Modeling Misclassification: Exploring the Relationship between Diabetes and Access to Healthy Foods

Ashley Mullan - Vanderbilt University - IDWSDS 2025


bitly

Motivation

Healthy Eating ➡️ Healthy Living

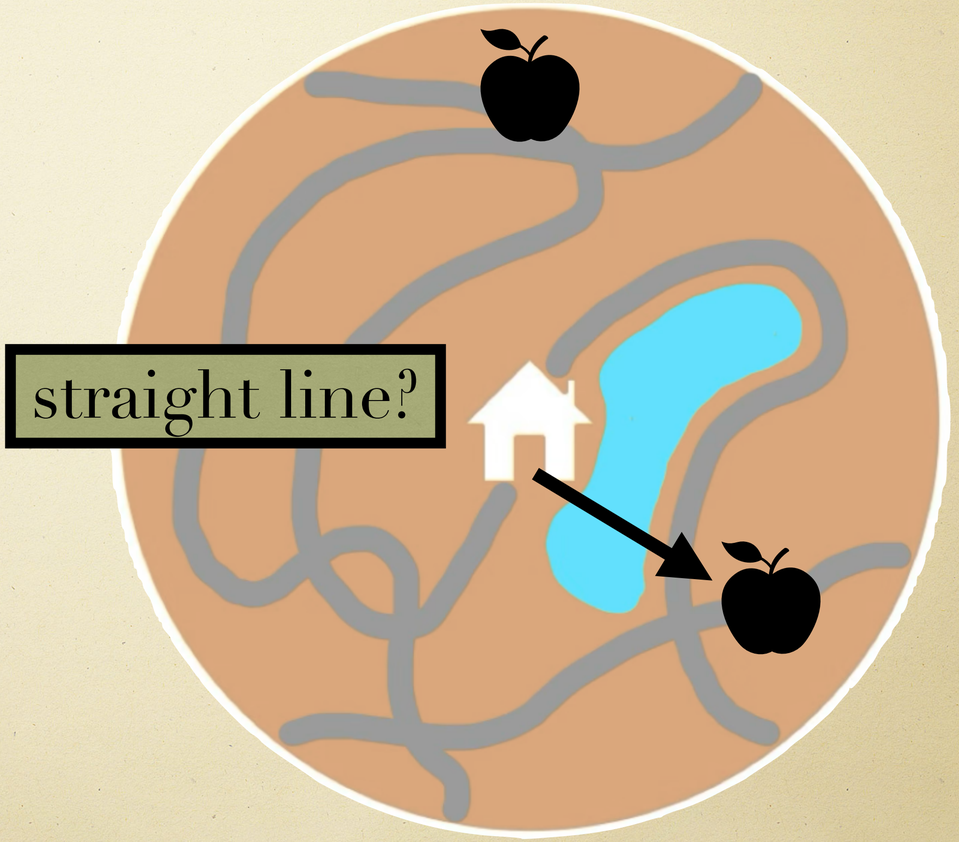
- A healthy diet increases the likelihood of good overall health and decreases risk of preventable illness.
- Maintaining a healthy diet requires access to healthy food, which may be hindered by geography, income, or social factors.
- Review studies found high prevalence of diabetes in food insecure households.

Measuring Food Access 🍏

- We define a **neighborhood** of interest with a **radius**, a **centroid**, and possibly some **healthy food retailers**.
- If there's **at least one** healthy food retailer within that radius, we define the neighborhood as having **access** to healthy food.
- We need a **distance metric** to make that decision.

Measuring Food Access 🍏

- We define a **neighborhood** of interest with a **radius**, a **centroid**, and possibly some **healthy food retailers**.
- If there's **at least one** healthy food retailer within that radius, we define the neighborhood as having **access** to healthy food.
- We need a **distance metric** to make that decision.

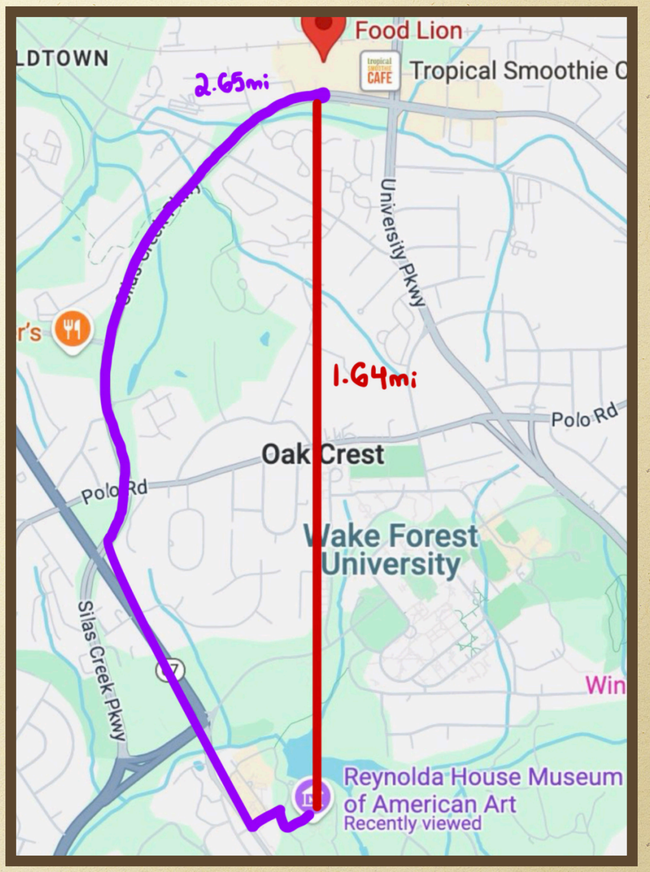


Measuring Food Access 🍏

- We define a **neighborhood** of interest with a **radius**, a **centroid**, and possibly some **healthy food retailers**.
- If there's **at least one** healthy food retailer within that radius, we define the neighborhood as having **access** to healthy food.
- We need a **distance metric** to make that decision.

The Distance Tradeoff

- The **Haversine distance** is easy to find but **ignores obstacles** and is an **error-prone** underestimate.
- The **route-based distance** is **more accurate** but computationally and financially **expensive**.
- We can compute route-based distances for some neighborhoods with the **ggmap** R package.
- If we want to study N neighborhoods, we can use a **two-phase design** to maximize our information, even if we only **query** n < N route-based distances.



Can we counteract missingness and misclassification issues in our data and still accurately estimate an association between access to healthy food and diabetes prevalence?

Methods

For each neighborhood i, we observe:

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

For each neighborhood i, we observe:

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

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

For each neighborhood i, we observe:

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

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

route-based, which is not always seen

the **exposure**, representing the food access in the neighborhood

Haversine, which might be wrong but is always seen

For each neighborhood i, we observe:

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

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

the **covariates**, assumed to be error-free

route-based, which is not always seen

the **exposure**, representing the food access in the neighborhood

Haversine, which might be wrong but is always seen

We use the N (independent) census tracts to build this likelihood and an EM algorithm to maximize it.

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

We use the N (independent) census tracts to build this likelihood and an EM algorithm to maximize it.

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

from the **queried** tracts

(factors into Y | X, **Z** ~ Poisson and X | X*, **Z** ~ Bernoulli)

We use the N (independent) census tracts to build this likelihood and an EM algorithm to maximize it.

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

from the **queried** tracts
(factors into Y | X, **Z** ~ Poisson
and X | X*, **Z** ~ Bernoulli)

from the **unqueried** tracts
(factors the same way but
marginalizes over X)

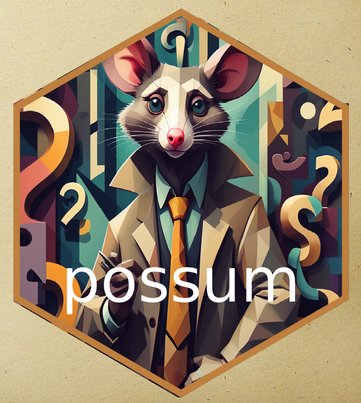
We use the N (independent) census tracts to build this likelihood and an EM algorithm to maximize it.

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

from the **queried** tracts
(factors into Y | X, **Z** ~ Poisson
and X | X*, **Z** ~ Bernoulli)

from the **unqueried** tracts
(factors the same way but
marginalizes over X)

this R package now
does it for you!

