BIOS 6612

Lecture 6

Covariate Adjustment in Logistic Regression

Review (Lecture 5) / Current (Lecture 6)/ Preview (Lecture 7)

- Lecture 5
    - Comparing logistic regression models
    - Effect modification

- Lecture 6: Covariate Adjustment in Logistic Regression
    - Confounding
        - Operational vs Classical Criteria (NOT the same)

- Lecture 7
        - Model Fit

Confounding in Logistic Regression Example

Two **randomized** clinical trials were performed to compare a drug versus placebo for prevention of disease. The baseline characteristics of the study participants are given in the tables below.

Table 1. Baseline characteristics of participants in study 1. (n=42)

	Drug *(n = 21)*	*Placebo* *(n =21)*	*p-value*
Age (mean±SD)	46.2 ± 6.9	52.2 ± 12.5	0.0653
Gender (% male)	42.9%	52.4%	0.5366
SBP (mean±SD)	147.4 ± 10.1	148.5 ± 10.0	0.7247

6.6 year age difference
1.1 mm/Hg difference in SBP

Table 2. Baseline characteristics of participants in study 2. (n=2000)

	Drug *(n = 1000)*	*Placebo* *(n =1000)*	*p-value*
Age (mean±SD)	47.7 ± 9.9	49.1 ± 10.2	0.0013
Gender (% male)	48.5%	50.0%	0.5023
SBP (mean±SD)	147.4 ± 10.2	148.3 ± 10.1	0.0475

1.4 year age difference
0.9 mm/Hg difference in SBP

- Age and SBP are known risk factors for the disease.
- In which of these two studies would you be most concerned about potential confounding?

Study-1 (N=42)
6.0 year age difference (p=0.0653)
1.1 mm/Hg SBP difference (p=0.7247)

Study-2 (N=2000)
1.4 year age difference (p=0.0013)
0.9 mm/Hg SPP difference (p=0.0475)

Variable Coding:

Group: 1 = "Placebo"
0 = "Drug"

Study 1 Results:

The LOGISTIC Procedure

Response Profile

Ordered Value	disease	Total Frequency
1	1	23
2	0	19

Crude Model
(Unadjusted)
(Not adjusting for Age)

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.4855	0.4494	1.1674	0.2799
group	1	1.4018	0.6597	4.5147	0.0336

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
group	**4.062**	1.115	14.804

Adjusted Model
(Adjusting for Age)

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-10.6953	3.2687	10.7065	0.0011
group	1	1.2161	0.9453	1.6551	0.1983
age	1	0.2189	0.0694	9.9473	0.0016

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
group	**3.374**	0.529	21.519
age	1.245	1.086	1.426

Adjusted OR is
~17% smaller

Study 2 Results:

Variable Coding:

Group: 1 = "Placebo"
0 = "Drug"

The LOGISTIC Procedure

Response Profile

Ordered Value	disease2	Total Frequency
1	1	831
2	0	1169

Crude Model
(Not adjusting for Age)

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.5841	0.0660	78.4005	<.0001
group	1	0.4759	0.0914	27.0877	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
group	**1.610**	1.345	1.925

Adjusted Model
(Adjusting for Age)

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-28.0924	1.4360	382.7285	<.0001
group	1	0.4631	0.1736	7.1162	0.0076
age	1	0.5497	0.0283	378.4551	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
group	**1.589**	1.131	2.233
age	1.733	1.639	1.831

Adjusted OR is
~1% smaller

Which Study Was Confounding More of an Issue? Why?

Study 1

- Magnitude of the age difference was greater in study 1
  - Even though the p-value was smaller in study 2
    - Due to larger N in study 2
- Age was a strong predictor of the outcome in both studies.

- You should look at the magnitude of the difference between drug and placebo
  - NOT the p-value!
  - p-values in these two studies are heavily influenced by the sample sizes (42 versus 2000)!

- The amount of confounding is related to the magnitude of the confounder-exposure and confounder-disease association
  - NOT the p-values.

- However, if the treatment effect was very small in study 2
  - Then a small imbalance in baseline covariates may lead to a spurious treatment effect

  - In addition, if age was a much stronger predictor in study 2
    - Could have had more confounding in study 2 even though the age difference was smaller

Why Adjust for Covariates?

