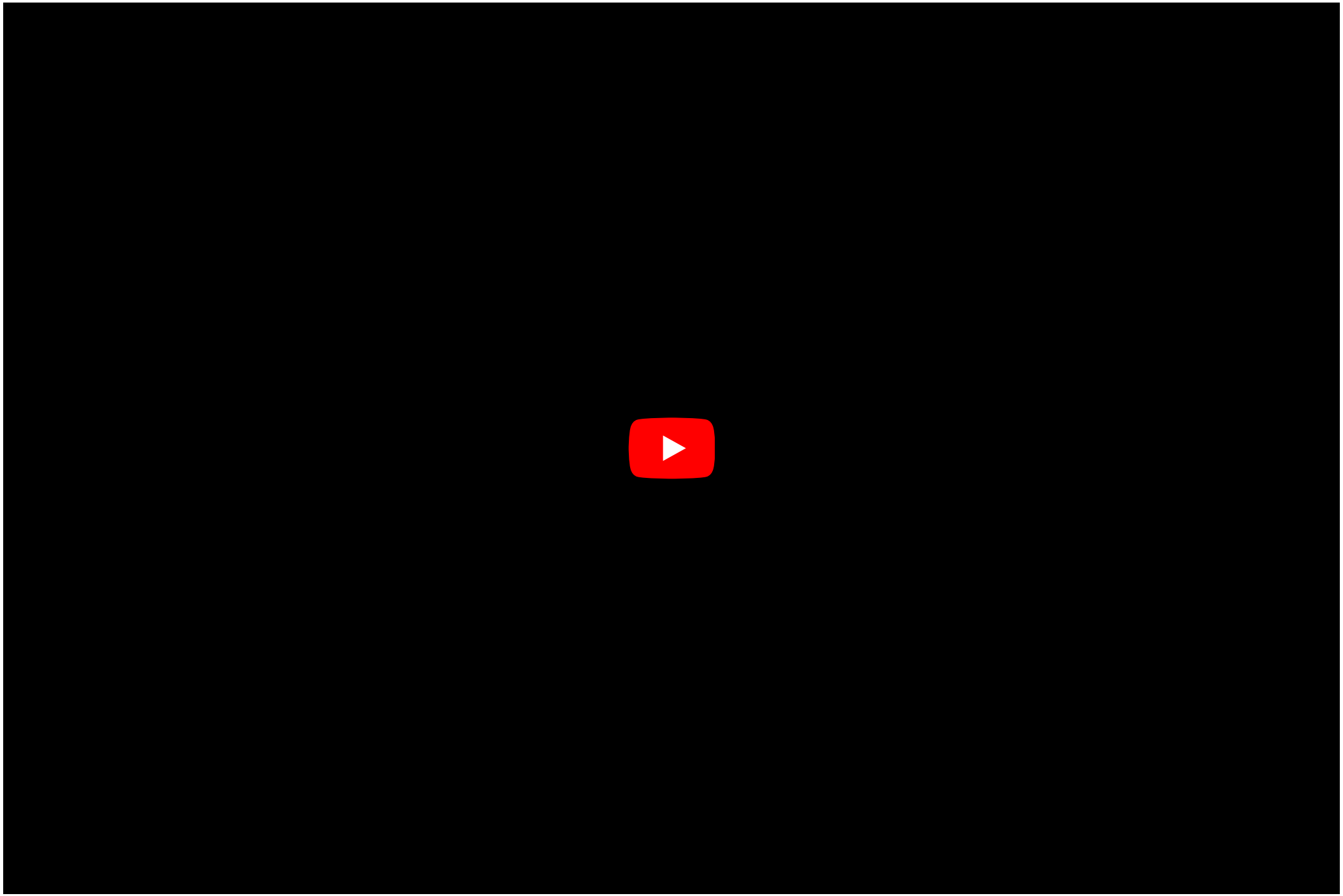
Depression Incidence and Vitamin D Concentration in NHANES Survey Data

A Case Study in Collaboration and Causal Inference

Julia Piaskowski & Yimin Chen

October 10, 2024

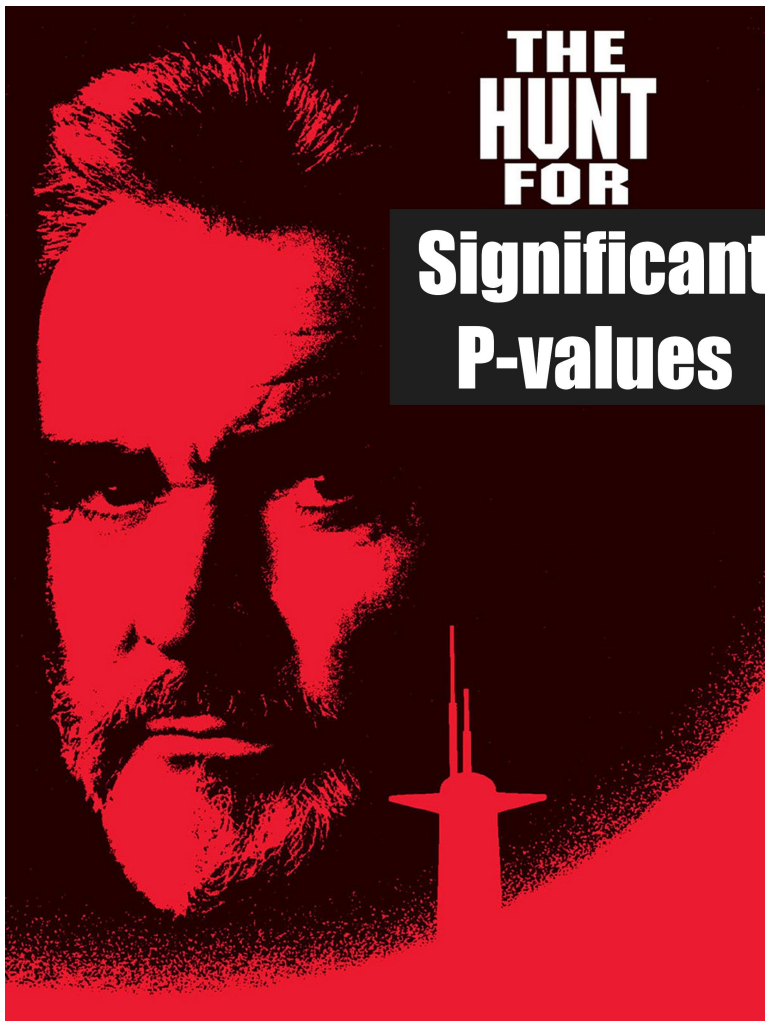
https://jpiaskowski.gitlab.io/talks/nhanes-magic/



Statistical Consulting

Frequently

- linear model -> ANOVA -> p-values
- correlation -> coefficient of correlation
- t-test -> p-values
- test every column in a spreadsheet



Problems with this Approach

- Single-minded focus on p-values and/or correlation at the expense of understanding systems or correctly reporting results.
- An underlying assumption that statistics (magically) extracts meaningful results.
- Fishing for any pairwise association without fully considering the implications of that relationship

Case Study of (Maybe) a Better Approach

Depression During Pregnancy and Postpartum

- New moms experience new stress
- Antepartum depression associated with stunted infant growth
- Postpartum depression associated with child behavioral issues and developmental challenges
- Vitamin D concentration negatively associated with depression, although literature is inconsistent

Vitamin D Mechanism

- Potential protective mechanism through serotonin, the “happy” neurotransmitter
- Serotonin has a consistent negative association with depression, anxiety, etc.
- Brain makes serotonin from tryptophan – which requires vitamin D to activate transcription factor
- Vitamin D inhibits monoamine oxidase (breaks down neurotransmitters)
- Vitamin D inhibits serotonin reuptake receptors (terminates serotonin signaling)

NHANES

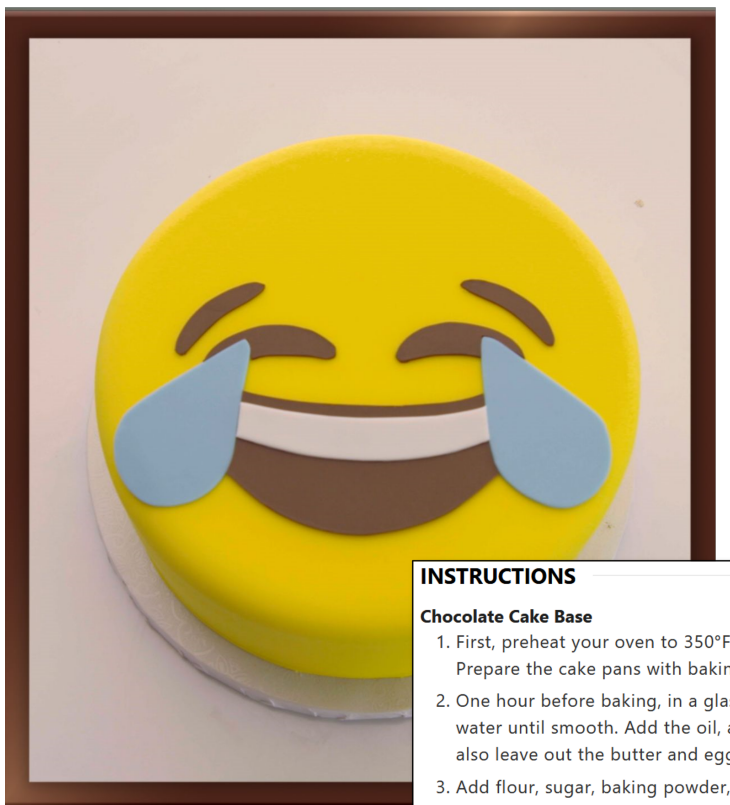
> The National Health and Nutrition Examination Survey (NHANES) is a program of studies designed to assess the health and nutritional status of adults and children in the United States. The survey is unique in that it combines interviews and physical examinations. -CDC Website

NHANES

- Semi-annual survey run by the CDC; they sample ~5,000 individuals/year since early 1960s
- Results intended to reflect diversity of the U.S. population.
- Allows assessment of nutritional status associated with health promotion & disease prevention
- Widely used as large epidemiological data set that's representative of the entire US population
- We used data from 2007-2018 because the assay protocol for Vitamin D was consistent in that time period

Define What You Want to Estimate

Estimand



Estimate

Estimator

INSTRUCTIONS

Chocolate Cake Base

1. First, preheat your oven to 350°F and set an oven rack in the lower third o
   Prepare the cake pans with baking spray and lined parchment paper.
2. One hour before baking, in a glass cup measurer, whisk cocoa powder wit
   water until smooth. Add the oil, and then let cool until room temperature.
   also leave out the butter and eggs to room temperature.
3. Add flour, sugar, baking powder, and salt into your stand mixer and then
   until blended.
4. Add the butter and cocoa mixture and manually mash the butter and coc
   mixing so it doesn't get everywhere. Then, mix on low speed until moister
   speed to medium and beat for 1½ minutes. Scrape down the bowl.
5. In a small bowl crack the eggs and add the vanilla extract. Whisk u
   combined. On medium-low speed, add the egg mixture in three parts, wai
   seconds between each part to let the egg incorporate gradually.

