Statistics foundations

Statistical Diversity Lab @ University of Washington
Amy Willis — @AmyDWillis — Associate Professor
Sarah V Teichman — Research Scientist
María Valdez — Postdoctoral Scholar
and
Sarah J Tucker — Postdoctoral Scholar (MBL)

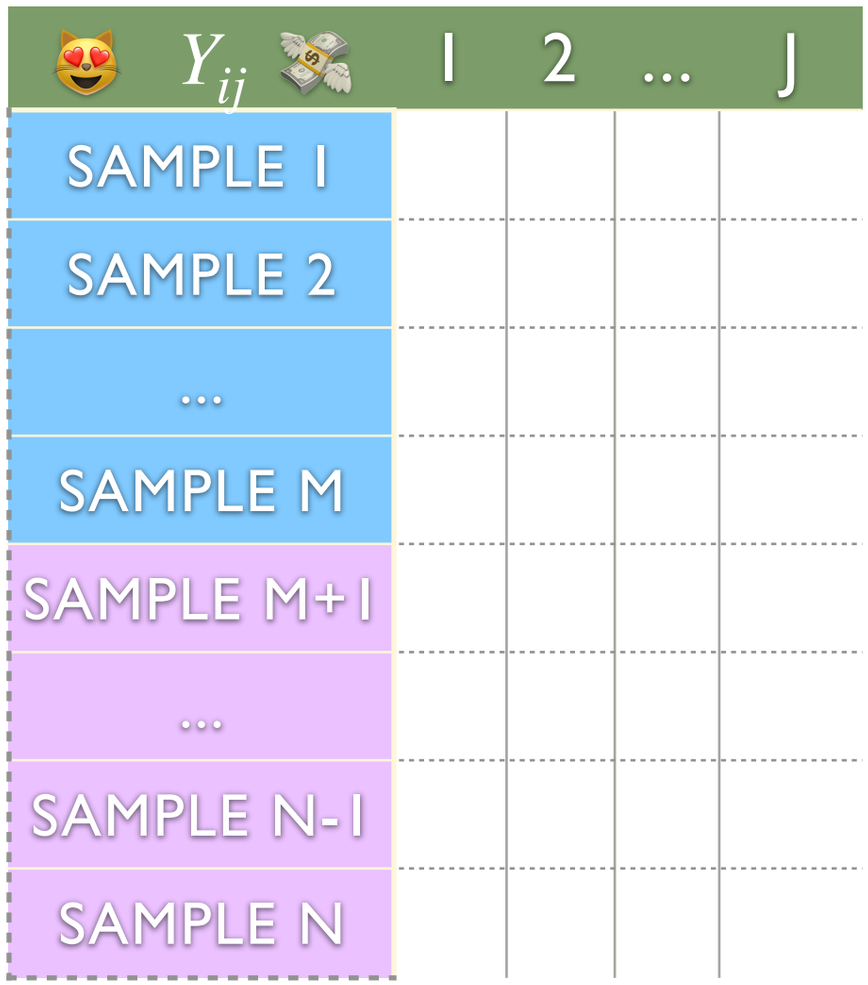
Slides at https://github.com/mblstamps/stamps2025/wiki

Context

Microbial universe

- Y_{ij} = true number of unit j in sample i

- e.g., there are 7,455,469 16S DNA copies/ml of *S epidermidis* on my finger



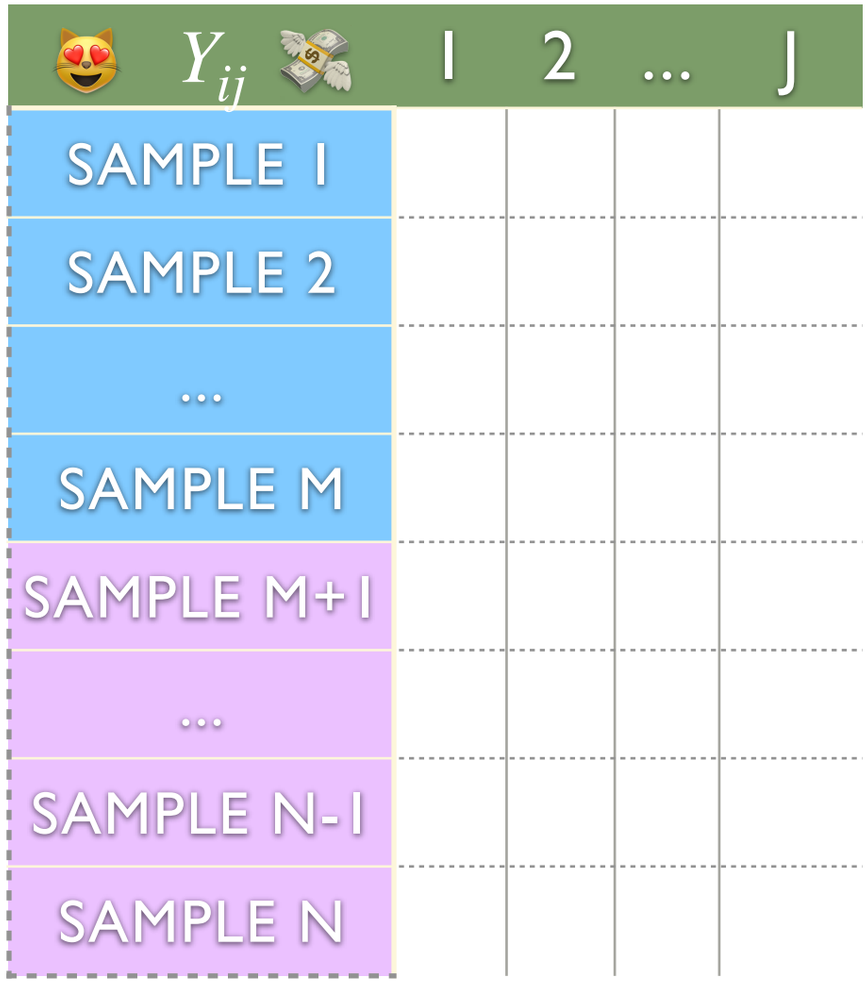
Question of the day

- What can (and can't) we learn about Y_{ij}'s from microbiome data?
  - qPCR and other "absolute" technologies
  - HTS
  - multiple data measurements

Review: Parameters

- **Parameters** are summaries of a data generating process

- *Functions* of Y_{ij}'s



Review: Data

- W_{ij} = number of times unit j observed in sample i

🌧️ W_{ij} 😿	1	2	…	J
SAMPLE 1				
SAMPLE 2				
…				
SAMPLE M				
SAMPLE M+1				
…				
SAMPLE N-1				
SAMPLE N				

🦉 *The* question🦚: How do we connect the W_{ij}'s to the $Y'_{ij}s$?

Review: Estimators

- Parameters are *unknown*

- We *estimate* parameters using our data

- We call these functions of our data *estimators*

Example:
Shannon diversity

- Shannon diversity is a *parameter*

$$\alpha_i := -\sum_{j=1}^{J} p_{ij} \log p_{ij} \qquad \text{for } p_{ij} := \frac{Y_{ij}}{\sum_{j=1}^{J} Y_{ij}}$$

- A *function* of the true, unknown Y_{ij}'s… thus, a parameter!

Example: Shannon diversity

- We can *estimate* Shannon diversity
- The most common estimator is the “plug-in” estimator

$$\hat{\alpha}_i := -\sum_{j=1}^{J} \hat{p}_{ij} \log \hat{p}_{ij} \text{ for } \hat{p}_{ij} := \frac{W_{ij}}{\sum_{j=1} W_{ij}}$$

- A *function* of the observed W_{ij}’s… thus, an estimator!

Estimators: notation

- The parameter *Amy*:

Estimators: notation

- An estimator of the parameter *Amy*:

Example: differences in log-ratios

- Here's a different parameter: $\beta_{j,j'}$

average of **treatment** samples' $\log\left(\frac{Y_{ij}}{Y_{ij'}}\right)$

minus

average of **control** samples' $\log\left(\frac{Y_{ij}}{Y_{ij'}}\right)$

Example: differences in log-ratios

- Having defined the parameter $\beta_{j,j'}$ as

average of **treatment** samples' $\log\left(\frac{Y_{ij}}{Y_{ij'}}\right)$

minus

average of **control** samples' $\log\left(\frac{Y_{ij}}{Y_{ij'}}\right)$

…come up with an estimator of $\beta_{j,j'}$

Bonus points: *any* justification for your estimator

Example: differences in log-ratios

- Who proposed $\hat{\beta}_{j,j'}$ to be

average of **treatment** samples' $\log\left(\frac{W_{ij}}{W_{ij'}}\right)$

minus

average of **control** samples' $\log\left(\frac{W_{ij}}{W_{ij'}}\right)$

…?

Estimators: properties

- Congratulations! You just came up with a good estimator!

- What makes an estimator *good*?

Estimators: properties

- Here are three estimators of our difference-of-ratios parameter

  A. what you came up with

  B. the sample average of the first half of your observations

  C. the number “7”

- Contrast A and B. Contrast A and C.

Evaluating estimators

- We want estimators to be

- Accurate = correct on average = unbiased / consistent

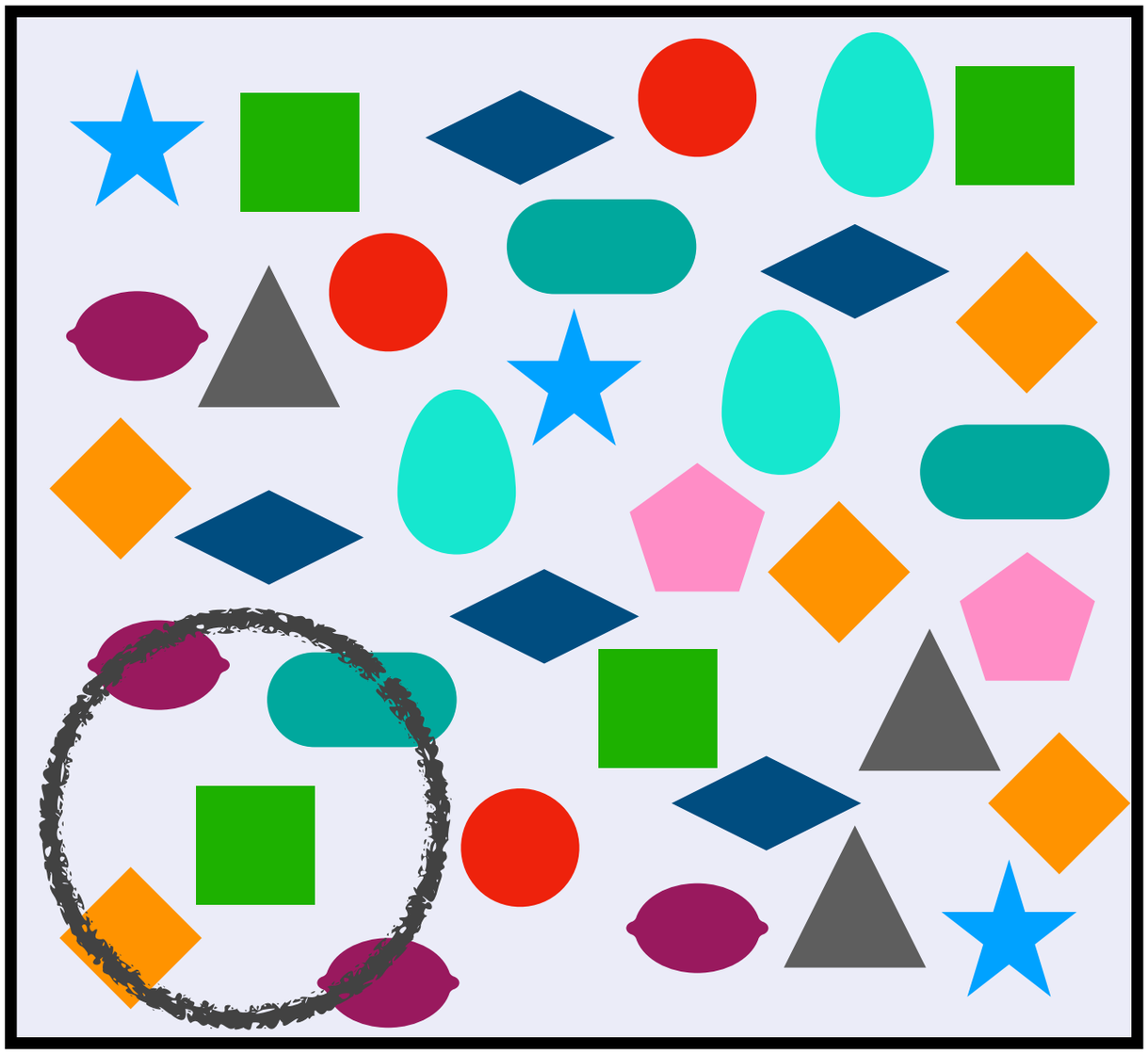
- Precise = usually close to their average = low variance

