Predictor Selection

February 18, 2020

Biostatistics 208

Predictor Selection (Chapter 10 of text)

- 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
- Methods for all 3 goals under continuing development, especially with recent advances in machine learning
- Biostatistics 210 & 215 offer more detailed coverage

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 of outcome & predictors 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 average outcomes, failure rates?
- Both discrimination and calibration are important

Prediction

- Bias-variance tradeoff and over-fitting
  - Parameter estimates & predicted values from regression:
    - biased when important predictors omitted from model
    - more variable when unimportant predictors are included
- Excess predictors yield "over-fitted" estimates reflecting minor data features
  - overfit models may yield poor prediction performance

Prediction example: "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 $3/4$ of the sample (the *training set*); validated in the remaining $1/4$ (the *validation set*)

```
. * split sample into two groups consisting of 3/4, 1/4 of observations
. splitsample, generate(group) split(0.75 0.25) rseed(2182020)
. * compare mean outcomes in 2 groups
. tabulate group, summarize(bmd)
```

group	Summary of BMD		
	Mean	Std. Dev.	Freq.
1	.73162908	.13622632	2,084
2	.7272446	.13360084	695
Total	.73053257	.13556385	2,779

Example prediction model from SOF

```
. * use backwards selection to select model, restricted to 3/4 training set

. 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   =     1,913
-------------+----------------------------------   F(10, 1902)     =    100.91
       Model |  12.0856305         10  1.20856305   Prob > F        =    0.0000
    Residual |  22.7791996      1,902  .011976446   R-squared       =    0.3466
-------------+----------------------------------   Adj R-squared   =    0.3432
       Total |  34.8648301      1,912  .018234744   Root MSE        =    .10944

------------------------------------------------------------------------------
         bmd |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         eeu |   .0030044   .0010282     2.92   0.004     .0009879    .0050209
         age |  -.0026494   .0006224    -4.26   0.000    -.0038701   -.0014287
     lweight |   .8454537   .0328716    25.72   0.000     .7809855    .9099219
_Iestrogen_1 |   .0046784   .0058427     0.80   0.423    -.0067803    .0161371
_Iestrogen_2 |   .0748497   .0069694    10.74   0.000     .0611812    .0885182
     calsupp |  -.0093625   .0054043    -1.73   0.083    -.0199614    .0012365
    caffeine |  -.0358523   .0091882    -3.90   0.000    -.0538722   -.0178324
        etid |   -.045589   .0236228    -1.93   0.054    -.0919183    .0007404
      momhip |  -.0200263   .0081115    -2.47   0.014    -.0359347    -.004118
     usearms |  -.0578298   .0092322    -6.26   0.000    -.0759361   -.0397235
       _cons |  -.5957026    .087274    -6.83   0.000    -.7668656   -.4245397
------------------------------------------------------------------------------
```

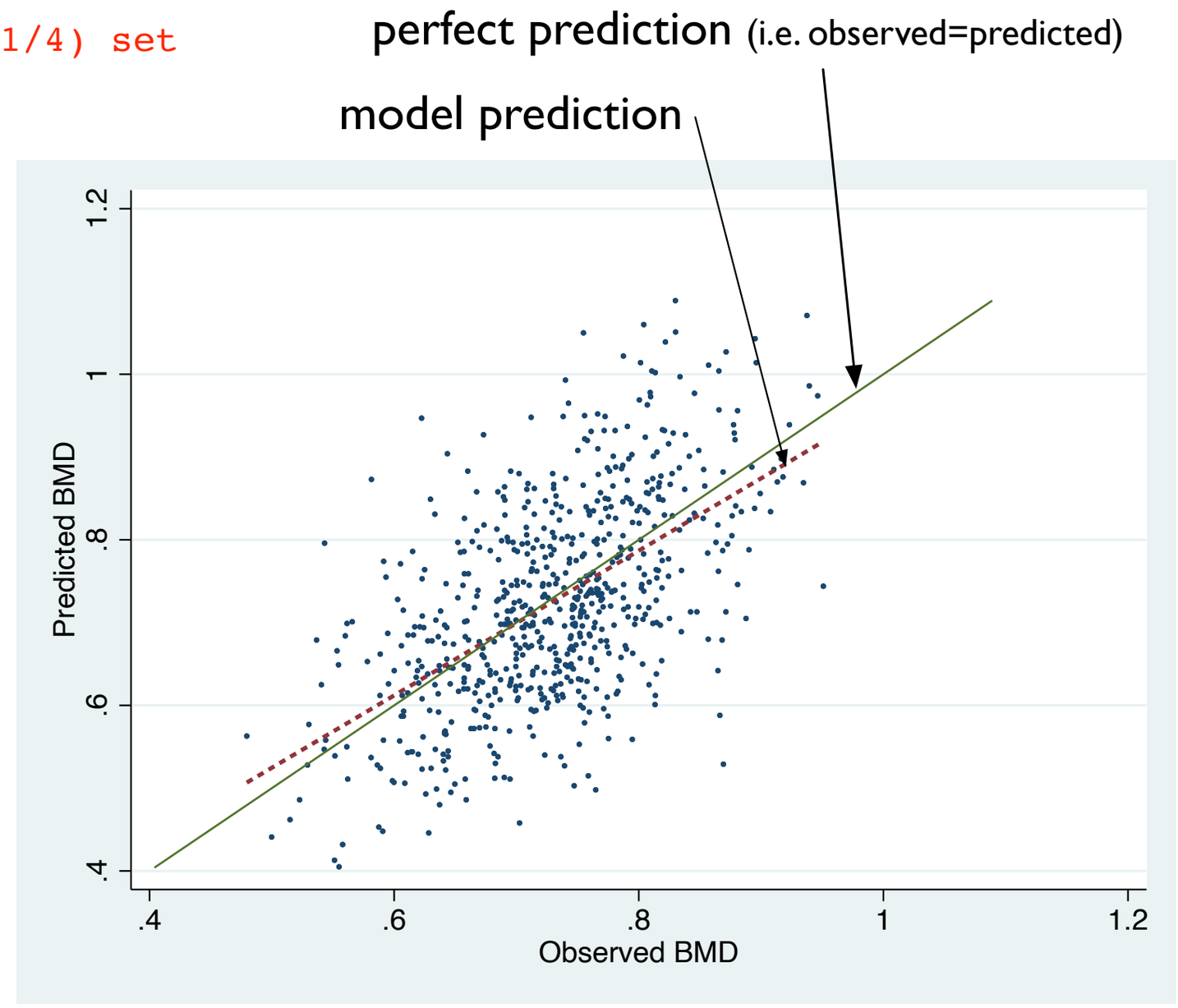
Note: old "`xi:`" syntax required for proper inclusion of categorical terms

Example prediction model from SOF

```
. * Predict outcomes for validation (1/4) set
. predict pr if group==2
. * Estimate R^2 for validation group
. * (to assess discrimination)
. corr bmd pr
(obs=684)

             |      bmd       pr
-------------+------------------
         bmd |   1.0000
          pr |   0.5772   1.0000

. display r(rho)^2
.33319165
* Plot observed vs predicted values
. * (to assess calibration)
. twoway (scatter bmd pr, msize(tiny))  (lfit bmd pr, sort lpattern(shortdash)
lwidth(0.5)) (line bmd bmd, sort lwidth(0.25) ytitle("Predicted BMD")
xtitle("Observed BMD") legend(off)) if inrange(bmd, 0.4, 1.1)
```



Problems with example (stepwise) approach

- Fitting procedure:
  - doesn't account for interaction & nonlinearity
  - backwards selection ignores many models
  - criteria for predictor selection limited to single training sample (i.e. overfitting a possibility)
- Validation assessment:
  - choice of 3/4, 1/4 split arbitrary (potentially inefficient)
  - validation limited to study pop. (no external validation)

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 in training data – lead to over-fitting, optimistic estimates of clinical utility
- Use of an optimism-corrected measure of PE helps avoid over-fitting

