



## Colliders and selection bias

- Being a firefighter is a collider
- Using it as an inclusion criterion for study
  - opens backdoor path between exercise and CHD
  - as if it was a variable and we controlled for it
- Controlling for SES reblocks path, removing selection bias
  - not a confounder in this DAG (but probably should be)

5

## Interpreting results for log-transformed variables

- Positive continuous variables commonly log-transformed
  - outcomes: normalize and equalize variance
  - predictors: get rid of non-linearity, interaction
  - *because this reflects how you think things work*
- Both  $\log_{10}$  and natural log transformations used
- How does this affect interpretation of regression coefficients?

6

## Log-transformed predictors

- Models for diminishing dose-response

$$E(y) = \beta_0 + \beta_1 \ln(x)$$

For example,  $\beta_1 = 10/\ln(2)$  models a 10-unit increase in  $E(y)$  for each dose doubling

$x$	$E(y)$
1	10
2	20
4	30
8	40
16	50

7

## Log-transformed predictors

- For natural-log transformed predictor  $x_j$ 
  - $\beta_j$  gives increase in  $E(y)$  for 1-unit increase in  $\log(x_j)$  – equivalently a 2.7-fold increase in  $x_j$
- If  $x_j$  is  $\log_{10}$ -transformed,  $\beta_j$  gives increase in  $E(y)$  for 1-unit increase in  $\log(x_j)$  or a 10-fold increase in  $x_j$
- How can we make this more flexible?

8

## Log-transformed predictors

- For natural-log transformed predictor  $x_j$ 
  - $\beta_j \ln(1 + k/100)$ : change in  $E(y)$  for a  $k\%$  increase in  $x_j$ .
  - e.g.,  $\beta_j \ln(1.5)$ : change in  $E(y)$  for a 50% increase in  $x_j$ .
- Use  $\beta_j \log_{10}(1 + k/100)$  if  $x_j$  is  $\log_{10}$ -transformed
- Use `nlcom` to get estimates with confidence intervals

9

## SBP and log-transformed creatinine

```
. reg sbp age10 lncreat diabetes
```

Source	SS	df	MS	Number of obs = 2761		
Model	62724.9657	3	20908.3219	F( 3, 2757) = 61.70		
Residual	934341.037	2757	338.897728	Prob > F = 0.0000		
				R-squared = 0.0629		
				Adj R-squared = 0.0619		
				Root MSE = 18.409		

sbp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
age	.444083	.0536661	8.27	0.000	.3388531	.5493129
lncreat	9.211717	1.682269	5.48	0.000	5.913081	12.51035
diabetes	6.426345	.8007529	8.03	0.000	4.856209	7.996481
_cons	103.2765	3.600996	28.68	0.000	96.21558	110.3374

```
. * increase in SBP associated with a 50% increase in creatinine
. nlcom _b[lncreat]*log(1.5)
```

sbp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
_nl_1	3.73503	.6821016	5.48	0.000	2.397548	5.072511

10

## Natural log-transformed outcomes

- Models increasing or decreasing exponential dose-response

$$E[\ln(y)] = \beta_0 + \beta_1 x$$

For example,  $\beta_1 = \pm \ln(10)$  models a 10-fold change in  $E(y)$  for each unit increase in  $x$

$x$	$E(y)$	
	$\beta_1 = \ln(10)$	$\beta_1 = -\ln(10)$
1	10	0.1
2	100	0.01
3	1,000	0.001
4	10,000	0.0001

- Interpretation is approximate:  $E[\ln(y)] \neq \ln(E[y])$

11

## Natural log-transformed outcomes

- $e^{\beta_j}$  gives *relative increase* in  $E(y)$  for a 1-unit increase in  $x_j$ 
  - if  $e^{\beta_j} = 1.5$ , a 1.5-fold increase in  $E(y)$ , from 1 to 1.5, 10 to 15, 100 to 150, ... for a 1-unit increase in  $x_j$
- $100(e^{\beta_j} - 1)$  gives *percent increase* in  $E(y)$ 
  - a 50% increase in  $E(y)$  for a 1-unit increase in  $x_j$

