Predictor Selection

February 18, 2014

Biostatistics 208

Predictor Selection

- Define three primary goals in selecting predictors for regression models
- Application of DAGs in selecting confounders
- Applications of model selection algorithms

Predictor Selection

- Given a (potentially large) number of available predictors, which ones should be included in a regression model?
- Depends on "inferential goal":
  1. predict future outcomes
  2. estimate causal effect of a primary predictor
  3. identify important risk factors for an outcome
- Controversial, with positions dogmatically held

Goal 1: Prediction

- Includes diagnosis and prognostic risk stratification
- Often used in making decisions at the level of the individual
- Causal relationships useful, but not the primary focus
- Only strong predictors useful: (e.g. OR for binary diagnostic variable with sensitivity and specificity of 90% is 81)
- Model validation based on assessment of prediction error (PE)

Prediction Error

- Prediction error measures how well a model predicts a new outcome – i.e., for new observations not used in fitting model
- Two aspects of PE:
  1. *discrimination*: how well does model distinguish outcomes between individuals (e.g. cases from controls, high-risk patients from low, early from late events)?
  2. *calibration*: how accurately does model estimate outcome values, failure rates?
- Both discrimination and calibration are important

Prediction

- Bias-variance tradeoff and over-fitting
- Parameter estimates, predicted values:
  - biased when important predictors omitted from model
  - noisier when unimportant predictors are included
- Too many predictors results in fitting idiosyncrasies in the current data – i.e., over-fitting
- However, some methods for eliminating weak predictors can also cause trouble

Prediction: "traditional" split-sample approach

- *Example*: prediction of bone min. density (BMD) in subsample ($n \approx 2600$) from the Study of Osteoporotic Fractures (SOF)
- Model developed in 1/2 of the sample; validated in the remaining 1/2
- Model based on baseline predictors, selected using stepwise selection

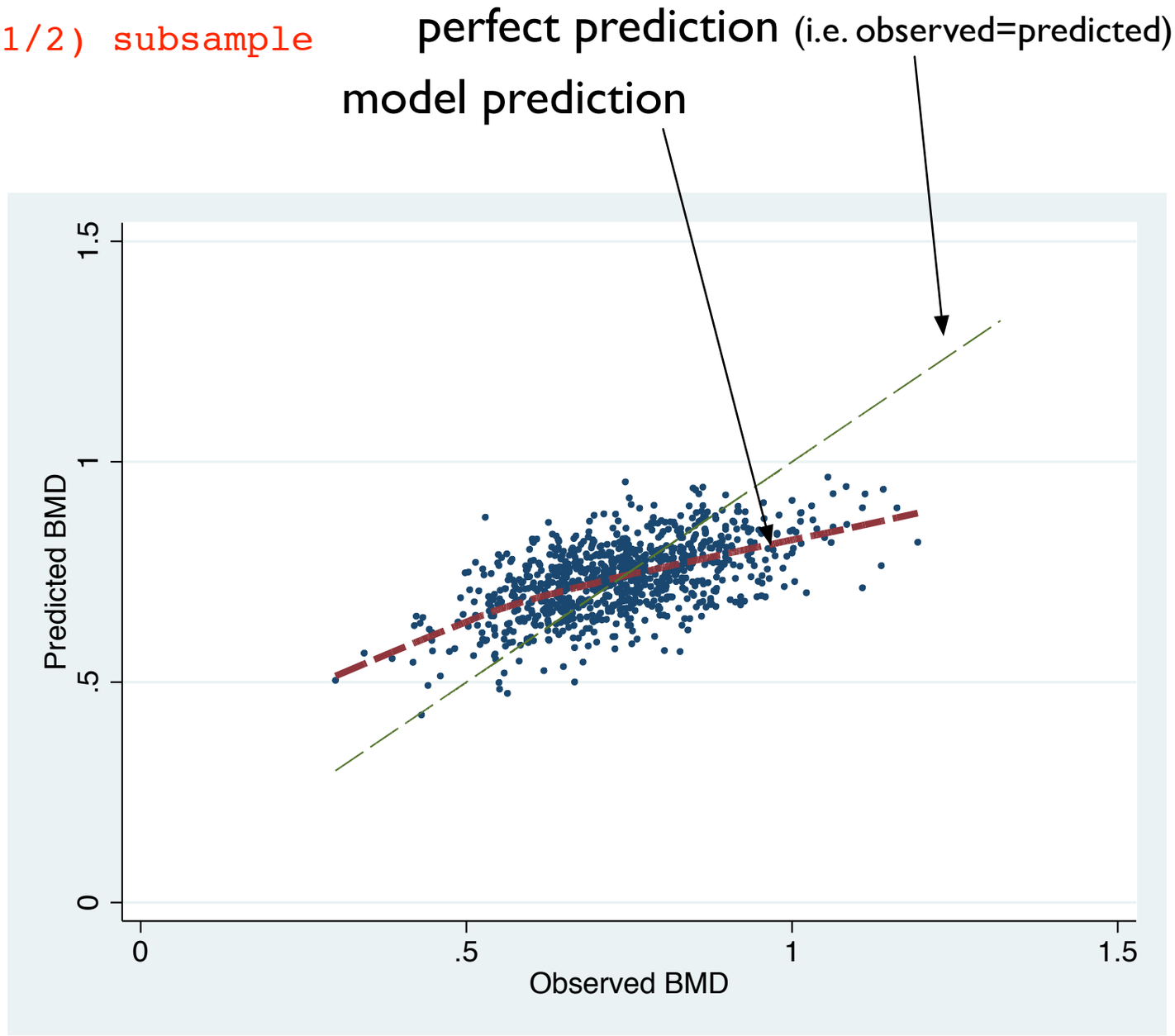
```
. * randomly order observations
. gen r = uniform()
. sort r
. * split sample into two groups consisting of 1/2, 1/2 of observations,
. * generate group indicator
. gen group = 1 if _n <= _N*0.5
(945 missing values generated)
. replace group=2 if group!=1
(945 real changes made)
. * results
. tabulate group, summarize(bmd)
            |           Summary of BMD
      group |        Mean   Std. Dev.       Freq.
------------+------------------------------------
          1 |   .73139813   .13575511        1389
          2 |   .72966763   .13541578        1390
------------+------------------------------------
      Total |   .73053257   .13556385        2779
```

Example prediction model from SOF

```
. * use backwards stepwise selection to select model, restricted to 1/2 of sample

. xi: stepwise, pr(.1): regress bmd eeu age lweight i.estrogen calsupp diuretic etid momhip usearms
(i.tandstnd) has10 gs10 gaitspd poorhlth caffeine drnkspwk smoker if group==1

      Source |       SS       df       MS              Number of obs =    1278
-------------+------------------------------           F( 13,  1264) =   52.89
       Model |  8.14515465    13  .626550358           Prob > F      =  0.0000
    Residual |  14.9745495  1264  .011846954           R-squared     =  0.3523
-------------+------------------------------           Adj R-squared =  0.3456
       Total |  23.1197041  1277  .018104702           Root MSE      =  .10884


------------------------------------------------------------------------------
         bmd |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    caffeine |  -.0396394     .01159    -3.42   0.001    -.0623772   -.0169017
         age |  -.0028671    .000776    -3.69   0.000    -.0043895   -.0013447
     lweight |   .8253971   .0396535    20.82   0.000     .7476031    .9031911
_Iestrogen_1 |   .0180801   .0070925     2.55   0.011     .0041657    .0319946
_Iestrogen_2 |   .0706062   .0086327     8.18   0.000     .0536702    .0875423
     calsupp |  -.0144882   .0066203    -2.19   0.029    -.0274763   -.0015002
    drnkspwk |   .0034561    .001295     2.67   0.008     .0009156    .0059967
        etid |  -.0674801   .0275688    -2.45   0.015    -.1215658   -.0133943
      smoker |  -.0313772   .0157915    -1.99   0.047    -.0623576   -.0003968
     usearms |  -.0519449   .0119332    -4.35   0.000    -.0753559   -.0285338
_Itandstnd_2 |  -.0159852   .0066144    -2.42   0.016    -.0289615   -.0030089
_Itandstnd_3 |  -.0193925   .0156356    -1.24   0.215    -.0500671    .0112821
_Itandstnd_4 |  -.0744849   .0394209    -1.89   0.059    -.1518225    .0028526
       _cons |  -.5337911   .1052804    -5.07   0.000    -.7403346   -.3272477
------------------------------------------------------------------------------
```

Example prediction model from SOF

```
. * Predict outcomes for validation (1/2) subsample
. predict pr if group==2
(option xb assumed; fitted values)
(1454 missing values generated)
. * Estimate R^2 for validation group
. * (to assess discrimination)
. corr bmd pr
(obs=1325)
             |      bmd       pr
-------------+------------------
         bmd |   1.0000
          pr |   0.5639   1.0000
. display r(rho)^2
.31802397
* Plot observed vs predicted values
. * (to assess calibration)
```




```
. twoway (scatter bmd pr, msize(vsmall)) (lowess bmd pr, sort lpattern(longdash)
lwidth(thick)), ytitle("Predicted BMD") xtitle("Observed BMD") legend(off)
```

