Causal Inference, Propensity Scores, and Odds Ratios

Travis Loux

Department of Statistics and Applied Probability
University of California, Santa Barbara

April 11, 2012
Outline

Part I: Causal Inference
► Defining cause and effect
► The counterfactual model
► Exchangeability
► Randomization
► Observational studies
► Propensity scores

Part II: Estimating the Marginal Odds Ratio
► Odds ratio definitions and estimation
► Estimation by propensity scores
► Simulations
Part I: Causal Inference
Causal Inference

What is a cause?

- “We may define a cause to be an object followed by another... where, if the first object had not been, the second never had existed”
  - Hume (1748)

- “One event is the cause of another if the appearance of the first event is followed with a high probability by the second, and there is no third event that we can use to factor out the probability relationship between the first and second events.”
  - Suppes (1970)

- Inference should focus on “studying the effects of causes rather than the traditional approach of trying to define what the cause of an effect is”
  - Holland (1986)
Causal Inference

What is an effect? The counterfactual model

The effect of an exposure is the outcome of that exposure relative to the outcome which would have been observed had another exposure occurred.

- For a unit \( u \), \( Z(u) \) is the binary exposure associated with \( u \)
  - \( Z(u) = 1 \) if \( u \) is exposed; \( Z(u) = 0 \) if \( u \) is not exposed
  - \( Y_{Z(u)}(u) \) is the response of unit \( u \) with exposure \( Z(u) \)
  - Exposure leads to the effect \( Y_1(u) - Y_0(u) \)

Causal inference is concerned with unbiased estimation of this effect

- We cannot observe both \( Y_1(u) \) and \( Y_0(u) \)
- We observe \( Y(u) \): \( Y = ZY_1 + (1 - Z)Y_0 \)
The Counterfactual Model

A Population

A hypothetical population would then look like the following:

<table>
<thead>
<tr>
<th>Unit</th>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$Z$</th>
<th>$Y_1$</th>
<th>$Y_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.09</td>
<td>-1.80</td>
<td>1.86</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>-0.51</td>
<td>0.25</td>
<td>1.62</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>-0.48</td>
<td>0.69</td>
<td>-0.35</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>-1.48</td>
<td>-1.76</td>
<td>-0.47</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>0.55</td>
<td>0.50</td>
<td>-0.82</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

... ... ... ... ... ...

where the observed $Y$ for each unit is colored blue.

- Here $Y$ is binary. This does not need to be the case.
Exchangeability

We focus on estimating $\Delta = E(Y_1 - Y_0) = E(Y_1) - E(Y_0)$, but we can only observe $Y_1$ for those units which are exposed. Since the counterfactual outcome is not observable from all units, it is necessary to substitute a reference group in order to estimate the unobservable quantities.

- One must use the non-exposed population as a proxy for what would have happened in the exposed population if they had not been exposed.

- Similarly, one must use the exposed population as a proxy for what would have happened in the non-exposed population had they been exposed.

The degree to which outcome under no exposure for the unexposed differs from the potential outcome under no exposure for the exposed is the degree to which bias enters the estimate of the causal effect.
Mathematically speaking, we can estimate

- $E(Y_1 \mid Z = 1)$, the average outcome under exposure for the exposed, and
- $E(Y_0 \mid Z = 0)$, the average outcome under no exposure for the unexposed

We encounter bias when $E(Y_1 \mid Z = 1) \neq E(Y_1)$ or $E(Y_0 \mid Z = 0) \neq E(Y_0)$.

If it’s true that $E(Y_1 \mid Z = 1) = E(Y_1 \mid Z = 0) = E(Y_1)$ and similarly for $Y_0$, the exposed and unexposed units are said to be *exchangeable* and their respective parameters *comparable*.
Randomization

In an experimental setting, exposure (now, treatment) can be assigned to units under the control of the researcher.

- By randomly assigning units to treatment or control, the researcher has reason to believe that the two groups are exchangeable.

- R.A. Fisher calls randomization “the reasoned basis for inference.”

