Bayesian Unit-level Modeling of Categorical Survey Data with a Longitudinal Design

---

Daniel Vedensky[1]

Joint work with Paul A. Parker[2] and Scott H. Holan[13]

JSM – 2024

[1]University of Missouri
[2]University of California, Santa Cruz
[3]U.S. Census Bureau

Disclaimer

This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed on statistical issues are those of the authors and not those of the NSF or U.S. Census Bureau.

Introduction

Overview

Motivation: Unit-level modeling for small area estimation has received significant attention in recent years, but important methodological gaps remain with

- Longitudinal designs
  - Household Pulse Survey (HPS)
  - Survey of Income and Program Participation (SIPP)
  - National Crime Victimization Survey, etc.
- Categorical responses,
  - Nominal (e.g., Type of health insurance, housing tenure)
  - Ordinal (e.g., Likert scale, frequency of symptoms)
- While accounting for survey design and high dimensionality

So far,

- Parker et al. (2022) cross-sectional binary and nominal models
- Vedensky et al. (2023) longitudinal Gaussian and binary models[1]

Gaps to fill in today:

- Longitudinal nominal
- Cross-sectional ordinal
- Longitudinal ordinal

[1] https://arxiv.org/abs/2304.07897v1

Background

Longitudinal design

Household Pulse was launched to track changes at the start of COVID pandemic and employed a rotating panel design

Panel / week	Panel 1	Panel 2	Panel 3	Panel 4
Week 1	Initial Full Sample			
Week 2	Respondents from week 1 only	New Sample		
Week 3	Respondents from week 2 only	Respondents from week 2 only	New Sample	
Week 4		Respondents from week 3 only	Respondents from week 3 only	...
Week 5			⋮	⋮

Table 1: HPS panel structure (adapted from HPS documentation)

Need to account for

- "temporal" correlation across weeks
- "longitudinal" correlation within person/household's responses

It is common for dependence to exist between response variable and probability of selection.

- Important to adjust for *informative sampling* (IS) in a unit-level model otherwise bias may occur.
- Parker et al. (2023) provide an overview of the problems that arise, as well as remedies.
- One remedy is to use a *pseudo-likelihood* (Binder, 1983; Skinner, 1989; Savitsky and Toth, 2016)

$$\text{PL}(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{i \in \mathcal{S}} f(y_i|\boldsymbol{\theta})^{w_i}.$$

Starting point: cross-sectional, binary unit-level model with Bayesian pseudo-likelihood (Parker et al., 2022)

$$\begin{aligned}
\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\eta} &\propto \prod_{i\in S} \text{Binom}(y_i|n_i,p_i)^{\widetilde{w}_i} \\
\text{logit}(p_i) &= \boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{\psi}_i'\boldsymbol{\eta} \\
\boldsymbol{\eta}|\sigma_\eta^2 &\sim N_r(0_r, \sigma_\eta^2 \boldsymbol{I}_r) \\
\boldsymbol{\beta} &\sim N_q(0_q, \sigma_\beta^2 \boldsymbol{I}_q) \\
\sigma_\eta^2 &\sim \text{IG}(a,b) \\
\sigma_\beta, a, b &> 0.
\end{aligned}$$

- $\boldsymbol{\psi}_i$ may be an incidence vector or a spatial basis function.
- *Pólya-Gamma data augmentation* (Polson et al., 2013) leads to conjugacy

Nominal

Stick-breaking (Fienberg, 1980; Linderman et al., 2015)

For nominal responses with K unordered categories, can write

$$\text{Multinomial}((y_1, \ldots, y_K)|n, \boldsymbol{p}) = \prod_{k=1}^{K-1} \text{Binomial}(y_k|n_k, \tilde{p}_k),$$

where $n_k = n - \sum_{j<k} y_j$ and $\tilde{p}_k = p_k/(1 - \sum_{j<k} p_j)$.

