causal salad

Ülo Maiväli

linear regression as a Bayesian procedure.

Normal by addition – drunkards walk (random steps either way).

we generate for each drunkard a list of 16 random numbers between −1 and 1. These are the individual steps. Then we add these steps together to get the position after 16 steps. Then we replicate this procedure 1000 times.

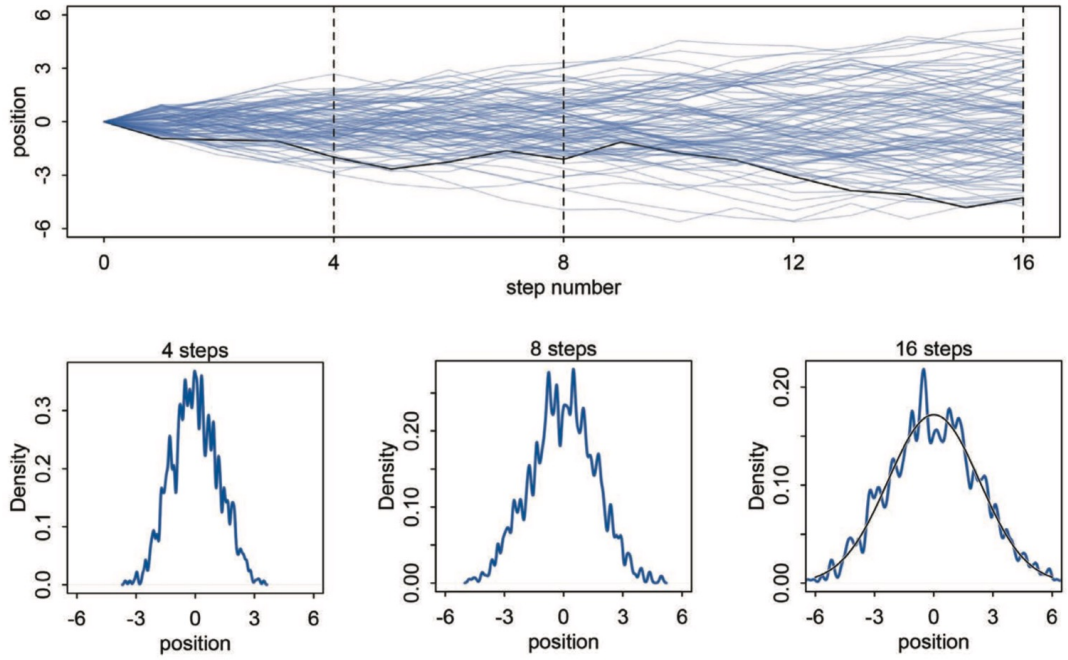
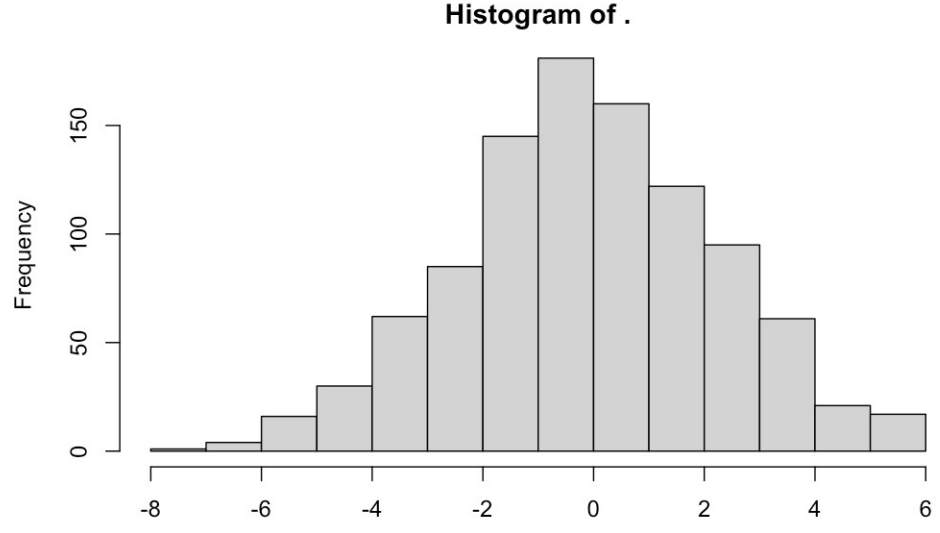


FIGURE 4.2. Random walks on the soccer field converge to a normal distribution. The more steps are taken, the closer the match between the real empirical distribution of positions and the ideal normal distribution, superimposed in the last plot in the bottom panel.

```{r}
replicate(1000, sum(runif(16, -1, 1))) %>% hist
```



Any process that adds together random values from the same distribution converges to a normal.

- Whatever the average value of the source distribution, each sample from it can be thought of as a fluctuation from that average value.
- When we add these fluctuations together, they also cancel one another out.
- The more terms in the sum, the more chances for each fluctuation to be cancelled by another, or by a series of smaller ones in the opposite direction.
- Eventually the most likely sum, in the sense that there are the most ways to realize it, will be a sum in which every fluctuation is cancelled by another, a sum of zero (relative to the mean).
- CLT: the underlying distribution doesn't matter. It could be uniform, like in our example above, or it could be (nearly) anything else. Depending upon the underlying distribution, the convergence might be slow, but it will be inevitable.

Normal by multiplication- **Multiplying small numbers is approximately the same as addition.**

Suppose the growth rate of an organism is influenced by a dozen loci, each with several alleles that code for more growth. these loci interact with one another, such that each increase growth by a percentage. This means that their effects multiply, rather than add.

```
replicate( 10000 , prod( 1 + runif(12,0,0.1) ) )
```

sample 12 random numbers between 1.0 and 1.1, each representing a proportional increase in growth. 1.0 means no additional growth and 1.1 means a 10% increase. The product of all 12 is returned. what distribution will these random products take?

We again get convergence towards a normal distribution, because the effect at each locus is quite small.

Normal by log-multiplication.

```
replicate( 10000 , log(prod(1 + runif(12,0,0.5))) )
```

- Yet another Gaussian distribution
- We get the Gaussian distribution back, because adding logs is equivalent to multiplying the original numbers.
- **even multiplicative interactions of large deviations can produce Gaussian distributions, once we measure the outcomes on the log scale**.
- Since measurement scales are arbitrary, there's nothing suspicious about this transformation.

  After all, it's natural to measure sound and earthquakes and even information on a log scale.

Using Gaussian distributions

- The justifications for using the Gaussian distribution fall into two broad categories: (1) **ontological** and (2) **epistemological**.
- ontologically, the world is full of Gaussian distributions, approximately.

  Measurement errors, variations in growth, and the velocities of molecules all tend towards Gaussian distributions. These processes, at their heart, add together fluctuations.
- **repeatedly adding fluctuations results in a distribution of sums that have shed all information about the underlying process, aside from mean and spread**.
- One consequence of this is that **Gaussian distributions cannot reliably model micro-process**.

  these models can do useful work, even when they cannot identify process. If we had to know the development biology of height before we could build a statistical model of height, human biology would be sunk.

By the epistemological justification, the Gaussian represents a particular state of ignorance.

- When all we know or are willing to say about a distribution of continuous values is their mean and variance, then the Gaussian distribution is the most consistent with our assumptions.
- the Gaussian distribution is the most natural expression of our state of ignorance, because if all we are willing to assume is that a measure has finite variance, the it is the shape that can be realized in the largest number of ways and does not introduce any new assumptions.
  - It is the least surprising and least informative assumption to make.
  - If you don't think the distribution should be Gaussian, then that implies that you know something else that would improve inference.
  - This justification is premised on information theory and maximum entropy

Heavy tails.

- The Gaussian distribution is common in nature. But there are risks in using it as a default data model.
- the Gaussian has some very thin tails—there is very little probability in them.
  - most of the probability mass lies within one standard deviation of the mean.
- Many natural (and unnatural) processes have much heavier tails.
- These processes have much higher probabilities of producing extreme events.

  An important example is financial time series—the ups and downs of a stock market can look Gaussian in the short term, but over medium and long periods, extreme shocks make the Gaussian model (and anyone who uses it) look foolish.

The linear model strategy is to make the parameter for the mean of a normal distribution (μ) into a linear function of the predictor variable and new parameters that we invent.

- the model assumes that the predictor variable has a constant and additive relationship to the mean of the outcome.
- some of the parameters stand for the strength of association between the mean of the outcome, μ, and the value of some x-variable. For each combination of parameter values, the machine computes the posterior probability, which is a measure of relative plausibility, given the model and data.
- **the posterior distribution ranks the infinite possible combinations of parameter values by their logical plausibility**.
  - We ask the model: "Consider all the lines that relate one variable to the other. Rank all of these lines by plausibility, given these data."

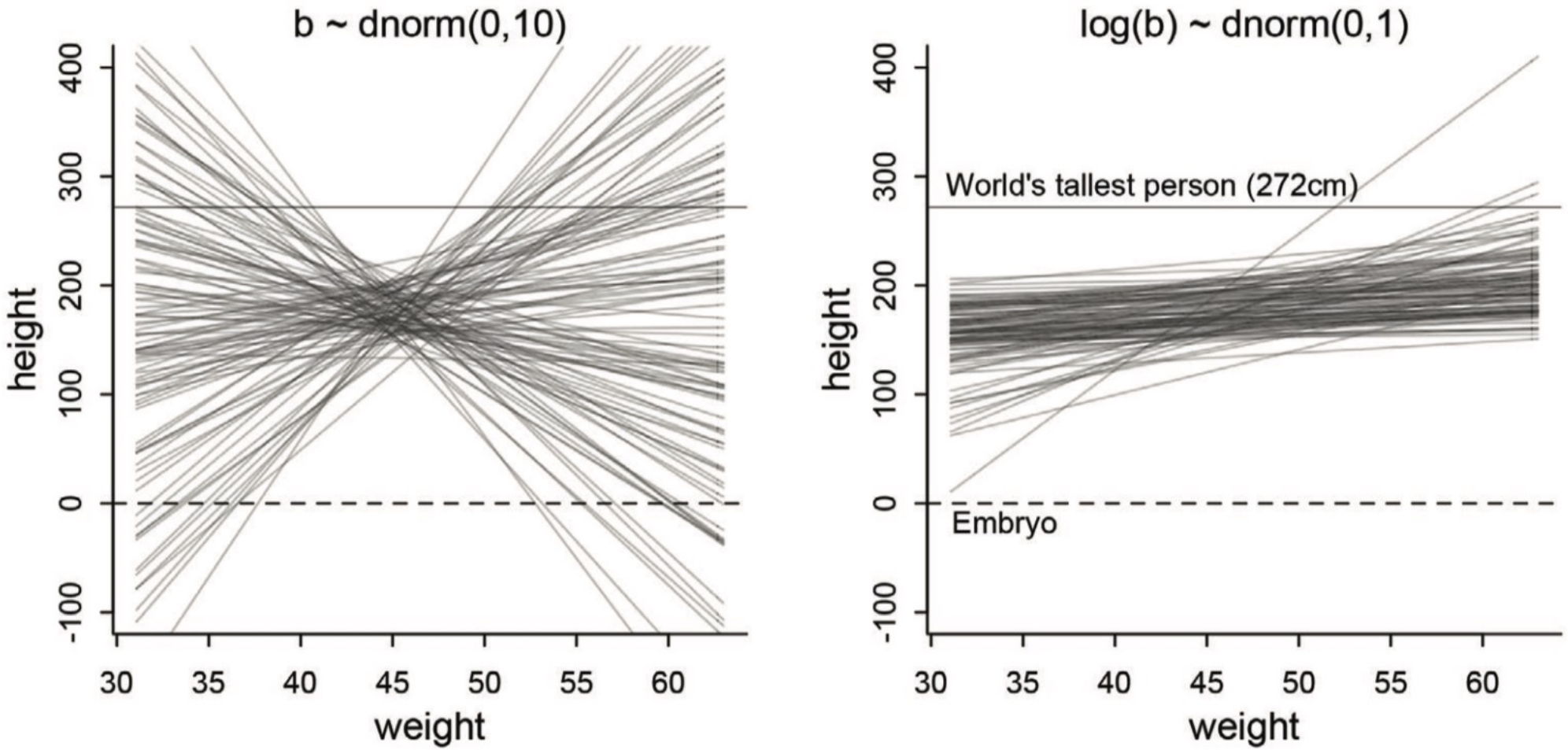
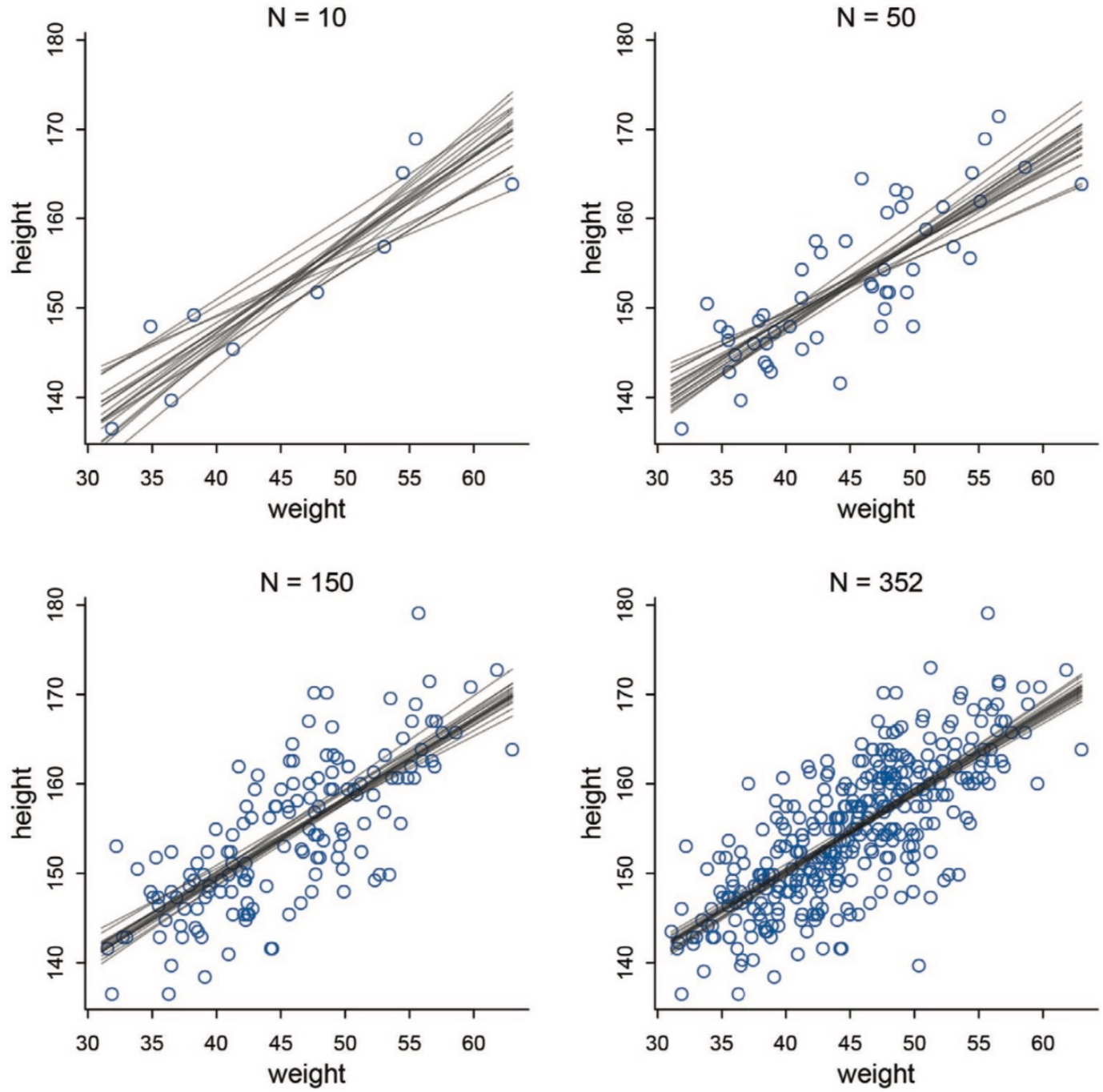


FIGURE 4.5. Prior predictive simulation for the height and weight model. Left: Simulation using the $\beta \sim \text{Normal}(0, 10)$ prior. Right: A more sensible $\log(\beta) \sim \text{Normal}(0, 1)$ prior.


