Three C's of causal consideration: confounding, collinearity and colliders

A useful talk on deciding about including or excluding variables in statistical models based on causal relationships or high levels of correlation.

Ben Maslen

Stats Central

Mark Wainwright Analytical Centre

May 23, 2019

Causal consideration

- Correlation $\neq$ causation
- Reasons for correlation
- Confounding
- Colliders
- Collinearity
- Take home tips

Correlation $\neq$ causation

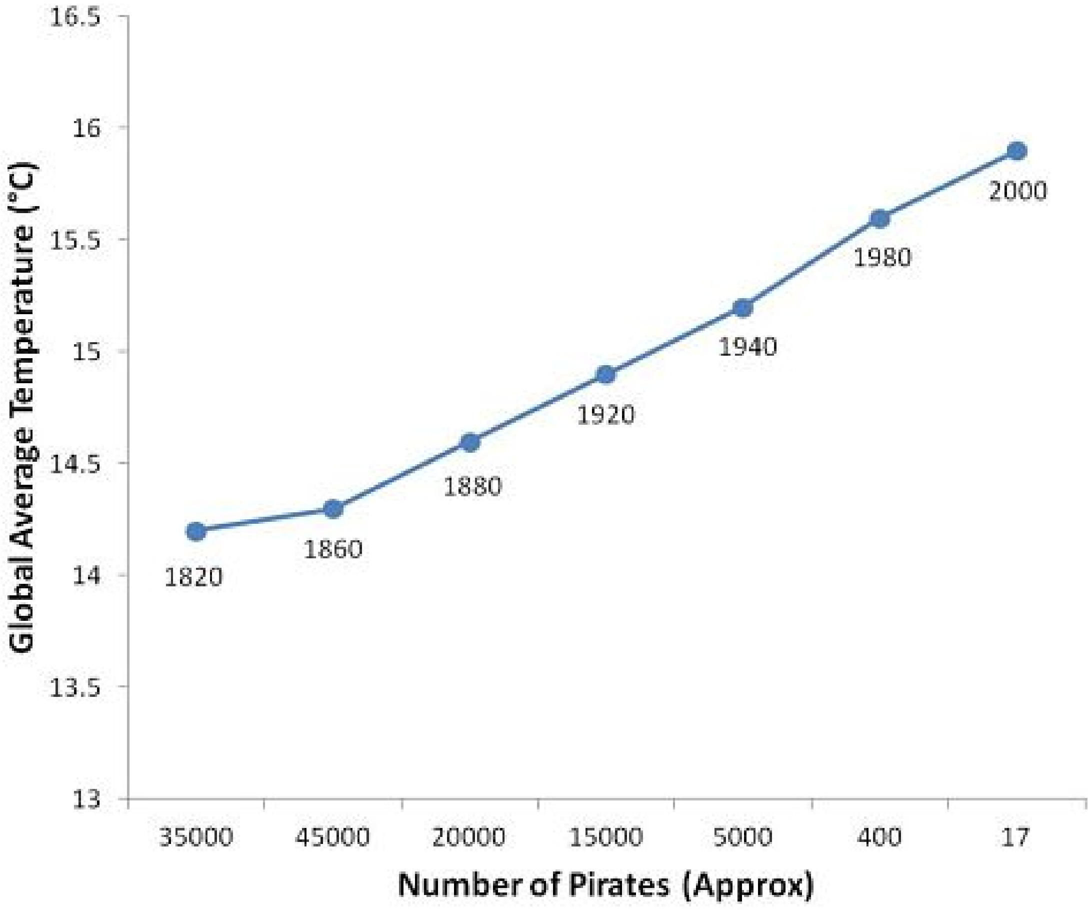
Correlation implies two variables are associated with eachother.

Causation implies that a change to one variable (the cause), creates a change to another variable (the effect).

Causation $\Rightarrow$ correlation, however correlation $\not\Rightarrow$ causation.

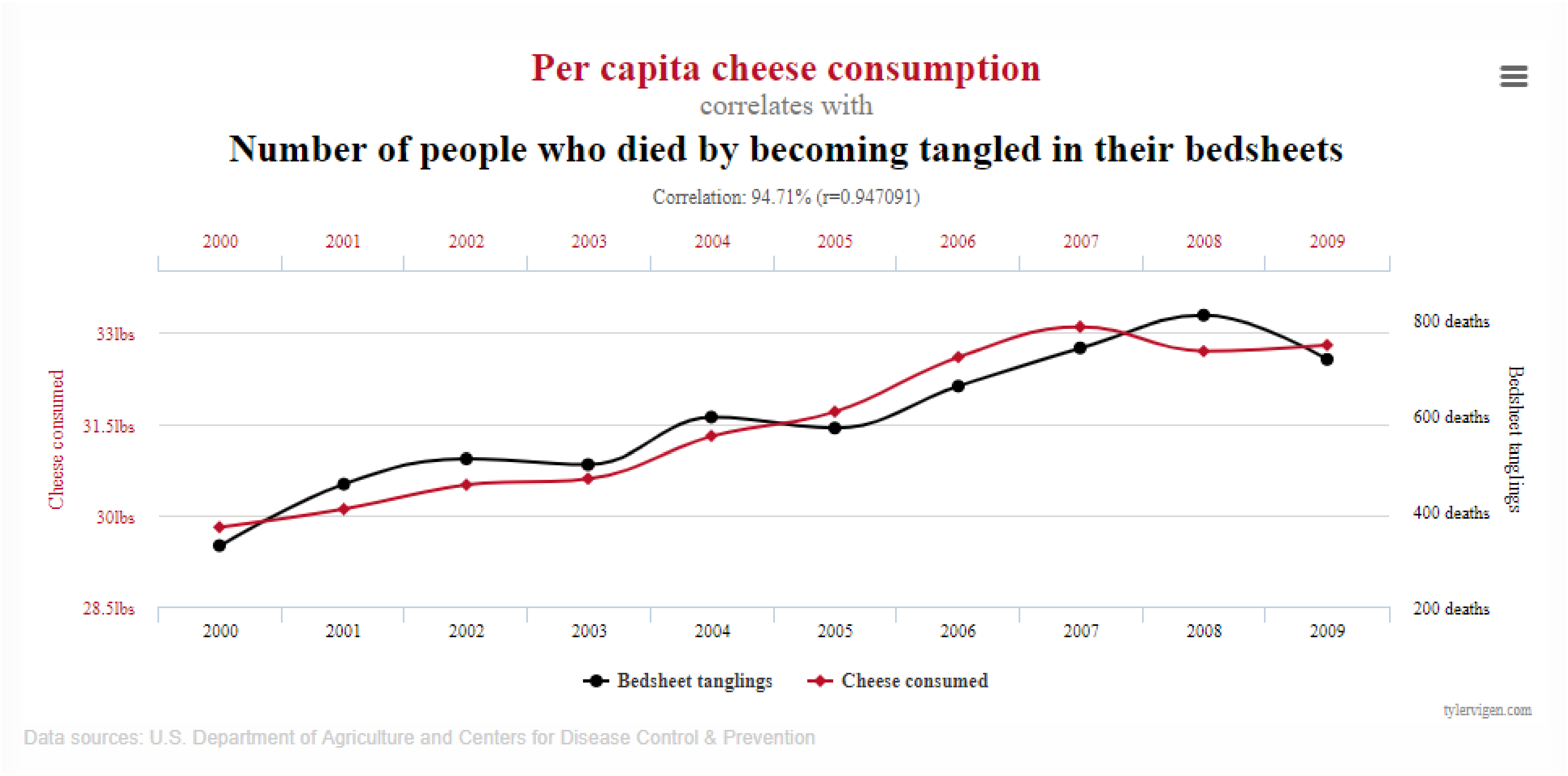
The false assumption that correlation = causation, particularly when events precede eachother, is called the post-hoc fallacy.

