Causal Inference II

PSY517 Quantitative Analysis III

Derek Powell

Module 4

Causal Inference II

2021-10-07

Causal Inference II

PSY517 Quantitative Analysis III

Derek Powell

Module 4

Failing to condition

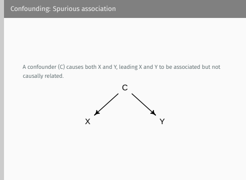
Failing to condition

Confounding: Spurious association

A confounder (C) causes both X and Y, leading X and Y to be associated but not causally related.

$$\begin{tikzcd}
 & C \arrow[dl] \arrow[dr] & \\
X & & Y
\end{tikzcd}$$

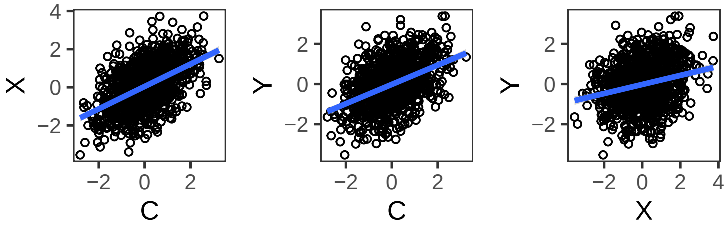
2021-10-07

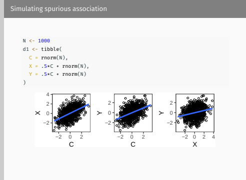


got to slide 33 in last lecture

Simulating spurious association

```r
N <- 1000
d1 <- tibble(
  C = rnorm(N),
  X = .5*C + rnorm(N),
  Y = .5*C + rnorm(N)
)
```





Regressions can de-confound

```
coef(lm(Y ~ X, data = d1))
```

```
## (Intercept)           X
## -0.02348444  0.22632165
```

```
coef(lm(Y ~ C + X, data = d1))
```

```
## (Intercept)           C           X
## -0.02093400  0.46475515  0.02750676
```


Regressions can de-confound

coef(lm(Y ~ X, data = d1))
(Intercept)   X
-0.02348444 0.22632165
coef(lm(Y ~ C + X, data = d1))
(Intercept)   C   X
-0.02093400 0.46475515 0.02750676

Confounding (masked relationship)

- Confounding can also mask relationships, making them appear weaker than they are
- Causal explanation of so-called “suppressor” variables in multiple regression

Confounding (masked relationship)
- Confounding can also mask relationships, making them appear weaker than they are
- Causal explanation of so-called “suppressor” variables in multiple regression

Primate milk

- Milk is a costly physiological investment for mammals
- Brains are also a costly physiological investment
- A popular hypothesis is that primates with larger brains produce more energetic milk to support brain growth.
- Humans are unique in having a larger brain and more developed neocortex than other primates (and mammals)

Primate milk
- Milk is a costly physiological investment for mammals
- Brains are also a costly physiological investment
- A popular hypothesis is that primates with larger brains produce more energetic milk to support brain growth.
- Humans are unique in having a larger brain and more developed neocortex than other primates (and mammals)

- comparative, evolutionary anthropology
- also unique for how long it takes human infants to develop
- primate milk is relatively dilute, because primate infants suckle frequently and for long periods of growth/development, and humans are a fairly extreme case of this

We will focus on 3 variables:

- `mass`: Average body mass of adult female (Kg)
- `neocortex.perc`: Percent of brain mass that is neocortex ("grey matter")
- `kcal.per.g`: Milk energy density (Kcal/g)

```
data("milk", package = "rethinking")
glimpse(milk)
```

```
## Rows: 29
## Columns: 8
## $ clade          <fct> Strepsirrhine, Strepsirrhine, Strepsirrhine, Strepsirrh~
## $ species        <fct> Eulemur fulvus, E macaco, E mongoz, E rubriventer, Lemu~
## $ kcal.per.g     <dbl> 0.49, 0.51, 0.46, 0.48, 0.60, 0.47, 0.56, 0.89, 0.91, 0~
## $ perc.fat       <dbl> 16.60, 19.27, 14.11, 14.91, 27.28, 21.22, 29.66, 53.41,~
## $ perc.protein   <dbl> 15.42, 16.91, 16.85, 13.18, 19.50, 23.58, 23.46, 15.80,~
## $ perc.lactose   <dbl> 67.98, 63.82, 69.04, 71.91, 53.22, 55.20, 46.88, 30.79,~
## $ mass           <dbl> 1.95, 2.09, 2.51, 1.62, 2.19, 5.25, 5.37, 2.51, 0.71, 0~
## $ neocortex.perc <dbl> 55.16, NA, NA, NA, NA, 64.54, 64.54, 67.64, NA, 68.85, ~
```

