# multilevel models

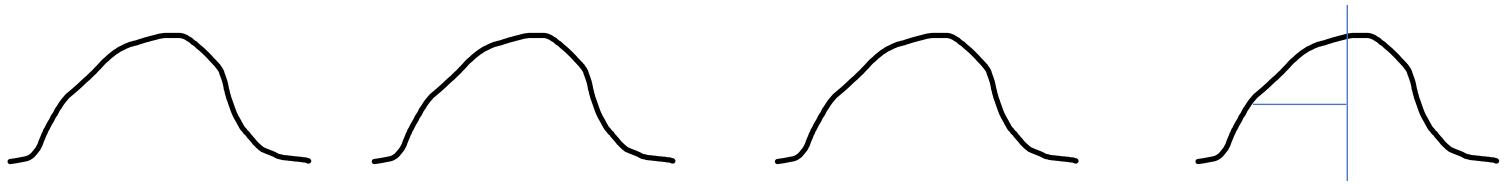
Ülo Maiväli

16.10.2021

Tallinn

| School ID (sch) | students score (Y) | students sex (X) | Schools quality index (Z) |
|---|---|---|---|
| A | 34 | F | 85 |
| A | 73 | M | 78 |
| B | 21 | M | 69 |

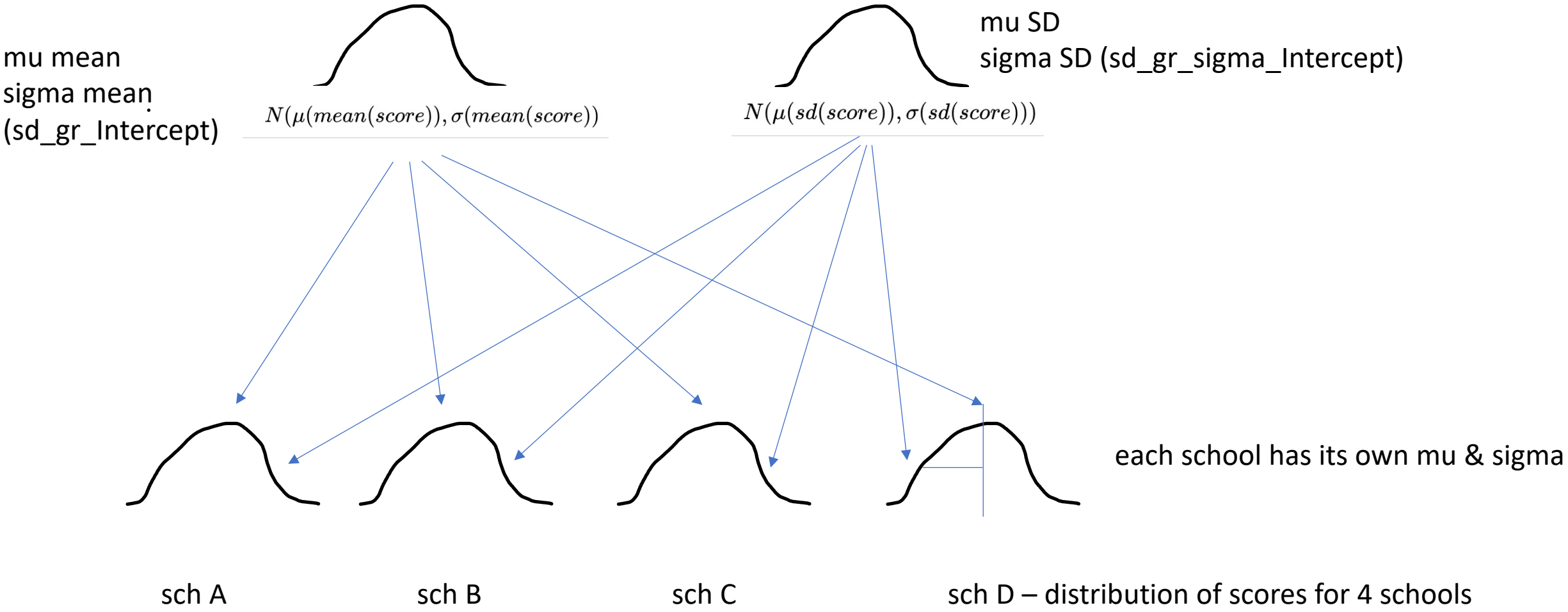
there is a single row per student



each school has its own mu & sigma

sch A sch B sch C sch D

4 distributions of scores for 4 schools




1. meta-distributions have meta-parameters which require meta-priors
2. meta-distributions are adaptive – their parameters are fitted from the lower-level pupil data. Information moves in both directions during model fitting.
3. This leads to partial pooling of information between schools & shrinkage of school estimates.
4. we can partition variation into between-schools sd(mean) and what remains at the school level (sigma_y).

# y ~ N(mu, sigma) generates y-values from a normal distribution of scores.

if sigma = 0, then all y-s are identical and equal mu. stochastic model disappears, leaving the process model

if sigma = inf, then mu loses meaning and y comes from runif(-Inf, Inf). process model disappears, leaving the stochastic model.

If Y is modelled in different groups, then each group has its mean(y) and sd(y).

we can build meta-distributions of schools means and SD-s like this

$$N(\mu(mean(score)), \sigma(mean(score))$$

$$N(\mu(sd(score)), \sigma(sd(score)))$$

we assume that schools come from 2 normal distributions of school means/sd-s.

If the meta-sigma for school means = 0, then all schools have identical means and we can ignore the sch variable in regression.

If the meta-sigma = Inf, then all schools are completely separate and there is no information about sch A in sch B and vice versa.

# single-level models are limiting cases for multilevel regression

$$y_i \sim N(\alpha, \sigma)$$

between-schools sd(mean) = 0
y ~ 1

$$within\ school\ j : y \sim N(\alpha_j, \sigma_y)$$

between-schools sd(mean) = Inf
y ~ 0 + sch

$$y_i \sim N(\alpha_j, \sigma_y)$$

$$\alpha_j \sim N(\mu_\alpha, \sigma_\alpha)$$

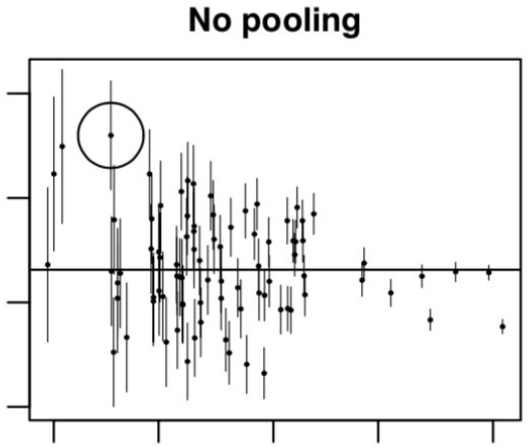
adaptive-prior contains parameters to be estimated.
Each such parameter needs a prior – a meta-prior.

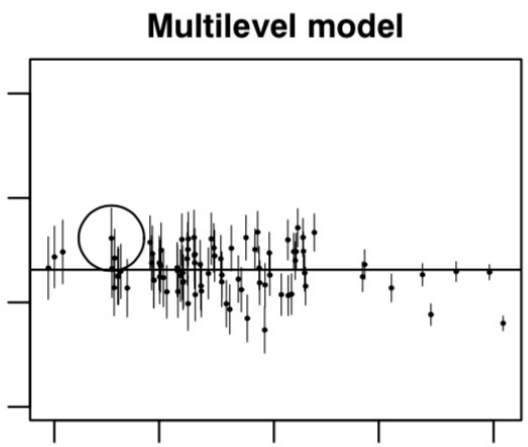
in brms this 2-level model reads as $y \sim 0 + (1|sch)$

pars that require us setting priors: sigma_y, mu_a, sigma_a (the last 2 are meta-priors for the adaptive prior for a_j)

# Suitable grouping factors

- categorical variable, whose levels are not ordered (ordered levels lead to gaussian process models)
- you must be able to imagine the groups emanating from a meta-population of groups modelled as a normal distribution
- individual-level modelling assumptions (linearity, etc.) also hold at the group level.



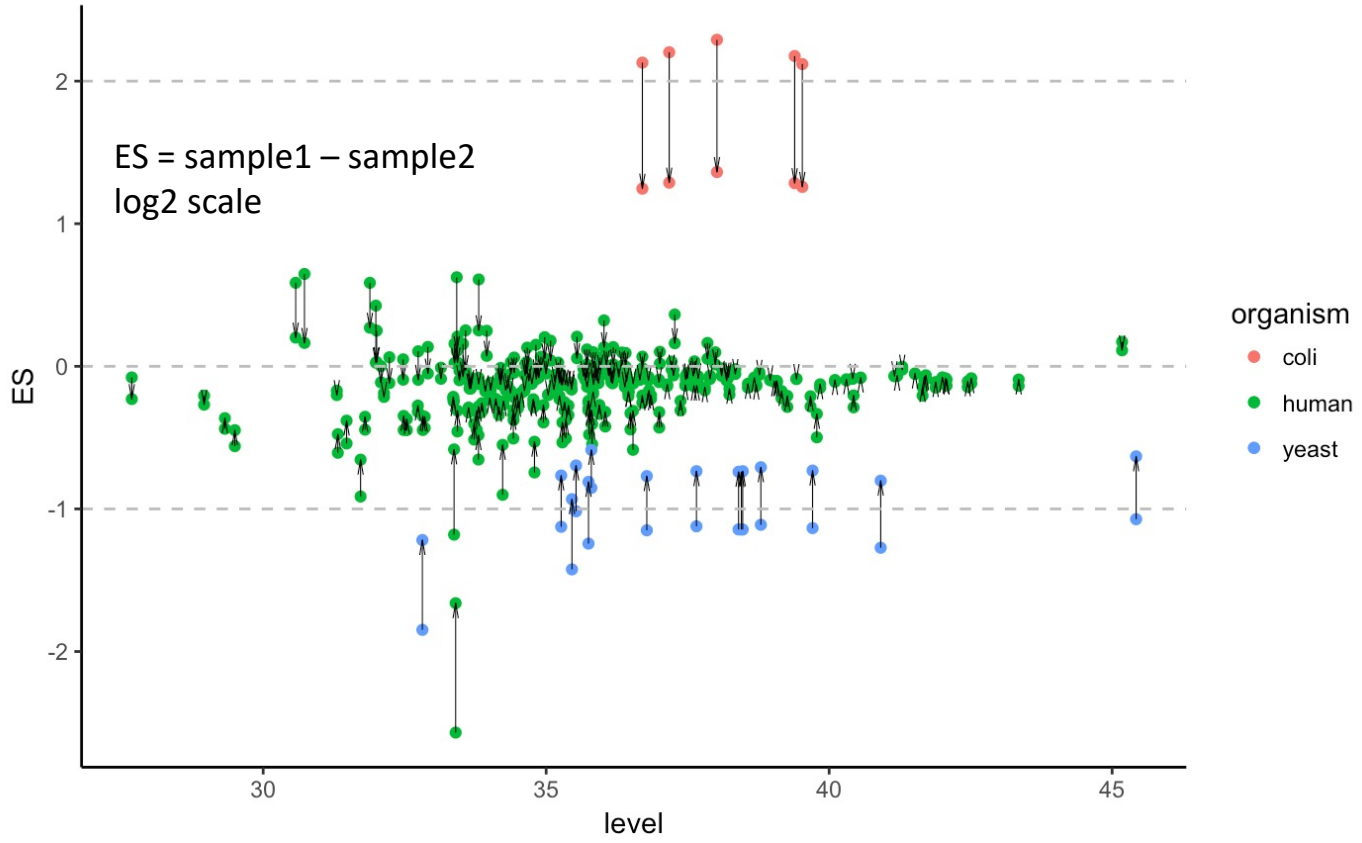

- Whereas complete pooling ignores variation between counties, the no-pooling analysis overstates it. the no-pooling analysis overfits the data within each group.
  - consider the group circled in the plot, which has the highest average level of all groups. This average is estimated using only two data points. given the variability in the data we would not have much trust in this estimate.
- looking at all the groups together, the estimates from the no-pooling model overstate the variation among groups and tend to make the individual groups look more different than they actually are.
- shrinkage estimates, instead of the raw estimates, provide more accurate estimates of the individual cluster means. They do a better job of trading off underfitting and overfitting.

