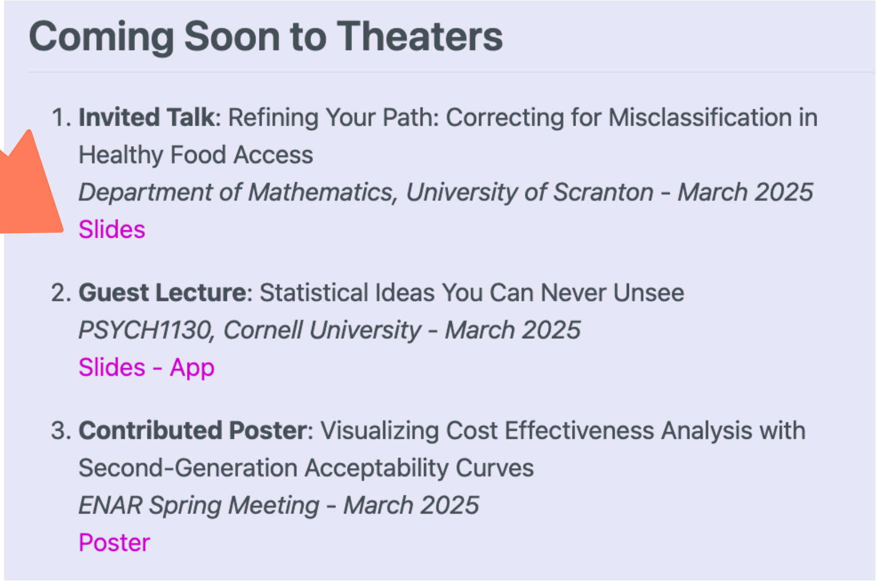
Refining Your Path: Correcting for Misclassification in Healthy Food Access

Ashley E. Mullan - March 7, 2025

Scan or search to follow along with me!

https://bit.ly/ash-talks



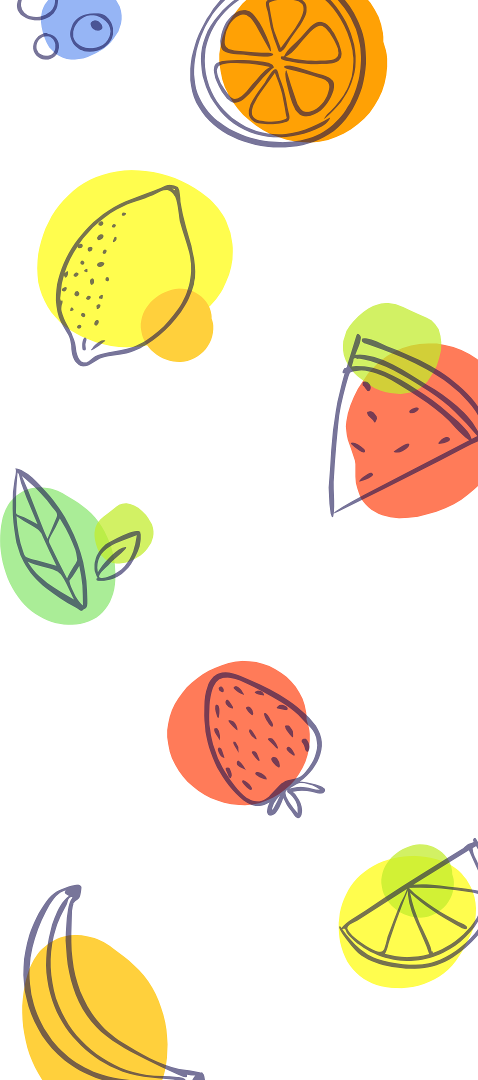
Coming Soon to Theaters

1. **Invited Talk**: Refining Your Path: Correcting for Misclassification in Healthy Food Access
   Department of Mathematics, University of Scranton - March 2025
   Slides

2. **Guest Lecture**: Statistical Ideas You Can Never Unsee
   PSYCH1130, Cornell University - March 2025
   Slides - App

3. **Contributed Poster**: Visualizing Cost Effectiveness Analysis with Second-Generation Acceptability Curves
   ENAR Spring Meeting - March 2025
   Poster

bitly

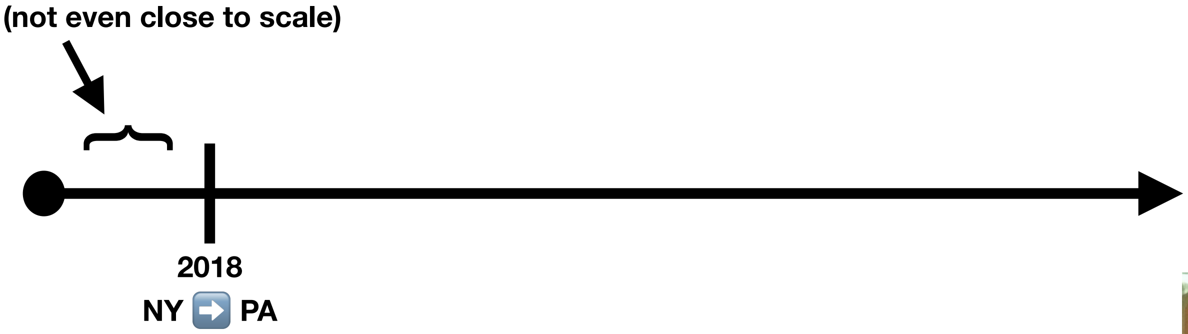


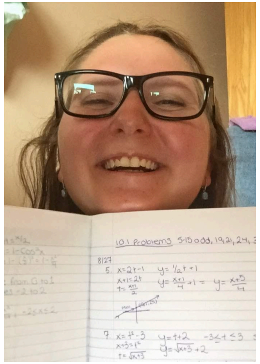
Part I.

Refining MY Path

How, exactly, did I end up on this side of department seminar?

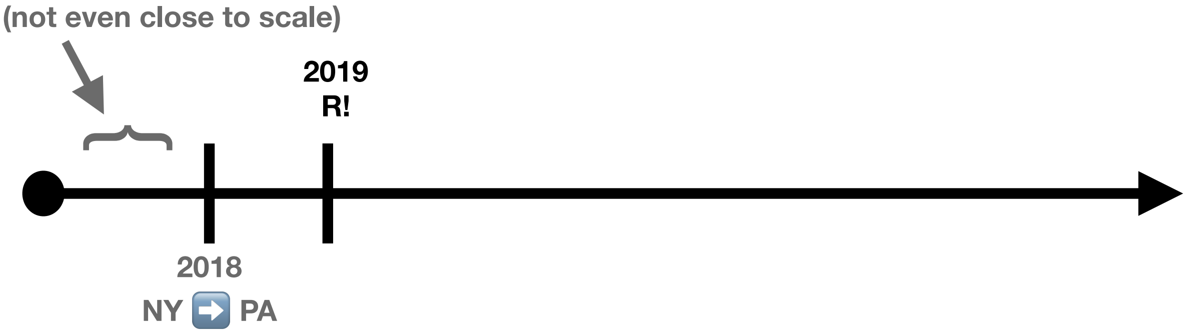
A Linear Combination* of Events





*tip from the pros, you're going to want to pay attention in linear algebra! it comes in handy :)

A Linear Combination* of Events



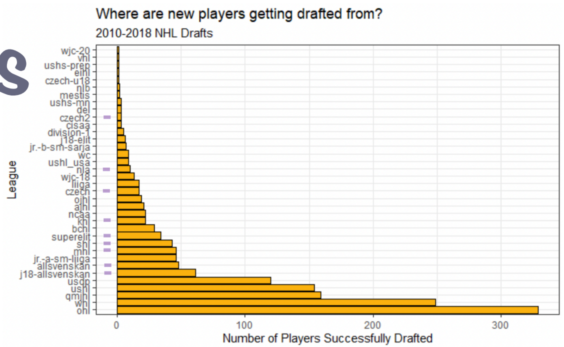
(not even close to scale)

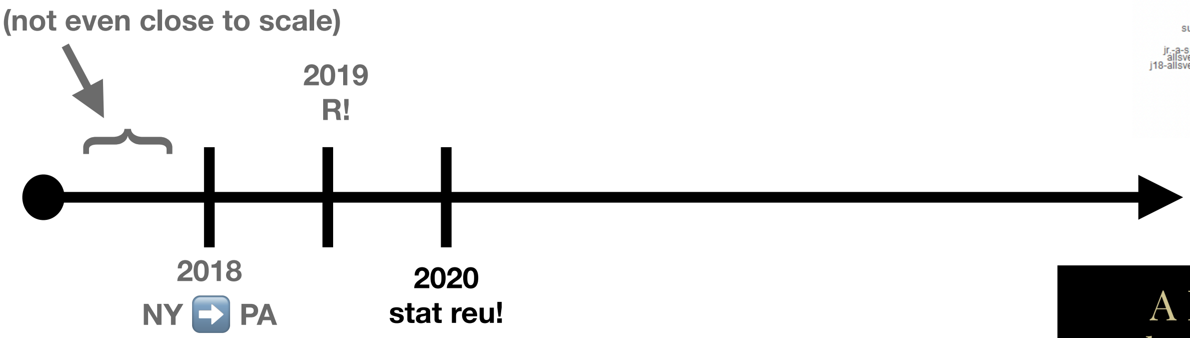
2019
R!

2018
NY ➡️ PA

*tip from the pros, you're going to want to
pay attention in linear algebra! it comes in handy :)

A Linear Combination* of Events



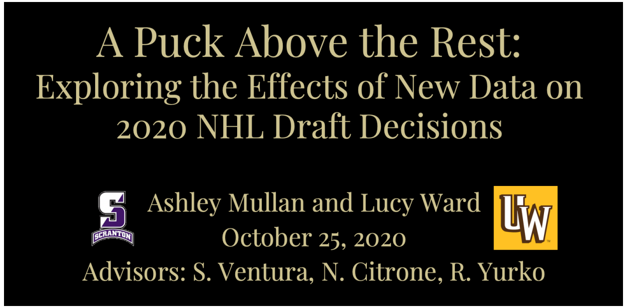


(not even close to scale)

2019
R!

2018
NY ➡️ PA

2020
stat reu!

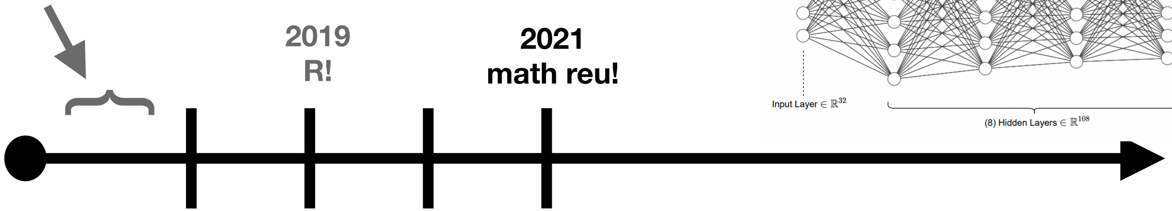


A Puck Above the Rest:
Exploring the Effects of New Data on 2020 NHL Draft Decisions

Ashley Mullan and Lucy Ward
October 25, 2020
Advisors: S. Ventura, N. Citrone, R. Yurko

*tip from the pros, you're going to want to pay attention in linear algebra! it comes in handy :)

A Linear Combination* of Events



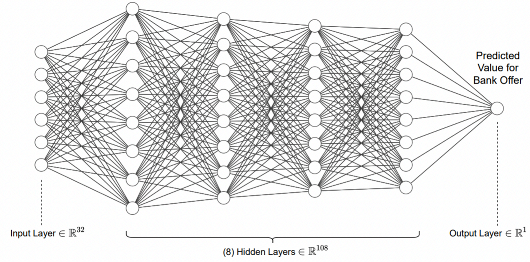
(not even close to scale)

2018
NY ➡️ PA

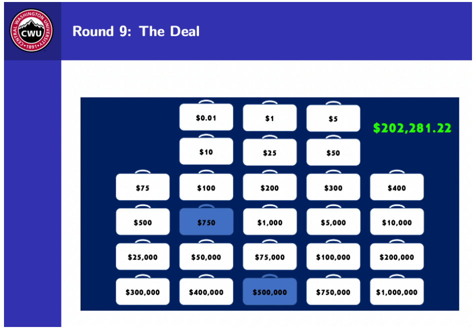
2019
R!

2020
stat reu!

2021
math reu!



Predicted Value for Bank Offer
Input Layer $\in \mathbb{R}^{32}$
(8) Hidden Layers $\in \mathbb{R}^{108}$
Output Layer $\in \mathbb{R}^{1}$



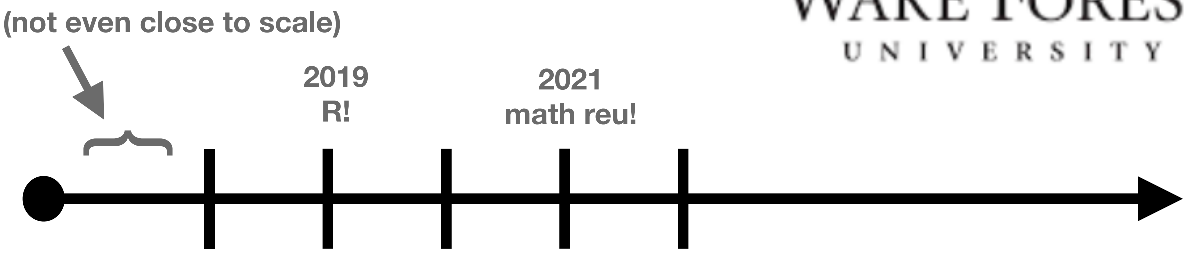
Round 9: The Deal

	$0.01	$1	$5	$202,281.22
	$10	$25	$50	
$75	$100	$200	$300	$400
$500	$750	$1,000	$5,000	$10,000
$25,000	$50,000	$75,000	$100,000	$200,000
$300,000	$400,000	$500,000	$750,000	$1,000,000

*tip from the pros, you're going to want to pay attention in linear algebra! it comes in handy :)

A Linear Combination* of Events

WAKE FOREST
UNIVERSITY


(not even close to scale)

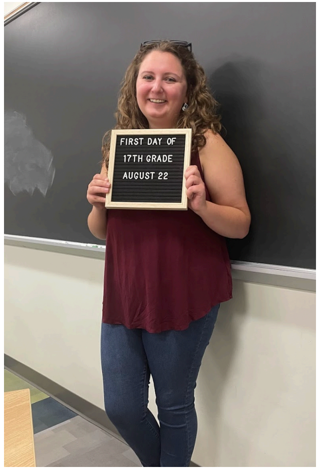
2019
R!

2021
math reu!

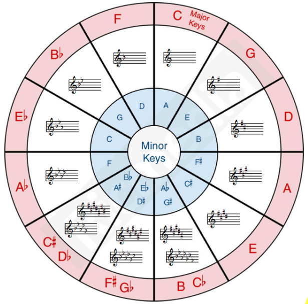
2018
NY ➡️ PA

2020
stat reu!

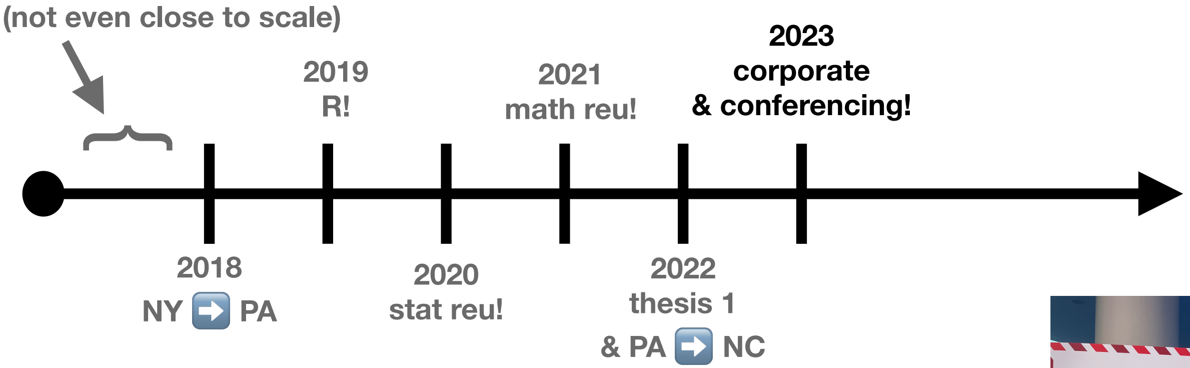
2022
thesis 1
& PA ➡️ NC



*tip from the pros, you're going to want to pay attention in linear algebra! it comes in handy :)



A Linear Combination* of Events



(not even close to scale)

2018
NY ➡️ PA

2019
R!

2020
stat reu!