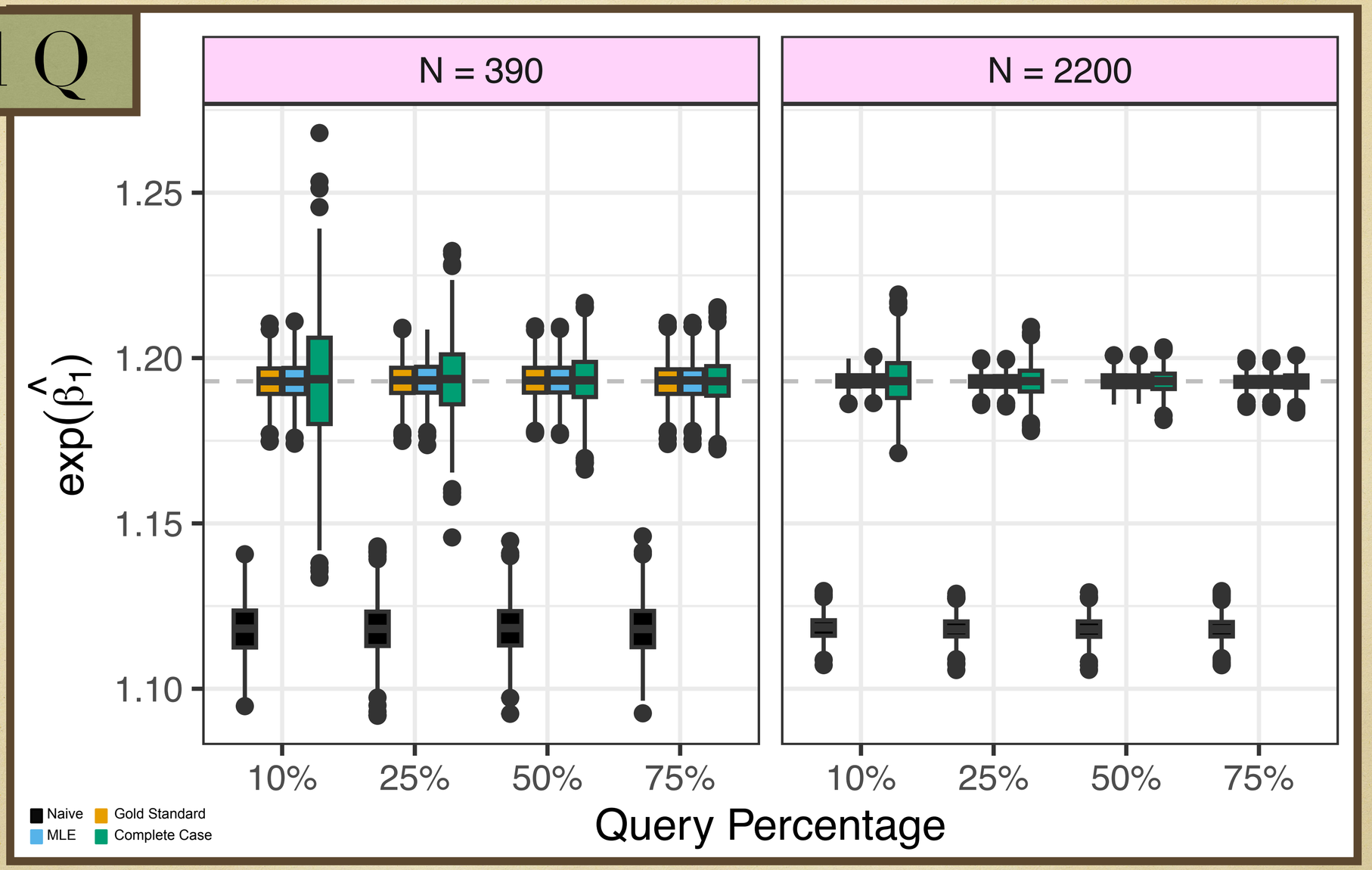


Simulations

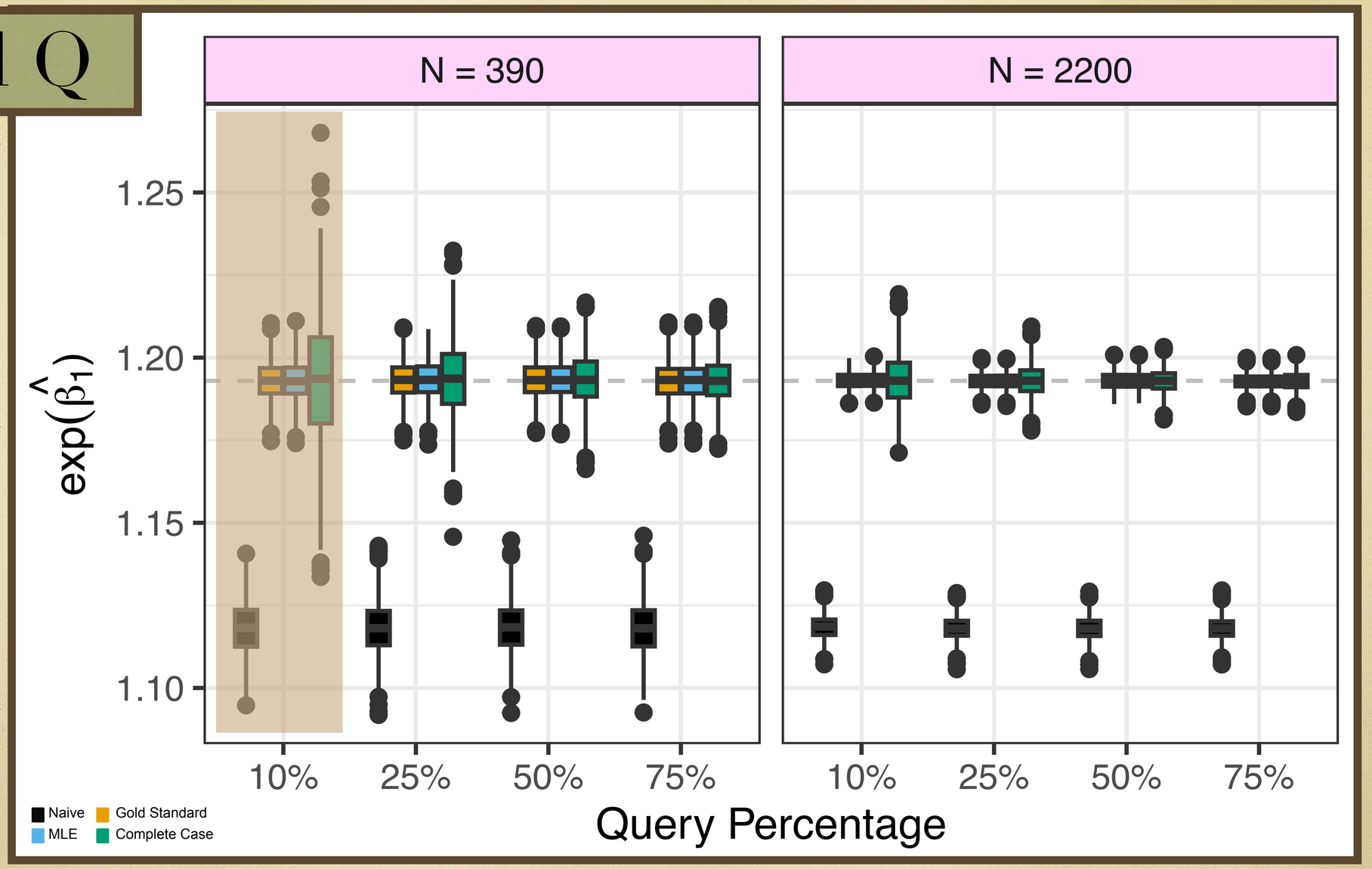
Simulation Roadmap 

- We want to estimate $\exp(\beta_1)$, the **prevalence ratio**.
- We demonstrate the results of two simulation studies that explore:
  - *What happens if we query fewer neighborhoods?*
    We fix everything but the **query proportion** and try N = 390 and N = 2200.
  - *What happens as the errors get more drastic?*
    We fix everything but the **positive predictive value** and try N = 390 and N = 2200.