Our Estimand

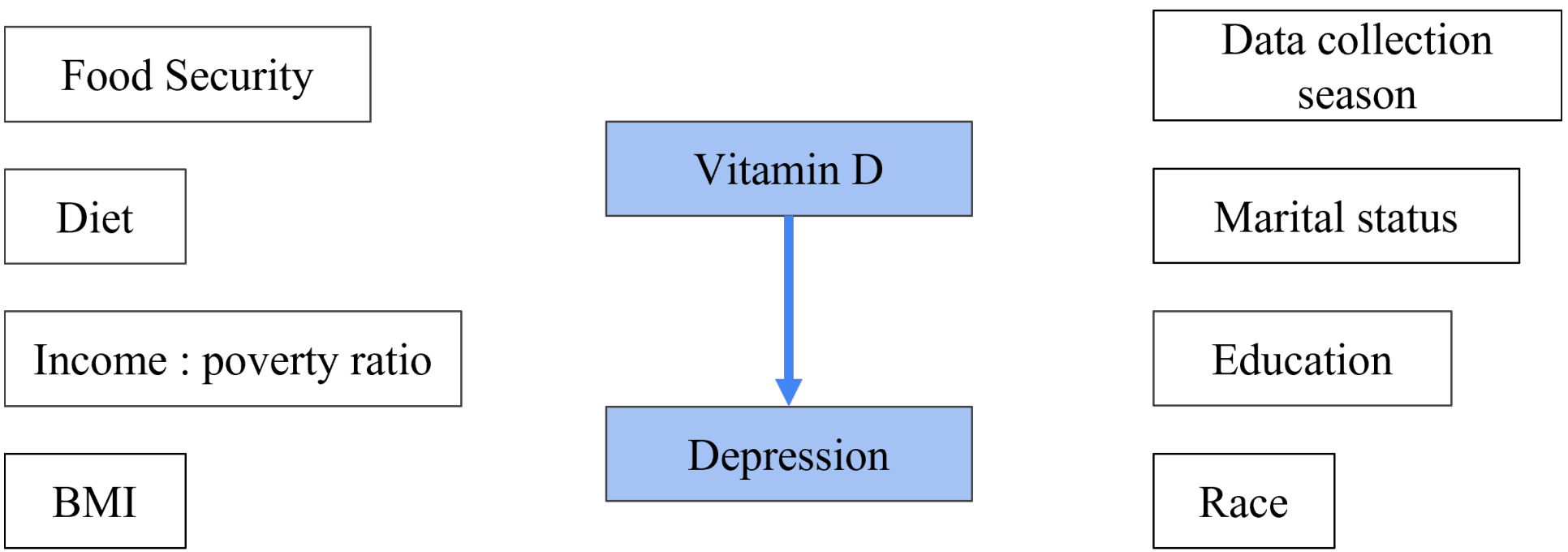
The impact of vitamin D on depression risk in pregnant and postpartum women

AND

The differential impact of vitamin D on depression risk in postpartum women, stratified by breastfeeding status

Understand the System

- Consider possible confounders that prevent proper estimation of a estimand.
- Confounders are things that influence both the ‘exposure’ and the outcome



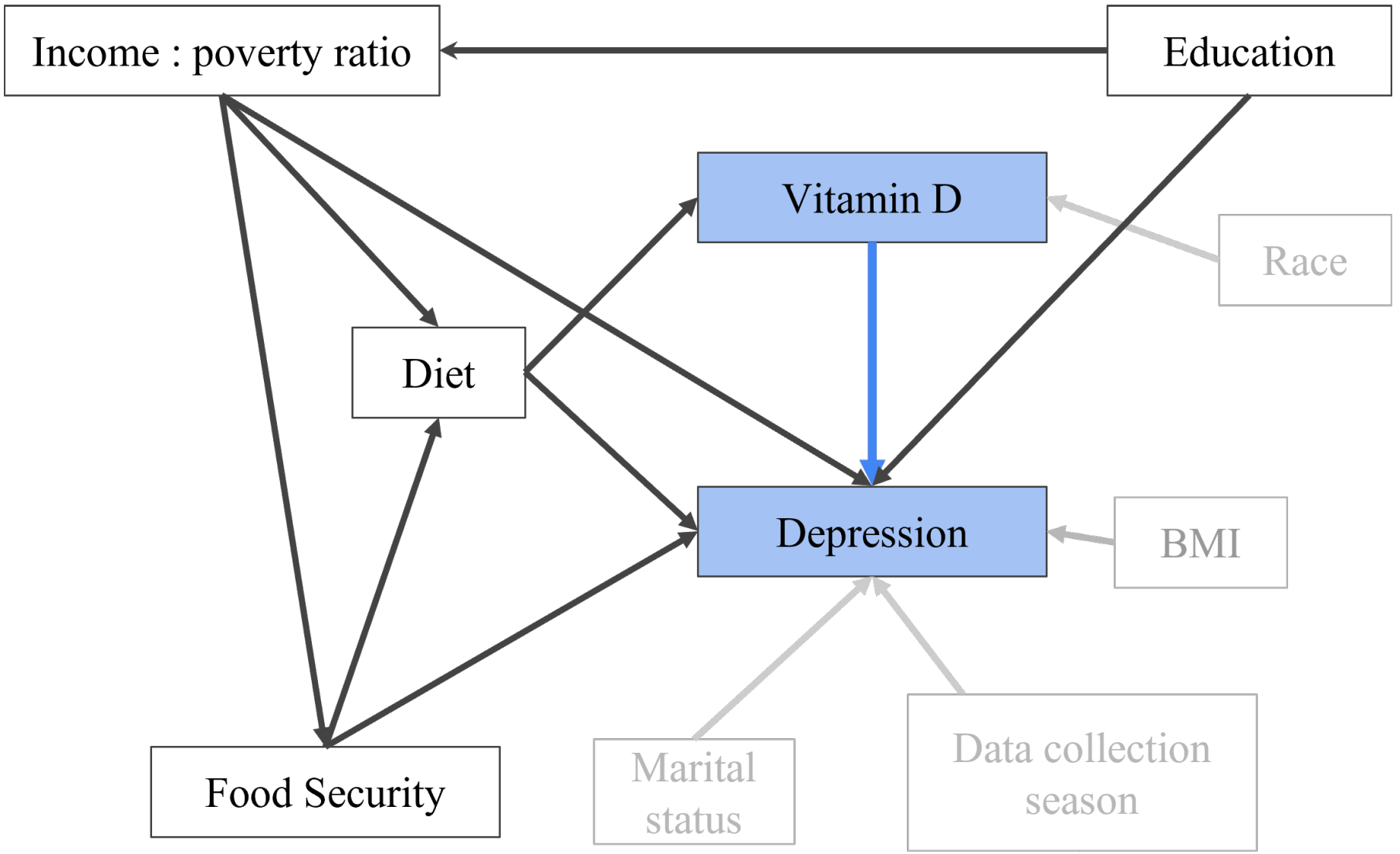
Avoid the 'Causal Salad'

```
Depression = VitaminD + Food Security + Diet + Income/Poverty Ratio +
             BMI +  Data Collection Season + Marital status +
             Education + Race
```

