Section 22

Lecture 8

Plan for today

- More on IPW estimators
- Say a few things about variance (so far we have talked quite a lot about bias).
  - Variance of estimators
  - bootstrap
- If time: begin introducing an estimator that combines both regression and IPW.

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 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.

Saturated model

This is just a notational remark. I will sometimes use the term ”saturated” model. A (parametric) model is saturated when the number of parameters (say, regression coefficients) is equal to the number of data points.

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, but you will show that this estimator is consistent in your homework.

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 $[0, 1]$.

Estimation when the propensity score is known

When $\pi(a \mid l)$ is a known function, the estimator of $\mathbb{E}(Y^a)$ is

$$\hat{\mu}_{IPW}(a) = \frac{1}{n} \sum_{i=1}^{n} \frac{I(A_i = a) Y_i}{\pi(A_i \mid L_i)}.$$

The propensity score $\pi(a \mid l)$, unlike the function $Q(l, a)$, is known in randomised experiments (determined by the investigator). However, in most observational data settings, the propensity score is unknown.
PS: This estimator has been known for a long time and is often called the Horvitz Thompson estimator in survey sampling.[35] It is not a maximum likelihood estimator of $\mathbb{E}(Y^a)$, but it is an M-estimator (you will see this from doing the homework).

[35] Daniel G Horvitz and Donovan J Thompson. "A generalization of sampling without replacement from a finite universe". In: *Journal of the American statistical Association* 47.260 (1952), pp. 663–685.

Estimation when the propensity score is unknown

More generally, we can propose a regression model $\pi(A \mid L; \gamma)$ for $\pi(A \mid L)$, and we can consider the estimator

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

For example, suppose that we fit a logistic regression model and find the MLE $\hat{\gamma}$ of γ, which is the solution to the estimating equation (See slide 201)

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

Estimation when the propensity is unknown

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

$$
\begin{aligned}
\sum_{i=1}^{n} \left(\frac{I(A_i = a) Y_i}{\pi(A_i \mid L_i; \gamma)} - \mu \right) &= 0, \\
\sum_{i=1}^{n} \begin{pmatrix} 1 \\ L_i \end{pmatrix} \left(A_i - \frac{\exp(\gamma_1 + \gamma_2^T L_i)}{1 + \exp(\gamma_1 + \gamma_2^T L_i)} \right) &= 0,
\end{aligned}
$$

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.

Example: The IPW estimator and variance

Suppose that we are in the randomised experiment, such that γ is known: let $P(A = 1 \mid L) = 0.5$, so $A \perp\!\!\!\perp L$. Suppose also that we adapt the correctly specified model $\pi(1 \mid l; \gamma) = \gamma$. In particular, the truth is $\gamma_0 = 0.5$.
Statistician 1 suggests using the true value $\gamma_0 = 0.5$ because it is known.
Statistician 2 suggests using the MLE $\pi(1 \mid l; \hat{\gamma}) = \hat{\gamma} = \frac{1}{n} \sum_{i=1}^{n} A_i$.
Who selected the most efficient estimator?

Statistician 1

The estimator for $\mu_1 = \mathbb{E}(Y^{a=1}) = \mathbb{E}(Y \mid A = 1)$ is

$$\hat{\mu_1} = \frac{1}{n}\sum_{i=1}^{n} \frac{A_i Y_i}{\gamma_0} = \sum_{i=1}^{n} \frac{A_i Y_i}{n/2}.$$

The estimator $\hat{\mu}_1$ is consistent because $\mathbb{E}(YA) = \mathbb{E}(A\mathbb{E}(Y \mid A)) = \frac{\mu_1}{2}$ and thus $n^{-1}\sum_{i=1}^{n} A_i Y_i \xrightarrow{P} \frac{\mu_1}{2}$. After some algebra,

$$\sqrt{n}(\hat{\mu}_1 - \mu_1) = 2n^{-1/2}\sum_{i=1}^{n}(A_i Y_i - \mu_1/2).$$

Define $\sigma_1^2 = var(Y \mid A = 1)$,

$$\begin{aligned} var(AY) &= \mathbb{E}(var(AY \mid A)) + var(\mathbb{E}(AY \mid A)) && (8) \\ &= \mathbb{E}(A\sigma_1^2) + var(A\mu_1) = \frac{\sigma_1^2}{2} + \frac{\mu_1^2}{4}. && (9) \end{aligned}$$

CLT: $\sqrt{n}(\hat{\mu}_1 - \mu_1) \xrightarrow{D} \mathcal{N}(0, 2\sigma_1^2 + \mu_1^2)$.