N = 10
N = 50
N = 150
N = 352
height
weight
140
150
160
170
180
30
35
40
45
50
55
60

PREDICTIONS from the model

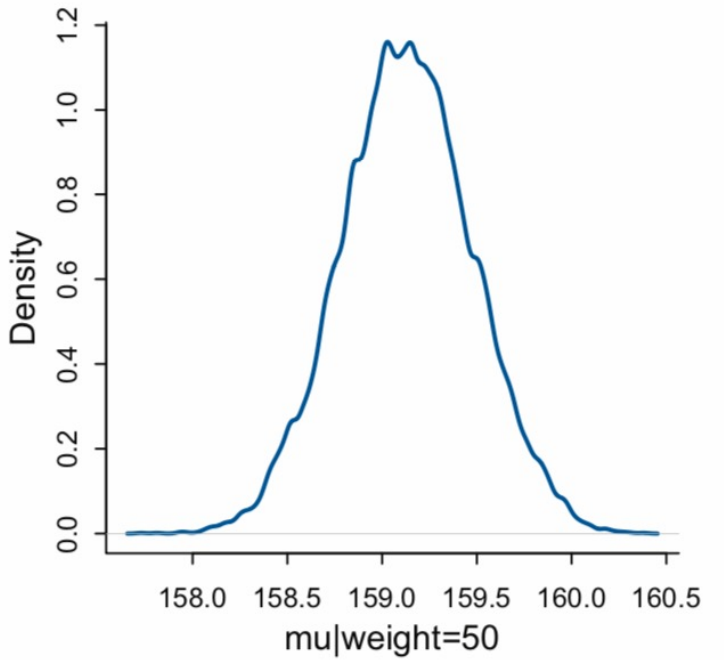


FIGURE 4.8. The quadratic approximate posterior distribution of the mean height, μ, when weight is 50 kg. This distribution represents the relative plausibility of different values of the mean.

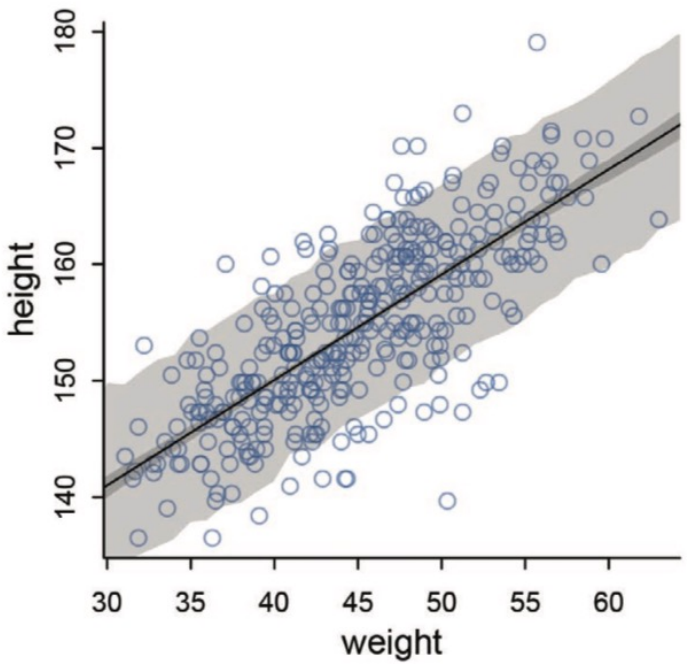


FIGURE 4.10. 89% prediction interval for height, as a function of weight. The solid line is the average line for the mean height at each weight. The two shaded regions show different 89% plausible regions. The narrow shaded interval around the line is the distribution of μ. The wider shaded region represents the region within which the model expects to find 89% of actual heights in the population, at each weight.

we encounter both uncertainty in parameter values and uncertainty in a sampling process.

- These are distinct concepts, even though they blend together in the posterior predictive simulation.
- The posterior distribution is a ranking of the relative plausibilities of every possible combination of parameter values.
- The distribution of simulated outcomes (height) is instead a distribution that includes sampling variation from some process that generates Gaussian random variables.

  This sampling variation is still a model assumption. It's no more or less objective than the posterior distribution.
- Both kinds of uncertainty matter, at least sometimes. But it's important to keep them straight, because they depend upon different model assumptions.

  it's possible to view the Gaussian likelihood as a purely epistemological assumption (a device for estimating the mean and variance of a variable), rather than an ontological assumption about what future data will look like. In that case, it may not make complete sense to simulate outcomes.

multiple regression

- Statistical “control” for confounds. A confound is something that misleads us about a causal influence. confounds are diverse. They can hide important effects just as easily as they can produce false ones.
- Multiple and complex causation. A phenomenon may arise from multiple simultaneous causes, and causes can cascade in complex ways. And since one cause can hide another, they must be measured simultaneously.
- Interactions. The importance of one variable may depend upon another.

  plants benefit from both light and water. But in the absence of either, the other is no benefit at all.

does marriage cause divorce?

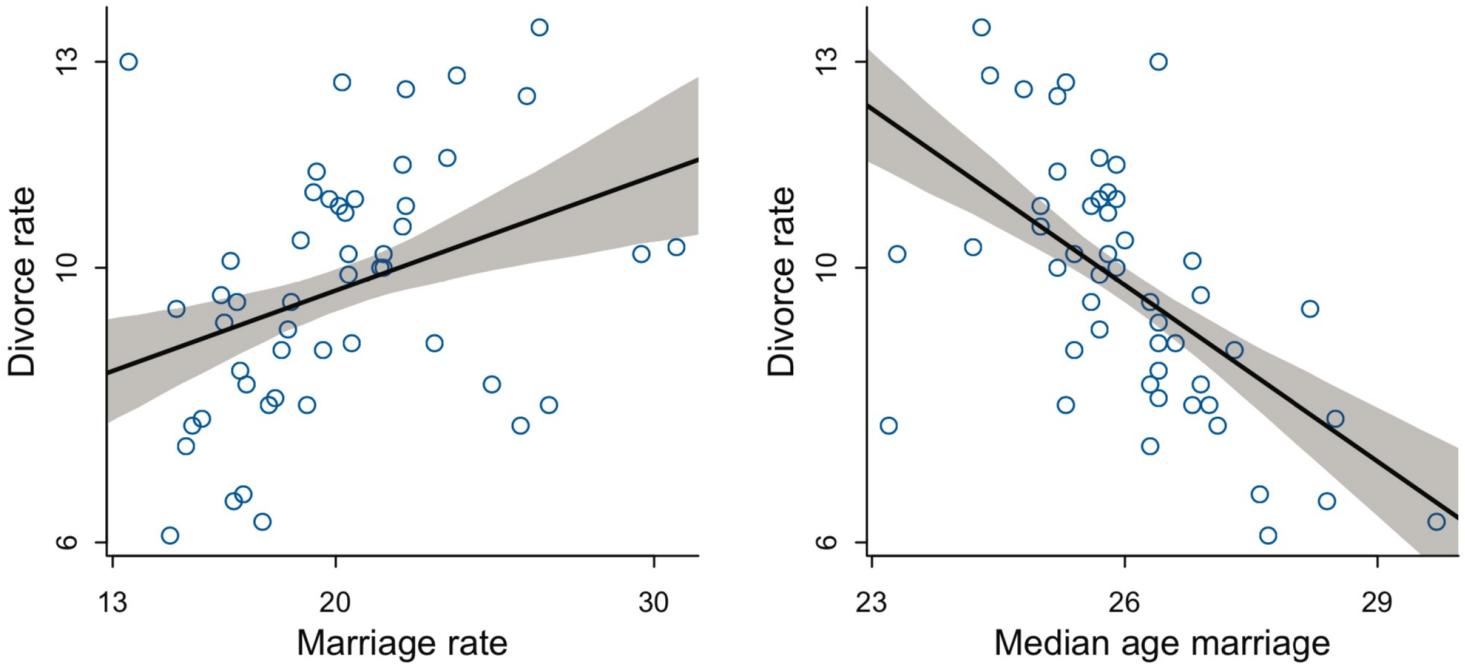


FIGURE 5.2. Divorce rate is associated with both marriage rate (left) and median age at marriage (right). Both predictor variables are standardized in this example. The average marriage rate across States is 20 per 1000 adults, and the average median age at marriage is 26 years.

multiple regression can address a descriptive question: Is there any additional value in knowing a variable, once I already know all of the other predictor variables?

- So once you fit a multiple regression to predict divorce using both marriage rate and age at marriage, the model addresses the questions:

1. After I know marriage rate, what additional value is there in also knowing age at marriage?
2. After I know age at marriage, what additional value is there in also knowing marriage rate?

main effects are additive combinations of variables, the simplest type of multiple variable model.

these models can help us with:

1. revealing spurious correlations
2. revealing important correlations that may be masked by unrevealed correlations with other variables.

- multiple regression can be worse than useless, if we don't know how to use it. Just adding variables to a model can do a lot of damage.

What's a cause?

- Knowing a cause in statistics means being able to correctly predict the consequences of an intervention.
- There are contexts in which this is complicated. For example, changing someone's body weight would mean intervening on another variable, like diet, and that variable would have other causal effects in addition. But being underweight can still be a legitimate cause of disease, even when we can't intervene on it directly.

Causality, statistics and scientific method

- the goal is the mechanism (of cancer, of love, of whatever)

  alternative view: scientific theory should be expressed in terms of predictive, mathematical relations. A nice equation expresses so much more clearly than a complex causal story that often entails a variety of exceptions.

  Russell: ‘‘The law of causality, I believe, like much that passes muster among philosophers, is a relic of a bygone age, surviving, like the monarchy, only because it is erroneously supposed to do no harm.’’
- the mechanism is always causal
- the mechanism produces data
- statistical model is a tool for elucidating the mechanism
- but the stat model is acausal (describes co-dependencies and predicts new data)
- therefore, we need an additional component to use the model in a causal way
  - both to build “right” models and to draw causal inferences from fitted models
- we need a formal description of our causal beliefs

a mechanism is a process that takes causes and produces effects

- statistical model is fitted on data to model the mechanism that generated these data
- the model is stochastic, so it models the uncertainty involved
- but its underlying structure is deterministic. This part models our hypothesis about the data generating mechanism.
  - linear models are used, because they work surprisingly well – although in terms of modelling the data generating process they are abit like Ptolemaic epicycles.

conclusions are drawn from experiments through comparison

- the value of an effect must be compared in two different worlds: the world in which the cause is at one value and one in which the cause is at a different value, because a causal statement is a claim about two (or more) worlds simultaneously.

Causal relations imply counterfactuals (statements that concerns other worlds)

- a causal relation doesn't merely imply that events happened together, but that there's some **generating mechanism** that produces an event (data) when engaged by an event of another type.
  - in some other world, in which the mechanism had not been engaged, the effect would not have resulted.
  - counterfactual dependence: the effect would not have occurred if the cause had not.
- Causal processes generate possibilities - actual or counterfactual.
  - Causal relations tell how to get there from here, whether you're going or not.
- As **generative processes**, they differ fundamentally from correlations, for a correlation is merely a **description** of what's been observed.

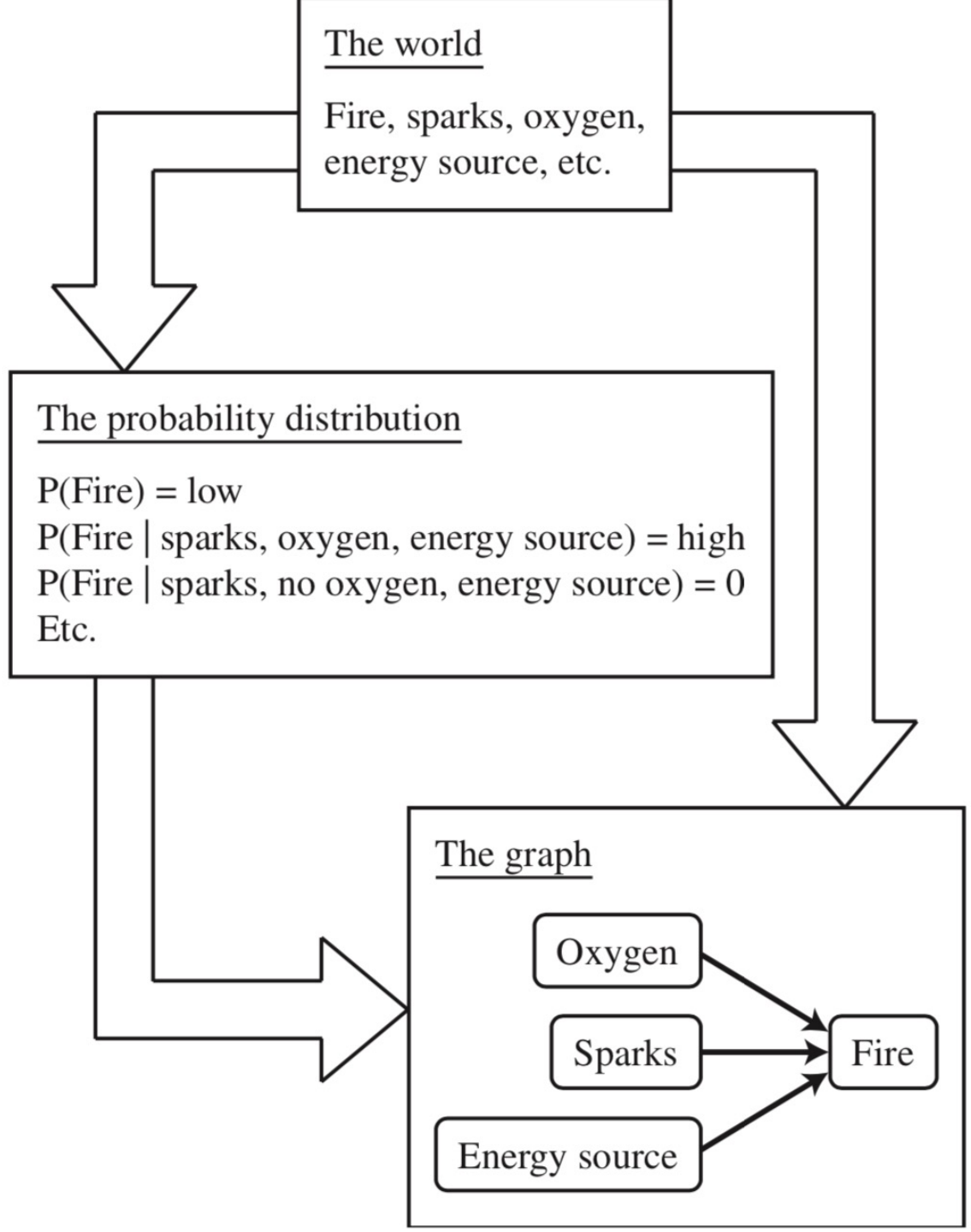
causal DAGs (**D**irected **A**cyclic **G**raphs)

