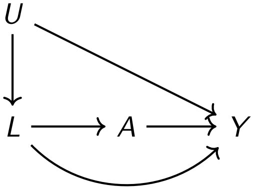
Separating causal questions from statistics with application to dietary interventions in the Project Viva cohort offspring

Soren Harnois-Leblanc, Jessica G. Young and Lan Wen

SPER Advanced Methods Workshop
April 28, 2026

Recall: notation and data structure

- Treatment/Exposure A
- Outcome of interest Y
- Measured baseline covariates L (unmeasured: U)
- Observe (L, A, Y) for each subject
- n denotes sample size
- Superscripts for counterfactual variable.



Methods for strict treatment regimes

- Parametric g-formula

Methods for strict treatment regimes

- Parametric g-formula
- Inverse probability weighting (IP Weighting)

Methods for strict treatment regimes

- Parametric g-formula
- Inverse probability weighting (IP Weighting)
- Doubly robust (DR) estimation

Example (strict treatment regime)

- Suppose for now that $A \in \{0, 1\}$;
- Interest is in estimating $E(Y^{a=1})$ – mean counterfactual outcome under 'always treat'

Example (strict treatment regime)

- Suppose for now that $A \in \{0, 1\}$;
- Interest is in estimating $E(Y^{a=1})$ – mean counterfactual outcome under 'always treat'
- Sufficient conditions:
  - Consistency: If $A = 1$, then $Y^{a=1} = Y$;
  - Conditional exchangeability: $Y^{a=1} \perp\!\!\!\perp A \mid L$; and
  - Positivity: $P(A = 1 \mid L = l) > 0$ for all levels of l observed.

Example (strict treatment regime)

- Suppose for now that $A \in \{0,1\}$;
- Interest is in estimating $E(Y^{a=1})$ – mean potential outcome under 'always treat'
- Under usual conditions (consistency, positivity and exchangeability),

$$E(Y^{a=1}) = \sum_{\forall l} E(Y \mid A=1, L=l)p(l),$$

where $p(l) = P(L=l)$.

Example (strict treatment regime)

Under usual conditions:

$$E(Y^{a=1}) = \sum_{\forall l} E(Y \mid A = 1, L = l) p(l).$$

where $p(l) = P(L = l)$.

Example (strict treatment regime)

Under usual conditions:

$$E(Y^{a=1}) = \sum_{\forall l} E(Y \mid A = 1, L = l)p(l).$$

where $p(l) = P(L = l)$.

Double expectation representation:

$$\begin{aligned} E(Y^{a=1}) &= \sum_{\forall l} E(Y \mid A = 1, L = l)p(l) \\ &= E\Big[E(Y \mid A = 1, L)\Big] \end{aligned}$$

Example (strict treatment regime)

Under usual conditions:

$$E(Y^{a=1}) = \sum_{\forall l} E(Y \mid A = 1, L = l)p(l).$$

where $p(l) = P(L = l)$.

Inverse probability (IP) weighted representation:

$$\begin{aligned} E(Y^{a=1}) &= \sum_{\forall l} E(Y \mid A = 1, L = l)p(l) \\ &= E\left[\frac{I(A = 1)}{P(A = 1 \mid L)} Y\right] \end{aligned}$$

Parametric g-formula estimator

Double expectation representation:

$$E(Y^{a=1}) = E\Big[E(Y \mid A = 1, L)\Big]$$

Parametric g-formula estimator

Double expectation representation:

$$E(Y^{a=1}) = E\Big[E(Y \mid A = 1, L)\Big]$$

Parametric g-formula estimator:

1. Regress Y on L in those whose $A = 1$.

Parametric g-formula estimator

Double expectation representation:

$$E(Y^{a=1}) = E\Big[E(Y \mid A = 1, L)\Big]$$

Parametric g-formula estimator:

1. Regress Y on L in those whose $A = 1$.
2. Using the fitted model in step 1, predict $E(Y \mid A = 1, L)$ for all subjects.
   - Denote the predicted value as $\hat{E}(Y \mid A = 1, L)$.

