Statistical Methods for Causal Inference in Observational and Randomized Studies

Mark J. van der Laan[1], Maya L. Petersen[1], Sherri Rose[2]

[1]University of California, Berkeley School of Public Health
[2]Johns Hopkins Bloomberg School of Public Health

laan@berkeley.edu · mayaliv@berkeley.edu · srose@jhsph.edu
stat.berkeley.edu/∼laan/
works.bepress.com/maya_petersen/
drsherrirose.com

targetedlearningbook.com

September 26, 2011

DAY ONE: LECTURE TWO

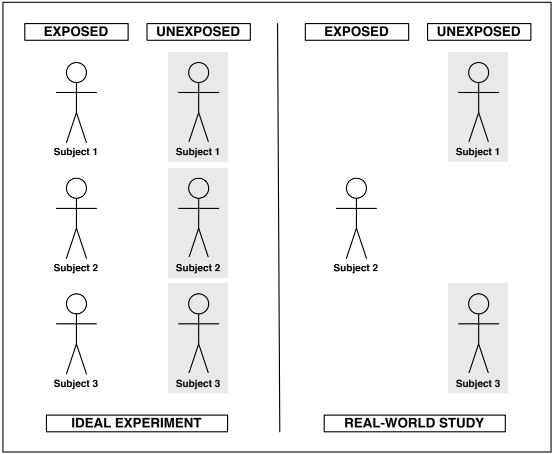
Defining the Research Question: Data, Model, Parameter

The question

How does one translate the results from studies, how do we take the information in the data, and draw effective conclusions?

Learning from Data

Just what type of studies are we conducting? The often quoted "ideal experiment" is one that cannot be conducted in real life.



Experimental Studies

- The randomization in RCTs suggests that we can estimate the causal effect of the treatment.
- Indeed, this randomization of treatment in RCTs allows us to go from the observed data to the causal effect of interest.
- The difference in means can be estimated using a saturated regression of the outcome on treatment in a parametric statistical model where covariates are ignored.

Experimental Studies

- Suppose randomization did not occur perfectly due to chance (as is common), and there is a single covariate that is predictive of the outcome.
- Why not run regressions in parametric statistical models (incorporating all covariates) for RCTs? The short answer is simple: the FDA does not allow it. We will expand on this issue later.

Examples of RCTs.

Observational Studies

- Recall that observational studies do not involve randomization to treatment or exposure.
- In most observational studies, standard practice for effect estimation involves assuming a parametric statistical model and using maximum likelihood estimation to estimate the parameters in that statistical model.

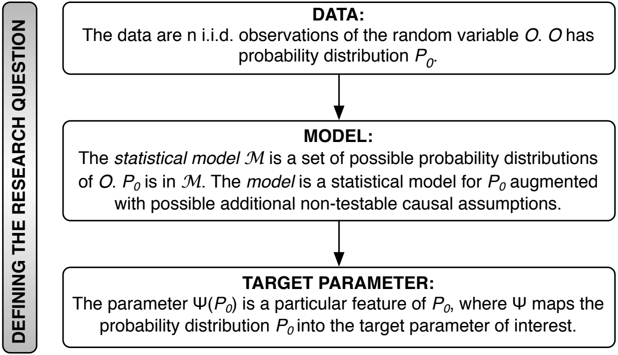
Examples of observational studies, including large comprehensive databases.

Observational Studies

The use of parametric statistical models in observational studies is troublesome for several main reasons.

1. The statistical models are always misspecified in practice since we do not know the underlying data-generating distribution and we handle complex problems with many covariates.
2. The target parameter is not defined as a parameter of the true probability distribution that generated the data.
3. The traditional approach does not typically make causal assumptions allowing us to define the desired causal effect, and often neglects other key assumptions, such as the positivity assumption, that are part of the statistical model.

Road map - Defining the Research Question



DEFINING THE RESEARCH QUESTION

DATA:
The data are n i.i.d. observations of the random variable O. O has probability distribution P_0.

