Introduction to Simulation Studies

Dr. Andy Bell & **Dr. Calum Webb**
Sheffield Methods Institute, School of Education, the University of Sheffield
andrew.j.d.bell@sheffield.ac.uk
c.j.webb@sheffield.ac.uk

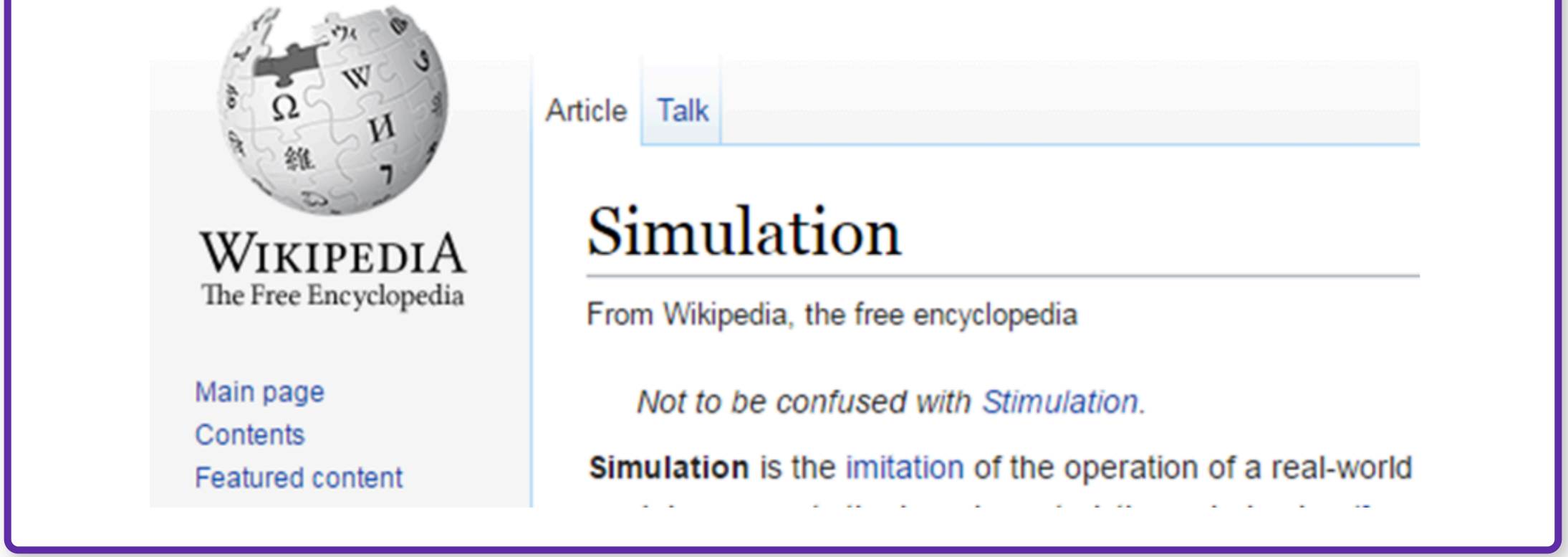
Plan for the day:

- Lecture: What are simulations good for?
- Lecture: The logic of a simulation study.
- Practical exercise: A simple simulation of omitted variable bias.

Lunch

- Lecture: Simulation quantities of interest
- Lecture: Presenting simulation results
- Practical exercise: Conducting many simulations and creating a dataset of the quantities of interest

Lecture 1: Introduction to Simulation Studies


WIKIPEDIA
The Free Encyclopedia
Main page
Contents
Featured content
Article
Talk
Simulation
From Wikipedia, the free encyclopedia
Not to be confused with Stimulation.
Simulation is the imitation of the operation of a real-world

What are simulation studies?

- Lots of different meanings: including as a methodology in itself (e.g. spatial micro-simulation)
- Here, we are interested in using simulated data to test whether a model / a methodology works as we expect it to
- Difficult with real data, as we don't know how it came about

What are simulation studies?

- Lots of different meanings: including as a methodology in itself (e.g. spatial micro-simulation)
- Here, we are interested in using simulated data to test whether a model / a methodology works as we expect it to
- Difficult with real data, as we don't know how it came about

So:

1. Take a dataset that we know the truth about
2. Run a model using that dataset
3. See whether the results that are produced are accurate
4. Repeat this multiple times to check results weren't by chance.
5. See under what conditions the model is accurate

Why would we want to spend an entire day learning about simulation?

- A geeky interest in the models themselves
- To test whether a model that you want to use for a substantive topic works
- To really understand what a model is really doing
- As a way of (re)learning basic statistics that we should know already, but don't.

Why would we want to spend an entire day learning about simulation?

- A geeky interest in the models themselves
- To test whether a model that you want to use for a substantive topic works
- To really understand what a model is really doing
- As a way of (re)learning basic statistics that we should know already, but don't.

But...

- Simulations seem a bit artificial - how do we know they have anything to do with the real world.
- This was my attitude when a reviewer first asked me to conduct a simulation. (And it's not just me...)

Clarifying hierarchical age–period–cohort models: A rejoinder to Bell and Jones

Eric N. Reither [a,*], Kenneth C. Land [b], Sun Y. Jeon [a], Daniel A. Powers [c], Ryan K. Masters [d], Hui Zheng [e], Melissa A. Hardy [f], Katherine M. Keyes [g], Qiang Fu [h], Heidi A. Hanson [i], Ken R. Smith [j], Rebecca L. Utz [k], Y. Claire Yang [l]

[a] Department of Sociology and the Yun Kim Population Research Laboratory, Utah State University, 0730 Old Main Hill, Logan UT 84322-0730, USA
[b] Department of Sociology and Center for Population Health and Aging, Duke University, USA
[c] Department of Sociology, Population Research Center, The University of Texas at Austin, USA
[d] Department of Sociology and Institute of Behavioral Science, University of Colorado at Boulder, USA
[e] Department of Sociology, The Ohio State University, USA
[f] Department of Sociology and Criminology, Population Research Institute, The Pennsylvania State University, USA
[g] Department of Epidemiology, Mailman School of Public Health, Columbia University, USA
[h] Department of Sociology, The University of British Columbia, Canada
[i] Department of Family and Preventive Medicine and the Huntsman Cancer Institute, The University of Utah, USA
[j] Department of Family and Consumer Studies and the Huntsman Cancer Institute, The University of Utah, USA
[k] Department of Sociology, The University of Utah, USA
[l] Department of Sociology and the Lineberger Comprehensive Cancer Center, The University of North Carolina at Chapel Hill, USA

ARTICLE INFO

Article history:
Received 8 July 2015
Accepted 12 July 2015
Available online 31 July 2015

Keywords:
Age–period–cohort models
Cohort effects
Simulation models
Hierarchical modeling
Random effects
Body mass index
Obesity epidemic
Social change

ABSTRACT

Previously, Reither et al. (2015) demonstrated that hierarchical age–period–cohort (HAPC) models perform well when basic assumptions are satisfied. To contest this finding, Bell and Jones (2015) invent a data generating process (DGP) that borrows age, period and cohort effects from different equations in Reither et al. (2015). When HAPC models applied to data simulated from this DGP fail to recover the patterning of APC effects, B&J reiterate their view that these models provide "misleading evidence dressed up as science." Despite such strong words, B&J show no curiosity about their own simulated data—and therefore once again misapply HAPC models to data that violate important assumptions. In this response, we illustrate how a careful analyst could have used simple descriptive plots and model selection statistics to verify that (a) period effects are not present in these data, and (b) age and cohort effects are conflated. By accounting for the characteristics of B&J's artificial data structure, we successfully recover the "true" DGP through an appropriately specified model. We conclude that B&Js main contribution to science is to remind analysts that APC models will fail in the presence of exact algebraic effects (i.e., effects with no random/stochastic components), and when collinear temporal dimensions are included without taking special care in the modeling process. The expanded list of coauthors on this commentary represents an emerging consensus among APC scholars that B&J's essential strategy—testing HAPC models with data simulated from contrived DGPs that violate important assumptions—is not a productive way to advance the discussion about innovative APC methods in epidemiology and the social sciences.

To what extent do simulations reflect reality? To what extent do they need to reflect reality?

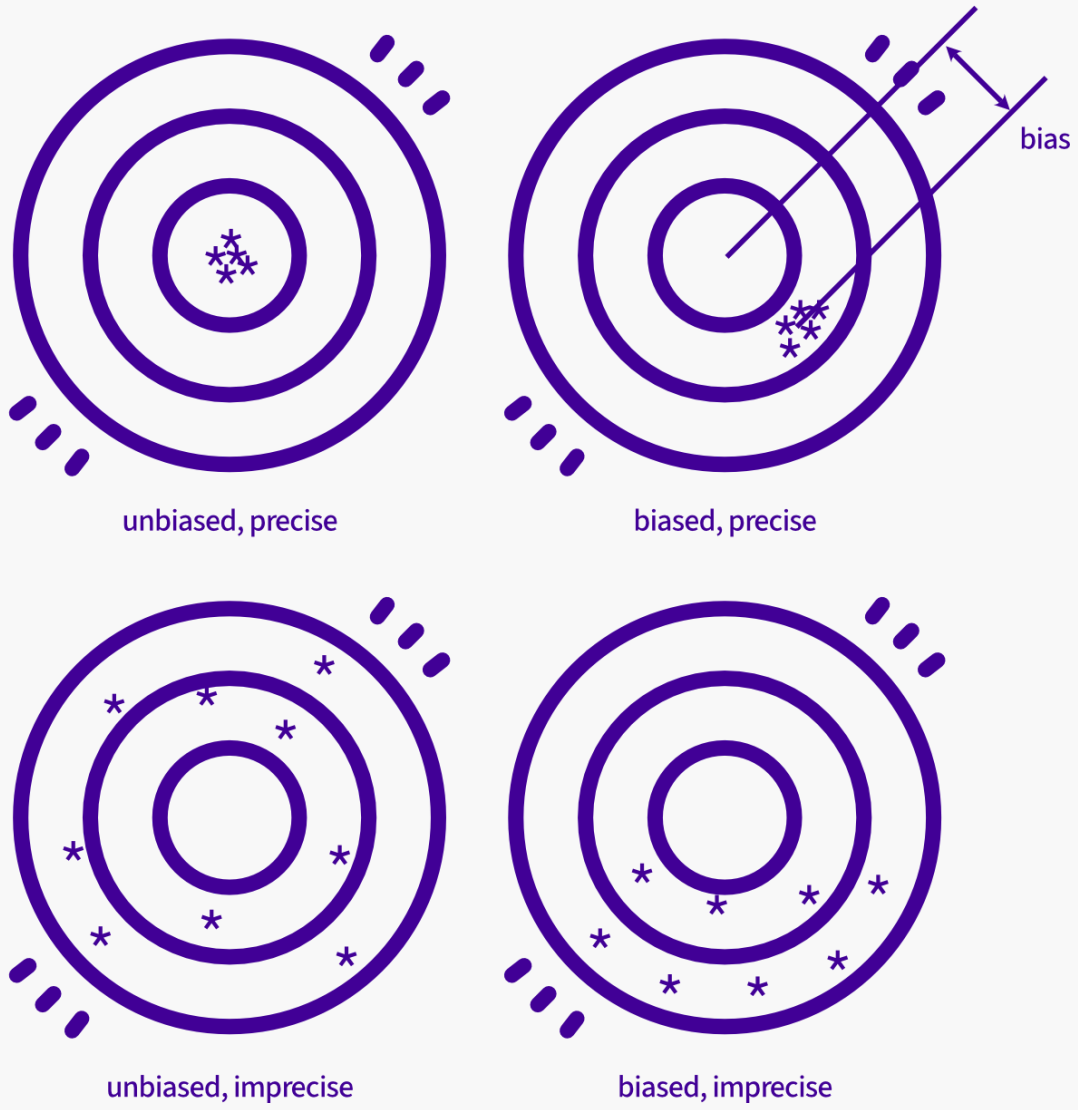
To what extent do they need to reflect reality?

My view: If a mathematical model can draw out interesting things about society (I think it can) then a mathematically simulated dataset can tell us interesting things about the model's relation to real data.

That doesn't mean models, or simulation, can tell us everything about the world: simulations as an important initial test

What can be wrong with models?

- They give inaccurate answers on average (bias)
- They give correct answers on average, but with not as much precision as another model
- They give correct answers but the precision is inaccurately estimated
- They give the correct answer, but to a different question to what you thought
- A combination of the above


bias
unbiased, precise
biased, precise
unbiased, imprecise
biased, imprecise

What causes models to be wrong?

- Sample size too small
- Omitted variables
- Poor performing estimator
- Reverse causality
- Model specified wrong in some other way
- Missing data

What do we mean by wrong?

If a model is biased… what is it biased from? And is that thing what we were really expecting to find?

> Example: failing to control for ‘cultural differences’ in finding effect of race

Results should always be situated in some kind of reality!

FE versus RE & Age, Period Cohort Analysis

Fixed Effects and Random Effects

- Conventional wisdom: FE 'gold standard', RE affected by omitted variables
- But FE is limited (e.g no higher level variables can be estimated)
- Bell and Jones (2015) – shows RE with group means produces same results
- Also showed problems with SEs for FEVD

Result: a change in advice of what researchers should do

Bring into question results of some models (eg using the FEVD)

FE versus RE & Age, Period Cohort Analysis

Fixed Effects and Random Effects

- Conventional wisdom: FE 'gold standard', RE affected by omitted variables
- But FE is limited (e.g no higher level variables can be estimated)
- Bell and Jones (2015) – shows RE with group means produces same results
- Also showed problems with SEs for FEVD

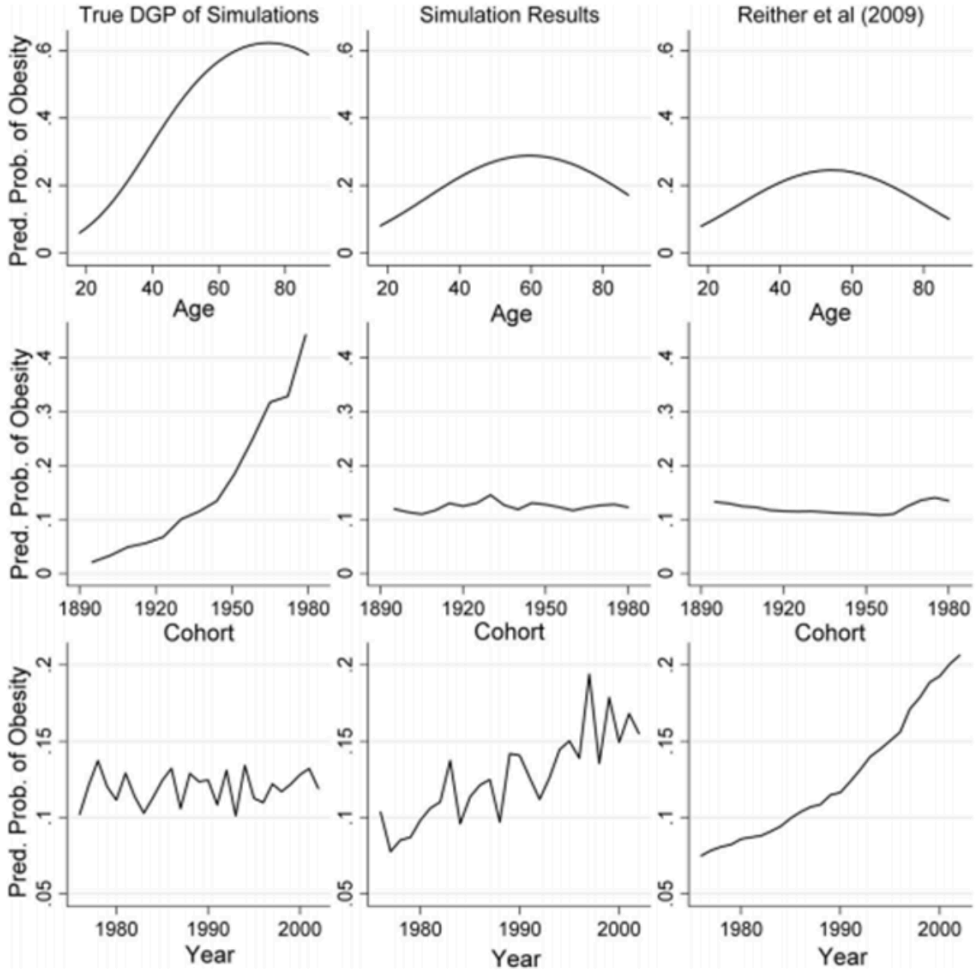
Result: a change in advice of what researchers should do

Bring into question results of some models (eg using the FEVD)

Age-Period-Cohort Models

- Age = Period – Cohort, so model cannot include all three (exactly colinearity)
- Various proposed 'solutions' to the identification problem, and justified with simulations
- Bell and Jones (various) show situations where these models do not work (by changing the DGP used in simulations

Result A change in academic practice using these models?

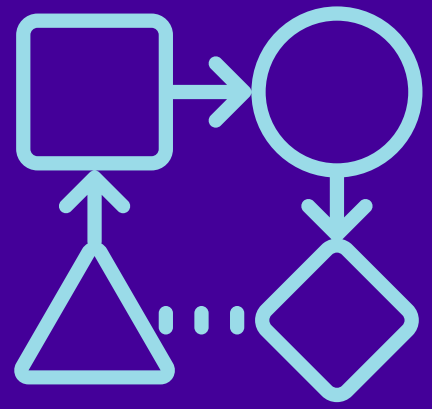

True DGP of Simulations
Simulation Results
Reither et al (2009)
Pred. Prob. of Obesity
0
.2
.4
.6
20
40
60
80
Age
Pred. Prob. of Obesity
0
.1
.2
.3
.4
1890
1920
1950
1980
Cohort
Pred. Prob. of Obesity
.05
.1
.15
.2
1980
1990
2000
Year

Summary

- These sorts of studies can have a big impact in challenging standard practice
- But are also an important way to understand what we are doing with our models

I didn't really understand SEs until I started simulating things!

Lecture 2: The logic of a simulation study



What really is a regression model anyway?

A regression model is a **description** of data that we have observed. However, we can use that description as a way to generate data.

This is an oversimplified representation of the world / the "truth". Literally: a model.

Linear regression model:

$$Y = B_0 + B_1 X_1 + e$$

What really is a regression model anyway?

A regression model is a **description** of data that we have observed. However, we can use that description as a way to generate data.

This is an oversimplified representation of the world / the "truth". Literally: a model.

Linear regression model:

$$Y = B_0 + B_1 X_1 + e$$

B_1: The average change in Y for a one-unit change in X. From a frequentist perspective, every repeat study of a given sample size will have some error in this value that is normally distributed around the true (population) value.

From a Bayesian perspective, multiple values of B_1 are plausible and these are usually normally distributed.

What really is a regression model anyway?

A regression model is a **description** of data that we have observed. However, we can use that description as a way to generate data.

This is an oversimplified representation of the world / the "truth". Literally: a model.

Linear regression model:

$$Y = B_0 + B_1 X_1 + e$$

X_1: Our observations of the variable X from our sample will be the product of its own data generating process that results in a distribution, and will have some measurement error (e.g. a sample of incomes from the wider population will reflect a positively skewed normal distribution/log-normal distribution).

Every set of values for X we get will look slightly different depending on who is in our sample and the error in our measurement.

What really is a regression model anyway?

A regression model is a **description** of data that we have observed. However, we can use that description as a way to generate data.

This is an oversimplified representation of the world / the "truth". Literally: a model.