Optimism-corrected estimates of PE

- Naïve measures penalized for number of predictors: adjusted R^2, AIC, BIC (obtained via `estat ic` postestimation command in Stata)
  - retain disadvantage that they are based on the estimation sample

- Use different data to estimate parameters, evaluate PE
  - measure out of sample performance, include
    - 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 predictions; repeat
- Calculated optimism-corrected PE by averaging over subset results
- 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: 10-fold cross-validation for SOF example

- Downloadable Stata utility `crossfold` uses *k*-fold cross-validation

```
crossfold: regress bmd eeu age lweight i.estrogen calsupp caffeine etid momhip usearms
caffeine, k(10) r2
```

```
             | Pseudo-R2
-------------+-----------
        est1 |  .4089059
        est2 |  .3534381
        est3 |   .317163
        est4 |  .3441949
        est5 |  .3768065
        est6 |  .3354325
        est7 |  .3012609
        est8 |  .2727431
        est9 |  .3432689
       est10 |  .4281787
```

```
. svmat double r(est), name(cvr2))
. mean cvr2
Mean estimation                   Number of obs   =         10
--------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
       cvr21 |   .3481392   .0149038      .3144244    .3818541
--------------------------------------------------------------
```

Optimism-corrected R^2 (compare to R^2 of 0.3505 obtained from fitting model to full sample)

Example: 10-fold cross-validation for SOF example

- Previous example estimates optimism-corrected prediction error for a *single model* fitted to one training sample
  - Performance of resulting model is likely still *over-optimistic* because variable selection was limited
- In practice, alternate models would be subjected to the same procedure and the model with the lowest optimism-corrected PE would be selected for further validation

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
  - This will be illustrated for logistic regression in lecture 11

Application: Point score prediction models based on regression

- Individual outcome predictions from model-based "risk scores"
- "Points" derived by rounding/scaling 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)
- See section 10.1.6 of text for an example using Cox regression

Recommendations for Goal 1

- For clinical use, select easily available predictors
- Pay attention to nonlinearity and interactions in fitting candidate models
- 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 (Altman & Royston, What do we mean by validating a prognostic model? *Stat Med*, 2000;19:453-73)
- Consider applying modern "supervised" *machine learning* tools (e.g. lasso, random forests, super learner) in applications where number of predictors and interpretability are not of primary concern, or for sensitivity analyses of conventional regression-based approaches
  - *Reference*: James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. 2013; Springer, NY.
- These topics are treated in more detail in Biostatistics 210

Goal 2: Assessing a predictor of primary interest

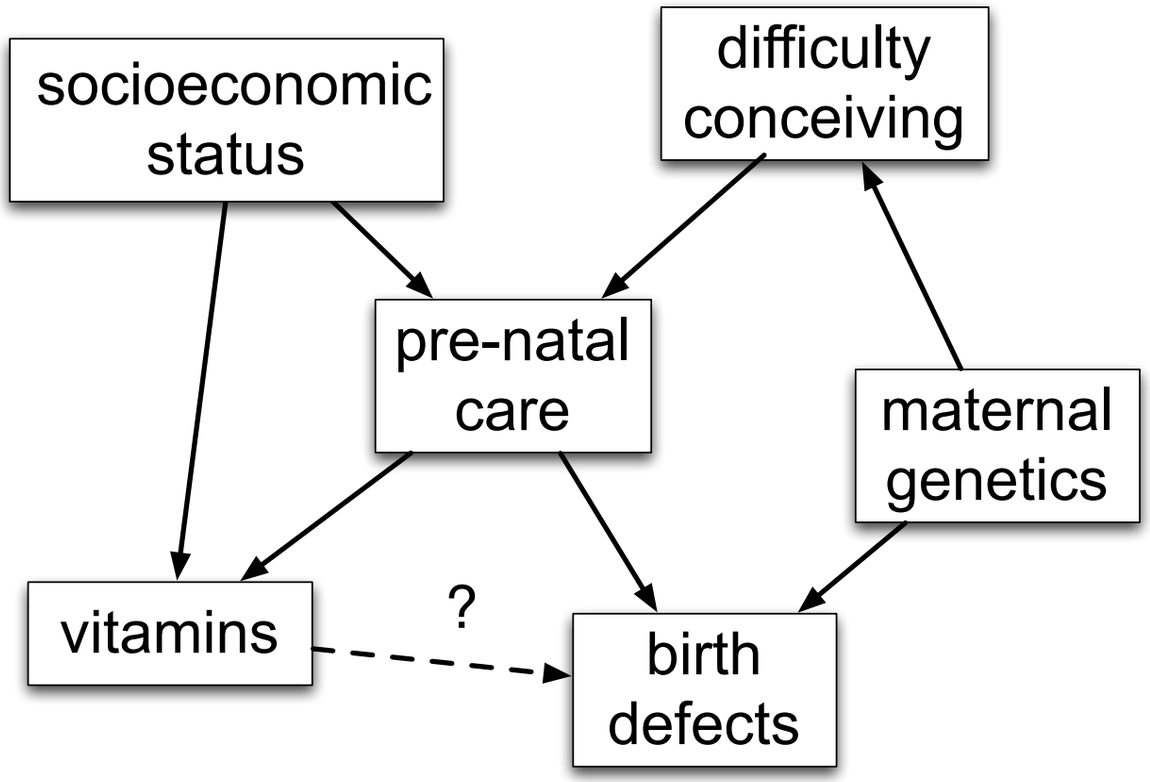
- Research question focuses on a single predictor
  - *Example*: Does maternal vitamin use reduce risk of birth defects?
- Ruling out confounding is key
- Minimizing PE is not critical, so over-fitting not a central issue
- Directed acyclic graphs (DAGs) are central in model selection for this goal

1. DAGs: Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology, 1999;10:37–48.

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
- Structural Causal Models (SCMs) provide and alternative & equivalent approach (Reference: Pearl J, Glymour M, Jewell NP. Causal Inference in Statistics: A Primer. 2016: Wiley.)

Maternal vitamin use and birth defects

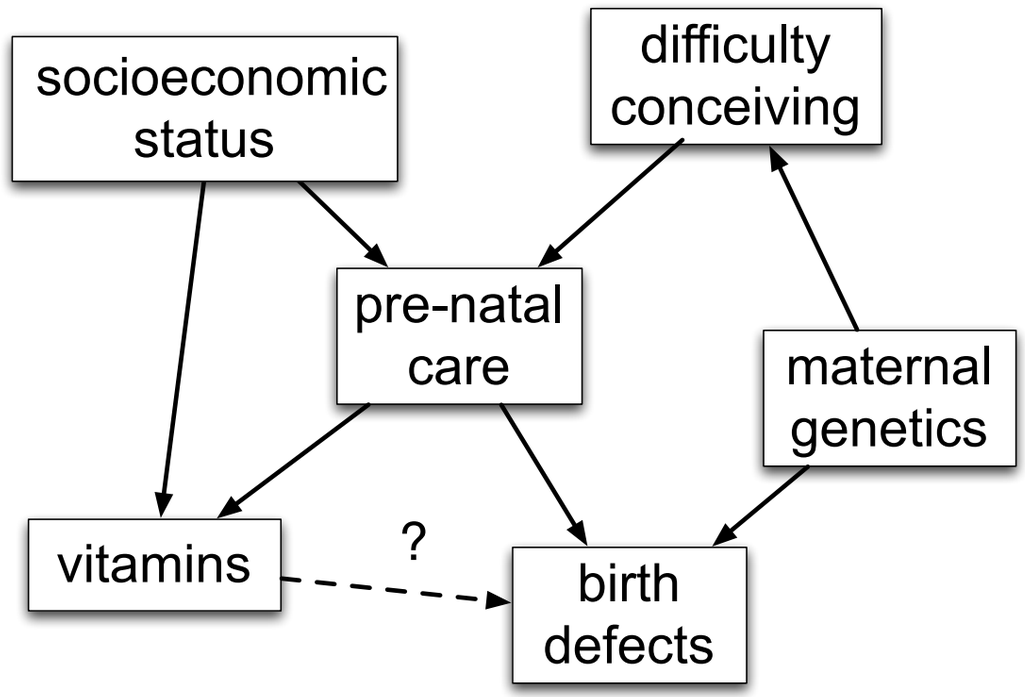


Assumptions about causal pathways:

- 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

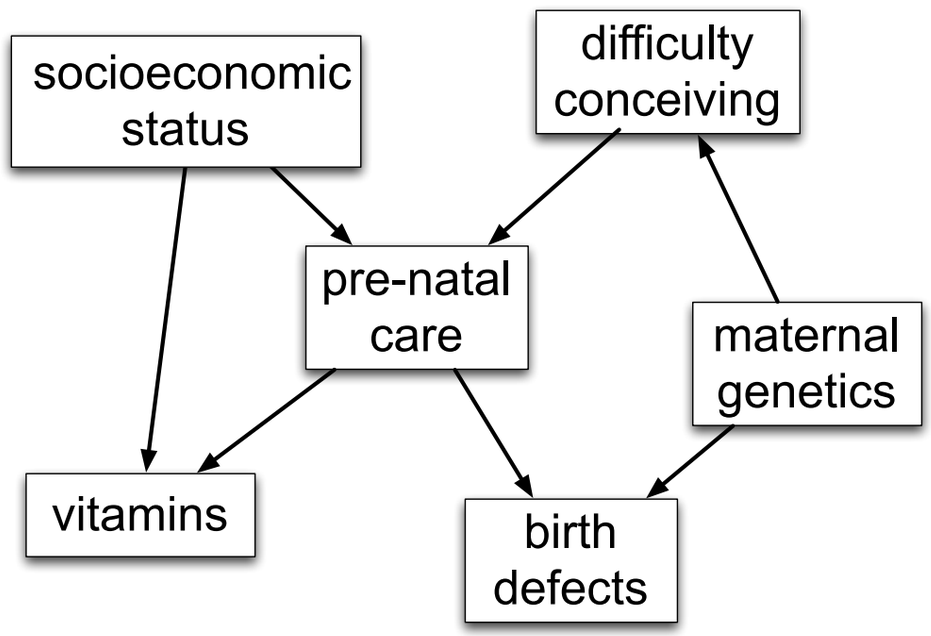
- SES affects birth defects only via PNC, vitamin use
- Difficulty conceiving affects vitamin use only through PNC
- No other common causes of vitamin use and birth defects
- In summary, no excluded confounders or causal links



Backdoor paths

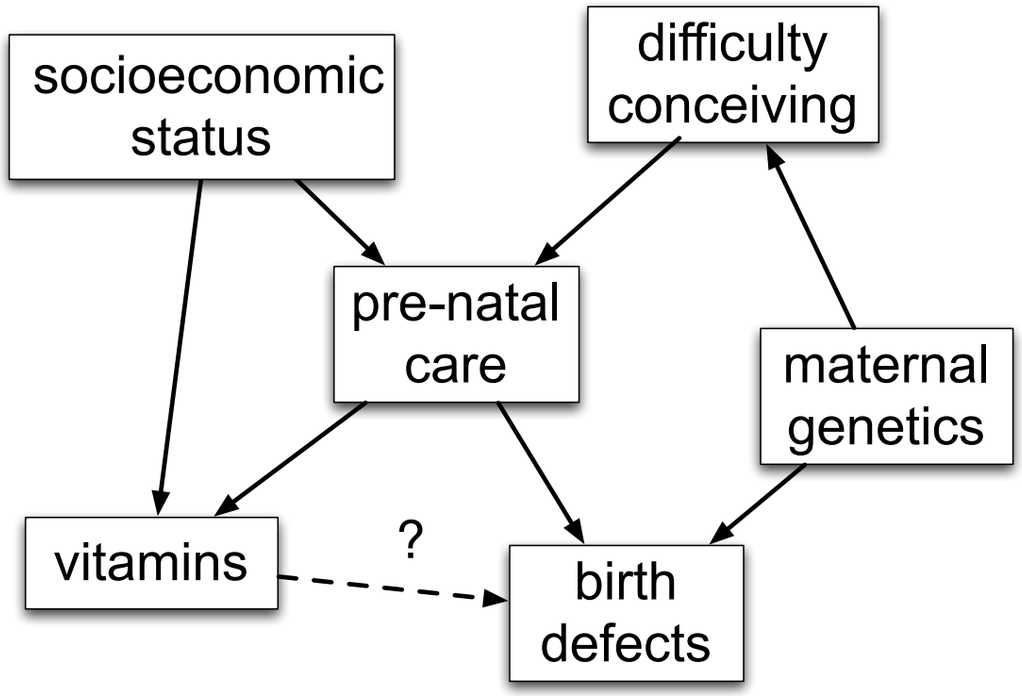
- After removing directed edge from vitamin use to birth defects, four backdoor paths remain between them:

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



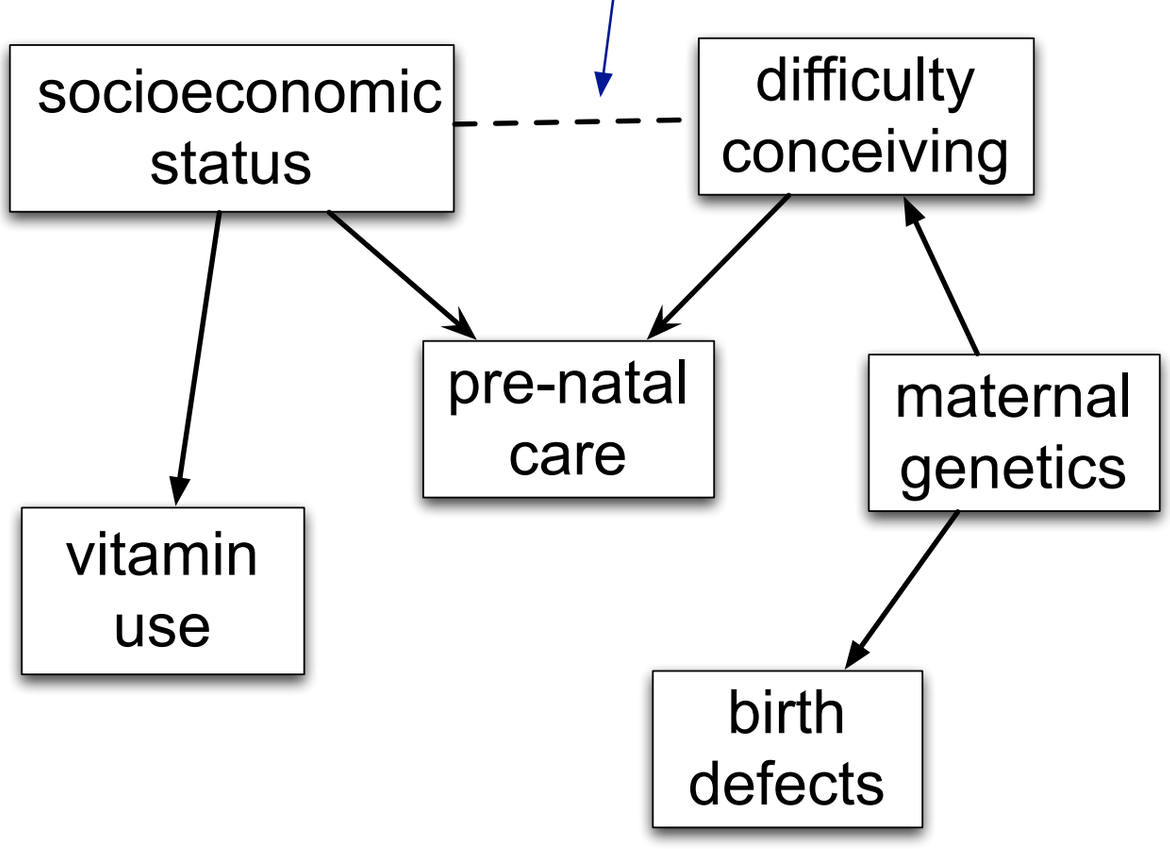
Colliders

- 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 via induced association of SES and difficulty conceiving

