w + \beta_4 az + \beta_5 aw$$

Again we compute the integrand of the integral in the theorem, and get

$$E(Y \mid A = a, Z = z, W = w) - E(Y \mid A = 0, Z = z, W = w) = \beta_1 a + \beta_4 az + \beta_5 aw$$

Using the theorem we get

$$\begin{aligned}
\mathrm{DF}(a) &= E_W \int_{\mathcal{Z}} \beta_1 a + \beta_4 az + \beta_5 aw \, dP_{Z_0|W}(z) \\
&= \beta_1 a + \beta_5 a E(W) + \beta 4 a E_W \int z \, dP_{Z_0|W}(z) \\
&= \beta_1 a + \beta_5 a E(W) + \beta 4 a E_W \Big(E(Z_0 \,|\, W) \Big) \\
&= \beta_1 a + \beta_5 a \underbrace{E(W)}_{=c_1} + \beta 4 a \underbrace{E_W \Big(E(Z|A=0,W) \Big)}_{=c_2} \\
&= a(\beta_1 + \beta_4 c_1 + \beta_5 c_2)
\end{aligned}$$

where the next-to-last equation follows by using the standard causal identification.

Example 3 : Suppose we have a (more general) model of the form

$$E(Y \,|\, A = a, Z = z, W = w) = m_0(z, w) + a m_1(z, w)$$

where m_0, m_1 are fixed functions. In that case

$$E(Y \,|\, A = a, Z = z, W = w) - E(Y \,|\, A = 0, Z = z, W = w) = a m_1(z, w)$$

And we have

$$\begin{aligned}
\mathrm{DF}(a) &= E_W \int_{\mathcal{Z}} a m_1(z, w) \, dP_{Z_0|W}(z) \\
&= a \underbrace{E_W \int_{\mathcal{Z}} m_1(z, w) \, dP_{Z_0|W}(z)}_{=c} \\
&= ac
\end{aligned}$$

Estimation of direct causal effects with MSMí¯s.

Marginal structural models are models for marginal distributions of treatment specific counterfactuals, possibly conditional on baseline covariates. We are looking at an observed data structure $O = (W, A, Z, Y)$, where W is the baseline covariates, A is the treatment, Z is the intermediate variable and Y is the outcome. They are random variables and time ordering $W \mapsto A \mapsto Z \mapsto Y$. In the definition of a direct effect of A on Y one blocks the effect of A mediated through the intermediate variable Z.

The full data structure is $X^{full} = (W, (Z_a, a \in A), (Y_{az}, a \in A, z \in Z))$ in terms of the counterfactuals $(Z_a, a \in A)$ and $(Y_{az}, a \in A, z \in Z)$ for the intermediate variable and outcome, respectively. X denotes the full data structure on a randomly sampled subject, one would like to observe. The direct effect of A (on Y) is defined to be $DF(a) = E(Y_{aZ_0}) - E(Y_{0Z_0})$. A direct effect is thus defined as the population mean of individual direct effects, where an individual direct effect is the difference between the outcome when an individual is treated and the intermediate variable is set at its value under no treatment, and the outcome when the same individual is not treated.

In the full data world we would observe, for each individual, the value of the intermediate variable over time resulting from each possible treatment history, and the value of the covariate process over time, including the outcome process, resulting from each combination of possible treatment history and possible intermediate variable history. Instead, our observed data is only a subset of this full data, consisting of a single treatment history and the corresponding intermediate variable and outcome.

Under the following assumptions, we are able to build our new marginal direct effect model.

CA (consistency assumption): We assume the existence of counterfactuals such that the observed data are $O = (W, A, Z_A, Y_{AZ_A})$.

RA (randomization assumption): We assume $(A, Z) \perp ((Y_{az}; a \in A, z \in Z|)W)$, $A \perp (Z_a, a \in A)|W$

DEA (direct effect assumption): $E(Y_{az} - Y_{0z}|Z_0 = Z, W) = E(Y_{az} - Y_{0z}|W)$ for all z.

So **marginal direct effect model** is $E(Y_{az_0} - Y_{0z_0}|V) = m(a, V|B_0)$ where V is a subset of baseline covariates W and B_0 is our parameter of interest of the full data distribution F_X, where the parametrization m satisfies $m(0, V|B) = 0$ for all B.

We will propose a class of estimating functions for this parameter B_0. The proof that this class of estimating functions is indeed unbiased relies on the previously established identifiability result of $m(a, V|B_0)$, which we state her for convenience. **Theorem**

$$\begin{aligned} DF(a, v) &= E_0(Y_{az_0} - Y_{0z_0}|V) \\ &= E_{W|V} \int_Z (E_0(Y|A = a, Z = z, W) - E_0(Y|A = 0, Z = a, W))dP_0(Z = z|A = 0, W) \end{aligned}$$

We are using 0 as an index for true distribution. $E_{W|V}$ denotes the conditional expectation of W, given V.

How to estimate B using thIS result?

Firstly, we note that

$$\begin{aligned} E_0(Y_a, Z_0|V) &= E(Y_{0Z_0}|V) + E(Y_{aZ_0} - Y_{0Z_0}|V) \\ &= \alpha(V|\eta_0) + m(a, V|B_0) \end{aligned}$$

Let $\theta = (B, \eta)$ be the combined parameter. We propose the following class of estimating functions for $\theta = (B, \eta)$ indexed by a vector function h of A, V of the same dimension as the dimension of θ:

$$D_h(O|\theta, g_{Z|A,W}, g_{A|W}) = \frac{h(A, V)g(Z|A = 0, W)(Y - \alpha(V|\eta) - m(A, V|B))}{g(A|W)g(Z|A, W)}$$

We propose as particular choice

$$h^*(A, V) \equiv \frac{d}{d\theta}(\alpha(V \mid \eta) + m(A, V \mid B).$$

Given estimators $g_{n,Z|A,W}$ and $g_{n,A|W}$ of $g_{Z|A,W}$ and $g_{A|W}$, respectively, and a possibly data dependent index h_n (estimating h^*), let θ_n be the solution of the corresponding estimating equation:

$$0 = \sum_{i=1}^{n} D_{h_n}(O \mid \theta, g_{n,Z|A,W}, g_{n,A|W}).$$

It is of interest to compare this estimator with an estimator of β_0 directly based on the identifiability result stated in the theorem above. That is, one substitutes estimates of $E(Y \mid A, Z, W)$ and $g_{Z|A,W}$ into the identifiability mapping for $E(Y_{aZ_0} - Y_{0Z_0} \mid W)$, and one regresses this on a, V according to the regression model $m(a, V \mid \beta)$. We refer to the latter estimator as the likelihood based estimator. In the likelihood based estimator, one needs to estimate

(1) $E(\overline{Y}|A, Z, W)$
(2) $g(Z|A, W)$

On the contrast, the IPTW-estimator of β relies on estimation of

(1) $g(A|W)$
(2) $g(Z|A, W)$

The IPTW estimator:

Suppose we observe $(W, A, Y = Y_A)$, and the randomization assumption $A \perp (Y_a : a) \mid W$ (RA) holds. Suppose that we are interested in estimation of β_0 in the MSM $E(Y_a \mid V) = m(a, V \mid \beta_0)$.

Definition:The IPTW estimating function for β_0 with nuisance parameter $g(A \mid X) = g(A \mid W)$ is defined as:

$$D_h(O|g,\beta) = \frac{h(A,V)\epsilon(\beta)}{g(A|X)}$$

where $\epsilon(\beta) = Y_A - m(A,V|\beta)$

Note that the IPTW estimating function is indeed a function of the observed data under RA. We denote the estimator of the nuisance parameter g with g_n.

The IPTW estimator of β is defined as the solution of the estimating equation associated with the observed data O and the IPTW estimating function at g_n:
$\sum_{i=1}^{n} D_h(O_i|g_n,\beta) = 0$ where O_i for $i = 1, ..., n$ represents the n i.i.d. observations in the observed data.

The IPTW estimating function is **unbiased** at the true β: $E_{P_{F_X},g} D_h(O|\beta,g) = 0$ if the ETA assumption holds for $g(A \mid W)$:

$$max_{\bar{a}\in A}\frac{h(\bar{a},V)}{g(\bar{a}|X)} < \infty, F_x - ae$$

Thus, if g_n is consistent for $g(A \mid W)$, and the ETA holds, then the IPTW estimator is asymptotically linear and thus consistent.

Proof:

$$\begin{aligned}
E_{P_{F_X},g}[D_h(O \mid g,\beta)] &= EE\left(\frac{h(\bar{A},V)\epsilon(\beta)}{g(\bar{A}|X)}|X\right) \\
&= E_{F_X}\left(\sum_{\bar{a}:g(\bar{a} \mid X)\neq 0} \frac{h(\bar{a},V)\epsilon_{\bar{a}}(\beta)}{g(\bar{a}|X)} g(\bar{a}|X)\right) \\
&\overset{\text{ETA}}{=} E_{F_X}\big(\sum_{\bar{a}} h(\bar{a},V)\epsilon_{\bar{a}}(\beta)\big) \\
&= \sum_{\bar{a}} E_{F_X}(h(\bar{a},V)\epsilon_{\bar{a}}(\beta)) \\
&= \sum_{\bar{a}\in\mathcal{A}} E_{F_V}(h(\bar{a},V)\epsilon_{\bar{a}}(\beta) \mid V) \\
&= \sum_{\bar{a}\in\mathcal{A}} E_{F_V}(h(\bar{a},V)E(\epsilon_{\bar{a}}(\beta)|V)) \\
&= 0
\end{aligned} \tag{4}$$

Where $X = (T_{\bar{a}}, \bar{W}_a(T_{\bar{a}}))_{\bar{a}} \in A \sim F_X$ for the full data.

The IPTW estimate of β can be obtained in practice by performing a weighted least squares regression of Y on $\bar{A}$ and V using the MSM and weights inversely proportional to the treatment mechanism:
$w(\bar{A},V) = \frac{\lambda(\bar{A},V)}{g_n(\bar{A}|X)} \overset{\text{SRA}}{=} \frac{\lambda(\bar{A},V)}{\Pi_{t=0}^{K} g_n(A(t)|A(t-1),\bar{L}(t))}$, where λ can be any non-null function of $\bar{A}$ and V.

It can indeed be shown that the resulting estimate is a solution of the IPTW estimating equation where $h(\bar{a},V) = \lambda(\bar{a},V)\frac{d}{d\beta}(\bar{a},V|\beta)$.

If we substitute λ by $g^{'}(\bar{A}|V)$ where $g^{'}$ is the conditional distribution of $\bar{A}$ given V,the resulting weights are called stabilized weights: $w(\bar{A},V) = \frac{g_n^{'}(\bar{A}|V)}{g_n(A|V)}$.

In addition, it was shown, see van der Laan and Robins (2002), that the treatment mechanism g should always be estimated even when g is known, as would be the case in a randomized trial. As a result, the

IPTW estimator may gain in efficiency by taking into account possible empirical confounding, without loss of consistency.

Estimating Functions, Kelly Moore, September 20, 2004.

Why do we want to use another procedure than maximum likelihood?

Motivating example:

$T_1, ..., T_n$ I.I.D $T \sim f_0$ (unspecified)
Parameter of interest: $\mu_0 = S_0(t)$ (Survival function at point t)
Likelihood: $L(f) = \prod_{i=1}^{n} f(x_i)$
Use piecewise linear densities indexed by knot points $(x_1, ...x_n)$
Sieve: Sequence of subspaces of the whole space which approximates the whole model

1) Set n equally spaced knot points
2) Maximum Likelihood over piecewise linear densities

Let sieve be $M(m) \subset \mathcal{M}$ where $\mathcal{M}$ is all densities.
$\hat{f}_m(P_n)$ is Maximum Likelihood of f_0 based on the model indexed by m knot points $M(m)$. P_n is the empirical distribution of $T_1, ..., T_n$.

How do we choose the number of knots?
Likelihood Cross-Validation. Given the sieve, this is a data adaptive procedure for choosing m.

Let $B_n \in \{0, 1\}^n$

$$B_n(i) = \begin{cases} 1 & \text{if } T_i \text{ is in validation sample} \\ 0 & \text{if } T_i \text{ is in training sample} \end{cases}$$

Given B_n, let P^0_{n,B_n}, P^1_{n,B_n} be the empirical distributions of the training and validation samples respectively.

