Section 28

Lecture 10

Marginal Structural Models

- We have learned that a *statistical* models puts restrictions on laws, that is, it puts restriction on conditional distributions (densities).
- Thus, the statistical model puts restrictions on the observed data distributions.
- A causal model puts restrictions on *counterfactual* densities, e.g. based on independence ($\perp\!\!\!\perp$) restrictions.
- We can make a causal (structural) model parametric by imposing parametric models for counterfactuals. Examples of such models are marginal structural models. Note that these models cannot be fitted directly to the data, because we don't directly observe the counterfactuals (see next slide)

Marginal structural models

An alternative way of weighting by the propensity scores is to define a so-called marginal structural model, which is a *statistical* model that parameterizes a functional of a *marginal* counterfactual Y^a (not the *joint* counterfactual $(Y^{a=1}, Y^{a=0})$).

- An example of a marginal structural model is

$$\mathbb{E}(Y^a) = \eta_0 + \eta_1 a.$$

- This model is saturated[39] for a binary A and implies that

$$\begin{aligned} \mathbb{E}(Y^0) &= \eta_0 \\ \mathbb{E}(Y^1) &= \eta_0 + \eta_1 \\ \mathbb{E}(Y^1) - \mathbb{E}(Y^0) &= \eta_1 \end{aligned}$$

- You can think about this as a regression model that is fitted to a (pseudo)population where A is randomly assigned.

[39] it does not impose restrictions on the data.

Estimator in marginal structural model

The estimator in a marginal structural model will look like

$$\hat{\mu}_{MSM}(a) = \frac{\frac{1}{n}\sum_{i=1}^{n} \frac{I(A_i=a)Y_i}{\pi(A_i|L_i;\gamma)}}{\frac{1}{n}\sum_{i=1}^{n} \frac{I(A_i=a)}{\pi(A_i|L_i;\gamma)}}.$$

I have omitted a proof.

PS: you can also try to show that, under our identifiability assumptions, $\hat{\mu}_{MSM}(a)$ is a consistent estimator of $\mathbb{E}(Y^a)$ by using results for weighted least square regressions. Both $\hat{\mu}_{IPW}(a)$ and $\hat{\mu}_{MSM}(a)$ are consistent. If Y is binary, only $\hat{\mu}_{MSM}(a)$ ensures that the estimate of $\mathbb{E}(Y^a)$ is in $[0, 1]$.

Further intuition on inverse probability weighting

- We can think of IPTW as creating an imaginary pseudopopulation in which there is no confounding: informally, we have a population where each individual i is represented by themselves and $w_i - 1$ other individuals, where w_i is the weight of individual i.
    - More formally, we consider a new law defined by a likelihood ratio
- Indeed, this is the way many applied researchers (including applied statisticians) think about this way of modelling. Formally, we do not need the concept of a pseudopopulation, but it is sometimes a useful motivation for the math and gives us some direction to come up with solutions.
- To be explicit, let us use the subscript "ps" to denote probability and expectation in the pseudopopulation (P_{ps} and $\mathbb{E}_{ps}$), while P and $\mathbb{E}$ without subscripts refer to the actual population. Consider the observed data $(Y\overline{A}, \overline{L})$.

*Estimation when the propensity is unknown

Define $\theta = (\mu, \gamma^T)^T$, and solve the stacked estimating equations

$$\sum_{i=1}^{n} \left(\frac{I(A_i = a) Y_i}{\pi(A_i \mid L_i; \gamma)} - \mu \right) = 0$$

$$\sum_{i=1}^{n} \begin{pmatrix} 1 \\ L_i \end{pmatrix} \left(A_i - \frac{\exp(\gamma_1 + \gamma_2^T L_i)}{1 + \exp(\gamma_1 + \gamma_2^T L_i)} \right) = 0,$$

The solution $\hat{\mu}_{IPW}$ to this system is an M-estimator, and therefore it is consistent (under our regularity conditions). We can use M-estimator theory to argue that the estimator is asymptotically normal.
In the next slide, we will study an interesting special case.

Outcome prediction for predictive purposes

Outcome regression is often used for purely *predictive* purposes.

- Online stores would like to predict which customers are more likely to purchase their products. The goal is not to determine whether your age, sex, income, geographic origin, and previous purchases have a causal effect on your current purchase. Rather, the goal is to identify those customers who are more likely to make a purchase so that specific marketing programs can be targeted to them. It is all about association, not causation. Similarly, doctors use algorithms based on outcome regression to identify patients at high risk of developing a serious disease or dying.
- A study found that Facebook Likes predict sexual orientation, political views, and personality traits (Kosinski et al, 2013). Low intelligence was predicted by, among other things, a "Harley Davidson" Like. This is purely predictive, not necessarily causal.