Parametric g-formula estimator

Double expectation representation:

$$E(Y^{a=1}) = E\Big[E(Y \mid A = 1, L)\Big]$$

Parametric g-formula estimator:

1. Regress Y on L in those whose $A = 1$.
2. Using the fitted model in step 1, predict $E(Y \mid A = 1, L)$ for all subjects.
   - Denote the predicted value as $\hat{E}(Y \mid A = 1, L)$.
3. Compute the average: $\frac{1}{n}\sum_{i=1}^{n}\hat{E}(Y_i \mid A_i = 1, L_i)$.

IP Weighting

Inverse probability (IP) weighted representation(s):

$$E(Y^{a=1}) = E\left[\frac{I(A=1)}{P(A=1 \mid L)}Y\right]$$

IP Weighting

Inverse probability (IP) weighted representation(s):

$$E(Y^{a=1}) = E\left[\frac{I(A=1)}{P(A=1 \mid L)}Y\right]$$

IP weighted estimator:

1. Fit a logistic regression model to predict $P(A=1 \mid L)$ for all subjects; denote predicted value as $\hat{P}(A=1 \mid L)$.

IP Weighting

Inverse probability (IP) weighted representation(s):

$$E(Y^{a=1}) = E\left[\frac{I(A=1)}{P(A=1 \mid L)}Y\right]$$

IP weighted estimator:

1. Fit a logistic regression model to predict $P(A=1 \mid L)$ for all subjects; denote predicted value as $\hat{P}(A=1 \mid L)$.
2. Compute the average: $\frac{1}{n}\sum_{i=1}^{n} \frac{I(A_i=1)}{\hat{P}(A_i=1|L_i)} Y_i$.

Singly robust estimator – a comparison

- Parametric g-formula and IP weighted estimators are modeling different components of the observed data distribution (**different statistical assumptions**)

Singly robust estimator – a comparison

- Parametric g-formula and IP weighted estimators are modeling different components of the observed data distribution (**different statistical assumptions**)
- Parametric g-formula tends to exhibit higher precision than IP weighting

Singly robust estimator – a comparison

- Parametric g-formula and IP weighted estimators are modeling different components of the observed data distribution (**different statistical assumptions**)
- Parametric g-formula tends to exhibit higher precision than IP weighting
- Misspecification of the treatment model (IP weighting) and the outcome model (parametric g-formula) $\rightarrow$ will not generally result in the same magnitude and direction of bias in the effect estimate

Doubly robust estimator

Can also show that:

$$E(Y^{a=1}) = E\left[\frac{I(A=1)}{P(A=1 \mid L)}\left\{Y - E(Y \mid A=1, L)\right\} + E(Y \mid A=1, L)\right].$$

("doubly robust" representation)

Doubly robust estimator

Can also show that:

$$E(Y^{a=1}) = E\left[\frac{I(A=1)}{P(A=1 \mid L)}\left\{Y - E(Y \mid A=1, L)\right\} + E(Y \mid A=1, L)\right].$$

("doubly robust" representation)

Motivates a doubly robust estimator called an augmented IP weighted (AIPW) estimator:

1. Predict $E(Y \mid A=1, L)$ and $P(A=1 \mid L)$ for all subjects;
2. Compute the average:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{I(A_i=1)}{\hat{P}(A_i=1 \mid L_i)}[Y_i - \hat{E}(Y_i \mid A_i=1, L_i)] + \hat{E}(Y_i \mid A_i=1, L_i).$$

Doubly robust estimator

- The AIPW estimator has nice theoretical guarantees.
- It is doubly robust: it's *consistent* as long as the model for $P(A_i = 1 \mid L_i)$ or the model for $E(Y_i \mid A_i = 1, L_i)$ is correctly specified (don't need both!).

Consistent = gets closer and closer to the true value as sample size increases.

Doubly robust estimator

- The AIPW estimator has nice theoretical guarantees.
- It is doubly robust: it's *consistent* as long as the model for $P(A_i = 1 \mid L_i)$ or the model for $E(Y_i \mid A_i = 1, L_i)$ is correctly specified (don't need both!).
- There are variations of doubly robust estimators: Bang and Robins' (2005) estimator; weighted standardized mean estimator; targeted minimum loss-based estimator.

Consistent = gets closer and closer to the true value as sample size increases.

