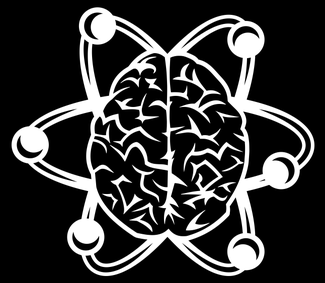


CHOOSING PREDICTORS TO LINEAR REGRESSION MODELS

Turku PET Centre Brain Imaging Course 2025

Tuulia Malen, Turku PET Centre
tukama@utu.fi

‘Blindly tossing variables into the causal salad is never a good idea.’

- R. McElreath

Reference

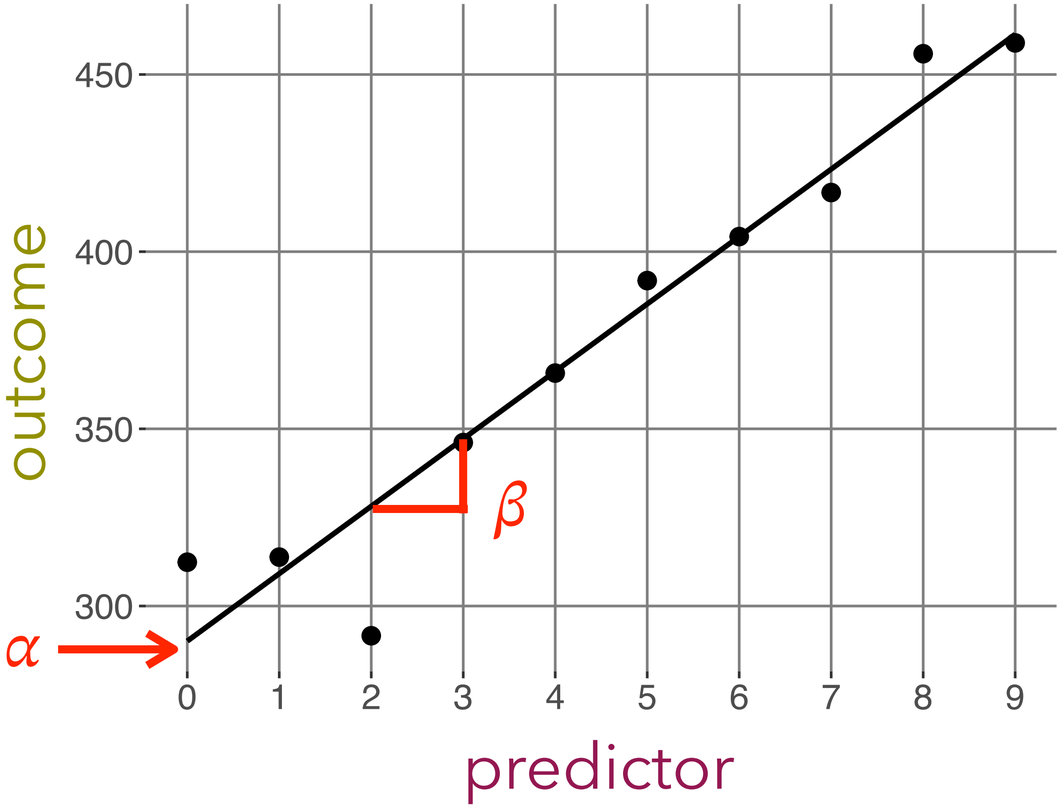
- Richard McElreath, an American professor of Anthropology, Max Planck Institute, Germany
- Statistical Rethinking. *A Bayesian Course with Examples in R and Stan* (2nd edition, 2020). Chapman and Hall/ CRC.
  - Chapters 5-7 (mainly)
  - The 2024 edition of the course, including slides, lectures and exercise material:
    - https://github.com/rmcelreath/stat_rethinking_2024
    - https://www.youtube.com/watch?v=mBEA7PKDmiY&list=PLDcUM9US4XdPz-KxHM4XHt7uUVGWWVSus&index=5
- Visualization
  - Posit team (2025). RStudio: Integrated Development Environment for R. Posit Software, PBC, Boston, MA. URL http://www.posit.co/.
  - OpenAI's DALL·E via ChatGPT

Introduction: Linear regression modeling

- Estimate associations between one or more predictors (independent variables) and an outcome (dependent variable)
- Predictor effect on outcome
  - How age, sex, and chronic pain affect the opioid receptor availability?
  - How dopamine synthesis capacity affects the dopamine receptor availability?

Introduction: General linear regression with main effects

- If one predictor:
  - outcome = α + β*predictor + error
- α and β are estimated by the regression model based on our data so that error (distances between observations and the fit= line) is minimized
- α = intercept (outcome when predictor = 0)
- β = **the effect** = regression coefficient = the change in outcome with one-unit increase in predictor
- Syntax often something like: outcome ~ predictor



More in Linear mixed effects models -lecture from the 2024 course

Introduction

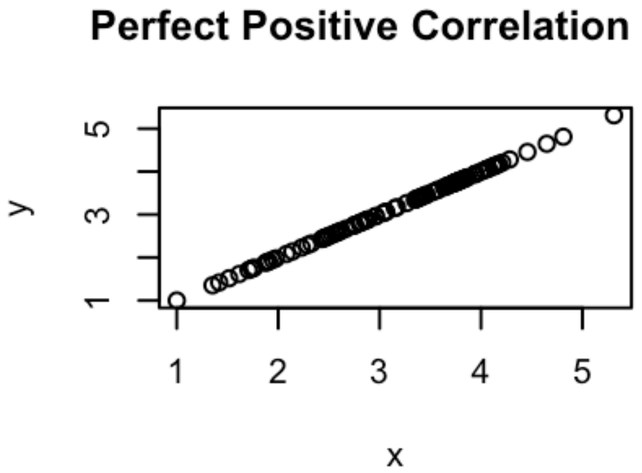
- If one predictor:
  - outcome = α + β*predictor + error

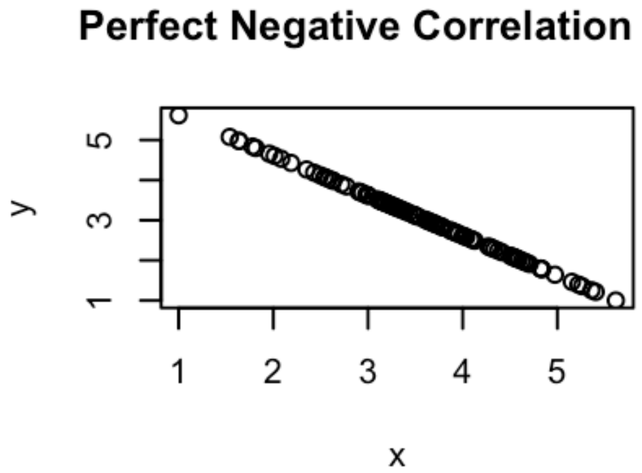
- If two (or more) predictors:
  - outcome = α + β_1*$predictor_1$ + β_2*$predictor_2$ + … + error
  - Independent effects
  - The effect of $predictor_1$, when $predictor_2$ is adjusted (‘controlled’) for

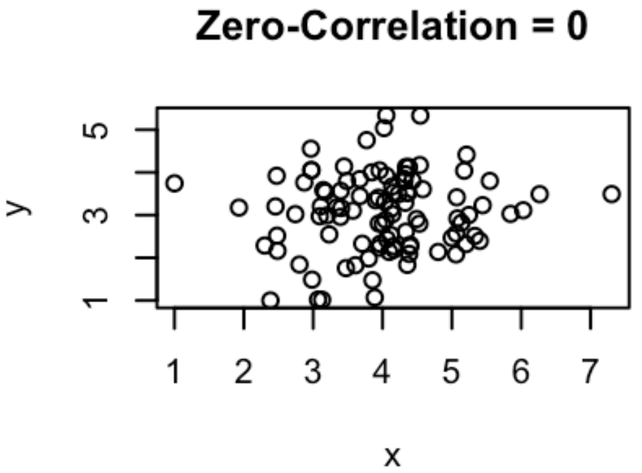
Correlation & Causality

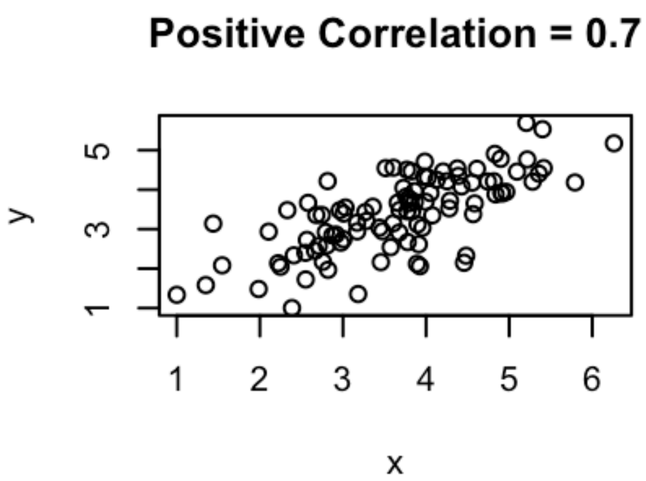
- Linear regression is based on correlation between the variables (predictors and outcome)
- **Correlation is common and it does not necessarily reveal causality, but only association**
  - Thus, the *effect* is only an association, although it sounds causal!
  - https://www.tylervigen.com/spurious-correlations
    - Probiotics and problems: Yogurt consumption and Google searches for "I can't even"

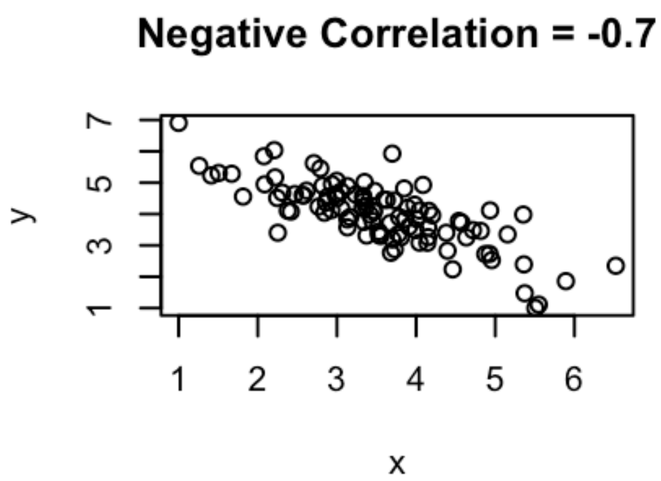
Correlation & Causality











Correlation & Causality

- Many times causality (the effect!) is the key interest in our studies
  - Which factors cause the disease?
  - Predictor= the factor that ***predicts*** the outcome

Which predictors to include in the model?

Which predictors to include in the model?

