Introduction to Causal Inference

Lecture 7

Unconfounded Treatment Assignment & Estimating Propensity Score

邓婉璐

WANLUDENG@TSINGHUA.EDU.CN

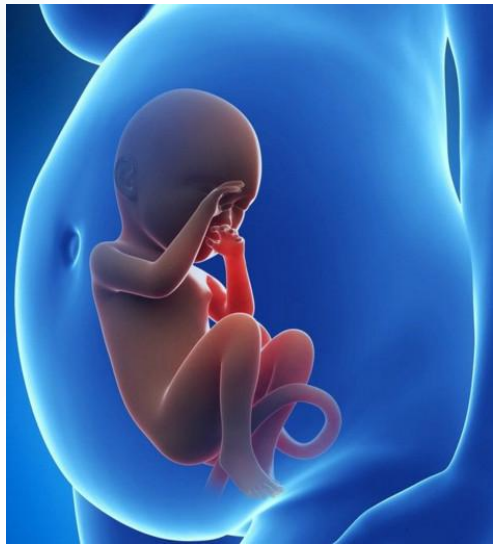


清华大学统计与数据科学系

The Barbituate Exposure Data

- A study of the effect of prenatal exposure to barbituates (Reinisch, Sanders, Mortenson, and Rubin, 1995)
  - ✓ 巴比妥类的药物又称之为巴比妥酸盐，是一种作用于中枢神经系统的镇静剂
- Treatment and control: exposed *in utero* to substantial amounts of barbituates or not
- The outcomes of interest: reflecting cognitive development







Outline

- Assumptions under Observational Study
- Balancing Score and Propensity Score
- Overview for Inference Strategies
- Estimating Propensity Score (PS)

Assumptions under Observational Study



Unconfoundedness is an assumption

Unconfoundedness

$$\Pr(W_i = 1 | Y_i(0), Y_i(1), X_i) = \Pr(W_i = 1 | X_i) \quad \Leftrightarrow \quad W_i \perp\!\!\!\perp \big(Y_i(0), Y_i(1)\big) \,\big|\, X_i$$

$$\begin{aligned}
\big(Y_i(0) \,\big|\, W_i = 1, X_i\big) &\sim \big(Y_i(0) \,\big|\, W_i = 0, X_i\big), \quad \text{for } i = 1, \ldots, N, \\
\big(Y_i(1) \,\big|\, W_i = 0, X_i\big) &\sim \big(Y_i(1) \,\big|\, W_i = 1, X_i\big), \quad \text{for } i = 1, \ldots, N.
\end{aligned}$$

Review: Randomized Experiments



- **Randomized Experiment**: *A randomized experiment is an assignment mechanism that*
  1) *is probabilistic, and* ----▸ random experiment
  2) *has a known functional form that is controlled by the researcher.*
- **Classical Randomized Experiment**: a special case of randomized experiment whose *assignment mechanism is*
  1) *individualistic, and*
  2) *unconfounded.*
- Four important special cases:
  1) Bernoulli trials
  2) completely randomized experiments
  3) stratified randomized experiments (randomized blocks)
  4) paired randomized experiments

Review: Regular Assignment Mechanisms



- **Observational Study**: An assignment mechanism corresponds to an observational study if the functional form of the assignment mechanism is **unknown**.
  - → The key difference from experiments
- **Regular Assignment Mechanism**: An assignment mechanism is **regular** if it is
  1) individualistic:
  2) probabilistic:
  3) unconfounded:
- **Irregular Assignment Mechanism**:
  - ✧ Assignment to treatment may differ for some units from the receipt of treatment, i.e., noncompliance
  - ✧ We often assume that assignment to treatment itself is unconfounded, but allow receipt of treatment to be confounded

Review: Assignment Mechanism/Probability

Definition 3.1 (Assignment Mechanism)
Given a population of N units, the assignment mechanism is a row-exchangeable function $\Pr(\mathbf{W}|\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1))$, *taking on values in* $[0, 1]$, *satisfying*

$$\sum_{\mathbf{W}\in\{0,1\}^N} \Pr(\mathbf{W}|\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) = 1, \quad \textit{for all } \mathbf{X}, \mathbf{Y}(0), \textit{ and } \mathbf{Y}(1).$$

Definition 3.2 (Unit Assignment Probability)
The unit-level assignment probability for unit i is

$$E(W_i|\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) = p_i(\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) = \sum_{\mathbf{W}:W_i=1} \Pr(\mathbf{W}|\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)).$$

Review: Restrictions on Assignment Mechanism (I)

- Individualistic assignment: a unit's assignment probability does not depend on the values of covariates and potential outcomes for other units.

Definition 3.4 (Individualistic Assignment)

An assignment mechanism $\Pr(\mathbf{W}|\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1))$ *is individualistic if, for some function* $q(\cdot) \in [0, 1]$,

$$p_i(\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) = q(X_i, Y_i(0), Y_i(1)), \quad \text{for all } i = 1, \ldots, N,$$

and

$$\Pr(\mathbf{W}|\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) = c \cdot \prod_{i=1}^{N} q(X_i, Y_i(0), Y_i(1))^{W_i} \left(1 - q(X_i, Y_i(0), Y_i(1))\right)^{1-W_i},$$

for $(\mathbf{W}, \mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) \in \mathbb{A}$, *for some set* $\mathbb{A}$, *and zero elsewhere (c is the constant that ensures that the probabilities sum to unity).*



Review: Restrictions on Assignment Mechanism (II)

- Probabilistic assignment:
  - every unit has the possibility of being assigned to the active treatment and the possibility of being assigned to the control treatment.