Bias

- Bias = average value of estimator − true value of parameter
  - e.g., average $\hat{\beta}_{j,j'} - \beta_{j,j'}$

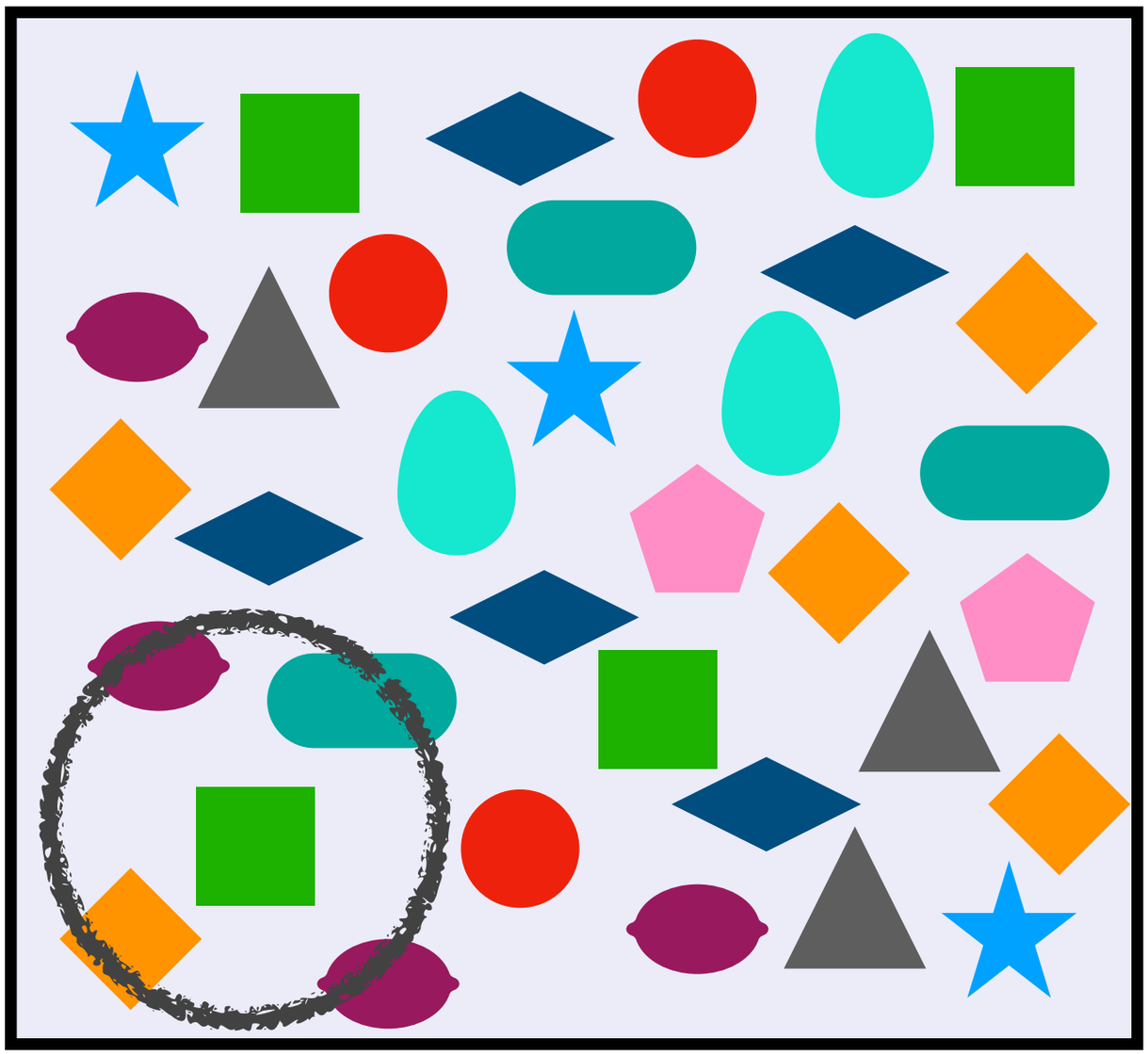
Bias:
species richness

- Parameter: total species richness
  - $C = 10$
- Estimator: observed species richness



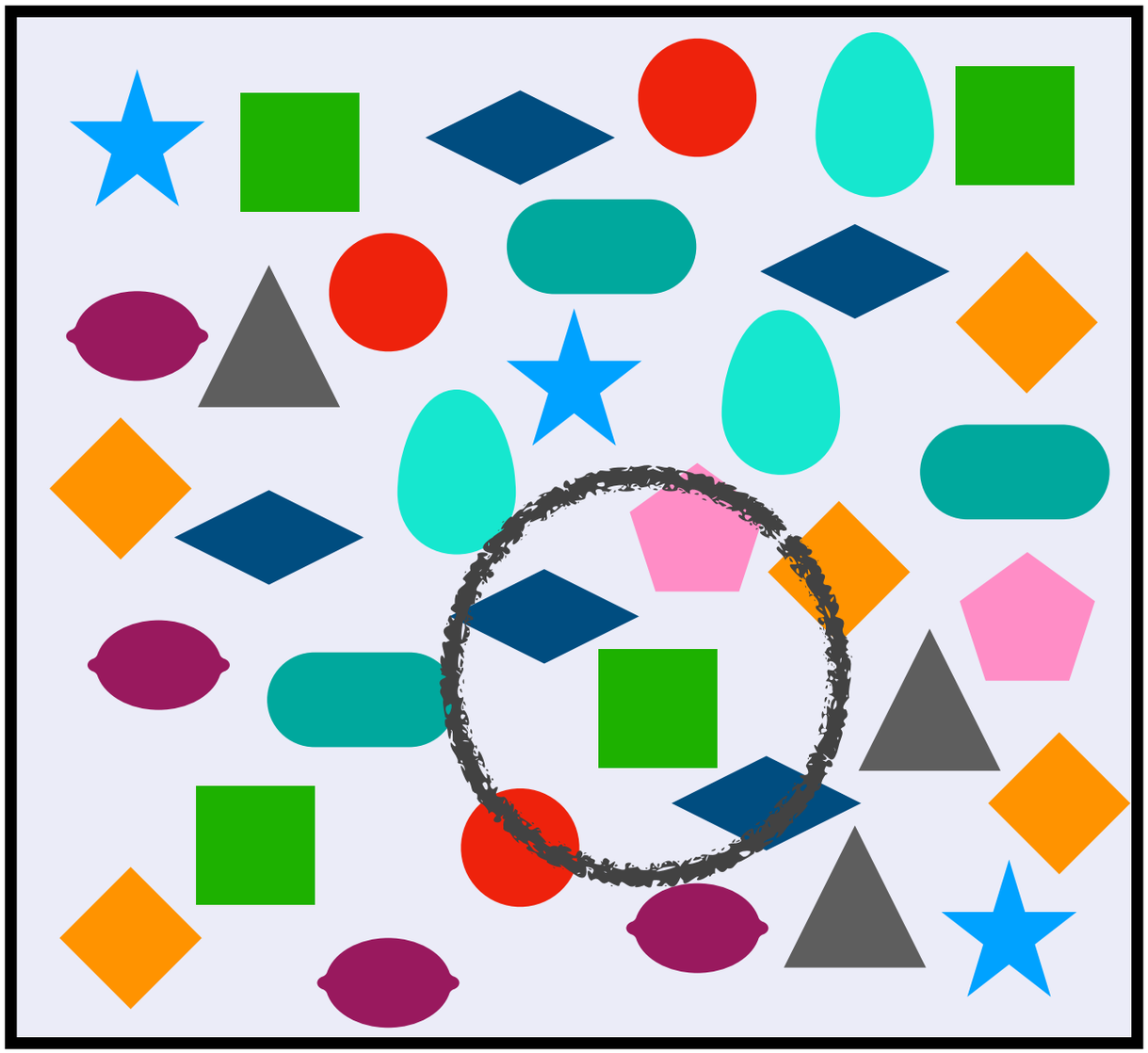
Bias:
species richness

- Parameter: total species richness
  - $C = 10$
- Estimator: observed species richness
  - $\hat{C} = 4$



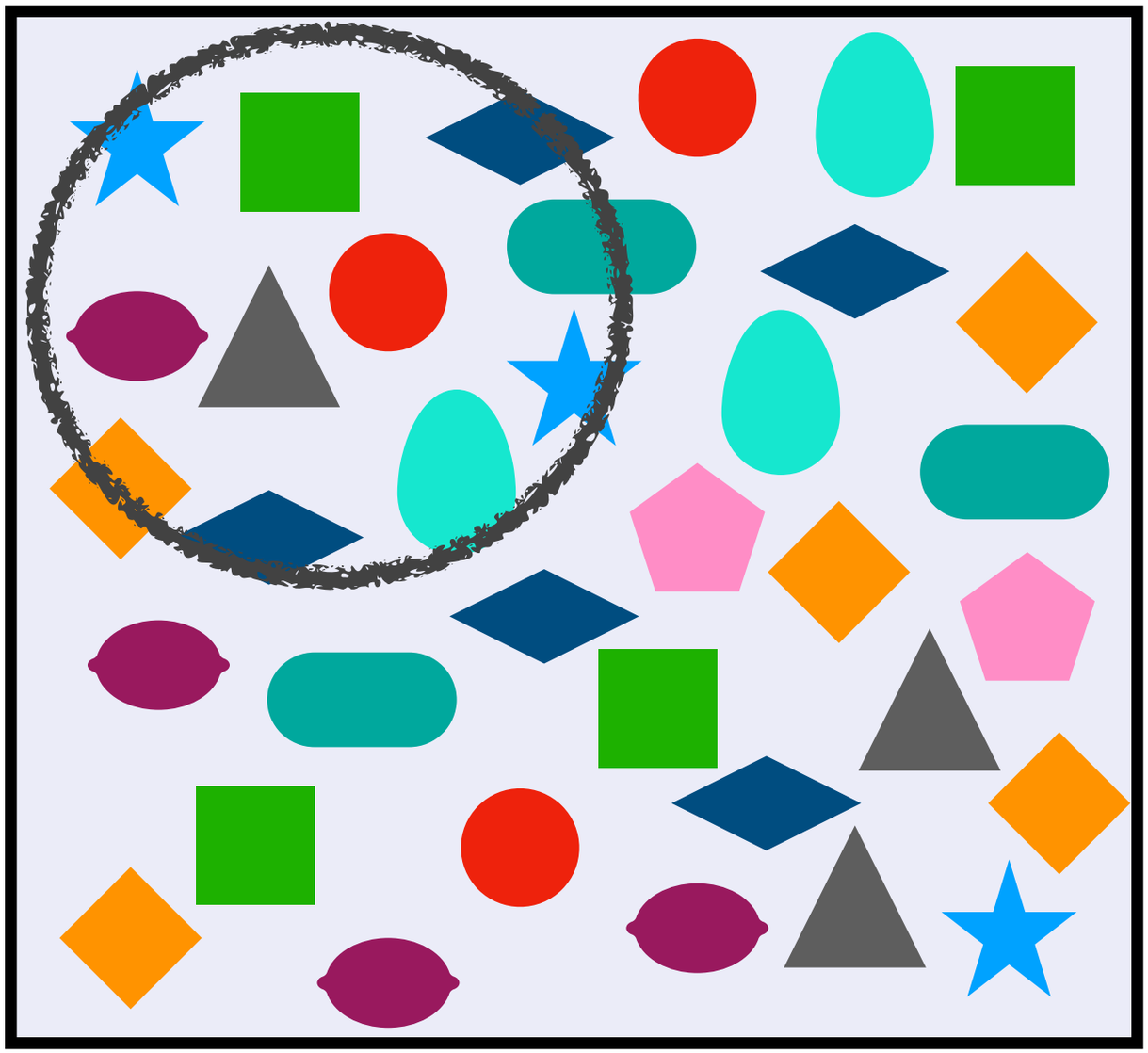
Bias:
species richness

- Parameter: total species richness
  - $C = 10$
- Estimator: observed species richness
  - $\hat{C} = 5$