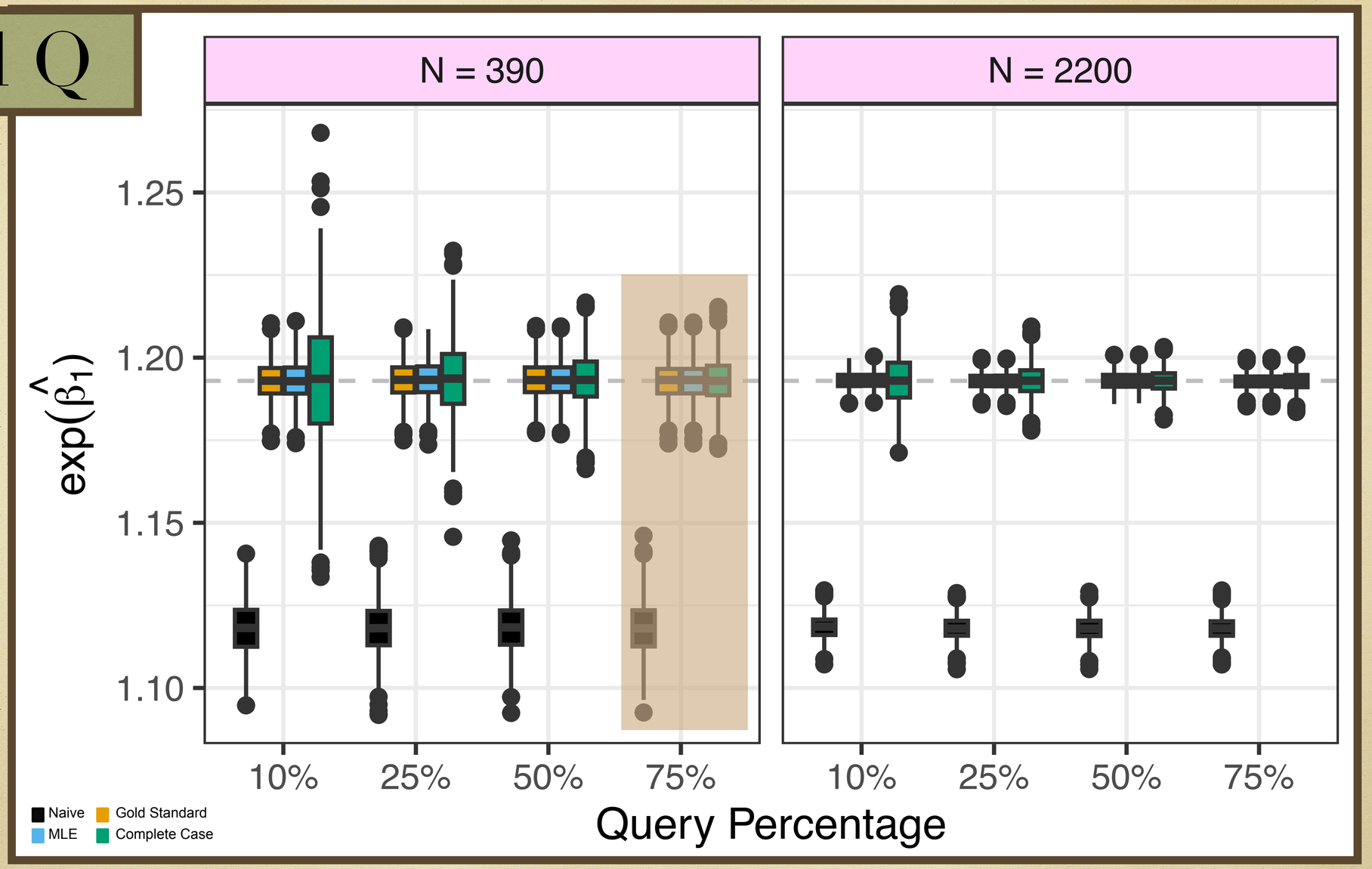
Varied Q


N = 390
N = 2200
1.25
1.20
1.15
1.10
$\exp(\hat{\beta}_1)$
10%
25%
50%
75%
Query Percentage
Naive
Gold Standard
MLE
Complete Case

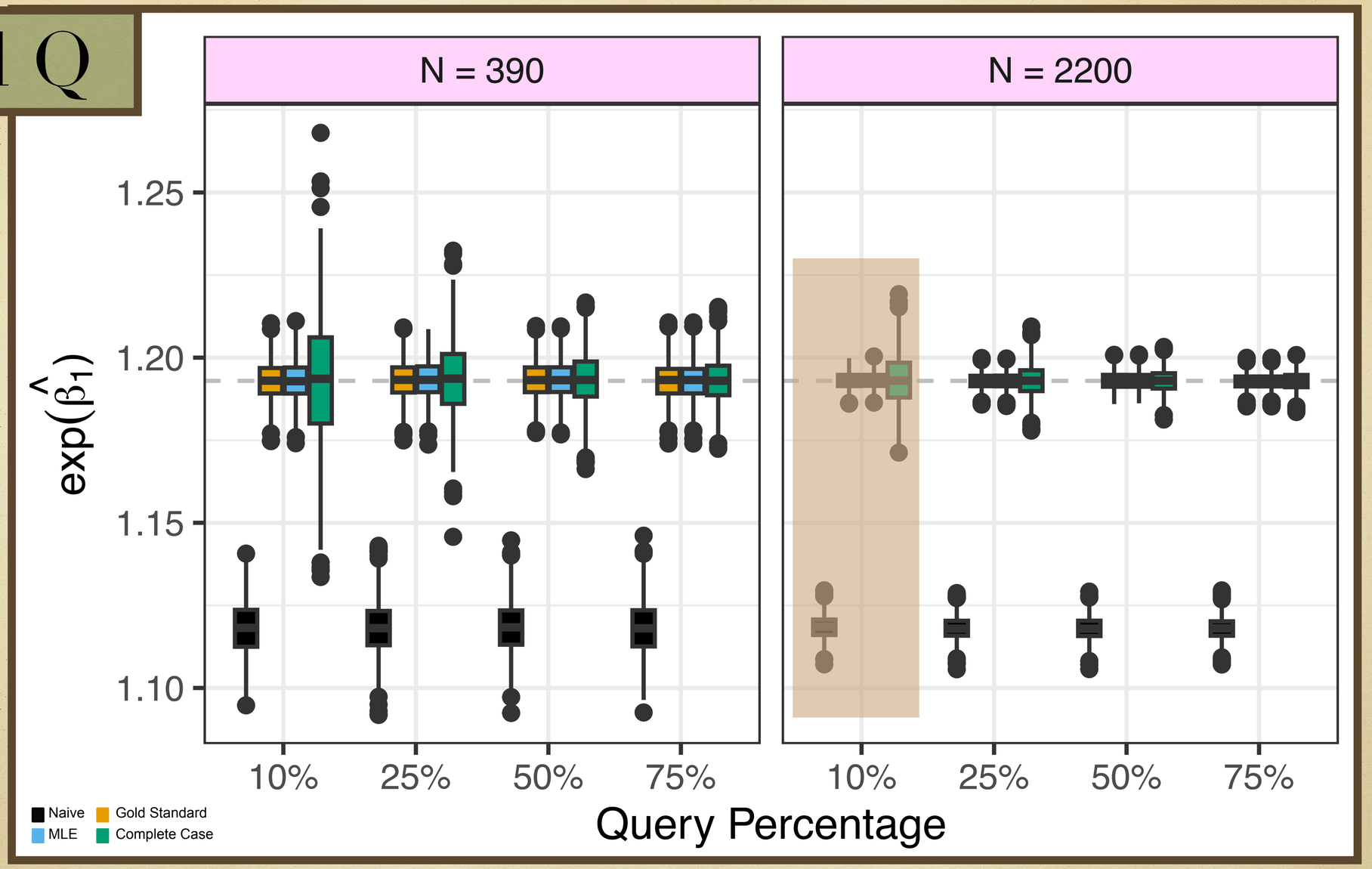
Varied Q


N = 390
N = 2200
1.25
1.20
1.15
1.10
$\exp(\hat{\beta}_1)$
10%
25%
50%
75%
Query Percentage
Naive
Gold Standard
MLE
Complete Case

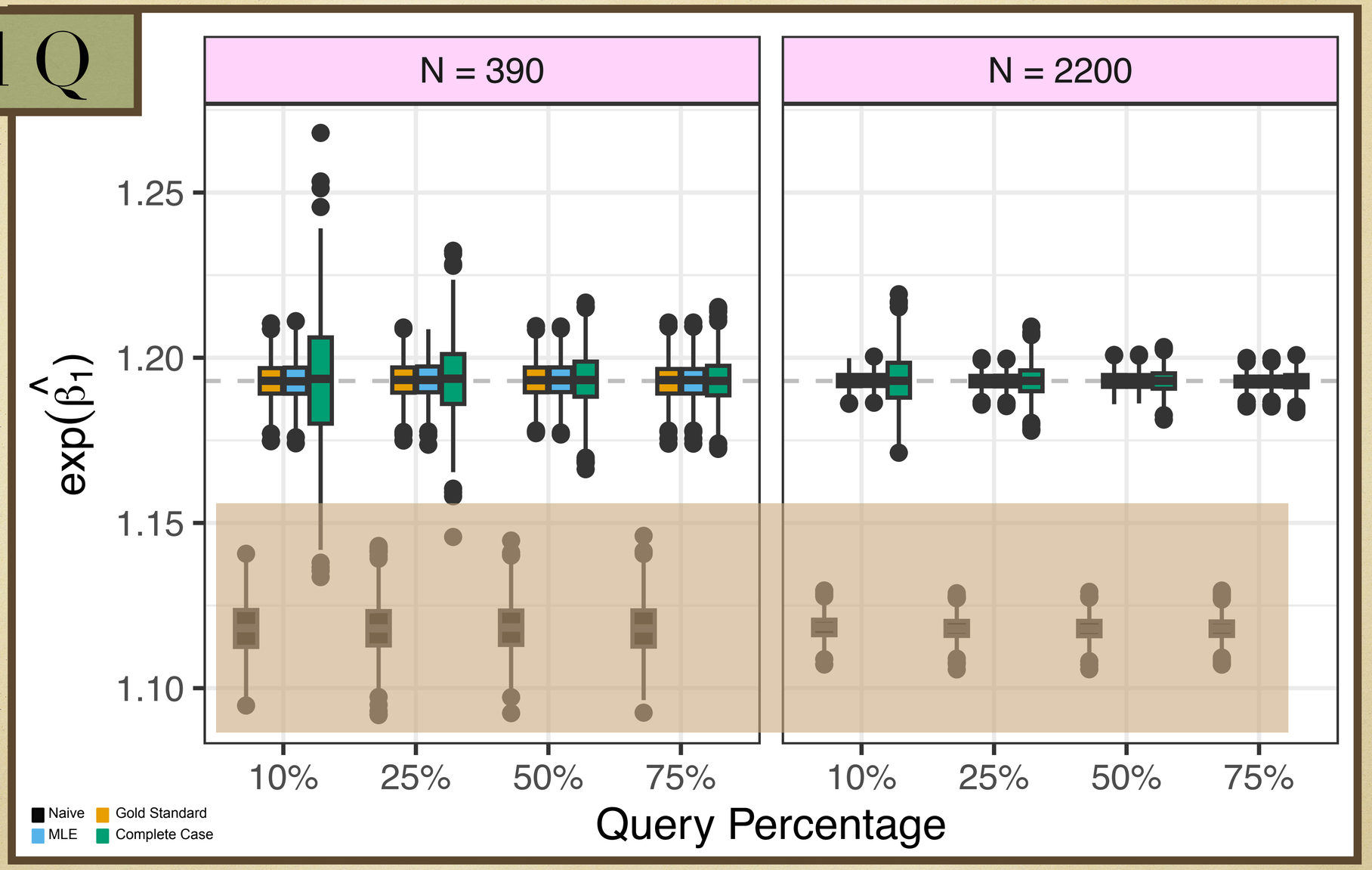
Varied Q


N = 390
N = 2200
1.25
1.20
1.15
1.10
exp(β̂1)
10%
25%
50%
75%
Query Percentage
Naive
Gold Standard
MLE
Complete Case