- Just the ones that are interesting (research question)?
- Everything that we could possibly get?

- Variables that are known to systematically explain variance in the outcome
  - Including the interesting + 'uninteresting' (covariates*) ones
  - Disorder, age and sex: Disorder effect that is independent of the effects of age and sex

*Btw the model treats them just like any other predictors

Which predictors to include in the model?

- Just the ones that are interesting (research qu
- Everything that we
- Variables that are known to systematically explain variance in the outcome
  - Including the interesting + 'uninteresting' (covariates*) ones
  - Disorder, age and sex: Disorder effect that is independent of the effects of age and sex

'Blindly tossing variables into the causal salad is never a good idea.'

*Btw the model treats them just like any other predictors

Which predictors to include in the model?

- Just the ones that are interesting (research question)?
- Everything that we could possibly get?

- Variables that are known to systematically explain variance in the outcome
  - Including the interesting + 'uninteresting' (covariates*) ones
  - Disorder, age and sex: Disorder effect that is independent of the effects of age and sex

*Btw the model treats them just like any other predictors

Which predictors to include in the model?

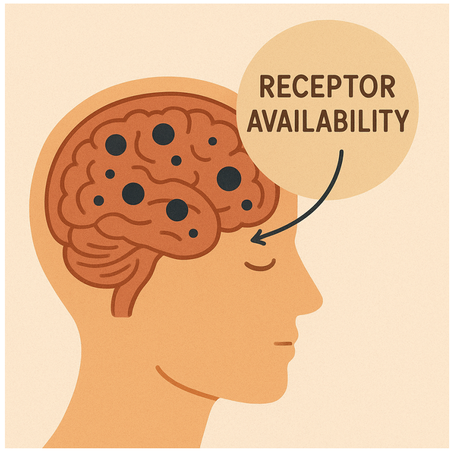
- But with caution…
- Let's think of causality between our modeling variables **to avoid confounds**
  - Although many times difficult, we can try to interpret (or rule out) causality between our variables
    - Longitudinal data: Happiness today cannot cause happiness yesterday
    - Receptor availability cannot cause age

Confounds

- Let’s think of causality between our modeling variables **to avoid confounds**
- Features of our data and model that will **mislead us about the effects**
  - Produce false effects
  - Hide existing effects
- Fork, pipe, collider, descendant

Confounds: The Fork

- What is the effect of IBS severity on receptor availability?

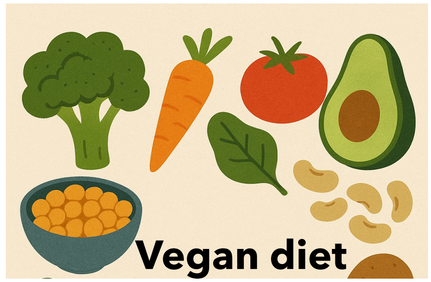


IBS= Irritated bowel syndrome

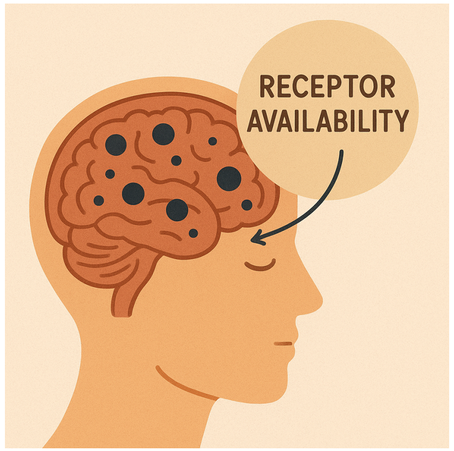
Confounds: The Fork

- Let’s assume…

- Many vegan proteins, e.g. beans induce IBS symptoms in vulnerable individuals



- Many vegan products are healthy for the brain

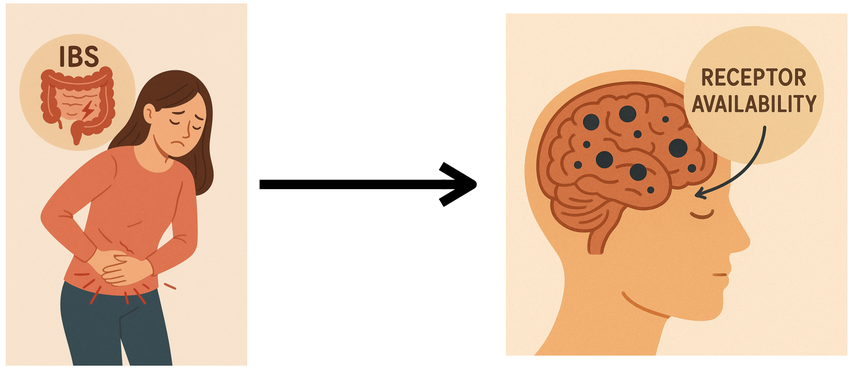


Directed acyclic graph (DAG), arrows point the direction of the causality

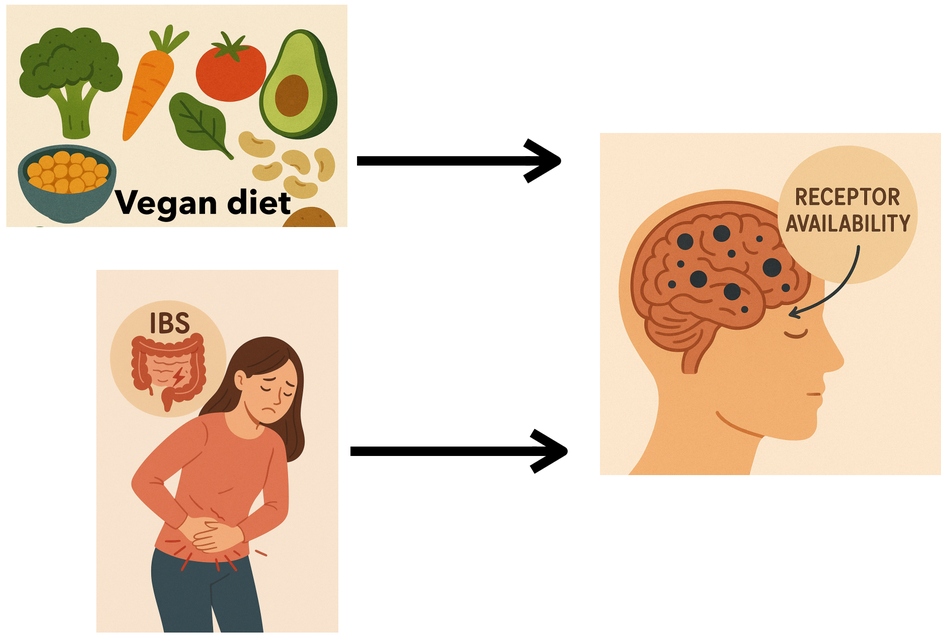
Confounds: The Fork

- What is the effect of IBS severity on receptor availability?

Receptor availability ~ IBS severity



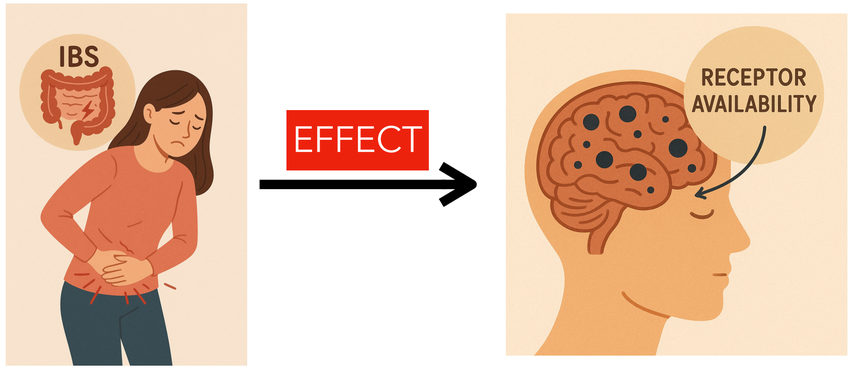
Receptor availability ~ IBS severity + vegan diet



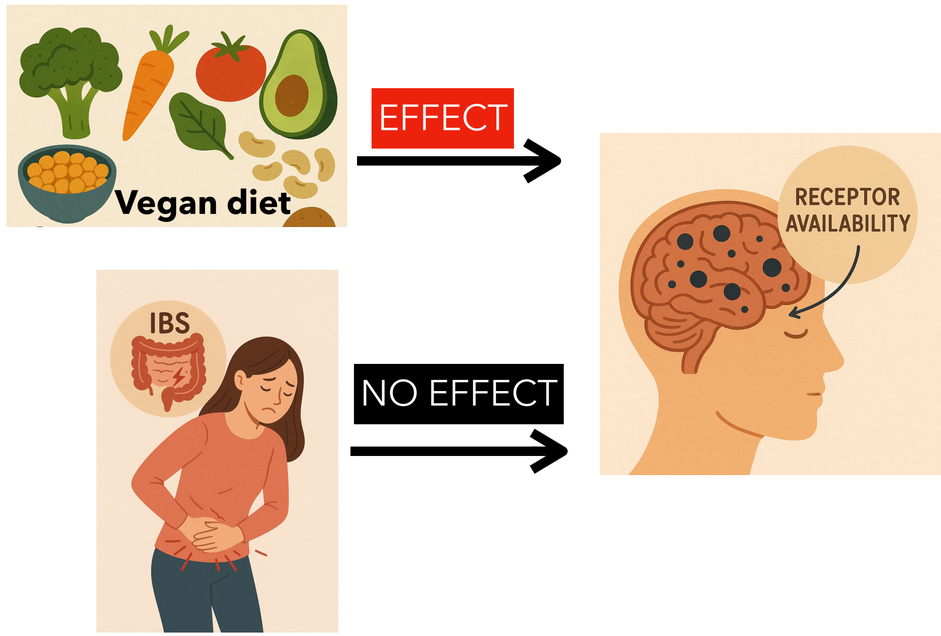
Confounds: The Fork

- What is the effect of IBS severity on receptor availability?