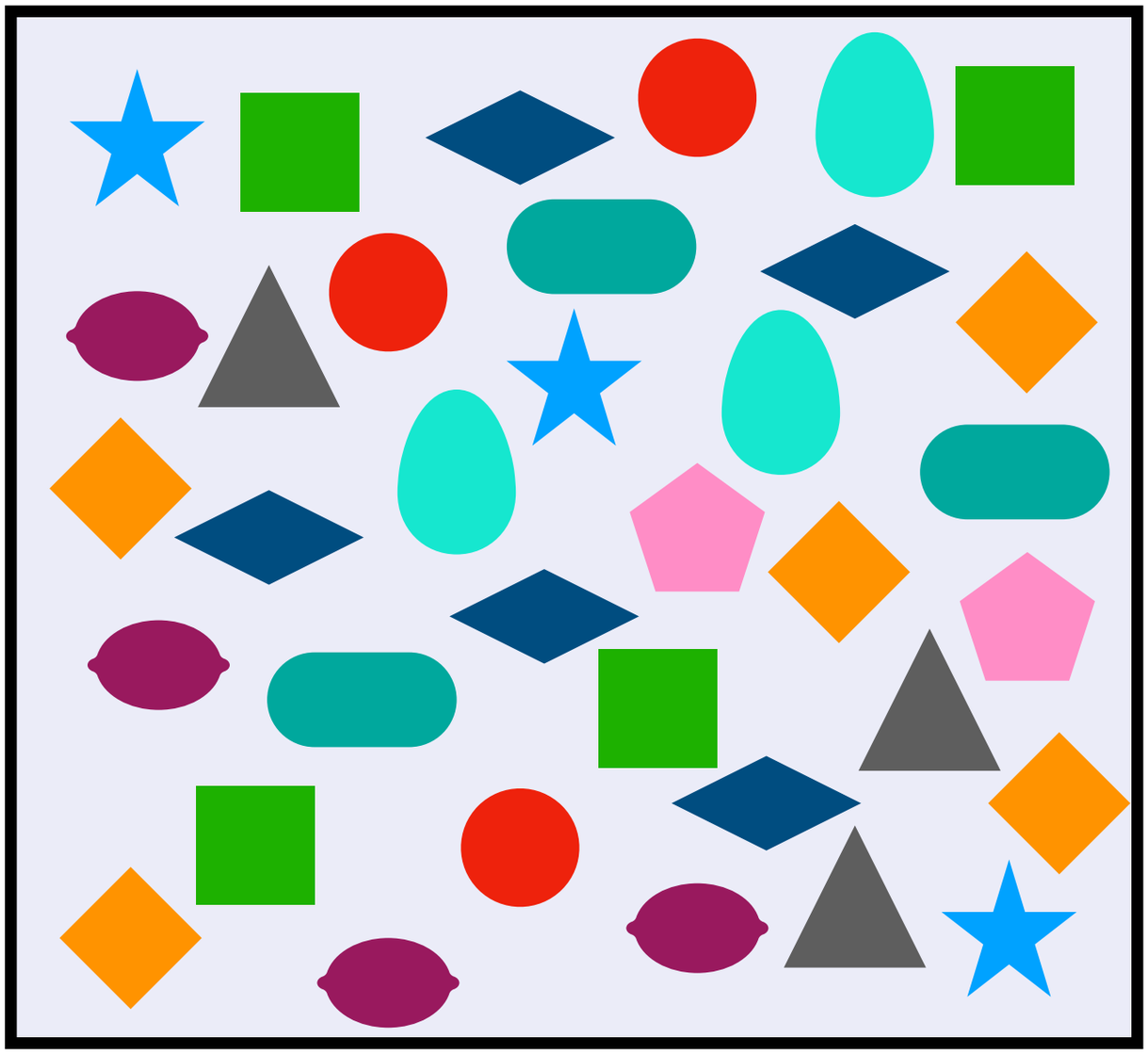
Bias:
species richness

- Parameter: total species richness
  - $C = 10$
- Estimator: observed species richness
  - $\hat{C} = 9$

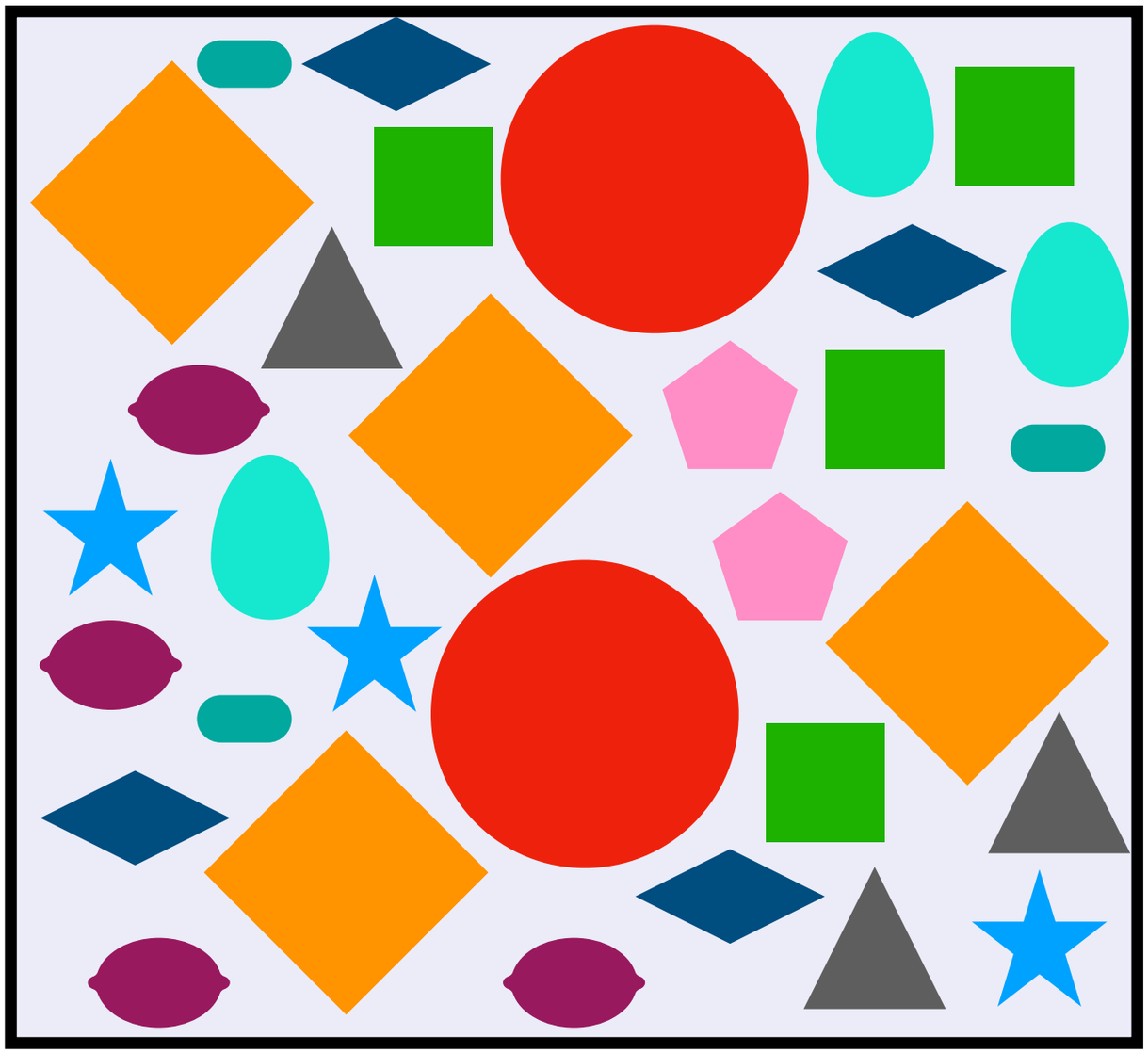


Bias:
species richness

Observed species richness is *negatively* biased = too small on average



Bias:
relative abundance



- Parameter: true relative abundance of

- Estimator: observed relative abundance of

Activity: Estimate the bias

Hint: there are 32 members of this community

Variance

variance = average of (estimator − average value of estimator)^2

- If estimates from repeated experiments are
  - 12, 12, 12, 12, 12… ➡️ variance is 0
  - 12, 12, 12, 13, 12… ➡️ variance is ~0.2
  - 12, 12, 12, 13013, 12… ➡️ variance is ~27 000 000

Variance

$$\text{variance} = \text{average of } (\text{estimator} - \text{average value of estimator})^2$$

- Hard to compare size of variance relative to estimate itself

$$\text{Standard deviation} = \sqrt{\text{variance}}$$

Variance

$$\text{variance} = \text{average of } (\text{estimator} - \text{average value of estimator})^2$$

$$\text{Standard deviation} = \sqrt{\text{variance}}$$

- Variances are *unknown*
- *True* variance vs *estimate* of variance

  Standard error = estimate of the standard deviation

Evaluating estimators

- We want estimators to be
  - Accurate = ~~unbiased~~ consistent
  - Precise = low variance
- It only makes sense to compare estimators of the same parameter
- Different estimators may be optimal under different *assumptions*

We'll come back to some specific cases: diversity, differential abundance…

Examples of assumptions

- We saw all the species that were present
- All species are equally easy to detect
- Amplicon counts follow a zero-inflated Negative Binomial distribution
- …

Examples of assumptions

- Taxa are consistently over/underdetected within a sequencing batch

- Measurements taken from different participants are independent

- The more deeply I sequence, the more likely I am to see something that’s present

Identifiability

- Assumptions aren't bad… they're *necessary*
- You need to make assumptions to make a parameter *identifiable*
  - Identifiable = able to be learned from the data

Identifiability

- Why can't we estimate Y_{ij} from W_{ij}? W_{ij} from HTS
- Because the assumptions needed to estimate Y_{ij} from W_{ij} aren't plausible…
  - They don't allow us to *identify* Y_{ij}
- We'll talk about this more this afternoon!

Y = true abundances, W = observed data

The life of a statistician

- Statisticians do the following
  - Choose assumptions
  - Show that the parameter is identifiable using the data + assumptions
  - Derive an estimator
  - Write software & make it useful for others
- These steps allow us to *learn about the universe* while understanding the *limitations of our methods*

The life of a microbial ecologist

- Choose a parameter meaningful & identifiable under reasonable assumptions

- Choose a sensible estimator

- Communicate the estimate of the parameter, and a measure of its uncertainty

- (If appropriate) Perform a valid test about the value of the parameter

Recap

- Everyone here cares about different things…
    - Presence of ARGs
    - Abundance of *Fusobacteria*
    - The diversity of protists
    - …
- These are different *parameters*

Recap

- The difference between parameters and estimators is not widely appreciated
    - This makes it hard to have a rational conversation about better and worse approaches
- Microbial ecologists may take for granted that there is only one way to estimate parameters…
    - Plug-in estimates
    - Black box estimates

    …and are often left out of the conversation about what assumptions are needed and reasonable

The plan

- This framework will guide us for the next 48 hours or the rest of your lives…
- I've used examples to illustrate specific concepts
- I haven't recommended specific estimators… I will!

The plan

- Now
    - Regression models
- This afternoon
    - Inference
    - Abundance
- Tomorrow
    - Trees
    - Expression & abundance
    - Diversity

Questions?