Potential problems with example approach

- Fitting procedure:
  - doesn't account for interaction & nonlinearity
  - backwards selection ignores many models
  - criteria for predictor selection limited to estimation sample, based on observed associations with outcome (i.e. overfitting a possibility)
- Validation assessment:
  - choice of 1/2, 1/2 split arbitrary (potentially inefficient)
  - single split provides limited assessment of PE
  - validation applies to study pop.
- Available predictors do not adequately support outcome predictions
- "Shrink" coefficient estimates for poor performing predictors

Model selection strategies to avoid over-fitting

1. Pre-specify well-motivated predictors, how to model them
2. Eliminate predictors without using the outcome
3. Select model to minimize "optimism"-corrected PE measure
4. "Shrink" coefficient estimates for poor performing predictors (e.g. "LASSO" regression methods)

Selection to minimize optimism-corrected PE

- Naïve methods – e.g., selecting model to maximize R^2 – lead to over-fitting, optimistic estimates of clinical utility
- Use of an optimism-corrected measure of PE helps avoid over-fitting
- External validation also needed (Altman & Royston, What do we mean by validating a prognostic model? *Stat Med*, 2000;19:453-73)

Optimism-corrected estimates of PE

- Naïve measures penalized for number of predictors: adjusted R^2, AIC, BIC (`estat ic` postestimation command in Stata)
- Use different data to estimate parameters, evaluate PE
  - split-samples (as in SOF example)
  - h-fold cross-validation
  - bootstrap

Optimism-corrected PE: h-fold cross-validation

- Divide data into h = 5 to 10 subsets, then for each subset:
  - fit model to the remaining subsets combined
  - obtain predictions for excluded subset
- Calculated PE from complete set of predictions; repeat
- Do this for each candidate model, select model with minimum cross-validated PE
- More efficient than single split-sample or leave-one-out methods

Example: 2-fold cross-validation for SOF example

- Downloadable Stata utility `regvalidate` uses two-fold cross-validation

```
. quietly xi: stepwise, pr(.1): regress bmd eeu age lweight i.estrogen calsupp diuretic
etid momhip usearms (i.tandstnd) has10 gs10 gaitspd poorhlth caffeine drnkspwk smoker

. regvalidate, reps(50)

original sample size = 2542   reps = 50

regression model: stepwise , pr(.1): regress bmd eeu age lweight _Iestrogen_* calsupp
diuretic etid momhip usearms (_Itandstnd> _*) has10 gs10 gaitspd poorhlth caffeine
drnkspwk smoker

                orig          train          test          diff          orig adj

R-squared       0.3475        0.3506         0.3339        0.0167        0.3309
rss/n           0.0118        0.0119         0.0121       -0.0002        0.0120
fit slope       0.9925        0.9936         0.9668        0.0268        0.9657
fit _cons       0.0064        0.0051         0.0252       -0.0200        0.0264
```

- adjusted estimates are optimism-corrected (compared to original estimates)
- in this case, results are equivalent to repeating the above split-sample approach 50 times and averaging the results
- in general, 5-10 fold cross validation is preferable

Optimism-corrected PE: bootstrapping

- In each of, say, 200 bootstrap samples
  - fit model to bootstrap sample, obtain predictions
  - evaluate PE in bootstrap, original samples
  - average difference in paired PEs estimates optimism
- Fit model to original data, calculate naïve PE, penalize by average bootstrap optimism
- Do this for each candidate model, select model with minimum penalized PE

Application: Point score prediction models

- Individual outcome predictions from model-based risk scores
- Points derived by rounding model coefficients
- Motivation is clinical utility
- Examples: TIMI (Thrombolysis In Myocardial Infarction), versions of Framingham
- Discrimination, calibration can suffer (Gordon et al. Coronary risk assessment by point-based and equation-based Framingham models: significant implications for clinical care. Journal of General Internal Medicine. 2010)

Recommendations for Goal 1

- For clinical use, select easily available predictors
- To avoid over-fitting:
  - eliminate candidate predictors without using outcome
  - select model using cross-validation or bootstrap
- Check loss of discrimination, calibration before going to point score models
- Validate model in external test set

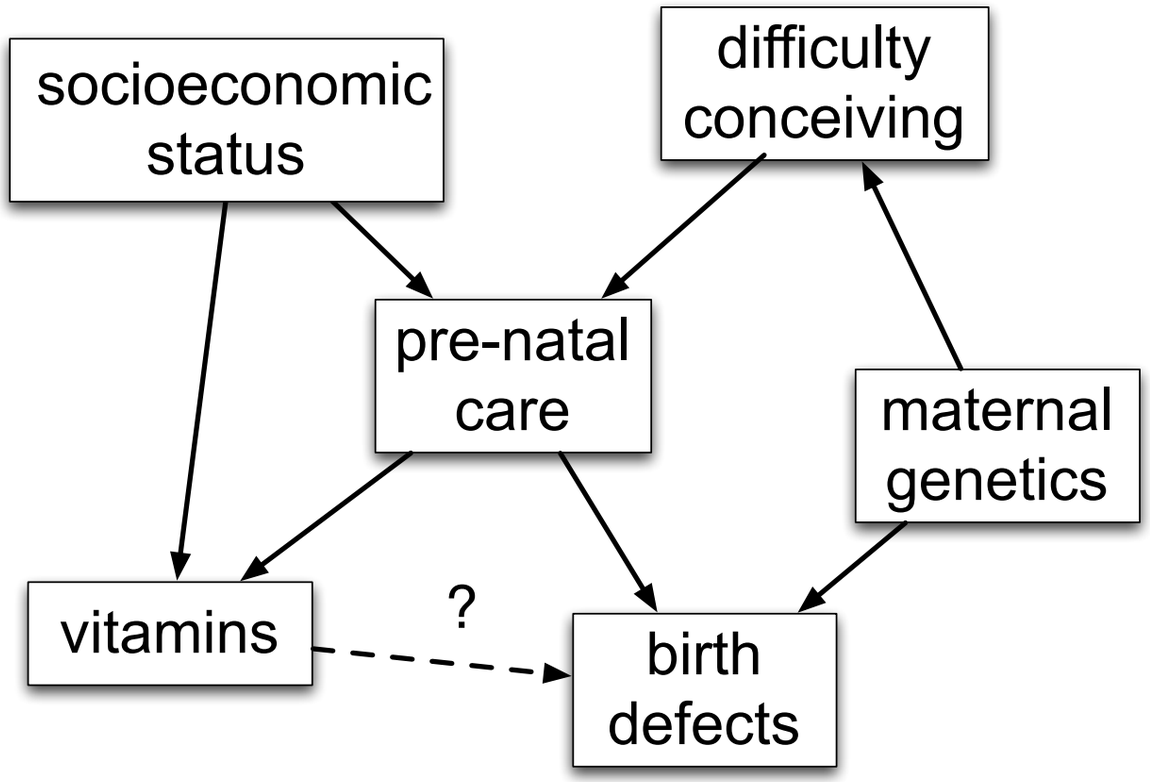
Goal 2: Assessing a predictor of primary interest

- Research question focuses on a single predictor
  - *Example*: Does maternal vitamin use reduce risk of birth defects?
- Ruling out confounding is key
- Minimizing PE is not, so over-fitting not a central issue
- Directed acyclic graphs (DAGs) central in model selection for this goal

Directed Acyclic Graphs (DAGs)

- Help identify what we do and *do not* need to control for to avoid confounding
- Primary predictor, outcome, potential confounders and mediators represented as *nodes* of the DAG
- Causal relationships depicted as *directed edges*
- No factor can cause itself, hence graph is *acyclic*
- A useful DAG requires *a lot* of prior substantive information about what causes what – and what doesn't

Maternal vitamin use and birth defects

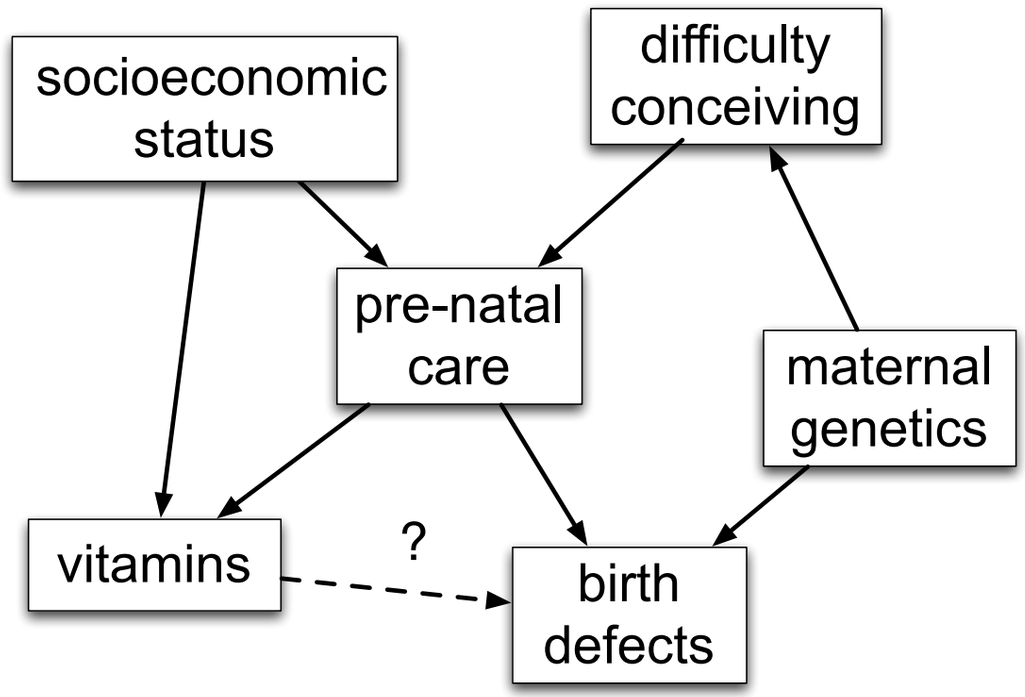


Assumptions about causal pathways:

- Prenatal care (PNC) increases vitamin use, and reduces risk of birth defects directly
- Difficulty conceiving may lead mother to seek PNC
- Maternal genetics affects birth defects and also difficulty conceiving
- SES affects both access to PNC and vitamin use

Additional assumptions in this DAG

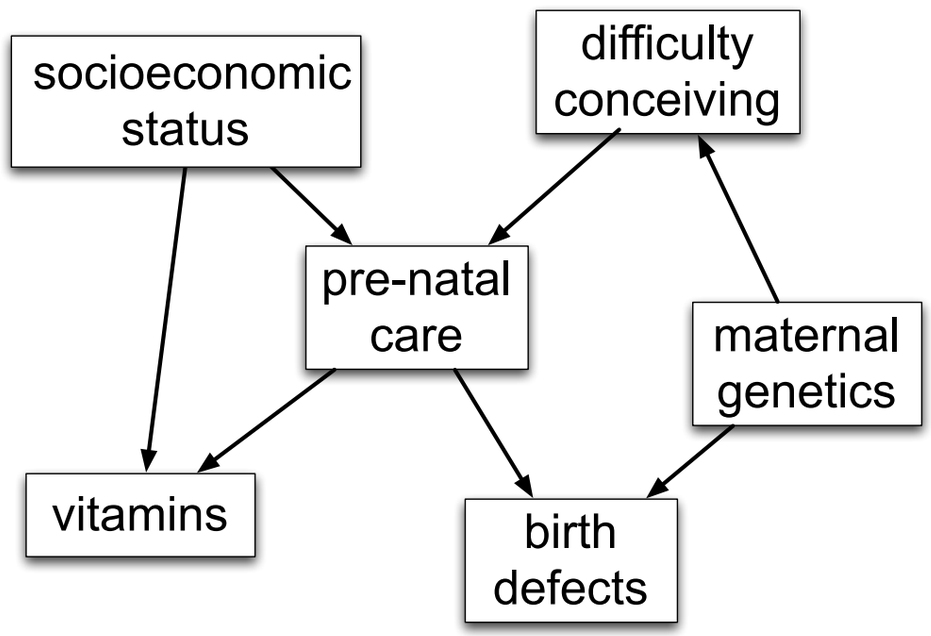
- SES affects birth defects only via PNC, vitamin use
- Difficulty conceiving affects vitamin use only through PNC
- No other common causes of vitamin use and birth defects
- In summary, no excluded confounders or directed edges



Backdoor paths

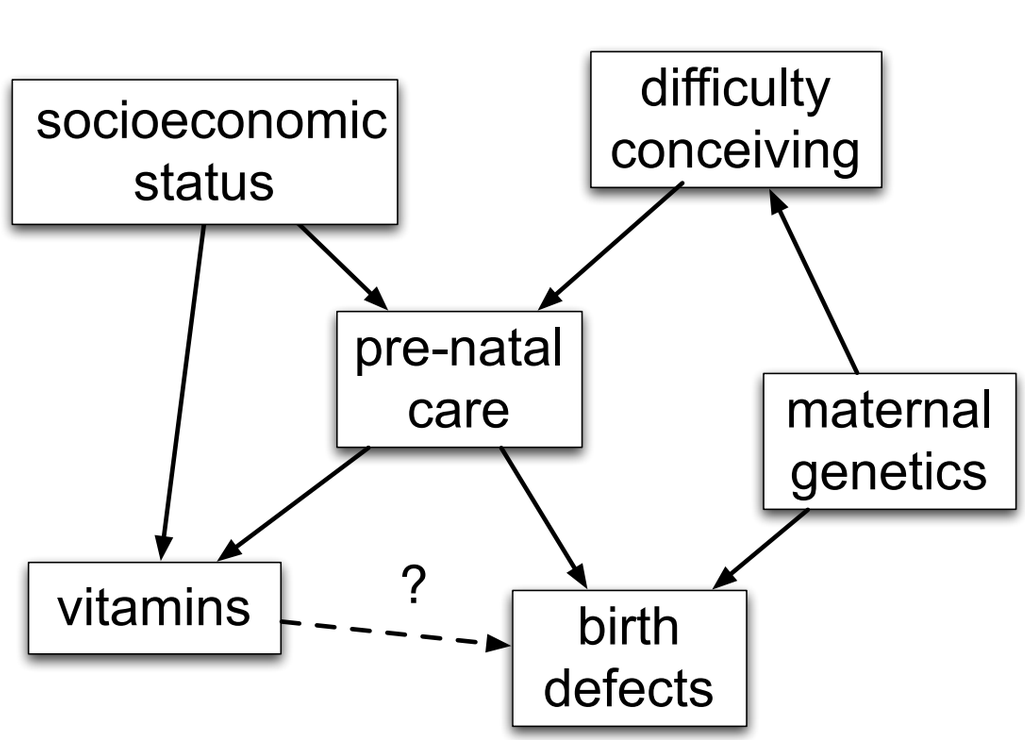
- After removing directed edge from vitamin use to birth defects, four backdoor paths remain between them:

1. Vitamin use ← pre-natal care → birth defects
2. Vitamin use ← SES → pre-natal care → birth defects
3. Vitamin use ← pre-natal care ← difficulty conceiving ← maternal genetics → birth defects
4. Vitamin use ← SES → pre-natal care ← difficulty conceiving ← maternal genetics → birth defects



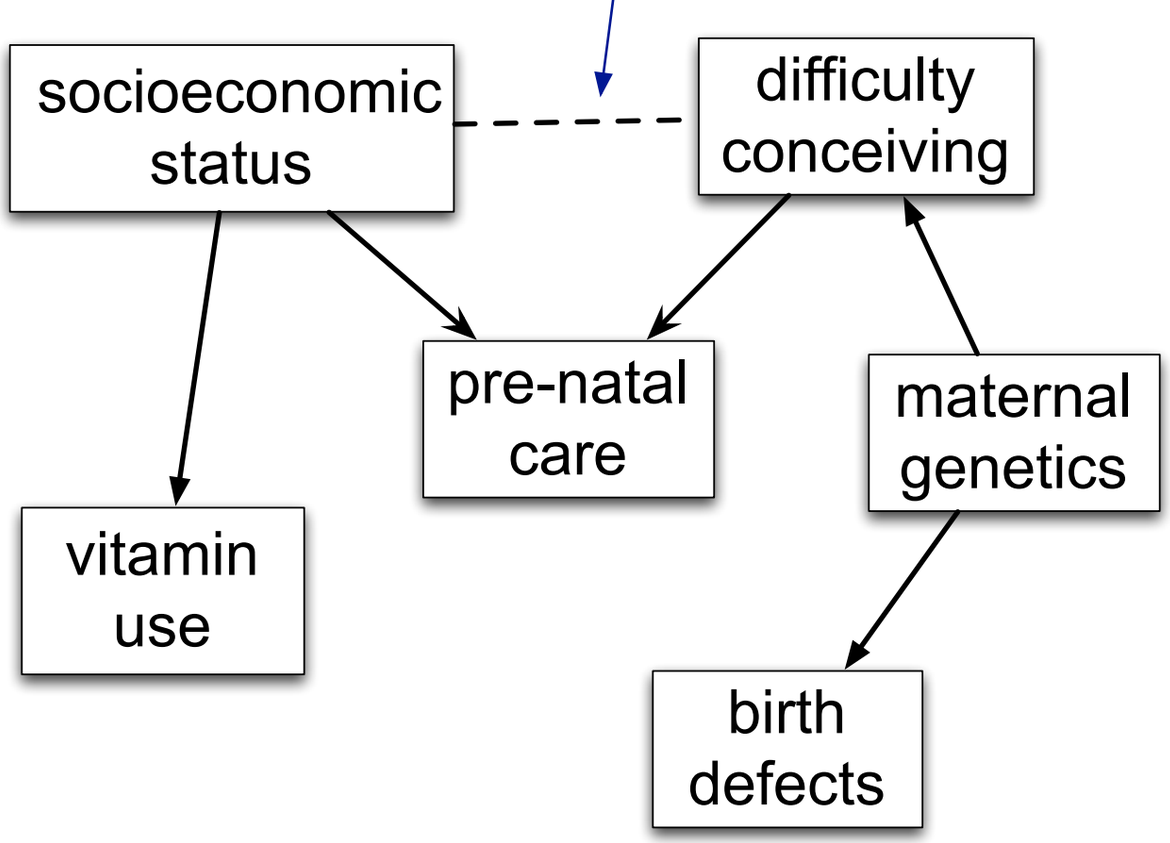
Colliders

- A *collider* on a backdoor path between exposure and outcome is a node with incoming arrows from both directions
- Pre-natal care is a collider of the fourth backdoor path
- It is not a collider on any other backdoor path
- Controlling for pre-natal care would induce a correlation between its common causes, SES and difficulty conceiving, opening a fifth backdoor path



