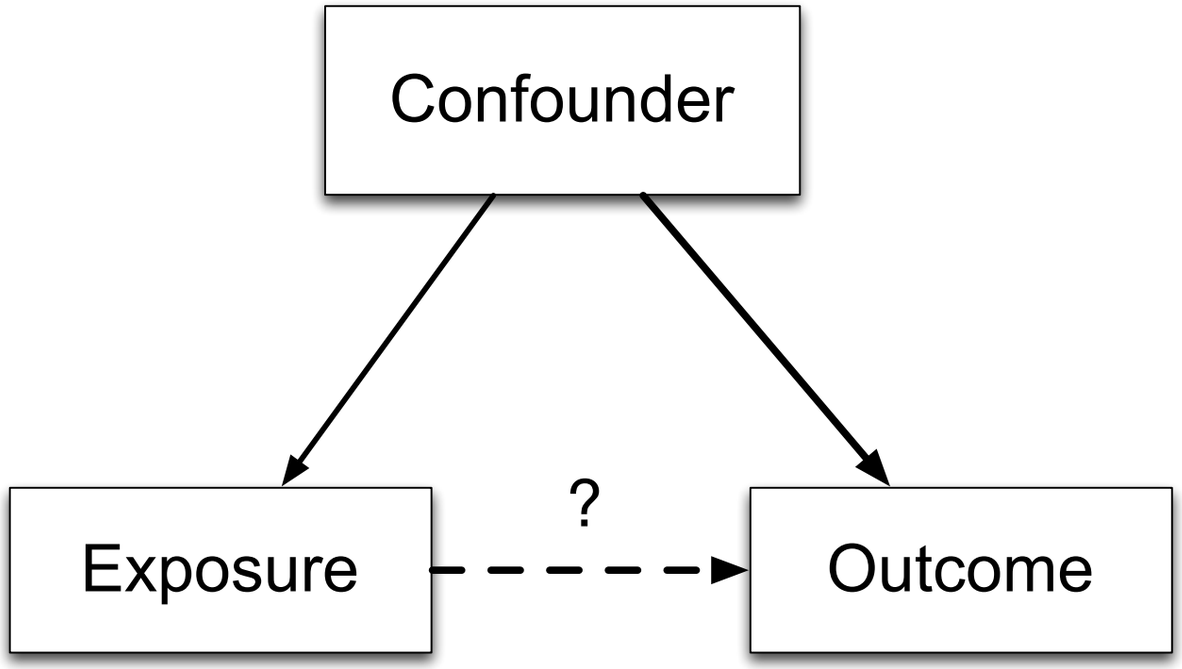
Confounding & Mediation

January 31, 2017
Biostatistics 208

Confounding

- *Confounder*: a common cause of both exposure (treatment) and outcome, or a surrogate for a common cause (e.g., age)
- Estimation of causal effects of an exposure/treatment from observational data must account for possible confounding to avoid bias
- Possible analytic strategies include
  - restriction, stratification, matching, propensity scores, inverse probability weighting, marginal structural models, targeted maximum likelihood, …
- Our focus: *adjust for confounding effects using regression models*
  - how to adjust adequately
  - when it will work

Confounding DAG



Potential outcomes and causal effects

- Suppose we could observe outcomes for each individual under exposure *and* no exposure, holding everything else constant

- For an outcome y & exposure e, the potential (counterfactual) outcomes are

$$y(e=1), \quad y(e=0)$$

- Within-person differences between potential outcomes are what we mean by individual *causal effects*

  - by definition, free of confounding influences

  - Problem: in general, only one of the potential outcomes is observable on an individual

- Overall causal effect can be summarized by a marginal average causal effect (ACE):

$$\mathrm{E}\left[y(e=1)\right] - \mathrm{E}\left[y(e=0)\right]$$

- Two common approaches to estimating ACE: randomization, and statistical modeling (e.g. regression, propensity scores, etc.)

Causal effects, RCTs, and linear regression modeling

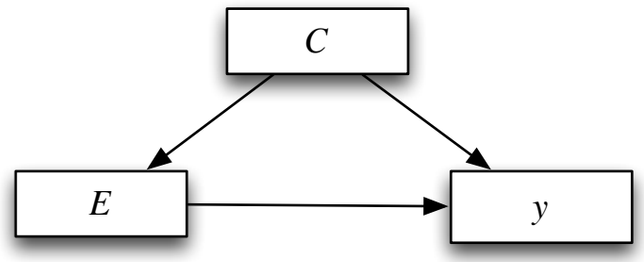
- RCTs use randomization to break the confounder → exposure link, so treatment is unconfounded
  - treatment groups are *exchangeable*, at least in large trials
  - causal effect estimated by comparing average outcomes between treatment groups
- Exchangeability in observational study data potentially achieved by accurately *modeling* all confounding effects
  - *within modeled strata*, exposed, unexposed exchangeable
- Adjusted estimate unbiased for causal effect if all necessary confounders (i.e., *minimum sufficient adjustment set*, or *MSAS*) included and *accurately modeled*

Why adjustment works to remove confounding

- A simple example based on simulated data
  - continuous exposure E
  - binary confounder C, correlated with E
  - continuous outcome $y = \beta_0 + \beta_1 E + \beta_2 C + \varepsilon$
  - in this example, $\beta_2 > 0$ and E, C are positively correlated
- Because E and C are correlated, model must include C to estimate β_1 – the causal effect of exposure – without bias

Simulation code

- Example simulation in Stata - follows the hypothesized DAG:
  - generate a binary confounder C
  - generate a continuous exposure E, correlated with C
  - generate a continuous outcome $y = \beta_0 + \beta_1 E + \beta_2 C + \varepsilon$
  - assume $\beta_2 > 0$ and E, C are positively correlated




```
program define regsim
drop _all
set obs 50
* generate a binary confounder with equal numbers in each group
gen confounder = 0
replace confounder = 1 if _n<=25
* generate exposure
gen exposure = 62 + 7*confounder +  rnormal(0,4)
* generate an error variable for the outcome (mean 0, SD 30)
gen ro = rnormal(0,30)
* generate the outcome as a function of both exposure & confounder
gen outcome = -25 + exposure*2.5 + confounder*30 + ro
end
```

Why adjustment works to remove confounding

simulated data example based on the set-up in previous slide (assumed: $\beta_1 = 2.5$)

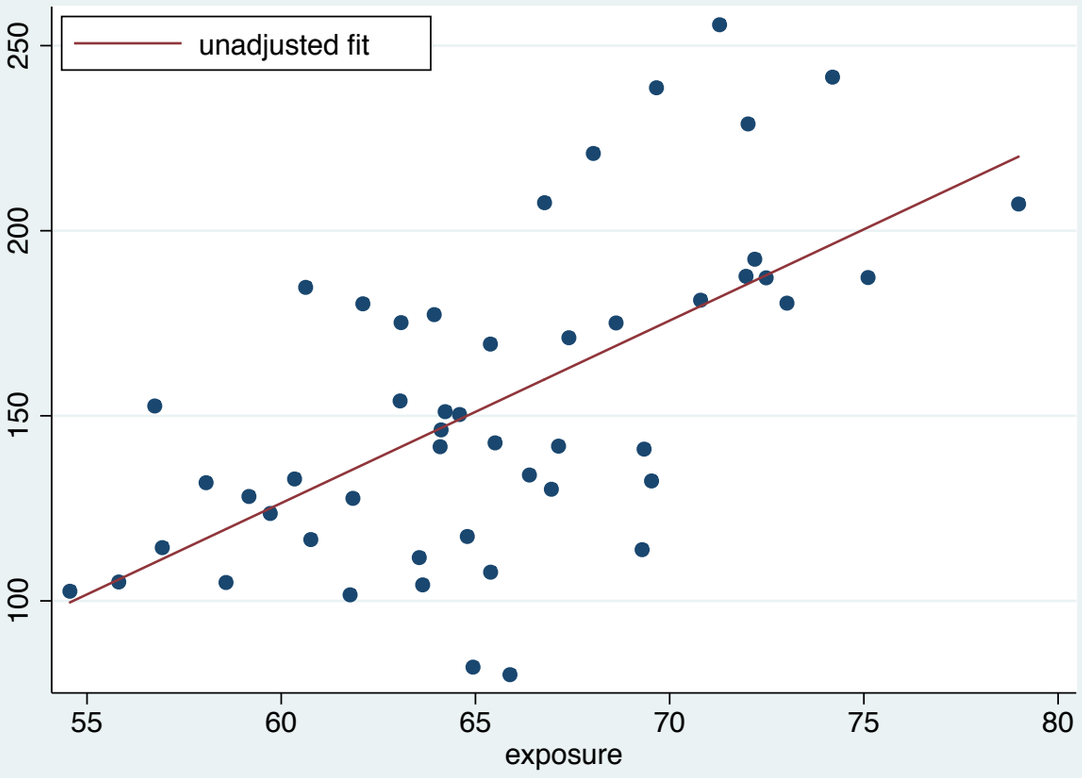
```
. tab confounder, sum(exposure)

           |        Summary of exposure
confounder |        Mean   Std. Dev.       Freq.
-----------+------------------------------------
         0 |   62.316344   3.7140415          25
         1 |   68.662241   4.9223134          25
-----------+------------------------------------
     Total |   65.489293   5.3755658          50
```

```
. tab confounder, sum(outcome)

           |        Summary of outcome
confounder |        Mean   Std. Dev.       Freq.
-----------+------------------------------------
         0 |    127.6001   27.565811          25
         1 |   179.34827   39.539979          25
-----------+------------------------------------
     Total |   153.47418   42.673949          50
```

