University of Antwerp
Faculty of Social Sciences

DAGs and PP

or commonly know as:
"a bit more transparent way to state
your research assumptions and questions"

Jose Rivera
(Josema, for the friends)

May 5, 2022

What are we going to talk about? I

1. About research
   - A typical scientific lab
   - Research hypothesis production
2. DAGs and PP
3. Example cases
   - Experimental design: the panacea
   - Simulation conventions
   - Fork bias: spurious relationships
   - Fork bias: masked relationships (a)
   - Fork bias: masked relationships (b)
   - Fork bias: multicollinearity

What are we going to talk about? II

- No more fork bias: neutral control
- Pipe bias: precision parasite
- Pipe bias: post-treatment
- Pipe bias: masked relationships
- Pipe/Fork bias: bias amplification
- Collider bias: M-bias
- Descendant bias: case control

4 Concluding remarks

5 Do you wanna know more???

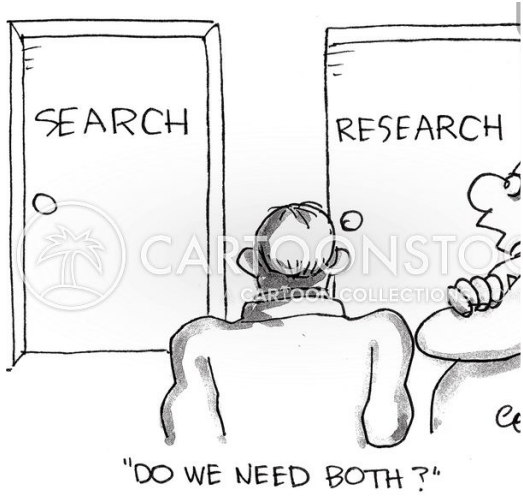
1. About research

A typical scientific lab

A typical scientific lab[1]

What is needed?

1. Quality of theory
2. Quality of data
3. Reliable procedures and code
4. Quality of data analysis
5. Documentation
6. Reporting



[1] McElreath [12], lecture 20 and McElreath [13], chapter 17

A typical scientific lab

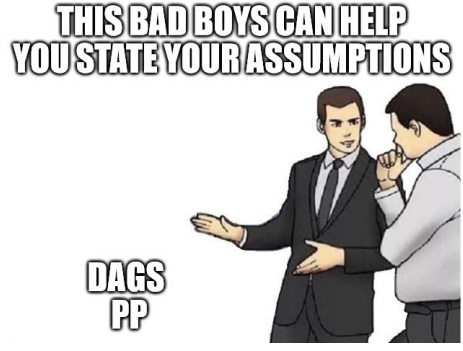
What we “normally” focus on?

1. Quality of theory
2. Quality of data
3. Reliable procedures and code
4. Quality of data analysis
5. Documentation
6. Reporting

A typical scientific lab

What can be improved?
(with DAGs and PP)

1. Quality of theory
2. Quality of data
3. Reliable procedures and code
4. Quality of data analysis
5. Documentation
6. Reporting



1. About research

Research hypothesis production

Research hypothesis production

Well known challenges[a]

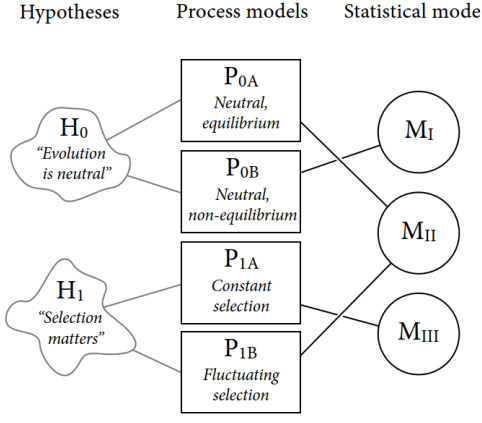
- Insufficient data
- Wrong population
- Measurement error
- Selection bias
- Confounding

[a] Hernán [8], lesson 4

Research hypothesis production

but we should not forget[a]

- No one-to-one relationship exists between our process models and statistical models,
- Nor between our hypothesis and a process models

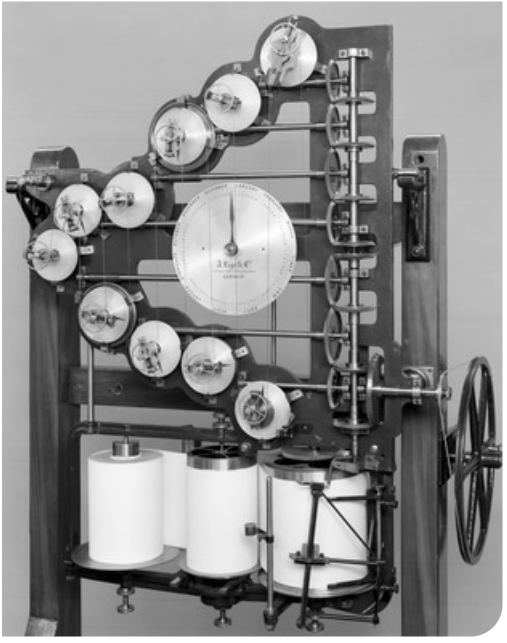


[a] Figure 1.2 reproduced from chapter 1 McElreath [13]

Research hypothesis production

and also

- statistical models are just "machines to find association", not a reliable reflection of the theory (I can prove it!!).



Research hypothesis schematics[2]

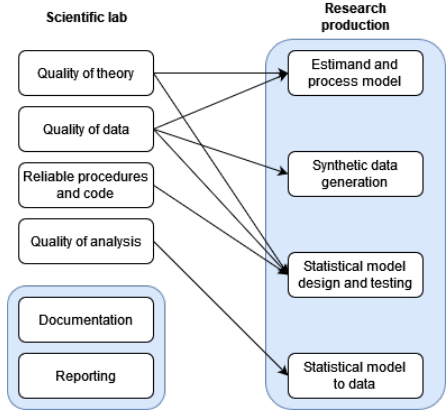
a. Estimand and process model
b. Synthetic data generation
c. Statistical model design and testing
d. Apply statistical model to data

[2] McElreath [13], lecture 20, Pearl [16]. Follow Fogarty et al. [6] on item (c).

Research hypothesis schematic

Where does it match with the previous?

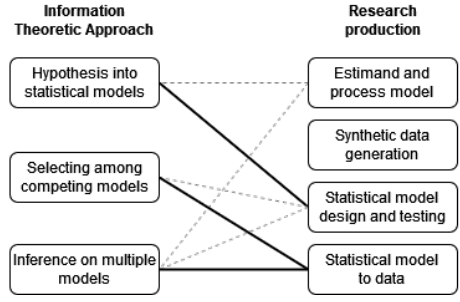
a. Estimand and process model maps 1 (theory) and 2 (data) to a heuristic model.
b. Synthetic data generation maps 2 (data) to an idealized data.
c. Statistical model design and testing maps 1 (theory), 2 (data), and 3 (reliable code) to an statistical model.
d. Apply statistical model to data maps 4 (analysis) onto a result.



Where does the ITA fit?

Information Theoretic Approach (ITA) is framework to select among competing models [1, 3]:

1. Hypothesis into statistical models,
   (how about a process model?)
2. Select among competing models,
   (do the code works as intended?)
3. Make inferences based on one or multiple models.
   (do the code works as intended?, are there variables that can bias our results?)



2. DAGs and PP

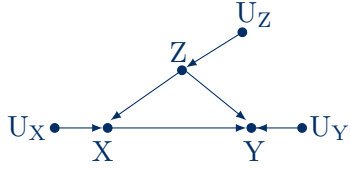
DAGs and PP

- Directed acyclic graphs (DAGs), are a type of structural causal model (SCM) [15, 4]
- DAGs can be represented by a structural model, and its associated causal diagram[a].
- we put distributional assumptions to the structural model through probabilistic programming (PP) [10]. (more in part 3)

[a] reproduced from Cinelli et al. [4].

$$\mathrm{M} = \begin{cases} \mathrm{Z} \leftarrow \mathrm{f_Z}(\mathrm{U_z}) \\ \mathrm{X} \leftarrow \mathrm{f_X}(\mathrm{Z}, \mathrm{U_x}) \\ \mathrm{Y} \leftarrow \mathrm{f_Y}(\mathrm{X}, \mathrm{Z}, \mathrm{U_y}) \\ \mathrm{U} \sim \mathrm{P}(\mathbf{U}) \end{cases}$$

(a) structural model



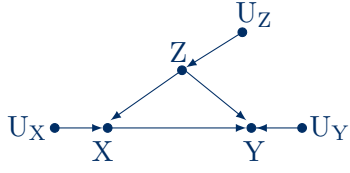
(b) causal diagram

DAGs and PP

- $\mathbf{V} = \{Z, X, Y\}$ are called endogenous variables.
- $\mathbf{U} = \{U_Z, U_X, U_Y\}$ are called exogenous variables.
  (drawn when strictly required)
- $\mathbf{F} = \{f_Z, f_X, f_Y\}$ are called structural equations.

$$M = \begin{cases} Z \leftarrow f_Z(U_Z) \\ X \leftarrow f_X(Z, U_X) \\ Y \leftarrow f_Y(X, Z, U_Y) \\ U \sim P(\mathbf{U}) \end{cases}$$

(a) structural model



(b) causal diagram

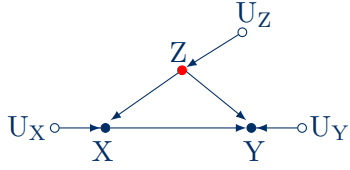
DAGs and PP

Causal diagram conventions [4],

- black nodes are observed variables.
- white nodes are unobserved variables.
- red nodes are variables for which we will decide its inclusion or not.

$$M = \begin{cases} Z \leftarrow f_Z(U_Z) \\ X \leftarrow f_X(Z, U_X) \\ Y \leftarrow f_Y(X, Z, U_Y) \\ U \sim P(\mathbf{U}) \end{cases}$$

(a) structural model



(b) causal diagram

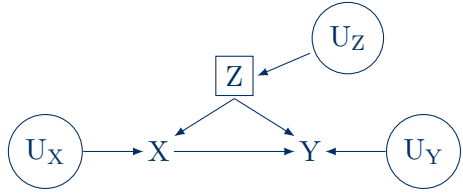
DAGs and PP

Other causal diagram conventions,

- no circle nodes are observed variables.
- circled nodes are unobserved variables.
- squared nodes are variables for which we will decide its inclusion or not.

$$M = \begin{cases} Z \leftarrow f_Z(U_Z) \\ X \leftarrow f_X(Z, U_X) \\ Y \leftarrow f_Y(X, Z, U_Y) \\ U \sim P(\mathbf{U}) \end{cases}$$

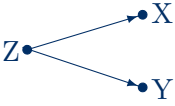
(a) structural model



(b) causal diagram

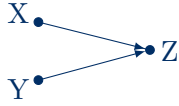
The benign case of DAG elementals

For everything can be depicted with them

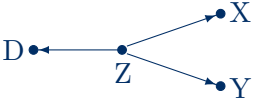


(a) fork

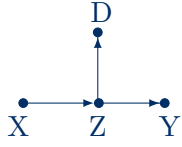
(b) pipe



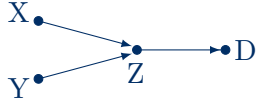
(c) collider



(d) descendant on fork



(e) descendant on pipe



(f) descendant on collider

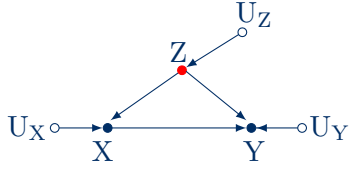
About D-separation

Causal graph theory [14, 15, 17, 18, 19],

1. descendant (child, grandchild), parent (grandparent).
   (path specific)
2. paths (directional, non-directional).
3. paths are blocked or open according to the D-separation rules.
   (also path specific)
4. there are only four (4) D-separation rules.

$$M = \begin{cases} Z \leftarrow f_Z(U_Z) \\ X \leftarrow f_X(Z, U_X) \\ Y \leftarrow f_Y(X, Z, U_Y) \\ U \sim P(\mathbf{U}) \end{cases}$$

(a) structural model



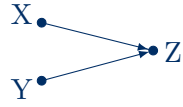
(b) causal diagram

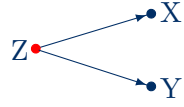
About D-separation

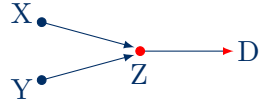
The D-separation (Directional) rules [8],

1. If no variables being conditioned on, a path is blocked if and only if, two arrowheads on the path collide at some variable on the path.
2. Any path that contains a noncollider that has been conditioned on, is blocked (backdoor path)[a].
3. A collider that has been conditioned on does not block a path.
4. A collider that has a descendant that has been conditioned on does not block a path.

[a] there is also a front-door path (if you wonder).