Estimating comparative parameters

aka regression

Comparative parameters

- Parameters can summarise one or many groups, e.g…
  - One group: average Shannon diversity
  - Two groups:

average Shannon diversity in group 1

minus

average Shannon diversity in group 2

Regression models

- Regression models take the form

  functional of outcome variable = function of predictor variables

- Common regression models
  - Linear, logistic, Poisson…
  - …

Regression models

- Regression models take the form

  functional of outcome variable = function of predictor variables

- e.g.,

- expected diversity$_i = \beta_0 + \beta_1 \times \mathbf{1}_{\{i \text{ is from lakewater (not seawater)}\}}$

  - $\hat{\beta}_0$ is an estimate of the average diversity in seawater environments

  - $\hat{\beta}_1$ is an estimate of the difference in average diversity in lake vs seawater environments

Regression models

- Regression models take the form

  functional of outcome variable = function of predictor variables

- e.g.,

- expected diversity$_i$ $= \beta_0 + \beta_1 \times \mathbf{1}_{\{i \text{ is from lakewater (not seawater)}\}}$

  - $\hat{\beta}_0$ is an estimate of the average diversity in seawater environments

  - $\hat{\beta}_1$ is an estimate of the difference in average diversity in lake vs seawater environments

Regression models

functional of outcome variable = function of predictor variables

- e.g.,
- expected diversity$_i = \gamma_0 \times e^{\gamma_1 \times \mathbf{1}_{\{i \text{ is from lakewater}\}}}$
  - $\hat{\gamma}_0$ is an estimate of the average diversity in seawater environments
  - $\hat{\gamma}_1$ is an estimate of the *log fold-difference* in average diversity in lake vs seawater environments

Regression models

- expected diversity$_i = \beta_0 + \beta_1 \times \mathbf{1}_{\{i \text{ is from lakewater (not seawater)}\}}$

$\beta_1 =$ expected diversity in lakewater
minus
expected diversity in seawater

- expected diversity$_i = \gamma_0 \times e^{\gamma_1 \times \mathbf{1}_{\{i \text{ is from lakewater}\}}}$

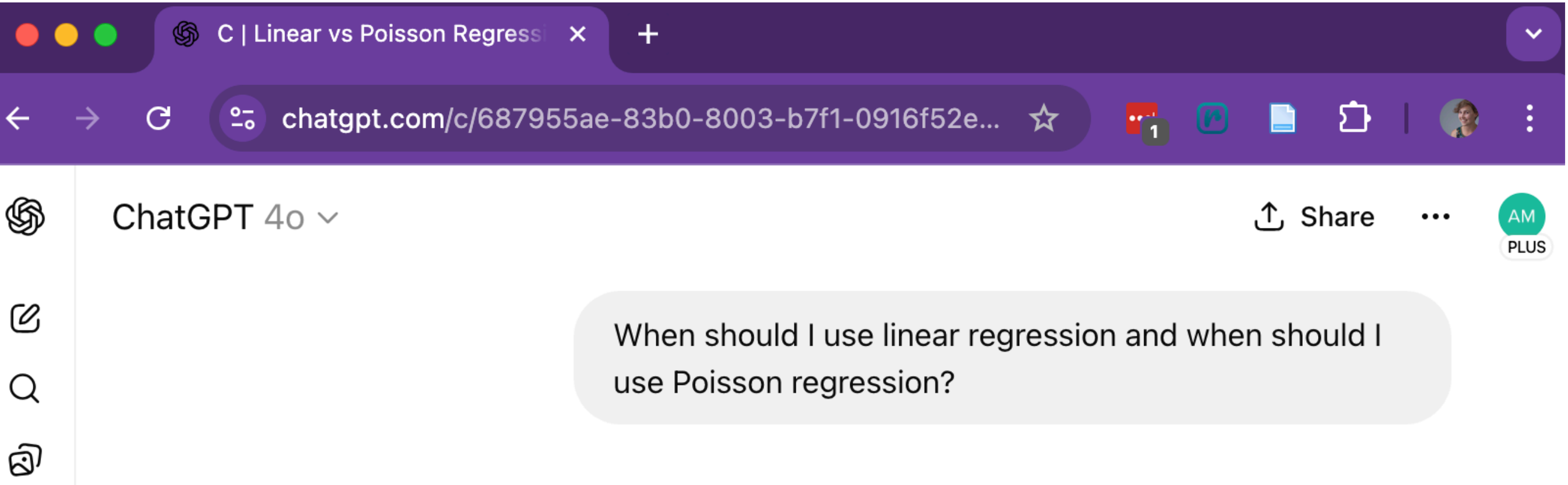
$$e^{\gamma_1} = \frac{\text{expected diversity in lakewater}}{\text{expected diversity in seawater}}$$

Regression models

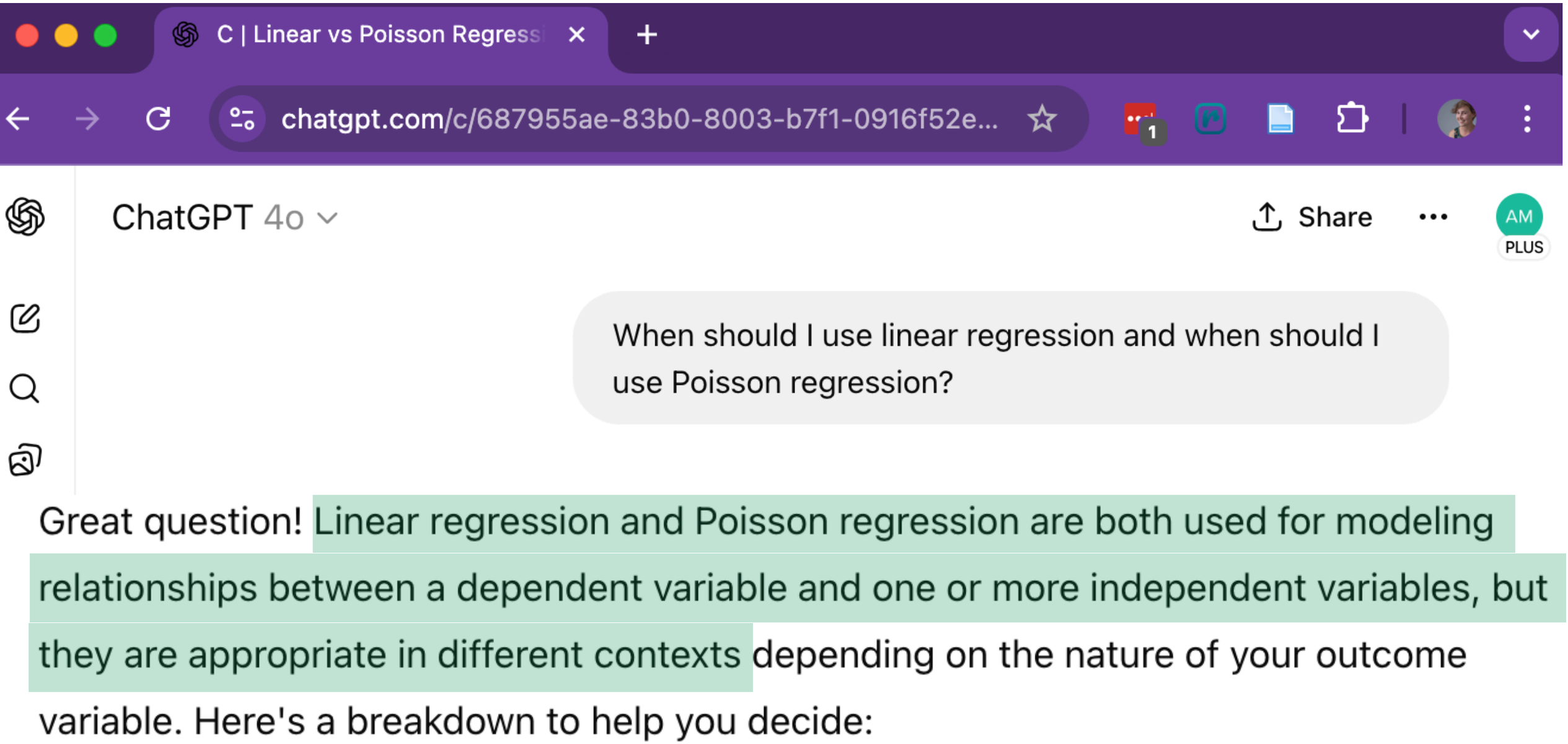
β_1 = expected diversity in lakewater
minus
expected diversity in seawater

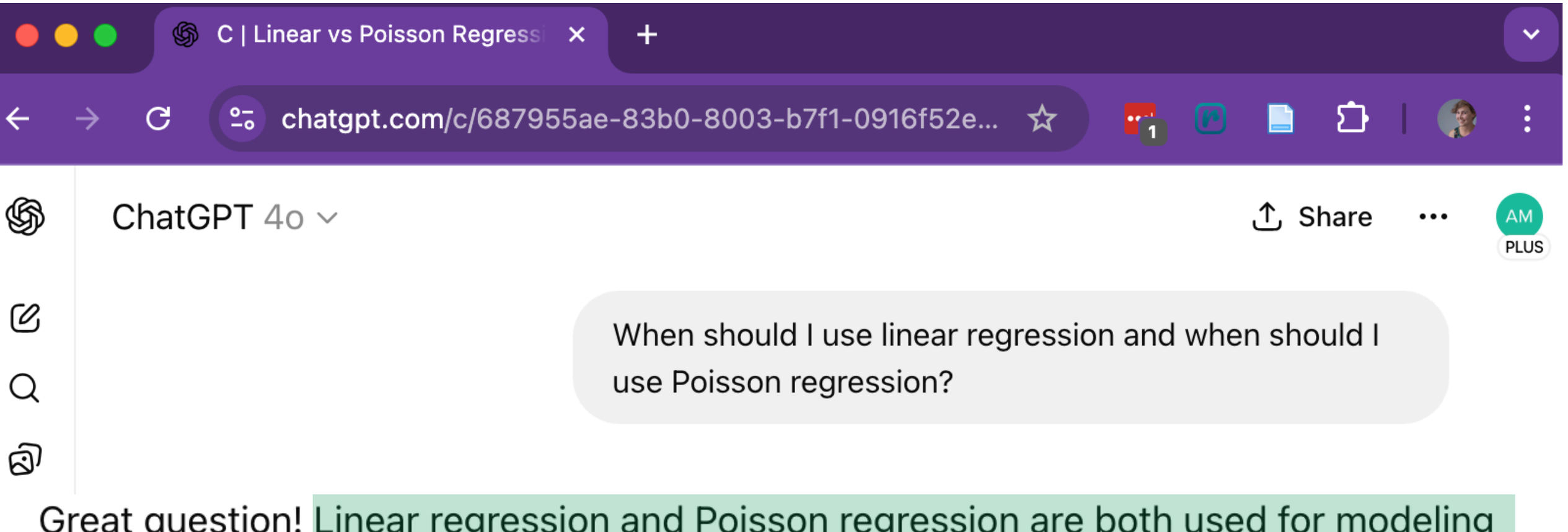
$$e^{\gamma_1} = \frac{\text{expected diversity in lakewater}}{\text{expected diversity in seawater}}$$

- You can use linear regression to estimate β_1
- You can use Poisson regression to estimate γ_1 (or e^{γ_1})