Construct a Directed Acyclical Graph

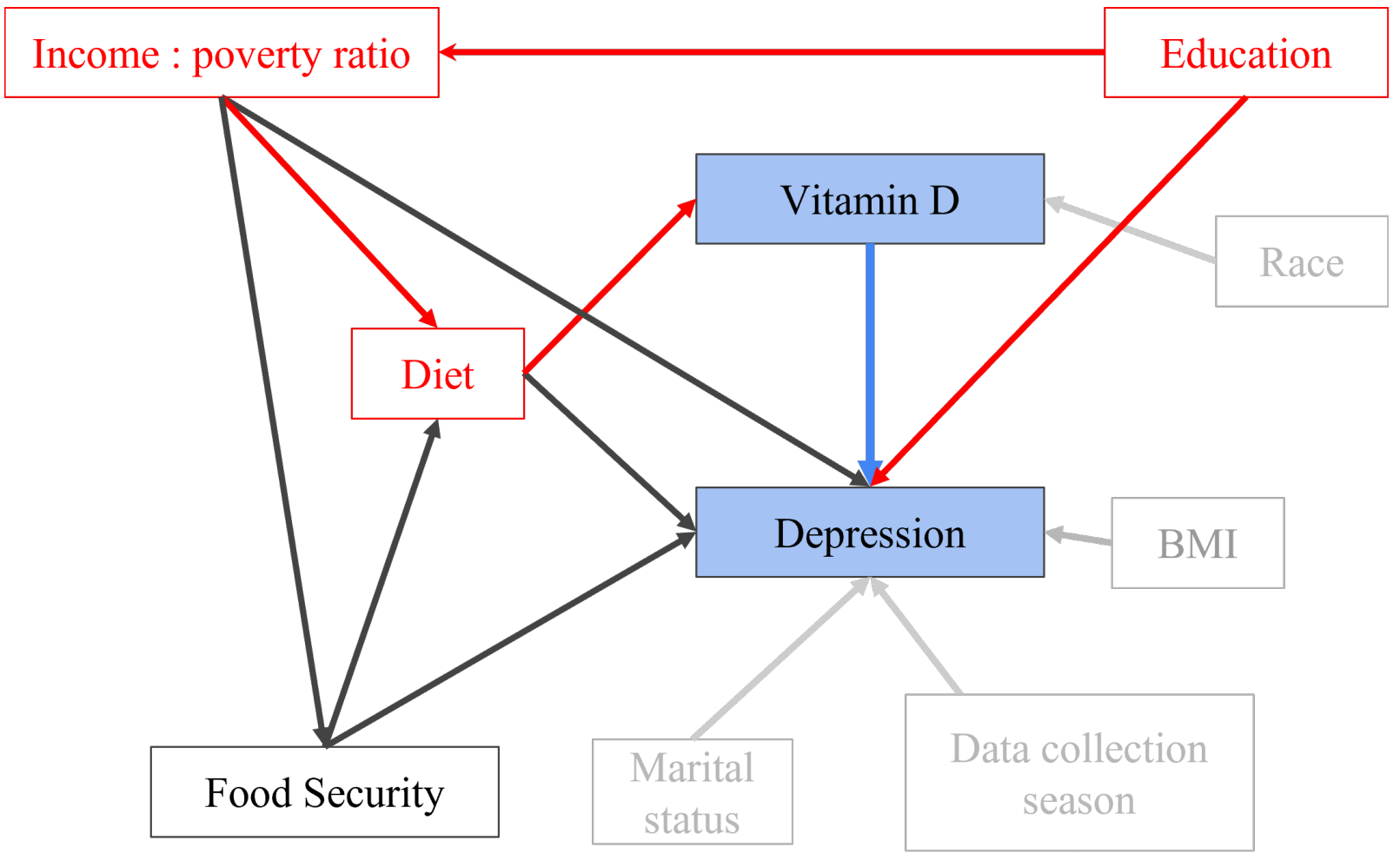
- Use existing information to construct a Directed Acyclical Graph (DAG): a set of directed, causal pathways.
- DAGs are an explicit statement of relationships and assumptions.
- There is no mathematical/statistical statements from a DAG, only causal paths
- DAGs and counfounders are rich area of inquiry!

Our DAG


Income : poverty ratio
Education
Vitamin D
Race
Diet
Depression
BMI
Food Security
Marital status
Data collection season

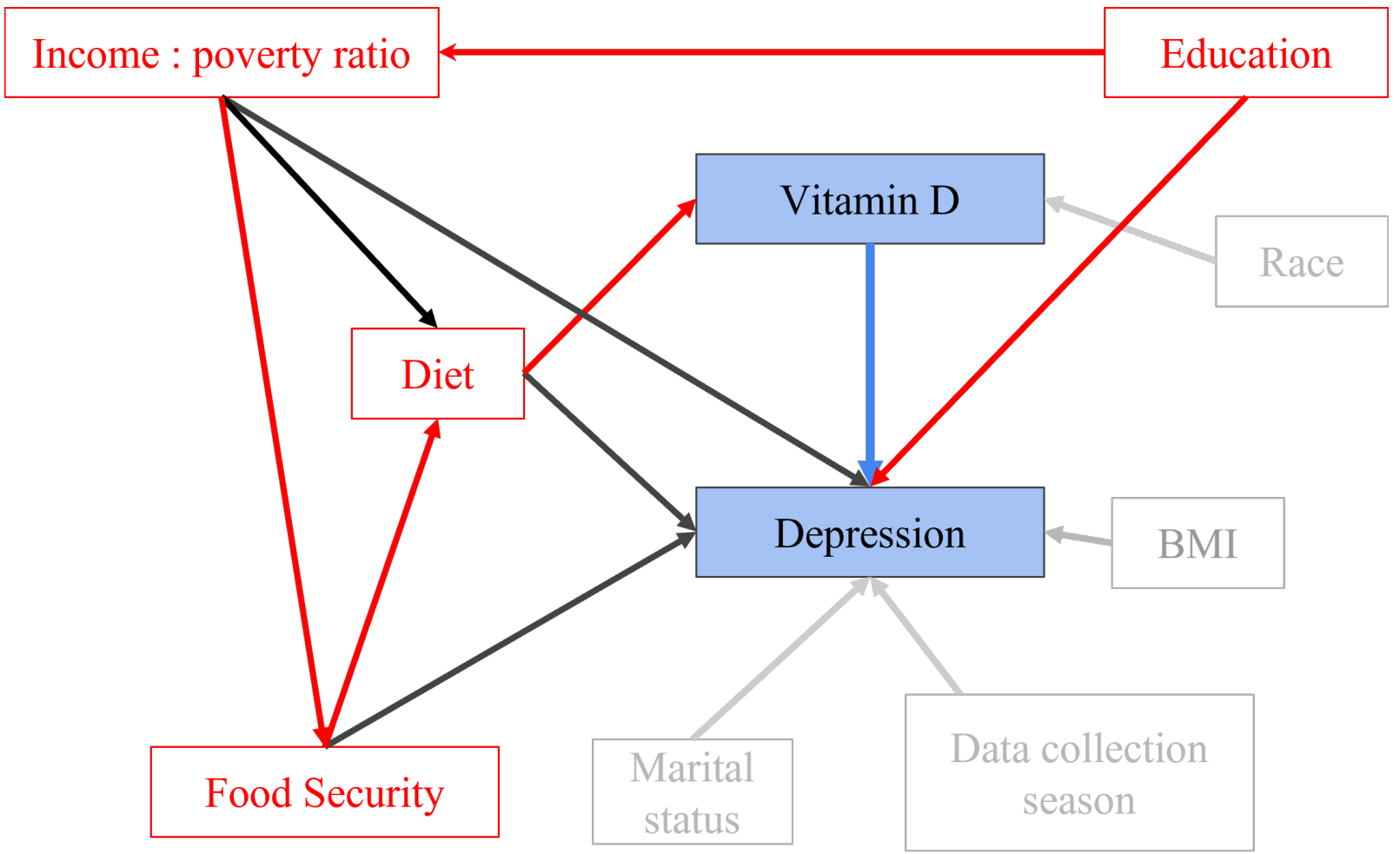
Look for Backdoor Paths

That connect the exposure and outcome variables



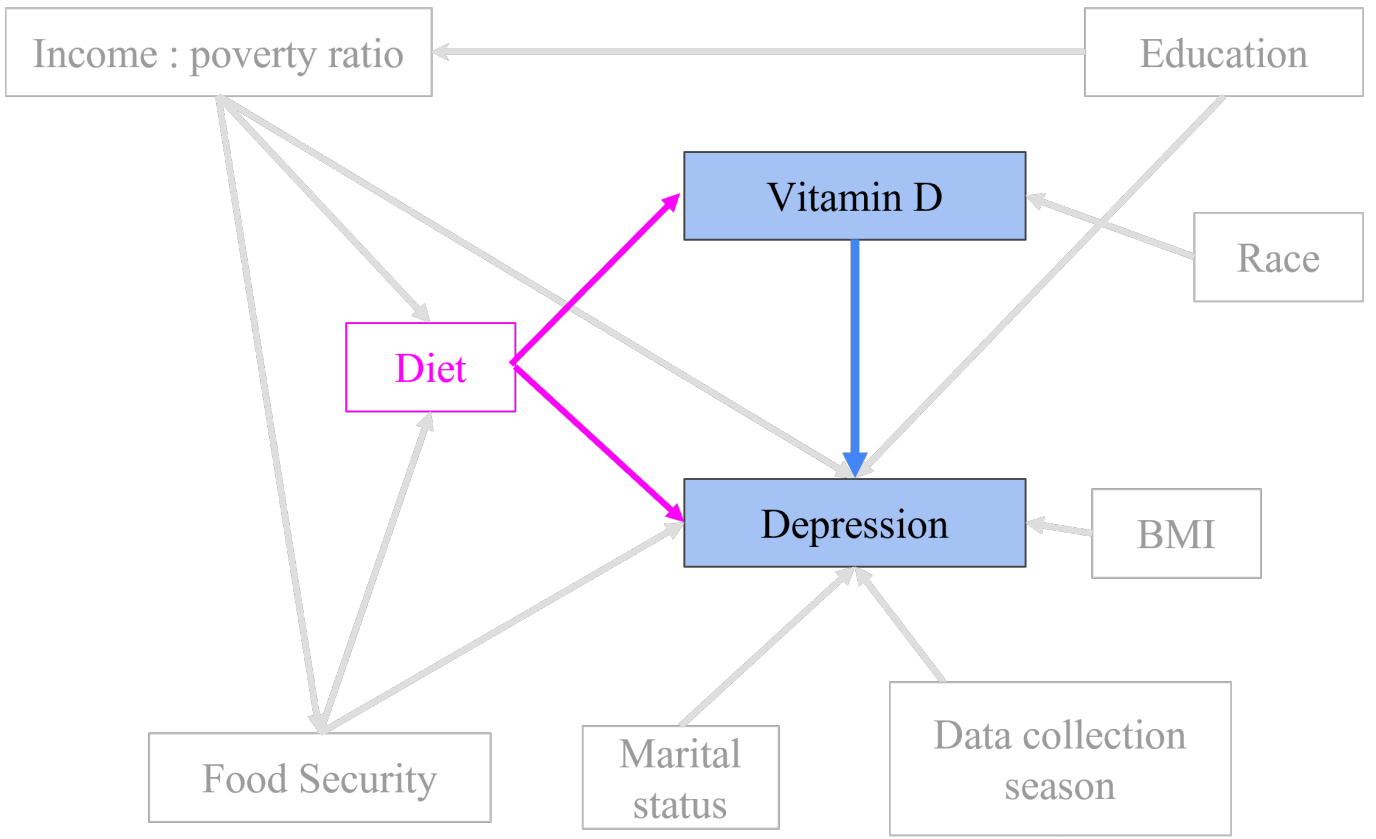
Look for Backdoor Paths

That connect the exposure and outcome variables



Create the Final Adjustment Set

- To eliminate backdoor paths
- Good news: our estimand is estimable!

