AGENDA

Review of handout on confounding from c678

Key Points in / Commentary on ALR Ch 3.5 —

Worked Examples

- Salaries of PhD's versus MSc's
  - Y measured in $ .. i.e. interval scale
  - same principles whether confounder is binary or interval
  - extreme confounding : example of Simpson's paradox
- Who is more likely to receive death penalty?
  - 2 binary "X" variables
  - extreme confounding: example 2 of Simpson's paradox
- Neonatal Mortality (Brand and Keirse article)
  - good e.g. of adjustment formula (crude -> adjusted logit)
  - age confounds M:F comparison, but not vice versa!
- Down's syndrome in relation to parity & maternal age
  - using age as an interval versus categorical variable
- Autism & MMR
  - modeling age-specific dx-rates as log-linear doesn't reduce the confounding (see exercise)

Demo

- Resting on a knife-edge: collinearity
  - Excel spreadsheet in session 4 of course 678

Readings (* = ?insufficient)

* Chapter 3.5 of Hosmer and Lemeshow

Other Resources

texts

Statistical methods for comparative studies: methods for bias reduction, by Anderson et al. -- 6 authors, whom I abbreviate to 'aahovw'. Extract from the relevant chapter 3 on the c622 website. Diagram for c678 adapted from these authors

articles

- Brand & Keirse: pair of very good expository articles (from Pediatric and Perinatal Epidemiology 1990; 4: 22-38) on logistic regression [c678, Resources/Material for sessions 9-11]
- "Appropriate Uses of Multivariate Analysis" (JH) from the Annual Review of Public Health in 1983.
  under Other Resources in the c622 website,
  See in particular the sections on two major uses of multiple regression:- to make comparisons
  FAIRER (reduce bias)
  SHARPER (increase precision).
- Figure in the Blood Pressure and Altitude article under resources for session 5 of course 678

Clear and present dangers

1 In the article CLINICAL PRACTICE GUIDELINE: ENDPOINTS OF RESUSCITATION, Tisherman reviews a number of studies, including Maynard, N 1993 "Assessment of splanchnic oxygenation by gastric tonometry in patients with acute circulatory, failure. JAMA 270:1203-10." and tells us that

> Although a variety of resuscitation endpoints correlated with surviving critical illness, only pHi at 24h proved an independent predictor of *death by logistic regression*

2 A title of a Table in another article..

> ADJUSTED ESTIMATES OF ASSOCIATION OF PATIENT CHARACTERISTICS WITH INITIALLY MISSED DIAGNOSIS, DELAYED TREATMENT, AND ***DEATH FROM MULTIVARIATE LOGISTIC REGRESSION***

Key Points in / Commentary on ALR Ch 3.5 —

- **The multivariable model *(3.5 page 65, para 1)***

  "One goal of such an analysis is to statistically adjust the estimated effect of ***each*** variable in the model for differences in the distributions of, and associations among the other independent variables " (italics by JH)

  Remark 1: yes, in some analyses (but certainly not in all, and not in cases of diagnostic and prognostic functions) this might be ***a*** goal

  Remark 2: in many studies, there is not the 'symmetry of role' among the X's that the authors imply: more often, there is ***one*** X variable ('exposure' for want of a better word) that is of ***primary*** interest, and the ***other*** X's are a ***nuisance***. The authors' use of the word ***each*** makes the aim broader than it really is. Indeed, in situations where the focus is on one X, it would be helpful to rewrite the equation as

  $$g[\ \mu \text{ or } \pi \mid \mathbf{X}\ \mathbf{Z}\ ] = \beta_0 + \beta_X X + \gamma_Z \mathbf{Z}$$

  where X is the 'eXposure' cariable, and **Z** is the collection of the confounding and other explanatory variables. [H&L do so in group & age e.g on p66)

  Mathematically speaking, there is a certain ***symmetry*** to the regression equations. In particular, in c621 situations, X2 being a confounder of the observed X1=>Y relationship implies that X1 is ***also*** a confounder of the observed X2=>Y relationship. For example, when we study the effect of duration of occupational exposure to noise on the prevalence hearing loss, we are, because of practical research design constraints, forced to confront, and 'adjust for', the simultaneous effect of age. But we would not add as a study aim "to adjust the estimated effect of ***age*** for differences in the distributions of exposure". Here, age is an extraneous variable, of no intrinsic interest, one that -- if we could do the study experimentally, on say animals, we would control directly, rather than by regression techniques. [This is not necessarily so in c681, where rules that apply in the 'regular' or identity scale, do not always hold in the logit scale: for example, in Brand and Keirse's first example (Table 9), age confounds the sex=>mortality relationship (if work with odds) but sex does ***not*** confound the age=>mortality relationship! This goes back to the rules for 'collapsibility, and confounding: these rules are different for different effect measure scales (i.e. the rules for confounding are different for odds ratios than they are for risk differences. I expect you will cover this again in the more advanced epidemiology classes).