From Hernan and Robins, *Causal inference: What if?*

Prediction and procedures for model selection

- Model selection is a different endeavour when the aim is prediction.
- Investigators who seek to do pure predictions may want to include any variables that, when used as covariates in the model, improve its predictive ability.
- This motivates the use of selection procedures, such as forward selection, backward elimination, stepwise selection and new developments in machine learning.
- However, using these procedures for causal inference tasks can be unnecessary and harmful. Both bias and inflated variance may be the result.
- For example, we do not fit a propensity score model to predict the treatment A as good as possible: we just fit the model to guarantee exchangeability. Indeed, covariates that strongly associated with treatment, but are not necessary to guarantee exchangeability, do not reduce bias. Adjustment for these variables can lead to larger variance...

Standard error and variance for estimators in causal inference

- We can sometimes obtain variance estimators from M-estimator theory (see next slide).
- However, I do suggest using the bootstrap for the settings we consider here (see next slide for a brief introduction to bootstrap).
  - Computer intensive but convenient.
  - Simple in practice, but rigorous theory behind

*On the variance of M-estimators

Under regularity conditions, the asymptotic properties of an M-estimator $\hat{\theta}$ can be derived from Taylor series approximations, the law of large numbers, and the central limit theorem. Here is a brief outline.

- Let θ_0 and $\dot{M}(Z_i, \theta) = \partial M(Z_i, \theta)/\partial \theta^\intercal$ (This is a $k \times k$ matrix).
- $C(\theta_0) = E[-\dot{M}(Z_i, \theta_0)]$, and
- $B(\theta_0) = E[M(Z_i, \theta_0)M(Z_i, \theta_0)^\intercal]$. Then under suitable regularity assumptions, $\hat{\theta}$ is consistent and asymptotically Normal, i.e.,

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, \Sigma(\theta_0)) \text{ as } n \to \infty,$$

  where $\Sigma(\theta_0) = C(\theta_0)^{-1} B(\theta_0) \{C(\theta_0)^{-1}\}^\intercal$.
- This can be seen by a first-order Taylor series expansion of each row of the estimating equation $\sum_{i=1}^n M(Z_i; \hat{\theta}) = 0$ in $\hat{\theta}$ about θ_0,

$$0 = \sum_{i=1}^n M(Z_i; \theta_0) + \sum_{i=1}^n \left[\dot{M}(Z_i, \theta^*)\right](\hat{\theta} - \theta_0),$$

  where θ^* is a value between $\hat{\theta}$ and θ_0.

*Variance continues

- The sandwich form of $\Sigma(\theta_0)$ suggests several possible large sample variance estimators.
- For some problems, the analytic form of $\Sigma(\theta_0)$ can be derived and estimators of θ_0 and other unknowns simply plugged into $\Sigma(\theta_0)$.
- Alternatively, $\Sigma(\theta_0)$ can be consistently estimated by the empirical sandwich variance estimator, where the expectations in $C(\theta)$ and $B(\theta)$ are replaced with their empirical counterparts.
- Let $C_i = -\dot{M}(Z_i, \theta)|_{\theta=\hat{\theta}}$, $C_n = n^{-1}\sum_{i=1}^n C_i$, $B_i = M(Z_i, \hat{\theta})M(Z_i, \hat{\theta})^\intercal$, and $B_n = n^{-1}\sum_{i=1}^n B_i$. The empirical sandwich estimator of the variance of $\hat{\theta}$ is:

$$\hat{\Sigma} = C_n^{-1} B_n \{C_n^{-1}\}^\intercal / n.$$

Bootstrap

Bootstrap is a method for estimating the variance of a parameter.
Let $U_n = g(X_1, \ldots, X_n)$ be a statistic, i.e. a function of data. For example, $\hat{\mu}_{IPW}(a) = \frac{1}{n}\sum_{i=1}^{n} \frac{I(A_i=a)Y_i}{\pi(A_i|L_i;\hat{\gamma})}$, where in this case $X_i = (L_i, A_i, Y_i)$.
We want to estimate $VAR(U_n)$, and the bootstrap is motivated by two steps

1. Estimate $VAR(U_n)$ by $VAR_{\hat{\mathbb{P}}_n}(U_n)$, where $\hat{\mathbb{P}}_n$ is the empirical distribution.
2. Approximate $VAR_{\hat{\mathbb{P}}_n}(U_n)$ using simulations.

