Section 31

Lecture 11

Pros/Cons of g-computationo (from Shpitser)

Positives:

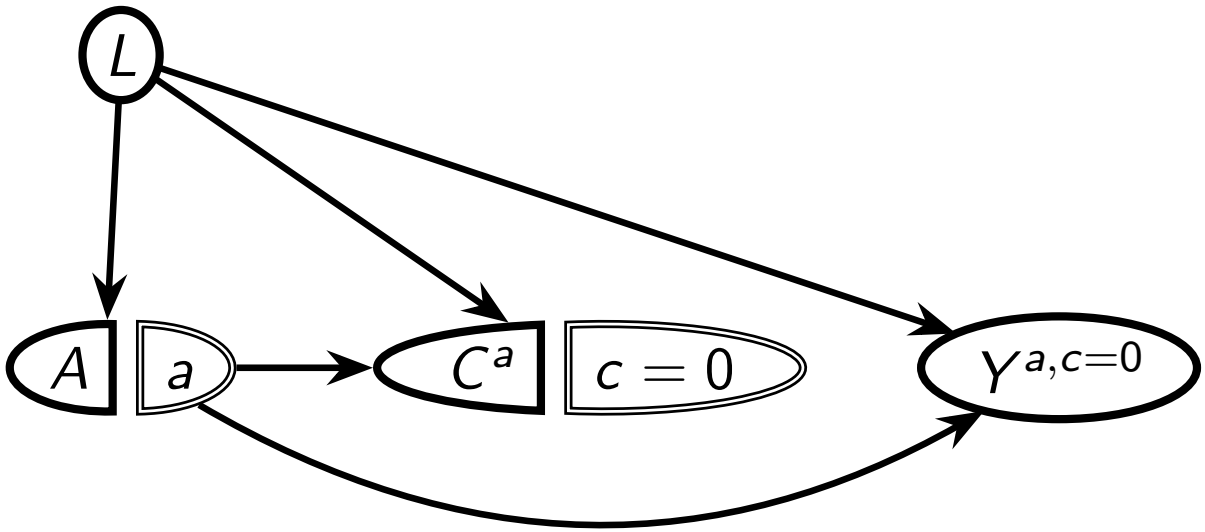
- Efficient if the models are correctly specified.
- In practice, people have reported that the approach is fairly robust to misspecification in practice.
- Conceptually, this is simple.

Negatives:

- Have to do a lot of (parametric) modeling, which means a risk of model misspecification.
- In general settings, this could be intractable and very slow.
- Sampling is computationally intensive
- Sampling trajectories can be unstable.

Study on weight gain continues

Slightly extended graph



Example: Smoking Cessation A on weight gain Y.

1566 cigarette smokers aged 25-74 years. The outcome weight gain measured after 10 years.

Mean baseline characteristics	A = 1	A = 0
Age, years	46.2	42.8
Men, %	54.6	46.6
White, %	91.1	85.4
University, %	15.4	9.9
Weight, kg	72.4	70.3
Cigarettes/day	18.6	21.2
Years smoking	26.0	24.1
Little exercise, %	40.7	37.9
Inactive life, %	11.2	8.9

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

Example; 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 can be correctly described as

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

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

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

- 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 -formula estimator, see the next slide.

Estimation using the g-formula in the smoking example

We can estimate $\hat{\mathbb{E}}(Y^{a,c=0})$ by a parametric g-formula estimator,

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

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 the model $Q(l, a; \beta)$ 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.

Section 32

More on IPW

General positivity definition

Before we continue, here is a more general definition of positivity that I include for your reference. This allows estimation with the parametric g-formula estimator and IPW. The function $g_{j_l}(\cdot)$ is the function that 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.

Censoring: weight gain study continues with IPW

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

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

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

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

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

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 use the short hand notation $\overline{A}_K = \overline{A}$ and $\overline{L}_K = \overline{L}$. 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, most 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, we have that

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

- When using stabilised weights,
  $P_{ps}(A_k = a_k \mid \overline{A}_{k-1} = \overline{a}_{k-1}, \overline{L}_k = \overline{l}_k) = P(A_k = a_k)$.
  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 or not,

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

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

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

IPW theorem

We will give a theorem that shows feature 2 [41]:
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 shown.

