Linking Potentially Misclassified Healthy Food Access to Diabetes Prevalence

Ashley E. Mullan

WSDS 2025

Scan or search to follow along with me!

https://bit.ly/ash-talks

Coming Soon to Theaters

1. **Invited Talk**: Linking Potentially Misclassified Healthy Food Access to Diabetes Prevalence
   Women in Statistics and Data Science Conference - November 2025
   Slides

Today, we'll cover:

1. Motivation
2. Methods
3. Simulations
4. Case Study
5. Wrap Up

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.

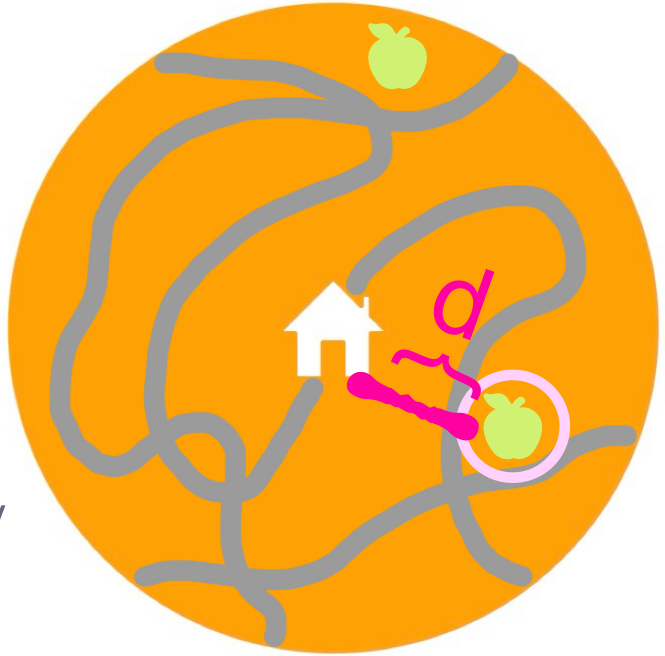
Measuring Food Access 🍎

- 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**!

Draw a straight line?

Measuring Food Access 🍎

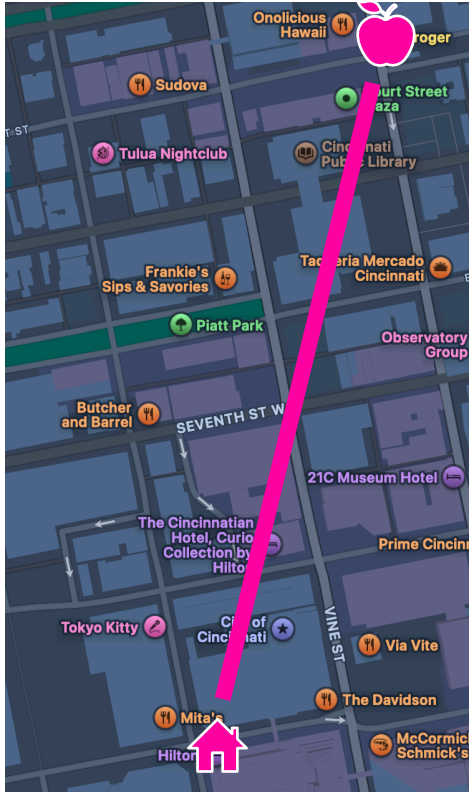
- All of these methods require some notion of **distance**!

Draw a straight line?

Follow the road?

Distance Computations

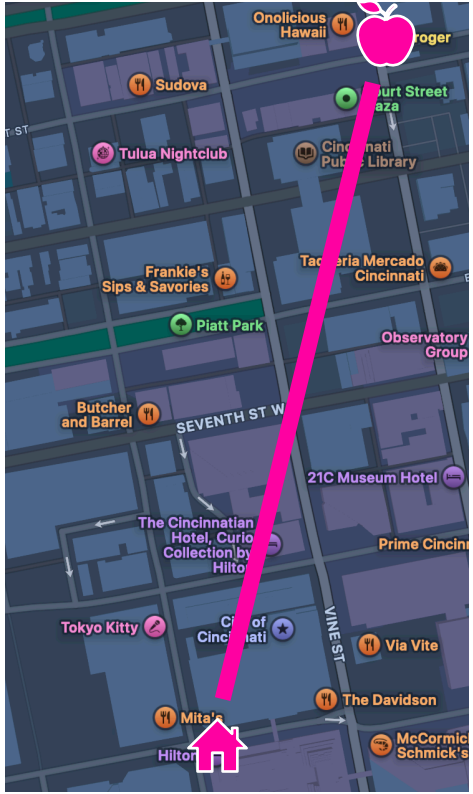
- The **Haversine distance** is a trigonometric function of latitude and longitude.
- It ignores physical obstacles, so it **underestimates** the true distance between two points and is considered **error-prone**.



0.4 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.



0.7 miles

We're left with a data challenge!

- 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!

2

Methods

For neighborhood i, we observe:

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

For neighborhood i, we observe:

outcome, number of diabetes cases in the neighborhood

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

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

For neighborhood i, we observe:

outcome, number of diabetes cases in the neighborhood

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

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

exposure, representing the food access indicator for that neighborhood

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

For neighborhood i, we observe:

outcome, number of diabetes cases in the neighborhood

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

correct but only available if query indicator $Q_i = 1$

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

exposure, representing the food access indicator for that neighborhood

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

error-prone but always available

For neighborhood i, we observe:

outcome, number of diabetes cases in the neighborhood

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

correct but only available if query indicator $Q_i = 1$

covariate vector (assumed to be error-free)

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

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

exposure, representing the food access indicator for that neighborhood

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

error-prone but always available

How do we query?

- Having X for **some** of the **N** neighborhoods is better than having it for none!
- **Two phase design** maximizes the utility of the available information.
- 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**.

We break down each queried observation.

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

We break down each queried observation.

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

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

$\exp(\beta_1)$ gives us the target association

outcome model

We break down each queried observation.

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

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

$\exp(\beta_1)$ gives us the target association

outcome model

error model

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

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

We break down each unqueried observation.

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

We break down each unqueried observation.

marginalize over X

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

We break down each unqueried observation.

marginalize over X

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

outcome model (same Poisson)

We break down each unqueried observation.

marginalize over X

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

outcome model (same Poisson)

error model (same Bernoulli)

We stack the observations to build the likelihood.

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

We stack the observations to build the likelihood.

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

queried neighborhoods

We stack the observations to build the likelihood.

unqueried neighborhoods

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

queried neighborhoods

We now compute an MLE.

- We **maximize** the likelihood function via an **EM algorithm** (Dempster et. al, 1977).
- We estimate **standard errors** via numerical differentiation.

3

Simulations

We follow this estimation procedure.

Generate:
1. Z
2. X* | Z
3. X | X*, Z
4. Y | X, Z

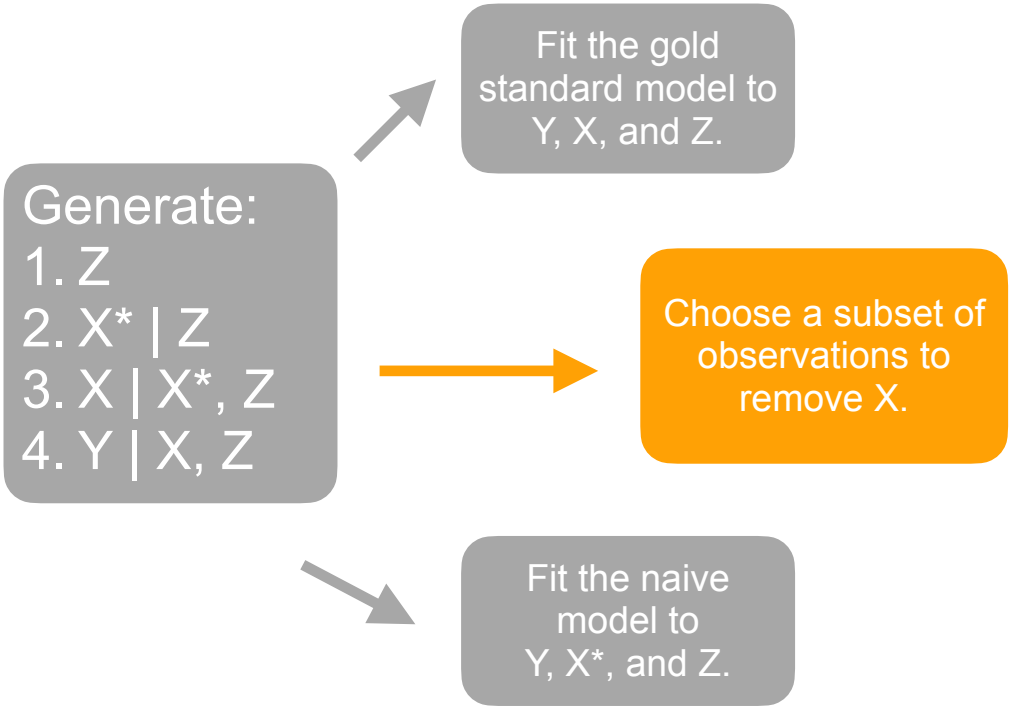
We follow this estimation procedure.

