# Chapter 7 Observational Study

(Optional topic)

## 7.1 Causality in observational studies

- Features of observational studies
- Selection bias

## 7.2 Analysis with no latent confouding

- Assumptions
  - I.I.D.
  - Ignorability
  - Overlap
- Estimation
  - Stratification
  - Outcome regression
- Propensity score
  - Definition and key properties
  - Propensity scores: matching
  - Propensity scores: weighting
  - Doubly-robust regression
- Covariance balancing

## 7.3 Instrumental variable

- Definition and assumptions
- Key properties of IV

- Survey Samples  self-reporting bias
- Cohort Studies  survival bias
- Case-control Studies  ① efficient cheap
  $Y=1$  $Y=0$  ② unbalanced - balanced

Randomized Response Technique  Wavuer wodel 1965

I Subject belongs to A ?  $\pi_A$  Proba  $Prob(I) = p$

II — — — - $A^c$ ?  $1-\pi_A$  $Prob(II) = 1-p$

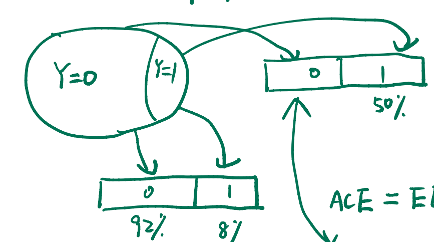
$$Prob(Y=1) = P(Y=1 \mid I)P(I) + P(Y=1 \mid II)P(II)$$
$$= \pi_A \cdot p + (1-\pi_A)(1-p)$$

$\{Y_1 \cdots Y_n\} \in \{0,1\}$  $Prob(Y=0)$

$$L(\pi_A \mid \vec{Y}) = \prod_{i=1}^{n} Prob(Y_i=1)^{Y_i} Prob(Y_i=0)^{1-Y_i}$$

→ number of $Y_i=1$

$$\hat{\pi}_{A,MLE} = \frac{\frac{n_1}{n} - (1-p)}{2p-1}$$ unbiased estimator




$ACE = E[Y=1] - E[Y=0]$

population → log odds ratio

Logistic Regression [case-controlled studies] not bec. a binary treatment

Q should we use a logistic Regression to analyze data from an RCT?

Data$(X_i, Y_i)$, $Y_i \in \{0,1\}$ $X_i \in \mathbb{R}^p$

logistic regression

$$\text{logit}(\pi_i) = X_i^T\beta \quad \text{where } \pi_i = \mathbb{P}(Y_i=1 \mid X_i) = E[Y_i \mid X_i] \in [0,1]$$

and $\text{logit}(a) = \log\left(\frac{a}{1-a}\right)$ : log odd

$$\beta_1 = \log\left\{\frac{\frac{\pi(X_1=b_1+1, X_2, \cdots, X_p)}{1-\pi(X_1=b_1+1, X_2 \cdots X_p)}}{\frac{\pi(X_1=b_1, X_2 \cdots X_p)}{1-\pi(X_1=b_1 \cdots X_p)}}\right.$$

$$L = \prod_{i=1}^{n} \pi_i^{Y_i}(1-\pi_i)^{1-Y_i} \qquad P(Y_i=1) = \pi_i$$

$$\log L = \sum_{i=1}^{n} Y_i \log \pi_i + \sum_{i=1}^{n}(1-Y_i)\log(1-\pi_i)$$

$$= n_1 \log \pi_i + (n-n_1)\log(1-\pi_i) \quad \text{where } \text{logit}(\pi_i) = \beta_0 \qquad \pi_i = \exp(\beta_0)$$

- Estimation

## 7.4 Missing data

- Missing mechanisms
- Multiple imputation

$= n_1\{\beta_0 - \log[\exp(\beta_0)+1]\} + (n-n_1)\{0 - \log[\exp(\beta_0)+1]\}$

$\frac{\partial \log L}{\partial \beta_0} = 0 \quad n_1 - n\frac{\exp(\hat\beta_0)}{\exp(\hat\beta_0)+1} = 0 \Rightarrow \hat\beta_0 : \text{logit}\left(\frac{n_1}{n}\right)$

$U(\beta) = \frac{\partial \log L}{\partial \beta}$ score function if $\beta \in \mathbb{R}^p$

At MLE $(\hat\beta_0)$, $U(\hat\beta) = 0$

$I(\beta) = \mathbb{E}\left[-\frac{\partial^2 \log L}{\partial \beta^2}\right]$ Fisher information $\in \mathbb{R}^{p\times p}$ | $\boxed{\text{var}(\hat\beta) = I^{-1}(\hat\beta)}$

$\sqrt{n}(\hat\beta - \beta) \longrightarrow N(0, I^{-1}(\beta))$

Confidence Interval for $\hat\beta_1$

$\hat{se}^2(\hat\beta_1) =$ 1st diagnos (entry in $I^{-1}(\hat\beta)$)

$H_0: \beta_1 = 0$ vs $H_a$ $\beta_1 \neq 0$

$t = \frac{\hat\beta_1}{\hat{se}(\beta_1)} \sim N(0,1)$ | wald test

$S = \frac{U_1(\beta_1=0)^2}{[I^{-1}(\beta_1=0)]_1} \sim \chi^2$ with d.f. = 1 | score test

$LR = -2[\log L(\text{residual}) - \log L(\text{full})] \sim \chi^2_{\Delta df}$

Likelihood ratio test

F test is a special case $F(a_1, a_2) \xrightarrow{a_2 \to \infty} \chi^2_{a_1}$

Model Diagnostics

Pearson test

Deviance test