C.V. Likelihood indexed by m
Here we compute likelihood over validation sample(s) using the likelihood from the training sample.

$$L_{CV}(m) = E_{B_n}\left(\log\left(\prod_{i:B_n(i)=1} \hat{f}_m(P^0_{n,B_n}(T_i))\right)\right)$$

Let

$$\hat{m} = \text{argmax}_m\{L_{CV}(m)\}$$

Estimate f_0 with $\hat{f}_m(P_n)$.
Estimate μ_0 with estimator from above:

$$\hat{\mu}_{MLE} = \int_t^\infty \hat{f}_{\hat{m}}(P_n)(s)ds$$

Why is this not the best method in this case?
This procedure uses the best trade-off of bias and variance. (m is small then you have a small variance and large bias, m is large then you have a large variance and small bias, as $m \longrightarrow \infty$ then we approach empirical) (If f_0 is what you want then this is a good procedure)

$$\|\hat{f}_{\hat{m}}(P_n) - f_0\| = O(n^{-2/5})$$

$$\text{Bias}(\hat{f}_{\hat{m}}(P_n) = O(n^{-2/5}) \text{ (SE behaves in the same way)}$$

Thus,

$$E(\hat{\mu}_{MLE} - \mu_0) = \int_t^\infty E(\hat{f}_{\hat{m}}(P_n)(s) - f_0(s))ds = O(n^{-2/5})$$

Averaging reduces the variability but not necessarily the bias. Bias should be converging at a rate $< 1/\sqrt{n}$ to be as good as the estimator using the empirical.
But here $\sqrt{n}(Bias) = \sqrt{n}O(n^{-2/5} = O(n^{1/10})$
Thus, likelihood is not the best procedure here. (In literature this is referred to as the curse of dimensionality)

Estimating Function Approach
1) $O_1, ...O_n$ $P_0 \in \mathcal{M}$. Our parameter of interest is: μ: $\mathcal{M} \to \mathbb{R}$, $\mu_0 = \mu(P_0)$
Example: $O = (Y, x)$
$\mathcal{M} = \{P : E_p[Y|X] = m(X|\beta)\}$
Parameter of interest: $\mu_0 = \beta_0$
2) At any $P_0 \in \mathcal{M}$: Define a class of $\{P_\varepsilon : \varepsilon = (-\delta, \delta)\} \subset \mathcal{M}, P_{\varepsilon=0} = P_0$.
We call a score

$$S(0) = \frac{d}{d\varepsilon} \log(P_\varepsilon)|_{\varepsilon=0} = 0 \text{ for which } \frac{d}{d\varepsilon}\mu(P_\varepsilon)|_{\varepsilon=0} = 0$$

a nuissance score.
Let $T_{NUIS}(P_0)$ be the Hilbert space generated by these nuissance scores within the Hilbert space.

$$L_0^2(P_0) = \{S(0) : E_0(S(0)) = 0, E_0(S^2(0)) < \infty\}$$

endowed with inner product $< S_1, S_2 >= E_0(S_1(0)S_2(0))$
Othogonal Complement of the nuissance tangent space:

$$T_{NUIS}^{\perp}(P_0) = \{S \in L_0^2(P_0) : S \perp T_{NUIS}(P_0)\}$$

3) From this we can get the class of all estimating functions. $D_n(0|\mu, \eta)$ such that $D_h(0|\mu_0, \eta_0) \in T_{NUIS}^{\perp}(P_0)$

Example: $O = (Y, X)$, $\beta \in \mathbb{R}^d$, Y is kx1
$\mathcal{M} = \{P : E_P[Y|X] = m(X|\beta)\}$
$T_{NUIS}^{\perp}(P_0) = h(X)(\bar{Y} - \bar{m}(X|\beta_0))\}$ (h is 1xk)

Estimating functions: $D_h(x, \bar{Y}|\beta) = h(X)(\bar{Y} - m(X|\beta_0))$
Find optimal h:
$h_{opt}(x) = \frac{d}{d\beta}\bar{m}(X|\beta)_{dxk}E(\varepsilon(\beta)\varepsilon(\beta)^T|X)^{-1}$, $\varepsilon(\beta) = \bar{Y} - m(X|\beta_0)$

10/11/2004 Lecture Notes, Dan Rubin

This lecture concerns estimating function methods in marginal structural models for point treatment longitudinal studies. The data is $O = (W, A, Y)$, where W, A, Y represent covariates, treatment, and an outcome. Let $X = (W, Y)$, and let $g(\cdot|W)$ denote the conditional distribution of A, given W. As usual, we make the consistency assumption, so assume there is an unobserved full data structure of counterfactuals $X^{Full} = (W, (Y_a : a \in \mathcal{A}))$, for $\mathcal{A}$ a finite set of possible treatments. We also make the randomization assumption, which here states that $g(a|X^{Full}) = g(a|W)$, and the experimental treatment assumption (ETA) that $g(a|W) > 0$ for all $a \in \mathcal{A}$ with probability one.

Suppose the parameter of interest is $E[Y_a|V]$, where $V \subseteq W$. In general, this will be hard to estimate for high dimensional data without additional assumptions, which is why marginal structural models were introduced. A marginal structural model assumes that $E[Y_a|V] = m(a, V|\beta)$, for a known function m, where β is an unknown k-dimensional Euclidean parameter. When we assume such a model, β becomes the parameter of interest.

In previous lectures, we studied how estimating functions can be used to estimate β when the full data is observed. In particular, the class of full data estimating functions was given by:

$$\{D_h(X^{Full}|\beta) = \sum_{a\in\mathcal{A}} h(a,V)(Y_a - m(a,V|\beta)) \in L_0^2(X^{Full}) : h\} \tag{5}$$

We next define the class of inverse probability of treatment weighted (IPTW) estimators as follows:

$$\{D_{h,IPTW}(O|\beta,g) = \frac{h(A,V)}{g(A|W)}(Y - m(A,V|\beta)) \in L_0^2(O) : h\} \tag{6}$$

Note that unlike in the full data estimating functions, the IPTW estimating functions depend on a nuisance parameter (g), which must be estimated from the data. It is trivial to check that under the consistency, SRA and ETA assumptions (unlike G-computation methods, ETA is necessary here), $E[D_{h,IPTW}(O|\beta,g)|X^{Full}] = D_h(X^{Full}|\beta)$. So by first conditioning on X^{Full}, the IPTW estimating functions are indeed unbiased estimating functions for β.

Although the IPTW estimators are simple and easy to implement, we can improve upon them in terms of both efficiency and robustness, using estimating functions described below. Before giving these functions, we need a few definitions and facts. Consider the Hilbert space $L_0^2(O) = \{h(O) : E[h(O) = 0], E[h^2(O)] < \infty\}$ endowed with the inner product $(h_1(O), h_2(O)) = E[h_1(O)h_2(O)]$. Let T_{RA} denote the linear closure in this Hilbert space of all scores of one dimensional regular parametric submodels P_ϵ of the data generating distribution, where the submodels only fluctuate the treatment mechanism g, and pass through the true data generating distribution at $\epsilon = 0$. Such submodels can be represented as $P_\epsilon(O) = P(Y|A,W)P(W)g_\epsilon(A|W)$, where $-1 < \epsilon < 1$ and $g_{\epsilon=0}(A|W) = g(A|W)$. Consider the submodels $g_\epsilon(A|W) = (1 + \epsilon s(A,W))g(A|W)$, where $s(A,W) \in L_0^2(O)$ and $E[s(A,W)|W] = 0$. Taking the log of $P_\epsilon(0)$ and differentiating, it is easy to see that $s(A,W)$ is the score. From this, it can be shown that $T_{RA} = \{s(A,W) \in L_0^2(O) : E[s(A,W)|W] = 0\}$. We now need a few further facts.

(1) From Theorem 1.3 in Mark's book with Jamie Robins, the class of all possible estimating is given by $\{D_{h,IPTW}(O|\beta,g) - \Pi(D_{h,IPTW}(O|\beta,g)|T_{RA})\}$, where Π represents the projection operator in Hilbert space.
(2) For any function $f(O) \in L_0^2(O)$, $\Pi(f(O)|T_{RA}) = E[f(O)|A,W] - E[f(O)|W]$.

To see part (2), note that clearly the proposed $\Pi(f(O)|T_{RA})$ is in T_{RA}, because:

$$E[\Pi(f(O)|T_{RA})] = E[E[f(O)|A,W]|W] - E[E[f(O)|W]|W] = E[f(O)|W] - E[f(O)|W] = 0 \tag{7}$$

And if $s(A,W) \in T_{RA}$ then $E[s(A,W)|W] = 0$, so:

$$E[(f - E[f|A,W] + E[f|W])s(A,W)] = E[fs(A,W)] - E[E[f|A,W]s(A,W)] + E[E[f|W]s(A,W)] \tag{8}$$

$$= E[s(A,W)E[f|A,W]] - E[s(A,W)E[f|A,W]] + E[E[f|W]E[s(A,W)|W]] = 0 \tag{9}$$

Hence, subtracting the IPTW estimating functions from their projections on T_{RA}, we obtain the following class of all possible estimating functions. These depend on two nuisance parameters: the treatment mechanism $g(a|W)$ and $E[Y - m(A,V|\beta)|A,W]$. These are called *doubly robust* estimating functions, because they are unbiased if either of the two nuisance parameters is correctly specified.

$$\{\frac{h(A,V)}{g(A|W)}(Y - m(A,V|\beta)) - \frac{h(A,V)}{g(A|W)}E[Y - m(A,V|\beta)|A,W] + \sum_{a\in\mathcal{A}} h(a,V)E[Y - m(a,V|\beta)|A = a,W] : h\} \tag{10}$$

10/11/2004 and 10/13/2004 Lecture Notes, Dan Rubin

This lecture concerns estimating function methods in marginal structural models for point treatment longitudinal studies. The data is $O = (W,A,Y)$, where W,A,Y represent covariates, treatment, and an outcome. Let $X = (W,Y)$, and let $g(\cdot|W)$ denote the conditional distribution of A given W. As usual, we make the consistency assumption, so assume there is an unobserved full data structure of counterfactuals $X^{Full} = (W,(Y_a : a \in \mathcal{A}))$, for $\mathcal{A}$ a finite set of possible treatments, and that $Y = Y_A$. We also make the randomization assumption, which here states that $g(a|X^{Full}) = g(a|W)$, and the experimental treatment

assumption (ETA) that $g(a|W) > 0$ for all $a \in \mathcal{A}$.

Suppose the parameter of interest is $E[Y_a|V]$, where $V \subseteq W$. In general, this will be hard to estimate for high dimensional data without additional assumptions, which is why marginal structural models were introduced. A marginal structural model assumes that $E[Y_a|V] = m(a, V|\beta)$, for a known function m, where β is an unknown k-dimensional Euclidean parameter. When we assume such a model, β becomes the parameter of interest.

In previous lectures, we studied how estimating functions can be used to estimate β when the full data is observed. In particular, the class of full data estimating functions was given by:

$$\{D_h(X^{Full}|\beta) = \sum_{a \in \mathcal{A}} h(a, V)(Y_a - m(a, V|\beta)) \in L_0^2(X^{Full}) : h\} \tag{11}$$

We next define the class of inverse probability of treatment weighted (IPTW) estimators as follows:

$$\{D_{h,IPTW}(O|\beta, g) = \frac{h(A, V)}{g(A|W)}(Y - m(A, V|\beta)) \in L_0^2(O) : h\} \tag{12}$$

Note that unlike in the full data estimating functions, the IPTW estimating functions depend on a nuisance parameter (g), which must be estimated from the data. It is trivial to check that under the consistency, SRA and ETA assumptions (unlike G-computation methods, ETA is necessary here), $E[D_{h,IPTW}(O|\beta, g)|X^{Full}] = D_h(X^{Full}|\beta)$. So by first conditioning on X^{Full}, the IPTW estimating functions are indeed unbiased estimating functions for β.

Although the IPTW estimators are simple and easy to implement, we can improve upon them in terms of both efficiency and robustness, using estimating functions described below. Before giving these functions, we need a few definitions and facts. Consider the Hilbert space $L_0^2(O) = \{h(O) : E[h(O) = 0], E[h^2(O)] < \infty\}$ endowed with the inner product $(h_1(O), h_2(O)) = E[h_1(O)h_2(O)]$. Let T_{RA} denote the linear closure in this Hilbert space of all scores of one dimensional regular parametric submodels P_ϵ of the data generating distribution, where the submodels only fluctuate the treatment mechanism g, and pass through the true data generating distribution at $\epsilon = 0$. Such submodels can be represented as $P_\epsilon(O) = P(Y|A, W)P(W)g_\epsilon(A|W)$, where $-1 < \epsilon < 1$ and $g_{\epsilon=0}(A|W) = g(A|W)$. Consider the submodels $g_\epsilon(A|W) = (1 + \epsilon s(A, W))g(A|W)$, where $s(A, W) \in L_0^2(O)$ and $E[s(A, W)|W] = 0$. Taking the log of $P_\epsilon(0)$ and differentiating, it is easy to see that $s(A, W)$ is the score. From this, it can be shown that $T_{RA} = \{s(A, W) \in L_0^2(O) : E[s(A, W)|W] = 0\}$. We now need a few further facts.