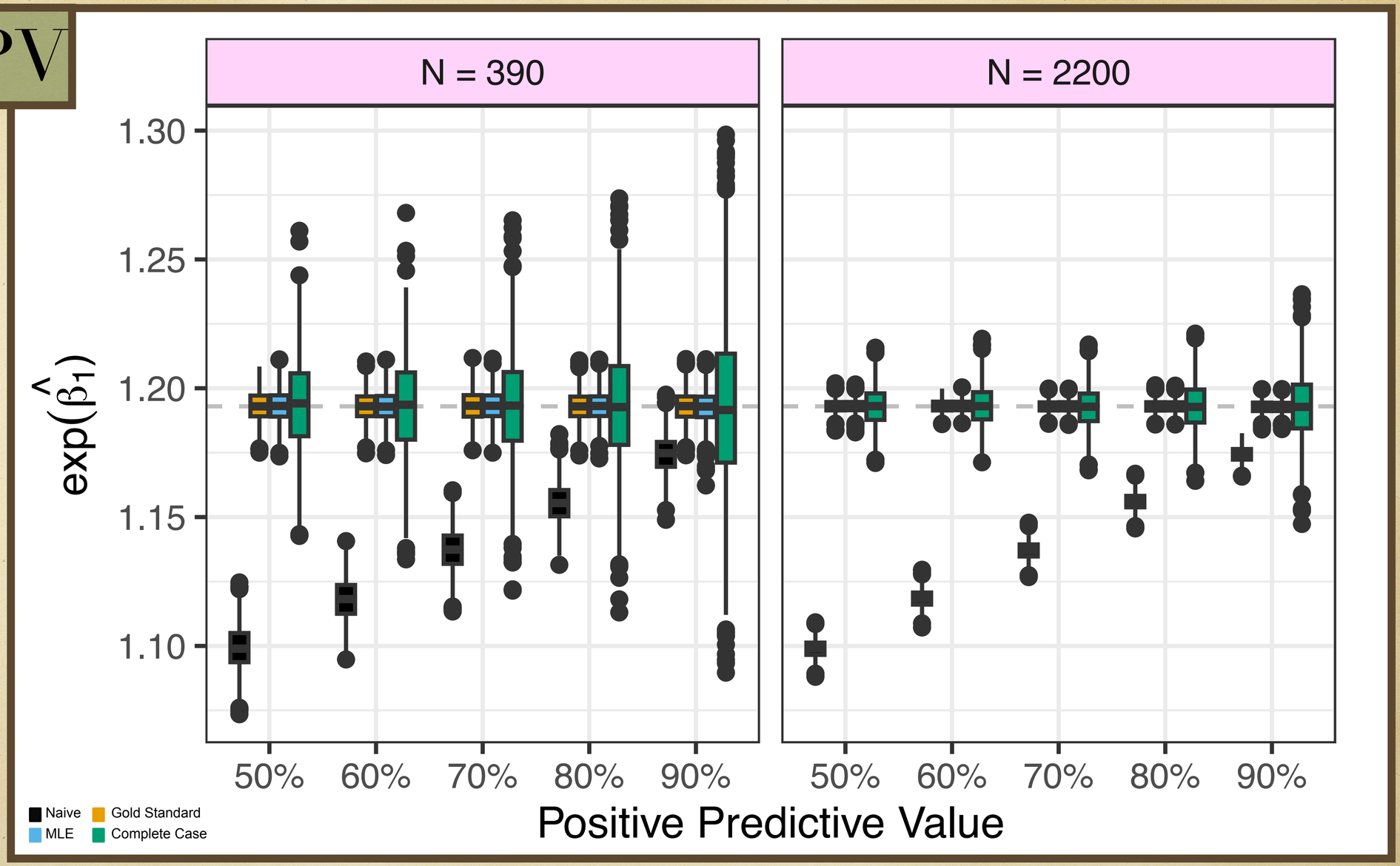
Varied Q


N = 390
N = 2200
1.25
1.20
1.15
1.10
$\exp(\hat{\beta}_1)$
10%
25%
50%
75%
Query Percentage
Naive
Gold Standard
MLE
Complete Case

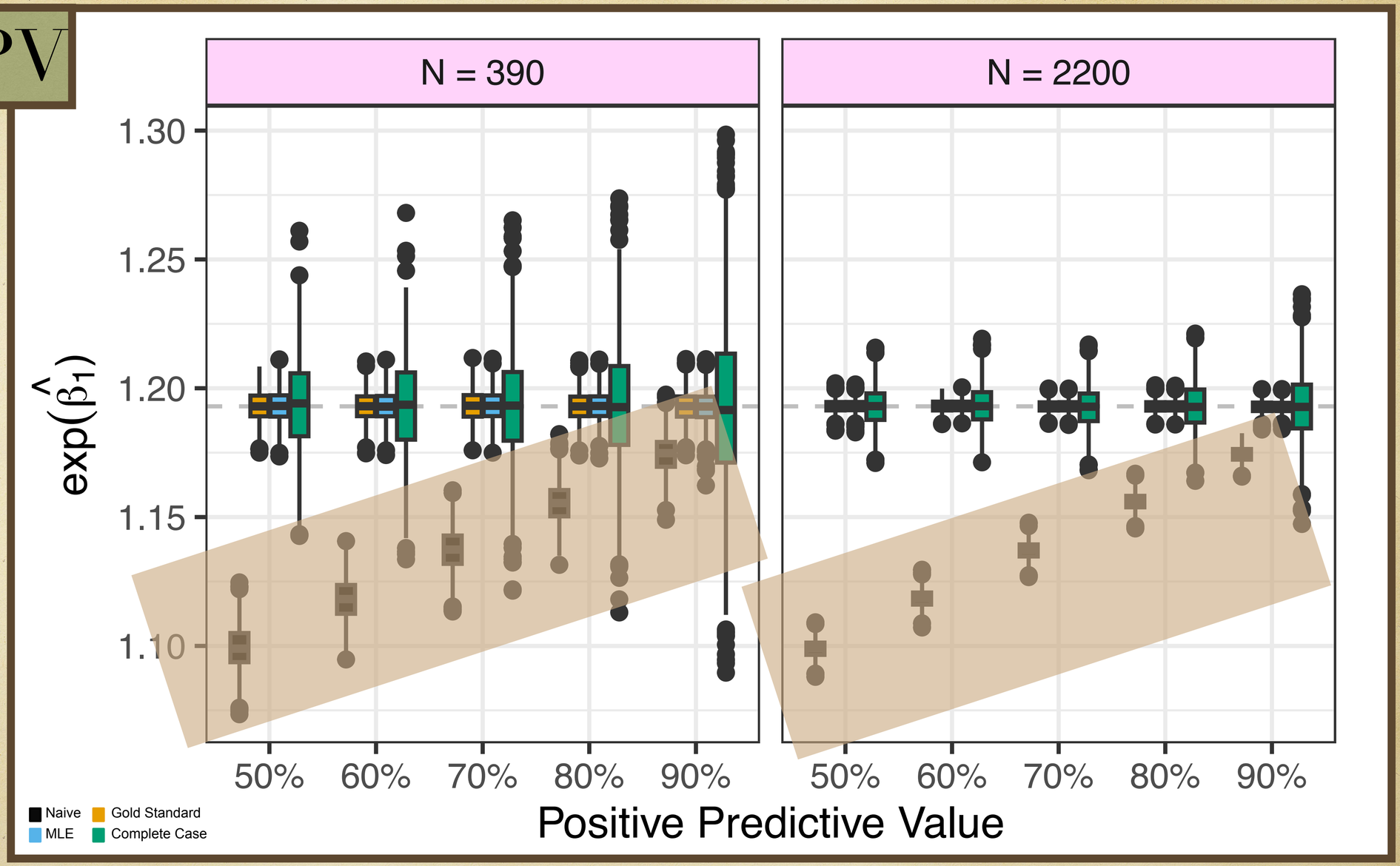
Varied Q


N = 390
N = 2200
1.25
1.20
1.15
1.10
exp(β̂1)
10%
25%
50%
75%
Query Percentage
Naive
Gold Standard
MLE
Complete Case