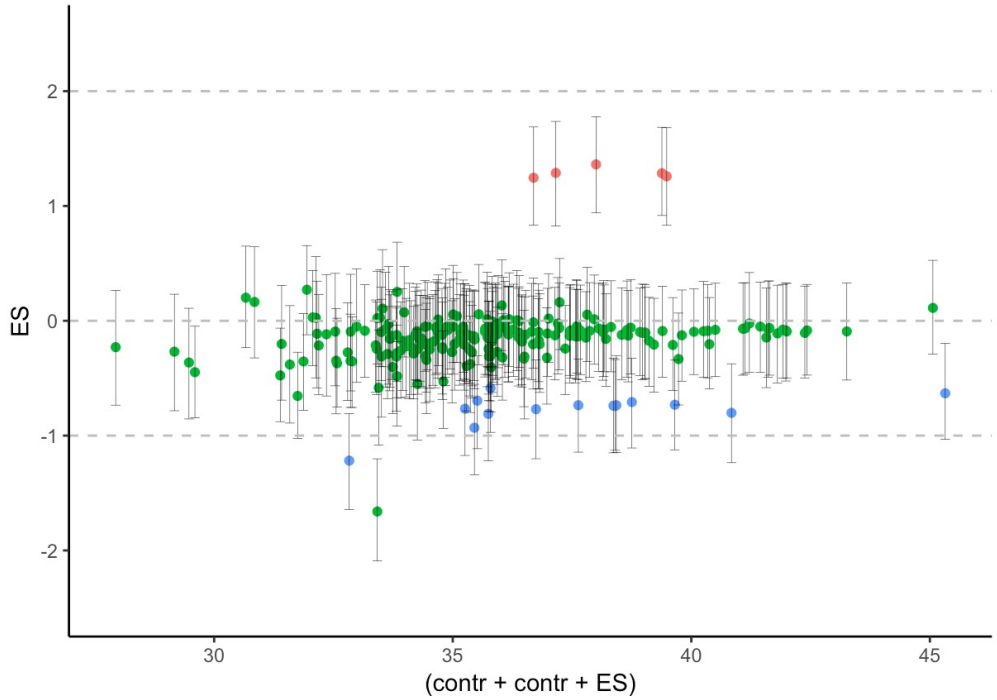
(1) **complete pooling** of information - equivalent to assuming that all groups are identical. use the overall mean to make predictions for each group. A lot of data contributes to your estimate, and so it can be quite precise. If groups differ, then the estimate is unlikely to exactly match the mean of any particular group → the total sample mean **underfits** the data.

(2) **no pooling**: no information is shared across groups. assumes that the variation among groups is infinite, so nothing you learn from one helps you predict another. In each group, little data contributes to estimates → estimates are imprecise. As a consequence, the error of these estimates is high, and they **overfit** the data.

(3) **partial pooling** produces estimates for each cluster that are less underfit than the grand mean and less overfit than the no-pooling estimates. they tend to be better estimates of the true per-cluster means. This will be especially true when groups have small N, because then the no pooling estimates will be especially overfit.

# partial pooling → shrinkage of estimates




Groups are better separated than before, but inter-group distances are smaller!

ES-s are systematically underestimated, because the assumption of a single meta-population doesn't exactly hold (here we have 3 metapopulations, in fact).

# a slightly different formulation: y ~ 1 + (1 | gr)

$$y_i \sim N(\mu, \sigma)$$

$$\mu = \alpha + \alpha_j$$

$$\alpha \sim N(.,.)$$

$$\alpha_j \sim normal(0, \sigma_{sch})$$

now we need meta-priors only for sigma_sch (and not for sch_mean, which is fixed at 0 and is therefore not estimated)

$$\sigma_{school} \sim cauchy(.)$$

$$\sigma \sim exponential()$$

value~ 0 + gr

| | Estimate | Est.Error | Q2.5 | Q97.5 |
|---|---|---|---|---|
| b_grA | -0.61128339 | 0.95023701 | -2.5134125 | 1.2127616 |
| b_grB | -0.33034594 | 0.64693662 | -1.5940719 | 0.9243307 |
| b_grC | 0.36838101 | 0.54713792 | -0.6923836 | 1.4331725 |
| b_grD | 0.37403851 | 0.47430470 | -0.5291760 | 1.3000063 |
| b_grE | 0.03373373 | 0.41208635 | -0.7591175 | 0.8211311 |
| b_grexp | 2.01003411 | 0.32191958 | 1.3751089 | 2.6376587 |
| b_grF | 0.53608821 | 0.38197606 | -0.2203967 | 1.2855347 |
| b_grG | -0.31674231 | 0.34599762 | -0.9903902 | 0.3610302 |
| b_grH | -0.03411949 | 0.31709661 | -0.6622295 | 0.5935553 |
| b_grI | 0.17440462 | 0.31658176 | -0.4444041 | 0.7936911 |
| b_grJ | 0.16032638 | 0.29218084 | -0.4302068 | 0.7312304 |
| sigma | 0.93335488 | 0.09156539 | 0.7758722 | 1.1319056 |

value~ 0 + (1|gr)

| | Estimate | Est.Error | Q2.5 | Q97.5 |
|---|---|---|---|---|
| sd_gr__Intercept | 0.70816384 | 0.22843789 | 0.3687608 | 1.2540327 |
| sigma | 0.92730507 | 0.09119723 | 0.7711151 | 1.1282743 |
| r_gr[A,Intercept] | -0.22859242 | 0.54929934 | -1.3478283 | 0.8178401 |
| r_gr[B,Intercept] | -0.18096756 | 0.46945507 | -1.1383463 | 0.7154338 |
| r_gr[C,Intercept] | 0.22726488 | 0.41484651 | -0.5623825 | 1.0503486 |
| r_gr[D,Intercept] | 0.25230562 | 0.38579430 | -0.4884552 | 1.0384005 |
| r_gr[E,Intercept] | 0.01814229 | 0.35101609 | -0.6640571 | 0.7016961 |
| r_gr[exp,Intercept] | 1.58827551 | 0.36249983 | 0.8497891 | 2.2663154 |
| r_gr[F,Intercept] | 0.40205595 | 0.34094523 | -0.2546394 | 1.0854438 |
| r_gr[G,Intercept] | -0.24090495 | 0.30564050 | -0.8278326 | 0.3637736 |
| r_gr[H,Intercept] | -0.02597655 | 0.28668904 | -0.5935078 | 0.5558989 |
| r_gr[I,Intercept] | 0.13352221 | 0.27868477 | -0.4204472 | 0.6742415 |
| r_gr[J,Intercept] | 0.13490450 | 0.27012199 | -0.4081973 | 0.6553667 |

bf(value~ 0 + (1|gr), sigma~ 0 + (1|gr))

| | Estimate | Est.Error |
|---|---|---|
| sd_gr__Intercept | 0.53796369 | 0.2847134 |
| sd_gr__sigma_Intercept | 0.28683674 | 0.1735652 |
| r_gr[A,Intercept] | -0.15630107 | 0.4763422 |
| r_gr[B,Intercept] | -0.12832313 | 0.4204251 |
| r_gr[C,Intercept] | 0.14956887 | 0.4113258 |
| r_gr[D,Intercept] | 0.20776922 | 0.3360905 |
| r_gr[E,Intercept] | 0.02509215 | 0.3600505 |
| r_gr[exp,Intercept] | 1.06281365 | 0.6563337 |
| r_gr[F,Intercept] | 0.33945511 | 0.3148671 |
| r_gr[G,Intercept] | -0.18018166 | 0.3053726 |
| r_gr[H,Intercept] | -0.02218129 | 0.2602802 |
| r_gr[I,Intercept] | 0.11979363 | 0.2412245 |
| r_gr[J,Intercept] | 0.10797926 | 0.2457000 |
| r_gr__sigma[A,Intercept] | -0.04982904 | 0.3077639 |
| r_gr__sigma[B,Intercept] | -0.06717050 | 0.2839776 |

bf(value~ 0 + (1|id|gr), sigma~ 0 + (1|id|gr))

are sch means & sch_SD-s correlated?

| | Estimate | Est.Error |
|---|---|---|
| sd_gr__Intercept | 0.53014319 | 0.2900424 |
| sd_gr__sigma_Intercept | 0.28039012 | 0.1859456 |
| cor_gr__Intercept__sigma_Intercept | 0.08150944 | 0.5055205 |
| r_gr[A,Intercept] | -0.15005580 | 0.4827728 |
| r_gr[B,Intercept] | -0.10852872 | 0.4131661 |
| r_gr[C,Intercept] | 0.14660226 | 0.3823311 |
| r_gr[D,Intercept] | 0.19086728 | 0.3341030 |
| r_gr[E,Intercept] | 0.03063959 | 0.3429154 |
| r_gr[exp,Intercept] | 0.99870194 | 0.6847423 |
| r_gr[F,Intercept] | 0.31646960 | 0.3118614 |
| r_gr[G,Intercept] | -0.16819192 | 0.2954687 |
| r_gr[H,Intercept] | -0.02729854 | 0.2698466 |
| r_gr[I,Intercept] | 0.10396778 | 0.2467358 |
| r_gr[J,Intercept] | 0.09940924 | 0.2471664 |
| r_gr__sigma[A,Intercept] | -0.04506816 | 0.3087082 |
| r_gr__sigma[B,Intercept] | -0.05915612 | 0.2699066 |

- Multilevel models allow modeling of data measured on different levels at the same time, thus taking complex dependency structures into account.
- it is desirable to allow for prediction of all response parameters at the same time. Models doing that are referred to as distributional models or models for location, scale and shape.
- non-linearity of predictor terms can be handled in at least two ways:

  (1) by fully specifying a non-linear predictor term with corresponding parameters each of which can be predicted using MLMs

  (2) estimating the form of the non-linear relationship on the fly using splines or Gaussian processes (*generalized additive models or* GAMs)

# *ANOVA-like 2-level model*

- $y_i \sim N(u, \sigma_y)$ likelihood
- $u = a_j$ j process models for j schools
- $\sigma_y \sim N()$ *prior*
- $a_j \sim N(\gamma, \sigma_a)$ *adaptive prior*
- $\gamma \sim N()$ *metaprior for meta-parameter gamma*
- $\sigma_a \sim N()$ *metaprior*

# now let us run some code…

## 6.2. Multilevel regression models: free intercepts

Now we add two predictors to our schools example: X - is students sex (male = 1 or female = 0) and S - is the school type (gymnasium or not). S is a group-level predictor, so there is only one S value per school, which is repeated in the table for all pupils in this school.

We start with converting the simple regression y = a + b1x + b2s + b3sch + error into a 2-level model. (In brms this model is $y \sim x + s + (1|sch)$.)

$$Within\ school_j : y_i \sim N(\alpha_j + \beta x_i, \sigma_y)$$

$$\alpha_j \sim N(\gamma_0 + \gamma_1 s_j, \sigma_\alpha)$$

- The multilevel model combines the J local models in two ways:
  - first, the local slopes are the same in all J models.
  - Second, the J intercepts are connected through the group-level model.
- The meaning of $\sigma_y$ and $\sigma_\alpha$ is the same as before, except the variance is estimated after conditioning for various predictors.
  - Group-level predictors reduce the group-level SD of $\sigma_\alpha$, leading to more shrinkage and to more precise estimates of the group means, especially for groups with small N.
  - Adding predictors at the individual and group levels typically reduces the unexplained variance at each level (adding a group-level predictor can also increase the unexplained variance).

## 6.4. Free intercepts and slopes

First the model without school-level variable S – $y \sim x + (x|sch)$ in brms syntax.

$$y_i \sim N(\alpha_{j[i]} + \beta_{j[i]} x_i, \sigma_y)$$
$$\begin{pmatrix} \alpha_j \\ \beta_j \end{pmatrix} \sim N(\begin{pmatrix} \mu_\alpha \\ \mu_\beta \end{pmatrix}, \begin{pmatrix} \sigma_\alpha & \rho\sigma_\alpha\sigma_\beta \\ \rho\sigma_\alpha\sigma_\beta & \sigma_\beta \end{pmatrix})$$

where $\rho$ models between-group correlation. This means that we model the school specific intercepts and slopes together so that there

We now expand this 2nd part of the model to include the school-level predictor S.

$$\begin{pmatrix} \alpha_j \\ \beta_j \end{pmatrix} \sim N(\begin{pmatrix} {\gamma_0}^\alpha + {\gamma_1}^\alpha u_j \\ {\gamma_0}^\beta + {\gamma_1}^\beta u_j \end{pmatrix}, \begin{pmatrix} \sigma_\alpha & \rho\sigma_\alpha\sigma_\beta \\ \rho\sigma_\alpha\sigma_\beta & \sigma_\beta \end{pmatrix})$$

Its brms version reads $y \sim x + s + x : s + (x|sch)$, so we have an interaction between the school type and sex of the pupil. Varying slopes can be considered as interactions between group indicators and an individual-level predictor.