2021
math reu!

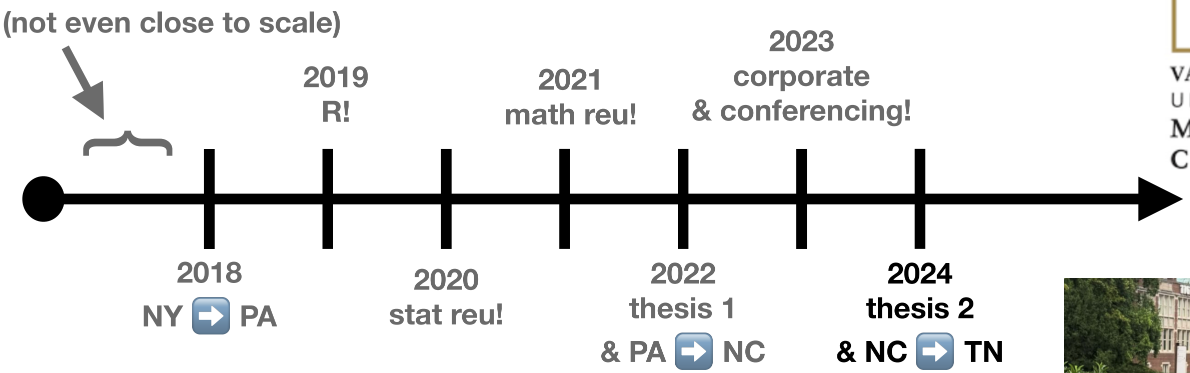
2022
thesis 1
& PA ➡️ NC

2023
corporate
& conferencing!

*tip from the pros, you're going to want to pay attention in linear algebra! it comes in handy :)

Hanes®

A Linear Combination* of Events



(not even close to scale)

2018
NY ➡️ PA

2019
R!

2020
stat reu!

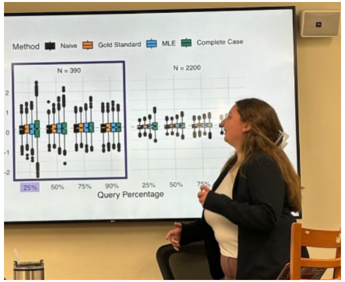
2021
math reu!

2022
thesis 1
& PA ➡️ NC

2023
corporate
& conferencing!

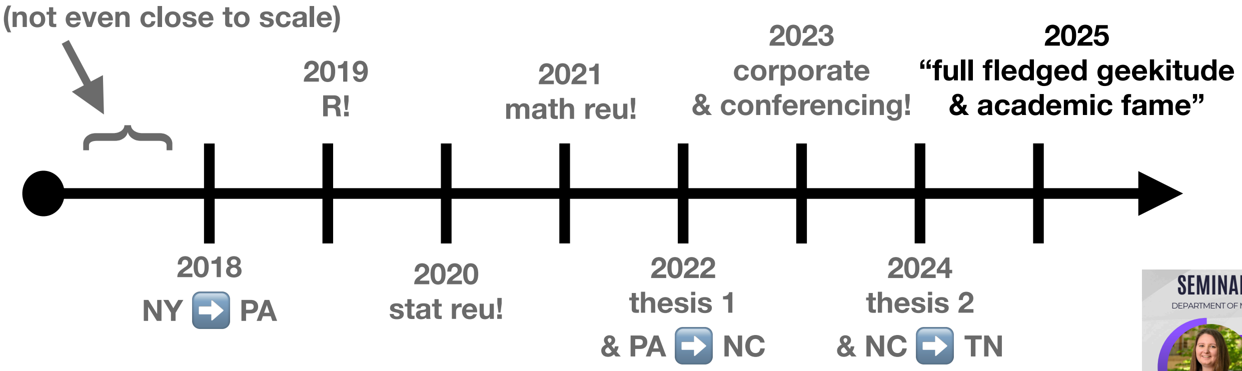
2024
thesis 2
& NC ➡️ TN

VANDERBILT
UNIVERSITY
MEDICAL
CENTER



*tip from the pros, you're going to want to pay attention in linear algebra! it comes in handy :)

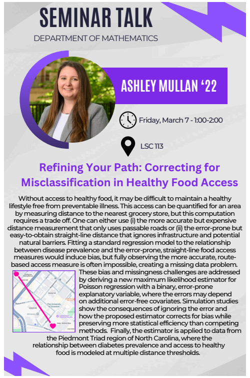
A Linear Combination* of Events



SIGNIFICANCE

a direct quote from Dr. Dougherty

*tip from the pros, you're going to want to pay attention in linear algebra! it comes in handy :)



SEMINAR TALK

DEPARTMENT OF MATHEMATICS

ASHLEY MULLAN '22

Friday, March 7 - 1:00-2:00

LSC 113

Refining Your Path: Correcting for Misclassification in Healthy Food Access

Without access to healthy food, it may be difficult to maintain a healthy lifestyle free from preventable illness. This access can be quantified for an area by measuring distance to the nearest grocery store, but this computation requires a trade off. One can either use (i) the more accurate but expensive distance measurement that only uses passable roads or (ii) the error-prone but easy-to-obtain straight-line distance that ignores infrastructure and potential natural barriers. Fitting a standard regression model to the relationship between disease prevalence and the error-prone, straight-line food access measures would induce bias, but fully observing the more accurate, route-based access measure is often impossible, creating a missing data problem. These bias and missingness challenges are addressed by deriving a new maximum likelihood estimator for Poisson regression with a binary, error-prone explanatory variable, where the errors may depend on additional error-free covariates. Simulation studies show the consequences of ignoring the error and how the proposed estimator corrects for bias while preserving more statistical efficiency than competing methods. Finally, the estimator is applied to data from the Piedmont Triad region of North Carolina, where the relationship between diabetes prevalence and access to healthy food is modeled at multiple distance thresholds.

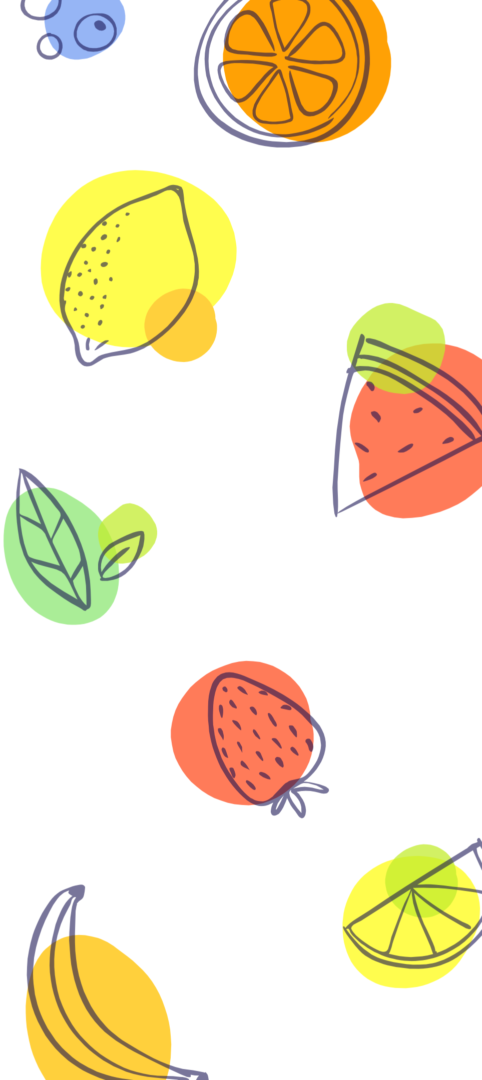
Part 2.

Refining Actual Paths

Y’all were promised some research!

Here's the plan!

1. Motivation
2. The Math
3. Battle Royale (Methods Edition)
4. Case Study
5. Odds & Ends



1

Motivation

Healthy Eating ➡️ Healthy Living

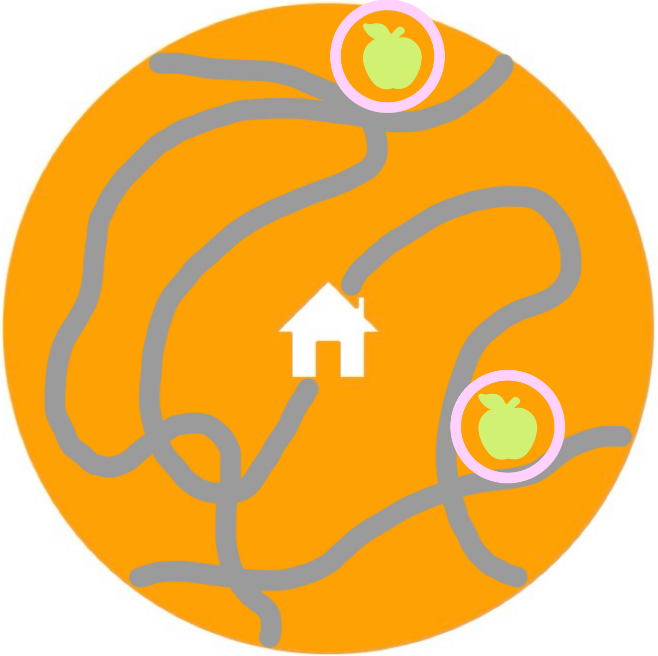
- A **healthy diet** is full of fruits, vegetables, whole grains, and other high-nutrient foods.
- A healthy diet increases the likelihood of good overall health and **decreases risk of preventable illness** (World Health Organization, 2019).
- Maintaining a healthy diet requires **consistent access to healthy food**, which may be hindered by physical or social barriers like geography or income.
- Review studies found **high prevalence of diabetes** in food-insecure households (Gucciardi et al., 2014).

Measuring Food Access 🍎

- Define a **neighborhood** of interest with a **radius**, a **centroid**, and possibly some **healthy food retailers**.
- We pick one of **three common methods** to quantify food access.

Measuring Food Access 🍎

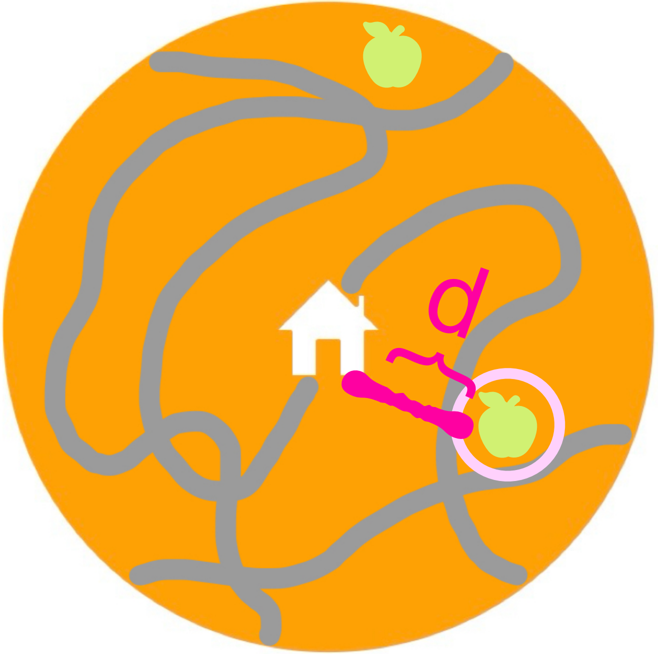
- Define a **neighborhood** of interest with a **radius**, a **centroid**, and possibly some **healthy food retailers**.
- We pick one of **three common methods** to quantify food access.



density: 2

Measuring Food Access 🍎

- Define a **neighborhood** of interest with a **radius**, a **centroid**, and possibly some **healthy food retailers**.
- We pick one of **three common methods** to quantify food access.



proximity: d

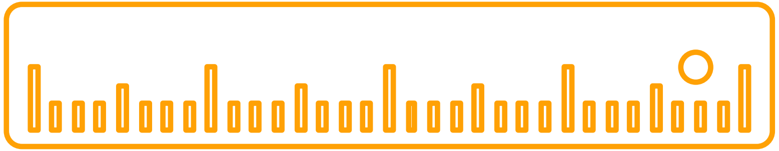
Measuring Food Access 🍎

- Define a **neighborhood** of interest with a **radius**, a **centroid**, and possibly some **healthy food retailers**.
- We pick one of **three common methods** to quantify food access.

indicator: 👍

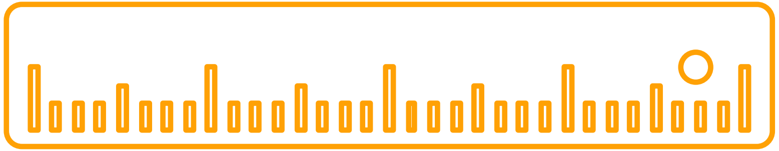
Measuring Food Access 🍎

- All of these methods require some notion of **distance**!



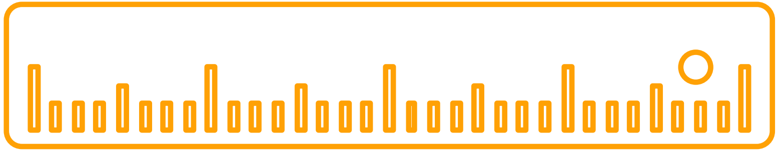
Measuring Food Access 🍎

- All of these methods require some notion of **distance**!

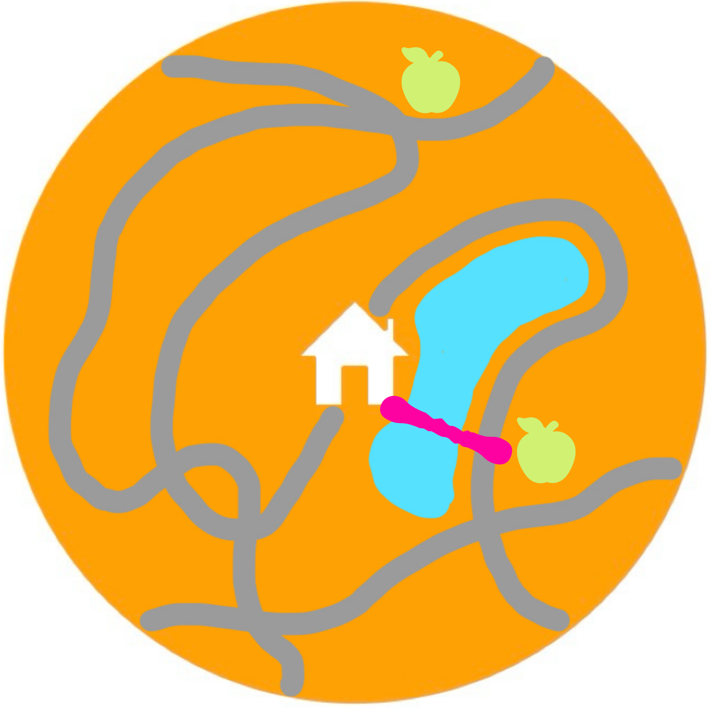


Measuring Food Access 🍎

× All of these methods require some notion of **distance**!