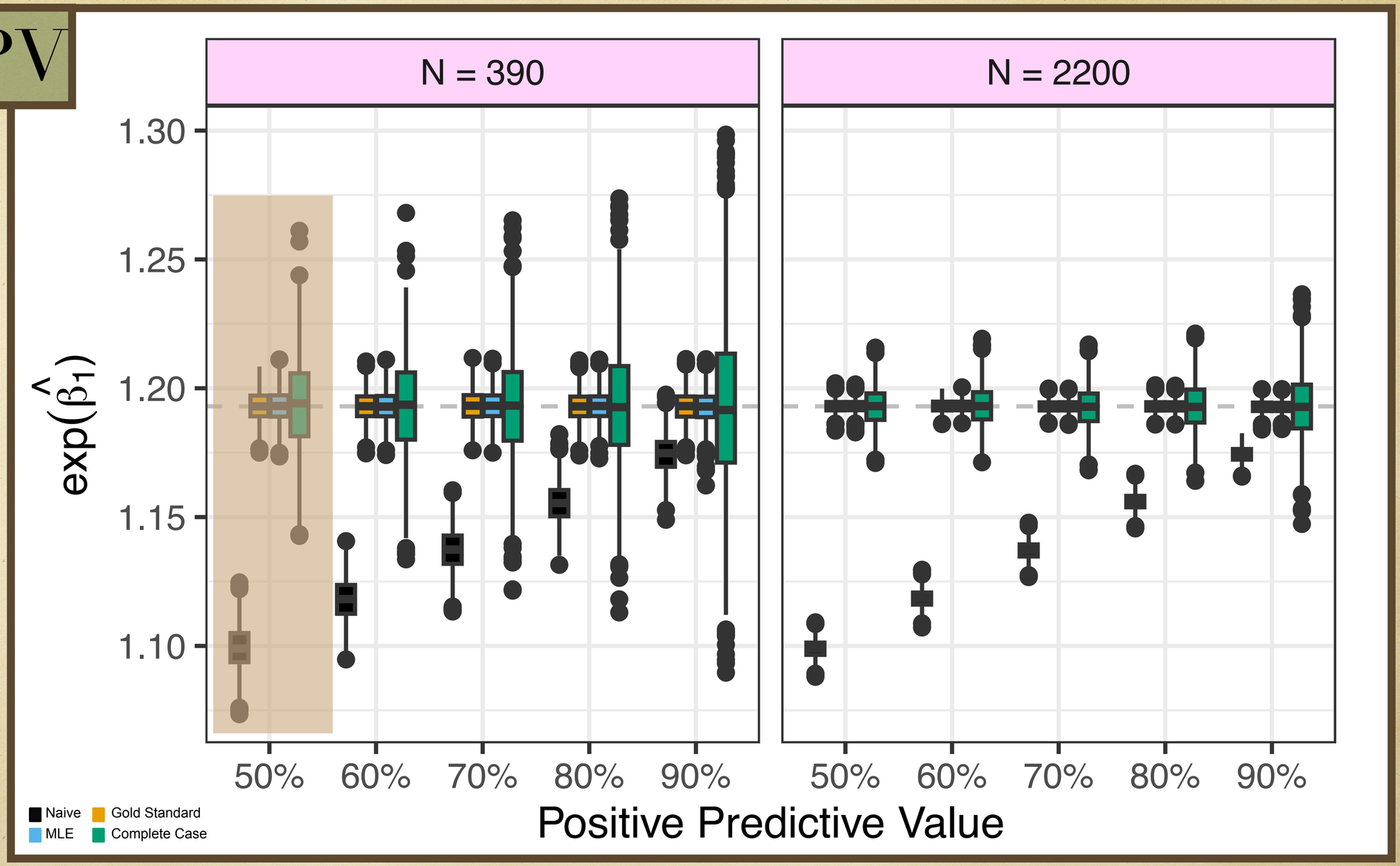
Varied PPV


N = 390
N = 2200
1.30
1.25
1.20
1.15
1.10
$\exp(\hat{\beta}_1)$
50%
60%
70%
80%
90%
Positive Predictive Value
Naive
Gold Standard
MLE
Complete Case

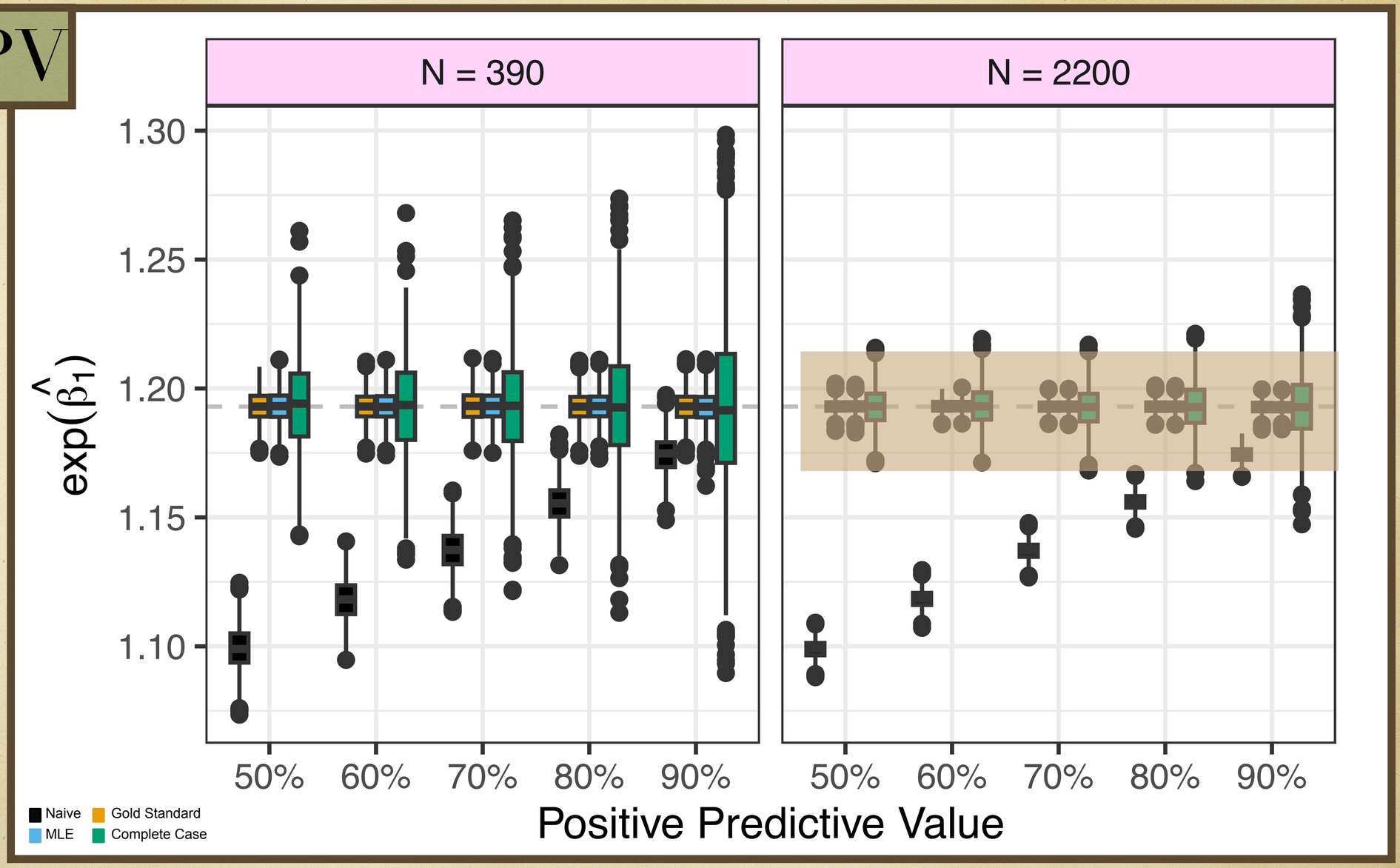
Varied PPV


N = 390
N = 2200
1.30
1.25
1.20
1.15
1.10
$\exp(\hat{\beta}_1)$
50%
60%
70%
80%
90%
Positive Predictive Value
Naive
Gold Standard
MLE
Complete Case