MODEL:
The *statistical model* $\mathcal{M}$ is a set of possible probability distributions of O. P_0 is in $\mathcal{M}$. The *model* is a statistical model for P_0 augmented with possible additional non-testable causal assumptions.

TARGET PARAMETER:
The parameter $\Psi(P_0)$ is a particular feature of P_0, where Ψ maps the probability distribution P_0 into the target parameter of interest.

Data

Our random variable O, which we observe n times, could be defined in a simple case as $O = (W, A, Y) \sim P_0$ if we are without common issues such as missingness and censoring.

- W: vector of covariates
- A: exposure or treatment
- Y: outcome

This data structure makes for effective examples, but data structures found in practice are frequently more complicated.

Censored Data

We define $O = (W, A, \tilde{T}, \Delta) \sim P_0$.

- T: time to event Y
- C: censoring time
- $\tilde{T} = \min(T, C)$: represents the T or C that was observed first
- $\Delta = I(T \leq \tilde{T}) = I(C \geq T)$: indicator that T was observed at or before C

Missing Outcome

We define $O = (W, A, \Delta, \Delta Y) \sim P_0$.

- Δ: Indicator of missingness

Other Data Structures to Discuss for Pt Treatment?

-

COFFEE BREAK

Defining the Research Question: Data, Model, Parameter

Model

We are considering the general case that one observed n i.i.d. copies of a random variable O with probability distribution P_0.

The data-generating distribution P_0 is also known to be an element of a statistical model $\mathcal{M}$: $P_0 \in \mathcal{M}$.

A **statistical model** $\mathcal{M}$ is the set of possible probability distributions for P_0; it is a collection of probability distributions.

If all we know is that we have n i.i.d. copies of O, this can be our statistical model, which we call a nonparametric statistical model

Model

A statistical model can be augmented with additional (nontestable causal) assumptions, allowing one to enrich the interpretation of $\Psi(P_0)$.

This does not change the statistical model.

We refer to the statistical model augmented with a possibly additional assumptions as the **model**.

- Causal assumptions made by the structural causal model (SCM)

Defining the SCM

- We first specify a set of endogenous variables $X = (X_j : j)$.
- Endogenous variables are those variables for which theSCM will state that it is a (typically unknown) deterministic function of some of the other endogenous variables and an exogenous error.
- Typically, the endogenous variables X include the observables O, but might also include some nonobservables that are meaningful and important to the scientific question of interest. Perhaps there was a variable you did not measure, but would have liked to, and it plays a crucial role in defining the scientific question of interest. This variable would then be an unobserved endogenous variable.

Defining the SCM

- In a very simple example, we might have $j = 1, \ldots, J$, where $J = 3$. Thus, $X = (X_1, X_2, X_3)$.
- We can rewrite X as $X = (W, A, Y)$ if we say $X_1 = W$, $X_2 = A$, and $X_3 = Y$.
- Let W represent the set of baseline covariates for a subject, A the treatment or exposure, and Y the outcome.
- All the variables in X are observed.

Defining the SCM

- For each endogenous variable X_j one specifies the parents of X_j among X, denoted $Pa(X_j)$.
- The specification of the parents might be known by the time ordering in which the X_j were collected over time: the parents of a variable collected at time t could be defined as the observed past at time t.
- We can see the time ordering involved in this process: the baseline covariates occurred before the exposure LTPA, which occurred before the outcome of death: $W \rightarrow A \rightarrow Y$.

Defining the SCM

- We denote a collection of exogenous variables by $U = (U_{X_j} : j)$.
- These variables in U are never observed and are not affected by the endogenous variables in the model, but instead they affect the endogenous variables.
- One assumes that X_j is some function of $Pa(X_j)$ and an exogenous U_{X_j}:
$$X_j = f_{X_j}(Pa(X_j), U_{X_j}),\ j = 1 \ldots, J.$$
- The collection of functions f_{X_j} indexed by all the endogenous variables is represented by $f = (f_{X_j} : j)$.
- Together with the joint distribution of U, these functions f_{X_j}, specify the data-generating distribution of (U, X) as they describe a deterministic system of structural equations (one for each endogenous variable X_j) that deterministically maps a realization of U into a realization of X.