Correlation $\neq$ causation e.g.



- Truth, Lies & Statistics: How to Lie with Statistics

Correlation $\neq$ causation e.g.



Per capita cheese consumption

correlates with

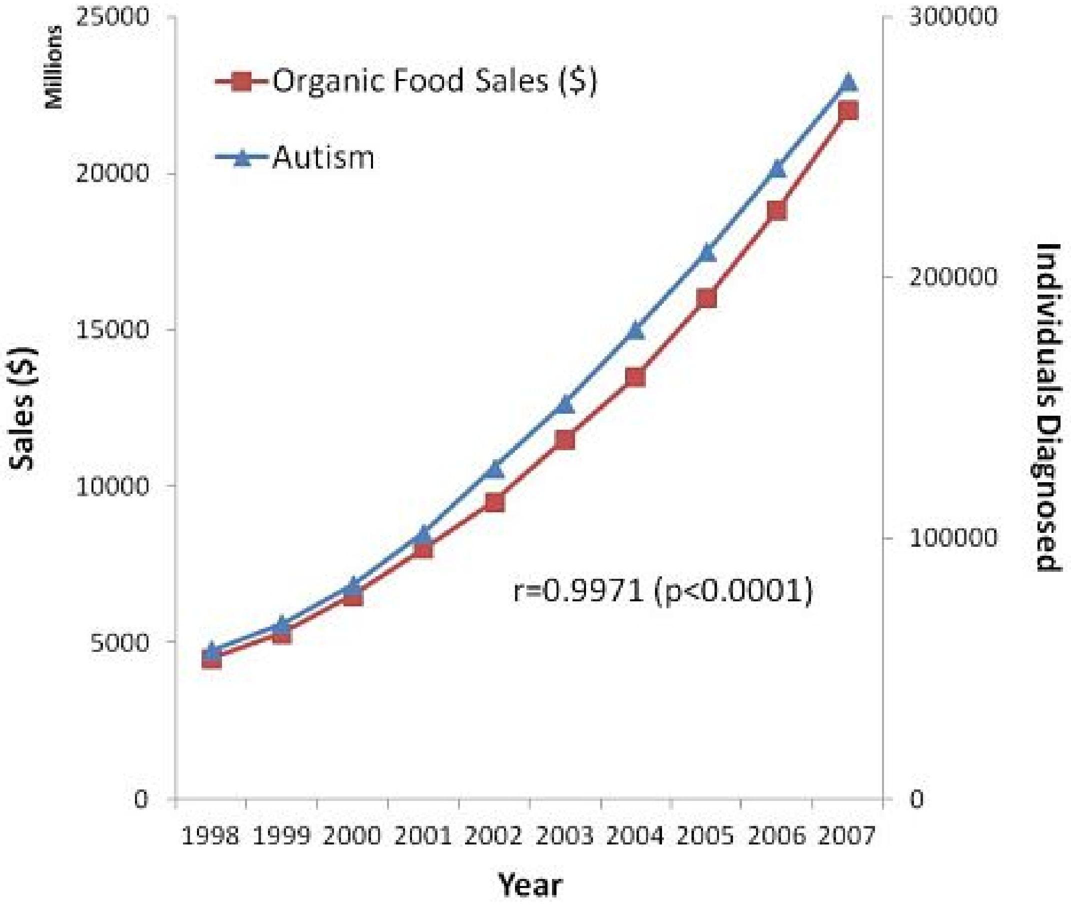
Number of people who died by becoming tangled in their bedsheets

Correlation: 94.71% (r=0.947091)

Data sources: U.S. Department of Agriculture and Centers for Disease Control & Prevention

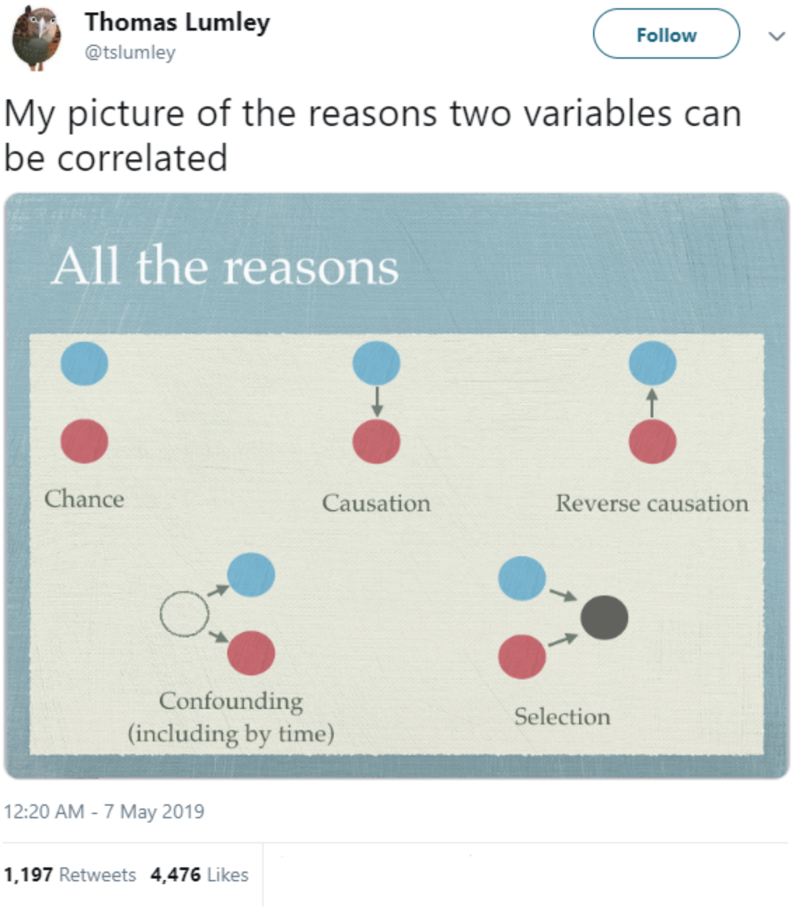
- https://tylervigen.com/spurious-correlations

Correlation $\neq$ causation e.g.