- causes don't always have to produce their effects; they only have to produce them sometimes; if the cause and effect are related deterministically, that's all right, too.
- an acyclic graph composed of nodes and links represents the causal structure of a system, with nodes corresponding to events or variables and arrows between nodes corresponding to causal relations.
- In this kind of scheme, three entities are involved:

  (i) the causal system being represented,

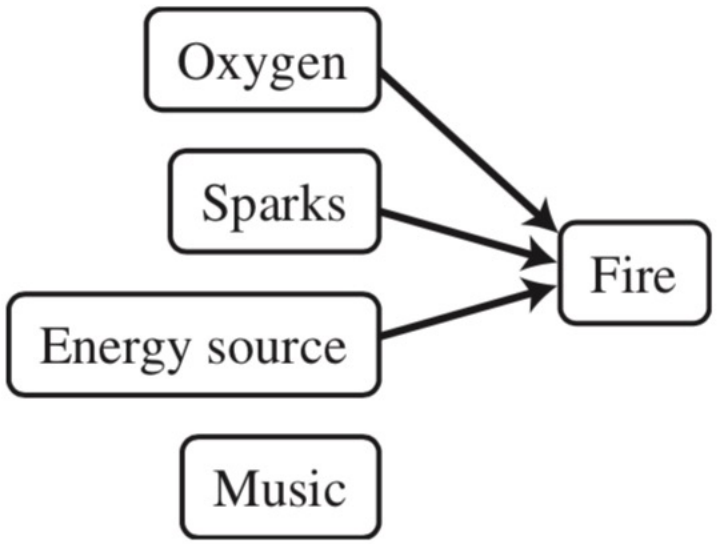
  (ii) the probability distribution that describes how likely events are to happen and how likely they are to occur with other events, and

  (iii) a graph that depicts the causal relations in the system


The world
Fire, sparks, oxygen, energy source, etc.
The probability distribution
P(Fire) = low
P(Fire | sparks, oxygen, energy source) = high
P(Fire | sparks, no oxygen, energy source) = 0
Etc.
The graph
Oxygen
Sparks
Energy source
Fire

- The probability distribution and graph both represent the world, and the graph also represents the probability distribution.
- A representation has to be simpler than the thing being represented, otherwise it would be the thing itself.
- Probability distributions are representations of the world because they specify how much confidence we can have that an event will occur if we know that another one has.
- Graphs are representations of probability distributions because they specify the causal relations among events that are implicit in the probabilities.
  - they depict the causal relations responsible for the probabilistic ones.
  - They show the structure of the causal mechanism that generates the probability distribution.

if there's no path between two events in the graph, then this should be reflected in the probability distribution. For instance, music is unrelated to fire, so if the presence of music were included in the causal system, it would be represented as a disconnected node in the graph:



This relation between fire and music is called **independence** because the probability of fire is the same whether music is playing or not.

Every time we say events are independent, we have said something important because it simplifies the graph by removing links and makes it easier to do calculations with probabilities.

Figure 4.3

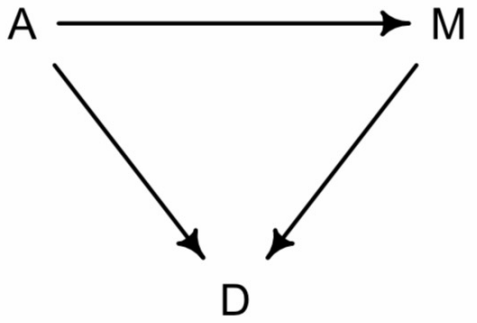
The probability distribution would reflect this by making sure that

$$P(\text{fire}) = P(\text{fire} \mid \text{music}) = P(\text{fire} \mid \text{no music}).$$

directed acyclic graph (DAG) is a way of describing qualitative causal relationships among variables

- Graph means it is nodes and connections.
- Directed means the connections have arrows that indicate directions of causal influence.
- Acyclic means that causes do not eventually flow back on themselves.
- A DAG isn't as detailed as a full model description, but it contains information that a purely statistical model does not.

  Unlike a statistical model, a DAG will tell you the consequences of changing a variable.



(1) A directly influences D
(2) M directly influences D
(3) A directly influences M

Structural Equations

- to think about the graph and the probabilities as parts of a single representation associate each node of a causal model with a set of conditional probabilities.
- Each node is the joint effect of all links pointing into it.
  - So the effect (fire) is fully described by stating the probability of fire under all possible combinations of its causes:

$$
\begin{aligned}
&P(\text{Fire} \mid \text{sparks, oxygen, energy source}) = \text{high} \\
&P(\text{Fire} \mid \text{sparks, oxygen, no energy source}) = 0 \\
&P(\text{Fire} \mid \text{sparks, no oxygen, energy source}) = 0 \\
&P(\text{Fire} \mid \text{sparks, no oxygen, no energy source}) = 0 \\
&P(\text{Fire} \mid \text{no sparks, oxygen, energy source}) = \text{very low} \\
&P(\text{Fire} \mid \text{no sparks, oxygen, no energy source}) = 0 \\
&P(\text{Fire} \mid \text{no sparks, no oxygen, energy source}) = 0 \\
&P(\text{Fire} \mid \text{no sparks, no oxygen, no energy source}) = 0
\end{aligned}
$$

- assuming only two possible values, $2^3 = 8$ conditional probabilities are required to describe this mechanism producing fire.
- To describe the system fully, we also need to know the marginal probabilities of sparks, of oxygen, and of an energy source.
  - 8 + 3 numbers would be a complete representation of this causal system.
- structural equation modeling makes the representation smaller by specifying how the causes combine to produce the effect.
- Structural equation modeling expresses the functional relation representing the mechanism using a different representational scheme than Pr-s and graphs, but it carries isomorphic representation.
- Each effect is associated with a function expressing how it is produced by its causes.

  Fire = f(spark, oxygen, energy); f means that all causes are required.
- the function is probabilistic, including an error variable.
- it describes the mechanism by expressing how the causes are combined to produce the effect.
- Sources of randomness can sometimes be reduced to other causal mechanisms that we happen to be ignoring.

Causal Structure Produces a Probabilistic World

- We posit that stable probabilistic relations between the observed variables of a system are generated by an underlying causal structure.
- the world can be attributed to the operation of a big, complicated network of causal mechanisms.
- On a smaller scale, **particular causal structures lead to particular patterns of probability in the form of particular patterns of dependence and independence**.

causal salad

postacute_therapy ~ age + sex + comorbidity +dementia + management_method + acute_length_of_stay + acute_therapy + (dementia | county),

hu ~ age + sex + comorbidity + dementia + management_method + acute_length_of_stay + acute_ therapy + (dementia | county)),

family=hurdle_lognormal()

European Geriatric Medicine
https://doi.org/10.1007/s41999-020-00348-5

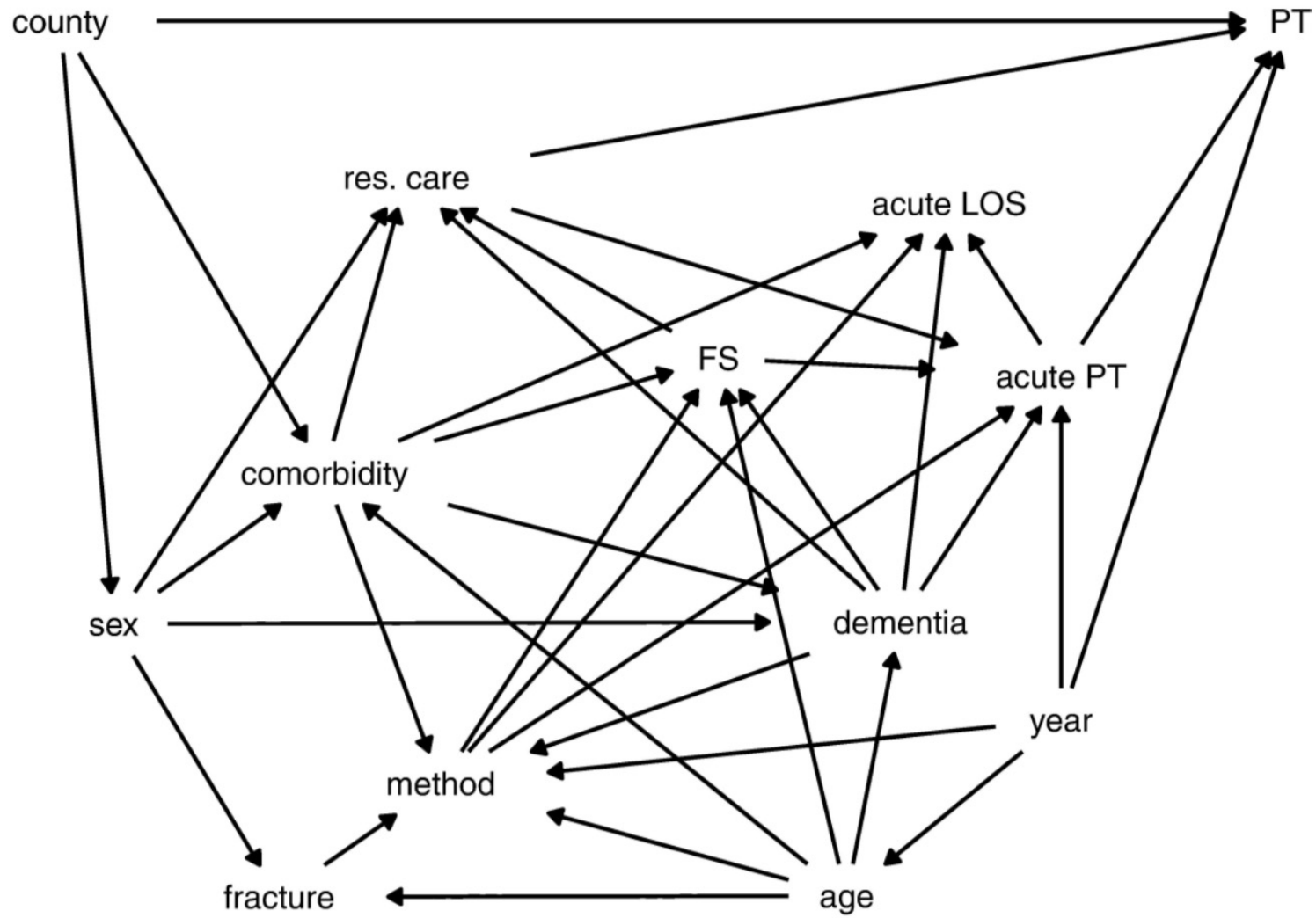
RESEARCH PAPER

Check for updates

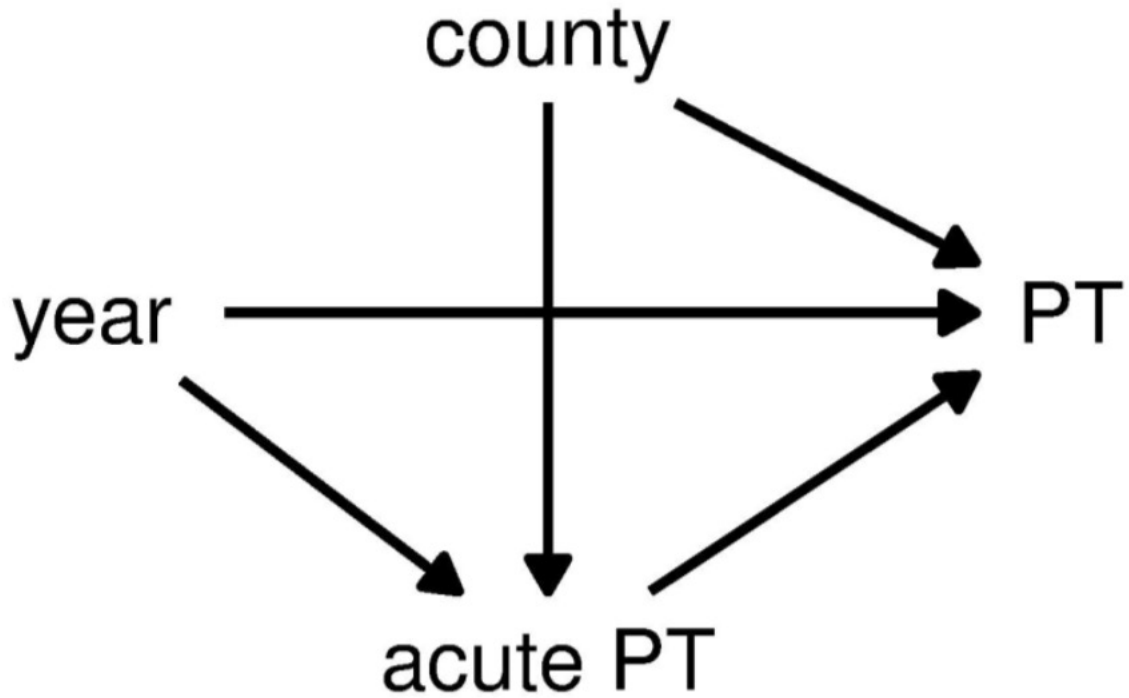
High variability in hip fracture post-acute care and dementia patients having worse chances of receiving rehabilitation: an analysis of population-based data from Estonia

Pärt Prommik[1,2,4] · Helgi Kolk[1,2] · Ülo Maiväli[3] · Mati Pääsuke[4] · Aare Märtson[1,2]

Received: 9 January 2020 / Accepted: 7 June 2020



Supplementary Figure 3. Directed acyclic graph of the association between the county of residence and received post-acute physical therapy after hip fracture. Available variables: county of residence (county), post-acute physical therapy (PT), age, sex, comorbidity status identified as CCI score, fracture type (fracture), fracture management method (method), received physical therapy during acute care (acute PT), acute length of stay (acute LOS), and study year (year). Included unobserved variables were residential care status (res. care) and functional status (FS).