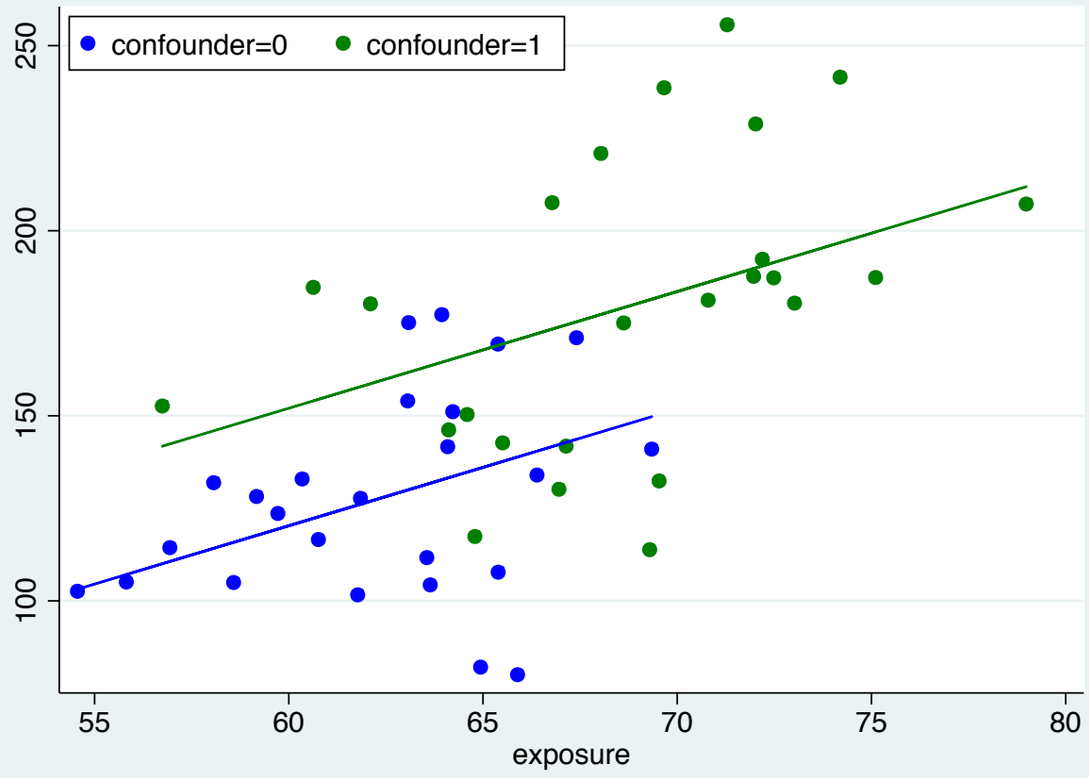
```
. corr exposure outcome
(obs=50)

             | exposure  outcome
-------------+------------------
    exposure |   1.0000
     outcome |   0.6211   1.0000
```

Why adjustment works to remove confounding

Unadjusted



Adjusted



Model : $\mathrm{E}[\, y \mid E, C] = \beta_0 + \beta_1 E + \beta_2 C$

Why adjustment works to remove confounding

- Unadjusted effect of E partly reflects increasing prevalence of C with increasing exposure
- Adjustment fixes the problem by modeling effect of C
- Model allows us to estimate effect of E within strata with $C = 0$ and $C = 1$
- No confounding by C within strata, because C does not vary

Unadjusted model

```
. regress outcome exposure

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  1,    48) =   30.15
       Model |  34425.9351     1  34425.9351            Prob > F      =  0.0000
    Residual |  54806.2966    48  1141.79785            R-squared     =  0.3858
-------------+------------------------------           Adj R-squared =  0.3730
       Total |  89232.2317    49  1821.06595            Root MSE      =   33.79

------------------------------------------------------------------------------
     outcome |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    exposure |    4.930837   .8979918     5.49   0.000     3.125303    6.736371
       _cons |   -169.4428   59.00268    -2.87   0.006    -288.0757      -50.81
------------------------------------------------------------------------------
```

Interpretation: without modeling the effect of the confounder, the estimated effect of exposure partly reflects the effect of the confounder. Here, confounding is bad enough that the 95% CI for the exposure effect excludes the true value of $\beta_1 = 2.5$.

Model adjusting for the confounder

```
. regress outcome exposure i.confounder

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  2,    47) =   21.41
       Model |  42543.0681     2  21271.5341            Prob > F      =  0.0000
    Residual |  46689.1635    47  993.386458            R-squared     =  0.4768
-------------+------------------------------           Adj R-squared =  0.4545
       Total |  89232.2317    49  1821.06595            Root MSE      =  31.518

------------------------------------------------------------------------------
     outcome |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    exposure |    3.152571   1.043347     3.02   0.004     1.053628    5.251514
1.confounder |    31.74228   11.10442     2.86   0.006     9.403068     54.0815
       _cons |   -68.85659   65.32243    -1.05   0.297    -200.2684    62.55518
```

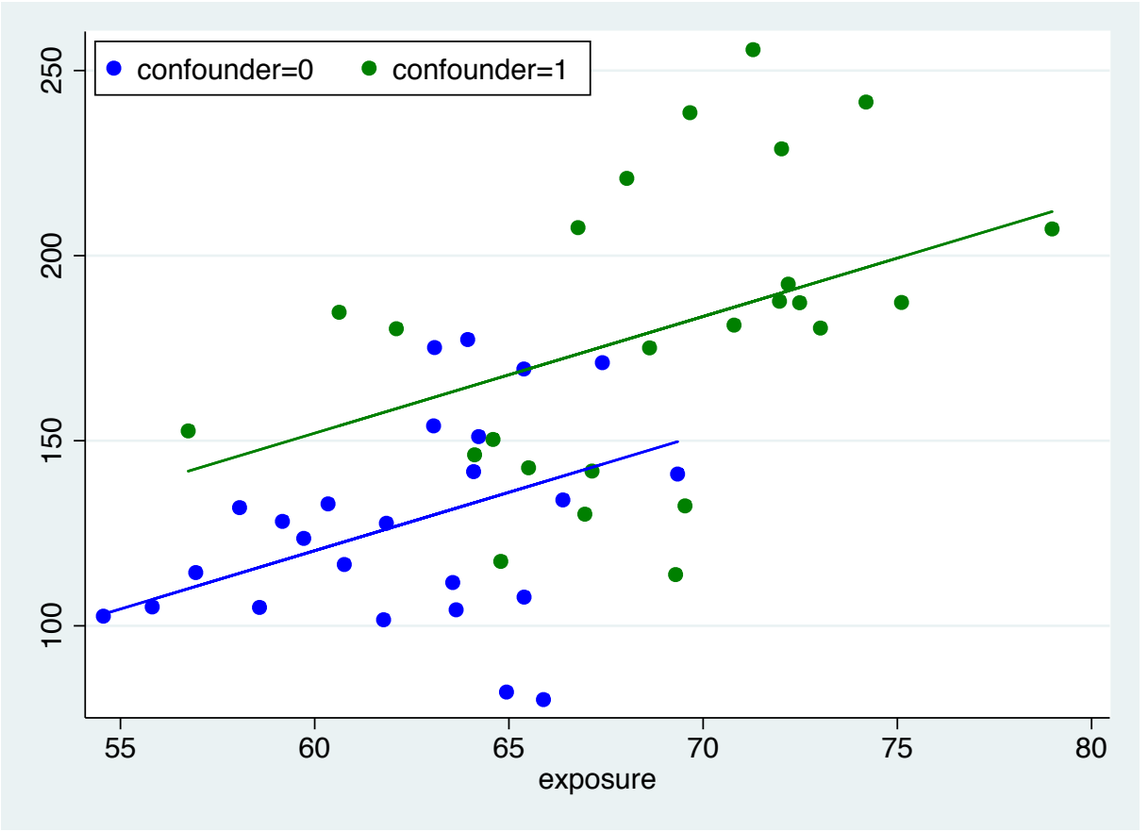
Interpretation: after we model the effect of the confounder, the estimated effect of exposure on the outcome is unbiased (and attenuated), and the 95% CI includes the true value of 2.5.