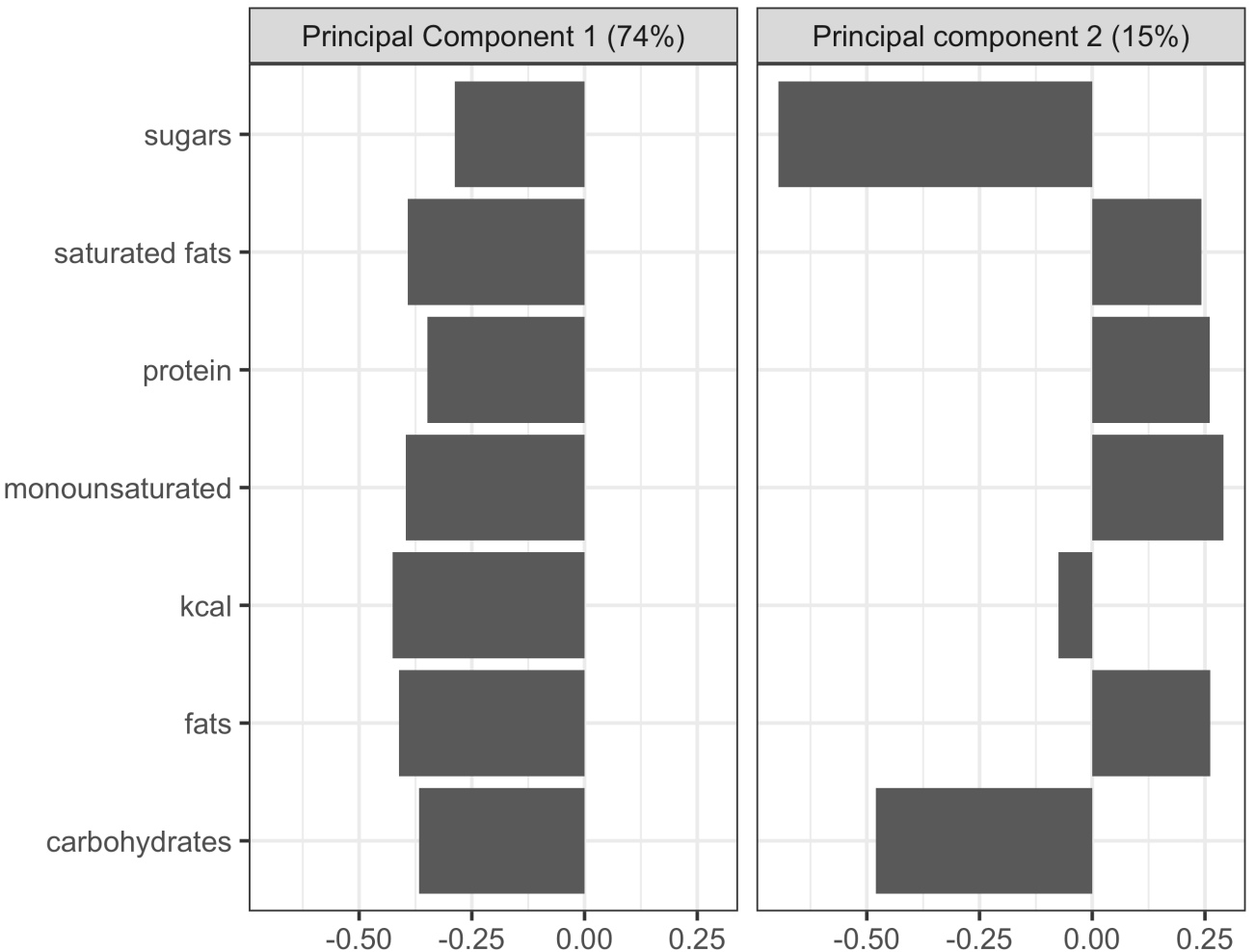


Diet Data

Sorry, what is 'diet', exactly?

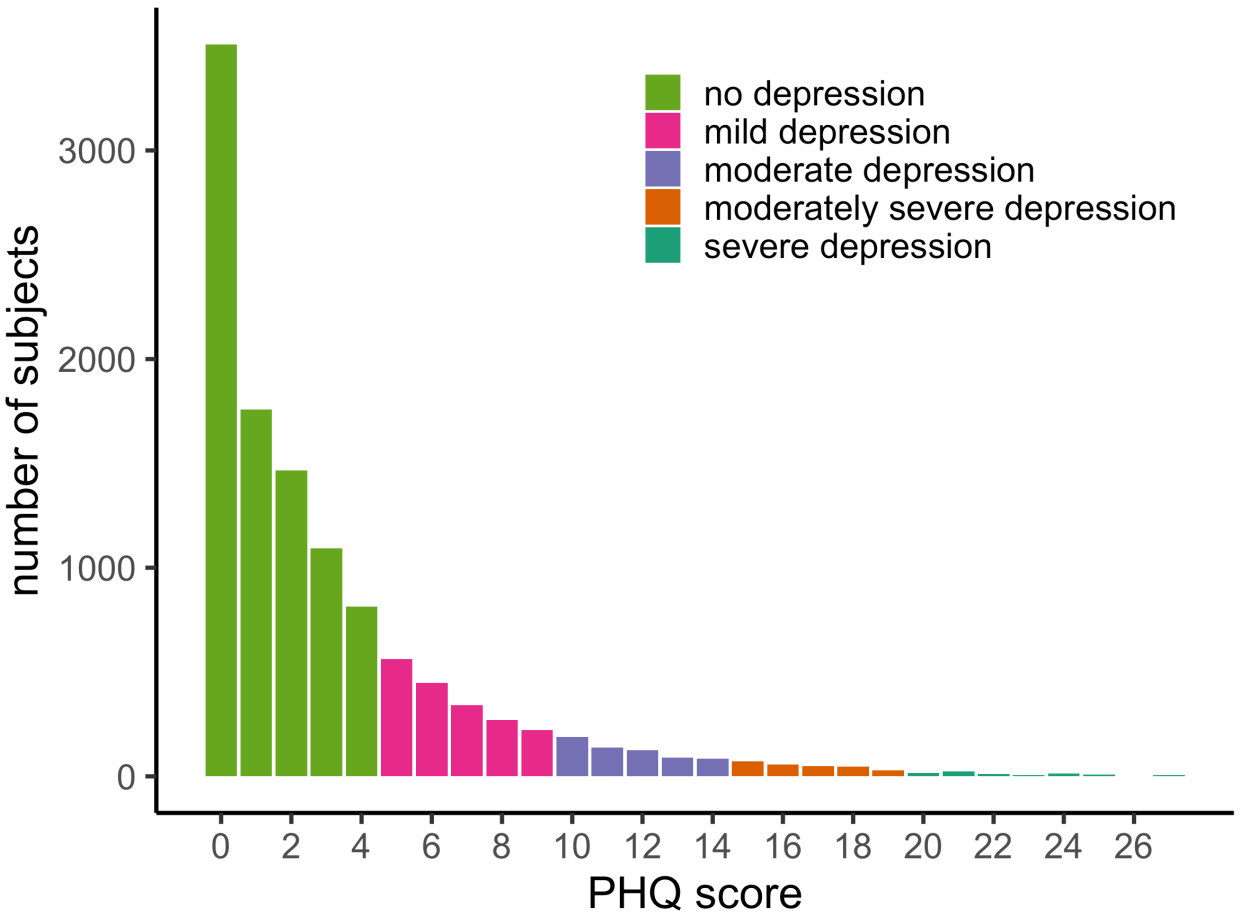
- Food intake data from the *What We Eat in America* dietary interview (24 hour recall)
- Data on 8 different food components:
  - total calories
  - fats (total, saturated, polyunsaturated and monounsaturated)
  - carbohydrates
  - sugars
  - protein

Diet Latent Variables



- PCA-derived
- 2 latent variables:
  - total consumption
  - ‘paleo’