Defining the SCM

- In an SCM one also refers to some of the endogenous variables as intervention variables.
- The SCM assumes that intervening on one of the intervention variables by setting their value, thereby making the function for that variable obsolete, does not change the form of the other functions.
- The functions f_{X_j} are often unspecified, but in some cases it might be reasonable to assume that these functions have to fall in a certain more restrictive class of functions.
- Similarly, there might be some knowledge about the joint distribution of U.

Defining the SCM

- The set of possible data-generating distributions of (U, X) can be obtained by varying the structural equations f over all allowed forms, and the distribution of the errors U over all possible error distributions defines the SCM for the full-data (U, X), i.e., the SCM is a statistical model for the random variable (U, X).
- An example of a fully parametric SCM would be obtained by assuming that all the functions f_{X_j} are known up to a finite number of parameters and that the error distribution is a multivariate normal distribution with mean zero and unknown covariance matrix. Such parametric structural equation models are not recommended.

Defining the SCM

The corresponding SCM for the observed data O also includes specifying the relation between the random variable (U, X) and the observed data O, so that the SCM for the full data implies a parameterization of the probability distribution of O in terms of f and the distribution P_U of U. This SCM for the observed data also implies a statistical model for the probability distribution of O.

Defining the SCM: Translation

We have the functions $f = (f_W, f_A, f_Y)$ and the exogenous variables $U = (U_W, U_A, U_Y)$. The values of W, A, and Y are deterministically assigned by U corresponding to the functions f. We specify our structural equation models, based on investigator knowledge, as

$$\begin{aligned} W &= f_W(U_W), \\ A &= f_A(W, U_A), \\ Y &= f_Y(W, A, U_Y), \end{aligned} \tag{1}$$

where no assumptions are made about the true shape of f_W, f_A, and f_Y. These functions f are nonparametric as we have not put a priori restrictions on their functional form.

Defining the SCM: Translation

- We may assume that U_A is independent of U_Y, given W, which corresponds with believing that there are no unmeasured factors that predict both A and the outcome Y: this is often called the no unmeasured confounders assumption.
- This SCM represents a semiparametric statistical model for the probability distribution of the errors U and endogenous variables $X = (W, A, Y)$.
- We assume that the observed data structure $O = (W, A, Y)$ is actually a realization of the endogenous variables (W, A, Y) generated by this system of structural equations.

This now defines the SCM for the observed data O.

Defining the SCM: Translation

We have assumed that the underlying data were generated by the following actions:

1. Drawing unobservable U from some probability distribution P_U ensuring that U_A is independent of U_Y, given W,
2. Generating W as a deterministic function of U_W,
3. Generating A as a deterministic function of W and U_A,
4. Generating Y as a deterministic function of W, A, and U_Y.

Defining the SCM

- Any probability distribution of O can be obtained by selecting a particular data-generating distribution of (U, X) in this SCM.
- Thus, the statistical model for P_0 implied by this SCM is a nonparametric model.
- As a consequence, one cannot determine from observing O if the assumptions in the SCM contradict the data.
- One states that the SCM represents a set of nontestable causal assumptions we have made about how the data were generated in nature.

Causal Graphs

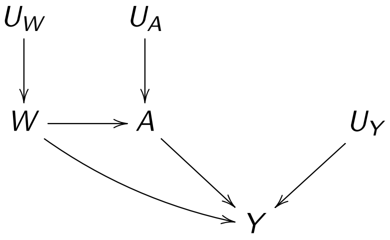


Figure: A possible causal graph for (1).