  There may be ***some situations*** where ***two or more X's (e.g. risk factors) are of simultaneous interest***, and one is keenly interested in the true ***net*** effect of each. An example is the article "Parental periconceptional smoking and male: female ratio of newborn infants (on the 678 website under resources for session 9-11), where the authors assessed whether the smoking habits of (1) the mother and (2) the father around the time of conception affects the likelihood of the offspring being male or female, and one would like to ***isolate*** the two effects (mother, father) from each other.. i.e. have narrow interval estimates for each beta, i.e.. we seek simultaneous CI's for the two betas such that the two estimates are not highly correlated (often negatively, if the two 'exposures' are positively correlated in their distribution in the sample, and their true effects are in the same direction), or even if there are, are within narrow ranges

- **The multivariable model *(3.5 page 65, para 4)***

  "If the age (Z) distribution is also the same for the two groups, then a univariate analysis ***would suffice*** and we would compare the mean weight (Y) in the two groups (X=0 and X=1, " (italics by JH)

  Remark 1: YES, in c621, one would get the same estimate whether one omitted age or not. BUT including age would (if age contributes to the variation in Y, and presumably it does or else one would not be considering it) reduce the standard error (i.e. increase the precision) of the estimate.

  Remark 2: Paradoxically, this again is one of the places where the message from c621 does not automatically carry over to c681.. Because, at any X, the variance of a binary Y is very large relative to its mean, and because this variance [$\pi(1-\pi)$] changes with the mean [i.e., with π], it is sometimes difficult to say whether including a Z in a logistic model will decrease or increase the SE of the beta for the X of interest. Given that, and given that a sample of n binary Y's does not mean we have n effective degrees of freedom to fit as many parameters as we wish, we have to be more prudent about inclusion of Z's that in c621 "might not do much good, but would not do much harm either."

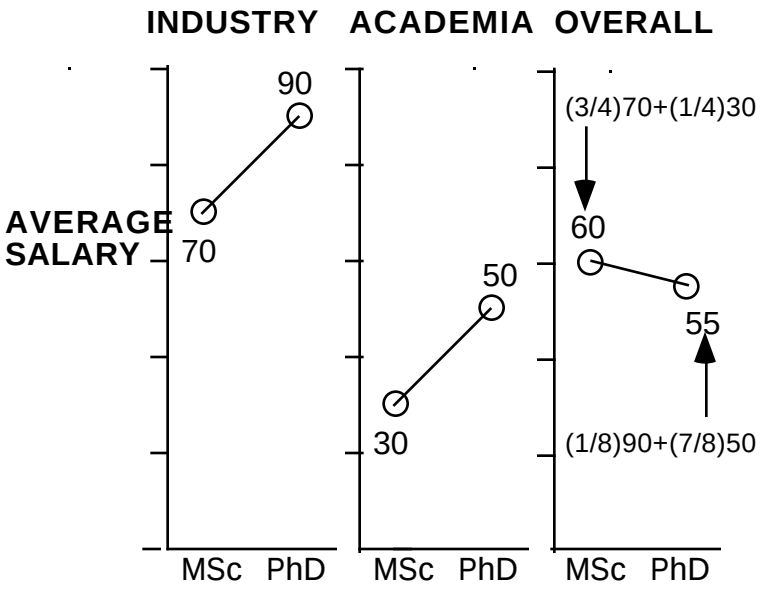
- **The multivariable model** (***analysis of covariance***) ***(3.5 page 65, para 5)***

  "This situation is described graphically in Figure 3.1"

  Fig 3.1 is AT THE CORE of the concept of statistical adjustment, and is repeated with colour, and lines that reflect ***where the data are***, on the right hand side of page 2 of what was handed out on confounding from c678. JH gives other examples later in the c681 notes that you are reading now, and points here to a textbook and articles where this is explained with greater graphical clarity and impact. The text is Statistical methods for comparative studies: methods for bias reduction, by Anderson et al. -- 6 authors, whom I abbreviate to 'aahovw'. I have put an extract from the relevant chapter 3 on the c622 website. Also, under Other Resources in the c622 website, is JH's article "Appropriate Uses of Multivariate Analysis" from the Annual Review of Public Health in 1983. It has two sections that explain -- graphically and with a few data points-- two big uses of multiple regression.. to make comparisons FAIRER (reduce bias) and SHARPER (increase precision). Finally, to see Fig 3.1 in a real situation, with much more striking graphics, you could examine the Figure in the Blood Pressure and Altitude article under resources for session 5 of course 678.

