Causal Inference with a Continuous Treatment and Outcome: Alternative Estimators for Parametric Dose-Response Functions

Joe Schafer
Office of the Associate Director for Research and Methodology
U.S. Census Bureau
joint work with Doug Galagate

May 12, 2017
Outline

1. Motivation and setup
2. Methods
3. Simulation
4. Real example
5. Extensions

Disclaimer: This presentation is intended to inform interested parties of ongoing research and to encourage discussion. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.
1. Motivation

- Methods for causal inference from nonrandomized studies typically assume the treatment is binary; continuous-dose treatments are less well studied
  - Two different generalizations of the propensity score (Imai and van Dyk, 2004; Hirano and Imbens, 2004)
  - Marginal structural models (Robins, Hernan and Brumback, 2000) could be applied

- Start with the simplest, plain-vanilla version
  - Suppose that the average causal dose-response relationship in the population follows a simple parametric form (e.g., linear)
  - Is there a potential-outcome version of ordinary linear regression?
  - How would one estimate this “causal regression” line?
  - Which ways are best?
Example: Smoking and Medical Expenditures

- Previously analyzed by Imai and van Dyk (2004)
- Data from National Medical Expenditure Survey (NMES)
- Multistage cluster sample from area frame, with oversampling of certain groups
- Interviews in 1987 from persons in sampled HHs, with expenditures obtained from health care providers
- Use smokers over the age of 18 (roughly 10,000 in sample)
- Examine relationship between $Y_i =$ expenditures and $T_i =$ packyears, a measure of cumulative lifetime smoking
• Red lines are loess curves
• Plot on RHS omits top 5% of Y values
• Relationship is distorted by many potential confounders (e.g., age) and by sample design
• We will return to this example later
Usual Setup for Binary Treatment

\[ T_i = \text{treatment received by unit } i \]
\[(0=\text{control, } 1=\text{experimental})\]

\[ \mathcal{Y}_i = \{ Y_i(0), Y_i(1) \} \] set of potential outcomes

\[ Y_i(1) - Y_i(0) = \text{treatment effect for unit } i \] (unobservable)

\[ Y_i = Y_i(T_i) \]
\[ = T_i Y_i(1) + (1 - T_i) Y_i(0) \] observed outcome

\[ E( Y_i(1) - Y_i(0) ) = \text{population average treatment effect (PATE)} \]

- Data available for estimating PATE are \( (T_i, Y_i, X_i), \)
  \( i = 1, \ldots, N, \) where \( X_i = (X_{i1}, \ldots, X_{ip})^\top \) is a vector of
  pre-treatment covariates
- For review and history of this binary case, see Rubin (2005)
Setup for Continuous Treatment

\[ T_i \in \mathcal{T} = (t_{\text{min}}, t_{\text{max}}) \text{ real interval} \]
\[ \mathcal{Y}_i = \{ Y_i(t) : t \in \mathcal{T} \} \]

- \( Y_i(t) \) is function or path indexed by \( t \) that describes all treatment effects for unit \( i \)
- Our goal is to estimate the average dose-response function (ADRF) in the population,
  \[ \mu(t) = E(Y_i(t)) \text{ for } t \in \mathcal{T} \]
- We cannot observe \( Y_i(t) \); we see only one randomly chosen point along the path, \( Y_i = Y_i(T_i) \), along with \( T_i \) and \( X_i \).
- The regression of \( Y_i \) on \( T_i \) estimates
  \[ \mu^*(t) = E(Y_i(t) | T_i = t) \]

which, in general, is not the same as the ADRF.
Simulated example

Simulated sample of $N = 200$ observed points $(T_i, Y_i)$, with representative potential-outcome paths (gray lines), average causal dose response function $\mu(t) = E(Y_i(t))$, and regression curve $\mu^*(t) = E(Y_i(t) \mid T_i = t)$. 
Parameterizing the dose-response relationship