Confounding: interpretation of results

- Unadjusted estimate for exposure
  - estimates an observable trend in whole population
  - usually lacks a causal interpretation
- Adjusted estimate
  - unbiased for causal effect (provided confounding is adequately controlled for and appropriate modeled)
  - adjustment can impact precision of estimates

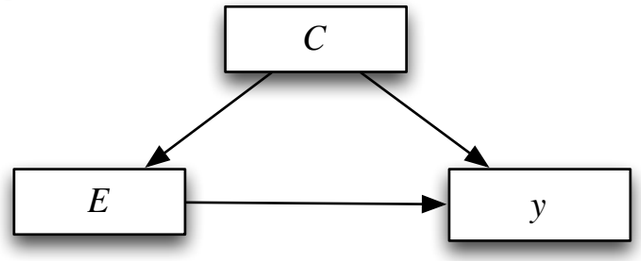
Confounding and interaction

- Regression lines for strata with $C = 0$ and $C = 1$:
  - slopes estimate exposure/outcome association within each stratum
  - assumed parallel: no *interaction* – E has same effect within both strata
- Next week's topic: how to check and relax this assumption



Range of confounding patterns

For an outcome y, binary exposure E & continuous confounder C:



Model for unadjusted effect: $\mathrm{E}[y|E] = \alpha_0 + \alpha_1 E$

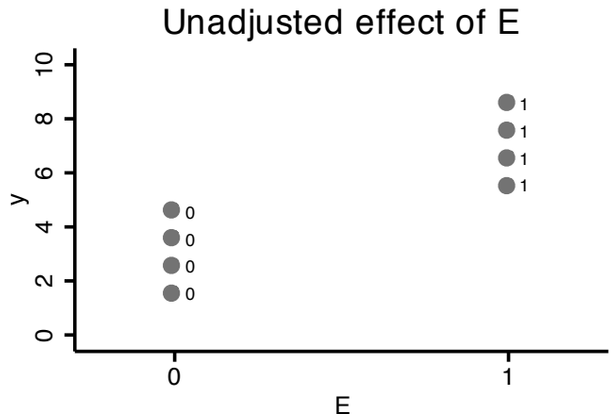
Unadjusted effect: $\mathrm{E}[y|E=1] - \mathrm{E}[y|E=0] = \alpha_1$

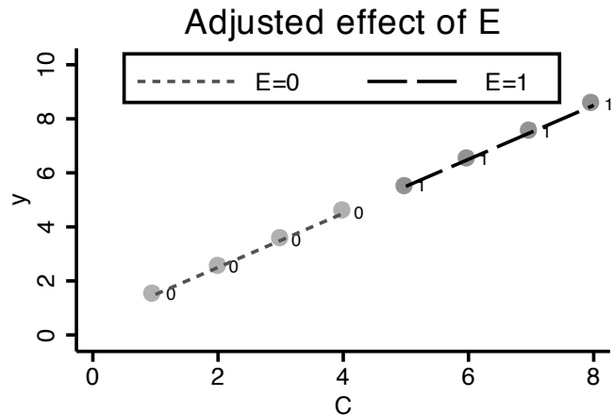
Model for adjusted effect: $\mathrm{E}[y|E, C] = \beta_0 + \beta_1 E + \beta_2 C$

Adjusted effect: $\mathrm{E}[y|E=1, C=c] - \mathrm{E}[y|E=0, C=c] = \beta_1$

Possible confounding if : $\alpha_1 \neq \beta_1$

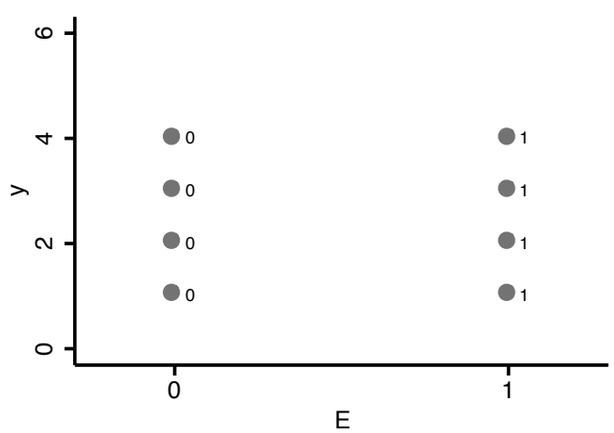
Examples :

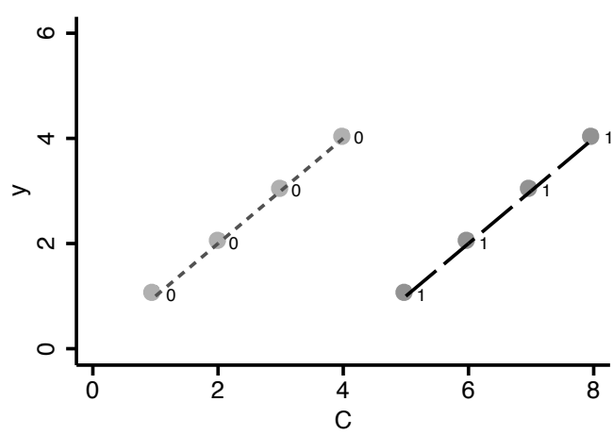




A: positive confounding

$$\alpha_1 > 0;\ \beta_1 = 0$$





B: negative confounding

$$\alpha_1 = 0;\ \beta_1 < 0$$

Positive confounding: two scenarios

Positive confounding may arise between predictors that are

1. *Positively correlated, with similar effects on outcome*
2. *Inversely correlated, with opposite effects on outcome*

- Adjusted coefficient for exposure is attenuated
- This is the typical pattern for confounding

Negative confounding: two scenarios

Negative confounding may arise between predictors that are

1. *Positively correlated, with opposite effects on outcome*

2. *Inversely correlated, with similar effects on outcome*

Negative confounding

- Example: needlestick injuries
  - Post-exposure prophylaxis (PEP) with AZT not marginally associated with protection against HIV transmission
    * PEP associated with severity of injury
    * severity increases transmission risk
  - PEP protection unmasked after controlling for severity

    Confounding by indication masks protective effect

- Example: BMI decreases with age in HERS cohort (-0.13 kg/m2 per year on average), but increases in both age & BMI are associated with higher systolic blood pressure (SBP)

Unadjusted association of BMI and SBP in HERS

```
. regress SBP BMI

      Source |       SS       df       MS              Number of obs =    2758
-------------+------------------------------           F(  1,  2756) =    9.91
       Model |  3573.21875     1  3573.21875            Prob > F      =  0.0017
    Residual |  993677.692  2756  360.550687           R-squared     =  0.0036
-------------+------------------------------           Adj R-squared =  0.0032
       Total |  997250.911  2757  361.715963           Root MSE      =  18.988

------------------------------------------------------------------------------
         SBP |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         BMI |   .2063226    .065539     3.148   0.002      .077812    .3348332
       _cons |   129.1723   1.907635    67.713   0.000      125.4318    132.9129
------------------------------------------------------------------------------
```

BMI and SBP adjusted for age

```
. regress SBP BMI age

      Source |       SS       df       MS                   Number of obs =    2758
-------------+------------------------------                F(  2,  2755) =   48.54
       Model |  33945.7704     2  16972.8852                 Prob > F      =  0.0000
    Residual |   963305.14  2755  349.657038                 R-squared     =  0.0340
-------------+------------------------------                Adj R-squared =  0.0333
       Total |  997250.911  2757  361.715963                 Root MSE      =  18.699

------------------------------------------------------------------------------
         SBP |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         BMI |   .3034786   .0653778     4.642   0.000     .1752842     .431673
         age |   .5056204   .0542507     9.320   0.000     .3992443    .6119965
       _cons |   92.69495   4.341354    21.352   0.000     84.18231    101.2076
------------------------------------------------------------------------------
```

Interpretation: effect of BMI on SBP is larger – 0.30 mmHg per kg/m^2 increase in BMI – after adjustment for age, compared to unadjusted effect of 0.21 mmHg per kg/m^2.

Summary: negative confounding

- Negative confounding can mask an independent association in unadjusted analysis
- Implications for predictor selection: univariate screening, forward selection procedures may miss some negatively confounded predictors
  - backwards deletion preferred

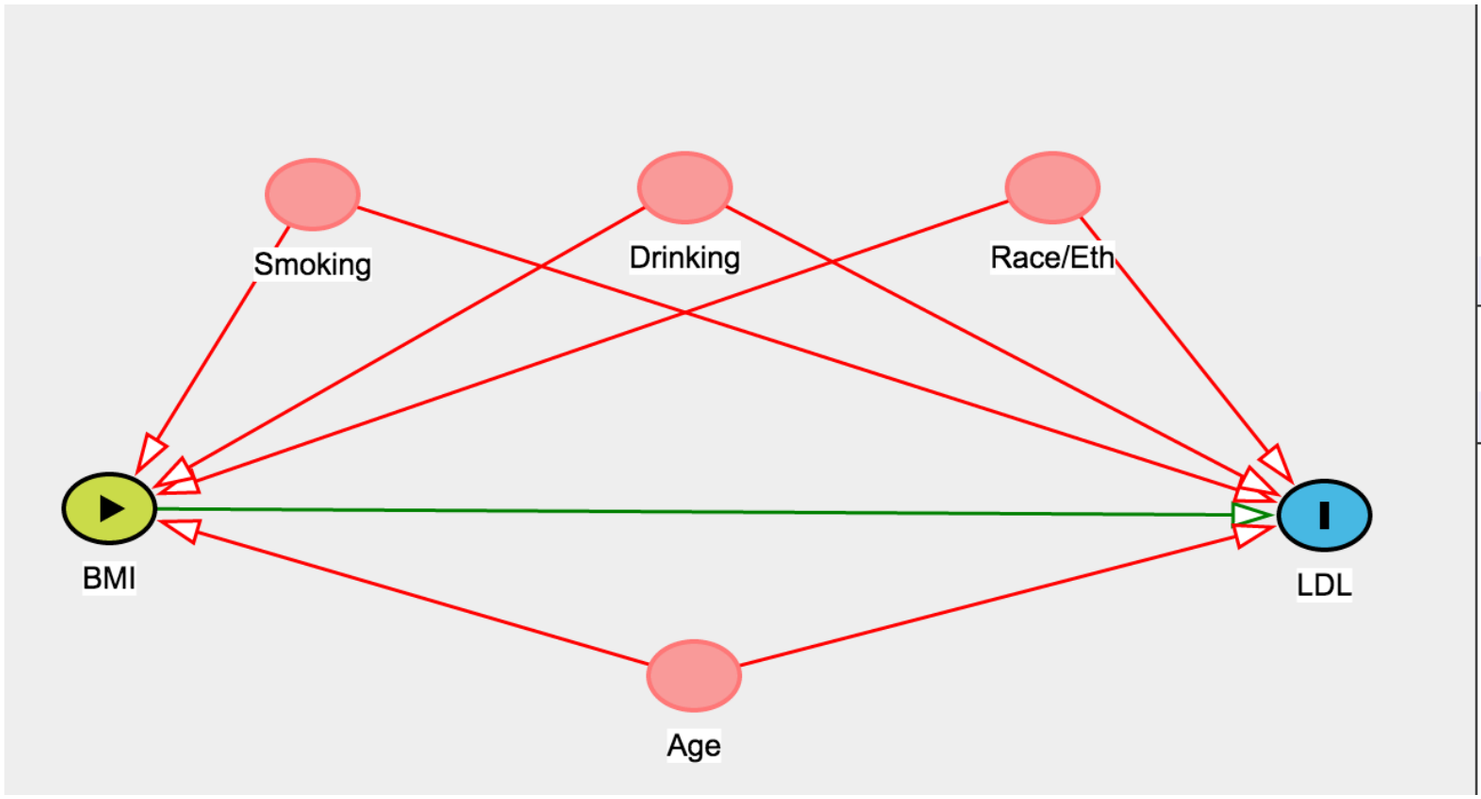
Interpreting adjusted regression effect estimates as causal