Methods for regimes that depend on natural value of exposure

- Now suppose that $A \in \{0, 1, ..., k\}$;
- Interest: estimating mean counterfactual outcome under regime g that depends on *natural value of treatment*: $E(Y^g)$

E.g.,

$$a = \begin{cases} 0, & \text{if do not eat fast-food in a week} \\ 1, & \text{if eat fast-food } 1-2 \text{ times a week} \\ 2, & \text{if eat fast-food } \geq 3 \text{ times a week} \end{cases}$$

Methods for regimes that depend on natural value of exposure

Treatment/exposure under intervention could depend on a. E.g.,

1. If eat fast-food 3 times or more per week ($a = 2$), reduce and eat fast-food 1-2 times per week ($a = 1$).
2. If eat fast-food 1-2 times per week ($a = 1$), reduce intake to zero times per week ($a = 0$).
3. If do not eat fast-food ($a = 0$), do the same.

$$A^{g+} = \begin{cases} 1, & \text{if } a = 2 \\ 0, & \text{if } a \leq 1 \end{cases}$$

where A^{g+}: exposure level specified under g.

Identification results

Under sufficient conditions, g-specific consistency, conditional exchangeability and positivity:

Identification results

Under sufficient conditions, g-specific consistency, conditional exchangeability and positivity:

$$E(Y^g) = \sum_l \sum_a E(Y|A=a, L=l)p^g(a \mid l)p(l)$$
$$= \sum_l \sum_a E(Y \mid A = A^{g+}, L=l)p(a,l).$$

$p(a,l) = P(A=a, l=w)$; $p^g(a \mid l) = P(A^{g+} = a \mid L = l)$ depends on the definition of g; e.g.,

$$p^g(a \mid l) = \mathrm{I}(a=1)P(A=2|L) + \mathrm{I}(a=0)\Big[1 - P(A=2|L)\Big]$$

Methods for regimes that depend on natural value of exposure

- Parametric g-formula
- Inverse probability weighting (IP Weighting)
- Doubly robust (DR) estimation

Parametric g-formula

$$
\begin{aligned}
E(Y^g) &= \sum_w \sum_a E(Y \mid A = A^{g+}, L = l) p(a, l) \\
&= E[E(Y \mid A^{g+}, L)] \;\; (A^{g+}: \text{ exposure level under } g)
\end{aligned}
$$

Parametric g-formula

$$E(Y^g) = \sum_w \sum_a E(Y \mid A = A^{g+}, L = l)p(a, l)$$
$$= E[E(Y \mid A^{g+}, L)] \;\; (A^{g+}: \text{ exposure level under } g)$$

Parametric g-formula estimator:

1. Regress Y on L and A.

Parametric g-formula

$$E(Y^g) = \sum_w \sum_a E(Y \mid A = A^{g+}, L = l)p(a, l)$$
$$= E[E(Y \mid A^{g+}, L)] \;\; (A^{g+}: \text{ exposure level under } g)$$

Parametric g-formula estimator:

1. Regress Y on L and A.
2. Using fitted model in step 1, predict $E(Y \mid A = A^{g+}, L)$ for all subjects.
   - Plug in $A^{g+} = 1$ if $A = 2$;
   - Plug in $A^{g+} = 0$ otherwise (i.e., if $A = 0$ or $A = 1$)
   - Denote the predicted value as $\hat{E}(Y \mid A = A^{g+}, L)$.

Parametric g-formula

$$E(Y^g) = \sum_w \sum_a E(Y \mid A = A^{g+}, L = l) p(a, l)$$
$$= E[E(Y \mid A^{g+}, L)] \;\; (A^{g+}: \text{ exposure level under } g)$$

Parametric g-formula estimator:

1. Regress Y on L and A.
2. Using fitted model in step 1, predict $E(Y \mid A = A^{g+}, L)$ for all subjects.
   - Plug in $A^{g+} = 1$ if $A = 2$;
   - Plug in $A^{g+} = 0$ otherwise (i.e., if $A = 0$ or $A = 1$)
   - Denote the predicted value as $\hat{E}(Y \mid A = A^{g+}, L)$.
3. Compute the average: $\frac{1}{n} \sum_{i=1}^{n} \hat{E}(Y_i \mid A_i^{g+}, L_i)$.