The matrix of variances and covariances is arranged like this:

$$\begin{pmatrix} \text{variance of intercepts} & \text{covariance of intercepts \& slopes} \\ \text{covariance of intercepts \& slopes} & \text{variance of slopes} \end{pmatrix}$$

And now in mathematical form:

$$\begin{pmatrix} \sigma_\alpha^2 & \sigma_\alpha \sigma_\beta \rho \\ \sigma_\alpha \sigma_\beta \rho & \sigma_\beta^2 \end{pmatrix}$$

The variance in intercepts is $\sigma_\alpha^2$, and the variance in slopes is $\sigma_\beta^2$. These are found along the *diagonal* of the matrix. The other two elements of the matrix are the same, $\sigma_\alpha \sigma_\beta \rho$. This is the covariance between intercepts and slopes. It's just the product of the two standard deviations and the correlation. It might help to imagine an ordinary variance as the covariance of a variable with itself. If you are rusty on the definition of a covariance—it's okay, most people

# varying slopes & varying intercepts.

- enables pooling that improves estimates. varying slopes models are massive interaction machines. They allow every unit in the data to have its own response to any treatment, while also improving estimates via pooling.
- When the variation in slopes is large, the average slope is of less interest. Sometimes, the pattern of variation in slopes provides hints about omitted variables that explain why some units respond more or less.
- The machinery that makes such complex varying effects possible will extend the varying effects strategy to Gaussian processes.
- Ordinary varying effects work only with discrete, unordered groups, where each group is equally different from all of the others. But it is possible to use pooling with categories such as age or location, where some ages and some locations are more similar than others.

# How to pool information across intercepts & slopes?

- By modeling the covariance of intercepts and slopes by a joint multivariate Gaussian distribution for all of the varying effects, both intercepts and slopes.
- So instead of having two independent Gaussian distributions of intercepts and of slopes, assign a 2D Gaussian distribution to both the intercepts (first dimension) and the slopes (second dimension).
- The variance-covariance matrix, vcov, describes how each parameter's posterior probability is associated with each other parameter's posterior probability. Now we'll use the same kind of distribution to describe the variation within and covariation among different kinds of varying effects. Varying intercepts have variation, and varying slopes have variation. Intercepts and slopes covary.

# multilevel models are regressions with coefficients that are themselves modeled, a.k.a. regressions with coefficients that can vary by group.

- With grouped data, a regression that includes indicators for groups is called a varying-intercept model because it can be interpreted as a model with a different intercept within each group. a model with one continuous predictor x and indicators for J = 5 groups.

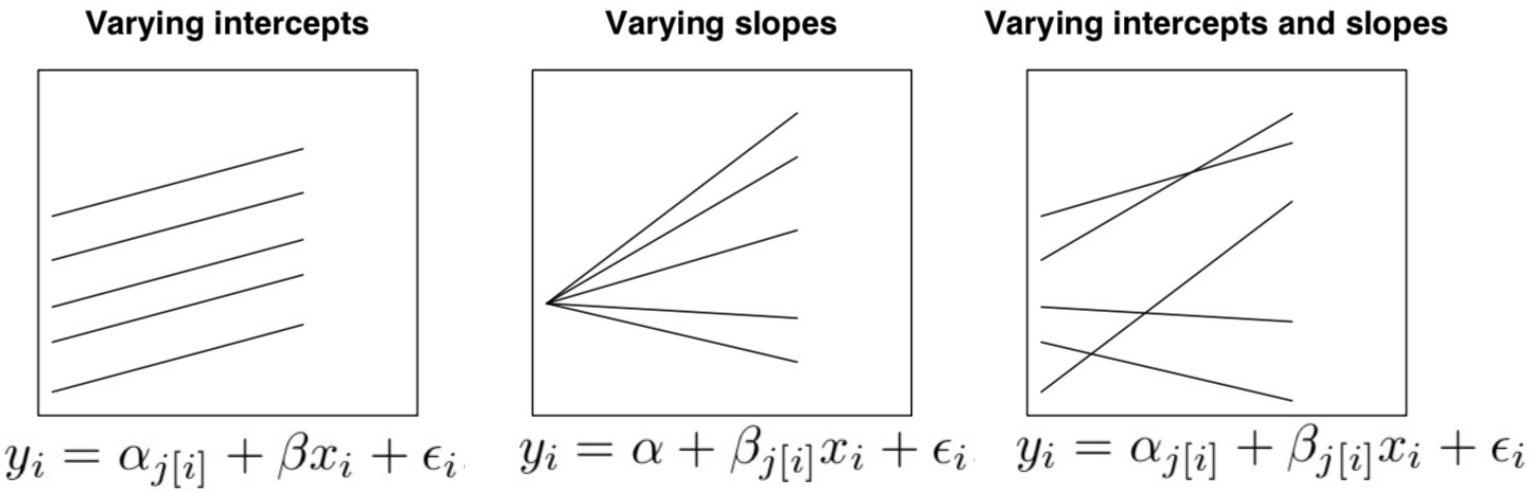



Figure 11.1 *Linear regression models with (a) varying intercepts ($y = \alpha_j + \beta x$), (b) varying slopes ($y = \alpha + \beta_j x$), and (c) both ($y = \alpha_j + \beta_j x$). The varying intercepts correspond to group indicators as regression predictors, and the varying slopes represent interactions between $x$ and the group indicators.*

# multilevel models give

- Improved estimates for repeat sampling. When more than one observation arises from the same individual, location, or time, then traditional, single-level models either maximally underfit or overfit the data.
- Improved estimates for imbalance in sampling. When some individuals, locations, or times are sampled more than others, multilevel models automatically cope with differing uncertainty across these clusters. This prevents over-sampled clusters from unfairly dominating inference.
- Estimates of variation. If our research questions include variation among individuals or other groups within the data, then multilevel models model variation explicitly.

  pre-averaging data to construct variables removes variation, and there are also different ways to perform the averaging. Averaging manufactures false confidence and introduces arbitrary data transformations.

multilevel regression deserves to be the default approach. There are contexts in which it would be better to use a single-level model. But the contexts in which multilevel models are superior are much more numerous. It is better to begin to build a multilevel analysis, and then realize it’s unnecessary, than to overlook it.

# Varying intercepts as over-dispersion

- the beta-binomial and negative binomial models were presented as ways for coping with over-dispersion of count data.
- Varying intercepts accomplish the same thing, allowing count outcomes to be over-dispersed. They accomplish this, because when each observed count gets its own unique intercept, but these intercepts are pooled through a common distribution, the predictions expect over-dispersion just like a beta-binomial or gamma-Poisson model would.
- Compared to a beta-binomial or negative binomial, a binomial or Poisson model with a varying intercept on every observed outcome will often be easier to estimate and easier to extend.

# Mixed-effects location scale models (MELSMs)

- Not only are parameters from the mean structure allowed to vary across groups, but parameters applied to σ are allowed to vary across groups, too.

$$
\begin{aligned}
\mathrm{NA}_{ij} &\sim \mathrm{Normal}(\mu_{ij}, \sigma_i) \\
\mu_{ij} &= \beta_0 + \beta_1 \mathrm{time}_{ij} + u_{0i} + u_{1i}\mathrm{time}_{ij} \\
\log(\sigma_i) &= \eta_0 + u_{2i} \\
\begin{bmatrix} u_{0i} \\ u_{1i} \\ u_{2i} \end{bmatrix} &\sim \mathrm{MVNormal}\left( \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \mathbf{SRS} \right) \\
\mathbf{S} &= \begin{bmatrix} \sigma_0 & 0 & 0 \\ 0 & \sigma_1 & 0 \\ 0 & 0 & \sigma_2 \end{bmatrix} \\
\mathbf{R} &= \begin{bmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{21} & 1 & \rho_{23} \\ \rho_{31} & \rho_{32} & 1 \end{bmatrix} \\
\beta_0 &\sim \mathrm{Normal}(0, 0.2) \\
\beta_1 \mathrm{and}\ \eta_0 &\sim \mathrm{Normal}(0, 1) \\
\sigma_0, \ldots, \sigma_2 &\sim \mathrm{Exponential}(1) \\
\mathbf{R} &\sim \mathrm{LKJ}(2).
\end{aligned}
$$

we now refer to σi, meaning the levels of variation not accounted for by the mean structure can vary across participants (hence the i subscript).
Two lines down, we see the formula for log($\sigma_i$) contains population-level intercept, $\eta_0$, and participant-specific deviations around that parameter, $u_{2i}$.
In the next three lines, we see that all 3 participant-level deviations, $u_{0i},\ldots,u_{2i}$ are multivariate normal with means set to zero and variation expressed in the parameters $\sigma_0,\ldots,\sigma_2$ of the S matrix.
In the R matrix, we now have three correlation parameters,
with $\rho_{31}$ and $\rho_{32}$ allowing us to assess the correlations among individual differences in variability and individual differences in starting points and change over time, respectively

```
b14.13 <-
  brm(data = dat,
      family = gaussian,
      bf(N_A.std ~ 1 + day01 + (1 + day01 |i| record_id),
         sigma ~ 1 + (1 |i| record_id)),
```

```
##  Family: gaussian
##   Links: mu = identity; sigma = log
## Formula: N_A.std ~ 1 + day01 + (1 + day01 | i | record_id)
##          sigma ~ 1 + (1 | i | record_id)
##    Data: dat (Number of observations: 13033)
## Samples: 4 chains, each with iter = 3000; warmup = 1000; thin = 1;
##          total post-warmup samples = 8000
##
## Group-Level Effects:
## ~record_id (Number of levels: 193)
##                               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk
## sd(Intercept)                     0.76      0.04     0.69     0.84 1.00
## sd(day01)                         0.60      0.04     0.53     0.69 1.00
## sd(sigma_Intercept)               0.69      0.04     0.63     0.77 1.00
## cor(Intercept,day01)             -0.34      0.08    -0.49    -0.17 1.01
## cor(Intercept,sigma_Intercept)    0.61      0.05     0.51     0.70 1.00
## cor(day01,sigma_Intercept)       -0.10      0.08    -0.26     0.05 1.00
##
## Population-Level Effects:
##                Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept          0.04      0.05    -0.07     0.14 1.01      193      274
## sigma_Intercept   -0.78      0.05    -0.88    -0.68 1.01      285      424
## day01             -0.16      0.05    -0.26    -0.06 1.00      596     1071
##
```

# randomization deletes all arrows that point to cause (fertilizer)

yield ~ fertilizer + soil fertility + … + Other

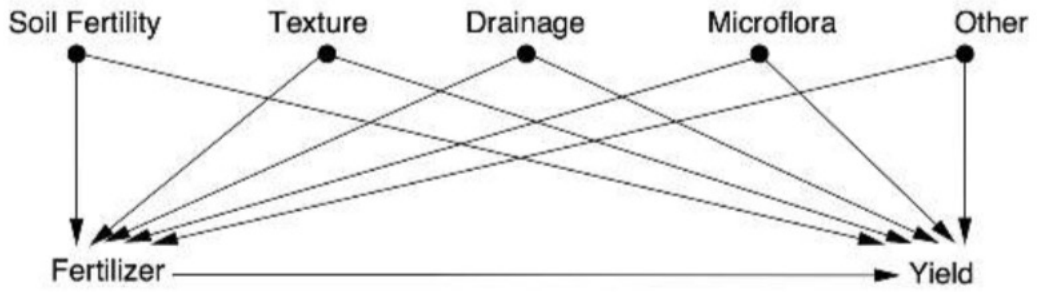



FIGURE 4.4. Model 1: an improperly controlled experiment.

yield ~ fertilizer

fertilizer ~ random card + error, where error = 0

so if random card = 1, then fertilizer = 1 (process model)

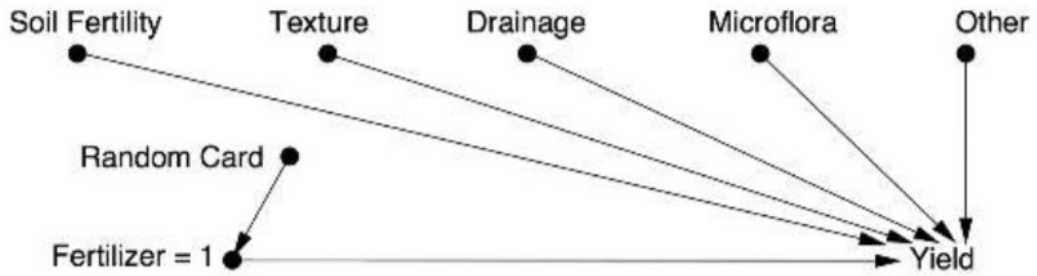



FIGURE 4.6. Model 3: the world simulated by a randomized controlled trial.

# John Snow and birth of epidemiology

- In 1853 and 1854, England was in the grips of a cholera epidemic.
- a person who drinks cholera-tainted water can die within 24 hours.
- in 1853, disease-causing germs had never yet been seen under a microscope for any illness. The prevailing wisdom held that a “miasma” of unhealthy air caused cholera, a theory supported by the fact that the epidemic hit harder in the poorer sections of London, where sanitation was worse.
- Dr. John Snow, who had taken care of cholera victims for more than twenty years, was skeptical of the miasma theory. He argued that since the symptoms manifested themselves in the intestinal tract, the body must first come into contact with the pathogen there.