Definition 3.5 (Probabilistic Assignment)
An assignment mechanism $\Pr(\mathbf{W}|\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1))$ *is probabilistic if the probability of assignment to treatment for unit i is strictly between zero and one:*

$$0 < p_i(\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) < 1, \quad \text{for each possible } \mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1),$$

for all $i = 1, \ldots, N$.

Review: Restrictions on Assignment Mechanism (III)

- Unconfounded assignment:
  - disallows dependence of the assignment mechanism on the potential outcomes.

Definition 3.6 (Unconfounded Assignment)

An assignment mechanism is unconfounded if it does not depend on the potential outcomes:

$$\Pr(\mathbf{W}|\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) = \Pr(\mathbf{W}|\mathbf{X}, \mathbf{Y}'(0), \mathbf{Y}'(1)),$$

$$\Rightarrow \quad \Pr(\mathbf{W}|\mathbf{X},\mathbf{Y}(0),\mathbf{Y}(1)) = \Pr(\mathbf{W}|\mathbf{X})$$

for all $\mathbf{W}$, $\mathbf{X}$, $\mathbf{Y}(0)$, $\mathbf{Y}(1)$, $\mathbf{Y}'(0)$, *and* $\mathbf{Y}'(1)$.

Unconfoundedness Is Not Testable

Then can we still use randomized experiments for golden standard?

- There is no information in the data that can tell us that unconfoundedness does not hold, even in settings with a large number of units
- Any assertion that unconfoundedness does not hold must rely on additional information

Unconfoundedness

$$\Pr(W_i = 1 | Y_i(0), Y_i(1), X_i) = \Pr(W_i = 1 | X_i) \iff W_i \perp\!\!\!\perp \big(Y_i(0), Y_i(1)\big) \,\big|\, X_i$$

$$\Big(Y_i(0) \,\Big|\, W_i = 1, X_i\Big) \sim \Big(Y_i(0) \,\Big|\, W_i = 0, X_i\Big), \quad \text{for } i = 1, \ldots, N,$$
$$\Big(Y_i(1) \,\Big|\, W_i = 0, X_i\Big) \sim \Big(Y_i(1) \,\Big|\, W_i = 1, X_i\Big), \quad \text{for } i = 1, \ldots, N.$$

$$\Big(Y_i^{\text{mis}} \,\Big|\, W_i = w, X_i\Big) \sim \Big(Y_i^{\text{obs}} \,\Big|\, W_i = 1 - w, X_i\Big), \quad \text{for } i = 1, \ldots, N.$$

- All involves Y^{mis}
- Not testable

Regular Assignment Mechanisms



- **Observational Study**: An assignment mechanism corresponds to an observational study if the functional form of the assignment mechanism is **unknown**.

  → The key difference from experiments

- **Regular Assignment Mechanism**: An assignment mechanism is regular if it is

  1) individualistic: $0 < p_i(\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) < 1, \quad \text{for } i = 1, \ldots, N.$

  2) probabilistic: $p_i(\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) = q(X_i, Y_i(0), Y_i(1)), \quad \text{for } i = 1, \ldots, N.$

  $$\Pr(\mathbf{W} \mid \mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) = c \cdot \prod_{i=1}^{N} q(X_i, Y_i(0), Y_i(1))^{W_i} \cdot (1 - q(X_i, Y_i(0), Y_i(1)))^{1-W_i}$$

  3) unconfounded:

  $$\Pr(W_i = 1 \mid Y_i(0), Y_i(1), X_i) = \Pr(W_i = 1 \mid X_i) = e(X_i) \iff W_i \perp\!\!\!\perp \big(Y_i(0), Y_i(1)\big) \,\big|\, X_i$$

Propensity score: $e(x) = q(x)$ → Probabilistic assignment ⟺ $0 < e(x) < 1$

Assignment mechanism: $\Pr(\mathbf{W} \mid \mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) = c \cdot \prod_{i=1}^{N} e(X_i)^{W_i} \cdot (1 - e(X_i))^{1-W_i}$

RAMs vs SREs

- Common features
  - Base inference on strata defined by covariates: $X_i = x$
  - Stratum-level causal effects are identifiable & estimable

$$\hat{\tau}^{\text{dif}}(f) = \bar{Y}_{\text{t}}^{\text{obs}}(f) - \bar{Y}_{\text{c}}^{\text{obs}}(f) \quad \Rightarrow \quad \tau_{\text{fs}}(f) = \frac{1}{N(f)} \sum_{i:X_i=f} \big(Y_i(1) - Y_i(0)\big)$$

$$\hat{\tau}^{\text{dif}}(m) = \bar{Y}_{\text{t}}^{\text{obs}}(m) - \bar{Y}_{\text{c}}^{\text{obs}}(m) \quad \Rightarrow \quad \tau_{\text{fs}}(m) = \frac{1}{N(m)} \sum_{i:X_i=m} \big(Y_i(1) - Y_i(0)\big)$$

$$\hat{\tau}^{\text{strat}} = \frac{N(f)}{N(f)+N(m)} \cdot \hat{\tau}^{\text{dif}}(f) + \frac{N(m)}{N(f)+N(m)} \cdot \hat{\tau}^{\text{dif}}(m). \quad \Rightarrow \quad \tau_{\text{fs}} = \frac{N(f)}{N(f)+N(m)} \cdot \tau_{\text{fs}}(f) + \frac{N(m)}{N(f)+N(m)} \cdot \tau_{\text{fs}}(m)$$

- Key differences
  - $e(j) = N_t(j)/N(j)$ are known in a SRE
  - But unknown in a RAM $\longrightarrow$ can be estimated by $\hat{e}(j) = N_t(j)/N(j)$

RAM can be treated as SREs with unknown propensity score in each stratum

Review: Propensity Score

Definition 3.3 (Finite Population Propensity Score)

The propensity score at x is the average unit assignment probability for units with $X_i = x$,

$$e(x) = \frac{1}{N(x)} \sum_{i:X_i=x} p_i(\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1))$$

where $N(x) = \#\{i = 1, \ldots, N | X_i = x\}$ *is the number of units with* $X_i = x$. *For values x with* $N(x) = 0$, *the propensity score is defined to be zero.*

- the average of the unit-level assignment probabilities for subpopulations with a common value of the covariates $X_i = x$.



