

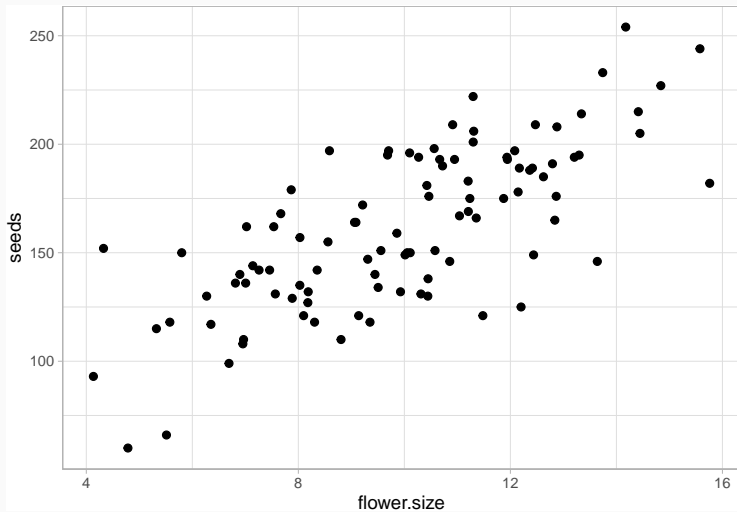
From causal salads to causal inference

Francisco Rodríguez-Sánchez

<https://frodriguezsanchez.net>



Larger flowers produce more seeds



Larger flowers produce more seeds

```
lm(seeds ~ flower.size)
```

<i>Variable</i>	<i>Beta</i>	<i>SE</i>	<i>p.value</i>
(Intercept)	57	10.1	<0.001
flower.size	11	0.978	<0.001

Abbreviations: SE = Standard Error, NA

Does flower size

really **cause**

increased seed production?

Shall we select plants with large flowers
to increase seed production?

Shall we select plants with large flowers
to increase seed production?

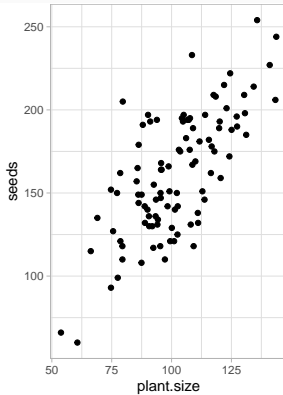
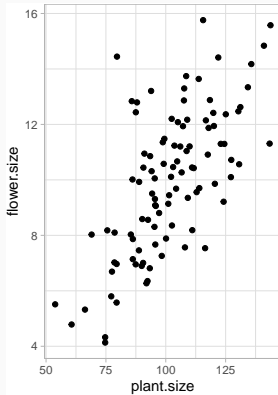
We tried but didn't get the expected benefits

Maybe **large plants** (e.g. growing on better soil)
have **large flowers** AND produce **more seeds**?

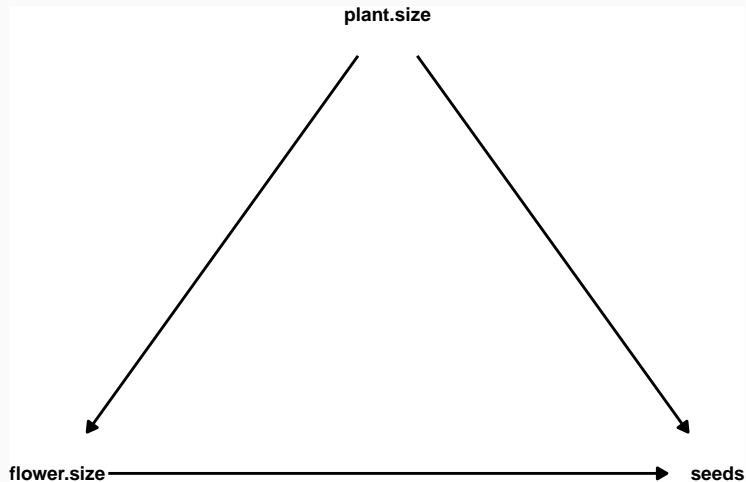


Maybe **large plants** (e.g. growing on better soil)

have **large flowers** AND produce **more seeds**?



Maybe plant size is a **CONFOUNDER**?



Adjusting for plant size (confounding)

```
lm(seeds ~ flower.size + plant.size)
```

<i>Variable</i>	<i>Beta</i>	<i>SE</i>	<i>p.value</i>
(Intercept)	12	12.9	0.4
flower.size	6.6	1.18	<0.001
plant.size	0.82	0.168	<0.001

Abbreviations: SE = Standard Error, NA

Including pollinators (bees)