Linear regression model:

$$Y = B_0 + B_1 X_1 + e$$

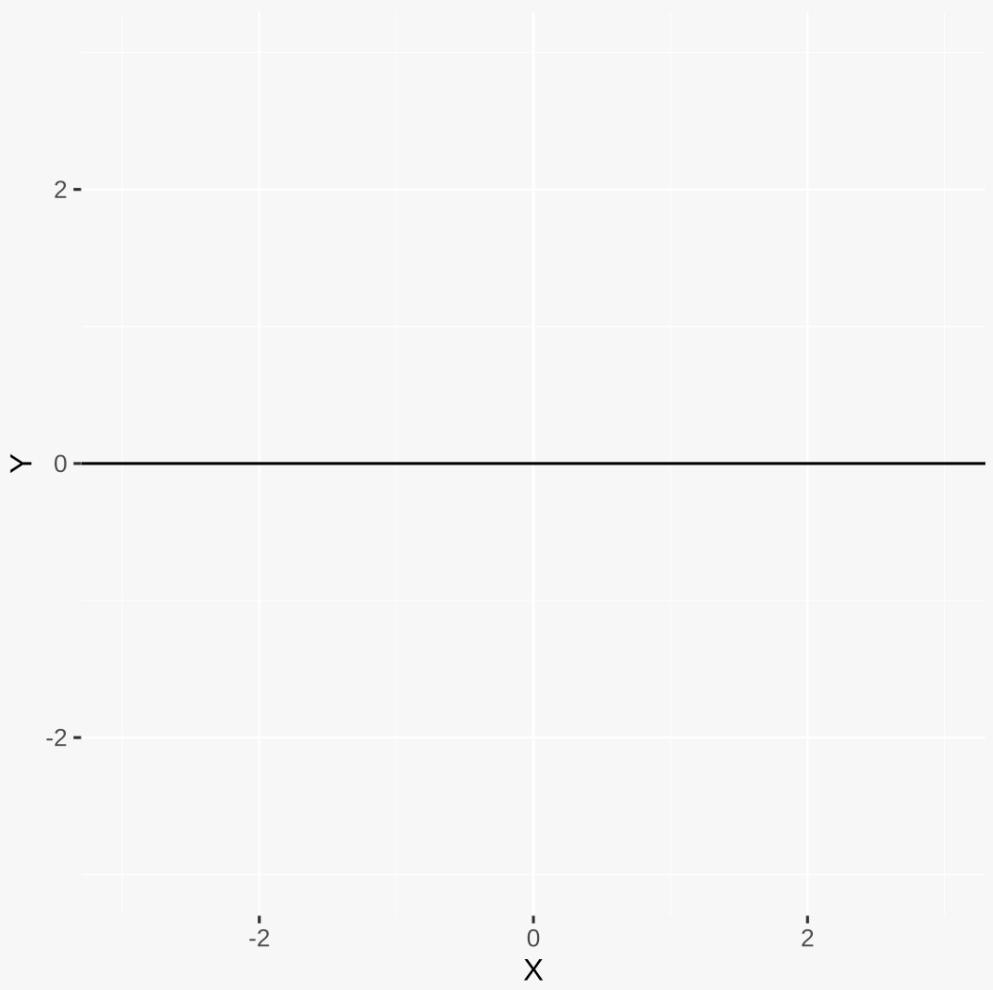
e: Sometimes called the error, e can also be considered the considered the **unexplained residual** - that is, the difference between the true value of Y and that expected by the model. It is everything that cannot be predicted by X_1. In linear regression, it is assumed that e is normally distributed.

Working backwards from a linear regression to simulated data...

With these pieces of information we can simulate what Y might look like.

$$Y = B_0 + B_1 X_1 + e$$

B_0 = 0, B_1 = ???, X_1 = ???, e = ???



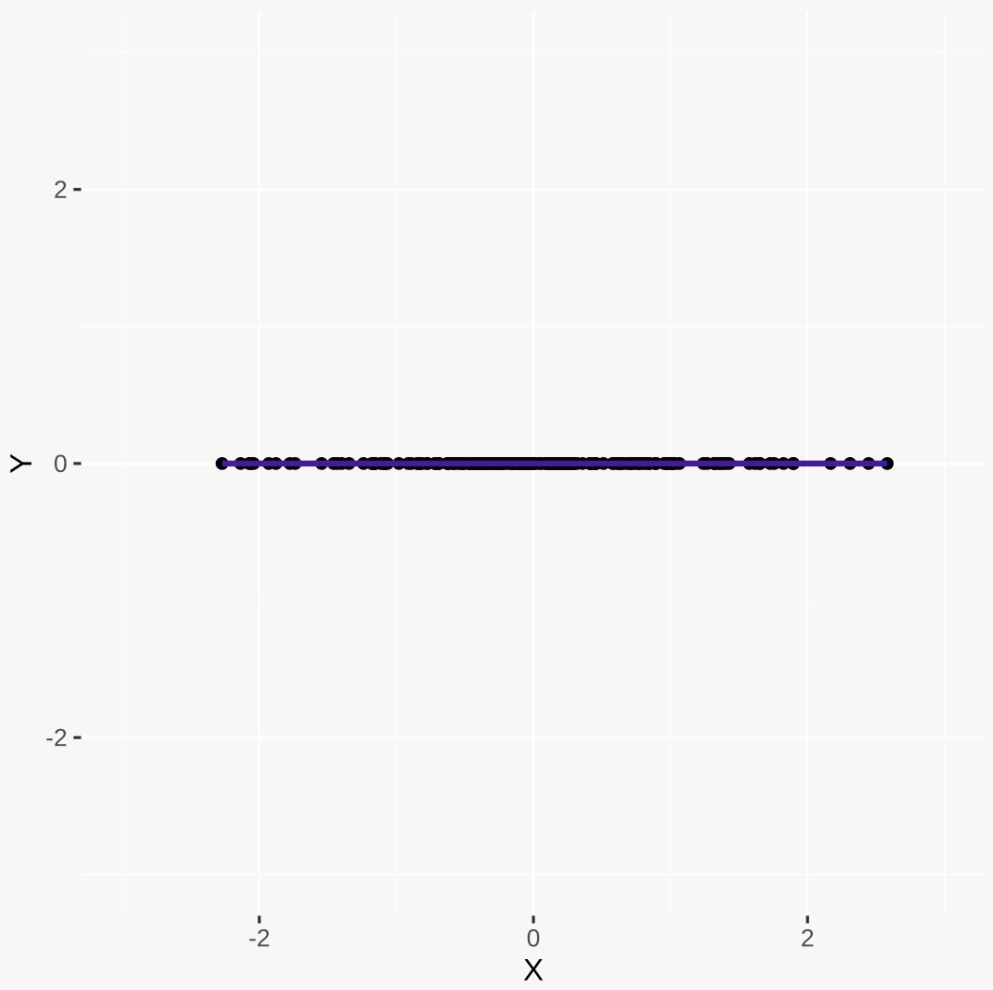
Working backwards from a linear regression to simulated data...

With these pieces of information we can simulate what Y might look like.

$$Y = B_0 + B_1 X_1 + e$$

B_0 = 0, B_1 = ???, X_1 = $N(0, 1)$, e = ???

* $N(0, 1)$ means "A normally distributed variable with a mean of 0 and a standard deviation of 1".



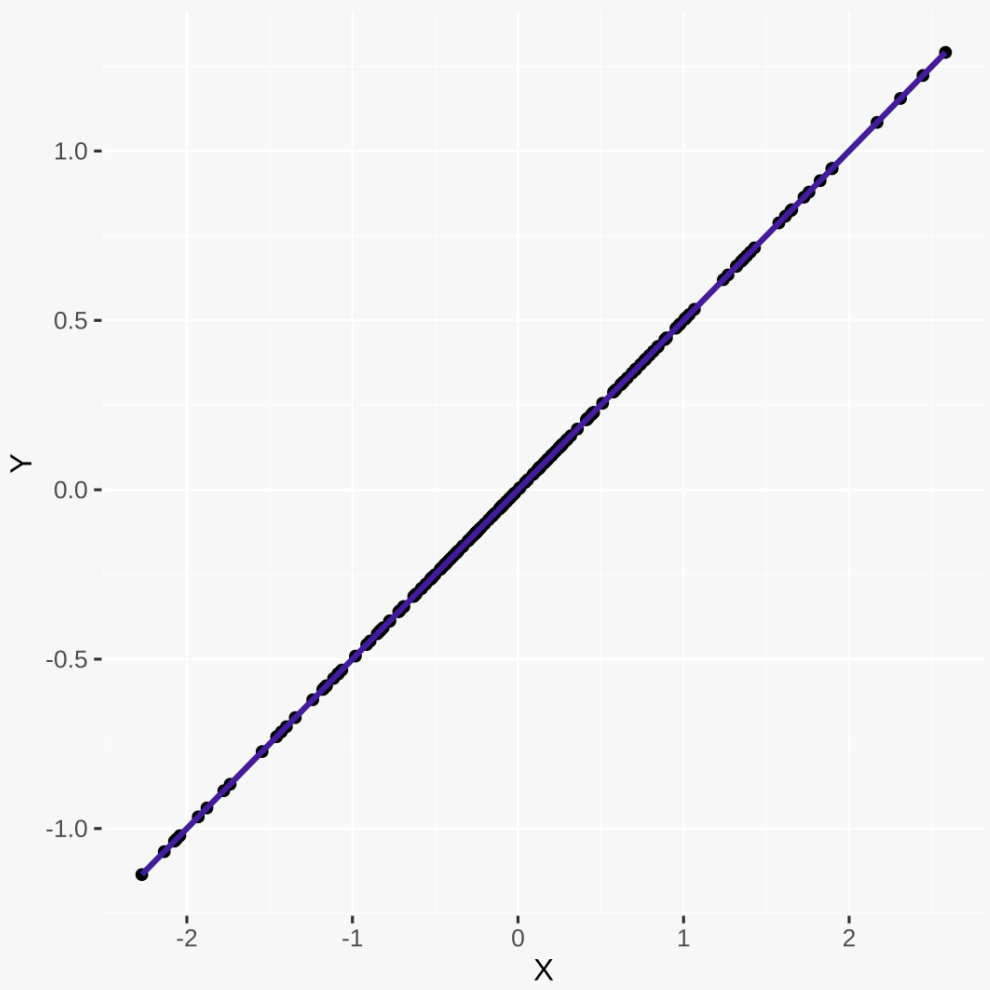
Working backwards from a linear regression to simulated data...

With these pieces of information we can simulate what Y might look like.

$$Y = B_0 + B_1 X_1 + e$$

B_0 = 0, B_1 = 0.5, X_1 = $N(0, 1)$, e = ???

* $N(0, 1)$ means "A normally distributed variable with a mean of 0 and a standard deviation of 1".



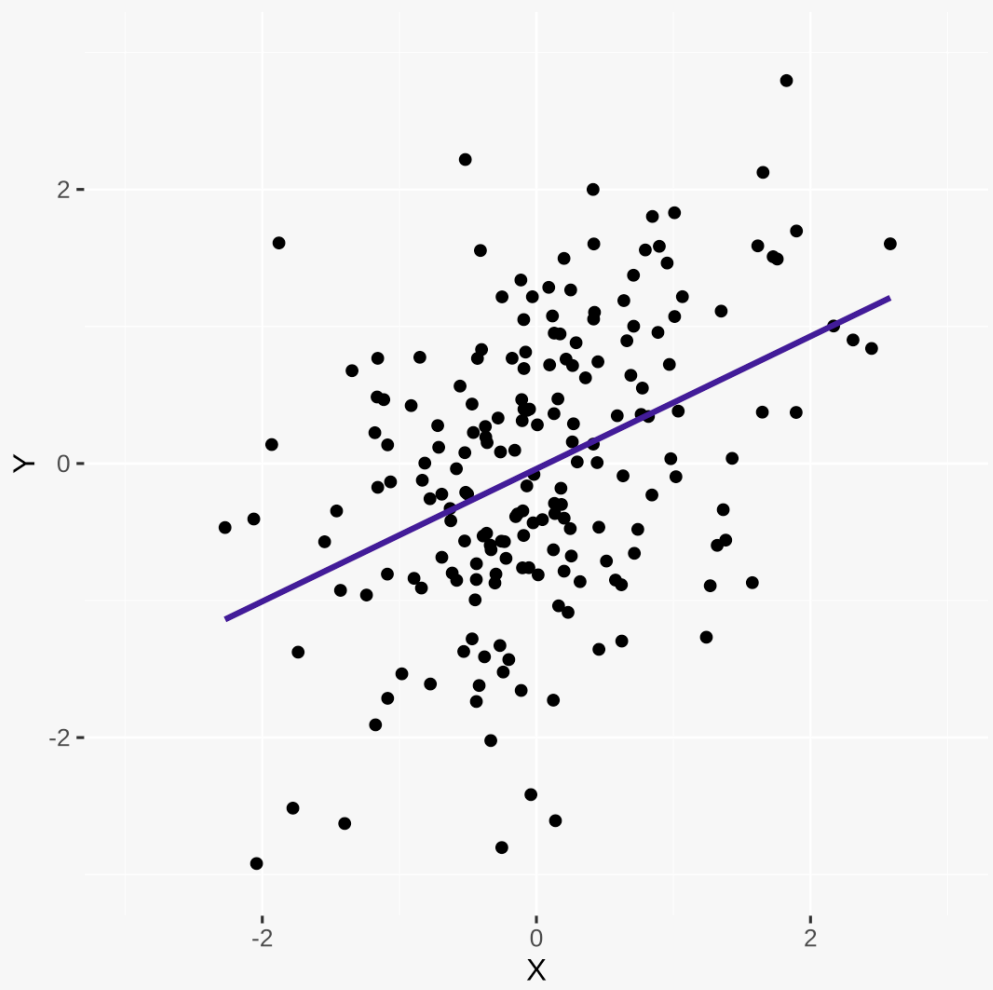
Working backwards from a linear regression to simulated data...

With these pieces of information we can simulate what Y might look like.

$$Y = B_0 + B_1 X_1 + e$$

B_0 = 0, B_1 = 0.5, X_1 = $N(0, 1)$, e = $N(0, 1)$

* $N(0, 1)$ means "A normally distributed variable with a mean of 0 and a standard deviation of 1".



Working backwards from a linear regression to simulated data...

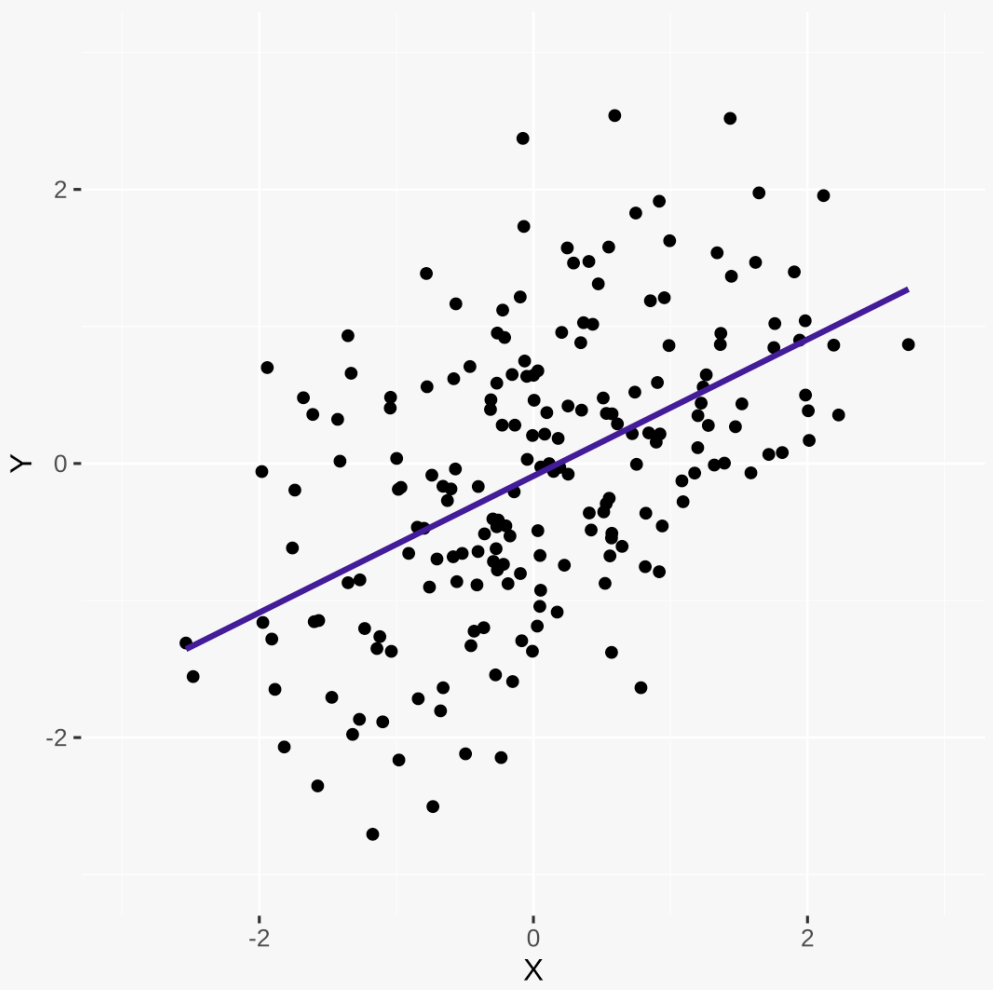
With these pieces of information we can simulate what Y might look like.

$$Y = B_0 + B_1 X_1 + e$$

B_0 = 0, B_1 = 0.5, X_1 = $N(0, 1)$, e = $N(0, 1)$

* $N(0, 1)$ means "A normally distributed variable with a mean of 0 and a standard deviation of 1".

Each time we simulate the data, the results will change a little bit because the values of X and e will be drawn at random from their distributions.



Working backwards from a linear regression to simulated data...

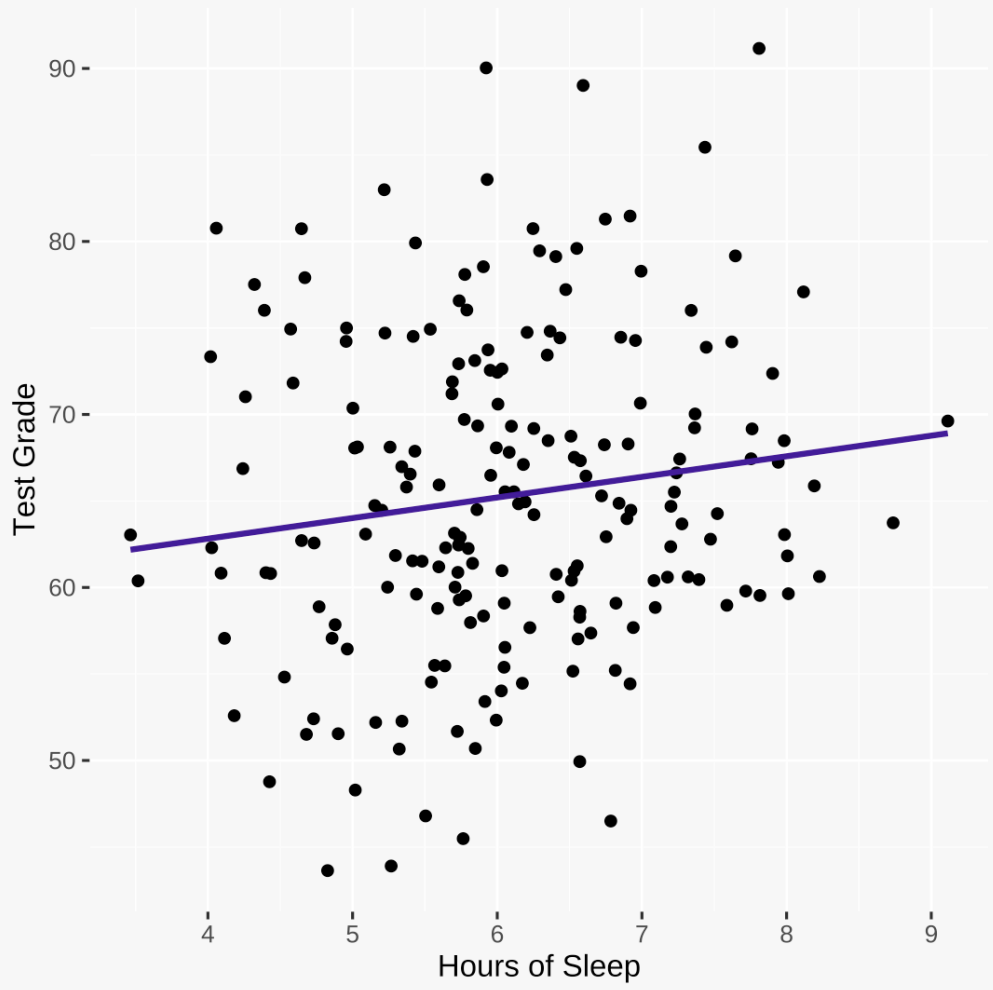
With these pieces of information we can simulate what Y might look like.

$$Y = B_0 + B_1 X_1 + e$$

B_0 = 60, B_1 = 1, $X_1 = N(6, 1)$, $e = N(0, 10)$

* $N(0, 1)$ means "A normally distributed variable with a mean of 0 and a standard deviation of 1".

