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)

April 28, 2022

What are we going to talk about? I

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

What are we going to talk about? II

- No more fork bias: neutral control
- Pipe bias: masked relationships

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

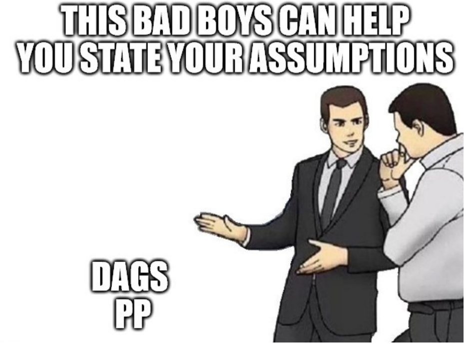
What we “normally” focus on?

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]

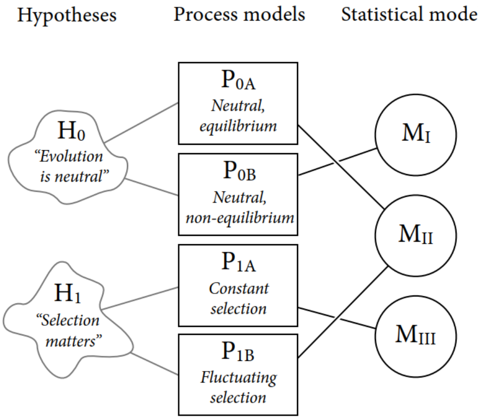
- Insufficient data
- Wrong population
- Measurement error
- Selection bias
- Confounding

[a] Hernán [8], lesson 4

Research hypothesis production

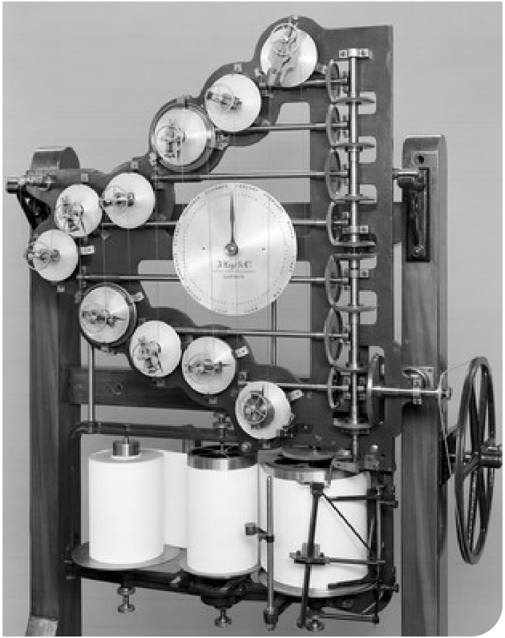
but we should not forget[a]

- 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]

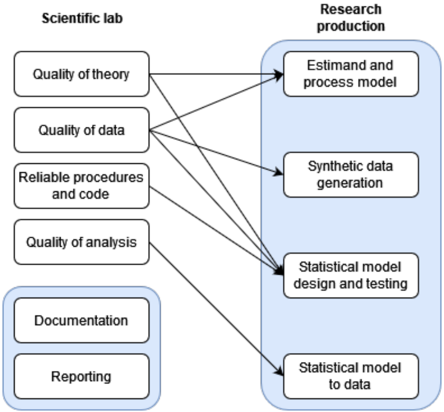
a. Estimand and process model
b. Synthetic data generation
c. Statistical model design and testing
d. Apply statistical model to data

[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?

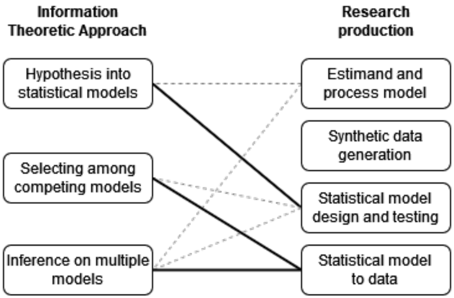
a. Estimand and process model
   maps 1 (theory) and 2 (data) to a heuristic model.
b. Synthetic data generation
   maps 2 (data) to an idealized data.
c. Statistical model design and testing
   maps 1 (theory), 2 (data), and 3 (reliable code) to an statistical model.
d. Apply statistical model to data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



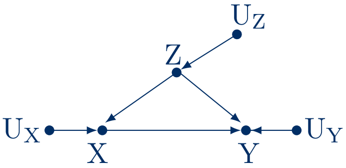
2. DAGs and PP

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)

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

(a) structural model



(b) causal diagram

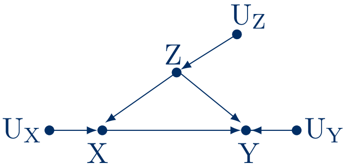
[a]reproduced from Cinelli et al. [4].