Step 2 is very useful when it is hard to express the closed form solution to the variance of U_n. Bootstrap variance estimation is done as follows:

1. Draw $X_1^*, \ldots, X_n^* \sim \hat{\mathbb{P}}_n$. (Sample with replacement from $(X_1, \ldots, X_n)$)
2. Compute $U_n^* = g(X_1^*, \ldots, X_n^*)$.
3. Repeat step 1 and 2 K times to get $U_{n,1}^*, U_{n,2}^*, \ldots, U_{n,K}^*$.[40]
4. $v_{\text{boot}} = \frac{1}{K}\sum_{k=1}^{K}\left(U_{n,k}^* - \frac{1}{K}\sum_{l=1}^{K} U_{n,l}^*\right)^2$

[40] Usually ≥ 1000 times.

Bootstrap

- Bootstrap is based on two approximations

$$VAR(U_n) \approx VAR_{\hat{\mathbb{P}}_n}(U_n) \approx v_{\text{boot}}.$$

- Bootstrap is very useful in practice and simple to implement: You just draw $X_1^*, \ldots, X_n^*$ with replacement from $(X_1, \ldots, X_n)$.

Bootstrap confidence intervals

Bootstrap confidence intervals can be created in several ways.

1. The normal intervals: $U_n \pm \eta_{\alpha/2}\hat{\text{se}}_{boot}$, $\sqrt{v_{\text{boot}}} = \hat{\text{se}}_{boot}$, where $\eta_{\alpha/2}$ is the $\alpha/2$ quantile of a standard normal variable. this requires U_n to be close to normal.
2. Percentile intervals: Define the interval $C_n = (U^*_{\eta/2}, U^*_{1-\eta/2})$, where U^*_ρ is the ρ sample quantile of $(U^*_{n,1}, U^*_{n,2}, \ldots, U^*_{n,K})$.
3. Studentised pivot intervals: Often perform better. A pivot is a random variable whose distribution does not depend on unknowns.

There are also many other ways of obtaining bootstrap confidence intervals. One high-level disclaimer: The bootstrap can, under certain data generating mechanisms, fail. If we have i.i.d. data an we study functionals that are reasonably smooth, which we study in the course the bootstrap will usually work. We will not consider violations in depth here.
For a detailed theory on the bootstrap, see Anthony Christopher Davison and David Victor Hinkley. *Bootstrap methods and their application*. 1. Cambridge university press, 1997

Section 29

Doubly robust estimators

Precision and IPW

- IPW estimators are often considered to be inefficient, that is, to have low precision.
- In principle, we can give two reasons why:
  - They give a more appropriate ("honest") reflection of the uncertainty, because they do not rely on implausible model assumptions.
  - They are truly inefficient, and we could impose the same model assumptions, and obtain a more efficient estimator.
- Asymptotic results from *semi-parametric efficiency theory* suggest that both these explanations can be true. We will not go into the details of semiparametric estimation theory, but we will show properties in some interesting examples.

Doubly robustness

- Natural way is to combine both regression and inverse probability weighting.
- Give a full factorization and see which terms are estimated in IPW and regression modelling.

Definition (Doubly robust estimator)

An estimator $\hat{\mu}$ of a parameter μ is doubly robust if it is a consistent estimator for μ if either of two models are correctly specified (e.g., the propensity model or the outcome regression model is correctly specified), but not necessarily both models are correctly specified.

Doubly robust estimator

Theorem (Doubly robust estimator of $\mathbb{E}(Y \mid L, A = a)$)

If either the propensity model $\pi(a \mid l; \gamma)$ or the outcome regression model $Q(l, a; \beta)$ is correctly specified, then

$$\mathbb{E}\left[\frac{I(A=a)Y}{\pi(a \mid L;\gamma)} + \left(1 - \frac{I(A=a)}{\pi(a \mid L;\gamma)}\right) Q(L, a; \beta)\right] = \mathbb{E}[\mathbb{E}(Y \mid L, A = a)].$$

Intuitively, the doubly robust estimator – unlike the simple inverse probability weighted estimator – exploits information from both treated and untreated. PS: note that we can re-write the expression in the theorem,

$$\begin{aligned}
&\mathbb{E}\left[\frac{I(A=a)Y}{\pi(a \mid L;\gamma)} + \left(1 - \frac{I(A=a)}{\pi(a \mid L;\gamma)}\right) Q(L, a; \beta)\right] \\
=&\mathbb{E}\left[Q(L, a; \beta) + \frac{I(A=a)}{\pi(a \mid L;\gamma)}\{Y - Q(L, a; \beta)\}\right]
\end{aligned}$$

Proof

Proof.