- Truth, Lies & Statistics: How to Lie with Statistics

Reasons for correlation



Thomas Lumley
@tslumley

Follow

My picture of the reasons two variables can be correlated

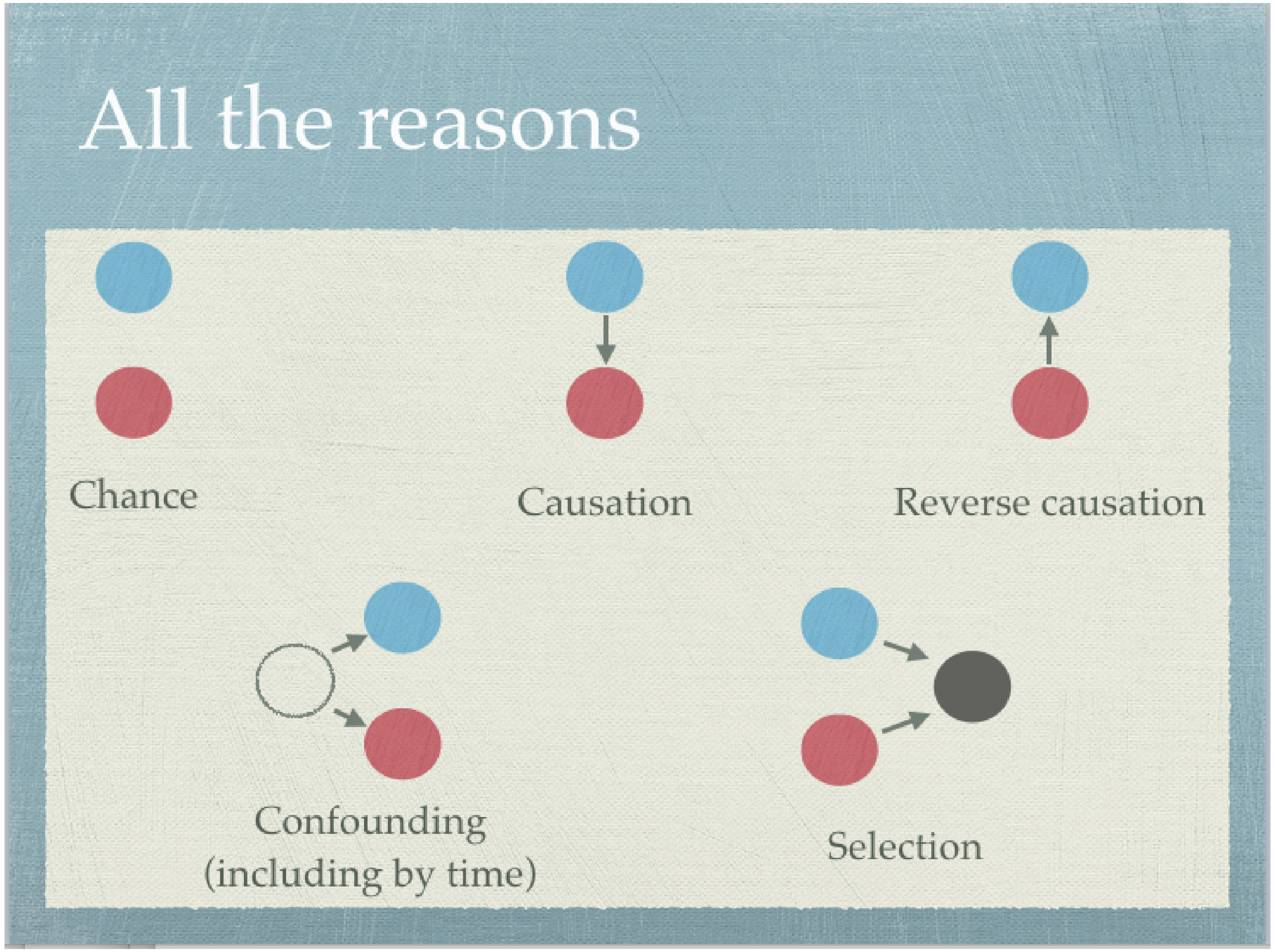
12:20 AM - 7 May 2019

1,197 Retweets 4,476 Likes

- https://twitter.com/tslumley/status/1125661624356954112 Professor and Chair in Biostatistics at the University of Auckland, New Zealand

Reasons for correlation

Using DAGs (Directed Acyclic Graphs):



All the reasons

Chance

Causation

Reverse causation

Confounding
(including by time)

Selection

- https://twitter.com/tslumley/status/1125661624356954112

Important for Statistical Inference!

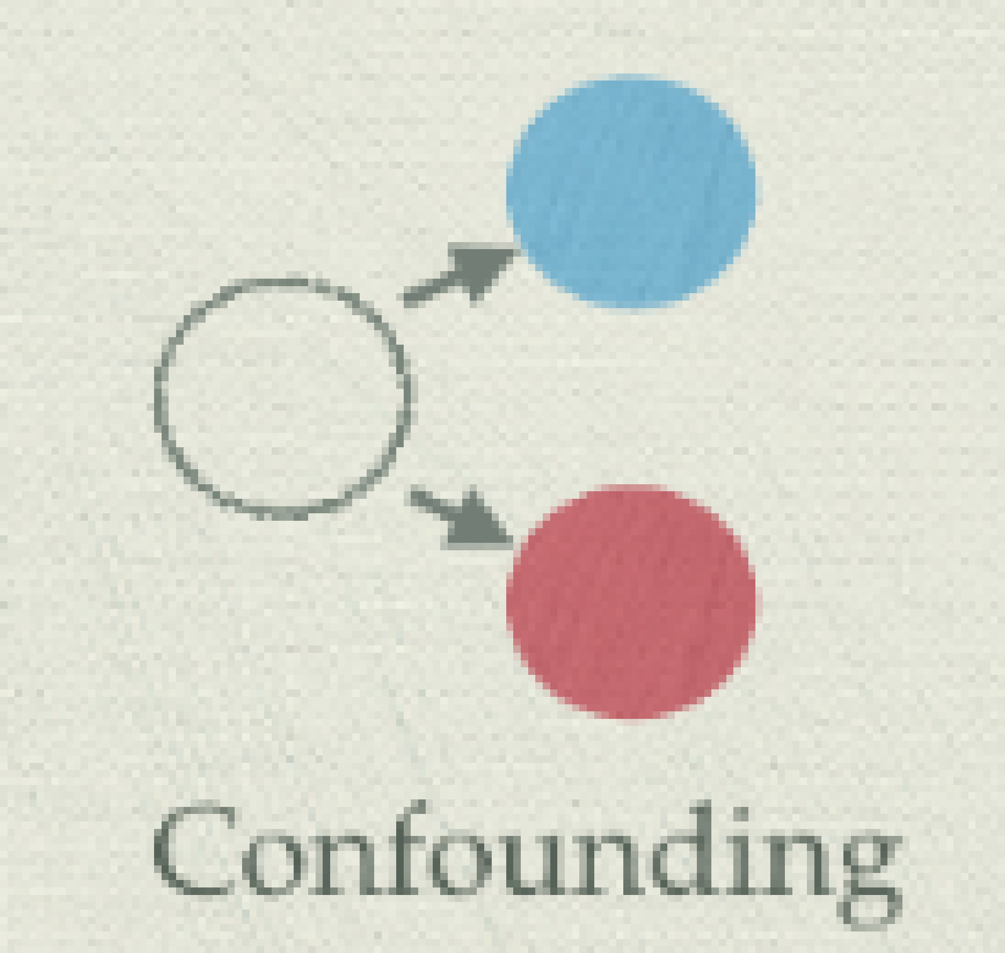
For instance in R;

```
fit <- lm(y~x,data=data)
anova(fit)
```

y $\sim$ x implies x $\Rightarrow$ y.

That is, when using the code y $\sim$ x we are interested in the effect of x on y.

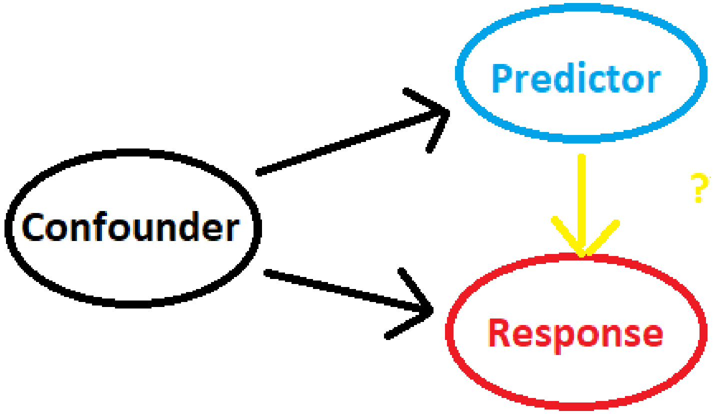
Confounding



- https://twitter.com/tslumley/status/1125661624356954112

Confounding

When a third variable, that affects both the response (dependent/outcome) and the predictor (independent/exposure) variable, distorts the underlying relaitionship between the predictor and response variables when it is not conditioned on.



Conditioning

Here, conditioning on a variable can mean:

- Adjusting for the variable in a multiple regression.
- Stratification at various levels of a variable.
- Restriction to only one level of a variable.

Confounding e.g.

Mr Kwan is a maths teacher at a local primary school and is interested in observing if there is a relationship between a student's score on a math test and their height.