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

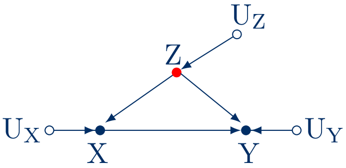
DAGs and PP

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

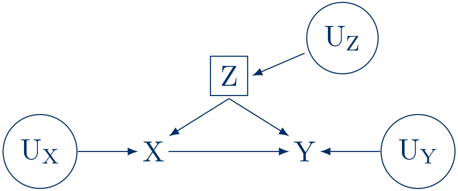
DAGs and PP

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.

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

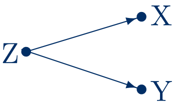
(a) structural model



(b) causal diagram

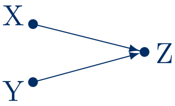
The benign case of DAG elementals

For everything can be depicted with them

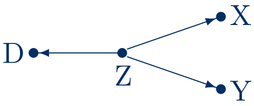


(a) fork

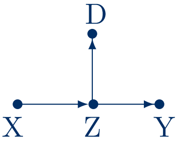
(b) pipe



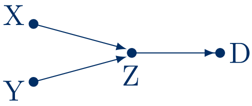
(c) collider



(d) descendant on fork



(e) descendant on pipe



(f) descendant on collider

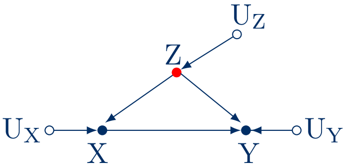
About D-separation

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

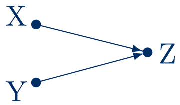


(b) causal diagram

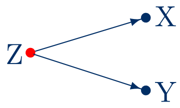
About D-separation

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.

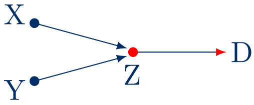


2. Any path that contains a noncollider that has been conditioned on, is blocked (backdoor path)[a].



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.



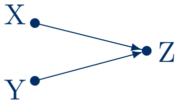
---

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

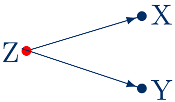
About D-separation

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

1. $X \perp\!\!\!\perp Y \Longrightarrow$
   $P(X, Y) = P(X) \cdot P(Y)$

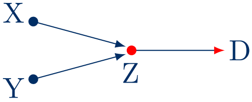


2. $X \perp\!\!\!\perp Y |\ Z \Longrightarrow$
   $P(X, Y|Z) = P(X|Z) \cdot P(Y|Z)$
   (same for fork or pipe)



3. $X \not\perp\!\!\!\perp Y |\ Z \Longrightarrow$
   $P(X, Y|Z) \neq P(X|Z) \cdot P(Y|Z)$

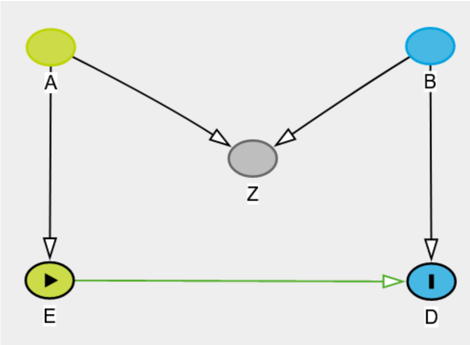
4. $X \not\perp\!\!\!\perp Y |\ D \Longrightarrow$
   $P(X, Y|D) \neq P(X|D) \cdot P(Y|D)$



University of Antwerp
Faculty of Social Sciences

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 $$$)



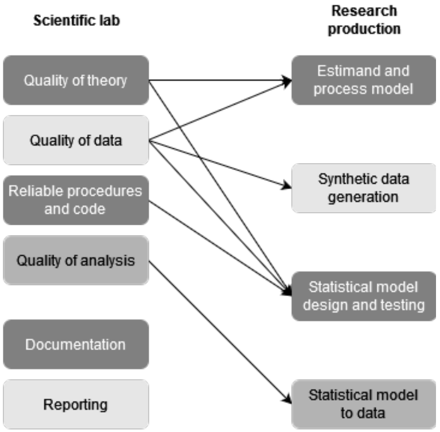
Where do DAGs and PP fit?