12

### Natural log-transformed outcomes

- For untransformed predictors:
  - use `eform("exp(beta)")` to have `regress` display  $e^{\beta_j}$
  - use `lincom` with `eform` option to get relative increase in outcome per  $k$ -unit increase in predictor
  - use `nlcom` to get percent increase in outcome per  $k$ -unit increase in predictor

### Outcome and predictor both log-transformed

- Models exponential response to proportional increases in dose

$$E[\ln(y)] = \beta_0 + \beta_1 \ln(x)$$

For example,  $\beta_1 = \ln(10)/\ln(2)$  models an approximate 10-fold increase in  $E(y)$  for each doubling of  $x$

x	E(y)
1	10
2	100
4	1,000
8	10,000
16	100,000

- Interpretation also approximate

### Outcome and predictor both log-transformed

- $e^{\beta_j \ln(1+k/100)}$  gives relative increase in  $E(y)$  for a  $k\%$  increase in  $x_j$
- $100(e^{\beta_j \ln(1+k/100)} - 1)$  gives percent increase in  $E(y)$
- Use `eform` with `regress`, `nlcom` for estimates with CIs
- See VGSM, Sect. 4.7.5

### Log-transformed creatinine and TG

```
. reg lncr bmi age i.raceth educ i.drinkamt lessact lntg, eform("exp(beta)")
```

Source	SS	df	MS	Number of obs = 2750		
Model	10.7976826	9	1.19974251	F( 9, 2740) = 29.05		
Residual	113.167246	2740	.041301915	Prob > F = 0.0000		
Total	123.964929	2749	.045094554	R-squared = 0.0871		
				Adj R-squared = 0.0841		
				Root MSE = .20323		

lncreat	exp(beta)	Std. Err.	t	P> t	[95% Conf. Interval]	
bmi	1.000855	.0007458	1.15	0.251	.9993939	1.002318
lntg	1.051728	.0105646	5.02	0.000	1.031215	1.072649

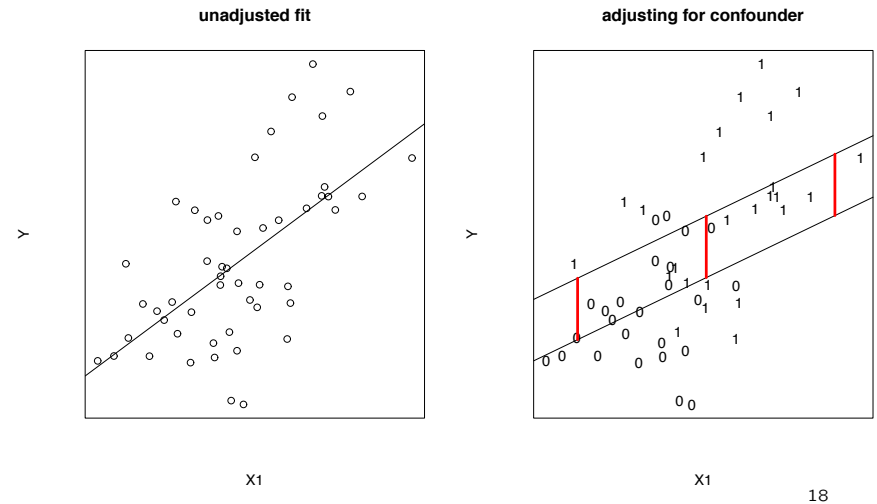
```
. * Percent increase in creatinine for a 25% increase in tryglycerides
. nlcom 100*(exp(_b[lntg]*log(1.25))-1)
       _nl_1: 100*(exp(_b[lntg]*log(1.25))-1)
```

lncreat	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
_nl_1	1.131766	.2266849	4.99	0.000	.6872758	1.576257