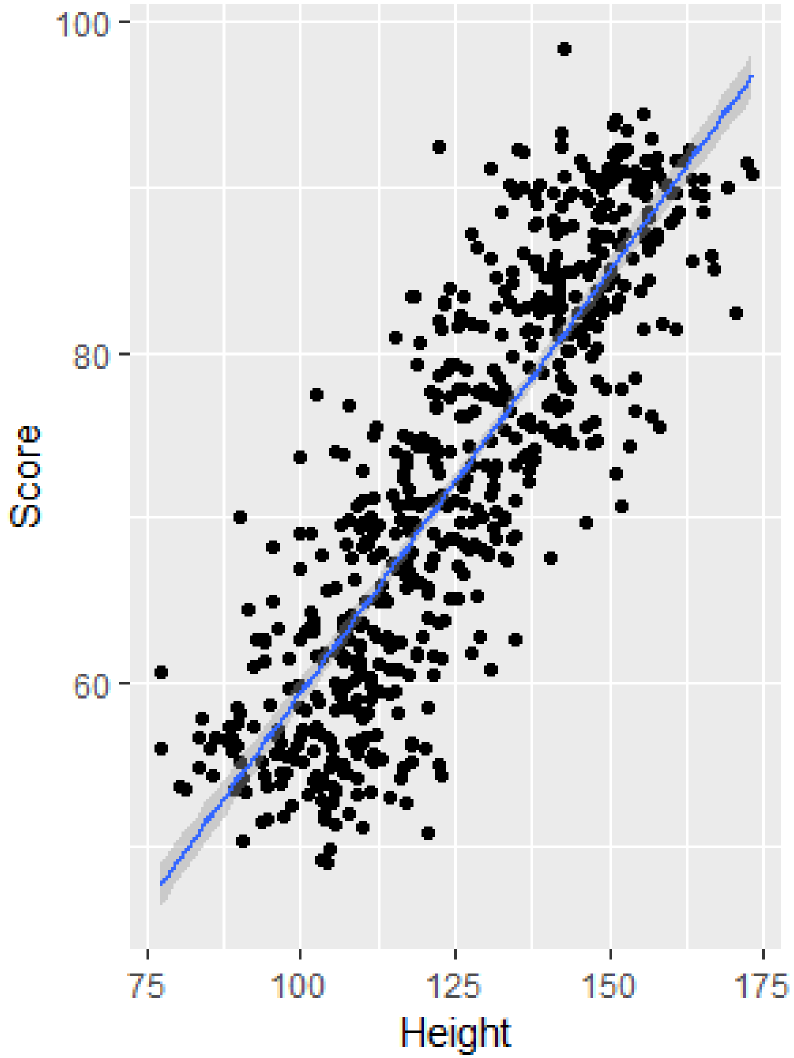
Score	55	56	55	63	60	59	66	64	65	...
Height	100	101	101	109	108	112	118	121	120	...

Confounding e.g.

Mr Kwan initially found a strong relaitionship between a student's height and their math test score.

```
> fit <- lm(Score ~ Height, data = Conf_data)
> anova(fit)
Analysis of Variance Table

Response: Score
Df Sum Sq Mean Sq F value    Pr(>F)
Height      1  64520   64520  1536.5 < 2.2e-16 ***
Residuals 598  25111      42
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Confounding e.g.

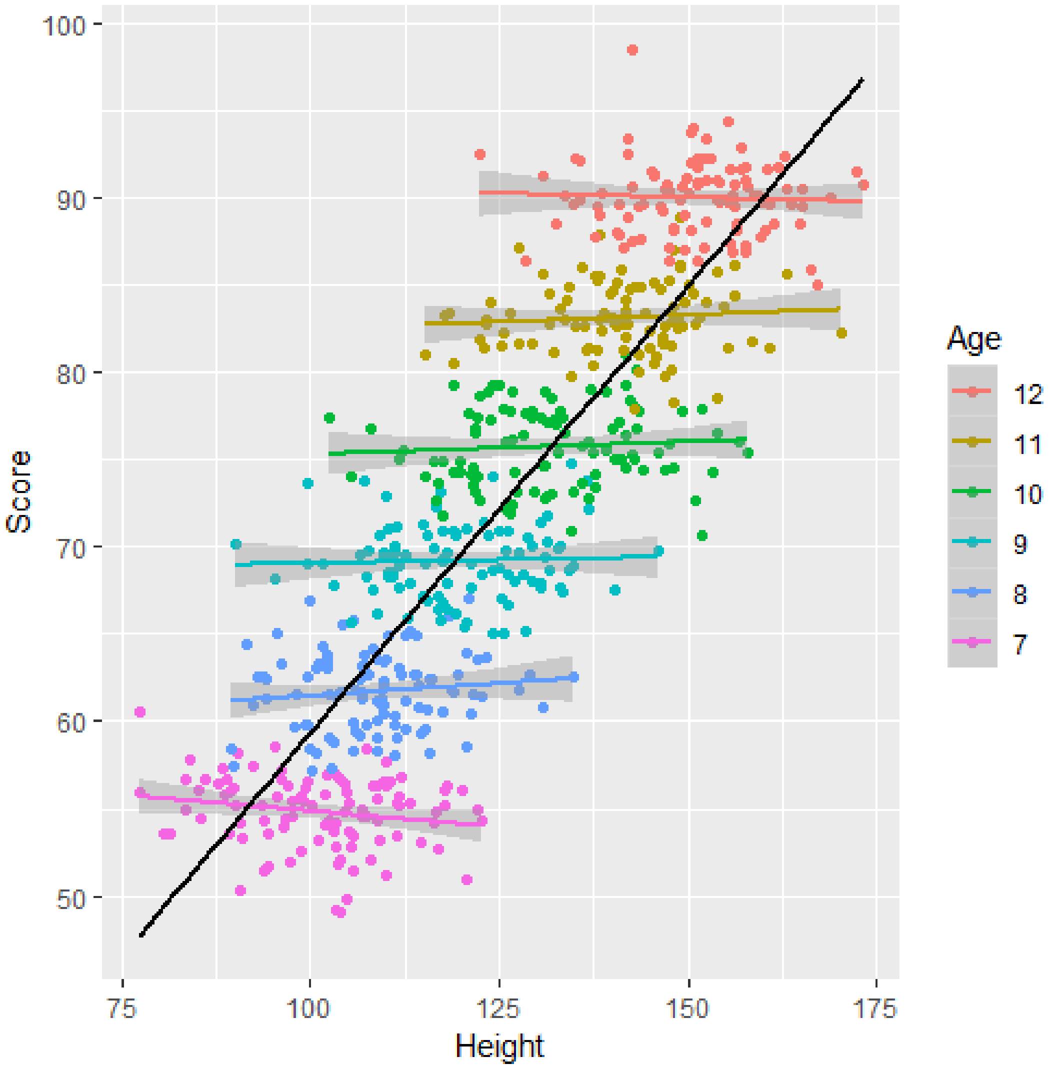


Confounding e.g.

Mr Kwan however forgot to account for a student's age. After accounting for age as another predictor, we observe a very different relationship...

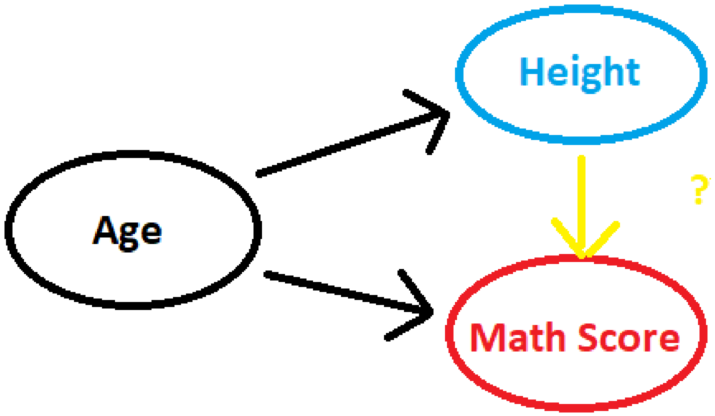
```
> fit <- lm(Score ~ Age + Height, data = Conf_data)
> anova(fit)
Analysis of Variance Table

Response: Score
Df Sum Sq Mean Sq   F value Pr(>F)
Age         5  86906 17381.2 3783.3416 <2e-16 ***
Height      1      1     0.5    0.1147 0.7349
Residuals 593   2724     4.6
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Confounding e.g.