Satisfied?

$\Pr(W_i = 1|X_i = x) = \mathbb{E}[W_i|X_i = x]$ be estimated by $\hat{e}(j) = N_t(j)/N(j)$

- What else influence the estimation / inference?
- Sample Size!
- If X_i takes on only a few values:
  - Looks fine
- If X_i takes on many values, such a saturated model is not feasible:
  - K binary covariates will result in 2^K subclasses
  - A continuous covariate will results in an infinite number of subclasses

Big Names

Randomized Experiments



R.A. Fisher
(1890-1962)



Jerzy Neyman
(1894-1981)

Observational Study



Donald B. Rubin
(1943-)

Balancing Score and Propensity Score



Balancing Scores

Definition 12.1 (Balancing Scores)
A balancing score $b(x)$ is a function of the covariates such that $W_i \perp\!\!\!\perp X_i \mid b(X_i)$.

Lemma 12.2 (Unconfoundedness Given a Balancing Score)
Suppose assignment to treatment is unconfounded. Then assignment is unconfounded given any balancing score: $W_i \perp\!\!\!\perp Y_i(0), Y_i(1) \mid b(X_i)$.

$$
\begin{aligned}
\Pr_W(W_i = 1 | Y_i(0), Y_i(1), b(X_i)) &= \mathbb{E}_W[W_i | Y_i(0), Y_i(1), b(X_i)] \\
&= \mathbb{E}\left[\mathbb{E}_W[W_i | Y_i(0), Y_i(1), X_i, b(X_i)] \middle| Y_i(0), Y_i(1), b(X_i)\right] \\
&\qquad \leftarrow W_i \perp\!\!\!\perp (Y_i(0), Y_i(1)) \mid X_i \\
&\qquad \mathbb{E}[W_i | X_i, b(X_i)] = \mathbb{E}[W_i | b(X_i)] \\
&= \mathbb{E}\left[\mathbb{E}_W[W_i | b(X_i)] \middle| Y_i(0), Y_i(1), b(X_i)\right] \\
&= \mathbb{E}[W_i | b(X_i)] = \Pr(W_i = 1 | b(X_i))
\end{aligned}
$$



Balancing Property of Propensity Score

Definition 12.1 (Balancing Scores)
A balancing score $b(x)$ is a function of the covariates such that $W_i \perp\!\!\!\perp X_i \mid b(X_i)$.

Lemma 12.1 (Balancing Property of the Propensity Score)
The propensity score is a balancing score.

Definition of propensity score $\Rightarrow$ $\Pr(W_i = 1|X_i, e(X_i)) = \Pr(W_i = 1|X_i) = e(X_i)$

$$\Pr(W_i = 1|e(X_i)) = \mathbb{E}[W_i|e(X_i)] = \mathbb{E}\left[\mathbb{E}[W_i|X_i, e(X_i)]|e(X_i)\right] = \mathbb{E}[e(X_i)|e(X_i)] = e(X_i)$$

$\Rightarrow$ $\Pr(W_i = 1|X_i, e(X_i)) = \Pr(W_i = 1|e(X_i))$ $\Leftrightarrow$ $W_i \perp\!\!\!\perp X_i \mid e(X_i)$

Toy Example

- Large class vs Small class (W). Influence on score (Y). Covariate is grade.(X) Sample size is listed in the table.
- PS is a balancing score.
- Why do we need balancing score?

W\X	1	2	3	4	5
0	40 (50)	20 (60)	60 (70)	80 (80)	160 (90)
1	60 (51)	30 (61)	90 (71)	20 (81)	40 (91)
e(x)	3/5	3/5	3/5	1/5	1/5



Coarsest Balancing Score

Lemma 12.3 (Coarseness of Balancing Scores)
The propensity score is the coarsest balancing score. That is, the propensity score is a function of every balancing score.

Let $b(x)$ be a balancing score

If we can *not* write the propensity score $e(x)$ as a function of the balancing score