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. Anecdotal cases

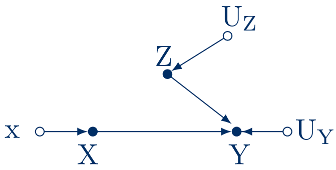
Experimental design: the panacea

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)

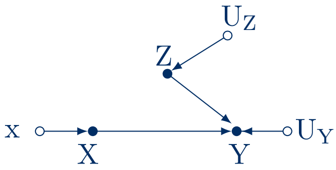
Experimental design

- 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,

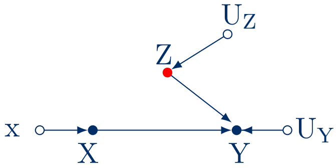
1. Average causal effect:
   $\text{ACE}(x) = E[Y|do(x+1)] - E[Y|do(x)]$
2. Controlled direct effect:
   $\text{CDE}(x, z) = E[Y|do(x+1), do(z)] - E[Y|do(x), do(z)]$

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. Anecdotal cases

Simulation conventions

Simulation conventions

one way to defined it,

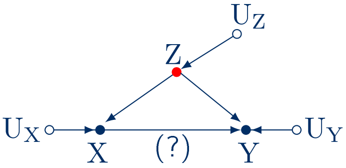
$$
\begin{aligned}
Z &= U_Z && ;U_Z \sim N(0, \sigma_Z)\\
X &= \beta_Z Z + U_X && ;U_X \sim N(0, \sigma_X)\\
Y &= \beta_Z Z + \beta_X X + U_Y && ;U_Y \sim N(0, \sigma_Y)
\end{aligned}
$$

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. Anecdotal cases

Fork bias: spurious relationships

Spurious relationships[4]

also known as,

- spurious association
- confounder
- an instance of fork bias

research question,

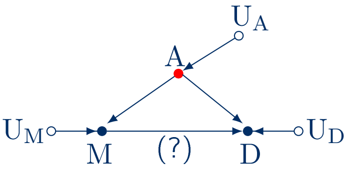
- Does M has a (direct) effect on D?

variables,

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

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

(a) structural model



(b) causal diagram

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

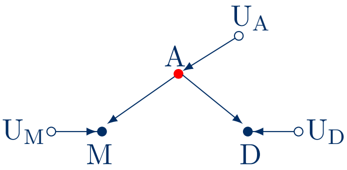
Simulation setting

```
# 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

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

(a) structural model



(b) causal diagram

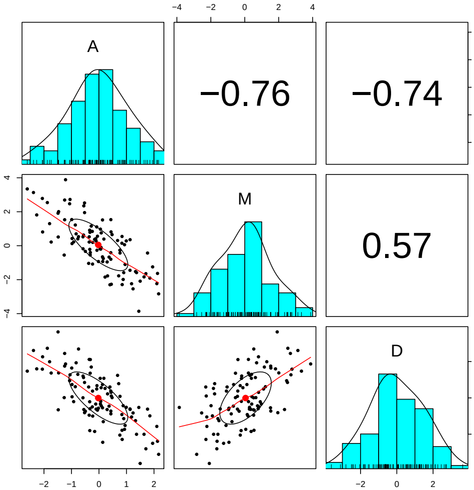
Implications,

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

"Eyeballing" analysis

based on correlation analysis,

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



Regression, regression!!

based on statistical analysis,

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

```
> 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    
```

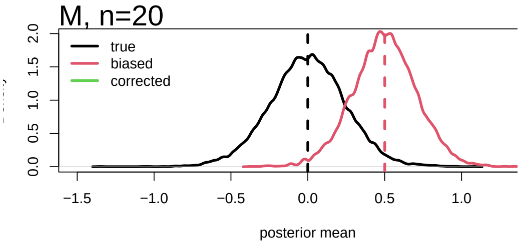
I'll get more data!!

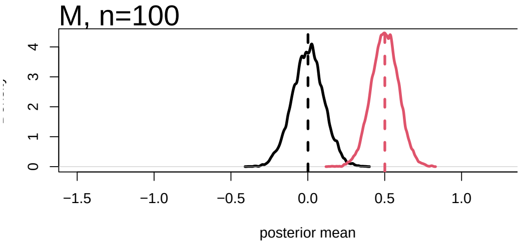
imagine we can continue sampling,

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

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 winner's curse)





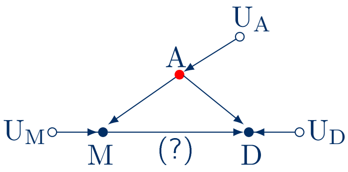
The dream team!!

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    
```