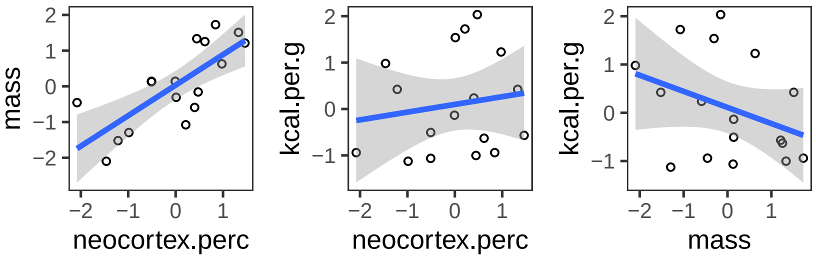
Transforming variables

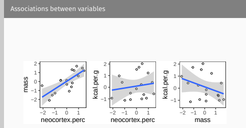
```r
milk <- milk %>%
  mutate(
    mass = log(mass)
  ) %>%
  mutate_at(vars(mass,neocortex.perc,kcal.per.g), standardize) %>%
  drop_na(neocortex.perc)
```


Transforming variables

```r
milk <- milk %>%
  mutate(
    mass = log(mass)
  ) %>%
  mutate_at(vars(mass,neocortex.perc,kcal.per.g), standardize) %>%
  drop_na(neocortex.perc)
```

Associations between variables

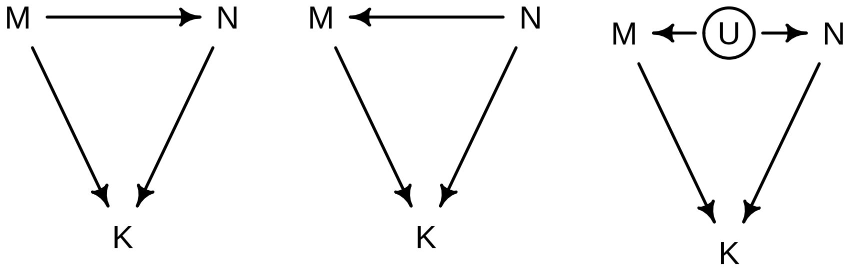




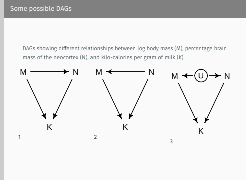
- Remember, independence is the strong assumption
- So even though the relationships look weak or uncertain, there could be associations among all these variables, so not clearly independent.

Some possible DAGs

DAGs showing different relationships between log body mass (M), percentage brain mass of the neocortex (N), and kilo-calories per gram of milk (K).



1 2 3



Comparing regressions

Model 1: `brm(kcal.per.g ~ mass, data = milk)`

```
##           Estimate Est.Error   Q2.5 Q97.5
## Intercept    0.107     0.274 -0.420 0.653
## mass        -0.338     0.247 -0.827 0.141
```

Model 2: `brm(kcal.per.g ~ neocortex.perc, data = milk)`

```
##                Estimate Est.Error   Q2.5 Q97.5
## Intercept         0.096     0.289 -0.478 0.663
## neocortex.perc    0.162     0.307 -0.462 0.783
```

Model 3: `brm(kcal.per.g ~ neocortex.perc + mass, data = milk)`

```
##                Estimate Est.Error   Q2.5  Q97.5
## Intercept         0.135     0.217 -0.291  0.587
## neocortex.perc    1.033     0.337  0.370  1.721
## mass             -1.013     0.296 -1.599 -0.437
```

What is happening?

Regression is asking:

- do species that have high neocortex percent *for their body mass* have high milk energy?
- do species with high body mass *for their neocortex percent* have higher milk energy?

What is happening?
Regression is asking:
- do species that have high neocortex percent *for their body mass* have high milk energy?
- do species with high body mass *for their neocortex percent* have higher milk energy?