- two main water companies: the Southwark and Vauxhall Company and the Lambeth Company. The former drew its water from the area of the London Bridge, which was downstream from London's sewers. The latter had moved its water intake several years earlier.
- Districts supplied by the Southwark and Vauxhall Company had a death rate eight times higher. A proponent of the miasma theory could argue that the miasma was strongest in those districts.

- Snow noticed that in those districts served by both companies, the death rate was still much higher in the households that received Southwark water. Yet these households did not differ in terms of miasma or poverty.

  "The mixing of the supply is of the most intimate kind. The pipes of each Company go down all the streets, and into nearly all the courts and alleys.... Each company supplies both rich and poor, both large houses and small; there is no difference either in the condition or occupation of the persons receiving the water of the different Companies."

- Even though RCT was still in the future, it was as if the water companies had conducted a randomized experiment on Londoners. Snow even notes this:

  "No experiment could have been devised which would more thoroughly test the effect of water supply on the progress of cholera than this, which circumstances placed ready made before the observer. The experiment, too, was on the grandest scale. No fewer than three hundred thousand people of both sexes, of every age and occupation, and of every rank and station, from gentlefolks down to the very poor, were divided into two groups without their choice, and in most cases, without their knowledge."

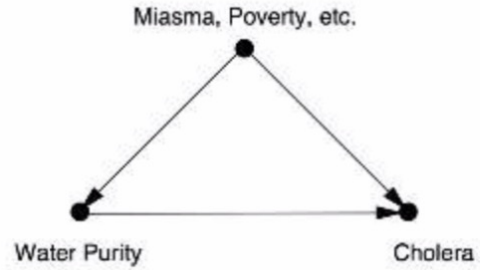
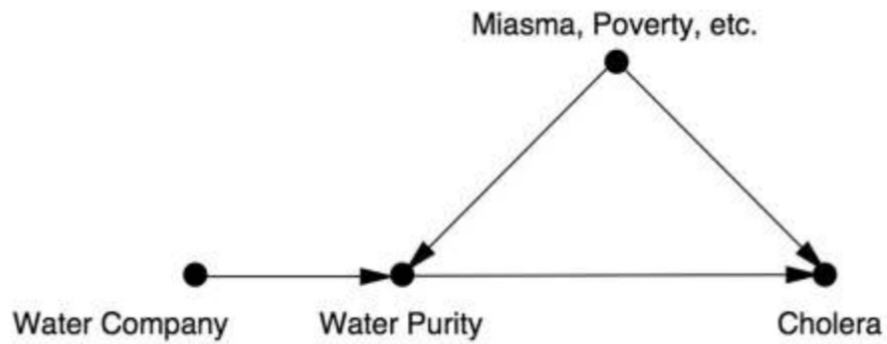



FIGURE 7.7. Causal diagram for cholera (before discovery of the cholera bacillus).

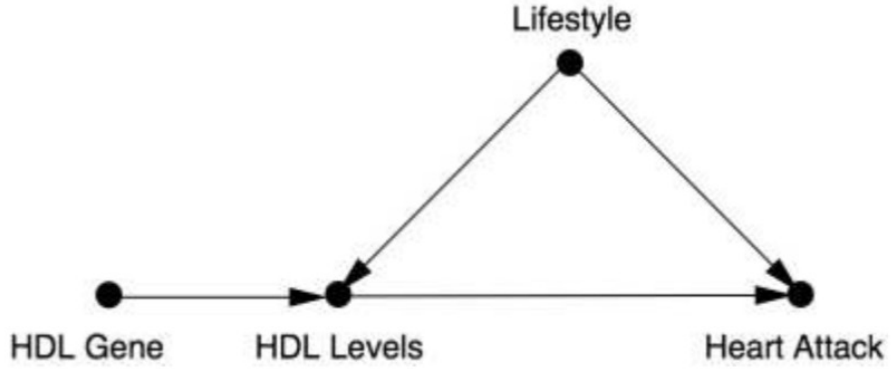
Snow's observations introduced a new variable into the causal diagram, which now looks like Figure 7.8. Snow's painstaking detective work had showed two important things: (1) there is no arrow between Miasma and Water Company (the two are independent), and (2) there is an arrow between Water Company and Water Purity. Left unstated by Snow, but equally important, is a third assumption: (3) the absence of a direct arrow from Water Company to Cholera, which is fairly obvious to us today because we know the water companies were not delivering cholera to their customers by some alternate route.



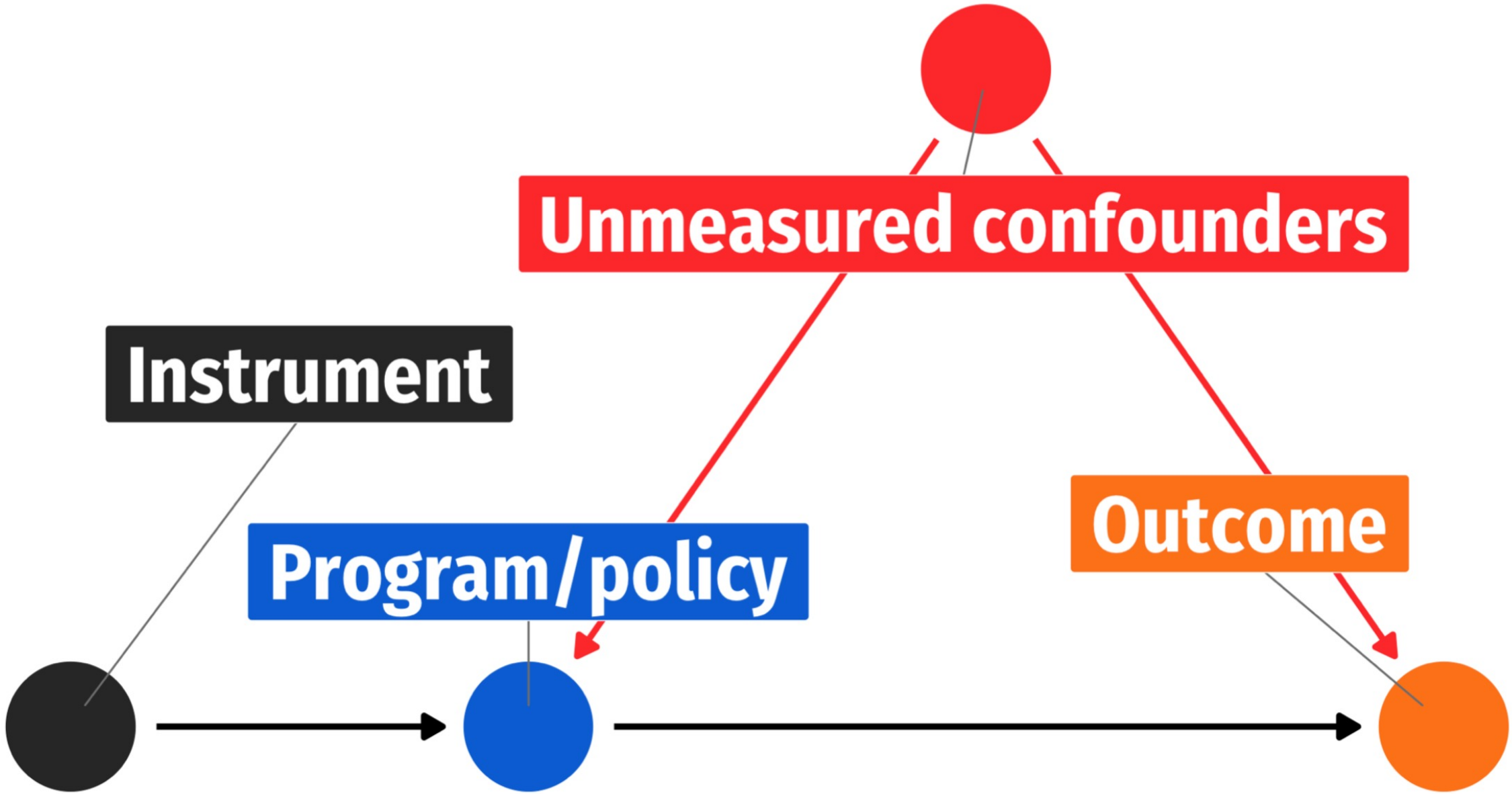
- Clearly Snow thought of this variable as similar to a coin flip, which simulates a variable with no incoming arrows. Because there are no confounders of the relation between Water Company and Cholera, any observed association must be causal. Likewise, since the effect of Water Company on Cholera must go through Water Purity, we conclude (as did Snow) that the observed association between Water Purity and Cholera must also be causal. Snow stated his conclusion in no uncertain terms: if the Southwark and Vauxhall Company had moved its intake point upstream, more than 1,000 lives would have been saved.
- He printed a pamphlet at his own expense, and it sold 56 copies.

# Mendelian randomization

- Although the effect of LDL, or “bad,” cholesterol is now settled, there is still considerable uncertainty about high-density lipoprotein (HDL), or “good,” cholesterol. Early observational studies suggested that HDL had a protective effect against heart attacks. But high HDL often goes hand in hand with low LDL, so how can we tell which lipid is the true causal factor?
- suppose we knew of a gene that caused people to have higher HDL levels, with no effect on LDL. Then we could set up the causal diagram. it is advantageous, as in Snow’s example, to use an instrumental variable that is randomized. If it’s randomized, no causal arrows point toward it. For this reason, a gene is a perfect instrumental variable: randomized at the time of conception



- In 2012, a giant collaborative study led by Sekar Kathiresan of Massachusetts General Hospital showed that there was no observable benefit from higher HDL levels. On the other hand, the researchers found that LDL has a very large effect on heart attack risk. According to their figures, decreasing your LDL count by 34 mg/dl would reduce your chances of a heart attack by about 50 percent. So lowering your “bad” cholesterol levels, whether by diet or exercise or statins, seems to be a smart idea. On the other hand, increasing your “good” cholesterol levels, despite what some fish-oil salesmen might tell you, does not seem likely to change your heart attack risk at all.
- As always, there is a caveat. The second study, published in the same year, pointed out that people with the lower-risk variant of the LDL gene have had lower cholesterol levels for their entire lives. Mendelian randomization tells us that decreasing your LDL by thirty-four units over your entire lifetime will decrease your heart attack risk by 50 percent. But statins can’t lower your LDL cholesterol over your entire lifetime


Unmeasured confounders
Instrument
Program/policy
Outcome

**Relevance**
**Correlated with policy**

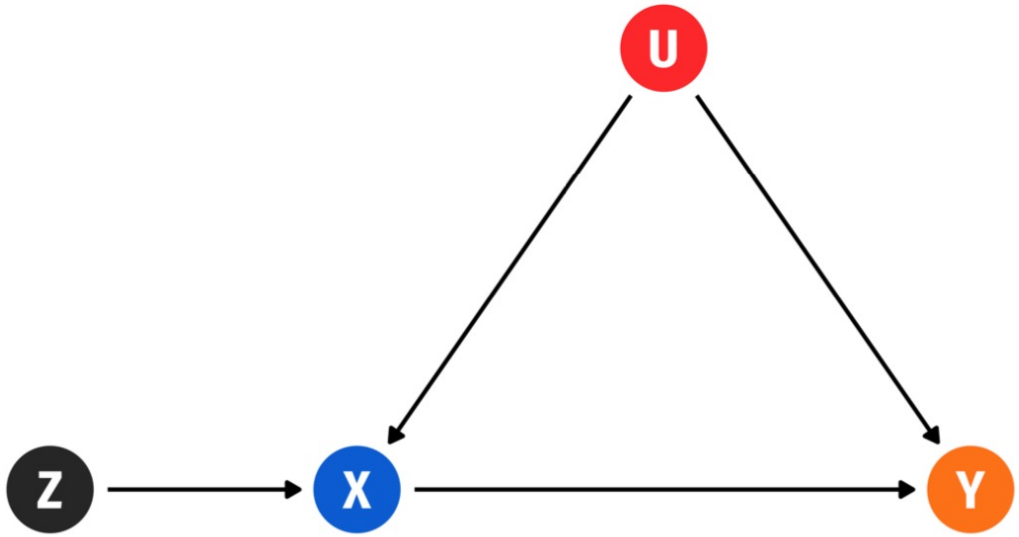
$Z \rightarrow X \quad Cor(Z, X) \neq 0$

**Excludability**
**Correlated with outcome *only through* policy**

$Z \rightarrow X \rightarrow Y \quad Z \not\rightarrow Y \quad Cor(Z, Y \mid X) = 0$

**Exogeneity**
***Not* correlated with omitted variables**

$U \not\rightarrow Z \quad Cor(Z, U) = 0$




**Relevance** testable with stats

**Excludability** testable with stats + story

**Exogeneity** requires story, no stats

# The huh? factor

"A necessary but not a sufficient condition for having an instrument that can satisfy the exclusion restriction is if people are confused when you tell them about the instrument's relationship to the outcome."