⇩

Exist two values x and x' s.t. $b(x) = b(x')$ and $e(x) \neq e(x')$

⇩

$\Pr(W_i = 1|X_i = x) = e(x) \neq e(x') = \Pr(W_i = 1|X_i = x')$

⇩

W_i and X_i are not independent given $b(X_i) = b(x)$

⇩

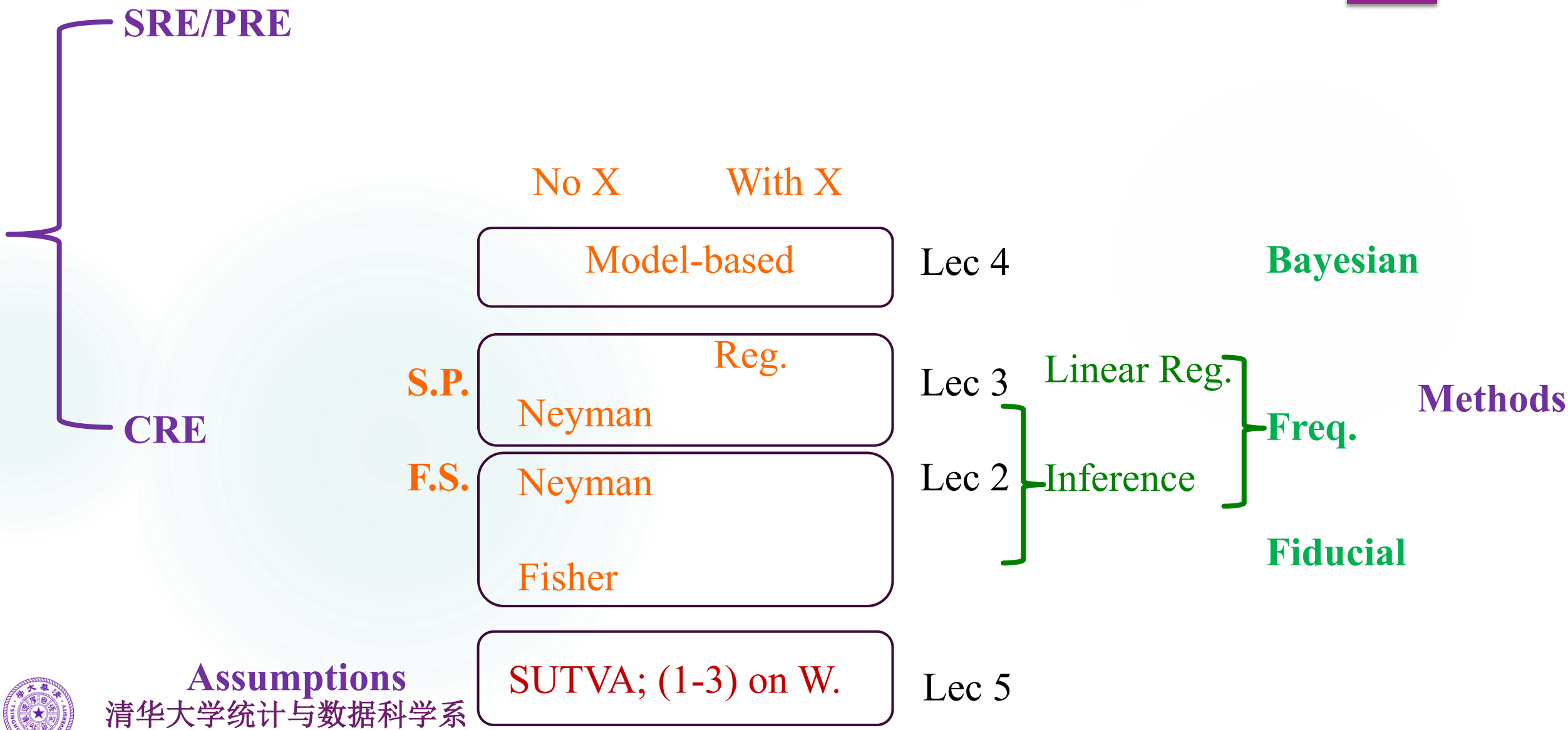
Violation the assumption that $b(x)$ is a balancing score



Overview for Inference Strategies



Review


SRE/PRE
CRE
No X
With X
Model-based
Lec 4
Bayesian
S.P.
Reg.
Neyman
Lec 3
Linear Reg.
F.S.
Neyman
Fisher
Lec 2
Inference
Freq.
Methods
Fiducial
SUTVA; (1-3) on W.
Lec 5
Assumptions

Strategies for Estimation

1. Model-based imputation
2. Regression estimators
3. Weighting estimators using the propensity score
4. Blocking estimators using the propensity score
5. Matching estimators using a distance metric
6. Mixed estimators: combine 2-5



M4: Blocking Estimators

- The sample is partitioned into subclasses based on the value of the estimated propensity score
- Within each subclass, the data can be analyzed as if they arose from a CRE

Subclass indicator: $B_i(j) = I(b_{j-1} < \hat{e}(X_i) < b_j)$ ⇒ Units within each subclass have similar propensity score

estimated propensity score

subclass boundaries

Data from a subclass ≈ data from a CRE

$$N(j) = \sum_{i=1}^{N} B_i(j)$$

$$\hat{\tau}^{\text{strat}} = \sum_{j=1}^{J} \frac{N(j)}{N} \cdot \hat{\tau}^{\text{dif}}(j), \longrightarrow \hat{\tau}^{\text{dif}}(j) = \frac{\sum_{i:B_i(j)=1} Y_i \cdot W_i}{\sum_{i:B_i(j)=1} W_i} - \frac{\sum_{i:B_i(j)=1} Y_i \cdot (1 - W_i)}{\sum_{i:B_i(j)=1} (1 - W_i)}$$

- This method is more robust than the weighting estimator to the presence of units with extreme values of the estimated propensity score
- We still do not recommend it until we reduce the bias and increase the precision further by using covariance adjustment within the subclasses.

M5: Matching Estimators

- Find direct comparisons (i.e., matches) for each unit
  - If exact matches can be found, this is a particularly attractive and simple strategy.
  - If no exact matches can be found (when the number of covariates is large), one needs to assess the trade-offs of different violations of exact matching.