Controlling for a collider induces an association

- Controlling for PNC is tantamount to stratifying on it
- Within PNC user stratum:
  - SES, difficulty conceiving competing explanations for PNC
  - higher SES women less likely to have had difficulty conceiving, and vice versa
- Controlling for PNC opens a backdoor path (representing a negative correlation) via induced association of SES and difficulty conceiving

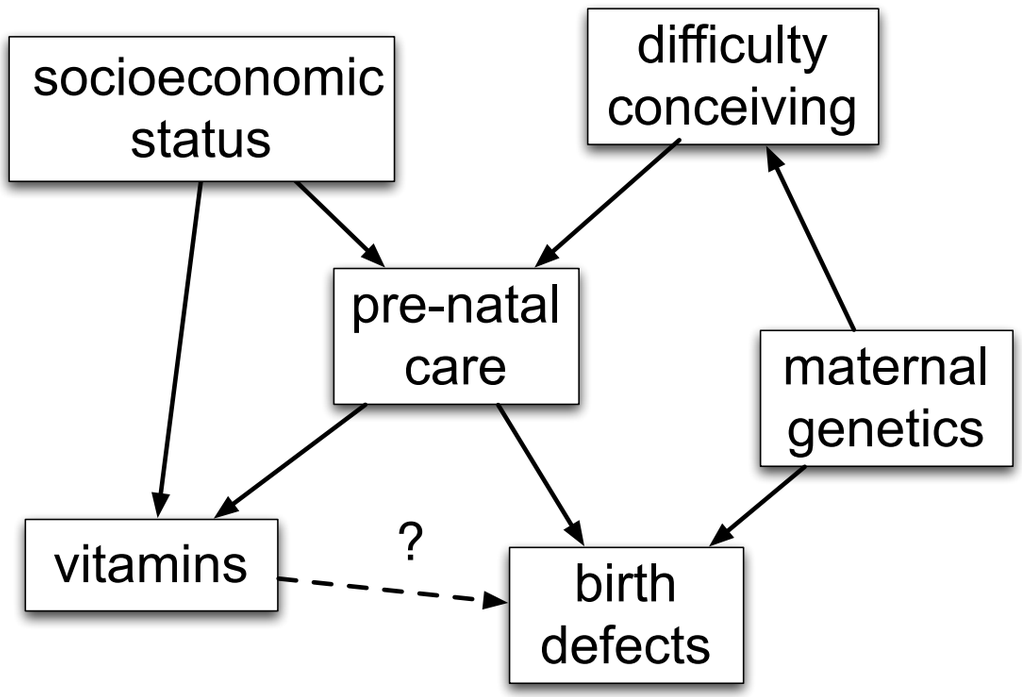


Open and blocked backdoor paths

- If any of the backdoor paths between vitamin use and birth defects remains open, we would expect to find an association between them, even if there was no causal effect
- A backdoor path is blocked, provided we control for at least one non-collider on the path.
- A backdoor path including a collider is blocked provided we do *not* control for the collider
- If we do control for the collider, we need to control for a non-collider on the backdoor path to block it

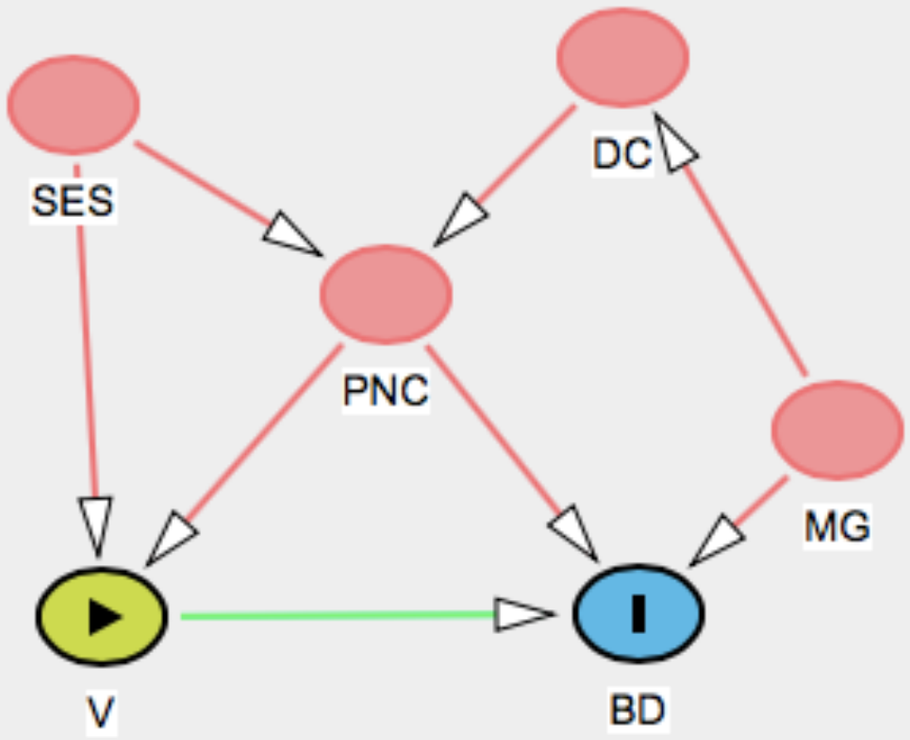
What do we need to control for?

- Controlling for pre-natal care blocks the first three backdoor paths, but it is also a collider
- Controlling for it opens the fourth backdoor path, but controlling for any non-collider on that path will block it
- Controlling for SES or difficulty conceiving would be cheaper and easier than maternal genetics



Determination of MSAS using dagitty.net

Layout | **Help**



Adjustment for total effect

Minimal sufficient adjustment sets for estimating the total effect of V on BD:

- {DC, PNC}
- {MG, PNC}
- {PNC, SES}

Adjustment for direct effect

Total and direct effects are equal in this model.

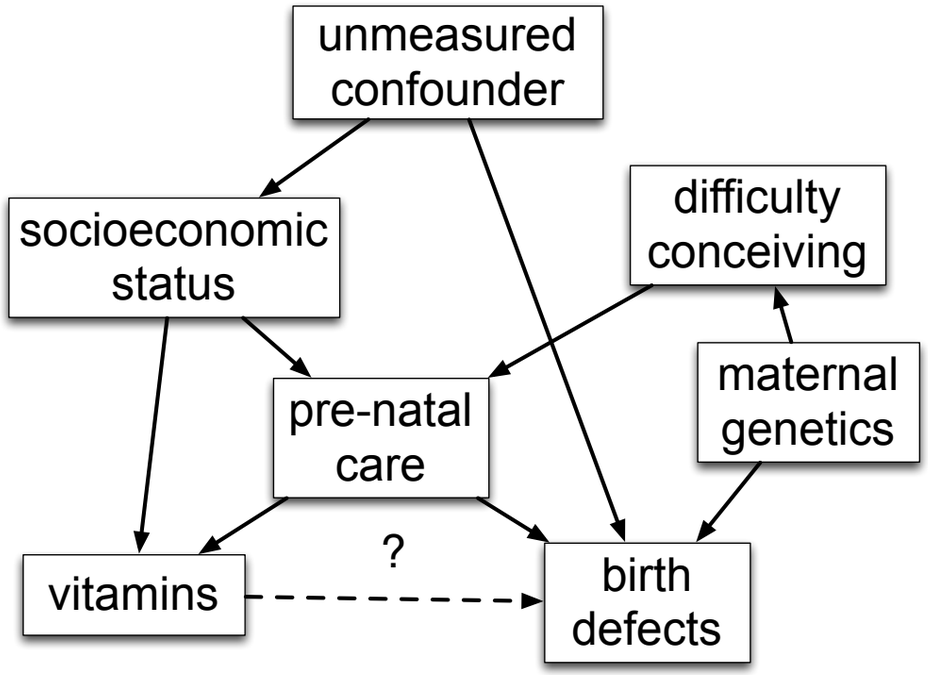
Testable implications

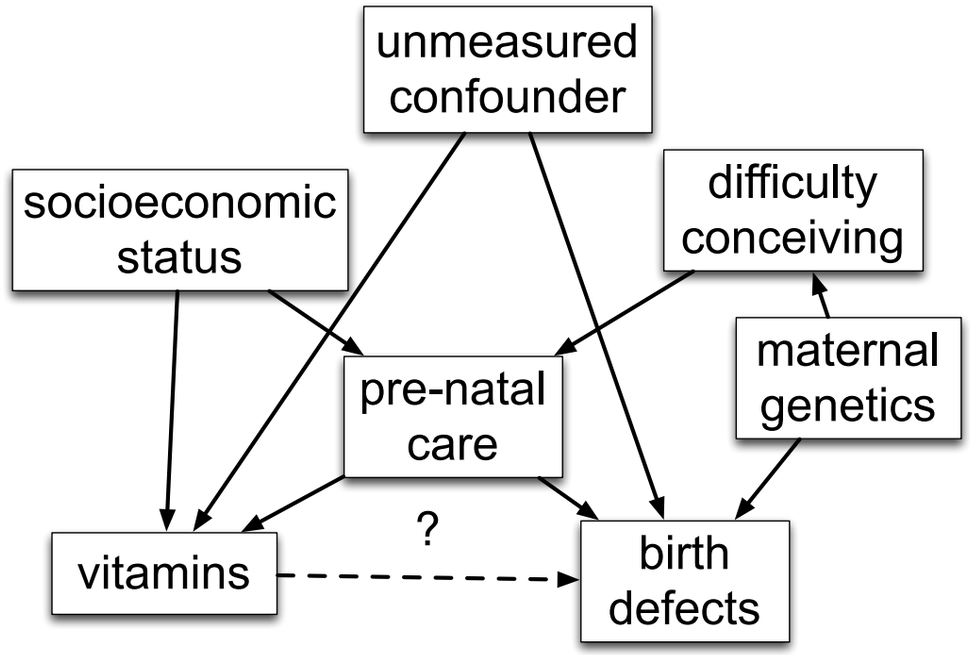
The model implies the following conditional independences:

- V ⊥ DC | PNC, SES
- V ⊥ MG | DC
- V ⊥ MG | PNC, SES
- BD ⊥ SES | DC, PNC, V
- BD ⊥ SES | MG, PNC, V
- BD ⊥ DC | MG, PNC, SES
- BD ⊥ DC | MG, PNC, V
- SES ⊥ DC
- SES ⊥ MG

Remaining issues to consider in confounding control

Unmeasured confounders

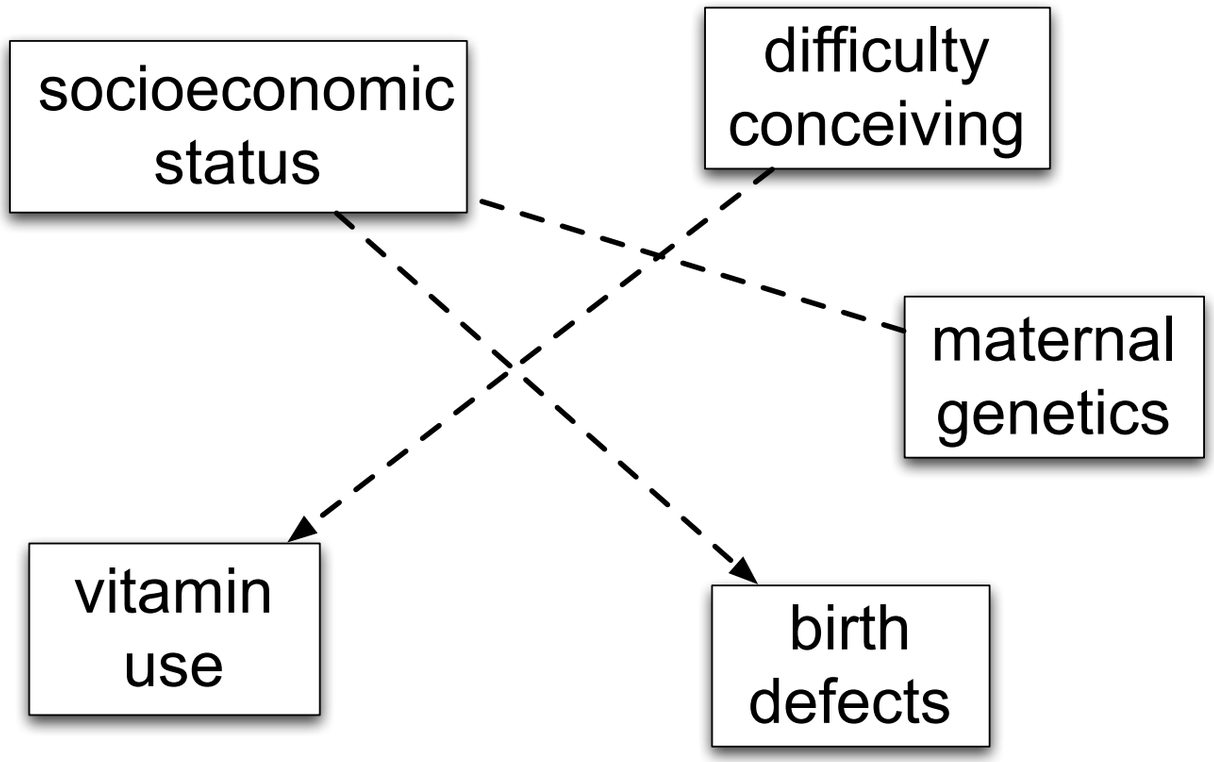




- DAGS can help determine vulnerability to unmeasured confounders
  - provided we have sufficient information about their effects
- Depends on what the measured confounder affects

Remaining issues to consider in confounding control

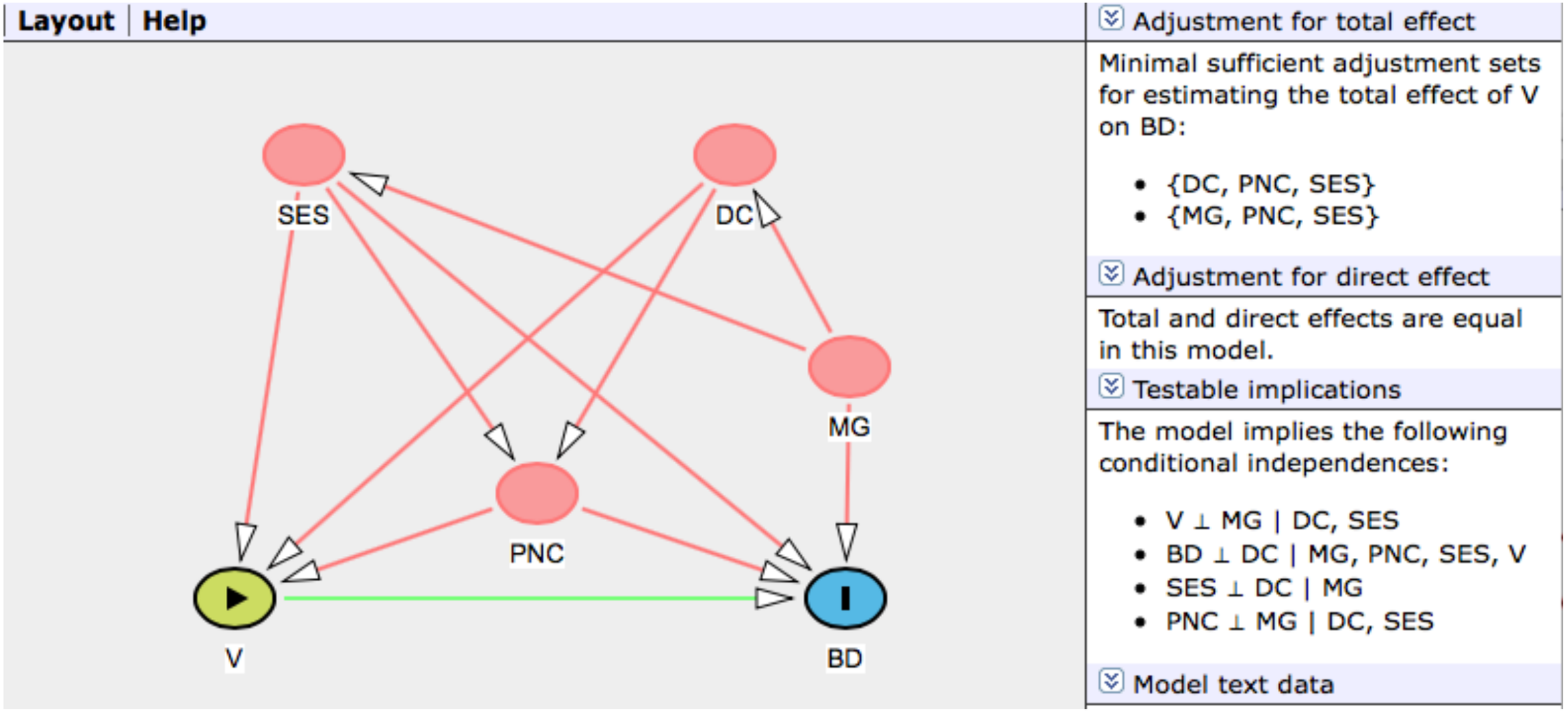
Other plausible causal pathways (edges) to consider:



- SES may affect birth defects via environmental exposures
- Discrimination may link maternal genetics and SES
- Difficulty conceiving could lead directly to vitamin use
- If all these paths are open, we would need to control for SES and difficulty conceiving or maternal genetics to block them
- (Succinct DAGs are not too hard to second guess)