Scott Cunningham, *Causal Inference: The Mixtape*, p. 123

# Instruments are hard to find!

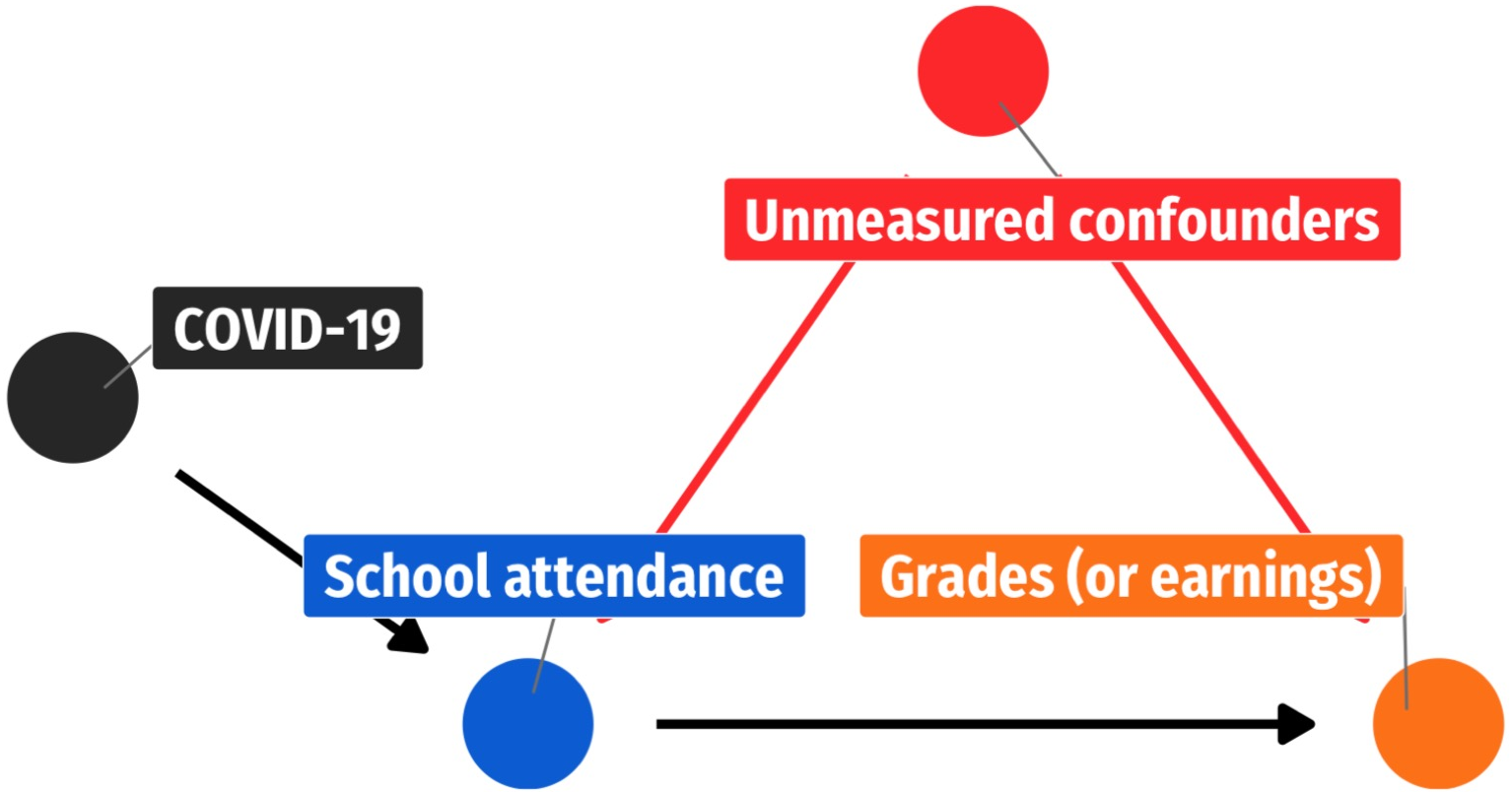
The trickiest thing to prove is the exclusion restriction

**Instrument causes the outcome *only through* the policy**

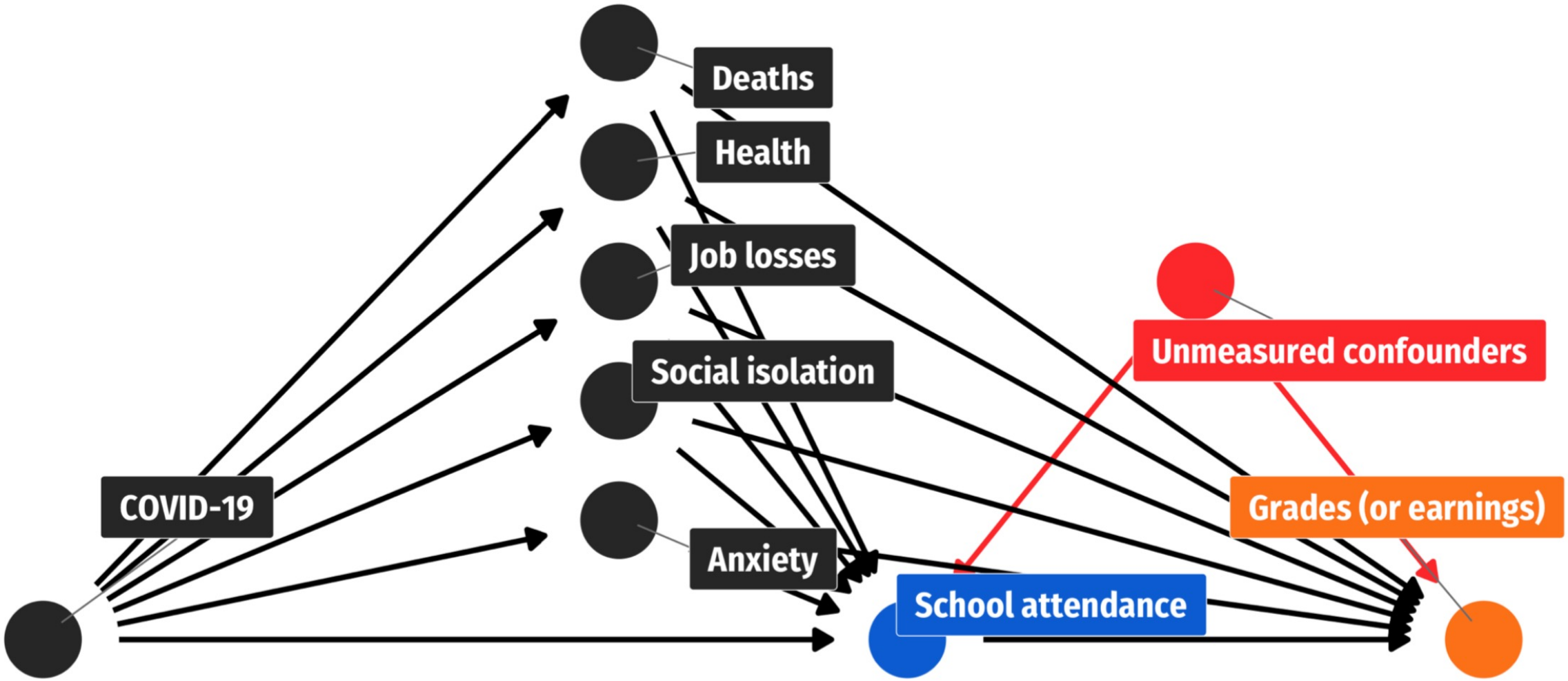
Most proposed instruments fail this!

# COVID-19 as an instrument

**What effect does closing schools have on student performance or lifetime earnings?**



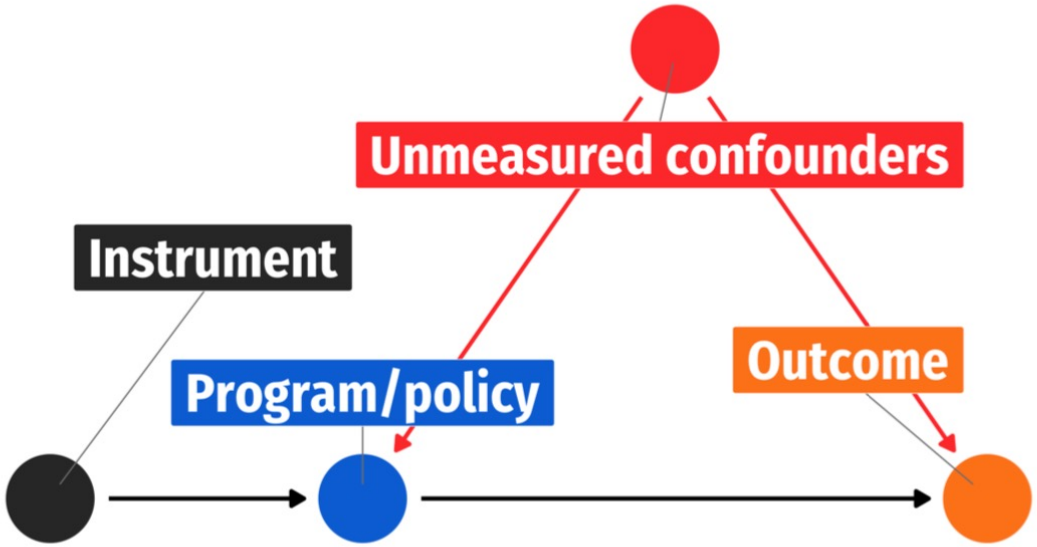
# lolnope


Deaths
Health
Job losses
Social isolation
Anxiety
COVID-19
Unmeasured confounders
Grades (or earnings)
School attendance

# Falsifying exclusion assumptions

**Can you think of some other way that the instrument can cause the outcome outside of the policy?**

**If so, the instrument doesn't meet exclusion restriction**



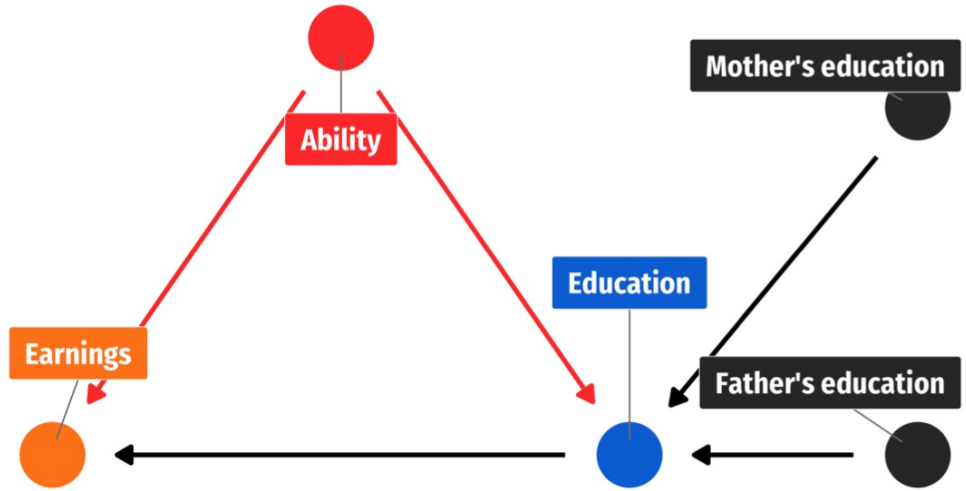

**Instrument → ?? → outcome?**

**Rainfall → ?? → civil war?**

**Tobacco taxes → ?? → health?**

**Scrabble score → ?? → Labor market success?**

You can use multiple instruments to explain more of the endogeneity in the policy node




# Multiple instruments

$$\widehat{\text{Education}}_i = \gamma_0 + \gamma_1 \text{Father's education}_i + \gamma_2 \text{Mother's education}_i + \upsilon_i$$

$$\text{Earnings}_i = \beta_0 + \beta_1 \widehat{\text{Education}}_i + \varepsilon_i$$

# Other control variables

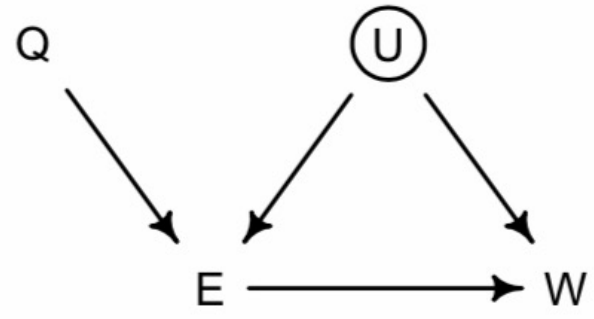
**You can use control variables too!**

**For mathy reasons,
all exogenous controls need to go in both stages**

$$\widehat{\text{Education}}_i = \gamma_0 + \gamma_1 \text{Father's education}_i + \gamma_2 \text{Mother's education}_i + \gamma_3 \text{SES}_i + \gamma_4 \text{State}_i + \gamma_5 \text{Year}_i + \upsilon_i$$

$$\text{Earnings}_i = \beta_0 + \beta_1 \widehat{\text{Education}}_i + \beta_2 \text{SES}_i + \beta_3 \text{State}_i + \beta_4 \text{Year}_i + \varepsilon_i$$

## 14.3. Instruments and causal designs




- What is the impact of education E on wages W?
- we cannot condition on U, since we haven't observed it.
- an instrumental variable Q (quarter of birth) acts like a natural experiment on E.
- an instrumental variable satisfies these criteria:

(1) Independent of U ($Q \perp U$) – cannot be tested

(2) Not independent of E ($Q \not\perp E$) – regression slope E ~ Q is not 0.