Confounding e.g.

Here a student's age is confounding the effect of a student's height on their test score.



How to deal with Confounding

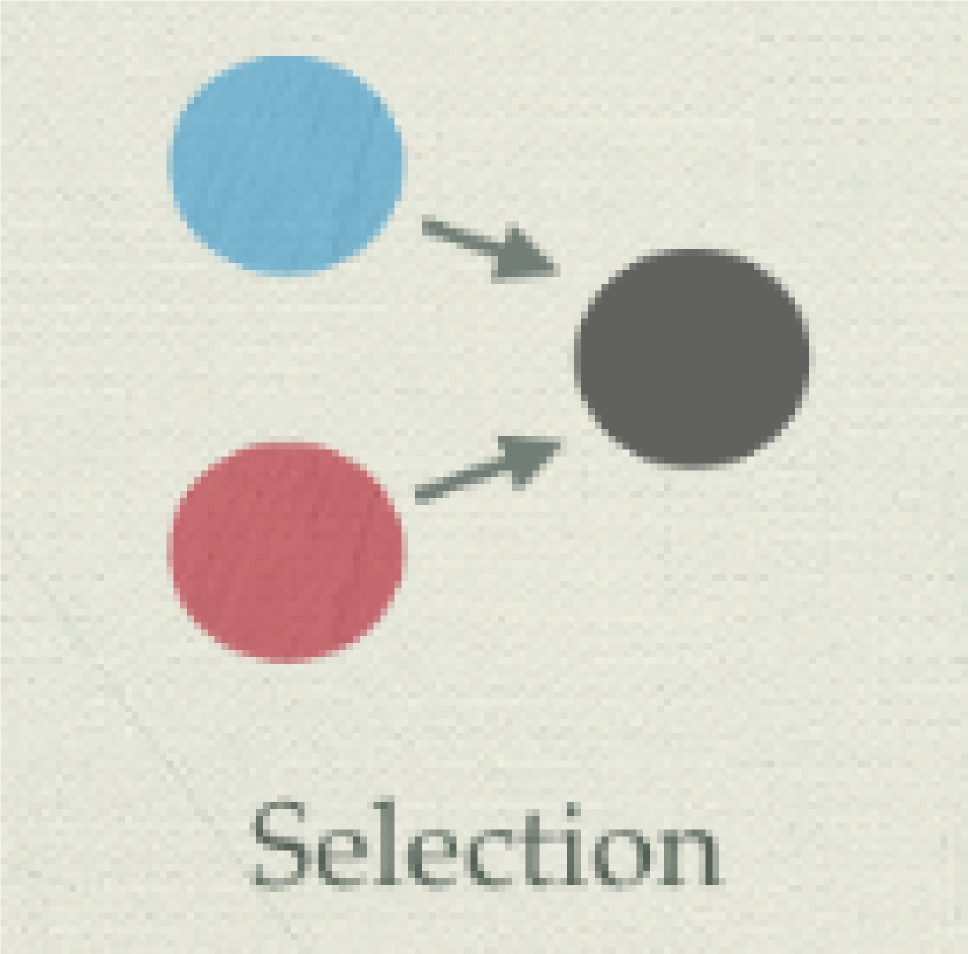
1. Random sampling and/or a controlled experiment to remove the effect of the confounding variable.

2. Measure possible confounding variables and adjust for them in a multiple regression analysis.

3. Stratified sampling, or testing at each level of the confounding variable seperately.

4. Restricted sampling, at only one level of the confounding variable, so that the effect of the confounding variable is consistent accross all samples.

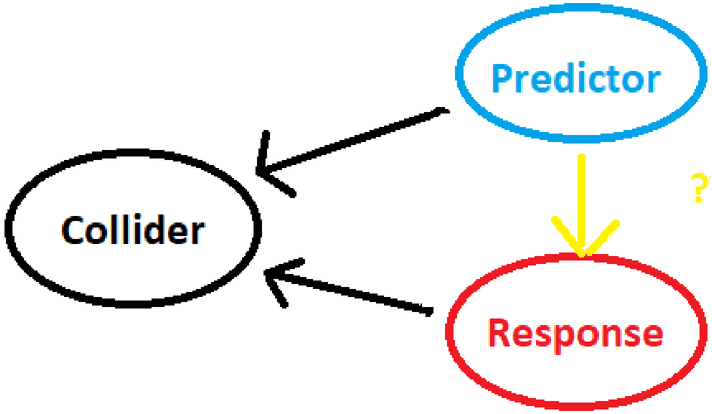
Colliders



- https://twitter.com/tslumley/status/1125661624356954112

Collider

When a third variable, that is affected by both the response (dependent/outcome) and the predictor (independent/exposure) variable, distorts the underlying relaitionship between the predictor and response variables when it is conditioned on.



Collider e.g.

Professor Langley is interested in measuring the relationship between 'little fish' and the 'big fish' that consumes them in lakes across Australia. She managed to count the number of little fish (*smalliformes fishidae*) and big fish (*bigiformes fishidae*) in a random sample of lakes across Australia.

Little Fish	8523	12004	11059	10456	10705	9873	8956	...
Big Fish	310	645	640	543	555	473	423	...

Collider e.g.

Professor Langley heard about Mr Kwan's experiment and is worried about possible confounding variables, so she also measured the levels of nitrogen in each of these lakes.

Previous studies have found that the defecation of fish produces ammonia, which increase the lakes nitrogen levels.

Little Fish	8523	12004	11059	10456	10705	9873	8956	...
Big Fish	310	645	640	543	555	473	423	...
Nitrogen	Low	High	High	Med	Med	Med	Low	...

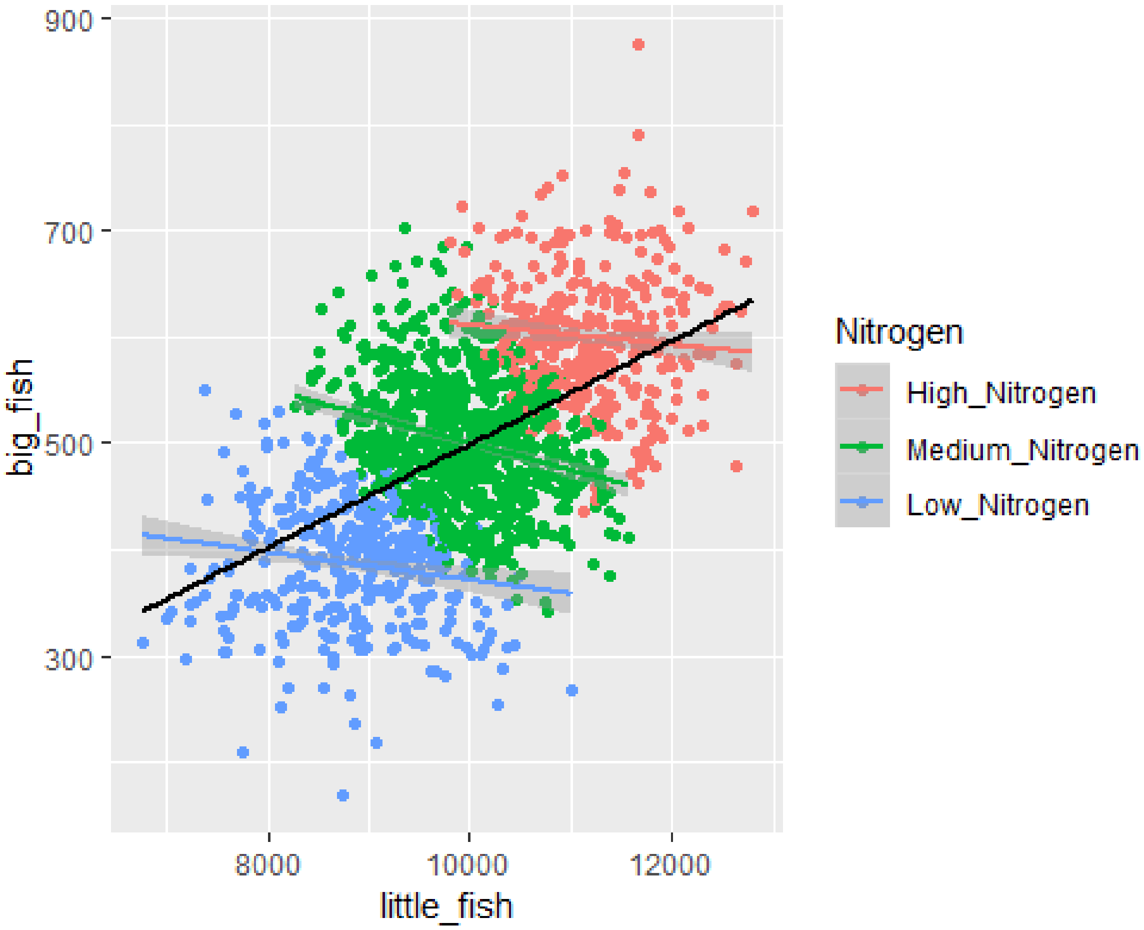
Collider e.g.