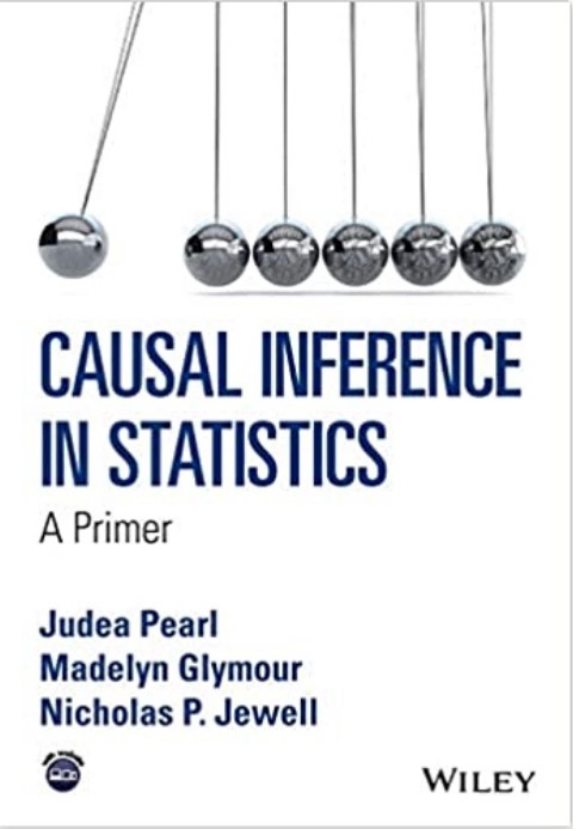
Supplementary Figure 4. Directed acyclic graph of the association between study year and received post-acute physical therapy after hip fracture. Variables are: study year (year), post-acute physical therapy (PT), county of residence (county), and received physical therapy during acute care (acute PT).

why do regression

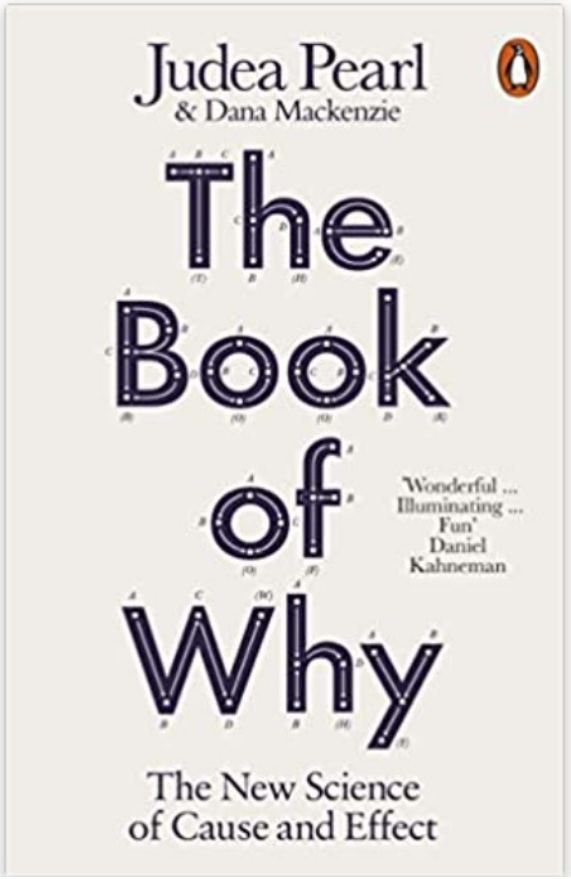
1. Test causal hypotheses
2. Prediction – put in the vars that improve out-of-sample fit and exclude vars whose slope is reliably 0 (use CI, not p value). main danger is overfitting
3. select predictors: (i) does var_x predict Y?, (ii) how big is var_x-s influence?, (iii) what is the set of best predictors? We prefer simple models with a couple of vars, which we interpret very carefully.
4. Study individuals, using model predictions & residuals. Adjusted models (single measurement per unit) or multilevel models (repeat measurement).

predictive vs causal regression

- predictive – you predict mean y, or new data points, or new data groupings
  - search for the best regression equation for predictive purposes
  - care about out-of-sample fit, a.k.a. predictive accuracy (training set vs test set, AIC, WAIC, LOO, cross-validation, loo_R2)
  - do not care about collinearity, different biases, including post-treatment vars, etc.
- Causal – you interpret coefficients (thereby testing/falsifying causal models), compare, develop causal models
  - you specify 2 mathematical structures: regression equation & DAG (a qualitative causal scheme describing your causal model). Then you use the 1st to test the 2nd.
  - out-of-sample fit is (almost) irrelevant
  - you care about included, non-included, unmeasured & unknown predictors, model structure vis-a-vis data generating mechanism



causal regression



1. cause precedes effect
2. changing the cause will change the effect (experiment)

proving a causal effect through counterfactuals: had the cause been different, the effect would have been different too.

counterfactuals kind of assume a multiverse, where in universe 1 everybody gets the drug and in universe 2 everybody gets the placebo – this is **the fundamental problem of causation**

conditioning (stratifying) the linear model

- Y – height, X_1 – weight, X_2 - sex
- model1: $Y = b_0 + b_1X_1 + Normal(0, sigma)$ $lm(Y \sim X_1)$
- model2: $Y = b_0 + b_1X_1 + b_2X_2 + Normal(0, sigma)$ $lm(Y \sim X_1 + X_2)$

in model1 we estimate the average height for every (in)conceivable exact value of weight (from – Inf to Inf). We assume that relationship is a straight line in our measurement scales of height & weight, and that variation in individual heights (sigma) is the same for all weights.

In model 2 we separate our data by sex, but still assume that for all sexes and all weights the variation is equal. Now we estimate 2 lines, one for females and other for males. These lines are constrained to be parallel.

b_1 tells, how much we expect the mean Y to change, if X_1 changes by 1 unit, and X_2 is kept constant.

b_2 tells, how much we expect the mean Y to change, if X_2 changes by 1 unit, and X_1 is kept constant.

Lets suppose that we want to give a causal interpretation to X_2 (sex causes height – if we do a sex change, we expect height to change too). Now, if there are hidden in the error term unmeasured variables that affect Y and independently of that affect X_2, then these vars can bias the b_2, on which we are going to base our causal conclusions.

This means that we need to know the full complement of causal factors, in order to interpret X_2 causally (we do not necessarily need to known their values, just their existence).

Drug vs no drug – 1.29 RR
men – 0.53 RR
women – 0.87 RR

	Drug	No drug
Men	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
Women	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined data	273 out of 350 recovered (78%)	289 out of 350 recovered (83%)

If we know the patient's sex we should prescribe the drug, but if sex is unknown we should not?

How can lack of knowledge of the patient's sex make the drug harmful?

The answer is nowhere to be found in simple statistics.

- **additional facts**:
  - women are less likely to recover than men.
  - women are more likely to take the drug than men are.
- the reason the drug appears to be harmful overall is that, if we select a drug user at random, that person is more likely to be a woman.
- Being a woman is a common cause of both drug taking and failure to recover.
  - Therefore, we need to compare subjects of the same sex, thereby ensuring that any difference in recovery rates between those who take the drug and those who do not is not ascribable to sex.
  - We assume for both sexes separately that the people who got the drug are otherwise similar to those, who did not.
- **we should consult the segregated data, which shows that the drug works.**

- numbers are the same, but now we are recording BP & blood pressure.
- Drug vs no drug – 1.29 RR
  Low BP – 0.53 RR
  High BP – 0.87 RR

Table 1.2 Results of a study into a new drug, with posttreatment blood pressure taken into account

	No drug	**Drug**
Low BP	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
High BP	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined data	273 out of 350 recovered (78%)	289 out of 350 recovered (83%)

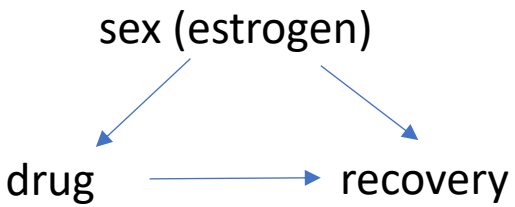
lowering blood pressure is one of the mechanisms by which treatment affects recovery.

though the numbers are the same in the gender and blood pressure examples, the correct result lies in the segregated data for the former and the aggregate data for the latter.

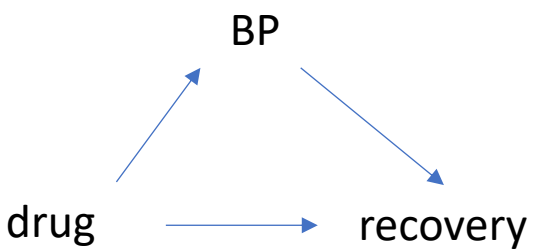
- **None of the information that allowed us to make the decision is in the data**.
- In our gender example we assume that treatment cannot affect sex. If it could, the causal story behind the data would assume the same structure as in our blood pressure example.
- Trivial though the assumption “treatment does not cause sex” may seem, there is no way to test it in the data.

The mechanism that generated the data determines data analysis – you must incorporate causal structure into your regression equation.

If your causal model is correct, then your regression tests the strength of causation



Recovery ~ drug + sex



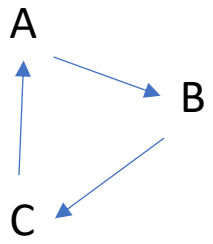
Recovery ~ drug

Siin huvitab meid ravimi summaarne mõju mõlemaid teid mööda
Aga mis siis, kui meid huvitab ainult ravimi otsemõju (BP-st sõltumatu mõju)?

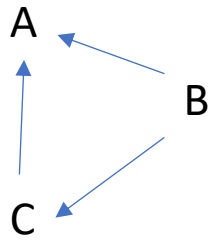
DAG – directed acyclic graph

A – B is not a DAG

A -> B is a DAG



is not a DAG



is a DAG

- Directed graph - all edges are directed.
- The node that a directed edge starts from is the **parent** of the node that the edge goes into; the node that the edge goes into is the **child** of the node it comes from.
- A directed path can be traced along the arrows (no node on the path has two edges on the path directed into it, or two edges directed out of it).
- If two nodes are connected by a directed path, then the first node is the **ancestor** of every node on the path, and every node on the path is the **descendant** of the first node.

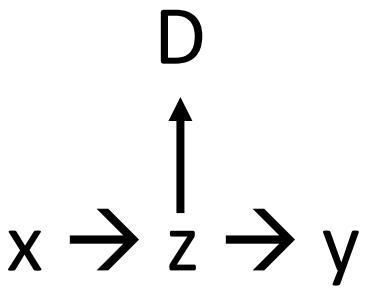
4 elementary structures

A -> B -> C **pipe** – conditioning on B closes path (information flow between A & C)

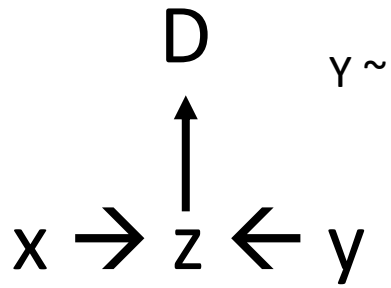
A <- B -> C **fork** – conditioning on B closes path

A -> B <- C **collider** – conditioning on B opens path

Y ~ X + D closes x -> y pipe



descendant



Y ~ X + D opens x -> y collider

- if the two variables are disconnected, those variables should be unrelated, or probabilistically **independent**.
- if a causal arrow points from one variable to another, they should be related, or **dependent**.
- **Pipe *A → B → C***, A and C are dependent, as A is an indirect cause of C. Changes in A should produce changes in C.
  - if we hold B fixed, A and C are independent (A tells us nothing about C when we already know the value of B).
  - A and C are **conditionally independent,** given B.
- **Fork A <-- B --> C,** B is a common cause of A and C.
  - A and C are dependent and this relation is again mediated by B.
  - A and C are **conditionally independent ,** given B.
- **Collider A → B ← C**, B is a common effect of both A and C
  - A and C are independent, if the value of B is unknown as they have no common causes and they don't cause each other.
  - A and C are **conditionally dependent**, given B.

- causal graph allows us to make inferences about dependence and independence.
- Conversely, given the observation of (conditional) dependencies, one can infer something about the underlying causal structure.
- A and B could change together for many reasons; you could produce many causal models, some of which could be tested by examining other dependencies.
- If A and B are dependent and A and C are dependent but independent given all possible values of B, then 3 DAGs are possible to express these relations.

$$A \rightarrow B \rightarrow C;$$
$$C \rightarrow B \rightarrow A;$$
$$A \leftarrow B \rightarrow C$$

- These cannot be distinguished using relations of independence/dependence, or any kind of probabilistic information alone.
- Some other kind of information is required, like the temporal order, or the effect of an intervention

every causal DAG is built out of these 4 elements.

- you know how to open and close each, so you can figure out which variables you need to include or not include.

1. List all of the paths connecting X (the potential cause of interest) and Y.
2. Classify each path by whether it is open or closed. A path is open unless it contains a collider.
3. Classify each path by whether it is a backdoor path. A backdoor path has an arrow entering X.
4. If there are any open backdoor paths, decide which variable(s) to condition on to close it (if possible).

d (directional)-separation: some vars on a DAG are independent of others, there is no connecting path.

- *d*-separation is a criterion for deciding, from a DAG, whether a set *X* of variables is independent of another set *Y,* given a third set *Z.*
- **if Y and X are independent, then in the model Y ~ X, slope = 0.**
- **conditional independence**: Y is independent on X, if we condition on Z.
- **Rule 1:** *x* and *y* are *d*-connected if there is an unblocked path between them.
- **Rule 2:** *x* and *y* are d-separated by the conditioning set Z if there is no collider-free path between *x* and *y* (every path between *x* and *y* is "blocked" by *Z).*
- **Rule 3:** If a collider is a member of the conditioning set *Z,* or has a descendant in *Z,* then it no longer blocks any path that traces this collider.

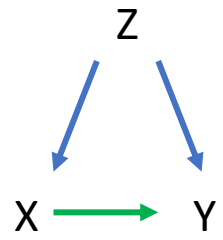
DAGs are not enough

- If you don’t have a real, mechanistic model of your system, DAGs are fantastic. They make assumptions transparent and easier to critique. And they highlight the danger of using multiple regression as a substitute for theory.
- But once you have a dynamical model of your system, you don’t need a DAG. In fact, many dynamical systems have complex behavior that is sensitive to initial conditions, and so cannot be usefully represented by DAGs. But these models can still be analyzed and causal interventions designed from them.
- domain specific structural causal models can make causal inference possible even when a DAG with the same structure cannot decide how to proceed. Additional assumptions give us power.