Generate:
1. Z
2. X* | Z
3. X | X*, Z
4. Y | X, Z

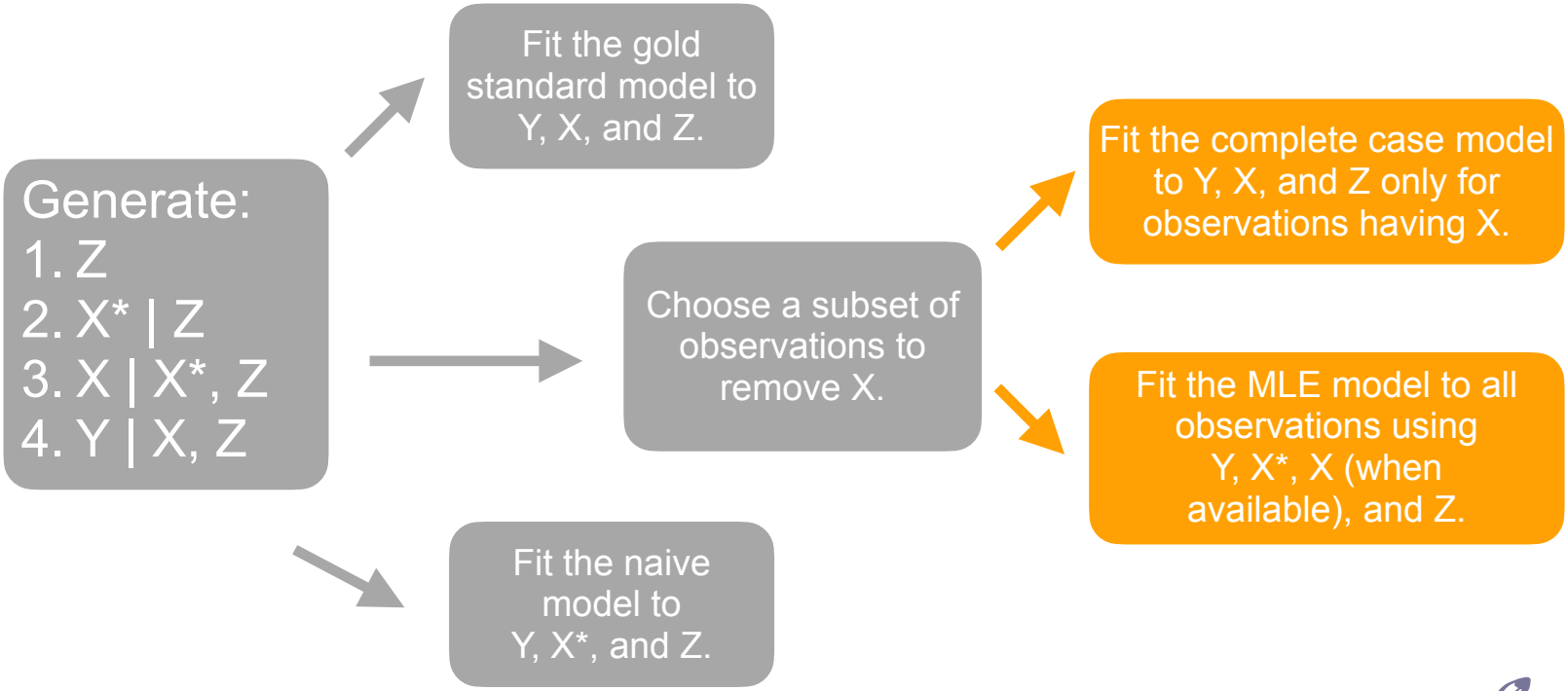
Fit the gold standard model to Y, X, and Z.

Fit the naive model to Y, X*, and Z.

We follow this estimation procedure.



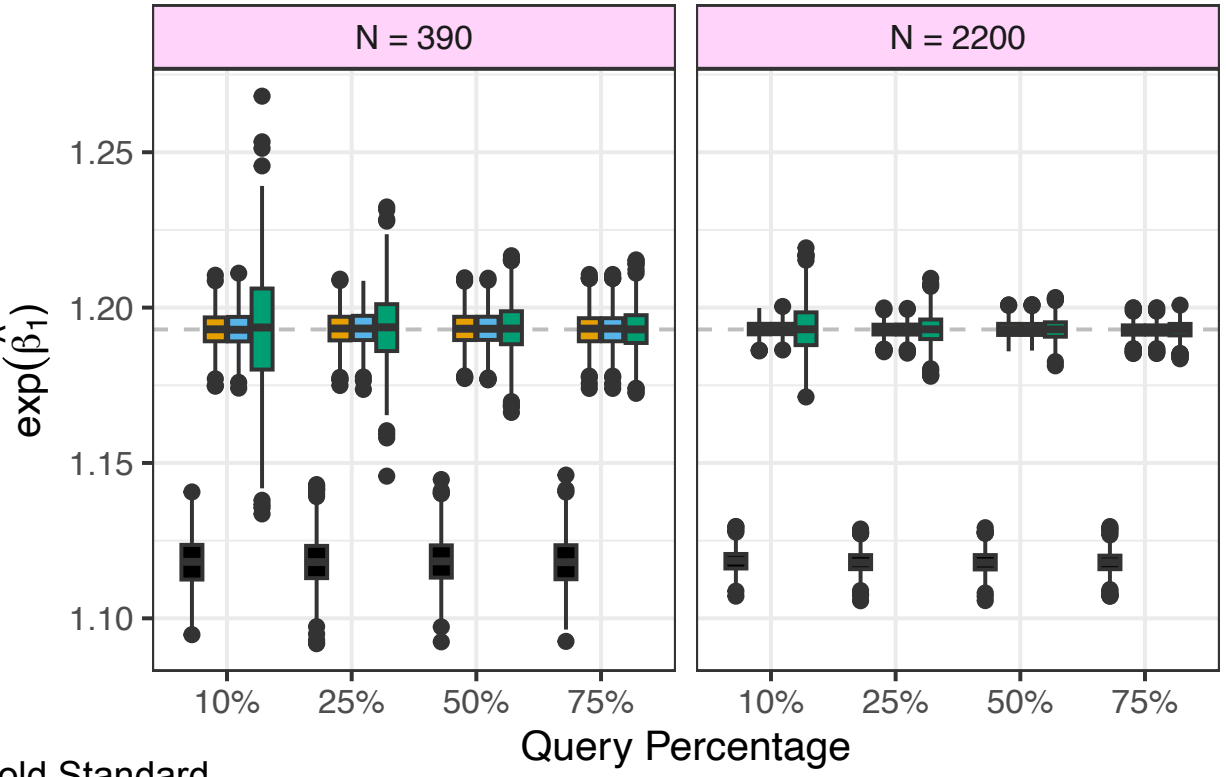
We follow this estimation procedure.



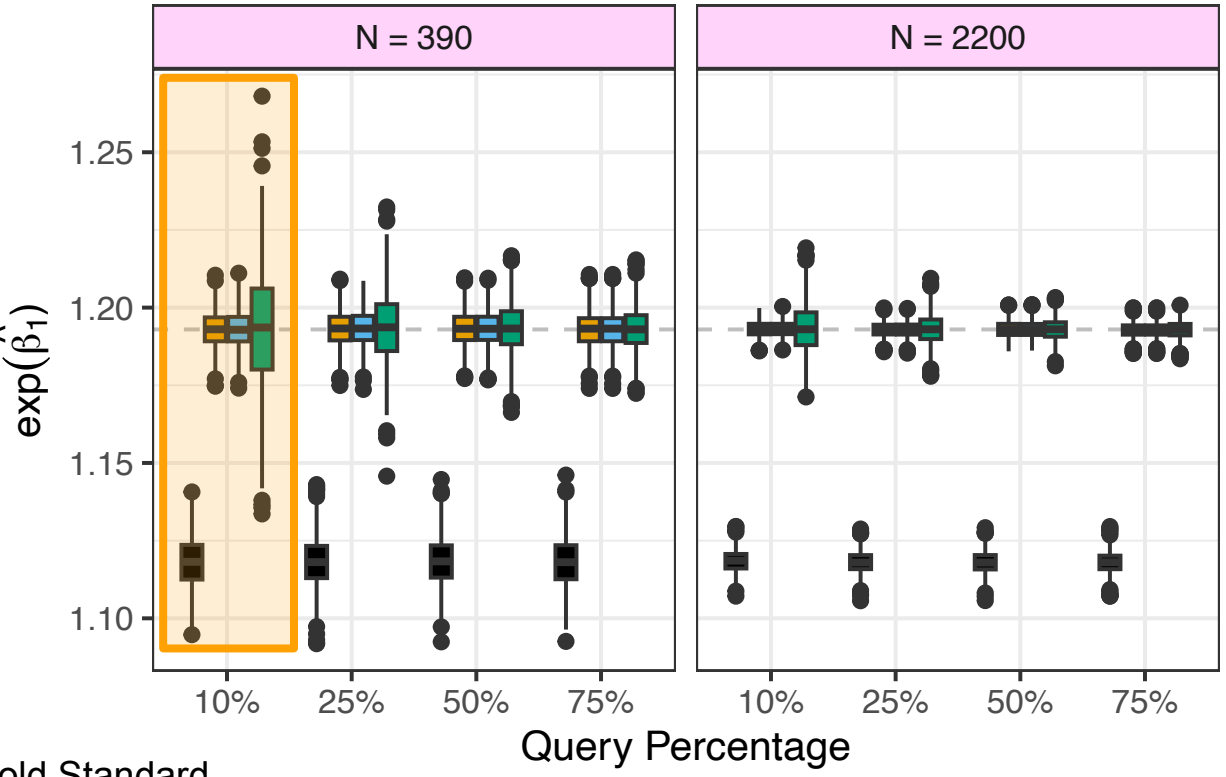
We focus on two variations.