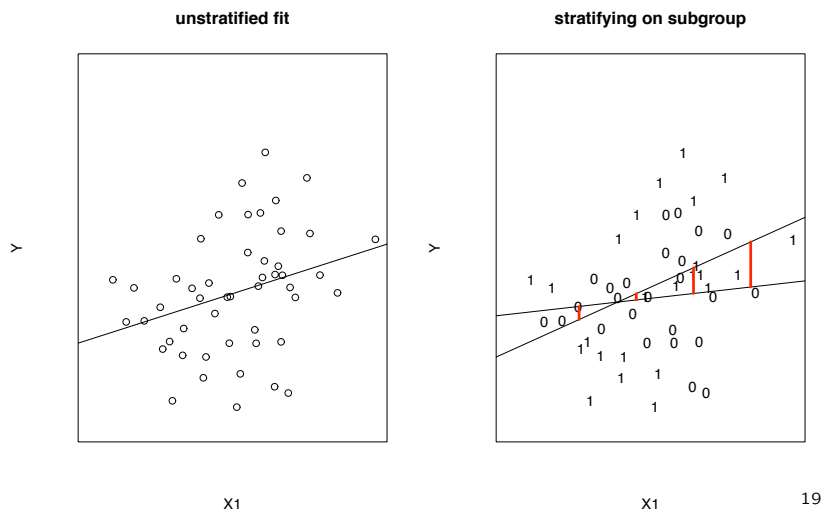
## Interaction

- Confounding and mediation: association of primary predictor with outcome *differs after adjustment*
  - adjustment variable *explains* or *masks* unadjusted effect of predictor
- Interaction: association of primary predictor with the outcome *differs across levels of the interaction variable*
  - interaction variable *modifies* effect of primary predictor

17



18



19

## Example: interaction of BMI and physical activity (PA) effects on SBP

- Excess weight increases SBP, but PA may blunt this effect
- Does higher BMI lead to higher SBP only among inactive?
- Does protective effect of PA on SBP increase with BMI?

20

## Modeling interaction of BMI and PA

- “Main effects”:
  - Centered BMI:  $BMI_c = BMI - 28.6$  (sample mean)
  - PA: active (0 = inactive, 1 = active)
- Interaction (product) term  $BMI_c \times active$ :
  - $BMI_c \times active = BMI_c \times active$
  - $= 0$  if  $active = 0$
  - $= BMI_c$  if  $active = 1$

21

## Interaction model: effects of PA

- Model for SBP:
  - $E[SBP|x] = \beta_0 + \beta_1 active + \beta_2 BMI_c + \beta_3 BMI_c \times active$
- Difference in mean SBP by PA:
  - $E[SBP|active = 1, BMI_c] = \beta_0 + \beta_1 + (\beta_2 + \beta_3) BMI_c$
  - $E[SBP|active = 0, BMI_c] = \beta_0 + \beta_2 BMI_c$
- Difference is  $\beta_1 + \beta_3 BMI_c$
- Effect of PA depends on BMI

22

## Interaction model: effects of BMI

- $E[SBP|x] = \beta_0 + \beta_1 active + \beta_2 BMI_c + \beta_3 BMI_c \times active$
- For each 1-unit increase in BMI among the *inactive*,
  - $E[SBP|active = 0, BMI_c = (k + 1)] = \beta_0 + \beta_2(k + 1)$
  - $E[SBP|active = 0, BMI_c = k] = \beta_0 + \beta_2 k$
- Increment in mean SBP is  $\beta_2$

23

## Interaction model: effects of BMI

- For each 1-unit increase in BMI among the *active*,
  - $E[SBP|active = 1, BMI_c = (k + 1)] = \beta_0 + \beta_1 + (\beta_2 + \beta_3)(k + 1)$
  - $E[SBP|active = 1, BMI_c = k] = \beta_0 + \beta_1 + (\beta_2 + \beta_3)k$
- Increment in mean SBP is  $\beta_2 + \beta_3$
- Effect of BMI depends on PA