(3) Q cannot influence W except through E - cannot be tested & often implausible

If by conditioning on other variables, you satisfy 1-3, then you have an instrument.

W~E +Q opens the non-causal path Q -> E <- U -> W (non-causal, as changing Q results in no change in W through this path). the coef on Q picks up the association between U and W → the coef on E can get even more confounded.

- Q indicates which quarter of the year a person was born in. people born earlier in the year tend to get less schooling, as they are older when they start school.
- a simple generative version of the DAG has 4 sub-models.

1) how wages W are caused by education E and the unobserved confound U.

$$W_i \sim \text{Normal}(\mu_{\text{W},i}, \sigma_{\text{W}})$$
$$\mu_{\text{W},i} = \alpha_{\text{W}} + \beta_{\text{EW}} E_i + U_i$$

2) how education levels E are caused by quarter of birth Q—our instrument —and the same U.

$$E_i \sim \text{Normal}(\mu_{\text{E},i}, \sigma_{\text{E}})$$
$$\mu_{\text{E},i} = \alpha_{\text{E}} + \beta_{\text{QE}} Q_i + U_i$$

3) 1/4 of all people are born in each quarter of the year.

$$Q_i \sim \text{Categorical}([0.25, 0.25, 0.25, 0.25])$$

4) the unobserved confound U is normally distributed

$$U_i \sim \text{Normal}(0, 1)$$

- translate this generative model into a statistical model. Define W and E as coming from a common multivariate normal distribution.

$$\begin{pmatrix} W_i \\ E_i \end{pmatrix} \sim \text{MVNormal}\left( \begin{pmatrix} \mu_{\text{W},i} \\ \mu_{\text{E},i} \end{pmatrix}, \text{S} \right)$$
$$\mu_{\text{W},i} = \alpha_{\text{W}} + \beta_{\text{EW}} E_i$$
$$\mu_{\text{E},i} = \alpha_{\text{E}} + \beta_{\text{QE}} Q_i$$

  The matrix S is the error covariance between wages and education. It's not the descriptive covariance between these variables, but rather the matrix equivalent of the typical σ we stick in a Gaussian regression.

- The above is a **multivariate linear model** with multiple simultaneous outcomes, all modeled with a joint error structure. Each variable gets its own linear model, yielding the two μ definitions.
- education E is both an outcome and a predictor, as an implication of the DAG.

  The DAG says that E might influence W and that also pairs of W and E values might have some residual correlation, which arises through U.

The statistical solution to this mess is to express the data-generating DAG as a multivariate statistical model following the form

$$
\begin{aligned}
\begin{bmatrix} W_i \\ E_i \end{bmatrix} &\sim \text{MVNormal}\left( \begin{bmatrix} \mu_{\text{W},i} \\ \mu_{\text{E},i} \end{bmatrix}, \Sigma \right) \\
\mu_{\text{W},i} &= \alpha_{\text{W}} + \beta_{\text{EW}} E_i \\
\mu_{\text{E},i} &= \alpha_{\text{E}} + \beta_{\text{QE}} Q_i \\
\Sigma &= \begin{bmatrix} \sigma_{\text{W}} & 0 \\ 0 & \sigma_{\text{E}} \end{bmatrix} \text{R} \begin{bmatrix} \sigma_{\text{W}} & 0 \\ 0 & \sigma_{\text{E}} \end{bmatrix} \\
\text{R} &= \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \\
\alpha_{\text{W}} \text{ and } \alpha_{\text{E}} &\sim \text{Normal}(0, 0.2) \\
\beta_{\text{EW}} \text{ and } \beta_{\text{QE}} &\sim \text{Normal}(0, 0.5) \\
\sigma_{\text{W}} \text{ and } \sigma_{\text{E}} &\sim \text{Exponential}(1) \\
\rho &\sim \text{LKJ}(2).
\end{aligned}
$$

the estimated influence of education on wages is the correct causal inference.

the correlation Rho between the two outcomes, wages and education reflects the common influence of U. This correlation is conditional on E (for W) and Q (for E). It isn't the raw empirical correlation, but rather the residual correlation.

```r
e_model <- bf(e ~ 1 + q)
w_model <- bf(w ~ 1 + e)

b14.6 <-
  brm(data = dat_sim,
      family = gaussian,
      e_model + w_model + set_rescor(TRUE),
      prior = c(# E model
                prior(normal(0, 0.2), class = Intercept, resp = e),
                prior(normal(0, 0.5), class = b, resp = e),
                prior(exponential(1), class = sigma, resp = e),

                # W model
                prior(normal(0, 0.2), class = Intercept, resp = w),
                prior(normal(0, 0.5), class = b, resp = w),
                prior(exponential(1), class = sigma, resp = w),

                # rho
                prior(lkj(2), class = rescor)),
```

we will make sure to set set_rescor(TRUE). priors for residual correlations are of class = rescor.

```
Population-Level Effects:
            Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
e_Intercept    -0.00      0.04    -0.07     0.07 1.00     2866     2597
w_Intercept    -0.00      0.04    -0.09     0.09 1.00     2895     2703
e_q             0.59      0.04     0.52     0.66 1.00     2809     2527
w_e            -0.05      0.08    -0.21     0.09 1.00     1969     2492

Family Specific Parameters:
        Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma_e     0.81      0.03     0.76     0.86 1.00     3419     2895
sigma_w     1.02      0.05     0.94     1.12 1.00     2055     2223

Residual Correlations:
            Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
rescor(e,w)     0.54      0.05     0.44     0.64 1.00     1983     2373
```

Now the parameter for E→W, w_e, is just where it should be– near zero. The residual correlation between E and Q, rescore(e,w), is positive and large in magnitude, indicating their common influence from the unmeasured variable U.

One can use the `dagitty()` and `instrumentalVariables()` functions from the **dagitty** package to first define a DAG and then query whether there are instrumental variables for a given exposure and outcome.

```
library(dagitty)

dagIV <- dagitty("dag{Q -> E <- U -> W <- E}")

instrumentalVariables(dagIV, exposure = "E", outcome = "W")
```

```
##  Q
```

> The hardest thing about instrumental variables is believing in any particular instrument. If you believe in your DAG, they are easy to believe. But should you believe in your DAG?…

## 14.5. Continuous categories and the Gaussian process

- It doesn't make sense to estimate a unique varying intercept for all individuals of the same age, ignoring the fact that individuals of similar ages should have more similar intercepts.
- complexity of tool kits among historic Oceanic societies. If all of your neighbors are small islands, then high rate of contact with them may not do much at all to tool complexity. Second, if tools were exchanged among societies then the total number of tools for each are truly not independent of one another, even after we condition on all of the predictors. we expect close geographic neighbors to have more similar tool counts, because of exchange. Third, closer islands may share unmeasured geographic features like sources of stone or shell that lead to similar technological industries. So space could matter in multiple ways.
- We'll define a distance matrix among the societies. Then we can estimate how similarity in tool counts depends upon geographic distance.

- the first part of the model is a Poisson probably of the outcome variable. Then there is a model-derived expected number of tools:

$$T_i \sim \text{Poisson}(\lambda_i)$$

$$\lambda_i = \alpha P_i^{\beta} / \gamma$$

- We'd like to have λ values adjusted by a varying intercept ($k_{society[i]}$). We could just add the intercept, but then $\lambda_i$ might end up negative. So let's make the $k_{society[i]}$ multiplicative:

The heart of the Gaussian process is the multivariate prior for these intercepts:

$$\begin{pmatrix} k_1 \\ k_2 \\ k_3 \\ \dots \\ k_{10} \end{pmatrix} \sim \text{MVNormal}\left( \begin{pmatrix} 0 \\ 0 \\ 0 \\ \dots \\ 0 \end{pmatrix}, \mathbf{K} \right)$$

[prior for intercepts]

$$\mathbf{K}_{ij} = \eta^2 \exp(-\rho^2 D_{ij}^2) + \delta_{ij}\sigma^2$$

[define covariance matrix]

$$T_i \sim \text{Poisson}(\lambda_i)$$

$$\lambda_i = \exp(k_{\text{SOCIETY}[i]})\alpha P_i^{\beta} / \gamma$$

The first line is the 10D Gaussian prior for the intercepts (there are 10 societies in the distance matrix). The vector of means is all zeros, which means that inside the linear model the average society will multiply λ by exp(0) = 1. Negative k values will reduce λ, and positive k values will increase it.

| | Ml | Ti | SC | Ya | Fi | Tr | Ch | Mn | To | Ha |
|---|---|---|---|---|---|---|---|---|---|---|
| Malekula | 0.0 | 0.5 | 0.6 | 4.4 | 1.2 | 2.0 | 3.2 | 2.8 | 1.9 | 5.7 |
| Tikopia | 0.5 | 0.0 | 0.3 | 4.2 | 1.2 | 2.0 | 2.9 | 2.7 | 2.0 | 5.3 |
| Santa Cruz | 0.6 | 0.3 | 0.0 | 3.9 | 1.6 | 1.7 | 2.6 | 2.4 | 2.3 | 5.4 |
| Yap | 4.4 | 4.2 | 3.9 | 0.0 | 5.4 | 2.5 | 1.6 | 1.6 | 6.1 | 7.2 |
| Lau Fiji | 1.2 | 1.2 | 1.6 | 5.4 | 0.0 | 3.2 | 4.0 | 3.9 | 0.8 | 4.9 |
| Trobriand | 2.0 | 2.0 | 1.7 | 2.5 | 3.2 | 0.0 | 1.8 | 0.8 | 3.9 | 6.7 |
| Chuuk | 3.2 | 2.9 | 2.6 | 1.6 | 4.0 | 1.8 | 0.0 | 1.2 | 4.8 | 5.8 |
| Manus | 2.8 | 2.7 | 2.4 | 1.6 | 3.9 | 0.8 | 1.2 | 0.0 | 4.6 | 6.7 |
| Tonga | 1.9 | 2.0 | 2.3 | 6.1 | 0.8 | 3.9 | 4.8 | 4.6 | 0.0 | 5.0 |
| Hawaii | 5.7 | 5.3 | 5.4 | 7.2 | 4.9 | 6.7 | 5.8 | 6.7 | 5.0 | 0.0 |

- The covariance matrix for these intercepts is K, and the covariance between any pair of societies i and j is $K_{ij}$.
- This covariance is defined by the formula on the second line. This formula uses three parameters—η, ρ, and σ—to model how covariance among societies changes with distances among them.
- The part of the formula that gives the covariance model its shape is $\exp(-\rho^2 D^2_{ij})$. $D_{ij}$ is the distance between the i-th and j-th societies. So the covariance between any two societies i and j declines exponentially with the squared distance.
- ρ determines the rate of decline. If it is large, then covariance declines rapidly.
- the squared distance is the most common assumption, both because it is easy to fit to data and has the property of allowing covariance to decline more quickly as distance grows.
- $\eta^2$ is the maximum covariance between any two societies i and j.
- $\delta_{ij}\sigma^2$, provides for extra covariance beyond $\eta^2$ when i = j. It does this because the function $\delta^{ij} = 1$ when i = j but is 0 otherwise. In these data, this term will not matter, because we only have one observation for each society. But if we had more than one observation per society, σ here describes how these observations covary.

# Missing Data and Other Opportunities

- With measurement error, the insight is to realize that any uncertain piece of data can be replaced by a distribution that reflects uncertainty. But sometimes data are simply missing— no measurement is available at all. At first, this seems like a lost cause. What can be done when there is no measurement at all, not even one with error?
- The most common treatment of missing values is just to drop all cases with any missing values (complete case analysis). It is the default and silent behavior of most statistical software.
- Another common response is to replace missing values with some assumed value, like the mean of the variable or a reference value like zero.
- Complete case analysis is at best inefficient. But it can also produce bias, depending upon the causal details. Replacing missing values with static values is never warranted—we do not know those values, and if you fix them, the model will think it knows them with certainty.
- We can think causally about missingness, and we can use the model to impute missing values. A generative model tells you whether the process that produced the missing values will also prevent the identification of causal effects. Luckily, we can add missingness to a DAG to figure out whether it produces confounding.