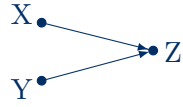


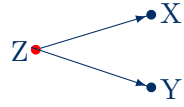


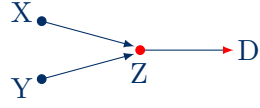
About D-separation

The D-separation rules implications,
(independent of distributional assumptions)

1. $X \perp\!\!\!\perp Y \implies$
   $P(X, Y) = P(X) \cdot P(Y)$
2. $X \perp\!\!\!\perp Y|\ Z \implies$
   $P(X, Y|Z) = P(X|Z) \cdot P(Y|Z)$
   (same for fork or pipe)
3. $X \not\perp\!\!\!\perp Y|\ Z \implies$
   $P(X, Y|Z) \neq P(X|Z) \cdot P(Y|Z)$
4. $X \not\perp\!\!\!\perp Y|\ D \implies$
   $P(X, Y|D) \neq P(X|D) \cdot P(Y|D)$

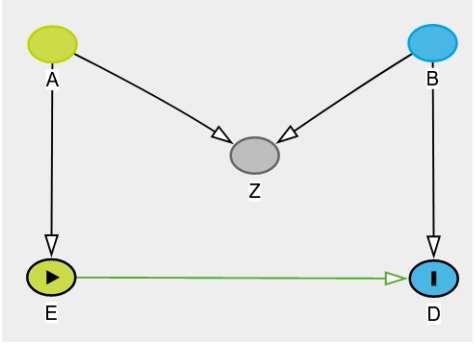






Oh DAGitty!! mijn vriendin

- browser (R package) environment for creating, editing, and analyzing causal diagrams [20].
- available online: http://dagitty.net
- But there are more fish in the sea: http://www.causalfusion.net [2] (b**** better have my $$$)



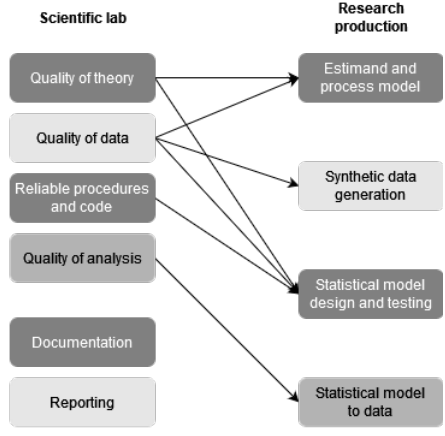
Where do DAGs and PP fit?

starts with:

- A clear definition of the estimand and process model (assumptions).
- An improved the reliability of your procedures.
- As a documentation procedure.

and leads to:

- A sound analysis, and result (even when we cannot have an answer to our question)
- An improved planning to get data.



3. Example cases

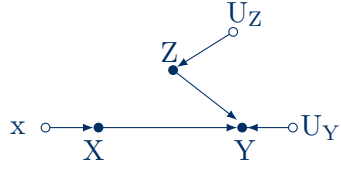
Experimental design: the panacea

Experimental design[3]

- Purpose: to control all factors responsible for the outcome's variation.
  (understand the system)
- It is modeled by modifying the structural model (and causal diagram).

$$M = \begin{cases} Z \leftarrow f_Z(U_Z) \\ X \leftarrow f_X(x) \\ Y \leftarrow f_Y(X, Z, U_Y) \\ U \sim P(\mathbf{U}) \end{cases}$$

(a) structural model



(b) causal diagram

[3] Cinelli et al. [4], appendix A (p. 15)

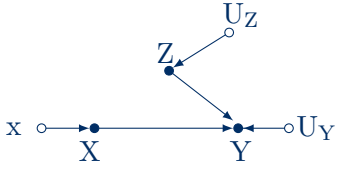
Experimental design

- intervention on X can be written in do-calculus[a] as: $P(\mathbf{V} \mid do(X = x))$.
- remember:
  $\mathbf{V} = \{Z, X, Y\}$,
  $\mathbf{U} = \{U_Z, U_X, U_Y\}$, and
  $\mathbf{F} = \{f_Z, f_X, f_Y\}$.

[a] an appropriate treatment can be found with the usual suspects [14, 15, 17, 18])

$$M = \begin{cases} Z \leftarrow f_Z(U_Z) \\ X \leftarrow f_X(x) \\ Y \leftarrow f_Y(X, Z, U_Y) \\ U \sim P(\mathbf{U}) \end{cases}$$

(a) structural model



(b) causal diagram

Effects of interest

two types of effects,

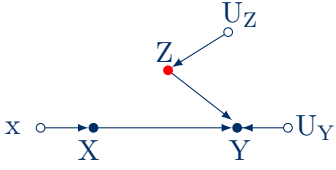
1. Average causal effect:
   $\text{ACE}(x) = \text{E}[Y|\text{do}(x+1)] - \text{E}[Y|\text{do}(x)]$
2. Controlled direct effect:
   $\text{CDE}(x, z) = \text{E}[Y|\text{do}(x+1), \text{do}(z)] - \text{E}[Y|\text{do}(x), \text{do}(z)]$

points to consider:

- CDE takes a particular relevance with observational data.
- There is also a distinction between total effect and direct effect.

$$M = \begin{cases} Z \leftarrow f_Z(U_Z) \\ X \leftarrow f_x(x) \\ Y \leftarrow f_Y(X, Z, U_Y) \\ U \sim P(\mathbf{U}) \end{cases}$$

(a) structural model



(b) causal diagram

3. Example cases

Simulation conventions

Simulation conventions

one way to defined it,

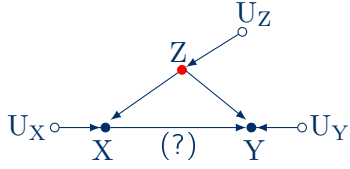
$$
\begin{aligned}
Z &= U_Z && ; U_Z \sim N(0, \sigma_Z) \\
X &= \beta_Z Z + U_X && ; U_X \sim N(0, \sigma_X) \\
Y &= \beta_Z Z + \beta_X X + U_Y && ; U_Y \sim N(0, \sigma_Y)
\end{aligned}
$$

a more succinct way,

$$
\begin{aligned}
Z &\sim N(0, \sigma_Z) \\
X &\sim N(\beta_Z Z, \sigma_X) \\
Y &\sim N(\beta_Z Z + \beta_X X, \sigma_Y)
\end{aligned}
$$

$$
M = \begin{cases}
Z \leftarrow f_Z(U_Z) \\
X \leftarrow f_X(Z, U_X) \\
Y \leftarrow f_Y(Z, X, U_Y) \\
U \sim P(\mathbf{U})
\end{cases}
$$

(a) structural model



(b) causal diagram

3. Example cases

Fork bias: spurious relationships

Spurious relationships[4]

also known as,

- spurious association
- confounder
- an instance of fork bias

research question,

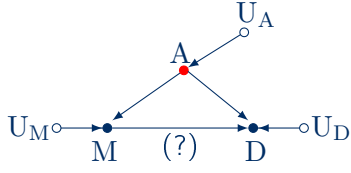
- Does M has a (direct) effect on D?

variables,

- A, median age at marriage
- M, marriage rate
- D, divorce rate

$$M = \begin{cases} A \leftarrow f_A(U_A) \\ M \leftarrow f_M(A, U_M) \\ D \leftarrow f_D(A, M, U_D) \\ U \sim P(\mathbf{U}) \end{cases}$$

(a) structural model



(b) causal diagram

[4] McElreath [12], chapter 05 (p. 125)

Simulation setting

```r
# sim
A = rnorm( 100 )
M = rnorm( 100 , mean=-1*A )
D = rnorm( 100 , mean=-1*A + 0*M )
d = data.frame(A=A,M=M,D=D)
```

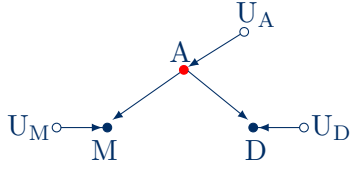
(c) R code

$$M = \begin{cases} A \leftarrow f_A(U_A) \\ M \leftarrow f_M(A, U_M) \\ D \leftarrow f_D(A, U_D) \\ U \sim P(\mathbf{U}) \end{cases}$$

(a) structural model

Implications,

- $M \not\perp\!\!\!\perp D$
- $M \perp\!\!\!\perp D \mid A$

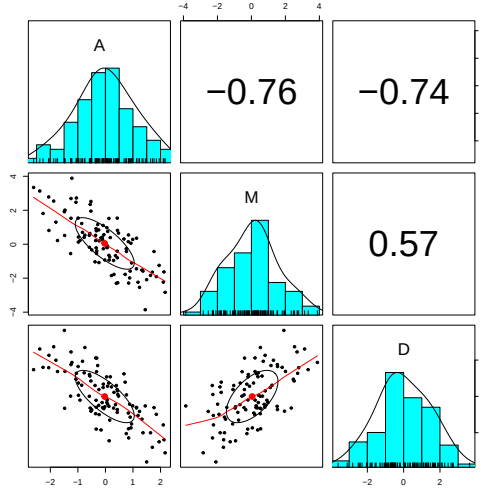


(b) causal diagram

“Eyeballing” analysis

based on correlation analysis,

- $\text{cor}(\text{A}, \text{D}) < 0$ and $\text{cor}(\text{M}, \text{D}) > 0$ goes in line of our “rudimentary” understanding of the data.
- why there is $\text{cor}(\text{M}, \text{D}) > 0$?
  (hint: univariate correlation)
- we include M as a covariate in our statistical model
  (is our research hypothesis)



Regression, regression!!

based on statistical analysis,

- we have two different stories, (which one is the “truth”?)

```
> summary(lm(D ~ M, data=d)) # spurious relation

Call:
lm(formula = D ~ M, data = d)

Residuals:
     Min       1Q   Median       3Q      Max
-2.80012 -0.90447 -0.03866  0.80220  2.82970

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.23298    0.12412  -1.877   0.0635 .
M            0.40233    0.08986   4.477 2.04e-05 ***
> summary(lm(D ~ A + M, data=d)) # controlled relation

Call:
lm(formula = D ~ A + M, data = d)

Residuals:
     Min       1Q   Median       3Q      Max
-2.27295 -0.68174  0.03781  0.78885  2.95320

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.18854    0.09871  -1.910   0.0591 .
A           -1.03121    0.13483  -7.648 1.49e-11 ***
M           -0.06134    0.09362  -0.655   0.5139
```

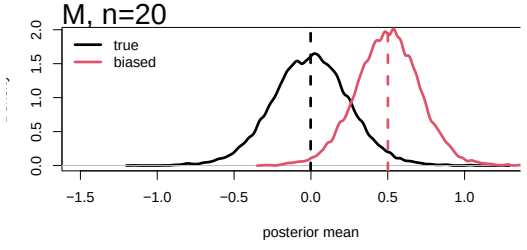
I'll get more data!!

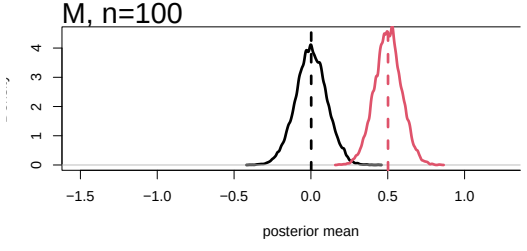
imagine we can continue sampling,

- top: $10,000$ samples $n = 20$
- bottom: $10,000$ samples $n = 100$

under the incorrect model,
the larger the sample size,

- the more certain you are about your biased estimates
  (the winner's curse)





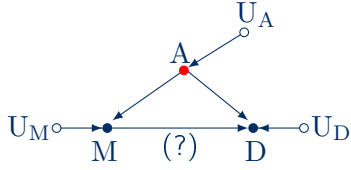
The dream team!!

based on DAG and statistical model,

- the 2nd D-separation rule requires you to control any noncollider to block the backdoor path, i.e. $M \perp\!\!\!\perp D \mid A$
- conditioning on A we can find, $E[D|do(m)] = E[\, E[D|M = m, A]\,]$ (law of total expectation)
- then we can find the $ACE(m) = E[D|do(m+1)] - E[D|do(m)]$ (Frisch-Waugh-Lovell theorem)

$$M = \begin{cases} A \leftarrow f_A(U_A) \\ M \leftarrow f_M(A, U_M) \\ D \leftarrow f_D(A, M, U_D) \\ U \sim P(\mathbf{U}) \end{cases}$$

(a) structural model



(b) causal diagram

the dream team!!

based on DAG and statistical analysis,

- the less biased model is the second, (assuming our DAG is true)

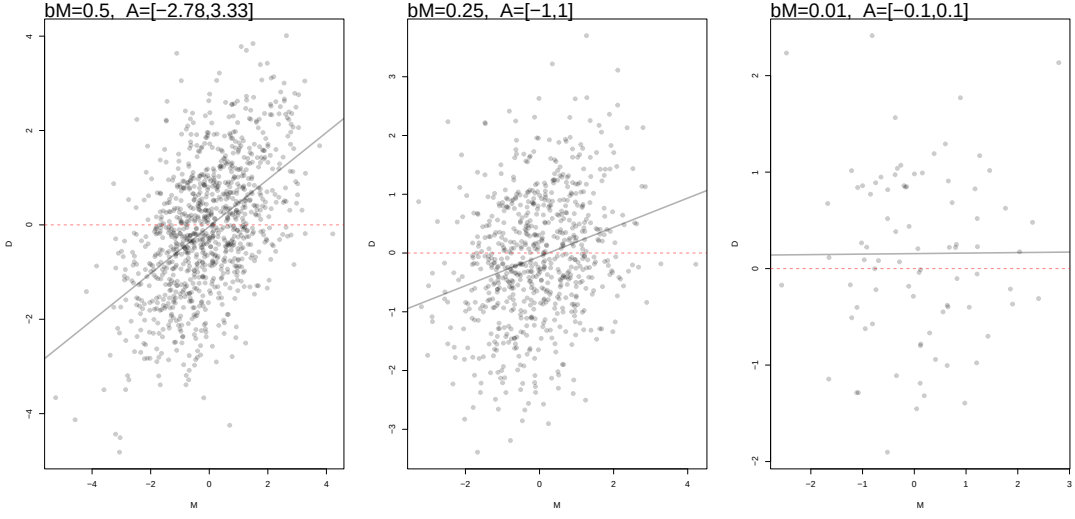
```
> summary(lm(D ~ A + M, data=d)) # controlled relation

Call:
lm(formula = D ~ A + M, data = d)

Residuals:
     Min       1Q   Median       3Q      Max
-2.27295 -0.68174  0.03781  0.78885  2.95320

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.18854    0.09871  -1.910   0.0591 .
A           -1.03121    0.13483  -7.648 1.49e-11 ***
M           -0.06134    0.09362  -0.655   0.5139
```

So, what is going on?