24

```

. reg SBP BMic active BMiCActive
-----
      SBP |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----+-----
   _cons |   135.5883   .7508309    180.58  0.000    134.116   137.0605
   active |   -0.5468103   .8611566    -0.63  0.525   -2.235388   1.141768
   BMic   |    .0512451   .1138269     0.45  0.653   -0.1719495   .2744397
   BMiCActive |  .2293768   .1406966     1.63  0.103   -0.0465047   .5052583
-----+-----

. lincom active + 1.4 * BMiCActive
-----
      sbp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----+-----
   (1) |  -0.2256828   .8505544    -0.27  0.791   -1.893472   1.442106
-----+-----

. lincom BMic + BMiCActive
-----
      SBP |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----+-----
   (1) |    .2806219   .0826981     3.39  0.001    .1184653   .4427785
-----+-----

```

25

### Interpretation of coefficients: active ( $\beta_1$ )

- Interpretation: effect of being active, if  $\text{BMiC} = 0$

$$E[\text{SBP}|x] = \beta_0 + \beta_1 \text{active} + \beta_2 \text{BMiC} + \beta_3 \text{BMiCActive}$$

$$E[\text{SBP}|\text{active} = 1, \text{BMiC} = 0] = \beta_0 + \beta_1$$

$$E[\text{SBP}|\text{active} = 0, \text{BMiC} = 0] = \beta_0$$

- Difference is  $\beta_1$
- At any other value of  $\text{BMiC}$ , difference would involve  $\beta_3$
- At BMI of  $28.6 \text{ kg/m}^2$ , PA reduces SBP by  $0.55 \text{ mmHg}$

27

### Interpretation of coefficients: \_cons ( $\beta_0$ )

- Interpretation: average SBP in inactive participants with  $\text{BMiC} = 0$  (i.e., BMI at mean of  $28.6 \text{ kg/m}^2$ )

$$E[\text{SBP}|x] = \beta_0 + \beta_1 \text{active} + \beta_2 \text{BMiC} + \beta_3 \text{BMiCActive}$$

$$E[\text{SBP}|\text{active} = \text{BMiC} = \text{BMiCActive} = 0] = \beta_0$$

- Average SBP among inactive with  $\text{BMI} = 28.6 \text{ kg/m}^2$  is  $135.6 \text{ mmHg}$

26

### Interpretation of first lincom result:

active + 1.4\*BMiCActive

- Interpretation: effect of being active, if  $\text{BMiC} = 1.4$

$$E[\text{SBP}|x] = \beta_0 + \beta_1 \text{active} + \beta_2 \text{BMiC} + \beta_3 \text{BMiCActive}$$

$$E[\text{SBP}|\text{active} = 1, \text{BMiC} = 1.4] = \beta_0 + \beta_1 + 1.4(\beta_2 + \beta_3)$$

$$E[\text{SBP}|\text{active} = 0, \text{BMiC} = 1.4] = \beta_0 + 1.4\beta_2$$

- Difference is  $\beta_1 + 1.4\beta_3$
- At BMI of  $30 \text{ kg/m}^2$ , PA reduces SBP by  $0.23 \text{ mmHg}$

28

**Interpretation of coefficients:  $\text{bmic}$  ( $\beta_2$ )**

- $\text{bmic}$  ( $\beta_2$ ): effect of 1-unit increase in BMI, if inactive

$$E[\text{SBP}|x] = \beta_0 + \beta_1\text{active} + \beta_2\text{BMic} + \beta_3\text{BMicXactive}$$

$$E[\text{SBP}|\text{active} = 0, \text{BMic} = k + 1] = \beta_0 + (k + 1)\beta_2$$

$$E[\text{SBP}|\text{active} = 0, \text{BMic} = k] = \beta_0 + k\beta_2$$

- Difference is  $\beta_2$  for any value of  $k$
- Among inactive, a 1-unit increase in BMI increases SBP by 0.05 *mmHg*