(1) From Theorem 1.3 in Mark's book with Jamie Robins, the class of all possible estimating is given by $\{D_{h,IPTW}(O|\beta, g) - \Pi(D_{h,IPTW}(O|\beta, g)|T_{RA})\}$, where Π represents the projection operator in Hilbert space.
(2) For any function $f(O) \in L_0^2(O)$, $\Pi(f(O)|T_{RA}) = E[f(O)|A, W] - E[f(O)|W]$.

To see part (2), note that clearly the proposed $\Pi(f(O)|T_{RA})$ is in T_{RA}, because:
$E[\Pi(f(O)|T_{RA})] = E[E[f(O)|A, W]|W] - E[E[f(O)|W]|W] = E[f(O)|W] - E[f(O)|W] = 0$
And if $s(A, W) \in T_{RA}$ then $E[s(A, W)|W] = 0$, so:
$E[(f - E[f|A, W] + E[f|W])s(A, W)] = E[fs(A, W)] - E[E[f|A, W]s(A, W)] + E[E[f|W]s(A, W)]$
$= E[s(A, W)E[f|A, W]] - E[s(A, W)E[f|A, W]] + E[E[f|W]E[s(A, W)|W]] = 0$

Hence, subtracting the IPTW estimating functions from their projections on T_{RA}, we obtain the following class of all possible estimating functions. These depend on two nuisance parameters: the treatment mechanism $g(a|W)$ and $E[Y - m(A, V|\beta)|A, W]$. These are called *doubly robust* estimating functions, for reasons we will discuss below. Here, $Q(A, W|\beta) = E[Y - m(A, V|\beta)|A, W]$.

$$\{D_{h,DR}(O|\beta, g, Q) = \frac{h(A, V)}{g(A|W)}(Y - m(A, V|\beta)) - \frac{h(A, V)}{g(A|W)}Q(A, W|\beta) + \sum_{a \in \mathcal{A}} h(a, V)Q(a, W|\beta) : h\} \tag{13}$$

Now consider the model where $g = g_0$ is known, as would be the case in a randomized trial. Intuitively, this constrained model allows for less submodels through the data generating distribution varying the nuisance parameter g, so the nuisance tangent space should be smaller than before. Hence, the orthocomplement of the nuisance tangent space (meaning the class of all estimating functions) should be larger than before. This is indeed the case. In this constrained model, the orthogonal complement of the nuisance tangent space at the data generating distribution P_0 is:

$$T^{\perp}_{NUIS}(P_0) = \{D_{h,IPTW}(O|\beta, g_0) + \varphi(A, W) : h, \varphi(A, W) \in T_{RA}\} \tag{14}$$

For given h, it can be shown that the function $D_{h,IPTW}(O|\beta, g_0) + \varphi(A, W)$ in $T^{\perp}_{NUIS}$ with the smallest variance (and hence the best estimating function) has $\varphi(A, W) = \Pi(D_{h,IPTW}(O|\beta, g)|T_{RA})$. However, this choice of φ leads to the estimating function $D_{h,DR}(O|\beta, g_0, Q)$. So when g is known, minimizing the variance over (h, φ) leads to the same optimal estimating function as in the model where g is unknown, and the variance is minimized over the index h. In practice, a general recommendation is to take $h(A, V) = \frac{d}{d\beta} m(A, V|\beta) g(A|V)$.

We will now explain why the estimating function $D_{h,DR}(O|\beta, g, Q)$ is called doubly robust. Note that the function depends on two nuisance parameters, g and Q, and each parameter is an unknown function. Here g is the treatment mechanism, while Q only depends on the full data distribution. If *either* of the two nuisance parameters is correctly specified, then the estimating function will be unbiased. Suppose statistician A models g and implements the IPTW estimator, statistician B models Q (so models $E[Y|A, W]$) from which he/she implements the G-computation estimator, and statistician C implements the double robust estimator from the g and Q fits of statisticians A and B. If g and Q are correctly modelled, all three statisticians will have consistent estimators. If g is correctly modelled but Q is not, statisticians A and C will be consistent. If Q is correctly modelled but g is not, statisticians B and C will be consistent. Therefore, statistician C is said to implement a doubly robust estimator, because he/she is consistent whenever statisitician A or B is consistent. Note that when g is correctly modelled but Q is not, the doubly robust estimator still has smaller asymptotic variance than the IPTW estimator, so it can improve efficiency as well as robustness.

Lecture of October 25, 2004, Oliver Bembom

Causal attributable risk models

We consider again a data structure consisting of baseline covariates W, point treatment A, and outcome Y: $O = (W, A, Y)$. Under the consistency assumption, there exist full data $X = (W, \{Y_a : a \in \mathcal{A}\})$ that allow us to view this as a missing data problem by thinking of the observed data as $O = (W, A, Y) = (W, A, Y_A) \sim P_{F_{X_0}, g_0}$ where F_{X_0} is the distribution of the full data and g_0 is the conditional distribution of the treatment mechanism given the full data: $g_0(a|X) \equiv P(A = a|X)$. Under the randomization assumption, treatment assignment is independent of the full data X given the baseline covariates W, i.e. $g_0(a|X) = g_0(a|W)$.

The class of causal attributable risk models arises when one is interested in comparing the observed marginal distribution of Y to the marginal distribution of the counterfactuals Y_{a_0} for different treatments a_0. If desired, this comparison can be stratified based on a subset V of the baseline covariates W. These models may be of interest, for example, when one wishes to estimate the effect of different interventions on the marginal distribution of Y.

The two distributions of interest can be compared through a number of different summary measures that then yield natural parameters of interest. If we wish to compare the two distributions on an additive scale, a natural parameter of interest is $E[Y_{a_0} - Y|V]$. If Y is binary, a comparison on a multiplicative scale may be preferred. In this case, one natural parameter of interest is the relative risk

$$\frac{E[Y_{a_0}|V]}{E[Y|V]} = \frac{P[Y_{a_0}=1|V]}{P[Y=1|V]}$$

If we wish to model the relative risk as a function of a_0 and the covariates of interest V, we need to ensure that our model does not allow these two probabilities to lie outside the interval $[0,1]$. If, for example, we were to use the model

$$\log \frac{P[Y_{a_0}=1|V]}{P[Y=1|V]} = \beta_0 + \beta_1 a_0 + \beta_2 v$$

we would have that

$$P[Y_{a_0}=1|V] = P[Y=1|V]\exp(\beta_0 + \beta_1 a_0 + \beta_2)$$

which could easily fall outside the range $[0,1]$. A commonly used approach is to model the log odds rather than the probabilities themselves. In the example above, one might use the model

$$\log \frac{P[Y_{a_0}=1|V]}{1-P[Y_{a_0}=1|V]} = \log \frac{P[Y=1|V]}{1-P[Y=1|V]} + \beta_0 + \beta_1 a_0 + \beta_2 v$$

which always constrains $P[Y_{a_0}=1|V]$ to lie in the interval $[0,1]$. Such models,however, have the disadvantage that they are not robust with respect to the estimation of the nuisance parameter $\log \frac{P[Y=1|V]}{1-P[Y=1|V]}$.

A class of models that not only respects the constraints on the probabilities of interest, but also achieves greater robustness focuses on modelling the switch relative risk given by

$$\frac{E[Y_{a_0}|V]}{E[Y|V]} I\left(\frac{E[Y_{a_0}|V]}{E[Y|V]} \leq 1\right) + \frac{1-E[Y_{a_0}|V]}{1-E[Y|V]} I\left(\frac{E[Y_{a_0}|V]}{E[Y|V]} > 1\right)$$

where $I(\cdot)$ is the indicator function. Such models do not rely on estimating $E[Y|V]$ consistently. In fact, even if one simply plugs in 0.5 for $E[Y|V]$, the switch relative risk is still estimated consistently.

Another approach to comparing the observed marginal distribution of Y to the marginal distribution of the counterfactuals Y_{a_0} is to use quantile-quantile functions. If $X_1 \sim F_1$ and $X_2 \sim F_2$, then the quantile-quantile function $F_2^{-1}F_1(\cdot)$ maps X_1 to a random variables that is distributed as F_2. This approach leads to the class of structural nested models. Note that the quantile-quantile function as defined above does not work for discrete random variables. In this case, one may use the function $F_2^{-1}[\Delta F_1(X_1) + (1-\Delta)F_1(X_1^-)]$, where $\Delta \sim U(0,1)$.

We will now focus on additive models and write $E[Y_{a_0} - Y|V] = m(a_0, V|\beta_0)$, where $m(\cdot)$ is a function that is known up to β_0. Our parameter of interest thus becomes β_0. In order to estimate β_0, we identify $\eta_0(V) \equiv E_0[Y|V]$ as an additional nuisance paramter and write

$$E[Y_{a_0}|V] = \eta_0(V) + m(a_0, V|\beta_0)$$

Suppose $\eta_0(V)$ were known. Then the class of IPTW estimating functions would be given by all

$$D^*_{h,IPTW}(O|g,\eta_0,\beta) = \frac{h(A,V)}{g(A|W)}\left(Y_A - \eta_0(V) - m(A,V|\beta)\right)$$

such that h is any function of A,V. The class of double robust estimating functions is obtained by subtracting the projection of $D^*_{h,IPTW}(O|g,\eta_0,\beta)$ onto the nuisance tangent space under the randomization assumption, T_{RA}. Since $T_{RA} = \{s(A,W) : E[s(A,W)|W] = 0\}$, the projection of $D^*_{h,IPTW}(O|g,\eta_0,\beta)$ onto T_{RA} can be obtained by first projecting onto the larger space of functions $\{s(A,W) : s\}$, yielding $E[D^*_{h,IPTW}(O|g,\eta_0,\beta)|A,W]$, and then completing the projection onto T_{RA} by subtracting the conditional mean of $E[D^*_{h,IPTW}(O|g,\eta_0,\beta)|A,W]$ given W. Letting $Q_0(A,W) = E_0[Y|A,W]$, we have that the class of double robust estimating functions is given by all

$$D^*_{h,DR}(O|g,Q,\eta_0,\beta) = \frac{h(A,V)}{g(A|W)}\left(Y_A - \eta_0(V) - m(A,V|\beta)\right) \quad - \quad \frac{h(A,V)}{g(A|W)}(Q_0(A,W) - \eta_0(V) - m(A,V|\beta)) \quad +$$
$$\sum_{a\in\mathcal{A}} h(a,V)\left(Q_0(a,W) - \eta_0(V) - m(a,V|\beta)\right)$$

such that h is any function of A,V. If the nuisance parameter $\eta_0(V)$ needs to be estimated, $D^*_{h,IPTW}(O|g,\eta_0,\beta)$ and $D^*_{h,DR}(O|g,Q,\eta_0,\beta)$ are not orthogonal to $\eta_0(V)$. We thus would like to find T_{NUIS} under the assumption that $\eta(V)$ is unknown in order to orthogonalize these estimating functions with respect to T_{NUIS}. If we achieve this, our estimator β_n of β will no longer depend on $\eta(V)$ in first order. Furthermore, in some cases, an estimating function that has been orthogonalized to a nuisance parameter remains unbiased even if that nuisance parameter is mis-specified. In order to find T_{NUIS}, we need the following definition and lemma:

Definition: An estimator β_n of β is asymptotically linear with influence function $IC(O|\beta)$ if

$$\beta_n - \beta = \frac{1}{n}\sum_{i=1}^{n} IC(O_i|\beta) + o_p\left(\frac{1}{\sqrt{n}}\right),$$

i.e. if $\beta_n - \beta$ can be written as an empirical mean of a function of the data plus a term that tends to zero in probability even when multiplied by $\sqrt{n}$.

Lemma 1.3 (van der Laan & Robins): Under certain regularity conditions, $T^{\perp}_{NUIS}$ can be identified with the set of all influence functions corresponding to the class of asymptotically linear estimators β_n of β:

$$T^{\perp}_{NUIS} = \{IC_{\beta_n}(O|\beta) : \beta_n \text{ is asymptotically linear}\}$$

Thus we can identify $T^{\perp}_{NUIS}$ by finding the set of all influence functions of the estimators obtained by solving the estimating equation corresponding to $D^*_{h,DR}(O|g,Q,\eta,\beta)$. Having identified $T^{\perp}_{NUIS}$ in this way, we find the projections of $D^*_{h,IPTW}(O|g,\eta,\beta)$ and $D^*_{h,DR}(O|g,Q,\eta,\beta)$ onto $T^{\perp}_{NUIS}$ as

$$D_{h,IPTW}(O|g,\eta,\beta) = D^*_{h,IPTW}(O|g,\eta,\beta) - \sum_a h(a,V)Y + E\sum_a h(a,V)\eta(V)$$
$$D_{h,DR}(O|g,Q,\eta,\beta) = D^*_{h,DR}(O|g,Q,\eta,\beta) - \sum_a h(a,V)Y + E\sum_a h(a,V)\eta(V)$$

We wish to check whether these estimating functions remain unbiased even if the specified $\eta_1(V)$ is wrong. In the case of the IPTW estimating function, suppose that g is specified correctly. Then

$$ED_{h,IPTW}(O|g_0,\eta_1,\beta) = E[ED_{h,IPTW}(O|g_0,\eta_1,\beta)|X] =$$

$$E_{F_{X_0}}\left[\sum_a \frac{h(a,V)}{g(a|W)}\left(Y_A - \eta_1(V) - m(a,V|\beta)\right)g(a|W) - \sum_a h(a,V)Y + E\sum_a h(a,V)\eta_1(V)\right] =$$

$$E_{F_{X_0}}\left[\sum_a h(a,V)\left(Y_A - m(a,V|\beta)\right) - \sum_a h(a,V)Y\right] = EE\left[\sum_a h(a,V)\left(Y_A - Y - m(a,V|\beta)\right)|V\right] = 0$$

Thus the IPTW estimating function will remain unbiased even if η is mis-specified as long as g is estimated consistently.