3. Example cases

Fork bias: masked relationships (a)

Masked relationships (a)[5]

also known as,

- omitted variable bias
- an instance of fork bias

research question,

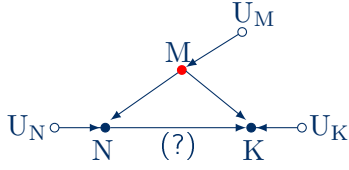
- Does N has a (direct) effect on K?

variables,

- M, mammal mass in kg.
- N, ratio neocortex over total brain mass
- K, Kcal. per gram of milk

$$\mathrm{M} = \begin{cases} \mathrm{M} \leftarrow \mathrm{f_M}(\mathrm{U_M}) \\ \mathrm{N} \leftarrow \mathrm{f_N}(\mathrm{M}, \mathrm{U_N}) \\ \mathrm{K} \leftarrow \mathrm{f_K}(\mathrm{M}, \mathrm{N}, \mathrm{U_K}) \\ \mathrm{U} \sim \mathrm{P}(\mathbf{U}) \end{cases}$$

(a) structural model



(b) causal diagram

[5] McElreath [12], chapter 05 (p. 144)

Simulation setting

```r
# sim
M = rnorm( 100 )
N = rnorm( 100 , 1*M )
K = rnorm( 100 , 1*N + -1*M )
d = data.frame(N=N,M=M,K=K)
```

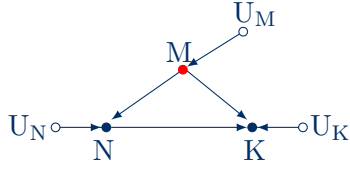
(c) R code

$$M = \begin{cases} M \leftarrow f_M(U_M) \\ N \leftarrow f_N(M, U_N) \\ K \leftarrow f_K(M, N, U_K) \\ U \sim P(\mathbf{U}) \end{cases}$$

(a) structural model

Implications,

- $N \not\perp K$
- $N \not\perp K \mid M$

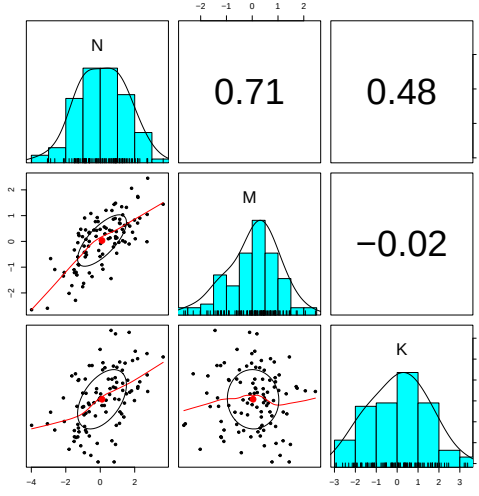


(b) causal diagram

“Eyeballing” analysis

based on correlation analysis,

- $\text{cor}(\text{N}, \text{K}) > 0$ goes in line of our “rudimentary” understanding of the data.
- but why there is $\text{cor}(\text{M}, \text{k}) \approx 0$? (hint: univariate correlation)
- we might not include M as a covariate in our statistical model



Regression, regression!!

based on statistical analysis,

- we have two different stories, (which one is the “truth”?)

```
> summary(lm(K ~ N, data=d)) # biased estimate

Call:
lm(formula = K ~ N, data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8355 -0.8110  0.0188  0.7897  3.4276

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.01401    0.12057   0.116    0.908
N            0.53002    0.09332   5.680 1.38e-07 ***
> summary(lm(K ~ N + M, data=d)) # less biased estima

Call:
lm(formula = K ~ N + M, data = d)

Residuals:
     Min       1Q   Median       3Q      Max
-2.50873 -0.72626 -0.01968  0.69016  2.93000

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.22096    0.09845   2.244   0.0271 *
N            0.95510    0.10089   9.466 1.91e-15 ***
M           -1.06246    0.15462  -6.871 6.14e-10 ***
```

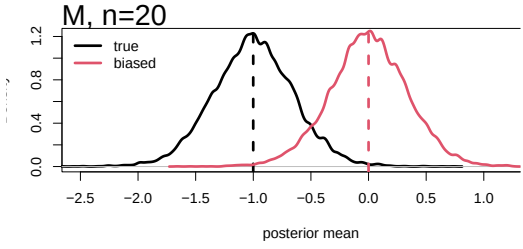
I'll get more data!!

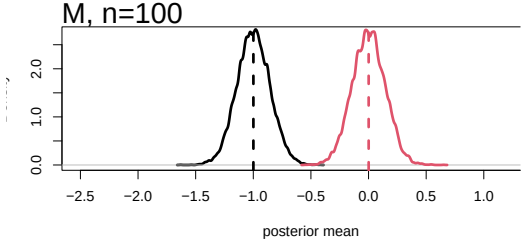
imagine we can continue sampling,

- top: $10,000$ samples $n = 20$
- bottom: $10,000$ samples $n = 100$

under the incorrect model,
the larger the sample size,

- the more certain you are about your biased estimates





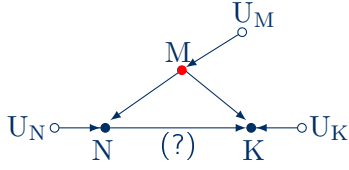
The dream team!!

based on DAG and statistical model,

- the 2nd D-separation rule requires you to control any noncollider to block the backdoor path, i.e. $N \not\perp\!\!\!\perp K \mid M$
- conditioning on M we can find, $E[K|do(n)] = E[\, E[K|N = n, M]\,]$ (law of total expectation)
- then we can find the $ACE(n) = E[D|do(n+1)] - E[D|do(n)]$ (Frisch-Waugh-Lovell theorem)

$$M = \begin{cases} M \leftarrow f_M(U_M) \\ N \leftarrow f_N(M, U_N) \\ K \leftarrow f_K(M, N, U_K) \\ U \sim P(\mathbf{U}) \end{cases}$$

(a) structural model



(b) causal diagram

the dream team!!

based on DAG and statistical analysis,

- the less biased model is the second, (assuming our DAG is true)

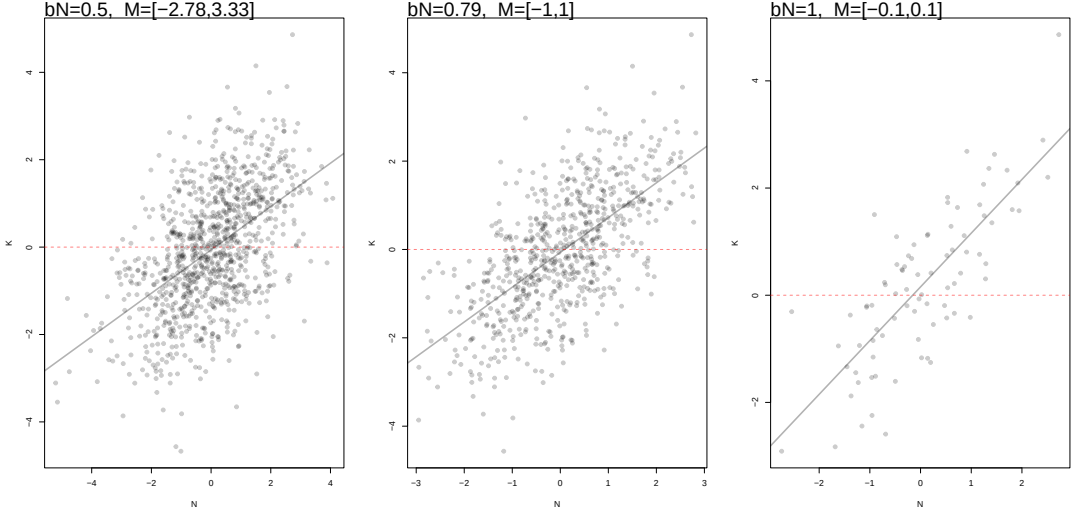
```
> summary(lm(K ~ N + M, data=d)) # less biased estima

Call:
lm(formula = K ~ N + M, data = d)

Residuals:
     Min       1Q   Median       3Q      Max
-2.50873 -0.72626 -0.01968  0.69016  2.93000

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.22096    0.09845   2.244   0.0271 *
N            0.95510    0.10089   9.466 1.91e-15 ***
M           -1.06246    0.15462  -6.871 6.14e-10 ***
```

So, what is going on?



3. Example cases

Fork bias: masked relationships (b)

Masked relationships (b)[6]

also known as,

- (unobserved) omitted variable bias
- an instance of fork bias

research question,

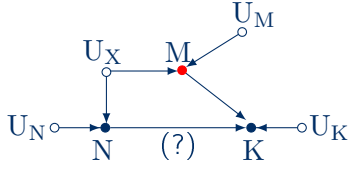
- Does N has a (direct) effect on K?

variables,

- U_X, unobservable (e.g. genetics)
- M, mammal mass in kg.
- N, neocortex over total brain mass
- K, Kcal. per gram of milk

$$M = \begin{cases} N \leftarrow f_N(U_N, U_X) \\ M \leftarrow f_M(U_M, U_X) \\ K \leftarrow f_K(M, N, U_K) \\ U \sim P(\mathbf{U}) \end{cases}$$

(a) structural model



(b) causal diagram

[6] McElreath [12], chapter 05 (p. 144)

Simulation setting

```r
# sim
U = rnorm( 100 )
N = rnorm( 100 , 1*U )
M = rnorm( 100 , 1*U )
K = rnorm( 100 , 1*N + -1*M )
d = data.frame(U=U,N=N,M=M,K=K)
```

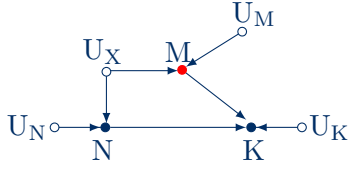
(c) R code

$$\mathrm{M} = \begin{cases} \mathrm{N} \leftarrow \mathrm{f_N}(\mathrm{U_N}, \mathrm{U_X}) \\ \mathrm{M} \leftarrow \mathrm{f_M}(\mathrm{U_M}, \mathrm{U_X}) \\ \mathrm{K} \leftarrow \mathrm{f_K}(\mathrm{M}, \mathrm{N}, \mathrm{U_K}) \\ \mathrm{U} \sim \mathrm{P}(\mathbf{U}) \end{cases}$$

(a) structural model

Implications,

- $\mathrm{N} \not\perp\!\!\!\perp \mathrm{K}$
- $\mathrm{N} \not\perp\!\!\!\perp \mathrm{K} \mid \mathrm{M}$

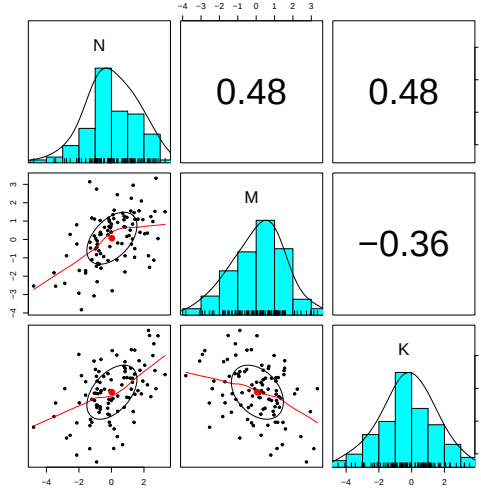


(b) causal diagram

“Eyeballing” analysis

based on correlation analysis,

- $\text{cor}(\text{N}, \text{K}) > 0$ goes in line of our “rudimentary” understanding of the data.
- $\text{cor}(\text{M}, \text{K}) < 0$ does NOT goes in line of our “rudimentary” understanding of the data.
  (hint: univariate correlation)
- we include M as a covariate in our statistical model
  (by chance?)