Receptor availability ~ IBS severity



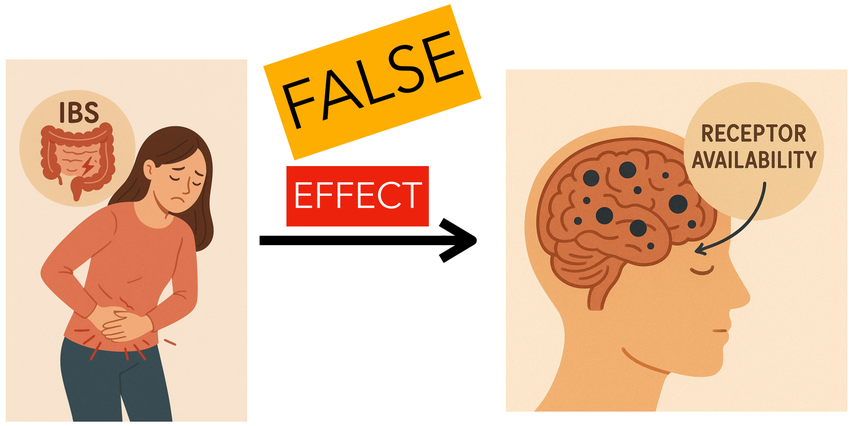
Receptor availability ~ IBS severity + vegan diet



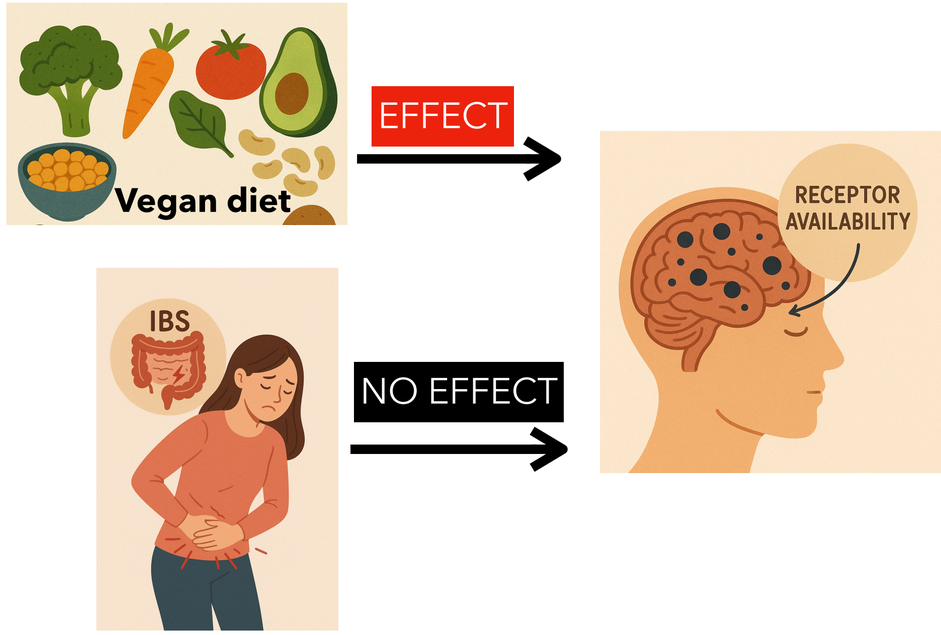
Confounds: The Fork

- What is the effect of IBS severity on receptor availability?

Receptor availability ~ IBS severity



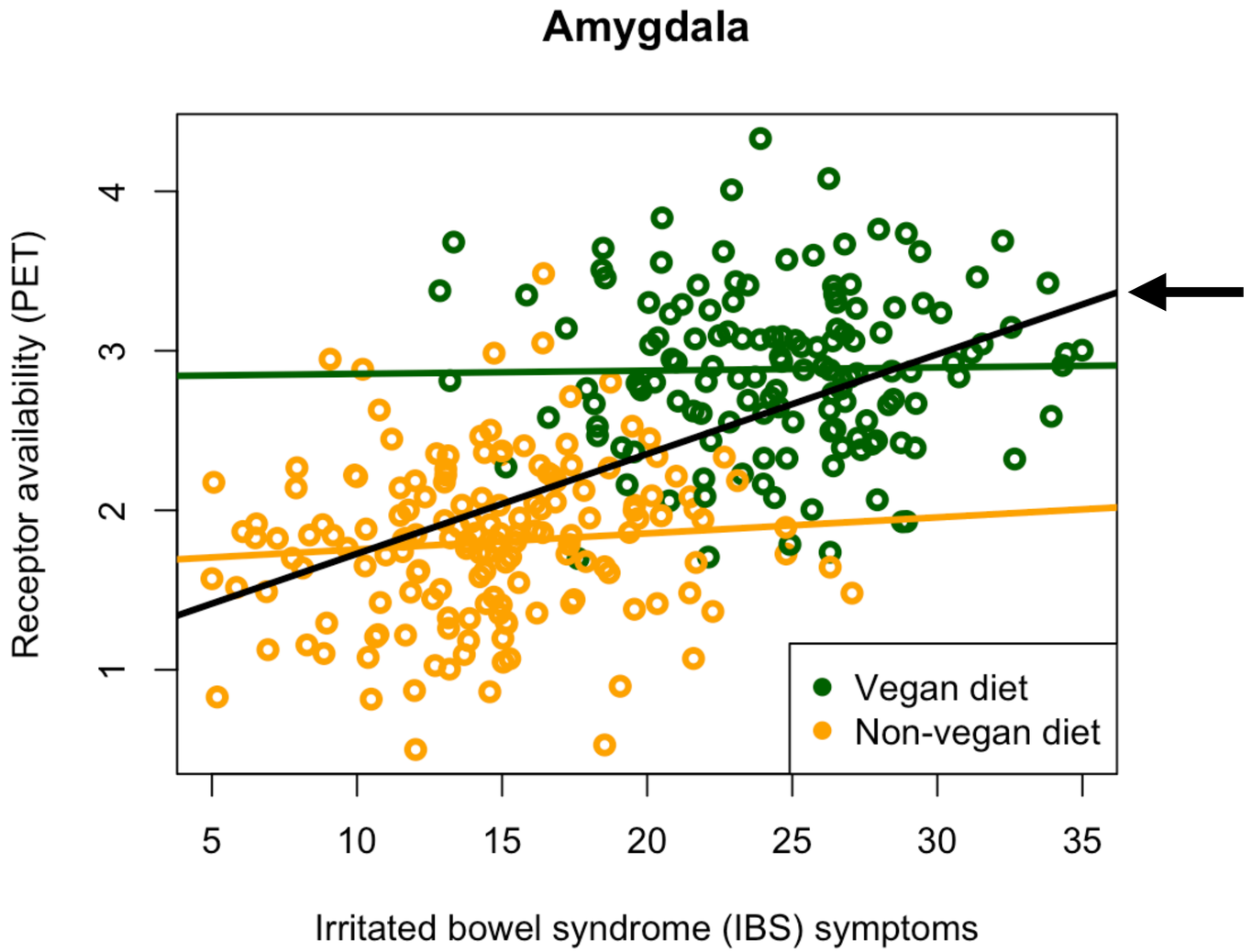
Receptor availability ~ IBS severity + vegan diet



- The vegan effect misinterpreted as the IBS effect

Confounds: The Fork

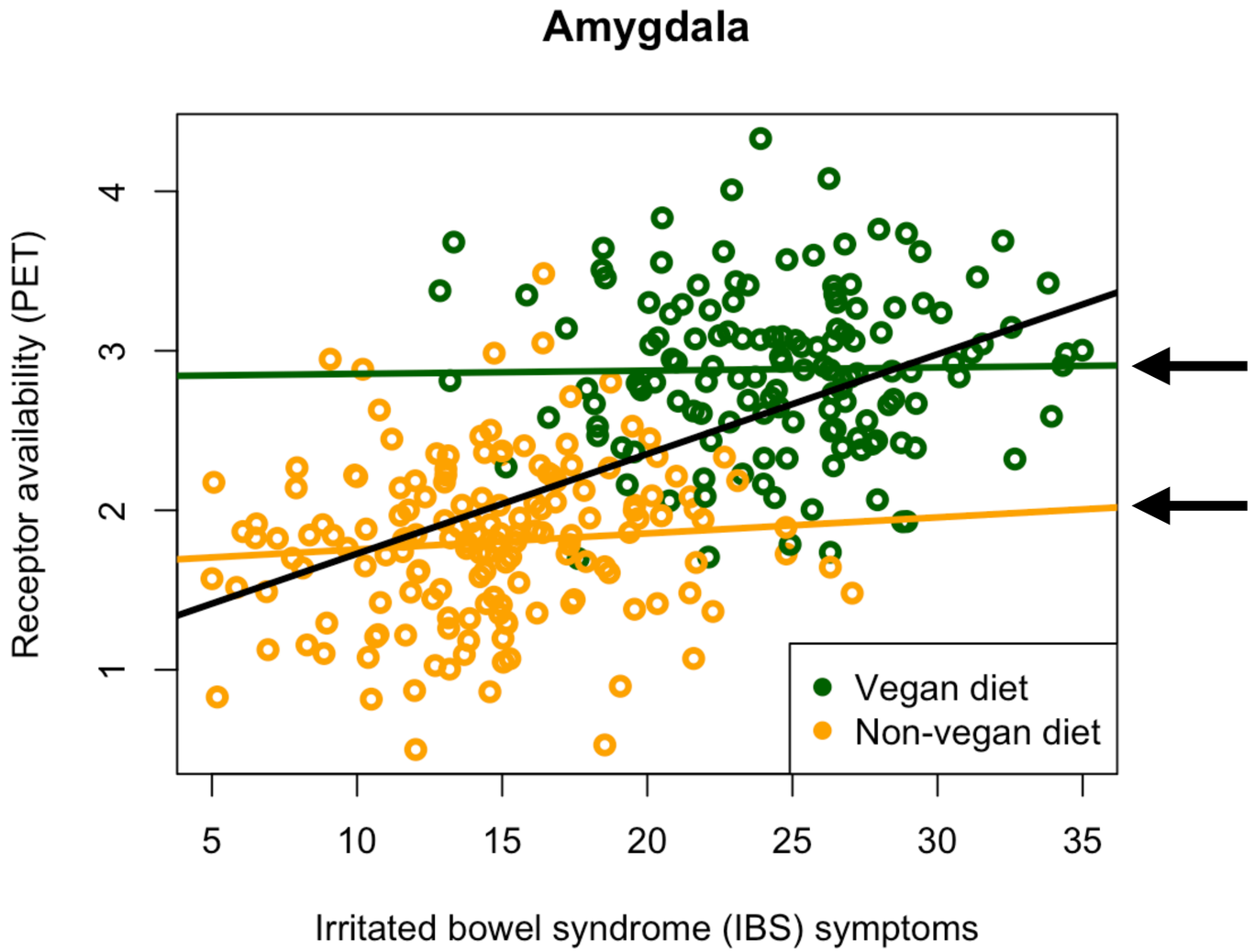
- What is the effect of IBS severity on receptor availability?



Confounds: The Fork

Disclaimer: These are not the regression lines from the model (we do not get IBS effect separately for vegans and non-vegans from main effects model), but this is describing the characteristics of the data that induces the confound! (Considers also the upcoming examples)

- What is the effect of IBS severity on receptor availability?