- Example: Association between BMI & LDL in HERS

- Unadjusted estimate:

```
. regress LDL BMI, noheader
------------------------------------------------------------------------------
         LDL |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         BMI |   .4151123   .1303648     3.18   0.001     .1594893    .6707353
       _cons |   133.1913     3.7939    35.11   0.000     125.7521    140.6305
------------------------------------------------------------------------------
```

- Goal: estimate adjusted association, accounting for possible confounders



Minimal sufficient adjustment sets for estimating the total effect of BMI on LDL:

- {Age, Drinking, Race/Eth, Smoking}

Adjustment for direct effect

Total and direct effects are equal in this model.

Testable implications

The model implies the following conditional independences:

- Age ⊥ Smoking
- Age ⊥ Race/Eth
- Age ⊥ Drinking
- Smoking ⊥ Race/Eth
- Smoking ⊥ Drinking
- Race/Eth ⊥ Drinking

Example: Association between BMI & LDL in HERS

- Adjusted estimate:

```
. regress LDL BMI age10 nonwhite smoking drinkany

      Source |       SS       df       MS              Number of obs =    2745
-------------+------------------------------           F(  5,  2739) =    5.97
       Model |  42279.1873     5  8455.83745            Prob > F      =  0.0000
    Residual |   3881903.3  2739  1417.27028            R-squared     =  0.0108
-------------+------------------------------           Adj R-squared =  0.0090
       Total |  3924182.49  2744  1430.09566            Root MSE      =  37.647

------------------------------------------------------------------------------
         LDL |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         BMI |    .3591038   .1341047     2.68   0.007      .0961472    .6220605
      age10 |   -1.897166   1.130776    -1.68   0.094     -4.114426    .3200947
   nonwhite |    5.219436   2.323673     2.25   0.025       .663108    9.775764
    smoking |    4.750738   2.210391     2.15   0.032      .4165362     9.08494
   drinkany |   -2.722354   1.498854    -1.82   0.069     -5.661352    .2166445
      _cons |    147.3153   9.256449    15.91   0.000       129.165    165.4656
------------------------------------------------------------------------------
```

Interpretation: each kg/m^2 increase in BMI is independently associated with an increase of 0.36 mg/dL in LDL (95% CI 0.10-0.62, p=0.007), adjusting for age, race/ethnicity, smoking, and alcohol use. (Analogous interpretations hold for the other effect estimates.)

Example: Association between BMI & LDL in HERS

- If the following conditions hold
  - All confounders are accounted for
  - The model is correct, and confounder distributions are comparable at diff. BMI levels
- Adjusted regression estimate based on the linear regression model provides an estimate of the average causal effect (it also estimates a conditional effect)
  - The mean change in LDL per unit increase in BMI, for fixed levels of (i.e. conditional on) the adjustment variables (output on previous slide)
  - The adjusted marginal mean change in LDL per unit increase in BMI:

```
. margins, dydx(BMI)
Average marginal effects                        Number of obs   =       2745
Model VCE    : OLS
Expression   : Linear prediction, predict()
dy/dx w.r.t. : BMI
------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         BMI |   .3591038   .1341047     2.68   0.007     .0961472    .6220605
------------------------------------------------------------------------------
```

  - marginal and conditional effects equal in linear model case

Confounding is difficult to rule out

- Were all important confounders adjusted for?
- Were they measured accurately?
- Were their effects modeled adequately – no un-modeled nonlinearities or omitted interactions?
- Was there enough "overlap" between exposed and unexposed to avoid extrapolation?

Getting the model right

- Non-linearities and interactions
  - default linear, no-interaction model may be too simple, allow residual confounding
  - departures hard to detect in small samples
- Avoiding extrapolation may require *restriction* of sample and scope of inference to subset of exposed with comparable unexposed controls (e.g., treatment effect in the treated)
- Difficulty of getting the model right a primary motivation for methods based on *propensity scores* (Biostat 215)

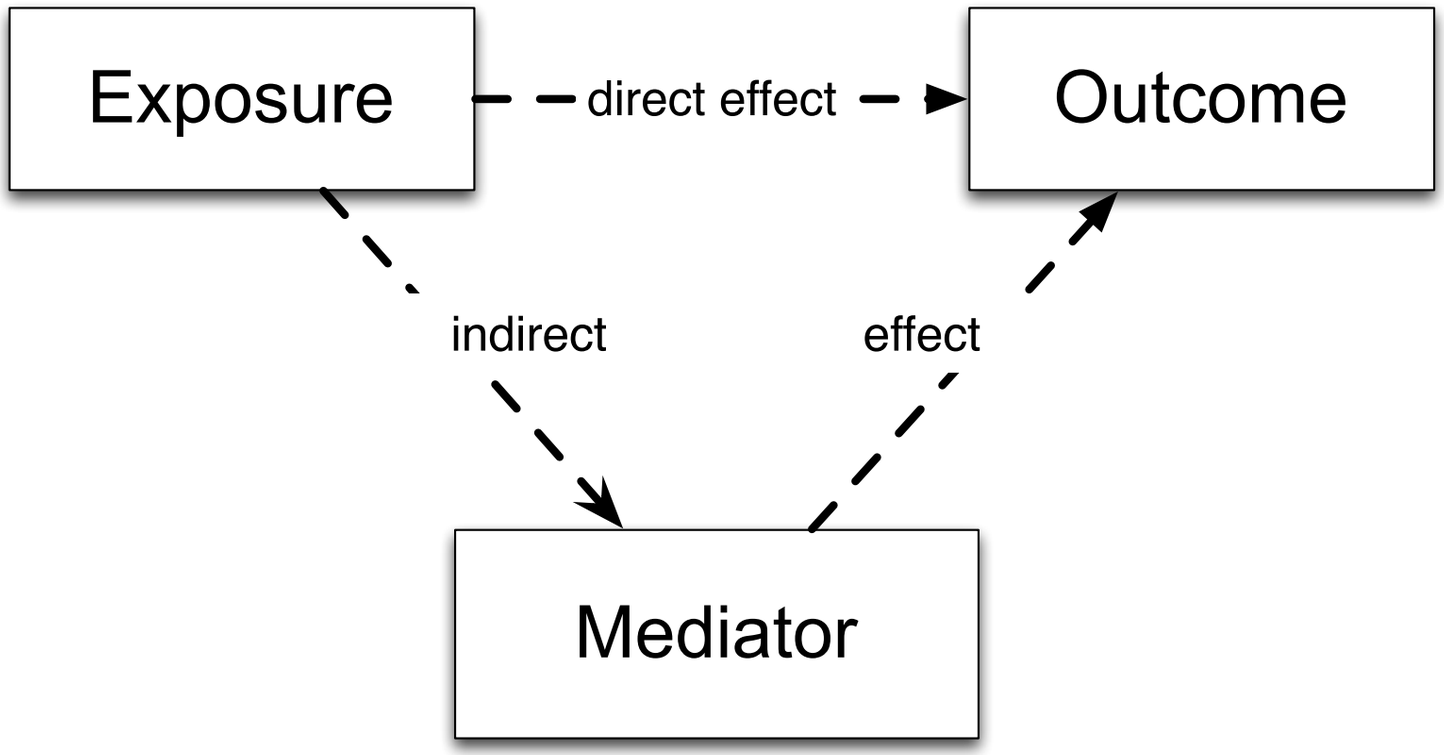
Confounding: summary

- Ideally, regression controls for confounding by jointly modeling effects of exposure and confounders
- Unbiased estimation of causal effects depends on getting the model right
- Bigger samples don't reduce un-modeled confounding, but do facilitate correct modeling of well-measured confounders
- Residual confounding is difficult to rule out