- In most previous work, authors made few direct assumptions about the form of $Y_i(t)$ or $\mu(t)$
- We will suppose that

$$Y_i(t) = \theta_i^\top b(t),$$

where $b(t) = (b_1(t), \ldots, b_k(t))^\top$ is a vector of known basis functions, and $\theta_i = (\theta_{i1}, \ldots, \theta_{ik})^\top$ is a vector of unknown coefficients specific to unit $i$
- An important special case is $b(t) = (1, t)^\top$, which specifies linear paths whose intercepts and slopes may vary
- Inference for $\mu(t) = E(Y_i(t))$ becomes a matter of estimating a $k$-dimensional parameter $\xi = E(\theta_i) = (\xi_1, \ldots, \xi_k)^\top$.
- We cannot observe $\theta_i$, but only

$$Y_i = Y_i(T_i) = \theta_i^\top B_i,$$

where $B_i = b(T_i) = (B_{i1}, \ldots, B_{ik})^\top$. Thus $Y_i$ is a randomly selected linear combination of the components of $\theta_i$, a type of coarsened data (Heitjan and Rubin, 1991)
Additional assumptions

- **Stable Unit Treatment Value Assumption (SUTVA):** Treatment applied to any unit has no impact on the outcome for any other unit (Rubin, 1980)
- **Strong ignorability:** $T_i \perp \theta_i \mid X_i$
- **Positivity:** $P(T_i \in T_0 \mid X_i) > 0$ for every $X_i$ in the population and every set $T_0 \subset T$ with positive measure
- Each technique for estimating $\xi$ will require us to model the distribution of $T_i$ given $X_i$ and/or the distribution of $\theta_i$ given $X_i$. Consistency will require at least some aspects of these models to be correctly specified.
- We envision situations where none of these models are precisely true, and look for estimators that perform well under moderate amounts of misspecification.
2. Methods

View from function space

- Under our parametric assumptions, each path $Y_i(t)$ is represented by a point $\theta_i = (\theta_{i1}, \ldots, \theta_{ik})^\top$ in $k$-dimensional space.

- Once we observe $Y_i$, we know that $\theta_i$ lies in the $(k - 1)$-dimensional hyperplane

$$L_i = \{\theta_i : \theta_i^\top b(t) = Y_i \quad \forall t \in T\}$$

- If we could observe the exact locations of the $\theta_i = (\theta_{i1}, \theta_{i2})^\top$, then the center of the point cloud, $\hat{\xi} = N^{-1} \sum_{i=1}^{N} \theta_i$, would be an unbiased estimate for $\xi$.

- Fortunately, there are some observable “magic vectors” whose expectations are equal to $E(\theta_i) = \xi$. 
Magic vector #1. Under a model for \( P(T_i \mid X_i) \), the vector
\[
\tilde{\theta}_i = W_i B_i Y_i,
\]
where \( W_i = E(B_i B_i^\top \mid X_i)^{-1} \), has expectation \( \xi \) if that \( T \)-model is correct. (Matrix generalization of the Horvitz-Thompson element \( \pi_i^{-1} I_i Y_i \).

Magic vector #2. Under a model for \( P(\theta_i \mid X_i) \), the vector
\[
\hat{\theta}_i = E(\theta_i \mid X_i)
\]
has expectation \( \xi \) if that \( Y \)-model is correct.

Magic vector #3. Under models for \( P(T_i \mid X_i) \) and \( P(\theta_i \mid X_i) \), the vector
\[
\tilde{\theta}_i = \tilde{\theta}_i + (I_k - W_i B_i B_i^\top) \hat{\theta}_i
\]
has expectation \( \xi \) if either the \( T \)-model or the \( Y \)-model is correct.
Method #0: The naive approach

Regress $Y_i$ on $B_i = b(T_i)$

$$
\hat{\xi} = \left( \sum_{i=1}^{N} B_i B_i^\top \right)^{-1} \left( \sum_{i=1}^{N} B_i Y_i \right).
$$

- Solves $U(\xi) = \sum_{i=1}^{N} U_i = 0$, where $U_i = U_i(\xi) = B_i(Y_i - B_i^\top \xi) = B_i B_i^\top (\theta_i - \xi)$ is a vector of estimating functions.

- Adapting terminology from Holland (1980), we call this the *prima facie* estimator.

- Ignores information in $X_i$.

- Would be consistent if $T_i$ and $\theta_i$ were independent.

- Performs poorly, not recommended.
What if we toss in the covariates?