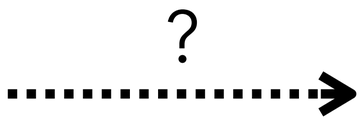
Confounds: The Fork

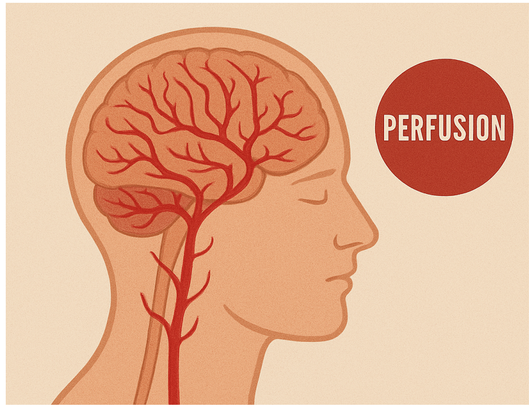
The general rule of thumb

- If a certain variable affects both the interesting predictor and the outcome, its effect should be ‘cleaned’ from the data by adding it as a predictor

Confounds: The Pipe

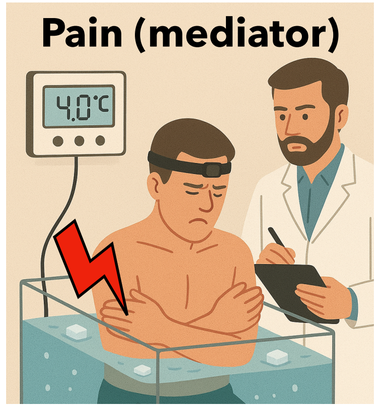
- What is the effect of cold exposure on brain perfusion?





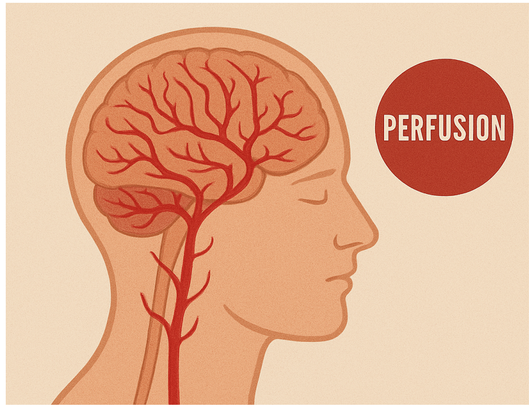
Confounds: The Pipe

- Let’s assume…



- The colder the water, the more likely the individual experiences pain

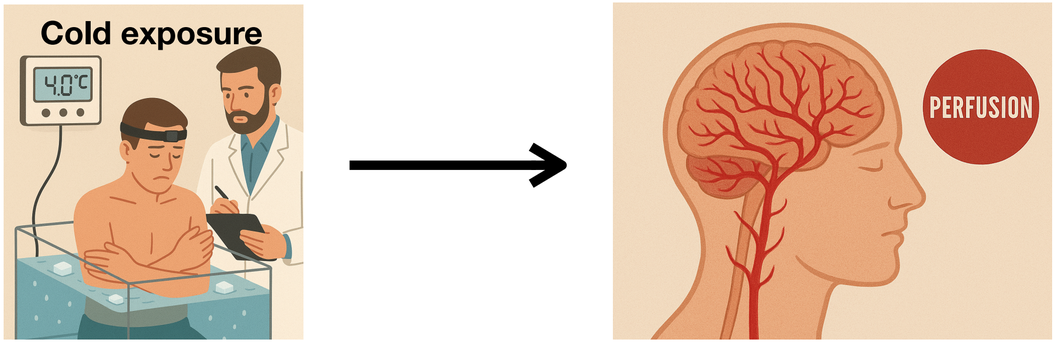
- Pain increases perfusion



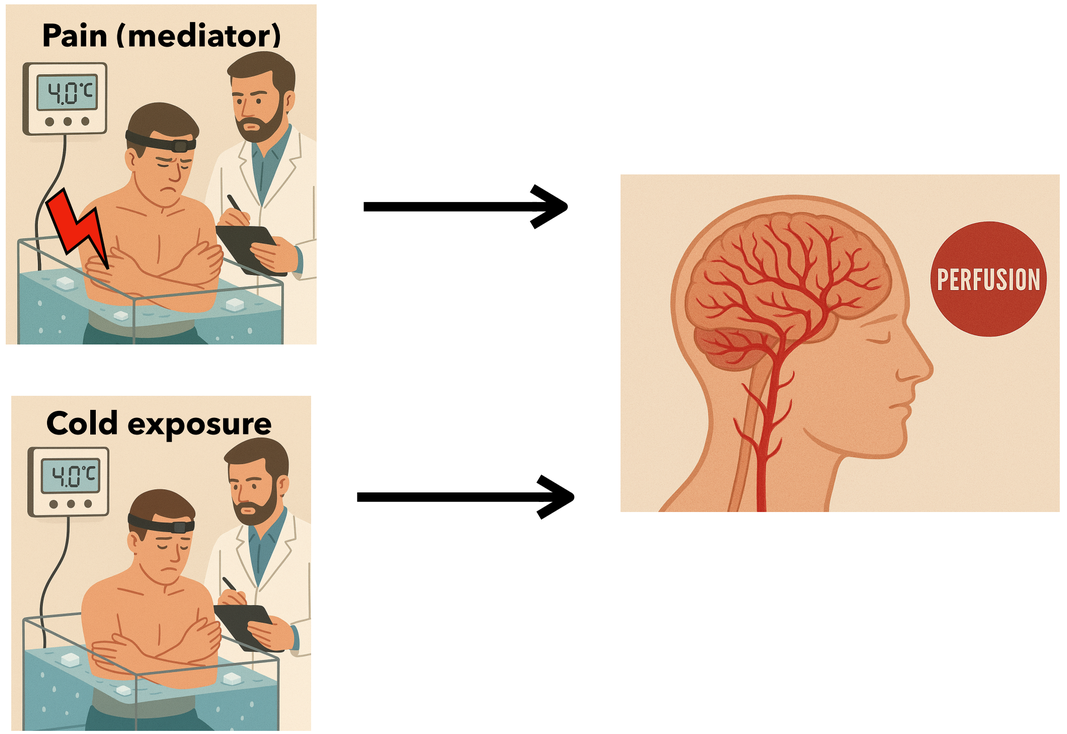
Confounds: The Pipe

- What is the **total effect of cold exposure** (including the pain-mediated effect) on brain perfusion?

Perfusion ~ cold exposure



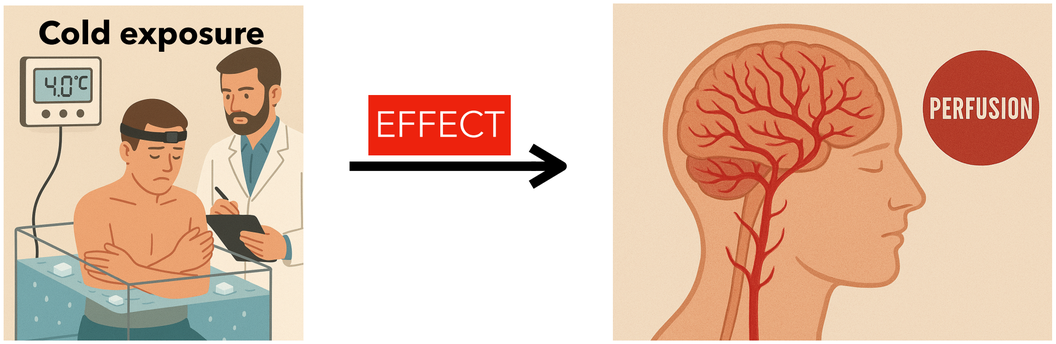
Perfusion ~ cold exposure + pain



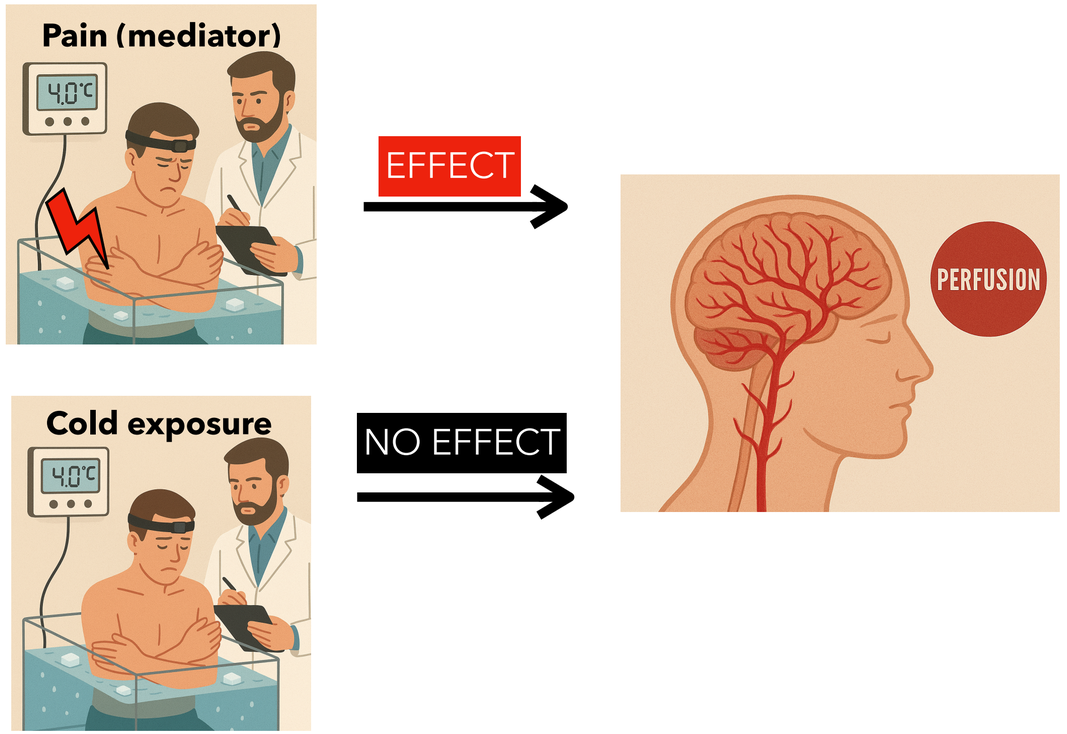
Confounds: The Pipe

- What is the **total effect of cold exposure** (including the pain-mediated effect) on brain perfusion?