Mediation

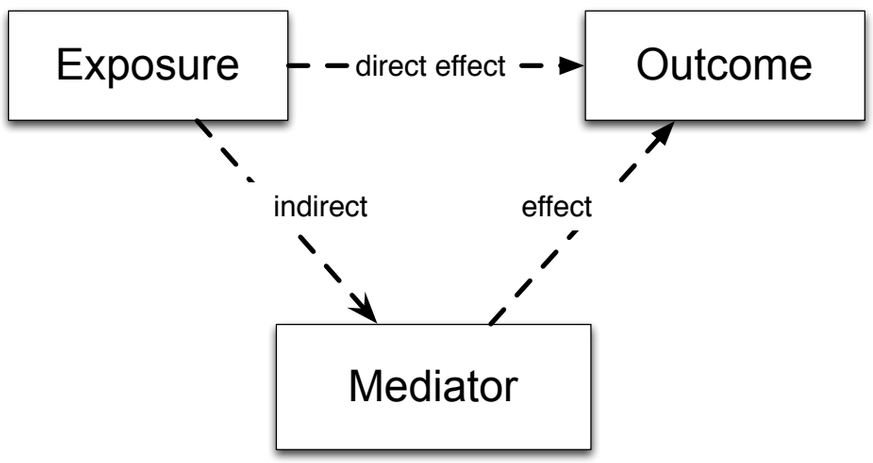
- *Confounder*: a common cause of both exposure and outcome (or a surrogate for a common cause)
- *Mediator*: on causal pathway from predictor to outcome
- In regression models, mediation and confounding behave alike
  - must be distinguished on substantive grounds
- Goal in mediation analysis: estimate *indirect* effect of exposure via mediator, *direct* effect via other pathways
- Examples of mediation
  - direct effect of cigarette smoking on coronary artery disease not mediated by blood pressure
  - mediation of the effect of treatment with bisphosphonates on fracture risk by changes in bone mineral density
  - mediation of effects of smoking on infant mortality by low birth weight
  - mediation of effects of assignment to a barrier contraceptive method (diaphragms) on HIV infection risk by condom use

Mediation DAG



Examining mediation

- Use a series of regression models to figure out whether:
  - exposure has indirect effect via mediator
    * exposure independently affects mediator
    * mediator affects outcome independently of exposure
  - exposure has a direct effect via other pathways
  - what proportion of overall effect of exposure is indirect



Examining mediation

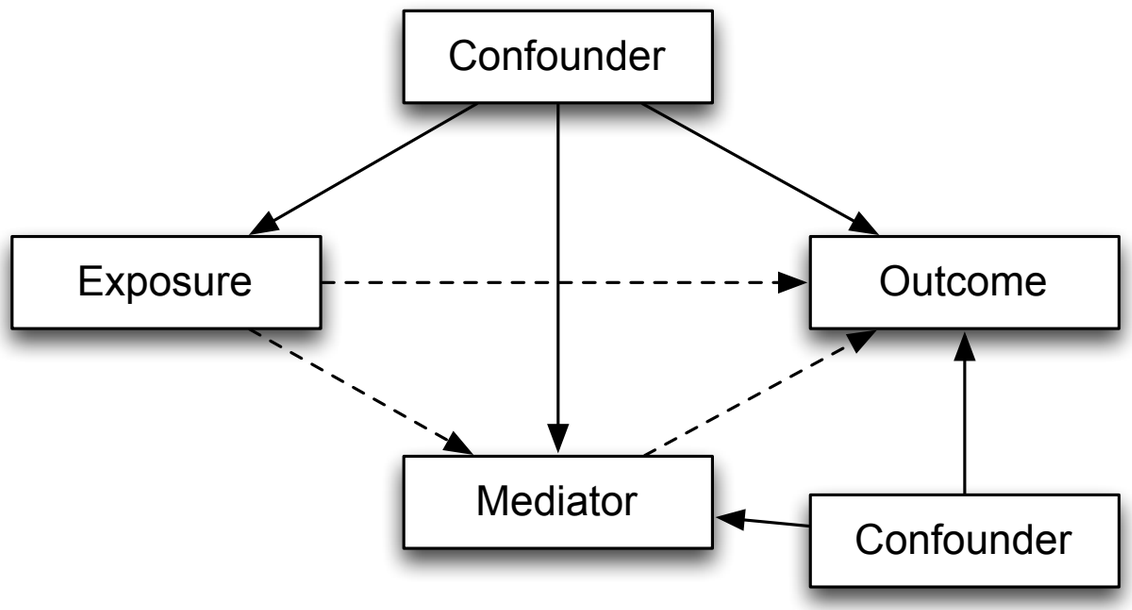
- Three regression models:

  1. *dependence of mediator on exposure and confounders*

  2. *dependence of outcome on exposure and confounders*

  3. *dependence of outcome on exposure, confounders, and mediator*

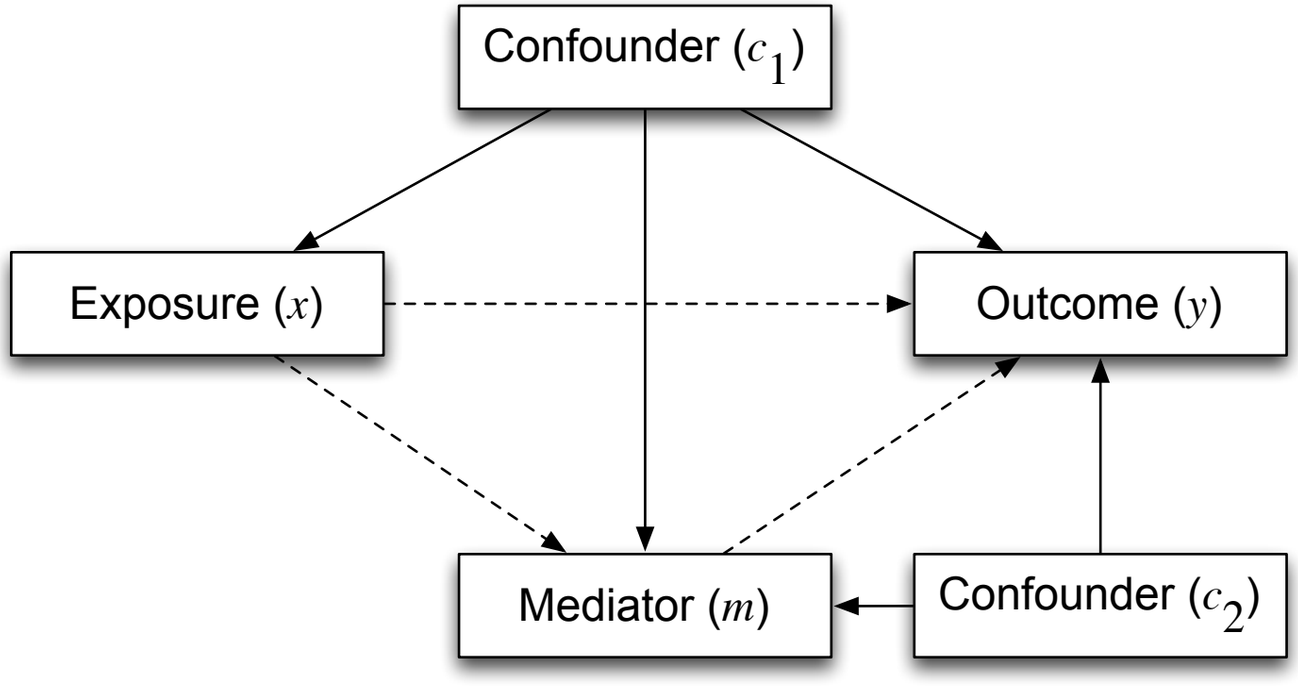
- *Note:* Models 1-3 should be estimated using the same set of observations (i.e., with complete – possibly imputed – data on outcome, exposure, mediator, and confounders)



Examining mediation

- Three models verifying mediation

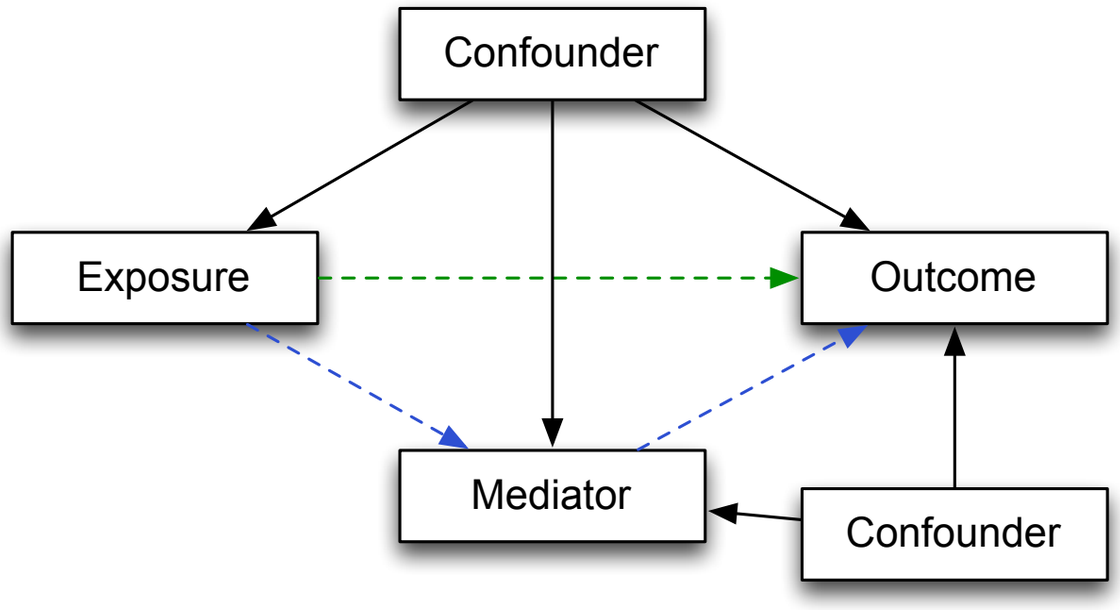
1. $\mathrm{E}\left[m|x, c_1\right] = \alpha_0 + \alpha_1 x + \alpha_2 c_1$
2. $\mathrm{E}\left[y|x, c_1\right] = \gamma_0 + \beta_{overall} x + \gamma_1 c_1$
3. $\mathrm{E}\left[y|x, m, c_1, c_2\right] = \beta_0 + \beta_{direct} x + \beta_2 m + \beta_3 c_1 + \beta_4 c_2$