Regression, regression!!

based on statistical analysis,

- we have two different stories, (which one is the “truth”?)

```
> summary(lm(K ~ N, data=d)) # unobserved path still

Call:
lm(formula = K ~ N, data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7763 -0.8480  0.1497  0.9874  3.3530

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.24867    0.14573  -1.706   0.0911 .
N            0.51406    0.09502   5.410 4.46e-07 ***
> summary(lm(K ~ N + M, data=d)) # unobserved path cl

Call:
lm(formula = K ~ N + M, data = d)

Residuals:
     Min       1Q   Median       3Q      Max
-2.58218 -0.58434 -0.00579  0.72016  1.78724

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.19978    0.09375  -2.131   0.0356 *
N            0.90893    0.06958  13.064   <2e-16 ***
M           -0.89676    0.07572 -11.843   <2e-16 ***
```

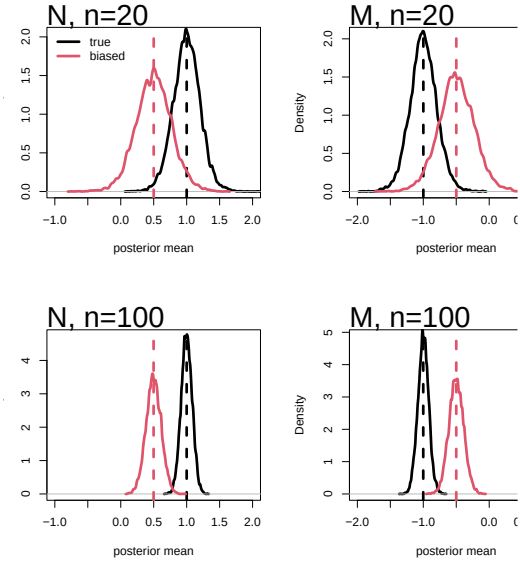
I'll get more data!!

imagine we can continue sampling,

- top: $10,000$ samples $n = 20$
- bottom: $10,000$ samples $n = 100$

under the incorrect model,
the larger the sample size,

- the more certain you are about your biased estimates



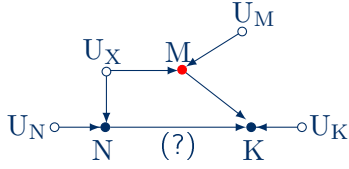
The dream team!!

based on DAG and statistical model,

- the 2nd D-separation rule requires control on any noncollider to block the backdoor path,
  i.e. $N \not\perp\!\!\!\perp K \mid U_X$
  (but it is unobservable)
- still we use the 2nd D-separation rule by controlling for M,
  i.e. $N \not\perp\!\!\!\perp K \mid M$
- conditioning on M we can still find,
  $E[K|do(n)] = E[\, E[K|N = n, M] \,]$
  (law of total expectation)
- then we can find the
  $ACE(n) = E[D|do(n+1)] - E[D|do(n)]$
  (Frisch-Waugh-Lovell theorem??)

$$M = \begin{cases} N \leftarrow f_N(U_N, U_X) \\ M \leftarrow f_M(U_M, U_X) \\ K \leftarrow f_K(M, N, U_K) \\ U \sim P(\mathbf{U}) \end{cases}$$

(a) structural model



(b) causal diagram

the dream team!!

based on DAG and statistical analysis,

- the less biased model is the second, (assuming our DAG is true)

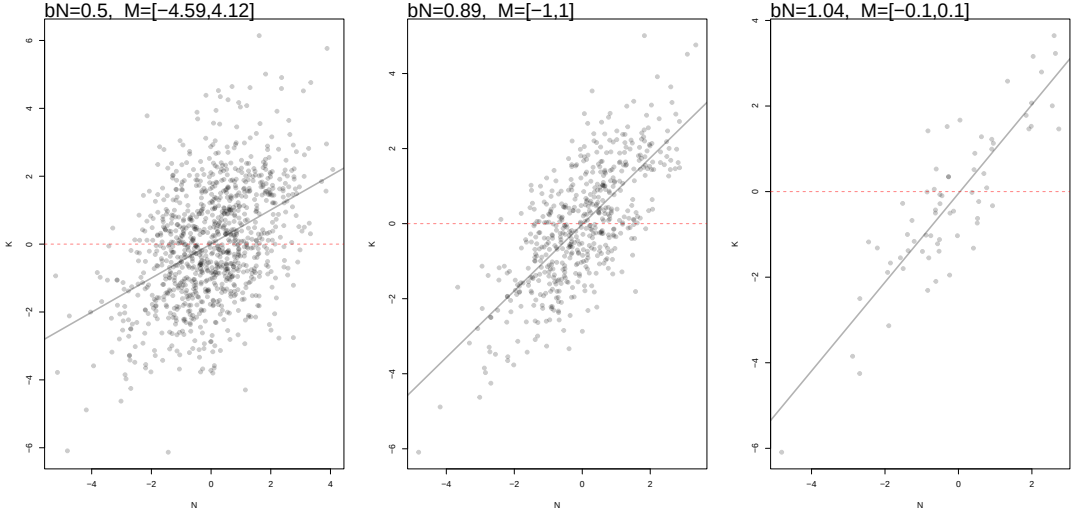
```
> summary(lm(K ~ N + M, data=d)) # unobserved path cl

Call:
lm(formula = K ~ N + M, data = d)

Residuals:
     Min       1Q   Median       3Q      Max
-2.58218 -0.58434 -0.00579  0.72016  1.78724

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.19978    0.09375  -2.131   0.0356 *
N            0.90893    0.06958  13.064   <2e-16 ***
M           -0.89676    0.07572 -11.843   <2e-16 ***
```

So, what is going on?



Similar scenario[7]

research question,

- Does E has a (direct) effect on I?

variables,

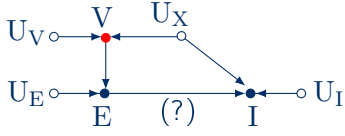
- U_X, unobservable (e.g. family context)
- V, personal values
- E, education
- I, income

then,

- we need to control by V to get an unbiased estimate of $E \to I$

$$M = \begin{cases} V \leftarrow f_M(U_V, U_X) \\ E \leftarrow f_E(V, U_E) \\ I \leftarrow f_I(E, U_X, U_I) \\ U \sim P(\mathbf{U}) \end{cases}$$

(a) structural model



(b) causal diagram

[7] Cinelli et al. [4] (p. 3), McElreath [13], lecture 6

3. Example cases

Fork bias: multicollinearity

Multicollinearity[8]

also known as,

- extreme case of masked relationships
- an instance of fork bias

research question,

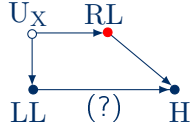
- Should we include RL in our model?

variables,

- U_X, unobservable (e.g. genetics and context)
- LL, individual's left leg
- RL, individual's right leg
- H, individual's height

$$M = \begin{cases} LL \leftarrow f_L(U_X) \\ RL \leftarrow f_L(U_X) \\ H \leftarrow f_K(RL, LL) \\ U \sim P(\mathbf{U}) \end{cases}$$

(a) structural model



(b) causal diagram

[8] McElreath [12], chapter 06 (p. 163)

Simulation setting

```r
# backward simulation
H = round( rnorm( 100 , 170, 2), 1)
Lp = runif( 100 , 0.5-0.05, 0.5+0.05)
LL = round( Lp*H + rnorm( 100 , 0, 1 ), 1)
RL = round( Lp*H + rnorm( 100 , 0, 1 ), 1)
d = data.frame(LL,RL,H)
```

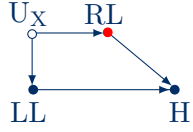
(c) R code

Implications,

- $\mathrm{LL} \not\perp \mathrm{RL}$

$$M = \begin{cases} \mathrm{LL} \leftarrow f_L(U_X) \\ \mathrm{RL} \leftarrow f_L(U_X) \\ \mathrm{H} \leftarrow f_K(\mathrm{RL}, \mathrm{LL}) \\ \mathrm{U} \sim P(\mathbf{U}) \end{cases}$$

(a) structural model

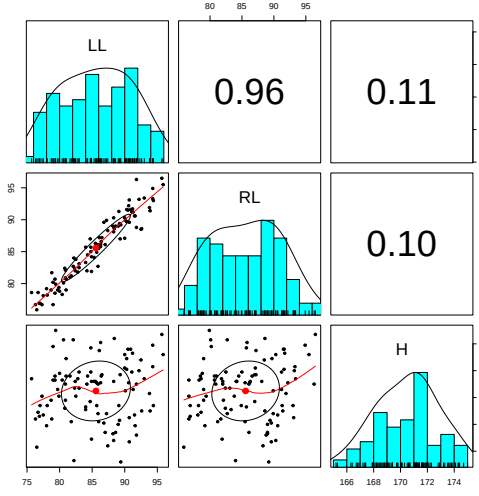


(b) causal diagram

“Eyeballing” analysis

based on correlation analysis,

- $\text{cor}(\text{LL}, \text{H}) > 0$, $\text{cor}(\text{RL}, \text{H}) > 0$ and $\text{cor}(\text{LL}, \text{RL}) > 0$ goes in line of our “understanding” of the data.
- we might not include RL as a covariate in our statistical model (based on univariate correlation)



Regression, regression!!

based on statistical analysis,

- the second regression show a smaller effect of LL,
- the second regression show way larger SE values
  (not rejecting the null)

```
> summary(lm(H ~ -1 + LL, data=d)) # unbiased,

Call:
lm(formula = H ~ -1 + LL, data = d)

Residuals:
     Min       1Q   Median       3Q      Max
-17.9704  -8.3662   0.7494  10.5256  21.1464

Coefficients:
   Estimate Std. Error t value Pr(>|t|)
LL  1.98486    0.01229   161.6   <2e-16 ***
> summary(lm(H ~ -1 + LL + RL, data=d)) # ine

Call:
lm(formula = H ~ -1 + LL + RL, data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-16.780  -8.592   0.532  10.253  18.299

Coefficients:
   Estimate Std. Error t value Pr(>|t|)
LL   1.0094     0.6969   1.448    0.151
RL   0.9757     0.6970   1.400    0.165
```

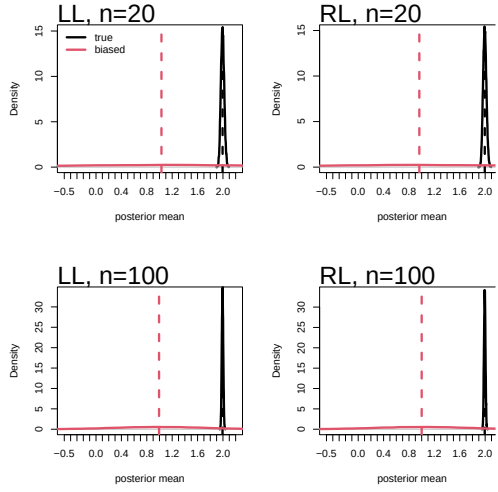
I'm sure data won't help

imagine we can continue sampling,

- top: $10,000$ samples $n = 20$
- bottom: $10,000$ samples $n = 100$

under the incorrect model,
the larger the sample size,

- the less certain are your biased estimates



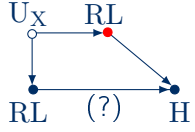
Not so great now??

based on DAG and statistical model,

- the 2nd D-separation rule requires control on any noncollider to block the backdoor path,
  i.e. $\mathrm{LL} \not\perp\!\!\!\perp \mathrm{H} \mid \mathrm{U_X}$
  (but it is unobservable)
- we still use the 2nd D-separation rule by controlling for RL, but still we have $\mathrm{LL} \not\perp\!\!\!\perp \mathrm{H} \mid \mathrm{RL}$
- issue goes beyond the backdoor path
  the issue is that RL and LL provide the same information,
  i.e. they form a singular matrix,
  (is like having a causal model like b)

$$\mathrm{M} = \begin{cases} \mathrm{LL} = \mathrm{RL} \\ \mathrm{RL} \leftarrow \mathrm{f_L}(\mathrm{U_X}) \\ \mathrm{H} \leftarrow \mathrm{f_K}(\mathrm{RL}) \\ \mathrm{U} \sim \mathrm{P}(\mathbf{U}) \end{cases}$$

(a) structural model

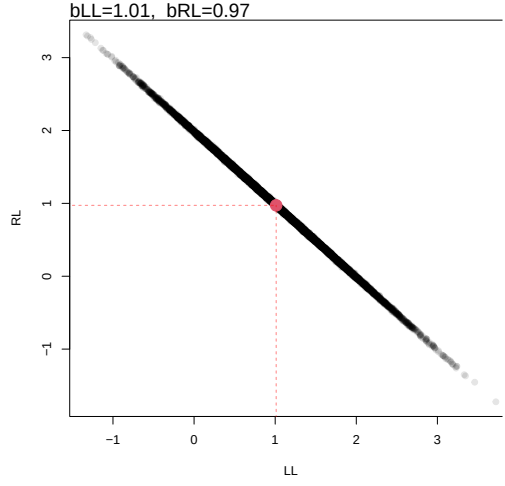


(b) causal diagram

So, what is going on then??

the parameters under the second regression,

- are in a thin ridge $\text{Cor}(\text{bLL}, \text{bRL}) \approx -1$
- statistical model does not find only one solution (red dot), the thin ridge implies there are infinite solutions for the model,
  (related to the singular matrix thing)



Not so great now??

based on DAG and statistical analysis,

- the less biased and more precise model is the first,
  (assuming our DAG is true)

```
> summary(lm(H ~ -1 + LL, data=d)) # unbiased

Call:
lm(formula = H ~ -1 + LL, data = d)

Residuals:
     Min       1Q   Median       3Q      Max
-17.9704  -8.3662   0.7494  10.5256  21.1464

Coefficients:
   Estimate Std. Error t value Pr(>|t|)
LL  1.98486    0.01229   161.6   <2e-16 ***
```