Statistician 2

The estimator for $\mu_1 = \mathbb{E}(Y^{a=1}) = \mathbb{E}(Y \mid A = 1)$ is

$$\hat{\mu}_1^* = \frac{1}{n} \sum_{i=1}^{n} \frac{A_i Y_i}{\hat{\gamma}} = \frac{\sum_{i=1}^{n} A_i Y_i}{\sum_{i=1}^{n} A_i}.$$

Indeed, $\hat{\mu}_1^*$ is consistent, $\mathbb{E}(YA) = \mathbb{E}(A\mathbb{E}(Y \mid A)) = \frac{\mu_1}{2}$, so that $n^{-1} \sum_{i=1}^{n} A_i Y_i \xrightarrow{P} \frac{\mu_1}{2}$ and $n^{-1} \sum_{i=1}^{n} A_i \xrightarrow{P} \frac{1}{2}$
After some algebra,

$$\sqrt{n}(\hat{\mu}_1^* - \mu_1) = \frac{n^{-1/2} \sum_{i=1}^{n} A_i(Y_i - \mu_1)}{n^{-1} \sum_{i=1}^{n} A_i}$$

$$var(A(Y - \mu_1)) = \mathbb{E}(Avar(Y - \mu_1 \mid A)) + var(A\mathbb{E}(Y - \mu_1 \mid A)) = \frac{\sigma_1^2}{2} + 0$$

CLT and Slutsky's theorem : $\sqrt{n}(\hat{\mu}_1^* - \mu_1) \xrightarrow{D} \mathcal{N}(0, 2\sigma_1^2)$.

Interesting insight from Statistician 1 vs Statistician 2

- It is more efficient to estimate the propensity score, even if the true propensity is known. (This is a more general result; not just a special case we have considered here.)
- Does this contradict what we know from MLE theory, where including more known information, leads to lower variance? **No**, this is not a contradiction because the IPW estimator is not an MLE for μ_1.

Outcome estimation 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 IPW estimators

- We can sometimes obtain variance estimators from M-estimator theory.
- 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 the scenes.

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 $\dot{M}(Z_i, \theta) = \partial M(Z_i, \theta)/\partial\theta^\intercal$ ($k \times k$ 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

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

Bootstrap

Bootstrap is a method for estimating the variance of a parameter.
Let $U_n = g(X_1, \dots, X_n)$ be a statistic, i.e. a function of data (that does not depend on unknown parameters). For example, $\hat{\mu}_{IPW}(a) = \frac{1}{n} \sum_{i=1}^{n} \frac{I(A_i=a)Y_i}{\pi(A_i|L_i;\hat{\gamma})}$.
We would like 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 $X_1^*, \dots, X_n^* \sim \hat{\mathbb{P}}_n$. (Sample with replacement from $(X_1, \dots, X_n)$)
2. Compute $U_n^* = g(X_1^*, \dots, X_n^*)$.
3. Repeat step 1 and 2 K times to get $U_{n,1}^*, U_{n,2}^*, \dots, U_{n,K}^*$.[36]
4. $v_{\text{boot}} = \frac{1}{K} \sum_{k=1}^{K} \left(U_{n,k}^* - \frac{1}{K} \sum_{l=1}^{K} U_{n,l}^* \right)^2$

[36] Usually $\geq$ 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 $(U^*_{n,1}, U^*_{n,2}, \dots, U^*_{n,K})$.
3. Studentised pivot intervals: Often perform better. A pivot is a random variable whose distribution does not depend on unknowns. You will see in the homework...