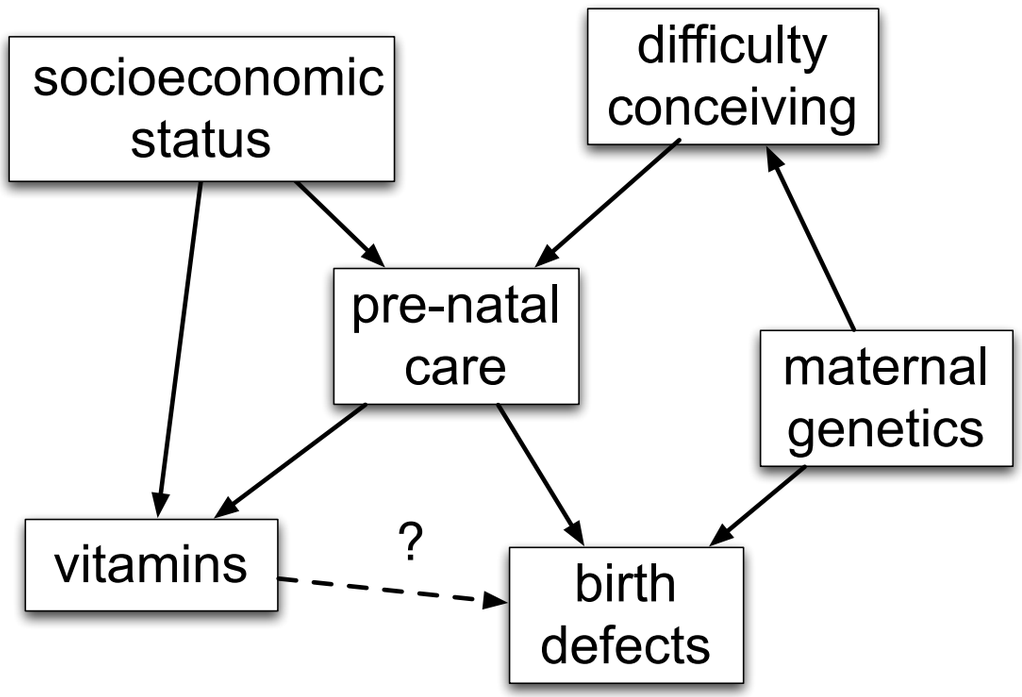


Open and blocked backdoor paths

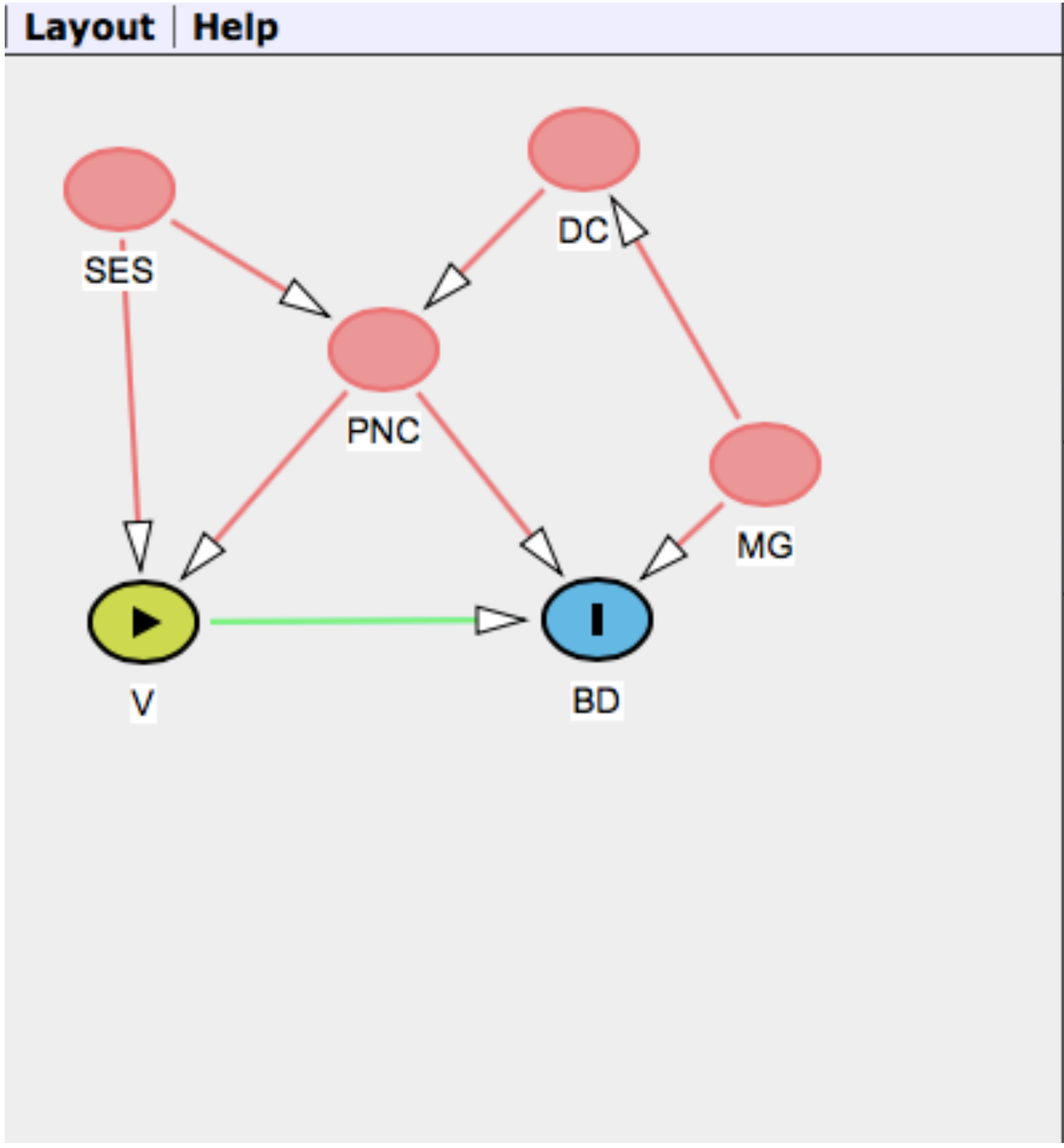
- If any of the backdoor paths between vitamin use and birth defects remains open, we would expect to find an association between them, even if there was no causal effect
- A backdoor path is blocked, provided we control for at least one non-collider on the path.
- A backdoor path including a collider 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



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.