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

*Lemma of IPW theorem

Lemma

If the weights take the form

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

then

$$b_{\overline{a}}(y) = \frac{1}{g(\overline{a})} \mathbb{E}\left\{ \frac{g(\overline{A}) I(\overline{A} = \overline{a})}{\prod_{k=0}^{K} p(A_k \mid \overline{L}_k, \overline{A}_{k-1})} p(y \mid \overline{L}_K, \overline{A}_K) \right\}.$$

Proof.

$$
\begin{aligned}
&b_{\bar{a}}(y) \\
&= \sum_{\bar{l}_K} p(y \mid \bar{l}_K, \bar{a}_K) \prod_{j=0}^{K} p(l_j \mid \bar{l}_{j-1}, \bar{a}_{j-1}) \\
&= \sum_{\bar{l}_K} p(y \mid \bar{l}_K, \bar{a}_K) \frac{\prod_{k=0}^{K} p(a_k \mid \bar{l}_k, \bar{a}_{k-1})}{\prod_{k=0}^{K} p(a_k \mid \bar{l}_k, \bar{a}_{k-1})} \prod_{j=0}^{K} p(l_j \mid \bar{l}_{j-1}, \bar{a}_{j-1}) \\
&= \sum_{\bar{l}_K} \frac{1}{\prod_{k=0}^{K} p(a_k \mid \bar{l}_k, \bar{a}_{k-1})} p(y \mid \bar{l}_K, \bar{a}_K) \prod_{k=0}^{K} p(a_k \mid \bar{l}_k, \bar{a}_{k-1}) \prod_{j=0}^{K} p(l_j \mid \bar{l}_{j-1}, \bar{a}_{j-1}) \\
&= \sum_{\bar{l}_K} \frac{1}{\prod_{k=0}^{K} p(a_k \mid \bar{l}_k, \bar{a}_{k-1})} p(y, \bar{l}_K, \bar{a}_K).
\end{aligned}
$$

Proof.

$$= \sum_{\bar{l}_K} \frac{1}{\prod_{k=0}^{K} p(a_k \mid \bar{l}_k, \bar{a}_{k-1})} p(y, \bar{l}_K, \bar{a}_K)$$

$$= \sum_{\bar{l}_K} \sum_{\bar{a}^*} \frac{I(\bar{a}^* = \bar{a})}{\prod_{k=0}^{K} p(a_k^* \mid \bar{l}_k, \bar{a}_{k-1}^*)} p(y, \bar{l}_K, \bar{a}_K^*)$$

$$= \frac{1}{g(\bar{a})} \sum_{\bar{l}_K} \sum_{\bar{a}^*} \frac{g(\bar{a}^*) I(\bar{a}^* = \bar{a})}{\prod_{k=0}^{K} p(a_k^* \mid \bar{l}_k, \bar{a}_{k-1}^*)} p(y \mid \bar{l}_K, \bar{a}_K^*) p(\bar{l}_K, \bar{a}_K^*)$$

$$= \frac{1}{g(\bar{a})} \mathbb{E}\left\{ \frac{g(\overline{A}) I(\overline{A} = \bar{a})}{\prod_{k=0}^{K} p(A_k \mid \overline{L}_k, \overline{A}_{k-1})} p(y \mid \overline{L}_K, \overline{A}_K) \right\}.$$

where the expectation is taken over $\overline{A}_K, \overline{L}_K$ under the distribution that generated the observed data, and positivity is used in the last line. □

So the lemma from Slide 312 is shown.

*A corollary

Proof.

$$\int y b_{\bar{a}}(y) dy$$

$$= \int y \frac{1}{g(\bar{a})} \mathbb{E}\left\{ \frac{g(\bar{A}) I(\bar{A} = \bar{a})}{\prod_{k=0}^{K} p(A_k \mid \bar{L}_k, \bar{A}_{k-1})} p(y \mid \bar{L}_K, \bar{A}_K) \right\} dy$$

$$= \frac{1}{g(\bar{a})} \int \mathbb{E}\left\{ \frac{g(\bar{A}) I(\bar{A} = \bar{a})}{\prod_{k=0}^{K} p(A_k \mid \bar{L}_k, \bar{A}_{k-1})} y p(y \mid \bar{L}_K, \bar{A}_K) \right\} dy$$

$$= \frac{1}{g(\bar{a})} \mathbb{E}\left\{ \frac{g(\bar{A}) I(\bar{A} = \bar{a})}{\prod_{k=0}^{K} p(A_k \mid \bar{L}_k, \bar{A}_{k-1})} Y \right\} \text{ (by def of expectation)}$$