The target unit

A 30-year-old woman
with 2 children under the age of 6
with a high school education
with 4 months of work experience in the past 12 months who went through this training program

A exact match

A 30-year-old woman
with 2 children under the age of 6
with a high school education
with 4 months of work experience in the past 12 months who did not go through this training program

Match I (not exact)

A 34-year-old woman
with 2 children under the age of 6
with a high school education
With 6 months of work experience in the past 12 months who did not go through this training program

Match II (not exact)

A 28-year-old woman
With 1 child under the age of 6
with a high school education
With 2 months of work experience in the past 12 months who did not go through this training program

May assess by a **distance metric** on the space of covariates

⇦ Match I or II, which one is a better match?



M1: Model-Based Imputation

- Building a model for the missing potential outcomes
- Using the model to predict/impute the missing outcomes

Model for the missing potential outcomes

Unit-level joint model (the working model)

$$\mathbf{Y}^{\text{mis}} \mid \mathbf{Y}^{\text{obs}}, \mathbf{X}, \mathbf{W} \Leftarrow \mathbf{Y}(0), \mathbf{Y}(1) \mid \mathbf{W}, \mathbf{X}, \theta \Leftrightarrow \mathbf{Y}(0), \mathbf{Y}(1) \mid \mathbf{X}, \theta \Leftrightarrow (Y_i(0), Y_i(1)) \mid X_i, \theta$$

unconfoundedness

exchangeability of the units
De Finetti's Theorem

- With no or a few covariates, it is relatively easy to specify a flexible functional form for this conditional distribution.
- If there are many covariates, such a specification is more difficult, and the results can be sensitive to alternative choices.
- For CREs, such sensitivity is often a minor issue as the covariate distributions in treatment and control groups are similar.

We do NOT recommend this approach for general cases

M2: Regression Estimators

Note that here $\boldsymbol{X}$ include a constant term

Model: $\begin{pmatrix} Y_i(0) \\ Y_i(1) \end{pmatrix} \Big| X_i, \theta \sim \mathcal{N}\left(\begin{pmatrix} X_i\beta_c \\ X_i\beta_t \end{pmatrix}, \begin{pmatrix} \sigma_c^2 & \sigma_c \cdot \sigma_t \\ \sigma_c \cdot \sigma_t & \sigma_t^2 \end{pmatrix} \right)$ Parameter: $\theta = (\beta_c, \beta_t, \sigma_c^2, \sigma_t^2)$

Estimator: $\hat{\beta}_c^{\text{ols}} = \arg\min_{\beta} \sum_{i:W_i=0} (Y_i - X_i\beta)^2$ $\hat{\beta}_t^{\text{ols}} = \arg\min_{\beta} \sum_{i:W_i=1} (Y_i - X_i\beta)^2$

$$\hat{\tau}^{\text{ols}} = \frac{1}{N}\sum_{i=1}^{N}\left(W_i \cdot (Y_i^{\text{obs}} - X_i\hat{\beta}_c^{\text{ols}}) + (1-W_i)\cdot(X_i\hat{\beta}_t^{\text{ols}} - Y_i^{\text{obs}})\right)$$

$$= \frac{N_{\text{t}}}{N_{\text{t}}+N_{\text{c}}}\cdot\hat{\tau}_t^{\text{ols}} + \frac{N_{\text{c}}}{N_{\text{t}}+N_{\text{c}}}\cdot\hat{\tau}_c^{\text{ols}}$$

$$\hat{\tau}_t^{\text{ols}} = \frac{1}{N_{\text{t}}}\sum_{i:W_i=1}\left(Y_i^{\text{obs}} - X_i\hat{\beta}_c\right) = \overline{Y}_{\text{t}}^{\text{obs}} - \overline{X}_t\hat{\beta}_c^{\text{ols}} = \overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}} - (\overline{X}_t - \overline{X}_c)\hat{\beta}_c^{\text{ols}}$$

$$\hat{\tau}_c^{\text{ols}} = \frac{1}{N_{\text{c}}}\sum_{i:W_i=0}\left(X_i\hat{\beta}_t - Y_i^{\text{obs}}\right) = \overline{X}_c\hat{\beta}_t^{\text{ols}} - \overline{Y}_{\text{c}}^{\text{obs}} = \overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}} - (\overline{X}_t - \overline{X}_c)\hat{\beta}_t^{\text{ols}}$$

Close to zero in CREs & PREs
Can be large in other cases

We do NOT recommend this approach for general cases without substantial modifications

M3: Weighting Estimators

$$\mathbb{E}\left[\frac{Y_i^{\text{obs}} \cdot W_i}{e(X_i)}\right] = \mathbb{E}_{\text{sp}}\left[\mathbb{E}\left[\frac{Y_i^{\text{obs}} \cdot W_i}{e(X_i)}\middle| X_i\right]\right] = \mathbb{E}_{\text{sp}}\left[\mathbb{E}\left[\frac{Y_i(1) \cdot W_i}{e(X_i)}\middle| X_i\right]\right]$$
$$= \mathbb{E}_{\text{sp}}\left[\frac{\mathbb{E}_{\text{sp}}[Y_i(1)|X_i] \cdot \mathbb{E}_W[W_i|X_i]}{e(X_i)}\right] = \mathbb{E}_{\text{sp}}\left[\mathbb{E}_{\text{sp}}[Y_i(1)|X_i]\right] = \mathbb{E}_{\text{sp}}\left[Y_i(1)\right]$$

$$\mathbb{E}\left[\frac{Y_i^{\text{obs}} \cdot W_i}{e(X_i)}\right] = \mathbb{E}_{\text{sp}}\left[Y_i(1)\right] \quad \longleftrightarrow \quad \mathbb{E}\left[\frac{Y_i^{\text{obs}} \cdot (1-W_i)}{1-e(X_i)}\right] = \mathbb{E}_{\text{sp}}\left[Y_i(0)\right]$$

$$\hat{\tau}^{\text{ht}} = \frac{1}{N}\sum_{i=1}^{N}\frac{W_i \cdot Y_i^{\text{obs}}}{e(X_i)} - \frac{1}{N}\sum_{i=1}^{N}\frac{(1-W_i) \cdot Y_i^{\text{obs}}}{1-e(X_i)} = \frac{1}{N}\sum_{i:W_i=1}\lambda_i \cdot Y_i^{\text{obs}} - \frac{1}{N}\sum_{i:W_i=0}\lambda_i \cdot Y_i^{\text{obs}}$$