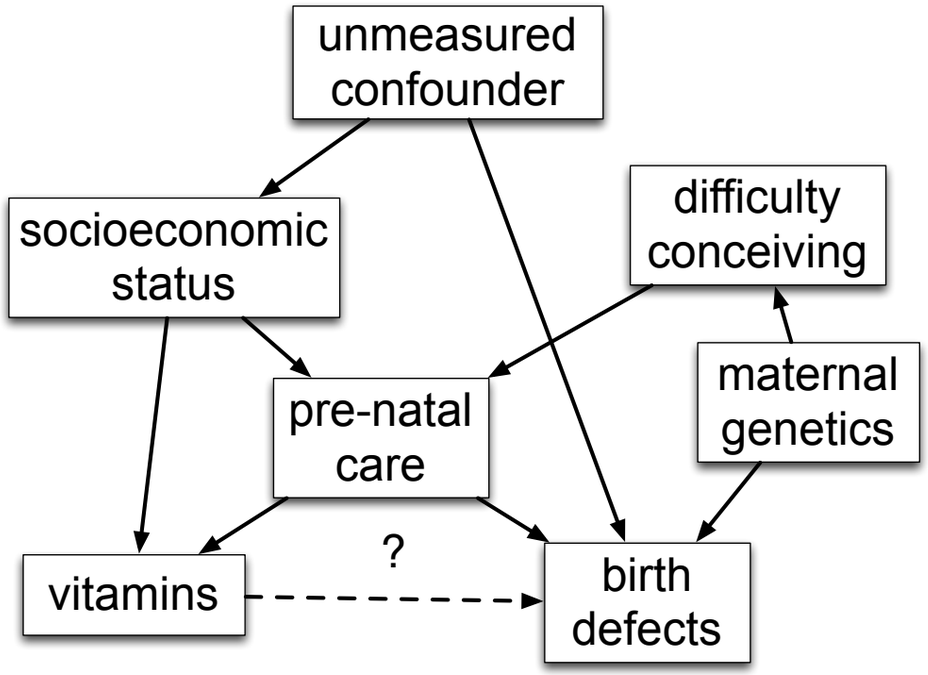
Testable implications

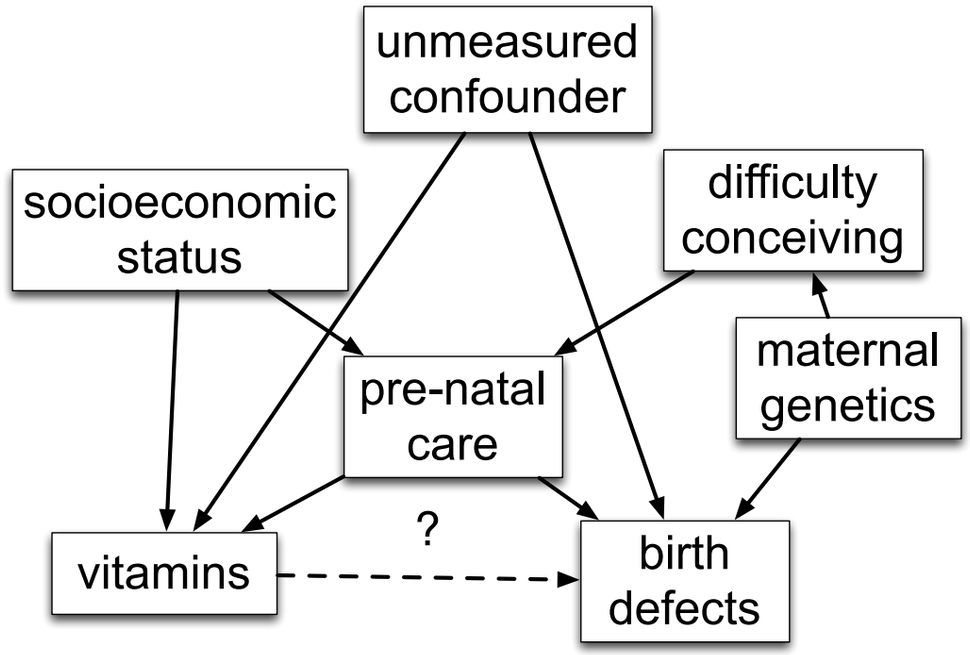
The model implies the following conditional independences:

- 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

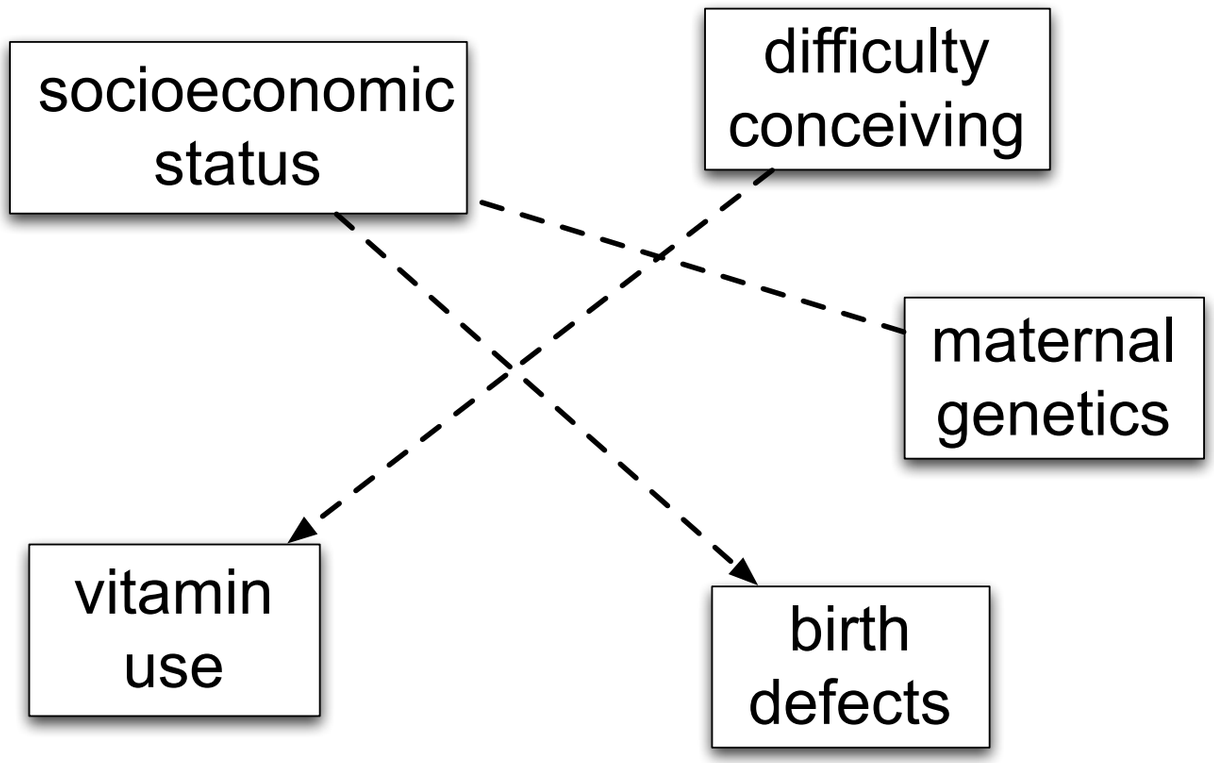




- DAGS can help determine vulnerability to unmeasured confounders
  - provided we have sufficient information about their effects
- In some cases, assessment of expected direction and magnitude of resulting bias possible via simulation or analytic approaches