IP Weighting

$$E(Y^g) = \sum_l \sum_a E(Y|A=a, L=l)p^g(a \mid l)p(l)$$
$$= E\left[\frac{p^g(A \mid L)}{p(A \mid L)} Y\right] \quad (A^{g+} : \text{ exposure level specified under } g)$$

e.g., $p^g(A \mid L) = \mathrm{I}(A=1)P(A=2|L) + \mathrm{I}(A=0)\Big[1 - P(A=2|L)\Big]$

IP Weighting

$$E(Y^g) = \sum_l \sum_a E(Y|A=a, L=l)p^g(a \mid l)p(l)$$
$$= E\left[\frac{p^g(A \mid L)}{p(A \mid L)} Y\right] \quad (A^{g+}: \text{ exposure level specified under } g)$$

e.g., $p^g(A \mid L) = \mathrm{I}(A=1)P(A=2|L) + \mathrm{I}(A=0)\Big[1 - P(A=2|L)\Big]$

IP weighted estimator:

1. Fit a model for $p(a \mid l) = P(A=a \mid L=l)$ to predict $p(a \mid l)$ and $p^g(a \mid l)$ for all subjects; denote predicted value as $\hat{p}(a \mid l)$ and $\hat{p}^g(a \mid l)$, respectively.

IP Weighting

$$E(Y^g) = \sum_l \sum_a E(Y|A=a, L=l) p^g(a \mid l) p(l)$$
$$= E\left[\frac{p^g(A \mid L)}{p(A \mid L)} Y\right] \quad (A^{g+}: \text{ exposure level specified under } g)$$

e.g., $p^g(A \mid L) = \mathrm{I}(A=1)P(A=2|L) + \mathrm{I}(A=0)\Big[1 - P(A=2|L)\Big]$

IP weighted estimator:

1. Fit a model for $p(a \mid l) = P(A=a \mid L=l)$ to predict $p(a \mid l)$ and $p^g(a \mid l)$ for all subjects; denote predicted value as $\hat{p}(a \mid l)$ and $\hat{p}^g(a \mid l)$, respectively.
2. Compute the average: $\frac{1}{n}\sum_{i=1}^{n} \frac{\hat{p}^g(A_i|L_i)}{\hat{p}(A_i|L_i)} Y_i$.

AIPW estimator

$$E(Y^g) = E\left[\frac{p^g(A \mid L)}{p(A \mid L)}[Y - E(Y \mid A, L)] + E(Y \mid A^{g+}, L)\right]$$

1. Predict $E(Y \mid A, L)$, $p^g(A \mid L)$ and $p(A \mid L)$ for all subjects;
2. Compute the average:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\hat{p}^g(A_i \mid L_i)}{\hat{p}(A_i \mid L_i)}[Y_i - \hat{E}(Y_i \mid A_i, L_i)] + \hat{E}(Y_i \mid A_i^{g+}, L_i).$$

Nuisance model estimation

- Parametric models vs machine learning

Nuisance model estimation

- Parametric models vs machine learning
- AIPW-based estimators are less prone to bias from overfitting compared with singly robust estimators

Nuisance model estimation

- Parametric models vs machine learning
- AIPW-based estimators are less prone to bias from overfitting compared with singly robust estimators
- Sample splitting + cross-fitting + AIPW $\rightarrow$ allows one to use flexible machine learning algorithms without worrying about smoothness of functions

Nuisance model estimation

- Parametric models vs machine learning
- AIPW-based estimators are less prone to bias from overfitting compared with singly robust estimators
- Sample splitting $+$ cross-fitting $+$ AIPW $\rightarrow$ allows one to use flexible machine learning algorithms without worrying about smoothness of functions
- Soren will demonstrate applications to some of these methods in R later!

References

- Haneuse, S. and Rotnitzky, A. (2013). *Estimation of the effect of interventions that modify the received treatment*. Statistics in Medicine, 32:5260–5277.
- Diaz, I., Williams, N., Hoffman, K. L., and Schenck, E. J. (2023). *Nonparametric causal effects based on longitudinal modified treatment policies*. Journal of the American Statistical Association, 118(542):846–857.
- Chiu, Y. H., and Wen, L. (2025). *Identification and estimation of the average causal effects under dietary substitution strategies*. Statistics in Medicine, 44(5), e70007