Overall and direct effect estimates

- Model 2: *overall effect* (all dashed lines) of exposure on outcome, including effect via mediator

- Model 3: *direct effect* of exposure on outcome, via other pathways not involving mediator



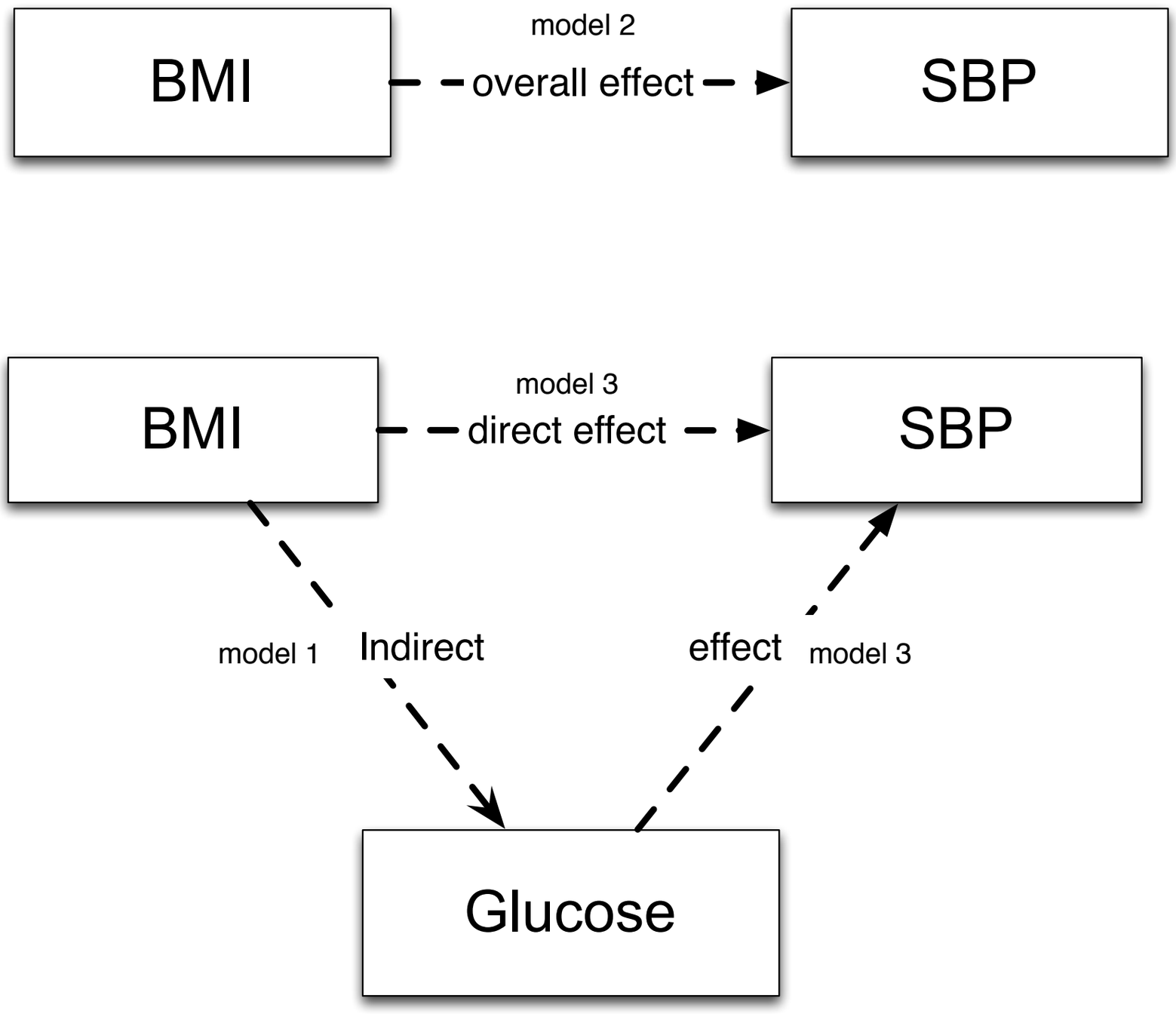
- *indirect effect* of exposure on outcome estimated from models 1 & 2 (i.e. estimates the degree to which exposure is mediated)

Indirect effect estimates

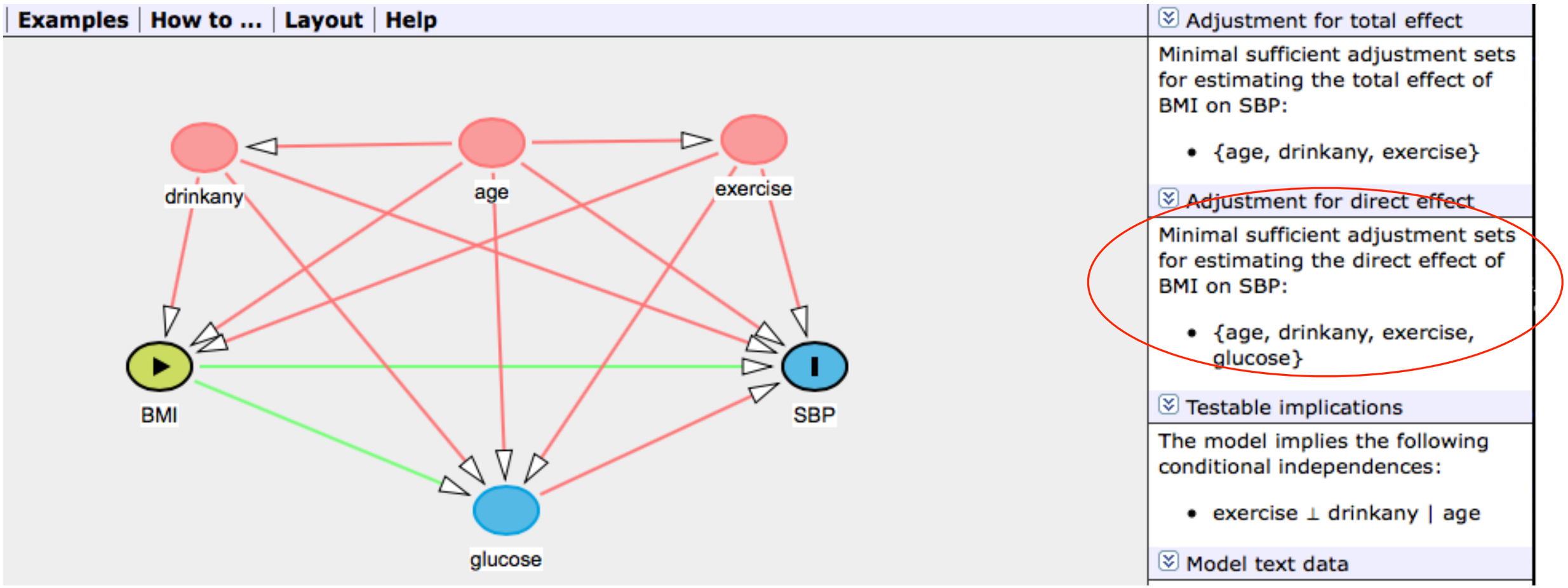
- Mediator, outcome both continuous (i.e. linear regression model applies): we can use either[1]
  - difference of exposure effects in Model 2 and Model 3 ($\beta_{overall} - \beta_{direct}$),
  - exposure effect in Model 1 × mediator effect in Model 3 ($\alpha_1 \times \beta_2$)
- Mediator and/or outcome is binary
  - use downloadable mediation package `medeff`
- For linear models, can also quantify mediation by *proportion explained* (*PE*), defined as $(\beta_{overall} - \beta_{direct})/\beta_{overall}$

1. Mackinnon DP et al. A Simulation Study of Mediated Effect Measures. *Multivariate Behav Res.* (1995)

Mediation of BMI → SBP by glucose



Mediation of BMI → SBP by glucose



Model 1: BMI predicts higher glucose

```
* model 1
. regress glucose BMI age exercise drinkany

      Source |       SS       df       MS
-------------+------------------------------
       Model |  321116.455     4  80279.1137
    Residual |  3426880.46  2751  1245.68537
-------------+------------------------------
       Total |  3747996.91  2755  1360.43445
Number of obs = 2756 F( 4, 2751) = 64.45 Prob>F = 0.0000 R-squared = 0.0857 Adj R-
squared = 0.0843 Root MSE = 35.294
------------------------------------------------------------------------------
     glucose |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         BMI |      1.6902   .1256441    13.45   0.000     1.443833    1.936566
         age |   -.2144159    .102671    -2.09   0.037    -.4157359   -.0130959
    exercise |   -2.774221   1.400946    -1.98   0.048    -5.521234   -.0272081
    drinkany |    -7.15434   1.385168    -5.16   0.000    -9.870414   -4.438266
       _cons |    82.01119    8.38539     9.78   0.000      65.5689    98.45349
------------------------------------------------------------------------------
* save for calculating indirect effect
. scalar link1 = _b[bmi]
```

Interpretation: each unit increase in BMI results in an increase of 1.7 mg/dL (95% CI 1.4-1.9, $p < 0.0005$) in average glucose, assuming the model includes an MSAS for this relationship