- Body mass is positively correlated with neocortex percentage
- Body mass is negatively correlated with milk energy (Kcal)
- Neocortex percentage is positively correlated with milk energy

Overadjustment: conditioning too much

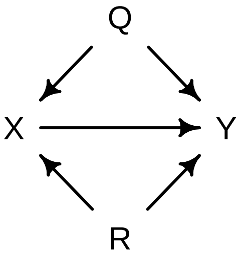
Overadjustment: conditioning too much

Causal salad

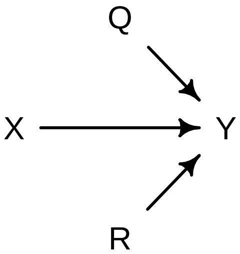
- Some researchers were trained to include as many covariates in a regression model as possible in order to “control for” as much as possible—toss everything into the salad
- But this is wrong!
- Controlling for the wrong covariates can be just as bad as failing to control for the right ones

Causal salad
- Some researchers were trained to include as many covariates in a regression model as possible in order to “control for” as much as possible—toss everything into the salad
- But this is wrong!
- Controlling for the wrong covariates can be just as bad as failing to control for the right ones

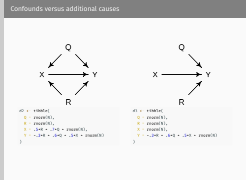
Confounds versus additional causes



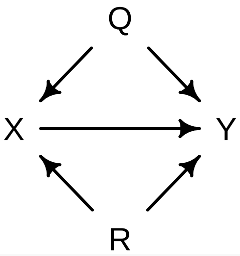

```r
d2 <- tibble(
  Q = rnorm(N),
  R = rnorm(N),
  X = .5*R + .7*Q + rnorm(N),
  Y = -.3*R + .6*Q + .5*X + rnorm(N)
)
```




```r
d3 <- tibble(
  Q = rnorm(N),
  R = rnorm(N),
  X = rnorm(N),
  Y = -.3*R + .6*Q + .5*X + rnorm(N)
)
```

Only need to control for confounds



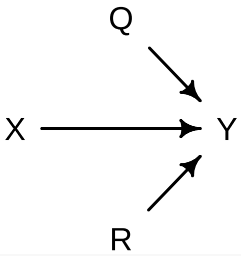

```r
fit_conf1 <- brm(Y ~ X, data = d2)
fit_conf2 <- brm(Y ~ X + Q + R , data = d2)
```

```r
fixef(fit_conf1) %>% round(3)
```

```
##           Estimate Est.Error   Q2.5 Q97.5
## Intercept   -0.011     0.038 -0.084 0.063
## X            0.659     0.028  0.604 0.714
```

```r
fixef(fit_conf2) %>% round(3)
```

```
##           Estimate Est.Error   Q2.5  Q97.5
## Intercept   -0.031     0.031 -0.092  0.031
## X            0.528     0.031  0.468  0.591
## Q            0.625     0.038  0.549  0.698
## R           -0.337     0.035 -0.405 -0.268
```



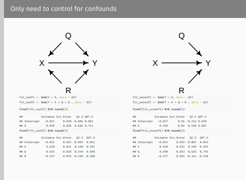

```r
fit_unconf1 <- brm(Y ~ X, data = d3)
fit_unconf2 <- brm(Y ~ X + Q + R , data = d3)
```

```r
fixef(fit_unconf1) %>% round(3)
```

```
##           Estimate Est.Error   Q2.5 Q97.5
## Intercept   -0.037      0.04 -0.114 0.040
## X            0.529      0.04  0.450 0.607
```

```r
fixef(fit_unconf2) %>% round(3)
```

```
##           Estimate Est.Error   Q2.5  Q97.5
## Intercept   -0.022     0.033 -0.087  0.043
## X            0.530     0.031  0.469  0.593
## Q            0.690     0.033  0.626  0.754
## R           -0.277     0.034 -0.344 -0.210
```

- but including additional causes as covariates does reduce our uncertainty in our estimates

To control or not to control

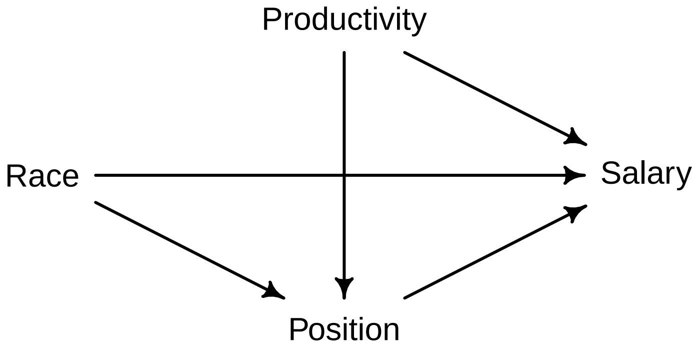
- Suppose we are interested in how race (white or non-white) affects salary within a firm.
- We can improve the fit of our model of salaries by adding covariates:
  - Each employee's productivity
  - Each employee's position within the company (manager and non-manager).
- Should we control for these factors to estimate the causal effect of race on salary?