**Interpretation of second  $\text{lincom}$  result:**

$$\text{BMic} + \text{BMicXactive} (\beta_2 + \beta_3)$$

- Interpretation: effect of 1-unit increase in BMI, if active

$$E[\text{SBP}|x] = \beta_0 + \beta_1\text{active} + \beta_2\text{BMic} + \beta_3\text{BMicXactive}$$

$$E[\text{SBP}|\text{active} = 1, \text{BMic} = k + 1] = \beta_0 + \beta_1 + (k + 1)(\beta_2 + \beta_3)$$

$$E[\text{SBP}|\text{active} = 1, \text{BMic} = k] = \beta_0 + \beta_1 + k(\beta_2 + \beta_3)$$

- Difference is  $\beta_2 + \beta_3$  for any value of  $k$
- Among active, a 1-unit increase in BMI increases SBP by 0.28 *mmHg*

**Interpretation of coefficients:  $\text{BMicXactive}$  ( $\beta_3$ )**

1. Difference between BMI effects in active and inactive

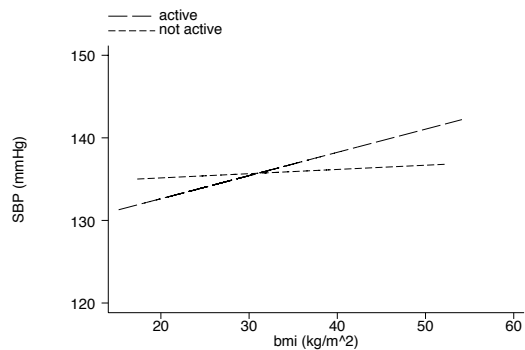
- among active, a 1-unit increase in BMI predicts a 0.23 *mmHg* greater increase in SBP than among inactive

2. Change in effect of PA for every 1-unit increase in BMI

- for every 1-unit increase in BMI, the beneficial effect of PA on SBP decreases by 0.23 *mmHg*

3.  $t$ -test and  $P$ -value for  $\text{BMicXactive}$  refer to test of interaction between BMI and PA

```
. reg SBP BMic active BMicXactive
.....
-----+-----
      SBP |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----+-----
      _cons |    135.5883   .7508309    180.58  0.000     134.116     137.0605
      active |   -0.5468103  .8611566    -0.63  0.525    -2.235388     1.141768
      BMic   |    0.0512451  .1138269     0.45  0.653    -0.1719495     0.2744397
      BMicXac`e |  0.2293768  .1406966     1.63  0.103    -0.0465047     0.5052583
-----+-----
. * Effect of PA if BMI = 30
. lincom active + 1.4 * BMicXactive
-----+-----
      sbp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----+-----
      (1) |   -0.2256828  .8505544    -0.27  0.791    -1.893472     1.442106
-----+-----
. * Effect of 1-unit increase in BMI if active
. lincom BMic + BMicXactive
-----+-----
      SBP |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----+-----
      (1) |    0.2806219  .0826981     3.39  0.001     0.1184653     0.4427785
-----+-----
```



### Figuring out what lincom statement to use

To estimate effect of PA at BMI 1.4  $kg/m^2$  above sample mean:

$$E[SBP|x] = \beta_0 + \beta_1 \text{active} + \beta_2 \text{BMIC} + \beta_3 \text{BMIC} \times \text{active}$$

Active	BMIC	_cons	active	BMIC	BMICXactive
yes	1.4	1	1	1.4	1.4
no	1.4	1	0	1.4	0
	Diff.	0	1	0	1.4

Use lincom active + 1.4\*BMICXactive

33

34

### Figuring out what lincom statement to use

To estimate effect of additional 1.4  $kg/m^2$  among the active:

$$E[SBP|x] = \beta_0 + \beta_1 \text{active} + \beta_2 \text{BMIC} + \beta_3 \text{BMIC} \times \text{active}$$

Active	BMIC	_cons	active	BMIC	BMICXactive
yes	$k + 1.4$	1	1	$k + 1.4$	$k + 1.4$
yes	$k$	1	1	$k$	$k$
	Diff.	0	0	1.4	1.4