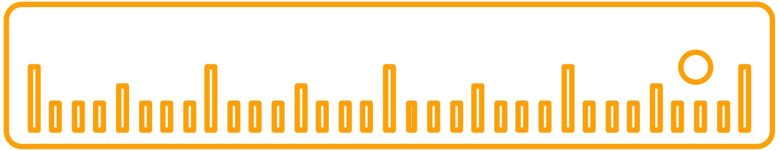


straight line: easy to see but sometimes a little tough to do



Measuring Food Access 🍎

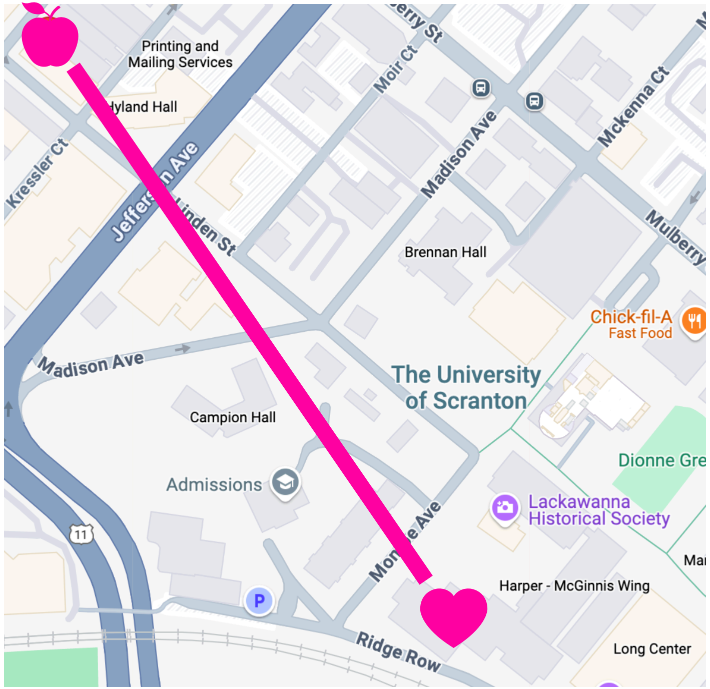
- All of these methods require some notion of **distance**!



route-based: tougher to find but sometimes easier to walk

Distance Computations ✏️

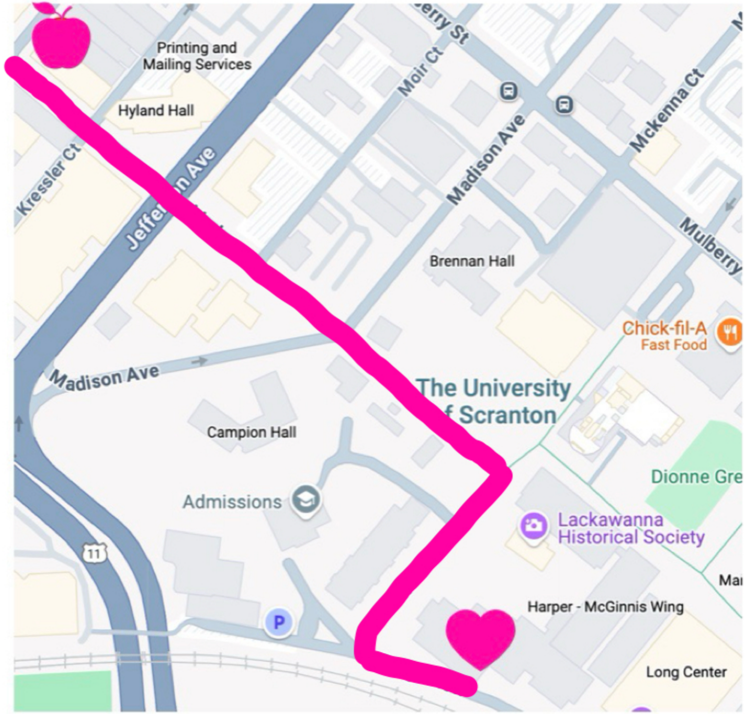
- The **Haversine distance** is a trigonometric function of latitude and longitude.
- It ignores physical obstacles, so it **underestimates** the true distance between two points and is considered **error-prone**.



0.2 miles

Distance Computations ✏️

- The **route-based** distance works around obstacles.
- It is **more accurate** than the Haversine distance, but it is **computationally and financially expensive**.
- These distances can be found with the **ggmap** package in R, which queries Google Maps.



0.4 miles

The real world stinks, but we do our best! 🌎

- In a **perfect world**, we'd have the route-based access indicators for **every neighborhood** of interest.
- In **real life**, we have a cap on how many route-based distances we can get, so **not every neighborhood** has an available route-based access indicator.
- Luckily, we do have a (**potentially misclassified**) Haversine-based access indicator for **every neighborhood**. Sometimes, we have both!

The Big Question*

- Normally, when you put in bad data, your models spit out bad results. Garbage in, garbage out!
- Can we sidestep the **misclassification and missingness issues** in our data and still accurately estimate an association between **access to healthy food** and **diabetes prevalence**?

*spoiler alert, we can! it just takes a little math 😊

2

The Math

First, the buzzwords!

- **Poisson distribution**: a probability model used for count data
- **Bernoulli distribution**: a probability model used for binary data, parameterized by success rate *p*
- **Joint likelihood function**: estimates how reasonable a parameter vector is given your specific observed data
- **Conditional probability**: incorporates data about one variable to tell you about another, used to factor joint probabilities

For neighborhood i, we observe:

$$(Y_i, X_i, X_i^*, \mathbf{Z}_i)$$

For neighborhood i, we observe:

$$(Y_i, X_i, X_i^*, Z_i)$$

the **outcome**, representing the number of diabetes cases in the neighborhood

$$Y_i \in \{1,2,3,....\}$$

For neighborhood i, we observe:*

$X_i \in \{0,1\}$

$$(Y_i, X_i, X_i^*, Z_i)$$

the **exposure**, representing the route-based food access indicator for that neighborhood

*terms and conditions apply, only known if **query indicator** $Q_i = 1$

For neighborhood i, we observe:

$X_i^* \in \{0,1\}$

$$(Y_i, X_i, X_i^*, Z_i)$$

the **error-prone exposure**, representing the straight line food access indicator for that neighborhood

For neighborhood i, we observe:

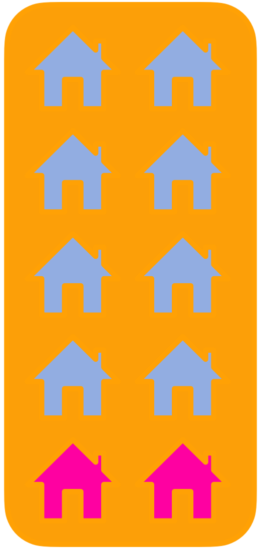
$$(Y_i, X_i, X_i^*, \boxed{\mathbf{Z}_i})$$

the **covariate vector**, representing everything else we know about that neighborhood (assumed to be error-free)

$$\mathbf{Z}_i \in \mathbb{R}^k$$

Back to those terms and conditions...

- Having X for **some** of the N neighborhoods is better than having it for none!
- **Two phase design** gives us the most bang for our buck.
- We **only** have X^* for $N - n$ neighborhoods, but we have **both** X^* and X for n of them!
- X^* is subject to **misclassification**, but X is subject to **missingness**.



In my Barbie dreamworld, we always have X$_i$!

$$P(Y_i, X_i, \mathbf{Z}_i) = P_\beta(Y_i \mid X_i, \mathbf{Z}_i) P_\eta(X_i \mid \mathbf{Z}_i) P(\mathbf{Z}_i)$$

In my Barbie dreamworld, we always have X_i!

$$P(Y_i, X_i, \mathbf{Z}_i) = P_\beta(Y_i \mid X_i, \mathbf{Z}_i) P_\eta(X_i \mid \mathbf{Z}_i) P(\mathbf{Z}_i)$$

This is the **outcome model!**

$$Y_i \mid X_i, \mathbf{Z}_i \sim \text{Poisson}(\beta_0 + \beta_1 X_i + \beta_2 \mathbf{Z}_i + \log(O_i))$$

an offset term representing the population size

In my Barbie dreamworld, we always have X_i!

$$P(Y_i, X_i, \mathbf{Z}_i) = P_\beta(Y_i \mid X_i, \mathbf{Z}_i) P_\eta(X_i \mid \mathbf{Z}_i) P(\mathbf{Z}_i)$$

This is the **outcome model!**

$$Y_i \mid X_i, \mathbf{Z}_i \sim \text{Poisson}(\beta_0 + \beta_1 X_i + \beta_2 \mathbf{Z}_i + \log(O_i))$$

$\exp(\beta_1)$ gives us the association we want!

In my Barbie dreamworld, we always have X_i!

$$P(Y_i, X_i, \mathbf{Z}_i) = P_\beta(Y_i \mid X_i, \mathbf{Z}_i) P_\eta(X_i \mid \mathbf{Z}_i) P(\mathbf{Z}_i)$$

This is going to become the **error model!**

In my Barbie dreamworld, we always have X_i!

$$P(Y_i, X_i, \mathbf{Z}_i) = P_\beta(Y_i \mid X_i, \mathbf{Z}_i) P_\eta(X_i \mid \mathbf{Z}_i) P(\mathbf{Z}_i)$$

nobody cares, it drops out later

Sadness, we consider error now.

$$P(Y_i, X_i, \overset{\downarrow}{X_i^*}, \mathbf{Z}_i) = P_\beta(Y_i \mid X_i, \mathbf{Z}_i)P_\eta(X_i \mid \overset{\downarrow}{X_i^*}, \mathbf{Z}_i)P(\overset{\downarrow}{X_i^*}, \mathbf{Z}_i)$$

Sadness, we consider error now.

$$P(Y_i, X_i, \overset{\downarrow}{X_i^*}, \mathbf{Z}_i) = \boxed{P_\beta(Y_i \mid X_i, \mathbf{Z}_i)}P_\eta(X_i \mid \overset{\downarrow}{X_i^*}, \mathbf{Z}_i)P(\overset{\downarrow}{X_i^*}, \mathbf{Z}_i)$$

This is the same **outcome model!**

Sadness, we consider error now.

$$P(Y_i, X_i, \overset{\downarrow}{X_i^*}, \mathbf{Z}_i) = P_\beta(Y_i \mid X_i, \mathbf{Z}_i)P_\eta(X_i \mid \overset{\downarrow}{X_i^*}, \mathbf{Z}_i)P(\overset{\downarrow}{X_i^*}, \mathbf{Z}_i)$$

This is now the **error model!**

$$X_i \mid X_i^*, \mathbf{Z}_i \sim \text{Bernoulli}(p_i)$$

$$p_i = \text{expit}(\eta_0 + \eta_1 X_i + \eta_2 \mathbf{Z}_i)$$

Sadness, we consider error now.

$$P(Y_i, X_i, \overset{\downarrow}{X_i^*}, \mathbf{Z}_i) = P_\beta(Y_i \mid X_i, \mathbf{Z}_i)P_\eta(X_i \mid \overset{\downarrow}{X_i^*}, \mathbf{Z}_i)\boxed{P(\overset{\downarrow}{X_i^*}, \mathbf{Z}_i)}$$

Some good news, we still don’t care!

Rats, I spilled hot cocoa all over X_i, and it's **all gone!**

$$P(Y_i,\overset{\downarrow}{\vphantom{X}} X_i^*, \mathbf{Z}_i) = \sum_{x=0}^{\overset{\downarrow}{1}} P_\beta(Y_i \mid X_i = x, Z) P_\eta(X_i = x \mid \mathbf{Z}_i) P(X_i^*, \mathbf{Z}_i)$$

Rats, I spilled hot cocoa all over X_i, and it's **all gone!**

$$P(Y_i, X_i^*, \mathbf{Z}_i) = \sum_{x=0}^{1} \boxed{P_\beta(Y_i \mid X_i = x, Z)} P_\eta(X_i = x \mid \mathbf{Z}_i) P(X_i^*, \mathbf{Z}_i)$$

This is the same* **outcome model!**

*except we fix $X_i = x$ and iterate over both possible X_i values

Rats, I spilled hot cocoa all over X$_i$, and it's **all gone!**

$$P(\overset{\downarrow}{Y_i}, X_i^*, \mathbf{Z}_i) = \sum_{x=0}^{\overset{\downarrow}{1}} P_\beta(Y_i \mid X_i = x, Z) P_\eta(X_i = x \mid \mathbf{Z}_i) P(X_i^*, \mathbf{Z}_i)$$

This is the same* **error model!**

*except we fix X$_i$ = x and iterate over both possible X$_i$ values

Rats, I spilled hot cocoa all over X$_i$, and it's **all gone!**

$$P(Y_i, X_i^*, \mathbf{Z}_i) = \sum_{x=0}^{1} P_\beta(Y_i \mid X_i = x, Z)P_\eta(X_i = x \mid \mathbf{Z}_i)\boxed{P(X_i^*, \mathbf{Z}_i)}$$

you get the picture, we still don't care

Now, to put it all together!

$$\mathscr{L}_N(\beta, \eta) = \prod_{i=1}^{N} \{P(X_i, X_i^*, Y_i, \mathbf{Z}_i)\}^{Q_i} \{P(X_i^*, Y_i, \mathbf{Z}_i)\}^{1-Q_i}$$

Now, to put it all together!

$$\mathscr{L}_N(\beta,\eta) = \prod_{i=1}^{N} \{P(X_i, X_i^*, Y_i, \mathbf{Z}_i)\}^{Q_i}\{P(X_i^*, Y_i, \mathbf{Z}_i)\}^{1-Q_i}$$

Now, we’re calling that joint probability a likelihood, or a function of the parameters given the data!

Now, to put it all together!

$$\mathscr{L}_N(\beta, \eta) = \prod_{i=1}^{N} \{P(X_i, X_i^*, Y_i, \mathbf{Z}_i)\}^{Q_i}\{P(X_i^*, Y_i, \mathbf{Z}_i)\}^{1-Q_i}$$

We assume the neighborhoods are independent, so we can take the product over all neighborhoods!

Now, to put it all together!

neighborhoods with the missing X$_i$ case

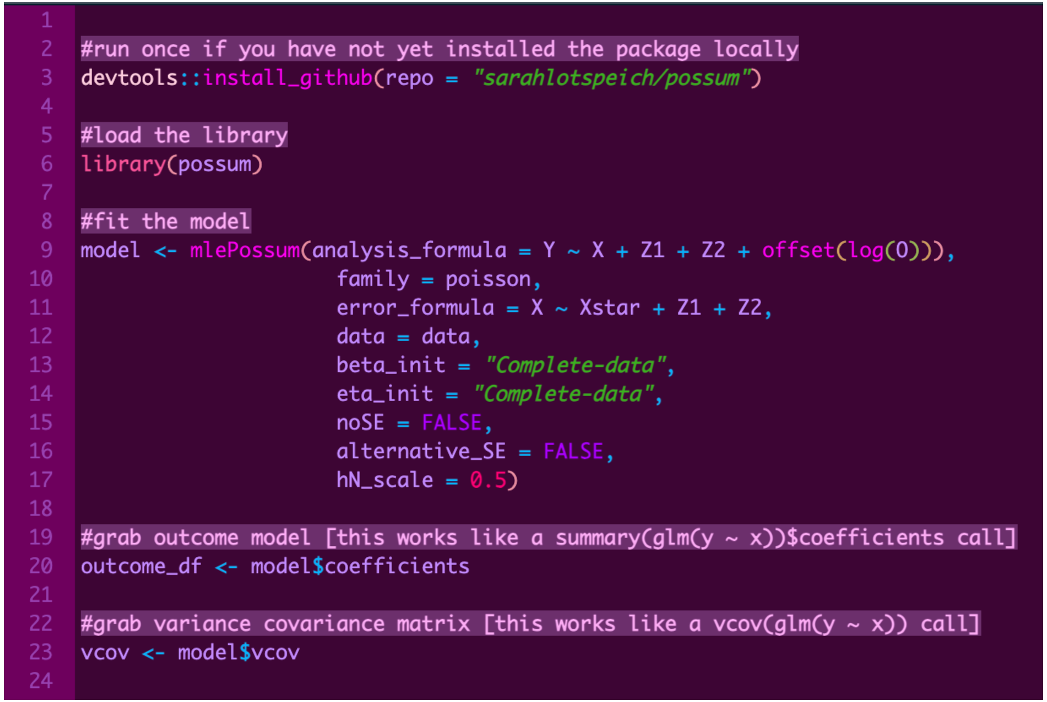
$$\mathscr{L}_N(\beta, \eta) = \prod_{i=1}^{N} \{P(X_i, X_i^*, Y_i, \mathbf{Z}_i)\}^{Q_i}\{P(X_i^*, Y_i, \mathbf{Z}_i)\}^{1-Q_i}$$

neighborhoods with the sadness case

We just outlined more cases than a law student, now what?