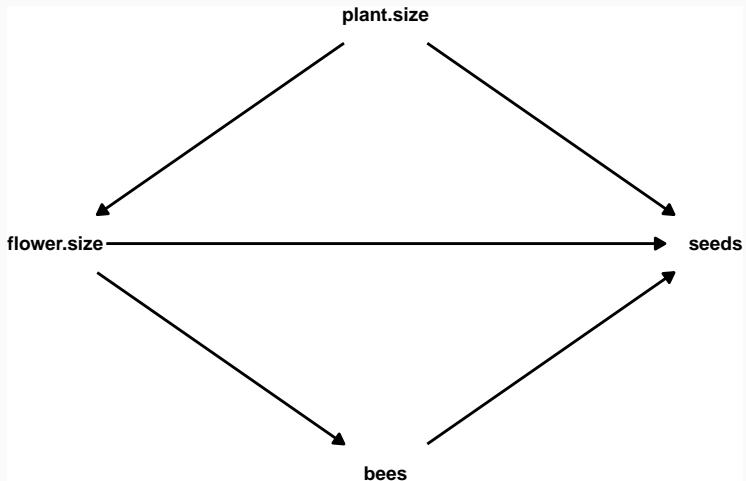
Including pollinators (bees)

```
lm(seeds ~ flower.size + plant.size + bees)
```

<i>Variable</i>	<i>Beta</i>	<i>SE</i>	<i>p.value</i>
(Intercept)	5.2	12.1	0.7
flower.size	2.1	1.56	0.2
plant.size	0.90	0.157	<0.001
bees	8.8	2.14	<0.001

Abbreviations: SE = Standard Error, NA

Pollinators are a **MEDIATOR**



Including beetles
(pollen & seed predators)

Including beetles (pollen & seed predators)

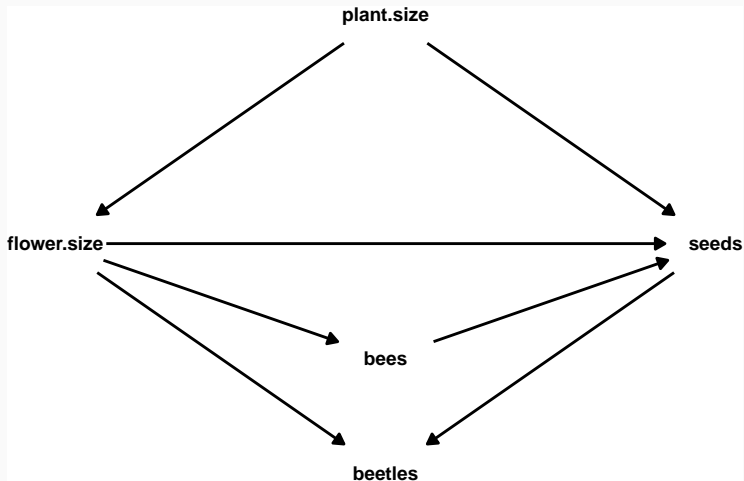
```
lm(seeds ~ flower.size + plant.size + bees +  
beetles)
```

<i>Variable</i>	<i>Beta</i>	<i>SE</i>	<i>p.value</i>
(Intercept)	-11	8.67	0.2
flower.size	-3.8	1.25	0.003
plant.size	0.47	0.118	<0.001
bees	4.8	1.56	0.003
beetles	5.2	0.529	<0.001

Abbreviations: SE = Standard Error, NA

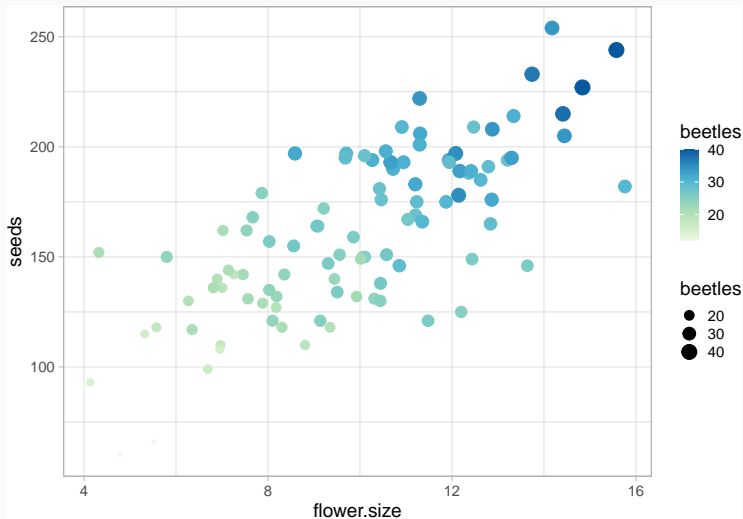
Now flower.size has negative coefficient!!

Beetles are a COLLIDER

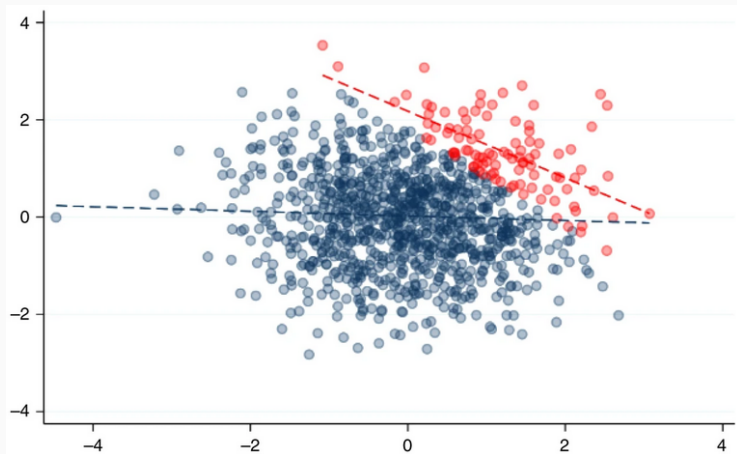


Collider bias

Colliders induce non-causal negative relation between treatment (*flower.size*) and outcome (*seeds*)