Exchangeability does not hold in every case, but only as an expectation over all possible randomizations.
Randomization

An example

Suppose we have four units, two of which we can randomize to treatment. Since treatment assignment is under an experimenter’s control, we can make sure all treatment permutations are equally likely. If treatment leads to an increase of 2 in the response, the possible results can be summarized in the table:

<table>
<thead>
<tr>
<th>$u_1$</th>
<th>$u_2$</th>
<th>$u_3$</th>
<th>$u_4$</th>
<th>$\hat{\Delta}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>9</td>
<td>3</td>
<td>6</td>
<td>3.5</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>5</td>
<td>6</td>
<td>-0.5</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>3</td>
<td>8</td>
<td>2.5</td>
</tr>
<tr>
<td>5</td>
<td>9</td>
<td>5</td>
<td>6</td>
<td>1.5</td>
</tr>
<tr>
<td>5</td>
<td>9</td>
<td>3</td>
<td>8</td>
<td>4.5</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
<td>5</td>
<td>8</td>
<td>0.5</td>
</tr>
</tbody>
</table>

Averaging over treatment permutations:

$$E(\hat{\Delta}) = \sum_{r \in R} \hat{\Delta}(r) P(r) = \frac{1}{6} \sum_{r \in R} \hat{\Delta}(r) = 2$$
Observational Studies

When experiments are infeasible, observational studies are conducted

▶ Exposures are observed and recorded, but the researcher has no influence over assignment

The exposure mechanism is unknown

▶ Probability of exposure is unknown, so expectations cannot be calculated

▶ Inference is conditional on observed distributions

▶ Must account for confounders - variables which are associated with both exposure and outcome
Observational Studies

An example

Suppose outcome follows the model $Y_i = \beta_0 + \beta_1 x_i + \Delta z_i + \epsilon_i$. Comparing the averages of the exposed and unexposed samples, we get

$$\bar{Y}_1 - \bar{Y}_0 = \frac{1}{n_1} \sum_{i: Z_i=1} (\beta_0 + \beta_1 x_i + \Delta + \epsilon_i) - \frac{1}{n_0} \sum_{i: Z_i=0} (\beta_0 + \beta_1 x_i + \epsilon_i)$$

$$= \Delta + \beta_1 (\bar{x}_1 - \bar{x}_0) + (\bar{\epsilon}_1 - \bar{\epsilon}_0)$$

- This estimate is biased for $\Delta$ if $X$ is unbalanced in the two populations:

$$E(\bar{Y}_1 - \bar{Y}_0) = \Delta + \beta_1 (\mu_1 - \mu_0)$$

where

$$\mu_z = E(X \mid Z = z)$$
Measuring covariates can help reduce the bias due to confounding. We have *strongly ignorable treatment assignment* (SITA) conditional on a set of covariates $X$ when

1. $(Y_1, Y_0) \perp Z \mid X$, and
2. $0 < P(Z_i = 1 \mid X = x_i) < 1$ for all $i$

Intuitively, SITA is the assumption that $X$ contains all variables confounding the relationship between $Z$ and $Y$. 
Propensity Score
Definition and consequences

The propensity score (Rosenbaum and Rubin 1983) is defined as the probability of exposure conditional on observed covariates:

\[ e(x) = P(Z = 1 \mid X = x) \]

Exposed and unexposed populations with the same PS have the same distribution of observed covariates.
When we have SITA conditional on \( X \), then we also have SITA on \( e(x) \):

1. \((Y_1, Y_0) \perp Z \mid e(X)\), and
2. \(0 < e(x_i) < 1\) for all \( i \)

In this sense, we can think of the PS as a dimension-reduction which contains all the information about confounding due to \( X \).

- The PS does not balance unobserved covariates.
Propensity Score
Estimation and Uses

The most common method for estimating the PS is logistic regression

- Other methods include GAMs, CART, random forests, boosting, neural networks
- Essentially a classification problem

Once the PS is estimated, units can be matched or subclassified on the PS to reduce the bias

- Using as few as five subclasses is common and has been shown to work relatively well
Propensity Score
Estimation and Uses

The inverse PS can also be used as a weight:

\[ \hat{\mu}_{IPW_{1,1}} = n^{-1} \sum \frac{Z_i Y_i}{e(X_i)} \]

is an unbiased estimator of \( E(Y_1) \) and

\[ \hat{\mu}_{IPW_{1,0}} = n^{-1} \sum \frac{(1 - Z_i) Y_i}{1 - e(X_i)} \]

is an unbiased estimator of \( E(Y_0) \)