- We need to find the vector of parameters (β, η) that **maximizes** this likelihood function, its **MLE**.
- Doing that analytically is a calculus mess, so we turn to the **EM algorithm** (Dempster et. al, 1977), a numerical option.
- The EM algorithm bounces back and forth between taking the **expected values** of the missing variables and then using them to **update** the parameter guesses until we converge to $(\hat{\beta}, \hat{\eta})$.
- We also estimate the **standard error** for the parameter vector.

All the cool kids are using open source software!



```r
#run once if you have not yet installed the package locally
devtools::install_github(repo = "sarahlotspeich/possum")

#load the library
library(possum)

#fit the model
model <- mlePossum(analysis_formula = Y ~ X + Z1 + Z2 + offset(log(0))),
                   family = poisson,
                   error_formula = X ~ Xstar + Z1 + Z2,
                   data = data,
                   beta_init = "Complete-data",
                   eta_init = "Complete-data",
                   noSE = FALSE,
                   alternative_SE = FALSE,
                   hN_scale = 0.5)

#grab outcome model [this works like a summary(glm(y ~ x))$coefficients call]
outcome_df <- model$coefficients

#grab variance covariance matrix [this works like a vcov(glm(y ~ x)) call]
vcov <- model$vcov
```

Ok, this new method is cool! What's it up against?

The MLE:

👍 Uses every possible bit of information from both queried and unqueried neighborhoods

👍 Has nice properties as N goes to infinity

👎 Nobody had done the math yet or implemented the software!

Gold Standard:

$Y = f(X,Z)$

👍 Has the lowest bias and variance we can get

👎 Impossible, since we're missing X_i for N - n neighborhoods!

Ok, this new method is cool! What's it up against?

The MLE:

👍 Uses every possible bit of information from both queried and unqueried neighborhoods

👍 Has nice properties as N goes to infinity

👎 Nobody had done the math yet or implemented the software!

Naive Analysis:

Y = f(X*,Z)

👍 Easy to fit and at least uses all of the information available from the error prone exposure

👎 Is biased, meaning we give up accuracy (Shaw et. al, 2020)

Ok, this new method is cool! What's it up against?

The MLE:

👍 Uses every possible bit of information from both queried and unqueried neighborhoods

👍 Has nice properties as N goes to infinity

👎 Nobody had done the math yet or implemented the software!

Complete Case: $Y = f(X,Z)$

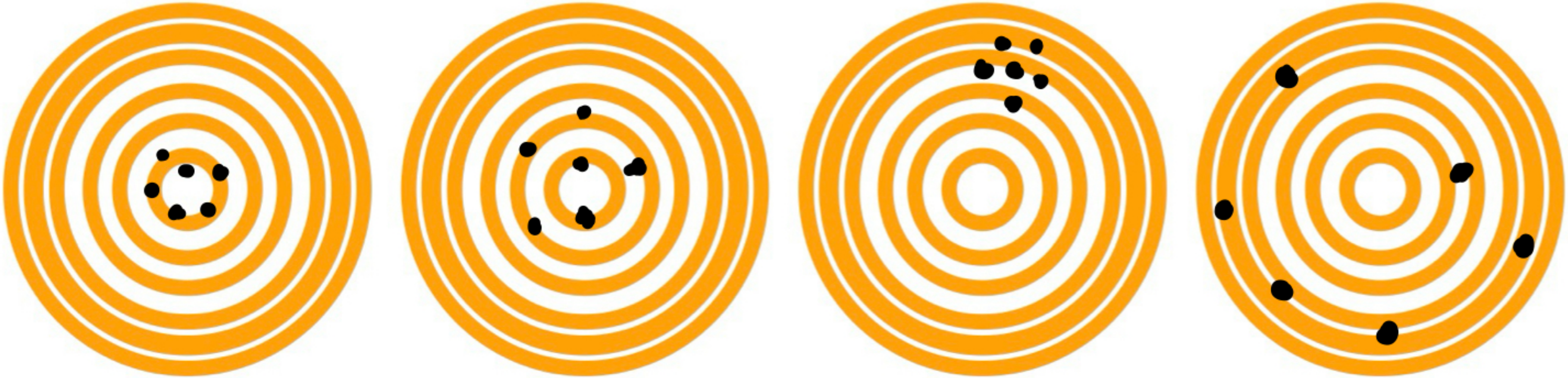
👍 Unbiased, since it uses X

👎 We lose efficiency, since we throw out any neighborhood that's missing X

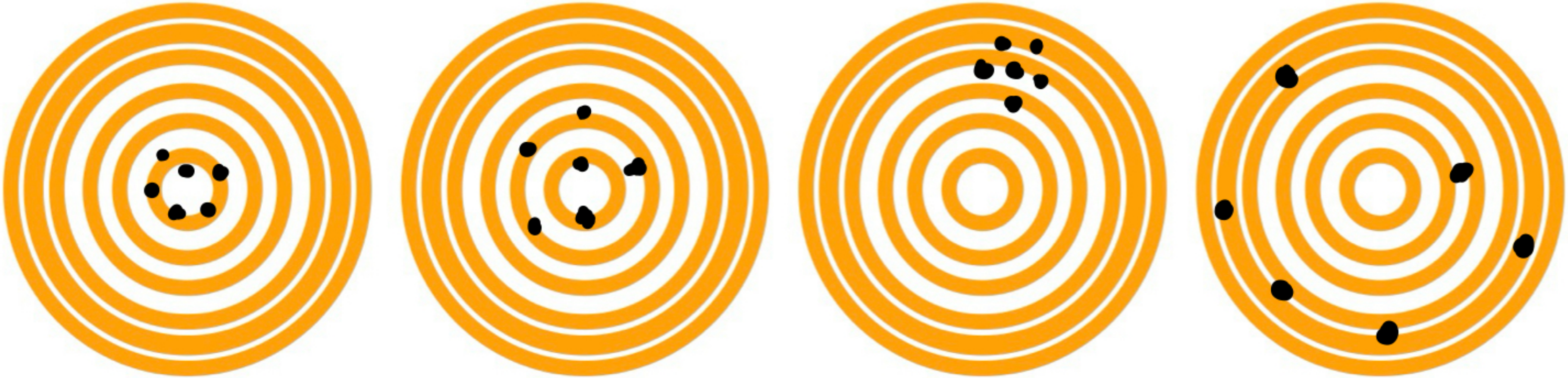
3

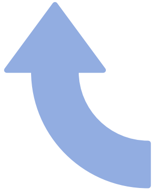
Battle Royale (Methods Edition)

How do we pick a winner? Treat them like darts!



How do we pick a winner? Treat them like darts!





This is the method you want in your toolbox!

Let's formalize the goals here.

- An **unbiased** method has an expected value of the true parameter value you're trying to catch!
- As the **variance** of your method drops, it gets more **efficient**!
- In the Barbie dream world, we're hoping for a method that has **low bias** and **low variance**.
- In the real world, sometimes improving one costs you the other. It's a **balance** between the two!

How do we find that balance?

- In (bio)statistics, we use **simulation** to investigate how our methods perform when we actually know the truth!
- We **randomly generate** data from a specific setup and **repeat many times** to get a sense of the method's behavior in the long run.
- As the simulation overlords, we get to choose and **fix our ground truth** and **vary one setting** to see how it changes our results!

Here's the simulation plan.

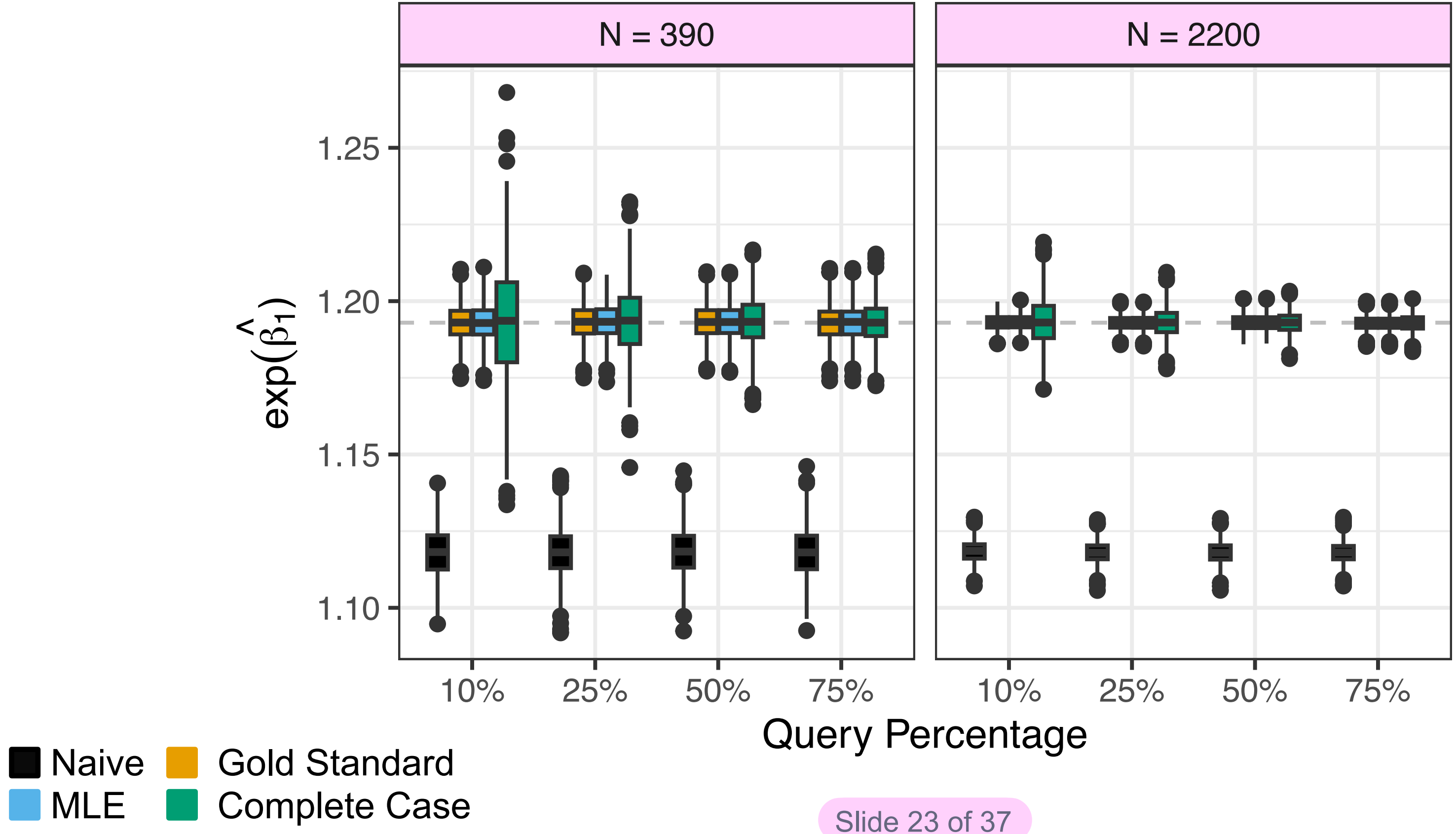
- We're looking to estimate **exp(β_1)**, the **prevalence ratio**.
- Study 1: *What happens if we query fewer observations?* We'll fix everything but the **query proportion** and then repeat the study with more neighborhoods.
- Study 2: *What happens if our errors get ugly?* We'll fix everything but the **positive predictive value** and then repeat the study with more neighborhoods.

Here's the simulation plan.

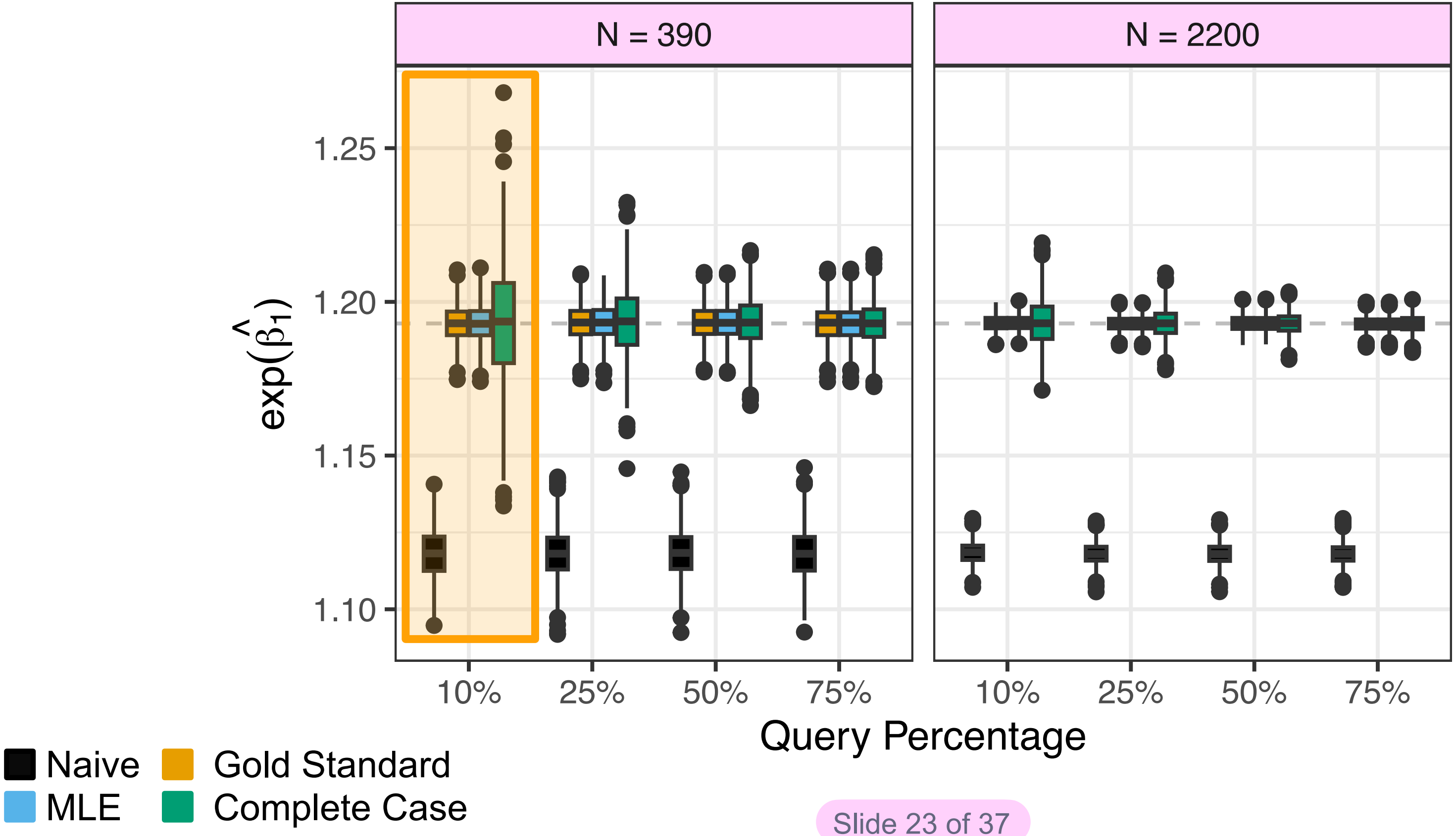
- We're looking to estimate **exp(β_1)**, the **prevalence ratio**.
- Study 1: *What happens if we query fewer observations?* We'll fix everything but the **query proportion** and then repeat the study with more neighborhoods.
- Study 2: *What happens if our errors get ugly?* We'll fix everything but the **positive predictive value** and then repeat the study with more neighborhoods.