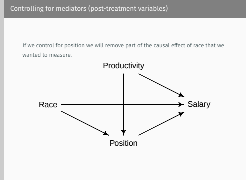
To control or not to control

- Suppose we are interested in how race (white or non-white) affects salary within a firm.
- We can improve the fit of our model of salaries by adding covariates:
  - Each employee's productivity
  - Each employee's position within the company (manager and non-manager).
- Should we control for these factors to estimate the causal effect of race on salary?

Controlling for mediators (post-treatment variables)

If we control for position we will remove part of the causal effect of race that we wanted to measure.





- remember, we don't need to control for covariates that are not confounders, so fine to leave position out of our regression to get the total effect of race

Sidebar: Race as a causal force

- We are thinking of causes in terms of counterfactuals: we mentally imagine changing that one factor and only that factor.
  - If $P(Y^{a=0}) \neq P(Y^{a=1})$ we say a is a cause
- Often it is really ***structural racism*** rather than ***race*** that is the cause
- Race causes lower salaries ***in a world with structural racism***.

Sidebar: Race as a causal force

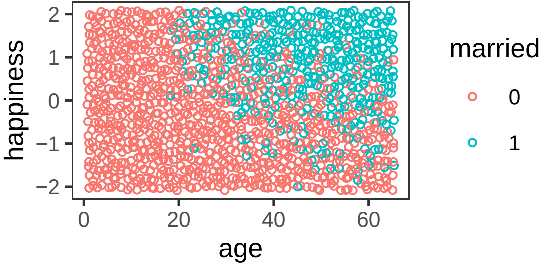
- We are thinking of causes in terms of counterfactuals: we mentally imagine changing that one factor and only that factor.
  - If $P(Y^{a=0}) \neq P(Y^{a=1})$ we say a is a cause
- Often it is really ***structural racism*** rather than ***race*** that is the cause
- Race causes lower salaries ***in a world with structural racism***.

- We are addressing only the counterfactual notion of causation, but also good as scientists to remember and consider more mechanistic views of causation.

$$H \longrightarrow M \longleftarrow A$$

Suppose

- Happiness is determined at birth, and never changes
- Each year, 100 people are born
- After age 18, have some probability of getting married each year
- Happier people are more likely to get married
- Once married, stay married



- performed an agent-basd simulation obeying these rules.
- So simulated people, simulated each year, etc.

Can see there is no association between happiness and age—in our simulation happiness is fixed across the lifespan.

Conditioning on a collider

```r
fit_happy <- brm(happiness ~ age + married, data = d5_reg)
```

```r
fixef(fit_happy)
```

```
##             Estimate  Est.Error       Q2.5      Q97.5
## Intercept -0.6106191 0.04146660 -0.6918851 -0.5304377
## age       -0.2250128 0.03382995 -0.2915924 -0.1601813
## married    1.4984901 0.06674276  1.3678832  1.6267622
```


Conditioning on a collider

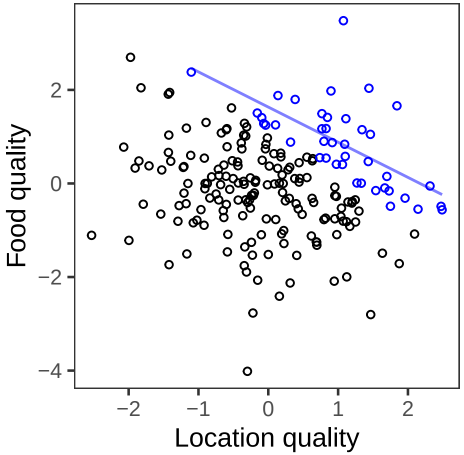
```r
fit_happy <- brm(happiness ~ age + married, data = d5_reg)
fixef(fit_happy)
```

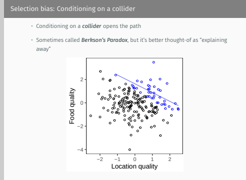
```
##             Estimate  Est.Error       Q2.5      Q97.5
## Intercept -0.6106191 0.04146660 -0.6918851 -0.5304377
## age       -0.2250128 0.03382995 -0.2915924 -0.1601813
## married    1.4984901 0.06674276  1.3678832  1.6267622
```


but if we condition on a collider by including marriage in our regression, age becomes associated with happiness