*Another (simple) lemma

Lemma (individuals with $\overline{A} = \bar{a}$ in the psedopopulation)

$$\mathbb{E}\left\{ \frac{g(\overline{A})I(\overline{A} = \bar{a})}{\prod_{k=0}^{K} p(A_k \mid \overline{L}_k, \overline{A}_{k-1})} \right\} = g(\bar{a}).$$

Proof.

We use that the g-formula is a density, i.e. that $\int b_{\bar{a}}(y)dy = 1$,

$$1 = \int b_{\bar{a}}(y)dy = \int \frac{1}{g(\bar{a})} \mathbb{E}\left\{ \frac{g(\overline{A})I(\overline{A} = \bar{a})}{\prod_{k=0}^{K} p(A_k \mid \overline{L}_k, \overline{A}_{k-1})} p(y \mid \overline{L}_K, \overline{A}_K) \right\} dy$$

$$g(\bar{a}) = \mathbb{E}\left\{ \frac{g(\overline{A})I(\overline{A} = \bar{a})}{\prod_{k=0}^{K} p(A_k \mid \overline{L}_k, \overline{A}_{k-1})} \right\},$$

where we used that integrals of sums are sums of integrals. □

*PS: Pseudopopulation vs observed population

Just a PS: the lemma allows us to characterize the number of treated in the pseudopopulation vs the original population. Recall that $\mathbb{E}(I(\overline{A} = \overline{a}))$ is the fraction of individuals with $\overline{A} = \overline{a}$ in the observed population. Let n be the total size of the observed population. Then

$$n \times \mathbb{E}(I(\overline{A} = \overline{a}))$$

is the expected number of individuals with $\overline{A} = \overline{a}$ in the observed population and

$$n \times \mathbb{E}\left\{ \frac{g(\overline{A})I(\overline{A} = \overline{a})}{\prod_{k=0}^{K} p(A_k \mid \overline{L}_k, \overline{A}_{k-1})} \right\} = n \times g(\overline{a})$$

is the expected number of individuals with $\overline{A} = \overline{a}$ in the pseudopopulation.

*Finally: A poof of the Theorem

Proof.

plugging in for $g(\bar{a})$ in the Expression from the Corollary on slide 315,

$$= \frac{\mathbb{E}\left\{ \frac{g(\overline{A})I(\overline{A}=\bar{a})}{\prod_{k=0}^{K} p(A_k|\overline{L}_k,\overline{A}_{k-1})} Y \right\}}{\mathbb{E}\left\{ \frac{g(\overline{A})I(\overline{A}=\bar{a})}{\prod_{k=0}^{K} p(A_k|\overline{L}_k,\overline{A}_{k-1})} \right\}} \quad \text{(i.e. an IPW formula)}$$

$$= \frac{\mathbb{E}_{ps}(I(\overline{A}=\bar{a})Y)}{P_{ps}(\overline{A}=\bar{a})}$$

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

□

This allows us to say "association is causation" in the pseudopopulation.

We can encode various assumptions in MSMs

- Suppose we hypothesize that the causal effect of treatment history $\bar{a}$ on the mean of Y is a linear function of the cumulative exposures, i.e.

$$\text{cum}(\bar{a}) = \sum_{k=0}^{K} a_k.$$

- This hypothesis is included in the MSM

$$\mathbb{E}(Y^{\bar{a}}) = \mathbb{E}_{ps}(Y \mid \bar{A} = \bar{a}) = \eta_0 + \eta_1 \text{cum}(\bar{a}).$$

  That is, we model the marginal mean of the counterfactuals $Y^{\bar{a}}$. Whereas there are 2^K treatment combinations (unknowns on the left-hand side of the equation), we have now reduced the model such that there are only two unknowns on the right-hand side of the equation.

- Obviously, like a statistical model, this model could also be misspecified, e.g. if the counterfactual outcome depends on some other function of the regime or if the outcome depends nonlinearly on the cumulative exposure.