comments/corrections welcomed .. jh feb 9 2004

Here I adopt, and illustrate **confounding**, & "**adjustment by regression**" for an extreme 'continuous' Y" example [607Ch2, Simpson's paradox]



1.

```
DATA salary;
INPUT phd industry salary number;
LINES;
       0     1       70    750
       1     1       90    125
       0     0       30    250
       1     0       50    875
;
```

2.

```
PROC MEANS  DATA = salary MEAN;
 CLASS PHD; VAR salary;
 FREQ number;
```

PHD	N Obs	Mean_SALARY
No 0	1000	60
Yes 1	1000	55

3a

```
PROC REG DATA = salary ;
  MODEL salary = PhD ;
  FREQ number;
```

3b. Dep. Var: SALARY

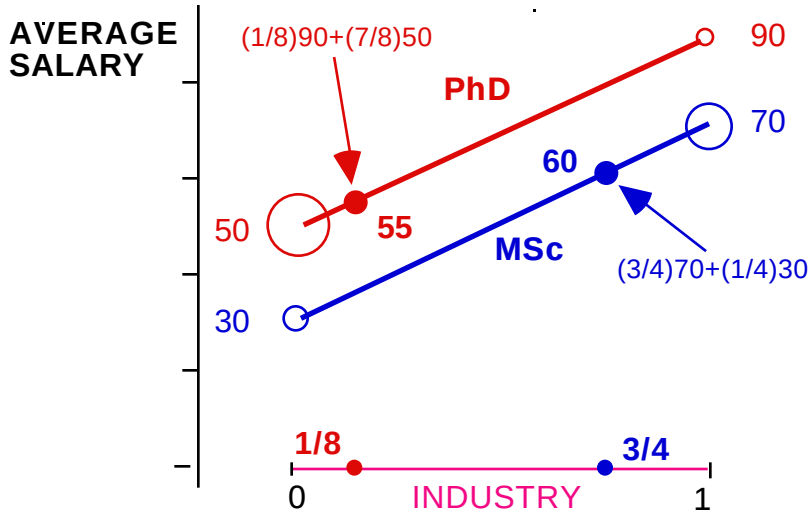
	Estimate	SE	
INTERCEP	60	0.49	
PHD	**-5**	0.69	**CRUDE Δ**

4a.

```
PROC SORT DATA = salary;
            by industry;
PROC REG  DATA = salary;
          by industry;
MODEL salary = PhD;
  FREQ number;
```

4b.

	INDUSTRY = 0		INDUSTRY = 1	
	Estimate	SE	Estimate	SE
INTERCEP	30	0	70	0
PHD	20	0	20	0



5a.

```
PROC REG  DATA = salary ;
MODEL salary = PhD industry;
  FREQ number;
```

5b.

	Estimates	SE
INTERCEP	30	0
PHD	20	0
INDUSTRY	40	0

Average Salary = 30 + 20 if PhD + 40 if Industry

6a.

```
PROC MEANS DATA = salary MEAN;
CLASS PHD; VAR industry;
FREQ number;
```

6b. Distrn. of CONFOUNDER

PHD	Mean of (Indicator) variable INDUSTRY	
No 0	0.750	(3/4)
Yes 1	0.125	(1/8)

From **2** or from **3b** ("CRUDE") $value of PhD: –5
From **4b** or from **5b** ("NET") $value of PhD: +20

Net (adjusted) value of PhD =

"CRUDE" value

MINUS

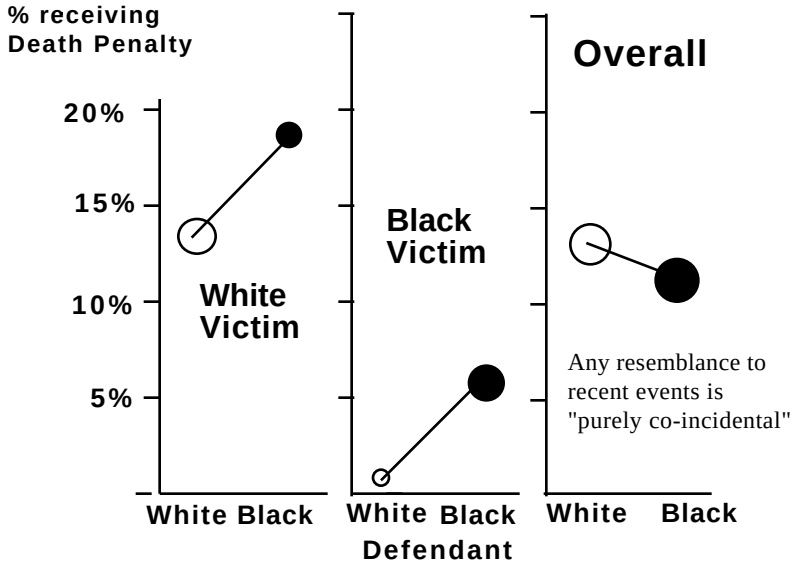
NET value of Industry × (π[Industry|PhD] – π[Industry|non-PhD])