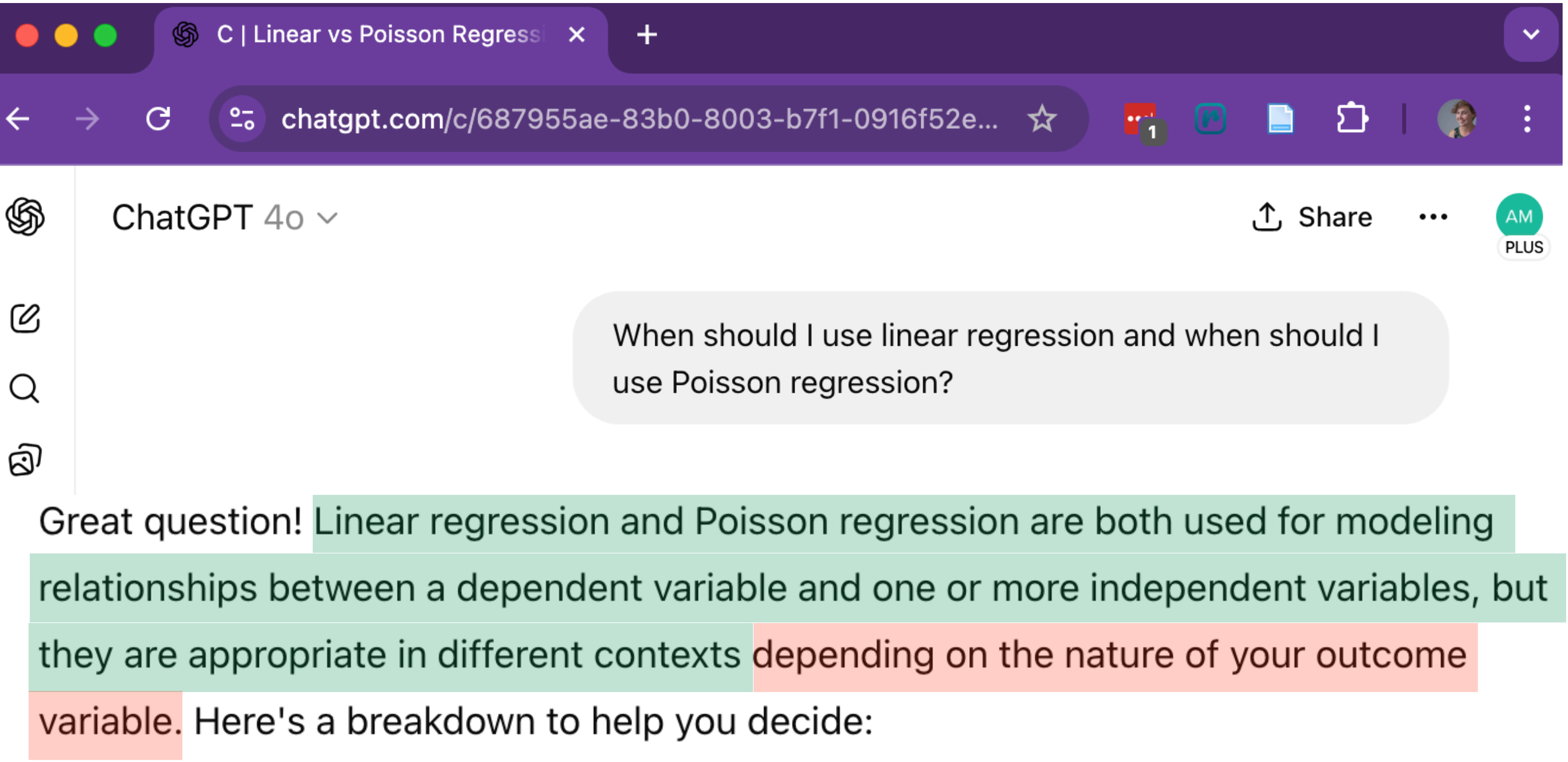


Great question! Linear regression and Poisson regression are both used for modeling relationships between a dependent variable and one or more independent variables, but they are appropriate in different contexts depending on the nature of your outcome variable. Here's a breakdown to help you decide:

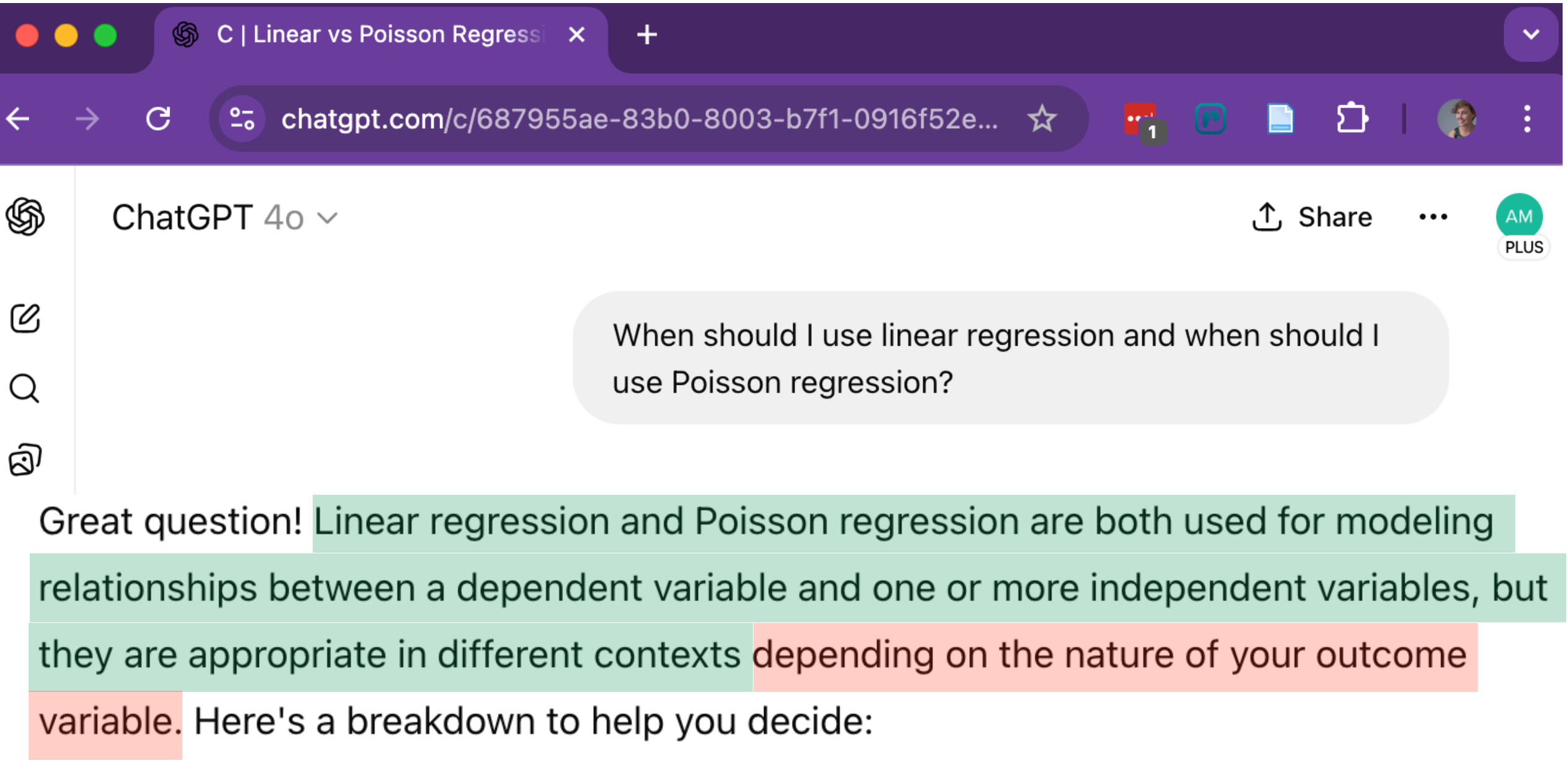

C | Linear vs Poisson Regressi
chatgpt.com/c/687955ae-83b0-8003-b7f1-0916f52e...
ChatGPT 4o
Share
AM
PLUS
When should I use linear regression and when should I use Poisson regression?
Great question! Linear regression and Poisson regression are both used for modeling relationships between a dependent variable and one or more independent variables, but they are appropriate in different contexts depending on the nature of your outcome variable. Here's a breakdown to help you decide:


C | Linear vs Poisson Regressi
chatgpt.com/c/687955ae-83b0-8003-b7f1-0916f52e...
ChatGPT 4o
Share
AM
PLUS
When should I use linear regression and when should I use Poisson regression?
Great question! Linear regression and Poisson regression are both used for modeling relationships between a dependent variable and one or more independent variables, but they are appropriate in different contexts depending on the nature of your outcome variable. Here's a breakdown to help you decide:

Yes — true!


C | Linear vs Poisson Regressi
chatgpt.com/c/687955ae-83b0-8003-b7f1-0916f52e...
ChatGPT 4o
Share
AM
PLUS
When should I use linear regression and when should I use Poisson regression?
Great question! Linear regression and Poisson regression are both used for modeling relationships between a dependent variable and one or more independent variables, but they are appropriate in different contexts depending on the nature of your outcome variable. Here's a breakdown to help you decide:

Yes — true!



Yes — true!

No, that’s rubbish, that doesn’t matter

🔷 Use Linear Regression When:

- **Your outcome (dependent) variable is continuous**
    - Examples: height, weight, blood pressure, test scores, income, temperature.
- **The residuals (errors) are approximately normally distributed and homoscedastic** (i.e., constant variance).
- **Negative and positive values are both possible**
    - Linear regression allows the outcome to be any real number.

🔷 Use Linear Regression When:

- **Your outcome (dependent) variable is continuous**
    - Examples: height, weight, blood pressure, test scores, income, temperature.
- **The residuals (errors) are approximately normally distributed and homoscedastic** (i.e., constant variance).
- **Negative and positive values are both possible**
    - Linear regression allows the outcome to be any real number.

false, not needed

🔷 Use Linear Regression When:

- **Your outcome (dependent) variable is continuous**
    - Examples: height, weight, blood pressure, test scores, income, temperature.
- **The residuals (errors) are approximately normally distributed and homoscedastic** (i.e., constant variance).
- **Negative and positive values are both possible**
    - Linear regression allows the outcome to be any real number.

false, not needed

🔷 Use Linear Regression When:

- **Your outcome (dependent) variable is continuous**
    - Examples: height, weight, blood pressure, test scores, income, temperature.
- **The residuals (errors) are approximately normally distributed and homoscedastic** (i.e., constant variance).
- **Negative and positive values are both possible**
    - Linear regression allows the outcome to be any real number.

false, not needed

🔷 Use Linear Regression When:

- **Your outcome (dependent) variable is continuous**
    - Examples: height, weight, blood pressure, test scores, income, temperature.
- **The residuals (errors) are approximately normally distributed and homoscedastic** (i.e., constant variance).
- **Negative and positive values are both possible**
    - Linear regression allows the outcome to be any real number.

false, not needed

false, not needed

🔷 Use Linear Regression When:

- **Your outcome (dependent) variable is continuous**
    - Examples: height, weight, blood pressure, test scores, income, temperature.
- **The residuals (errors) are approximately normally distributed and homoscedastic** (i.e., constant variance).
- **Negative and positive values are both possible**
    - Linear regression allows the outcome to be any real number.

false, not needed

false, not needed

false, not needed

🔷 Use Linear Regression When:

- **Your outcome (dependent) variable is continuous**
    - Examples: height, weight, blood pressure, test scores, income, temperature.
- **The residuals (errors) are approximately normally distributed and homoscedastic** (i.e., constant variance).
- **Negative and positive values are both possible**
    - Linear regression allows the outcome to be any real number.

fine

false, not needed

false, not needed

🔷 Use Linear Regression When:

- **Your outcome (dependent) variable is continuous**
    - Examples: height, weight, blood pressure, test scores, income, temperature.
- **The residuals (errors) are approximately normally distributed and homoscedastic** (i.e., constant variance).
- **Negative and positive values are both possible**
    - Linear regression allows the outcome to be any real number.

fine

false, not needed

Contemporary view:

Use linear regression when you want to estimate the difference between averages

🔶 Use Poisson Regression When:

- **Your outcome variable is a count**
    - Examples: number of doctor visits, number of crimes in a neighborhood, number of website clicks.
- **Counts are non-negative integers (0, 1, 2, ...)**.
- **The mean of the outcome is roughly equal to the variance** (an assumption of the Poisson distribution).
- **You're modeling rates** (e.g., events per time or population), often using an *offset* term.