Unbiased estimator when λ_is are known function of X_i.

$$\lambda_i = \frac{1}{e(X_i)^{W_i} \cdot (1-e(X_i))^{1-W_i}} = \begin{cases} 1/(1-e(X_i)) & \text{if } W_i = 0, \\ 1/e(X_i) & \text{if } W_i = 1. \end{cases}$$

Replaced by

$$\hat{\lambda}_i = \begin{cases} N \cdot (1-\hat{e}(X_i))^{-1} / \sum_{j:W_j=0}(1-\hat{e}(X_i)^{-1} & \text{if } W_i = 0, \\ N \cdot \hat{e}(X_i)^{-1} / \sum_{j:W_j=1}\hat{e}(X_i)^{-1} & \text{if } W_i = 1. \end{cases}$$

When $\hat{e}(X_i)$ is close to zero or one for some units, $\hat{\tau}^{\text{ht}}$ can be sensitive to minor changes in the specification of the model for the propensity score.

Strategies for Estimation

1. Model-based imputation
2. Regression estimators

} $P(\mathbf{Y}|\cdots)$ Imputation

✓ 3. Weighting estimators using the propensity score — $P(W|\cdots)$

✓ ?4. Blocking estimators using the propensity score

✓ ?5. Matching estimators using a distance metric

} Balance the covariates

?6. Mixed estimators: combine 2-5

Another Key difference: before / after seeing the outcome

Design | Analysis



PS is the Key to Causal Inference

- Many of the procedures for estimating and assessing causal effects under unconfoundedness involve the propensity score

Strategies for Estimation

1. Model-based imputation
2. Regression estimators
3. Weighting estimators using the propensity score
4. Blocking estimators using the propensity score
5. Matching estimators using a distance metric
6. Mixed estimators: combine 2-5



Estimating Propensity Score



Cases where PS is Known

- Propensity score is known in randomized experiments
  - CREs: $e(x) = N_t/N$ for all x
  - SREs: $e(x) = N_t(j)/N(j)$ for $x \in S_j$
  - PREs: $e(x) = 0.5$ for each pair

- It's rare that we know the propensity score in other settings

Cases where PS is Unknown

- A more common situation:
  - given available pre-treatment variables, a researcher views unconfoundedness as a reasonable approximation to the actual assignment mechanism
  - with only vague *a priori* information about the form of the dependence of the propensity score on the observed pre-treatment variables
- An example in typical medical settings:
  - decisions are based on a set of clinically relevant patient characteristics observed by doctors and entered in patients' medical records
  - but, there is typically no explicit rule that requires physicians to choose a specific treatment based on particular values of the pre-treatment variables



Challenges in Estimating PS

- If X_i takes on only a few values, estimating PS is straightforward:
  - we can simply partition the sample into subsamples that are homogeneous in the values of the covariates
  - then estimate the propensity score for each subsample as the proportion of treated units in that subsample
- If X_i takes on many values, such a saturated model is not feasible:
  - many strata defined by unique values of the covariates in the sample contain only a single unit
  - some form of smoothing is essential

Does the Estimation Accuracy Matter?

- Yes to methods that rely heavily on an accurate approximation of the true PS
  - ✓ Method A: weighting by the inverse of the estimated propensity score
    - ➢ Can eliminate all biases in arising from differences in covariates distributions
    - ➢ Estimated propensity score close to zero will greatly inflate variance of weights

- Not to methods that are not sensitive to the approximation of the true PS
  - ✓ Method B: construct strata or subclasses based on estimated propensity score, within which further adjustment methods (e.g., regression) are used
    - ➢ Estimated propensity score serves as a reference to construct strata or subclasses
    - ➢ Units within each subclass are expected to have similar true propensity scores
    - ➢ Further adjustment methods within each subclass can effectively remove bias

Be Clear about the Goal

- The goal is to obtain estimates of PS that balance the covariates between treated and control subsamples, i.e.,
  - within subsamples with similar values of the estimated propensity score
  - the distribution of covariates among the treated and control units are similar
- The goal is *not* simply to get the "best" estimate of the propensity score
  - in terms of MSE or other criterions that minimize the difference between the estimated and true propensity score
- For our purposes, it is often preferable to use the estimated propensity score
  - using the estimated score may lead to superior covariate balance in the sample compared to that achieved when using the true propensity score
  - For example, in a completely randomized experiment with a single binary covariate
    - using the estimated propensity score to stratify units would lead to perfect within-stratum balance on the covariates in the sample
    - whereas using the true propensity score generally would not