- This estimate is highly sensitive to extreme PS values and misspecified PS models
Propensity Score
Estimation and Uses

We can improve on $\hat{\mu}_{IPW1,1}$ and $\hat{\mu}_{IPW1,0}$ with the following:

$$\hat{\mu}_{IPW2,1} = (\sum_i \frac{Z_i}{e(X_i)})^{-1} \sum_i \frac{Z_i Y_i}{e(X_i)}$$

and

$$\hat{\mu}_{IPW2,0} = (\sum_i \frac{1 - Z_i}{1 - e(X_i)})^{-1} \sum_i \frac{(1 - Z_i) Y_i}{1 - e(X_i)}$$

- Less sensitive to extreme PS
- If $Y$ is binary, $\hat{\mu}_{IPW2,1}$ and $\hat{\mu}_{IPW2,0}$ are constrained to be between 0 and 1
Another weighting-based option is doubly robust estimation (Robins et al., 1994) where

\[
\hat{\mu}_{DR,1} = n^{-1} \sum_i \frac{Z_i Y_i - (Z_i - e(X_i)) m_1(X_i)}{e(X_i)}
\]

and

\[
\hat{\mu}_{DR,0} = n^{-1} \sum_i \frac{(1 - Z_i) Y_i + (Z_i - e(X_i)) m_0(X_i)}{1 - e(X_i)}
\]

estimate \( E(Y_1) \) and \( E(Y_0) \), respectively. The models \( m_z(X) \) model the outcome in the subgroup defined by \( Z = z \) on the basis of \( X \).
Doubly robust estimators have two useful properties:

- Only need one of $e(X)$ or $m_z(X)$ to be correctly specified for estimates to be unbiased
- If both models are unbiased, $\hat{\mu}_{DR,1}$ and $\hat{\mu}_{DR,0}$ are efficient in the class of semi-parametric estimators
  - If one model is misspecified, efficiency is not guaranteed
Part II: Estimating the Marginal Odds Ratio
Odds Ratios
Marginal Odds Ratio

The marginal odds ratio can be obtained by comparing the odds of response in the population if everyone is exposed

$$Odds_{exp} = \frac{P(Y_1 = 1)}{P(Y_1 = 0)}$$

to the odds of response if everyone is not exposed

$$Odds_{unexp} = \frac{P(Y_0 = 1)}{P(Y_0 = 0)}$$

Looking at the ratio $Odds_{exp} / Odds_{unexp}$, the marginal odds ratio is

$$\psi_{marg} = \frac{P(Y_1 = 1) \times P(Y_0 = 0)}{P(Y_1 = 0) \times P(Y_0 = 1)}$$
The marginal odds ratio is often approximated by estimating the crude odds ratio:

$$\psi_{\text{crude}} = \frac{P(Y_1 = 1 \mid Z = 1) \times P(Y_0 = 0 \mid Z = 0)}{P(Y_1 = 0 \mid Z = 1) \times P(Y_0 = 1 \mid Z = 0)}$$

- When confounding is present

$$P(Y_1 = 1 \mid Z = 1) \neq P(Y_1 = 1)$$
$$P(Y_0 = 1 \mid Z = 0) \neq P(Y_0 = 1)$$

so the crude and marginal odds ratio can be different values.
The conditional odds ratio is defined as

\[ \psi_{\text{cond}}(x) = \frac{P(Y = 1 \mid Z = 1, X = x) \times P(Y = 0 \mid Z = 0, X = x)}{P(Y = 0 \mid Z = 1, X = x) \times P(Y = 1 \mid Z = 0, X = x)} \]

With the assumption of strongly ignorable treatment assignment, we can simplify:

\[ \psi_{\text{cond}}(x) = \frac{P(Y_1 = 1 \mid X = x) \times P(Y_0 = 0 \mid X = x)}{P(Y_1 = 0 \mid X = x) \times P(Y_0 = 1 \mid X = x)} \]
In the linear model $E(Y \mid X, Z) = \beta^T X + \Delta Z$, the conditional effect of exposure is equal to the marginal effect:

$$
E(Y_1 \mid X) - E(Y_0 \mid X) = (\beta^T X + \Delta) - (\beta^T X)
$$

$$
= \Delta
$$

$$
E(Y_1) - E(Y_0) = E(E_X(\beta^T X + \Delta)) - E(E_X(\beta^T X))
$$

$$
= (\beta^T \mu_X + \Delta) - (\beta^T \mu_X)
$$

$$
= \Delta
$$

- The linear model is *collapsible*

In the logistic model, this property does not hold

- The difference between the marginal and conditional effects means estimators are needed for each
Odds Ratios
Standard Estimation