In the case of the DR estimating function, suppose first that g is specified correctly. Then

$$ED_{h,DR}(O|g_0,Q_1,\eta_1,\beta) = ED_{h,IPTW}(O|g_0,\eta_1,\beta) \quad -$$
$$EE\left[\frac{h(A,V)}{g(A|W)}(Q_1(A,W) - \eta_1(V) - m(A,V|\beta)) + \sum_{a\in\mathcal{A}} h(a,V)\left(Q_1(a,W) - \eta_1(V) - m(a,V|\beta)\right)|W\right] = 0{-}0 = 0$$

Now suppose that Q is specified correctly. Then

$$ED_{h,DR}(O|g_1,Q_0,\eta_1,\beta) =$$
$$EE\left[\frac{h(A,V)}{g(A|W)}\left(Y_A - \eta_1(V) - m(A,V|\beta)\right) - \frac{h(A,V)}{g(A|W)}(Q_0(A,W) - \eta_1(V) - m(A,V|\beta))|A,W\right]$$
$$+E\left[\sum_a h(a,V)\left(Q_0(a,W) - \eta_1(V) - m(a,V|\beta)\right) - \sum_a h(a,V)Y + E\sum_a h(a,V)\eta_1(V)\right] =$$
$$0 + EE\left[\sum_a h(a,V)\left(Q_0(a,W) - Y - m(a,V|\beta)\right)|V\right] = 0$$

We conclude that the DR estimator obtained as the solution of the estimating equation

$$0 = \frac{1}{n}\sum_{i=1}^n D_{h,DR}(O_i|g_n,Q_n,\eta_n,\beta)$$

retains its double robust quality even if η is mis-specified.

Marginal Structural Models for Time-Dependent Treatments: Doubly Robust Estimators

11/1/04-11/3/04

Introduction

Observed Data.

$$\begin{aligned} O &= (L(0), A(0), L(1), A(1), ..., L(K), A(K), Y \equiv L(K+1)) \\ &= (\bar{A}, \bar{L}) \end{aligned}$$

where $\bar{A} \equiv \bar{A}(K) = (A(0), ..., A(K))$ and $\bar{L} \equiv \bar{L}(K+1) = (L(0), ..., L(K+1))$.

Full Data.

$$X = (\bar{L}_{\bar{a}} : \bar{a} \in \mathcal{A})$$

where $\mathcal{A}$ is the set of possible treatments and $\bar{L}_{\bar{a}}$ is the counterfactual outcome process under treatment $\bar{a}$.

Parameter of Interest. We assume a **marginal structural model** (MSM)

$$E[Y_{\bar{a}}|V] = m(\bar{a}, V|\beta_0)$$

where $m()$ is a known function and $V \subseteq L(0)$.

β_0 is the parameter of interest.

Assumptions. Temporal Ordering We assume the time ordering that $L(0)$ precedes $A(0)$ precedes $L(1)$, etc. We also assume that $L_{\bar{a}}(j) = L_{\bar{a}(j-1)}(j)$ for $j = 0, ..., K$.
CA $O = (\bar{A}, \bar{L}_{\bar{A}})$
SRA $P(A(j)|\bar{A}(j-1), X) = P(A(j)|\bar{A}(j-1), \bar{L}_{\bar{A}(j-1)}(j))$, for $j = 0, ..., K$.
ETA $P(A = \bar{a}|X) > 0$ for all $\bar{a} \in \mathcal{A}$.

Factoring $P(O)$. $P(O)$ can be factored as

$$P(O) = (\prod_{j=0}^{K+1} P(L(j)|\bar{L}(j-1), \bar{A}(j-1))) * (\prod_{j=0}^{K} P(A(j)|\bar{A}(j-1), \bar{L}(j)))$$

By SRA, we have

$$\prod_{j=0}^{K} P(A(j)|\bar{A}(j-1), \bar{L}(j)) = P(\bar{A}|X) \equiv g_0(\bar{A}|X)$$

g_0 is called the **treatment assignment mechanism** (TAM). Recall that G-computation ignores the treatment assignment mechanism and uses the first factor of $P(O)$, which will be denoted Q_0 (see the Sept 8 notes). The distribution of O is thus determined by Q_0 and the treatment assignment mechanism g_0.

Recap of Point-Treatment Marginal Structural Models

In a point treatment study, the observed data $O = (W \equiv L(0), A(0), Y \equiv L(1))$, and thus is a special case of the type of studies we are considering here. To develop inverse probability of treatment weighted (IPTW) and doubly robust (DR) estimators for MSMs for time-dependent treatments, we will follow the same basic approach employed in developing the analogous estimators for point treatment MSMs. Therefore, a brief review of concepts and results from the point treatment case is provided; refer to the notes from Oct 11-13 for further details.

Recall that $\mathbf{L_0^2(O)}$ is the set of all functions of O with mean 0 and finite variance. If we define an inner product on $L_0^2(O)$ by $< d_1(O), d_2(O) > \equiv E[d_1(O)d_2(O)]$ (so the norm $\|d\| = Var^{1/2}(d)$), then $L_0^2(O)$ is a Hilbert space. That is, $L_0^2(O)$ is a complete inner product vector space. Completeness is a technical condition that guarantees the existence of projections: if $d \in L_0^2(O)$ and M is a closed subspace of $L_0^2(O)$, then there is a unique function $d^* \in M$ such that (i) $(d - d^*) \perp M$; ie, for all $h \in M, E[(d - d^*)h] = 0$, and (ii) $\|d - d^*\| < \|d - h\|$ for all $h \in M, h \neq d^*$. d^* is called the **projection** of d on M, denoted $\mathbf{\Pi(d|M)}$.

If f_θ is the density of a parametric model for O, then its **score**, denoted $\mathbf{S(O)}$, is

$$S(O) \equiv \frac{d}{d\theta} log f_\theta|_{\theta=\theta_0}$$

where $O \sim f_{\theta_0}$. Note that $E_0[S(O)] = 0$ and $S(O)$ has finite variance. Also note that in the following, it suffices to only consider scores for 1-dimensional parametric models.

Suppose $\mathcal{F}_{\beta,\eta}$ is a semiparametric model whose distributions are indexed by Euclidean parameter β and infinite-dimensional parameter η, where $O \sim F_{\beta_0,\eta_0}$; ie, the true distribution for O belongs to $\mathcal{F}_{\beta,\eta}$. Here, β_0 is the parameter of interest and η_0 is a nuisance parameter. Then $\mathcal{F}_\theta$ is a **parametric submodel** of $\mathcal{F}_{\beta,\eta}$ if (i) all distributions that belong to $\mathcal{F}_\theta$ also belong to $\mathcal{F}_{\beta,\eta}$, and (ii) $F_{\beta_0,\eta_0} \in \mathcal{F}_\theta$.

We are particularly interested in the scores of parametric submodels of $\mathcal{F}_{\beta_0,\eta}$. $\mathcal{F}_{\beta_0,\eta}$ is the set of distributions in $\mathcal{F}_{\beta,\eta}$ whose β parameter equals the true β_0 but whose ηs range over the allowable nuisance parameter values. The scores of parametric submodels of $\mathcal{F}_{\beta_0,\eta}$ are called **nuisance scores**, and the vector space generated by the nuisance scores from all parametric submodels of $\mathcal{F}_{\beta_0,\eta}$ is called the **nuisance tangent space**, denoted $\mathbf{T_{NUIS}}$. More exactly, the nuisance tangent space is the closure of the linear span of the nuisance scores from all parametric submodels of $\mathcal{F}_{\beta_0,\eta}$. This implies that T_{NUIS} is a closed subspace of $L_0^2(O)$.

An MSM is a semiparametric model, and one of its indices is infinite-dimensional nuisance parameter $g_{A|X}$, the treatment assignment mechanism. The nuisance tangent space for $g_{A|X}$ is denoted $\mathbf{T_{RA}}$. The "RA" subscript in T_{RA} indicates that we are only allowing TAMs that satisfy the RA assumption.

The class of IPTW estimating functions is $\{D_{h,IPTW}(O|\beta_0, g_0)$: all functions $h(A, V)\}$, where

$$D_{h,IPTW}(O|\beta_0, g_0) = \frac{h(A,V)}{g_0(A|X)}(Y - m(A, V|\beta_0))$$

The class of DR estimating functions is $\{D_{h,DR}(O|\beta_0, g_0, Q_0)$: all functions $h(A, V)\}$, where

$$D_{h,DR}(O|\beta_0, g_0, Q_0) = D_{h,IPTW}(O|\beta_0, g_0) - \Pi(D_{h,IPTW}(O|\beta_0, g_0)|T_{RA})$$

and $Q_0 = E[Y - m(A, V|\beta_0)|A, W]$.

Note that if $T_{RA}^\perp$ denotes the orthogonal complement of T_{RA} in $L_0^2(O)$, then

$$D_{h,IPTW}(O|\beta_0, g_0) - \Pi(D_{h,IPTW}(O|\beta_0, g_0)|T_{RA}) = \Pi(D_{h,IPTW}(O|\beta_0, g_0)|T_{RA}^\perp)$$

IPTW Estimating Functions for Time-Dependent MSMs

From the Oct 27 notes,

$$D_{h,IPTW}(O|\beta_0, g_0) = \frac{h(\bar{A}, V)}{g_0(\bar{A}|X)}(Y_{\bar{A}} - m(\bar{A}, V|\beta_0))$$

Doubly Robust Estimating Functions for Time-Dependent MSMs

Analogous to the point treatment case, the DR estimating function here is given by

$$\begin{aligned} D_{h,DR}(O|\beta_0, g_0, Q_0) &= \Pi(D_{h,IPTW}(O|\beta_0, g_0)|T_{SRA}^\perp) \\ &= D_{h,IPTW}(O|\beta_0, g_0) - \Pi(D_{h,IPTW}(O|\beta_0, g_0)|T_{SRA}) \end{aligned}$$

where T_{SRA} is the (more complicated) analogue of the point treatment T_{RA}.

Let's consider T_{SRA}. The distribution of O belongs to the semiparametric model $\mathcal{M}(g_{SRA}) \equiv \{P_{Q,g}$: $P_{Q,g} \equiv Qg$, g is a TAM that obeys SRA & $E[Y_{\bar{a}}|V] = m(\bar{a}, V|\beta_0)$, all $\bar{a}\}$ since $P(O) = Q_0 g_0 \equiv P_{Q_0,g_0}$. g is a nuisance parameter of $\mathcal{M}(g_{SRA})$, and we let $\mathbf{T_{SRA}}$ denote the nuisance tangent space for TAMs that obey SRA. That is, T_{SRA} is the subspace of $L_0^2(O)$ generated by scores of 1-dimensional parametric submodels that range over TAMs obeying SRA. Now, when we factored $P(O)$ above, we saw that an SRA-obeying TAM g factors as a product of "subTAMs", one per timepoint j, $j = 0, ..., K$:

$$g(\bar{A}|X) = \prod_{j=0}^{K} g_j(A(j)|\bar{A}(j-1), \bar{L}(j))$$

where

$$g_j(A(j)|\bar{A}(j-1), \bar{L}(j)) \equiv P(A(j)|\bar{A}(j-1), \bar{L}(j))$$

It is a useful fact that given this factorization,

$$T_{SRA} = T_{SRA,0} \oplus T_{SRA,1} \oplus ... \oplus T_{SRA,K}$$

where $\oplus$ denotes the direct sum of orthogonal subspaces and $\mathbf{T_{SRA,j}}$ is the nuisance tangent space for time point j "subTAMs" that obey SRA.