We observe opposite correlations after we account for Nitrogen!

```
> fit1 <- lm(big_fish~little_fish,data=col_data)
> summary(fit1)
Call:
lm(formula = big_fish ~ little_fish, data = col_data)

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.041983  20.011550   0.652    0.515
little_fish  0.048621   0.001996  24.360   <2e-16 ***

> fit2 <- lm(big_fish~Nitrogen+little_fish,data=col_data)
> summary(fit2)
Call:
lm(formula = big_fish ~ Nitrogen + little_fish, data = col_data)

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)             8.070e+02  2.693e+01   29.97  < 2e-16 ***
NitrogenMedium_Nitrogen -1.218e+02  4.795e+00  -25.40  < 2e-16 ***
NitrogenLow_Nitrogen    -2.576e+02  7.220e+00  -35.68  < 2e-16 ***
little_fish             -1.849e-02  2.388e-03   -7.74 1.82e-14 ***
```

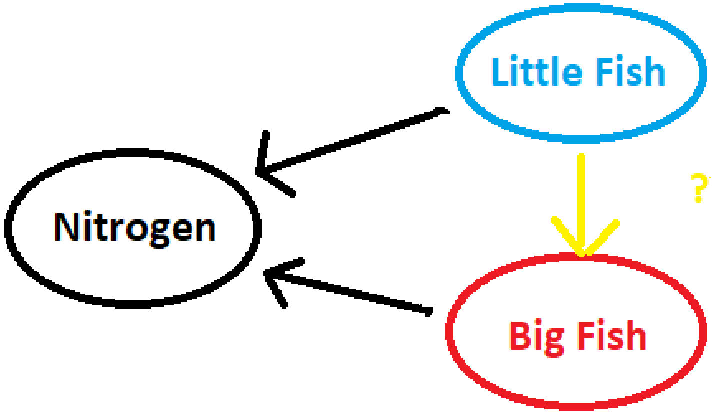
Collider e.g.

What is the correct interpretation?



Collider e.g.

Since both ’big fish’ and ’little fish’ (the predictor and response variables) impact Nitrogen, the variable is acting as a collider when we account for it in analysis. This distorts the underlying relationship between ’little fish’ and ’big fish’, and as such should **not** be accounted for in analysis.



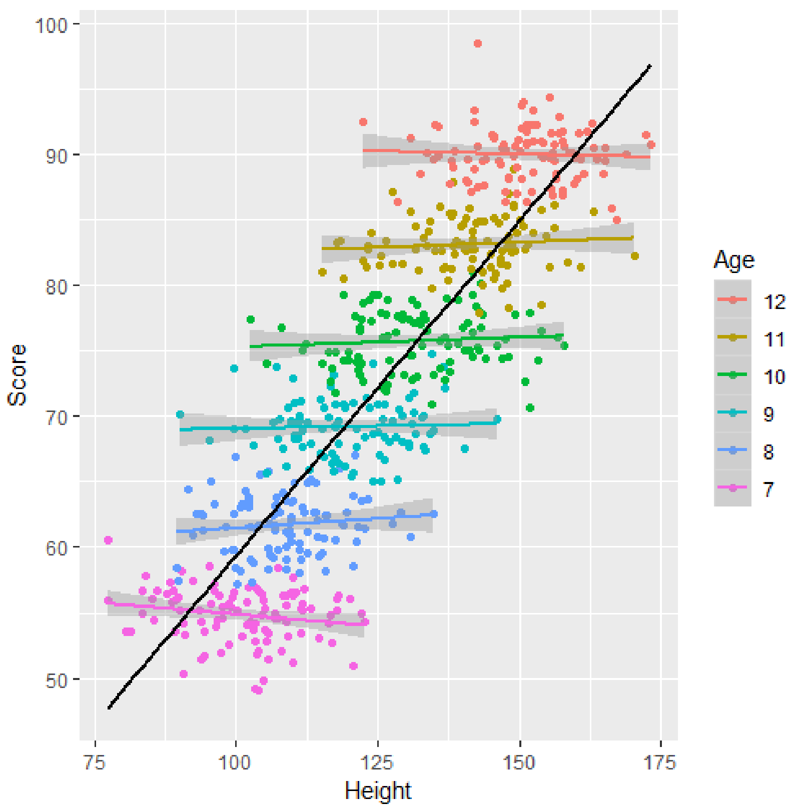
How to deal with Colliders

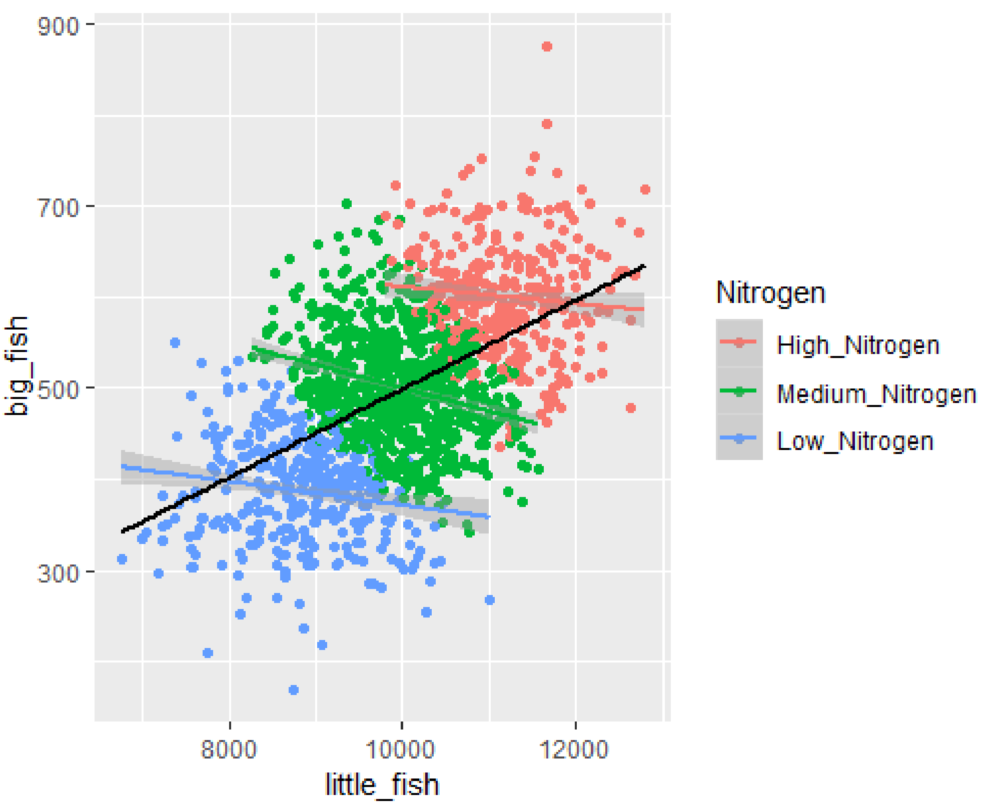
- Random sampling and/or a controlled experiment to remove the effect on a collider variable.