🔶 Use Poisson Regression When:

- **Your outcome variable is a count**
    - Examples: number of doctor visits, number of crimes in a neighborhood, number of website clicks.
- **Counts are non-negative integers (0, 1, 2, ...)**.
- **The mean of the outcome is roughly equal to the variance** (an assumption of the Poisson distribution).
- **You're modeling rates** (e.g., events per time or population), often using an *offset* term.

false, not needed

🔶 Use Poisson Regression When:

- **Your outcome variable is a count**
    - Examples: number of doctor visits, number of crimes in a neighborhood, number of website clicks.
- **Counts are non-negative integers (0, 1, 2, ...)**.
- **The mean of the outcome is roughly equal to the variance** (an assumption of the Poisson distribution).
- **You're modeling rates** (e.g., events per time or population), often using an *offset* term.

false, not needed

🔶 Use Poisson Regression When:

- **Your outcome variable is a count**
    - Examples: number of doctor visits, number of crimes in a neighborhood, number of website clicks.

integers: not needed

- **Counts are non-negative integers (0, 1, 2, ...)**.
- **The mean of the outcome is roughly equal to the variance** (an assumption of the Poisson distribution).
- **You're modeling rates** (e.g., events per time or population), often using an *offset* term.

false, not needed

🔶 Use Poisson Regression When:

- **Your outcome variable is a count**
  - Examples: number of doctor visits, number of crimes in a neighborhood, number of website clicks.
- **Counts are non-negative integers (0, 1, 2, ...)**.
- **The mean of the outcome is roughly equal to the variance** (an assumption of the Poisson distribution).
- **You're modeling rates** (e.g., events per time or population), often using an *offset* term.

integers: not needed

non-negative: yes

false, not needed

🔶 Use Poisson Regression When:

- **Your outcome variable is a count**
    - Examples: number of doctor visits, number of crimes in a neighborhood, number of website clicks.
- **Counts are non-negative integers (0, 1, 2, ...)**.

integers: not needed

non-negative: yes

- **The mean of the outcome is roughly equal to the variance** (an assumption of the Poisson distribution).
- **You're modeling rates** (e.g., events per time or population), often using an *offset* term.

false, not needed

🔶 Use Poisson Regression When:

- **Your outcome variable is a count**
    - Examples: number of doctor visits, number of crimes in a neighborhood, number of website clicks.
- **Counts are non-negative integers (0, 1, 2, ...)**.
- **The mean of the outcome is roughly equal to the variance** (an assumption of the Poisson distribution).
- **You're modeling rates** (e.g., events per time or population), often using an *offset* term.

integers: not needed

non-negative: yes

false, not needed

🔶 Use Poisson Regression When:

- **Your outcome variable is a count**
    - Examples: number of doctor visits, number of crimes in a neighborhood, number of website clicks.
- **Counts are non-negative integers (0, 1, 2, ...)**.
- **The mean of the outcome is roughly equal to the variance** (an assumption of the Poisson distribution).
- **You're modeling rates** (e.g., events per time or population), often using an *offset* term.

false, not needed

integers: not needed

non-negative: yes

false, not needed

false, not needed

🔶 Use Poisson Regression When:

false, not needed

- **Your outcome variable is a count**
    - Examples: number of doctor visits, number of crimes in a neighborhood, number of website clicks.

integers: not needed

non-negative: yes

- **Counts are non-negative integers (0, 1, 2, ...)**.
- **The mean of the outcome is roughly equal to the variance** (an assumption of the Poisson distribution).
- **You're modeling rates** (e.g., events per time or population), often using an *offset* term.

false, not needed

false, not needed

Contemporary view:

Use Poisson regression when you want to estimate the fold-difference between averages

⚠️ Additional Considerations:

- If your count data are **overdispersed** (i.e., variance >> mean), **Negative Binomial regression** might be more appropriate than Poisson.
- If your data include a **large number of zeros**, consider **zero-inflated Poisson** or **zero-inflated negative binomial** models.
- If your data are **binary** (0 or 1), use **logistic regression**, not linear or Poisson.

⚠️ Additional Considerations:

- If your count data are **overdispersed** (i.e., variance >> mean), **Negative Binomial regression** might be more appropriate than Poisson.
- If your data include a **large number of zeros**, consider **zero-inflated Poisson** or **zero-inflated negative binomial** models.
- If your data are **binary** (0 or 1), use **logistic regression**, not linear or Poisson.

Choosing between regression models

- "Traditional" wisdom
  - Linear for continuous, logistic for binary, Poisson for count
- "Contemporary" thinking
  - Linear for estimating differences in means
  - Logistic for estimating fold-differences in odds
  - Poisson for estimating fold-differences in means
- There are *many* other regression models… we'll talk about more this afternoon!

Choosing parameters: Think, pair, share

- Outcome: Total richness. Which do you like better?
    - “We estimate that forest soil microbiomes have 50 more species on average than desert sand microbiomes.”
    - “We estimate that forest soil microbiomes have 5% more species on average than desert sand microbiomes.”
- Outcome: Bacterial load. Which do you like better?
    - “We estimate that forest soil microbiomes have 4 million more bacterial DNA copies/ml on average than desert sand microbiomes.”
    - “We estimate that forest soil microbiomes have 3 times more bacterial DNA copies/ml on average than desert sand microbiomes.”

Summary of regression

functional of outcome variable = function of predictor variables

- So far, we have talked about standard regression models
- In microbiome science, these can be good for outcomes like
  - (estimated) diversity - species richness, Shannon, Simpson…
  - gene presence
  - bacterial load (eg q/ddPCR data)
  - …
- **They are <u>not good</u> for outcomes like compositions or counts**

Adjustment sets

functional of outcome variable = function of predictor variables

So far: decision making for the LHS

Next: decision making for the RHS!

Adjustment sets

$$\text{expected diversity}_i = \beta_0 + \beta_1 \times \mathbf{1}_{\{i \text{ is from lakewater }\}}$$

$$\text{expected diversity}_i = \delta_0 + \delta_1 \times \mathbf{1}_{\{i \text{ is from lakewater}\}} + \delta_2 \times \text{temp}$$

- Are β_1 and δ_1 the same?
    - Does adding additional variables change the meaning of the parameters?

Adjustment sets

$$\text{expected diversity}_i = \beta_0 + \beta_1 \times \mathbf{1}_{\{i \text{ is from lakewater }\}}$$

$$\text{expected diversity}_i = \delta_0 + \delta_1 \times \mathbf{1}_{\{i \text{ is from lakewater}\}} + \delta_2 \times \text{temp}$$

- β_1 = the difference in average diversity between lake and seawater environments

- δ_1 = the difference in average diversity between lake and seawater environments *of the same temperature*

Adjustment sets

expected value of $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} \ldots + \beta_p X_{ip}$

- $\hat{\beta}_k$ is an estimate of the difference in the average value of Y in environments that differ by 1 unit in $X_{\cdot k}$ but are alike in $X_{\cdot 1}, \ldots, X_{\cdot k-1}, X_{\cdot k+1}, \ldots, X_{\cdot p}$

- "We estimate that the difference in average microbial diversity between fresh- and seawater environments of the same temperature and light level is 32 species…"

Adjustment sets

expected value of $Y_i = e^{\beta_0+\beta_1 X_{i1}+\beta_2 X_{i2}\ldots+\beta_p X_{ip}}$

- $e^{\hat{\beta}_k}$ is an estimate of the fold-difference in the average value of Y in environments that differ by 1 unit in $X_{\cdot k}$ but are alike in $X_{\cdot 1}, \ldots, X_{\cdot k-1}, X_{\cdot k+1}, \ldots, X_{\cdot p}$

- "We estimate that the average microbial diversity in freshwater environments is 1.07 times greater than seawater environments of the same temperature and light level…"

What are good choices of adjustment variables?

Guidance for choosing adjustment sets

1. Choose based on the parameter you care about

    - “We estimate that the difference in average microbial diversity between fresh- and seawater environments of the same temperature and sunlight is 32 species…”

2. If you have beliefs about mechanism, and are curious about causal effects, choose based on a causal diagram

    - Adjust for confounders & precision variables

3. Choose a sensible comparison to make

    - “We estimate that Dialister is 49 times more abundant than typical in the gut metagenomes of CRC patients relative to non-CRC controls who are alike in gender, BMI, age and cohort.”

Categories of predictor variables

1. Predictor of interest
2. Precision variables
3. Confounders
4. Effect modifiers

Categories of predictor variables

1. Predictor of interest

- The main thing you set out to study

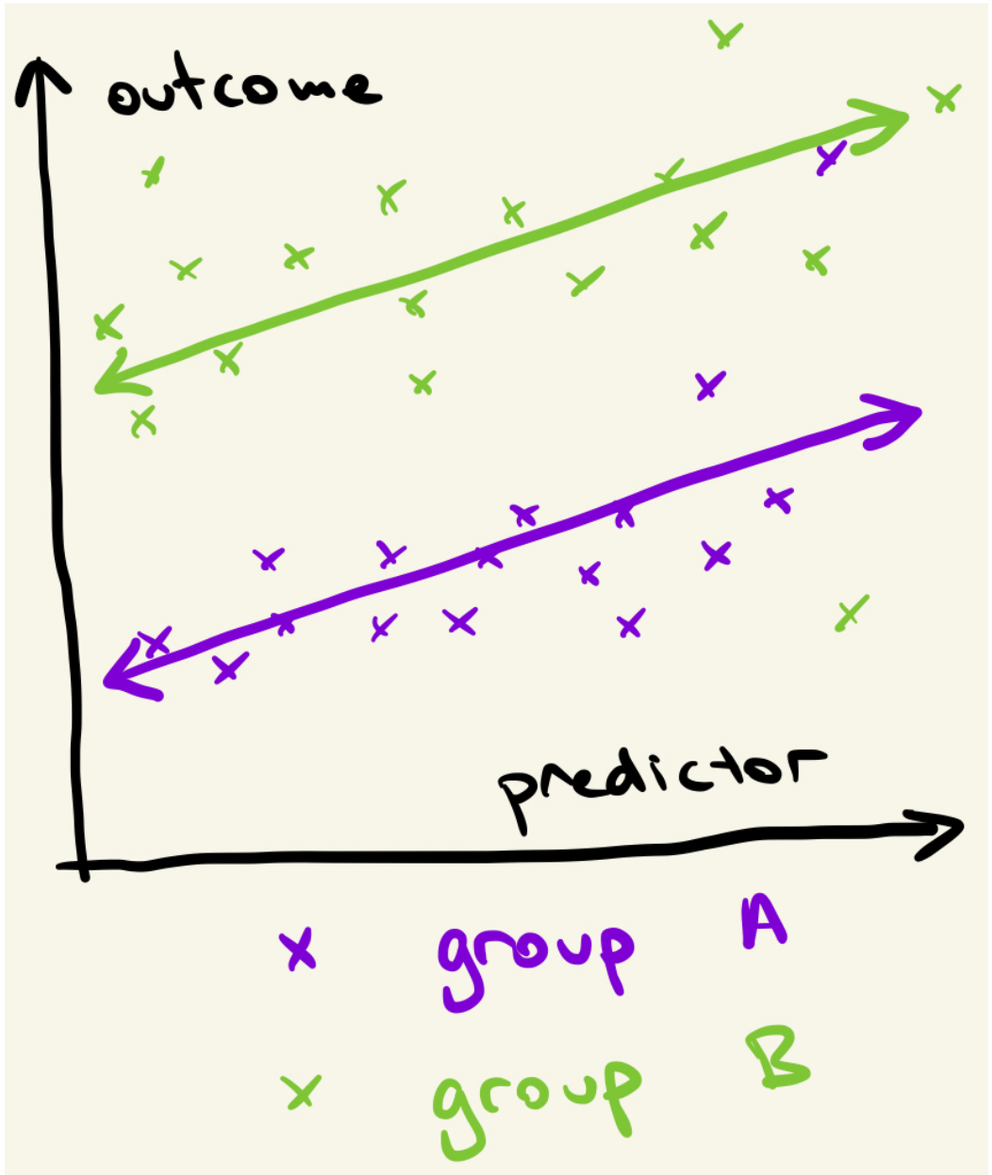
- Always include

Types of variables

2. Precision variables

- Associated with outcome
- Not associated with predictor of interest

- Helps to improve precision
  - e.g., batch effects, tank effects
  - e.g. in human microbiome: age, sex…
  - Often capture “technical variation”



Types of variables

3. Confounders

- "Common causes" of **predictor of interest** and **outcome**

- None of the following are confounders for true microbial abundances Y_{ij}
    - Batch
    - Sequencing technology
    - Any measurement variables (depth…)