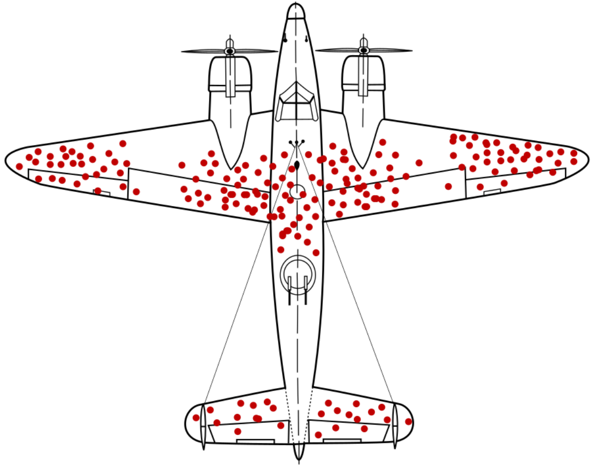
Selection bias: Conditioning on a collider

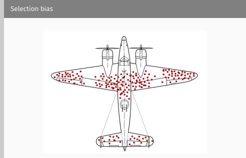
- Conditioning on a ***collider*** opens the path
- Sometimes called ***Berkson's Paradox***, but it's better thought-of as "explaining away"





Selection bias





Selection bias is generally a major problem that can do all kind of things to disrupt an analysis

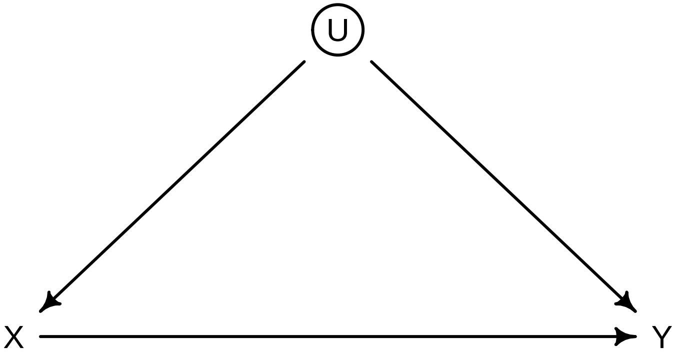
> *During World War 2, those in charge of Allied strategic bombing were trying to reduce the number of their bombers that were shot down by German fighter planes. Looking at bombers returning back from sorties they noticed that they typically had most received damage from flak and bullets in certain places, and decided that these areas should be reinforced with additional armour. However, a statistician, Abraham Wald, made the counterintuitive suggestion that instead they reinforce the areas where they saw no damage. As Wald pointed out, the bombers that made it home had survived the damage that they could see; damage to other parts of the plane meant that it never made it home to be inspected.* — source

- beware career advice from (wildly) successful people

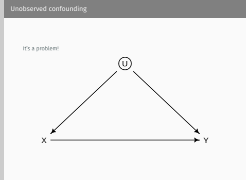
Unobserved confounding

Unobserved confounding

It’s a problem!



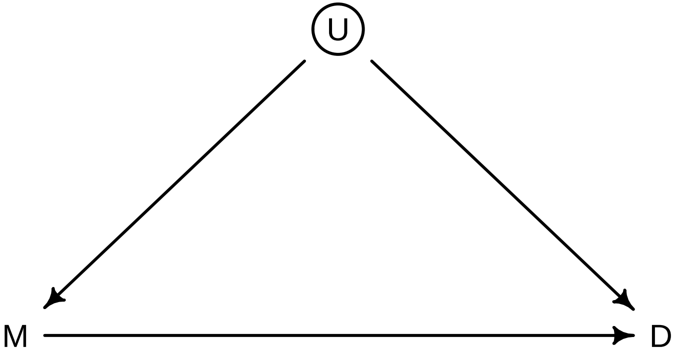
2021-10-07



can’t control for unobserved variables, so here can’t estimate effect of X on Y

Social transmission of family norms

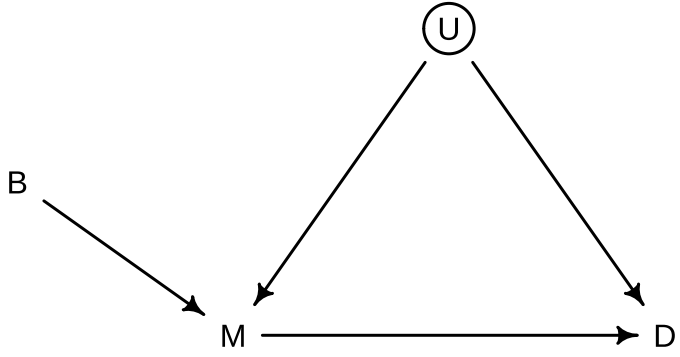
- How do family norms socially transmit within families?
- What is the influence of a mother's family size (M) on her daughter's family size (D)?
- Many potential unobserved variables (U) confound this relationship