Another useful fact is that the projection onto a direct sum of subspaces equals the sum of the projections onto each subspace, implying

$$\Pi(D_{h,IPTW}(O|\beta_0, g_0)|T_{SRA}) = \sum_{j=0}^{K} \Pi(D_{h,IPTW}(O|\beta_0, g_0)|T_{SRA,j})$$

This result, combined with the following two facts about $T_{SRA,j}$, give us $\Pi(D_{h,IPTW}(O|\beta_0, g_0)|T_{SRA})$: (1) $T_{SRA,j} = \{f(A(j), \bar{A}(j-1), \bar{L}(j)) : f \in L_0^2(O) \& E[f(A(j), \bar{A}(j-1), \bar{L}(j))|\bar{A}(j-1), \bar{L}(j)] = 0\}$. (2) The projection of D onto $T_{SRA,j}$, $D \in L_0^2(O)$, is given by

$$\begin{aligned}
\Pi[D|T_{SRA,j}] &= E_{Q_0,g_0}[D|\bar{A}(j), \bar{L}(j)] - E_{Q_0,g_0}[D|\bar{A}(j-1), \bar{L}(j)] \\
&= E_{Q_0,g_0}[D|\bar{A}(j), \bar{L}(j)] - \sum_{a(j)} E_{Q_0,g_0}[D|(a(j), \bar{A}(j-1)), \bar{L}(j)] * g_j(A(j) = a(j)|\bar{A}(j-1), \bar{L}(j))
\end{aligned}$$

To see that (2) follows from (1), see the discussion of the analogous two facts for the point treatment case in the Oct 11-13 notes.

In sum, then, we get

$$\begin{aligned}
D_{h,DR}(O|\beta_0, g_0, Q_0) &= \Pi(D_{h,IPTW}(O|\beta_0, g_0)|T_{SRA}^{\perp}) \\
&= D_{h,IPTW}(O|\beta_0, g_0) - \Pi(D_{h,IPTW}(O|\beta_0, g_0)|T_{SRA}) \\
&= D_{h,IPTW}(O|\beta_0, g_0) - \sum_{j=0}^{K} \Pi(D_{h,IPTW}(O|\beta_0, g_0)|T_{SRA,j}) \\
&= D_{h,IPTW}(O|\beta_0, g_0) - \\
&\quad \sum_{j=0}^{K} (E_{Q_0,g_0}[D_{h,IPTW}(O|\beta_0, g_0)|\bar{A}(j), \bar{L}(j)] \\
&\quad - E_{Q_0,g_0}[D_{h,IPTW}(O|\beta_0, g_0)|\bar{A}(j-1)), \bar{L}(j)]) \\
&= D_{h,IPTW}(O|\beta_0, g_0) - \\
&\quad \sum_{j=0}^{K} (E_{Q_0,g_0}[D_{h,IPTW}(O|\beta_0, g_0)|\bar{A}(j), \bar{L}(j)] - \\
&\quad \sum_{a(j)} E_{Q_0,g_0}[D_{h,IPTW}(O|\beta_0, g_0)|(a(j), \bar{A}(j-1)), \bar{L}(j)] * g_j(A(j) = a(j)|\bar{A}(j-1), \bar{L}(j)))
\end{aligned}$$

Implementing Doubly Robust Estimators

Let O_i denote the observed data from subject i, $i = 1, ..., n$. Estimate $\hat{\beta}_n$ of β_0 is the value of β that gives

$$\sum_{i=1}^{n} D_{h,DR}(O_i|\hat{\beta}_n, g_0, Q_0) = 0$$

An iterative procedure like Newton-Raphson will be used to find $\hat{\beta}_n$. More specifically, proceed via the following steps:

First, use MLE to estimate g_0 and Q_0 by g_n and Q_n, respectively.

Second, find $\beta(Q_n)$, the G-computation estimate of β based on Q_n.

Third, estimate $E_{Q_n,g_n}[D_{h,IPTW}(O_i|\beta(Q_n),g_n)|\bar{A}_i(j),\bar{L}_i(j)]$ for each subject i, $i=1,...,n$ and time point j, $j=0,...,K$. To do this, do the following a large number (say, $10,000$) times: (a) applying Q_n to $(\bar{A}_i(j),\bar{L}_i(j))$, generate a value $L^{MC}(j+1)$; (b) applying g_n to $(\bar{A}_i(j),(L_i^{MC}(j+1),\bar{L}_i(j)))$, generate a value $A_i^{MC}(j+1)$; continue alternately using Q_n and g_n until Y_i^{MC} is generated; and (c) compute

$$D_i^{MC} \equiv \frac{h(\bar{A}_i^{MC},V_i)}{g_n(\bar{A}^{MC}|X)}(Y_i^{MC}-m(\bar{A}^{MC},V|\beta(Q_n))$$

Then the mean of the $10,000$ Monte Carlo values of D_i^{MC} is the desired estimate.

Fourth, following an analogous procedure, estimate $E_{Q_n,g_n}[D_{h,IPTW}(O_i|\beta(Q_n),g_n)|\bar{A}_i(j-1),\bar{L}_i(j)]$ for all i and j.

Fifth, compute an estimate of $\Pi(D_{h,IPTW}(O_i|\beta_(Q_n),g_n)|T_{SRA})$ for each subject i using the results of the third and fourth steps. Call this $\hat{\Pi}_i$.

Finally, apply Newton-Raphson to solve

$$\sum_{i=1}^{n}\frac{h(\bar{A}_i,V_i)}{g_n(\bar{A}_i|X)}(Y_{\bar{A}_i}-m(\bar{A}_i,V_i|\beta_n))-\hat{\Pi}_i=0$$

Properties of Doubly Robust Estimators

Because $D_{h,DR}(O|\beta_0,g_0,Q_0)$ is an estimating function, we know that
$E_{Q_0,g_0}[D_{h,DR}(O|\beta_0,g_0,Q_0)]=0$.

Suppose $(Q_1,g_1)\neq(Q_0,g_0)$. Under what conditions does
$E_{Q_0,g_0}[D_{h,DR}(O|\beta_0,g_1,Q_1)]=0$?

This expectation equals zero if either

(a) $g_1=g_0$ and extended ETA holds: $max_{\bar{a}}\frac{g_1(\bar{a}|V)}{g_1(\bar{a}|X)}h(\bar{a},V)<\infty$ ae

or

(b) $Q_1=Q_0$ and extended ETA holds. To see (b), suppose that $g_1\neq g_0$ but condition (b) holds. If g_1 were the true TAM, then

$d(g_0,g_1)\equiv\frac{(g_0-g_1)(\bar{A}|X)}{g_1(\bar{A}|X)}$ is a nuisance score in T_{SRA} (where T_{SRA} is defined with respect to the (temporarily assumed) true TAM g_1). That $d(g_0,g_1)$ is a nuisance score follows from

$$d(g_0,g_1)=\frac{d}{d\varepsilon}(1-\varepsilon d(g_0,g_1))g_1|_{\varepsilon=0}$$

and

$$E_{Q_0,g_1}[d(g_0,g_1)|X]=E_{Q_0}E_{g_1}[d(g_0,g_1)|X]=0$$

See the Oct 11-13 notes for a related discussion.

Since $d(g_0,g_1)\in T_{SRA}$ and $D_{h,DR}(O|\beta_0,g_1,Q_0)=\Pi(D_{h,IPTW}(O|\beta_0,g_0)|T_{SRA}^{\perp})\in T_{SRA}^{\perp}$ we have $d(g_0,g_1)\perp D_{h,DR}(O|\beta_0,g_1,Q_0)$. Then

$$\begin{aligned}
0 &= E_{Q_0,g_1}[D_{h,DR}(O|\beta_0,g_1,Q_0)*d(g_0,g_1)] \\
&= E_{Q_0}\left[\sum_{\bar{a}}D_{h,DR}(O|\beta_0,g_1,Q_0)\frac{(g_0-g_1)(\bar{A}|X)}{g_1(\bar{A}|X)}g_1(\bar{A}|X)\right] \\
&= E_{Q_0,g_0}D_{h,DR}(O|\beta_0,g_1,Q_0)-E_{Q_0,g_1}D_{h,DR}(O|\beta_0,g_1,Q_0) \\
&= E_{Q_0,g_0}D_{h,DR}(O|\beta_0,g_1,Q_0)
\end{aligned}$$

as desired.

Finally, a note about another important property of DR estimators. We have proceeded above under the assumption that the true TAM was unknown. But in some studies, such as sequentially randomized clinical trials (eg, Hernan, Brumback, & Robins (2002, pp. 1698-1703)), we know the true TAM. In such a case,

can we do better than the DR estimator derived above? As in the point treatment case, the answer is no. This follows from the fact that, under the model with the true TAM g_0 known, the class of all estimating functions is

$\{D_{h,IPTW}(O|\beta_0, g_0) - s$: all $h; s \in T_{SRA}\}$

Now, the most efficient estimating function has the smallest variance. But, by the properties of projections discussed above, for $s \in T_{SRA}$, where $s \neq \Pi[D_{h,IPTW}(O|\beta_0, g_0)|T_{SRA}]$,

$$\begin{aligned} Var[D_{h,DR}(O|\beta_0, g_0, Q_0)] &= Var[D_{h,IPTW}(O|\beta_0, g_0) - \Pi[D_{h,IPTW}(O|\beta_0, g_0)|T_{SRA}]] \\ &= \|D_{h,IPTW}(O|\beta_0, g_0) - \Pi[D_{h,IPTW}(O|\beta_0, g_0)|T_{SRA}]\|^2 \\ &< \|D_{h,IPTW}(O|\beta_0, g_0) - s\|^2 \\ &= Var[D_{h,IPTW}(O|\beta_0, g_0) - s] \end{aligned}$$

Hence, the DR estimating function is still the one to use. In fact, it can be shown that when g_0 is known, it is still better to use its estimate g_n.

Censored Longitudinal Data and Causality: Notes for 10/27/04

Marginal Structural Models for Time-Dependent Treatment

As usual, we begin by specifying the observed data structure:
$O = (L(0), A(0), L(1), A(1), ..., L(K), A(k), Y(k+1))$
Note that, in our discussion today, we focus on outcome at a fixed point in time (vs. e.g., a survival time). However, see earlier lectures on formulating longitudinal data structures in generality (including survival times and data subject to censoring.) A brief review: We write the full data as $X = (X_{\bar{a}} : \bar{a})$ where $\bar{a}$ denotes an action history, which may include a censoring component as well as a treatment component. The full data consist of action-specific processes for every action regime. For example, if $\bar{a}_1$ is the treatment component, and $\bar{a}_2$ is the censoring component, then$X_{\bar{a}}(t) \equiv X_{\bar{a}}(min(t, c(a_2)))$ where $c(a_2)$ denotes the censoring time. The observed data are then denoted $O = (\bar{A}, \bar{X}_{\bar{A}})$. In general, marginal structural models are models of the marginal distribution of $X_{\bar{a}}$.

Returning to our example today, with no censoring and the outcome defined as the value of a covariate(s) at one point in time,
$\bar{a} = (a(0), a(1), ..., a(K))$ denotes the treatment process over time, composed of treatment at each time point ($j = 1, ..., K$), and $\bar{L}_{\bar{a}}(j)$ denotes the counterfactual treatment regime-specific process through time j. Note that when treatment is time-dependent it can change over time as a result of prior treatment and treatment-induced changes in the treatment-regime specific counterfactual processes. What does this mean? For example, prior treatment can affect the value of a covariate, which in turn can affect subsequent treatment. In this case, standard methods of analysis (i.e. multivariable regression) may lose their causal interpretation. For example, we might perform a standard regression of outcome on observed treatment and covariate history: $E[Y|\bar{A}(K), \bar{L}_(K)]$. One might then try to estimate the effect of treatment $\bar{a}$ as compared to no treatment throughout the study ($\bar{a} = 0$) as
$E[Y|\bar{A}(k) = \bar{a}(k), \bar{L}_{\bar{a}}(K)] - E[Y|\bar{A}(k) = 0, \bar{L}_{\bar{a}=0}(K)]$.

However, such an analysis can be difficult to interpret causally. If $\bar{L}$ is in fact affected by previous treatment, as is often the case, then $\bar{L}_{\bar{a}}$ and $\bar{L}_{\bar{a}=0}$ will not be comparable, and so when we try to compare the expectation of the outcome conditioning on $\bar{L}$, as above, we are comparing non-comparable quantities. This concept is familiar to many as the rule that one should not condition on covariates affected by the exposure (or treatment) of interest. However, some of the covariates L may also be confounders; i.e., they may predict both subsequent treatment assignment and also, independently predict outcome. If we do not somehow adjust for these confounders in our analysis, we will have a biased estimate of the causal effect of treatment. Covariates such as these, that are both affected by previous treatment, and themselves affect subsequent treatment and outcome, are called time-dependent confounders. Standard regression analysis does not give us an adequate tool to address them. This is one motivation for marginal structural models.

