

## Lecture 13

1. Confounding
2. Mediation
3. Reading spreadsheets with problems
4. CSV form of data
5. INFORMATS and FORMATS

1

We are interested in the effect of  $X$  on  $Y$ :

```
Proc GLM;  
  model Y = X;
```

We could add another predictor  $W$  to get an adjusted effect of  $X$  on  $Y$ :

```
Proc GLM;  
  model Y = X W;
```

What happens when we add  $W$ ?

2

$$(1) \quad Y = a_0 + a_1X + e_1$$

$$(2) \quad Y = b_0 + b_1X + b_2W + e_2$$

$a_1$  unadjusted (crude) effect of  $X$  from (1)

$b_1$  adjusted effect of  $X$  from (2)

$a_1 \neq b_1$  !

*adjusted  $b_1 < unadjusted a_1$*  is common,

*adjusted  $> unadjusted$*  can happen.

3

### Confounding

$$Y = b_0 + b_1X + b_2W + e$$

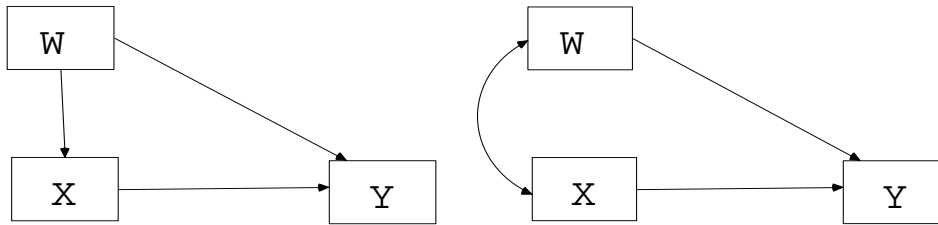
**Definition 1.** An adjustor  $W$  is *confounded* with  $X$ , the effect of interest, when  $W$  carries all or most of the same information in  $X$  about  $Y$ .

**Definition 2.** An adjustor  $W$  is *confounded* with  $X$ , the effect of interest, under any of these conditions:

- $W$  and  $X$  are both causal determinants of  $Y$
- $W$  is a causal determinant of  $X$
- another variable  $Z$  is a causal determinant of both  $X$  and  $W$

(Def 2 is from Vittinghoff, Shiboski, Glidden, McCulloch *Regression Methods in Biostatistics*, 2005, Chapter 4)

4



$$(1) \quad Y = a_0 + a_1X + e_1$$

$$(2) \quad Y = b_0 + b_1X + b_2W + e_2$$

$$(3) \quad W = c_0 + c_1X + e_3$$

5

Substitute (3) into (2):

$$\begin{aligned} Y &= b_0 + b_1X + b_2W + e_2 \\ &= b_0 + b_1X + b_2(c_0 + c_1X + e_3) + e_2 \\ &= (b_0 + b_2c_0) + (b_1 + b_2c_1)X + e_1 \end{aligned}$$

but this is the unadjusted regression

$$Y = a_0 + a_1X + e_1$$

6

so we have

$$a_1 = b_1 + b_2 c_1$$

$$a_1 - b_1 = b_2 c_1$$

$$\text{unadjusted effect} - \text{adjusted effect} = b_2 c_1$$

$$(1) \quad Y = a_0 + a_1 X + e_1$$

$$(2) \quad Y = b_0 + b_1 X + b_2 W + e_2$$

$$(3) \quad W = c_0 + c_1 X + e_3$$

7

Example 1. Buchanan vote in Florida, 2000

```
proc reg data=pred;
  title3 "Percent Bush alone";
  model log_buchanan = p_bush log_votes loghispanic
        income percent_65;
proc reg data=pred;
  title3 "Percent Bush + Percent Gore";
  model log_buchanan = p_bush p_gore log_votes loghispanic
        income percent_65;
proc reg data=pred;
  title3 "Percent Gore";
  model p_gore = p_bush ;
```

8

Percent Bush alone

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	-3.70402	0.44582	-8.31	<.0001
p_bush	1	0.01680	0.00507	3.31	0.0016
log_votes	1	0.93020	0.04490	20.72	<.0001
logphhispanic	1	-0.28132	0.04898	-5.74	<.0001
income	1	-0.05496	0.01270	-4.33	<.0001
percent_65	1	-0.01477	0.00675	-2.19	0.0325