- Confounder
  - Provide an unbiased estimate of an exposure- outcome association
  - $X \leftarrow Z \rightarrow Y$
    - Where X is the exposure, Z is the confounder, and Y is the outcome

- Effect Modifiers
  - Examine interactions

- Effect Mediators
  - Examine causal pathways
  - $X \rightarrow Z \rightarrow Y$

- Precision/ Efficiency Variables
  - Increase precision and/or efficiency of the exposure- outcome comparison

    - More Precise:
      - smaller SE and narrower CI

    - More Efficient/More Powerful:
      - Larger z-statistics, smaller p-value under H_1

Classical Criteria for Confounding

1. A confounding factor must be associated with the exposure (PEV) under study in the source population
2. A confounding factor must be a risk factor for the disease among the unexposed conditioning on exposure
3. A confounding factor must not be affected by the exposure or the disease.
   -It cannot be an intermediate step in the causal path between the exposure and the disease

Confounding:

$$\begin{tikzcd} C \arrow[dr] \arrow[dd] & \\ & Y \\ X \arrow[ur] & \end{tikzcd} \qquad \begin{tikzcd} & C \arrow[dr] & \\ U \arrow[ur] \arrow[dr] & & Y \\ & X \arrow[ur] & \end{tikzcd}$$

Mediation:

$$\begin{tikzcd} & C \arrow[dr] & \\ X \arrow[ur] \arrow[rr] & & Y \end{tikzcd}$$

Operational Criterion for Confounding

- Covariate is a confounding factor if the estimate of the exposure effect changes when the covariate is included in the analysis (i.e. $\beta_{crude} \neq \beta_{adj}$)
  - By stratification or regression methods

- What change is meaningful? $\dfrac{\beta_{crude} - \beta_{adj}}{\beta_{adj}}$
  - (1) Clinically meaningful change? (2) 10% change? (3) 20% change?

Criteria for Confounding

- Many people define confounding using the classical criteria
  - BUT use the operational definition to determine whether or not a covariate is a confounder in their data

- PROBLEM: classical and operational criteria for confounding do NOT always agree

Crude and Adjusted Estimates in Linear Regression

Crude Model: $E[Y] = \beta_{01} + \beta_{crude} X$

Adjusted Model: $E[Y] = \beta_{02} + \beta_{adj} X + \beta_z Z$

Covariate Model: $E[Z] = \gamma_0 + \gamma_X X$

X = PEV
Z = Covariate/ Potential Confounder
Y = Outcome/Response

For Linear regression, the classical and operational criteria agree

$$\beta_{crude} - \beta_{adj} = \gamma_X \times \beta_z$$

operational = classical

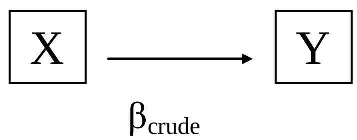
Classical Criterion #1 (X and Z association): $\gamma_x \neq 0$

Classical Criterion #2 (Z and Y associated given X): $\beta_z \neq 0$

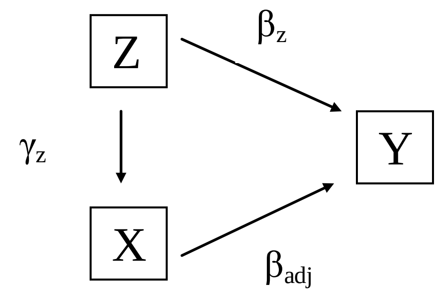
Classical Criterion #3 X ← Z → Y (i.e. Z is NOT a mediator)