Use lincom 1.4\*(BMIC + BMICXactive)

35

### Figuring out what lincom statement to use

To estimate effect of additional 1.4  $kg/m^2$  among inactive:

$$E[SBP|x] = \beta_0 + \beta_1 \text{active} + \beta_2 \text{BMIC} + \beta_3 \text{BMIC} \times \text{active}$$

Active	BMIC	_cons	active	BMIC	BMICXactive
no	$k + 1.4$	1	0	$k + 1.4$	0
no	$k$	1	0	$k$	0
	Diff.	0	0	1.4	0

Use lincom 1.4\*BMIC

36

## Average causal effect (ACE) of PA on SBP

- $\hat{\beta}_1 = -0.55 \text{ mmHg}$  estimates ACE because we centered BMI
- ACE: difference in mean SBP if everyone vs no one exercised  

$$E[\text{SBP}|x] = \beta_0 + \beta_1 \text{active} + \beta_2 \text{BMIc} + \beta_3 \text{BMIcXactive}$$
- Averaging over everyone,  $E[\text{BMIc}] = E[\text{BMIcXactive}] = 0$
- So ACE =  $\beta_0 + \beta_1 - \beta_0 = \beta_1$

37

## Interaction of hormone therapy and statins

- HT reduced LDL in HERS. Smaller effect in statin users?
- Interaction again modeled using three predictors:
  - statins: 0-1 indicator of statin use at baseline
  - ht: 0-1 indicator of assignment to HT
  - htstat = statins  $\times$  ht
- Centering does not help interpretability when predictors are binary or categorical

38

## HT and statins interaction model

$$E[\text{LDL}|x] = \beta_0 + \beta_1 \text{statins} + \beta_2 \text{ht} + \beta_3 \text{htstat}$$

group	statins	treatment	_cons	statins	ht	htstat	E[LDL x]
1	yes	HT	1	1	1	1	$\beta_0 + \beta_1 + \beta_2 + \beta_3$
2	yes	placebo	1	1	0	0	$\beta_0 + \beta_1$
3	no	HT	1	0	1	0	$\beta_0 + \beta_2$
4	no	placebo	1	0	0	0	$\beta_0$

- HT effect in users of statins: group 1 - group 2 =  $\beta_2 + \beta_3$
- HT effect in non-users of statins: group 3 - group 4 =  $\beta_2$
- Difference in HT effects by statin use:  $\beta_3$

39

## HT and statins interaction model

```
. reg ldl1 stat ht htstat
```

Source	SS	df	MS	Number of obs =
Model	220840.637	3	73613.5456	2608
Residual	3749008.16	2604	1439.71128	F( 3, 2604) = 51.13
Total	3969848.8	2607	1522.76517	Prob > F = 0.0000
				R-squared = 0.0556
				Adj R-squared = 0.0545
				Root MSE = 37.944

ldl1	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
_cons	145.1026	1.338998	108.37	0.000	142.477 147.7282
statins	-13.26844	2.142072	-6.19	0.000	-17.46878 -9.068105
ht	-17.81947	1.887797	-9.44	0.000	-21.5212 -14.11773
htstat	6.286759	3.062286	2.05	0.040	.2819978 12.29152

```
. lincom ht + htstat
( 1) ht + htstat = 0
```

ldl1	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
(1)	-11.53271	2.411186	-4.78	0.000	-16.26074 -6.804671

40

### Interpretation of coefficients: `_cons` ( $\beta_0$ )

- Mean year-1 LDL in placebo statin non-users

$$E[\text{LDL}|x] = \beta_0 + \beta_1\text{statins} + \beta_2\text{ht} + \beta_3\text{htstat}$$

$$E[\text{LDL}|\text{statins} = 0, \text{ht} = 0] = \beta_0$$

- Mean year-1 LDL among placebo participants not using statins was 145 *mg/dL*

41

### Interpretation of coefficients: `ht` ( $\beta_2$ )

- HT effect among statin *non*-users

$$E[\text{LDL}|x] = \beta_0 + \beta_1\text{statins} + \beta_2\text{ht} + \beta_3\text{htstat}$$

$$E[\text{LDL}|\text{statins} = 0, \text{ht} = 1] = \beta_0 + \beta_2$$

$$E[\text{LDL}|\text{statins} = 0, \text{ht} = 0] = \beta_0$$

- Among statin non-users, HT lowered LDL by 17.8 *mg/dL*
- Confounded?