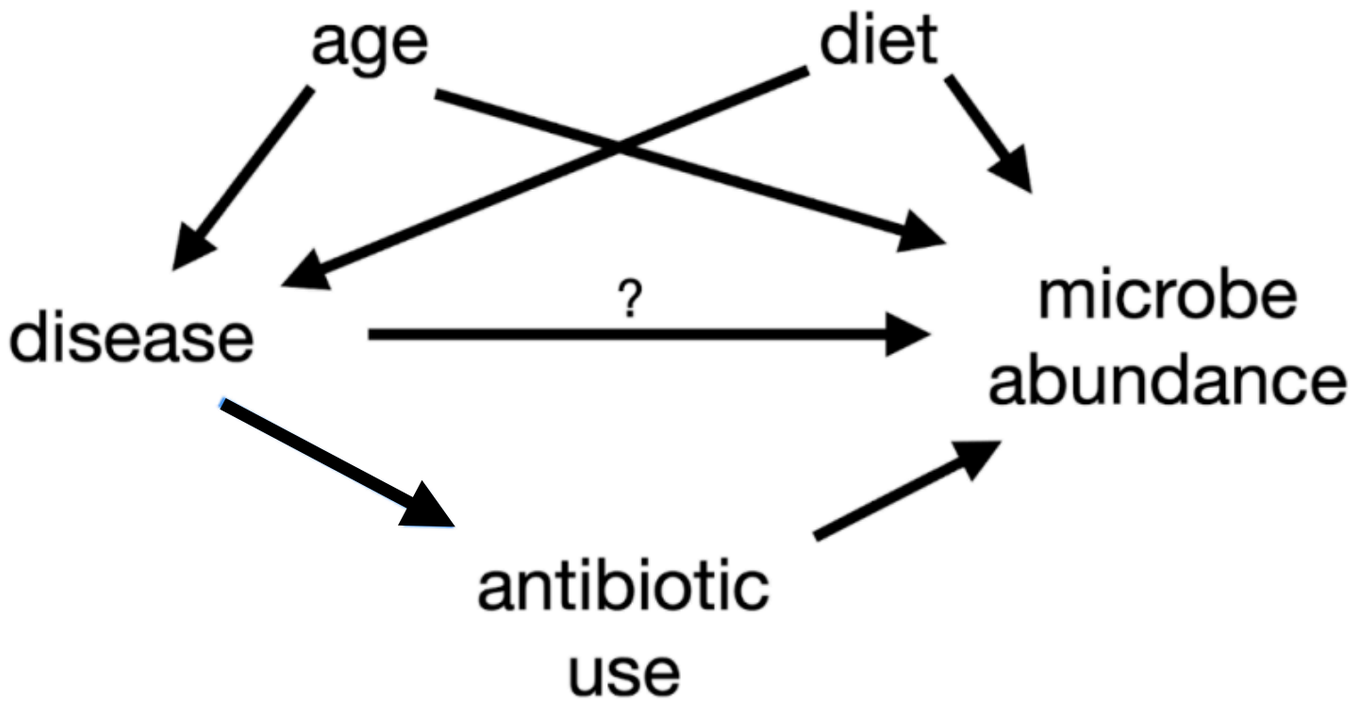
- Variables associated with the measurement process *cannot* be causally associated with outcome

- "Confounders" is more often misused than correctly used

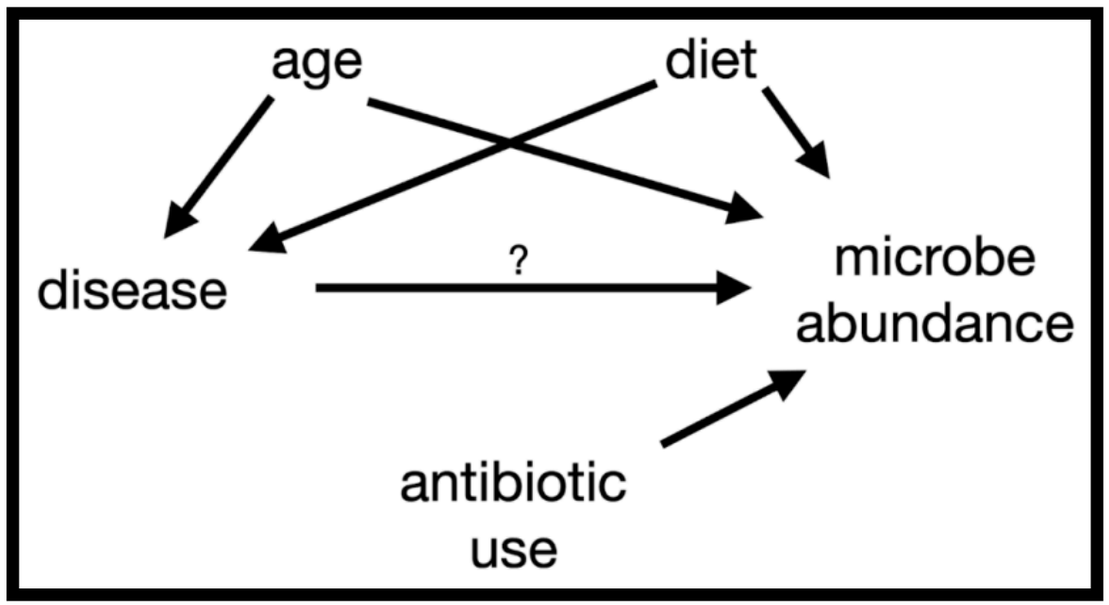
Types of variables

- “Common causes” requires you to write down causal assumptions

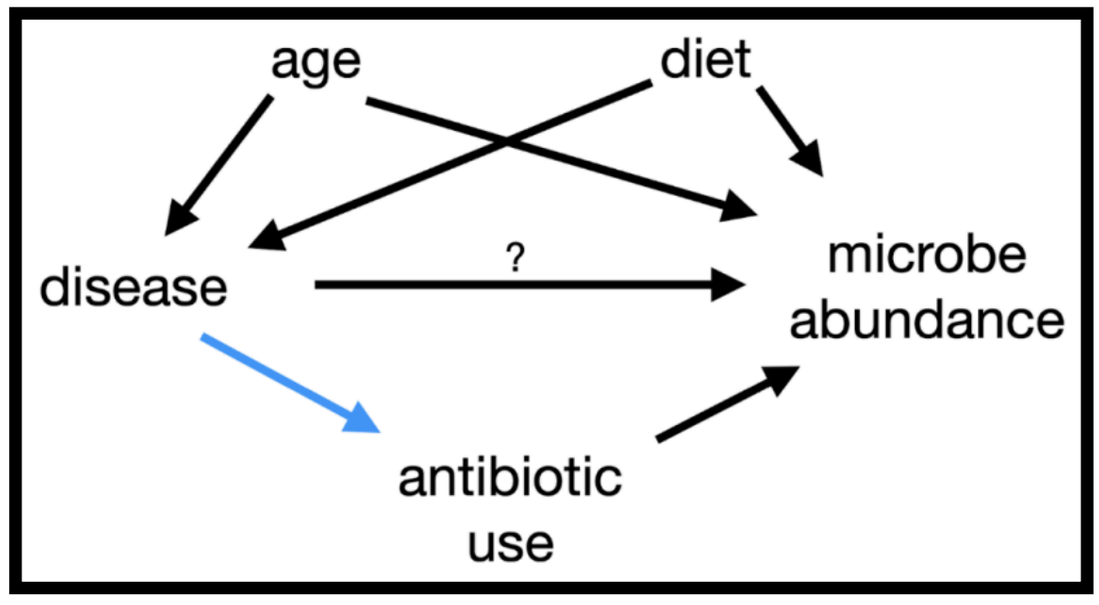
- Causal assumptions = a hypothesized list of causes and outcomes



Adjustment sets



Adjust for: age, diet, antibiotic use



Adjust for: age, diet

- Even if not attempting “causal inference,” write down causal assumptions to choose adjustment sets
    - https://www.r-causal.org/chapters/05-dags
    - dagitty::adjustmentSets()

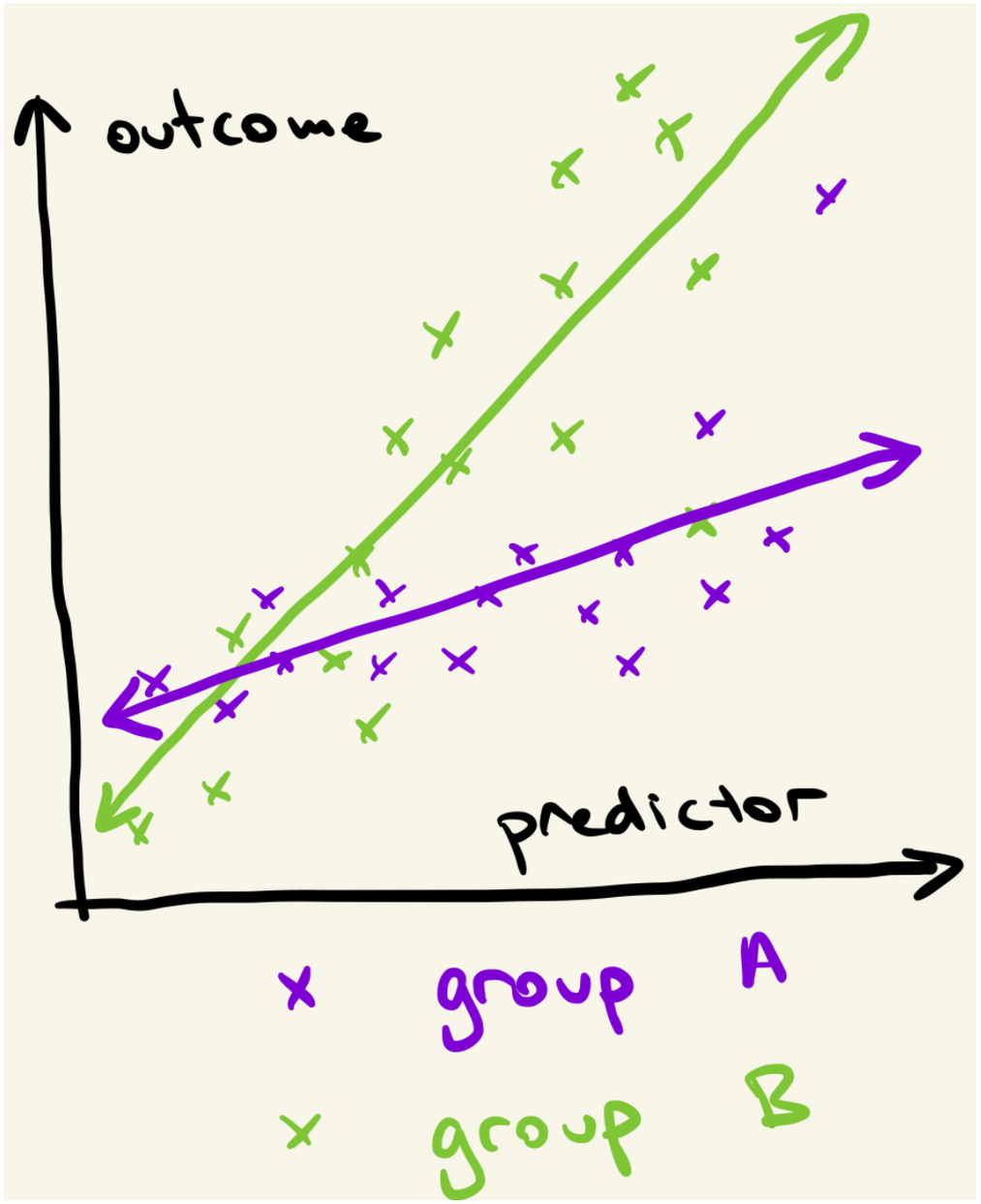
Types of variables

4. Effect modifiers

- Association b/w response & predictor of interest differs for different values of an effect modifier