Depression Data

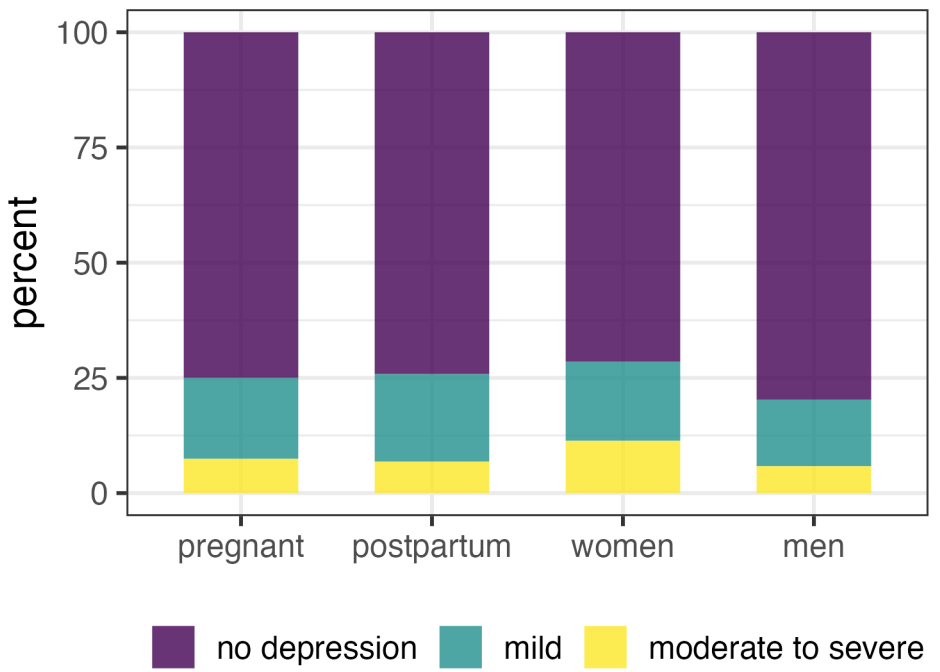


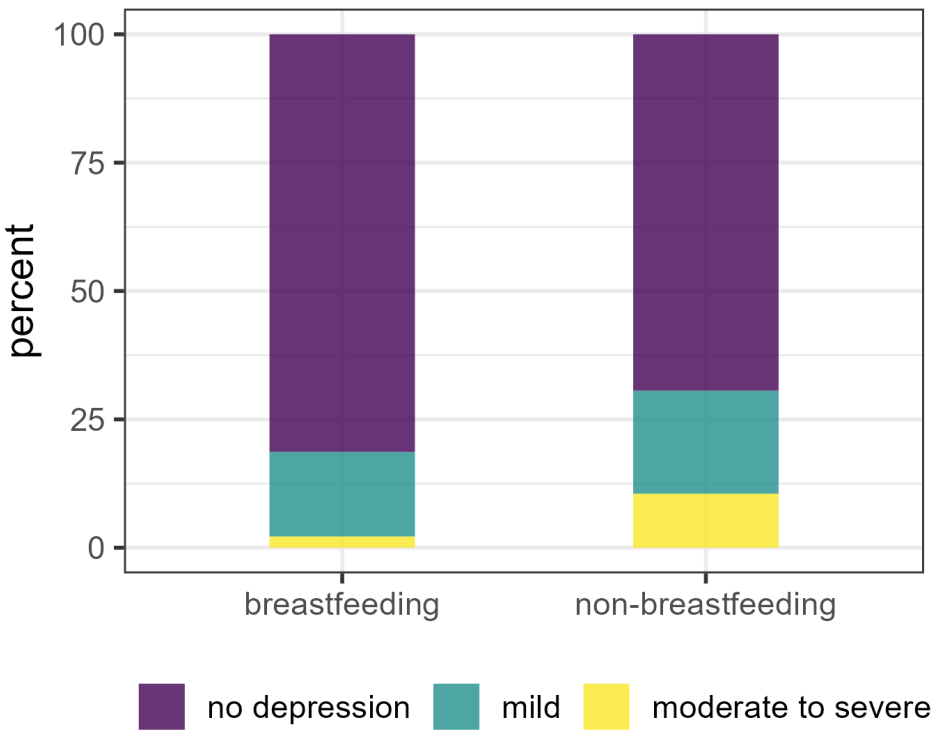
- Summed results from the *Patient Health Questionnaire*, PHQ-9
- 9 questions, each ranging from 0 to 3
- binned into categories based on existing standards

Subject Counts for Depression Categories

Depression Level	Pregnant	Postpartum	Non-pp	Men	Total
None	193	330	3489	4321	8333
Mild	59	82	881	777	1799
Moderate to severe	23	29	579	338	969
Total	275	441	5061	4949	11,101

Relative Distribution of Depression Categories

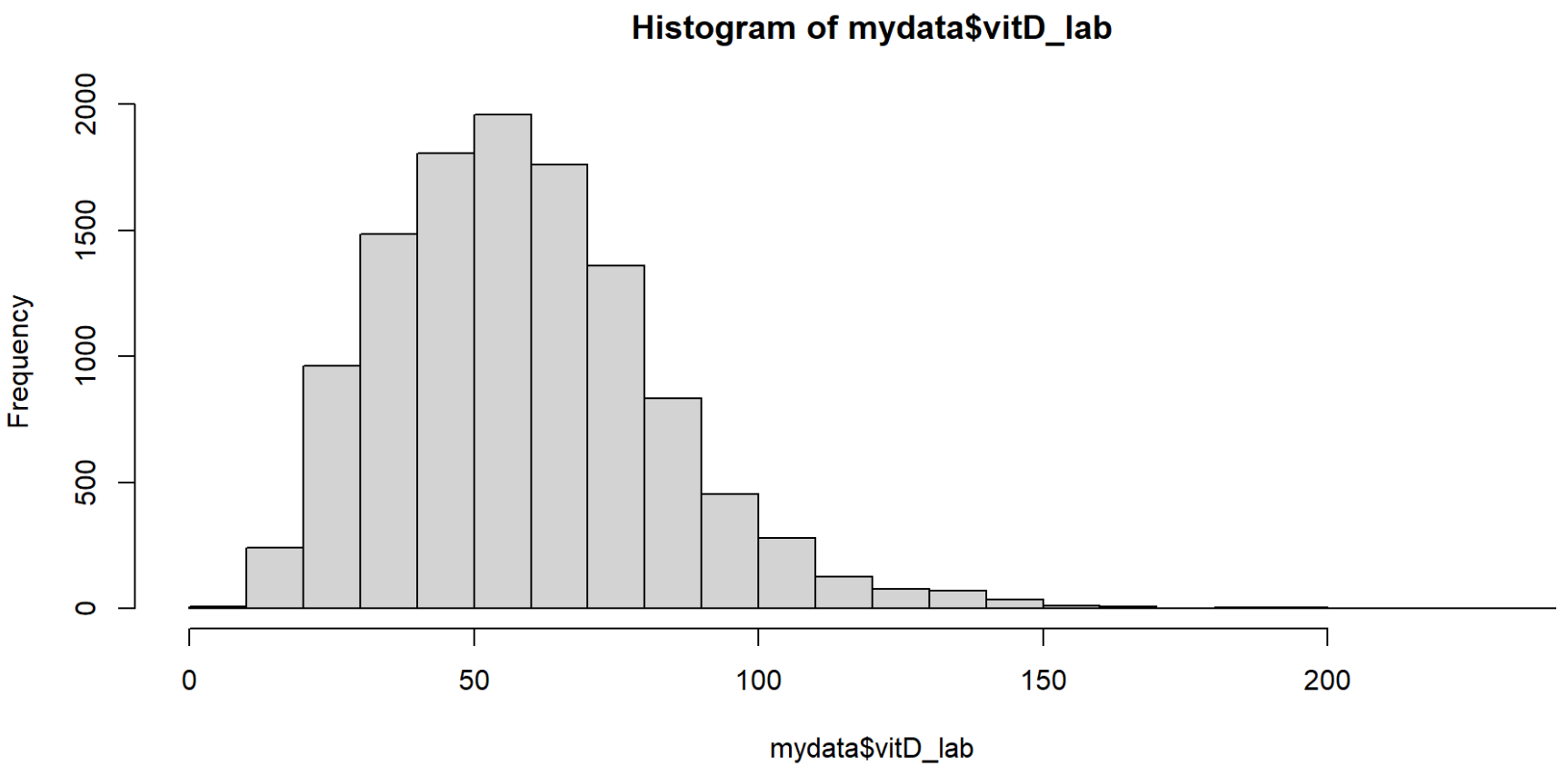




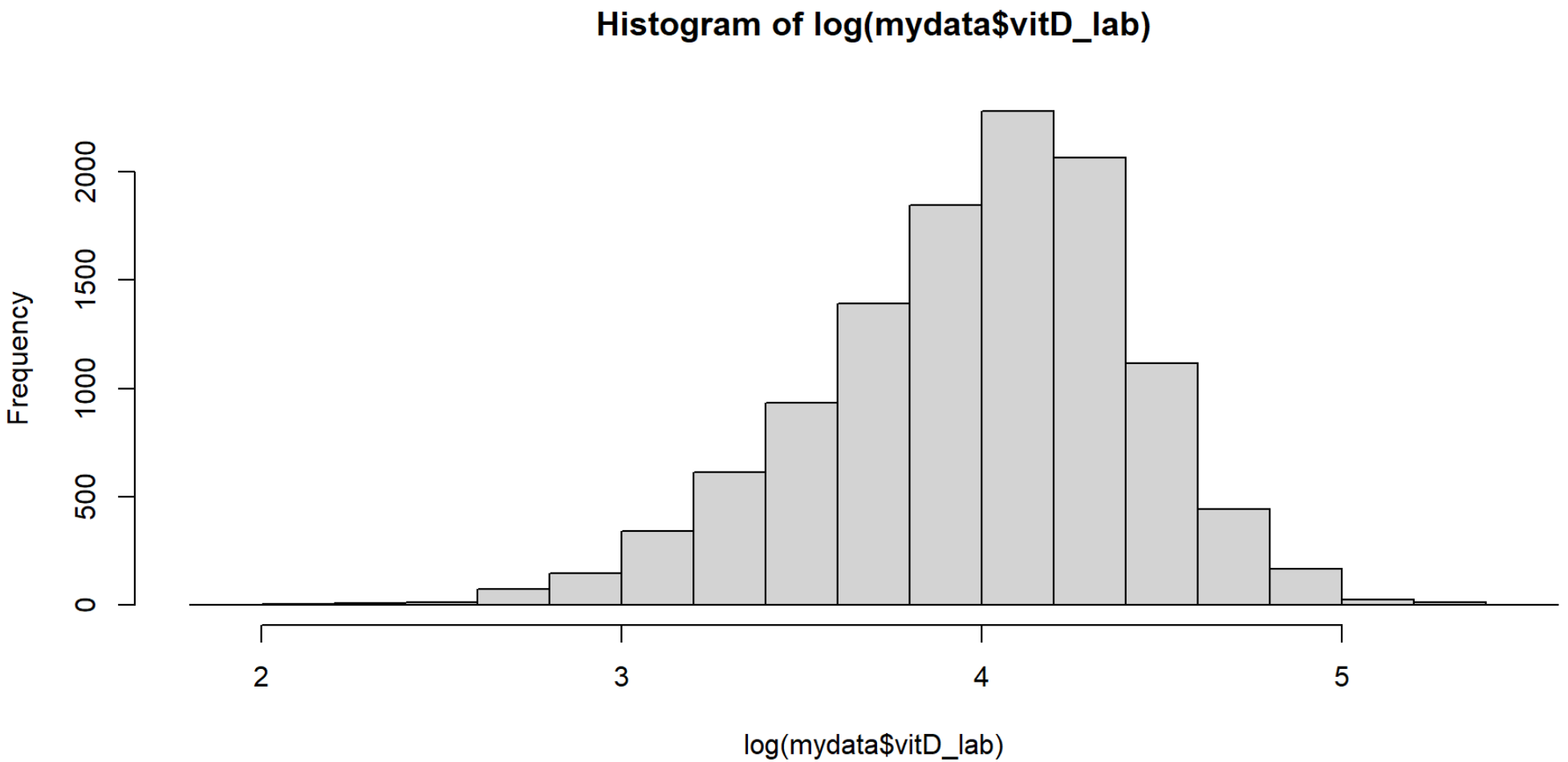
These are the relative percentages after applying survey weights. They reflect the overall population estimates.