We can change the values of the distribution to represent real-world variables better, e.g. Y = Test Grade, X_1 = Average Hours of Sleep.



Working backwards from a linear regression to simulated data...

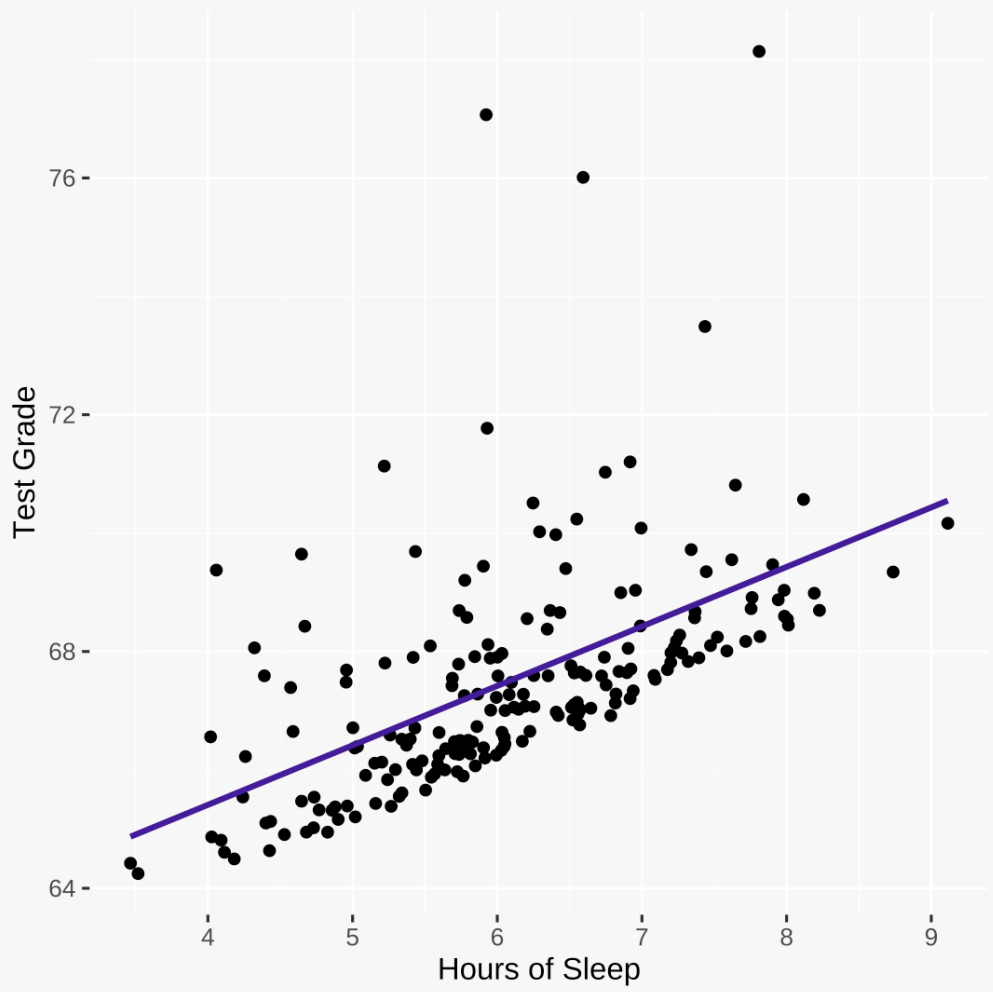
With these pieces of information we can simulate what Y might look like.

$$Y = B_0 + B_1 X_1 + e$$

$B_0 = 60$, $B_1 = 1$, $X_1 = N(0,1)$, $e = Lognormal(0,1)$

* $N(0,1)$ means "A normally distributed variable with a mean of 0 and a standard deviation of 1".

We could also change the distributions that we are using to see what kind of effect it has, for example, if the error was log-normal distributed... (but we won't really go into that in this course)



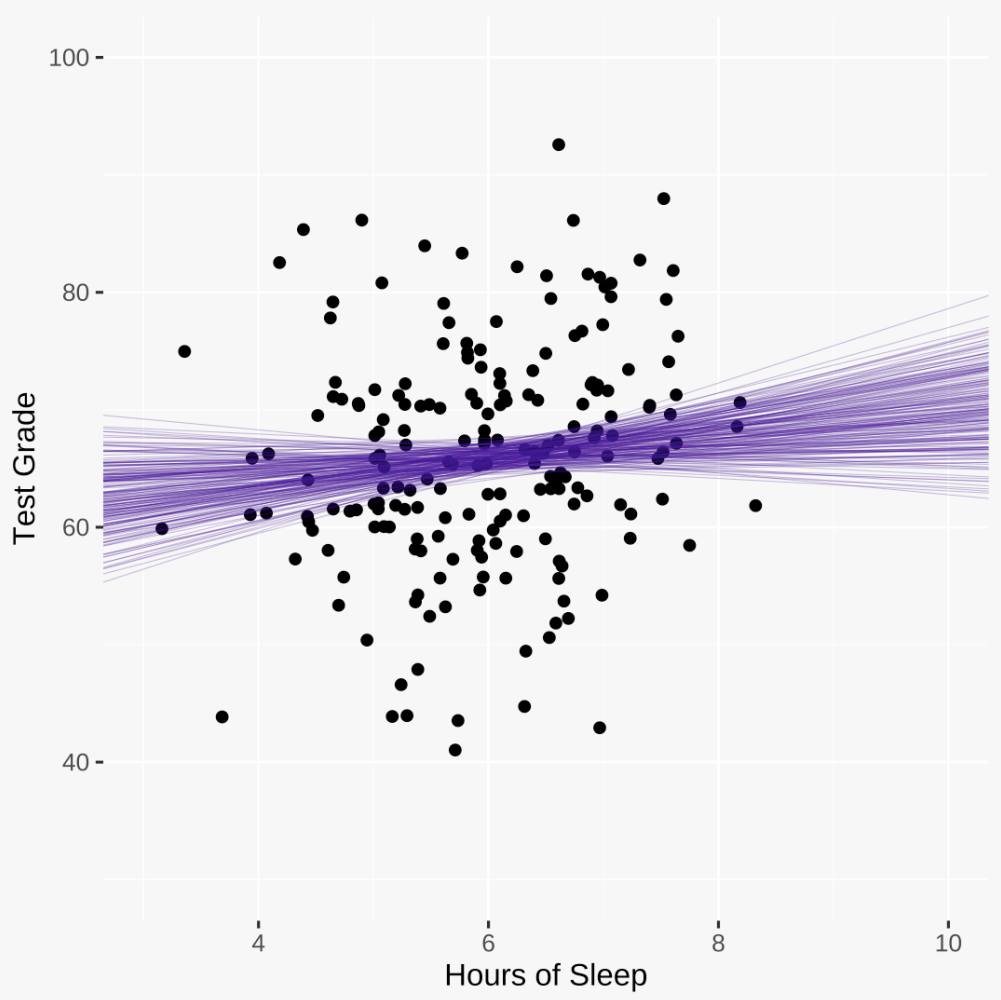
Simulation shows how uncertainty (of effect, Bayesian, or of sample, Frequentist) plays out

On the right are the regression lines for 200 simulations based on the regression equation in the previous slides with a sample size of 200 (with the last simulation's data overlaid on top).

- If our sample is truly random, our estimates of B0 and B1 should be normally distributed around the true value.

Here's the first five results for B_0 and B_1:

B_0	B_1
55.99	1.78
58.27	1.49
60.56	0.92
59.09	1.05
64.81	0.26



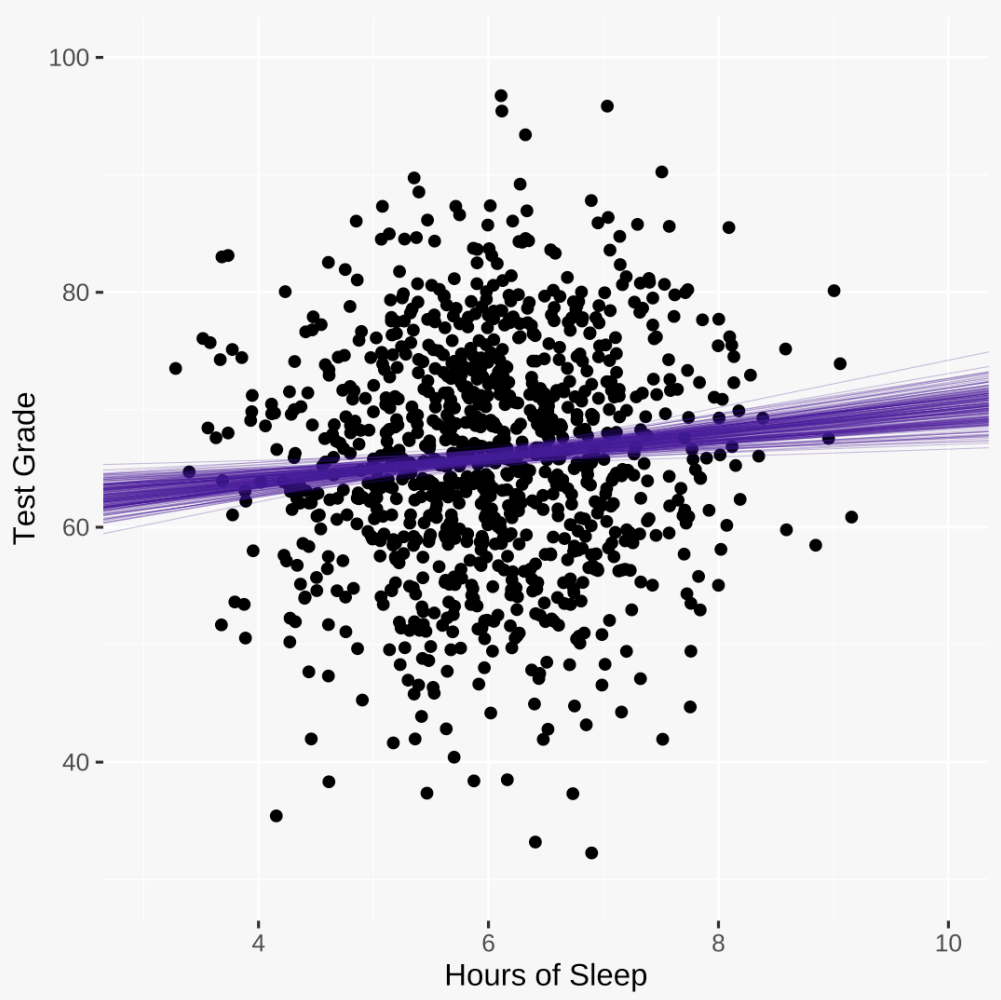
Simulation shows how uncertainty (of effect, Bayesian, or of sample, Frequentist) plays out

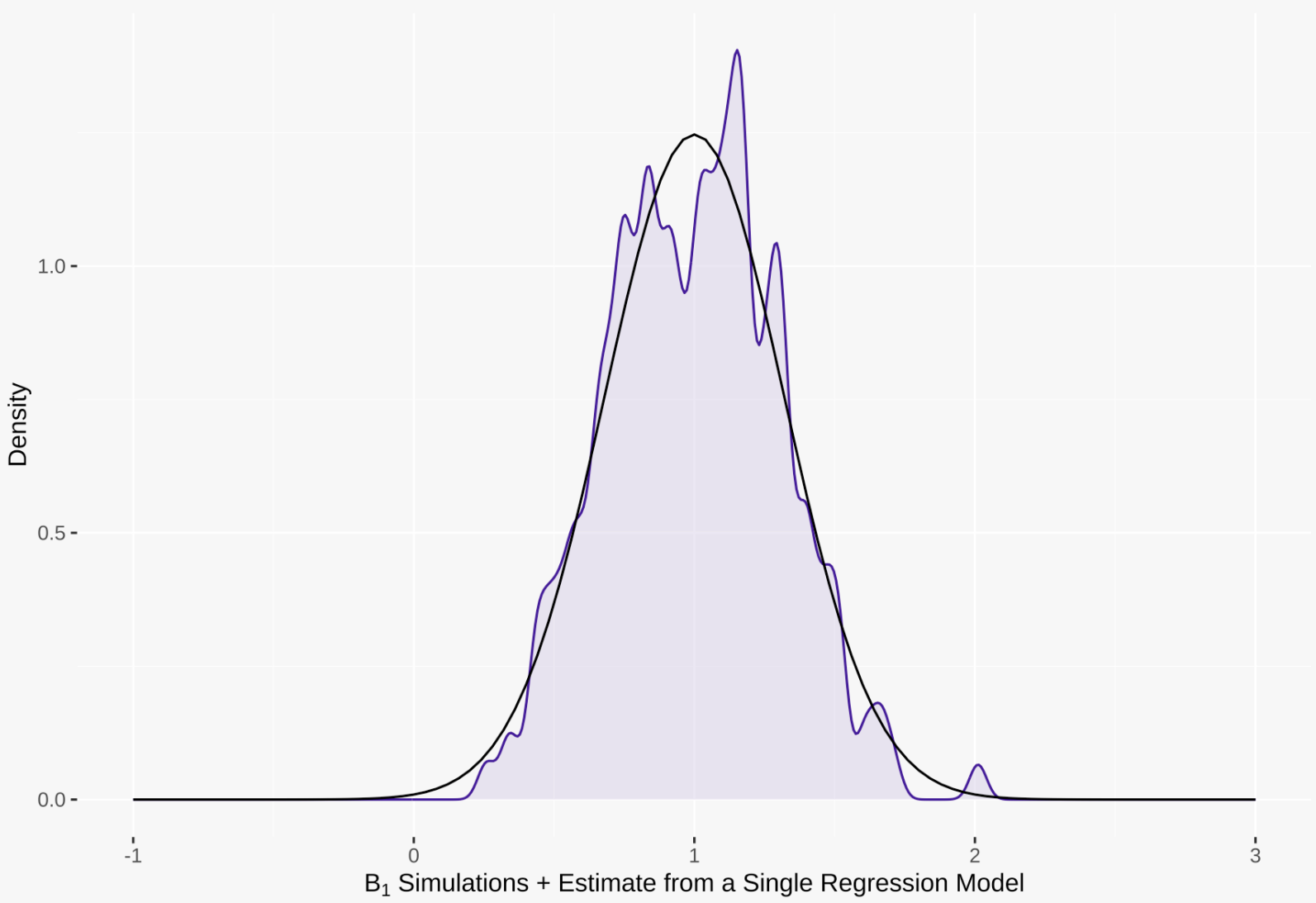
On the right are the regression lines for 200 simulations based on the regression equation in the previous slides with a sample size of 1000 (with the last simulation's data overlaid on top).

- If our sample is truly random, our estimates of B0 and B1 should be normally distributed around the true value.

Here's the first five results for B_0 and B_1:

B_0	B_1
55.94	1.66
58.49	1.31
62.52	0.60
58.60	1.16
62.71	0.59




Density
1.0
0.5
0.0
-1
0
1
2
3
B_1 Simulations + Estimate from a Single Regression Model

Building up the complexity

Most of the time we're going to be interested in models with more than one independent variable. Moreover, we're likely to be interested in know how **the presence or absence of variables that are associated with both our key predictors of interest and our outcome, and the strength of association, influences our results.**

- How bad would the difference in propensity to non-respond to a survey need to be to invalidate the results?

Building up the complexity

Most of the time we're going to be interested in models with more than one independent variable. Moreover, we're likely to be interested in know how **the presence or absence of variables that are associated with both our key predictors of interest and our outcome, and the strength of association, influences our results.**

- How bad would the difference in propensity to non-respond to a survey need to be to invalidate the results?

Cornfield Conditions Study:

- In the 1950s, many statisticians (including Karl Pearson) believed there was no good evidence that smoking caused cancer because an omitted, unmeasurable variable, whether someone has a **smoking gene** that makes them more likely to smoke and more likely to develop lung cancer, could exist.
- Jerome Cornfield demonstrated that the effect of such an omitted variable would have to be implausibly large in order to "explain away" the link between smoking and cancer.

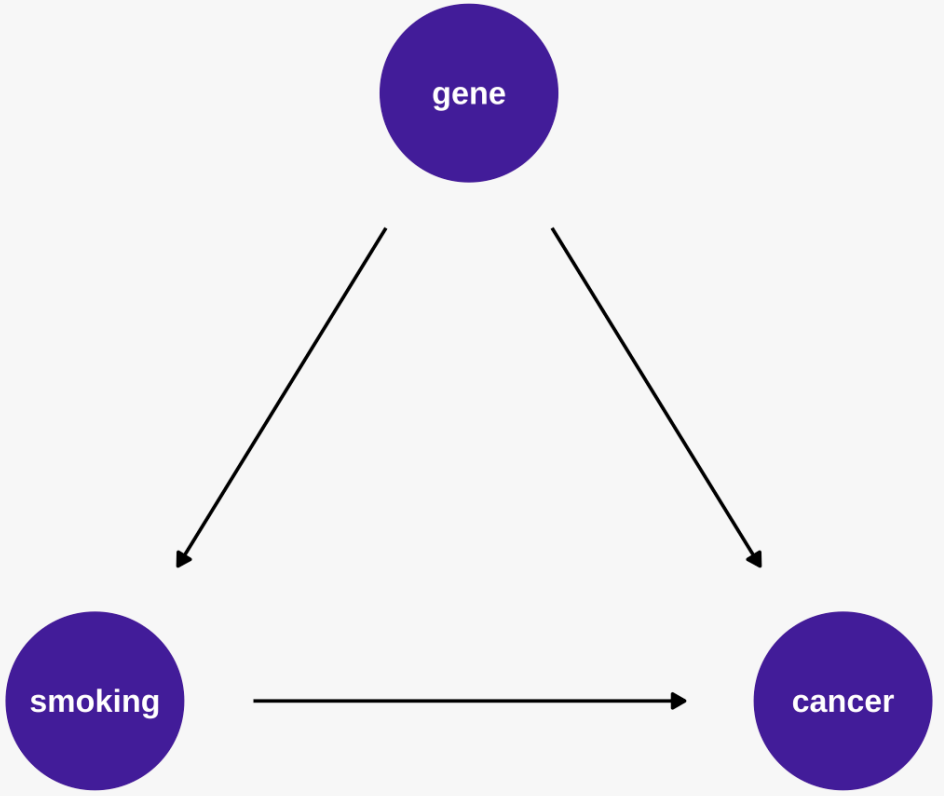
Building up the complexity

Most of the time we're going to be interested in models with more than one independent variable. Moreover, we're likely to be interested in know how **the presence or absence of variables that are associated with both our key predictors of interest and our outcome, and the strength of association, influences our results.**

- How bad would the difference in propensity to non-respond to a survey need to be to invalidate the results?