Observed Data In our example, the observed data are:
$O = (L(0), A(0), L(1), A(1), ..., L(K), A(K), Y(K+1))$
Full Data The full data are : $X = (\bar{L}_{\bar{a}} : \bar{a})$.
Temporal Ordering Assumption:

We assume that $L_{\bar{a}}(j) = L_{\bar{a}(j-1)}(j)$, or in other words, a covariate at time j is only affected by treatment that occurs before time j.

Consistency Assumption:

Under the consistency assumption, we can write the observed data as the counterfactual process that would have been observed under the observed treatment history: $O = (\bar{A}, \bar{L}_{\bar{A}}(K+1))$.

Under the Temporal Ordering Assumption, the observed data can further denoted, in chronological order:
$O = (L(0), A(0), L_{\bar{A}(0)}(1), A(1), L_{\bar{A}(1)}(2), A(2), ..., L_{\bar{A}(K-1)}(K), A(K), Y_{\bar{A}(K)} = L_{\bar{A}(K)}(K+1))$

Parameter of Interest

Our parameter of interest can be any parameter of the distribution of the treatment-regime-specific covariate process, $P_{L_{\bar{a}}}$. For our example, we define the parameter of interest to be $E[Y_a|V]$ for a $V \subset L(0)$.

Note: Traditional Marginal structural models are restricted to examining causal effects within subgroups defined by baseline covariates. We will talk later about a variation on marginal structural models (History Adjusted Marginal Structural Models) that allow estimation of causal effects conditional on time-varying covariates (such as the effect of future treatment, given a time-varying covariate's history up till that time point).

To make our parameter of interest identifiable from the observed data, we need an additional assumption.

Sequential Randomization Assumption (SRA):

$P(A(j) = a(j)|\bar{A}(j-1), X) = P(A(j) = a(j)|\bar{A}(j-1), \bar{L}_{\bar{A}(j-1)}(j)), j = 0, ..., K$ In other words, we assume that, at each time point, the probability of being assigned a specific treatment only depends on the observed past up till that time point.

The distribution of the observed data can be written: $O \sim P_{F_{X_0}, g_0}$ where $g_0(\bar{a}|X) \equiv P(\bar{A} = \bar{a}|X)$ denotes the true treatment mechanism.

We can write the treatment mechanism as
$g_0(\bar{a}|X) = \prod_{j=0}^{K} g_0(a(j)|\bar{a}(j-1), X)$

Further, under the SRA:
$g_0(\bar{a}|X) = \prod_{j=0}^{K} g_0(a(j)|\bar{a}(j-1), \bar{L}_{\bar{a}(j-1)}(j))$

The density of the observed data is:
$P(O) = \prod_{j=0}^{K+1} P(L(j)|\bar{L}(j-1), \bar{A}(j-1)) \prod_{j=0}^{K} P(A(j)|\bar{A}(j-1), \bar{L}(j))$

Under the SRA and CA, the first term of this density can be written using the G-computation formula:
$\prod_{j=0}^{K+1} P(L(j)|\bar{L}(j-1), \bar{A}(j-1) = P_{L_{\bar{a}}}|_{\bar{a}=\bar{A}(K+1)}$ (See previous lectures for this proof.)

Under the SRA, the second term of the density can be written:
$\prod_{j=0}^{K} P(A(j)|\bar{A}(j-1), \bar{L}(j)) = g_0(\bar{A}|X)$

Marginal Structural Model We assume a model for our parameter of interest (which, recall, is a parameter of the marginal distribution of the treatment-specific counterfactual outcome): $E[Y_{\bar{a}}|V] = m(\bar{a}, V|\beta_0)$, where β_0 refers to the true parameter (we use subscript 0 throughout to refer to the truth). For example, we might assume the following model:
$m(\bar{a}, V|\beta_0) = \beta(0) + \beta(1) \sum_{j=0}^{K} a(j) + \beta(2)V + \beta(3)(\sum_{j=0}^{K} a(j))V$ (Note: $\beta(0)$ refers to the first element of β, $\beta(1)$ to the second element, etc.)

Alternatively, we might also want to include, for example, a quadratic term in our model:
$m(\bar{a}, V|\beta_0) = \beta(0) + \beta(1) \sum_{j=0}^{K} a(j) + \beta(2)V + \beta(3)(\sum_{j=0}^{K} a(j))V + \beta(4)(\sum_{j=0}^{k} a(j))^2$

(Side Note: In choosing your model, think about both your subject matter and your objective. Do you expect your curve to be monotone? Are you interested in finding an optimum treatment? etc...)

Estimation using censored data: General Approach

A general result for any censored data structure: The class of estimating functions of the observed data is defined by the class of functions of the observed data, which, if you take their conditional expectation given the full data (or in other words, integrate with respect to the censoring mechanism), you get back the class of all estimating functions of the full data. To make this class of estimating functions maximally orthogonal to the nuisance parameter, we orthogonolize it with respect to the Hilbert Space generated by the scores of 1-dimensional submodels varying only the nuisance parameter (i.e. the nuisance tangent spece, T_{NUIS}. To do this, we subtract off the projection of the class of estimating functions onto the nuisance tangent space.

Carrying out this procedure to estimate β

Full data: $X : (L_{\bar{a}} : \bar{a})$

What would the estimating functions be in the full data?
For every $\bar{a}$, a regression of $L_{\bar{a}}$ on $\bar{a}, V$. So, for a fixed $\bar{a}$, the class of all estimating functions (or nuisance tangent space, $T_{NUIS}^{\perp,FULL}$) would be $h(\bar{a},V)(Y - m(\bar{a},V|\beta_0))$.
We take advantage of the result that, if we know the $T_{NUIS}^{\perp}$ for each $\bar{a}$-specific model, the $T_{NUIS}^{\perp}$ for the intersection of the $\bar{a}$-specific models is just the sum of the $T_{NUIS}^{\perp}$ for each $\bar{a}$-specific model.
So, the class of all estimating functions for the full data is

$$T_{NUIS}^{\perp,FULL} = \sum_{\bar{a}} h(\bar{a},V)(Y_{\bar{a}} - m(\bar{a},V|\beta_0)) \tag{15}$$

(Note: This is just the class of estimating functions for the repeated measures regression model)
We want the functions of the observed data that, if we take the conditional expectation given the full data, we get back the class of all estimating functions for the full data (15).
IPTW Estimating Function
$D_{h,IPTW}(O|g_0,\beta_0) = \frac{h(\bar{A},V)}{g(\bar{A}|X)}(Y_{\bar{A}} - m(\bar{A},V|\beta_0))$
Result: If

$$\max_{\bar{a}\in\mathcal{A}} \frac{h(\bar{A},V)}{g(\bar{A}|X)} < \infty \tag{16}$$

then, $E[D_{h,IPTW}(O|g_0,\beta_0)|X] = \sum_{\bar{a}} h(\bar{a},V)(Y_{\bar{a}} - m(\bar{a},V|\beta_0))$ (in other words, by taking the expectation conditional on the full data, we get back the full data estimating function), and $E[D_{h,IPTW}(O|g_0,\beta_0)] = 0$ (in other words, the IPTW estimating function is unbiased at the true data generating distribution).
Proof:

$$\begin{aligned}
E[E[D_{h,IPTW}(O|g_0,\beta_0)|X]] &= E\sum_{\bar{a}} \frac{h(\bar{a},V)}{g(\bar{a}|X)}(Y_{\bar{a}} - m(\bar{a},V|\beta_0)g(\bar{a}|X) \\
&=^{ETA(16)} E\sum_{\bar{a}} h(\bar{a},V)(Y_{\bar{a}} - m(\bar{a},V|\beta_0) \\
&= \sum_{\bar{a}} E[E[(h|a,V)(Y_{\bar{a}} - m(a,v|\beta_0))|V]] \\
&= E[\sum_{\bar{a}} h(\bar{a},V)(E(Y_{\bar{a}}|V) - m(\bar{a},V|\beta_0))] \\
&= 0
\end{aligned}$$

(17)

Note: (16) is a specific version of the ETA assumption. More generally, the ETA assumption states that any possible treatment at each time point, given past treatment and covariate history, needs to have a positive probability of occurring.
Side Note: Choice of an estimator (IPTW, DR-IPTW or G-comp) must depend on the problem at hand. For example, if you have a very high-dimensional treatment or treatment mechanism, the IPTW and DR-IPTW may not give you any extra-robustness (due to misspecification of the treatment mechanism). In this case, you may in fact be better off with the MLE-based G-comp estimator (but remember, this is different from the standard MLE approach in that you integrate out those covariates which are not part of your parameter of interest).
How to Implement
$h^*(\bar{A},V) = g(\bar{A}|V)\frac{\partial}{\partial\beta}m(\bar{A},V|\beta)$ is a good choice of $h(\bar{A},V)$ for linear regression.
$h^*(\bar{A},V) = g(\bar{A}|V)\frac{\frac{\partial}{\partial\beta}m(\bar{A},V|\beta)}{var(Y-m(\bar{A},V|\beta)|V)}$ is a good choice of $h(\bar{A},V)$ for logistic or Poisson regression, where, if $Y \in 0,1$ $var[y - m(\bar{A},V|\beta)|V] = \sigma^2(\bar{A},V) = m(A,V|\beta_0)(1 - m(A,V|\beta_0))$
These choices of $h(\bar{A},V)$ correspond to the weighted regressions performed by standard software.
The solution $\beta_{n,IPTW}$ of $0 = \sum_{i=1}^n D_h(O_i|g_n,\beta)$ equals
$\beta_{n,IPTW} = argmin_\beta \sum_{i=1}^n (Y_i - m(\bar{A}_i,V_i|\beta))^2 \frac{g_n(\bar{A}_i|V_i)}{g_n(A_i|X_i)\sigma^2(A_i,V_i)}$ or in other words, the solution to ordinary

weighted least squares. So we can implement this estimator using standard software and supplying a n-dimensional vector of weights, with the weight for each subject equal to $w_i = \frac{g_n(\bar{A}_i|V_i)}{g_n(A_i|X_i)}$
Note: The ETA makes this a dangerous estimator. At the very least, we must inspect the ETA (via the bootstrap technique presented in an earlier lecture). If there is enough experimentation in our data that we can estimate our parameter of interest without extrapolation, then this is a good estimator. In contrast, even if the ETA is violated, the DR-IPTW estimator will extrapolate to sparse areas of your data as well as MLE.
How to estimate g_n
Maximum likelihood, based on a model for the treatment:
$g_n = argmax_\theta \prod_{i=1}^n \prod_{j=1}^K g_\theta(A_i(j)|\bar{A}_i(j-1), \bar{L}_i(j))$, $g_n \equiv g_{\theta_n}$ For example, we could use a logistic regression model of $A(j)$ given past covariates and treatment. We could assume a single model for all time points and just run a logistic regression on the pooled data set, with each subject contributing k lines of data. Alternatively, we could assume a single model for some subset of time points, or have a separate model for each time point. You should just pool those time-points where a common model makes sense. For example, it may not make sense to model $A(0)$ with all the rest of the time points, since at baseline, treatment can't depend on previous measured covariates. Estimate g_n as non-parametrically as possible, using cross-validation. The more non-parametrically you estimate g_n, the more asymptotically efficient your estimator will be.

11/17/2004 Notes

History-adjusted MSMs

Let $\widetilde{T} = \min(T, C)$ and $\Delta = I(T \leq C)$, where T represents a time until death and C represents a censoring time. A general right-censored data structure for following a patient in a longitudinal study is $O = (\overline{A}_1(\widetilde{T}), \overline{L}(\widetilde{T}))$, where A_1 represents a treatment process, and L represents a covariate process. Here $L(t)$ is assumed to include $I(T \leq t)$ (an indicator of whether the patient has died by time t), and a time-dependent outcome $Y(t)$. We can define counterfactuals on this data structure, and fit marginal structural models, as will be discussed in subsequent lectures.

MSMs with Missing Data in Point Treatment

Using our common notation for point treatment studies, suppose that we are interested in modelling $E[Y_a|V]$ where $V \subset W$, the baseline covariates. Complications arise when there is missing data. Assume that V is always observed, but that the rest of W can occasionally be missing. So for $\Delta = I(W \text{ is observed})$, $O = (\Delta, \Delta W, V, A, Y)$ is the observed data. There are two ways to proceed.

Method I. The first method is to simply redefine the baseline covariates as $W' = (W\Delta, \Delta, V)$, and fit a marginal structural model (as in previous lectures) to the observed data model $O = (W', A, Y)$. In this case, the SRA assumption becomes $P(A = a|W', (Y_a : a \in \mathcal{A})) = P(A = a|W')$, which is different from the traditional SRA assumption.

Method II. The second method is to treat the unobserved (W, A, Y) as if it were the full data structure, and apply the general techniques of van der Laan and Robins for mapping full data estimating functions into observed data estimating functions in coarsening at random models. The coarsening at random assumption is here that $P(\Delta = 1|X) = P(\Delta = 1|V, A, Y)$, where X is the counterfactual process $(W, (Y_a : a \in \mathcal{A}))$. When making the usual SRA assumption that $P(A = a|X) = P(A = a|W)$, we saw in previous lectures how to derive all estimating functions $(D_h(W, A, Y) : h)$ for the "full" data model (observing (W,A,Y)).