3. Anecdotal cases

Fork bias: masked relationships (a)

Masked relationships (a)[5]

also known as,

- omitted variable bias
- an instance of fork bias

research question,

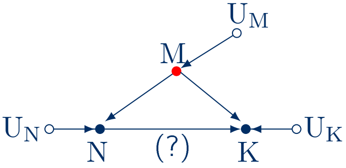
- Does N has a (direct) effect on K?

variables,

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

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

(a) structural model



(b) causal diagram

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

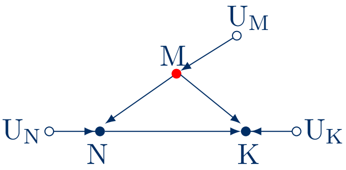
Simulation setting

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

(a) structural model

```
# 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



(b) causal diagram

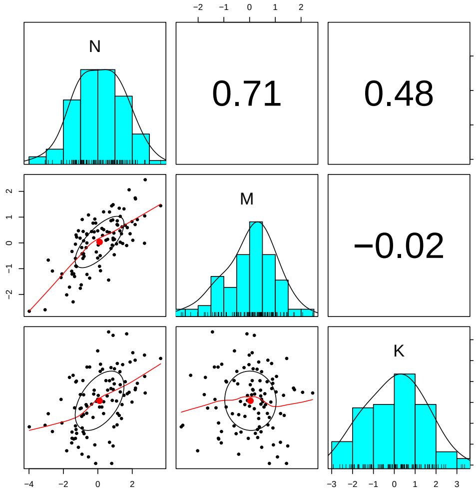
Implications,

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

“Eyeballing” analysis

based on correlation analysis,

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



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 
-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 ***
```

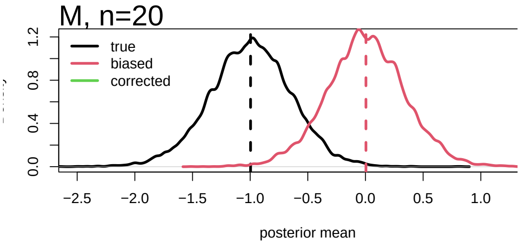
I'll get more data!!

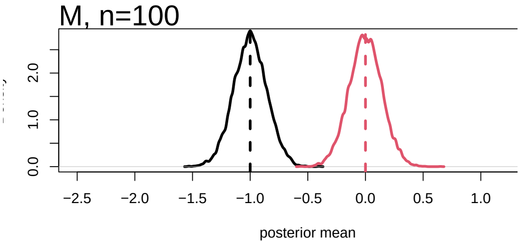
imagine we can continue sampling,

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

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)





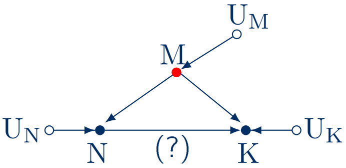
The dream team!!

based on DAG and statistical model,

- the 2nd D-separation rule requires you to control any noncollider to block the backdoor path, i.e. $N \not\perp\!\!\!\perp K \mid M$
- 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 ***
```

3. Anecdotal cases

Fork bias: masked relationships (b)

Masked relationships (b)[6]

also known as,

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

research question,

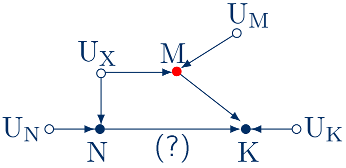
- Does N has a (direct) effect on K?

variables,

- U_X, unobservables (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