Cornfield Conditions Study:

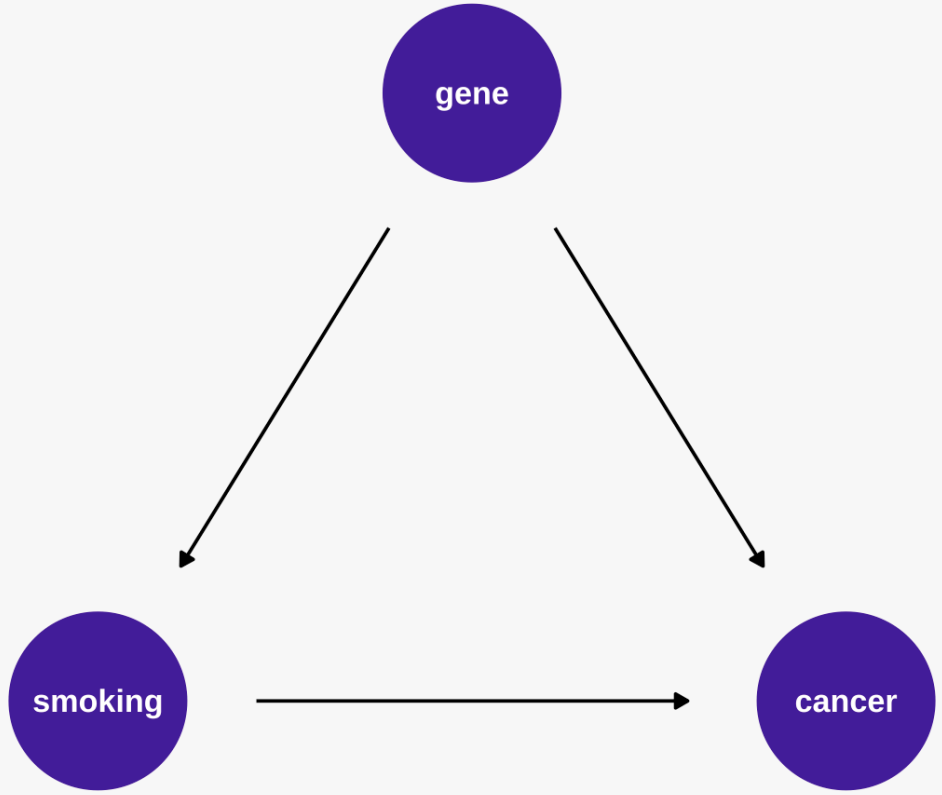
- In the 1950s, many statisticians (including Karl Pearson) believed there was no good evidence that smoking caused cancer because an omitted, unmeasurable variable, whether someone has a **smoking gene** that makes them more likely to smoke and more likely to develop lung cancer, could exist.
- Jerome Cornfield demonstrated that the effect of such an omitted variable would have to be implausibly large in order to "explain away" the link between smoking and cancer.



Building up the complexity

The order of simulation

- Identify a variable/variables that are not a product of any other variables (i.e. they only have arrows going away from them, not to them)

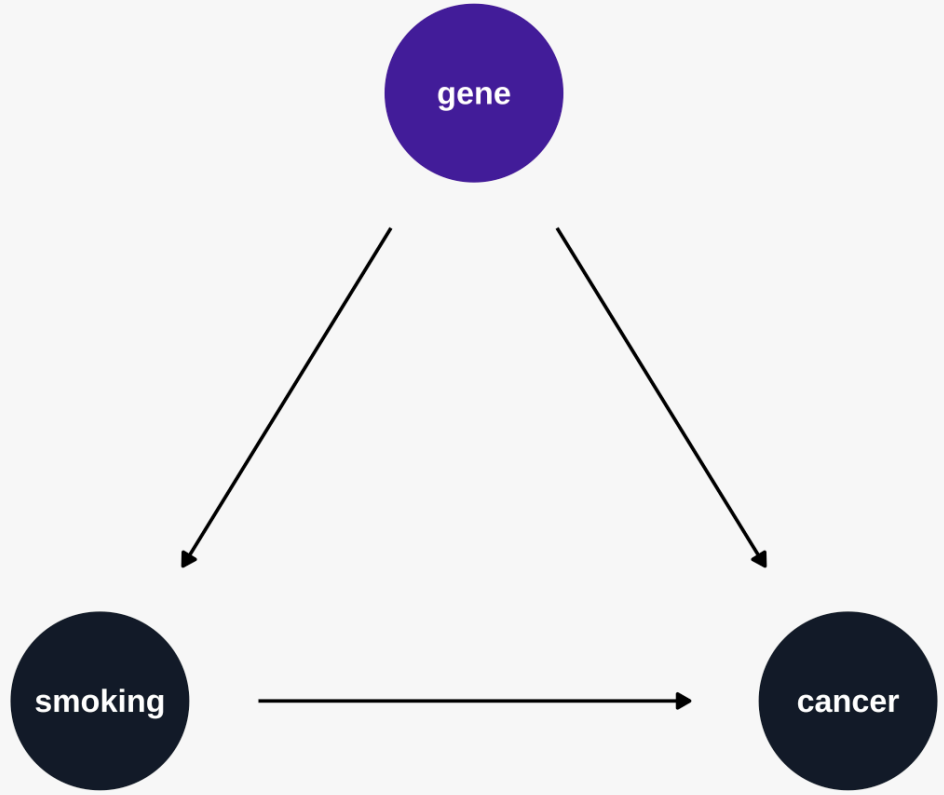


Building up the complexity

The order of simulation

- Identify a variable/variables that are not a product of any other variables (i.e. they only have arrows going away from them, not to them)
- Simulate those variables.

$$gene = N(\mu_1, \sigma_1)$$

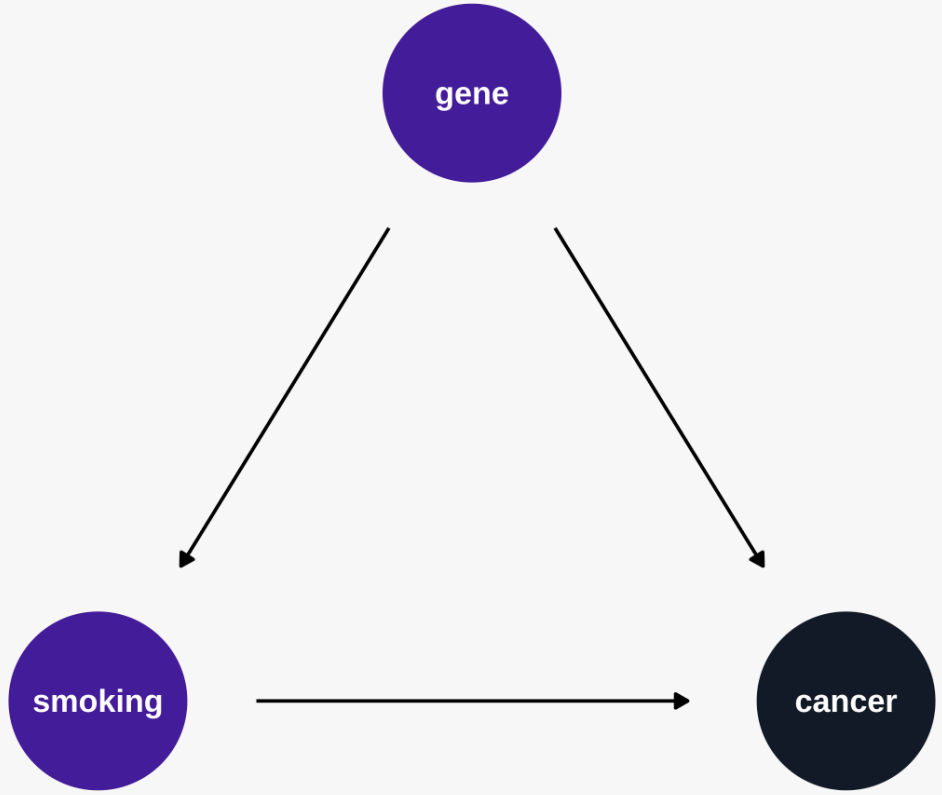


Building up the complexity

The order of simulation

- Identify a variable/variables that are not a product of any other variables (i.e. they only have arrows going away from them, not to them)
- Simulate those variables.
- Now you have a new variable, you can identify which other variables you now have the pieces of to simulate.

$$gene = N(\mu_1, \sigma_1)$$
$$smoking = B_0 + B_1 gene + e_1$$
$$e_1 = N(0, \sigma_2)$$

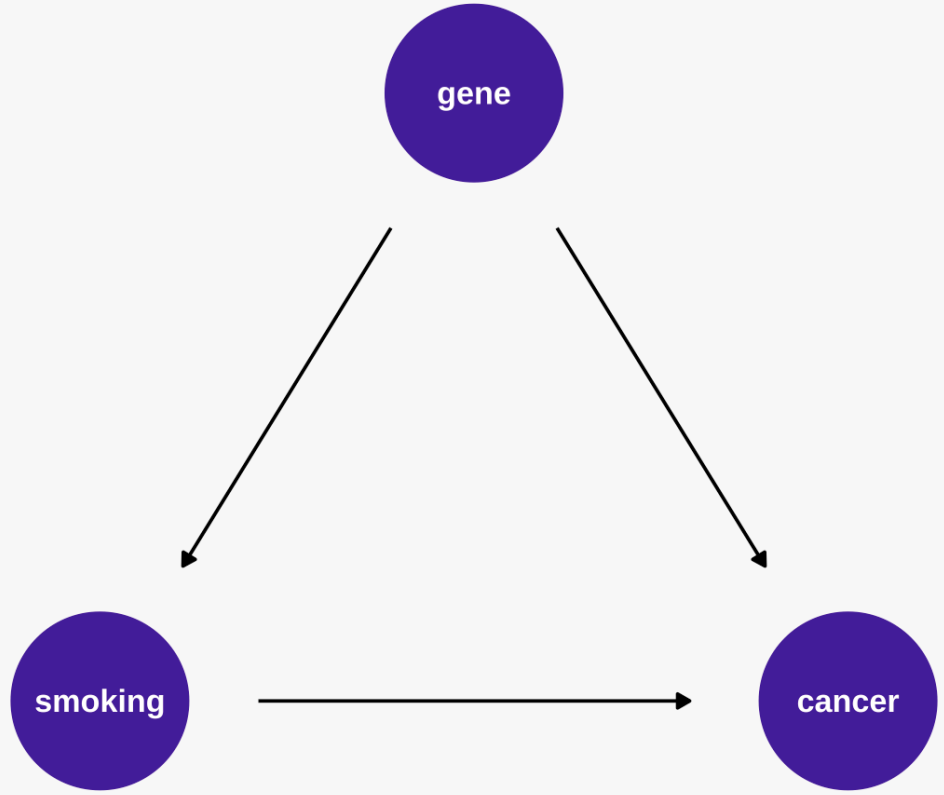


Building up the complexity

The order of simulation

- Identify a variable/variables that are not a product of any other variables (i.e. they only have arrows going away from them, not to them)
- Simulate those variables.
- Now you have a new variable, you can identify which other variables you now have the pieces of to simulate.
- Repeat until you have simulated all of your variables.

$$
\begin{aligned}
gene &= N(\mu_1, \sigma_1) \\
smoking &= B_0 + B_1 gene + e_1 \\
cancer &= B_3 + B_4 smoking + B_5 gene + e_2 \\
e_1 &= N(0, \sigma_2) \\
e_2 &= N(0, \sigma_3)
\end{aligned}
$$

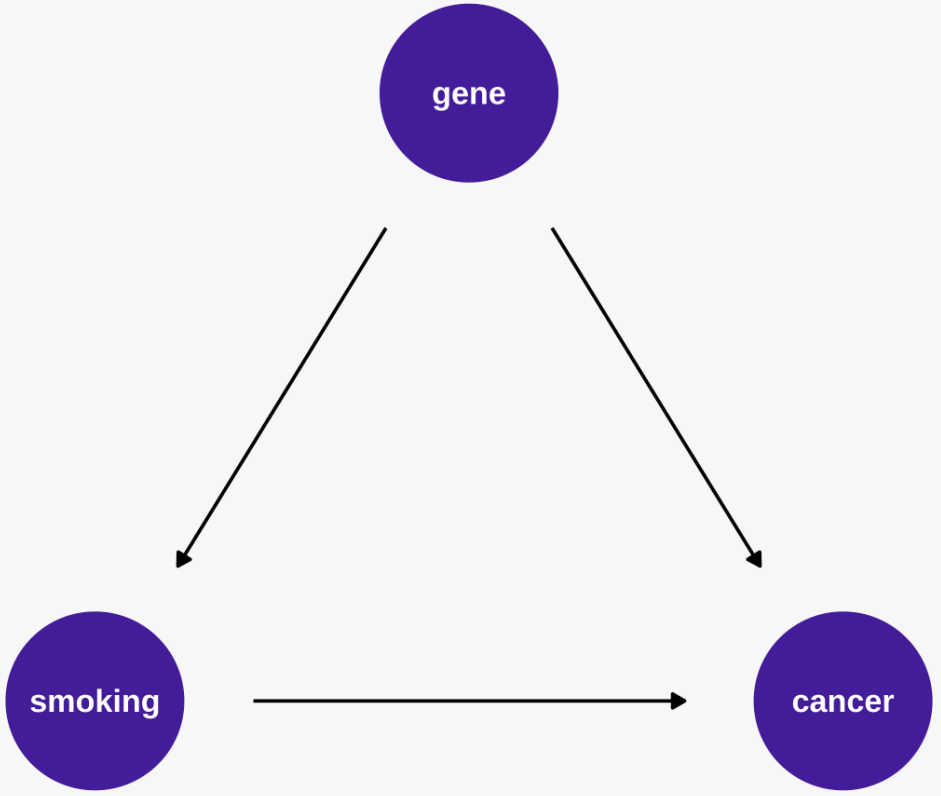


Building up the complexity

The order of simulation

- Identify a variable/variables that are not a product of any other variables (i.e. they only have arrows going away from them, not to them)
- Simulate those variables.
- Now you have a new variable, you can identify which other variables you now have the pieces of to simulate.
- Repeat until you have simulated all of your variables.

A warning: This does become a bit more complicated when it comes to more complex relationships between variables: simulation of trends, multilevel data structures, latent factors, bidirected relationships, etc. but the general principle stays the same.



One simulation to many...

Now that you have a simulation that works (represents the underlying data generating process), we need to repeat it many, many, times, to see what happens. Some things to keep in mind:

- Can someone reproduce your results?
- Is there a clear description of what each part of your code is intended to do?
- What things do you actually want to vary, how much work will that be, and how fine-grained do the conditions need to be:
  - Sample size?
  - Relationship between confounder and outcome?
  - Relationship between confounder and predictor?
  - Both?
  - Distribution of the error terms?

One simulation to many...

Now that you have a simulation that works (represents the underlying data generating process), we need to repeat it many, many, times, to see what happens. Some things to keep in mind:

- Can someone reproduce your results?
- Is there a clear description of what each part of your code is intended to do?
- What things do you actually want to vary, how much work will that be, and how fine-grained do the conditions need to be:
    - Sample size?
    - Relationship between confounder and outcome?
    - Relationship between confounder and predictor?
    - Both?
    - Distribution of the error terms?

Best practice

1. Random is not truly random. Randomness is generated based on starting conditions. These can be set to be the same across several instances of a script by **setting a seed** that determines the randomness.
2. **Generate your data** using your simulation code
3. **Estimate your model**
4. **Save the results** for the things you're interested in (parameters, SEs, etc.)
5. **Repeat** with different randomness (re-sampling from the same data generating process)
6. **Repeat** under different conditions (different confounding effects, different sample sizes, etc.)

Only change one thing at a time unless you're interested in how the two things interact with each other.

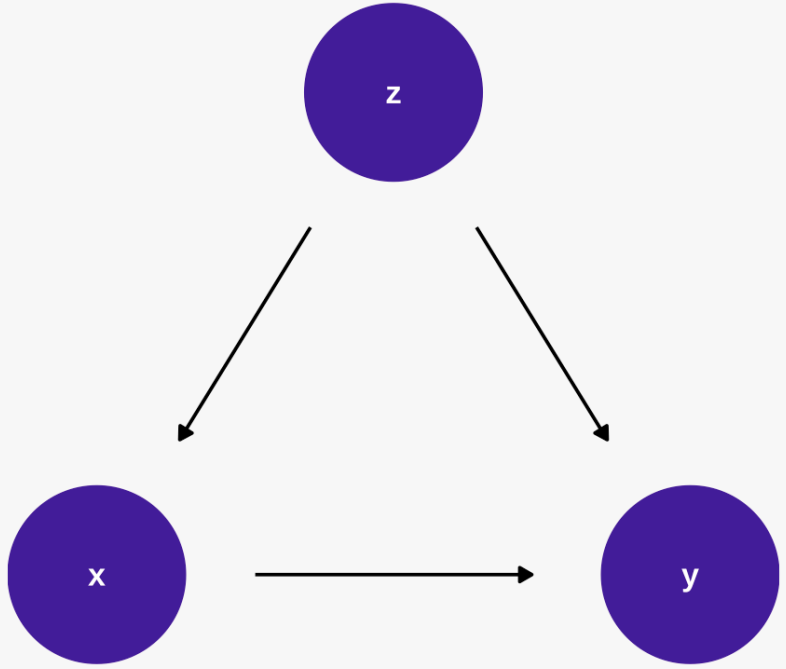
Exercise 1

In the first exercise, we'll be building up a simple simulation similar to the graph we've just looked at (an outcome (Y), a predictor of interest (X), and a confounder (Z)) and see **what happens to our estimates of our regression coefficient** if we exclude this confounder. We'll then look at how the strength of that confounder's relationship with the outcome affects that estimate.

You'll then be tasked with creating your own code seeing what happens if Z is a collider instead of a confounder, meaning a variable that is caused by both X and Y.

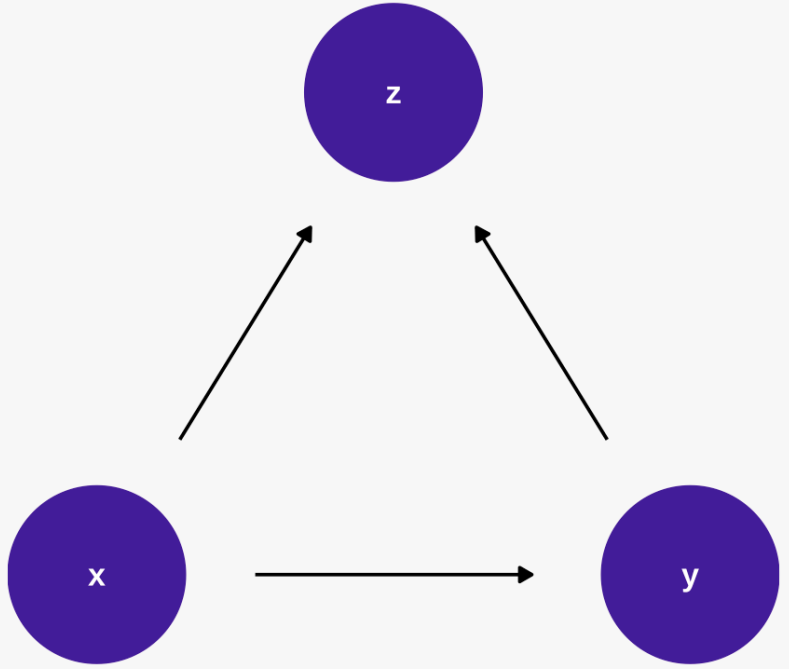
Demonstration

How does the treatment of a confounder change our predictions?



Exercise

How does the treatment of a collider change our predictions?



Lecture 3: Simulation quantities of interest

What have we done so far?

What simulations are good for, why the they are useful, the logic of how to simulate based on a data generating process (that you can draw graphically), how to run a (single) simulation.

What have we done so far?

What simulations are good for, why the they are useful, the logic of how to simulate based on a data generating process (that you can draw graphically), how to run a (single) simulation.

What's next?

How we can usefully summarise the results of our simulations in a standardised way, how to interpret and present those results, and how to run multiple simulations and store the results "on the fly".

What do we want to measure?

- What does the model usually say?
- Does the model generally get it “right” on average?
- Are the results very dispersed?
- Do the measures of dispersal (SE) accurately reflect the true dispersal?

What do we want to measure?

- What does the model usually say?
- Does the model generally get it “right” on average?
- Are the results very dispersed?
- Do the measures of dispersal (SE) accurately reflect the true dispersal?