Causal Graphs

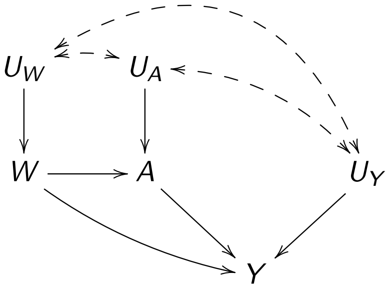


Figure: A causal graph for (1) with no assumptions on the distribution of P_U

Causal Graphs

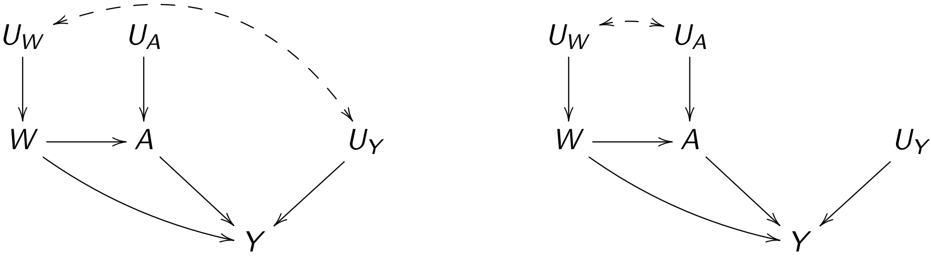


Figure: Causal graphs for (1) with various assumptions about the distribution of P_U

Defining the Causal Target Parameter

We can explicitly define the target parameter of the probability distribution P_0 as some function of P_0: $\Psi(P_0)$.

We are interested in estimating a parameter $\Psi(P_0)$ of the probability distribution $P_0 \in \mathcal{M}$, which is known to be an element of a non-parameteric (or semiparametric) statistical model $\mathcal{M}$.

Defining the Causal Target Parameter

- Formally, we denote the SCM for the full-data (U, X) by $\mathcal{M}^F$, a collection of possible $P_{U,X}$ as described by the SCM.
- In other words, $\mathcal{M}^F$, a model for the full data, is a collection of possible distributions for the underlying data (U, X).
- Ψ^F is a mapping applied to a $P_{U,X}$ giving $\Psi^F(P_{U,X})$ as the target parameter of $P_{U,X}$.

Defining the Causal Target Parameter

- This mapping needs to be defined for each $P_{U,X}$ that is a possible distribution of (U, X), given our assumptions coded by the posed SCM.
- We state $\Psi^F : \mathcal{M}^F \to \mathbb{R}^d$, where $\mathbb{R}^d$ indicates that our parameter is a vector of d real numbers.
- The SCM $\mathcal{M}^F$ consists of the distributions indexed by the deterministic function $f = (f_{X_j} : j)$ and distribution P_U of U, where f and this joint distribution P_U are identifiable from the distribution of the full-data (U, X).
- Thus the target parameter can also be represented as a function of f and the joint distribution of U.

Defining the Causal Target Parameter

- Recall our example with data structure $O = (W, A, Y)$ and SCM given in (1) with no assumptions about the distribution P_U.
- We can define $Y_a = f_Y(W, a, U_Y)$ as a random variable corresponding with intervention $A = a$ in the SCM.
- The marginal probability distribution of Y_a is thus given by

$$P_{U,X}(Y_a = y) = P_{U,X}(f_Y(W, a, U_Y) = y).$$

- The causal effect of interest for a binary A (suppose it is the causal risk difference) could then be defined as a parameter of the distribution of (U, X) given by

$$\Psi^F(P_{U,X}) = E_{U,X} Y_1 - E_{U,X} Y_0.$$

- In other words, $\Psi^F(P_{U,X})$ is the difference of marginal means of counterfactuals Y_1 and Y_0.

Interventions

- We will define our causal target parameter as a parameter of the distribution of the data (U, X) under an intervention on one or more of the structural equations in f.
- The intervention defines a random variable that is a function of (U, X), so that the target parameter is $\Psi^F(P_{U,X})$.
- Intervening on the system defined by our SCM describes the data that would be generated from the system at the different levels of our intervention variable (or variables).