Varied PPV

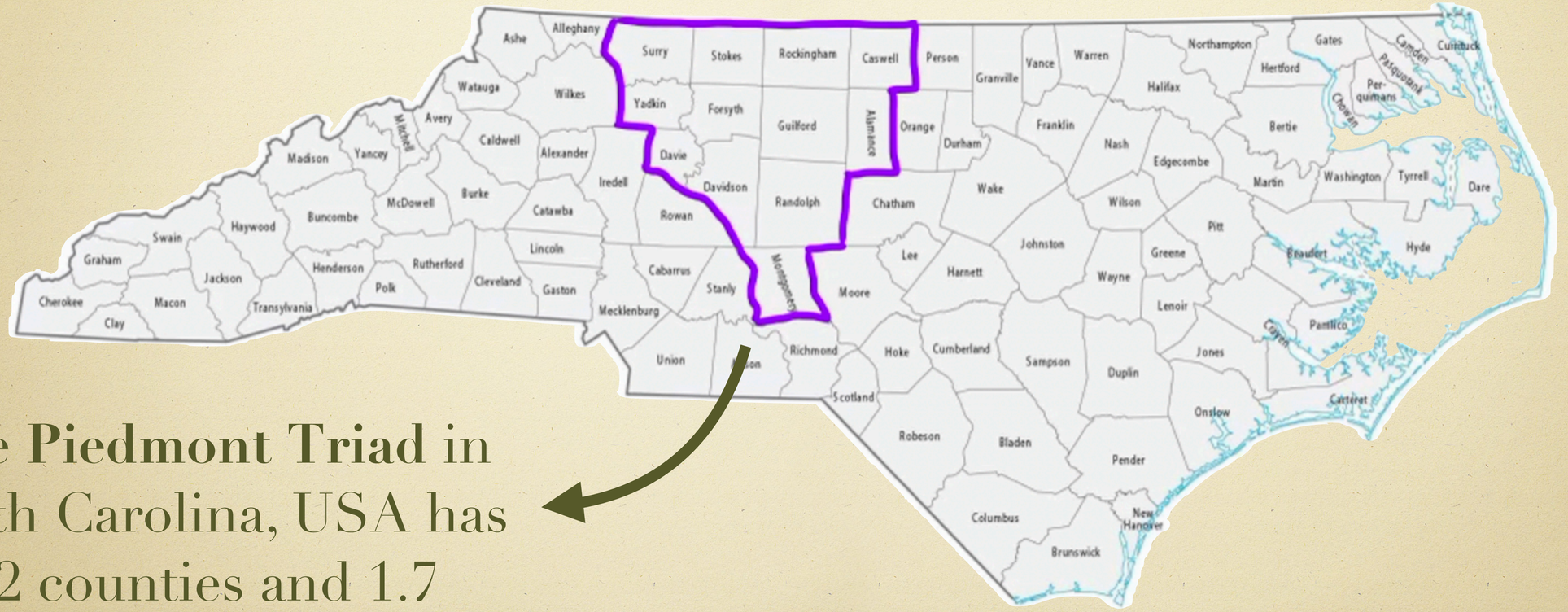

N = 390
N = 2200
1.30
1.25
1.20
1.15
1.10
$\exp(\hat{\beta}_1)$
50%
60%
70%
80%
90%
Positive Predictive Value
Naive
Gold Standard
MLE
Complete Case

Varied PPV


N = 390
N = 2200
1.30
1.25
1.20
1.15
1.10
$\exp(\hat{\beta}_1)$
50%
60%
70%
80%
90%
Positive Predictive Value
Naive
Gold Standard
MLE
Complete Case

Case Study

The Study Region



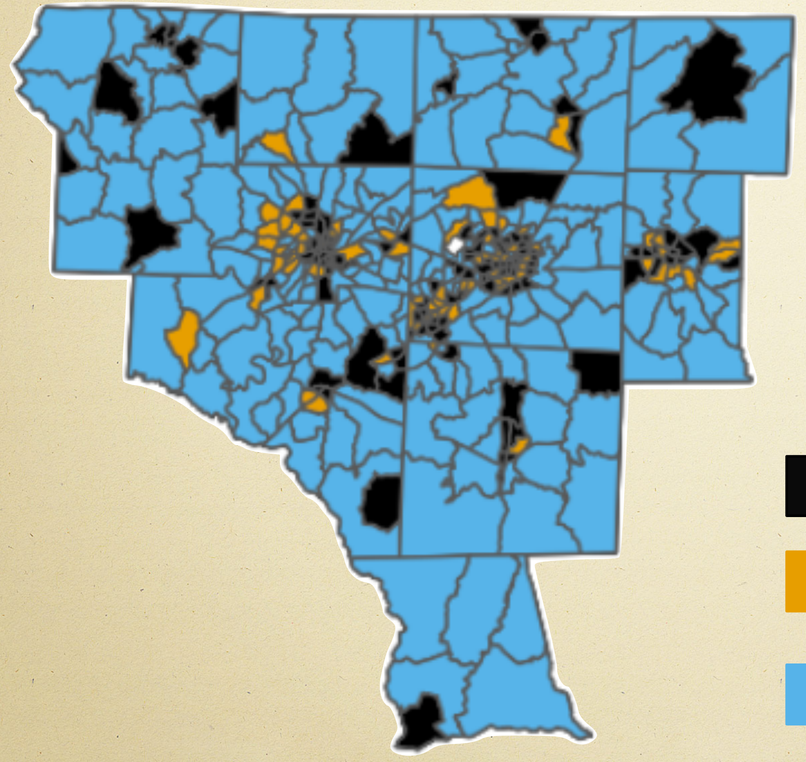
The **Piedmont Triad** in North Carolina, USA has 12 counties and 1.7 million people.

We have data describing:

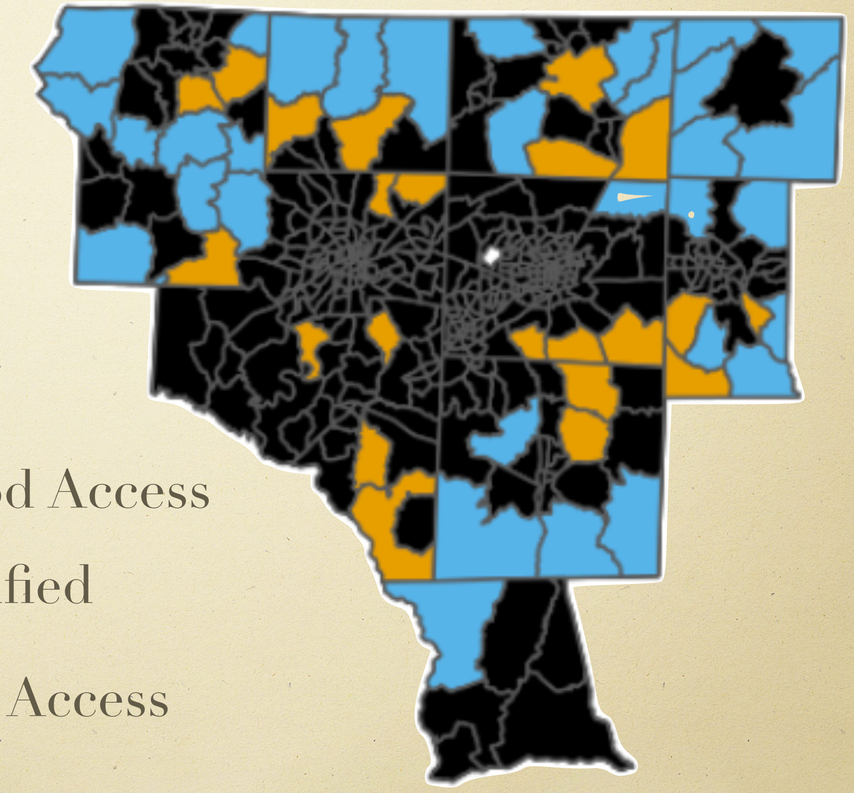
- N = 387 census tract **population centers** from the 2010 census.
- M = 701 **healthy food retailers** from the 2022 USDA SNAP release.
- **Population sizes** for each census tract and **diabetes prevalences** from the 2022 CDC PLACES release.
- **Metro indicators** for each census tract derived from the 2010 USDA RUCA release.

The Food Landscape in the Triad

One Mile



Five Miles



True Food Access

Misclassified

No Food Access

Our Poisson Outcome Model

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

Our Poisson 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 tract i given its food access at that radius (X_i) and whether it's metro or not (M_i)

Our Poisson Outcome Model

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

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

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

Our Poisson Outcome Model

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

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

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

add this to β_1 to adjust the prevalence ratio for a metro tract

Our Poisson Outcome Model

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

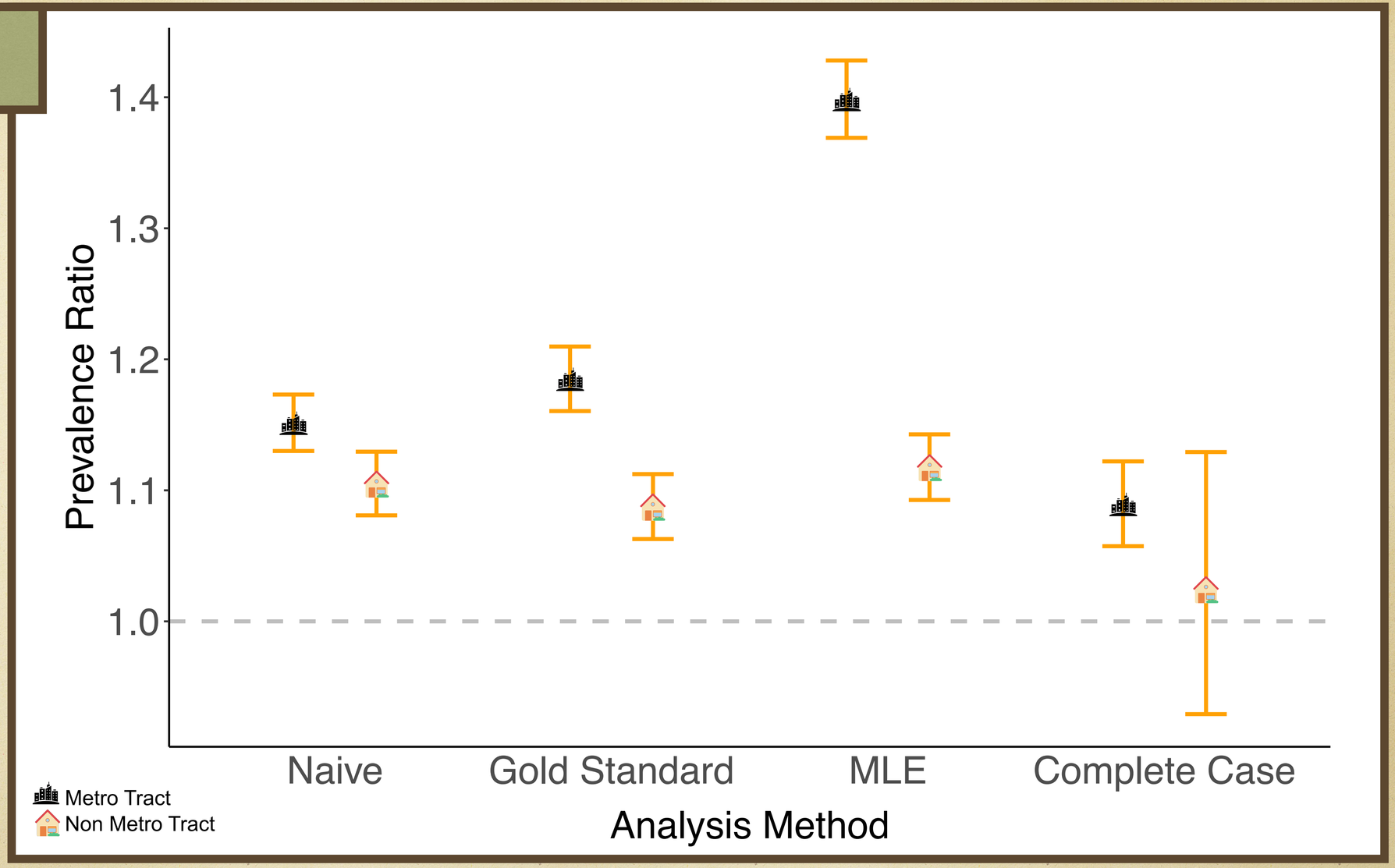
the offset turns case counts into prevalences