–5

MINUS

40 × (0.125 – **0.750**) = –5 -40(–0.625) = +20

On the "participation in industry" scale, PhDs are at 0.125, and non_PhD's at 0.750, a difference of 0.625; the salary gap from 'not in industry' (industry = 0, ie academia) to in industry (=) is 40K




```
data a;
INPUT
   de_black
      vi_black
           n_death
                n_spared;
n_cases = n_death + n_spared;
LINES;
      0    0  19   132
      0    1   0     9
      1    0  11    52
      1    1   6    97
;
```

```
data b; set a;
death=1; nmbr = n_death ; output;
death=0; nmbr = n_spared; output;
PROC PRINT;
VAR de_black vi_black death nmbr;
```

DE_BLACK	VI_BLACK	DEATH	NMBR
0	0	1	19
0	0	0	132
0	1	1	0
0	1	0	9
1	0	1	11
1	0	0	52
1	1	1	6
1	1	0	97

```
PROC FREQ data=b;
  TABLES death * de_black /
     cmh norow nopercent;
  WEIGHT nmbr;
                    "CRUDE"
Type of
Study  Method   OR       CI
------------------------------
C-C    M-H    0.847  0.42 1.69
      Logit   0.847  0.42 1.69
```

	DE_BLACK		
DEATH	0	1	Total
0	141 (88.1%)	149 (89.8%)	290
1	19 (11.9%)	17 (10.2%)	36 (11%)
Total	160	166	326

```
PROC LOGISTIC data=a;
MODEL n_death/n_cases
        = de_black ;
  "CRUDE"
```

-2 LOG L 226.513(null) 226.291(model)

	Estimate	SE	ChiSq	P-value	OR
INTERCPT	-2.00	0.24	67.3	0.0001	.
DE_BLACK	-0.16	0.35	0.2	0.6382	**0.847**

```
PROC FREQ data=b;
TABLES
 vi_black*death*de_black
 / measures cmh;
 WEIGHT nmbr;
```

VI_BLACK=0

DEATH	DE_BLACK 0	DE_BLACK 1	Tot
0	132 (87.42)	52 (82.54)	184
1	19 (12.58)	11 (17.46)	30
Tot	151	63	214

OR 1.47 (0.65,3.3)

VI_BLACK=1

DEATH	DE_BLACK 0	DE_BLACK 1	Total
0	9 (100.00)	97 (94.17)	106
1	0 (0.00)	6 (5.83)	6
Tot	9	103	112

OR not computed - 0 cell

Estimates of Common OR

M-H ** **1.574** 0.701 3.533
Logit* **1.454** 0.667 3.173

** test-based.

* correction of 0.5 in every cell of tables that contain a zero.

```
PROC LOGISTIC data=a;
MODEL n_death/n_cases =
   de_black vi_black ;
OUTPUT OUT=fitted
PREDICTED = fitted_p ;
```

*cf 0.266 OR (crude)
victim white 1 (ref)
victim black 0.347

-2 LOGL 226.5(B0) 219.1(B0,B1,B2)
ChiSq: 7.4 with 2 DF (p=0.0243)

Var.	Est.	SE	ChiSq	P	OR
INTERCPT	-1.95	0.25	63.9	0.00	.
DE_BLACK	0.44	0.40	1.2	0.27	**1.553**
VI_BLACK	-1.32	0.52	6.5	0.01	0.266*

```
PROC PRINT
    data=fitted;
VAR de_black
    vi_black
    fitted_p;
```

DE_BL	VI_BL	FITTED_P	OBS_P	#OBS	#FITTED
0	0	0.1237	0.1258	19	18.7
0	1	0.0362	0.0000	0	0.3
1	0	0.1798	0.1746	11	11.3
1	1	0.0551	0.0583	6	5.7

NB: 4 proportions modeled by 3 parameters: close to 'saturated' model.

Other eg's: neonatal mortality (Brand/Keirse); Down's, parity, age.