That is, we can fit $K - 1$ binomial models then reconstruct $\boldsymbol{p}$ as

$$p_k = \tilde{p}_k \prod_{j<k} (1 - \tilde{p}_j).$$

Nominal unit-level model

(Parker et al., 2022)

A nominal model can then take the form

$$\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\eta} \propto \prod_{i\in S}\prod_{k=1}^{K-1} \text{Binom}(y_{ik}|n_{ik},p_{ik})^{\widetilde{w}_i}$$

$$\text{logit}(\tilde{p}_{ik}) = \boldsymbol{x}_i'\beta_k + \boldsymbol{\psi}_i'\boldsymbol{\eta}_k$$

$$\boldsymbol{\eta}_k|\sigma^2_{\eta k} \sim N_r(0_r, \sigma^2_\eta \boldsymbol{I}_r), \quad k = 1,\ldots,K-1$$

$$\boldsymbol{\beta}_k \sim N_q(0_q, \sigma^2_\beta \boldsymbol{I}_q), \quad k = 1,\ldots,K-1$$

$$\sigma^2_{\eta k} \sim \text{IG}(a,b), \quad k = 1,\ldots,K-1$$

$$\sigma_\beta, a, b > 0.$$

Sequential ordinal models

(Tutz, 1990; Albert and Chib, 2001)

For ordered categories

- take our latent variable $\boldsymbol{\mu}_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{\psi}_i'\boldsymbol{\eta}$
- introduce unordered cutpoints $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{K-1})$
- suppose

$$P(y_i = k | y_i \geq k, \boldsymbol{\gamma}, \boldsymbol{\mu}_i) = F(\gamma_k - \boldsymbol{\mu}_i).$$

Then

$$P(y_i = k | \gamma_k, \boldsymbol{\mu}_i) = F(\gamma_k - \boldsymbol{\mu}_i) \prod_{j<k} (1 - F(\gamma_j - \boldsymbol{\mu}_i)).$$

Sequential ordinal models

(Tutz, 1990; Albert and Chib, 2001)

For ordered categories

- take our latent variable $\boldsymbol{\mu}_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{\psi}_i'\boldsymbol{\eta}$
- introduce unordered cutpoints $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{K-1})$
- suppose

$$P(y_i = k | y_i \geq k, \boldsymbol{\gamma}, \boldsymbol{\mu}_i) = F(\gamma_k - \boldsymbol{\mu}_i).$$

Then

$$\underbrace{P(y_i = k | \gamma_k, \boldsymbol{\mu}_i)}_{p_k} = \underbrace{F(\gamma_k - \boldsymbol{\mu}_i)}_{\tilde{p}_k} \prod_{j<k} (\underbrace{1 - F(\gamma_j - \boldsymbol{\mu}_i)}_{1-\tilde{p}_j}).$$

Ordinal unit-level model

Cross-sectional

Taking $F(\cdot) = \text{logit}^{-1}(\cdot)$ and placing a prior on $\boldsymbol{\gamma}$ leads to

$$\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\gamma} \propto \prod_{i \in \mathcal{S}} \prod_{k=1}^{K-1} \text{Binom}(y_{ik}|n_{ik}, \tilde{p}_{ik})^{\tilde{w}_i}$$

$$\text{logit}(\tilde{p}_{ik}) = \gamma_k - \boldsymbol{x}_i'\boldsymbol{\beta} - \boldsymbol{\psi}_i\boldsymbol{\eta}$$

$$\boldsymbol{\eta} \sim \text{N}_r(0, \sigma_\eta^2 I_r)$$

$$\boldsymbol{\beta} \sim \text{N}_q(0, \sigma_\beta^2 I_q)$$

$$\boldsymbol{\gamma} \sim \text{N}_{K-1}(0, \sigma_\gamma^2 I_{K-1})$$

$$\sigma_\eta^2 \sim \text{IG}(a, b)$$

- Pólya-Gamma data augmentation again yields conjugacy
- If $\tilde{w}_i = 1$ for all i, this is Bayesian ordinal logistic regression with a fully Gibbs sampler