If the data follow a logistic model with

\[
P(Y = 1 \mid X = x, Z = z) = \frac{\exp \left\{ \beta_0 + \sum_{j=1}^{p} \beta_j x_j + \alpha z \right\}}{1 + \exp \left\{ \beta_0 + \sum_{j=1}^{p} \beta_j x_j + \alpha z \right\}}
\]

the conditional odds ratio is constant across \( X \) with \( \psi_{\text{cond}}(x) = \psi_{\text{cond}} = e^\alpha \).

Logistic regression will lead to an unbiased and asymptotically efficient estimator of \( \psi_{\text{cond}}(x) \)

- In order for the estimate to be unbiased, \( X \) must contain all predictors of \( Y \), not just the confounders (Gail, Wieand, and Piantadosi, 1984)
Odds Ratios
Standard Estimation

Subclassifying observations based on covariates into multiple $2 \times 2$ tables defined by $X$, we can estimate the odds ratio by the Mantel-Haenszel (MH) estimator

- If the $k^{th}$ table takes the form

\[
\begin{array}{c|cc|c|c|l}
 & \multicolumn{2}{c|}{Z = 1} & Z = 0 & \\
\hline \hline
Y = 1 & a_k & b_k & \hline
Y = 0 & c_k & d_k & \hline
\hline
\end{array}
\]

the MH estimator is defined as

\[
\hat{\psi}_{MH} = \sum_k \frac{a_k d_k}{n_k} = \sum_k \frac{b_k c_k}{n_k} = \sum_k \frac{n_k}{b_k c_k} \hat{\psi}_k
\]

But what odds ratio are we estimating?
Odds Ratios
Standard Estimation

- If the covariates are constant within each subclass, $\hat{\psi}_{MH}$ estimates the conditional odds ratio, assuming $\psi_{cond}$ is constant
  
  - Examples: Subclassification defined by categorical covariates; perfect matching on covariates

- If covariates vary within the subclasses, $\hat{\psi}_{MH}$ will be biased due to non-collapsibility (eg. Greenland, Robins, and Pearl, 1999)
  
  - Extreme case: When data is summarized in one table, $\hat{\psi}_{MH}$ estimates the crude odds ratio

- Cochran (1968) shows that using 5 subclasses removes approximately 90% of bias due to confounding in linear models
Matching on the Propensity Score

In 1-to-1 matching, the MH estimator simplifies:

▶ The $k^{th}$ table has the form

<table>
<thead>
<tr>
<th></th>
<th>$Z = 1$</th>
<th>$Z = 0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y = 1$</td>
<td>$a_k$</td>
<td>$b_k$</td>
</tr>
<tr>
<td>$Y = 0$</td>
<td>$1 - a_k$</td>
<td>$1 - b_k$</td>
</tr>
<tr>
<td></td>
<td>$1$</td>
<td>$1$</td>
</tr>
<tr>
<td></td>
<td>$2$</td>
<td></td>
</tr>
</tbody>
</table>

the MH estimator becomes

$$\hat{\psi}_{MH} = \frac{\sum_k a_k (1 - b_k)}{\sum_k (1 - a_k) b_k}$$

▶ Similar simplifications hold for case-control designs
Matching on the Propensity Score

As the number of matched pairs increases, it can be shown that

\[ \hat{\psi}_{MH} \rightarrow \tilde{\psi} \]

where \( \tilde{\psi} \) depends on the strength of the relationship between the function of the covariates in the outcome and propensity score models.
Matching on the Propensity Score

When the exposure follows the model \( \logit P(Z = 1 \mid X) = \gamma^T X \), and the outcome follows the model \( \logit P(Y = 1 \mid X, Z) = \beta^T X + \alpha Z \), the value of \( \tilde{\psi} \) depends on the relationship between \( \gamma^T X \) and \( \beta^T X \):