3. Example cases

No more fork bias: neutral control

Neutral control[9]

also known as,

- precision "booster"
- similar to experimental design

research question,

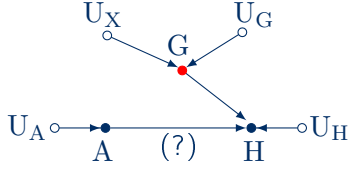
- Should we include G on our model?

variables,

- A, "hearing" age
- G, gender
- U_X, unobservable (e.g. no idea yet)
- H, inverse logit of entropy (approximate of speech intelligibility)

$$M = \begin{cases} G \leftarrow f_G(U_G, U_X) \\ A \leftarrow f_A(U_A) \\ H \leftarrow f_H(A, G, U_H) \\ U \sim P(\mathbf{U}) \end{cases}$$

(a) structural model



(b) causal diagram

[9] Cinelli et al. [4] (p. 4)

Simulation setting

```r
# sim
G = sample( 0:1, 100 , replace=T )
A = rnorm( 100 )
H = rnorm( 100 , -1*A + -1*G )
d = data.frame(G=G,A=A,SI=SI)
```

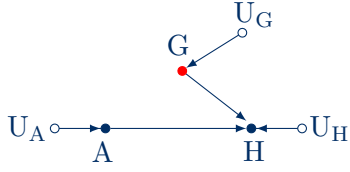
(c) R code

$$M = \begin{cases} G \leftarrow f_G(U_G) \\ A \leftarrow f_A(U_A) \\ H \leftarrow f_H(A, G, U_H) \\ U \sim P(\mathbf{U}) \end{cases}$$

(a) structural model

Implications,

- $A \perp\!\!\!\perp G$
- $A \not\perp\!\!\!\perp H$
- $G \not\perp\!\!\!\perp H$

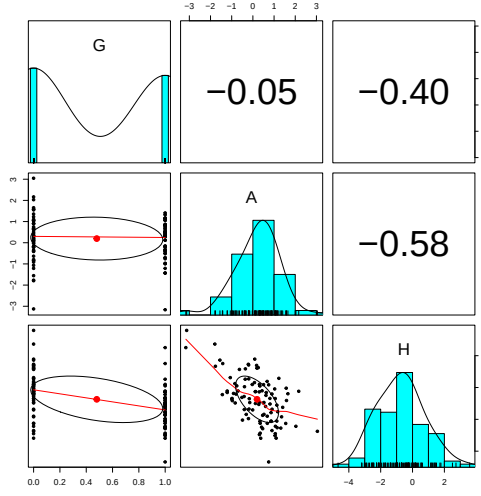


(b) causal diagram

“Eyeballing” analysis

based on correlation analysis,

- $\text{cor}(\text{G}, \text{H}) < 0$, $\text{cor}(\text{G}, \text{A}) \approx 0$ and $\text{cor}(\text{A}, \text{H}) < 0$ goes in line of our “rudimentary” understanding of the data.
- we include both as a covariate in our statistical model



Regression, regression!!

based on statistical analysis,

- almost no change on our estimates,
- lower standard errors for A when G is included
  (because we have explained variability in H)

```
> summary(lm(H ~ A, data=d)) # correct estimate

Call:
lm(formula = H ~ A, data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4714 -0.8797 -0.0633  0.8963  2.4346

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.5770     0.1216  -4.746 7.07e-06 ***
A            -0.8410     0.1183  -7.108 1.92e-10 ***
> summary(lm(H ~ A + G, data=d)) # correct estimate,

Call:
lm(formula = H ~ A + G, data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-2.7994 -0.6914  0.0579  0.7796  1.8274

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.03317    0.14312   0.232    0.817
A           -0.87360    0.10090  -8.658 1.05e-13 ***
G           -1.25786    0.20371  -6.175 1.55e-08 ***
```

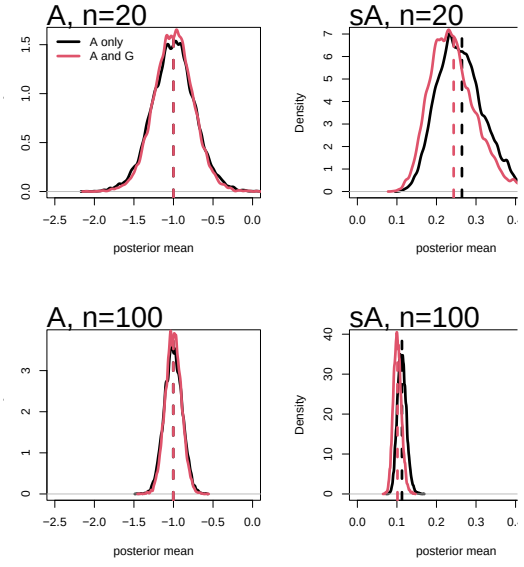
I'll get more data!!

imagine we can continue sampling,

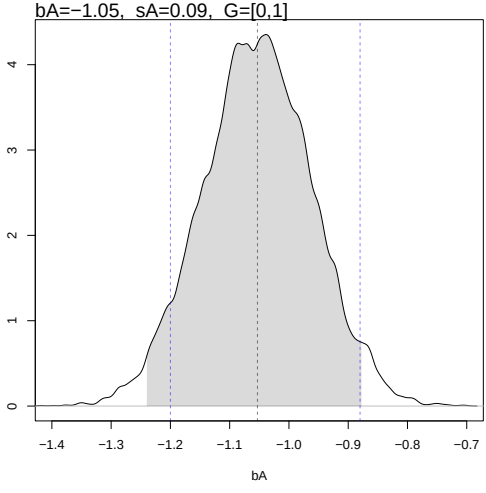
- top: $10,000$ samples $n = 20$
- bottom: $10,000$ samples $n = 100$

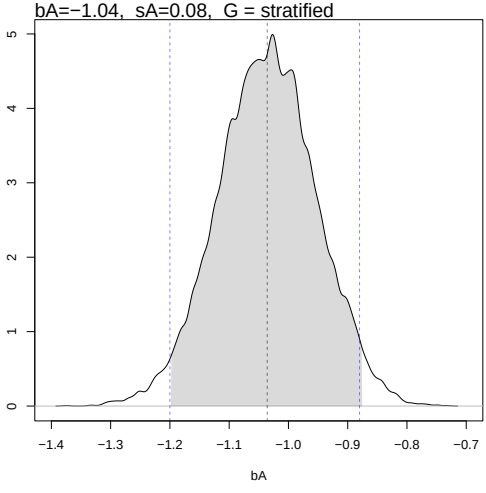
under the second model,
the larger the sample size,

- the more certain you are about your non-biased estimates (under the any model)



So, what is going on?





3. Example cases

Pipe bias: precision parasite

Precision parasite[10]

research question,

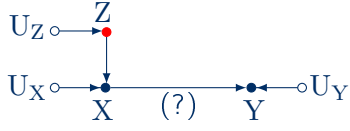
- Should we include Z in the model?

variables,

- Z,
- X,
- Y,

$$M = \begin{cases} Z \leftarrow f_Z(U_Z) \\ X \leftarrow f_X(Z, U_X) \\ Y \leftarrow f_Y(X, U_Y) \\ U \sim P(\mathbf{U}) \end{cases}$$

(a) structural model



(b) causal diagram

[10] McElreath [13], lecture 6, Cinelli et al. [4] (p. 5)

Simulation setting

```r
# sim
Z = rnorm( 100 )
X = rnorm( 100 , 1*Z )
Y = rnorm( 100 , 1*X )
d = data.frame(Z,X,Y)
```

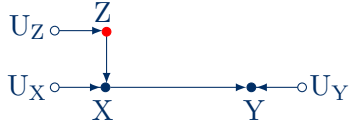
(c) R code

$$M = \begin{cases} Z \leftarrow f_Z(U_Z) \\ X \leftarrow f_X(Z, U_X) \\ Y \leftarrow f_Y(X, U_Y) \\ U \sim P(\mathbf{U}) \end{cases}$$

(a) structural model

Implications,

- $X \not\perp Y$
- $Z \not\perp Y \mid X$

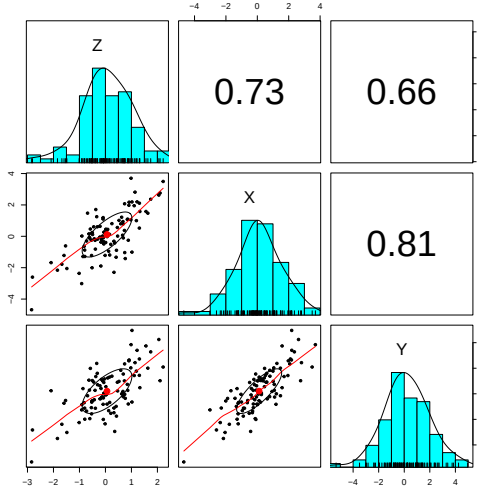


(b) causal diagram

“Eyeballing” analysis

based on correlation analysis,

- $\text{cor}(\text{Z}, \text{X}) > 0$ is not large enough to discard it as multicollinearity.
- $\text{cor}(\text{Z}, \text{Y}) > 0$ and $\text{cor}(\text{X}, \text{Y}) > 0$ indicate both should be in our model
  (it might be our research hypothesis)



Regression, regression!!

based on statistical analysis,

- no bias in parameter if Z is in,
- but we loose precision on X

```
> summary(lm(Y ~ X, data=d)) # unbiased effect, more

Call:
lm(formula = Y ~ X, data = d)

Residuals:
     Min       1Q   Median       3Q      Max
-2.41746 -0.73659 -0.09384  0.63812  2.10338

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.03433    0.10274   0.334    0.739
X            1.16908    0.06717  17.405   <2e-16 ***
> summary(lm(Y ~ X + Z, data=d)) # unbiased effects,

Call:
lm(formula = Y ~ X + Z, data = d)

Residuals:
     Min       1Q   Median       3Q      Max
-2.27019 -0.74072 -0.06355  0.66643  2.20770

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.04722    0.10350   0.456    0.649
X            1.08881    0.10358  10.512   <2e-16 ***
Z            0.15431    0.15159   1.018    0.311
```

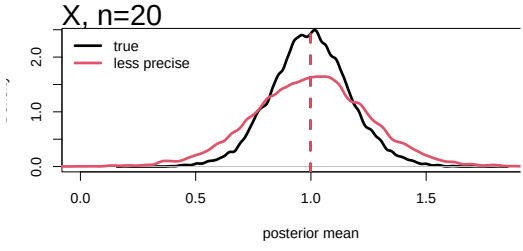
With more data??

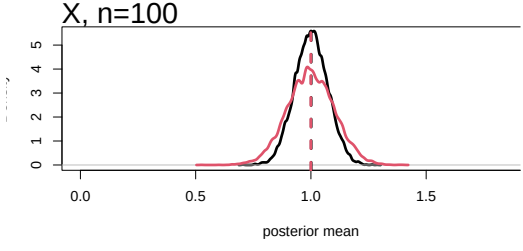
imagine we can continue sampling,

- top: $10,000$ samples $n = 20$
- bottom: $10,000$ samples $n = 100$

under the second model,
the larger the sample size,

- still less precise estimates
- more difficult to test hypothesis





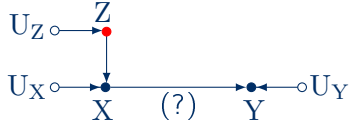
Now, what is going on here??

based on DAG and statistical model,

- conditioning on Z reduces variation on X

$$M = \begin{cases} Z \leftarrow f_Z(U_Z) \\ X \leftarrow f_X(Z, U_X) \\ Y \leftarrow f_Y(X, U_Y) \\ U \sim P(\mathbf{U}) \end{cases}$$

(a) structural model



(b) causal diagram

Now, what is going on here??

based on DAG and statistical analysis,

- the more appropriate model (for inference) is the first,
  (assuming our DAG is true)

```
> summary(lm(Y ~ X, data=d)) # unbiased effect, more

Call:
lm(formula = Y ~ X, data = d)

Residuals:
     Min       1Q   Median       3Q      Max
-2.41746 -0.73659 -0.09384  0.63812  2.10338

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.03433    0.10274   0.334    0.739
X            1.16908    0.06717  17.405   <2e-16 ***
```

So, what is going on?

pipe1_triptych.pdf

3. Example cases

Pipe bias: post-treatment

Post-treatment bias[11]

case of,

- full mediation

research question,

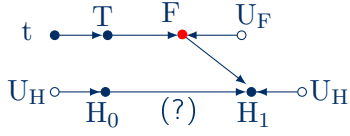
- Does the treatment T works?

variables,

- H_0, height of plant at $t = 0$
- T, antifungal treatment
- F, presence of fungus
- H_1, height of plant at $t = 1$

$$M = \begin{cases} H_0 \leftarrow f_H(U_H) \\ T \leftarrow f_T(t) \\ F \leftarrow f_F(T, U_F) \\ H_1 \leftarrow f_H(F, H_0, U_H) \\ U \sim P(\mathbf{U}) \end{cases}$$

(a) structural model



(b) causal diagram

[11] McElreath [12], chapter 6 (p. 170)

Simulation setting

```r
# sim
h0 = rnorm( 100 , 10, 2)
Tr = rep( 0:1 , each=100/2 )
Fu = rbinom( n , size=1 , prob=0.5 + -0.4*Tr )
h1 = h0 + rnorm( n , 5 + -3*Fu)
d = data.frame( h0=h0, h1=h1, Tr=Tr, Fu=Fu )
```