Quantities of Interest

In simulation studies, the key quantities of interest are:

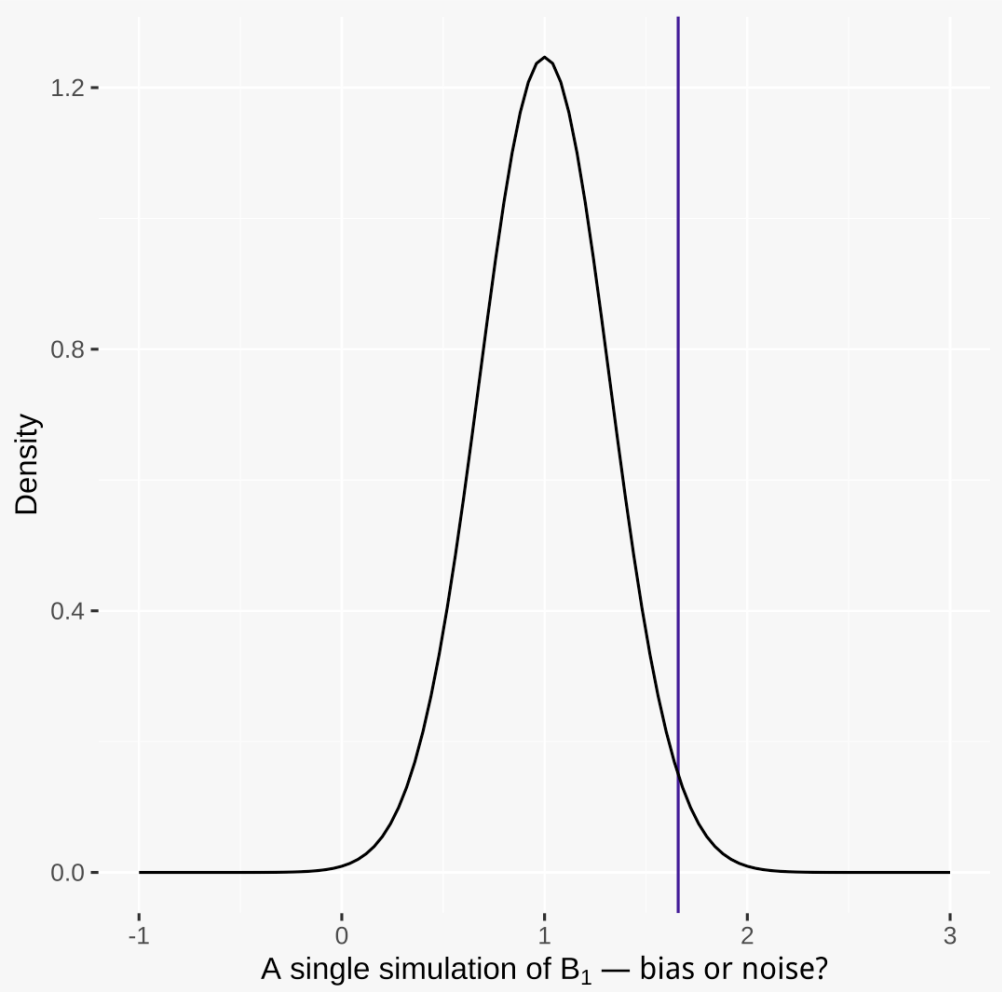
- Predictions
- Bias
- Root Mean Square Error
- Optimism
- 95% coverage

Repeated examples are needed

To do this, we can't just have a single simulation

- That would contain random white noise - so difficult to tell what is bias and what is white noise

Instead, simulate the data lots of time (e.g. 1000 times) and see how the results vary across those



unbiased, precise

biased, precise

unbiased, imprecise

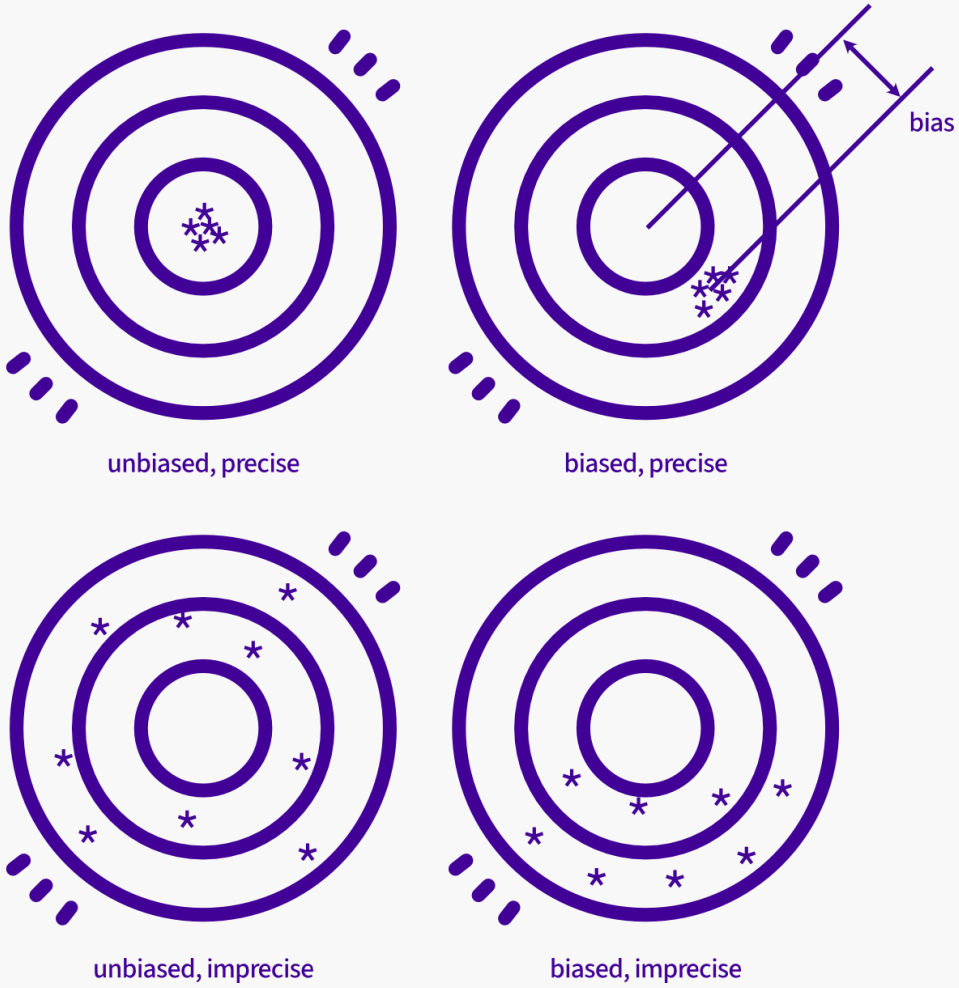
biased, imprecise

Predictions

We can make predictions from the model, based on some, or all, of the estimated parameters and compare these to predictions from the DGP

Useful for visualising problems (covered in the next lecture).

But can't quantify the problem in a standard way, difficult to compare models/specifications, etc.



Bias

> We want to estimate how far off the average of the simulations is

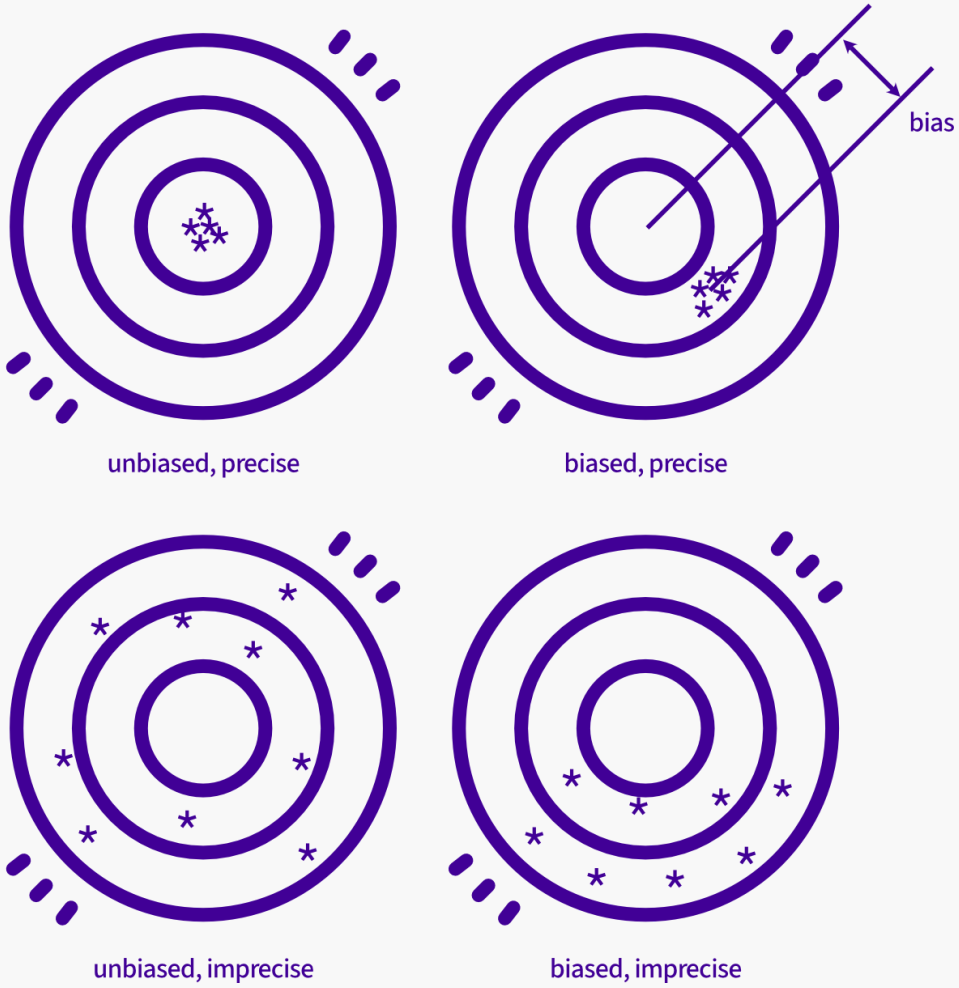
Two ways of estimating this:

- (1) as a ratio: Bias = estimate / truth
  - In this case, a perfect result is 1

$$\text{Bias} = 100 \times (\frac{\bar{B}}{B_{\text{true}}})$$

$\bar{B}$ = B-bar = The mean of all parameter estimates from the simulations

B_{true} = B true = The true value of B specified in the simulation



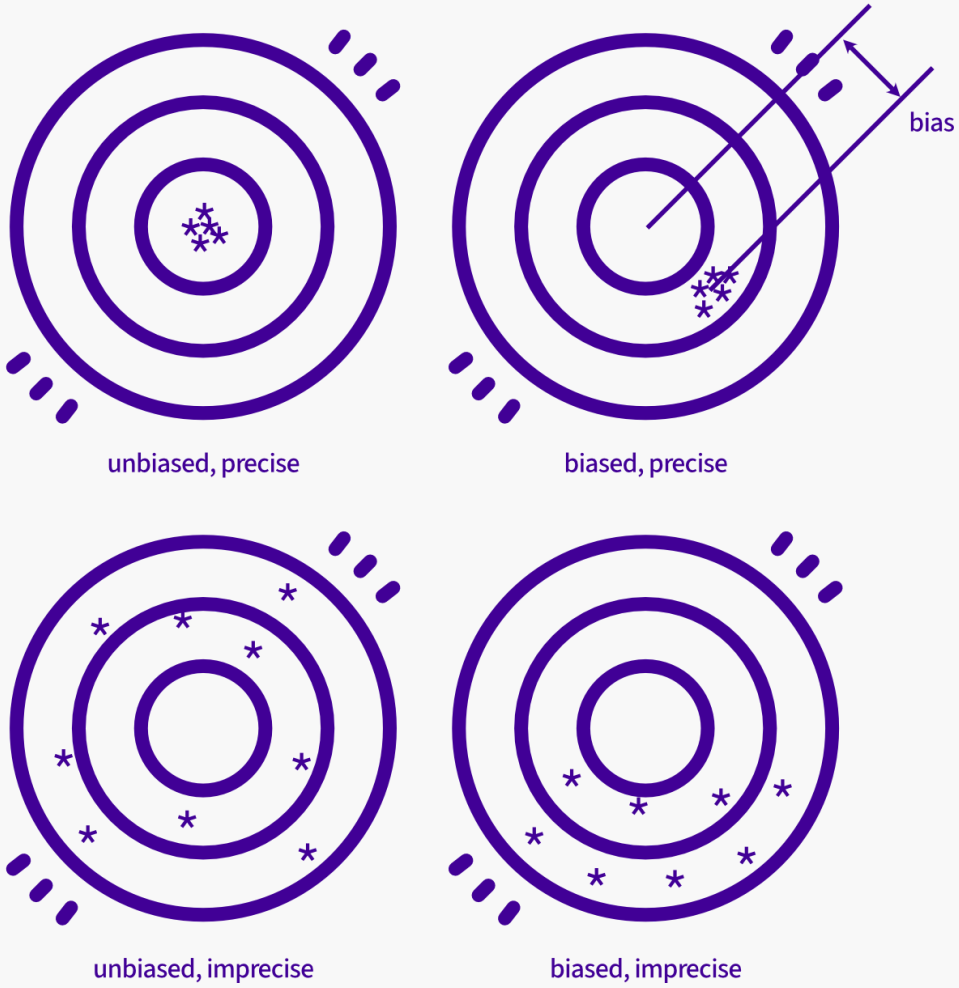
Bias

> We want to estimate how far off the average of the simulations is

Two ways of estimating this:

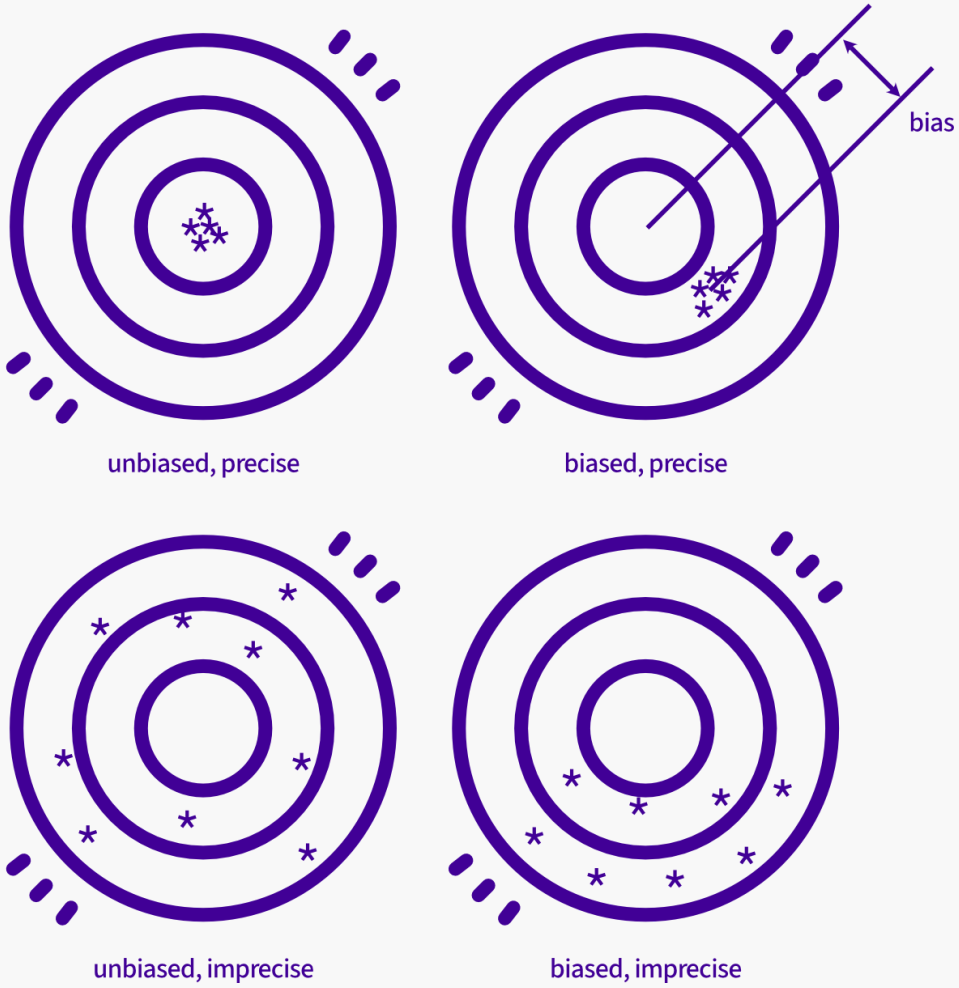
- (1) as a ratio: Bias = estimate / truth
  - In this case, a perfect result is 1
- (2) As a proportion: (estimate-truth)/truth
  - Here, the perfect result os 0

$$\frac{\bar{B} - B_{\text{true}}}{B_{\text{true}}} \times 100$$



Bias

- These measures tell us about the averages
- But not about the individual model deviations
- A model might be correct on average, but so wrong each individual time to be close to meaningless
- Often a bit of bias is worth having more precision



Root Mean Square Error

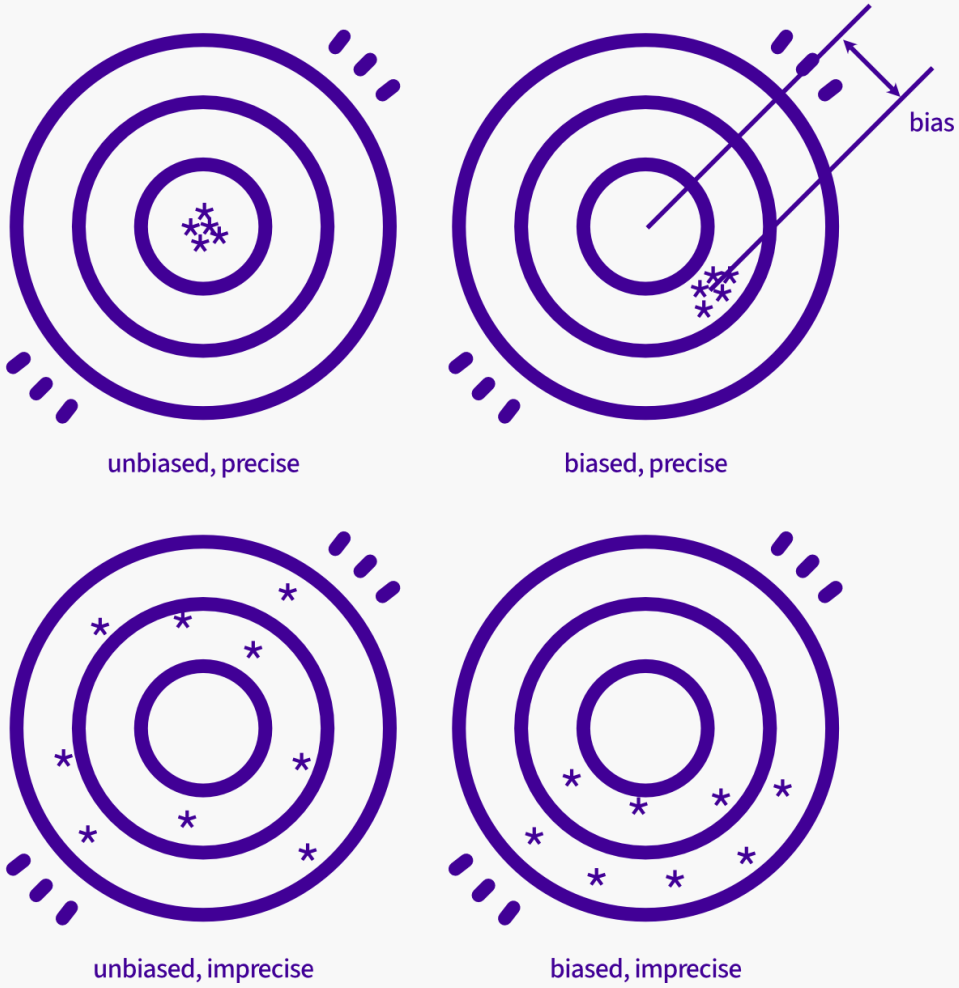
> We want a measure that incorporates both bias and precision

How close to the truth are the individual model runs rather than how close to the truth is the average of the model runs?

This effectively combines bias and precision/efficiency.