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

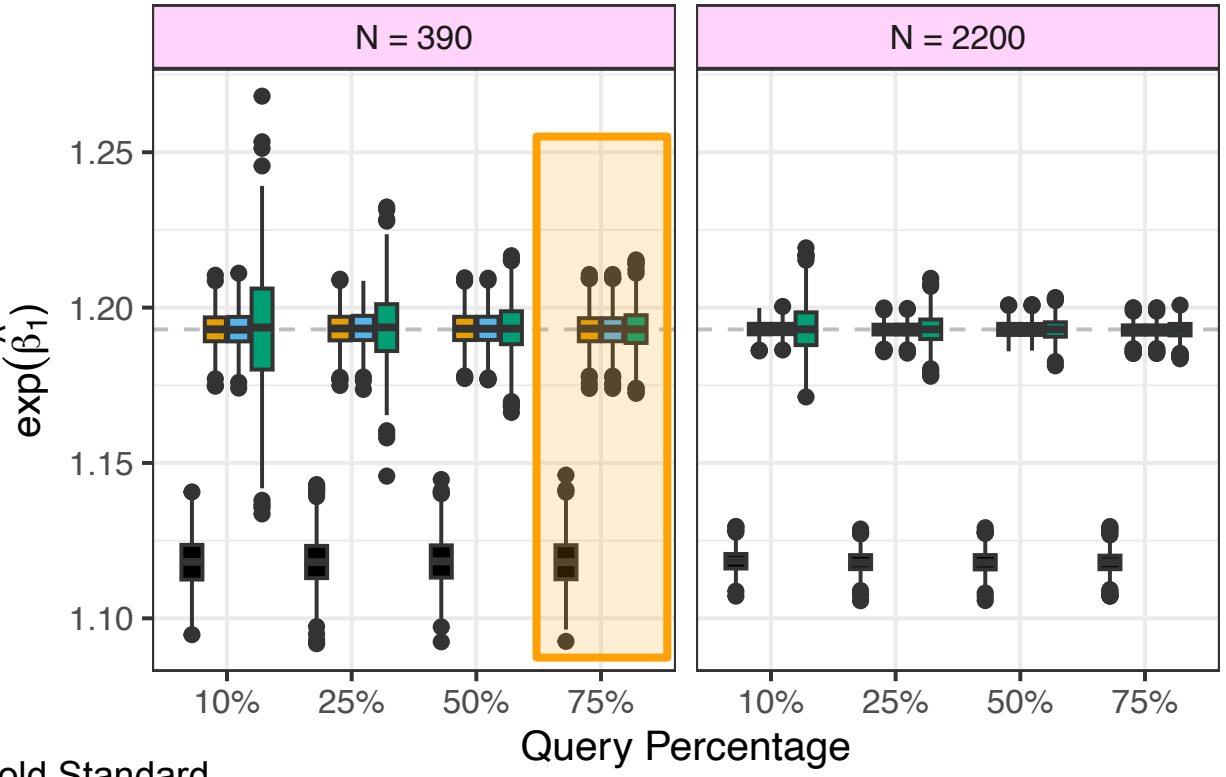
What happens as we vary the query size?



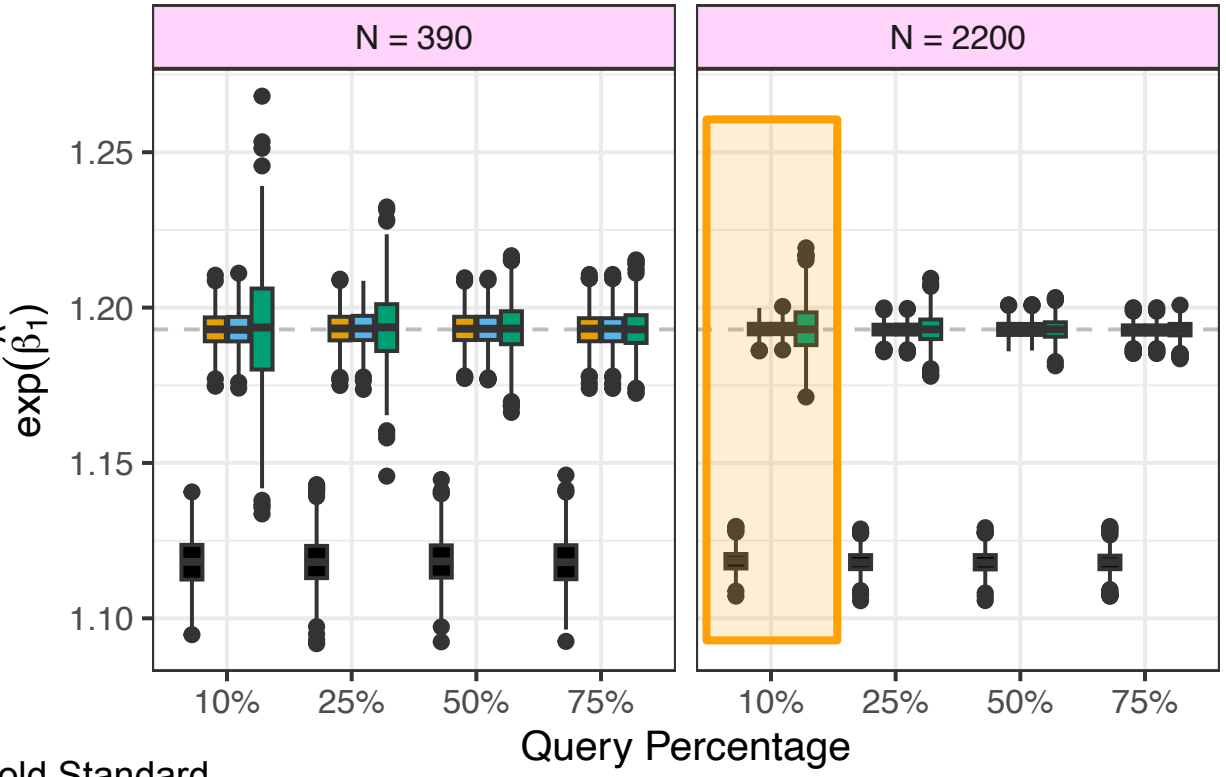
What happens as we vary the query size?



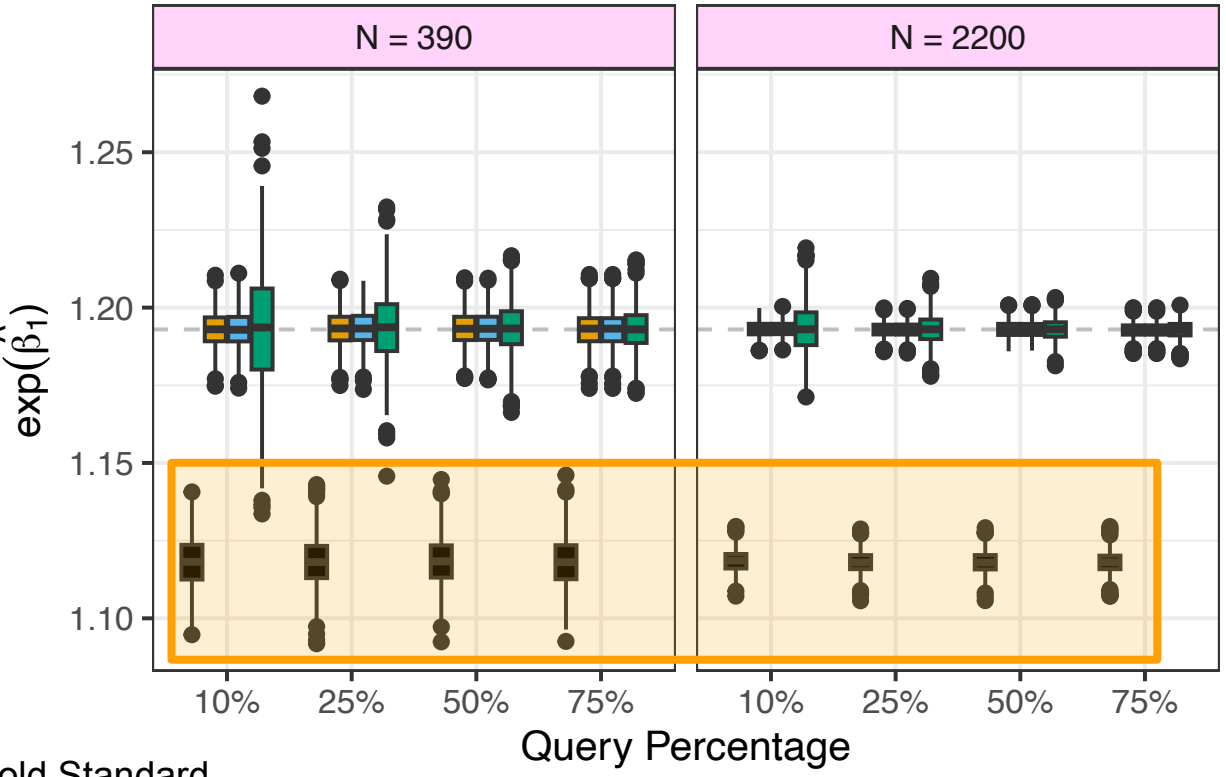
What happens as we vary the query size?