- If \( \beta^T X = f(\gamma^T X) \) for some \( f \), then \( \tilde{\psi} = \psi_{\text{cond}} \)
  - \( \beta^T X \) is constant in domains defined by \( e(X) \)

- If \( \gamma^T X \) and \( \beta^T X \) are orthogonal, then \( \tilde{\psi} = \psi_{\text{marg}} \)
  - Let \( H = h(X) = \beta^T X \). Then \( F_{H|e(X)} = F_H \)

- If \( 0 < |\rho(\gamma^T X, \beta^T X)| < 1 \), \( \tilde{\psi} \) falls between \( \psi_{\text{cond}} \) and \( \psi_{\text{marg}} \)
Weighting by the Propensity Score

We can use the weighted estimators of $E(Y_1)$ and $E(Y_0)$ shown previously to estimate the odds ratio:

$$\hat{\psi}_{IPW_1} = \frac{\hat{\mu}_{IPW_1,1}(1 - \hat{\mu}_{IPW_1,0})}{(1 - \hat{\mu}_{IPW_1,1})\hat{\mu}_{IPW_1,0}}$$

$$\hat{\psi}_{IPW_2} = \frac{\hat{\mu}_{IPW_2,1}(1 - \hat{\mu}_{IPW_2,0})}{(1 - \hat{\mu}_{IPW_2,1})\hat{\mu}_{IPW_2,0}}$$

$$\hat{\psi}_{DR} = \frac{\hat{\mu}_{DR,1}(1 - \hat{\mu}_{DR,0})}{(1 - \hat{\mu}_{DR,1})\hat{\mu}_{DR,0}}$$
Simulations

Mechanics

- Created population of 2 million units:
  - Standard normal covariates \( X = (X_1, X_2, X_3) \)
  - Exposure \( Z \): \( \logit P(Z = 1 \mid X) = \gamma^T X \)
  - Potential outcome \( Y_1 \): \( \logit P(Y_1 = 1 \mid X) = \beta^T X + \alpha \)
  - Potential outcome \( Y_0 \): \( \logit P(Y_0 = 1 \mid X) = \beta^T X \)

- Took 10,000 independent samples of 2,000 observations each
  - Crude and logistic estimation
  - Subclassified and matched on PS
  - Weighted by inverse propensity score, \( 0.005 < e(x) < 0.995 \)
Simulations
No Causal Effect

![Box plot showing OR estimates for different methods]

- Conditional: 1.0073
- Crude: 2.2059
- Marginal: 1.0004
Simulations

Strong Correlation: $\rho(\gamma^T X, \beta^T X) \approx 1$
Simulations
Weak Correlation: $\rho(\gamma^T X, \beta^T X) = 0.008$
Simulations

Moderate Correlation: $\rho(\gamma^T X, \beta^T X) = 0.6413$

<table>
<thead>
<tr>
<th>Method</th>
<th>Conditional</th>
<th>Crude</th>
<th>Marginal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crude OR</td>
<td>2.9965</td>
<td>4.0677</td>
<td>1.8019</td>
</tr>
<tr>
<td>Logistic</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 classes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PS-matched</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IPW1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IPW2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DR</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Simulations

Correlation and Matching Bias

Using scaled bias $= \frac{\hat{E}(\hat{\psi}_{MH}) - \psi_{marg}}{\psi_{cond} - \psi_{marg}}$ for 28 simulations with various $\gamma$ and $\beta$: 
Simulations
Omitting Confounders from Propensity Score

When omitting confounders, $\hat{\psi}_{IPW2}$ becomes biased while $\hat{\psi}_{DR}$ shows its double robustness property:
Simulations
Omitting Covariates from Propensity Score and Outcome Models

When omitting covariates from both models, both $\hat{\psi}_{IPW2}$ and $\hat{\psi}_{DR}$ are biased. Neither is categorically better:
Discussion of Marginal Odds Ratio

- Matching on the propensity score leads to an estimate which is consistent for neither the conditional nor marginal odds ratio.

- Inverse propensity weighting yields an estimate which is unbiased for the marginal odds ratio.

- $\hat{\psi}_{DR}$ can be as biased, or more so, than $\hat{\psi}_{IPW2}$ when both the propensity score and outcome models are misspecified.

- Doubly robust estimation is preferred in most cases with large ($n \geq 500$) sample sizes.