Colliders induce non-causal negative relation between treatment and outcome



Griffith et al 2020

Number of children is significant negative predictor of marital satisfaction

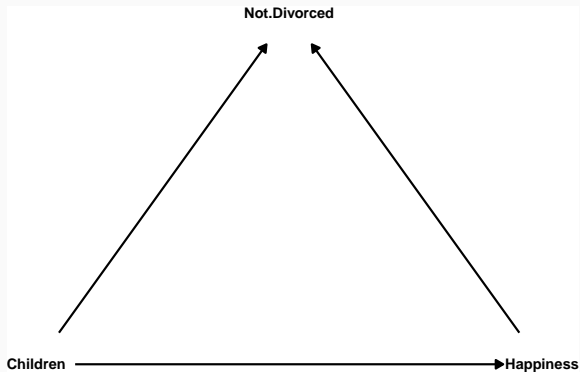
The more children, the more unhappy couples are



There is collider/selection bias

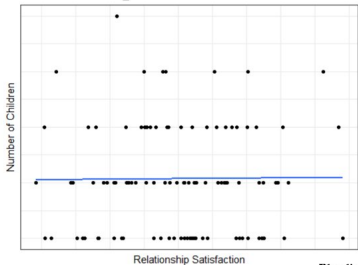
Selection bias: data only include married couples (not divorced)

And couples with children or happy are less likely to get divorced



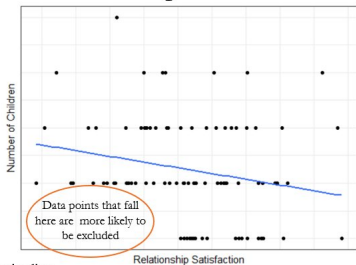
Collider induces negative correlation between number of children and happiness

Population Data



Blue line = regression line

Sample Data

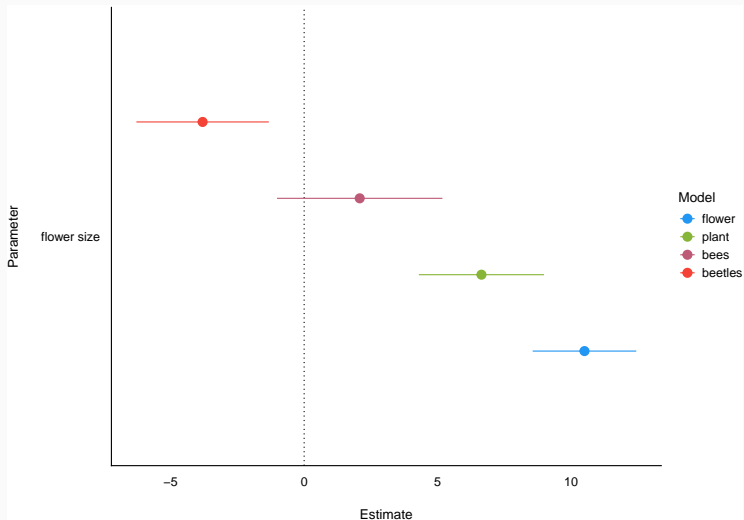


@ AnnaWysocki3

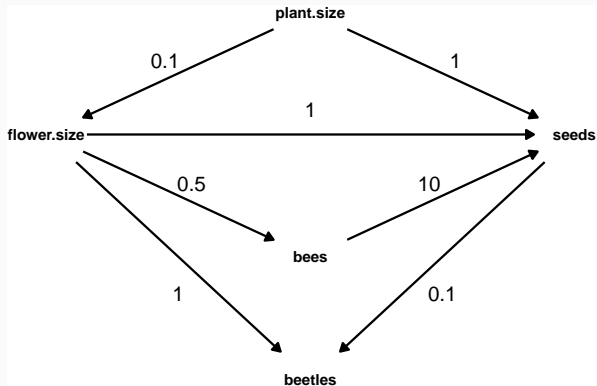
Recapitulating...

What is the real causal effect of flower size?

What is the real causal effect of flower size?

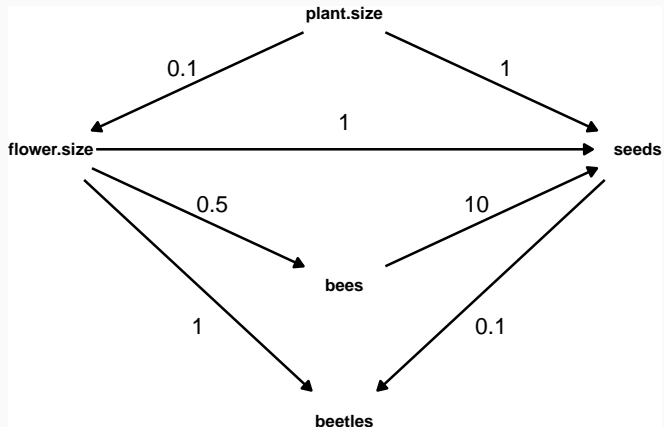


What is the real causal effect of flower size?