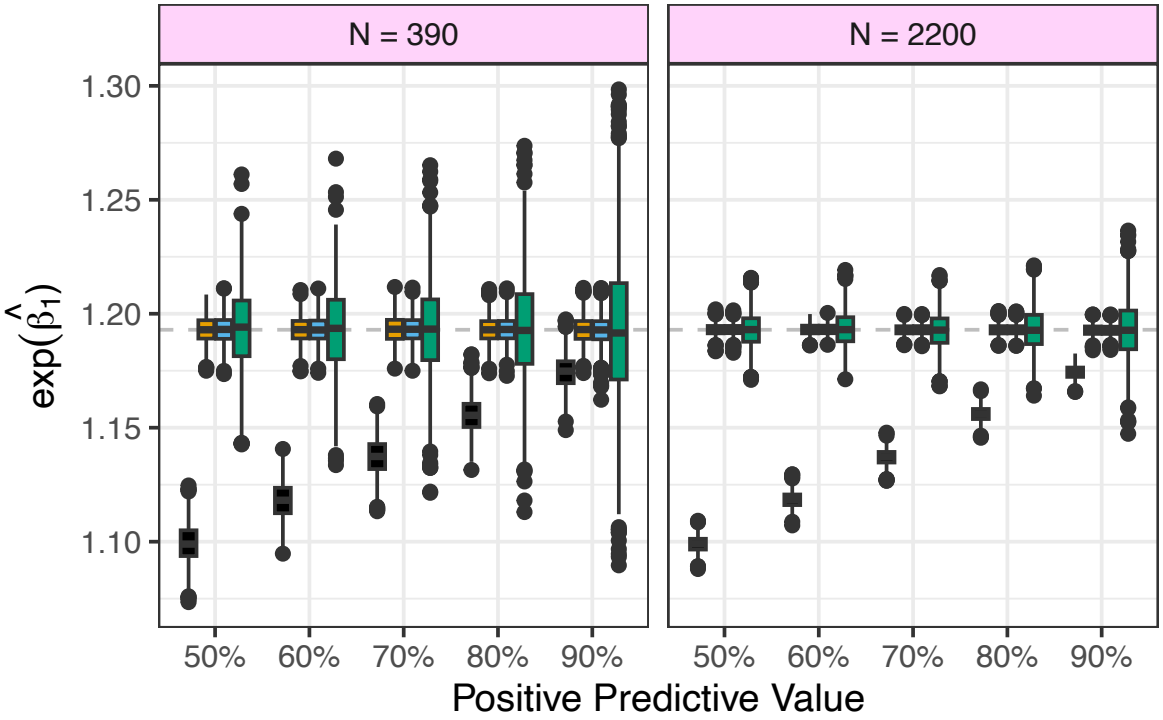
What happens as we vary the query size?



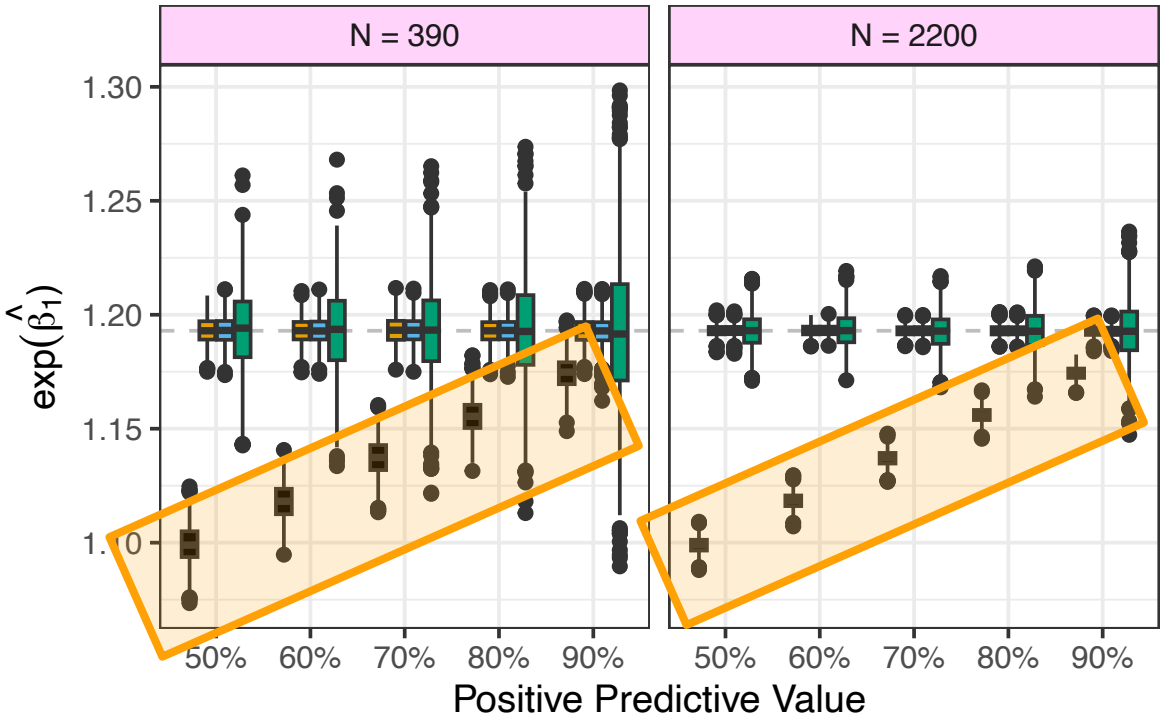
What happens as we vary the query size?



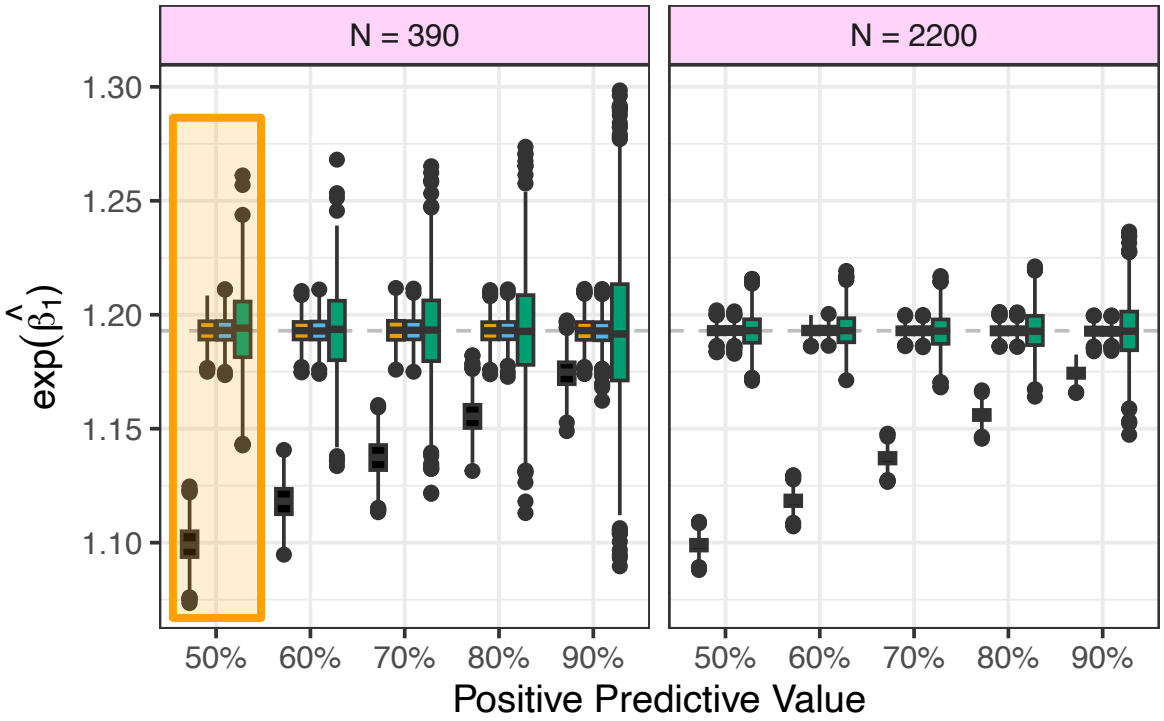
What happens as the errors get worse?