Interventions

By assumption, intervening and changing the functions f_{X_j} of the intervention variables does not change the other functions in f. With the SCM given in (1) we can intervene on f_A and set $a = 1$:

$$\begin{aligned} W &= f_W(U_W), \\ a &= 1, \\ Y_1 &= f_Y(W, 1, U_Y). \end{aligned}$$

We can also intervene and set $a = 0$:

$$\begin{aligned} W &= f_W(U_W), \\ a &= 0, \\ Y_0 &= f_Y(W, 0, U_Y). \end{aligned}$$

Counterfactuals

- We would ideally like to see each individual's outcome at all possible levels of exposure A. The study is only capable of collecting Y under one exposure, the exposure the subject experiences.
- Y_a represents the outcome that would have been observed under this system for a particular subject under exposure a.
- In our example, for each realization u, which might correspond with an individual randomly drawn from some target population, by intervening on (1), we can generate so-called counterfactual outcomes $Y_1(u)$ and $Y_0(u)$.

Counterfactuals

- These counterfactual outcomes are implied by our SCM; they are consequences of it.
- That is, $Y_0(u) = f_Y(W, 0, u_Y)$, and $Y_1(u) = f_Y(W, 1, u_Y)$, where $W = f_W(u_W)$ is also implied by u.
- The random counterfactuals $Y_0 = Y_0(U)$ and $Y_1 = Y_1(U)$ are random through the probability distribution of U.
- For example, the expected outcome of Y_1 is the mean of $Y_1(u)$ with respect to the probability distribution of U. Our target parameter is a function of the probability distributions of these counterfactuals: $E_0 Y_1 - E_0 Y_0$.

Establishing Identifiability

Are the assumptions we have already made enough to express the causal parameter of interest as a parameter of the probability distribution P_0 of the observed data?

We want to be able to write $\Psi^F(P_{U,X,0})$ as $\Psi(P_0)$ for some parameter mapping Ψ.

Since the true probability distribution of (U, X) can be any element in the SCM $\mathcal{M}^F$, and each such choice $P_{U,X}$ implies a probability distribution $P(P_{U,X})$ of O, this requires that we show that $\Psi^F(P_{U,X}) = \Psi(P(P_{U,X}))$ for all $P_{U,X} \in \mathcal{M}^F$.

Establishing Identifiability

This step involves establishing possible additional assumptions on the distribution of U, or sometimes also on the deterministic functions f, so that we can identify the target parameter from the observed data distribution.

Thus, for each probability distribution of the underlying data (U, X) satisfying the SCM with these possible additional assumptions on P_U, we have $\Psi^F(P_{U,X}) = \Psi(P(P_{U,X}))$ for some Ψ.

O is implied by the distribution of (U, X), such as $O = X$ or $O \subset X$, and $P = P(P_{X,U})$, where $P(P_{U,X})$ is a distribution of O implied by $P_{U,X}$.

Establishing Identifiability

Let us denote the resulting full-data SCM by $\mathcal{M}^{F*} \subset \mathcal{M}^{F}$ to make clear that possible additional assumptions were made that were driven purely by the identifiability problem, not necessarily reflecting reality.

Establishing Identifiability

Theorems exist that are helpful to establish such a desired identifiability result. For example, if $O = X$, and the distribution of U is such that, for each s, A_s is independent of L_d, given $Pa(A_s)$, then the well-known g-formula expresses the distribution of L_d in terms of the distribution of O:

$$P(L_d = l) = \prod_{r=1}^{R} P(L_r = l_r \mid Pa_d(L_r)) = Pa_d(l_r)),$$

where $Pa_d(L_r)$ are the parents of L_r with the intervention nodes among these parent nodes deterministically set by intervention d.

Commit to a Statistical Model and Target Parameter

The identifiability result provides us with a purely statistical target parameter $\Psi(P_0)$ on the distribution P_0 of O.

The full-data model $\mathcal{M}^{F*}$ implies a statistical observed data model $\mathcal{M} = \{P(P_{X,U}) : P_{X,U} \in \mathcal{M}^{F*}\}$ for the distribution $P_0 = P(P_{U,X,0})$ of O.

This now defines a target parameter $\Psi : \mathcal{M} \to \mathbb{R}^d$.