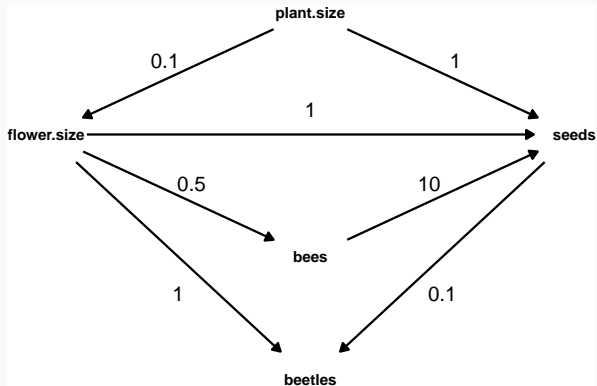
Variable	Beta	SE	p.value
(Intercept)	-11	8.67	0.2
flower.size	-3.8	1.25	0.003
plant.size	0.47	0.118	<0.001
bees	4.8	1.56	0.003
beetles	5.2	0.529	<0.001

What is the real causal effect of flower size?



Avoid **COLLIDERS** -> collider/selection bias

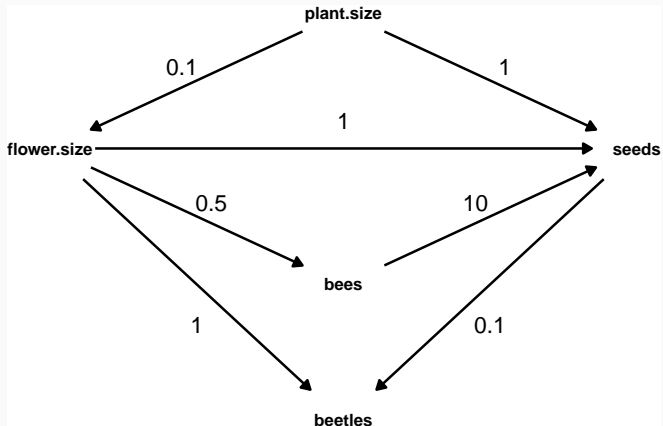
What is the real causal effect of flower size?



<i>Variable</i>	<i>Beta</i>	<i>SE</i>	<i>p.value</i>
(Intercept)	5.2	12.1	0.7
flower.size	2.1	1.56	0.2
plant.size	0.90	0.157	<0.001
bees	8.8	2.14	<0.001

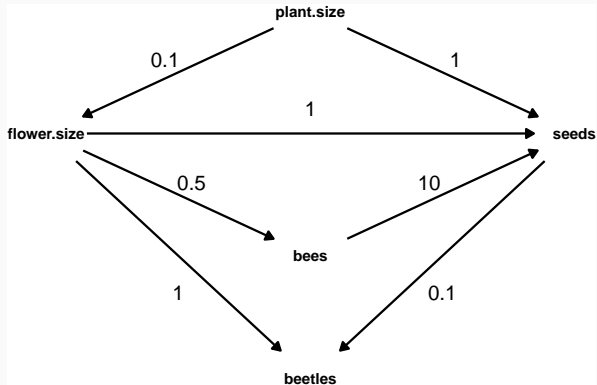
Abbreviations: SE

What is the real causal effect of flower size?



MEDIATORS split **total effect** into **direct** and **indirect** effects
(overcontrol bias)

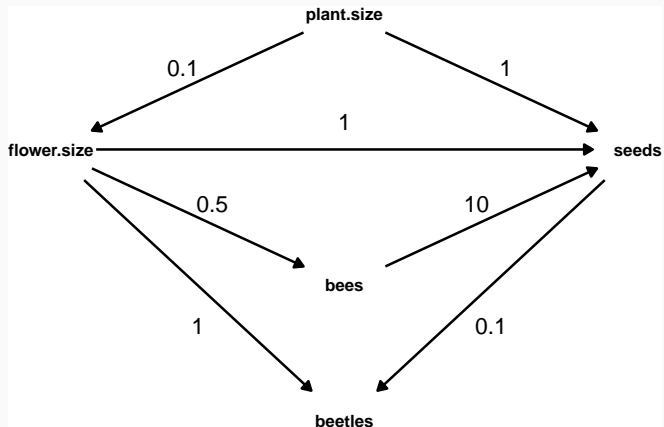
What is the real causal effect of flower size?



<i>Variable</i>	<i>Beta</i>	<i>SE</i>	<i>p.value</i>
(Intercept)	12	12.9	0.4
flower.size	6.6	1.18	<0.001
plant.size	0.82	0.168	<0.001

Abbreviations: SE
= Standard Error,

What is the real causal effect of flower size?