Perfusion ~ cold exposure



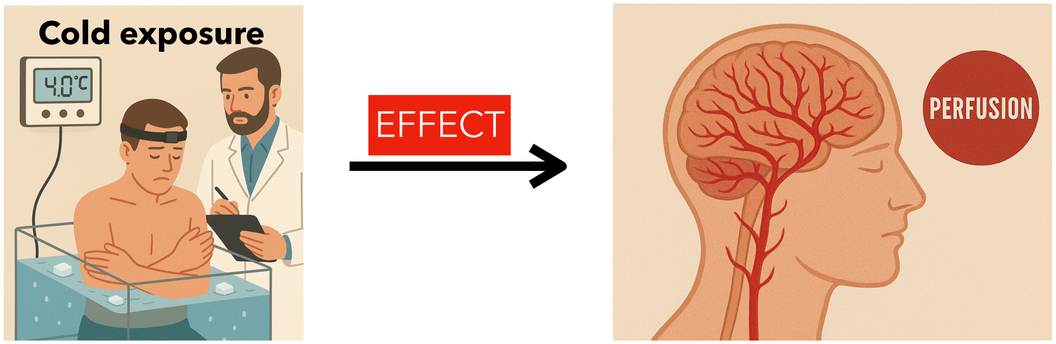
Perfusion ~ cold exposure + pain



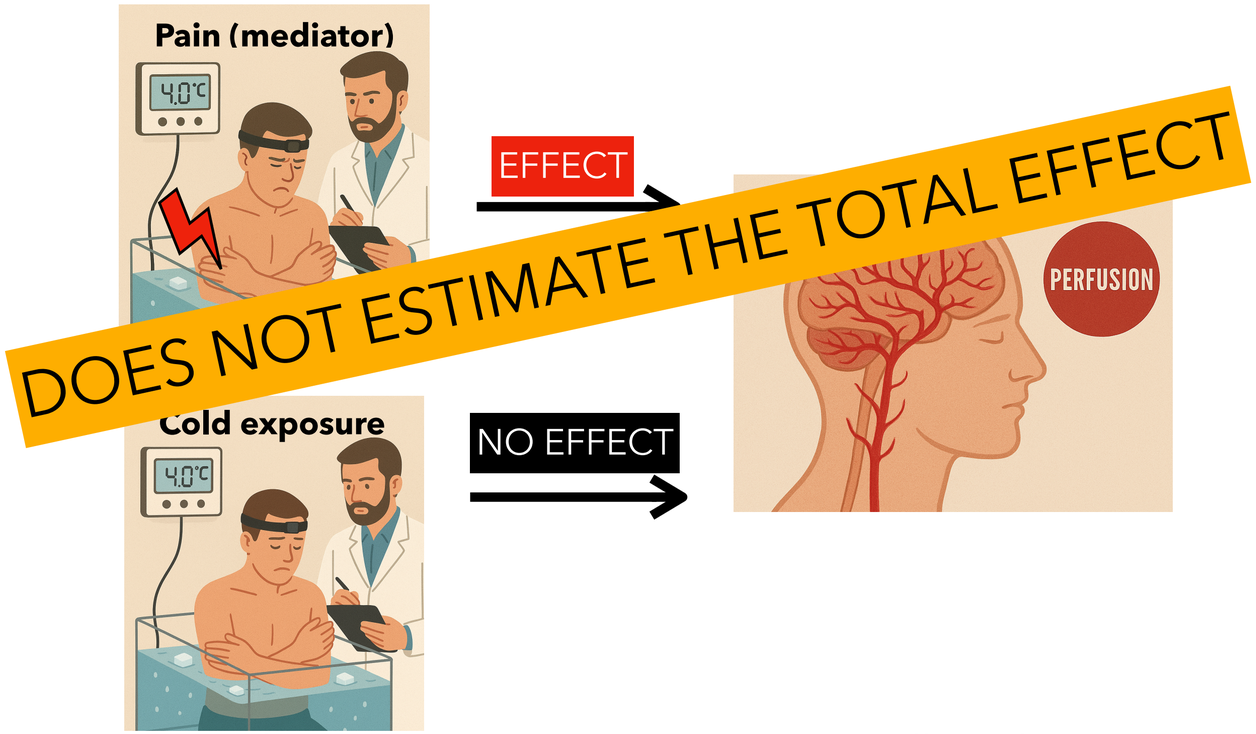
Confounds: The Pipe

- What is the **total effect of cold exposure** (including the pain-mediated effect) on brain perfusion?