Percent Bush + Percent Gore

Variable	DF	Estimate	SE	t Value	Pr >  t
Intercept	1	15.24473	7.29139	2.09	0.0409
p_bush	1	-0.17575	0.07412	-2.37	0.0210
p_gore	1	-0.19436	0.07466	-2.60	0.0117
log_votes	1	0.93612	0.04294	21.80	<.0001
logphhispanic	1	-0.25158	0.04815	-5.22	<.0001
income	1	-0.05565	0.01214	-4.59	<.0001
percent_65	1	-0.02031	0.00679	-2.99	0.0040

Percent Gore

Variable	DF	Estimate	SE	t Value	Pr >  t
Intercept	1	97.46656	0.40825	238.74	<.0001
p_bush	1	-0.99093	0.00730	-135.70	<.0001

9

$$\text{unadjusted effect} - \text{adjusted effect} = b_2c_1$$

$$\text{unadjusted effect} - \text{adjusted effect} = 0.01680 - (-0.17575) = 0.19255$$

$$b_2c_1 = (-0.19436) * (-0.99093) = 0.1925972$$

Dealing with confounding:

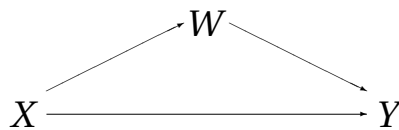
1. Make categories of the confounder, and stratify: look for effect of  $X$  when  $W$  is fixed
2. Check for interaction between  $X$  and  $W$ -categories: does effect of  $X$  change with category?

11

### Mediation

$$Y = b_0 + b_1X + b_2W + e$$

**Def** (Vittinghoff, *et.al.* p 96) A *mediating variable*  $W$  is hypothesized to lie on the causal pathway between  $X$  and  $Y$ , and thus to mediate the effect of  $X$  on  $Y$ .



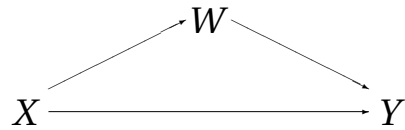
Example:

$X$  = pregnant mother's smoking status;  $W$  = gestational age;  $Y$  = birth weight.

This definition appeals to external information about the causal relations.

Mediation cannot distinguished from confounding by statistical results only.

12



(1)  $Y = a_0 + a_1X + e_1$

(2)  $Y = b_0 + b_1X + b_2W + e_2$

(3)  $W = c_0 + c_1X + e_3$

Checking whether a suspected mediator  $W$  may be acting as a mediator:

- $X$  predicts  $W$ , so  $c_1 \neq 0$
- $W$  predicts  $Y$ , adjusting for  $X$ , so  $b_2 \neq 0$
- adjusting for  $W$  attenuates the regression coefficient of  $X$ , so  $b_1 < a_1$

13

### Mediation terminology

(1)  $Y = a_0 + a_1X + e_1$

(2)  $Y = b_0 + b_1X + b_2W + e_2$

(3)  $W = c_0 + c_1X + e_3$

$a_1$  **total effect** of  $X$

$b_1$  **direct effect** of  $X$

$b_2c_1 = a_1 - b_1$  **mediated effect** of  $X$  (equality holds for linear regression only)

$(a_1 - b_1)/a_1$  proportion of total effect mediated

14

### (Baron-Kenny) Test for mediation

Null hypothesis of test: mediated effect of  $X = b_2c_1$  is zero

Estimate of mediated effect:  $\hat{b}_2\hat{c}_1$  (fitted regression coefficients)

Standard error of mediated effect:

$$SE(\hat{b}_2\hat{c}_1) = \sqrt{\hat{b}_2^2 SE(\hat{c}_1)^2 + \hat{c}_1^2 SE(\hat{b}_2)^2}$$

Test statistic:  $Z = (\hat{b}_2\hat{c}_1)/SE(\hat{b}_2\hat{c}_1)$ , compare to standard normal distribution.

15

This test follows the steps in Baron and Kenny (1986) The moderator-mediator variable distinction in social psychology research: conceptual, strategic and statistical considerations. *Journal of Personality and Social Psychology*, 51, 1173-1182.

Extensive review:

MacKinnon DP, *Introduction to Statistical Mediation Analysis*, 2008, Psychology Press, Taylor & Francis.