There are also many other ways of estimating 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, for example when the limiting distribution of U_n is normal and $X_1, \dots, X_n$ are iid, 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

Reminder on IPW

Remember that

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

is a consistent estimator. Indeed, it is an M-estimator because $\hat{\mu}_{IPW}(a)$ is a solution to the estimating equation,

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

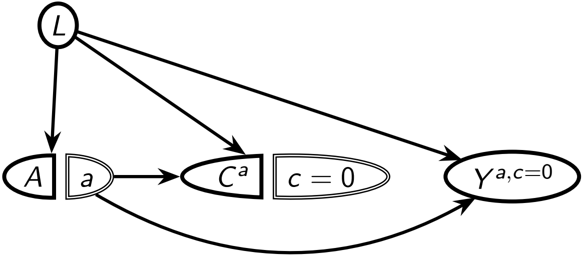
i.e., the solution to

$$\sum_{i=1}^{n} M(Z_i, \hat{\mu}_{IPW}(a)) = 0.$$

where $Z_i = (L_i, A_i, Y_i)$ and $M(Z_i, \mu) = \{\frac{I(A_i=a)Y_i}{\pi(A_i|L_i;\hat{\gamma})} - \mu\}$, where $\hat{\gamma}$ is an estimate of $\hat{\gamma}$ derived from another M-estimator, e.g. an MLE (thus, we have a stacked estimating equations).

Study on weight gain continues

Slightly extended graph



Censoring: weight gain study continues

- Suppose that there were 63 additional individuals who met our eligibility criteria but were excluded from the analysis because their weight in 1982 was not known. That is, their outcome was *censored*.
- Excluding the censored individuals will lead to selection bias due to conditioning on a collider.
- Then, the naive estimate is

$$\hat{\mathbb{E}}(Y \mid A = 1, C = 0) - \hat{\mathbb{E}}(Y \mid A = 0, C = 0) = 2.5 \text{ (95\% CI } : 1.7, 3.4).$$

- On the other hand, the causal effect of interest is

$$\hat{\mathbb{E}}(Y^{a=1,c=0}) - \hat{\mathbb{E}}(Y^{a=0,c=0}).$$

- From the exercises, we derived an identification formula $E[Y^{a,c=0}] = \sum_l E[Y \mid A = a, C = 0, L = l]P(L = l)$, that motivates a plug-in estimator, see the next slide.

Estimation using standardisation in the smoking example

We can estimate $\hat{\mathbb{E}}(Y^{a,c=0})$ by a plug-in g-formula estimator, also called a parametric g-formula estimator, because $\mathbb{E}(Y^{a,c=0}) = \mathbb{E}[\mathbb{E}(Y \mid A = a, C = 0, L)]$,

$$\frac{1}{n}\sum_{i=1}^{n} \hat{\mathbb{E}}(Y \mid A = a, C = 0, L_i),$$

where $\hat{\mathbb{E}}(Y \mid A = a, C = 0, L_i)$ is a regression model, like $Q(l, a; \beta)$ which is fitted to those who are uncensored $(C = 0)$.

- Suppose that we included a product term between smoking cessation A and intensity of smoking, but otherwise only main terms. This implies that our model imposes the restriction that each covariate's contribution to the mean is independent of that of the other covariates, except that the contribution of smoking cessation varies linearly.
- If we were interested in the average causal effect in a particular subset of the population, say characterised by V, we could have restricted our calculations to that subset.

Censoring: weight gain study continues with IPW

- Analogously, we can consider an IPW estimator in the presence of censoring.
- We multiply the original IPW weight with an inverse probability of censoring weight,

$$\pi_C(c \mid a, l) \equiv P(C = c \mid A = a, L = l).$$

  The proof that this work is essentially identical to the proof that IPW weighting works. Just replace $\pi(a \mid l)$ in the original proof with the product $\pi(a \mid l)\pi_C(0 \mid a, l) = P(A = a, C = 0 \mid L = l)$.

- Explicitly,

$$\hat{\mu}_{IPW}(a) = \frac{1}{n}\sum_{i=1}^{n} \frac{I(A_i = a, C_i = 0)Y_i}{\pi(A_i \mid L_i; \gamma_1)\pi_C(0 \mid a, L_i; \gamma_2)}.$$

- How would you obtain an estimate of $\pi_C(0 \mid a, l)$?

Just a few remarks on logistic regression models

- Logistic regression models have coefficients that are easily transformed to odds ratios. Thus, odds ratios are often reported in practice.
- Rarely a parameter of interest for decision making or causal inference more broadly (think back to the exercise on collapsibility too).