Suppose first that $\pi(a \mid l; \gamma)$ is correctly specified, but the outcome model $Q(l, a; \beta)$ is misspecified. Use iterative expectation,

$$
\begin{aligned}
\mathbb{E}\left\{\frac{I(A=a)Y}{\pi(a \mid L;\gamma)}\right\} =& \mathbb{E}\left\{\frac{I(A=a)}{\pi(a \mid L;\gamma)} E(Y \mid L, A)\right\} \\
=& \mathbb{E}\left\{\frac{I(A=a)}{\pi(a \mid L;\gamma)} E(Y \mid L, A=a)\right\} \\
=& \mathbb{E}\left\{\frac{\mathbb{E}(I(A=a) \mid L)}{\pi(a \mid L;\gamma)} E(Y \mid L, A=a)\right\} \\
=& \mathbb{E}\left\{\frac{(\pi(a \mid L)}{\pi(a \mid L;\gamma)} E(Y \mid L, A=a)\right\} \\
=& \mathbb{E}\left\{\mathbb{E}(Y \mid L, A=a)\right\}.
\end{aligned}
$$

Proof continues

Proof.

Next, consider the second term

$$\begin{aligned}
\mathbb{E}\left\{\left(1-\frac{I(A=a)}{\pi(a \mid L;\gamma)}\right)Q(L,a;\beta)\right\} &= \mathbb{E}\left\{\mathbb{E}\left[\left(1-\frac{I(A=a)}{\pi(a \mid L;\gamma)}\right)Q(L,a;\beta) \mid L\right]\right\} \\
&= \mathbb{E}\left\{\mathbb{E}\left(1-\frac{\mathbb{E}(I(A=a) \mid L)}{\pi(a \mid L;\gamma)}\right)Q(L,a;\beta)\right\} \\
&= \mathbb{E}\left\{(1-1)Q(L,a;\beta)\right\} = 0.
\end{aligned}$$

Proof continues (note: no reference to counterfactuals)

Proof.

Suppose now that $\pi(a \mid I; \gamma)$ is mis-specified, but the outcome model $Q(I, a; \beta)$ is correctly specified. After some algebra,

$$\begin{aligned}
&\mathbb{E}\left[\frac{I(A=a)Y}{\pi(a \mid L;\gamma)} + \left(1 - \frac{I(A=a)}{\pi(a \mid L;\gamma)}\right) Q(L, a; \beta)\right] \\
=&\mathbb{E}\left[Q(L, a; \beta) + \frac{I(A=a)}{\pi(a \mid L;\gamma)}\{Y - Q(L, a; \beta)\}\right]
\end{aligned}$$

Due to the correct specification, we know that the first term $\mathbb{E}[Q(L, a; \beta)] = \mathbb{E}[\mathbb{E}(Y \mid L, A = a)]$. Furthermore, using iterative expectation on the second term (conditional on L, similar to part 1 of the proof)

$$\begin{aligned}
&\mathbb{E}\left[\frac{I(A=a)}{\pi(a \mid L;\gamma)}\{Y - Q(L, a; \beta)\}\right] \\
=&\mathbb{E}\left[\frac{E(I(A=a) \mid L)}{\pi(a \mid L;\gamma)}\{E(Y \mid L, A=a) - Q(L, a; \beta)\}\right] = 0.
\end{aligned}$$

Some practical thoughts on estimation

- If we cannot guarantee that our model is correctly specified, we should in principle try to use different estimators (In practice it can be difficult).
- If all estimators give similar results, then there is some evidence (but not a guarantee!!) that we have modelled the problem correctly.
- If the estimators do not give the same results, try to understand why...
- In practice some degree of misspecification is inescapable in all models, and model misspecification will introduce some bias. But the misspecification of the treatment model (IP weighting) and the outcome model (standardization) will not generally result in the same magnitude and direction of bias in the effect estimate. Therefore the IP weighted estimate will generally differ from the standardised estimate because unavoidable model misspecification will affect the point estimates differently.
- The main advantage of doubly robust estimators is that they can have small bias, even when $Q(l, a)$ and $\pi(a \mid l)$ are estimated with machine learning methods. This has to do with the fact that the bias of the doubly robust estimator is a product of the errors in estimating $Q(l, a)$ and $\frac{1}{\pi(a|l)}$.

E-values

- However, the E-value method is controversial because "it uses no data information on observed confounders or prevalences, and no background information about uncontrolled confounders or their intercorrelations with controlled confounders. Instead it assumes only a worst case for the bias parameters, which tend to be as implausible as the best case assumed by conventional analyses. It seems then that use of E-values without more detailed confounding analysis is a clear violation of good practices".[49].

[49] Sander Greenland. "Dealing with the Inevitable Deficiencies of Bias Analysis–and All Analyses". In: *American Journal of Epidemiology* (2021).