Remaining issues to consider in confounding control

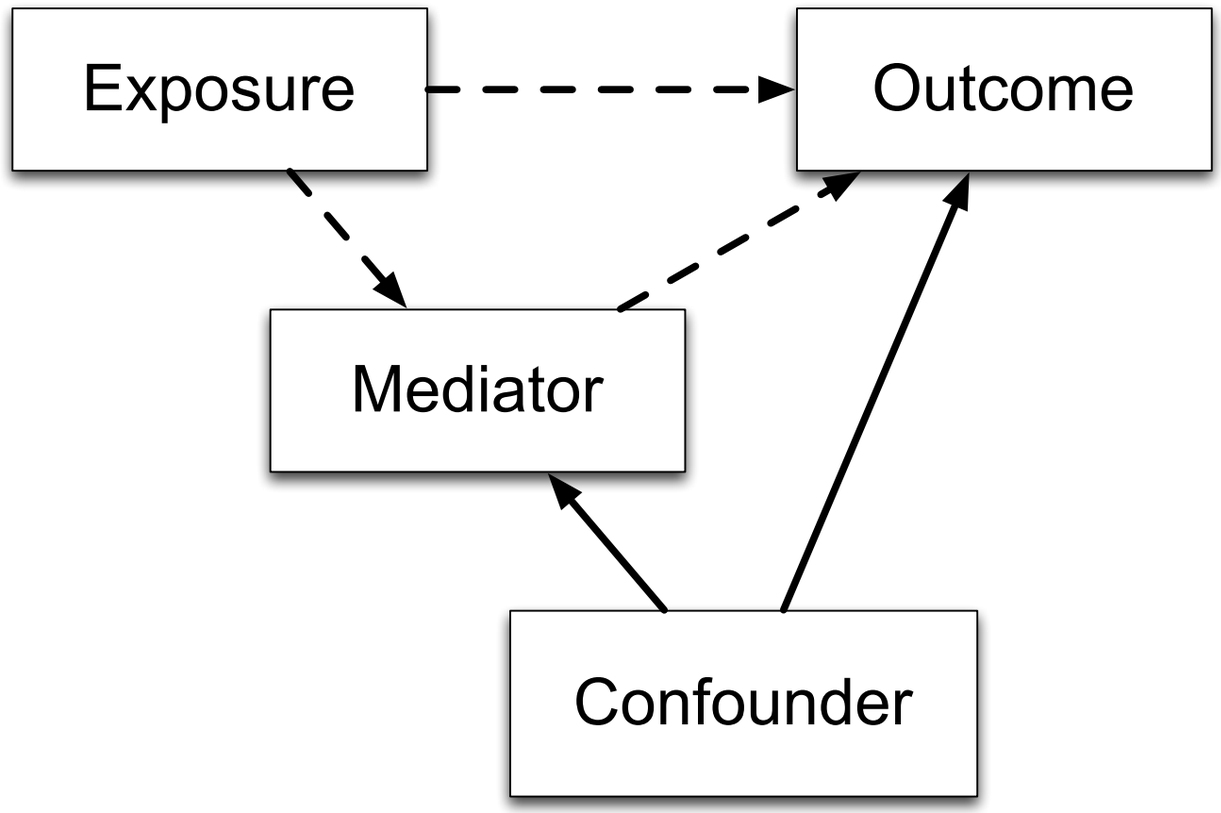
- MSAS options resulting from considering other plausible causal connections



Other insights from colliders

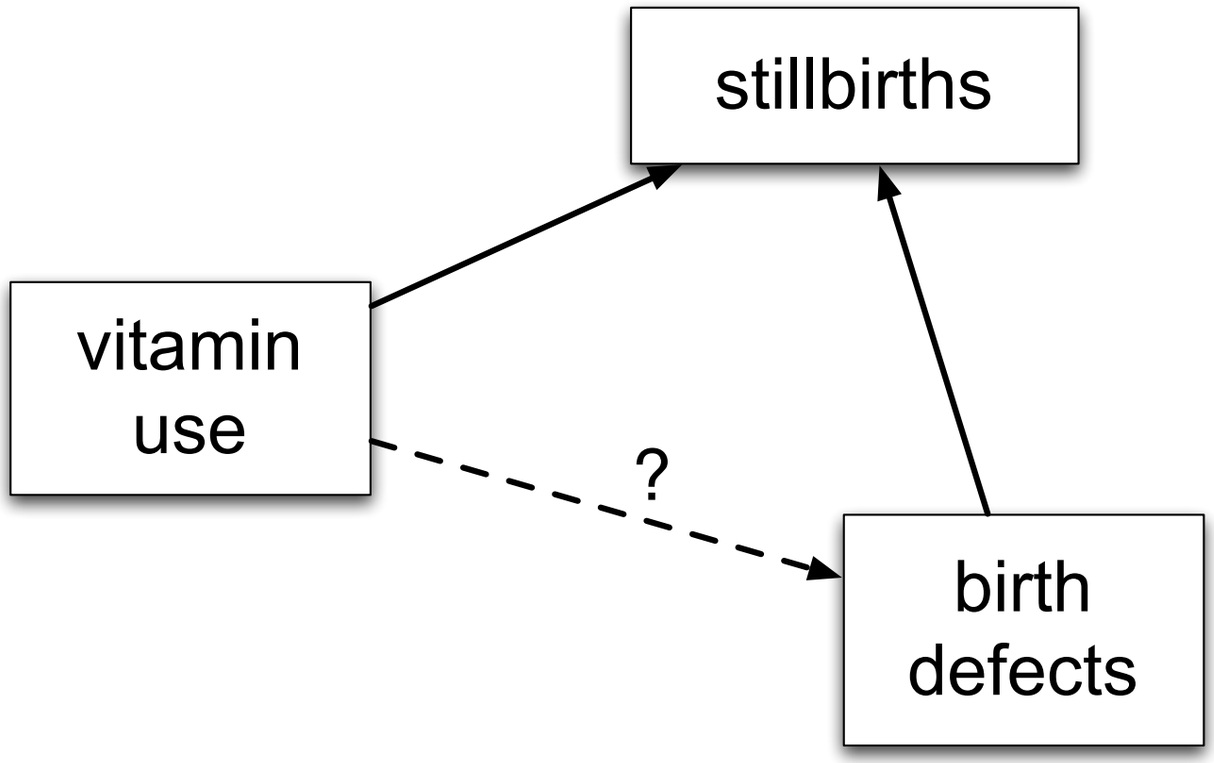
- Need to adjust for confounders of mediator/outcome relationship
- Don't adjust for a common effect of exposure and outcome
- Don't adjust for a common effect of unmeasured causes of exposure and outcome

Confounders of mediator/outcome relationship



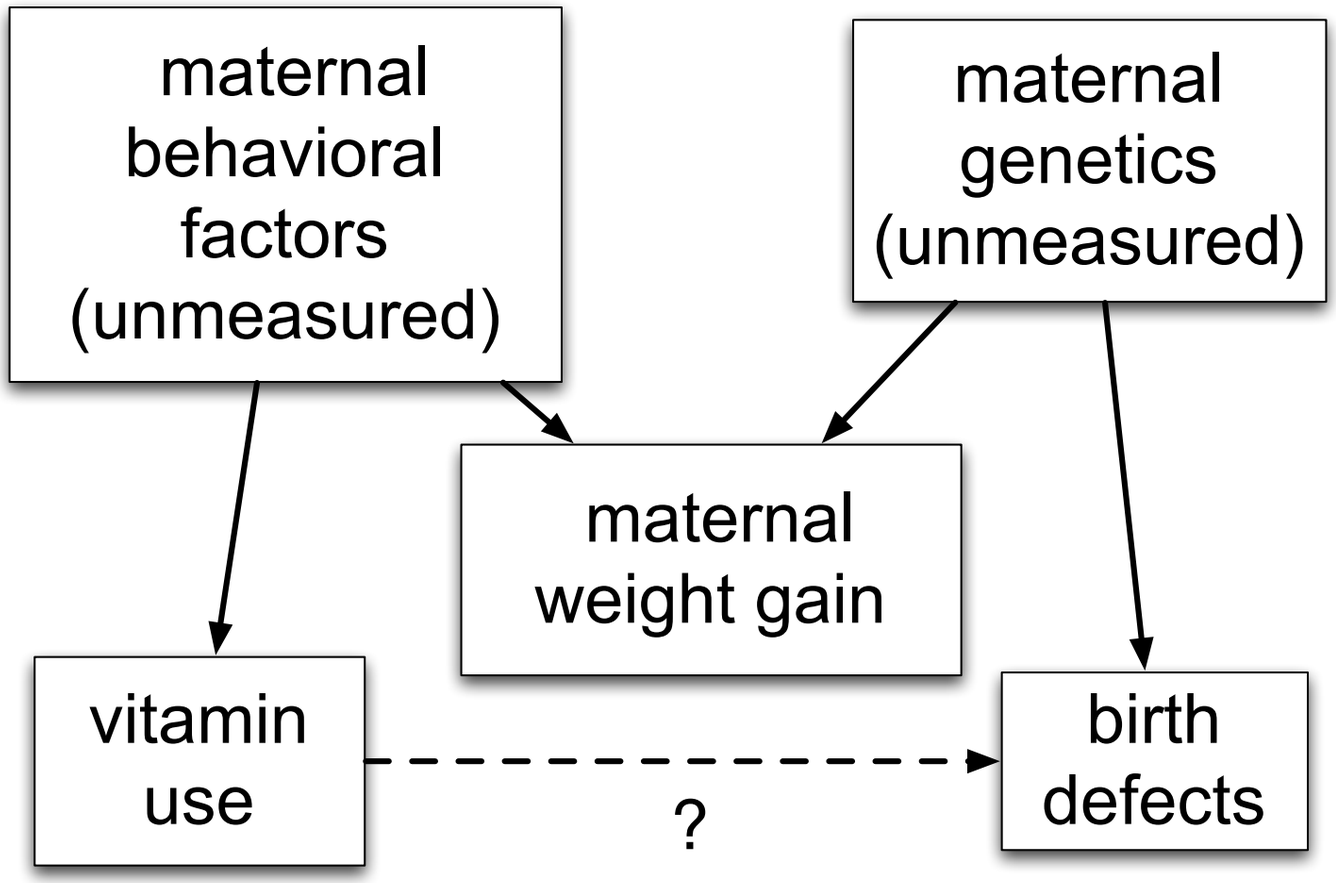
- Mediator is a collider - stratification on Med. will result in a spurious assoc. between Exp. & Out. (Cole SR, Hernán MA. Fallibility in estimating direct effects, Int J Epid, 2002;31:163-5)

Common effect of exposure, outcome



- Stillbirths a possible common cause of vit. use & birth def.