- In practice, many analysts adjust for confounders by including them as additional predictors, regressing $Y_i$ on $(B_i^T, X_i^T)^T$ and hoping for a good result.
- Can work surprisingly well in some cases, badly in others
- The ADRF describes the marginal mean of $Y_i(t)$, which requires averaging over covariates, not conditioning on them, and this difference often goes unappreciated.
- When analysts do this, the connection to potential outcomes is rarely made explicit.
- IMO, we should stop teaching people to do this.
Method #1: Importance weighting

Robins, Hernan and Brumback (RHB) (2000) use the estimating function of the form

$$U_i = \frac{P(T_i)}{P(T_i | X_i)} B_i (Y_i - B_i^\top \xi)$$

- Equivalent to weighted least-squares (WLS) regression of $Y_i$ on $B_i$, with weights $P(T_i)/P(T_i | X_i)$; RHB call them “stabilized weights”
- Related to importance sampling (Hammersley and Handscomb, 1964)
- Adjusts the expectation of the \textit{prima facie} $U_i$ to what it would be if $(X_i, T_i, Y_i)$ were sampled from a population in which $T_i$ is independent of $X_i$
• This is a classic case where importance sampling tends to fail, because the target density $P(T_i)$ is more diffuse than the actual density $P(T_i \mid X_i)$, causing the weights to be highly unstable.

• Sensitive to misspecification of $P(T_i \mid X_i)$ and $P(T_i)$ in the tails

• Sometimes works well when $T_i$ is discrete, but not recommended for continuous treatments
Method #2: Inverse second-moment weighting

Taking $U_i = (W_i B_i Y_i - \xi)$ for $W_i = [E(B_i B_i^\top | X_i)]^{-1}$ gives

$$\hat{\xi} = \frac{1}{N} \left( \sum_{i=1}^{N} W_i B_i Y_i \right).$$

- A natural extension of inverse probability of treatment weighting (IPTW) for a binary $T_i$ (Hirano and Imbens, 2001).
- A modified version that normalizes the weights is

$$\hat{\xi} = \left( \sum_{i=1}^{N} W_i B_i B_i^\top \right)^{-1} \left( \sum_{i=1}^{N} W_i B_i Y_i \right).$$

- Not a typical WLS regression; the weight $W_i$ is a matrix, and the “information” $\sum_i W_i B_i B_i^\top$ is asymmetric
- Relies on a model for $P(T_i | X_i)$, but is more stable and robust than importance weighting
Method # 3: Regression prediction

Taking $U_i = (\hat{\theta}_i - \xi)$, where $\hat{\theta}_i = E(\theta_i|X_i)$ under a model, gives

$$\hat{\xi} = N^{-1} \sum_{i=1}^{N} \hat{\theta}_i$$

- For example, suppose $\theta_i|X_i \sim N(\nu + \Gamma^T X_i, \Sigma)$. Under strong ignorability, we can fit this model by regressing $Y_i$ on $B_i$ and $(B_i \otimes X_i)$.
- Including the interactions between $B_i$ and $X_i$ seems key.
- The difficulty of doing this as the lengths of $B_i$ and $X_i$ grow is the “curse of dimensionality”
Method # 4: Regression prediction with residual bias correction

Taking

\[ U_i = W_i B_i (Y_i - B_i^\top \xi) + (I - W_i B_i B_i^\top) (\hat{\theta}_i - \xi). \]

leads to

\[ \hat{\xi} = \frac{1}{N} \sum_{i=1}^{N} \hat{\theta}_i + \frac{1}{N} \sum_{i=1}^{N} W_i B_i (Y_i - \hat{Y}_i), \]

where \( \hat{Y}_i = B_i^\top \hat{\theta}_i. \)

- Requires models for \( P(T_i \mid X_i) \) and \( P(\theta_i \mid X_i) \)
- Asymptotically unbiased if either model is correct
- Related to “generalized regression” (GREG) estimators from survey literature (Deville and Särndal, 1992), but with matrix weights
Method # 5: Prediction from weighted regression

Like regression prediction (Method 3), except that we apply the weight matrices $W_i = \left[ E(B_iB_i^\top | X_i) \right]^{-1}$ when estimating $\hat{\theta}_i = E(\theta_i | X_i)$.