Remaining issues to consider in confounding control

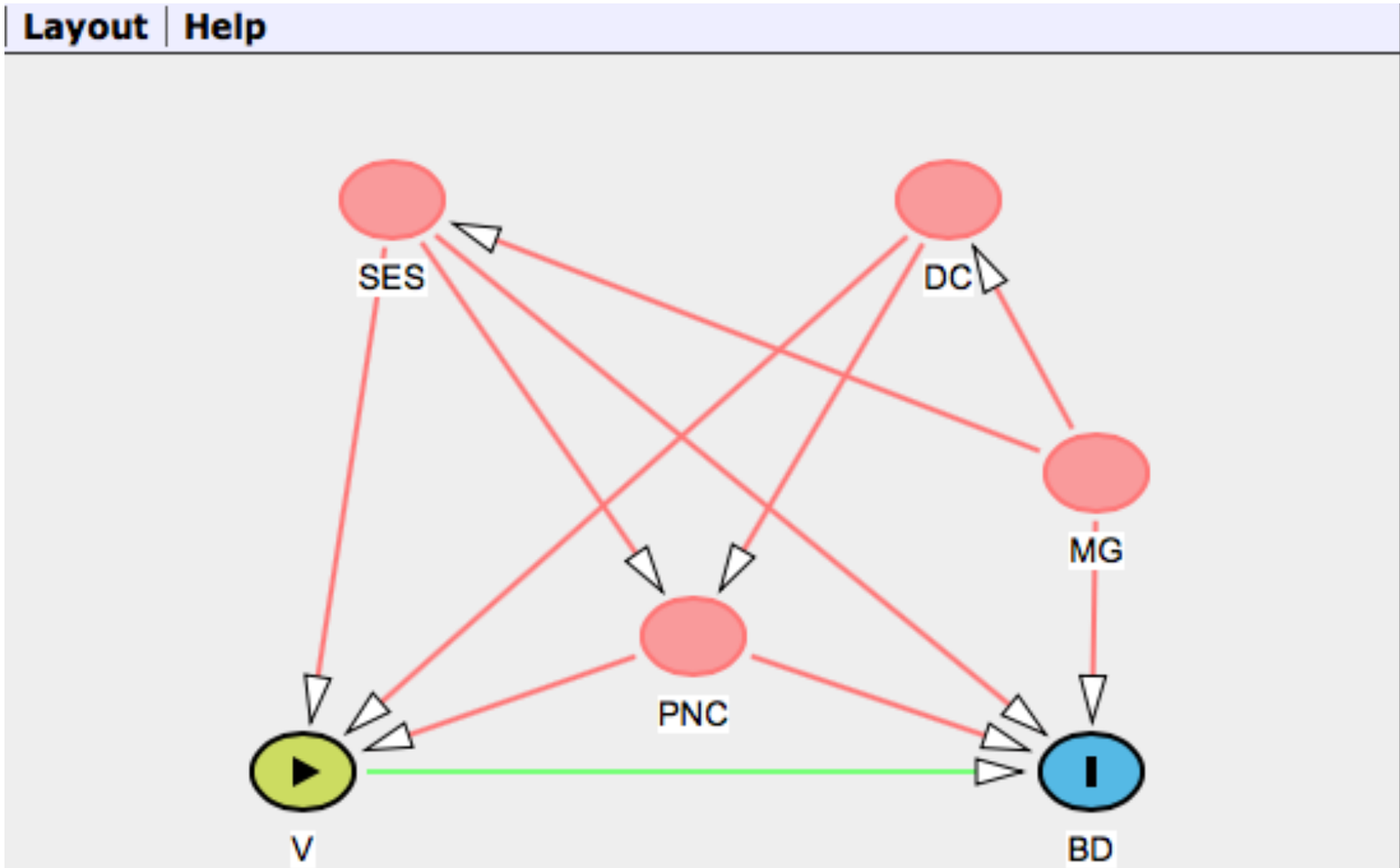
Other plausible causal pathways (edges) to consider:



- 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

Remaining issues to consider in confounding control

- MSAS options resulting from considering other plausible causal connections



Adjustment for total effect

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

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

Adjustment for direct effect

Total and direct effects are equal in this model.

Testable implications

The model implies the following conditional independences:

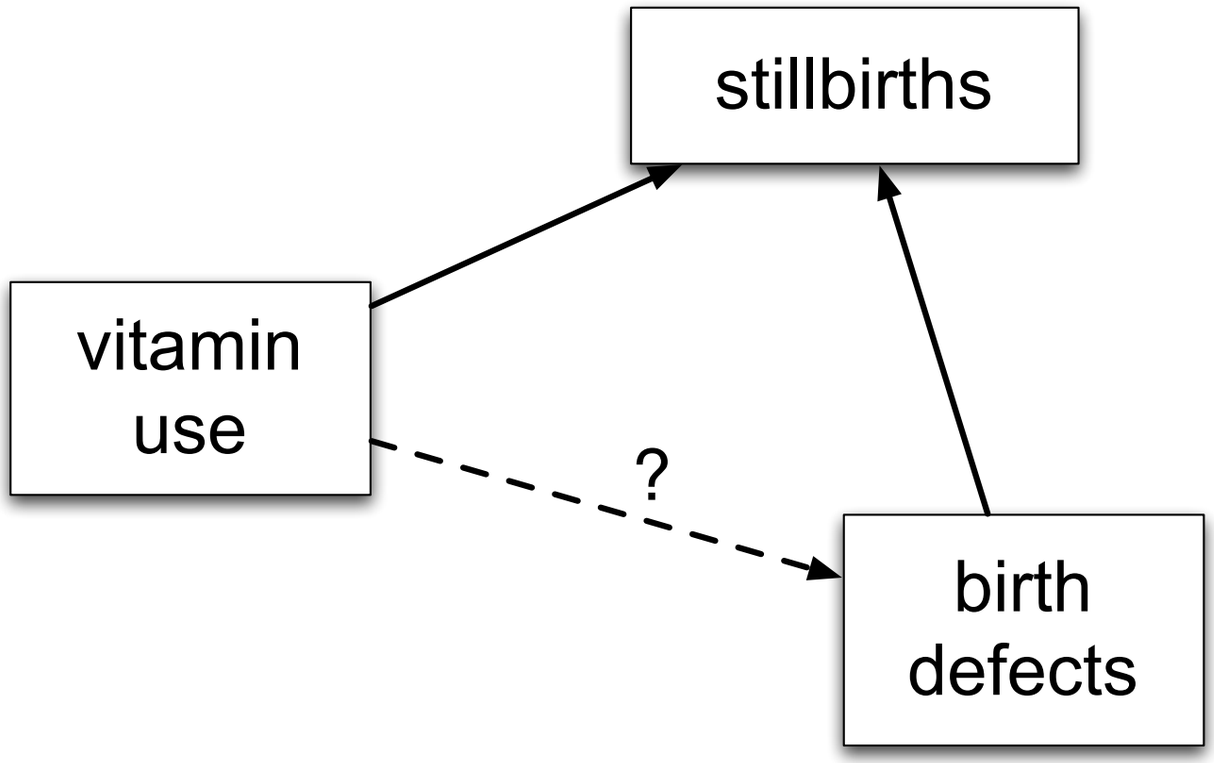
- V ⊥ MG | DC, SES
- BD ⊥ DC | MG, PNC, SES, V
- SES ⊥ DC | MG
- PNC ⊥ MG | DC, SES

Model text data

Other insights from colliders

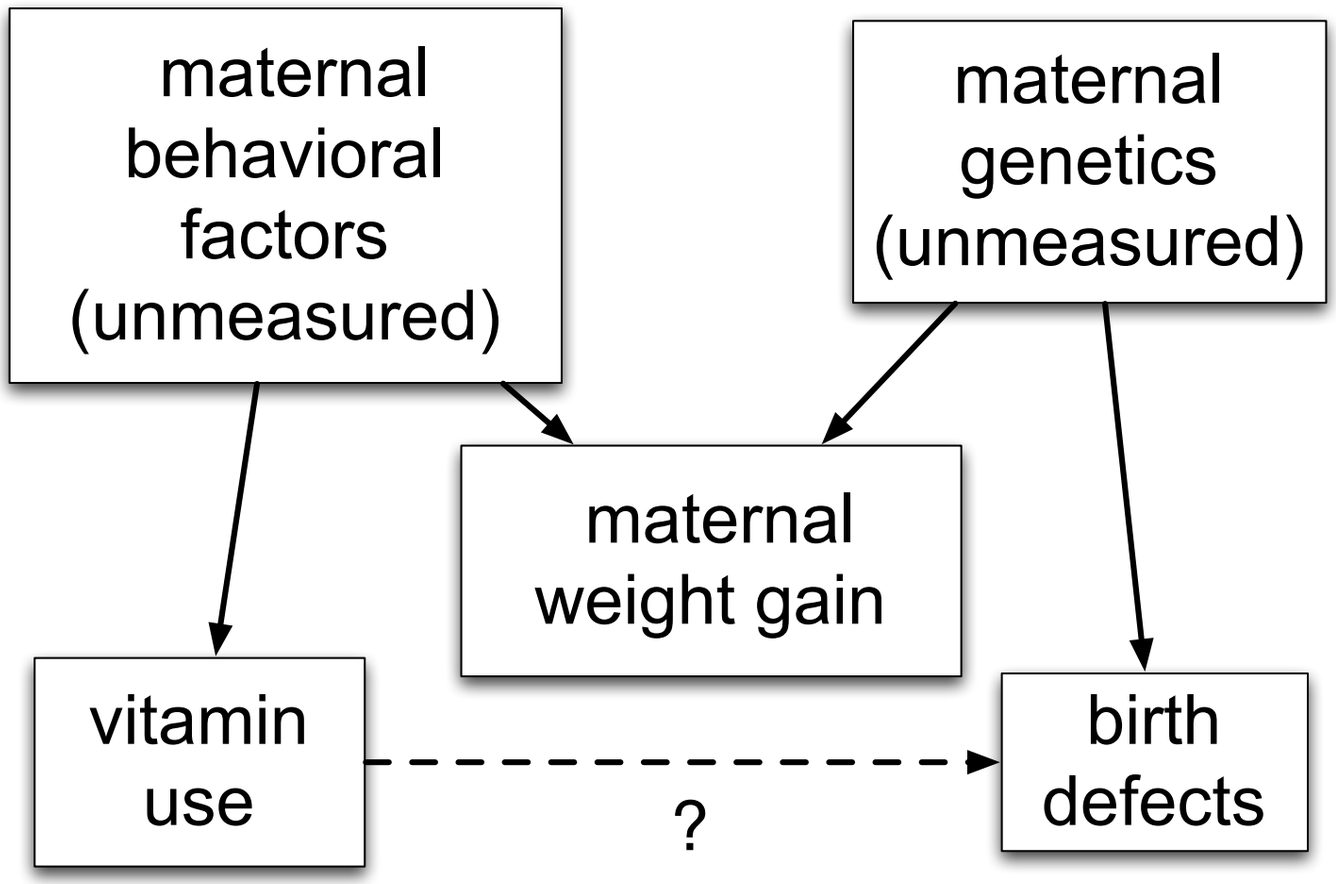
- Need to adjust for confounders of mediator/outcome relationships (lecture 4)
- Don’t adjust for a common effect of exposure and outcome
- Don’t adjust for a common effect of unmeasured causes of exposure and outcome