Perfusion ~ cold exposure

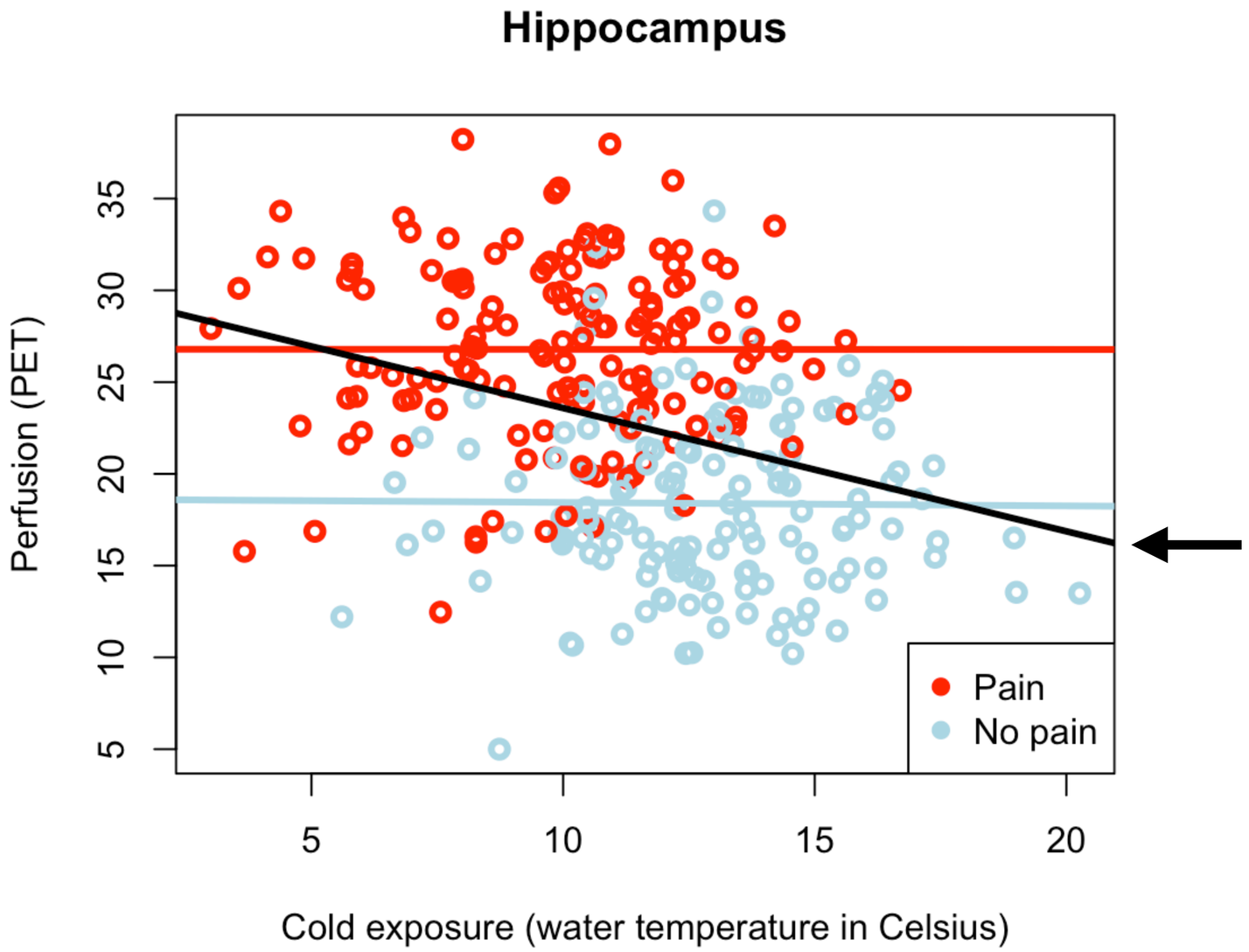


Perfusion ~ cold exposure + pain

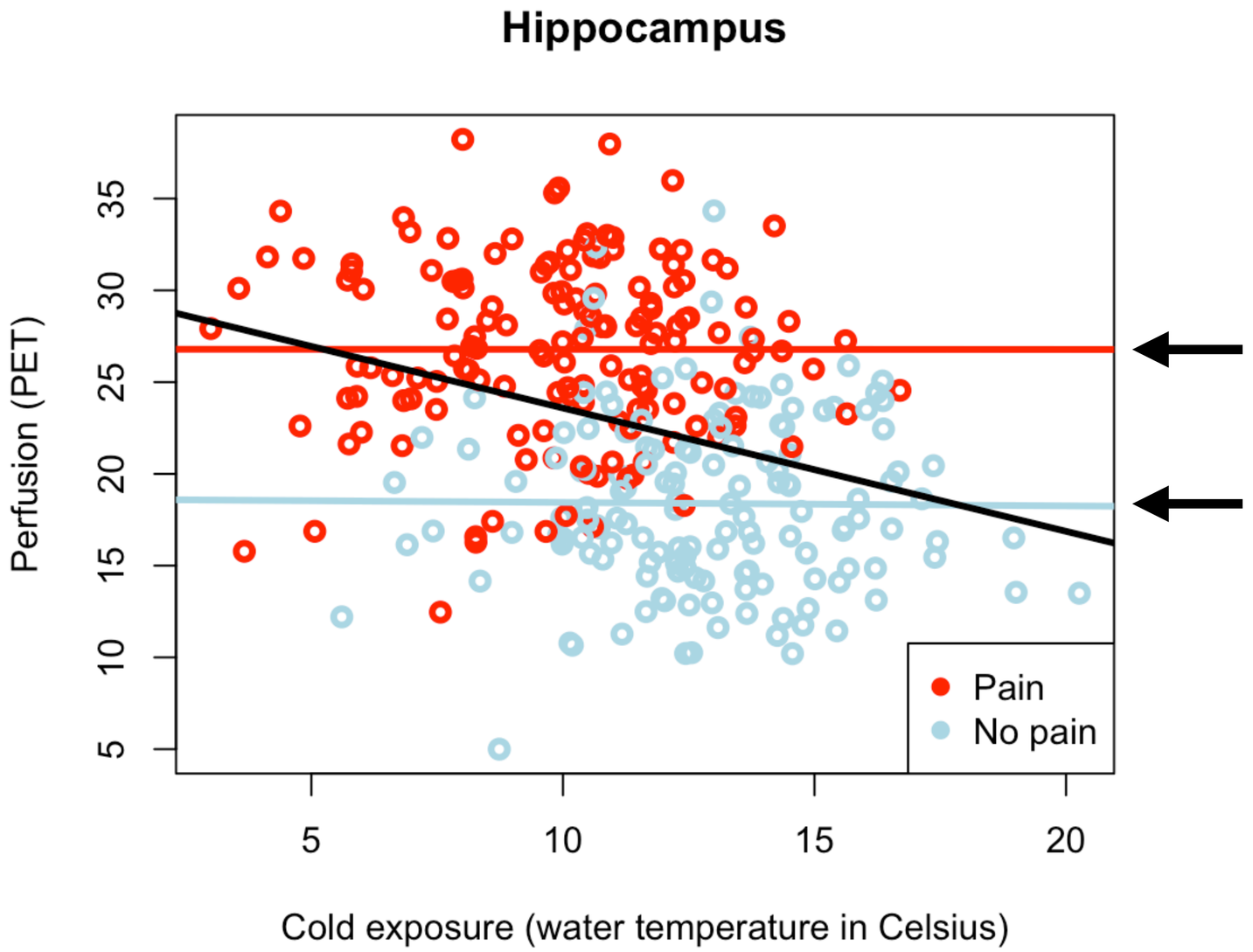


- Include pain if interested in the pain-independent effect of cold-exposure

Confounds: The Pipe



Confounds: The Pipe



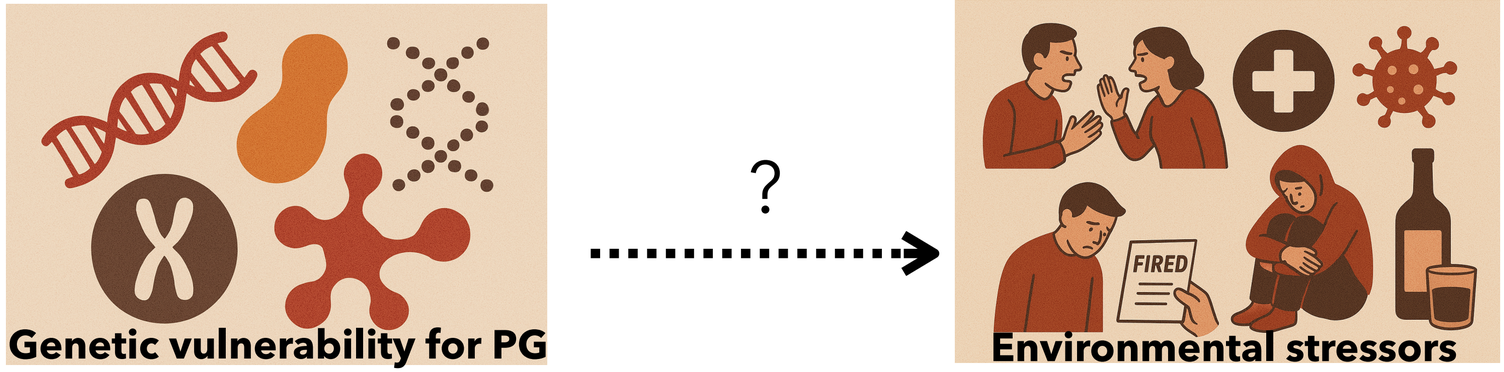
Confounds: The Pipe

The general rule of thumb

- **Let’s not include treatment consequences** as predictors if interested in the total treatment effect
  - Drug reduces heart rate that lowers anxiety
    - Anxiety ~ drug + heart rate
      - The drug doesn’t work (bad conclusion, post-treatment bias)

Confounds: The Collider

- What is the effect of genetic vulnerability for pathological gambling (PG) on environmental stressors?



Confounds: The Collider

- Let’s assume…

Genetic vulnerability for PG

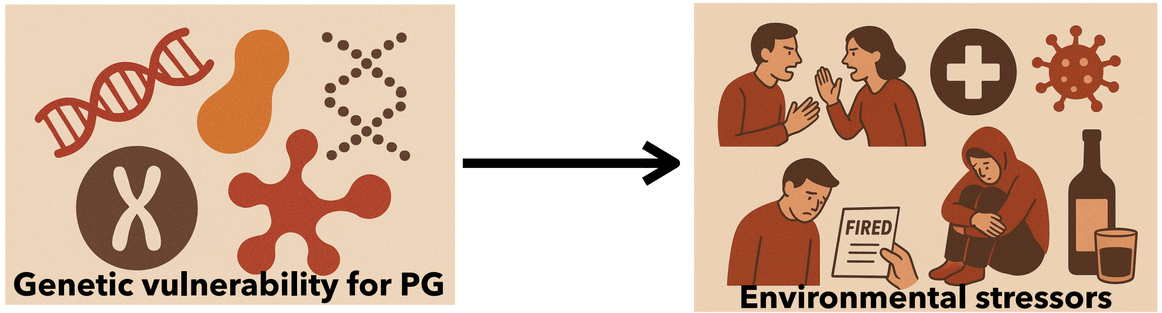
Environmental stressors

- Both genetic vulnerability and environmental stressors increase the likelihood of pathological gambling

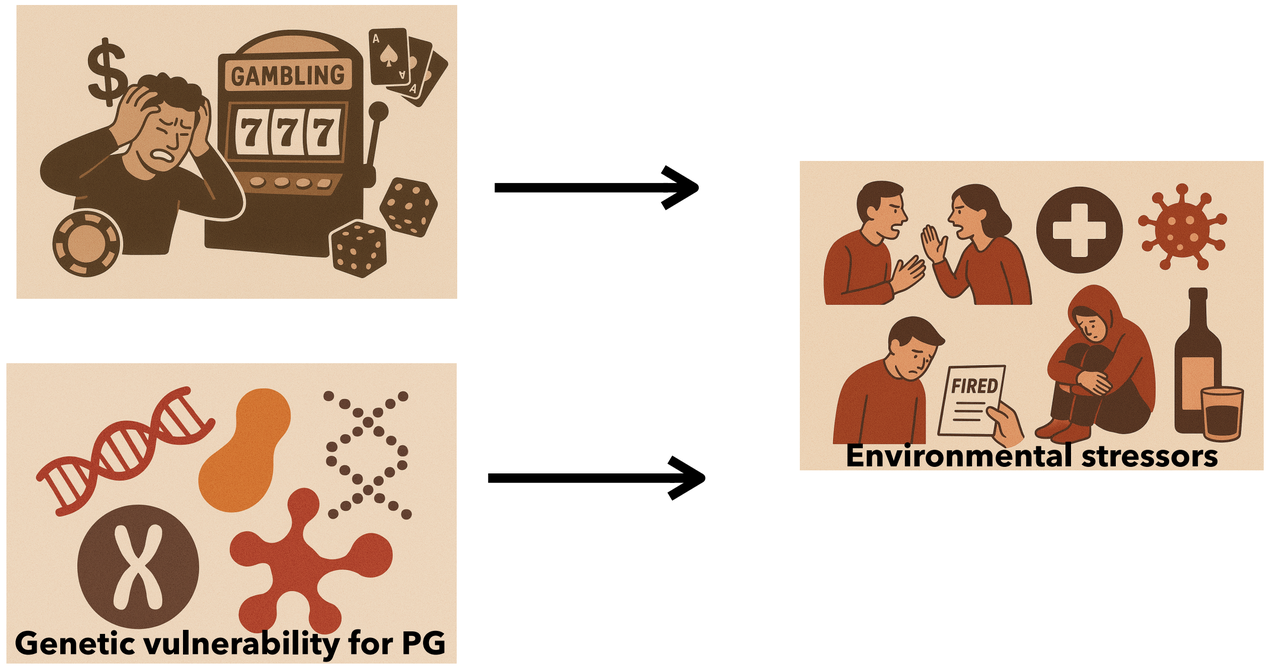
Confounds: The Collider

- What is the effect of genetic vulnerability on environmental stressors?

Environmental stressors ~ genetic vulnerability



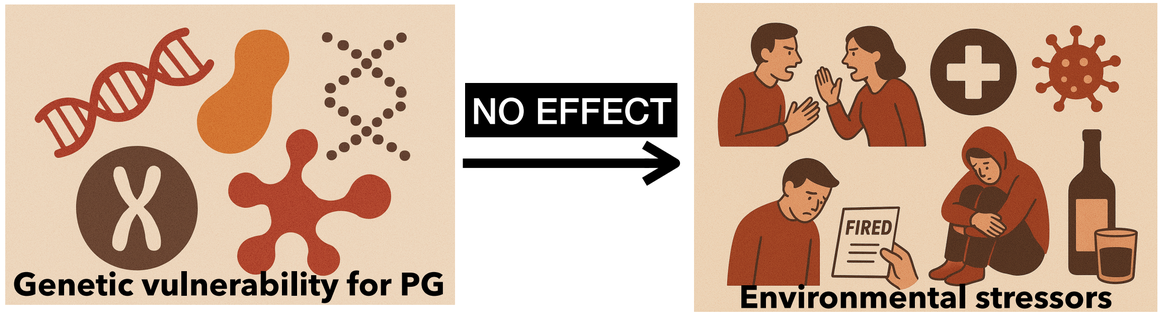
Environmental stressors ~ genetic vulnerability + clinical status



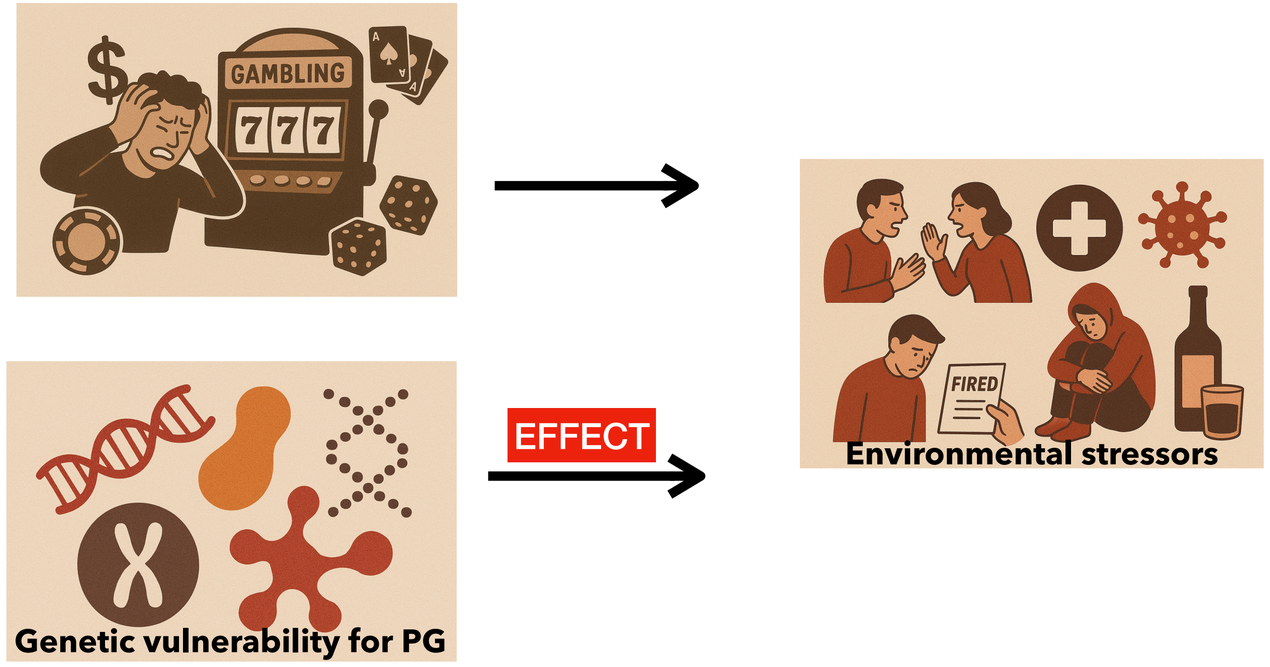
Confounds: The Collider

- What is the effect of genetic vulnerability on environmental stressors?