Alternative test: get 95% confidence interval for proportion of total effect mediated  $(a_1 - b_1)/a_1$  using bootstrap, see whether CI covers zero.

16

## Reading spreadsheets with problems

Download `bad_spreadsheet.xls` from the course website.

A	B	C	D	E	F	G	H	I	J
subject	city	gender	BMI	visit date	0 min	30 min	120 min	PredDose	Lymphs
1065	Mankato	F	25.6X	lost	12	NA	105	5	0.2
1066	Minneapolis	M	27.8		15		112	10	NA
1067	Cedar Rapids Iowa	male	31.2	11/13/2002	7	61.00	220	0	1.8
1068	Kansas City MO	F	29.4	10/24/81	19	143.00	386	10	0.7*

17

Import the spreadsheet into SAS as “import\_xls” using *SAS Import Wizard*.

Equivalent: Proc IMPORT (see *LSB* §2.16–2.17).

```
Proc PRINT data = bad_xls;
```

```
Proc CONTENTS data = bad_xls;
```

Proc Contents tells what SAS knows about a dataset, and the type of each variable.

18

Data Set Name		PH6470.BAD_XLS		Observations	5	
Member Type		DATA		Variables	12	
#	Variable	Type	Len	Format	Informat	Label
3	BMI	Num	8			BMI
10	F10	Char	1	\$1.	\$1.	F10
11	F11	Char	1	\$1.	\$1.	F11
12	F12	Char	1	\$1.	\$1.	F12
9	Lymphs	Num	8			Lymphs
8	PredDose	Num	8			PredDose
6	_0_min	Num	8			30 min
7	_20_min	Num	8			120 min
5	__min	Num	8			0 min
2	city	Char	17	\$17.	\$17.	city
1	subject	Num	8			subject
4	visit_date	Char	8	\$8.	\$8.	visit date

19

### Reading a .CSV file (LSB §2.15)

When the Data Import Wizard cannot correctly read an Excel spreadsheet, save it as a CSV file and tell SAS exactly how to read each variable.

CSV means *comma separated values*.

In Excel save the spreadsheet in CSV format as bad\_spreadsheet\_1.csv.

Use a data step to read in the CSV file:

```
Data new;
  INFILE "path to file" firstobs=2 DLM="," DSD missover
  lrecl=100;

  INPUT list of variables, with required INFORMATS;
```

20

These are the options for the INFILE statement:

`firstobs=2` skip the first line with variable names and start reading at line 2

`DLM =","` specifies the *delimiter*, the thing that separates variables (a comma)

`DSD` treats `, ,` as a missing value

`missover` If there are more variables to read at the end of the data line, set them to missing instead of continuing on to the next line for them.

`lrecl` = logical record length. If your data lines are longer than 100 characters, pick a large number for this. SAS log will tell you the actual maximum record length and you can correct it.

21

```
Data new;
```

```
  INFILE "path to file" firstobs=2 DLM="," DSD missover  
        lrecl=100;
```

```
  INPUT list of variables, with required INFORMATS ;
```

INFORMAT is a format for input data

FORMAT is a format for output data

See *LSB §2.8* for table of useful INFORMATS.

22

Character data: specify width = number  $w$  of characters. `$w.`

Trims leading blanks.

`Colon ( :$w. )` read until  $w$  characters or delimiter

Dates:

`DATEw.` reads dates in form *ddmmmyy* or *ddmmmyyyy*

e.g. `DATE10.` reads 05sept1998 or 1 JAN 07

`DDMMYYw.` read dates in form *ddmmmyy* or *ddmmmyyyy*

e.g. `ddmmyy10.` reads 05/09/1998

23

The date INFORMATs also work as FORMATs to write SAS dates as intelligible dates.

See the log file for a list of problems reading the data.

Use Proc Contents and Proc Print to check that SAS is reading the data correctly.

```
proc print;
  title3 "CSV version 2: NAs, X, * removed";

data read_csv3;
  infile "C: . . bad_spreadsheet_2.csv"
    DSD dlm="," firstobs=2 missover lrecl=300;
  input subject city :$17. gender $ BMI visit_date :mddy10.
    min0 min30 min120 PredDose Lymphs;
  if (subject NE .); * delete blank lines;
  format visit_date mddy10. ;
```