Vitamin D Distribution

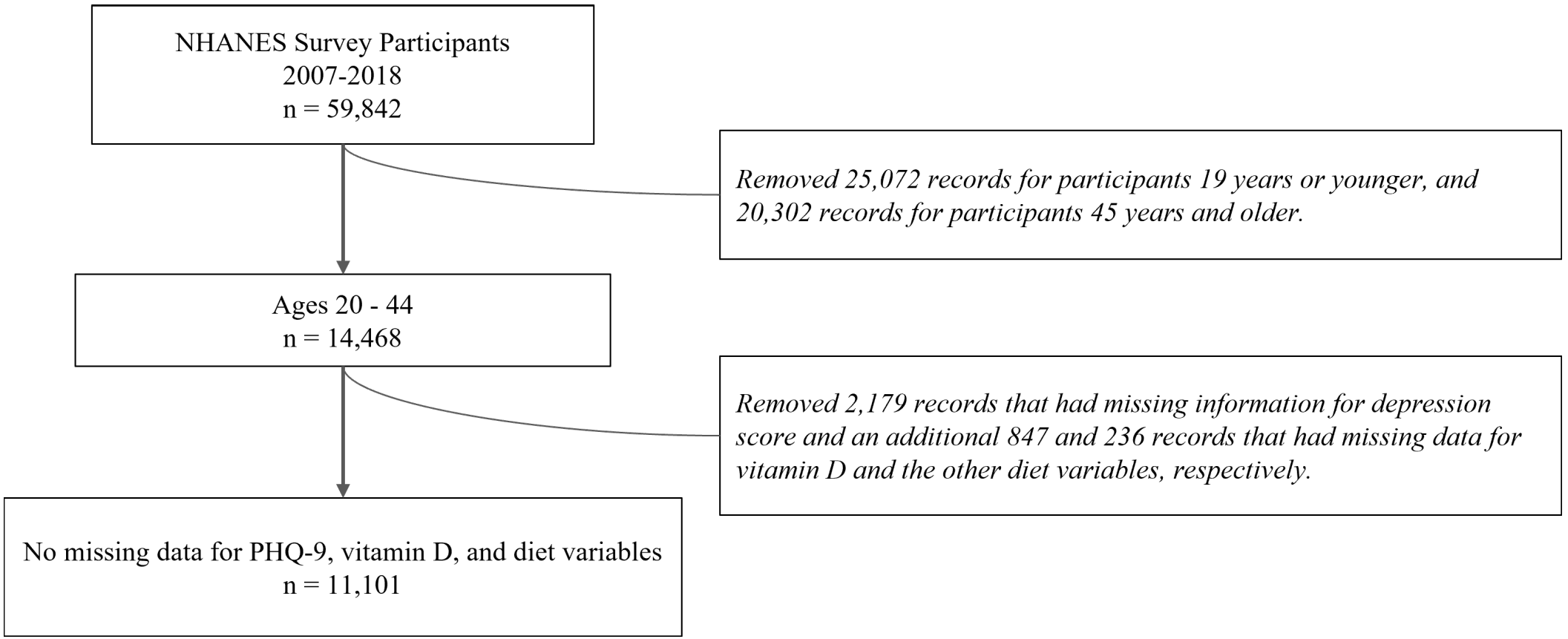
- Total (blood) serum vitamin D (D2 + D3) in nmol/L



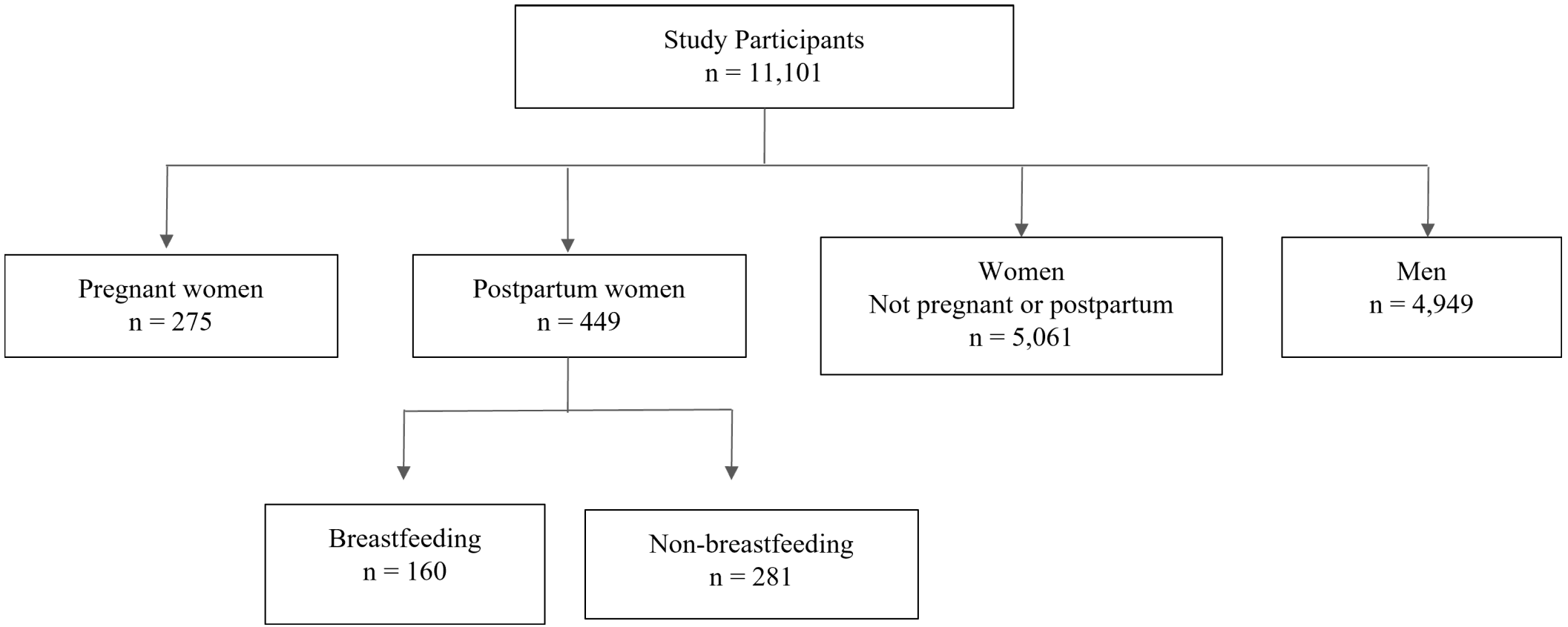
Log Transformed Vitamin D Distribution



Our Population



Our Population



Positivity & Propensity Scores

One assumption of causal inference: **Positivity**

- Within each level and combination of the study variables used, each individual has some chance of experiencing every available exposure level
- Statistically: there are no associations between the 'independent' variables
- This assumption is often not met in observational studies *a priori*, so we use propensity scores as weights to force the observations to mimic independence

Propensity Scores

For continuous exposures like vitamin D concentration

$$VitaminD_{2i} = \beta_0$$

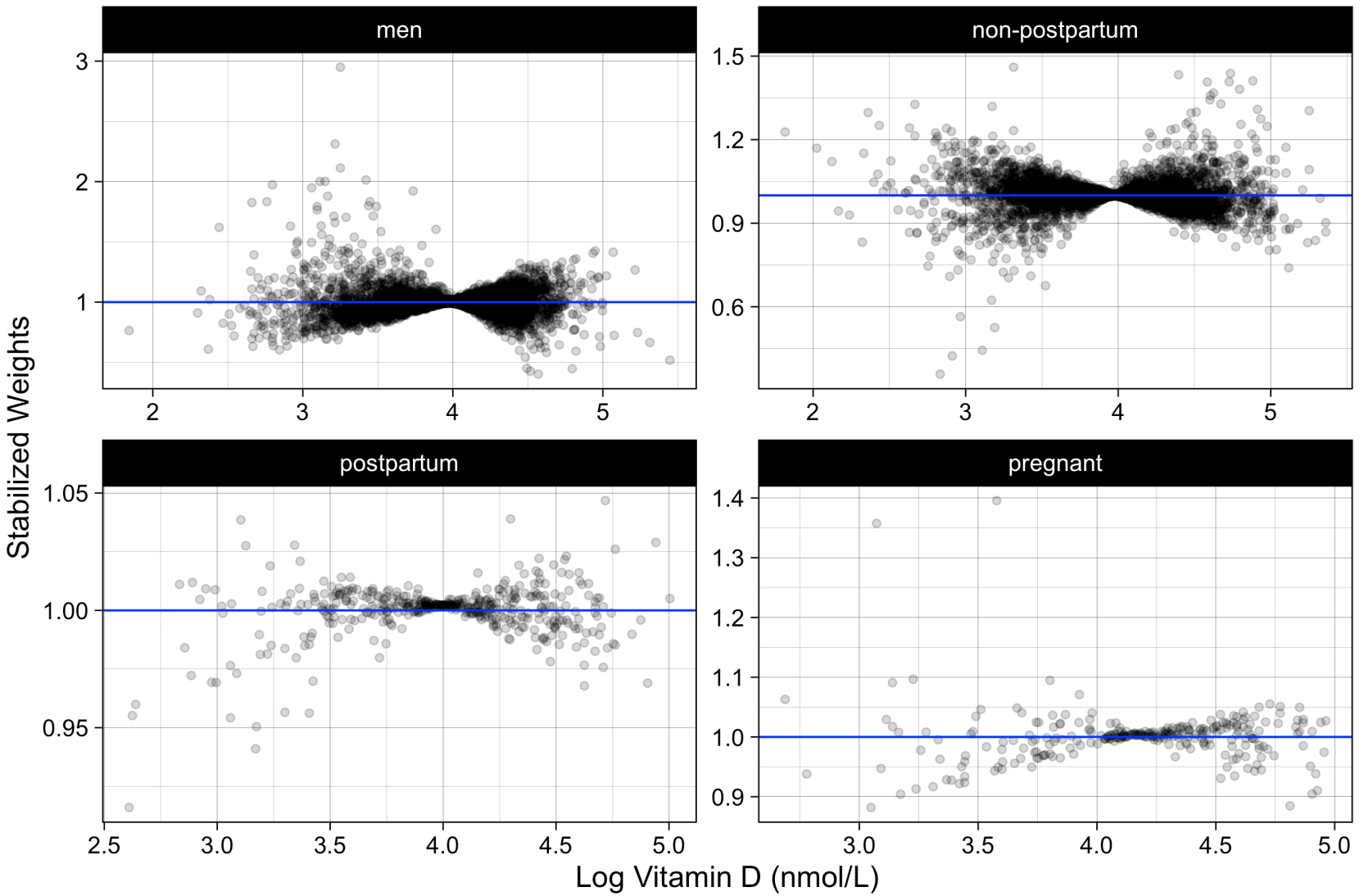
$$VitaminD_{1i} \sim N(\hat{Y_{1i}}, \hat{\sigma_1})$$

$$VitaminD_{2i} \sim N(\hat{Y_2}, \hat{\sigma_2})$$

Weights are the inverse of the propensity scores:

$$PS = \frac{f(normal\ PDF|Y_i, \hat{Y_{1i}}, \hat{\sigma_1})}{f(normal\ PDF|Y_i, \hat{Y_2}, \hat{\sigma_2})}$$