```
# 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)
```

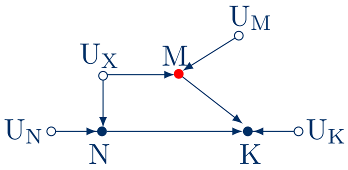
(c) R code

Implications,

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

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

(a) structural model

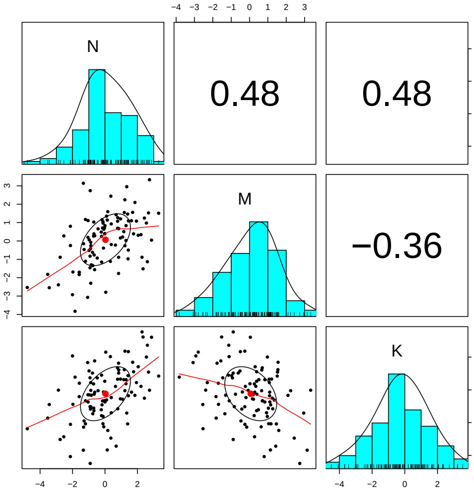


(b) causal diagram

"Eyeballing" analysis

based on correlation analysis,

- $\text{cor}(N, K) > 0$ goes in line of our "rudimentary" understanding of the data.
- $\text{cor}(M, K) < 0$ 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,

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

```
> 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 ***
```

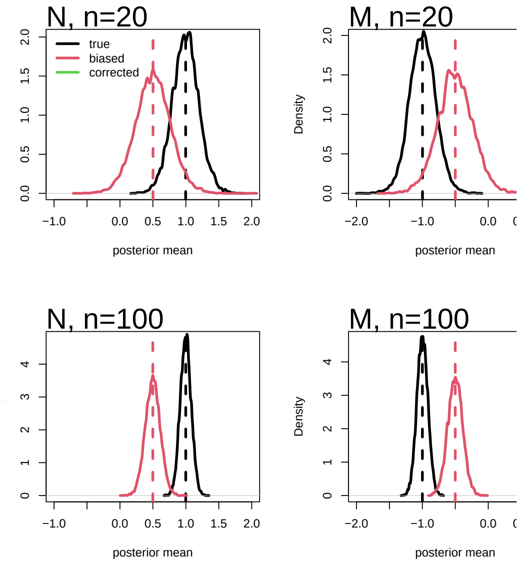
I'll get more data!!

imagine we can continue sampling,

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

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)



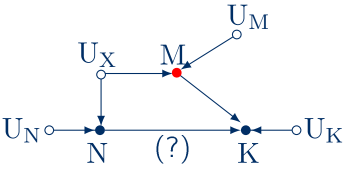
The dream team!!

based on DAG and statistical model,

- the 2nd D-separation rule requires 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(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 ***
```

Similar scenario[7]

research question,

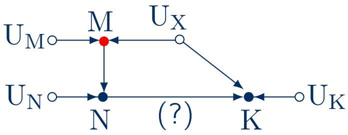
- Does N has a (direct) effect on K?

variables,

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

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

(a) structural model



(b) causal diagram

[7] Cinelli et al. [4] (p. 3)

3. Anecdotal cases

Fork bias: bias amplification

Bias amplification[8]

also known as,

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

research question,

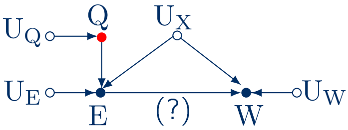
- Does E has a (direct) effect on W?

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

[8] 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)
```

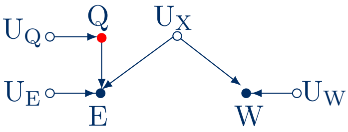
(c) R code

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

(a) structural model

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)

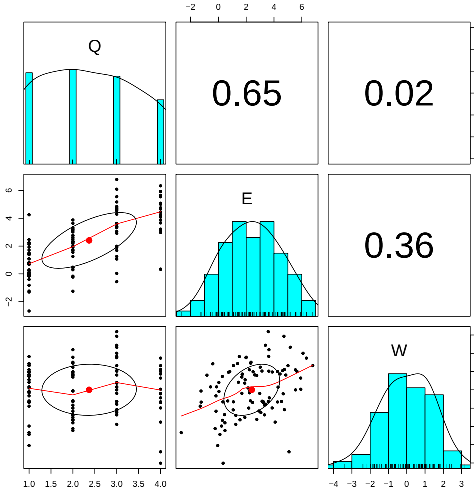


(b) causal diagram

"Eyeballing" analysis

based on correlation analysis,

- cor(Q, E) $>$ 0 and cor(E, W) $>$ 0 goes in line of our "rudimentary" understanding of the data.
- cor(Q, 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 regressions with two different results, which model is the "true"?
- one is "worse"/"better" than the other