Causal Models

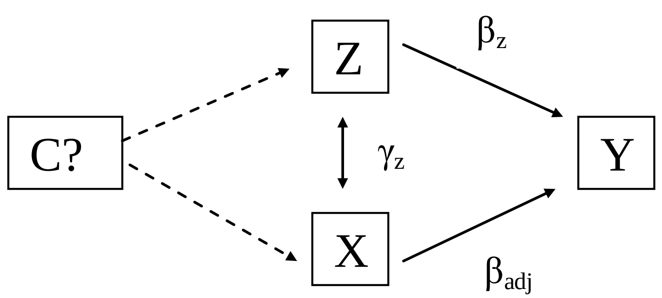
Unadjusted (Crude) Model



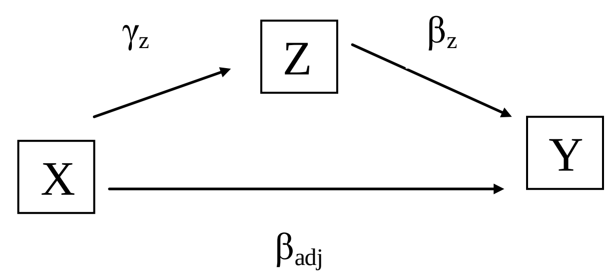
Confounding by Z



Confounding by unmeasured variable C, Z is a proxy measure



Mediation by Z



For Y & Z continuous:

$$\beta_{crude} - \beta_{adj} = \gamma_x \times \beta_z$$

operational = classical

Crude and Adjusted Estimates in Logistic Regression

Crude Model: $\log it(p) = \beta_{01} + \beta_{crude} X$

Adjusted Model: $\log it(p) = \beta_{02} + \beta_{adj} X + \beta_z Z$

Covariate Model: $E[Z] = \gamma_0 + \gamma_X X$ *or* $\log it(\mathrm{p}_z) = \gamma_0 + \gamma_X X$

X = PEV Z = Covariate/ Potential Confounder Y = Outcome/Response p=P(Y=1) & p_z=P(Z=1)

For logistic regression, the classical and operational criteria do NOT agree

$$\beta_{crude} - \beta_{adj} \neq \gamma_x \beta_z$$

- γ_x can be zero, yet $\beta_{crude} \neq \beta_{adj}$
  - β_{adj} will be further from the null
- γ_x and β_z may both be non-zero
  - Yet $\beta_{crude} = \beta_{adj}$

Covariate Adjustment in Logistic Regression: Example with No "Classical" Confounding

HIV ← City → Risky

Covariate → San Francisco | New York

Exposure → Known HIV +

Outcome ↓ Risky Behavior

San Francisco (Exposure: Known HIV +)

Risky Behavior	No	Yes	
No	90	75	165
Yes	10	25	35
	100	**100**	200

OR = 3.0

New York (Exposure: Known HIV +)

Risky Behavior	No	Yes	
No	50	25	75
Yes	50	75	125
	100	**100**	200

OR = 3.0

NOTE: No "Classical" confounding since City is not associated with exposure.

P(Expose|SF) = 100/200
P(Expose|NY) = 100/200

← Adjusted OR will be 3.0

Hypothetical study: Is knowledge of one's HIV-infection status (Exposure) related to high risk sexual behavior in prior month (Outcome).

Suppose data collected in NY and SFO (Covariate) and presume that risk behavior is rarer in SFO.
SOURCE: Hauck 1991

Example Details

SFO

p_1 = (90/165) = 0.5454
p_2 = (10/35) = 0.2857
OR= $(p_1*(1-p_2))/(p_2*(1-p_1))$ = (a*d)/(b*c) = (90*25)/(10*75) = 3
Pr(exposure) = (75+25)/200 = 0.5

NYC

p_1 = (50/75) = .6667
p_2 = (50/125) = 0.40
OR= $(p_1*(1-p_2))/(p_2*(1-p_1))$ = (a*d)/(b*c) = (50*75)/(50*25) = 3

Pr(exposure) = (25+75)/200 = 0.5

Mantel-Haenszel Estimate of the Common Odds Ratio

$$OR_{MH} = \frac{\sum_{i=1}^{k}\left(\frac{c_i \times b_i}{n_i} * OR_i\right)}{\sum_{i=1}^{k}\left(\frac{c_i \times b_i}{n_i}\right)} = \frac{\sum_{i=1}^{k}\left(a_i d_i / n_i\right)}{\sum_{i=1}^{k}\left(b_i c_i / n_i\right)}$$

Weighted-average of stratum-specific ORs,

weights are (c×b)/n

$$OR_{MH} = \frac{(90*25)/200 + (50*75)/200}{(10*75)/200 + (50*25)/200} = \frac{30}{10} = 3$$

Is City a Confounder?