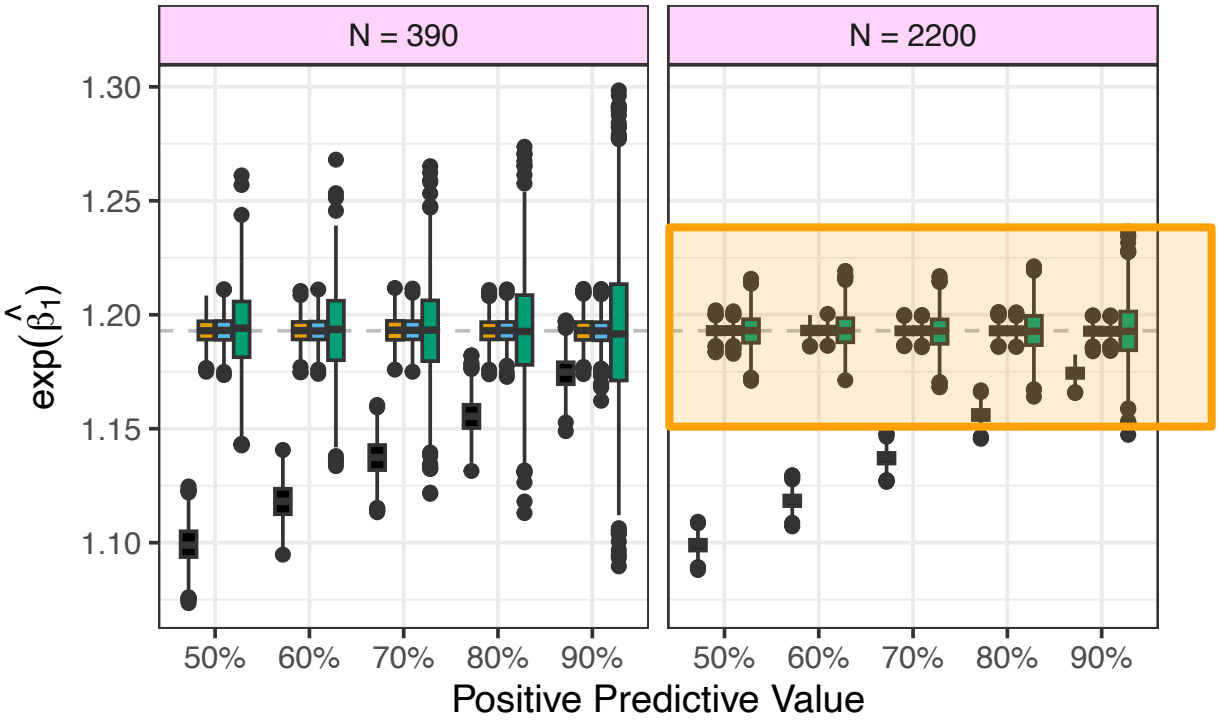
What happens as the errors get worse?



What happens as the errors get worse?



What happens as the errors get worse?



Takeaways

- 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 let asymptotic behavior take care of some of the **issues in the competitors**, the MLE still shows improved behavior.

4

Case Study

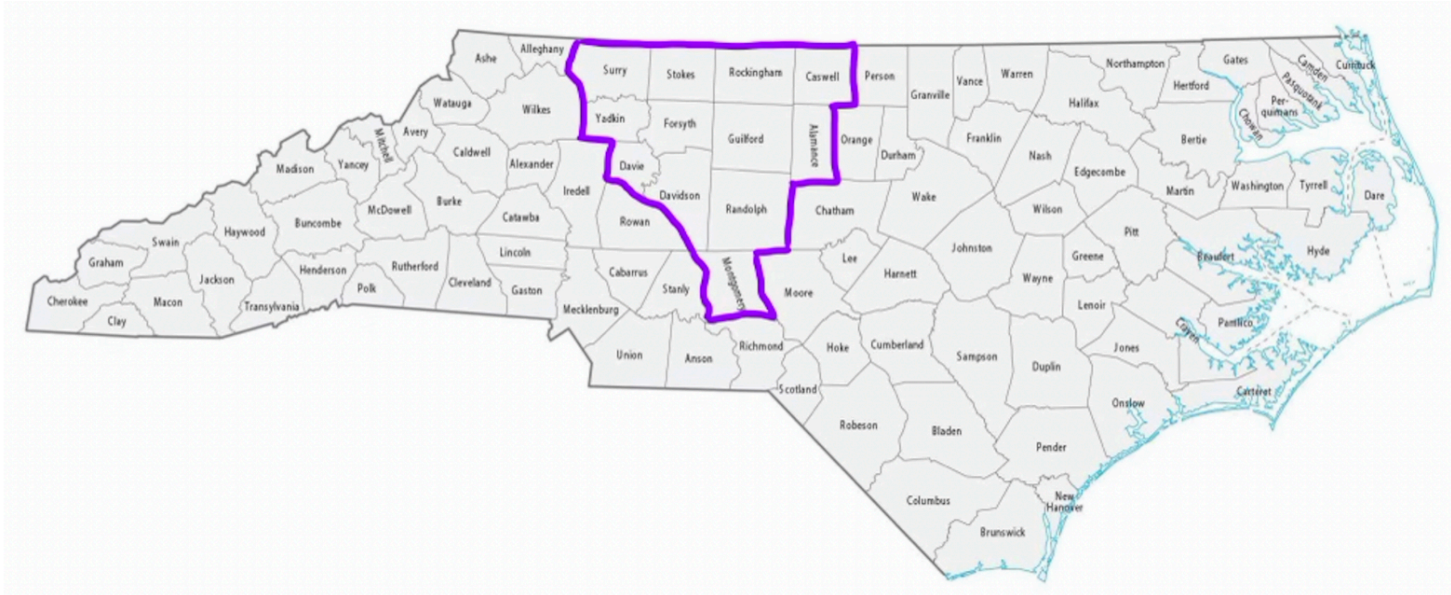
To reorient in context:

- 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**?”

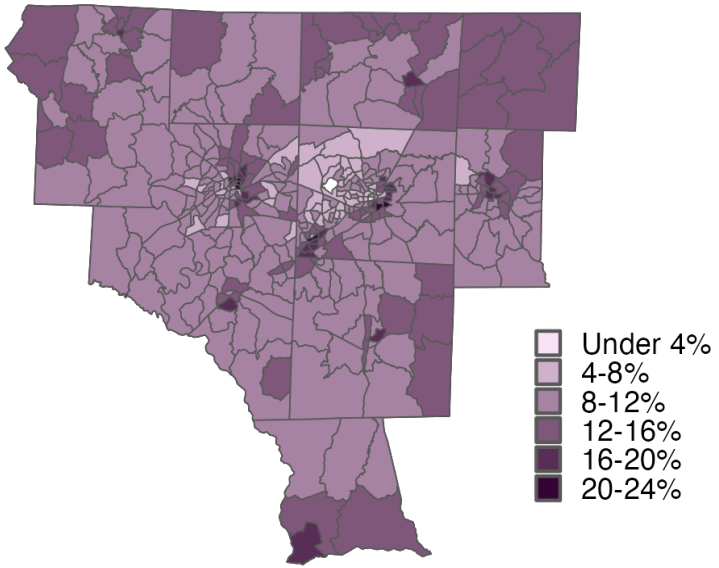
- We now apply our method to data from the **Piedmont Triad** in **North Carolina**.

What's the Piedmont 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**.



From Lotspeich et. al 2025, we adapt:

- N = 387 census tract **population centers and sizes** from the 2010 census
- F = 701 **healthy food retailers** from the 2022 USDA SNAP retailer locator release
- Tract level **diabetes prevalences** from the 2022 CDC PLACES release
- Indicators of **metro status** for each neighborhood derived from the 2010 USDA RUCA code release

The outcome model just got a bit more specific.

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

The outcome model just got a bit more specific.

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

the expected diabetes case count (Y_i) in a tract given its food access at that radius (X_i) and metro status (M_i)

The outcome model just got a bit more specific.

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

the expected diabetes case count (Y_i) in a tract given its food access at that radius (X_i) and metro status (M_i)

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

The outcome model just got a bit more specific.