(this is the probability that a neighborhood with food access according to X* really has it according to X, remember this idea!)

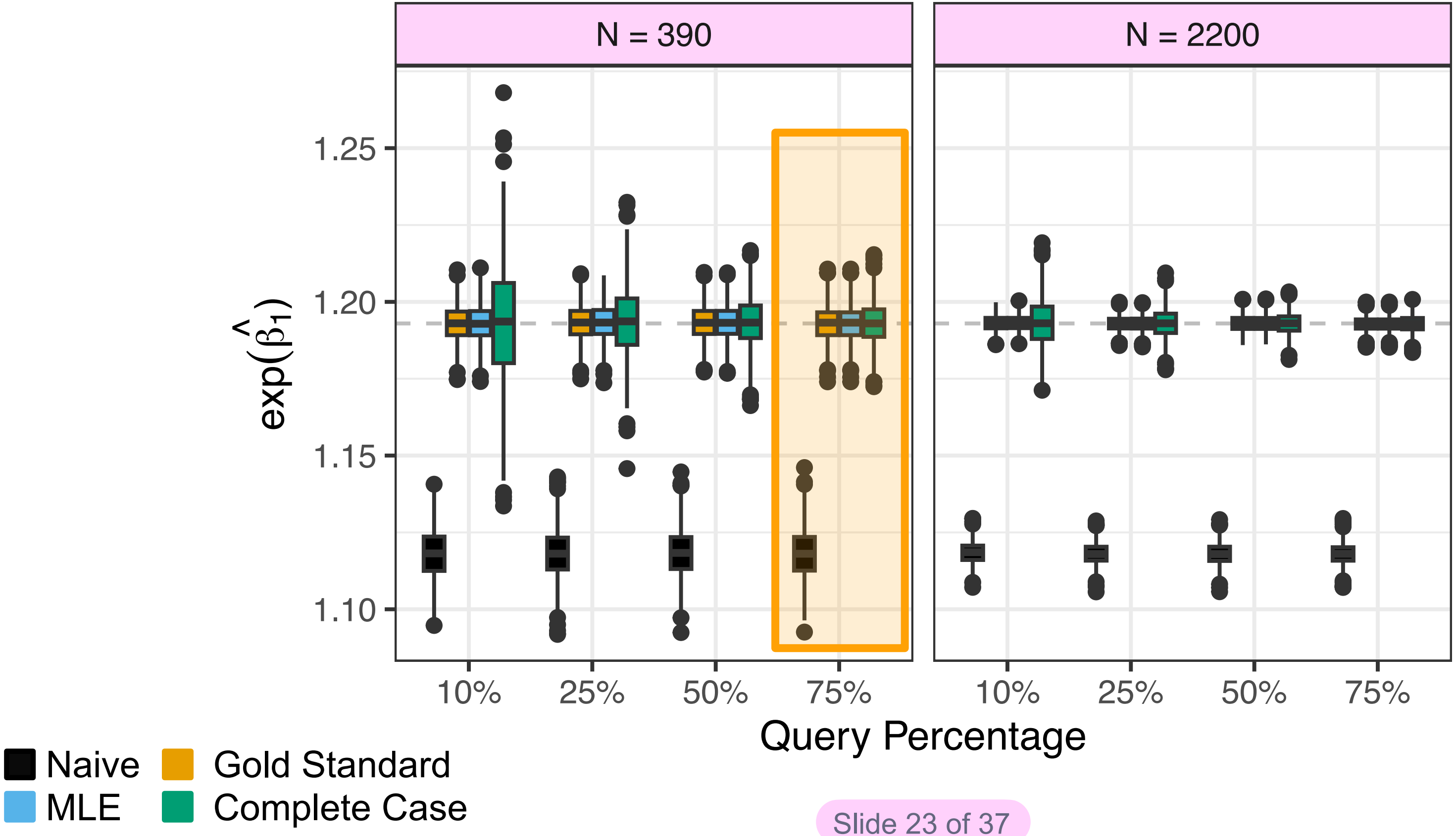
What happens as we vary the query size?



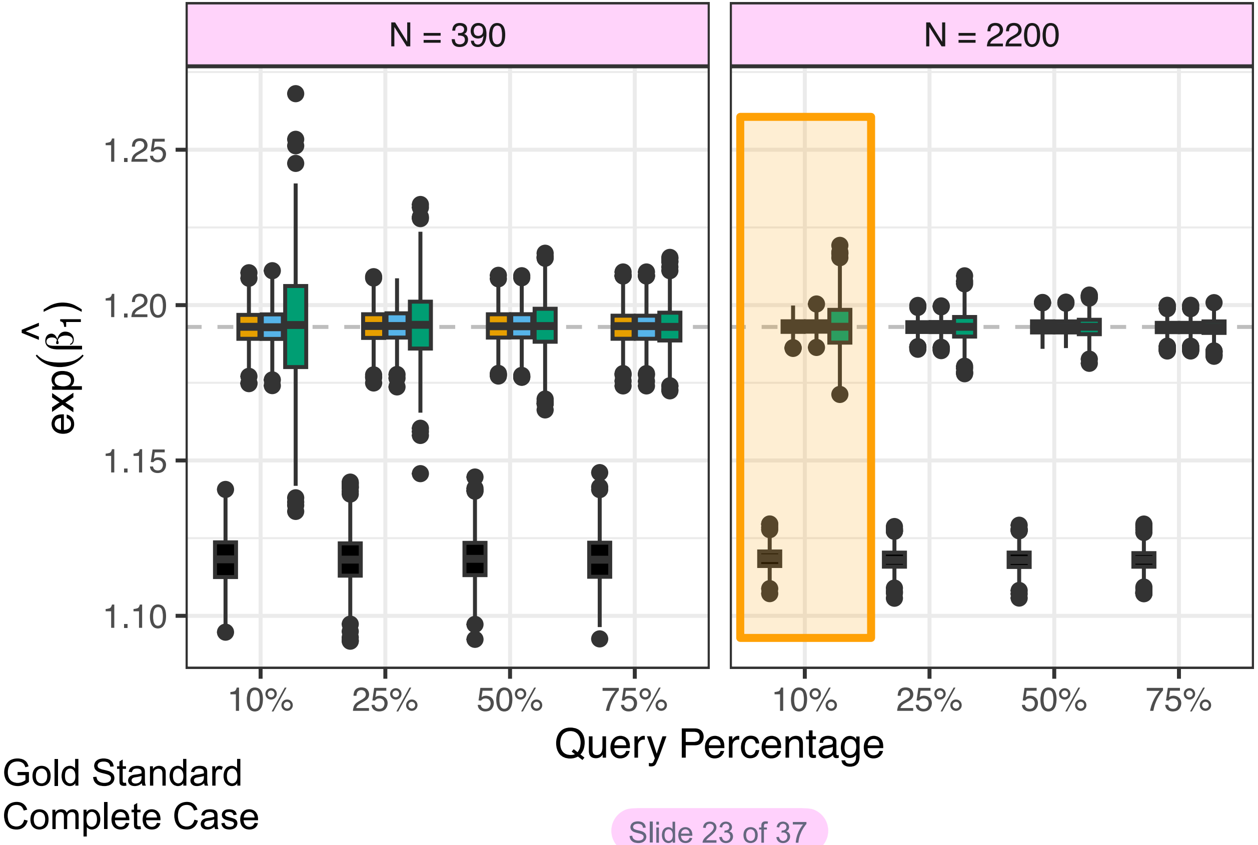
What happens as we vary the query size?



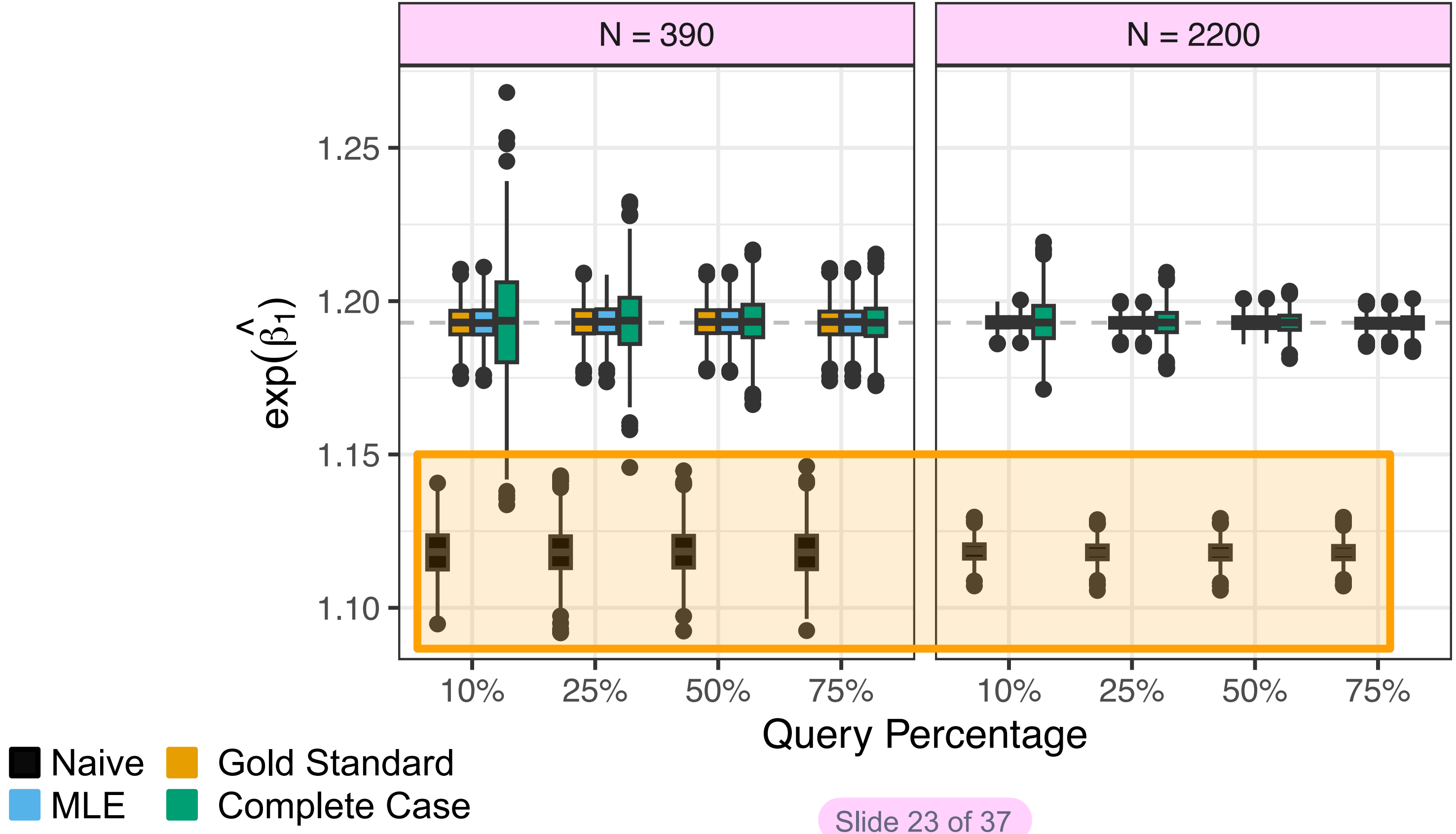
What happens as we vary the query size?



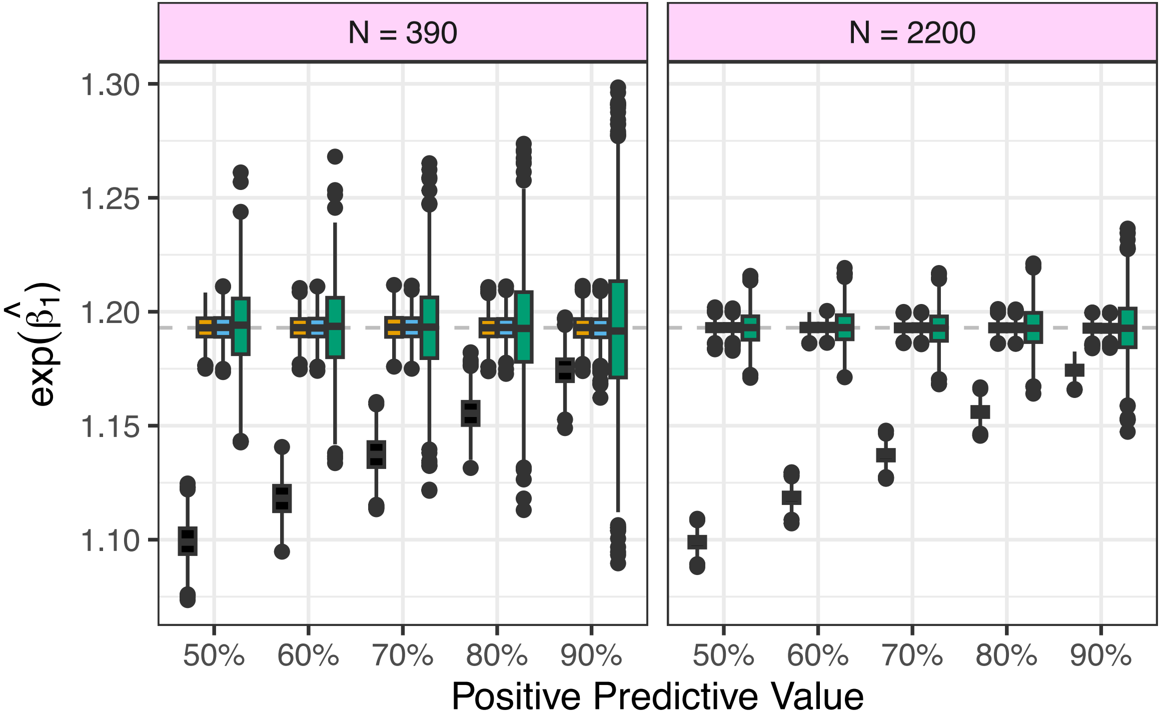
What happens as we vary the query size?



What happens as we vary the query size?

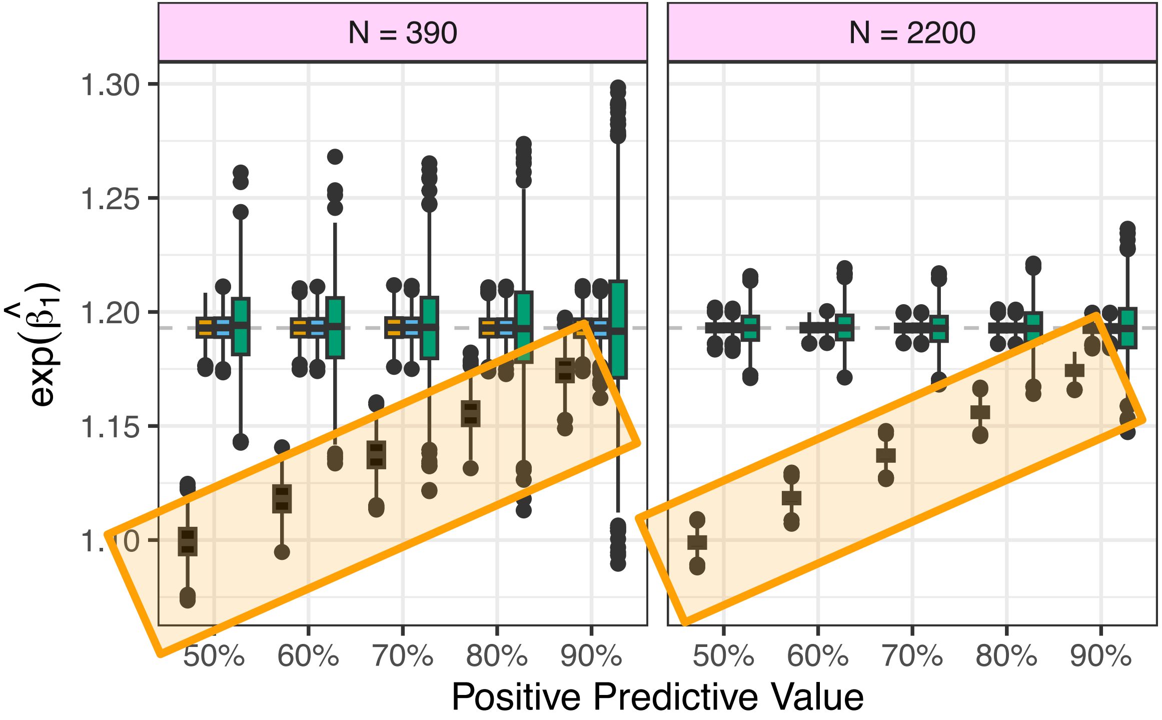


What happens as the errors get worse?