*Motivating the weighted regressions

Lemma (Result for weighted least squares)

Suppose excheangeability, consistency and positivity hold. Then $\mathbb{E}_{ps}(Y \mid \overline{A} = \overline{a}) = \int yb(\overline{a})dy = \mathbb{E}(Y^{\overline{a}})$. *Then,*

$$\mathbb{E}\left\{ \frac{g(\overline{A})}{\prod_{k=0}^{K} p(A_k \mid \overline{L}_k, \overline{A}_{k-1})} [Y - \mathbb{E}(Y^{\overline{A}})] \right\}$$

$$\mathbb{E}_{ps}\left\{ [Y - \mathbb{E}(Y^{\overline{A}})] \right\}$$

$$= \mathbb{E}_{ps}\left\{ \mathbb{E}_{ps}\left\{ [Y - \mathbb{E}(Y^{\overline{A}})] \mid \overline{A} \right\} \right\}$$

$$= 0, \text{ because the inner expectation above is 0.}$$

Consider now the estimating equations

We use the results from the previous slide and the parameterisation

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

Now, consider the (two-dimensional) estimating equation

$$\sum_{i=1}^{n} M(\overline{L}_{k,i}, \overline{A}_i; \eta_0, \eta_1) = 0,$$

where

$$M(\overline{L}_k, \overline{A}; \eta_0, \eta_1) = \frac{g(\overline{A})}{\prod_{k=0}^{K} p(A_k | \overline{L}_k, \overline{A}_{k-1}; \gamma)} \begin{pmatrix} 1 \\ \text{cum}(\overline{A}) \end{pmatrix} [Y - \eta_0 - \eta_1 \text{cum}(\overline{A})].$$

This is an estimating equation for the weighted least squares estimator, where we simultaneously also solve the estimating equations for the propensities. Together, we denote the estimating equations for the counterfactual model and the propensity scores a "stacked estimating equation".

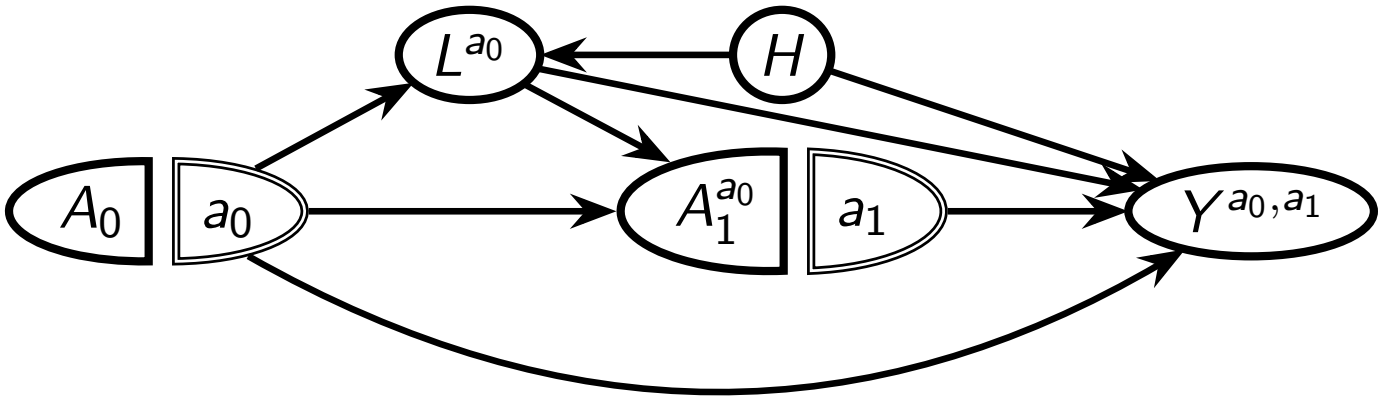
Null hypotheses in MSMs

Note that under the null hypothesis of no effect of any a_k, the MSM is correctly specified with

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

However, the standardisation estimator (parametric g-formula estimator) suffers from the so-called "g-null-paradox". That is, it is possible to show that it will always reject the null hypothesis – even if the null hypothesis is true – when the sample size grows.