Three Comments

- During the entire process, we do not use the outcome data
  - Thus, there is no way of deliberately biasing the final estimation results for the treatment effects
- In general, it is difficult to give a fully automatic procedure for specifying the propensity score
  - The specification may be much improved by incorporating knowledge on the role of the covariates in the treatment assignment decision *&* outcome process
- The primary goal is to find an adequate specification of the propensity score that achieves statistical balance in the covariates
  - We are *not* directly interested in a structural, behavioral, or causal interpretation of the propensity score



The Barbituate Exposure Data

Table 13.1. ***Summary Statistics Reinisch Data Set***

Label	Variable Description	Controls (N_c =7198)		Treated (N_t =745)		t-Stat
		Mean	(S.D.)	Mean	(S.D.)	Difference
sex	Sex of child (female is 0)	0.51	(0.50)	0.50	(0.50)	−0.3
antih	Exposure to antihistamine	0.10	(0.30)	0.17	(0.37)	4.5
hormone	Exposure to hormone treatment	0.01	(0.10)	0.03	(0.16)	2.5
chemo	Exposure to chemotherapy agents	0.08	(0.27)	0.11	(0.32)	2.5
cage	Calendar time of birth	−0.00	(1.01)	0.03	(0.97)	0.7
cigar	Mother smoked cigarettes	0.54	(0.50)	0.48	(0.50)	−3.0
lgest	Length of gestation (10 ordered categories)	5.24	(1.16)	5.23	(0.98)	−0.3
lmotage	Log of mother's age	−0.04	(0.99)	0.48	(0.99)	13.8
lpbc415	First pregnancy complication index	0.00	(0.99)	0.05	(1.04)	1.2
lpbc420	Second pregnancy complication index	−0.12	(0.96)	1.17	(0.56)	55.2
motht	Mother's height	3.77	(0.78)	3.79	(0.80)	0.7
motwt	Mother's weight	3.91	(1.20)	4.01	(1.22)	2.0
mbirth	Multiple births	0.03	(0.17)	0.02	(0.14)	−1.9
psydrug	Exposure to psychotherapy drugs	0.07	(0.25)	0.21	(0.41)	9.1
respir	Respiratory illness	0.03	(0.18)	0.04	(0.19)	0.7
ses	Socioeconomic status (10 ordered categories)	−0.03	(0.99)	0.25	(1.05)	7.0
sib	If sibling equal to 1, otherwise 0	0.55	(0.50)	0.52	(0.50)	−1.6

17 covariates



Logistic Regression for PS

The general model:

$$\Pr(W_i = 1|X_i = x;\phi) = \frac{\exp(h(x)\phi)}{1+\exp(h(x)\phi)}$$

$h(x) = x$

A simple model:

$$\Pr(W_i = 1|X_i = x;\phi) = \frac{\exp(x\phi)}{1+\exp(x\phi)}$$

Data likelihood:

$$\mathcal{L}(\phi|\mathbf{Y}^{\text{obs}},\mathbf{W},\mathbf{X}) = \prod_{i=1}^{N}\Pr(W_i = 1|X_i;\phi)^{W_i}\cdot(1-\Pr(W_i = 1|X_i;\phi))^{1-W_i} = \prod_{i=1}^{N}\frac{\exp(W_i\cdot X_i\phi)}{1+\exp(X_i\phi)}$$

Log-likelihood:

$$L(\phi|\mathbf{Y}^{\text{obs}},\mathbf{W},\mathbf{X}) = \sum_{i=1}^{N} W_i\cdot X_i\phi - \ln(1+\exp(X_i\phi))$$

globally concave if the matrix $\sum_{i=1}^{N} h(X_i)^T\cdot h(X_i)$ is positive definite

Get MLE by Newton-Raphson algorithm

Initial value: $\phi^0 = (\ln(N_t/N_c), 0^T)^T$

Updating function: $\phi^{k+1} = \phi^k - \left(\frac{\partial^2}{\partial\phi\partial\phi^T}L(\phi^k)\right)^{-1}\frac{\partial}{\partial\phi}L(\phi^k)$

Maximum likelihood estimator:

$$\hat{\phi}_{\text{ml}} = \arg\max_{\phi} L(\phi|\mathbf{Y}^{\text{obs}},\mathbf{W},\mathbf{X})$$

$$\hat{\mathbb{V}}\left(\hat{\phi}_{\text{ml}}\right) = -\left(\frac{\partial^2}{\partial\phi\partial\phi^T}L(\hat{\phi})\right)^{-1}.$$

Selecting Covariates & Interactions

- Step 1. select a set of K_B basic covariates
  - covariates that are *a priori* viewed as important for explaining the assignment
  - covariates that are *a priori* viewed as plausibly related to some outcome measures
  - covariates that are *a priori* viewed as to be strongly associated with the outcomes
- Step 2. select $K_L - K_B$ additional linear terms
  - Space: $K - K_B$ covariates not included yet
  - Method: forward logistic regression with the largest likelihood ratio statistic
- Step 3. select K_Q quadratic and interaction terms
  - Space: $K_L(K_L + 1)/2$ quadratic and interaction terms of the K_L covariates selected
  - Method: forward logistic regression with the largest likelihood ratio statistic



Example: Barbituate Exposure Data (I)

- Step 1: select 3 basic covariates based on domain knowledge

Table 13.2. ***Estimated Parameters of Propensity Score: Baseline Case; Barbituate Data***

Variable	EST	(s. e.)	t-Stat
Intercept	−2.38	(0.06)	−41.0
sex	−0.01	(0.08)	−0.2
lmotage	0.48	(0.04)	11.7
ses	0.10	(0.04)	2.6

- The child's sex, the mother's age, and mother's socio-economic status (sex, lmotage, and ses respectively) are particularly important covariates:
  - The first two because they are likely to be associated with the outcomes of interest,
  - and the last two because they are likely to be related to barbituate exposure.