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

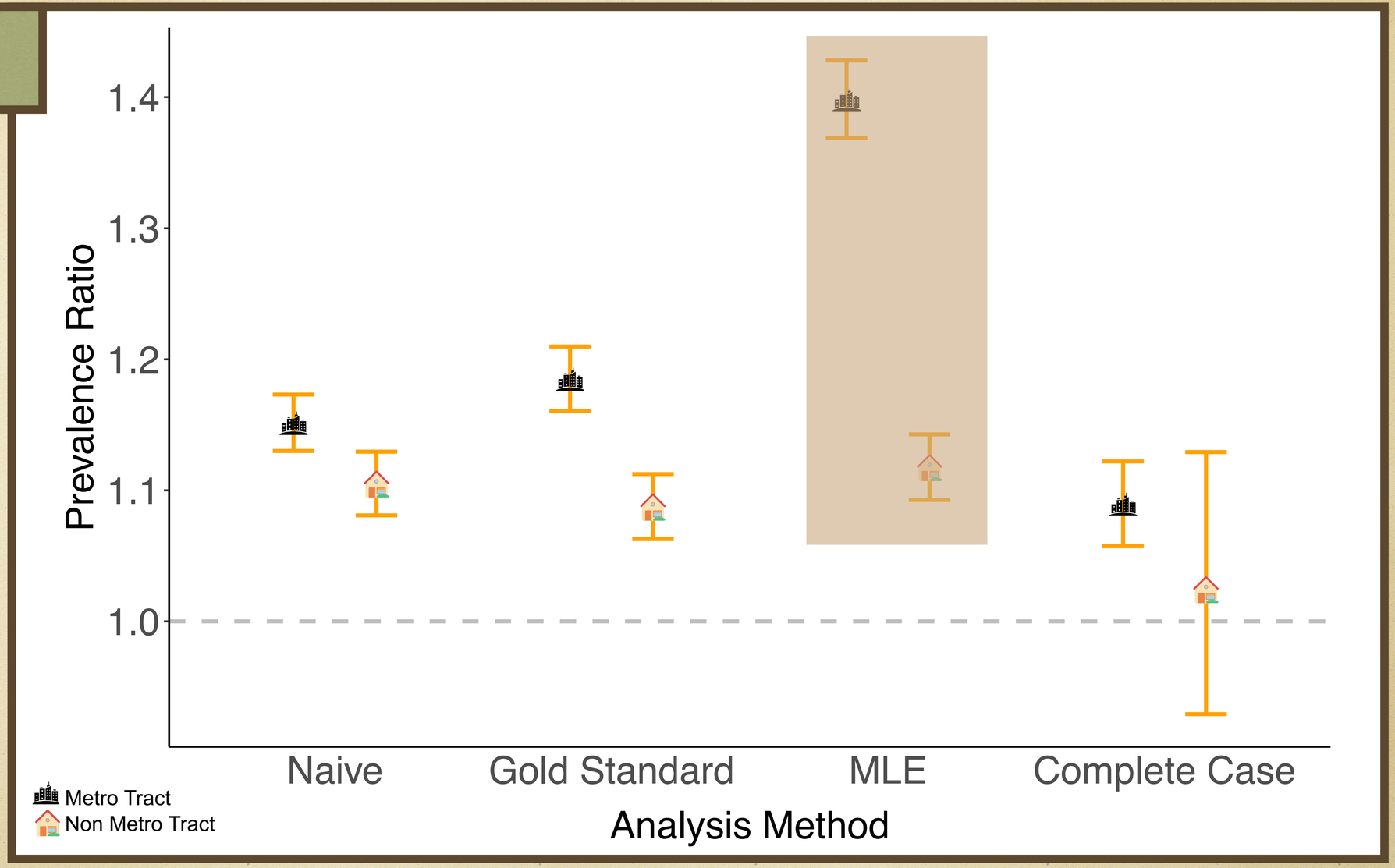
the expected number of diabetes cases (Y_i) in tract i given its food access at that radius (X_i) and whether it's metro or not (M_i)

add this to β_1 to adjust the prevalence ratio for a metro tract

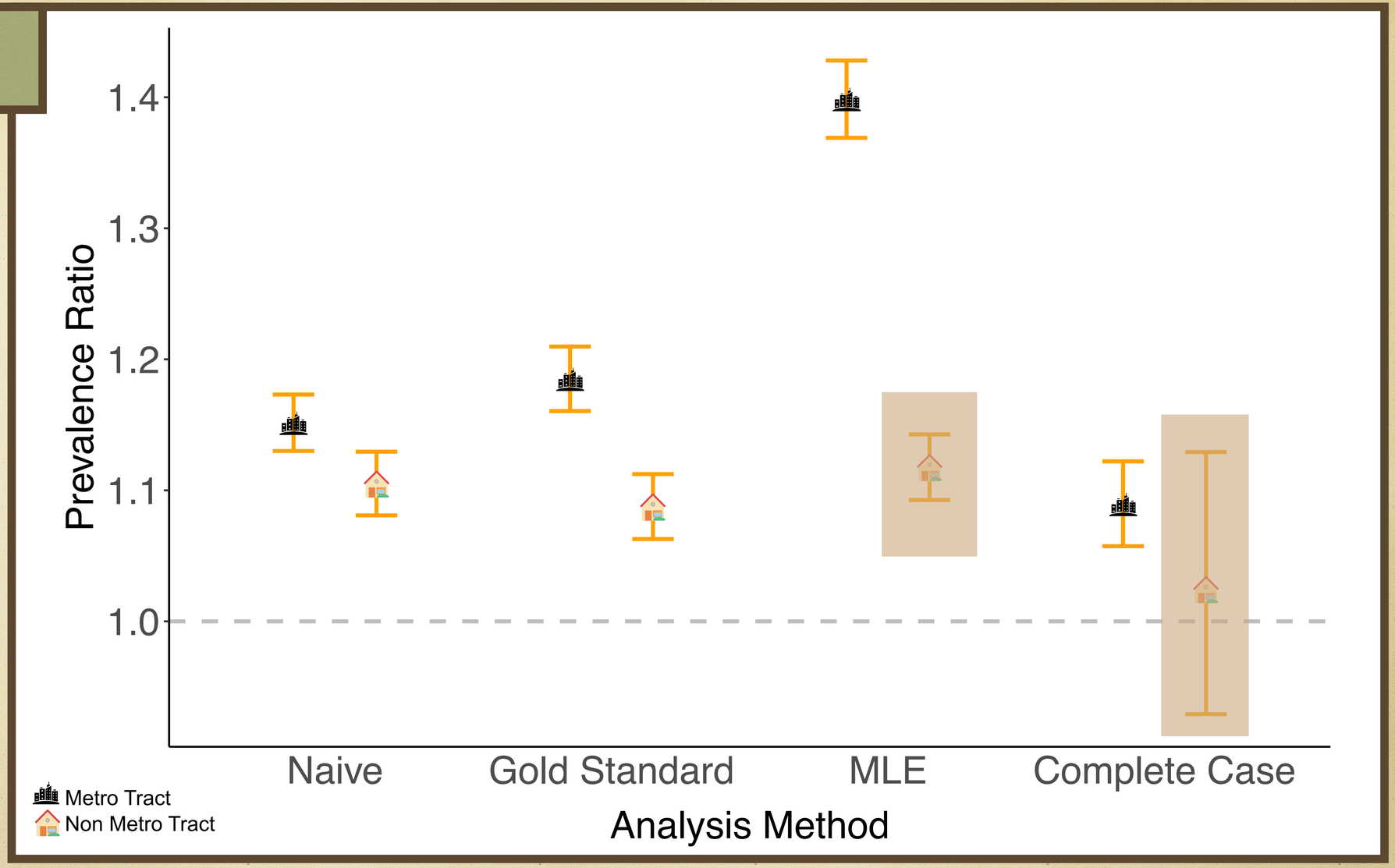
One Mile


1.4
1.3
1.2
1.1
1.0
Prevalence Ratio
Naive
Gold Standard
MLE
Complete Case
Analysis Method
Metro Tract
Non Metro Tract