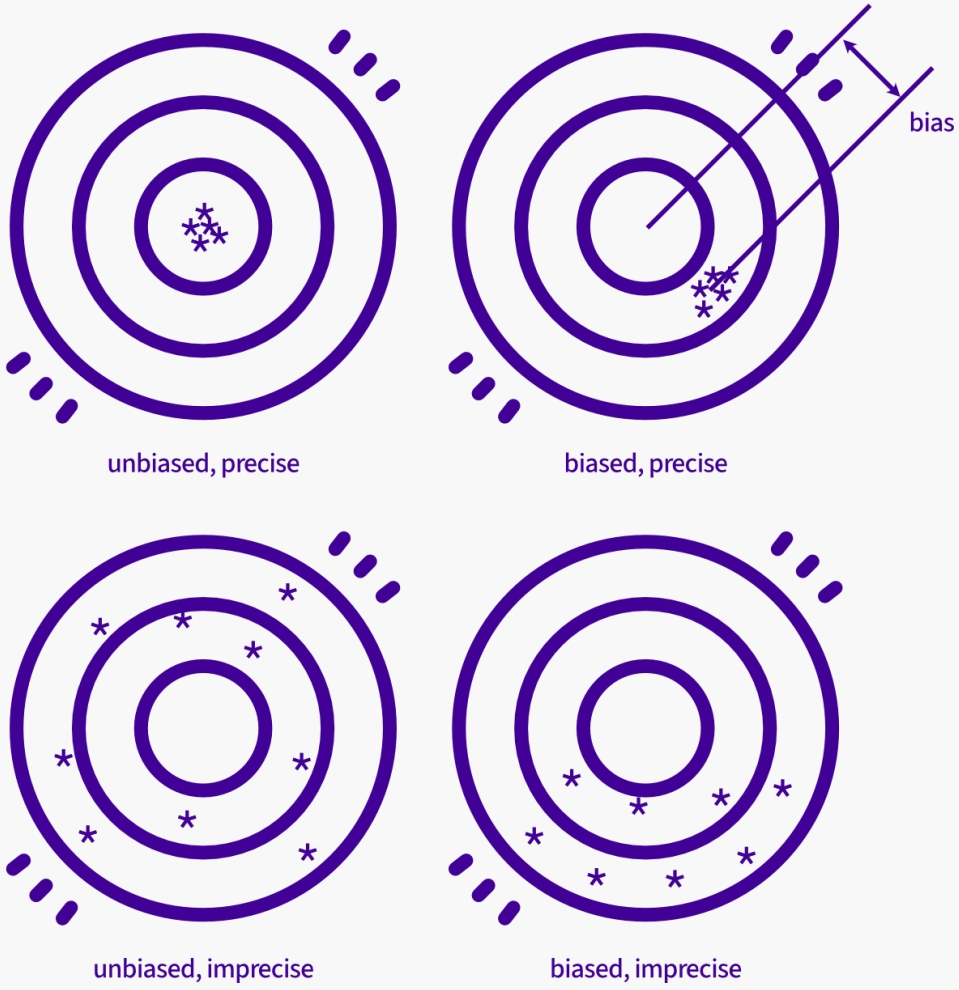
$$\text{RMSE} = \sqrt{\frac{\sum_{l=1}^{nsims}(\beta^{(l)} - \beta_{\text{true}})^2}{n}}$$

The square root os the sum of each parameter estimate for each simulation minus the true value of the parameter from the data generating process squared and divided by the number of observations.



What about standard errors?

- Everything so far has been about coefficient estimates
- But what about the SEs that go with them?
- These are just as important as the coefficients themselves
- Our models need to tell us the correct amount of uncertainty; they shouldn't be overly optimistic or overly pessimistic.



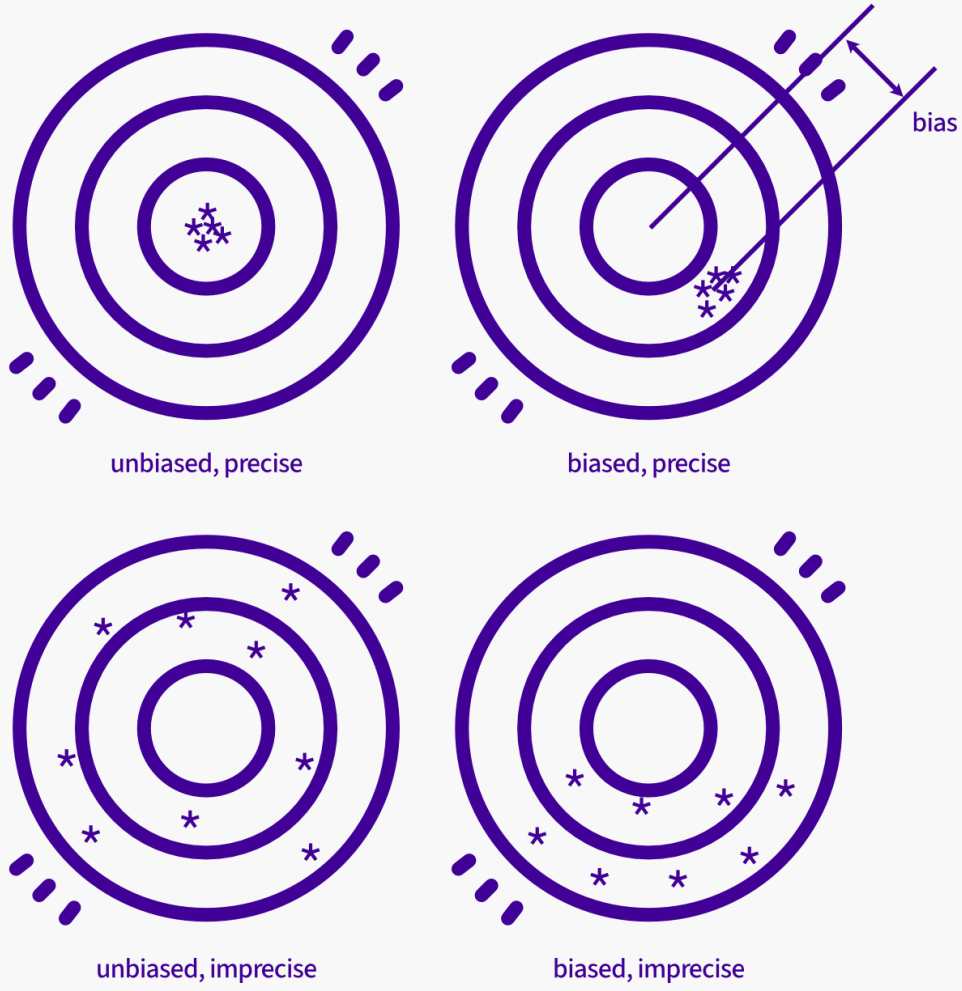
What about standard errors?

Fortunately, the distribution of our simulated results is exactly the uncertainty that our model should be estimating.

So we can compare our model's estimated uncertainty (SE) to the variability of our simulations

$$\text{Optimism} = 100 \times \frac{\sqrt{\sum_{l=1}^{\text{nsims}} (\beta^{(l)} - \bar{\beta})^2}}{\sqrt{\sum_{l=1}^{\text{nsims}} (\text{SE}(\beta^{(l)}))^2}}$$

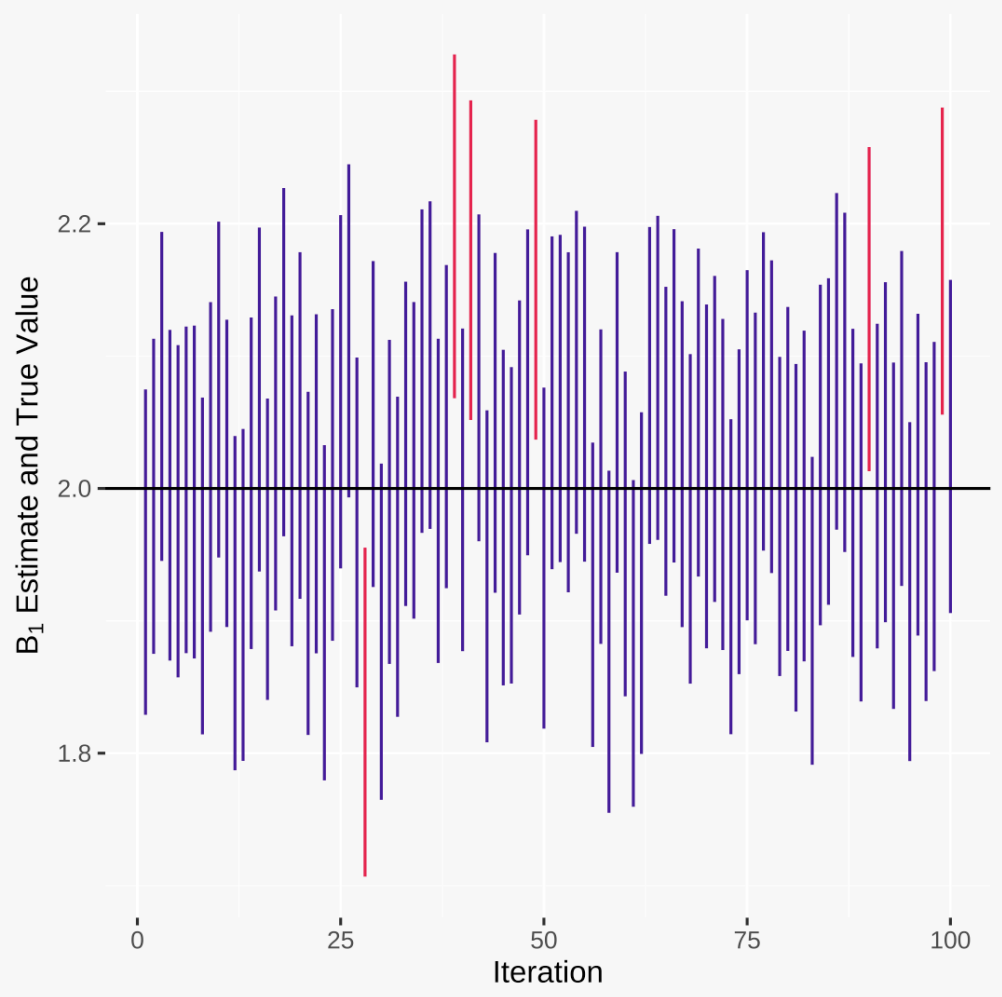
$\bar{\beta}$ = The mean of the parameter estimates from all of the simulations.



What about standard errors?

Alternatively, 95% Coverage.

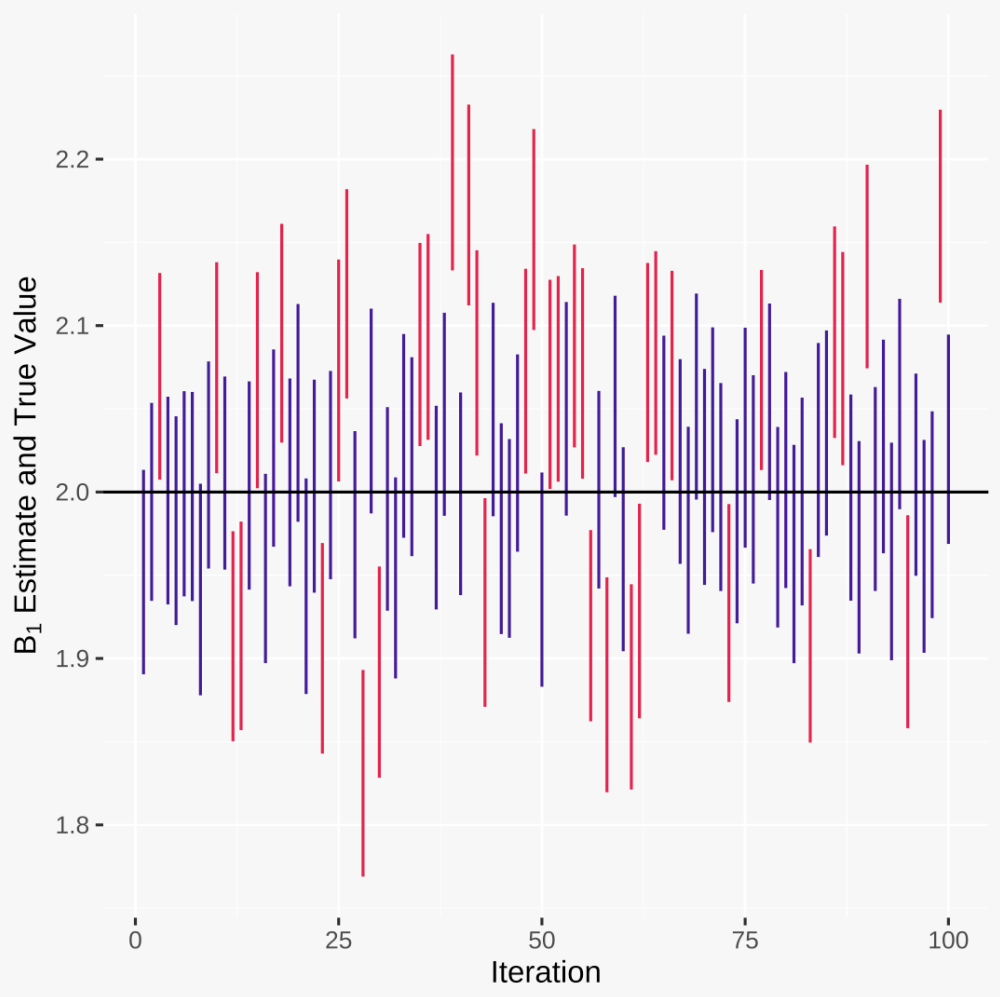
- Take the 95% confidence intervals for each simulation iteration
- See what proportion of those contain the true value
- Should be 95% of them (100% is not a good result!)



What about standard errors?

Alternatively, 95% Coverage.

- Take the 95% confidence intervals for each simulation iteration
- See what proportion of those contain the true value
- Should be 95% of them (100% is not a good result!)



Optimism versus 95% Coverage

- Both aim for much the same thing
- 95% coverage more constrained by 100% limit than 0% limit
- Optimism less constrained (although can't go below 0)
- 95% coverage useful for Bayesian models: doesn't require SE Normality assumptions

General Process for Repeat Simulations

1. Repeat the following*1000
    - Generate data
    - Run model
    - Save parameters of interest (beta, SE, etc)
2. Calculate quantities of interest for those 1000 parameters (e.g. bias, RMSE, etc)
3. Repeat for all modelling situations of interest (e.g. changes in sample size, omitted variable bias, nested data structures)

General Process for Repeat Simulations

1. Repeat the following*1000
    - Generate data
    - Run model
    - Save parameters of interest (beta, SE, etc)
2. Calculate quantities of interest for those 1000 parameters (e.g. bias, RMSE, etc)
3. Repeat for all modelling situations of interest (e.g. changes in sample size, omitted variable bias, nested data structures)

In Sum

We can work out general model performance by simulating data (and running models on those data) lots of times.

That allows us to work out things about the models performance that aren't clear from a single iteration.

We can measure the bias and efficiency of parameter estimates, and the bias of the SE estimates.

Lecture: Presenting Simulation Results

Principles of Data Visualisation

- Data storytelling
- Data visualisation and design principles
- Medium of publication

Principles of Data Visualisation

- **Data storytelling**
- Data visualisation and design principles
- Medium of publication

1. **Does it need to be a data visualisation?**
    - Can a table of summary statistics do the same job? Does the visualisation of the summary statistics add anything (e.g. making non-linear changes, diminishing returns, etc. easier to spot)?
2. **One story — one visualisation**
    - Is it clear what you are trying to communicate with the data visualisation or is too much being shown at once? Can the story be made clearer using an active title or by splitting important parts into facets?
3. **Who is your audience?**
    - Is your intended audience already familiar with simulations? If so, maybe it's more efficient to just go straight to visualising measures of bias/precision? If not, is illustrating the predictions and 95% confidence/credible intervals more appropriate?

Principles of Data Visualisation

- Data storytelling
- **Data visualisation and design principles**
- Medium of publication

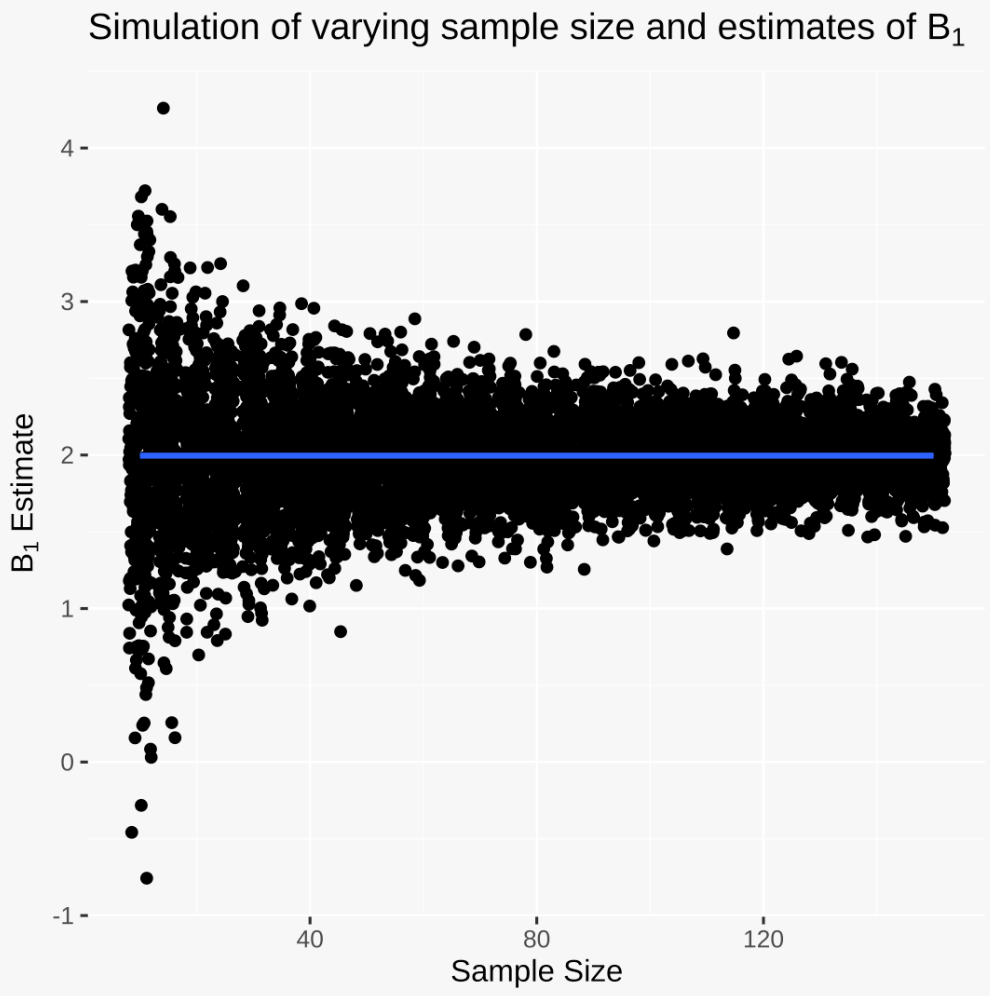
1. **Ensure your presentation of data is accessible**
    - Avoid colour schemes that have low accessibility for people with colour-blindness. Ensure that whatever you're showing in the data visualisation is explained in text in some form (even generally). Ensure text and annotations are clear. Use an accessible contrast ratio.
2. **Minimise the ink to data ratio**
    - Avoid too many distracting flourishes. Consider density contours, shaded areas (ribbons etc.) rather than many overlapping points if possible. Avoid junk (caps on the end of segments/bars). Remove extraneous gridlines.
3. **Make good use of colour**
    - Use colour and saturation (if appropriate to the publishing medium) to emphasize the most important feature of the visualisation.