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 expected diabetes case count (Y_i) in a tract given its food access at that radius (X_i) and metro status (M_i)

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

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

The outcome model just got a bit more specific.

add these!

the population offset

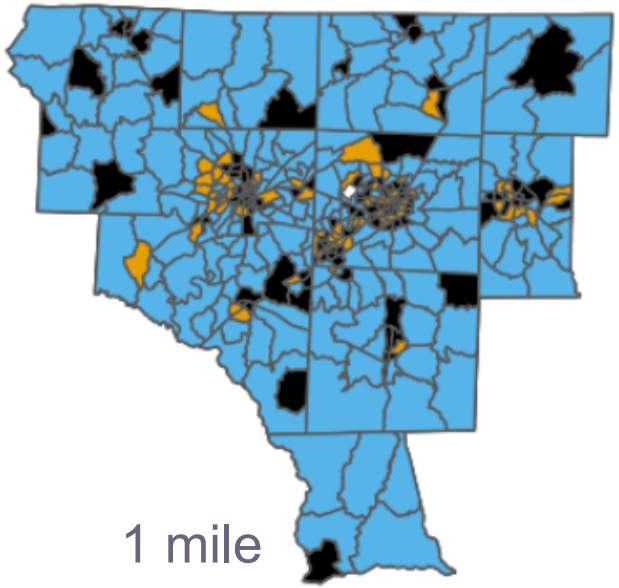
$$\log\{\mathrm{E}_{\beta}(Y_i \mid X_i, M_i)\} = \beta_0 + \beta_1 X_i + \beta_2 M_i + \beta_3 X_i \times M_i + \log(O_i)$$

the expected diabetes case count (Y_i) in a tract given its food access at that radius (X_i) and metro status (M_i)

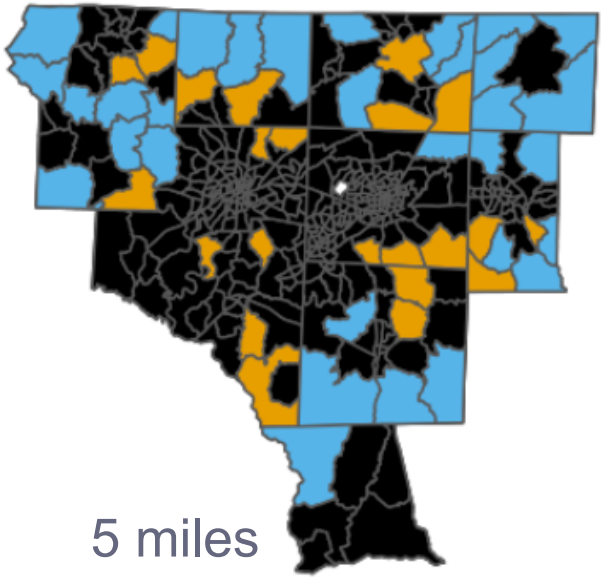
the (log) ratio of diabetes prevalence in a non-metro tract with food access at that radius compared to one without

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

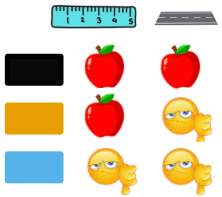
We can visualize food access at the tract level.



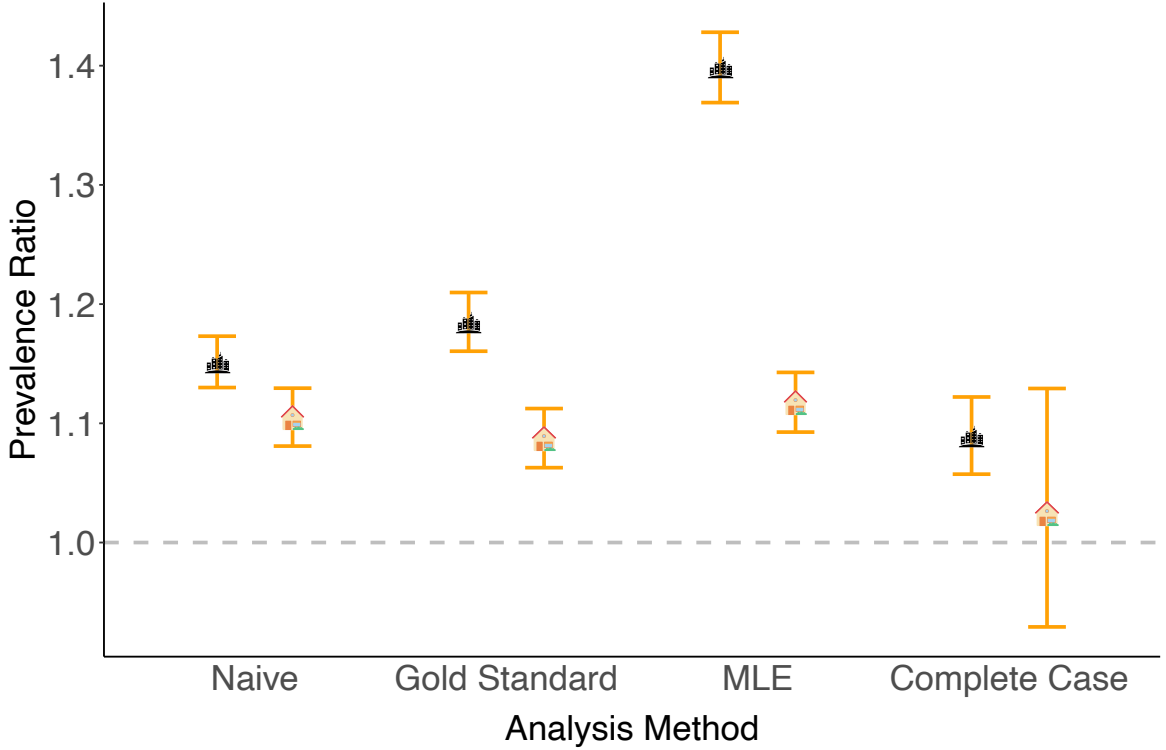
1 mile



5 miles

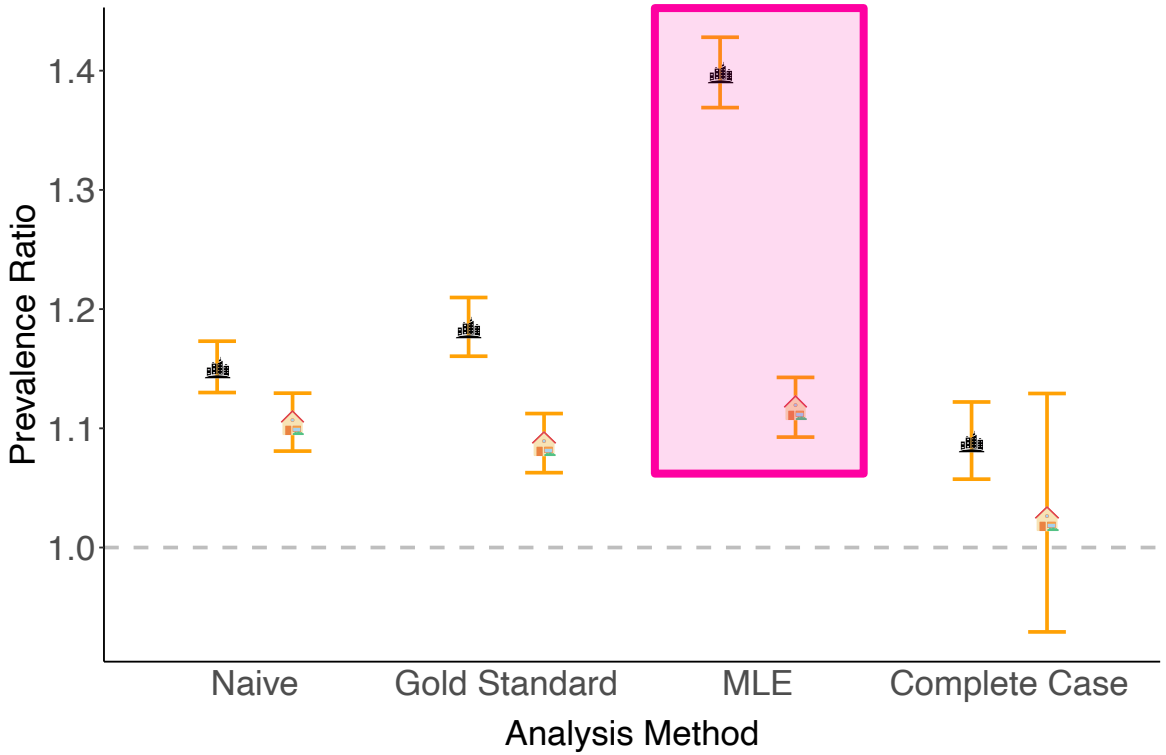


At the one mile radius:



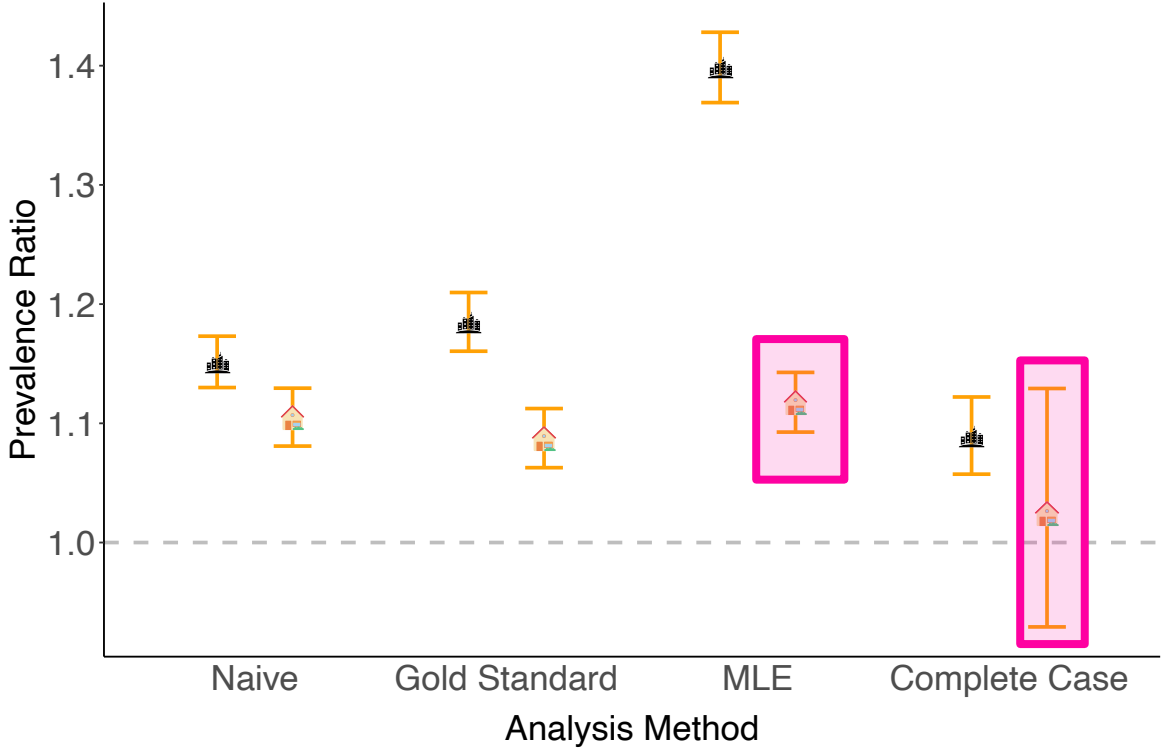
Metro Tract
Non Metro Tract

At the one mile radius:



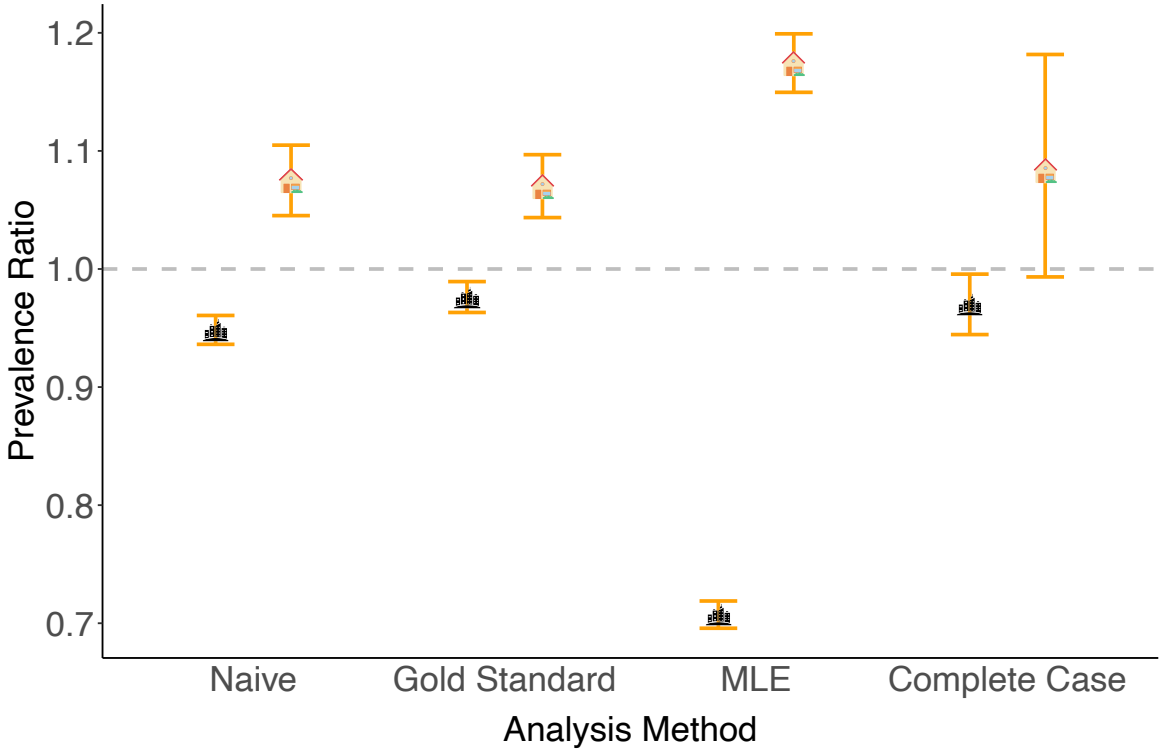
Metro Tract
Non Metro Tract

At the one mile radius:



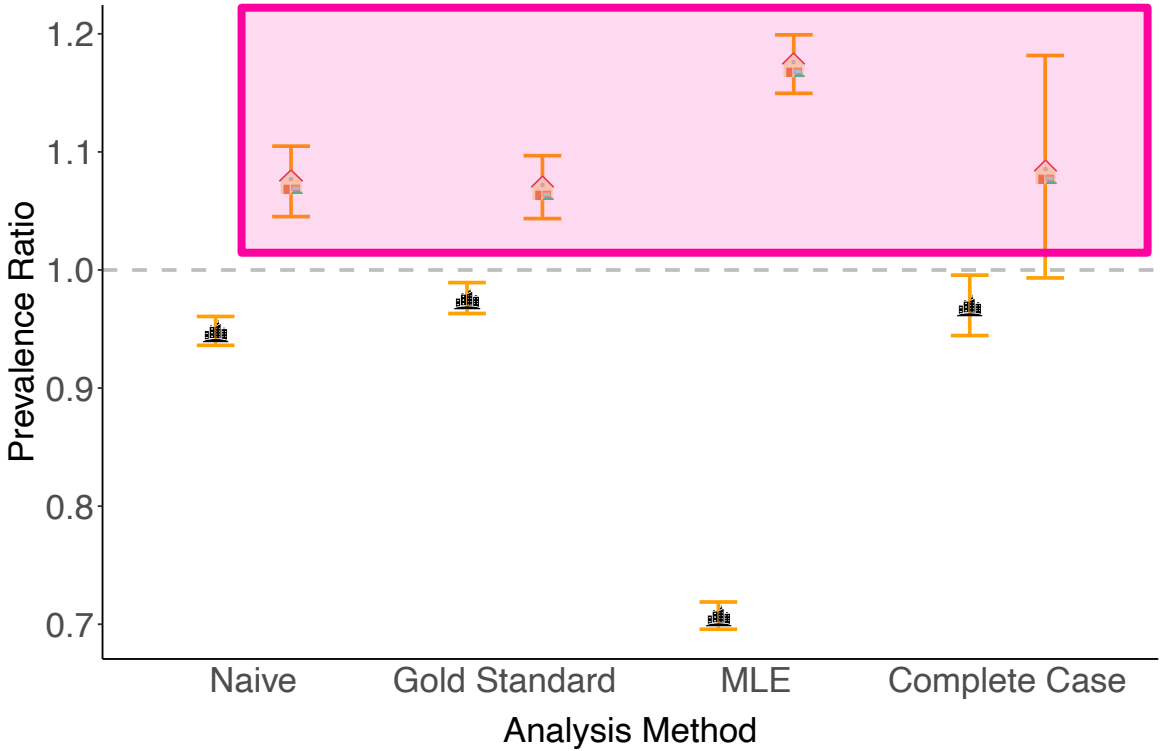
Metro Tract
Non Metro Tract

At the five mile radius:



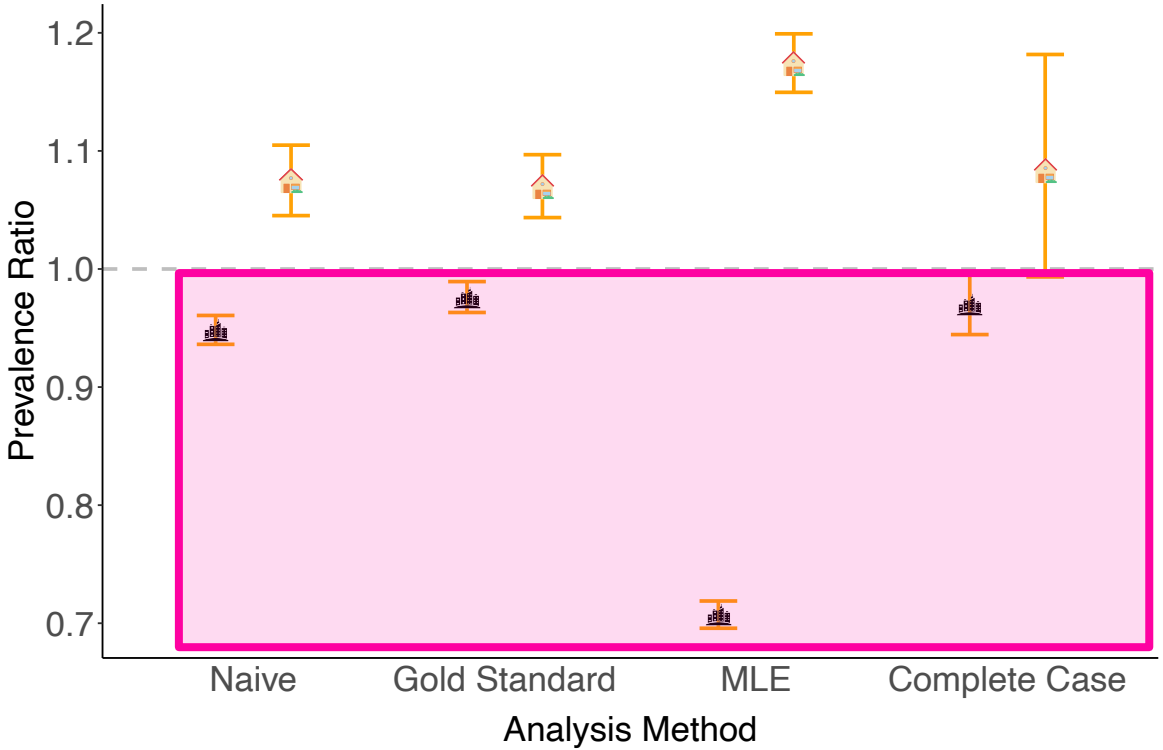
Metro Tract
Non Metro Tract

At the five mile radius:



