Section 23

Lecture 9

Precision and IPW

- IPW estimators are often considered to be inefficient, that is, to have low precision (You will see this in your homework).
- 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 of $\mathbb{E}(\mathbb{E}(Y \mid L, A = a)))$

An estimator $\hat{\mu}$ of a parameter μ is doubly robust if it is a consistent estimator for μ when either 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.

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

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

Proof continues (note: no reference to counterfactuals)

Proof.

Suppose now that $\pi(a \mid l; \gamma)$ is mis-specified, but the outcome model $Q(l, a; \beta)$ 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 twice on the second term (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.

New conceptual example, relevant for the analysis of simple randomised experiments (from Dukes)

- We aim to assess the effect of a simple treatment A (1 : treatment, 0 : control) on mortality Y (1: yes; 0: no) after one year.
- We have data from a trial in which A is randomly assigned.
- Suppose that, by chance, baseline disease severity L is more common on average in the treatment arm.
- Statistician 3 proposes to correct for the imbalance by fitting the outcome model regression model

$$\text{logit}\{Q(l, a; \beta)\} = \beta_1 + \beta_2 a + \beta_3 l.$$

Example continues

- Statistician 4 disagrees, because she is worried about model mis-specification.
- She suggests fitting the model

$$\text{logit}\{Q(a, \beta')\} = \beta'_1 + \beta'_2 a.$$

- Who is right?

By the way, Statistician 4's model is a saturated model, because it does not impose restrictions on the data; we just call it a model because it looks like a model, but the model does not put any restrictions on the data generating mechanism. Indeed, the number of parameters (here β'_1, β'_2) is equal to the number of conditional means that are estimated (here 2, $Q(a = 0, \beta')$ and $Q(a = 1, \beta')$)

Who is right: Statistician 3 or Statistician 4?

- Some will argue that statistician 4 is right that there is no need to account for the imbalance. [37]
- Statistician 4's model is saturated, and therefore guaranteed to be correct.
- So what are the advantage of statistician 3's approach? Imbalance creates noise, and heuristically statistician 3 filters this noise away.
- Thus, statistician 3's approach can drastically improve power, but the approach relies on correct parameterization.
- But let's take a look at Statistician 5's suggestion on the next slide.

[37] however, their strategy might violate the so-called conditionality principle, which you might have encountered in classes on statistical theory.

Statistician 5 suggests a tweak

A new statistician comes into the room and suggests an estimator for $\mathbb{E}(Y^{a=1})$ that is based on the same model as Statistician 3, but she suggests to do the following tweak $\frac{1}{n}\sum_{i=1}^{n} \text{expit}(\hat{\beta}_1 + \hat{\beta}_2 a + \hat{\beta}_3 L_i)$, where $\text{expit}(x) = 1/(1 + exp(-x))$.

Lemma (Consistent RCT estimator, even if misspecified)

The estimator $\frac{1}{n}\sum_{i=1}^{n} expit(\hat{\beta}_1 + \hat{\beta}_2 a + \hat{\beta}_3 L_i)$ based on MLE estimates from a logistic regression model

$$logit\{Q(l, a; \beta)\} = \beta_1 + \beta_2 a + \beta_3 l,$$

unbiasedly estimates $Q(l, a)$, even if the logistic regression model is misspecified.

Proof of Statistician 5

Proof.

We know that the following estimator is doubly robust, and hence consistent as long as $P(A_1 = 1)$ is correct:

$$\frac{1}{n}\sum_{i=1}^{n} \text{expit}(\hat{\beta}_1 + \hat{\beta}_2 a + \hat{\beta}_3 L_i) + \frac{I(A_i = a)}{P(A_i = a)}\{Y_i - \text{expit}(\hat{\beta}_1 + \hat{\beta}_2 a + \hat{\beta}_3 L_i)\},$$

where $\hat{\beta}$ are obtained via maximum likelihood estimation.
Now, we will show that the equation above is identical to the estimator suggested in the lemma. To see this, note that the MLE for a logistic regression model gives a (particular) score equation that solves (PS: this you have shown in an exercise)

$$0 = \sum_{i=1}^{n} A_i\{Y_i - \text{expit}[\hat{\beta}_1 + \hat{\beta}_2 a + \hat{\beta}_3 L_i]\},$$

and because $P(A_i = a)$ is a constant,
$0 = \sum_{i=1}^{n} \frac{A_i}{P(A_i=a)}\{Y_i - \text{expit}[\hat{\beta}_1 + \hat{\beta}_2 a + \hat{\beta}_3 L_i]\}$. □

Section 24

More on IPW

Further intuition on inverse probability weighting

- First, IPW is a *procedure*
- 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 (see next slide)
- 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 $(Y, \overline{A}_k, \overline{L}_k)$.

General positivity definition

Here is a more general definition of positivity that I include for your reference. The function $g_{j_l}(\cdot)$ gives a value to a_{j_l} under the counterfactual regime g of interest.