Principles of Data Visualisation

- Data storytelling
- Data visualisation and design principles
- **Medium of publication**

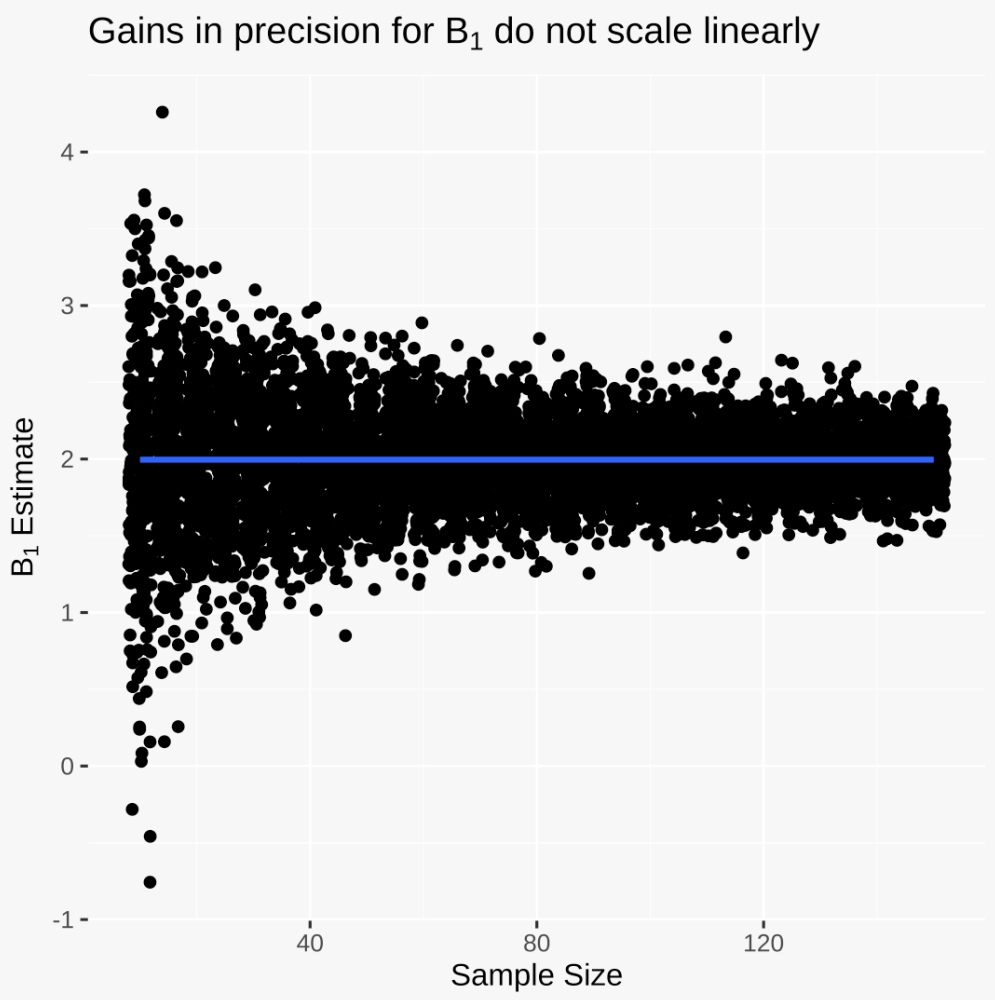
1. **Consider the limitations of the medium you're publishing in**
    - Do you need to use black-and-white? Does your file need to be submitted in a specific format? Do the formats you can use have specific restrictions (e.g. SVG/EPS does not support transparency)
2. **Do you need different versions for different mediums?**
    - Does a print medium with restrictions on figures require you to make a composite of plots? Do you need a landscape version for a full slide in a presentation/for half a page in a publication and a square version for half a slide? Can the same information be displayed on all versions?
3. **Resolution**
    - How high resolution does the image need to be (dpi/ppi) in order for everything to be reproduced well?

Let's improve this...



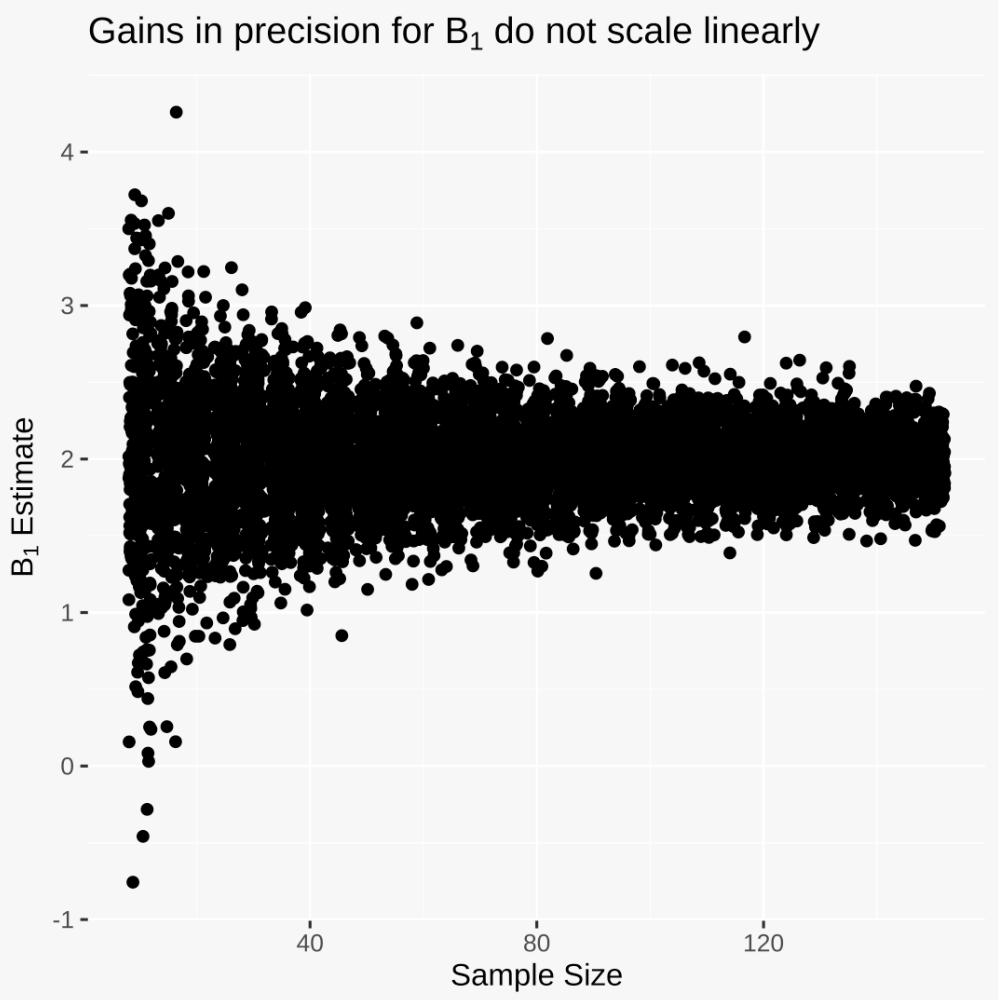
Let's improve this...

- Add an active title.



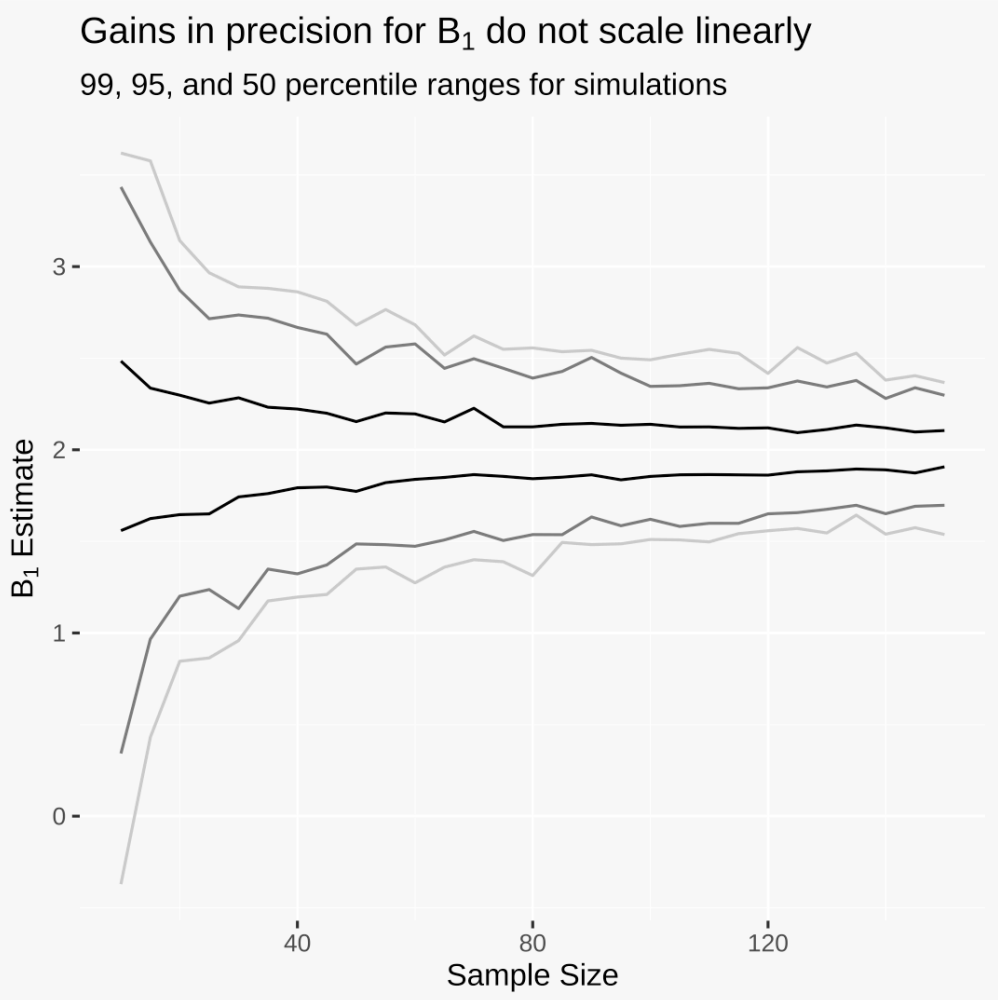
Let's improve this...

- Add an active title.
- Get rid of extraneous information (bias is not the focus here).



Let's improve this...

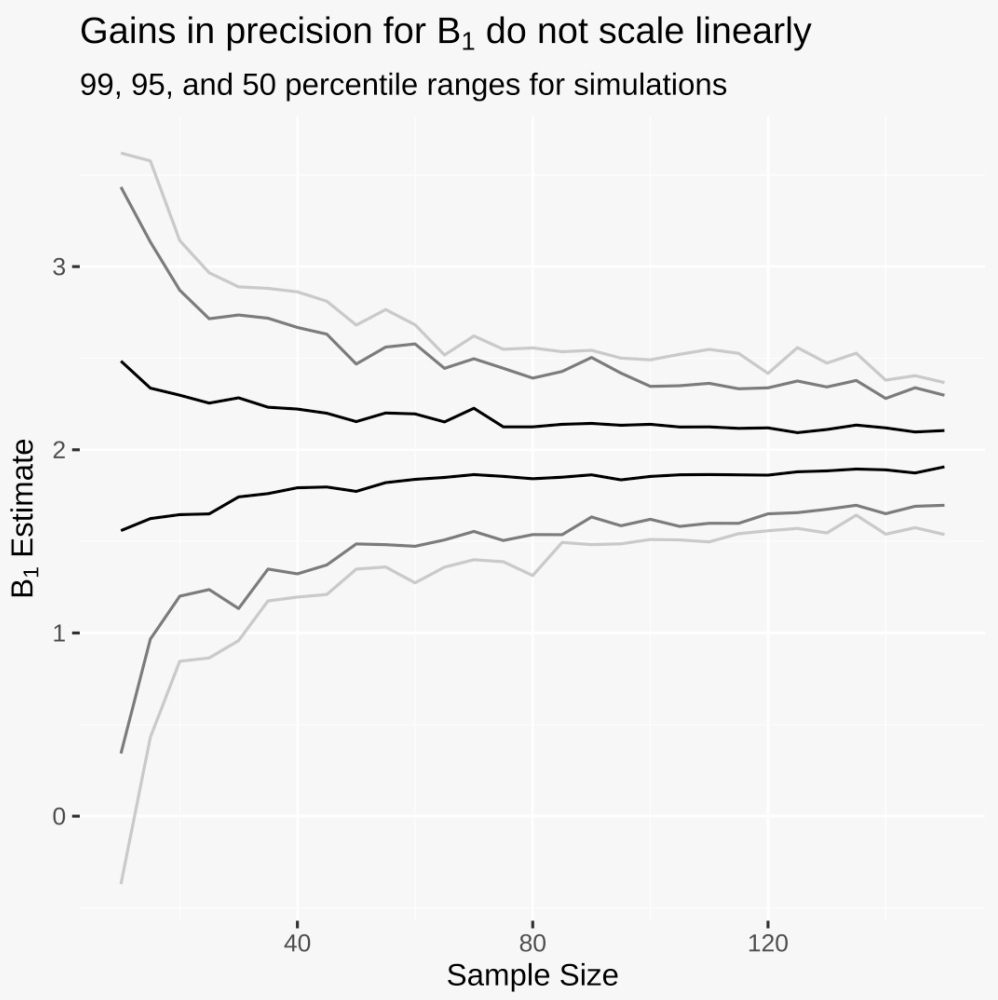
- Add an active title.
- Get rid of extraneous information (bias is not the focus here).
- Minimise our ink-to-data/ink-to-information ratio.



Let's improve this...

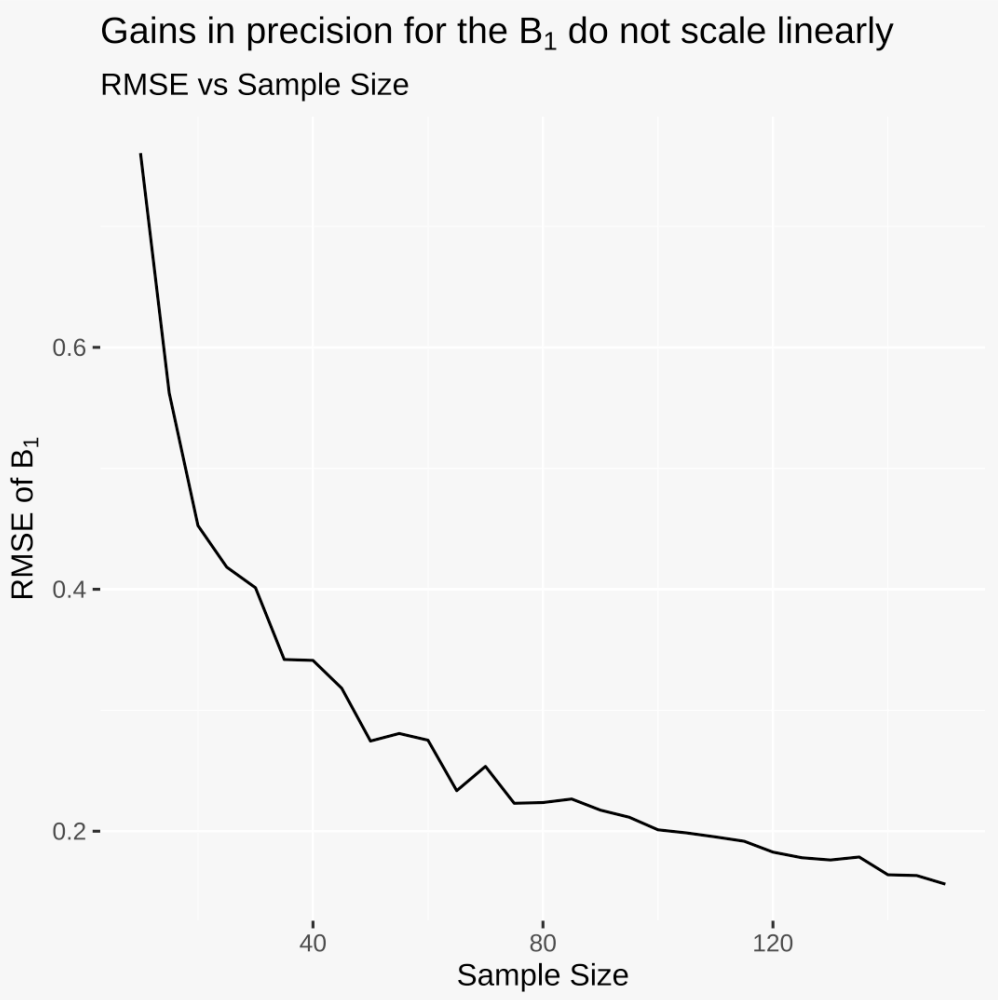
- Add an active title.
- Get rid of extraneous information (bias is not the focus here).
- Minimise our ink-to-data/ink-to-information ratio.

Keep in mind: Sometimes we *do* want to see every single iteration (or a large sample) of iterations to illustrate RMSE and poor coverage of CIs (dispersed results, with or without bias, that are overly confident, e.g. this plot from earlier). *Again, what is it you are trying to show.*



Let's improve this...

- Add an active title.
- Get rid of extraneous information (bias is not the focus here).
- Minimise our ink-to-data/ink-to-information ratio.
- Consider the audience (are they likely to be familiar with simulation terms?)



Some examples from existing literature...

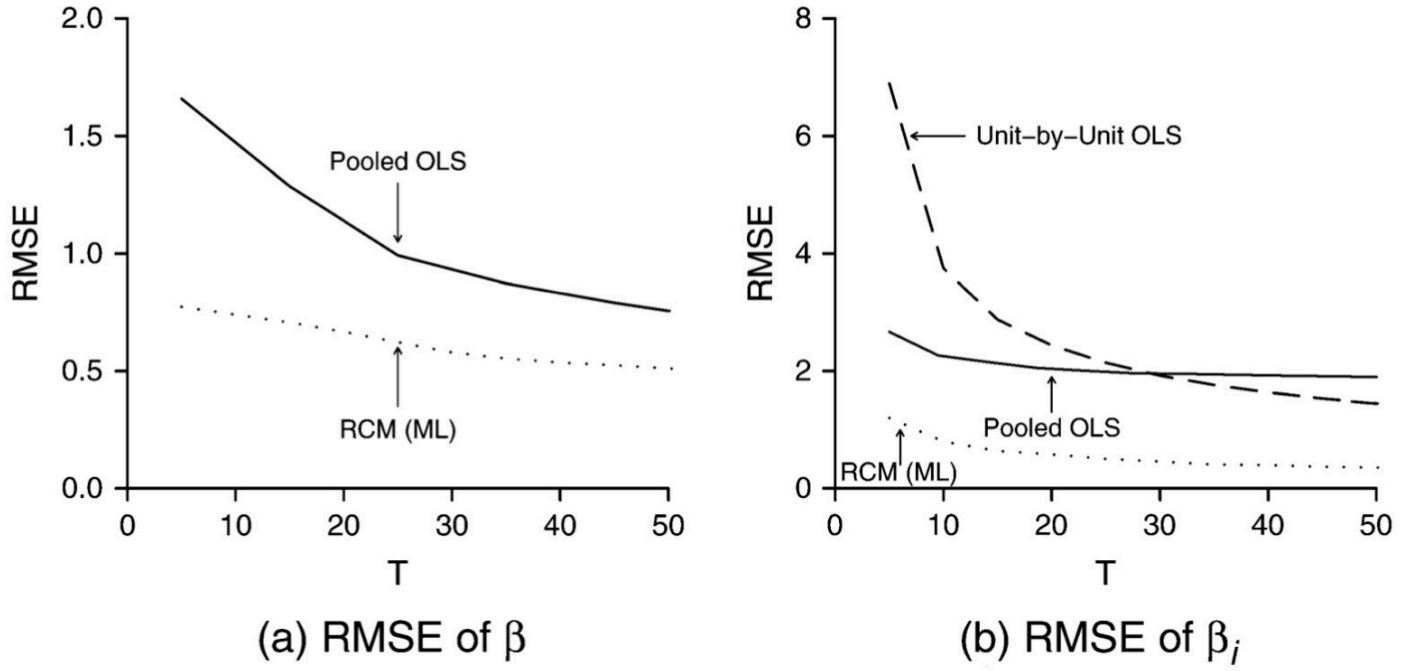


Fig. 1 Comparison of RMSE for RCM and OLS estimators of β and β_i as T varies from 5 to 50. For all runs of the experiment, $N = 20$, $\beta = 5$, $\gamma = 1.8$, $\sigma_\varepsilon^2 = 1$, and $\sigma_x^2 = 0.01$.

Beck, N., & Katz, J. N. (2007). Random coefficient models for time-series—cross-section data: Monte Carlo experiments. Political Analysis, 15(2), 182-195.