Include **CONFOUNDERS** to avoid 'omitted variable bias'

(use **backdoor criterion**)

Tools to identify correct causal structure

<https://daggity.net>

Variable

flower.size

- exposure
- outcome
- adjusted
- selected
- unobserved

View mode

- normal
- moral graph
- correlation graph
- equivalence class

Effect analysis

- atomic direct effects

Diagram style

- classic
- SEM-like

Coloring

- causal paths
- biasing paths
- ancestral structure

Legend

- exposure
- outcome
- ancestor of exposure
- ancestor of outcome
- ancestor of exposure and outcome
- adjusted variable
- unobserved (latent)
- other variable
- causal path

Model | Examples | How to ... | Layout | Help

```
graph TD; plant_size((plant.size)) --> flower_size((flower.size)); plant_size --> seeds((seeds)); flower_size --> seeds; flower_size --> beetles((beetles)); seeds --> beetles; beetles --> beetles; bees((bees)) --> beetles;
```

Causal effect identification

Adjustment (total effect)

Exposure: flower.size
Outcome: seeds
Selected: beetles
Adjusted: plant.size
Incorrectly adjusted.

Testable implications

The model implies the following conditional independences:

- plant.size \perp beetles | flower.size, seeds
- bees \perp beetles | flower.size, seeds

Model code

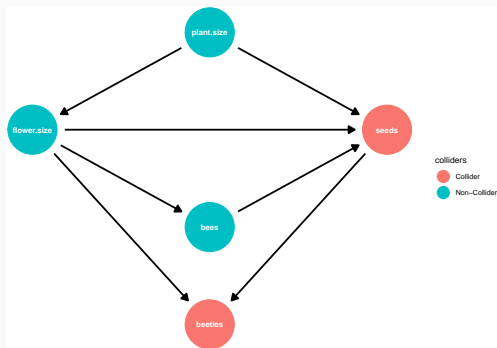
```
dag {
  bb="0,0,1,1"
  bees [pos="0.464,0.568"]
  beetles
  [selected,pos="0.467,0.812"]
  flower.size
  [exposure,pos="0.081,0.389"]
  plant.size
}
```

Summary

exposure(s) **flower.size**
outcome(s) **seeds**
covariates **3**
causal paths **2**

Tools to identify correct causal structure

```
dagify(  
  seeds ~ plant.size + flower.size + bees,  
  flower.size ~ plant.size,  
  bees ~ flower.size,  
  beetles ~ flower.size + seeds,  
  coords = coords  
) |>  
ggdag_collider(size = 2) + theme_dag_blank()
```



Tools to identify correct causal structure

```
library(easystats)
dag <- check_dag(
  seeds ~ flower.size + plant.size + bees,
  flower.size ~ plant.size,
  bees ~ flower.size,
  beetles ~ flower.size + seeds,
  outcome = "seeds",
  exposure = "flower.size",
  adjusted = c("plant.size", "bees", "beetles")
)
dag
```

```
# Check for correct adjustment sets
- Outcome: seeds
- Exposure: flower.size
- Adjustments: bees, beetles and plant.size
- Collider: beetles
```

Identification of direct and total effects

Incorrectly adjusted!

Your model adjusts for a potential collider. To estimate the direct and total effect, do not adjust for 'beetles'

Causal salads

Causal salads

You put everything into a regression equation, toss with some creative story-telling, and hope the reviewers eat it

R. McElreath



{Jerry Pank}

*Throwing predictor variables into a statistical model
hoping this will improve the analysis is a dreadful idea*

Jan Vanhove

Predictive criteria don't help for
causal inference

Predictive criteria don't help to choose correct causal model

Making good predictions doesn't require accurate causal model

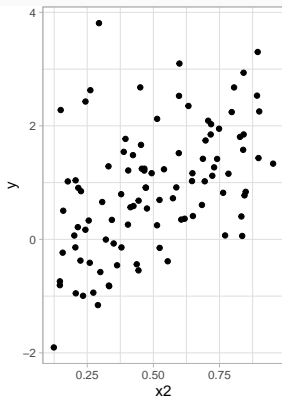
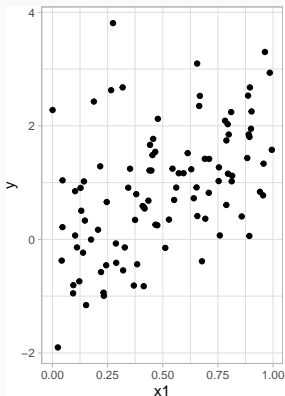
Model	AIC	R2
m.flower	933.3	0.5
m.flower.plant	913.2	0.6
m.flower.plant.bees	899.1	0.7
m.flower.plant.bees.beetles	829.9	0.8

'Best model' (based on AIC or R2) not good for causal inference

Simpler (best) model provides biased causal estimates

Simulate response depending on two correlated variables (Hartig 2022)

```
x1 = runif(100)
x2 = 0.8*x1 + 0.2*runif(100)
y = x1 + x2 + rnorm(100)
```



Simpler (best) model provides biased causal estimates

Simulate response depending on two correlated variables (Hartig 2022)

```
fullmodel = lm(y ~ x1 + x2)
```

Call:

```
lm(formula = y ~ x1 + x2)
```

Residuals:

Min	1Q	Median	3Q	Max
-1.8994	-0.6821	-0.1086	0.5749	3.3663

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.1408	0.2862	-0.492	0.624
x1	1.2158	1.5037	0.809	0.421
x2	0.8518	1.8674	0.456	0.649

Residual standard error: 0.9765 on 97 degrees of freedom

Multiple R-squared: 0.237, Adjusted R-squared: 0.2212

F-statistic: 15.06 on 2 and 97 DF, p-value: 2.009e-06

Simpler (best) model provides biased causal estimates

```
simplemodel = MASS::stepAIC(fullmodel, trace = 0)
```

Call:

```
lm(formula = y ~ x1)
```

Residuals:

Min	1Q	Median	3Q	Max
-1.9047	-0.6292	-0.1019	0.6077	3.3394

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.04633	0.19670	-0.236	0.814
x1	1.88350	0.34295	5.492	3.13e-07 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9725 on 98 degrees of freedom

Multiple R-squared: 0.2353, Adjusted R-squared: 0.2275

F-statistic: 30.16 on 1 and 98 DF, p-value: 3.134e-07

Automated model selection (dredge)

Simulating data with 10 random predictors

```
dat <- data.frame(y = rnorm(100),  
                  x = matrix(runif(1000), ncol = 10))
```

y	x.1	x.2	x.3	x.4	x.5	x.6	x.7	x.8	x.9	x.10
-0.1	0.6	0.6	0.3	0.8	0.2	0.0	0.4	0.4	0.3	0.2
0.8	0.4	0.4	0.9	0.2	0.5	0.1	0.6	0.2	0.0	0.0
-0.5	0.0	0.3	0.4	0.3	0.1	0.1	0.9	0.9	0.5	0.8
-0.6	0.7	0.7	0.4	0.5	0.2	0.7	0.7	0.8	0.5	0.3
0.7	0.0	0.6	0.9	0.1	0.2	0.4	0.8	0.6	0.6	0.1
-0.1	0.4	0.2	0.9	0.4	0.6	0.5	0.9	0.1	0.8	0.8

Automated model selection

Running `MuMIn::dredge` with 10 random predictors

```
full.model <- lm(y ~ ., data = dat)
dd <- MuMIn::dredge(full.model)
```

Best model:

Parameter	Coefficient	SE	p
(Intercept)	-1.50	0.36	0.00
x.2	0.78	0.36	0.03
x.5	0.59	0.32	0.07
x.6	0.61	0.35	0.09
x.9	0.87	0.34	0.01

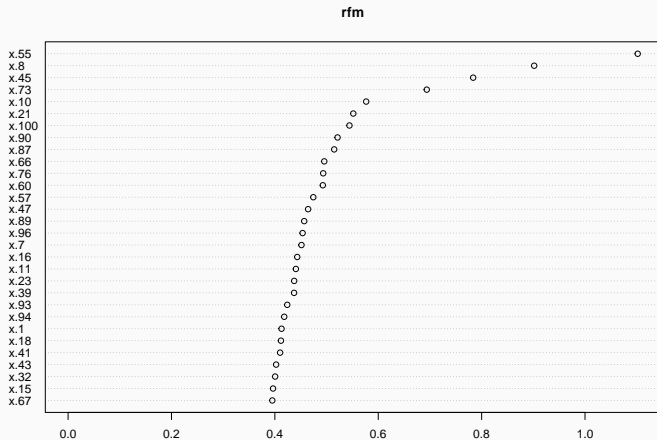
“Let the computer find out” is a poor strategy and usually reflects the fact that the researcher did not bother to think clearly about the problem of interest and its scientific setting

Burnham and Anderson 2002

Variable importance in machine learning

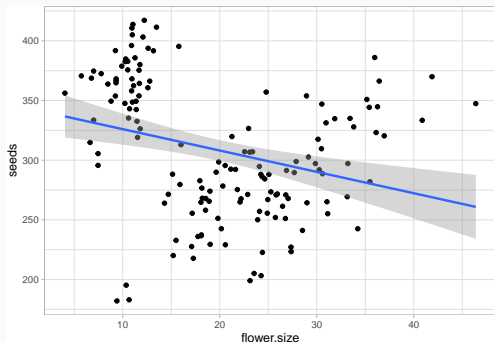
Random forest on 100 random predictors

```
dat <- data.frame(x = matrix(runif(50000), ncol = 100), y = runif(500))  
rfm <- randomForest::randomForest(y ~ ., data = dat)  
varImpPlot(rfm)
```



Simpson's paradox as a causal problem

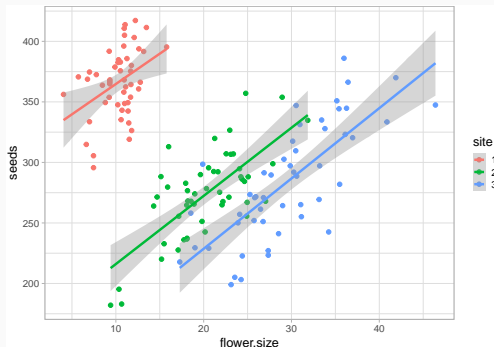
Simpson's paradox



<i>Variable</i>	<i>Beta</i>	<i>SE</i>	<i>p.value</i>
(Intercept)	344	10.7	<0.001
flower.size	-1.8	0.486	<0.001