Treatment-confounder feedback



- we cannot adjust for L using traditional methods, like stratification, outcome regression, and matching.
- But we read off that $Y^{a_0,a_1} \perp\!\!\!\perp A_0$ and $Y^{a_0,a_1} \perp\!\!\!\perp A_1^{a_0} \mid L_0^{a_0}, A_0 = a_0$, and we can fit MSMs, like the one on Slide 321.

MSMs and effect modification

Suppose that an investigator believes that for a particular component V of the vector of baseline covariates L_0, there might exist qualitative effect modification with respect to V. For example, suppose that $A = 1$ is harmful to subjects with $V = 0$ and beneficial to those with $V = 1$.
To examine this hypothesis, we would elaborate the MSM,

$$\mathbb{E}(Y^{\bar{a}} \mid V) = \eta_0 + \eta_1 \text{cum}(\bar{a}) + \eta_2 V + \eta_3 \text{cum}(\bar{a}) V.$$

Then we have qualitative effect modification if $\text{sign}(\eta_1) \neq \text{sign}(\eta_1 + \eta_3)$.
We can e.g. use the weights,

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

in a weighted least squares regression model.
One thing to notice: Here, IPW is used to adjust for confounding and regression modelling is used to study effect modification.

MSMs and direct effects

To illustrate a point, consider the saturated MSM for two binary treatments A_0, A_1,

$$\mathbb{E}(Y^{\bar{a}}) = \mathbb{E}(Y^{a_0,a_1}) = \eta_0 + \eta_1 a_0 + \eta_2 a_1 + \eta_3 a_0 a_1.$$

Now, the direct effect of A_0 when A_1 is set to 1 is $\mathbb{E}(Y^{1,1}) - \mathbb{E}(Y^{0,1})$.
How do we articulate the hypothesis that $\mathbb{E}(Y^{1,1}) = \mathbb{E}(Y^{0,1})$?

$$\begin{aligned} \mathbb{E}(Y^{1,1}) &= \mathbb{E}(Y^{0,1}) \\ \eta_0 + \eta_1 + \eta_2 + \eta_3 &= \eta_0 + \eta_2 \\ 0 &= \eta_1 + \eta_3 \end{aligned}$$

Optimal regimes and dynamic MSMs

Suppose that we aim to find the optimal treatment regime g^* in a given class of regimes $\{g = x : x \in \mathcal{X}\}$, where $|\mathcal{X}| = m$. Suppose that $x \in \{0, 1, \ldots, 999\}$. Let $n = 2000$ individuals.

- Suppose I come up with the following strategy: Run an experiment and randomly assign the regime g (In the experiment, we know association is causation)
- Maximize $\hat{\mathbb{E}}(Y \mid X = x)$
- Problem: We have m regimes, but only 2000 people so $\hat{\mathbb{E}}(Y \mid X = x)$ will be too variable...we will expect to have two people receiving the regime.
- Running example: Once we have started treatment (say, antiretroviral therapy in patients with HIV), then we never stop treatment. The question is: what is the best X to start treatment?

Dynamic MSMs

- Constructing an MSM allows us to impose assumptions, and then borrow strength across the regimes g, for example by assuming that $\mathbb{E}(Y \mid X = x) = \mathbb{E}(Y^x)$ is smooth in x.
- Note that we have to do this even if the data are from an experiment.
- Idea: for example, suppose we fit the model

$$\mathbb{E}(Y^x) = \eta_0 + \eta_1 x + \eta_2 x^2 + \eta_3 x^3.$$

- Then, we find the optimal regime g^* by maximising $\eta_1 x + \eta_2 x^2 + \eta_3 x^3$ over x.
- However, because there may be qualitative effect modification, we can expand the model to

$$\mathbb{E}(Y^x \mid V) = \eta_0 + \eta_1 x + \eta_2 x^2 + \eta_3 x^3 + \eta_4 x V,$$

and for each value of V maximize $\eta_1 x + \eta_2 x^2 + \eta_3 x^3 + \eta_4 x V$ over x,
$g(v) = \underset{x \in \mathcal{X}}{\arg\max}\ \eta_1 x + \eta_2 x^2 + \eta_3 x^3 + \eta_4 x v$

Advantages of MSMs

- Easy to understand
- Can be fitted with (weighted models) in standard software