Environmental stressors ~ genetic vulnerability



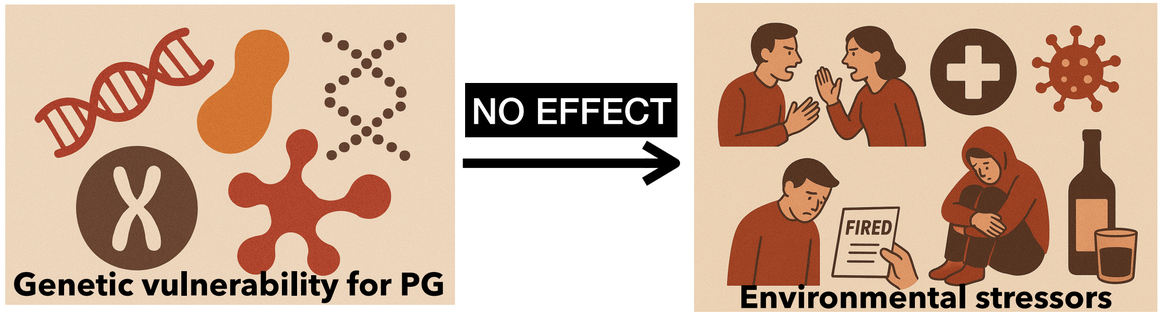
Environmental stressors ~ genetic vulnerability + clinical status



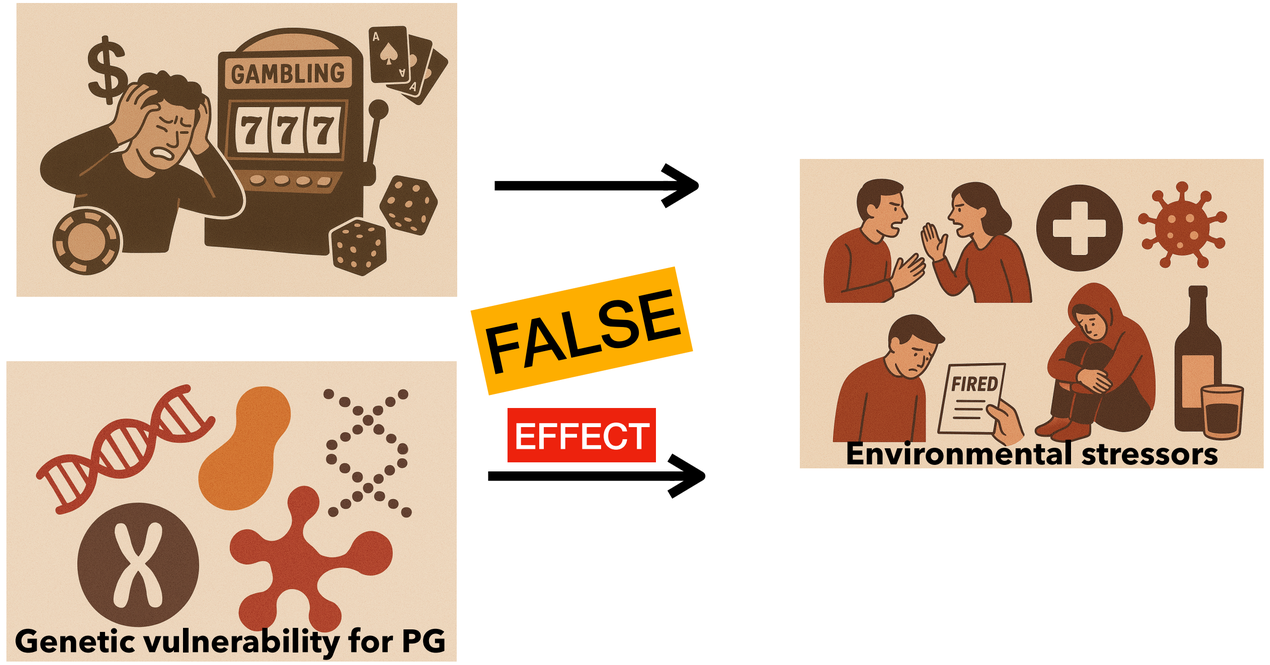
Confounds: The Collider

- What is the effect of genetic vulnerability on environmental stressors?

Environmental stressors ~ genetic vulnerability

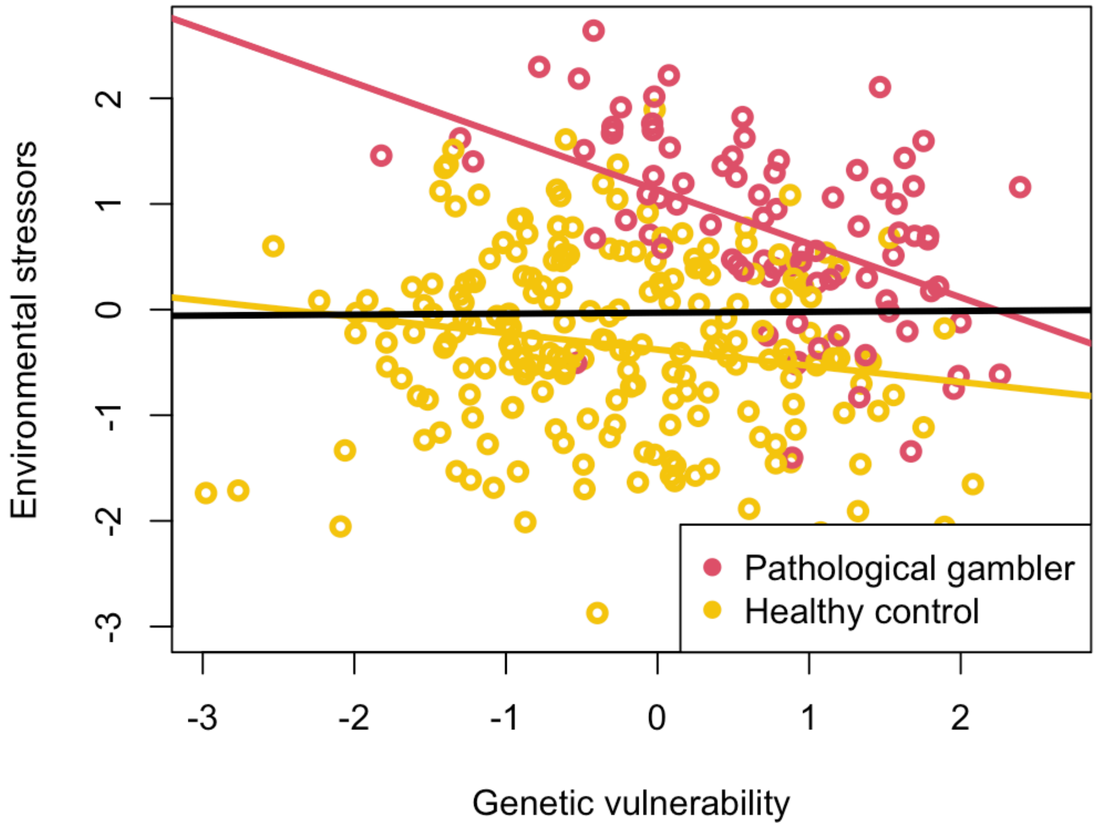


Environmental stressors ~ genetic vulnerability + clinical status



Confounds: The Collider

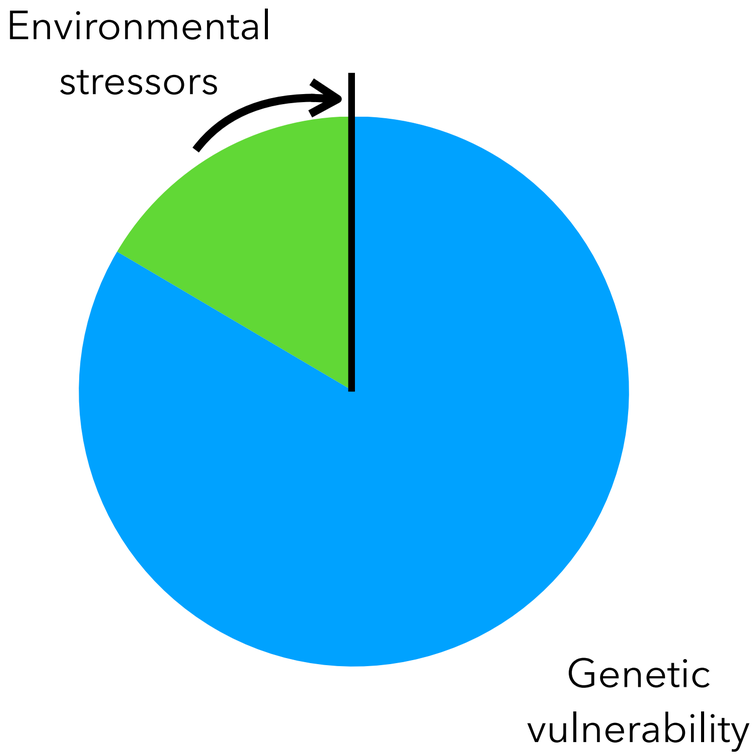
- What is the effect of genetic vulnerability on environmental stressors?