- Requires models for $P(T_i | X_i)$ and $P(\theta_i | X_i)$
- Asymptotically unbiased if either model is correct
- Performance is similar to Method 4 in large samples, but very unstable in small samples
Method # 6: Propensity-spline prediction

• If $P(T_i = t \mid X_i)$ depends on $X_i$ only through a $\psi(X_i)$ (typically of smaller dimension), then $\psi(X_i)$ is a propensity function (PF) (Imai and van Dyk, 2004)

• For example, if $T_i \mid X_i \sim N(X_i^\top \beta, \sigma^2)$, then $\psi = (\beta, \sigma^2)$, then the linear predictor $X_i^\top \beta$ is a PF.

• Inspired by Little and An (2004), build a rich prediction model for $\hat{\theta}_i = E(\theta_i \mid X_i)$ (Method 3), but include a spline basis for $\psi(X_i)$ as additional predictors

• Imai and van Dyk (2004) reduce the model to $\hat{\theta}_i = E(\theta_i \mid \psi(X_i))$, but this reduction is unnecessary and inefficient
3. Simulation study

Evaluate performance of estimators samples of $N = 200$ and $N = 1,000$ from a population where the true ADRF is constant.
Simulation study (continued)

- \( Y_i(t) = \theta_{i1} + \theta_{i2} t \)
- \( (\theta_{i1}, \theta_{i2}, T_i)^\top \) is multivariate normal with mean vector 
  \( (\xi_1, \xi_2, \kappa)^\top = (50, 0, 12)^\top \) and covariance matrix

\[
\begin{bmatrix}
\omega_{11} & \omega_{12} & \omega_{13} \\
\omega_{12} & \omega_{22} & \omega_{23} \\
\omega_{13} & \omega_{23} & \omega_{33}
\end{bmatrix}
= \begin{bmatrix}
51.0 & 3.80 & 5.92 \\
3.80 & 0.55 & 0.51 \\
5.92 & 0.51 & 2.02
\end{bmatrix}
\]

- \( \theta_{i1}, \theta_{i2} \) and \( T_i \) are linearly related to “true” normally distributed covariates \( A_{i1}^*, \ldots, A_{i8}^* \) which are hidden from view; analyst sees transformed versions \( A_{i1}, \ldots, A_{i8} \) which are skewed, bounded and binary, leading to model misspecification.
Simulation results

Performance of estimators for $\xi_2$ over 1,000 samples from the artificial population using incorrect $Y$-models and misspecified but rich $T$-models.

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Method</th>
<th>Bias</th>
<th>Var.</th>
<th>% Bias</th>
<th>RMSE</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N = 200$</td>
<td><em>Prima facie</em></td>
<td>7.05</td>
<td>0.35</td>
<td>1,190</td>
<td>7.08</td>
<td>7.06</td>
</tr>
<tr>
<td></td>
<td>Importance weighting</td>
<td>3.87</td>
<td>2.54</td>
<td>243</td>
<td>4.19</td>
<td>4.10</td>
</tr>
<tr>
<td></td>
<td>Inverse second-moment weighting</td>
<td>0.225</td>
<td>0.504</td>
<td>32</td>
<td>0.745</td>
<td>0.491</td>
</tr>
<tr>
<td></td>
<td>Regression prediction</td>
<td>0.616</td>
<td>0.500</td>
<td>87</td>
<td>0.938</td>
<td>0.660</td>
</tr>
<tr>
<td></td>
<td>Prediction + residual bias correction</td>
<td>0.268</td>
<td>0.496</td>
<td>38</td>
<td>0.753</td>
<td>0.493</td>
</tr>
<tr>
<td></td>
<td>Prediction from weighted regression</td>
<td>0.249</td>
<td>14.9</td>
<td>6</td>
<td>3.87</td>
<td>0.570</td>
</tr>
<tr>
<td></td>
<td>Propensity-spline prediction</td>
<td>0.166</td>
<td>0.585</td>
<td>22</td>
<td>0.783</td>
<td>0.531</td>
</tr>
<tr>
<td>$N = 1,000$</td>
<td><em>Prima facie</em></td>
<td>7.06</td>
<td>0.063</td>
<td>2,820</td>
<td>7.07</td>
<td>7.05</td>
</tr>
<tr>
<td></td>
<td>Importance weighting</td>
<td>2.75</td>
<td>1.89</td>
<td>200</td>
<td>3.07</td>
<td>3.07</td>
</tr>
<tr>
<td></td>
<td>Inverse second-moment weighting</td>
<td>0.144</td>
<td>0.083</td>
<td>50</td>
<td>0.322</td>
<td>0.215</td>
</tr>
<tr>
<td></td>
<td>Regression prediction</td>
<td>0.660</td>
<td>0.089</td>
<td>221</td>
<td>0.725</td>
<td>0.661</td>
</tr>
<tr>
<td></td>
<td>Prediction + residual bias correction</td>
<td>0.158</td>
<td>0.082</td>
<td>55</td>
<td>0.327</td>
<td>0.225</td>
</tr>
<tr>
<td></td>
<td>Prediction from weighted regression</td>
<td>0.133</td>
<td>0.087</td>
<td>45</td>
<td>0.324</td>
<td>0.219</td>
</tr>
<tr>
<td></td>
<td>Propensity-spline prediction</td>
<td>0.106</td>
<td>0.084</td>
<td>36</td>
<td>0.308</td>
<td>0.204</td>
</tr>
</tbody>
</table>