43

### Interpretation of coefficients: `statins` ( $\beta_1$ )

- Statin “effect” in placebo group

$$E[\text{LDL}|x] = \beta_0 + \beta_1\text{statins} + \beta_2\text{ht} + \beta_3\text{htstat}$$

$$E[\text{LDL}|\text{statins} = 1, \text{ht} = 0] = \beta_0 + \beta_1$$

$$E[\text{LDL}|\text{statins} = 0, \text{ht} = 0] = \beta_0$$

- In placebo group, year-1 LDL 13.3 *mg/dL* lower in statins users than non-users
- Confounded?

42

### Interpretation of `lincom` result:

$$\text{ht} + \text{htstat} (\beta_2 + \beta_3)$$

- HT effect on LDL in statin users

$$E[\text{LDL}|x] = \beta_0 + \beta_1\text{statins} + \beta_2\text{ht} + \beta_3\text{htstat}$$

$$E[\text{LDL}|\text{statins} = 1, \text{ht} = 1] = \beta_0 + \beta_1 + \beta_2 + \beta_3$$

$$E[\text{LDL}|\text{statins} = 1, \text{ht} = 0] = \beta_0 + \beta_1$$

- Among statin users, HT lowered LDL by 11.5 *mg/dL*
- Confounded?

44

### Interpretation of coefficients: htstat ( $\beta_3$ )

- First interpretation: difference in HT effects by statin use
  - $\beta_2 + \beta_3$ : effect of assignment to HT among statin users
  - $\beta_2$ : effect of assignment to HT among statin *non*-users
- HT effect on LDL 6.3 *mg/dL* smaller in statin-users

45

### Figuring out what lincom statement to use

Estimate HT effect in statin users:

$$E[LDL|x] = \beta_0 + \beta_1 \text{statins} + \beta_2 \text{ht} + \beta_3 \text{htstat}$$

Statins	HT	_cons	statins	ht	htstat
yes	yes	1	1	1	1
yes	no	1	1	0	0
	Diff.	0	0	1	1

Use `lincom ht + htstat`

47

### Interpretation of coefficients: htstat ( $\beta_3$ )

- Second interpretation: difference of treatment-specific statin differences
- We already showed that  $\beta_1$  gives difference in mean year-1 LDL by statin use, among women assigned to placebo
- Could easily show that  $\beta_1 + \beta_3$  gives corresponding difference by statin use among women assigned to HT
- Subtracting would give the result
- Confounded?

46

### Figuring out what lincom statement to use

Estimate statin "effect" in HT:

$$E[LDL|x] = \beta_0 + \beta_1 \text{statins} + \beta_2 \text{ht} + \beta_3 \text{htstat}$$

Statins	HT	_cons	statins	ht	htstat
yes	yes	1	1	1	1
no	yes	1	0	1	0
	Diff.	0	1	0	1

Use `lincom statins + htstat`

48

## Did we estimate ACE for hormone therapy?

- Covariate `statins` was not centered, for interpretability

$$E[LDL|x] = \beta_0 + \beta_1 \text{statins} + \beta_2 \text{ht} + \beta_3 \text{htstat}$$

- Without centering  $E[\text{statins}] = E[\text{htstat}] \neq 0$
- $\beta_2$ : causal effect of HT among statin non-users
- ACE for HT:  $\beta_2 + \beta_3 E[\text{statins}]$ 
  - $E[\text{statins}]$ : target population prevalence of statin use