Confounding Factor (City) must be associated with exposure (Know HIV) under study.

	Know HIV		
City	Yes	No	Total
SFO	100	100	**200**
NYC	100	100	**200**
Total	**200**	**200**	**400**

OR=1.0

A Confounding Factor (City) must be a Risk Factor for the Outcome (Risky Behavior) Among the Unexposed (Do Not Know HIV)

Do Not Know HIV Status

	Risky		
City	No	Yes	
SFO	90	10	**100**
NYC	50	50	**100**
	140	**60**	**200**

OR=9.0

City is a <u>Precision/Efficiency Variable</u>: Related to Outcome, but not Exposure

Example with No "Classical" Confounding

- By the classical definition of confounding:

1. A confounding factor must be associated with the exposure under the study in the source population

   - The odds of knowing one's HIV+ status among those living in San Francisco are equal to the odds of knowing one's HIV+ status among those living in New York (OR=1)

   - City is NOT related to the exposure (HIV+ status)

2. A confounding factor must be a risk factor for the outcome among the unexposed

   - Among those who do not know of their HIV+ status, the odss of risky behavior among those living in New York are 9 times higher than the odds of risky behavior among those living in San Francisco ($p<0.001$)

   - City is a risk factor for the outcome (i.e. risky behavior) among the unexposed (i.e. HIV+ status unknown)

Covariate Adjustment in Logistic Regression: Example with No "Classical" Confounding

Collapsed Across City

Exposure → Known HIV +

Outcome ↓ Risky Behavior

	No	Yes	
No	140	100	240
Yes	60	100	160
	200	**200**	400

OR = 2.3

Operational Confounding

The crude OR (2.3) is 22% *lower* than the adjusted OR (3.0), even though this covariate does *not* meet the classical definition of confounding.

Source: Hauck, 1991

Classical vs Operational Confounding in Logistic Regression

- According to the classical criteria:
  - City is NOT a confounder of the exposure-outcome association since city and exposure (knowledge of HIV+ status) are not associated in the source population

- According to the operational criteria:
  - City is a confounder of the exposure-outcome association since the exposure effect changes (29% increase) when we stratify by city in the analysis

Precision/ Efficiency Covariates

- A precision/ efficiency covariate is a variable that is
  1) Independent of exposure in the source population ($\gamma_x = 0$)
  2) But predictive of the outcome ($\beta_z \neq 0$)

- Precision/ efficiency covariates CANNOT be confounders according to the classical criteria

- Inclusion of a precision/efficiency variables can provide
  1) A more efficient test of the exposure- outcome association
  2) A more precise estimate of the exposure-outcome association

Precision/ Efficiency Covariates in Linear Regression

- The relative precision of the adjusted exposure estimate relative to the crude exposure estimate can be given as follows:

$$\frac{\mathrm{var}\left(\beta_{crude}\right)}{\mathrm{var}\left(\beta_{adj}\right)} = \left(\frac{1-\rho_{xz}^2}{1-\rho_{YZ|X}^2}\right)\left(\frac{n-3}{n-2}\right)$$

  - A strong association between X and Z (ρ_{XZ}^2) **decreases** the precision of β_{adj}
  - A strong association between Z and Y given X ($\rho_{YZ|X}^2$) **increases** the precision of β_{adj}

Precision/ Maverick Covariates in Logistic Regression

- **Maverick:** covariate that satisfies the operational but not the classical criteria for confounding
  - Hauck et al (1991) A consequence of omitted covariates when estimating odds ratios. Journal of Clinical Epidemiology. 44(1):77-81.