Propensity Score Weights



Statistical Model

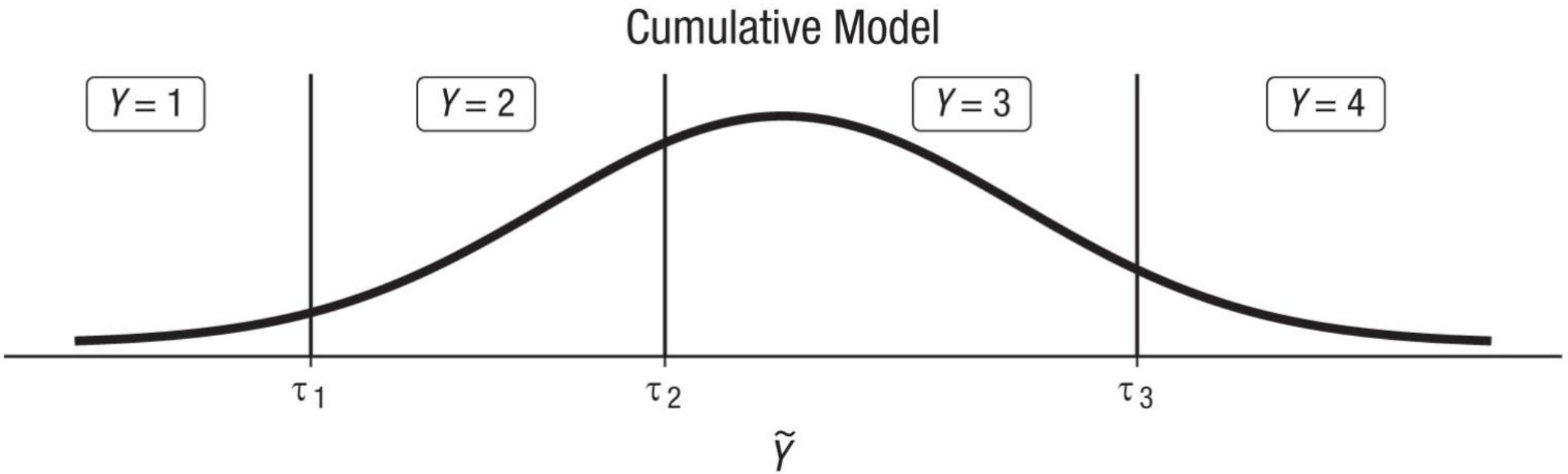
(for each cohort)

Proportional-odds cumulative logit model:

$$Depression = log(Vitamin\ D) + Diet\ V$$

$$log\left(\frac{Pr(Y_i > j)}{Pr(Y_i \leq j)}\right) = \beta_{0j} + \beta_j X$$

Cumulative Ordinal Models



Map proportions to a normal distribution and establish breakpoints between adjacent categories.

Model Details

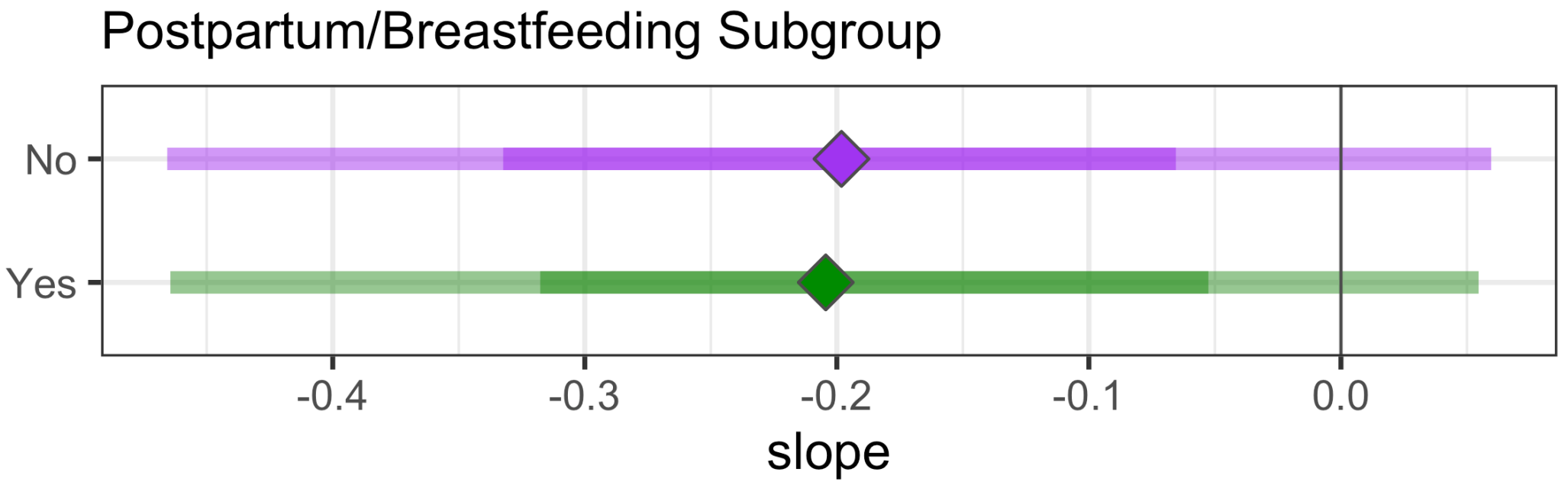
- Bayesian models fit using R version 4.4 and the package ‘brms’ (a stan wrapper)
- 4 chains, 2000 iterations with 1000 as ‘warmup’
- Compared cumulative logit and cumulative probit models
- Separate models fit for each study cohort (men, women, pregnant women, postpartum women)
- Joint model fit for postpartum breastfeeding subpopulation analysis assuming unequal variance for the cohorts

Results

Model Fit

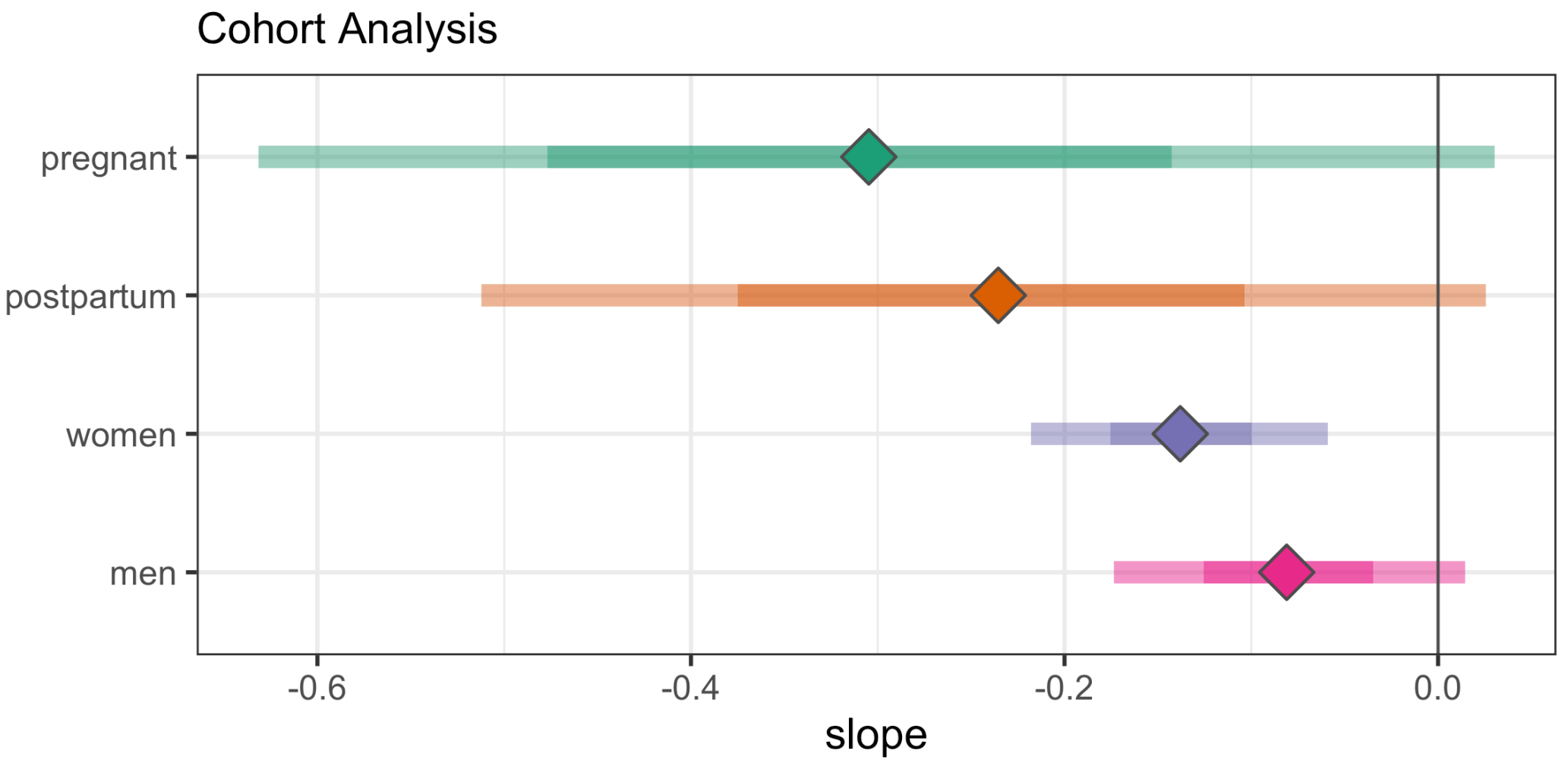
- All MCMC chains converged
- No meaningful difference in probit versus logit models according to WAIC
- and effective sample size ratio statistics in acceptable ranges $\hat{R}$

Estimates!