- “interaction” between variables

- Sometimes, effect modification is the predictor of interest



Types of variables

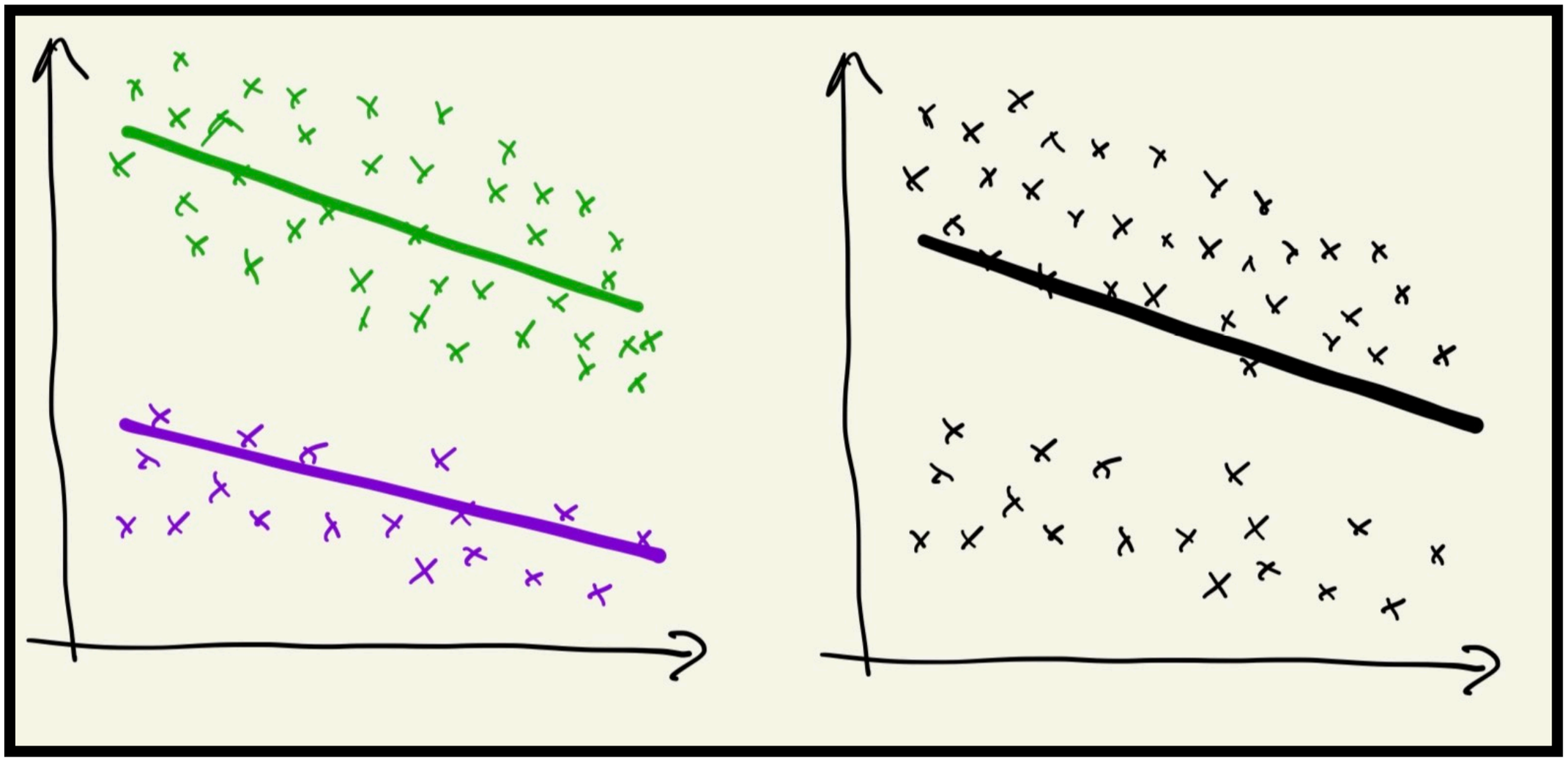
- Precision variables and effect modifiers
    - There almost always will be many unmeasured or unmeasurable precision variables
    - There almost always will be many unmeasured or unmeasurable effect modifiers
    - This is *fine!* You don’t need to include all PVs and EMs in your model!

What happens when we omit variables?

- Unmodeled precision variables and effect modifiers get “averaged over”
    - Not necessarily a problem

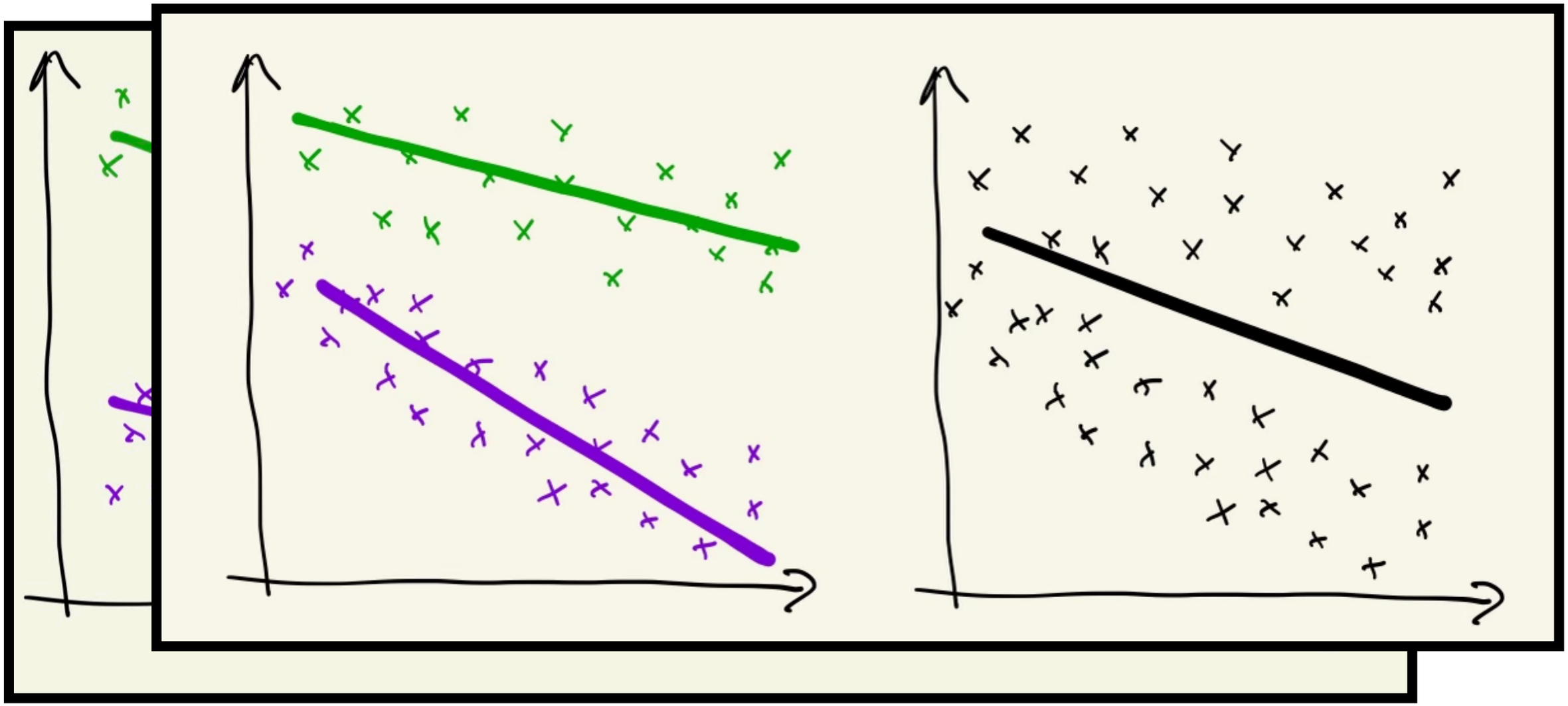
What happens when we omit variables?

- Unmodeled precision variables and effect modifiers get “averaged over”
  - Not necessarily a problem



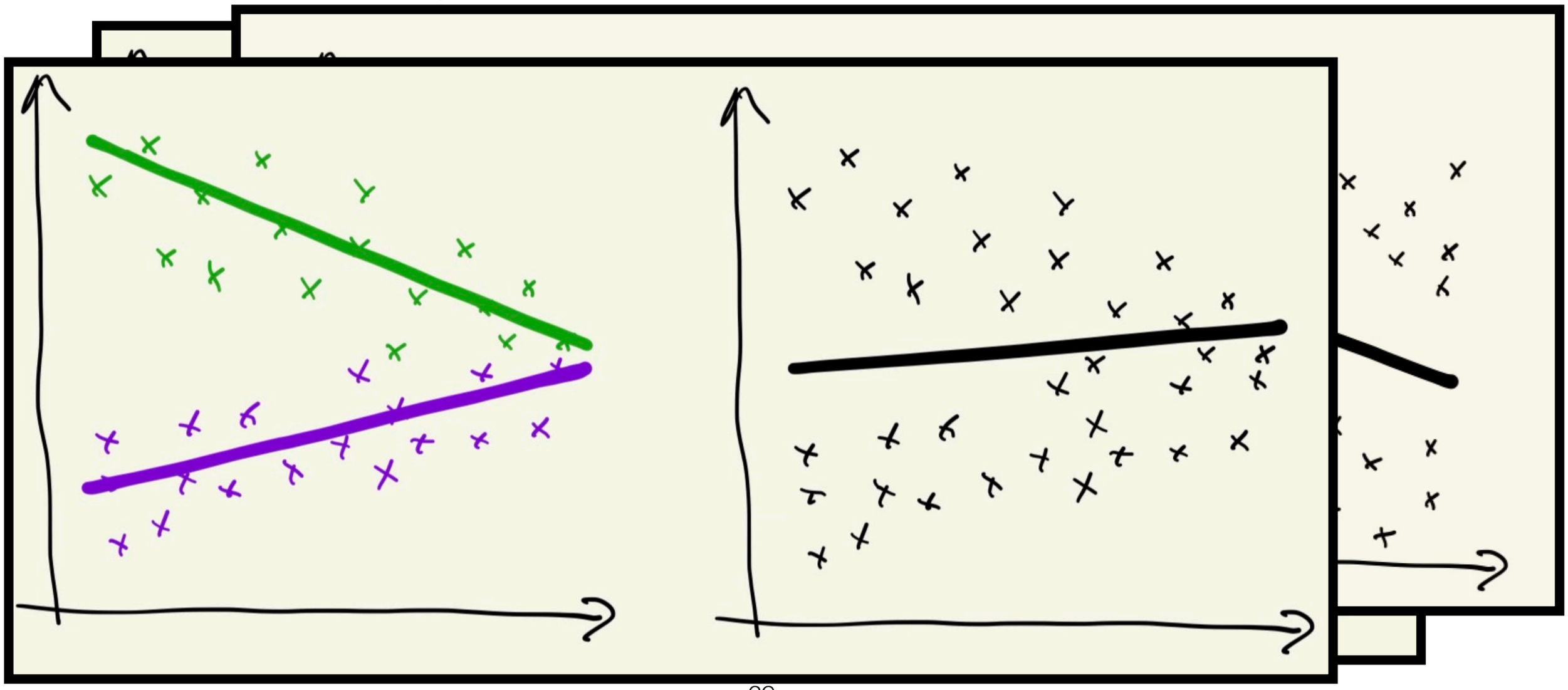
What happens when we omit variables?

- Unmodeled precision variables and effect modifiers get “averaged over”
  - Not necessarily a problem



What happens when we omit variables?

- Unmodeled precision variables and effect modifiers get “averaged over”
    - Not necessarily a problem



Summary of regression

functional of outcome variable = function of predictor variables

- So far, we have talked about standard regression models
- In microbiome science, these can be good for outcomes like
  - (estimated) diversity - species richness, Shannon, Simpson…
  - gene presence
  - bacterial load (eg q/ddPCR data)
  - …
- **They are <u>not good</u> for outcomes like compositions or counts**

Summary

- High-level ideas
  - parameters true unknown things about the universe
  - estimators & their properties better and worse guesses at parameters
- Examples
  - Estimating parameters using common regression models
    - differences in averages via linear regression
    - ratios of averages via Poisson regression
    - ratios of odds via logistic regression
  - Which are sensible comparisons to make? What should you adjust for

📉 Now: Regression lab 📈

1. Wiki ➡️ Schedule ➡️ “Statistics labs”
2. Copy the command under “regression lab”
3. Run the copied command in your RStudio Server console
4. Open the downloaded Rmd and work through the exercises

🧚 *This is no ordinary quest* 🐲

🦹 *There are errors in the code we gave you* 🧌

🪄 *You are to debug them* 🦄

👬 *You can do it! And we are here to help!* 👯