```
> 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 ** 
```

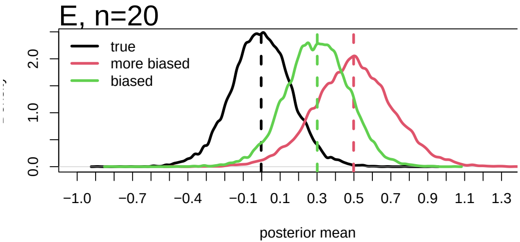
I'll get more data!!

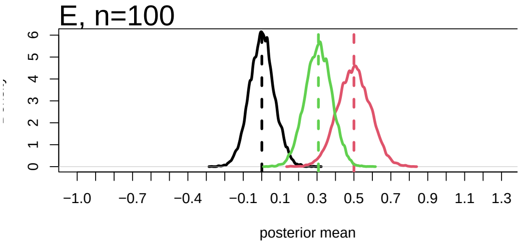
imagine we can continue sampling,

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

the larger the sample size,

- the more certain you are about your estimates
- the more mistaken you are about your research question
  (under the any model!!)
  (the winner's curse)





University of Antwerp
Faculty of Social Sciences

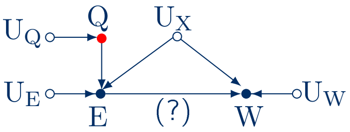
Yo, what is going on??

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 it is impossible)
- 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

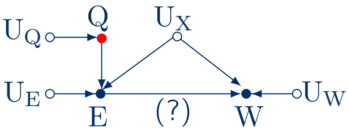
Yo, what is going on??

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

(a) structural model

open paths?:

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



(b) causal diagram

What should I do then??

$$\begin{pmatrix} W \\ E \end{pmatrix} \sim \mathrm{MVN}\left[\begin{pmatrix} \mu_W \\ \mu_E \end{pmatrix}, \boldsymbol{\Sigma}\right]$$

$$\mu_W = \alpha_W + \beta_{EW} E$$

$$\mu_E = \alpha_E + \beta_{QE} E$$

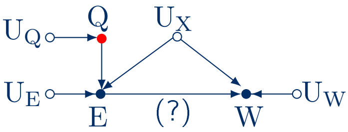
(c) probabilistic model

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

(a) structural model

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

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



(b) causal diagram

did it worked???

based on DAG and bayesian statistical analysis,

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

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

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

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

	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

3. Anecdotal cases

No more fork bias: neutral control

Neutral control[9]

also known as,

- precision "booster"
- similar to experimental design

research question,

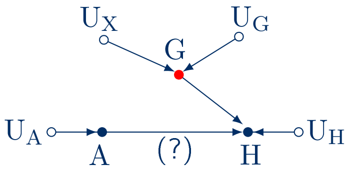
- Does N has a (direct) effect on K?

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

```
# 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)
```

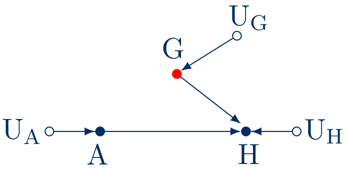
(c) R code

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

(a) structural model

Implications,

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

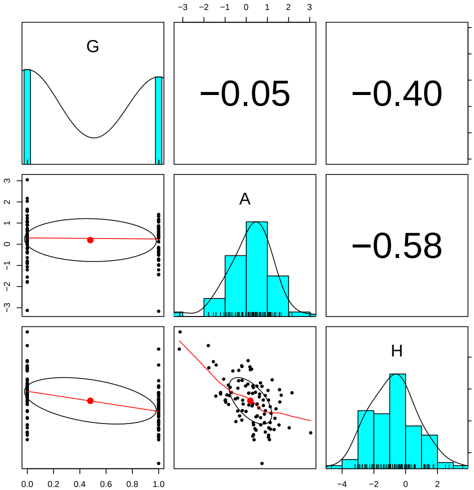


(b) causal diagram

“Eyeballing” analysis

based on correlation analysis,

- cor(G, H) ≈ 0, cor(G, A) ≈ 0 and cor(A, 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,

- now there is no severe biasing
- notice the standard errors, lower for A when G is included