Model 2: overall effect of BMI on SBP

```
* model 2
. regress SBP BMI age exercise drinkany

      Source |       SS       df       MS              Number of obs =    2756
-------------+------------------------------           F(  4,  2751) =   30.40
       Model |  42202.7194     4  10550.6799            Prob > F      =  0.0000
    Residual |  954711.522  2751  347.041629            R-squared     =  0.0423
-------------+------------------------------           Adj R-squared =  0.0409
       Total |  996914.241  2755  361.856349            Root MSE      =  18.629

------------------------------------------------------------------------------
         SBP |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         BMI |    .245644   .0663176     3.70   0.000     .1156067    .3756813
         age |   .4933303   .0541919     9.10   0.000     .3870694    .5995913
    exercise |  -2.911396   .7394488    -3.94   0.000    -4.361327   -1.461465
    drinkany |  -2.022715   .7311205    -2.77   0.006    -3.456316   -.5891144
       _cons |   97.07514   4.425984    21.93   0.000     88.39656    105.7537
------------------------------------------------------------------------------
* save for calculating PE
. scalar b_overall = _b[BMI] // for use in calculating PE
```

Interpretation: the overall effect of a 1-unit increase in BMI is to increase average SBP by 0.25 mmHg (95% CI 0.12-0.38, p < 0.0005), again assuming an MSAS

Model 3: direct BMI effect via other pathways

```
* model 3
. regress SBP BMI age exercise drinkany glucose
      Source |       SS       df       MS
-------------+------------------------------
       Model |  50555.9327     5  10111.1865
    Residual |  946358.308  2750  344.130294
-------------+------------------------------
       Total |  996914.241  2755  361.856349
Number of obs = 2756 F( 5, 2750) = 29.38 Prob>F = 0.0000 R-squared = 0.0507 Adj R-
squared = 0.0490 Root MSE = 18.551
------------------------------------------------------------------------------
         SBP |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         BMI |   .1621961   .0681763     2.38   0.017     .0285141    .2958781
         age |   .5039164   .0540069     9.33   0.000     .3980183    .6098145
    exercise |  -2.774428   .7368653    -3.77   0.000    -4.219294   -1.329563
    drinkany |  -1.669494   .7315689    -2.28   0.023    -3.103974   -.2350138
     glucose |   .0493716   .1002102     4.93   0.000     .2972212     .690211
       _cons |   93.02612   4.483349    20.75   0.000     84.23505    101.8172
------------------------------------------------------------------------------
* save for calculating indirect effect
. scalar link2 = _b[glucose]
* save for calculating PE
. scalar b_direct = _b[BMI]
```

Interpretation: the direct effect of a 1-unit increase in BMI is to increase average SBP by 0.16 mmHg (95% CI 0.03-0.30, p = 0.017), again assuming model includes an MSAS.

Estimates of indirect effect

```
* two ways of calculating indirect effect
. display link1*link2
.08344789
. display b_overall-b_direct
.08344789
* PE
. display round((b_overall - b_direct) / b_overall * 100)
34
```

Interpretation: the indirect effect of a 1-unit increase in BMI via glucose is to increase average SBP by 0.08 mmHg, assuming an MSAS (note: this only works when mediator, outcome continuous). 34% of the overall effect of BMI on SBP is mediated by its effect on glucose levels

Mediation of BMI by glucose levels

- BMI independently predicts glucose (Model 1); glucose independently predicts SBP (Model 3)
- BMI effects:
  - Overall: 0.25 mmHg per kg/m2 (Model 2)
  - Direct: 0.16 mmHg per kg/m2 (Model 3)
  - Indirect: 0.08 mmHg per kg/m2 (Models 1 and 3)
- PE: glucose explains (.25-.16)/.25*100 = 34% of BMI

Why test for mediation?

- We don't test for confounding, but need to validate indirect pathway
- Comparing overall and direct effect estimates from Models 2 and 3 is hard (they are correlated), can be unreliable
- Testing both links in indirect pathway more reliable, powerful:
  - i.e., test exposure effect in Model 1, mediator effect in Model 3
  - both must be statistically significant (avoiding multiple comparison problems)

Assessing evidence for indirect pathway

```
. * Model 1
. regress glucose BMI age exercise drinkany
------------------------------------------------------------------------------
     glucose |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         BMI |      1.6902   .1256441    13.45   0.000     1.443833    1.936566
   ...
------------------------------------------------------------------------------
. * Model 3
 . regress SBP BMI age exercise drinkany glucose
------------------------------------------------------------------------------
         SBP |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   glucose10 |   .4937161   .1002102     4.93   0.000     .2972212     .690211
   ...
------------------------------------------------------------------------------
```

Interpretation: both links in the indirect pathway are strongly supported.
Note: no multiple testing issue because both links must be significant

Bootstrap Confidence Interval for PE

- Apply bootstrap re-sampling using a simple Stata program to estimate PE
- Use ≥500 bootstrap samples and bias-corrected 95% CI

```
. capture program drop mediate
. program define mediate, rclass
  1. syntax varlist, outcome(varlist) mediator(varlist fv) covars(varlist fv)
  2. reg `outcome' `varlist' `covars'
  3. scalar b_overall = _b[`varlist']
  4. reg `outcome' `varlist' `covars' `mediator'
  5. scalar b_direct = _b[`varlist']
  6. return scalar pe = (b_overall - b_direct) / b_overall * 100
  7. end
```

```
. bootstrap pe=r(pe), reps(1000) notable nodots: mediate BMI, outcome(SBP)
mediator(glucose) covars(age exercise drinkany)

Bootstrap results                                   Number of obs      =      2756
                                                    Replications       =      1000

      command:  mediate BMI, outcome(SBP) mediator(glucose) covars(age exercise drinkany)
           pe:  r(pe)
```

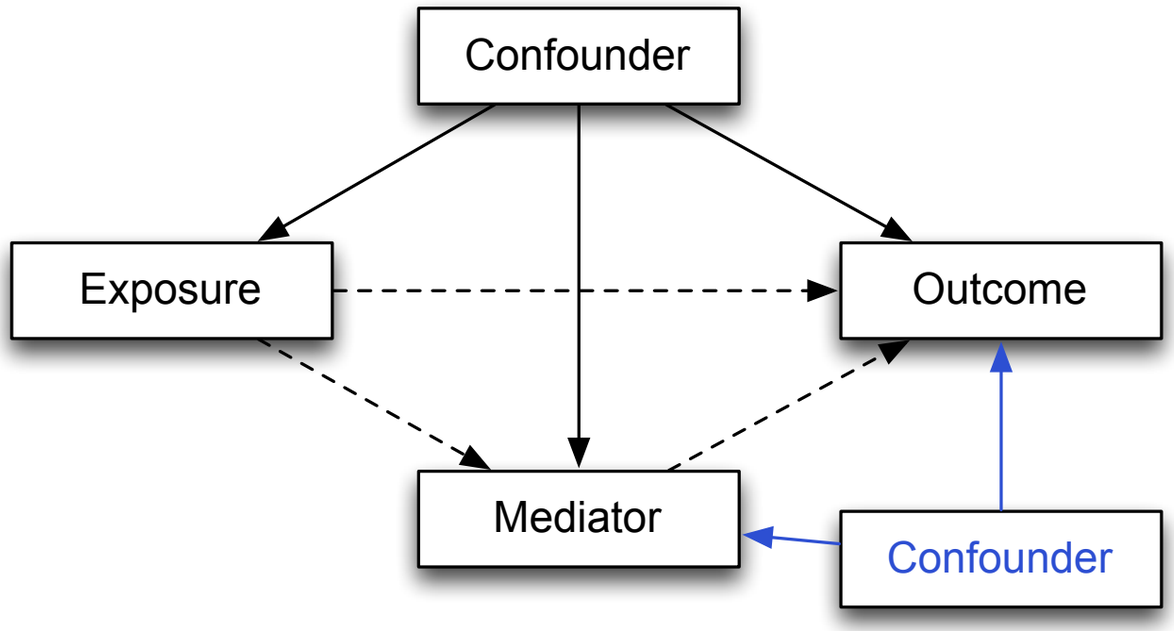
Bootstrap Confidence Interval for PE

- Apply bootstrap re-sampling using a simple Stata program to produce the estimated PE
- Use ≥500 bootstrap samples and bias-corrected 95% CI

```
. * get bias corrected bootstrap CI using the estat command
. estat bootstrap, all
Bootstrap results                                 Number of obs      =      2756
                                                  Replications       =      1000

command:  mediate BMI, outcome(SBP) mediator(glucose) covars(age exercise drinkany)
     pe:  r(pe)

------------------------------------------------------------------------------
             |    Observed                 Bootstrap
             |       Coef.       Bias    Std. Err.  [95% Conf. Interval]
-------------+----------------------------------------------------------------
          pe |   33.971067   2.727515   20.677067   -6.555239   74.49737   (N)
             |                                       17.10639   75.33008   (P)
             |                                       17.31348    76.9211  (BC)
------------------------------------------------------------------------------
(N)    normal confidence interval
(P)    percentile confidence interval
(BC)   bias-corrected confidence interval
```

Confounding of mediator/outcome relationship

- Failure to control for confounding of mediator - outcome relationship can lead to biased estimates of mediation effects[2]
  - In the presence of such confounding, the mediator is a *collider* on front door path from exposure to outcome