- Ensure collider variables are not adjusted for in analysis, as this will open up the causal pathway.

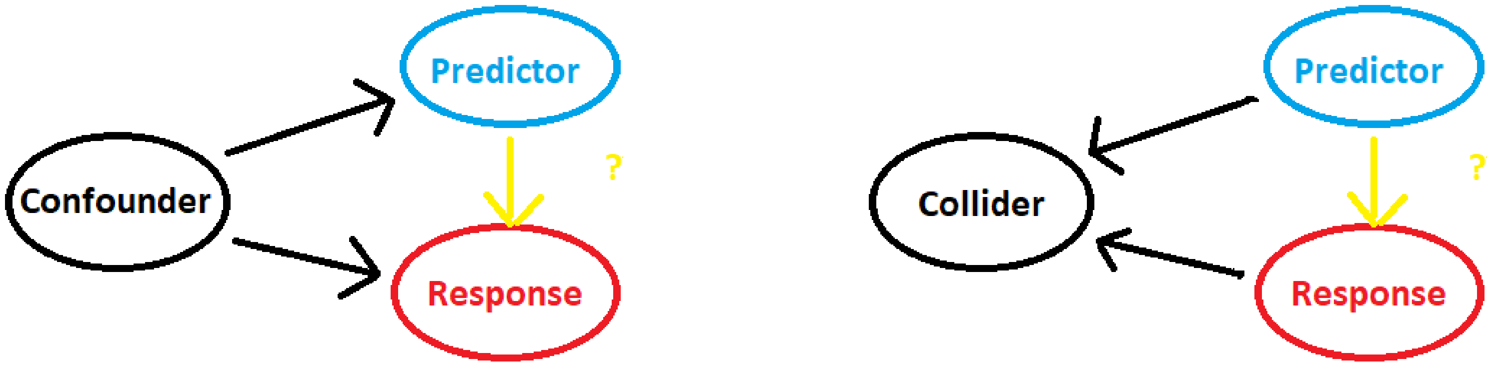
- Ensure samples are not being stratified or restricted to one or more levels of the collider variable as this could cause sampling bias.

Confounding vs. Collider





Confounding vs. Collider



Predictor

?

Confounder

Response

Predictor

?

Collider

Response

Confounding vs. Collider

Confounder is a third variable, that **affects** both the response (dependent/outcome) and the predictor (independent/exposure) variable, and can distort the underlying relaitionship between the predictor and response variables when it **is** conditioned on.

Collider is a third variable, that is **affected** by both the response (dependent/outcome) and the predictor (independent/exposure) variable, and can distort the underlying relaitionship between the predictor and response variables when it **is not** conditioned on.

DAGs as a useful tool for research

DAGs (directed acyclic graphs) can be used as a useful tool for researchers at two key stages of research:

1. **The experimental design phase** - to aid in visualising what variables affect the response and predictor variable of interest, to either remove them from the experiment or measure said variables to later condition on them (if they are not a collider).

2. **Prior to analysis** - to work out which variables to condition on, in order to only focus on the causal pathway between the predictor of interest (independent variable) and response (dependent) variable.

General Rules for DAGs

The following general rules can be used:

- All paths between the independent and dependent variable need to be blocked except for the causal path.
- 'Open' paths between the independent and dependent variable opens up the path of association.
- Paths are blocked by either colliders or by conditioning on a non-collider (e.g. a confounder).
- Conditioning on a collider opens up the causal pathway.
- Always a good idea to test the sensitivity of DAG assumptions, by modifying said assumptions and observing how the results change.
- Results can change depending upon the assumptions of the DAG, which comes from 'expert' field knowledge and previous studies.

Collinearity

Collinearity

Collinearity (multicollinearity/ill-conditioning) occurs when predictors in regression are so heavily correlated that it becomes difficult or even impossible to distinguish their individual effects on the response variable.

Collinearity e.g.

Dr Dedden is trying to create the ideal cupcake, and is interested in how ’sweetness’ impacts a muffins ’tastiness’.

Cupcakes were baked with a variety of sweetness levels, and an expert taster scored the cupcakes based on their ’tastiness’. Dr Dedden however has a few measures of cupcake sweetness...

Collinearity e.g.

Two measures; *sweetness 1* and *sweetness 2* are infact identical. If Dr Dedden attempted to fit the following regression:

$$\mu_{tastiness} = \alpha + \beta_1 sweetness1 + \beta_2 sweetness2$$

There would be no unique solution to this equation. As an example, the below solutions would all be equivalent:

$$\mu_{tastiness} = 30 + 5 sweetness1 + 0 sweetness2$$

$$\mu_{tastiness} = 30 + 3 sweetness1 + 2 sweetness2$$

$$\mu_{tastiness} = 30 + 2.5 sweetness1 + 2.5 sweetness2$$

and would all produce the same predicted tastiness!

Collinearity e.g.

We note that collinearity does not affect a model’s ability to predict the response accurately (all the equations on the previous slide make the same predictions), however it becomes problematic if the objective of the study is to measure the individual effects of each predictor.

Collinearity e.g.

In the case of perfect collinearity, most programs will either not include both variables or give a warning and stop analysis.

```
> fit2 <- lm(tastiness~sweetness1 +sweetness2,data=cupcake_dat)
> summary(fit2)
Call:
lm(formula = tastiness ~ sweetness1 + sweetness2, data = cupcake_dat)

Coefficients: (1 not defined because of singularities)
Estimate Std. Error t value Pr(>|t|)
(Intercept)  28.1905    10.1740   2.771  0.00669 **
sweetness1    5.1113     0.5078  10.065  < 2e-16 ***
sweetness2        NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.58 on 98 degrees of freedom
Multiple R-squared:  0.5083,Adjusted R-squared:  0.5033
F-statistic: 101.3 on 1 and 98 DF,  p-value: < 2.2e-16
```

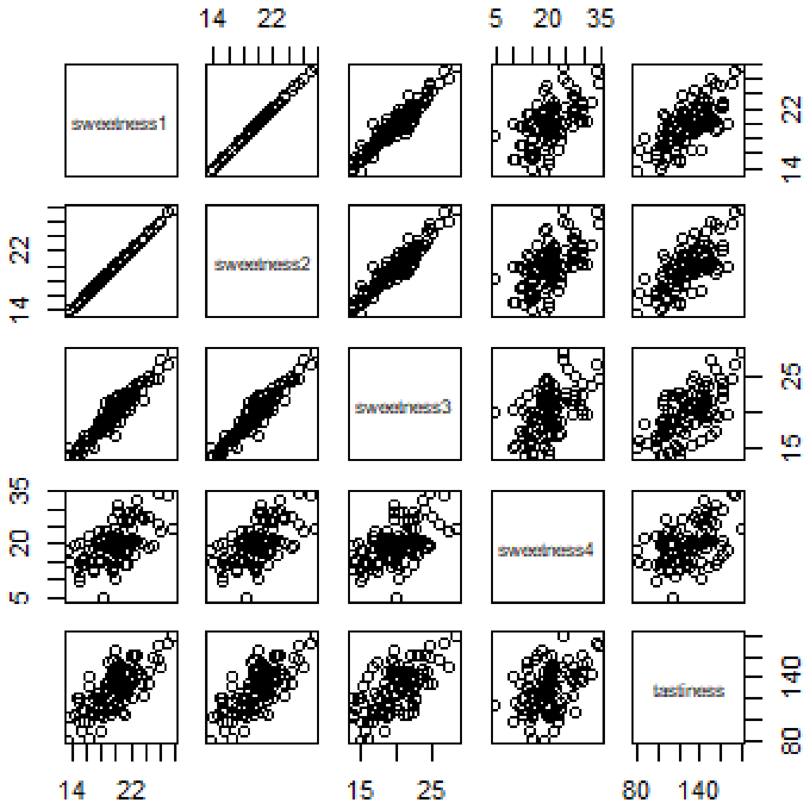
”near - collinearity”, where variables are not identical, but highly correlated still posses a problem...