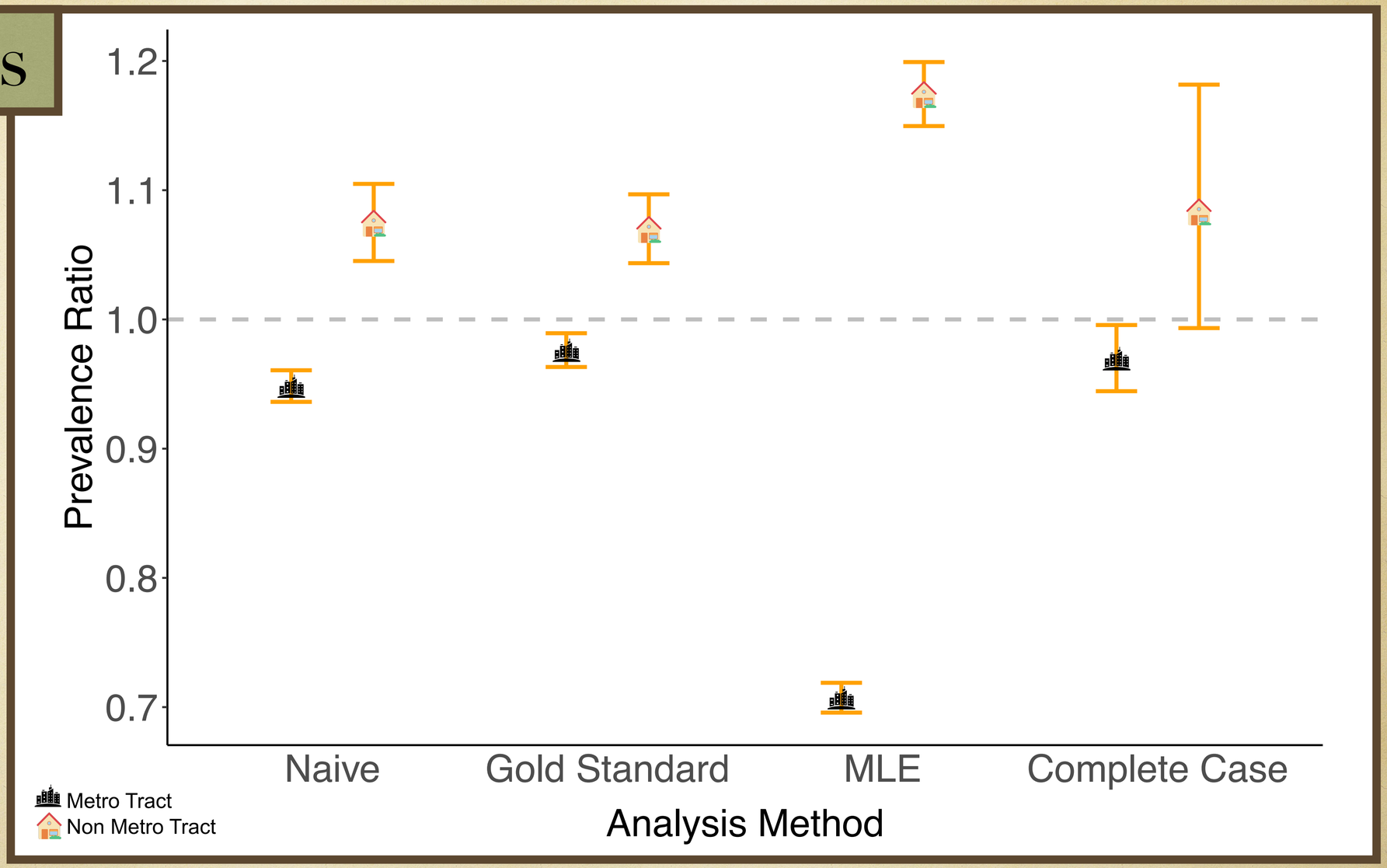
One Mile


Prevalence Ratio
1.4
1.3
1.2
1.1
1.0
Naive
Gold Standard
MLE
Complete Case
Analysis Method
Metro Tract
Non Metro Tract

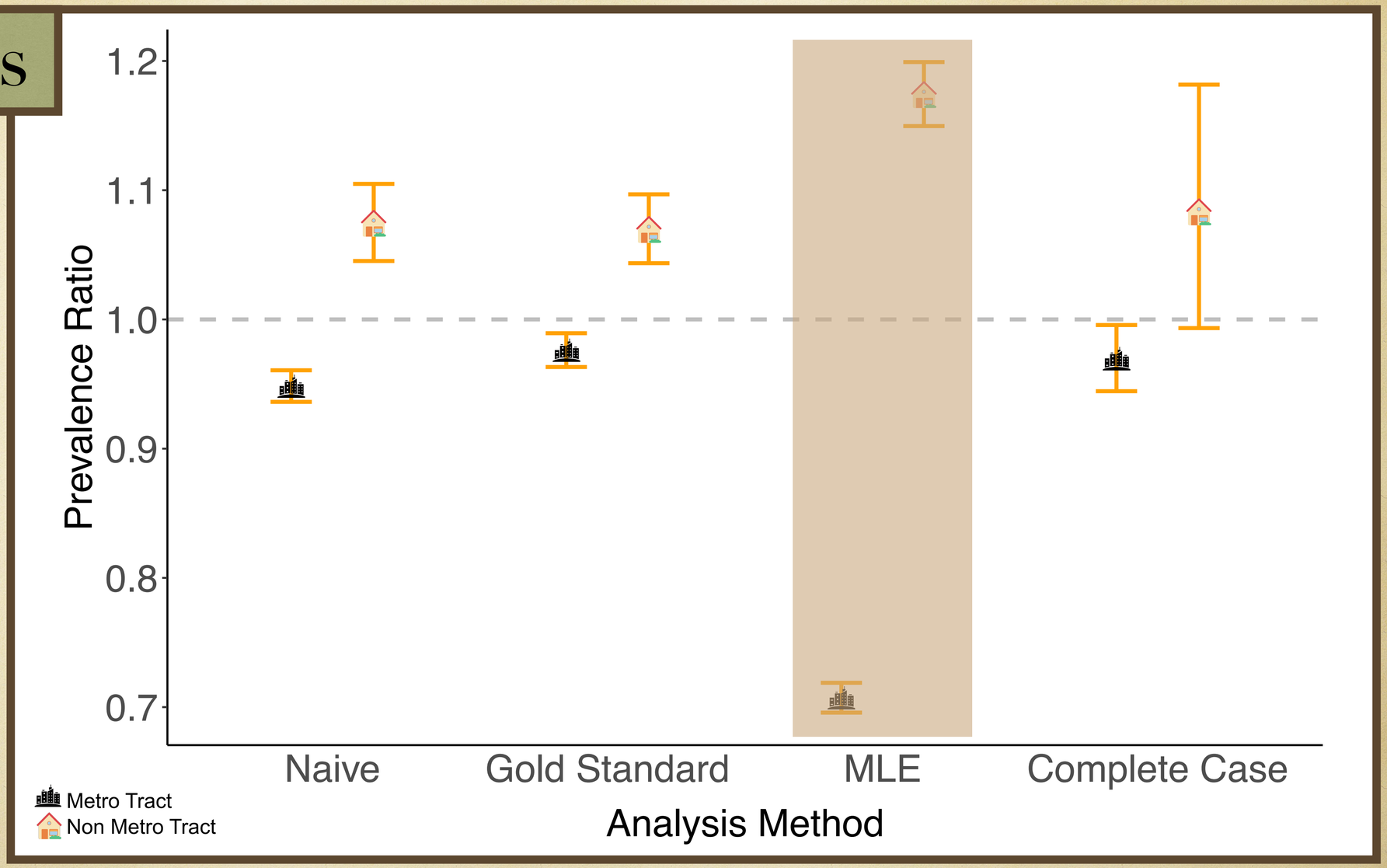
One Mile


1.4
1.3
1.2
1.1
1.0
Prevalence Ratio
Naive
Gold Standard
MLE
Complete Case
Analysis Method
Metro Tract
Non Metro Tract

Five Miles


1.2
1.1
1.0
0.9
0.8
0.7
Prevalence Ratio
Naive
Gold Standard
MLE
Complete Case
Analysis Method
Metro Tract
Non Metro Tract

Five Miles


1.2
1.1
1.0
0.9
0.8
0.7
Prevalence Ratio
Naive
Gold Standard
MLE
Complete Case
Analysis Method
Metro Tract
Non Metro Tract

Takeaways

From a bird's eye view, 

- The MLE avoids the heavy bias of the naive analysis and improves on the efficiency of the complete case analysis.
- Even with large enough sample sizes to fix issues in competitors, the MLE still wins!
- Tracts with **one mile food access** counterintuitively saw **higher** diabetes prevalences.
- Tracts with **five mile food access** had patterns dictated by **metro status**.
- The MLE model usually reported **stronger** effects more **efficiently** than the complete case.

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
- 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