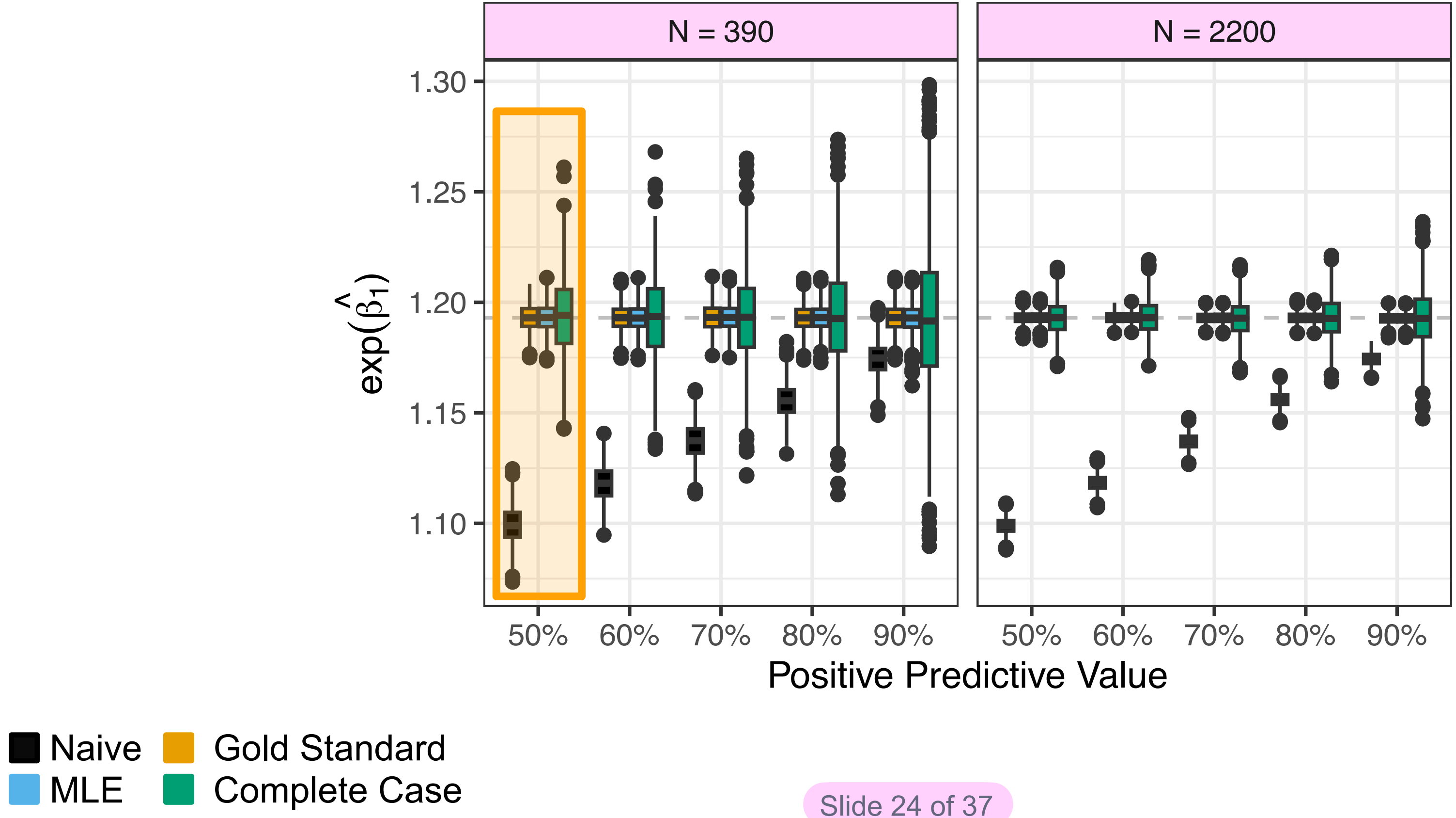
⬛ Naive 🟧 Gold Standard
🟦 MLE 🟩 Complete Case

What happens as the errors get worse?

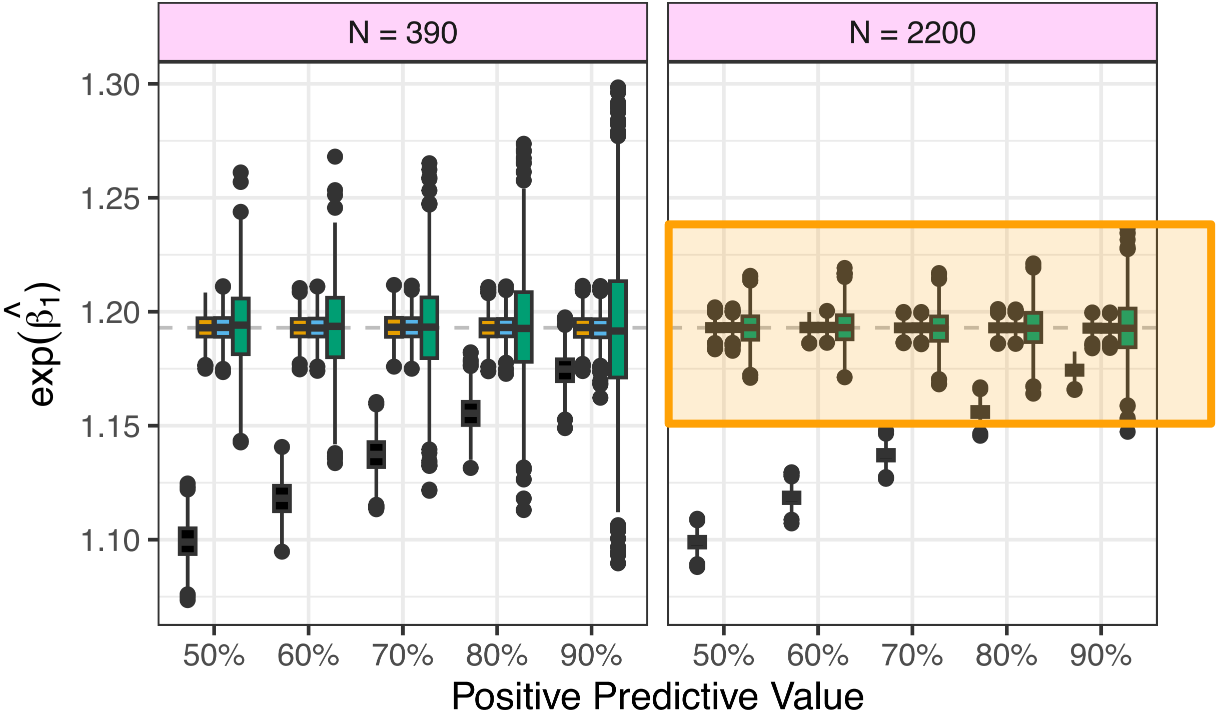


■ Naive ■ Gold Standard
■ MLE ■ Complete Case

What happens as the errors get worse?



What happens as the errors get worse?



⬛ Naive 🟧 Gold Standard
🟦 MLE 🟩 Complete Case

Whew, that was a lot! What's the TL;DR?

- Obviously, the gold standard is what you'd want if you lived in the Barbie dream world, but the MLE is **pretty close**.
- The MLE **avoids the heavy bias** of the naive analysis and **improves on the efficiency** of the complete case analysis.
- Even as we increase the sample size enough to take care of some of the **issues in the competitors**, it doesn't fix all of them!

4

Case Study

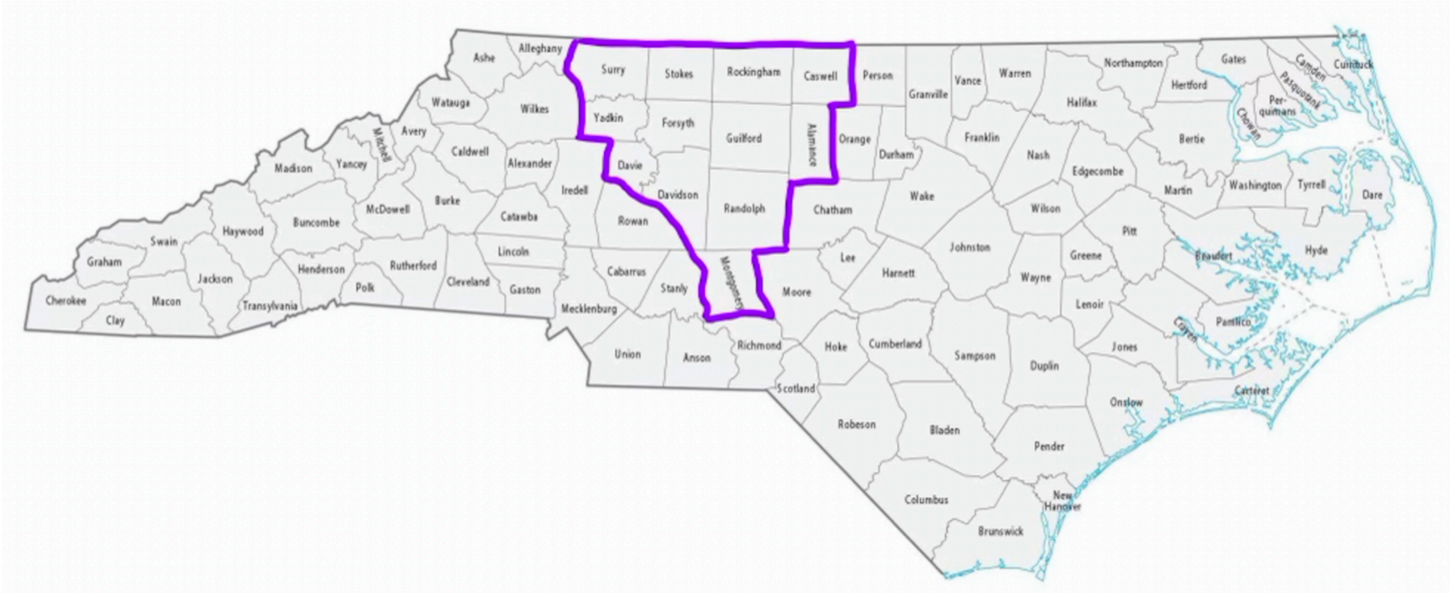
Why did we go through all that?

- Our original question was:

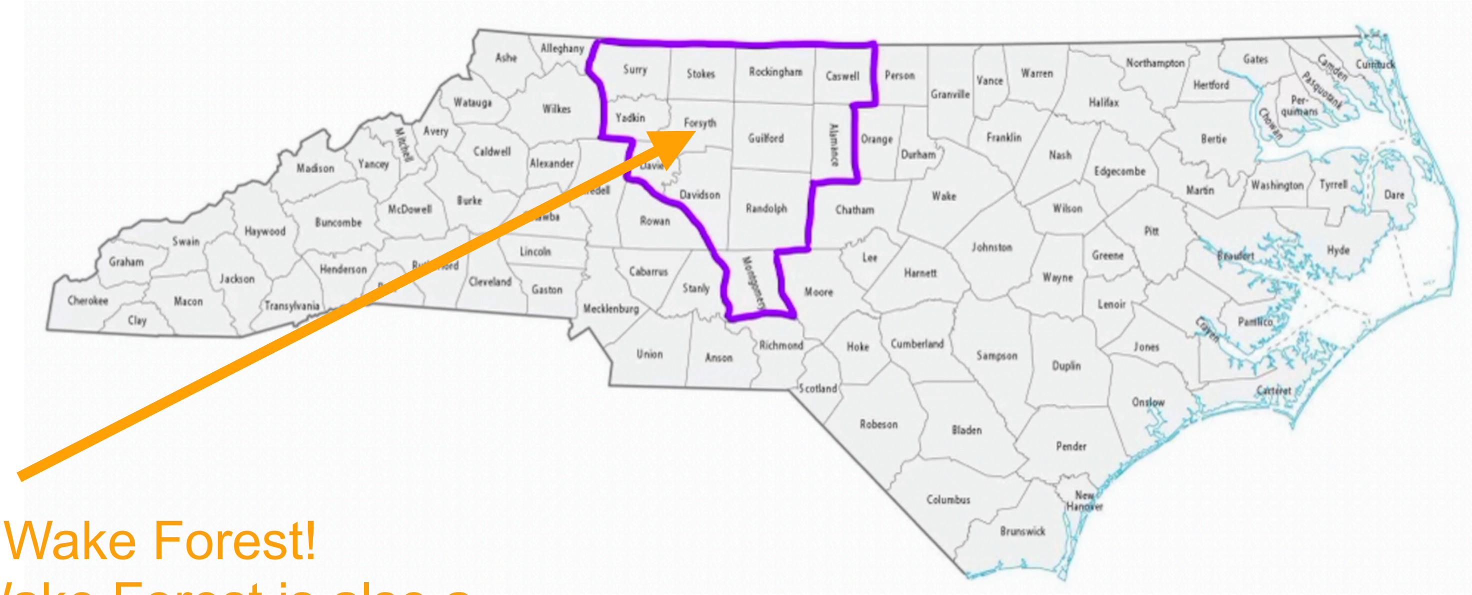
  “Can we sidestep the **misclassification and missingness issues** in our data and still accurately estimate an association between **access to healthy food** and **diabetes prevalence**?”

- We can apply the MLE method to data from the **Piedmont Triad** in **North Carolina** to help answer this question.

What's the Piedmont Triad?



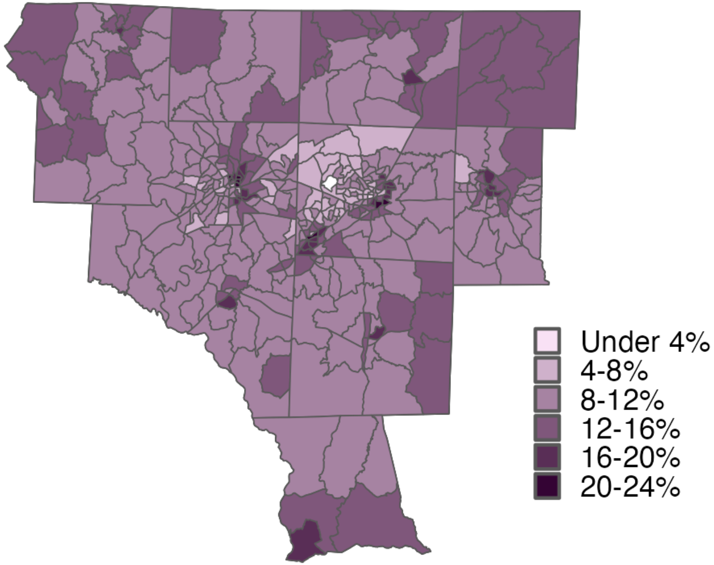
What's the Piedmont Triad?



There's Wake Forest!
(note: Wake Forest is also a town but isn't in the Triad!)

What does diabetes prevalence look like in the Triad?