- Clinical status is a problematic collider
  - Clinical status (pathological gambler, healthy control) as a predictor
    - Creates a **fake association between genetic vulnerability and environmental stressors**

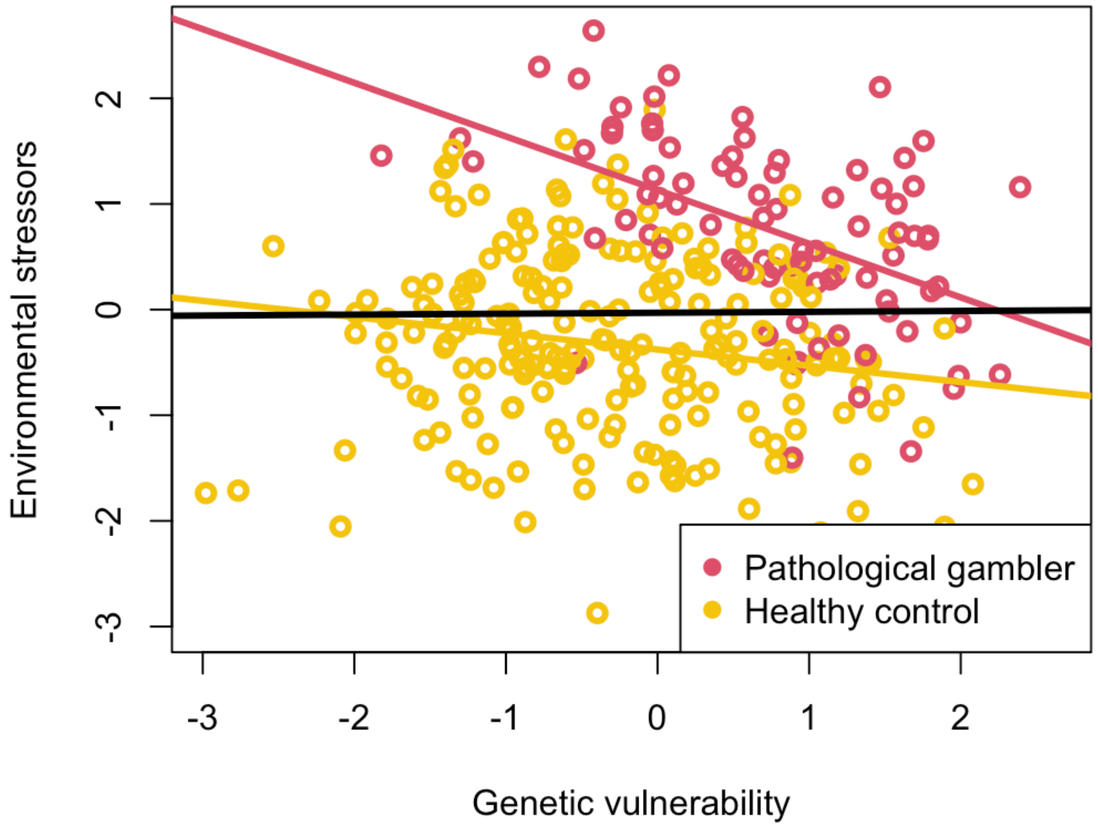
Confounds: The Collider

The thresholding effect



Confounds: The Collider

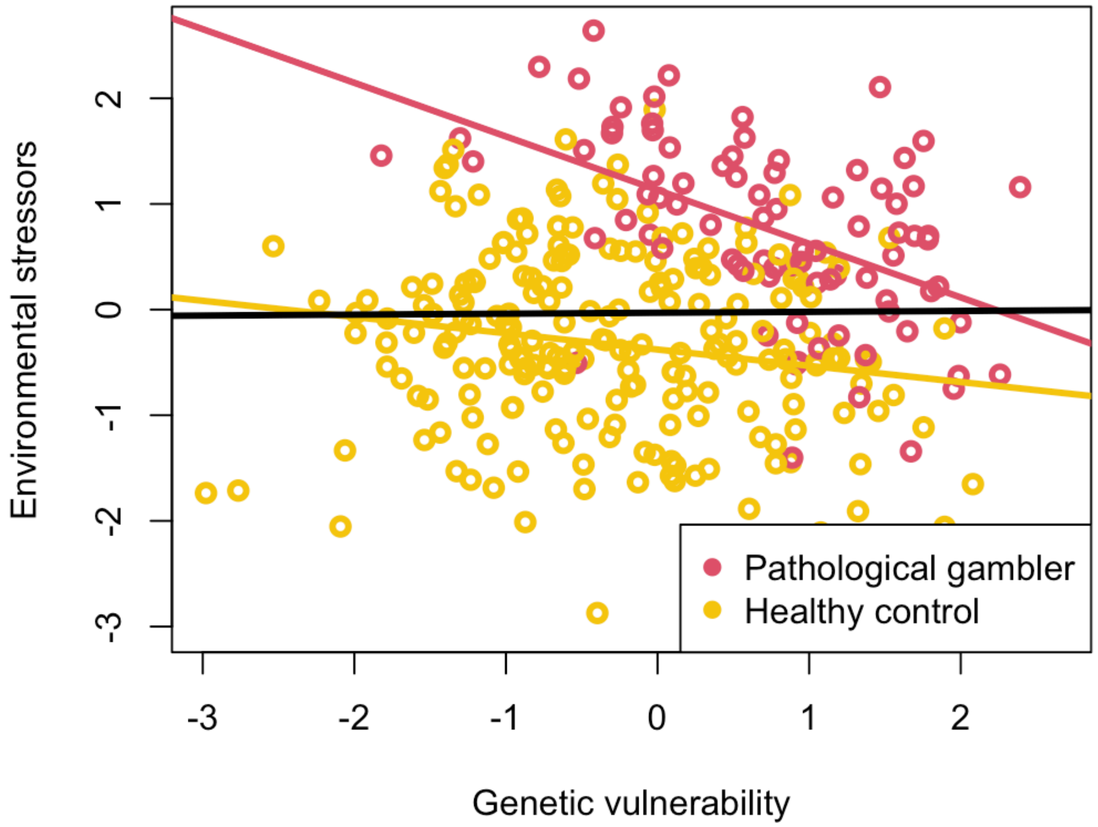
The thresholding effect



- If either genetic vulnerability or environmental stressors big enough, then onset of pathological gambling
- The higher the one, the lower the other needs to be for onset
- Either one must be high
- Looking at the genetic-environmental association at each level of clinical status, there is an arbitarily strong association

Confounds: The Collider

The thresholding effect

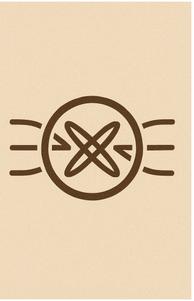


- So it can well be true, that subjects with PG show this ‘higher the one, lower the other’ -tendency
- Yet, it does not mean that genetic vulnerability for PG protects from environmental stressors (higher the vulnerability, lower the stressors)

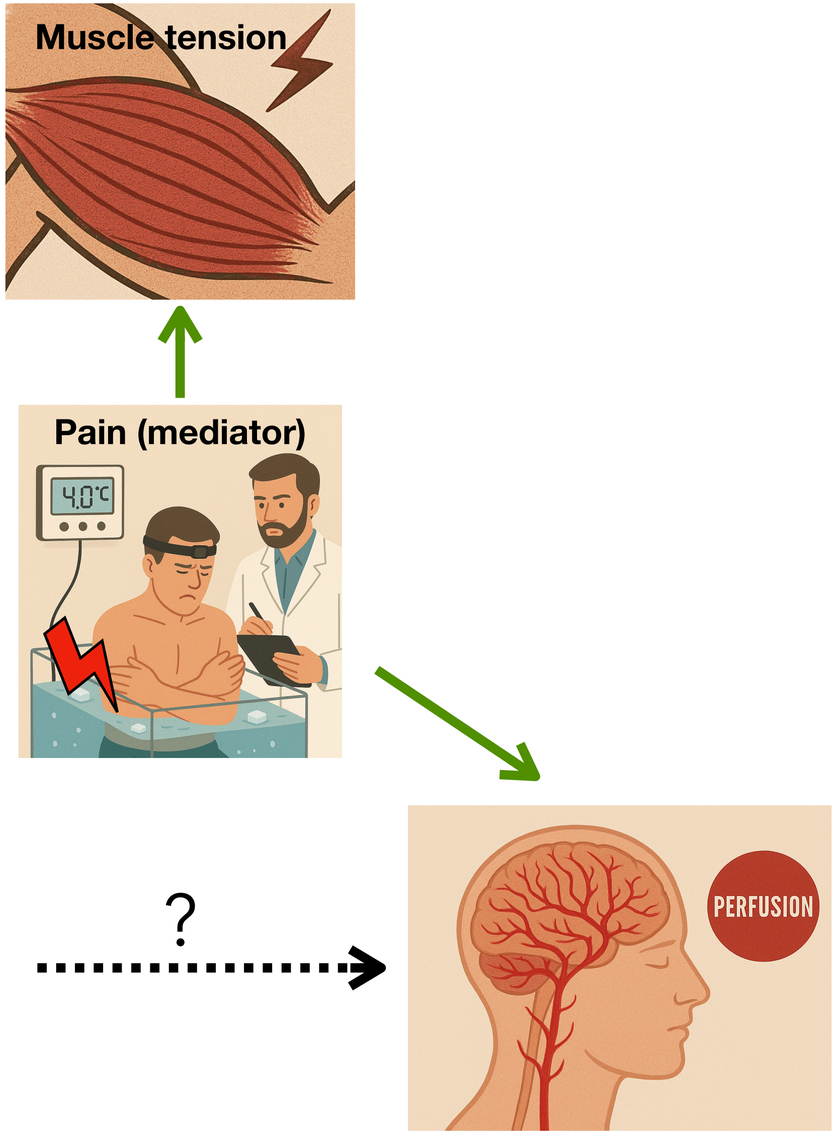
Confounds: The Collider

The general rule of thumb

- Let’s consider
  - The subpopulations we have in the data
  - In which population we want to assess the interesting effect
    - Visualizing data is always useful



Confounds: The Descendant



- Muscle tension is a descendant (child) for pain (parent)
- Behavior depends on where the descendant is attached to (here pipe)
- Including muscle tension like including its ‘parent’ pain, but a bit weaker - not a clone but includes some of the same information
- If interested in the total effect, leave out both the parent and the child
- If interested in the mediator-independent effect, I would include the parent but not child

Theory vs practice

- The DAGs make sense in theory. In practice, they may become extremely large and cryptic as the number of variables increases
- Often, we do not know the causal paths between the variables
  - If we did, would we study them?
- Awareness!

Useful practices

- **Correlation of the predictors, multicollinearity metrics (e.g. VIF)**
  - High correlations: Are the predictors measures of the same thing?
    - Yes: Can we choose only one of them or combine several variables into one umbrella variable (metabolic strain from body mass index, waist measurement, cholesterol…)?
- **Model comparison**
  - Modify the set of the predictors, then compare models: Essential changes in the findings?
    - No: No signs of major problems with the predictor combination, might be good to report the model comparison
    - Yes: Analyze more in detail and consider the causal paths

Underfitting & Overfitting

- Not too few, not too many predictors
- **Underfitting: Too few**
  - Too general
  - Is missing meaningful predictors that in real life influence the outcome
    - Missing the age effect in brain data, while neural functions and structures are clearly affected by age

Underfitting & Overfitting

- Not too few, not too many predictors
- **Overfitting: Too many**
  - Difficult to interpret: The causal paths become difficult to handle
  - Fits the current sample ’perfectly’ but does not generalize well in the population

‘Blindly tossing variables into the causal salad is never a good idea.’

- R. McElreath