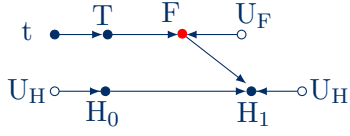
(c) R code

$$M = \begin{cases} H_0 \leftarrow f_H(U_H) \\ T \leftarrow f_T(t) \\ F \leftarrow f_F(T, U_F) \\ H_1 \leftarrow f_H(F, H_0, U_H) \\ U \sim P(\mathbf{U}) \end{cases}$$

(a) structural model

Implications,

- $T \perp\!\!\!\perp H_0$
- $T \not\perp\!\!\!\perp H_1$
- $T \not\perp\!\!\!\perp H_1 \mid F$



(b) causal diagram

Descriptive analysis

based on descriptive analysis,

- positive change in height with treatment.
- negative change in height with fungus.
- diluted relationship for T when both are in the model
  (hint: blocking path of information)

Tr	mean	sd	n	se
<fct>	<dbl>	<dbl>	<int>	<dbl>
0	2.96	1.75	50	0.248
1	4.53	1.45	50	0.205

Fu	mean	sd	n	se
<fct>	<dbl>	<dbl>	<int>	<dbl>
0	4.93	0.953	62	0.121
1	1.81	0.902	38	0.146

Tr	Fu	mean	sd	n	se
<fct>	<fct>	<dbl>	<dbl>	<int>	<dbl>
0	0	4.91	0.810	19	0.186
0	1	1.76	0.861	31	0.155
1	0	4.93	1.02	43	0.155
1	1	2.01	1.12	7	0.423

Again regression!!

based on statistical analysis we have two different stories (but not quite),

- treatment has a significant effect,
- but gets completely diluted when fungus is considered in the model

```
> summary(lm(h1-h0 ~ Tr, data=d)) # only treatment

Call:
lm(formula = h1 - h0 ~ Tr, data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1166 -1.0929  0.1755  1.2621  3.3990

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.9598     0.2273   13.02  < 2e-16 ***
Tr1           1.5656     0.3215    4.87 4.29e-06 ***
> summary(lm(h1-h0 ~ Tr + Fu, data=d)) # only fungus

Call:
lm(formula = h1 - h0 ~ Tr + Fu, data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-2.1691 -0.4823  0.0963  0.5315  2.0357

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.86397    0.19127  25.430   <2e-16 ***
Tr1          0.09138    0.21579   0.423    0.673
Fu1         -3.07122    0.22229 -13.816   <2e-16 ***
```

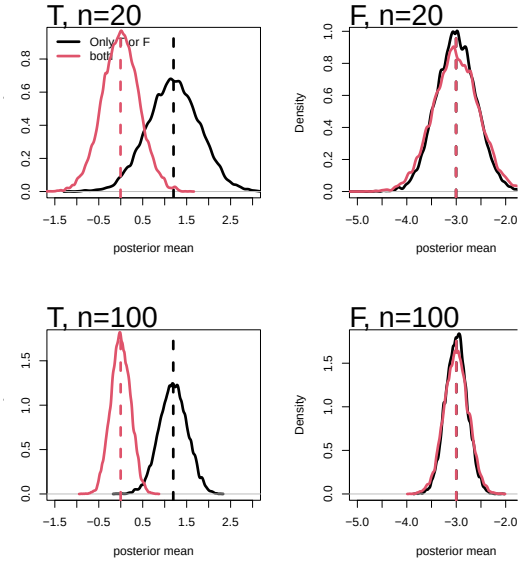
I can guess what happens with more data!!

imagine we can continue sampling,

- top: $10,000$ samples $n = 20$
- bottom: $10,000$ samples $n = 100$

under the "incorrect" model,
the larger the sample size,

- the more certain you are about your biased T estimates (not F) (this result is not wrong!)



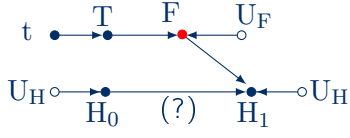
The dream team!!

based on DAG and statistical model,

- the 2nd D-separation states that if you to control any noncollider you block the backdoor path, i.e. $T \perp\!\!\!\perp H_1 \mid F$
- therefore if we want to find if $T = 1$ works, we should not stratify by F

$$M = \begin{cases} H_0 \leftarrow f_H(U_H) \\ T \leftarrow f_T(t) \\ F \leftarrow f_F(T, U_F) \\ H_1 \leftarrow f_H(F, H_0, U_H) \\ U \sim P(\mathbf{U}) \end{cases}$$

(a) structural model



(b) causal diagram

the dream team!!

based on DAG and statistical analysis,

- the model that answers our research question is the first one, (assuming our DAG is true)

```
> summary(lm(h1-h0 ~ Tr, data=d)) # only treatment

Call:
lm(formula = h1 - h0 ~ Tr, data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1166 -1.0929  0.1755  1.2621  3.3990

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.9598     0.2273   13.02  < 2e-16 ***
Tr1           1.5656     0.3215    4.87 4.29e-06 ***
```

So, what is going on?

pipe2_triptych.pdf

3. Example cases

Pipe bias: masked relationships

Masked relationships[12]

also known as,

- mediation
- Simpson's paradox
- an instance of pipe bias

research question,

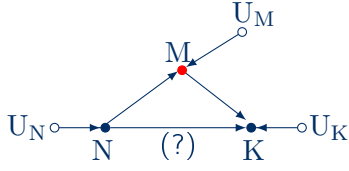
- Does N has a (direct) effect on K?

variables,

- M, mammal mass in kg.
- N, neocortex over total brain mass
- K, Kcal. per gram of milk

$$\mathrm{M} = \begin{cases} \mathrm{N} \leftarrow \mathrm{f_N}(\mathrm{U_N}) \\ \mathrm{M} \leftarrow \mathrm{f_M}(\mathrm{N}, \mathrm{U_M}) \\ \mathrm{K} \leftarrow \mathrm{f_K}(\mathrm{M}, \mathrm{N}, \mathrm{U_K}) \\ \mathrm{U} \sim \mathrm{P}(\mathbf{U}) \end{cases}$$

(a) structural model



(b) causal diagram

[12] McElreath [12], chapter 05 (p. 144)

Simulation setting

```r
# sim
N = rnorm( 100 )
M = rnorm( 100 , 1*N )
K = rnorm( 100 , 1*N + -1*M )
d = data.frame(N=N,M=M,K=K)
```

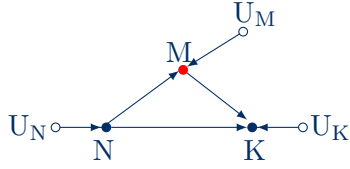
(c) R code

$$M = \begin{cases} N \leftarrow f_N(U_N) \\ M \leftarrow f_M(M, U_M) \\ K \leftarrow f_K(M, N, U_K) \\ U \sim P(\mathbf{U}) \end{cases}$$

(a) structural model

Implications,

- $N \not\perp K$
- $N \not\perp K \mid M$

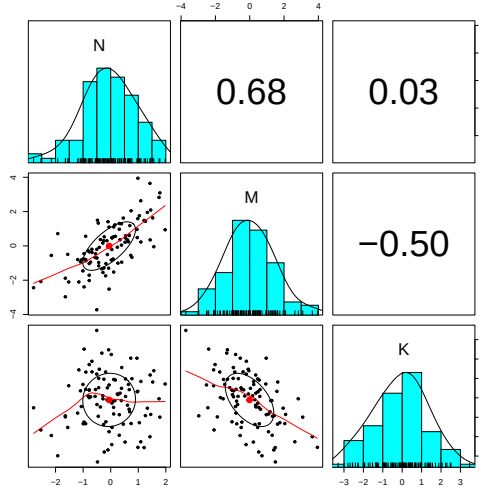


(b) causal diagram

“Eyeballing” analysis

based on correlation analysis,

- $\text{cor}(\text{M}, \text{K}) < 0$ does NOT goes in line of our “rudimentary” understanding of the data.
- and why there is $\text{cor}(\text{N}, \text{K}) \approx 0$? (hint: univariate correlation)
- we include N as a covariate in our statistical model (is our research hypothesis)



Regression, regression!!

based on statistical analysis,

- two regressions with two different results, which model is the “true”?

```
> summary(lm(K ~ N, data=d)) # biased estimate

Call:
lm(formula = K ~ N, data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1751 -0.9009  0.1519  0.8574  3.6041

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.10412    0.13808  -0.754    0.453
N            0.05005    0.14487   0.345    0.730
> summary(lm(K ~ N + M, data=d)) # less biased estimate

Call:
lm(formula = K ~ N + M, data = d)

Residuals:
     Min       1Q   Median       3Q      Max
-2.58484 -0.59175  0.04378  0.61175  2.43360

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.06181    0.09825  -0.629    0.531
N            0.98297    0.13994   7.024 2.98e-10 ***
M           -0.93107    0.09457  -9.846 2.89e-16 ***
```

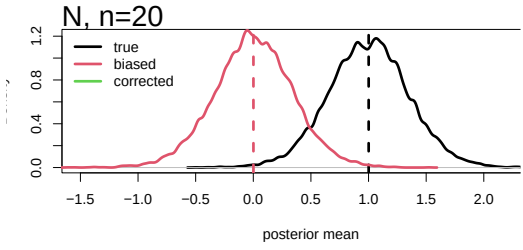
I'll get more data!!

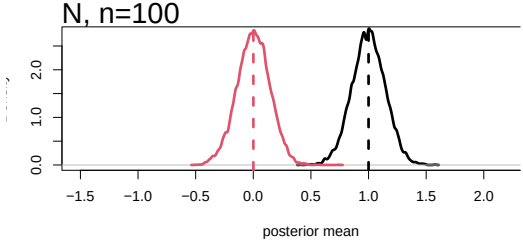
imagine we can continue sampling,

- top: $10,000$ samples $n = 20$
- bottom: $10,000$ samples $n = 100$

the larger the sample size,

- the more certain you are about your estimates
- the more mistaken you are about your research question
  (under the "incorrect" model)





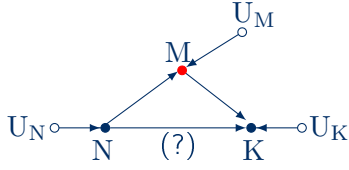
The dream team!!

based on DAG and statistical model,

- the 2nd D-separation rule requires you to control any noncollider to block the backdoor path, i.e. $N \not\perp K \mid M$
- conditioning on M we can find, $E[K|do(n)] = E[\, E[K|N = n, M]\,]$ (law of total expectation)
- then we can find the $ACE(n) = E[D|do(n+1)] - E[D|do(n)]$ (Frisch-Waugh-Lovell theorem)

$$M = \begin{cases} N \leftarrow f_N(U_N) \\ M \leftarrow f_M(M, U_M) \\ K \leftarrow f_K(M, N, U_K) \\ U \sim P(\mathbf{U}) \end{cases}$$

(a) structural model



(b) causal diagram

the dream team!!

based on DAG and statistical analysis,

- the less biased model is the second, (assuming our DAG is true)

```
> summary(lm(K ~ N + M, data=d)) # less biased estimate

Call:
lm(formula = K ~ N + M, data = d)

Residuals:
     Min       1Q   Median       3Q      Max
-2.58484 -0.59175  0.04378  0.61175  2.43360

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.06181    0.09825  -0.629    0.531
N            0.98297    0.13994   7.024 2.98e-10 ***
M           -0.93107    0.09457  -9.846 2.89e-16 ***
```

So, what is going on?

pipe3_triptych.pdf

3. Example cases

Pipe/Fork bias: bias amplification

Bias amplification[13]

also known as,

- (unobserved) omitted variable bias
- related to instrumental variables
- an instance of fork bias

research question,

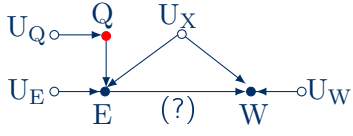
- Do E has a (direct) effect on W?

variables,

- Q, instrumental variable (e.g. quarter of the year)
- E, educational level
- U_X, unobservables (e.g. ability)
- W, future wages

$$M = \begin{cases} Q \leftarrow f_Q(U_Q) \\ E \leftarrow f_E(Q, U_X, U_E) \\ W \leftarrow f_W(E, U_X, U_W) \\ U \sim P(\mathbf{U}) \end{cases}$$

(a) structural model



(b) causal diagram

[13] McElreath [12], chapter 14 (p. 455), Cinelli et al. [4] (p. 5)

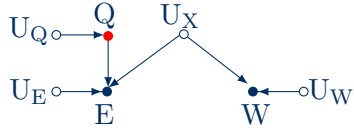
Simulation setting

```r
# sim
U = rnorm( 100 )
Q = sample( 1:4, 100, replace=T )
E = rnorm( 100 , 1*Q + 1*U )
W = rnorm( 100 , 0*E + 1*U )
d = data.frame(U=U,Q=Q,E=E,W=W)
```

(c) R code

$$M = \begin{cases} Q \leftarrow f_Q(U_Q) \\ E \leftarrow f_E(Q, U_X, U_E) \\ W \leftarrow f_W(U_X, U_W) \\ U \sim P(\mathbf{U}) \end{cases}$$

(a) structural model



(b) causal diagram

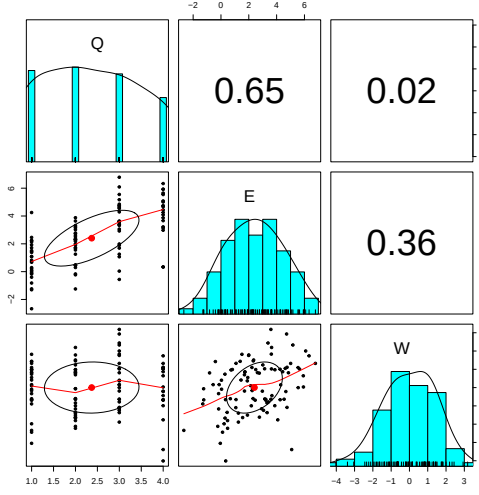
Implications,

- $E \not\perp\!\!\!\perp W$
- $E \perp\!\!\!\perp W \mid U_X$ (impossible)
- $Q \perp\!\!\!\perp U_X$ (cannot be tested)
- $Q \not\perp\!\!\!\perp E$
- $Q \perp\!\!\!\perp W \mid E$ (cannot be tested) (exclusion restriction)

“Eyeballing” analysis

based on correlation analysis,

- $\text{cor}(\text{Q}, \text{E}) > 0$ and $\text{cor}(\text{E}, \text{W}) > 0$ goes in line of our “rudimentary” understanding of the data.
- $\text{cor}(\text{Q}, \text{W}) > 0$ tells you about the exclusion restriction?
  (hint: No)
- we might NOT include Q as a covariate in our statistical model (but is the instrumental variable!!!)