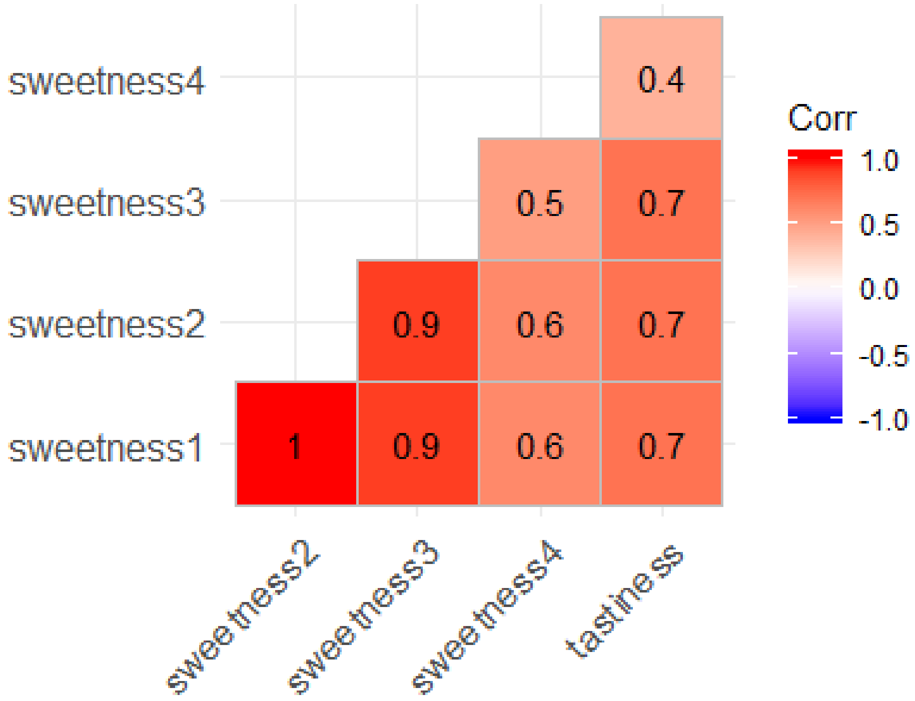
Collinearity e.g.

So... try looking at a correlation matrix of all variables prior to analysis, so that potential collinear variables can be flagged in advance!

```
plot(cupcake_dat)
library(ggcorrplot)
corr <- round(cor(cupcake_dat), 1)
ggcorrplot(corr, hc.order = FALSE, type = "lower",
lab = TRUE)
```

Collinearity e.g.





Collinearity e.g.

Having Sweetness1 on its own produces a regression coefficient of $\beta \approx 5$, with standard error $\sigma \approx 0.5$:

```
> fit1 <- lm(tastiness~sweetness1,data=cupcake_dat)
> summary(fit1)

Call:
lm(formula = tastiness ~ sweetness1, data = cupcake_dat)

Residuals:
Min      1Q  Median      3Q     Max
-32.714 -10.603   2.325  10.112  35.985

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  28.1905    10.1740   2.771  0.00669 **
sweetness1    5.1113     0.5078  10.065  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.58 on 98 degrees of freedom
Multiple R-squared:  0.5083,Adjusted R-squared:  0.5033
F-statistic: 101.3 on 1 and 98 DF,  p-value: < 2.2e-16
```

Collinearity e.g.

With Sweetness1 and Sweetness3, (that had a correlation coefficient of 0.9) we see that estimate change and standard errors inflate:

```
> fit3 <- lm(tastiness~sweetness1 +sweetness3,data=cupcake_dat)
> summary(fit3)

Call:
lm(formula = tastiness ~ sweetness1 + sweetness3, data = cupcake_dat)

Residuals:
Min      1Q  Median      3Q     Max
-34.028 -10.787   0.772  10.429  34.877

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)   26.169      9.978   2.623   0.0101 *
sweetness1     8.587      1.546   5.556 2.42e-07 ***
sweetness3    -3.368      1.418  -2.374   0.0195 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.25 on 97 degrees of freedom
Multiple R-squared:  0.5353,Adjusted R-squared:  0.5257
F-statistic: 55.87 on 2 and 97 DF,  p-value: < 2.2e-16
```

Collinearity e.g.

With sweetness1 and sweetness4, (that had a correlation coefficient of 0.6), the Sweetness1 regression coefficient and standard errors did not change too much, however Sweetness 4 does not come out significant:

```
> fit4 <- lm(tastiness~sweetness1 +sweetness4,data=cupcake_dat)
> summary(fit4)

Call:
lm(formula = tastiness ~ sweetness1 + sweetness4, data = cupcake_dat)

Residuals:
Min      1Q  Median      3Q     Max
-33.209 -10.489   2.016   9.606  36.297

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  28.4436    10.2320   2.780  0.00653 **
sweetness1    4.9550     0.6203   7.988 2.84e-12 ***
sweetness4    0.1432     0.3233   0.443  0.65885
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.64 on 97 degrees of freedom
Multiple R-squared:  0.5093,Adjusted R-squared:  0.4992
F-statistic: 50.33 on 2 and 97 DF,  p-value: 1.014e-15
```

Impacts of collinearity

Collinearity can have the following impacts in practice:

- Regression coefficients of collinear variables can change dramatically when they are both included, with non-sensical estimates able to be obtained.
- Inflated standed errors of collinear variables.
- Wider confidence intervals for regression coefficients of collinear variables.
- Lack of significance, as the effect becomes shared accross the two predictors. So each variables on its own may not come out significant, or selected in a model selection procedure.

Take home tips

If unsure about whether to include a variable or not based upon causal connections:

- Draw some DAGs of what you believe the causal pathway to be.
- Include in your DAG unobserved covariates, to gain a picture as to what variables need to be measured and conditioned upon to block certain causal pathways, or to get an idea about the bias that could be obtained by not doing so.
- Build a multiple regression based upon the general rules of your DAG.
- Don't worry about creating the 'perfect' DAG, weak correlations will likely only cause weak biases!
- Modify the assumptions of your DAG and observe how sensitive your assumptions are to analysis.

Take home tips

If unsure about whether you have collinear variables to exclude:

- Produce a correlation matrix of all variables prior to analysis, and flag potential collinear variables.
- Observe the effect of incuding both collinear variables (check for inflated standard errors).
- Only worry if correlations are really high, and analysis is highly affected by the inclusion of both variables (no hard and fast rule but I would start to be concerned if $r > 0.9$).

Further Reading..

- Truth, Lies & Statistics: How to Lie with Statistics
- http://www.medicine.mcgill.ca/epidemiology/Joseph/courses/EPIB-621/confounding.pdf
- https://twitter.com/tslumley/status/1125661624356954112
- Introduction to casual diagrams for confounder selection. Williamson and Aitken et. al. Respirology. 2013. doi: 10.1111/resp.12238
- Causal Inference Miguel A. Hernán, James M. Robins August 26, 2018
- Try making DAGs with ggdag: https://cran.r-project.org/web/packages/ggdag/vignettes/intro-to-ggdag.html