- Missing data are meaningful data. The fact that a variable has an unobserved value is still an observation. It is data, just with a very special value. The meaning of this value depends upon the context. Consider for example a questionnaire on personal income. If some people refuse to fill in their income, this may be associated with low (or high) income. Therefore a model that tries to predict the missing values can be enlightening. In ecology, the absence of an observation of a species is a subtle kind of observation. It could mean the species isn't there. Or it could mean it is there but you didn't see it. An entire category of models, occupancy models, exists to take this duality into account.

(a) A dog's decision to eat homework is not influenced by any relevant variable: there is no arrow entering D in the DAG. $(H \perp\!\!\!\perp D)$, because $H^*$ is a collider.

Since the missing values are completely random, missingness doesn't necessarily change the overall distribution of homework scores. missing homework doesn't necessarily bias our estimate of the causal effect of studying.

missing completely at random (MCAR),

(b) and (c) **MISSING AT RANDOM** (MAR)

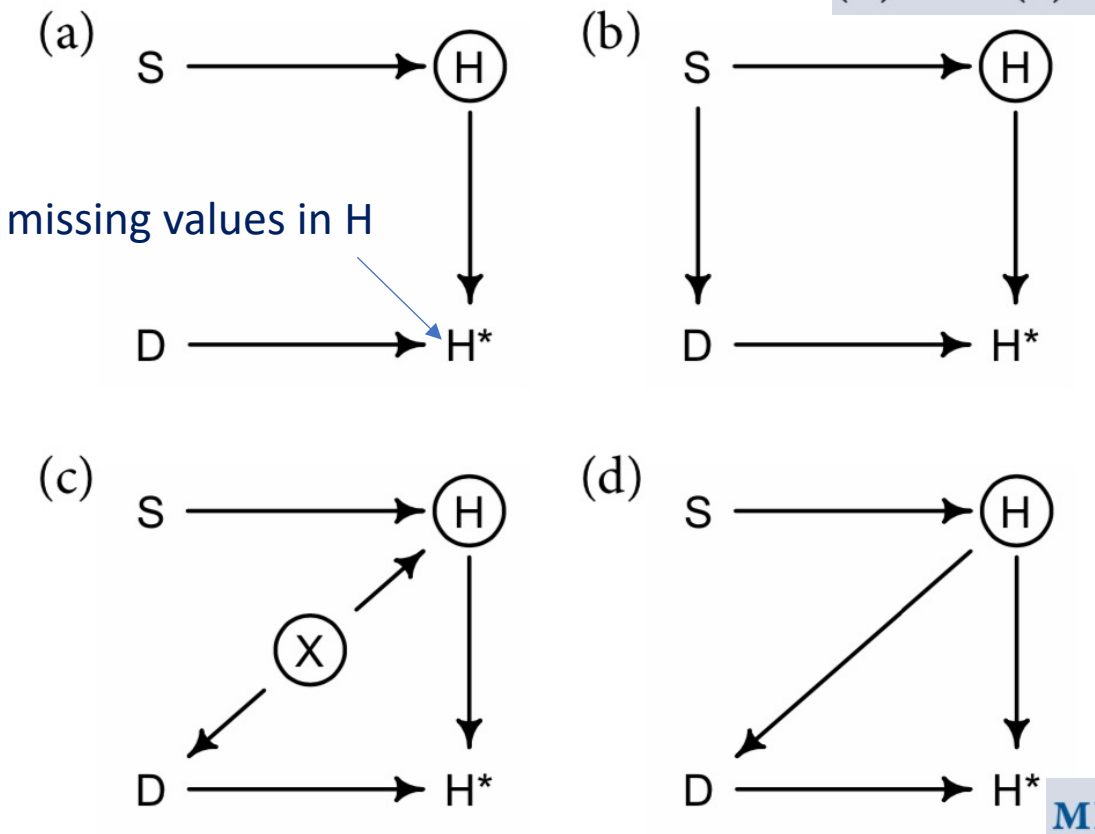
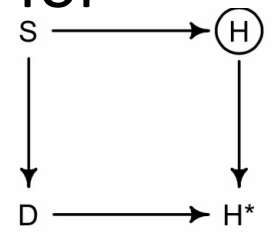



FIGURE 15.4. Four causal scenarios for the missing homework. See text for a complete explanation. (a) Dogs ($D$) eat homework ($H$) completely at random. (b) Dogs eat homework of students who study ($S$) too much. (c) Dogs eat more homework in noisy ($X$) homes, where the homework is also worse. (d) Dogs prefer to eat bad homework.

**MISSING NOT AT RANDOM** (MNAR)

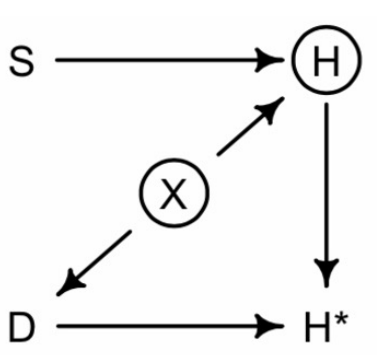
- (b) studying influences whether a dog eats homework, S → D. Suppose for example that students who study a lot do not play with their dogs.




- there is now a non-causal path (a backdoor path) from H → H∗ ← D ← S. If we don't close this path, it will confound inference along S → H. we can close it by conditioning on S, and we want to condition on S anyway.
- This doesn't mean there is no danger here. If we get the functions or distributions wrong, then we may get the wrong answer and the missing data may prevent us from seeing the absurdity of it in posterior predictive checks. Suppose for example that studying doesn't help at all until a student does more than the average amount. In that case, we never get to see homework from those students, so we can't possibly figure out the function that relates study effort to homework score.

(c), is more difficult. There is a variable that influences both H and D.

- it is a new variable X, the noise level of the student's house.
  - In a noisy house, students produce worse homework, X → H.
  - dogs in noisy houses tend to misbehave, X → D.
  - circle around X signals that it is unobserved.



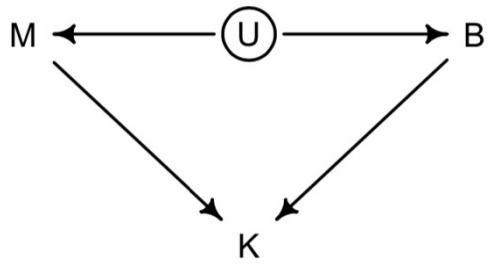

- when we regress $H^*$ on S, a new non-causal path: $H* \leftarrow D \leftarrow X \rightarrow H$.
- what effect this path has on our estimate of S → H?

(d), there is no X, but there is a path from H → D. Now dogs prefer to eat bad homework.

- there is nothing we can condition on to block the non-causal path S → H → D → $H*$. This type of missingness, in which the variable causes its own missing values, is the worst. Unless you know the mechanism that produces the missingness (D in this case), there is little hope. But even if you do know the mechanism, sometimes the only solution is to take better measurements.

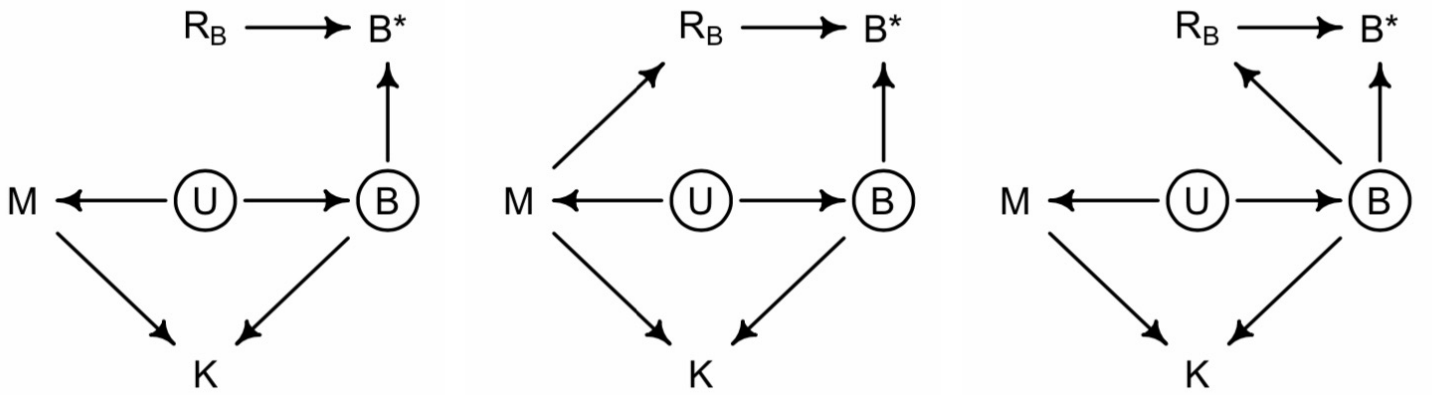
# We impute both to avoid biased estimation and so that we can use all of the observed data.

- any generative model necessarily contains information about variables that have not been observed.
- Some data go missing, but the model stays the same. In theory then imputing missing data is easy. In practice there can be challenges, as always.
- M is body mass,
- B is neocortex percent,
- K is milk energy, and




- U is some unobserved variable that renders M and B positively correlated.
- We want to add missingness to this graph.

- We haven't observed B (neocortex %). We've instead observed B∗, a partially observed set of values generated by B and some process.
- the observed values B∗ are a function of B and the "missingness" process.
- Whatever the process, it generates a variable RB that indicates which species have missing values. The crucial question is which variables influence RB.
- On the left, nothing influences $R_B$. there is no new non-causal path introduced. Dropping the species with missing brain values doesn't necessarily bias inference.
- In the middle, body mass M influences which species have missing values, creates a non-causal path B∗ ← $R_B$ ← M → K. conditioning on M blocks this path
- on the right, brain size B itself influencing $R_B$. This could happen because anthropologists are more interested in large-brained species. estimation of B → K will be biased by a non-causal path through $R_B$. It will not be possible to test, with these data, whether B influences $R_B$.
- In every DAG we want to impute. In the first and second, in order to not throw away corresponding values of M. In the third, to hope for any sensible estimate of B → K.



# The trick with Bayesian imputation is to model the variable that has missing values.

- Each missing value is assigned a unique parameter. The observed values give us information about the distribution of the values. This distribution becomes a prior for the missing values. This prior will then be updated by full model. So there will be a posterior distribution for each missing value.
- For every index i at which there is a missing value, there is also a parameter $B_i$ that will form a posterior distribution for it. The simplest model will impute B from its own normal distribution, ignoring that B and M are associated through U.

$$K_i \sim \text{Normal}(\mu_i, \sigma) \quad \text{[distribution for outcome } k\text{]}$$

$$\mu_i = \alpha + \beta_B B_i + \beta_M \log M_i \quad \text{[linear model]}$$

$$B_i \sim \text{Normal}(\nu, \sigma_B) \quad \text{[distribution for obs/missing } B\text{]}$$

when $B_i$ is observed, $B_i \sim \text{Normal}(\nu, \sigma_B)$ is a likelihood. The model learns the distributions of $\nu$ and $\sigma_B$ that are consistent with the data. But when $B_i$ is missing and therefore a parameter, same line is interpreted as a prior. Since the parameters $\nu$ and $\sigma_B$ are also estimated, the prior is learned from the data, just like the varying effects.