Commit to a Statistical Model and Target Parameter

The statistical observed data model for the distribution of O might be the same for $\mathcal{M}^F$ and $\mathcal{M}^{F*}$.

If not, then one might consider extending the Ψ to the larger statistical observed data model implied by $\mathcal{M}^F$, such as possibly a fully nonparametric model allowing for all probability distributions.

If the more restricted SCM holds, our target parameter would still estimate the target parameter, but one now also allows the data to contradict the more restricted SCM based on additional doubtful assumptions.

Commit to a Statistical Model and Target Parameter

The causal risk difference in our simple example, in terms of the corresponding statistical parameter $\Psi(P_0)$:

$$\begin{aligned}\Psi^F(P_{U,X,0}) &= E_0 Y_1 - E_0 Y_0 \\ &= E_0[E_0(Y \mid A = 1, W) - E_0(Y \mid A = 0, W)] \\ &\equiv \Psi(P_0)\end{aligned}$$

where the outer expectation in the definition of $\Psi(P_0)$ is the mean across the strata for W.

Commit to a Statistical Model and Target Parameter

This identifiability result for the additive causal effect as a parameter of the distribution P_0 of O required making the randomization assumption stating that A is independent of the counterfactuals (Y_0, Y_1) within strata of W.

This assumption might have been included in the original SCM $\mathcal{M}^F$, but, if one knows there are unmeasured confounders, then the model $\mathcal{M}^{F*}$ would be more restrictive by enforcing this "known to be wrong" randomization assumption.

Positivity

Another required assumption is that $P_0(A = 1, W = w) > 0$ and $P_0(A = 0, W = w) > 0$ are positive for each possible realization w of W. Without this assumption, the conditional expectations of Y in $\Psi(P_0)$ are not well defined. This positivity assumption is also called the experimental treatment assignment (ETA) assumption.

We will discuss positivity more on **DAY THREE**.

Target Parameter

To be very explicit about how this parameter corresponds with mapping P_0 into a number:

$$\begin{aligned}\Psi(P_0) &= \sum_w \Bigg[\sum_y y P_0(Y = y \mid A = 1, W = w) \\ &\quad - \sum_y y P_0(Y = y \mid A = 0, W = w) \Bigg] P_0(W = w),\end{aligned}$$

where

$$P_0(Y = y \mid A = a, W = w) = \frac{P_0(W = w, A = a, Y = y)}{\sum_y P_0(W = w, A = a, Y = y)}$$

is the conditional probability distribution of $Y = y$, given $A = a, W = w$, and

$$P_0(W = w) = \sum_{y,a} P_0(Y = y, A = a, W = w)$$

is the marginal probability distribution of $W = w$.

Interpretation of the Target Parameter

The observed data parameter $\Psi(P_0)$ can be interpreted in two possibly distinct ways:

1. $\Psi(P_0)$ with $P_0 \in \mathcal{M}$ augmented with the truly reliable additional nonstatistical assumptions that are known to hold (e.g., $\mathcal{M}^F$). This may involve bounding the deviation of $\Psi(P_0)$ from the desired target causal effect $\Psi^F(P_{U,X,0})$ under a realistic causal model $\mathcal{M}^F$ that is not sufficient for the identifiability of this causal effect.
2. The truly causal parameter $\Psi^F(P_{U,X}) = \Psi(P_0)$ under the more restricted SCM $\mathcal{M}^{F*}$, thereby now including all causal assumptions that are needed to make the desired causal effect identifiable from the probability distribution P_0 of O.

Example: Target Parameter

Causal risk difference:

$$\begin{aligned}\Psi(P_0) &= E_0[E_0(Y \mid A=1,W) - E_0(Y \mid A=0,W)] \\ &= E_0 Y_1 - E_0 Y_0\end{aligned}$$

Example: Target Parameter

FILL IN OTHER TARGET PARAMETER

Example: Target Parameter

FILL IN OTHER TARGET PARAMETER

Example: Target Parameter

FILL IN OTHER TARGET PARAMETER