Common effect of unmeasured causes of outcome "M-colliders"



- *Note*: in the case that causal links are weak, potential bias from adjustment is likely small (can try adjusting/not as a sensitivity analysis)

Insights from DAGs

- Exclude from model:
  - mediators (unless estimating direct effect)
  - common effects of outcome and exposure
  - redundant confounders
- Adjusting for a confounder/collider (e.g., PNC) may require adjusting for additional factors
- Often > 1 minimum sufficient adjustment set (MSAS)

Weakly supported ("*B*-list") confounders

- Frequently there are "*B*-list" measured variables with weakly-supported confounding roles
- If DAGs including, excluding *B*-list confounders have a common MSAS, go with it
- Otherwise:
  - use most feasible MSAS based on DAG including *B*-list
  - sequentially drop *B*-list confounders if adjusted coefficient estimate for primary predictor changes < 5% or 10%
- If uncertainty about causal direction (i.e., possible M-collider) affects estimate

Approach to dropping *B*-list confounders

- Suppose X_1 and X_2 are on *B*-list, and co-linear
- Whether each meets the change-in-coefficient criterion can depend on whether other is included
  - each can look unimportant if the other is included, important otherwise
- Requires a cyclical procedure like those used in Stata `stepwise (sw)` commands

Algorithm for dropping *B*-list confounders

- Starting from full model determined by MSAS including *B*-list
  1. Evaluate change-in-coefficient criterion for each confounder on *B*-list one at a time
  2. Drop the one with smallest change < 5% or 10%
  3. Repeat steps 1 and 2, starting from reduced model, until all remaining *B*-list confounders meet retention criterion

Algorithm for dropping *B*-list confounders

- Full *B*-list includes X_1, X_2, and X_3; X_1 and X_2 co-linear
- Iteration 1: dropping X_2 changes coefficient by 1%, dropping X_2 changes it by 2%, X_3 by 15% → drop X_1
- Iteration 2: dropping X_2 changes coefficient by 12%, X_3 by 17% → done
- This means fitting 5 extra models, calculating change-in-coefficient for primary predictor for each one
- We wrote Stata command to automate this procedure; introduced in lab

Limitations of DAGs

- No good way to represent interactions, no guidance about keeping or excluding them
- No representation of functional form (i.e., linearity)
- Co-linearity, allowable numbers of predictors ignored; more later about these issues
- Can be hard to specify convincingly

Recommendations for Goal 2

- Use DAG to identify MSASs, exclude mediators, common effects of exposure and outcome
  - use Dagitty.net for complicated DAGs
  - choose the most feasible MSAS
  - use change-in-estimate criterion, sensitivity analysis to deal with weakly supported potential confounders
- More later on number of predictors, interactions with main predictor, mediation, high correlation among confounders
- Alternative procedure when you can't draw a convincing DAG
  - identify an *A*-list of confounders strongly supported by the literature and/or face validity
  - identify a *B*-list of plausible but unclearly supported potential confounders
  - exclude mediators, common effects of exposure and outcome
  - use change-in-coefficient criterion to exclude unimportant *B*-list confounders

Recommendations for Goal 2 (*continued*)

- Hypothesis testing only of interest for primary predictor – so somewhat robust against inflation of type-I error
  - type-I errors for covariates are irrelevant
  - results for covariates can be left in the background
  - Over-fitting is primarily a problem for Goal 1, not Goal 2

Randomized Trials - a Special Case of Goal 2

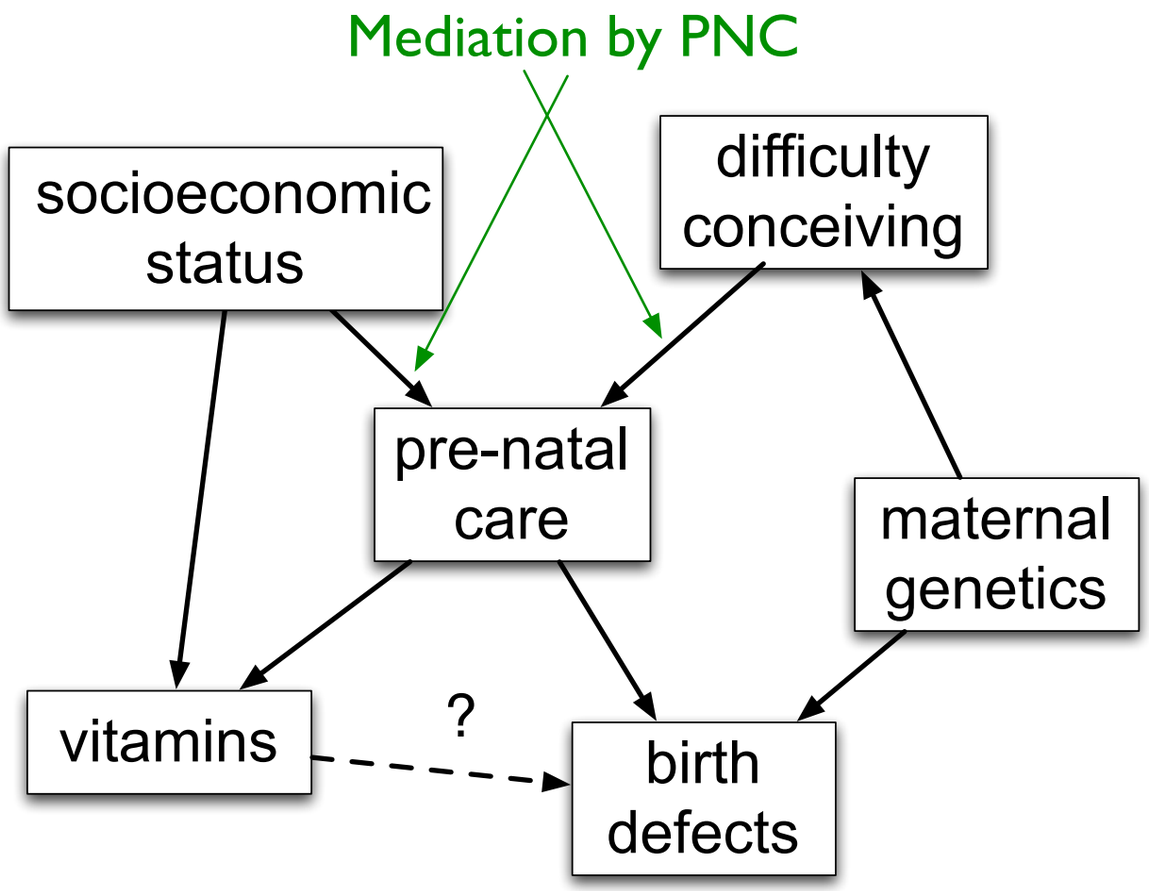
- Treatment the predictor of primary interest
- Confounding not an issue, provided randomization succeeds
- Include covariates to
  - account for clustering by clinical center
  - reduce variance using pre-specified stratification variables
- Do *not* adjust for any post-randomization variables

Goal 3: evaluating multiple predictors as risk factors for an outcome

- What are the risk factors for an outcome?
- Most difficult of three inferential goals
- Instead of one predictor of primary interest, several variables may be targeted

Goal 3: potential problems

- Many possible mediating, interaction relationships
- False positive findings, particularly for interactions
- No single model will summarize causal relationships
  - hard to achieve unconfoundedness for all included factors
  - mediation often an issue (e.g. in assessing effects of SS & DC in example below)



Recommendations for Goal 3

- Ruling out confounding is still central
- Best but labor intensive solution: treat each predictor as primary in turn, use DAG-based methods for Goal 2 (not the position taken in Section 10.3.2 of book)

An alternative approach

- Fit a single big model including potential confounders
  - needed for face validity, regardless of statistical criteria
  - if they meet liberal statistical inclusion criterion ($p < 0.2$)
    - *Note:* change-in-coefficient criterion not applicable (i.e. many coeff. involved)
- Cautiously interpret weaker, less plausible findings
- Multiple models may be required to deal with mediation

Additional topics in predictor selection

Techniques for selecting parsimonious models

- Retain only variables with $p < 0.05$ for goals 2 and 3?
- Motivations:
  - avoid over-fitting
  - Occam's razor: simpler explanations likelier – limit type-I errors
  - less to explain in the discussion
- Drawback: residual confounding, especially in small samples

Parsimonious models - over fitting

- Goal 1:
  - over-fitting increases prediction error
  - parsimonious models often have lower prediction error (predictors best chosen by cross-valid./bootstrap to minimize PE)
- Goal 2:
  - parsimony can lead to uncontrolled confounding
  - over-fitting of covariates usually not a problem

Number of predictors to include?