Example: Barbituate Exposure Data (II)

- Step 2: select additional linear terms by forward regression

Table 13.3. *Estimated Parameters of Propensity Score: Baseline Case with* `lpbc420` *Added; Barbituate Data*

Variable	EST	(s.e.)	t-Stat
Intercept	−3.71	(0.10)	−36.3
sex	0.07	(0.09)	0.8
lmotage	0.22	(0.05)	4.7
ses	0.15	(0.05)	3.3
lpbc420	2.11	(0.08)	27.2
LR statistic	1308.0		

Table 13.4. *Likelihood Ratio Statistics for Sequential Selection of Covariates to Enter Linearly; Barbituate Data*

Covariate	Step →										
sex	–	–	–	–	–	–	–	–	–	–	–
antih	17.5	0.5	1.6	1.3	2.1	1.8	1.6	1.6	1.7	1.3	–
hormone	3.9	0.3	0.7	0.7	0.4	0.8	0.7	0.7	0.7	0.8	0.9
chemo	10.0	36.6	41.9	–	–	–	–	–	–	–	–
cage	0.8	5.8	6.4	7.2	7.6	7.9	–	–	–	–	–
cigar	4.3	2.3	3.5	3.7	3.0	2.1	2.1	1.7	2.1	–	–
lgest	0.4	11.1	5.0	6.4	7.3	5.5	5.6	–	–	–	–
lmotage	–	–	–	–	–	–	–	–	–	–	–
lpbc415	0.6	0.0	0.2	0.2	0.0	0.0	0.1	0.1	0.0	0.0	0.0
lpbc420	1308.0	–	–	–	–	–	–	–	–	–	–
motht	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
motwt	6.1	1.5	0.6	1.2	2.5	2.7	2.4	3.4	–	–	–
mbirth	4.6	66.1	–	–	–	–	–	–	–	–	–
psydrug	93.1	29.8	38.9	46.8	–	–	–	–	–	–	–
respir	0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ses	–	–	–	–	–	–	–	–	–	–	–
sib	21.0	13.8	12.5	15.0	15.7	–	–	–	–	–	–

Threshold $C_L = 1$

Table 13.5. *Estimated Parameters of Propensity Score: Baseline Case with* `lpbc420` *and* `mbirth` *Added; Barbituate Data*

Variable	EST	(s.e.)	t-Stat
Intercept	−3.73	(0.10)	−35.9
sex	0.08	(0.09)	0.9
lmotage	0.21	(0.05)	4.5
ses	0.16	(0.05)	3.4
lpbc420	2.21	(0.08)	27.5
mbirth	−1.96	(0.30)	−6.6
LR statistic	66.1		

LR between the models with and without the target variable

Example: Barbituate Exposure Data (III)

- Step 3: select quadratic & interaction terms by forward regression

Table 13.6. ***Estimated Parameters of Propensity Score: Final Specification; Barbituate Data***

Variable	EST	(s.e.)	t-Stat
Intercept	−5.67	(0.23)	−24.4
Linear terms			
sex	0.12	(0.09)	1.3
lmotage	0.52	(0.11)	4.7
ses	0.06	(0.09)	0.6
lpbc420	2.37	(0.36)	6.6
mbirth	−2.11	(0.36)	−5.9
chemo	−3.51	(0.67)	−5.2
psydrug	−3.37	(0.55)	−6.1
sib	−0.24	(0.22)	−1.1
cage	−0.56	(0.26)	−2.2
lgest	0.57	(0.23)	2.5
motwt	0.49	(0.17)	2.9
cigar	−0.15	(0.10)	−1.5
antih	0.17	(0.13)	1.3
Second-order terms			
lpbc420 × sib	0.60	(0.19)	3.1
motwt × motwt	−0.10	(0.02)	−4.5
lpbc420 × psydrug	1.88	(0.39)	4.8
ses × sib	−0.22	(0.10)	−2.2
cage × antih	−0.39	(0.14)	−2.8
lpbc420 × chemo	1.97	(0.49)	4.0
lpbc420 × lpbc420	−0.46	(0.14)	−3.3
cage × lgest	0.15	(0.05)	3.0
lmotage × lpbc420	−0.24	(0.10)	−2.5
mbirth × cage	−0.88	(0.39)	−2.3
lgest × lgest	−0.04	(0.02)	−2.0
ses × cigar	0.20	(0.09)	2.2
lpbc420 × motwt	0.15	(0.07)	2.0
chemo × psydrug	−0.93	(0.46)	−2.0
lmotage × ses	0.10	(0.05)	1.9
cage × cage	−0.10	(0.05)	−1.8
mbirth × chemo	−∞	(0.00)	−∞

Summary



Summary

- Regular Assignment Mechanisms is a kind of observational study that we can deal with.
  - ✓ Similar to SRE
  - ✓ Inference strategies follows that in SRE tightly: Design + Analysis
- Propensity score (PS) is the coarsest balancing score, playing an important role in Inference on RAM.
- Estimating PS:
  - goal
  - method (logistic regression, or any other reasonable methods)



Something to Share

- When facing a new task and do not know what to do, imitation is a solution (fast way for learning to solve). 站在巨人的肩膀上

- When facing a task, if you want to control the cost, it is better to form a basic framework as the initial solution, and optimize it step by step. 建设性



Reference

- Reference: [T1] Chapter 12, 13

- ational Psycholog
- ord, F. M. (1967), "A Pa
  Comparisons", Psychologi