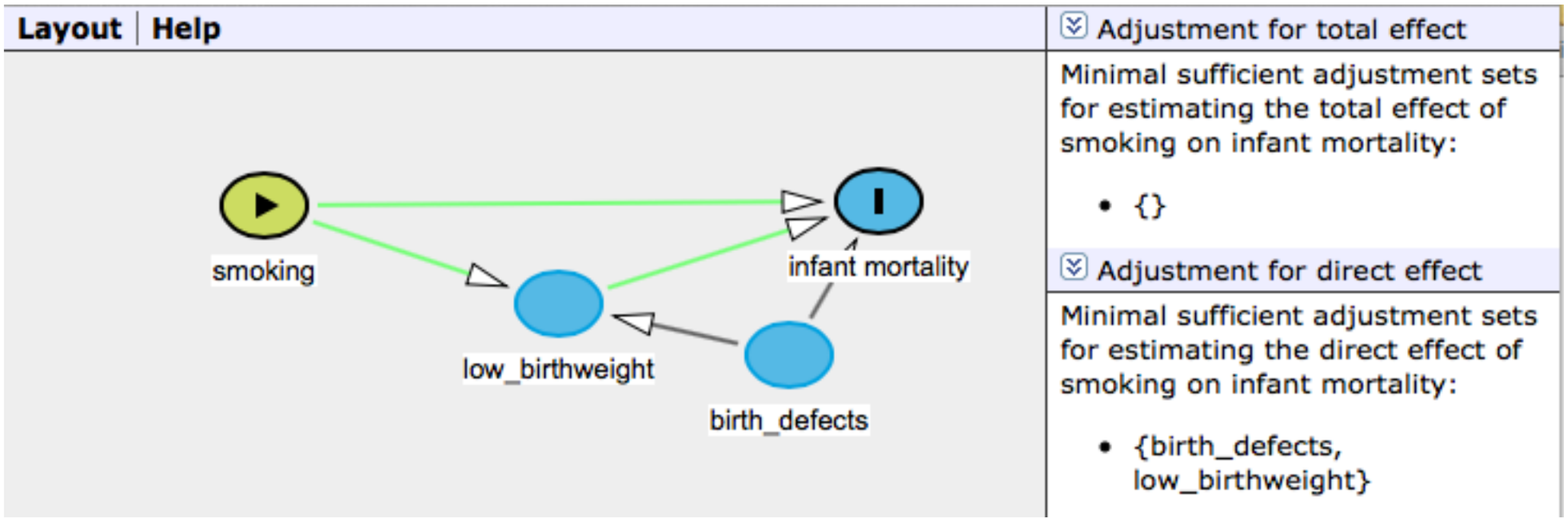


2. Cole SR, Hernán MA. Fallibility in estimating direct effects, Int J Epid, 2002;31:163-5

Confounding of mediator/outcome relationship

- Example: low birth weight as mediator of effects of smoking on infant mortality[3]

The "paradox" refers to the observation that low birthweight babies with mothers who smoke have been observed to do better than those with low birthweight & non-smoking mothers

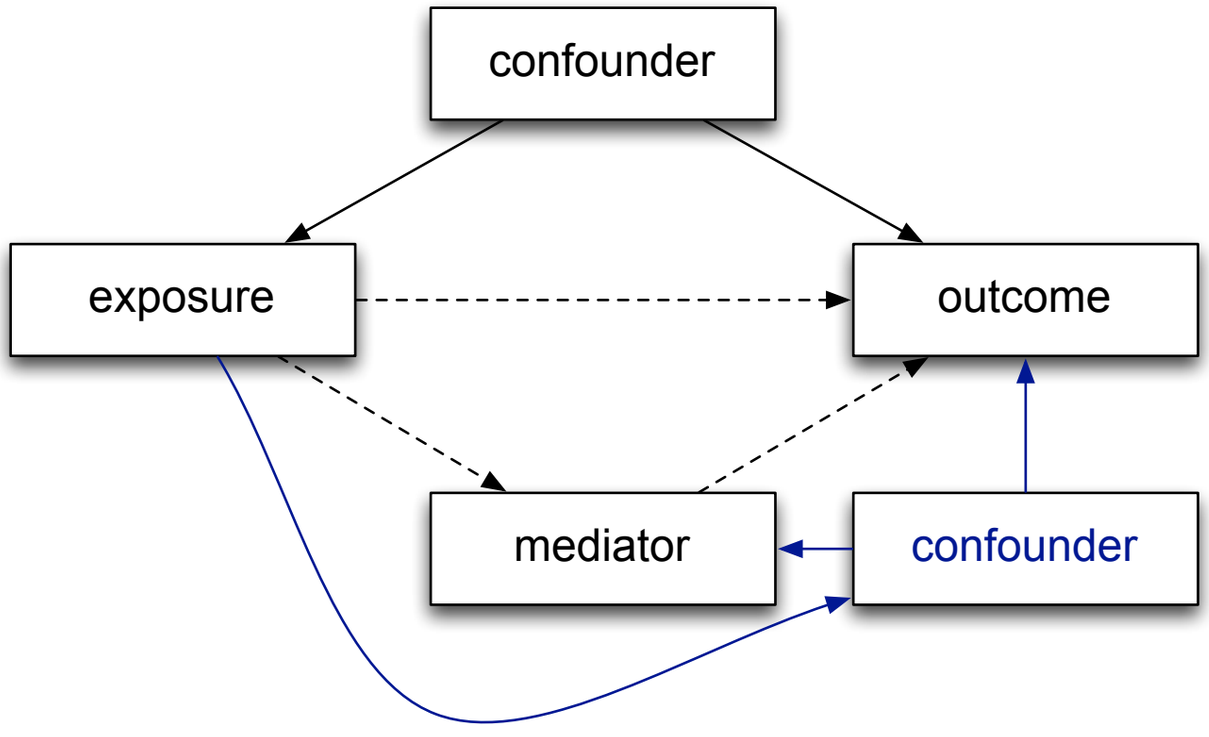


3. Whitcomb BW, Schisterman EF, Perkins NJ, Platt RW. Quantification of collider-stratification bias and the birthweight paradox. *Paediatr Perinat Epidemiol*. 2009;23:394-402.

Confounder/mediators of mediator/outcome relationship

- If a confounder of mediator/outcome relationship is also a mediator of exposure effects, *regression adjustment fails* to yield unbiased estimates (special methods using inverse weighting needed) [4]

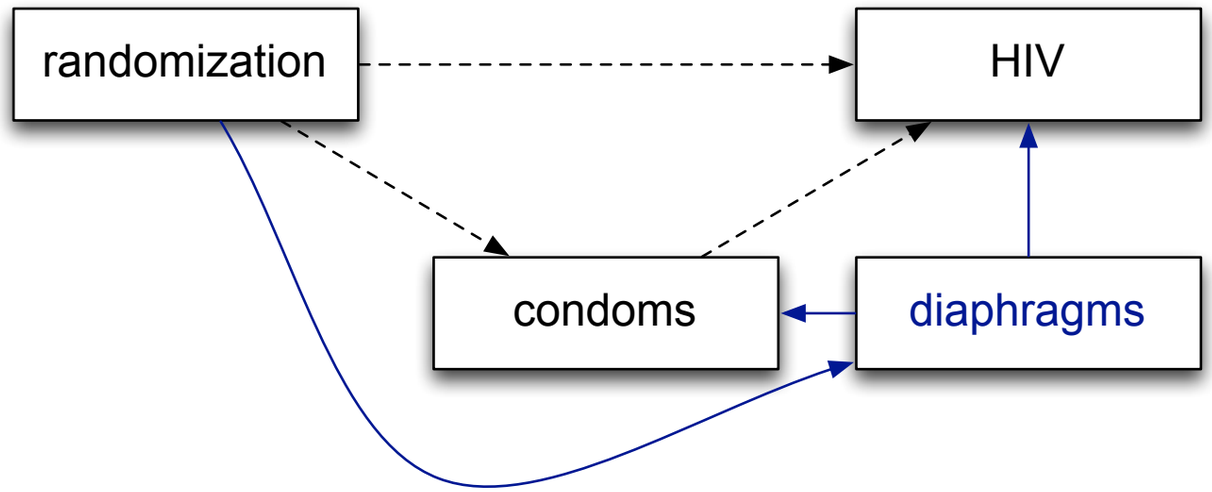
  - In this case, regression adjustment for the confounder/mediator removes some of the effect of exposure



4. VanderWeele TJ. Marginal structural models for direct and indirect effects, *Epidemiol*, 2009;20(1):18-26.

Confounder/mediators of mediator/outcome relationship

- MIRA study of diaphragms to prevent HIV infection in African women
  - randomization to either diaphragms + condoms / condoms only



- Comparison of outcomes by randomized assignment to the intervention (diaphragms) provides a valid estimate causal effect (none was observed)

- Possible between-group differences in post-randomization condom use [5]
  - *Direct effect* of randomization *not mediated through condoms* of interest
  - Diaphragm use as a possible confounder-mediator

5. Rosenblum M, Jewell NP, van der Laan M, Shiboski S, van der Straten A, Padian N. Analyzing Direct Effects in Randomized Trials with Secondary Interventions: An Application to HIV Prevention Trials. *J R Stat Soc Ser A Stat Soc*. 2009;17:443-465.

Additional issues with mediation analyses

- Potentially stronger conclusions about mediation possible with longitudinal data
  - Exposure affects mediator, mediator subsequently affects outcome. Example: Alendronate → BMD → fracture risk
  - However, effects may be effectively simultaneous: illness, frailty → depression → decreased cognitive function
- More complicated methods needed if
  - exposure and mediator interact; Stata package `medeff` needed
  - time-dependent covariates acting as confounder-mediators: *marginal structural models*
- Estimates of proportion explained (*PE*)
  - not really a proportion, complicated to estimate for nonlinear models (logistic, proportional hazards)

Mediation Summary

- Mediation distinguished from confounding on substantive grounds
- 3 models used to estimate overall, direct, indirect effects (and PE)
- Models must include confounders of mediator/outcome
- Estimating indirect effects often requires special methods