Confounding and experiment

- Z – confounder (segav muutuja)
- Nii x kui Z mõjutab Y-t segatud efektina.

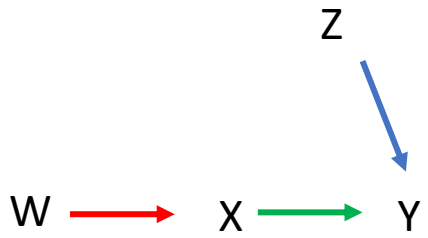
Y ~ X -- confounding: Z biases the slope of X
Y ~ X + Z -- backdoor is closed – no confounding



W - randomizer (exp treatment, medndelian randomization)

the value of X is now independent of Z, but it now depends on W (average Z is the same in all values of W)

randomized treatment W deletes all incoming arrows to X, thus removing all confounders



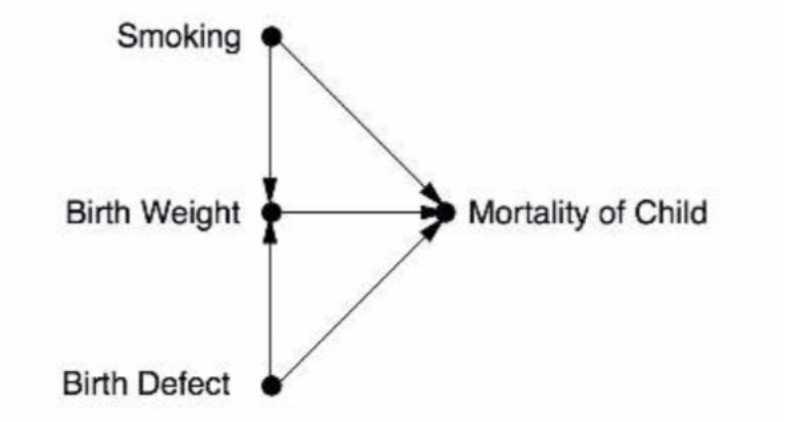
Y ~ W

Deconfounding & back door criterion

- ***back-door path****:* any path from *X* to *Y* that starts with an arrow pointing into *X*.
  - open backdoor paths bias (confound) your causal interpretation of X
- **direct path** starts with an arrow pointing out of X and eventually reaches X
  - direct paths reflect the effect of experimental treatment on Y
- you deconfound your model if (i) by conditioning on a set of confounders Z you close all backdoors, and (ii) Z contains no descendants of X
- to deconfound *X* and *Y*, we block every noncausal (backdoor) path between them without perturbing any causal paths
- If we do this by controlling for some set of variables *Z*, we also need to make sure that no member of *Z* is a descendant of *X* on a causal path; otherwise we might partly or completely close off that path.

Mothers smoking increases baby's survival

- In 1959, Yerushalmy launched a long-term public health study that collected data on >15,000 children.
- Several studies had already shown that the babies of smoking mothers weighed less at birth than the babies of nonsmokers, and it was natural to suppose that this would translate to poorer survival.
- the babies of smokers were lighter on average than the babies of nonsmokers (by seven ounces). **However, the low-birth-weight babies of smoking mothers had a better survival rate** than those of nonsmokers.
- epidemiologists argued about the paradox for forty years.



- birth-weight paradox is a perfect example of collider bias. The collider is Birth Weight.
- Backdoor: smoking → birth weight ← Birth defect → mortality
- Condition on birth defect & birth weight
- Smoking may be harmful in that it contributes to low birth weight, but certain other causes of low birth weight, such as genetic abnormalities, are much more harmful.
- There are two possible explanations for low birth weight in one particular baby: it might have a smoking mother, or it might be affected by one of those other causes. If we find out that the mother is a smoker, this explains away the low weight and consequently reduces the likelihood of a serious birth defect.

does marriage cause divorce?

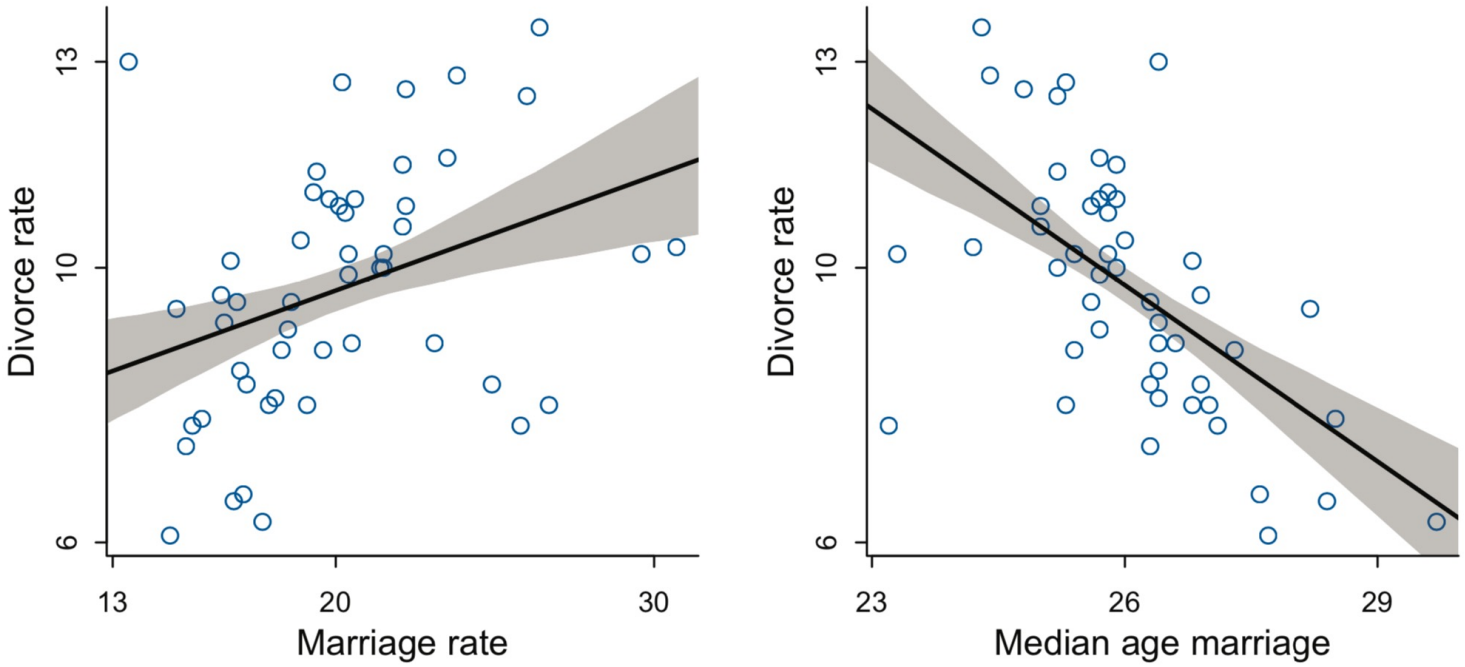
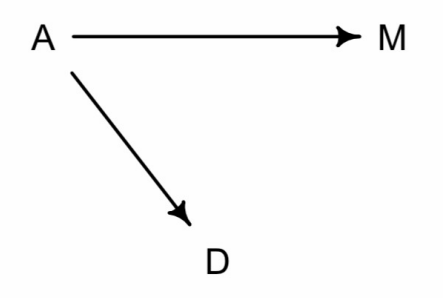


FIGURE 5.2. Divorce rate is associated with both marriage rate (left) and median age at marriage (right). Both predictor variables are standardized in this example. The average marriage rate across States is 20 per 1000 adults, and the average median age at marriage is 26 years.

- 3 observed variables:
- divorce rate (D),
- marriage rate (M),
- median age at marriage (A) in each State.
- only A has a causal impact on the outcome, D, even though both predictors are strongly associated with the outcome.

- These statements can have further implications. age of marriage influences divorce in two ways.
- First it has a direct effect, A → D.
  - Perhaps a direct effect would arise because younger people change faster than older people and are therefore more likely to grow incompatible with a partner.
- Second, it has an indirect effect by influencing the marriage rate, which then influences divorce, A → M → D.
  - If people get married earlier, then the marriage rate may rise, because there are more young people.
- the regression of D on A, tells us only that the total influence of age at marriage is strongly negative with divorce rate.
  - The “total” means we have to account for every path from A to D. There are two such paths: A → D, a direct path, and A → M → D, an indirect path.
  - In general, it is possible that a variable like could be associated with D entirely through the indirect path. That type of relationship is known as mediation.

- the indirect path does almost no work in this case. How can we show that? We know that marriage rate is positively associated with divorce rate. But that isn't enough to tell us that the path M → D is positive. It could be that the association between M and D arises entirely from A's influence on both M and D. Like this:

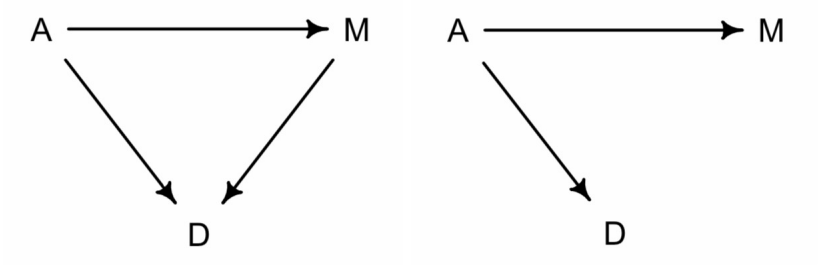


- This DAG is also consistent with models m5.1 and m5.2, because both M and D have information from A. So when you inspect the association between D and M, you pick up that common information.

testable implications of each model.

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

$\not\perp\!\!\!\perp$ - "not independent of."

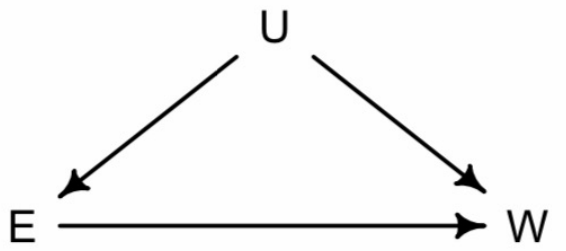


$D \perp\!\!\!\perp M|A$

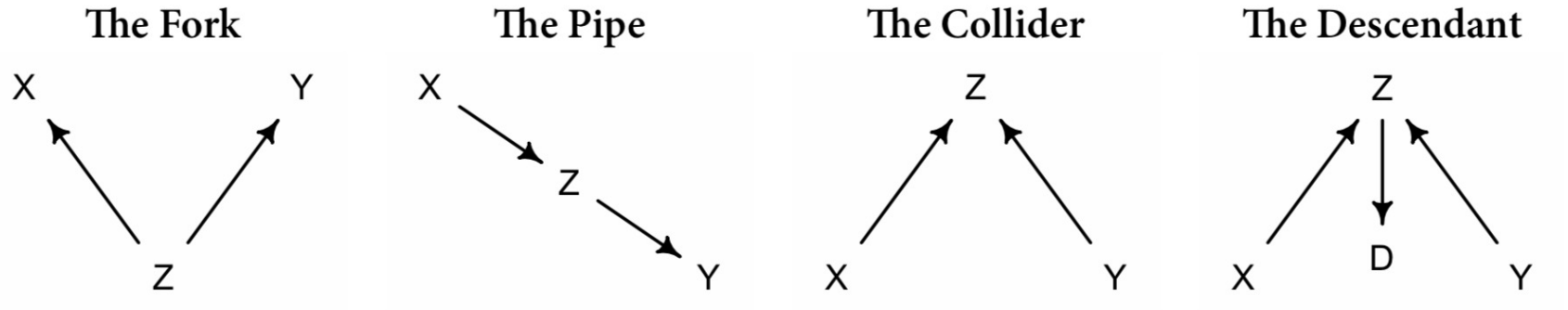
- Any DAG may imply that some variables are independent of others under certain conditions.
- These are the model's testable implications, its **conditional independencies**.
- First, conditional independencies are statements of which variables should be associated with one another (or not) in the data.
- Second, they are statements of which variables become disassociated when we condition on some other set of variables.
  - conditioning on Z means learning its value and then asking if X adds any additional information about Y. If learning X doesn't give any more information about Y, then Y is independent of X conditional on Z, or $Y \perp X|Z$.

Blocking confounding paths between predictor X and some Y is known as shutting the backdoor.

- We don't want any spurious association sneaking in through a non-causal path that enters the back of the predictor X.



- E ← U → W is a backdoor path, **because it enters E with an arrow and also connects E to W.** This path is non-causal—intervening on E will not cause a change in W through this path—but it produces an association between E and W.
- Given a DAG, it is always possible to say which, if any, variables to control for to shut all the backdoor paths.
- It is also possible to say which variables one must not control for, in order to avoid making new confounds.



- **fork** is the classic confounder, where variable Z is a common cause of X and Y, generating a correlation between them. If we condition on Z, then learning X tells us nothing about Y. X and Y are independent, conditional on Z.
- **pipe**: X influences Z which influences Y. If we condition on Z, we also block the path from X to Y. So conditioning of the middle variable again blocks the path.
- **collider**: now there is no association between X and Y unless you condition on Z. Conditioning on the collider variable opens the path and information flow between X and Y. However neither X nor Y has any causal influence on the other.
- A **descendent** is a variable influenced by another variable. Conditioning on a descendent D partly conditions on its parent Z, because D has some information about Z. this partially opens the path from X to Y, because Z is a collider. the consequence of conditioning on a descendent depends upon the nature of its parent. Descendants are common, because often we cannot measure a variable directly and instead have only some proxy for it.

Plotting multivariate posteriors.

1. Predictor residual plots show the outcome against residual predictor values. They are useful for understanding the statistical model, but not much else.

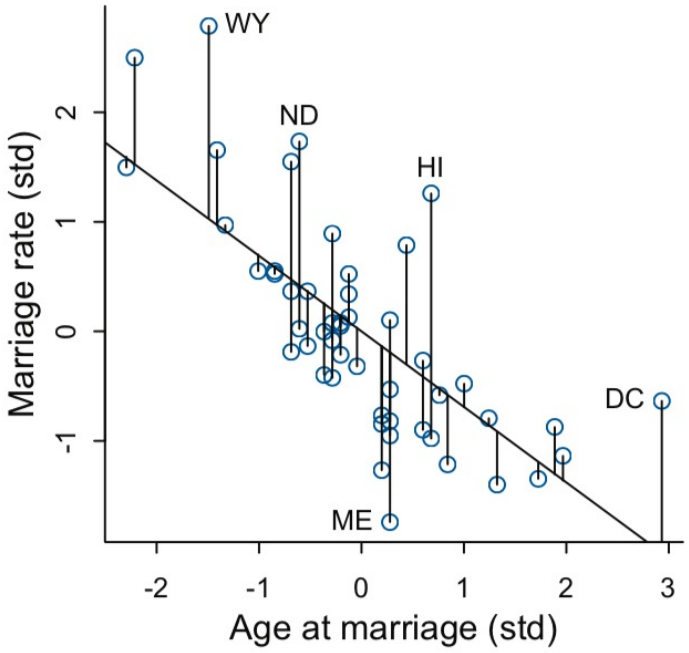
   A predictor residual is the average prediction error when we use all of the other predictor variables to model a predictor of interest.