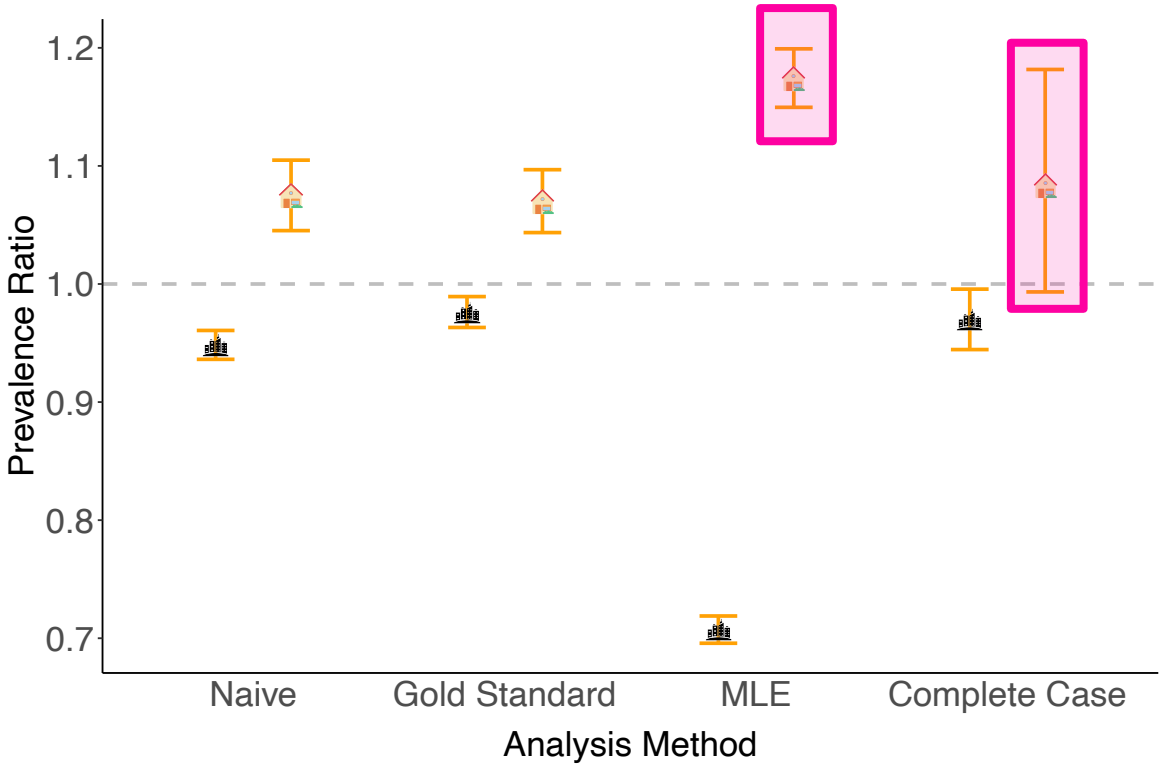
Metro Tract
Non Metro Tract

At the five mile radius:



Metro Tract
Non Metro Tract

At the five mile radius:



Metro Tract
Non Metro Tract

Takeaways

- Association patterns depend on the **radius**.
- 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.

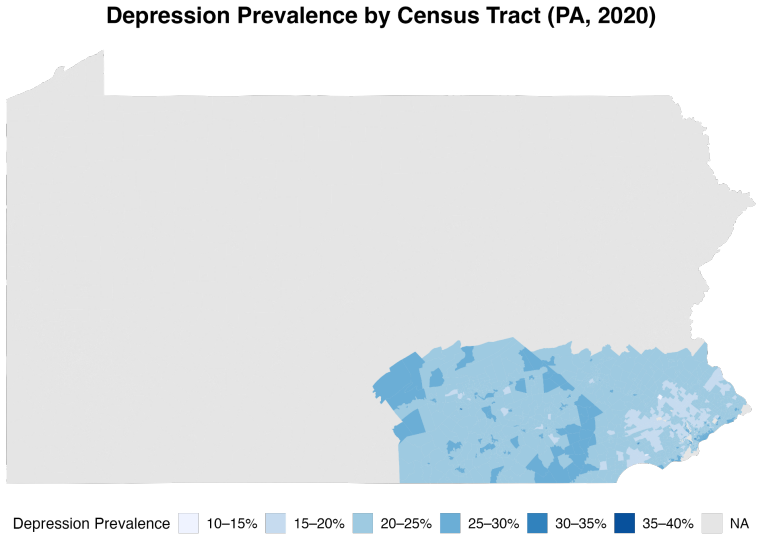
5

Wrap Up

Today, we:

- 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 an expectation-maximization algorithm in R
- Showcased its **asymptotic and small-sample advantages** via simulation studies
- Demonstrated its utility by estimating **diabetes prevalence** as a function of **food access and urbanicity** in North Carolina

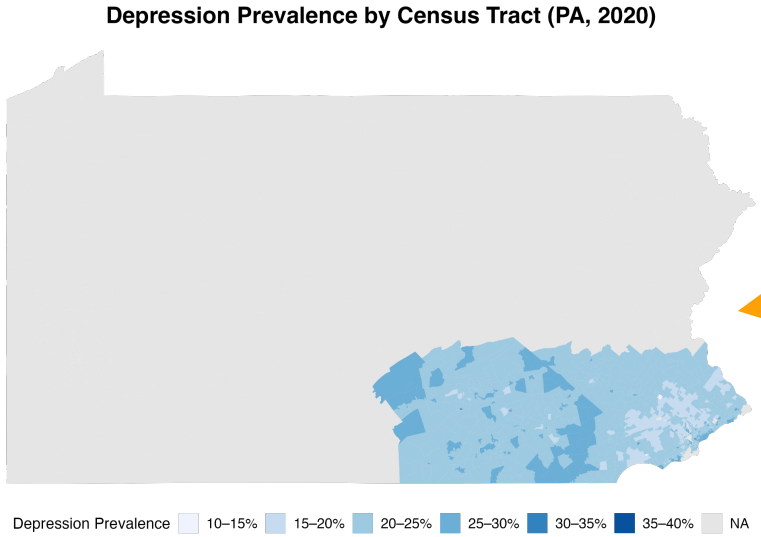
We can extend this framework!



$$X_i \in \{0,1\} \rightarrow X_i \in \mathbb{R}$$

Figure by Daniel Hoberman

We can extend this framework!

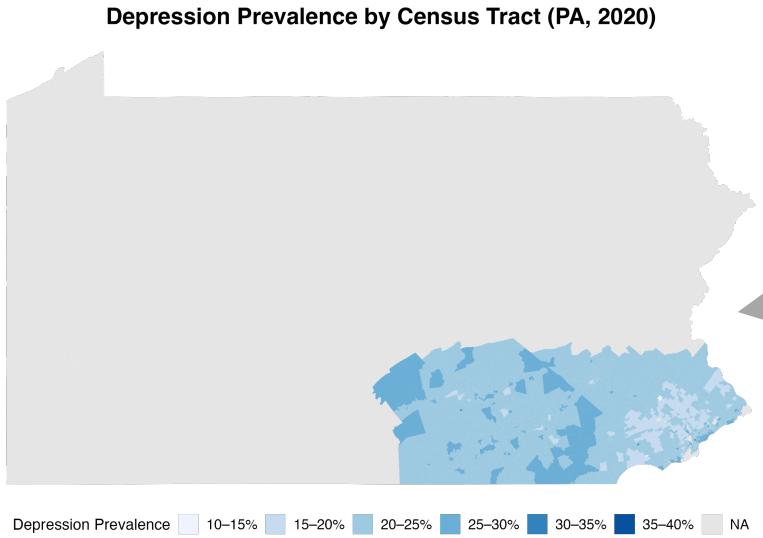


$$X_i \in \{0,1\} \rightarrow X_i \in \mathbb{R}$$

new communities and new context

Figure by Daniel Hoberman

We can extend this framework!



more flexible method settings

$$X_i \in \{0,1\} \rightarrow X_i \in \mathbb{R}$$

new communities and new context

Figure by Daniel Hoberman

Acknowledgements

ashley.e.mullan@vanderbilt.edu

ashleymullan.github.io

ashleymullan.bsky.social

https://arxiv.org/abs/2505.01465

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