- The plotting symbol is the mean estimate
- The light-colored line is the 95% HDI
- The dark line is the 68% HDI

Estimates!

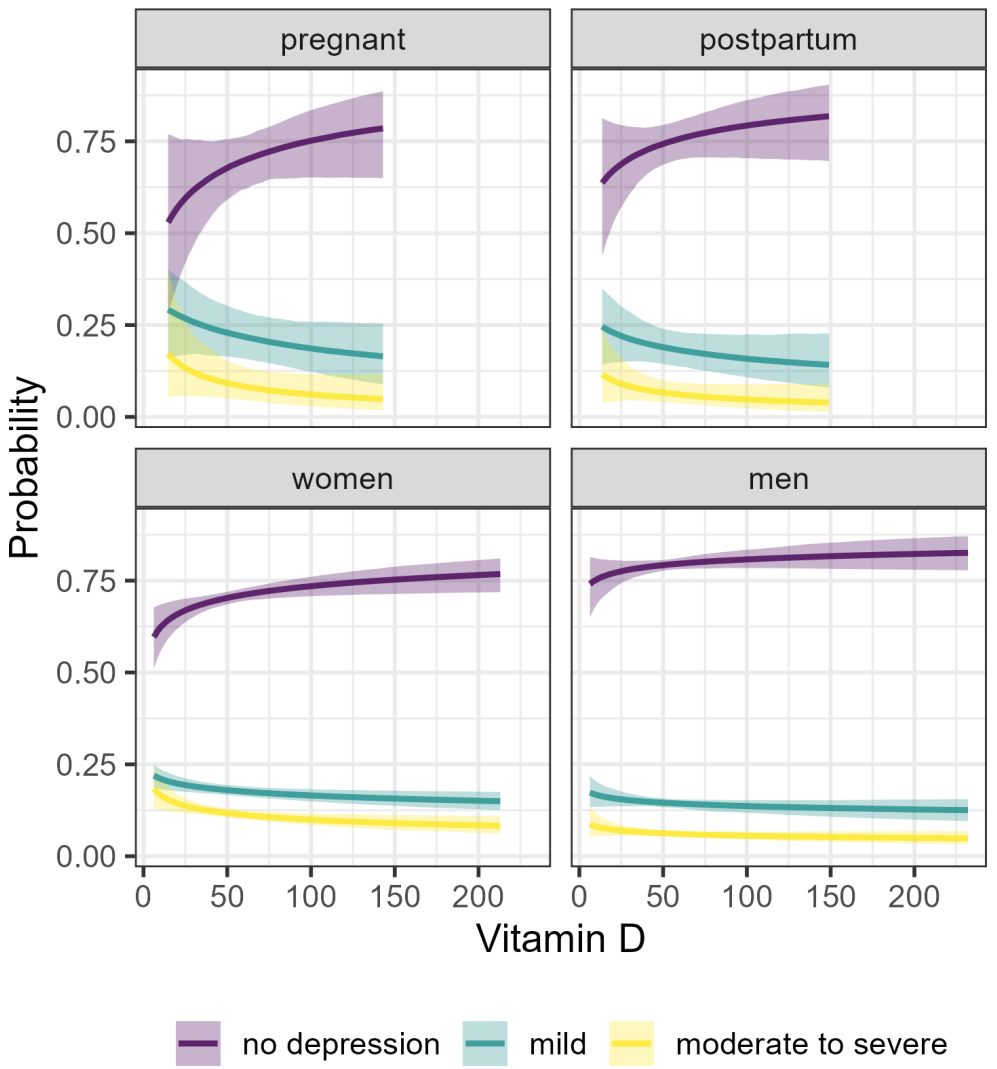

Cohort Analysis
pregnant
postpartum
women
men
-0.6
-0.4
-0.2
0.0
slope

Hypothesis Testing

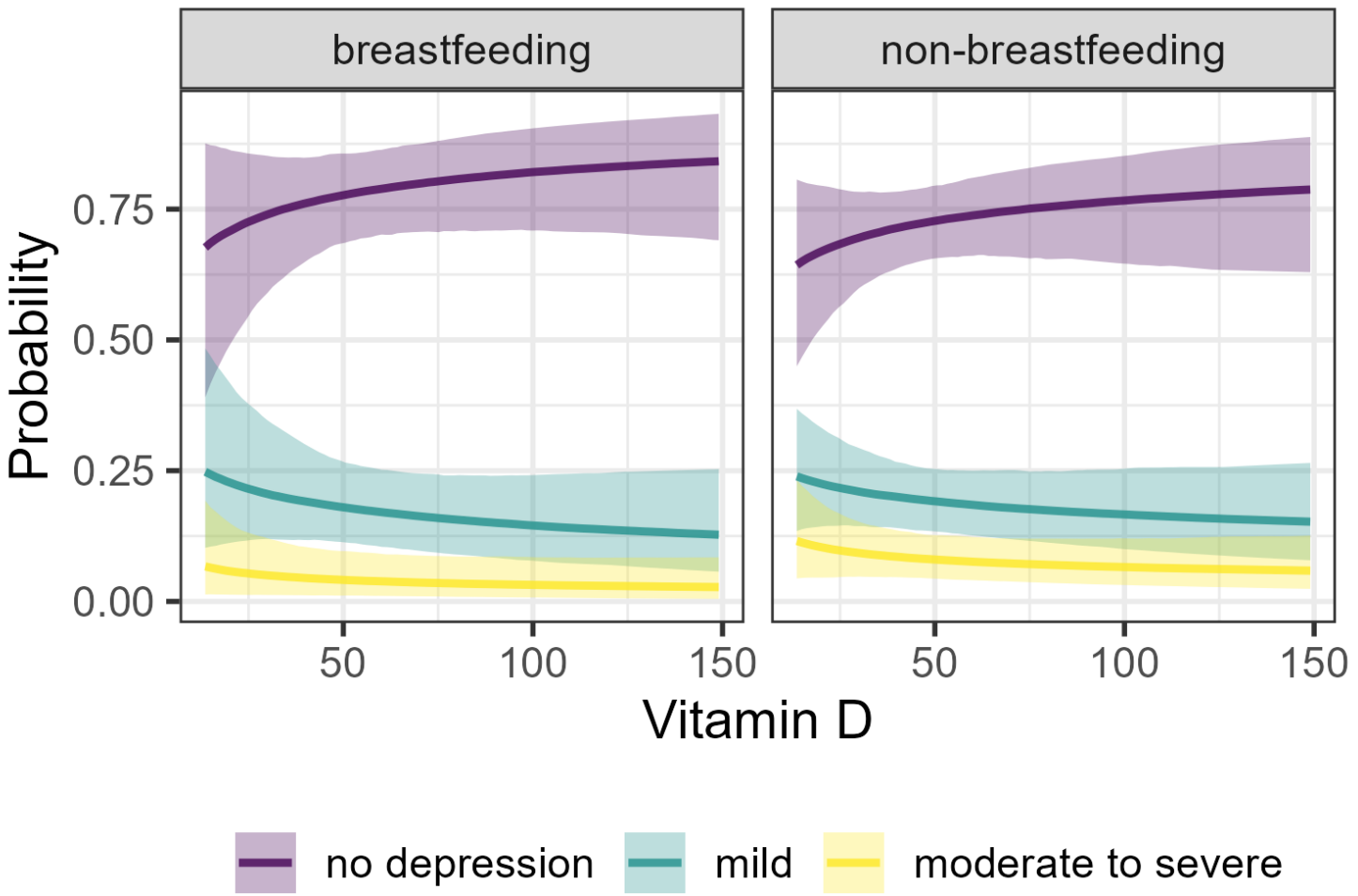
Using Posterior Probabilities

- Postpartum/breastfeeding subgroups: ? $B_{VitD/Yes} - B_{VitD/N}$
  - Posterior probability: 0.55
- All cohorts: ? $B_{VitD} < 0$
  - pregnant women: 0.97
  - postpartum women: 0.95
  - women: 1.00
  - men: 0.96

Estimates in Practice



Breastfeeding Subgroup Analysis



Conclusions on Vitamin D & Depression

- Vitamin D concentration does impact depression outcomes in this study population designed to reflect the full U.S. population. **As serum Vitamin D increases, the probability of depression increases.**
- Pregnant women have the most pronounced effect due to vitamin D exposure, followed by postpartum women, other women and men, however, the analysis did not quantify and evaluate those differences
- After adjusting for general diet effects, vitamin D is one causative contributing factor towards depression outcomes.
- Although the depression incidence differed between postpartum women who were breastfeeding compared to those who are not, the impact of Vitamin D on depression outcomes did not differ across the two populations.

The Moral of the Story

- Think through your analysis. Determine what is it you want to know and build your analytical plan around that.
- Make your model framework explicit. It is helpful for everyone to be clear about the modelling statement and assumptions.
- Always check the distributions of your data and make sure they fit model expectations.
- Thoughtful work and causal inference improves the quality of science. Statistics and science are not magic, but we can estimate certain things under certain conditions.
- Initiate collaborations early in the process.
- Statisticians/data scientists: domain knowledge is critical, so work with subject experts.
- Collaborations can bring amazing results.

Resources

- Code: GitHub Repo
- Our paper: Hollinshead, Piaskowski & Chen, 2024
- *Statistical Rethinking* by Richard McElreath, (book, YouTube, Github)
- Guide to Causal Inference in R
- *Causality: Models, Reasoning and Inference* (2nd Ed.) by Judea Pearl

> If we are very careful and try very hard, we might not completely mislead ourselves.
>
> – R. McElreath