- Prevalence in North Carolina was **12.4%** in 2021 (American Diabetes Association).
- Most tracts have between **8-12%** prevalence but this **varies** across the Triad.
- Tracts with **lower prevalences** tended to be smaller and **urban**.



(by the generosity of Lotspeich et al. 2025)

What kind of data do we ^ have?

- N = 387 neighborhood* **population centers** (and the number of people that live there) from the 2010 census
- 701 **healthy food retailers** from the 2022 USDA SNAP retailer locator release
- Diabetes **prevalences** for each neighborhood from the 2022 CDC PLACES release
- Metro indicators for each neighborhood derived from the 2010 USDA RUCA code release

*technically a census tract, about which (bio)statisticians have thoughts and feelings

Time to pick things out of the outcome model!

$$\log\{\mathrm{E}_{\beta}(Y_i \mid X_i, M_i)\} = \beta_0 + \beta_1 X_i + \beta_2 M_i + \beta_3 X_i \times M_i + \log(O_i)$$

Time to pick things out of the outcome model!

$$\log\{\mathrm{E}_\beta(Y_i \mid X_i, M_i)\} = \beta_0 + \beta_1 X_i + \beta_2 M_i + \beta_3 X_i \times M_i + \log(O_i)$$

the expected number of diabetes cases (Y_i) in neighborhood i given its food access at that radius (X_i) and whether it's metro (M_i) or not

Time to pick things out of the outcome model!

$$\log\{\mathrm{E}_{\beta}(Y_i \mid X_i, M_i)\} = \beta_0 + \boxed{\beta_1} X_i + \beta_2 M_i + \beta_3 X_i \times M_i + \log(O_i)$$

the (log of the) ratio of diabetes prevalence in a non-metro neighborhood with food access at that radius to one without

Time to pick things out of the outcome model!

add these!

$$\log\{\mathrm{E}_{\beta}(Y_i \mid X_i, M_i)\} = \beta_0 + \beta_1 X_i + \beta_2 M_i + \beta_3 X_i \times M_i + \log(O_i)$$

the extra effect tacked onto the (log of the) prevalence ratio for a metro neighborhood

Time to pick things out of the outcome model!

$$\log\{\mathrm{E}_{\beta}(Y_i \mid X_i, M_i)\} = \beta_0 + \beta_1 X_i + \beta_2 M_i + \beta_3 X_i \times M_i + \log(O_i)$$

the population offset that turns prevalences into case counts

What do our errors look like at each radius?

- We’re lucky enough to have the **route access measure** for **all** of the 387 neighborhoods in our study to check our work!

	🛣️🍎	🛣️😒
📏🍎	122	70
📏😒	0	195

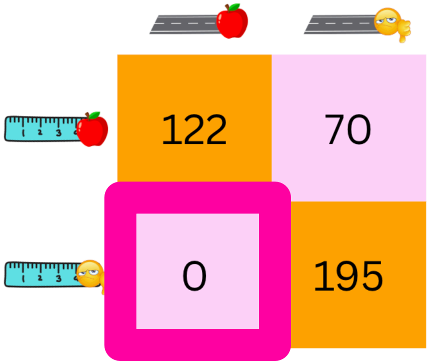
1 mile

	🛣️🍎	🛣️😒
📏🍎	333	22
📏😒	0	32

5 miles

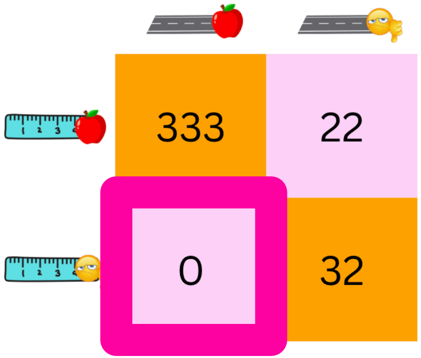
What do our errors look like at each radius?

- We're lucky enough to have the **route access measure** for **every** neighborhood in our study to check our work!



	🛣️🍎	🛣️😒
📏🍎	122	70
📏😒	0	195

1 mile

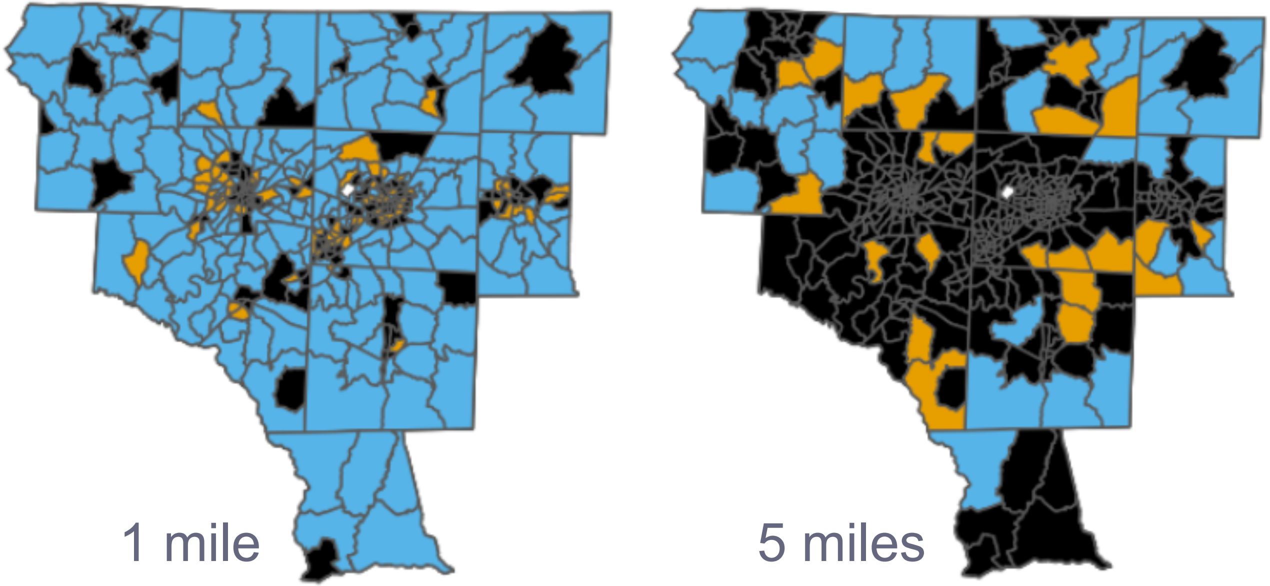


	🛣️🍎	🛣️😒
📏🍎	333	22
📏😒	0	32

5 miles

Did you notice these? This category always has zero neighborhoods by construction. This messes with our error model, so we modify our query scheme to fix it!

Can we map access out by neighborhood?

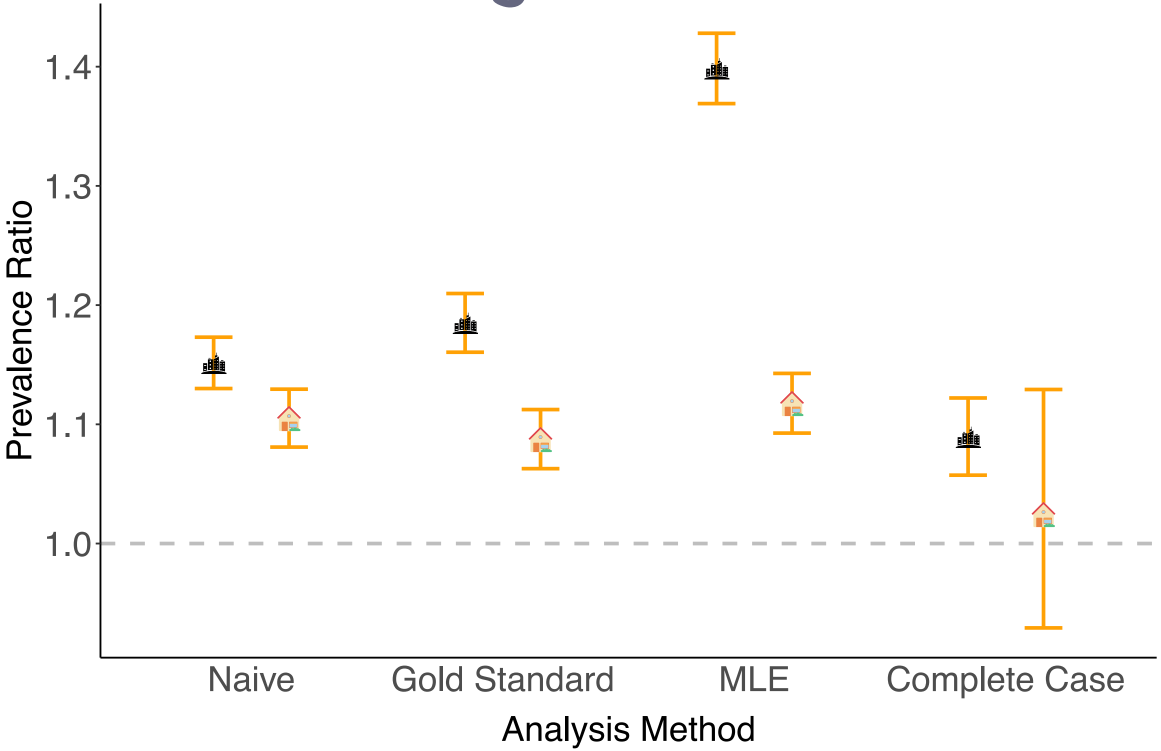


1 mile

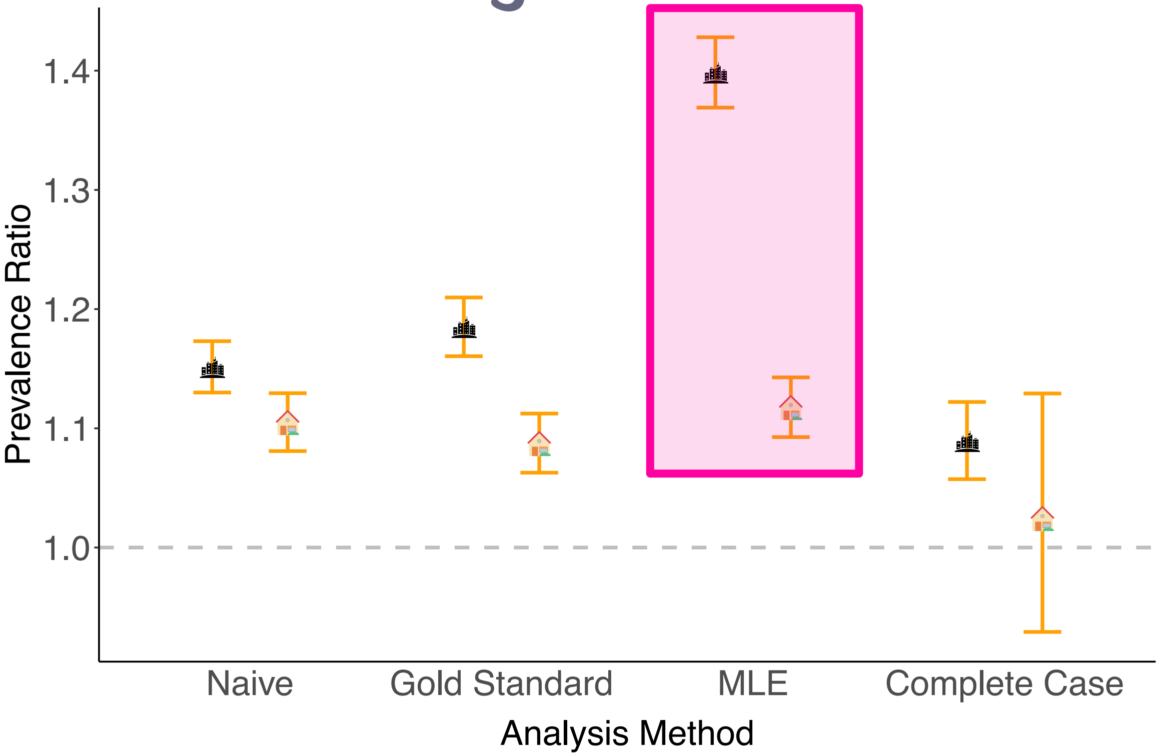
5 miles

	📏	🛣️
⬛	🍎	🍎
🟧	🍎	👎
🟦	👎	👎

What do the models say at a one mile radius?

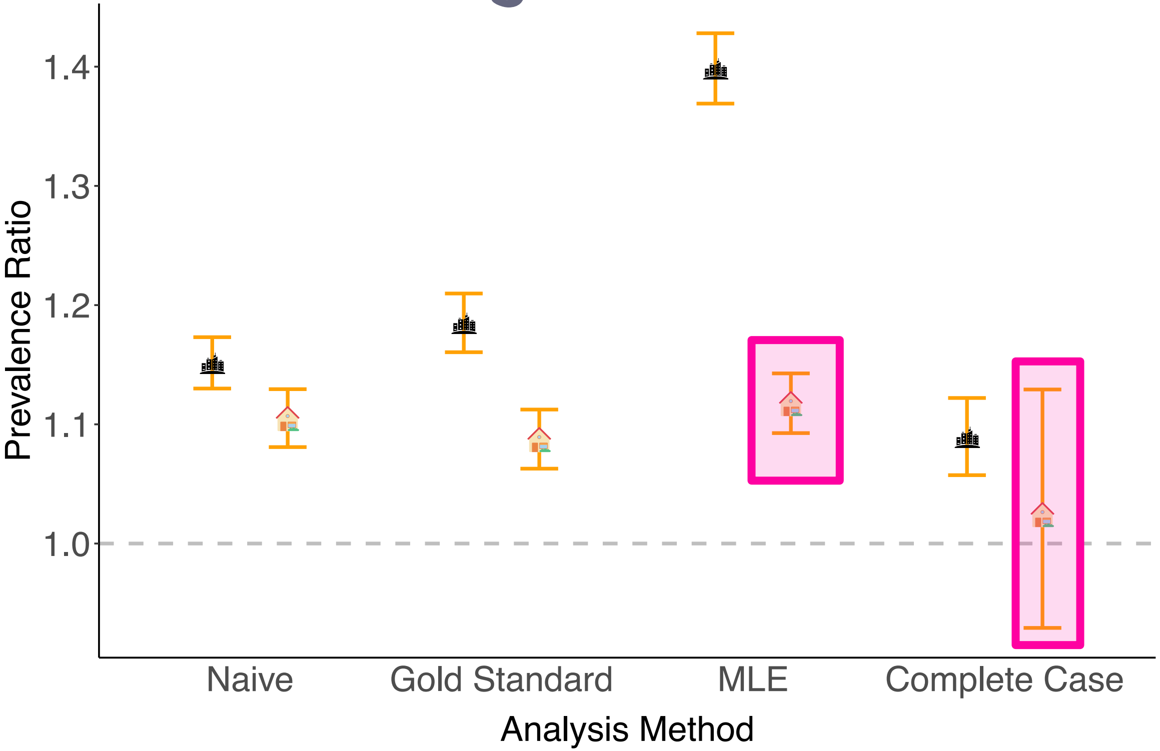


What do the models say at a one mile radius?