- The effect of omitting a maverick is to bias the odds ration towards no effect (i.e. the null)

  - The unadjusted estimate is a population-averaged estimate of the exposure effect

  - The adjusted estimate is a cluster or individual specific estimate of the exposure effect

  - The individual specific estimate will be larger in magnitude than the population averaged estimate

Positive and Negative Confounding in Logistic Regression

- A positive confounder is a confounder that is either
  - Positively related to both the exposure and disease
  - Negatively related to both the exposure and disease
- A negative confounder is a confounder that is either
  - Positively related to the disease and negatively related to the exposure
  - Negatively related to the disease and positively related to the exposure

Covariate Adjustment in Logistic Regression: Example with a "Positive" Confounder

Covariate → Absent / Present

Outcome → Disease

Exposed	Absent: Disease Yes	Absent: Disease No	Absent: Total	Present: Disease Yes	Present: Disease No	Present: Total
Yes	45	45	**90**	97	13	**110**
No	20	90	**110**	56	34	**90**
	65	135	200	153	47	200

OR = 4.50 (Absent) OR = 4.53 (Present)

Adjusted OR = 4.51

Covariate-Disease OR for the unexposed:

OR = (56/34)/(20/90) = 7.41

P(Exposed=Yes |Covariate=No)= 90/200 and P(Exposed=Yes |Covariate=Yes)= 110/200

OR =(Odds | covariate=Yes)/(Odds | covariate=No) =(110/90)/(90/110) = 1.5

Positive and Negative Confounding in Logistic Regression

- A confounding factor must be associated with the exposure under study in the source population
  - The odds of exposure among those with the covariate are 1.5 times the odds of exposure among those without the covariate ($p<0.05$)

- A confounding factor must be a risk factor for the disease among the unexposed
  - Among the unexposed, the odds of disease among those with the covariate are 7.41 times the odds of disease among those without the covariate ($p<0.001$)

Covariate Adjustment in Logistic Regression: Example with a "Positive" Confounder

Covariate → Pooled

Outcome → Disease

Exposure ↓ Exposed

Exposed	Disease Yes	Disease No	
Yes	142	58	**200**
No	76	124	**200**
	218	182	400

$\boxed{OR = 3.99}$

The crude OR (3.99) is 11% *lower* than the adjusted OR (4.51), even though this covariate would be called a "*positive*" confounder.

Summary: Conditions for $\beta_{crude} = \beta_{adj}$

Linear Regression

1) The covariate and exposure are independent (γ_X=0)
2) The covariate and outcome are independent given the exposure (β_z=0)

Logistic Regression

1) The covariate and exposure are independent **given the outcome**
2) The covariate and outcome are independent given the exposure
3) The covariate and exposure are independent

- The classical and operational definitions of confounding do NOT always agree in logistic regression
  - The classical criterion is NOT a sufficient condition for the absence of confounding when modeling binary data using logistic regression

    - A change in the estimate of the exposure of the exposure effect after adjustment for a covariate is NOT evidence of "classical" confounding in logistic regression
      - BUT this is evidence of confounding in linear regression

  - Must rely on the operational criterion
    - Then we say a variable is a confounder if its inclusion in a statistical model affects the estimated effect of exposure
    - Most use changes of 10% or more (Rothman & Greenland, Modern Epidemiology, 2nd Ed. 1998, pp 256-257)

Summary: Precision/ Efficiency Covariates in Logistic Regression

- A strong association between X and Z deceases the precision of Beta Adjusted

- A strong association between Z and Y also deceases the precision of Beta Adjusted

- Thus adjustment for any covariate in logistic regression results in an automatic loss of precision (unless the covariate is extraneous (i.e. jointly independent of X,Y))

- Why adjust for a precision covariate in logistic regression if it results in an automatic increase in the standard error of the exposure effect?

  - Recall the adjustment for a precision covariate in logistic regression also results in a shift of the exposure estimate away from the null.

  - The magnitude of this shift outweighs the loss in precision.

  - The adjusted estimate is more efficient (even though less precise) than the crude estimate

Summary: Adjusting for Confounders in Logistic Regression

- Adjustment for a "positive" confounder can sometimes **increase** the magnitude of the exposure effect estimate

- The effect of omitting a precision/ efficiency variable in logistic regression is to "bias" the estimate towards no effect.

    - The magnitude of this bias increases with the variance of the covariate and with magnitude of its effect on the outcome

- Adjusting for any non-extraneous covariates in logistic regression will result in a less precise estimate of the exposure effect (larger SE)

    - However, adjustment for a precision/ efficiency variable will result in a more efficient test of the exposure-outcome association

- When the outcome is rare, the effect of precision/efficiency variables on the exposure estimate becomes negligible