Regression, regression!!

based on statistical analysis,

- two different stories
  (which model is the “truth”?)
- one is “worse”/“better” than the other?
- are both wrong?

```
> summary(lm(W ~ E, data=d)) # biased

Call:
lm(formula = W ~ E, data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0726 -0.9674  0.1771  0.9234  2.8787

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.60816    0.20506  -2.966 0.003793 **
E            0.25408    0.06559   3.873 0.000194 ***
> summary(lm(W ~ E + Q, data=d)) # more biased

Call:
lm(formula = W ~ E + Q, data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7405 -0.9774  0.0879  0.9162  2.9825

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.12229    0.30650   0.399  0.69078
E            0.42054    0.08262   5.090 1.75e-06 ***
Q           -0.47716    0.15361  -3.106  0.00249 **
```

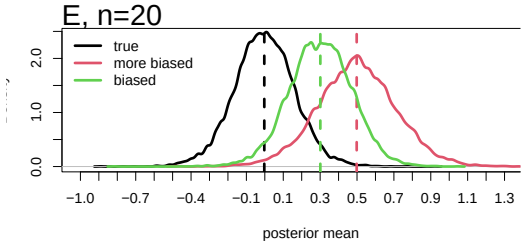
I'll get more data!!

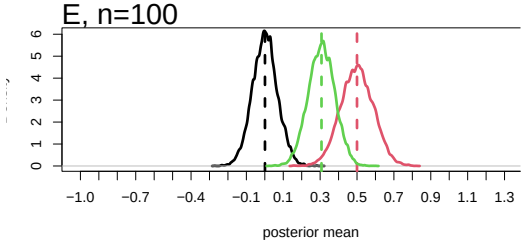
imagine we can continue sampling,

- top: $10,000$ samples $n = 20$
- bottom: $10,000$ samples $n = 100$

under the incorrect model,
the larger the sample size,

- the more certain you are about your biased estimates (under the any model!!)





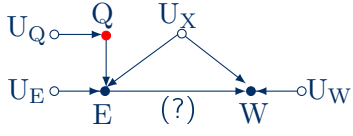
Yo, what is going on??

based on DAG and statistical model,

- the 2nd D-separation rule requires control on any noncollider to block the backdoor path,
  i.e. $E \perp\!\!\!\perp W \mid U_X$
  (but U_X is unobservable)
- if we use Q in the model, the 3rd D-separation rule kicks in:
  "A collider that has been conditioned on does not block a path."
  i.e. $Q \not\perp\!\!\!\perp U_X \mid E$
  (e.g. switch, electricity, and light bulb)

$$M = \begin{cases} Q \leftarrow f_Q(U_Q) \\ E \leftarrow f_E(Q, U_X, U_E) \\ W \leftarrow f_W(E, U_X, U_W) \\ U \sim P(\mathbf{U}) \end{cases}$$

(a) structural model



(b) causal diagram

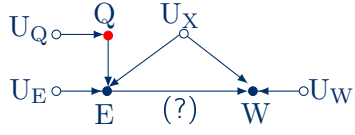
Yo, what is going on??

open paths?:

- $E \to W$
- $E \to U_x \to W$
- $E \to U_x \to Q \to E \to W$
- $E \to U_x \to Q \to E \to U_X \to W$

$$M = \begin{cases} Q \leftarrow f_Q(U_Q) \\ E \leftarrow f_E(Q, U_X, U_E) \\ W \leftarrow f_W(E, U_X, U_W) \\ U \sim P(\mathbf{U}) \end{cases}$$

(a) structural model



(b) causal diagram

What should I do then??

$$\begin{pmatrix} \mathrm{W} \\ \mathrm{E} \end{pmatrix} \sim \mathrm{MVN}\left[\begin{pmatrix} \mu_\mathrm{W} \\ \mu_\mathrm{E} \end{pmatrix}, \boldsymbol{\Sigma}\right]$$
$$\mu_\mathrm{W} = \alpha_\mathrm{W} + \beta_\mathrm{EW}\mathrm{E}$$
$$\mu_\mathrm{E} = \alpha_\mathrm{E} + \beta_\mathrm{QE}\mathrm{Q}$$

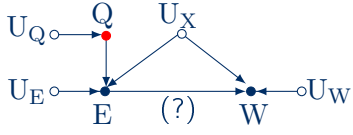
(c) probabilistic model

$$\mathrm{M} = \begin{cases} \mathrm{Q} \leftarrow \mathrm{f_Q}(\mathrm{U_Q}) \\ \mathrm{E} \leftarrow \mathrm{f_E}(\mathrm{Q}, \mathrm{U_X}, \mathrm{U_E}) \\ \mathrm{W} \leftarrow \mathrm{f_W}(\mathrm{E}, \mathrm{U_X}, \mathrm{U_W}) \\ \mathrm{U} \sim \mathrm{P}(\mathbf{U}) \end{cases}$$

(a) structural model

based on DAG and statistical model,
use the knowledge of the system

- one model for $\mathrm{Q} \to \mathrm{E}$
- one model for $\mathrm{E} \to \mathrm{W}$
- use the knowledge that $\mathrm{cov}(\mathrm{E}, \mathrm{W}) > 0$ due to unobserved confounder $\mathrm{U_X}$, (i.e. $\mathrm{cov}(\mathrm{E}, \mathrm{W}) = \boldsymbol{\Sigma} = \mathbf{SRS}$)



(b) causal diagram

did it worked???

based on DAG and bayesian statistical analysis,

- appropriate value estimated, (assuming our DAG is true)
- it picks up some of the unobserved correlation R[1, 2]

	mean	sd	5.5%	94.5%
aE	0.02	0.18	-0.26	0.30
aW	-0.14	0.16	-0.40	0.13
bQE	1.00	0.07	0.88	1.12
bEW	0.05	0.07	-0.06	0.16
R[1,1]	1.00	0.00	1.00	1.00
R[1,2]	0.33	0.11	0.15	0.50
R[2,1]	0.33	0.11	0.15	0.50
R[2,2]	1.00	0.00	1.00	1.00
S[1]	1.25	0.10	1.11	1.42
S[2]	1.39	0.10	1.24	1.56

did it worked???

frequentists guys apply
Two Stage Least Squares (2SLS)[a]:

- regress E ← Q,
- predict $\hat{E}$,
- regress W ← $\hat{E}$

```
s1 = lm( E ~ Q, data=d)
Ehat = s1$fitted.values
s2 = lm( W ~ Ehat, data=d)
# se not corrected

require(AER)
tsls = ivreg( W ~ E | Q, data=d)
# se corrected
```

```
> summary(s2)

Call:
lm(formula = W ~ Ehat, data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1911 -1.0670 -0.0643  1.3802  4.6813

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.5992     0.3667  -1.634    0.106
Ehat          0.2093     0.1263   1.658    0.101
> summary(tsls)

Call:
ivreg(formula = W ~ E | Q, data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-3.6571 -1.1852 -0.1819  1.0945  4.5328

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.5992     0.3325  -1.802   0.0746 .
E             0.2093     0.1145   1.829   0.0705 .
```

[a] Hanck et al. [7], section 12.1,
See McElreath [12] chapter 14 (p. 460) for a discussion on the method.

So, what is going on?

pipefork1_triptych.pdf

Similar case, contextual confounds[14]

research question,

- Does W has an effect on L?
- should we include H in our model?

variables,

- W, win the lottery
- H, happiness
- U_X, contextual confound
- L, lifespan

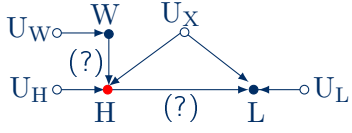
Short answer;

- for total effects: No
  (two question marks together)
- for direct effect of H → L:
  will be always counfounded

[14] McElreath [13], lecture 6

$$M = \begin{cases} W \leftarrow f_W(U_W) \\ H \leftarrow f_H(W, U_X, U_H) \\ L \leftarrow f_L(H, U_X, U_L) \\ U \sim P(\mathbf{U}) \end{cases}$$

(a) structural model



(b) causal diagram

3. Example cases

Collider bias: M-bias

M-bias[15]

case of,

- bias on pre-treatment variable

research question,

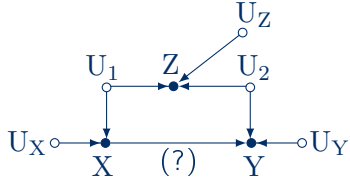
- Should we include Z in our model?

variables,

- Z, "health" quality of friends (defined as a continuum)
- X, health of individual 1
- U_1, hobbies of individual 1
- Y, health of individual 2
- U_2, hobbies of individual 2

$$\mathrm{M} = \begin{cases} \mathrm{X} \leftarrow \mathrm{f_X}(\mathrm{U_1}, \mathrm{U_X}) \\ \mathrm{Z} \leftarrow \mathrm{f_Z}(\mathrm{U_1}, \mathrm{U_2}, \mathrm{U_Z}) \\ \mathrm{Y} \leftarrow \mathrm{f_Y}(\mathrm{X}, \mathrm{U_2}, \mathrm{U_Y}) \\ \mathrm{U} \sim \mathrm{P}(\mathbf{U}) \end{cases}$$

(a) structural model



(b) causal diagram

[15] McElreath [13], lecture 6; Cinelli et al. [4] (p. 4)

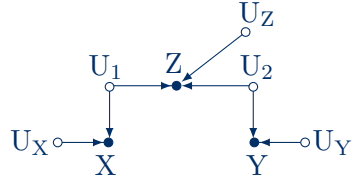
Simulation setting

```r
# sim
U1 = sample(1:5, 100 , replace=T)
U2 = sample(1:5, 100 , replace=T)
Z = rnorm( 100 , 0.5*U1 + 0.5*U2)
X = rnorm( 100 , 1*U1 + 0*Z)
Y = rnorm( 100 , 1*U2 + 0*X + 0*Z)
d = data.frame(U1,U2,Z,X,Y)
```

(c) R code

$$M = \begin{cases} X \leftarrow f_X(U_1, U_X) \\ Z \leftarrow f_Z(U_1, U_2, U_Z) \\ Y \leftarrow f_Y(X, U_2, U_Y) \\ U \sim P(\mathbf{U}) \end{cases}$$

(a) structural model



(b) causal diagram

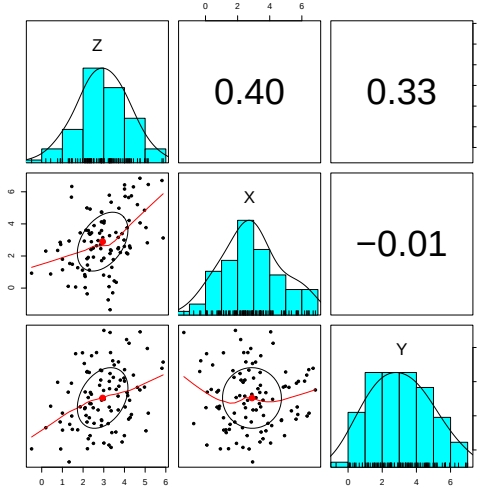
Implications,

- $X \perp\!\!\!\perp Y$
- $X \not\perp\!\!\!\perp Y \mid Z$

Descriptive analysis

based on descriptive analysis,

- $\text{Cor}(X, Y) \approx 0$, quite low
- larger $\text{Cor}(Z, Y)$, while $\text{Cor}(Z, X)$ is not high enough to discard it as a cause of multicollinearity,
- we might include Z rather than X (but the effect of X is our interest!!)
- then we include Z and X



Again regression!!

based on statistical analysis,

- X does not have an effect of Y, (in the first nor the second model)
- but X has an (non-negligible) effect when Z is in the model (but we do not reject the null)
- The increase of the X effect might lead you to think that with more data, we can reject the null (and you would be right!!)
- But is it correct to include Z?

```
> summary(lm(Y ~ X, data=d)) # unbiased effects (effic

Call:
lm(formula = Y ~ X, data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7452 -1.3300 -0.0253  1.2414  3.9479

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.067474   0.334586   9.168 7.74e-15 ***
X           -0.008255   0.097774  -0.084    0.933
> summary(lm(Y ~ X + Z, data=d)) # biased effects (ef

Call:
lm(formula = Y ~ X + Z, data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3756 -1.0898  0.0351  1.0540  3.7595

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.83324    0.45304   4.047 0.000105 ***
X           -0.15765    0.09991  -1.578 0.117835
Z            0.56590    0.14974   3.779 0.000272 ***
```

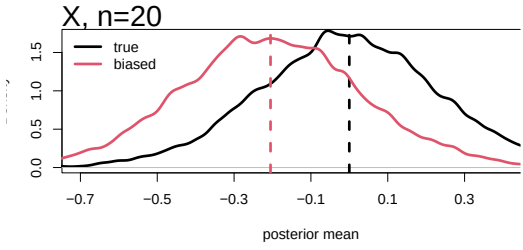
Ok, I get it!!, more data, more wrong!!