% Bias = 100 × Bias / $\sqrt{\text{Var}}$, MAE = median absolute error
Simulation summary

- Don’t use importance weighting
- Regression prediction isn’t terrible, but it seems more susceptible to model failure than other methods
- Inverse second-moment weighting uses only a $T$-model, but is surprisingly efficient and robust when the ADRF is linear
- Weighted regression prediction is unstable in small samples
- Other dual-model methods (prediction + bias correction, propensity-spline prediction) are competitive
Example: Smoking and Medical Expenditures

- Previously analyzed by Imai and van Dyk (2004)
- Data from National Medical Expenditure Survey (NMES)
- Multistage cluster sample from area frame, with oversampling of certain groups
- Interviews in 1987 from persons in sampled HHs, with expenditures obtained from health care providers
- Use smokers over the age of 18 (roughly 10,000 in sample)
- Examine relationship between $Y_i =$ expenditures and $T_i =$ packyears, a measure of cumulative lifetime smoking
4. Example: Smoking and Medical Expenditures

• Estimate the effect of $T_i = \text{packyears}$ on $Y_i = \text{expenditures}$, supposing that the ADRF is linear (!)

• Potential confounders include age, sex, race/ethnicity, marital status, education, region, income/poverty, seatbelt use
Complications

- Complex sample design, with strata, clusters and oversampling
  - Modify the estimating functions to $c_i w_i U_i$, where $c_i = 1$ for smokers and 0 otherwise, and $w_i$ is the survey weight (number of pop. persons represented by the sampled person)
  - Compute standard errors by a linearization (sandwich) method appropriate for general “with replacement” designs (Binder, 1983)

- Smoking status and packyears are missing for 11% of the sample
  - We multiply imputed them $M = 25$ times under a two-part regression model
  - Recovers an additional 1,000 smokers
Prima facie estimate

slope = 25.08  (SE=2.20)

• No adjustment for potential confounders
Inverse second-moment estimate

- Model the treatment using a heteroscedastic linear regression for cube root of packyears, with variance \( \propto \text{mean}^{1.8} \)
- Include all main effects and many two-way interactions (50 predictors)

\[
\text{slope} = 10.22 \quad (\text{SE}=3.06)
\]
5. Extensions

- Our parametric assumption maps each $Y_i(t)$ to a random vector $\theta_i \in \mathbb{R}^k$, leading to a wide variety of estimation methods.

- Assumed linear form $Y_i(t) = b(t) \theta_i$ is key, because the ADRF is completely determined by $E(\theta_i)$.

- Generalizations to nonlinear forms will be tricky, because they depend on other aspects of the distribution of $\theta_i$’s (e.g., covariances) which are more difficult to estimate from $Y_i$.

- Generalizations to discrete responses will also be tricky, because the $\theta_i$’s will no longer be coarsened to hyperplanes.

Thanks for listening!