Definition (Positivity)

for each $k \in \{0, \ldots, K\}$, suppose

$$p(v_{j_k} \mid \overline{v}_{j_k-1}) > 0 \; \forall \; \overline{v}_{j_k} \text{ s.t.}$$
$$p(\overline{v}_{j_k-1}) > 0 \text{ and } \overline{v}_{j_l} = g_{j_l}(\overline{v}_{j_l-1}), l = 1, \ldots, k.$$

The intuition is that covariates that will have positive probability in the counterfactual world must also have positive probability in the observed world. Otherwise, we cannot identify outcomes in the counterfactual world from the observed data distributions.

IPW more explicitly

- We define the law $P_{ps}(Y = y, \overline{A}_K = \overline{a}_K, \overline{L}_K = \overline{l}_K)$ by the likelihood ratio

$$\frac{p_{ps}(Y, \overline{A}_K, \overline{L}_K)}{p(Y, \overline{A}_K, \overline{L}_K)} = \frac{g(\overline{A}_K)}{\prod_{k=0}^{K} p(A_k \mid \overline{L}_k, \overline{A}_{k-1})},$$

  where we sometimes will use the short hand notation $\overline{A}_K = \overline{A}$ and $\overline{L}_K = \overline{L}$. We call this likelihood ratio a **weight**. Thus
  - $g(\overline{A}) = \prod_{k=0}^{K} p_{ps}(A_k \mid \overline{L}_k, \overline{A}_{k-1})$,
  - $p(Y \mid \overline{L}_K, \overline{A}_K) = p_{ps}(Y \mid \overline{L}_K, \overline{A}_K)$
  - $\prod_{j=0}^{K} p(L_j \mid \overline{L}_{j-1}, \overline{A}_{j-1}) = \prod_{j=0}^{K} p_{ps}(L_j \mid \overline{L}_{j-1}, \overline{A}_{j-1})$.

  That is, some of the conditional densities are identical in the pseudopopulation and the observed population, and, importantly, $g(\overline{A})$ is not a function of L.
- Intuitively, We can think of IPTW as a procedure to cut the arrows (in a DAG) from the covariate history ($\overline{L}_k$) into treatment (A_k). Indeed, *many* applied researchers like this heuristic way of thinking about the problems.

IPW continues: 2 features

Now we state two features of IPW.

Feature 1:

- When using unstabilised weights,
  $P_{ps}(A_k = a_k \mid \overline{A}_{k-1} = \overline{a}_{k-1}, \overline{L}_k = \overline{l}_k) = 0.5$.
  In the pseudopopulation,

$$(A_k \perp\!\!\!\perp \overline{A}_{k-1}, \overline{L}_k)_{ps}.$$

- When using stabilised weights,
  $P_{ps}(A_k = a_k \mid \overline{A}_{k-1} = \overline{a}_{k-1}, \overline{L}_k = \overline{l}_k) = P(A_k = a_k \mid \overline{A}_{k-1} = \overline{a}_{k-1})$.
  In the pseudopopulation, we have that

$$(A_k \perp\!\!\!\perp \overline{L}_k \mid \overline{A}_{k-1})_{ps}.$$

`PS: A pseudopopulation is defined differently than a counterfactual population, but the results in the next slide shows how they are related.`

IPW continues: 2 important features

Feature 2:
Suppose that exchangeability, positivity and consistency hold. Then, IPW creates a pseudopopulation characterised by the following:

- Regardless of whether we use unstabilised or stabilised weights,

$$\mathbb{E}(Y^{\bar{a}}) = \mathbb{E}_{ps}(Y^{\bar{a}}) = \mathbb{E}_{ps}(Y \mid \overline{A} = \bar{a}).$$

- Thus, the average causal effect is equal to association in the pseudopopulation, and

$$\mathbb{E}(Y^{\bar{a}}) - \mathbb{E}(Y^{\bar{a}'}) = \mathbb{E}_{ps}(Y \mid \overline{A} = \bar{a}) - \mathbb{E}_{ps}(Y \mid \overline{A} = \bar{a}').$$

IPW theorem

We will give a theorem that shows feature 2 [38]:
Remember that the g-formula for the *marginal* of $Y \equiv Y_K$ under treatment assignment $\bar{a} \equiv \bar{a}_K = (a_0, \ldots, a_K)$ is defined as

$$b_{\bar{a}}(y) = \sum_{\bar{l}_K} p(y \mid \bar{l}_K, \bar{a}_K) \prod_{j=0}^{K} p(l_j \mid \bar{l}_{j-1}, \bar{a}_{j-1}).$$

Theorem (IPW theorem)

Under positivity,

$$\int y b_{\bar{a}}(y) dy = \mathbb{E}_{ps}(Y \mid \overline{A} = \bar{a}).$$

You will see that the theorem is very similar to other IPW results we have already showed.

[38] Feature 1 follows from some of the steps in the proof of feature 2, but I haven't written out all the details here