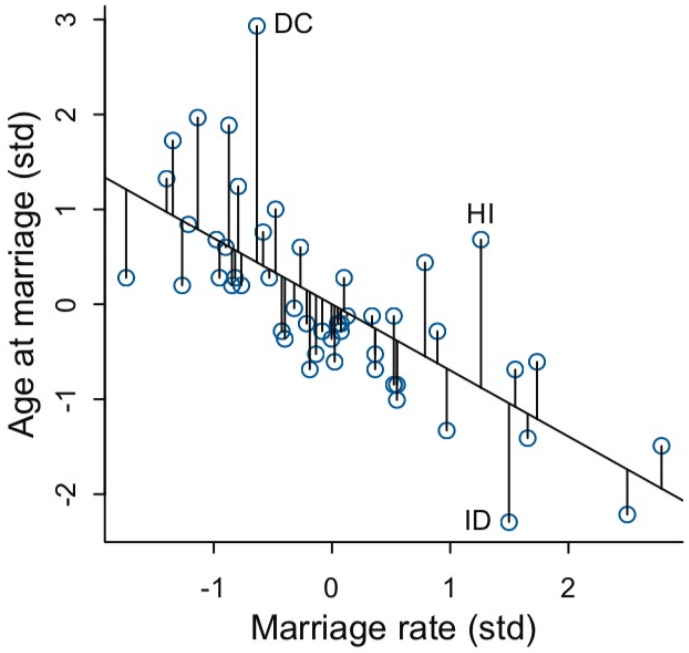
2. Posterior prediction plots show model-based predictions against raw data, or otherwise display the error in prediction. They are tools for checking fit and assessing predictions. They are not causal tools.

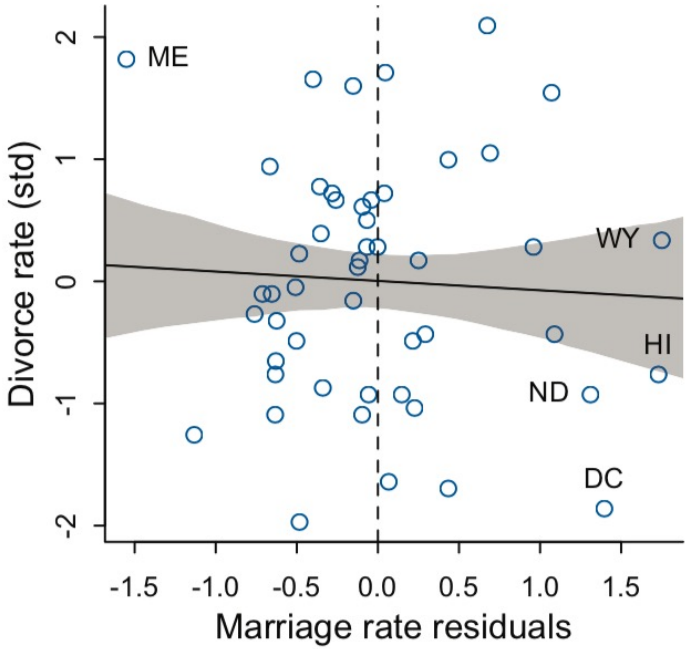
3. Counterfactual plots show the implied predictions for imaginary experiments. allow to explore the causal implications of manipulating one or more variables.

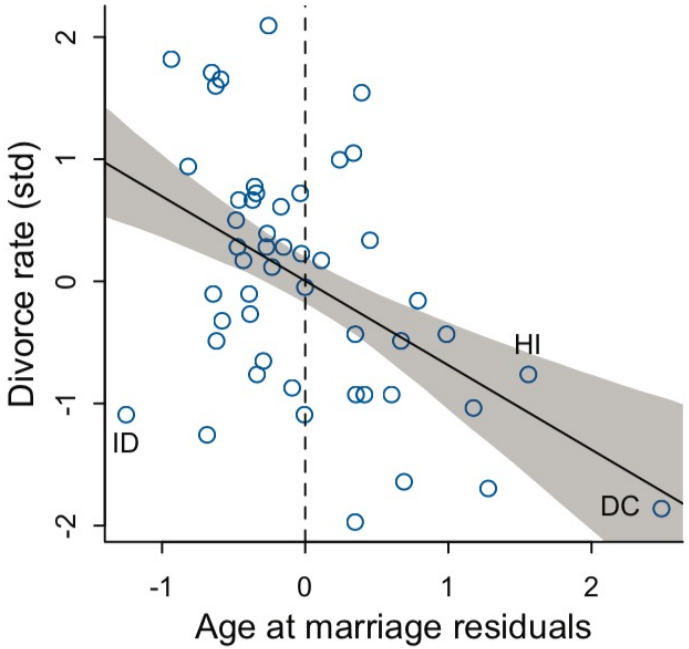
- Understanding multiple regression through residuals.
- The top row shows each predictor regressed on the other predictor.

  The lengths of the line segments connecting the model's expected value of the outcome, the regression line, and the actual value are the residuals.
- In the bottom row, divorce rate is regressed on the residuals from the top row.
  - Bottom left: Residual variation in marriage rate shows little association with divorce rate.
  - Bottom right: Divorce rate on age at marriage residuals, showing remaining variation, and this variation is associated with divorce rate.









- positive residuals have high marriage rates for their median age of marriage
- to use these residuals, put them on a horizontal axis and plot them against the y-var, divorce rate.
- this plot as displayes the linear relationship between divorce and marriage rates, having conditioned on median age of marriage.
- The vertical dashed line indicates marriage rate that matches the expectation from median age at marriage. States to the right of the line have higher marriage rates than expected.
  - marriage rate: average divorce rate on both sides of the line is about the same, and so the regression line demonstrates little relationship between divorce and marriage rates.
  - age at marriage: average divorce rate on the right is lower than on the left. States in which people marry older than expected for a given rate of marriage tend to have less divorce.
- There's conceptual value in seeing the model-based predictions displayed against the outcome, after subtracting out the influence of other predictors.
  - this procedure also brings home the message that regression models measure the remaining association of each predictor with the outcome, after already knowing the other predictors. regressions can behave in surprising ways as a result.

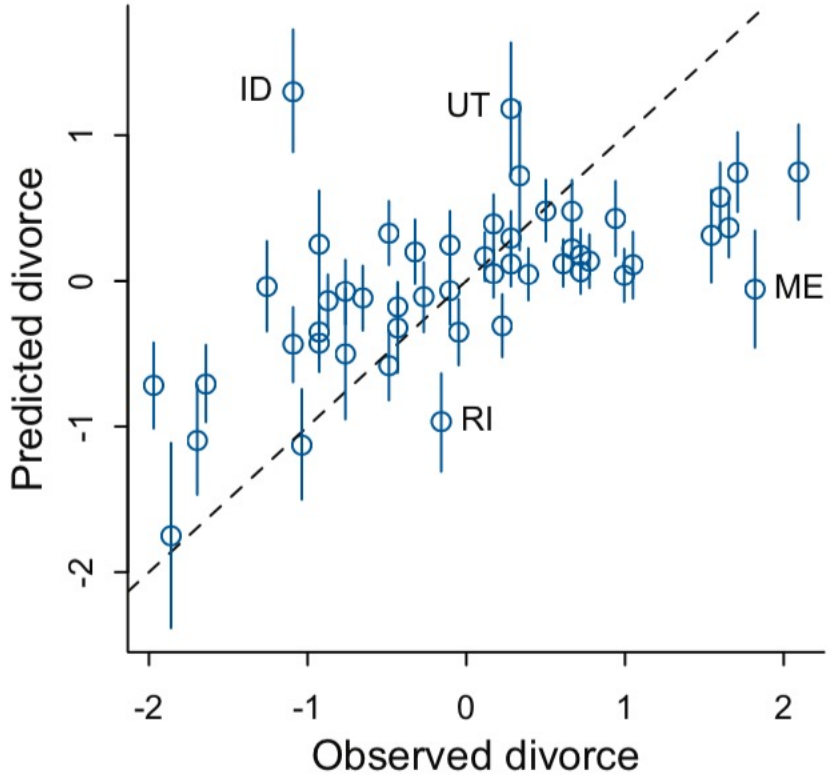


FIGURE 5.5. Posterior predictive plot for the multivariate divorce model, `m5.3`. The horizontal axis is the observed divorce rate in each State. The vertical axis is the model's posterior predicted divorce rate, given each State's median age at marriage and marriage rate. The blue line segments are 89% compatibility intervals. The diagonal line shows where posterior predictions exactly match the sample.

- is the model correct? Errors in fitting the model can be diagnosed by comparing implied predictions to the raw data.
- How does the model fail?
  - Sometimes, a model fits correctly but is still so poor for our purposes that it must be discarded.
  - More often, a model predicts well in some respects, but not in others. By inspecting the individual cases where the model makes poor predictions, you might get an idea of how to improve it. this process is essentially creative and relies upon the analyst's domain expertise. It also risks chasing noise.

- One way that spurious associations between a predictor and outcome can arise is when a truly causal predictor, call it x_{real}, influences both y and a spurious predictor, x_{spur}. here's a very basic simulation:

```
N <- 100                          # number of cases
x_real <- rnorm( N )              # x_real as Gaussian with mean 0 and stddev 1
x_spur <- rnorm( N , x_real )     # x_spur as Gaussian with mean=x_real
y <- rnorm( N , x_real )          # y as Gaussian with mean=x_real
d <- data.frame(y,x_real,x_spur) # bind all together in data frame
```

- Because x_real influences both y and x_spur, you can think of x_spur as another outcome of x_real, but one which we mistake as a potential predictor of y. As a result, both x_{real} and x_{spur} are correlated with y.
  - You can see this in the scatterplots from pairs(d).
- But when you include both x variables in a linear regression predicting y, the association between y and x_{spur} will be close to zero.

Counterfactual plots display the causal implications of the model.

- “What would Utah’s divorce rate be, if it’s median age at marriage were higher?”

1. counterfactual plots help you understand the model,
2. generate predictions for imaginary interventions
3. compute how much some observed outcome could be attributed to some cause.

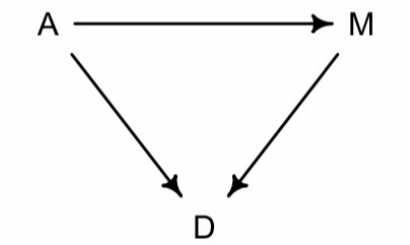
- **“counterfactual” indicates some computation that makes use of the structural causal model, going beyond the posterior distribution.** it could refer to both the past and the future.
- The simplest use is to see how the outcome would change as you change one predictor at a time. Changing just one X might also change other predictors, depending upon the causal model. Suppose for example that you pay young couples to postpone marriage until they are 35. Surely this will also decrease the number of couples who ever get married.

So let's see how to generate plots of model predictions that take the causal structure into account. The basic recipe is:

(1) Pick a variable to manipulate, the intervention variable.
(2) Define the range of values to set the intervention variable to.
(3) For each value of the intervention variable, and for each sample in posterior, use the causal model to simulate the values of other variables, including the outcome.

In the end, you end up with a posterior distribution of counterfactual outcomes that you can plot and summarize in various ways, depending upon your goal.

Let's see how to do this for the divorce model. Again we take this DAG as given:



To simulate from this, we need more than the DAG. We also need a set of functions that tell us how each variable is generated. For simplicity, we'll use Gaussian distributions for each variable, just like in model `m5.3`. But model `m5.3` ignored the assumption that *A* influences M. We didn't need that to estimate A → D. But we do need it to predict the consequences of manipulating A, because some of the effect of A acts through M.
To estimate the influence of A on M, all we need is to regress A on M. There are no other variables in the DAG creating an association between A and M.

```
d_model <- bf(d ~ 1 + a + m)
m_model <- bf(m ~ 1 + a)
```

```
brm(data = d,
    family = gaussian,
    d_model + m_model + set_rescor(FALSE),
    prior = c(prior(normal(0, 0.2), class = Intercept, resp = d),
              prior(normal(0, 0.5), class = b, resp = d),
              prior(exponential(1), class = sigma, resp = d),

              prior(normal(0, 0.2), class = Intercept, resp = m),
              prior(normal(0, 0.5), class = b, resp = m),
              prior(exponential(1), class = sigma, resp = m)),
```

we specify set_rescor(FALSE) to prevent **brms** from adding a residual correlation between d and m. each prior statement includes a resp argument. This clarifies which sub-model the prior refers to.

Counterfactual plots can be produced for any values of the predictor variables, even unobserved ones.

This predicted trend in D includes both paths: A → D and A → M → D.

M → D is very small, so the second path doesn't contribute much to the trend. But if M were to strongly influence D, the code would include the effect.