Abbreviations: SE = Standard Error, NA

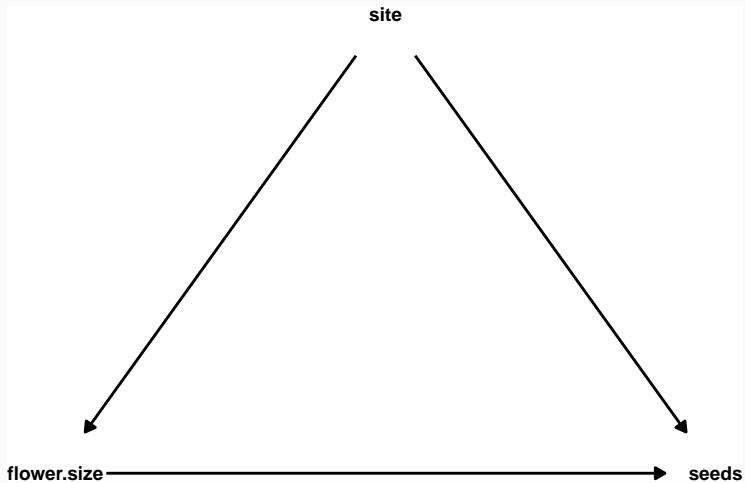
Simpson's paradox



Variable	Beta	SE	p.value
(Intercept)	308	6.50	<0.001
flower.size	5.7	0.500	<0.001
site			
1	—	—	
2	-149	7.50	<0.001
3	102	111	<0.001

Simpson's paradox

Site is a confounder!



From causal salads to causal inference

Causal interpretation requires **external knowledge**

To estimate causal effects accurately we require more information than can be gleaned from statistical tools alone

D'Agostino et al

Causal interpretation requires external knowledge

To estimate causal effects accurately we require more information than can be gleaned from statistical tools alone

D'Agostino et al

No amount of data reliably turns salad into sense

R. McElreath

From causal salad to causal inference

- Draw the **causal graph** (DAG) beforehand

From causal salad to causal inference

- Draw the **causal graph** (DAG) beforehand
- Control for **confounders**

From causal salad to causal inference

- Draw the **causal graph** (DAG) beforehand
- Control for **confounders**
- Avoid conditioning on **post-treatment variables**

From causal salad to causal inference

- Draw the **causal graph** (DAG) beforehand
- Control for **confounders**
- Avoid conditioning on **post-treatment variables**
 - Treatment -> Covariate -> Outcome

From causal salad to causal inference

- Draw the **causal graph** (DAG) beforehand
- Control for **confounders**
- Avoid conditioning on **post-treatment variables**
 - Treatment -> Covariate -> Outcome
- Beware of **collider bias**

From causal salad to causal inference

- Draw the **causal graph** (DAG) beforehand
- Control for **confounders**
- Avoid conditioning on **post-treatment variables**
 - Treatment -> Covariate -> Outcome
- Beware of **collider bias**
- **Predictive criteria** not fit for causal inference

To learn more

Suchinta Arif's papers

McElreath's workshop on causal inference

Byrnes & Dee 2024

<https://www.r-causal.org>

<https://theeffectbook.net>

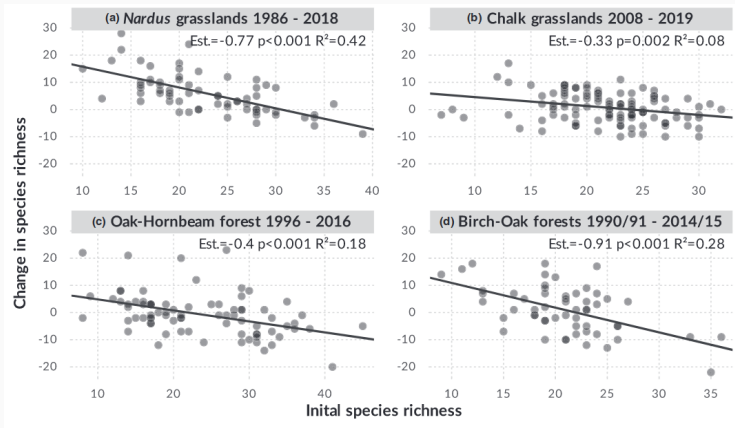
Regression to the mean

Francisco Rodríguez-Sánchez

<https://frodriguezsanchez.net>

The most biodiverse sites are losing more species

WHY??



Mazalla & Diekmann 2022

Most biodiverse sites are losing more species. Why?

- Stronger competition

Most biodiverse sites are losing more species. Why?

- Stronger competition
- Humans destroying most species-rich sites

Most biodiverse sites are losing more species. Why?

- Stronger competition
- Humans destroying most species-rich sites
- Establishment of new species favoured in poor sites

Most biodiverse sites are losing more species. Why?