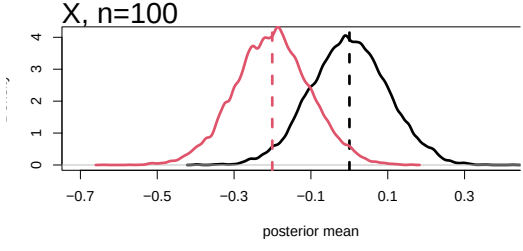
imagine we can continue sampling,

- top: $10,000$ samples $n = 20$
- bottom: $10,000$ samples $n = 100$

under the "incorrect" model,
the larger the sample size,

- the more certain you are about your biased estimates
  (with enough you could reject the null)





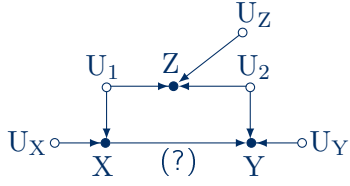
The dream team strikes back!!

based on DAG and statistical model,

- the 3rd D-separation states that a collider that has been conditioned on does not block a path,
  in this case: $X \to U_1 \to Z \to U_2 \to Y$
  i.e. $X \not\perp\!\!\!\perp Y \mid Z$
- therefore if we want to find the direct effect of $X \to Y$, we should not stratify by Z

$$M = \begin{cases} X \leftarrow f_X(U_1, U_X) \\ Z \leftarrow f_Z(U_1, U_2, U_Z) \\ Y \leftarrow f_Y(X, U_2, U_Y) \\ U \sim P(\mathbf{U}) \end{cases}$$

(a) structural model



(b) causal diagram

The dream team strikes back!!

based on DAG and statistical analysis,

- the model that answers our research question is the first one, (assuming our DAG is true)

```
> summary(lm(Y ~ X, data=d)) # unbiased effects (effic

Call:
lm(formula = Y ~ X, data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7452 -1.3300 -0.0253  1.2414  3.9479

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.067474   0.334586   9.168 7.74e-15 ***
X           -0.008255   0.097774  -0.084    0.933
```

So, what is going on?

collider4_triptych.pdf

3. Example cases

Descendant bias: case control

Case control[16]

also,

- an instance of descendant bias

research question,

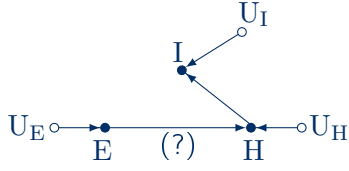
- Does E has a (direct) effect on H?
- Should we include I on our model?

variables,

- E, education
- H, hours in occupation (standardized)
- I, income

$$M = \begin{cases} E \leftarrow f_E(U_E) \\ H \leftarrow f_H(E, U_H) \\ I \leftarrow f_I(H, U_I) \\ U \sim P(\mathbf{U}) \end{cases}$$

(a) structural model



(b) causal diagram

[16] McElreath [13], lecture 06; Cinelli et al. [4] (p. 8)

Simulation setting

```r
# sim
E = rnorm( 100 )
H = rnorm( 100 , -1*E )
I = rnorm( 100 , -1*H )
d = data.frame(E,H,I)
```

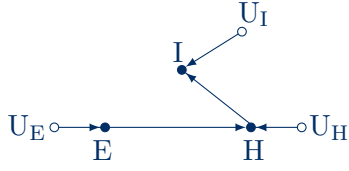
(c) R code

$$M = \begin{cases} E \leftarrow f_E(U_E) \\ H \leftarrow f_H(E, U_H) \\ I \leftarrow f_I(H, U_I) \\ U \sim P(\mathbf{U}) \end{cases}$$

(a) structural model

Implications,

- $E \not\perp\!\!\!\perp H$
- $E \perp\!\!\!\perp I \mid H$

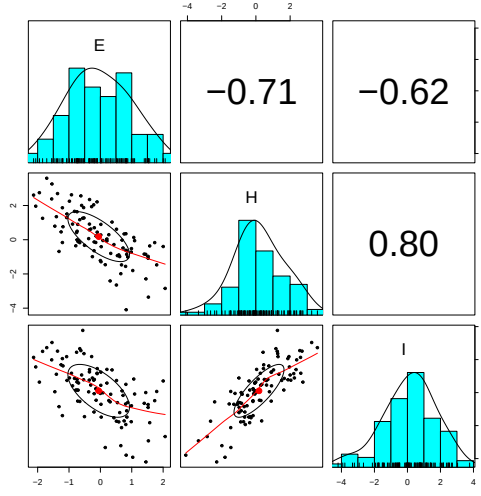


(b) causal diagram

“Eyeballing” analysis

based on correlation analysis,

- $\text{cor}(\text{E}, \text{I}) < 0$ does NOT goes in line of our “rudimentary” understanding of the data.
- while $\text{cor}(\text{H}, \text{I}) > 0$ indicate the more you work the more you gain (but is it the only way?)
- since $\text{cor}(\text{H}, \text{I})$ is high we might include it as a covariate in our statistical model (to improve the precision?)



Regression, regression!!

based on statistical analysis,

- we now have two models with two different “levels” of effects
- which one is the “truth”?

```
> summary(lm(H ~ E, data=d)) # unbiased effects

Call:
lm(formula = H ~ E, data = d)

Residuals:
     Min       1Q   Median       3Q      Max
-2.19319 -0.60621 -0.06694  0.55674  2.77776

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.04689    0.09534   0.492    0.624
E           -0.90223    0.08930 -10.104   <2e-16 ***
> summary(lm(H ~ E + I, data=d)) # biased effects

Call:
lm(formula = H ~ E + I, data = d)

Residuals:
     Min       1Q   Median       3Q      Max
-1.22991 -0.54882  0.01153  0.46886  1.79879

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.05036    0.06861   0.734    0.465
E           -0.45847    0.07915  -5.793 8.57e-08 ***
I            0.55017    0.05728   9.604 9.61e-16 ***
```

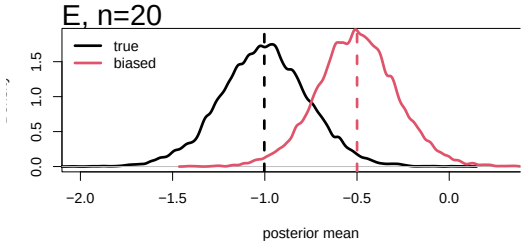
The data tell-tell story!!

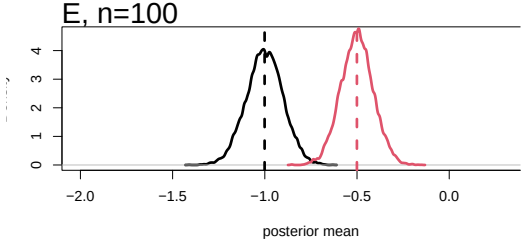
imagine we can continue sampling,

- top: $10,000$ samples $n = 20$
- bottom: $10,000$ samples $n = 100$

under the incorrect model,
the larger the sample size,

- the more certain you are about your biased estimates





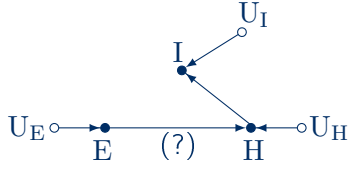
The dream team!!

based on DAG and statistical model,

- stratifying (controlling) on I, we are already explaining variability in H (I is a descendant of H)
- the remaining is explained by E (big chunk is already explained)

$$M = \begin{cases} E \leftarrow f_E(U_E) \\ H \leftarrow f_H(E, U_H) \\ I \leftarrow f_I(H, U_I) \\ U \sim P(\mathbf{U}) \end{cases}$$

(a) structural model



(b) causal diagram

the dream team!!

based on DAG and statistical analysis,

- the less biased model is the first,
  (assuming our DAG is true)

```
> summary(lm(H ~ E, data=d)) # unbiased effects

Call:
lm(formula = H ~ E, data = d)

Residuals:
     Min       1Q   Median       3Q      Max
-2.19319 -0.60621 -0.06694  0.55674  2.77776

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.04689    0.09534   0.492    0.624
E           -0.90223    0.08930 -10.104   <2e-16 ***
```

So, what is going on?

descendant1_triptych.pdf

4. Concluding remarks

Concluding remarks

- Research is filled with challenges
  (you: duh!!)
- Statistical models are not theory
  (you: so obvious again!!)
- Don't trust your statistical model when no DAG is involved
  (me: how about that?!)
- For explanation, without a DAG the (sample) size does not matter
  (me: sorry?!)
- For prediction, sometimes a DAG can help
  (me: did you expect this one?!)

Concluding remarks

Additionally, simulation can serve,

- a reflection of a hypothesis, and its research complexities
  (me: DAGs baby!!)
- a place where you can reflect the status of a population
  (test what happens if??, post-stratification??)
- as a data where you can test your statistical model on various purposes
  (parameter recovery, power?)

5. Do you wanna know more???

5. Do you wanna know more???

[1] Anderson, D. [2008]. Model Based Inference in the Life Sciences: A Primer on Evidence, Springer.

[2] Bareinboim, E. and Pearl, J. [2016]. Causal inference and the data-fusion problem, Proceedings of the National Academy of Sciences 113(27): 7345–7352. doi: https://doi.org/10.1073/pnas.1510507113.

[3] Chamberlain, T. [1965]. The method of multiple working hypotheses, Science 148(3671): 754–759.
url: https://www.jstor.org/stable/1716334.

[4] Cinelli, C., Forney, A. and Pearl, J. [2021]. A crash course in good and bad controls, Technical report.

[5] Cunningham, S. [2022]. Causal inference: The mixtape.
url: https://mixtape.scunning.com/index.html.

[6] Fogarty, L., Madeleine, A., Holding, T., Powell, A. and Kandler, A. [2022]. Ten simple rules for principled simulation modelling, PLOS Computational Biology 18(3): 1–8.
doi: https://doi.org/10.1371/journal.pcbi.1009917.

[7] Hanck, C., Arnold, M., Gerber, A. and Schmelzer, M. [2021]. Introduction to econometrics with r.
url: https://www.econometrics-with-r.org/index.html.

[8] Hernán, M. [2020]. Causal diagrams: Draw your assumptions before your conclusions.
url: https://www.edx.org/course/causal-diagrams-draw-your-assumptions-before-your.

[9] Hernán, M. and Robins, J. [2020]. Causal Inference: What If, 1 edn, Chapman and Hall/CRC.
url: https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book.

[10] Jaynes, E. [2003]. Probability Theory: The Logic of Science, Cambridge University Press.

[11] McElreath, R. [2019]. Statistical rethinking, 2019 course.
url: https://github.com/rmcelreath/statrethinking_winter2019.

[12] McElreath, R. [2020]. Statistical Rethinking: A Bayesian Course with Examples in R and STAN, Chapman and Hall/CRC.

[13] McElreath, R. [2022]. Statistical rethinking, 2022 course.
url: https://github.com/rmcelreath/stat_rethinking_2022.

[14] Pearl, J. [1988]. Probabilistic reasoning in intelligent systems: Networks of plausible inference, The Journal of Philosophy 88(8): 434–437.
doi: https://doi.org/10.2307/2026705.
url: https://www.jstor.org/stable/2026705.

[15] Pearl, J. [2009]. Causality: Models, Reasoning and Inference, Cambrige University Press.

[16] Pearl, J. [2019]. The seven tools of causal inference, with reflections on machine learning, Communications of the ACM 62(3): 54–60.
doi: https://doi.org/10.1177/0962280215586010.

[17] Pearl, J., Glymour, M. and Jewell, N. [2016]. Causal Inference in Statistics: A Primer, John Wiley Sons, Inc.

[18] Pearl, J. and Mackenzie, D. [2018]. The Book of Why: The New Science of Cause and Effect, 1st edn, Basic Books, Inc.

[19] Spirtes, P., Glymour, C. and Scheines, R. [1991]. From probability to causality, Philosophical Studies 64(1): 1–36.
url: https://www.jstor.org/stable/4320244.

[20] Textor, J., van der Zander, B., Gilthorpe, M., Liskiewicz, M. and Ellison, G. [2016]. Robust causal inference using directed acyclic graphs: the r package 'dagitty', Int J Epidemiol 45(6): 1887–1894.
doi: https://doi.org/10.1093/ije/dyw341.