Common effect of exposure, outcome



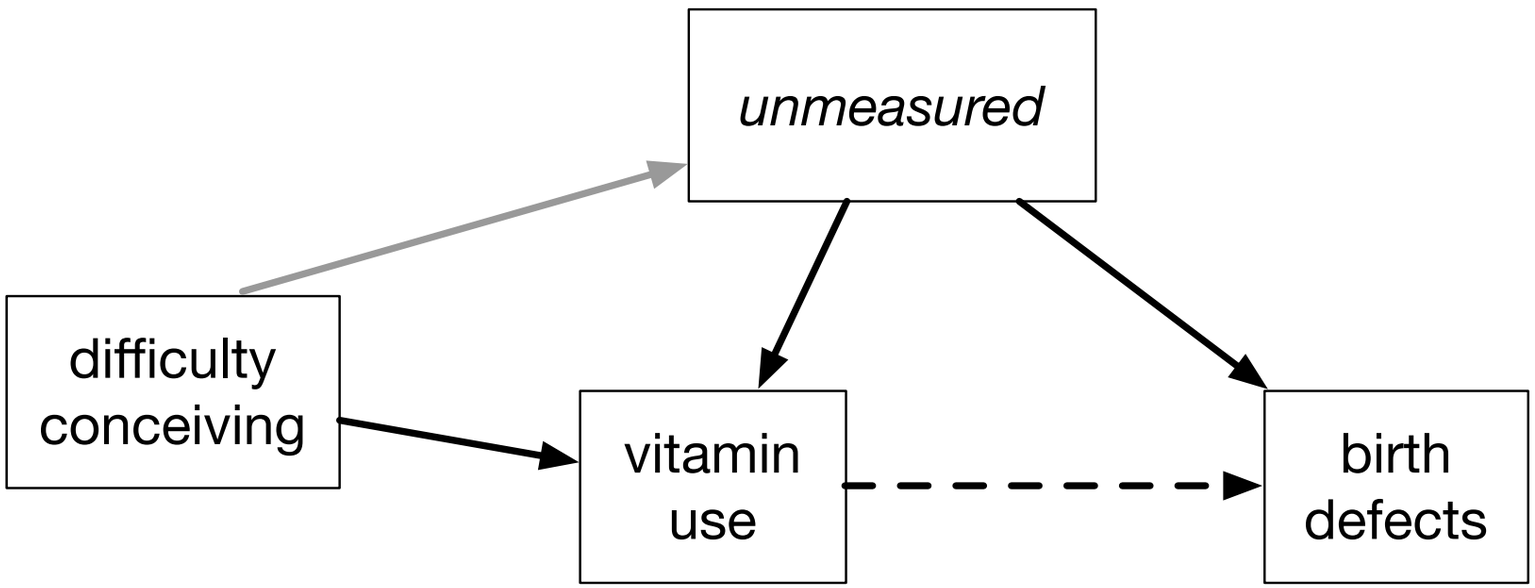
- Stillbirths a possible common effect of vit. use & birth def.
  - *Reference*: Cole SR, Platt RW, Schisterman EF, Chu H, Westreich D, Richardson D, Poole C. Illustrating bias due to conditioning on a collider. Int J Epidemiol. 2010;39:417-20.

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)

Control for variables strongly related to exposure in the presence of unmeasured confounders



- *Example*: if majority of effect of diff. conceiving (DC) acted through vitamin use, adjusting for DC may lead to *bias amplification* in the estimated effect of vitamin use
  - *References*:
    - Myers JA, Rassen JA, et al. Effects of adjusting for instrumental variables on bias and precision of effect estimates. Am J Epidemiol. 2011;174:1213-22.
    - Arah OA. Bias Analysis for Uncontrolled Confounding in the Health Sciences. Annu Rev Public Health. 2017;38:23-38.

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
- Be careful of control for "near-instrumental" variables
- 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) and adjustment affects estimate for primary predictor by > 5% or 10%, report and discuss implications

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

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_1 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
- A Stata command to automate this procedure is introduced in lab
- *Note*: Use of a change in coefficient criterion should respect DAG, and be used with appropriate caution

reference: Lee PH. Is a cutoff of 10% appropriate for the change-in-estimate criterion of confounder identification? J Epidemiol. 2014;24:161-7.

- Alternate approach for binary exposure/intervention variables:
  - develop propensity score to adjust for confounding, use shrinkage (e.g. lasso) regression to account for effects of multiple confounders
  - adjust for propensity score (e.g. via inverse weighting) in estimating marginal causal effect

Shortreed SM, Ertefaie A. Outcome-adaptive lasso: Variable selection for causal inference. Biometrics. 2017;73:1111-1122.

Limitations of DAGs

- No good way to represent interactions, no guidance about keeping or excluding them
  - interactions between primary predictor and important adjustment variables should be checked (text, section 10.2.3)
- 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 and adjust for identified confounders
  - use sensitivity analyses 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 (or shrinkage) 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 adjustment variables are irrelevant
  - effect estimates for adjustment variables are not main focus
  - over-fitting is primarily a problem for Goal 1, not Goal 2