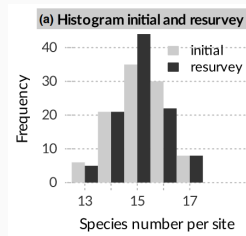
- Stronger competition
- Humans destroying most species-rich sites
- Establishment of new species favoured in poor sites

Most biodiverse sites are losing more species. Why?

- Stronger competition
- Humans destroying most species-rich sites
- Establishment of new species favoured in poor sites
- No ecological cause, but stochastic variation (**regression to the mean**)

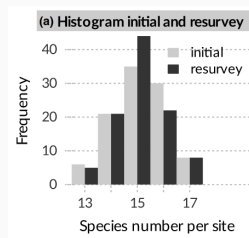
A simulation for 100 sites

- Simulate initial number of species:



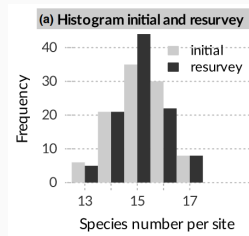
A simulation for 100 sites

- Simulate initial number of species:
 - `rnorm(n = 100, mean = 15, sd = 1)`



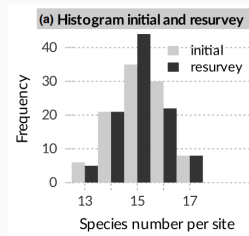
A simulation for 100 sites

- Simulate initial number of species:
 - `rnorm(n = 100, mean = 15, sd = 1)`
- Simulate number of species at resurvey:



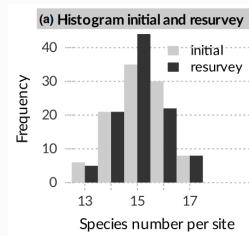
A simulation for 100 sites

- Simulate initial number of species:
 - `rnorm(n = 100, mean = 15, sd = 1)`
- Simulate number of species at resurvey:
 - `rnorm(n = 100, mean = 15, sd = 1)`



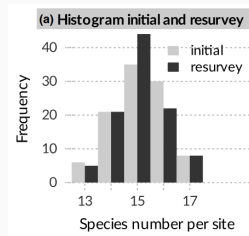
A simulation for 100 sites

- Simulate initial number of species:
 - `rnorm(n = 100, mean = 15, sd = 1)`
- Simulate number of species at resurvey:
 - `rnorm(n = 100, mean = 15, sd = 1)`
- No real change at all!



A simulation for 100 sites

- Simulate initial number of species:
 - `rnorm(n = 100, mean = 15, sd = 1)`
- Simulate number of species at resurvey:
 - `rnorm(n = 100, mean = 15, sd = 1)`
- **No real change at all!**
- (only stochastic variation)

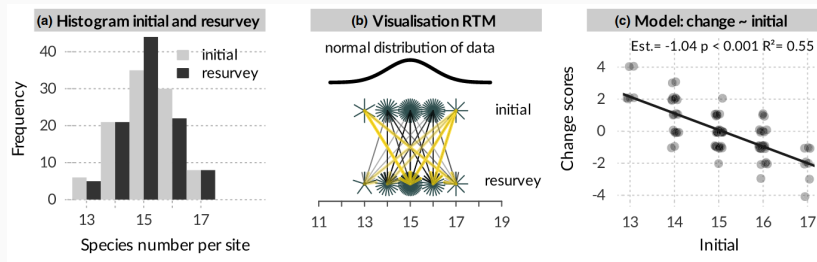


Regression to the mean

Species-rich sites lose more species

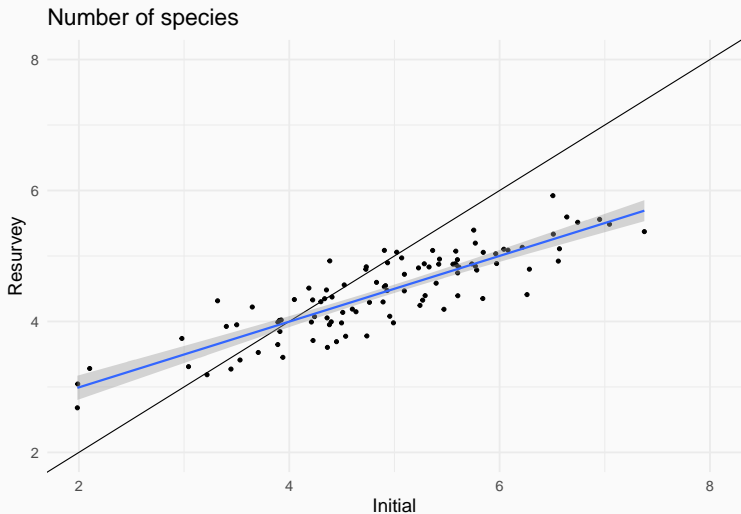
Species-poor sites gain more species

Negative trend against baseline



Mazalla & Diekmann 2022

Whenever two sets of measurements are not perfectly correlated there will be regression towards the mean



What to do?

- Model outcome ~ baseline

What to do?

- Model outcome ~ baseline
- If modelling Change, include baseline as predictor

- [Mazalla & Diekmann 2022](#)

To learn more

- Mazalla & Diekmann 2022
- Kelly & Price 2005