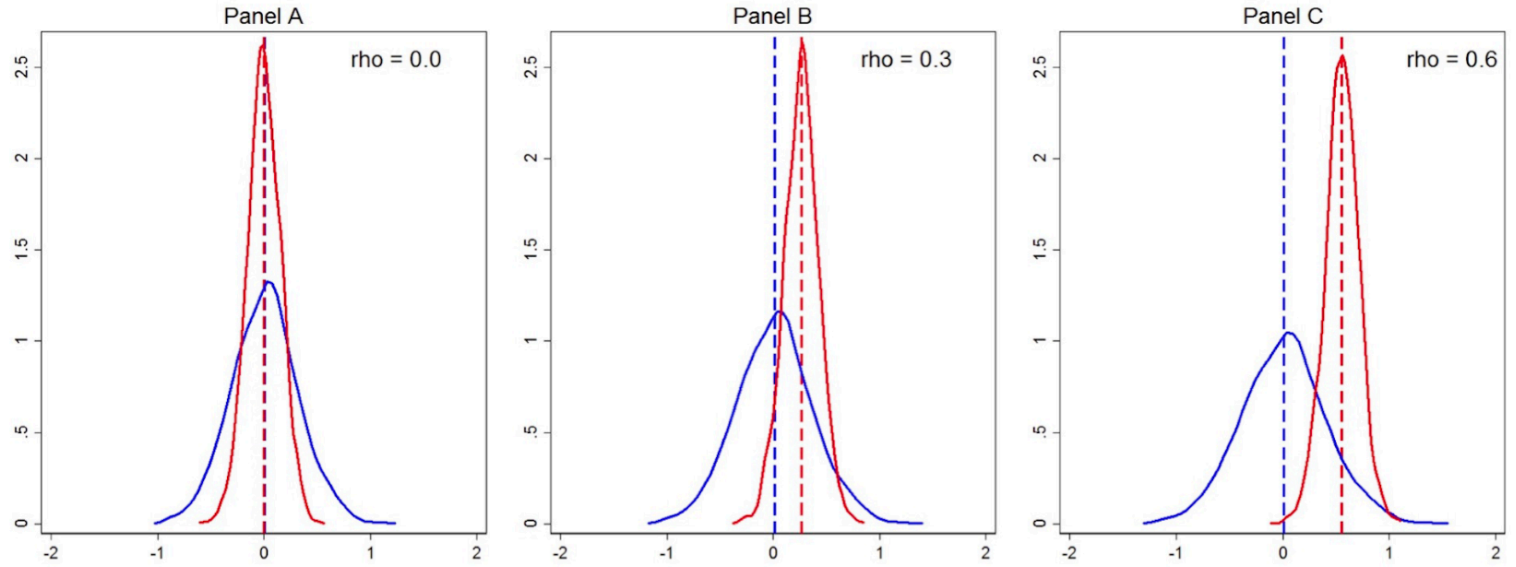
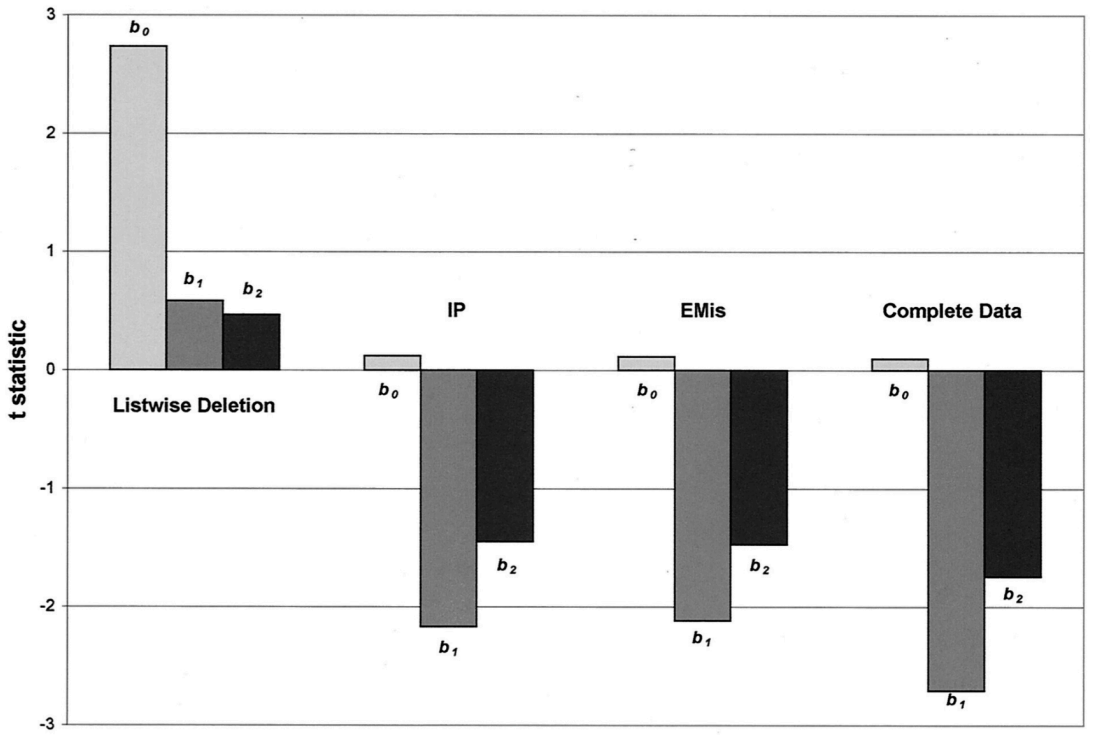


Figure 2. Distribution of errors. Red lines show the distribution of errors from RE estimation, while the blue lines show the distribution of errors from FE estimation. Each panel shows the correlation between the explanatory variable and the group-level effect set to a different value of ρ (0.0, 0.3, 0.6), increasing from left to right. Simulation based on the correct specification of a model with 50 groups and 10 observations per group. 50% of the variation of the outcome variable is explained by residuals, while only 10% of the variation in the explanatory variable is within-groups.
doi:10.1371/journal.pone.0110257.g002

Dieleman, J. L., & Templin, T. (2014). Random-effects, fixed-effects and the within-between specification for clustered data in observational health studies: a simulation study. PloS one, 9(10), e110257.

FIGURE 3. Monte Carlo Comparison of *t* Statistics



Note: *T* statistics are given for the constant (b_0) and the two regression coefficients (b_1, b_2) for the MAR-1 run in Figure 2. Listwise deletion gives the wrong results, whereas EMis and IP recover the relationships accurately.

King, G., Honaker, J., Joseph, A., & Scheve, K. (2001). Analyzing incomplete political science data: An alternative algorithm for multiple imputation. American political science review, 95(1), 49-69.

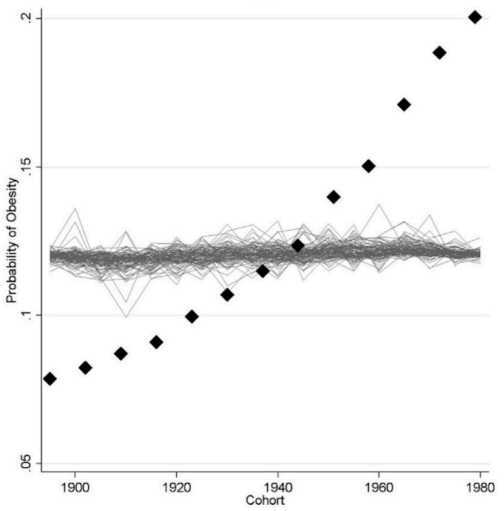
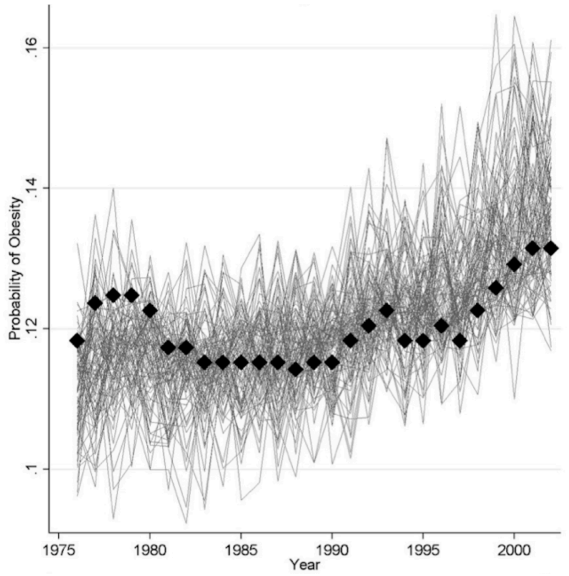
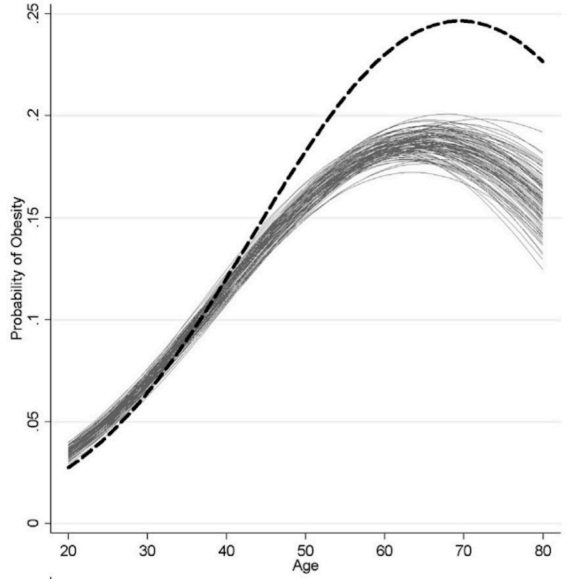
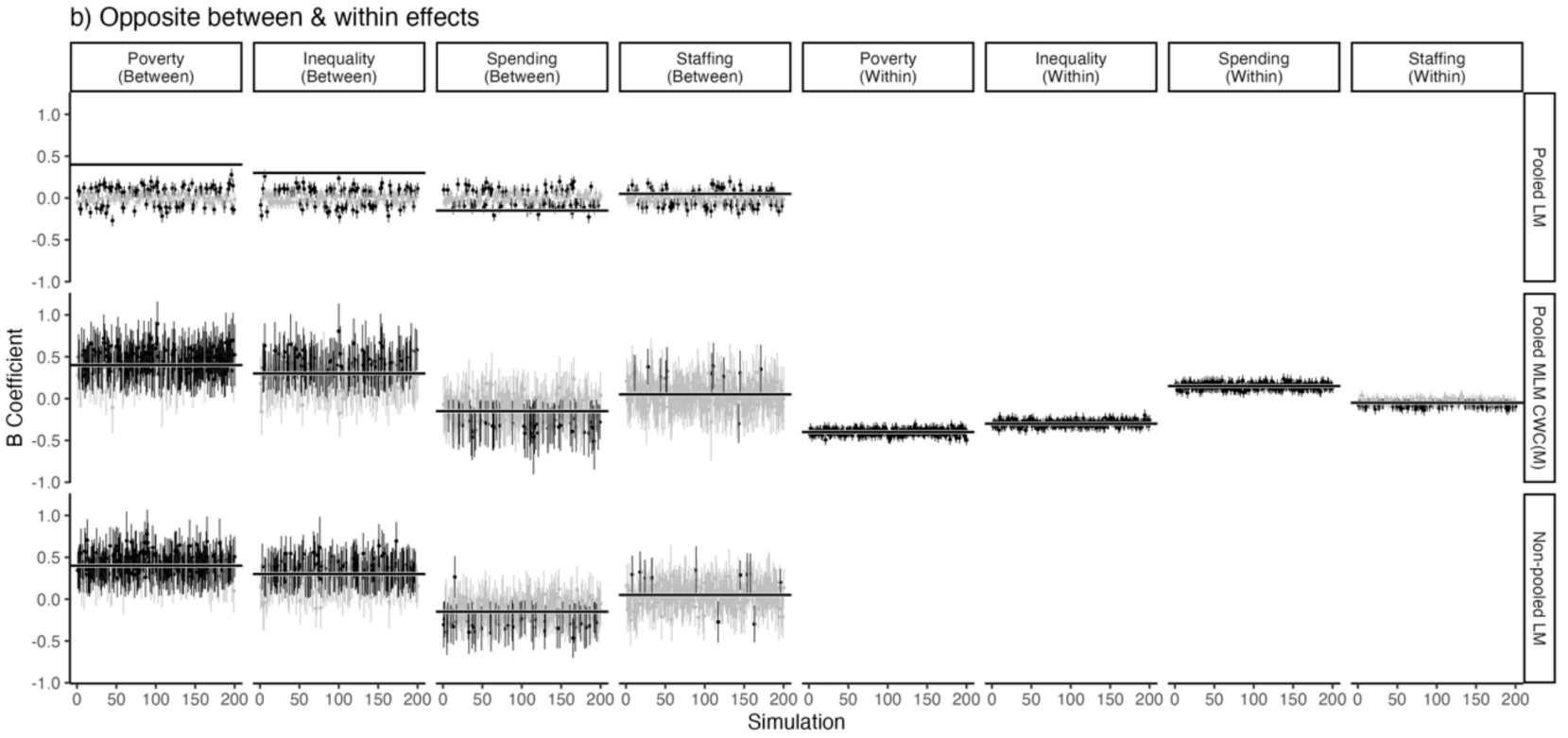


Fig. 1. The DGP (black) and results (grey) from fitting the HAPC model to 100 datasets generated as in Equation (1).





Bell, A., & Jones, K. (2015). Should age-period-cohort analysts accept innovation without scrutiny? A response to Reither, Masters, Yang, Powers, Zheng and Land. Social Science & Medicine, 128, 331-333.



Webb, C. (2023). Should we retire the Null Hypothesis Significance Test in (some) social policy research? Decisive versus descriptive approaches to statistical uncertainty in research on apparent populations with small or modest effects. Working Paper Presented at Social Policy Association Conference 2023

Table 1. Coverage of the 95% confidence interval by number of groups (0.9225 < C.I. < 0.9747, * = significant at 0.001).

		E0	U0	U1
Number of groups	30	0.9428	0.9104*	0.9120*
	50	0.9438	0.9261	0.9282
	100	0.9514	0.9404	0.9426

Maas, C. J., & Hox, J. J. (2004). Robustness issues in multilevel regression analysis. Statistica Neerlandica, 58(2), 127-137.

Table 1
Relative bias of the parameter estimates chi-squared residuals[a] ($\alpha = 0.001$)

	Relative bias	Population value	Estimate	*p*-value
Intercept	1.002	1.00	1.002	1.000
X	0.990	0.30	0.297	1.000
Z	0.997	0.30	0.299	1.000
XZ	1.002	0.30	0.301	1.000
E_0	0.984	0.50	0.492	0.001*
U_0	1.116	0.056	0.063	0.005
U_1	1.035	0.056	0.058	1.000

[a]Uniform and Laplace residuals: no difference from population value.
*sign.

Maas, C. J., & Hox, J. J. (2004). The influence of violations of assumptions on multilevel parameter estimates and their standard errors. Computational statistics & data analysis, 46(3), 427-440.

Table 1. Noncoverage of the 95% Confidence Interval by Number of Groups

Parameter	Number of groups			*p*-value[a]
	30	50	100	
U0	.089	.074	.060	.0000
U1	.088	.072	.057	.0000
E	.058	.056	.049	.0101
INT	.064	.057	.053	.0057
X	.060	.057	.050	.0058
Z	.052	.051	.050	.9205
XZ	.056	.052	.050	.2187

[a] *p*-value for the effect of number of groups on mean parameter estimate.

Maas, C. J., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. Methodology, 1(3), 86-92.

Presenting results

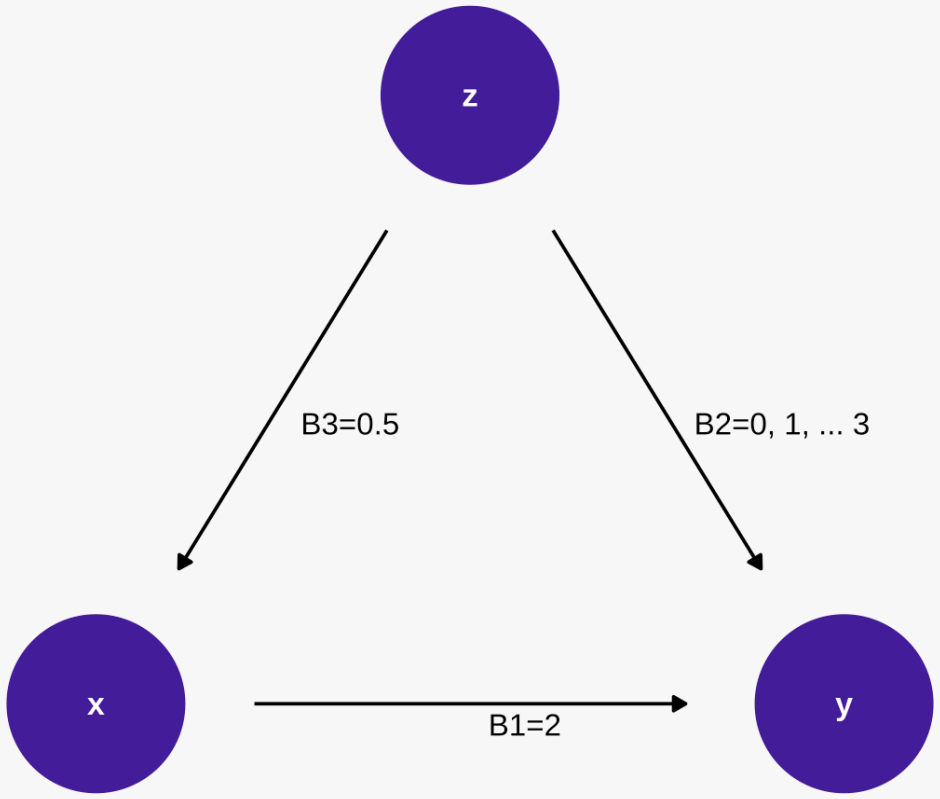
- It would be nice if someone had decided on some rules for simulations, and stuck to them.
- But they haven't.
- Which can be a really positive thing — you can be creative in how you display your results to do what you want the results to show.

Presenting results

- It would be nice if someone had decided on some rules for simulations, and stuck to them.
- But they haven't.
- Which can be a really positive thing — you can be creative in how you display your results to do what you want the results to show.

Exercise 2...

- In the next exercise we'll be building up to a full simulation study: deciding which value we want to vary, writing a loop to run many simulations for each value we are varying, calculating the relevant quantities of interest, and creating some simple tables and visualisations.



Exercise 2...

- In the next exercise we'll be building up to a full simulation study: deciding which value we want to vary, writing a loop to run many simulations for each value we are varying, calculating the relevant quantities of interest, and creating some simple tables and visualisations.

Final thoughts…

What could you do a simulation on?

- We've looked at the effect of confounders / colliders
- Could look at sample size (effectively do power analysis through simulation)
- What happens when you miss data's multilevel structure?
- What do different missing data patterns do?
- Different model choices (Bayesian vs Frequentist, Poisson vs Negative Binomial, Fixed vs random effects, etc)

Limits of simulations

- Simulation studies don't tell you for sure that models work well in real life
- Complex models will often struggle with messy real life data, but work OK with data designed as the model expects
- Can sometimes try to simulate messiness

Judging simulation studies

- What are the situations being compared? Are they realistic?
- Do they apply to your dataset? Might things be different for you?
- Do they test all likely eventualities in the DGP (or do they miss something important)?
- There are lots of different ways to do simulations. We've only shown you one way, but others work too!

Tips

- Be aware that some functions in some software will simulate "empirical" or "exact" simulated distributions rather than random simulated distributions (i.e. a simulated dataset that has values that return exactly the parameters specified), e.g. `corr2data` in Stata and with the `empirical = TRUE` argument in `faux` in R.
- More complex simulation (e.g. multilevel data, multivariate data, time-series) might require additional packages, e.g. `faux`, `MASS` (in R) `simsum` in Stata, but not always (often there's a logic order to follow)

Further reading...

- Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. Statistics in medicine, 38(11), 2074-2102.
- White, I. R., Pham, T. M., Quartagno, M., & Morris, T. P. (2024). How to check a simulation study. International Journal of Epidemiology, 53(1), dyad134.
- Måns Thulin (2025). *The role of simulation in modern statistics* https://www.modernstatisticswithr.com for more examples in R.
- Nick Huntington-Klein (2025). *Simulation* in *The Effect* https://theeffectbook.net/ch-Simulation.html (also in R)
- Adkins & Gade (2012) *'Monte Carlo Experiments Using Stata: A Primer with Examples'* in Terrell & Millimet (2012) Advances in Econometrics.