- Too many predictors can
  - degrade precision
  - in smaller datasets, swamp a real association
  - induce bias in logistic and Cox models, other GLMs

Recommendations: number of predictors

- Use 10-15 observations/predictor (10 events per predictor for binary/ survival outcomes) as a cautionary flag
- If you are close to 10, check for
  - high correlations between predictors
  - inflated SEs when a new covariate is added
  - inconsistency between Wald and likelihood ratio tests (logistic and Cox models)
  - gross inconsistency with smaller models
- If trouble is apparent:
  - tighten inclusion criterion to $p < 0.15$ or $p < 0.1$
  - omit variables included only for face validity
- With a binary primary predictor, many potential confounders, and a rare outcome, consider *propensity scores*
  - *Note:* propensity scores don't solve the problem with a rare predictor, or control for unmeasured confounders

Co-linearity between predictors

- Variance inflation factor (*VIF*): variability (SE) of β_j (coeff. for j^{th} predictor) increases with corr. (r_j) between x_j and other predictors in model (see lecture 2 & sect. 4.2.2.2):

$$VIF(\hat{\beta}_j) = \frac{1}{(1 - r_j^2)}$$

- Can make coefficient estimates very imprecise, even when F test for combined effects is statistically significant
- Co-linear predictors give similar information about outcome
  - p-values don't help distinguish between them

Diagnosing co-linearity

- High pairwise correlations (r) between predictors: caution if $r > 0.8$, trouble if $r > 0.9$
- Check variance inflation factors (`estat vif` postestimation command); watch out for *VIF* > 5
- If either of two co-linear predictors is included one at a time, *t*-test for predictor is significant
- In model with both predictors, *F*-test for joint effect significant, *t*-tests are not

Recommendations: dealing with co-linearity

- Goal 1: use cross-valid. to select one of co-linear predictors
- Goal 2: co-linearity between predictor of interest and a confounder in every MSAS: you may have to admit defeat
- Goal 3: see if the model solves the problem: e.g., one clearly dominates, or both are independently predictive. Or, if it makes sense, create summary variables or select among them
- Co-linearity between control variables (i.e. for goal 2): *not* a problem

Example co-linearity between control variables

- From lab 6: controlling for non-linearity in a continuous predictor by adding polynomial (e.g. squared) terms

```
* regression of BMD on age and weight (including square of weight)

. reg bmd age weight weight2

      Source |       SS       df       MS              Number of obs =     278
-------------+------------------------------           F(  3,   274) =   43.38
       Model |  1.67578732     3  .558595772            Prob > F      =  0.0000
    Residual |  3.52855652   274  .012877943            R-squared     =  0.3220
-------------+------------------------------           Adj R-squared =  0.3146
       Total |  5.20434383   277  .018788245            Root MSE      =  .11348

------------------------------------------------------------------------------
         bmd |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0047303   .0015033    -3.15   0.002    -.0076897   -.0017709
      weight |   .0182735   .0035289     5.18   0.000     .0113263    .0252207
     weight2 |  -.0000925   .0000239    -3.87   0.000    -.0001395   -.0000455
       _cons |   .3012021   .1826122     1.65   0.100    -.0582992    .6607034
------------------------------------------------------------------------------

. corr bmd weight weight2
             |      bmd   weight  weight2
-------------+---------------------------
         bmd |   1.0000
      weight |   0.5080   1.0000
     weight2 |   0.4742   0.9896   1.000

. estat vif
    Variable |       VIF       1/VIF
-------------+----------------------
      weight |     48.58    0.020585
     weight2 |     48.39    0.020665
         age |      1.07    0.938879
```

Example co-linearity between control variables

```
* regression of BMD on age and weight (including square of centered weight)

. quietly summ weight

. gen cweight2 = (weight - r(mean))^2

. corr weight weight2 cweight2
             |   weight  weight2 cweight2
-------------+---------------------------
      weight |   1.0000
     weight2 |   0.9896   1.0000
    cweight2 |   0.4528   0.5764   1.0000

. reg bmd age weight cweight2
      Source |       SS       df       MS              Number of obs =     278
-------------+------------------------------           F(  3,   274) =   43.38
       Model |   1.6757873     3  .558595768           Prob > F      =  0.0000
    Residual |  3.52855653   274  .012877944           R-squared     =  0.3220
-------------+------------------------------           Adj R-squared =  0.3146
       Total |  5.20434383   277  .018788245           Root MSE      =  .11348

------------------------------------------------------------------------------
         bmd |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0047303   .0015033    -3.15   0.002    -.0076897   -.0017709
      weight |   .0057653   .0005842     9.87   0.000     .0046152    .0069154
    cweight2 |  -.0000925   .0000239    -3.87   0.000    -.0001395   -.0000455
       _cons |    .724028   .1325897     5.46   0.000      .463004    .9850519
------------------------------------------------------------------------------
```

- Results for age unchanged, effect of weight more precisely estimated

Model selection algorithms

- Procedures for selecting predictors on statistical grounds, some implemented in Stata:
  - univariate screening; change-in-estimate criterion; backwards, forwards, and stepwise selection
- Epidemiologists and statisticians both uncomfortable with these methods
- For goal 1, screening with cross-validation works better
- For goals 2 and 3, DAG-based procedures preferable

Screening with single predictor models (AKA univariate screening)

- Screen using single-predictor models, retain predictors with p-value less than some criterion (e.g. $p<0.1$)
- Use a liberal retention criterion to avoid missing negatively confounded variables
- Initial list should reflect substantive considerations
- Can be useful for reducing the number of covariates (but better done a priori)

Change-in-estimate criterion

- Force in strongly motivated *A*-list predictors; delete *B*-list predictors if omission changes coefficient for primary predictor by < 5 or 10% (goal 2)
- Advantages:
  - reflects empirical criterion for confounding
  - can be integrated with DAG-based procedures – now implemented in Stata `cic` command (see lab 7)
- Disadvantage: defining A- and B-lists often challenging)

Backward stepwise selection

- Backward selection
  - Start from big model including all candidate predictors
  - Variables can be forced in
  - Sequentially delete predictor with biggest p-value and re-fit model
  - Stop when all remaining variables meet inclusion criterion (e.g., $p < 0.2$)

- Advantages: should do well with negative confounding
- Disadvantages:
  - initial model could be unstable if too large
  - selections subject to chance, especially between co-linear variables
  - sequential algorithm, so good models may be missed

Forward stepwise selection

- Start from minimum set of variables (i.e., the "*A*-list")
- Sequentially add omitted predictor with smallest *p*-value
- Stop when no more variables meet inclusion criterion
- Stepwise also allows variables to be deleted
- Advantages: Avoids problem with too-large initial model
- Disadvantages:
  - results may be affected by negative confounding

Categorical variables in stepwise selection procedures

- Selection procedures should keep indicator variables for all categories together - can use parentheses to force this in Stata:

```
xi:  stepwise, pr(.2):  regress outcome x1 (i.x2) x3 x4
```

  - uses F test to decide to keep or drop all levels of x_2
  - note that `xi:` prefix is required in Stata versions ≤ 13
  - command `stepwise` can be abbreviated as `sw`
- *Note*: collapsing categories post hoc inflates type-I error

Pitfalls of automated model selection

- *P*-values too small, CIs too narrow, but hard to fix (type-I error rate inflated by searching over many models)
- Testimation bias: coefficient estimates biased away from null, especially for weak predictors
- Only complete cases used in automated procedures (i.e., no missings on any predictors in initial list)
- Poorly motivated, unstable choices between predictors

A priori model selection (preferred)

- Select MSAS based on well-motivated DAG
- *Advantages*
  - avoids pitfalls of model selection
  - *p*-values retain their theoretical meaning – defensible on substantive grounds
- *Disadvantages*
  - more suitable for well-studied areas; DAG, MSAS may be controversial
  - opens door to residual confounding if important variables, arrows are omitted from DAG
- Still requires checking selected model for unmodeled nonlinearities, omitted interactions

Conducting sensitivity analyses

- Multiple plausible DAGs: show whether substantive results differ by MSAS
- For statistically driven approaches, check sensitivity to
  - model selection procedure (forwards, backwards, stepwise)
  - model size / retention criterion
  - a priori variable choices (e.g., between collinear variables)

Summary Recommendations

- *Goal 1* – prediction: select model using aggressive search plus cross-validation or AIC/BIC, validate in external sample
- *Goal 2* – predictor of primary interest: adjust for MSAS based on a strongly-motivated DAG; use change-in-estimate criterion or sensitivity analysis to deal with weakly motivated predictors
- *Goal 3* – identifying multiple independent predictors: treat each predictor as primary in turn, using methods for Goal 2; *or*, use an inclusive model, interpret weak findings cautiously
- Numbers of predictors: relax the 10-OBV/EPV rule of thumb if necessary to rule out confounding, but check for bad behavior
- Co-linearity:
  - between control variables, not a problem
  - between a predictor of primary interest and a non-ignorable confounder, may mean defeat
- Use sensitivity analyses