49

## Interactions between multi-category predictors

- Hard to detect and interpret
- Example: predictors with 3 and 4 categories:
  - $(3 - 1) \times (4 - 1) = 6$  interaction parameters
- Hand-made product terms like `htstat` don't work; must use
  - version 11: `i.catvar1##i.catvar2`
  - earlier versions: `i.catvar1 i.catvar2 i.catvar1*i.catvar2`

51

## Interactions between continuous predictors

- Be careful with interactions between continuous variables
  - non-linearity, influential points can be a problem
  - easier to interpret if one or both dichotomized, but information gets lost
- Restricted cubic splines work (next week), but results only interpretable by plotting

50

## Data dredging

- With 10 predictors
  - 45 possible 2-way interactions
  - 120 possible 3-way interactions
- Considering them all inflates type-I error, uncovers trivial interactions in big datasets

52

## Focused data dredging

- Focus on plausible interactions
  - treatment and severity of disease
  - treatment and calendar time
  - age and risk factors
  - measurement and state during measurement (Harrell *et al.*, *Stat Med*, 1996;15:361-87)
  - *with primary predictor*

53

## If we hadn't centered BMI ...

- In BMI and PA example, if we did not center BMI
  - $\beta_0$ : average SBP if inactive, BMI = 0
  - $\beta_1$ : effect of PA if BMI = 0
  - effect of PA at sample average BMI requires `lincom`
- With centered BMI:
  - $\beta_0$  and  $\beta_1$  refer to ppts. with average BMI
  - Coefficient for active estimates ACE

54

## Other issues: what p-value cutoff to use

- Keep interactions with  $p < 0.2$  because power is low, or ignore if  $p > 0.05$ ?
- Trying to prove something (e.g., demonstrate differential treatment effects in a subgroup and its complement)?
  - need strong evidence, especially for *ad hoc* finding, possibly penalized for multiple comparisons
- Weak evidence for interaction that undermines your controversial hypothesis?
  - good to report as a sensitivity analysis

55

## Summary: Interaction

- Interaction: association of one predictor with outcome modified by another predictor
- Modeled using product terms
- Subgroup-specific regression lines not parallel
- Centering makes “main effects” easier to interpret
- Power not always low, but can be

56

## Wang et al., NEJM, 2008;357(21):2189-94

- Not enough to show a within-subgroup effect; need to show effects significantly differ by subgroup
- Focus on *a priori* interactions, those with primary predictor
- Evaluate interactions for strength, biological plausibility
- Identify *post hoc* interactions
- In ambiguous cases, present both overall and stratified results

57

## Mediation vs confounding

- Mediators are on causal pathway from predictor to outcome
- In multi-predictor models, mediation and confounding behave alike, must be distinguished on substantive grounds
- Some problems too complex for this paradigm:
  - LDL both confounds, mediates statin effects on CHD
- You have the substantive expertise to draw the DAG

59

## Confounding, mediation, and interaction

- Confounders are a cause of both outcome and primary predictor, or surrogates for such a cause
- Potential confounders must be associated with predictor and independently with the outcome
- Confounding can account for the some or all of the unadjusted association between a predictor and an outcome

58

## Interaction vs confounding/mediation

- Important to distinguish between:
  - *change* in the coefficient for a primary predictor in two models *with and without adjustment for a confounder or mediator*, which we sometimes use to assess confounding and/or mediation
  - *differences* in the estimated effect of a primary predictor *across strata defined by an effect modifier* in a single model that includes interaction term(s)

60

## **Confounding, mediation, or interaction?**

1. Is alendronate effect on BMD reduced in vitamin-D deficient patients?
2. Is alendronate effect on vertebral fracture explained by changes in BMD?
3. Is HT effect on CHD explained by a healthier lifestyle?
4. Are beta-blockers less effective among African Americans?
5. Is diabetes a stronger risk factor for CHD events in women than in men?

61

## **Homework review sessions**

- Homework 2: Tuesday, February 9, 12-1, rm 6702
- Homework 3: Tuesday, February 24, 12-1, rm 6702

62