🏙 Metro Tract
🏠 Non Metro Tract

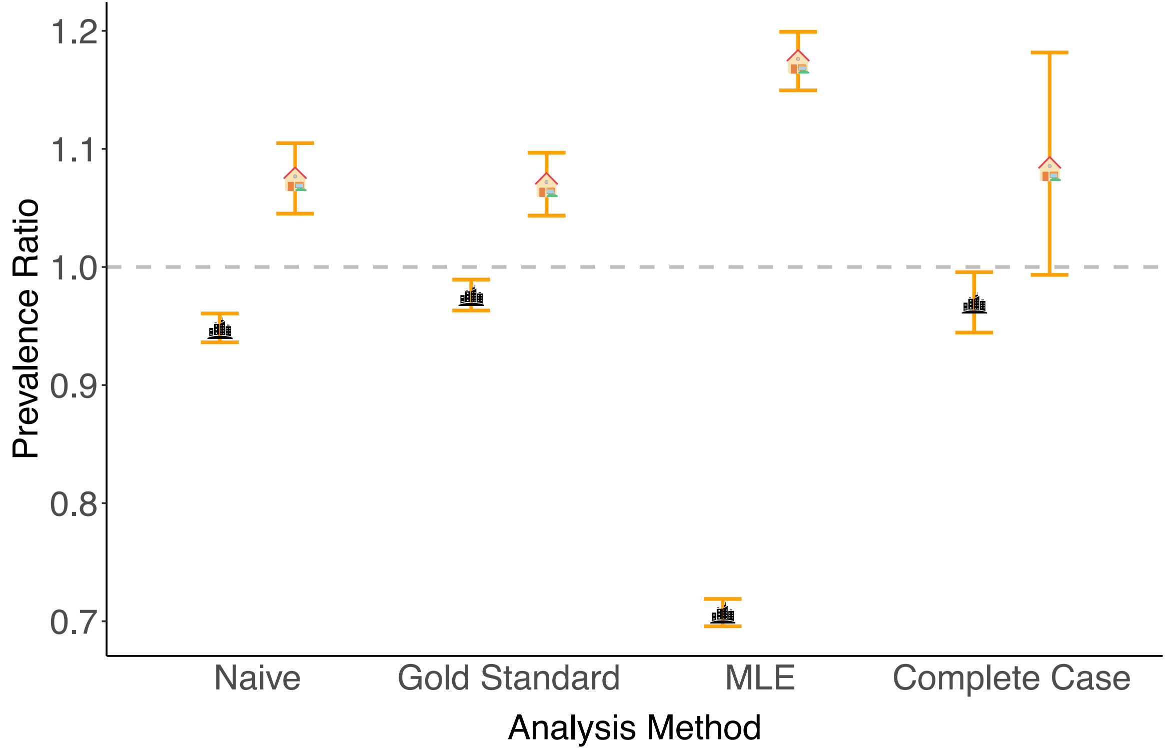
What do the models say at a one mile radius?



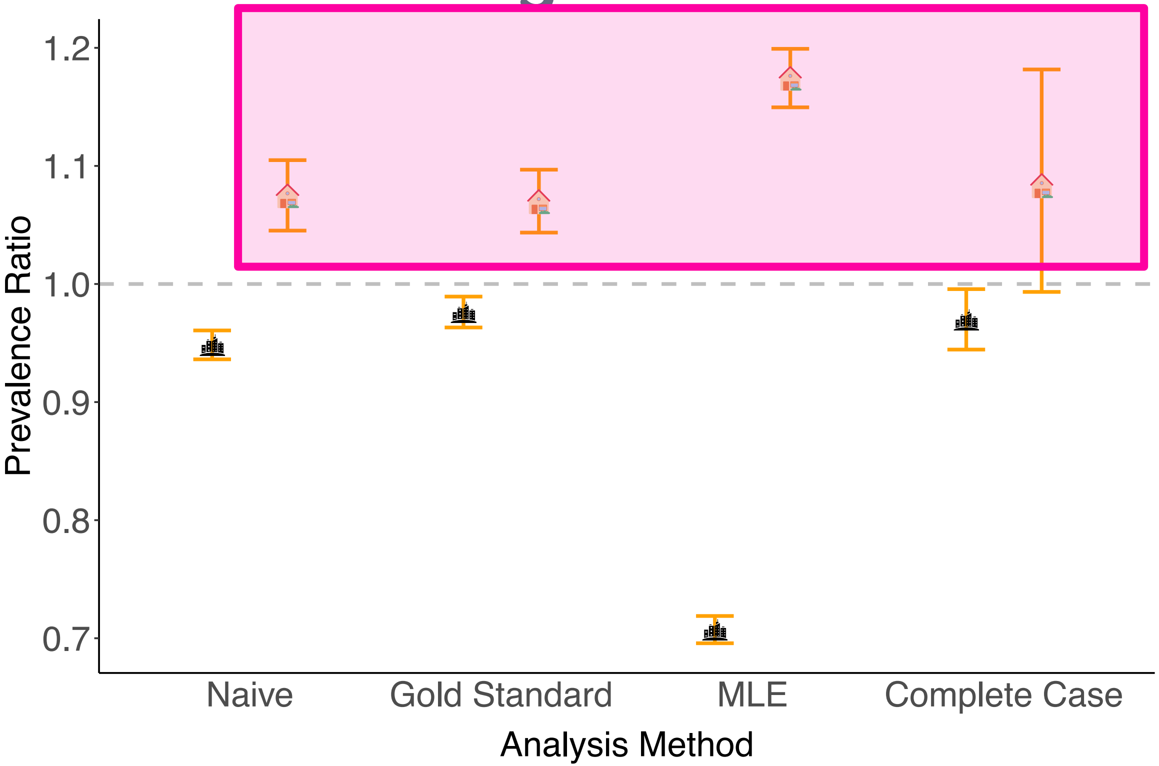
🏙️ Metro Tract

🏠 Non Metro Tract

What do the models say at a five mile radius?

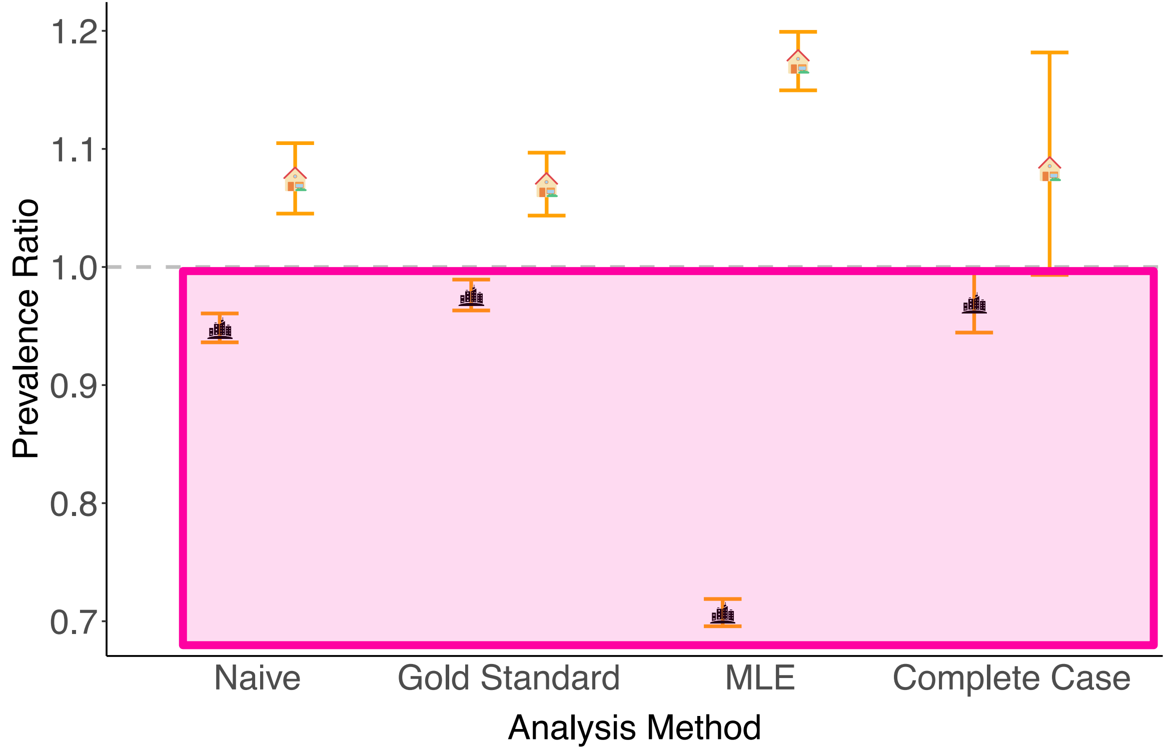


What do the models say at a five mile radius?



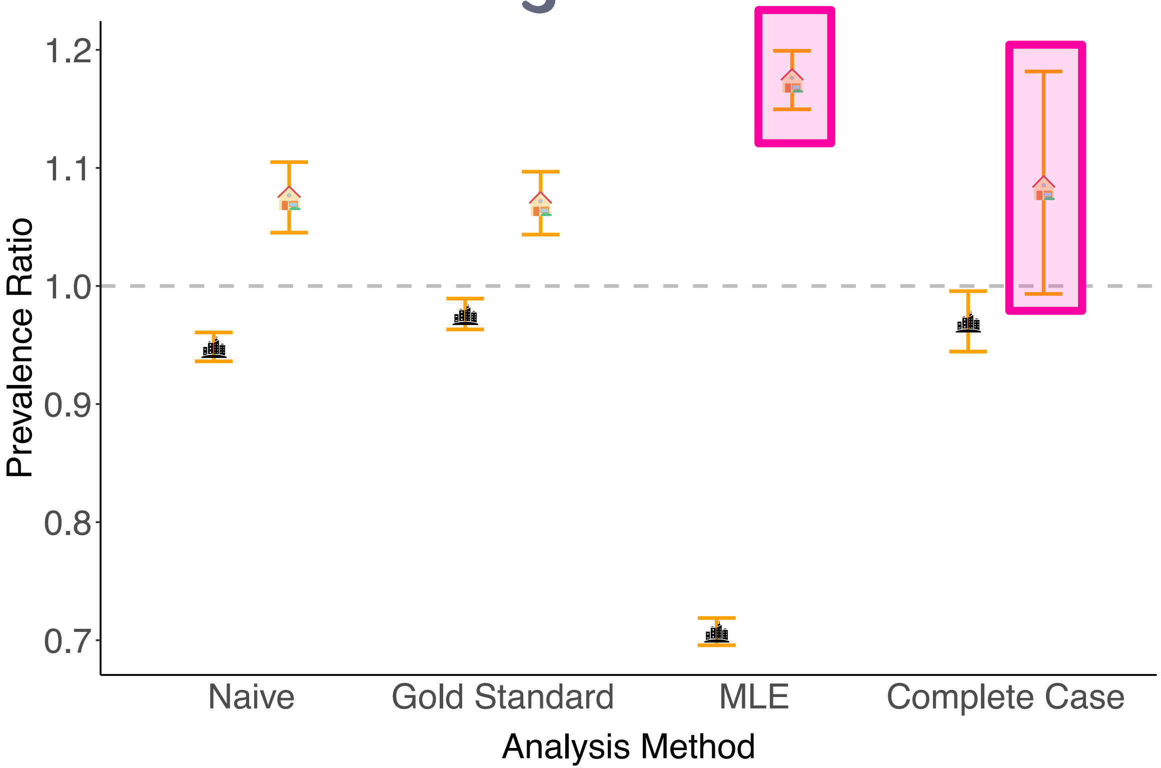
🏙️ Metro Tract
🏠 Non Metro Tract

What do the models say at a five mile radius?



🏙 Metro Tract
🏠 Non Metro Tract

What do the models say at a five mile radius?



🏙 Metro Tract
🏠 Non Metro Tract

For those who zoned out a bit...

- As the **radius** changed, so did the **patterns**.
- Overall, tracts with **food access** within a mile counterintuitively saw **higher** diabetes prevalences.
- Overall, tracts with **food access** within five miles had diabetes patterns dictated by **metro status**.
- The MLE model usually reported **stronger** effects more **efficiently** than the complete case.

5

Odds & Ends

Who else can POSSUM?

- POSSUM specifically handles a **count outcome** and a **problematic binary variable**, but these concepts can generalize!
- Broken lab equipment, **imprecise** measurement procedures, and **large health databases** like the **EHR** are some other places you might find misbehaving variables that you can handle similarly.

What did we do today in fancy statistics terms?

- Derived a novel maximum likelihood estimator for Poisson regression with a misclassified binary covariate and a two phase validation design
- Implemented the method and its standard error estimator using the expectation-maximization algorithm in R
- Showcased its asymptotic and small-sample advantages via simulation
- Demonstrated a use case by estimating diabetes prevalence as a function of food access and urbanicity in North Carolina
- Suffered through some extremely unserious slide titles and commentary

But maybe the real seminar takeaways were the math jokes we made along the way!!!

And now for the gratitude!

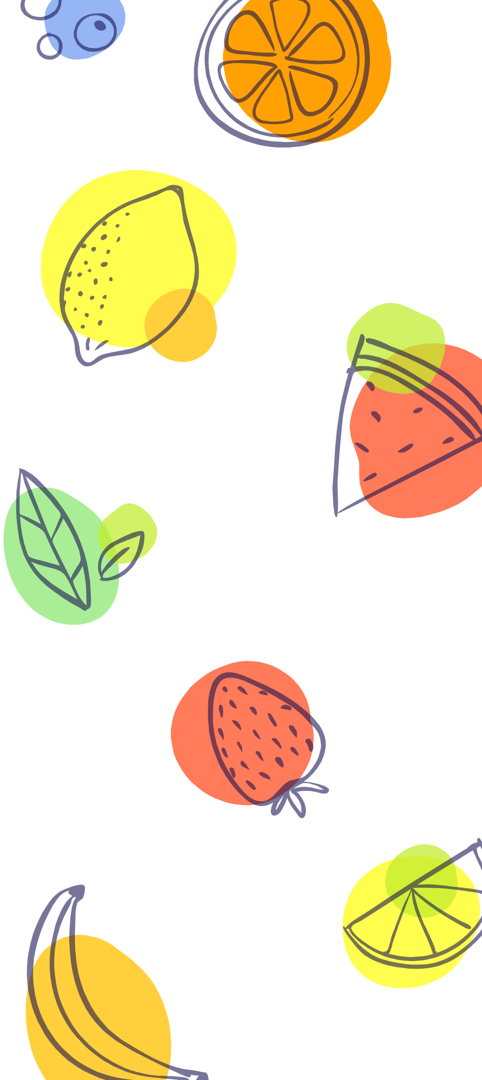
Want more of this? Consider applying to Wake Forest for an M.S. in Statistics!

🔗: www.stats.wfu.edu

- 2 year program with a small, personalized cohort.
- Research opportunities in faculty labs.
- All accepted students receive substantial financial aid!
- Most students receive assistantships with 100% tuition scholarship, and students work 15-18 hours per week to earn an additional 10-month stipend.
- All other students receive a partial scholarship, which covers >70% tuition.

References

- American Diabetes Association. About diabetes, 2021. URL https://diabetes.org/about-diabetes
- D. Kahle and H. Wickam. ggmap: Spatial Visualization with ggplot2. The R Journal, 5(1), 144-161. URL http://journal.r-project.org/archive/2013-1/kahle-wickham.pdf
- E. Gucciardi, M. Vahabi, N. Norris, J.P. Del Monte, and C. Farnum. The intersection between food insecurity and diabetes: a review: Current nutrition reports, 3:324-332, 2014
- P. A. Shaw, P. Gustafson, R. J. Carroll, V. Deffner, K. W. Dodd, R. H. Keogh, V. Kipnis, J. A. Tooze, M. P. Wallace, H. Küchenhoff, et al. STRATOS guidance document on measurement error and misclassification of variables in observational epidemiology: part 2—more complex methods of adjustment and advanced topics. Statistics in medicine, 39(16):2232–2263, 2020



References

- Walker K, Herman M (2024). _tidycensus: Load US Census Boundary and Attribute Data as 'tidyverse' and 'sf'-Ready Data Frames_. R package version 1.6, URL https://CRAN.R-project.org/package=tidycensus
- World Health Organization. Healthy diet, 2019. URL https://iris.who.int/handle/10665/325828
- Dempster, A. P., N. M., Laird, D. B., Rubin. "Maximum Likelihood from Incomplete Data Via the EM Algorithm". Journal of the Royal Statistical Society: Series B (Methodological) 39. 1(1977): 1-22.
- S.C. Lotspeich, A.E. Mullan, L.D. McGowan, S.A. Hepler. "Combining straight-line and map-based distances to investigate the connection between proximity to healthy foods and disease." (2025). URL https://arxiv.org/abs/2405.16385