Selection of predictors for adjustment in randomized trials

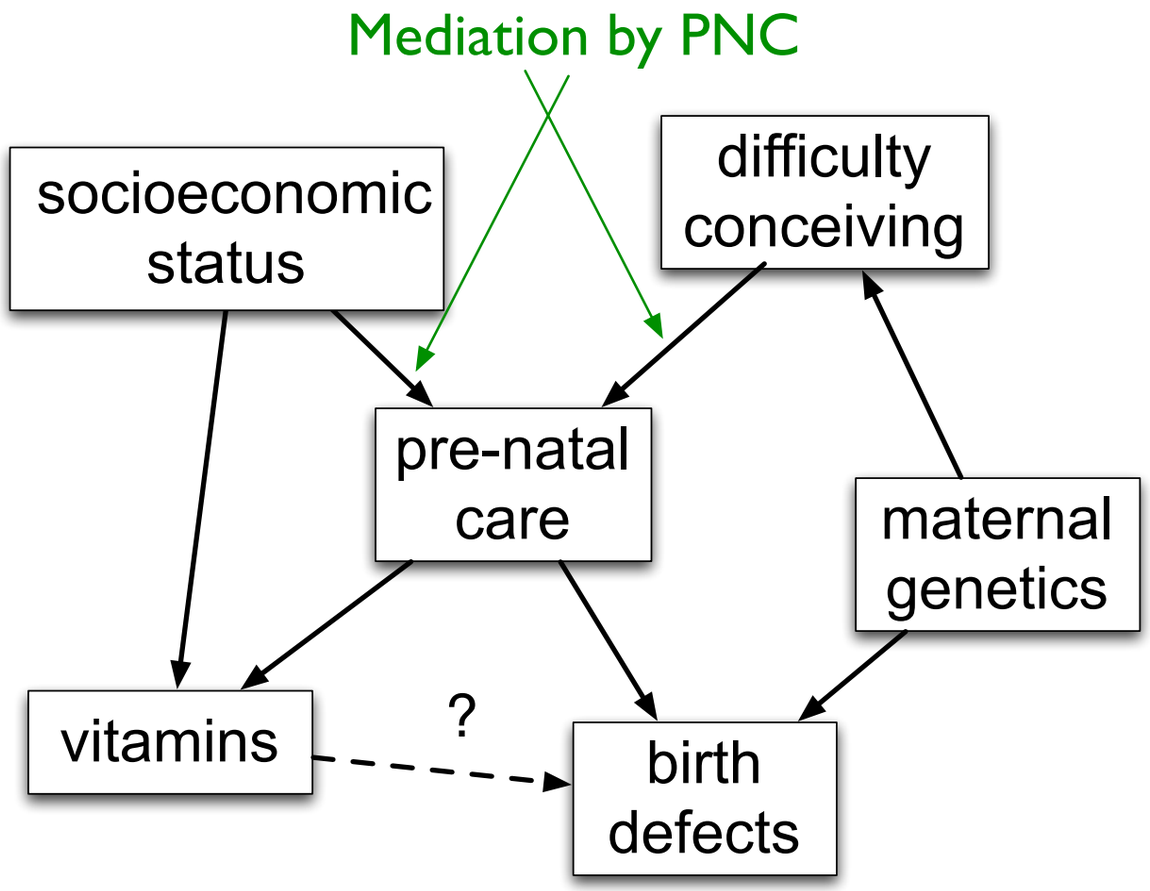
- Treatment assignment the predictor of primary interest
- Confounding not an issue, provided randomization succeeds
- Include covariates to
  - account for clustering by clinical center
  - reduce variance using pre-specified stratification variables
- Do *not* adjust for any post-randomization variables
- For binary outcomes, a standardized (marginal) estimate of treatment effect preferred (*reference*: Steingrimsson JA, Hanley DF, Rosenblum M. Improving precision by adjusting for prognostic baseline variables in randomized trials with binary outcomes, without regression model assumptions. Contemp Clin Trials. 2017;54:18-24.)

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
  - addressing potential confounding for all included factors difficult
  - mediation problematic if causal model misspecified (e.g. in assessing effects of SS & DC in example below)



Ref: Westreich D, Greenland S. The table 2 fallacy: presenting and interpreting confounder and modifier coefficients. *Am J Epidemiol.* 2013;177:292-8.

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
  - retain 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

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
  - retain 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

- Number of predictors
- Co-linearity
- Standard algorithms for predictor selection in regression models

Number of predictors to include in a model?

- Too many predictors can
  - degrade precision
  - in smaller datasets, swamp a real association
  - induce bias in estimates

Recommendations: number of predictors for regression models

- Use 10-15 observations/predictor (10 events per predictor for binary/ survival outcomes) as a cautionary flag
- If close to 10, check
  - high correlations between predictors
  - inflated SEs when a new covariate is added
  - inconsistency between *t*/Wald and likelihood ratio tests (logistic and Cox models)
  - gross inconsistency with smaller models
- If trouble is apparent:
  - use more stringent inclusion criterion
  - 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

- 2 predictors co-linear if their correlation is sufficient to substantially degrade precision of associated inferences about coefficients
- 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 individual 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
  - typically handled automatically by machine learning methods
- Goal 2: strong co-linearity between predictor of primary interest and a confounder in every MSAS is a problem
- Goal 3: see if the final model avoids the problem: e.g., one predictor clearly dominates, or both are independently predictive. Or, if it makes sense, create summary variables or select among them
- Co-linearity between adjustment variables (e.g. between confounders in goal 2, or between variables included in polynomial or spline terms): *not* a problem

Example: co-linearity between adjustment variables doesn't affect estimated effect of primary predictor

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

```
* regression of BMD on primary predictor estuse, age and weight (including
square of weight)

. recode estrogen (0=0) (1/2=1), g(estuse)

. reg bmd i.estuse age weight weight2

      Source |       SS           df       MS      Number of obs   =       278
-------------+----------------------------------   F(4, 273)       =     35.89
       Model |  1.79356154         4  .448390384   Prob > F        =    0.0000
    Residual |  3.41078229       273  .012493708   R-squared       =    0.3446
-------------+----------------------------------   Adj R-squared   =    0.3350
       Total |  5.20434383       277  .018788245   Root MSE        =    .11178

------------------------------------------------------------------------------
         bmd |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    1.estuse |   .0430352   .0140167     3.07   0.002     .0154407    .0706297
         age |  -.0038362    .001509    -2.54   0.012     -.006807   -.0008654
      weight |   .0162828   .0035358     4.61   0.000     .0093219    .0232438
     weight2 |   -.000079   .0000239    -3.30   0.001    -.0001261   -.0000319
       _cons |   .2820801   .1799751     1.57   0.118    -.0722354    .6363956
------------------------------------------------------------------------------
```

Example co-linearity between adjustment variables

```
. 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 i.estuse age weight cweight2
      Source |       SS           df       MS      Number of obs   =       278
-------------+----------------------------------   F(4, 273)       =     35.89
       Model |  1.79356153         4  .448390382   Prob > F        =    0.0000
    Residual |   3.4107823       273  .012493708   R-squared       =    0.3446
-------------+----------------------------------   Adj R-squared   =    0.3350
       Total |  5.20434383       277  .018788245   Root MSE        =    .11178
------------------------------------------------------------------------------
         bmd |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    1.estuse |   .0430352   .0140167     3.07   0.002     .0154407    .0706297
         age |  -.0038362    .001509    -2.54   0.012     -.006807   -.0008654
      weight |   .0056033   .0005778     9.70   0.000     .0044657    .0067408
    cweight2 |   -.000079   .0000239    -3.30   0.001    -.0001261   -.0000319
       _cons |   .6430895   .1332308     4.83   0.000     .3807993    .9053798
------------------------------------------------------------------------------
```

Centering reduces collinearity

Results for estrogen use *unchanged*, effect of weight more precisely estimated

Model selection algorithms for regression

- Procedures for selecting predictors on statistical grounds (e.g. *p*-values), some implemented in Stata:
  - univariate screening; change-in-estimate criterion; backwards, forwards, and stepwise selection; all subsets
- Epidemiologists and statisticians are often uncomfortable with these methods
- For goal 1, screening with cross-validation or many modern machine learning approaches may work better than alternatives above
- For goals 2 and 3, DAG-based procedures preferable

Backward 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 account for 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 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* version also allows variables to be deleted
- Advantages: Avoids problem with too-large initial model
- Disadvantages:
  - results may be affected by negative confounding

Pitfalls of automated model selection

- Only a subset of possible models considered
- *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 may overestimate effects in prediction applications (because they withstand the selection testing criterion), especially for predictors with weaker effects
- Only complete cases used in automated procedures (i.e., observations with missing values on any predictors in initial list are excluded)

Ref: Steyerberg EW, Eijkemans MJ, Habbema JD. Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. J Clin Epidemiol. 1999;52(10):935-42.

A priori model selection for Goal 2 (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 selection algorithms, check sensitivity to
  - model selection procedure (forwards, backwards, stepwise)
  - number of predictors retained / retention criterion
  - a priori variable choices (e.g., between collinear variables)

Summary Recommendations

- *Goal 1* – prediction: select model using aggressive search plus cross-validation/bootstrapping, 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 unless it occurs between a predictor of primary interest and a non-ignorable confounder
  - in prediction models, can be handled using methods like cross-validation
- Use sensitivity analyses!