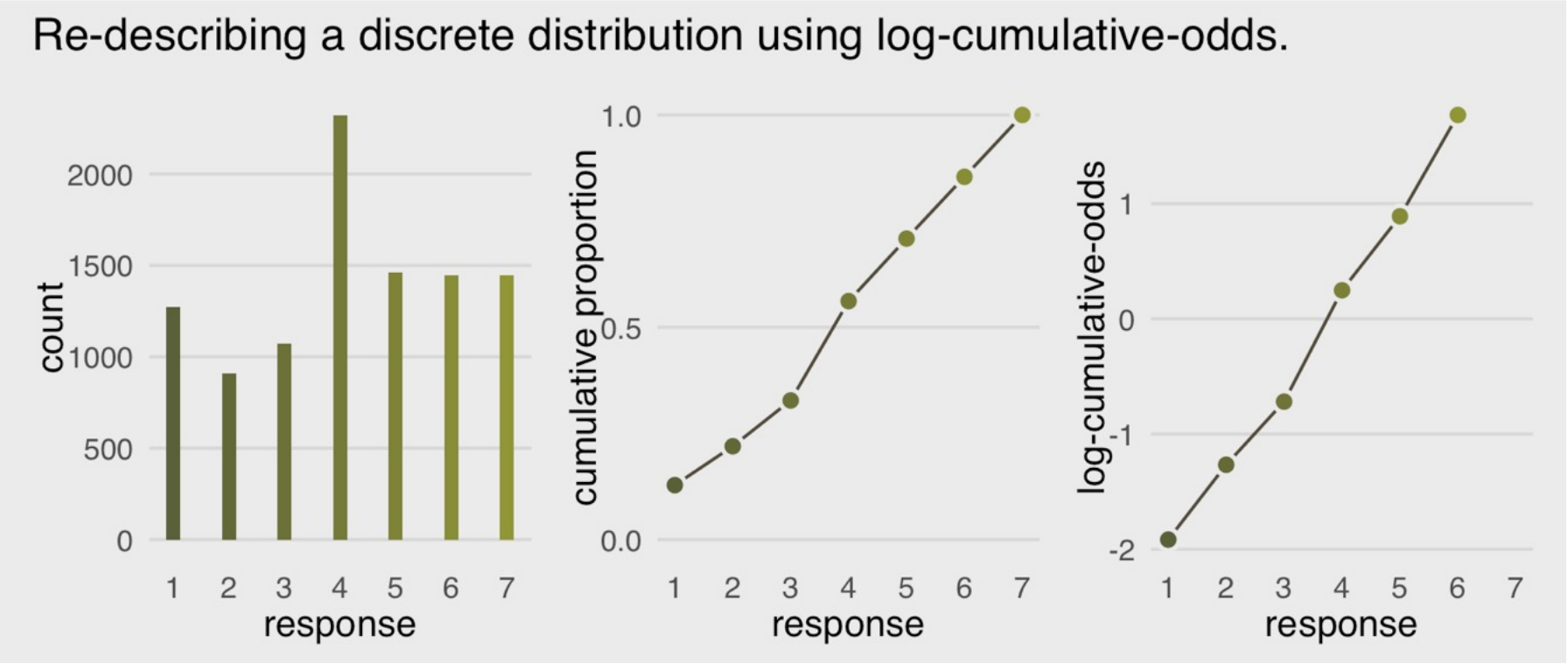
- this model assumes each B value has a standardized Gaussian uncertainty. But we know that these values are bounded between zero and one, because they are proportions. But assigning a Gaussian distribution doesn't really mean that the frequency distribution of the variable is a bell curve. It just means we will use only the mean and variance to describe it. The Gaussian is a very conservative choice, because it is the flattest unbounded distribution with a given variance.
- We can improve by changing the imputation model to estimate the relationship between the two predictors. This really just means that we use the entire generative model. In the DAG, B and M are associated as a result of U. If we can include that fact in the model, we might make better imputations and therefore better inferences. The is only to change the imputation line of the model from

  $B_i \sim Normal(v, \sigma_B)$ to

  $(M_i, B_i) \sim MVNormal((\mu_M, \mu_B), S)$
- The S is a covariance matrix, it measures the correlation between M and B, using the observed cases, and then uses that correlation to infer the missing B values.
- this is the simplest model of the association between M and B. It assumes that the covariance is sufficient to describe their relationship.

# ordered categories

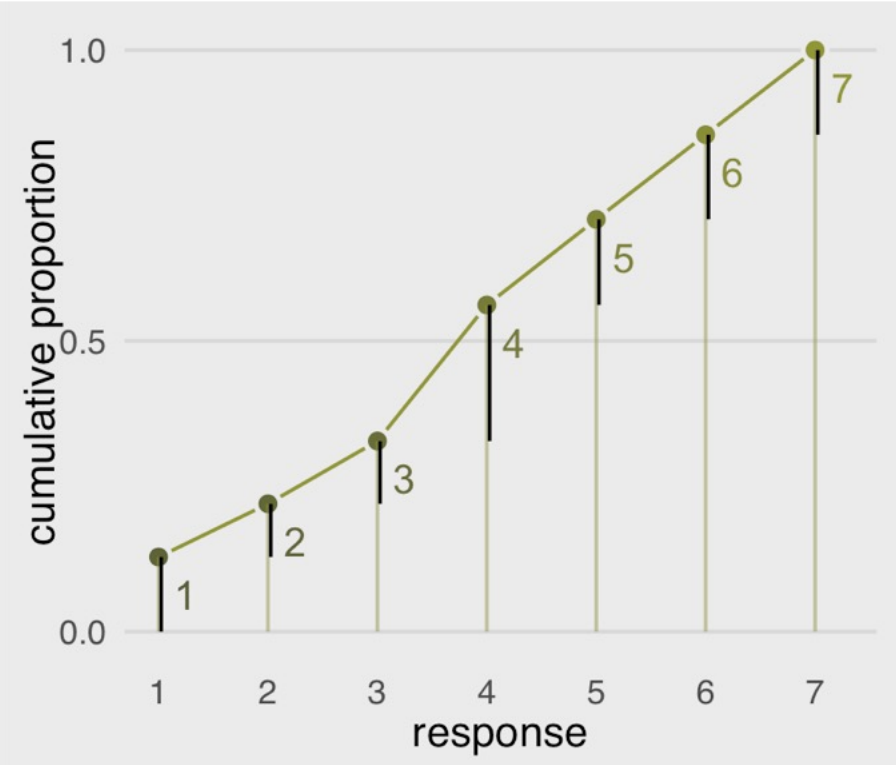
- an outcome variable is discrete, like a count, but the values merely indicate different ordered levels along some dimension.
  - how much you like to eat fish, on a scale from 1 to 7.
- The result is a set of **ordered categories**. Unlike a count, the differences in value are not necessarily equal
- an ordered categorical variable is a multinomial prediction problem. But the constraint that the categories be ordered demands special treatment.
- The conventional solution is to use a **cumulative link** function. The cumulative probability of a value is the probability of that value *or any smaller value*.
  - In the context of ordered categories, the cumulative probability of 3 is the sum of the probabilities of 3, 2, and 1. Ordered categories by convention begin at 1, so a result less than 1 has no probability at all.
- By linking a linear model to cumulative probability, it is possible to guarantee the ordering of the outcomes.




Then to re-describe the histogram as log-cumulative odds, we'll need a series of intercept parameters. Each intercept will be on the log-cumulative-odds scale and stand in for the cumulative probability of each outcome. So this is just the application of the link function. The log-cumulative-odds that a response value $y_i y_i$ is equal-to-or-less-than some possible outcome value $kk$ is:

$$\log \frac{\Pr(y_i \leq k)}{1 - \Pr(y_i \leq k)} = \alpha_k$$

where $\alpha_k$ is an "intercept" unique to each possible outcome value $k$




A compact way to express the formula for this first type of statistical model is

$$\begin{aligned} \text{response}_i &\sim \text{Categorical}(\mathbf{p}) \\ \text{logit}(p_k) &= \alpha_k - \phi \\ \phi &= 0 \\ \alpha_k &\sim \text{Normal}(0, 1.5), \end{aligned}$$

$\alpha_k$ denotes the $K-1$ intercepts (cut points or thresholds) we use to describe each possible outcome value $k$. $\phi$ is a stand-in for the potential terms of the linear model. If we have no predictors, $\phi = 0$.

- An ordered-logit distribution is really just a categorical distribution that takes a vector **p**={$p1$,$p2$,$p3$,$p4$,$p5$,$p6$} of probabilities of each response value below the maximum response (7 in this example). Each response value $k$ in this vector is defined by its link to an intercept parameter, $\alpha_k$.

```
# define the start values
inits <- list(`Intercept[1]` = -2,
              `Intercept[2]` = -1,
              `Intercept[3]` = 0,
              `Intercept[4]` = 1,
              `Intercept[5]` = 2,
              `Intercept[6]` = 2.5)

inits_list <- list(inits, inits, inits, inits)

b12.4 <-
  brm(data = d,
      family = cumulative,
      response ~ 1,
      prior(normal(0, 1.5), class = Intercept),
      iter = 2000, warmup = 1000, cores = 4, chains = 4,
      inits = inits_list,  # here we add our start values
```

```
##  Family: cumulative
##   Links: mu = logit; disc = identity
## Formula: response ~ 1
##    Data: d (Number of observations: 9930)
## Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup samples = 4000
##
## Population-Level Effects:
##             Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept[1]    -1.92      0.03    -1.98    -1.86 1.00     2667     2679
## Intercept[2]    -1.27      0.02    -1.31    -1.22 1.00     3787     3333
## Intercept[3]    -0.72      0.02    -0.76    -0.68 1.00     4319     3563
## Intercept[4]     0.25      0.02     0.21     0.29 1.00     4859     3437
## Intercept[5]     0.89      0.02     0.85     0.93 1.00     4731     3274
## Intercept[6]     1.77      0.03     1.72     1.83 1.00     5131     3576
##
## Family Specific Parameters:
##      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## disc     1.00      0.00     1.00     1.00 1.00     4000     4000
```

Intercept[k]-s are the $\alpha_k$ parameters (the thresholds).

We can use the `brms::inv_logit_scaled()` function to get these into the probability metric.

```
b12.4 %>%
  fixef() %>%
  inv_logit_scaled() %>%
  round(digits = 3)
```

```
##              Estimate Est.Error  Q2.5 Q97.5
## Intercept[1]    0.128     0.508 0.122 0.135
## Intercept[2]    0.220     0.506 0.212 0.229
## Intercept[3]    0.328     0.505 0.318 0.337
## Intercept[4]    0.562     0.505 0.552 0.571
## Intercept[5]    0.709     0.505 0.700 0.718
## Intercept[6]    0.854     0.507 0.848 0.861
```

But the posterior $SD$ (i.e., the 'Est.Error' values) are not valid using that approach.
If you really care about them, you'll need to work with the posterior_samples().

```r
posterior_samples(b12.4) %>%
  mutate_all(inv_logit_scaled) %>%
  pivot_longer(starts_with("b_")) %>%
  group_by(name) %>%
  summarise(mean = mean(value),
            sd   = sd(value),
            ll   = quantile(value, probs = .025),
            ul   = quantile(value, probs = .975))
```

# or use posterior_summary() instead

```
## # A tibble: 6 x 5
##   name            mean      sd    ll    ul
##   <chr>          <dbl>   <dbl> <dbl> <dbl>
## 1 b_Intercept[1] 0.128 0.00339 0.122 0.135
## 2 b_Intercept[2] 0.220 0.00426 0.212 0.229
## 3 b_Intercept[3] 0.328 0.00465 0.318 0.337
## 4 b_Intercept[4] 0.562 0.00481 0.552 0.571
## 5 b_Intercept[5] 0.709 0.00444 0.700 0.718
## 6 b_Intercept[6] 0.854 0.00350 0.848 0.861
```

# Adding predictor variables

- we define the generic linear model as $\phi_i = \beta x_i$.
- Accordingly, the formula for our cumulative logit model becomes

$$\log \frac{\Pr(y_i \leq k)}{1 - \Pr(y_i \leq k)} = \alpha_k - \phi_i$$
$$\phi_i = \beta x_i.$$

- This form ensures the correct ordering of the outcome values, while still morphing the likelihood of each individual value as the predictor $x_i$ changes value.
  - the linear model $\phi$ *is* subtracted from each intercept because if we decrease the log-cumulative-odds of every outcome value *k* below the maximum, this necessarily shifts probability mass upwards towards higher outcome values.
  - $\beta > 0$ means that increasing *x* also increases the mean *y*.

$$
\begin{aligned}
\text{response}_i &\sim \text{Categorical}(\mathbf{p}) \\
\text{logit}(p_k) &= \alpha_k - \phi_i \\
\phi_i &= \beta_1 \text{action}_i + \beta_2 \text{contact}_i + (\beta_3 + \beta_4 \text{action}_i + \beta_5 \text{contact}_i)\text{intention}_i \\
\alpha_k &\sim \text{Normal}(0, 1.5) \\
\beta_1, \ldots, \beta_5 &\sim \text{Normal}(0, 0.5),
\end{aligned}
$$

$$
\phi_i = \beta_1 \text{action}_i + \beta_2 \text{contact}_i + \beta_3 \text{intention}_i + \beta_4 (\text{action}_i \times \text{intention}_i) + \beta_5 (\text{contact}_i \times \text{intention}_i)
$$

```
brm(data = d,
    family = cumulative,
    response ~ 1 + action + contact + intention + intention:action + intention:contact,
    prior = c(prior(normal(0, 1.5), class = Intercept),
              prior(normal(0, 0.5), class = b)),
    iter = 2000, warmup = 1000, cores = 4, chains = 4,
    seed = 12,
    file = "fits/b12.05")
```

# Ordered categorical predictors

- We can handle ordered outcome variables using a categorical model with a cumulative link. That was the previous section.
- What about ordered predictor variables? We could just include them as continuous predictors like in any linear model. But this isn't ideal. Just like with ordered outcomes, we don't really want to assume that the distance between each ordinal value is the same.
- brms syntax for fitting models with monotonic predictors is simple. Just place your monotonic predictors within the mo() function and enter them into the formula.

```
brm(data = d,
    family = cumulative,
    response ~ 1 + action + contact + intention + mo(edu_new),
```