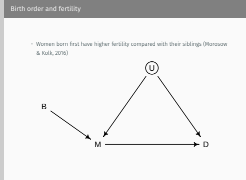


Social transmission of family norms
- How do family norms socially transmit within families?
- What is the influence of a mother's family size (M) on her daughter's family size (D)?
- Many potential unobserved variables (U) confound this relationship

Birth order and fertility

- Women born first have higher fertility compared with their siblings (Morosow & Kolk, 2016)





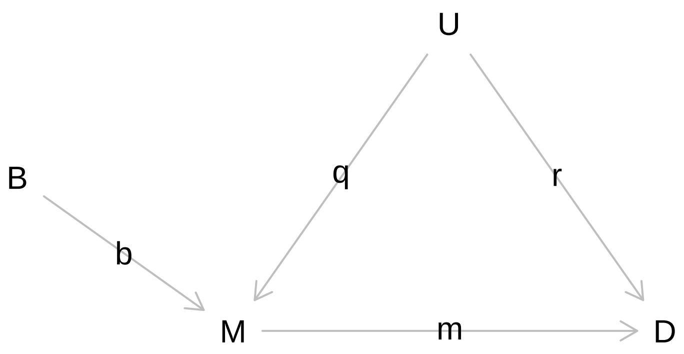
Instrumental variables

- Mother's birth order can be used as an "instrumental variable"
- IV is a parent of the cause of interest X and is independent from U and Y given X
- IV can be used to infer causal effects in the presence of unobserved confounder(s)
- **Caution:** they are often quite tricky to identify and many attempts to estimate causal effects with IVs are unsuccessful

Instrumental variables
- Mother's birth order can be used as an "instrumental variable"
- IV is a parent of the cause of interest X and is independent from U and Y given X
- IV can be used to infer causal effects in the presence of unobserved confounder(s)
- **Caution:** they are often quite tricky to identify and many attempts to estimate causal effects with IVs are unsuccessful

in economics, the weather is often used as an IV, but weather actually has many different effects, and so rarely serves the purpose correctly.

Using path tracing rules to calculate



- We want to know m, but can't estimate from regression because of unobserved confounding due to U.

Using path tracing rules to calculate
We want to know m, but can't estimate from regression because of unobserved confounding due to U.

Path tracing rules and Instrumental Variables

- Remember the path tracing rules, for instance:

$$\text{cov}(B, M) = \text{var}(B) \cdot b$$

- We can use the graph, and some algrebra, to estimate m

$$\text{cov}(B, D) = m \cdot \text{cov}(B, M)$$
$$m = \frac{\text{cov}(B, D)}{\text{cov}(B, M)}$$

Path tracing rules and Instrumental Variables
- Remember the path tracing rules, for instance:
$$\text{cov}(B, M) = \text{var}(B) \cdot b$$
- We can use the graph, and some algrebra, to estimate m
$$\text{cov}(B, D) = m \cdot \text{cov}(B, M)$$
$$m = \frac{\text{cov}(B, D)}{\text{cov}(B, M)}$$

Simulate and test

```r
set.seed(462626)
N <- 200
d6 <- tibble(
  U = rnorm(N), # unobserved confounder(s)
  B = rbernoulli(N, p = 0.5), # first-born or not
  M = rnorm(N, .5*B + .8*U ),
  D = rnorm(N, .5*M + 1.5*U )
)
```

```r
cov(d6$B, d6$D)/cov(d6$B, d6$M)
```

```
## [1] 0.5386191
```


Simulate and test
set.seed(462626)
N <- 200
d6 <- tibble(
U = rnorm(N), # unobserved confounder(s)
B = rbernoulli(N, p = 0.5), # first-born or not
M = rnorm(N, .5*B + .8*U),
D = rnorm(N, .5*M + 1.5*U)
)
cov(d6$B, d6$D)/cov(d6$B, d6$M)
[1] 0.5386191