The general methodology of van der Laan tells us that under the CAR assumption $P(\Delta = 1|X) = P(\Delta = 1|V, A, Y)$, the class of "observed" data estimating functions is given by $\{D_h \frac{\Delta}{P(\Delta=1|A,V,Y)} - \Pi(D_h \frac{\Delta}{P(\Delta=1|A,V,Y)}|T_{CAR}) : h\}$, where T_{CAR} is the Hilbert space in $L_0^2(O)$ containing all scores obtained from varying the missingness mechanism. Considering one-dimensional fluctuations through $P(\Delta = 1|V, A, Y)$ gives that $T_{CAR} = \{\phi(\Delta, V, A, Y) : E[\phi|V, A, Y] = 0, E\phi^2 < \infty\}$, and hence that $\Pi(U(O)|T_{CAR}) = E[U|\Delta, V, A, Y] - E[U|V, A, Y]$. Therefore, the class of all estimating functions in the "observed" data model is given by,

$\{U_h(O) - E[U_h(O)|\Delta, V, A, Y] + E[U_h(O)|V, A, Y] : U_h(O) = D_h(W, A, Y)\frac{\Delta}{P(\Delta=1|A,V,Y)}\}$. From these estimating functions, we can proceed as usual in fitting marginal structural models.

Censored Longitudinal Data and Causality: Notes for 11/22/04

Marginal Structural Models for Survival Time Outcomes

For Reference, see Mark and Jamie's book, section 6.4.

Let $L_{\bar{a}_1}(t) = (Y_{\bar{a}_1}(t), W_{\bar{a}_1}(t))$, indexed by treatment $\bar{a}_1$, where $L_{\bar{a}_1}(t) = L_{\bar{a}_1}(min(t, T_{\bar{a}_1}))$. Let $Y_{\bar{a}_1} = I(T_{\bar{a}_1} \leq t)$, where $T_{\bar{a}_1}$ is a treatment-specific survival time. Note: we treat time as discrete here, so $t = 0, 1, ...$

However, we have censoring, so that for some individuals we do not observe their survival time. Let $A_2(t) = I(C \leq t)$, where $C = \infty$ if $T_{\bar{A}_1} < C$. (Note we make this convention so that C is always observed, even if an individual is not censored.). We can define the full data in terms of action-specific processes indexed by treatment and censoring actions: $L_{\bar{a}_1,\bar{a}_2}(t) = L_{\bar{a}_1}(min(t, c(\bar{a}_2)))$ where $c(\bar{a}_2)$ is the time at which a_2 jumps from zero to one. Let $A(t) = (A_1(min(t, C, T)), A_2(t))$. We can then write the Full data as: $X = (L_{\bar{a}} : \bar{a} \in \mathcal{A})$.

Temporal Ordering assumption: We assume $L_{\bar{a}_1}(t) = L_{\bar{a}_1(t-1)}(t)$

Consistency Assumption: The observed data can be written chronologically as: $(L(0), A(0), L(1), A(1), ..., L(min(T-1), C), A(min(T-1), C), L(min(T, C)))$. Under the consistency assumption, we can write the observed data as: $O = (\bar{A}, L_{\bar{A}}) \sim P_{F_X, g(\bar{a}|X)}$ where $g(\bar{a}|X) = P(\bar{A} = \bar{a}|X)$

Sequential Randomization Assumption: We assume the SRA on the action process $A(t) = (A_1(min(t, C)), A_2(t))$:

$$\begin{aligned} g(\bar{A}|X) &= \prod_t g(A(t)|A(t-1), X) \\ &= \prod_{t=0}^{min((T-1),C)} g(A(t)|\bar{A}(t-1), \bar{L}_{\bar{A}(T-1)}(t)) \text{ (Under the SRA)} \\ &= \prod_{t=0}^{min(T-1,C)} g_1(A_1(t)|A_2(t), \bar{A}(t-1), \bar{L}(t)) \prod_{t=0}^{min(T-1,C)} g_2(A_2(t)|\bar{A}(t-1), \bar{L}(t)) \end{aligned}$$

Note that the last equality represents factorization of the action mechanism, into the first component (treatment) and the second (censoring). To shorten our notation, let $(A_2(t), \bar{A}(t-1), \bar{L}(t)) = \mathcal{F}_1(t)$ and let $(\bar{A}(t-1), \bar{L}(t)) = \mathcal{F}_2(t)$.

Parameter of interest: In this case the full data consist of counterfactual treatment-specific survival times under possible treatment regimes. We might be interested in, e.g., the treatment specific hazard. Recall that $Y_{\bar{a}_1}(t) = I(T_{\bar{a}_1} \leq t)$. So we might be interested in:

$$E(dY_{\bar{a}_1}(t)|Y_{\bar{a}_1}(t-1), V) = I(T_{\bar{a}_1} \geq t)P(T_{\bar{a}_1} = t|T_{\bar{a}_1} \geq t, V) \tag{18}$$

We might then assume a model for $P(T_{\bar{a}_1} = t|T_{\bar{a}_1} \geq t, V) = m(\bar{a}_1(t-1), t, V|\beta)$. For example:

$$m(\bar{a}_1(t-1), t, V|\beta) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 sum(\bar{a}_1(t-1)) + \beta_2 t + \beta_3 V)}} \tag{19}$$

Note that the hazard is a specific case of the intensity of a counting process, in which we are interested in the expectation of a jump in a counting process ($dN(t)$) given a past. While a survival function can only jump once, we can imagine many counting processes which could jump mulitple times (eg number of hospitalizations, etc...).

Estimation: Recall that the density of the observed data can be factorized as:

$$P(O) = \prod_{j=0}^{min(T,C)} p(L(j)|\bar{L}(j-1), \bar{A}(j-1)) g(\bar{A}|X) \tag{20}$$

The first term of this factorization gives us the G-comp formula. If we wished to use a likelihood-based estimation approach, we could evaluate this expression at $C > j$ and $\bar{A}_1$ equal to the treatment of interest to get $L_{\bar{a}_1,\bar{a}_2=0}$.

Estimating Function Approach: Alternatively, we can use the estimating function-based approach. We begin by asking what is the class of estimating functions for the type of model that defines our parameter

of interest (in this case an intensity model) in the full data world?
General Result: Say we are interested in
$E(dN(t)|\mathcal{F}(t)) = Y(t)P(dN(t) = 1|\mathcal{F}(t), Y(t) = 1)$, where

- $\mathcal{F}(t)$ is the past, including the past of the counting process $\bar{N}(t-1)$,
- $Y(t) = I(N(t)$at risk of jumping at time t), and
- $P(dN(t) = 1|\mathcal{F}(t), Y(t) = 1)$ is either a probability (if time is discrete) or an intensity (if time is continuous)

$E(dN(t)|\mathcal{F}(t))$ can be modelled accordingly, as, e.g. $Y(t)m(t, \mathcal{F}(t)|\beta)$. Then,

$$T^{\perp}_{NUIS} = h(t, \mathcal{F}(t))(dN(t) - Y(t)m(t, \mathcal{F}(t)|\beta)) \tag{21}$$

for each time point t
However, we are interested in the intersection of these models over all time points, so the nuisance tangent space is just the sum over time points of the tangent space for each time point:

$$T^{\perp}_{NUIS} = \sum_t h(t, \mathcal{F}(t))(dN(t) - E(dN(t)|\mathcal{F}(t)) \tag{22}$$

Because when $Y(t)$ goes to zero, $E(dN(t)|\mathcal{F}(t)) = 0$, (22) can also be written as:

$$\sum_t Y(t)h(t, \mathcal{F}(t))(dN(t) - E(dN(t)|\mathcal{F}(t))) \tag{23}$$

The equivalent for continuous time is:

$$\int h(t, \mathcal{F}(t))dM(t) \tag{24}$$

where $dM(t)$ is the martingale for $dN(t) - E(dN(t)|\mathcal{F}(t))$
Result applied to our parameter of interest: So, the class of all estimating functions for any given (fixed) $\bar{a}_1$ is:

$$T^{FULL,\perp}_{NUIS} = \sum_t^{T_{\bar{a}_1}} h(t, \mathcal{F}(t))(dY_{\bar{a}_1}(t) - E(dY_{\bar{a}_1}(t)|\bar{Y}_{\bar{a}_1}(t-1), V) \tag{25}$$

where we can replace $\mathcal{F}(t)$ with V because we know $Y(t-1) = 0$.
However, our MSM is really an intersection over all $\bar{a}_1 \in \mathcal{A}_1$, so

$$T^{FULL,\perp}_{NUIS} = \sum_{\bar{a}_1 \in \mathcal{A}_1} \sum_{t=0}^{T_{\bar{a}_1}} h(t, \bar{a}_1(t-1), V)(dY_{\bar{a}_1}(t) - E(dY_{\bar{a}_1}(t)|\bar{Y}_{\bar{a}_1}(t-1), V) \tag{26}$$

A standard choice of h would be

$$h^* = \frac{\frac{\partial}{\partial \beta} m(\bar{a}_1(t-1), t, V|\beta)}{m(1-m)} \tag{27}$$

Class of estimating functions for our parameter of interest using observed data: We are interested in treatment specific survival times in the absence of censoring. So we can write our MSM as:

$$E(dY_{\bar{a}_1,0}(t)|\bar{Y}_{\bar{a}_1,0}(t-1), V) = I(Y_{\bar{a}_1,0}(t-1) = 0)P(T_{\bar{a}_1,0} = t|T_{\bar{a}_1,0} \geq t, V) \tag{28}$$

In addition, because we are only interested in our outcomes in the absence of censoring we are only looking at the censoring mechanism where $\bar{a}_2 = 0$ We can now write our Inverse Probability of Action Weighted estimating function for this MSM as a function of the observed data:

$$(\sum_{t=0}^{T} h(t, \bar{A}_1(t-1), V)(I(T = t) - I(T \geq t)m(\bar{A}_1(t-1), t, V|\beta)) \frac{I(\bar{A}_2 = 0)}{g(\bar{A}_1, \bar{A}_2 = 0|X)} \tag{29}$$

where we note that $I(\bar{A}_2 = 0) = I(C > T)$ and we note that $I(T = t) - I(T \geq t)m(\bar{A}_1(t-1), t, V|\beta)$ is a martingale (in other words, the change in a counting process minus the expectation of a change given the past: $dY(t) - E(dY(t)|past)$.
We now have a class of estimating functions that is indeed a function of the observed data. If we take the conditional expectation of this class of estimating functions given the full data, we get back the class of

estimating functions for the full data.
In (29) we only use those individuals who are not censored ($\bar{A}_2 = 0$ throughout). Alternatively, we can write the IPAW estimating function as:

$$\sum_{t=0}^{T}(h(t,\bar{A}_1(t-1),V)(I(T=t)-I(T\geq t)m(\bar{A}_1(t-1),t,V|\beta))\frac{I(\bar{A}_2(t)=0)}{g(\bar{A}_1(t-1),\bar{A}_2(t)=0|X)}) \tag{30}$$

where now each individual get a weight for every time point until they are censored (so we use subjects even if they are censored before time T).

If we make our standard choice $h^* = g(\bar{A}_1(t-1),\bar{A}_2(t)=0|V)\frac{\frac{\partial}{\partial\beta}(m(\bar{A}(t-1),t,V))}{m(1-m)}$, we can estimate our parameter of interest using standard logistic regression of the binary outcome of death or not at each time point on our model, with each subject contributing one weighted line of data for each time point until being censored. Using the above choice of h^*, the weights are equal to:

$$\frac{g(\bar{A}_1(t-1),\bar{A}_2(t)=0|V)}{g(\bar{A}_1(t-1),\bar{A}_2(t)=0|X)} \tag{31}$$

We then run a standard weighted logistic regression on our pooled sample. If there is no confounding beyond V, this will be equivalent to doing a standard IPCW analysis.

Estimating the treatment and censoring mechanisms Factorization of the action mechanism means we can estimate g_1 and g_2 separately (ie with MLE):

$$\begin{aligned} g_{1n} &= \arg\max_{g_1\in G_1}\prod_{i=1}^{n}\prod_{t=0}^{min(T_i-1,C_i)} g_1(A_{1i}(t)|\mathcal{F}_{1i}(t)) \\ g_{2n} &= \arg\max_{g_2\in G_2}\prod_{i=1}^{n}\prod_{t=0}^{min(T_i-1,C_i)} g_2(A_{2i}(t)|\mathcal{F}_{2i}(t)) \end{aligned}$$

For example, if treatment is binary we could fit a logistic regression model for the probability of being treated at each time point given the observed past and not being censored by that time point, in which each subject would contribute one line of data for each time point. Similarly, we could fit a logistic regression model for the probability of being censored at each time point given the observed past.