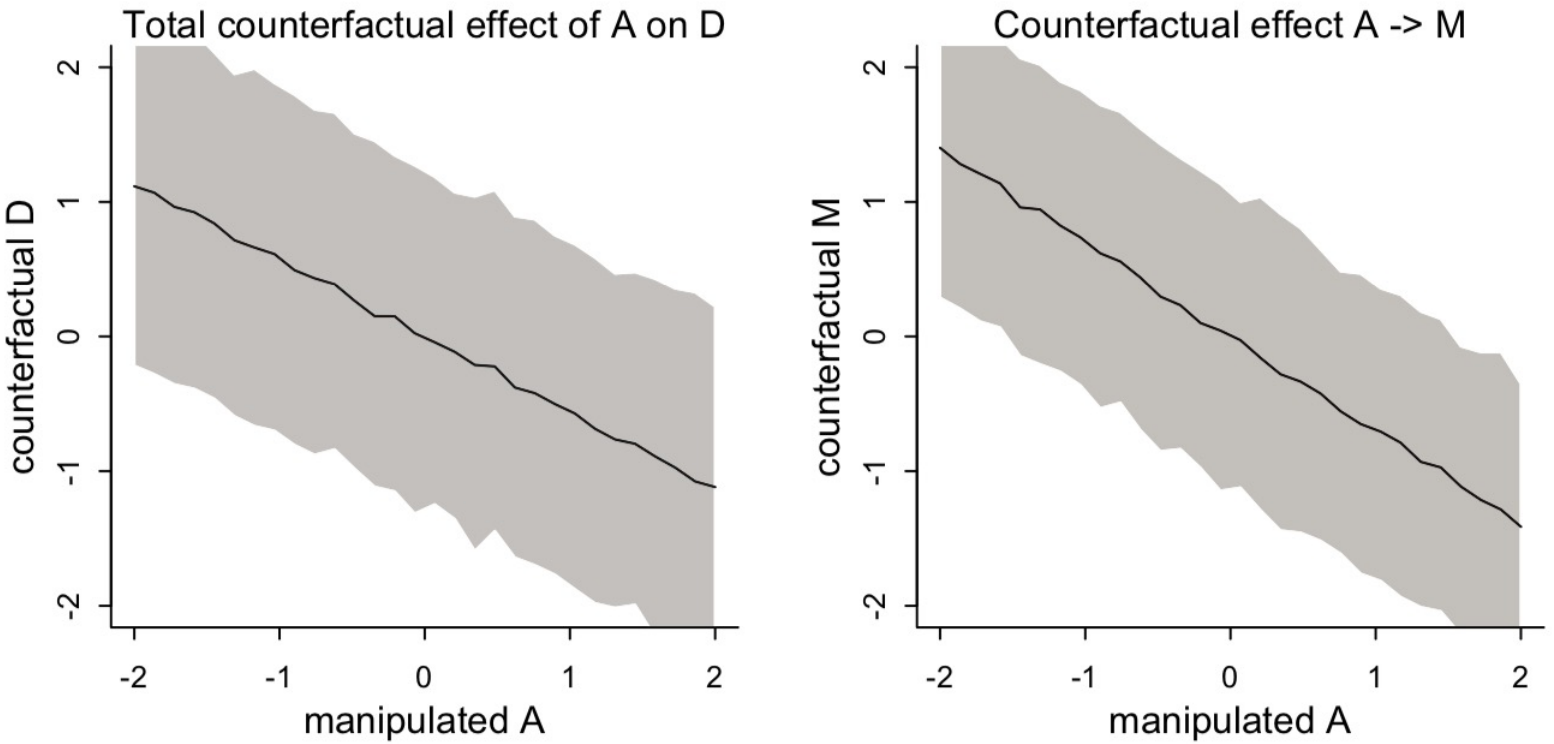
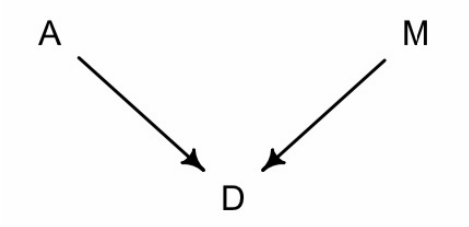


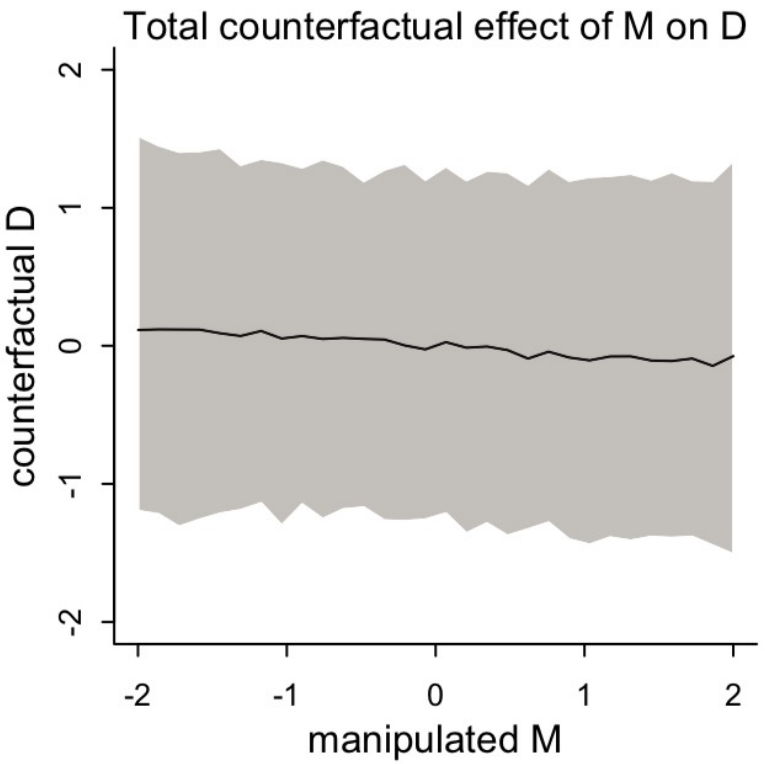
FIGURE 5.6. Counterfactual plots for the multivariate divorce model, `m5.3`. These plots visualize the predicted effect of manipulating age at marriage A on divorce rate D. Left: Total causal effect of manipulating A (horizontal) on D. This plot contains both paths, $A \rightarrow D$ and $A \rightarrow M \rightarrow D$. Right: Simulated values of M show the estimated influence $A \rightarrow M$.

The trick with simulating counterfactuals is to realize that when we manipulate some variable X, we break the causal influence of other variables on X. This is the same as saying we modify the DAG so that no arrows enter X. Suppose for example that we now simulate the effect of manipulating M. This implies the DAG:



The arrow $A \rightarrow M$ is deleted, because if we control the values of M, then A no longer influences it. It's like a perfectly controlled experiment. Now we can modify the code above to simulate the counterfactual result of manipulating M. We'll simulate a counterfactual for an average state, with $A = 0$, and see what changing M does.

We only simulate D now—note the vars argument to sim() in the code above. We don't simulate A, because M doesn't influence it. I show this plot in Figure 5.7. This trend is less strong, because there is no evidence for a strong influence of M on D.



Masked relationship

- multiple predictor variables are useful for knocking out spurious association.
- A second reason to use more than one predictor is to measure the direct influences of multiple factors on an outcome, when none of those influences is apparent from bivariate relationships.
- This kind of problem tends to arise when there are two predictor variables that are correlated with one another. However, one of these is positively correlated with the outcome and the other is negatively correlated with it.
- You'll consider this kind of problem in a new data context, information about the composition of milk across primate species, as well as some facts about those species, like body mass and brainsize.

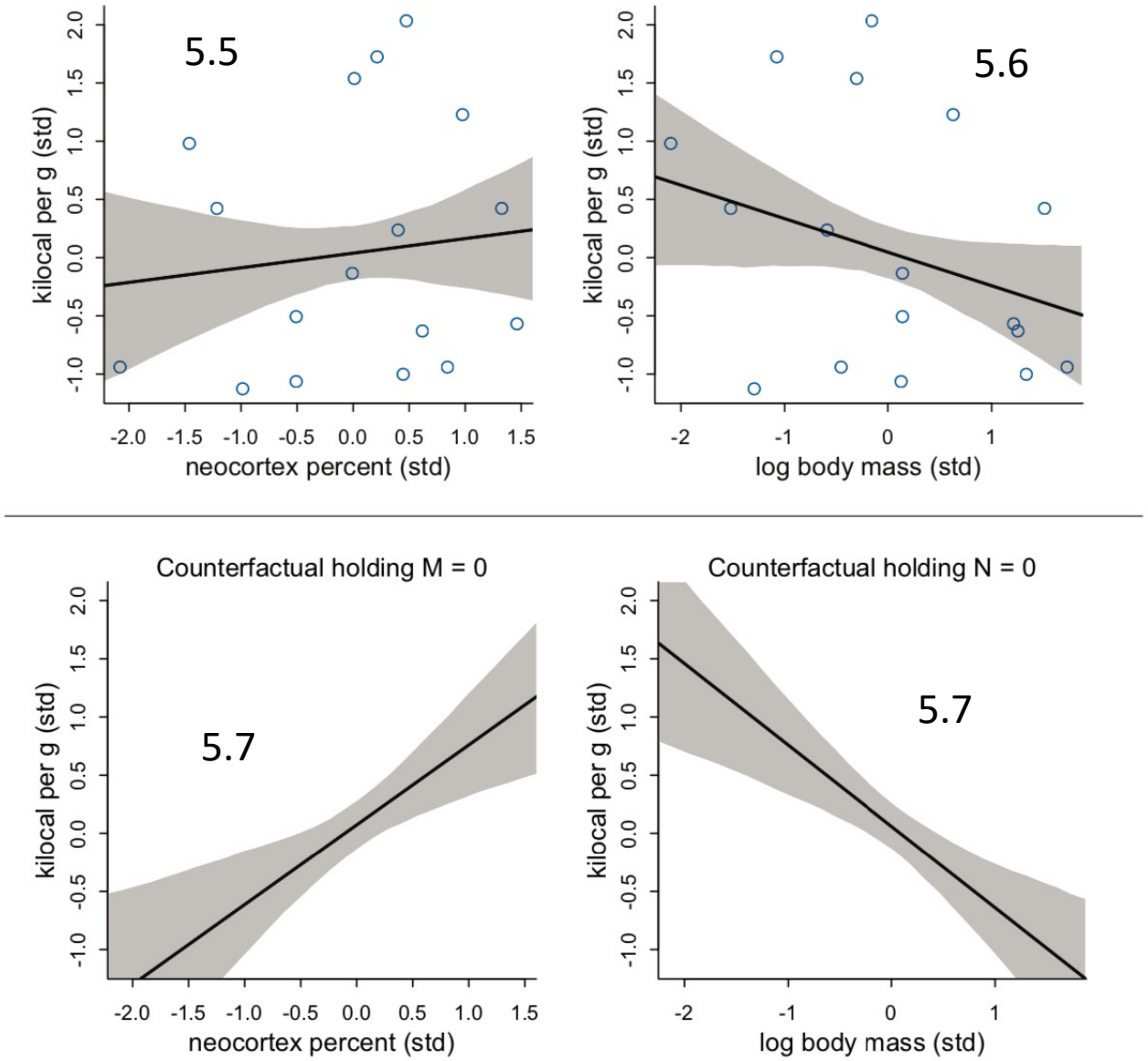
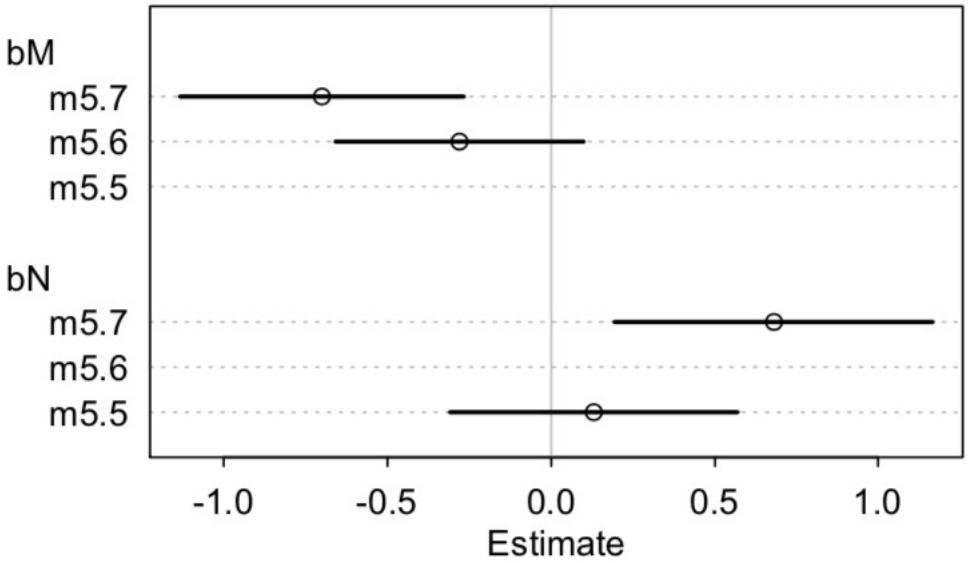
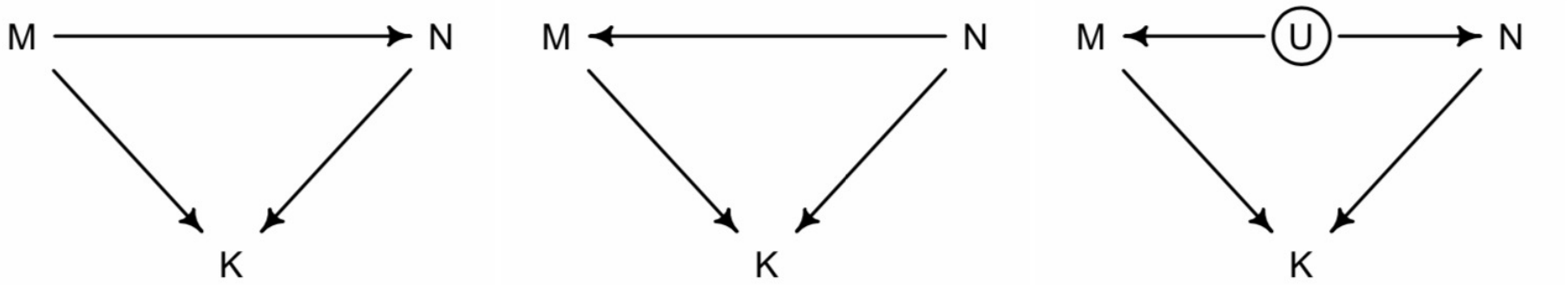


FIGURE 5.9. Milk energy and neocortex among primates. In the top two plots, simple bivariate regressions of kilocalories per gram of milk (K) on (left) neocortex percent (N) and (right) log female body mass (M) show weak associations. In the bottom row, a model with both neocortex percent (N) and log body mass (M) shows stronger associations.

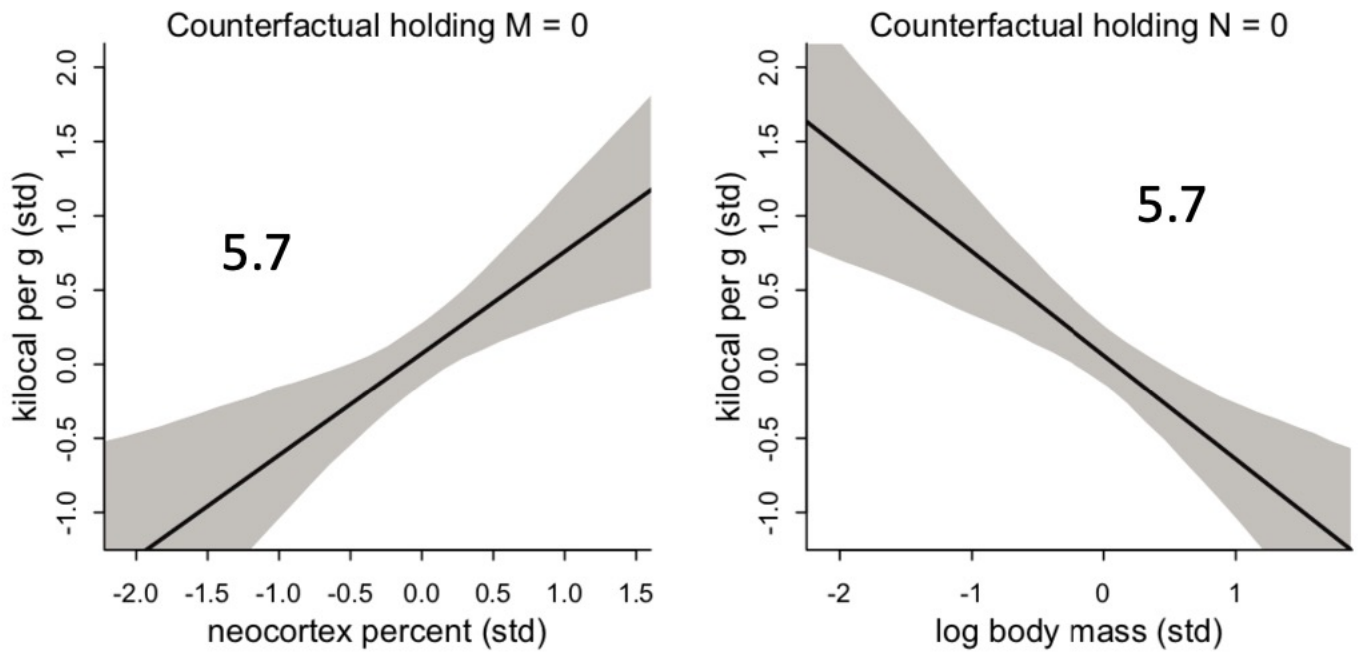


- Why did adding neocortex and body mass to the same model lead to stronger associations for both?
  - there are two variables correlated with the outcome, but one is positively correlated with it and the other is negatively correlated with it.
  - In addition, both of the explanatory variables are positively correlated with one another.
- What the model does is ask if species that have high neocortex percent for their body mass have higher milk energy.
- Likewise, the model asks if species with high body mass for their neocortex percent have higher milk energy.
- Bigger species, like apes, have milk with less energy. But species with more neocortex tend to have richer milk. The fact that body size and neocortex, are correlated across species makes it hard to see these relationships, unless we account for both.