```
> 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 ***
```

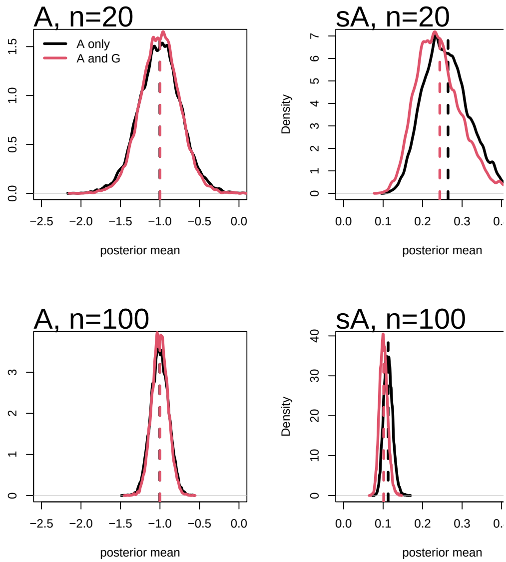
I'll get more data!!

imagine we can continue sampling,

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

the larger the sample size,

- the (more) certain you are about your estimates
- the more correct you are about your research question (under the any model)



3. Anecdotal cases

Pipe bias: masked relationships

Masked relationships[10]

also known as,

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

research question,

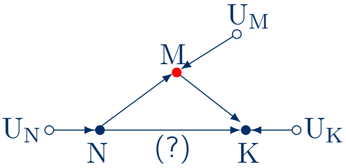
- Does N has a (direct) effect on K?

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

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

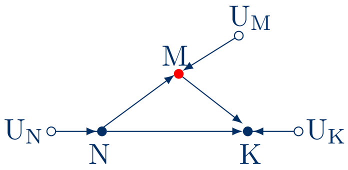
Simulation setting

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

(a) structural model

```
# 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



(b) causal diagram

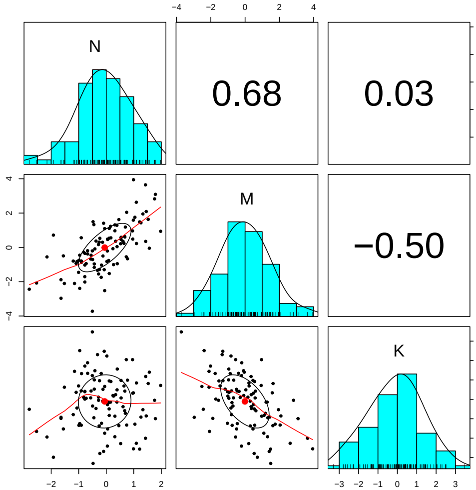
Implications,

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

"Eyeballing" analysis

based on correlation analysis,

- $\operatorname{cor}(M, K) < 0$ does NOT goes in line of our "rudimentary" understanding of the data.
- and why there is $\operatorname{cor}(N, K) \approx 0$? (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 ***
```

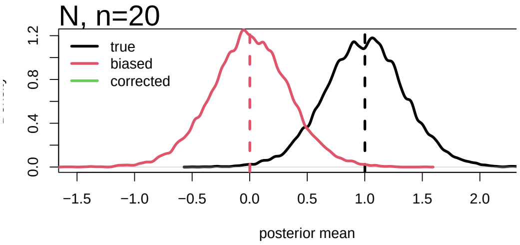
I'll get more data!!

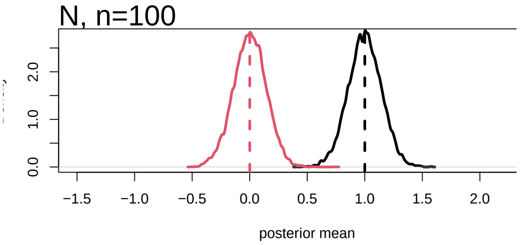
imagine we can continue sampling,

- 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)





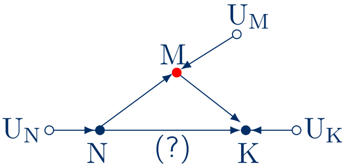
The dream team!!

based on DAG and statistical model,

- the 2nd D-separation rule requires you to control any noncollider to block the backdoor path, i.e. $N \not\perp\!\!\!\perp K \mid M$
- 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)

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

(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 ***
```

4. Concluding remarks

Concluding remarks

- Research is filled with challenges, some obvious, some not
  (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, no sample size can save you when no DAG is involved
  (me: booya?!)
- For prediction, sometimes a DAG can help
  (me: did you expect this one?!)

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.