General Approach to regression Above, we have been focusing on modeling the hazard, which implies the whole survival function. However, we note that this is not the only possible parameter of treatment-specific survival times we could be interested in. We might instead be happy with just mean survival ($E(T_{\bar{a}_1}|V) = m(\bar{a}_1,V|\beta)$), or median survival ($med(T_{\bar{a}_1}|V) = m(\bar{a}_1,V|\beta)$. To understand how we might model these parameters we introduce a general result for the class of estimating functions for any type of regression.

General Result:

$$E(\kappa(\epsilon|X)|X) = 0 \text{ at } \beta = \beta_0 \tag{32}$$

where $(\epsilon|X) = Y - m(X|\beta)$ and $\epsilon \to \kappa(\epsilon)$ is monotone increasing crossing zero.
We can choose different κ for different regression models. E.g.

- $\kappa(\epsilon) = \epsilon$, then $E(Y|X) = m(X|\beta_0)$ (mean regression)
- $\kappa(\epsilon) = (I(\epsilon > 0) - 1/2)$, then $median(Y|X) = m(X|\beta_0)$ (median regression)
- $\kappa(\epsilon) = (I(\epsilon > 0) - p)$ (pth quantile regression)
- $\kappa(\epsilon) = \epsilon$ for $\epsilon > \tau$, $-\tau$ for $\epsilon < -\tau$, and τ for $\epsilon > \tau$ (truncated mean regression)

So, the class of estimating functions for general regression is:

$$T_{NUIS}^{\perp} = \{h(X)\kappa(Y - m(X|\beta)) : h\} \tag{33}$$

To estimate the general regression parameter β, we set the empirical mean to zero and solve for β.

Censored Longitudinal Data and Causality: Notes for 11/24/04
History-Adjusted Marginal Structural Models for Survival Time Outcomes

In this lecture we will be concerned with estimating the treatment-specific hazard of survival conditional the past at a certain time-point.

The Data: The data will consist of $\tilde{T} = (min(T, C), \Delta = I(T \leq C), \bar{A}_1(\tilde{T}), \bar{L}(\tilde{T}))$ where $L(t) = (Y(t), W(t))$ and $Y(t) = I(T \leq t)$. (We note, however, that as in our previous lecture, $Y(t)$ could also be a general counting process that can jump more than once, in which case we would be interested in the intensity of the counting process rather than the hazard of survival that will be the parameter of interest in this lecture).

Our general approach in this lecture will be to construct counterfactuals under a joint action-mechanism, consisting of both censoring and treatment mechanisms, where the only counterfactuals we care about are those where censoring equals zero. An alternative approach would be to sequentially construct the estimating functions, as discussed in the prior lecture on missing data.

We introduce the following definitions:

- $L_{\bar{a}_1}(min(t, T_{\bar{a}_1}))$
- $L_{\bar{a}_1,\bar{a}_2}(t) = L_{\bar{a}_1}(min(t, c(\bar{a}_2)))$, where $c(\bar{a}_2)$ is the time at which censoring occurs and $\bar{a}_2$ jumps from zero to one.
- $A_2(t) = I(C \leq t)$, and if $T < C$ then $C = \infty$ to ensure that C is always observed.
- $A_1(t) = A_1(min(t, C, T))$
- $A(t) = (A_1(t), A_2(t))$

The full data can thus be written as $X^F = (L_{\bar{a}}, \bar{a} \in \mathcal{A})$, and the observed data can be written chronologically as $L(0), A(0), L(1), A(1), ..., L(j), A(j), ...L(\tilde{T})$.

Temporal Ordering: We assume $L_{\bar{a}}(t) = L_{\bar{a}(t-1)}(t)$, or that covariates are not affected by treatment that occurs after they are measured.

Consistency assumption: We assume that the observed data consist of the counterfactuals corresponding to the observed joint action history: $O = (\bar{A}, L_{\bar{A}}) \sim P_{F_X, g(\bar{A}|X)}$.

Sequential Randomization Assumption: We assume that the joint treatment and censoring action at each time point is only a function of the observed past

$$\begin{aligned} g(\bar{A}|X) &= \prod_t g(A_1(t)|\bar{A}(t-1), A_2(t), X) \prod_t g(A_2(t)|\bar{A}(t-1), X) \text{ (Always)} \\ &= \prod_t g(A_1(t)|\bar{A}(t-1), \bar{L}_{\bar{A}(t-1)}(t), A_2(t)) \prod_t g(A_2(t)|\bar{A}(t-1)\bar{L}_{\bar{A}(t-1)}(t)) \text{ (Under SRA)} \end{aligned}$$

where the first term represents the treatment mechanism and the second term represents the censoring mechanism. Let $(\bar{A}(t-1), \bar{L}_{\bar{A}(t-1)}(t), A_2(t)) = \mathcal{F}_1$ and let $(\bar{A}(t-1)\bar{L}_{\bar{A}(t-1)}(t)) = \mathcal{F}_2$. Since censoring (in this example) is a binary variable, we can write the censoring mechanism as the partial likelihood for a counting process:

$$\prod_t E(dA_2(t)|\mathcal{F}_2(t))^{dA_2(t)}(1 - E(dA_2(t)|\mathcal{F}_2)^{1-dA_2(t)} \tag{34}$$

and can model this using logistic regression:

$$\frac{1}{1 + e^{f(\mathcal{F}_2(t), \alpha_2)}} \tag{35}$$

Similarly, if treatment is binary we can model it using logistic regression; if it is categorical we can model it using multinomial logistic regression. We can do separate MLE for each partial likelihood, giving us an estimate of $g(\bar{A}|X)$ (possibly using cross-validation or other flexible approaches to our model).

Parameter of Interest: We are interested in the treatment-specific hazard under no censoring. The same framework could be used if we were interested in a general intensity rather than a hazard, or in some other parameter of the treatment specific survival time (such as median survival). Define $\bar{V}(j) = (\bar{A}(j-1), \bar{S}(j))$ where $\bar{S}(j) \subset \bar{L}(j)$, so that $\bar{V}(j)$ is a subset of the observed past up to time j, including the history of the treatment and censoring process and the history of some subset of measured covariates that are of interest. Define $\underline{a}_1(j) = a_1(j), a_1(j+1), ...$ as the future treatment regimen $\underline{a}_1$ beginning at time j. Our parameter of

interest is then:

$$E[dY_{\bar{A}_1(j-1),\underline{a}(j)}(t)|\bar{Y}_{\bar{A}_1(j-1),\underline{a}(j)},\bar{V}(j)] = \tag{36}$$

$$I(\tilde{T}>j, T_{\bar{A}_1(j-1),\underline{a}(j)} \geq t)E[dY_{\bar{A}_1(j-1)\underline{a}(j)}(t)|Y_{\bar{A}_1(j-1),\underline{a}(j)}(t-1)=0, \tilde{T}>j, \bar{V}(j)] = \tag{37}$$

$$I(\tilde{T}>j, T_{\bar{A}_1(j-1)\underline{a}(j)} \geq t)P(T_{\bar{A}_1(j-1),\underline{a}(j)}=t|T_{\bar{A}_1(j-1)\underline{a}(j)} \geq t, \tilde{T}>j, \bar{V}(j)) = \tag{38}$$

$$I(\tilde{T}>j, T_{\bar{A}_1(j-1)\underline{a}(j)} \geq t)\lambda_\beta(t, \bar{A}_1(j-1), \underline{a}_1(j), \bar{V}(j)) \tag{39}$$

For example, we might model the hazard as:

$$\lambda_\beta = \frac{1}{1+e^{-(\beta_0+\beta_1(summary(\bar{A}_1(j-1)))+\beta_2(summary(\underline{a}_1(j)))+\beta_3(V(j))+\beta_4(t-j))}} \tag{40}$$

(Note that, in defining our parameter of interest, we may be more interested in the hazard as a function of time elapsed after j (ie $t-j$)) We can pool the above model (40) across time points j. If we choose not to pool across time points, we are back to having a standard MSM for each time point j, treating the covariates up till that time point as baseline V. By including a term $(t-j)$, we can allow for different slopes and intercepts at different time points j, while still pooling our data to estimate a single parameter of interest. Note that in modeling the hazard, we are modeling the entire survival function ($S(t) = \prod_{s=0}^{t}(1-\lambda(s))$).

$$\bar{F}(t,|\bar{V}(j)) = P(T_{\bar{A}_1(j-1),\underline{a}_1(j)} \geq t|\bar{V}(j)) = I(\tilde{T}>j)\prod_{s=j}^{t}[1-\lambda_\beta(s, \bar{A}_1(j-1), \underline{a}_1(j), \bar{V}(j))] \tag{41}$$

Using our model of the treatment-specific survival function, we can define an optimal future treatment regimen beginning at time j that maximizes (or minimizes) our outcome of interest. For example, we could choose the future treatment regimen that maximizes the median survival time:

$$\underline{a}^*_{1\beta}(j)(j|\bar{V}(j)) = \arg\max_{\underline{a}_1(j)} \bar{F}^{-1}_{T_{\bar{A}_1(j-1),\underline{a}_1(j)}|\bar{V}(j),\tilde{T}>j}(0.5) \tag{42}$$

This gives us a dynamic treatment regimen, in which, when each subject comes in, we can look at their $\bar{V}(j)$ and plot the corresponding survival function under the future treatment regimes $\underline{a}_1$ of interest, and choose the future treatment regimen that maximizes the median survival time (ie choose the optimal future static treatment regime based on the observed past up till that time point) and take the first action of this optimal treatment regime. However, the optimal future treatment regime can then be updated when the subject is next observed, based on new $V(j)$ values. The fact that treatment decisions are made based on changing values of $V(j)$ makes this a dynamic treatment regime: $d(j|\bar{V}(j)) = a^*_{1,\beta}(j)$

Class of all Estimating Functions: We begin by writing down the IPTW estimating function for our MSM treating $\bar{A}_1(j-1), \bar{L}(j)$ as baseline covariates (39).

$$D^{IPTW}_{h,j}(O|\beta, g) = \sum_{t=j}^{\tilde{T}} h(t, \bar{V}(j), \bar{A}_1(j, t-1))$$

$$(dY_{\bar{A}_1(j-1),\underline{A}_1(j)(t)} - I(\tilde{T}>j, T \geq t)\lambda_\beta(t, \bar{A}_1(j-1), \underline{A}_1(j), \bar{V}(j)))\frac{I(\tilde{T} \geq t)}{g(\bar{A}_1(j,t-1), \bar{A}_2(j,t)=0|X)}$$

where we define $\bar{x}(j,t) = (x(j), x(j+1), ..., x(t))$ and note that $dY_{\bar{A}_1(j-1),\underline{A}_1(j)}(t) = dY_{\bar{A}} = dY$ and $I(\tilde{T} \geq t) = I(C \geq t)$. In addition, we can factorize the action mechanism as:

$$g(\bar{A}_1(j,t-1), \bar{A}_2(j,t)=0|X) = \prod_{s=j}^{t-1} g_1(A_1(s)|\mathcal{F}_1(s)) \prod_{s=j}^{t} g_2(0|\mathcal{F}_2(s)) \tag{43}$$

We make our usual choice of h:

$$h^* = \frac{\frac{\partial}{\partial\beta}\lambda_\beta g(\bar{A}_1(j,t-1), \bar{A}_2(j,t)=0|\bar{V}(j))}{\lambda_\beta(1-\lambda_\beta)} \tag{44}$$

To get the double robust estimating function at time j, we subtract the projection from the two parts of the action mechanism onto the nuisance tangent space. Because the action mechanism is a product of products,

we can project each piece separately and sum them.

$$D_{h,j}^{DR}(O|\beta,Q,g) = D_{h,j}^{IPTW}(O|\beta,g) - \sum_t [E(D_{h,j}^{IPTW}(O)|A_1(t),\mathcal{F}_1(t)) - E(D_{h,j}^{IPTW}(O)|\mathcal{F}_1(t))] \\ - \sum_t [E(D_{h,j}^{IPTW}(O)|A_2(t),\mathcal{F}_2(t)) - E(D_{h,j}^{IPTW}(O)|\mathcal{F}_2(t))]$$

where the first sum over t is the projection of the treatment mechanism and the second sum over t is the projection of the censoring mechanism. We can now write the IPTW and Double Robust estimating functions for the survival HA-MSM as the intersection of (45) over all time points j.

$$\begin{aligned} D_h^{IPTW}(O|\beta,g) &= \sum_{j=0}^{\tilde{T}} D_{h,j}^{IPTW}(O|\beta,g) \\ D_H^{DR}(O|\beta,g,Q) &= \sum_{j=0}^{\tilde{T}} D_{h,j}^{DR}(O|\beta,g,Q) \end{aligned}$$