Some DAGs will help. There are at least three graphs consistent with these data.



- Which of these graphs is right? We can't tell from the data alone, because these graphs imply the same set of conditional independencies.
- In this case, there are no conditional independencies—each DAG above implies that all pairs of variables are associated, regardless of what we condition on.
- A set of DAGs with the same conditional independencies is known as a Markov equivalence set.
- Suppose the third DAG above is the right one. Then imagine manipulating M and N, breaking the influence of U on each.
  - In the real world, such experiments are impossible. If we change an animal's body size, natural selection would then change the other features to match it. But these counterfactual plots do help us see how the model views the association between each predictor and the outcome.


Counterfactual holding M = 0
5.7
kilocal per g (std)
neocortex percent (std)
-2.0 -1.5 -1.0 -0.5 0.0 0.5 1.0 1.5
-1.0 -0.5 0.0 0.5 1.0 1.5 2.0
Counterfactual holding N = 0
5.7
kilocal per g (std)
log body mass (std)
-2 -1 0 1
-1.0 -0.5 0.0 0.5 1.0 1.5 2.0

over- and underfitting

- overfitting leads to poor prediction by learning too much from the data
- underfitting leads to poor prediction by learning too little from the data
- confounded models can produce better predictions than models that correctly measure a causal relationship. The consequence is that, when we design any particular statistical model, we must decide whether we want to understand causes or rather just predict.

To navigate among these monsters there are two approaches (which can be used in combination).

1. to use a regularizing prior to tell the model not to get too excited by the data. This is the same device that non-Bayesian methods refer to as “penalized likelihood.”
2. to use some scoring device, like information criteria or cross-validation, to estimate predictive accuracy.

p-values are not designed to help navigate between underfitting and overfitting.

- predictor variables that improve prediction are not always statistically significant.
- It is also possible for variables that are statistically significant to do nothing useful for prediction.
- Since the conventional 5% threshold is purely conventional, we shouldn’t expect it to optimize anything.

There are two related problems with adding variables

1. adding parameters—making the model more complex—nearly always improves the fit of a model to the data.

- "fit"is a measure of how well the model can retrodict the data used to fit the model. In the context of linear Gaussian models, R^2 is the most common measure of this kind. Often described as "variance explained," R^2 is defined as:

$$R^2 = \frac{\text{var}(\text{outcome}) - \text{var}(\text{residuals})}{\text{var}(\text{outcome})} = 1 - \frac{\text{var}(\text{residuals})}{\text{var}(\text{outcome})}$$

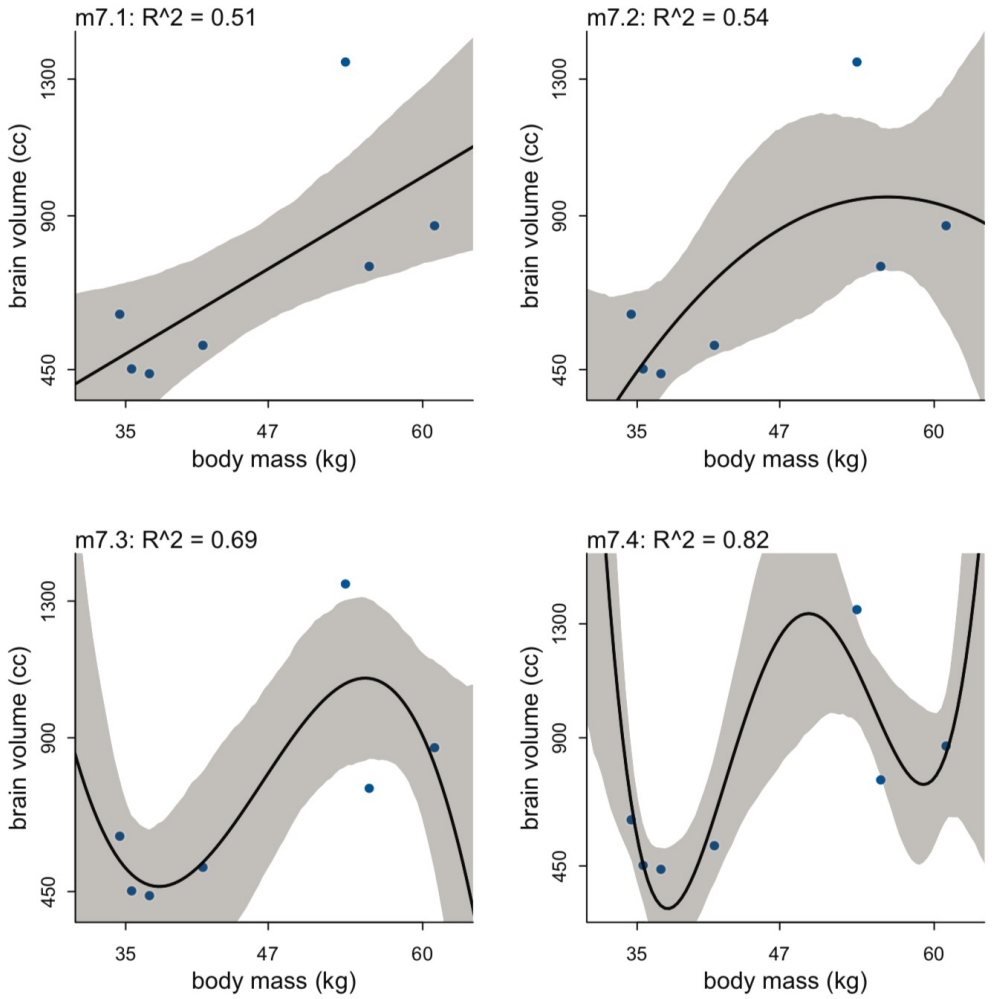
- Like other measures of fit to sample, R^2 increases as more predictors are added. This is true even when the variables you add are just random numbers. So it's no good to choose among models using only fit to the data.

2. while more complex models fit the data better, they often predict new data worse.

- Models that have many parameters tend to overfit. a complex model will be sensitive to the sample used to fit it, leading to potentially large mistakes when future data is not exactly like the past data.
- But models with few parameters tend to underfit, systematically over-predicting or under-predicting the data, regardless of how well future data resemble past data.
- So we can't always favor either simple models or complex models.

Overfitting occurs when a model learns too much from the sample.

- there are both regular and irregular features in every sample. The regular features are the targets of our learning, because they generalize well or answer a question of interest. The irregular features are aspects of the data that do not generalize and so may mislead us.
- Overfitting happens automatically. In ordinary models adding parameters will always improve the fit of a model to the sample.
- For multilevel models adding parameters does not necessarily improve fit to the sample, but may improve predictive accuracy.



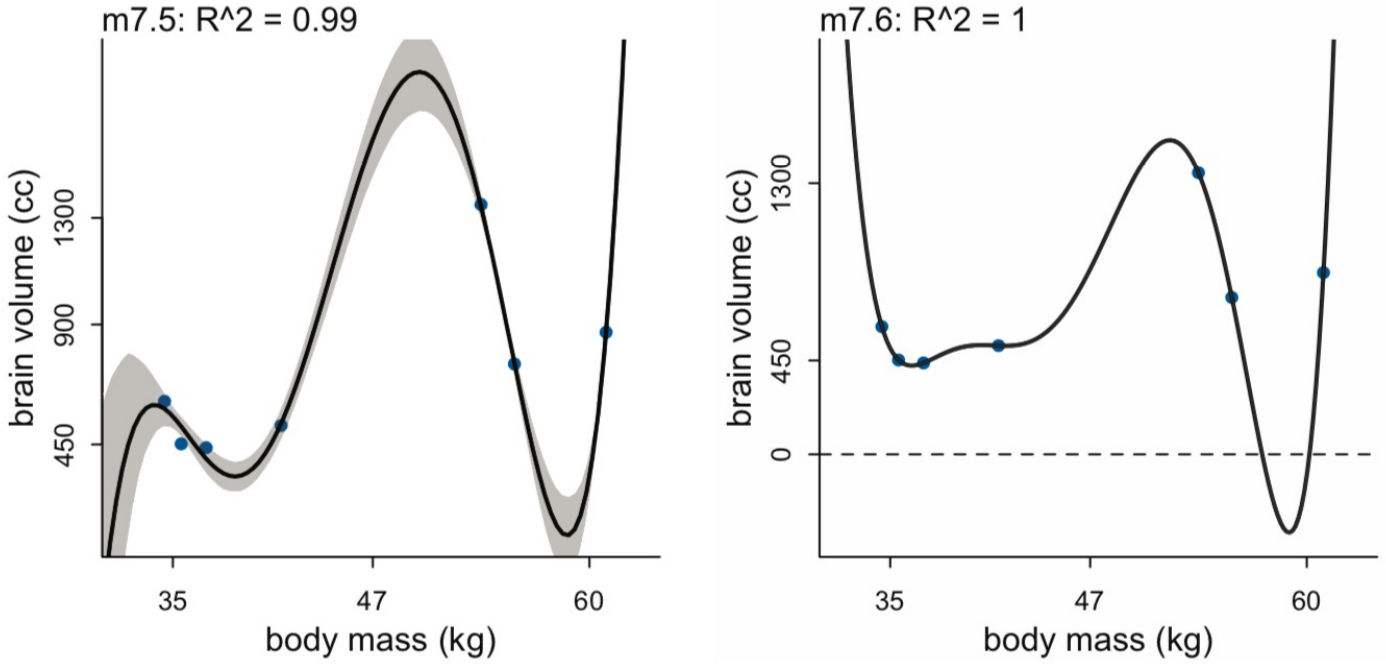


FIGURE 7.3. Polynomial linear models of increasing degree for the hominin data. Each plot shows the posterior mean in black, with 89% interval of the mean shaded. R^2 is displayed above each plot. In order from top-left: First-degree polynomial, second-degree, third-degree, fourth-degree, fifth-degree, and sixth-degree.

model fitting can be considered a form of data compression. Parameters summarize relationships among the data, compressing the data into a simpler form, with loss of information about the sample. The parameters can then be used to generate new data, effectively decompressing the data.

When a model has a parameter to correspond to each datum then there is no compression. The model just encodes the raw data in a different form, using parameters instead. we learn nothing about the data. This view of model selection is Minimum Description Length (MDL).

- underfitting produces models that are inaccurate both within and out of sample. They learn too little, failing to recover regular features of the sample.
- Another way to conceptualize an underfit model is to notice that it is insensitive to the sample. We could remove any one point from the sample and get almost the same regression line. In contrast. m7.6 is verv sensitive to the sample.

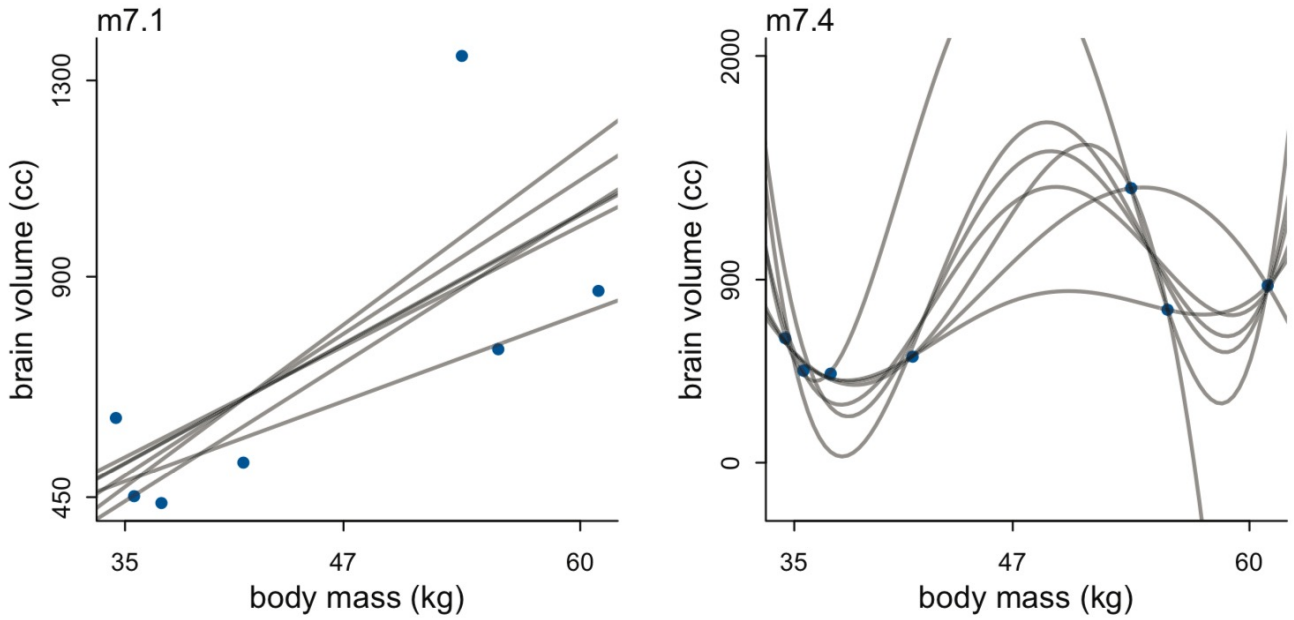


FIGURE 7.4. Underfitting and overfitting as under-sensitivity and over-sensitivity to sample. In both plots, a regression is fit to the seven sets of data made by dropping one row from the original data. Left: An underfit model is insensitive to the sample, changing little as individual points are dropped. Right: An overfit model is sensitive to the sample, changing dramatically as points are dropped.