Ordinal unit-level model

With time dependence

To capture time dependence, add AR(1) structure to random effects

$$\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\eta},\boldsymbol{\gamma} \propto \prod_{t=1}^{T}\prod_{i\in\mathcal{S}_t}\prod_{k=1}^{K-1}\text{Binom}(y_{itk}|n_{itk},\tilde{p}_{itk})^{\tilde{w}_{it}}$$

$$\text{logit}(\tilde{p}_{itk}) = \boldsymbol{c}_{it}'\boldsymbol{\gamma}_k - \boldsymbol{x}_i'\boldsymbol{\beta} - \boldsymbol{\psi}_i\boldsymbol{\eta}_t$$

$$\boldsymbol{\gamma} \sim \text{N}_g(0,\sigma_\gamma^2 I_g)$$

$$\boldsymbol{\eta}_t|\boldsymbol{\eta}_{t-1},\phi,\sigma_\eta^2 \sim \text{N}_r(\phi\boldsymbol{\eta}_{t-1},\sigma_\eta^2 I_r),\quad t=2,\ldots,T$$

$$\boldsymbol{\eta}_1|\sigma_\eta^2 \sim \text{N}_r(0,\sigma_\eta^2 I_r)$$

$$\boldsymbol{\beta} \sim \text{N}_q(0,\sigma_\beta^2 I_q)$$

$$\phi \sim \text{Unif}(-1,1)$$

$$\sigma_\eta^2 \sim \text{IG}(a,b).$$

Ordinal unit-level model

With time dependence

To capture time dependence, add AR(1) structure to random effects

$$\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\eta},\boldsymbol{\gamma} \propto \prod_{t=1}^{T}\prod_{i\in\mathcal{S}_t}\prod_{k=1}^{K-1}\text{Binom}(y_{itk}|n_{itk},\tilde{p}_{itk})^{\tilde{w}_{it}}$$

$$\text{logit}(\tilde{p}_{itk}) = \boldsymbol{c}_{it}'\boldsymbol{\gamma}_k - \boldsymbol{x}_i'\boldsymbol{\beta} - \boldsymbol{\psi}_i\boldsymbol{\eta}_t$$

$$\boldsymbol{\gamma} \sim \text{N}_g(0,\sigma_\gamma^2 I_g)$$

$$\boldsymbol{\eta}_t|\boldsymbol{\eta}_{t-1},\phi,\sigma_\eta^2 \sim \text{N}_r(\phi\boldsymbol{\eta}_{t-1},\sigma_\eta^2 I_r),\quad t=2,\ldots,T$$

$$\boldsymbol{\eta}_1|\sigma_\eta^2 \sim \text{N}_r(0,\sigma_\eta^2 I_r)$$

- Cutpoints vary with time
- Covariate with $K+1$ levels indexes whether prev. response = NA, $1,\ldots,K$

Nominal unit-level model

With time dependence

Similarly for the nominal model

$$\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\eta} \propto \prod_{t=1}^{T}\prod_{i\in\mathcal{S}_t}\prod_{k=1}^{K-1} \text{Binomial}(y_{itk}|n_{itk},\tilde{p}_{itk})^{\widetilde{w}_{it}}$$

$$\text{logit}(\widetilde{p}_{itk}) = \boldsymbol{x}_i'\boldsymbol{\beta}_k + \boldsymbol{\psi}_i'\boldsymbol{\eta}_{tk}$$

$$\boldsymbol{\eta}_{tk}|\boldsymbol{\eta}_{(t-1)k},\phi_k,\sigma^2_{\eta k} \sim \text{N}_m(\phi_k\boldsymbol{\eta}_{(t-1)k},\sigma^2_{\eta k}I_m), \quad k=1,\ldots,K-1$$

$$\boldsymbol{\eta}_{1k}|\sigma^2_{\eta_1 k} \sim \text{N}_m(0,\sigma^2_{\eta_1 k}I_m), \quad k=1,\ldots,K-1$$

$$\boldsymbol{\beta}_k \sim \text{N}_q(0,\sigma^2_\beta I_q), \quad k=1,\ldots,K-1$$

$$\phi_k \sim \text{Unif}(-1,1)$$

$$\sigma^2_{\eta k}, \overset{ind.}{\sim} IG(a,b)$$

$$\sigma_\beta, a, b > 0.$$

Empirical simulation

Setup

To assess the proposed models, we conduct an empirical simulation study

- Treat entire HPS Phase 1 data as population
- Take 100 informative subsamples
- Fit models to each sample and generate predictions for population
- Compare direct estimates, cross-sectional, and longitudinal models

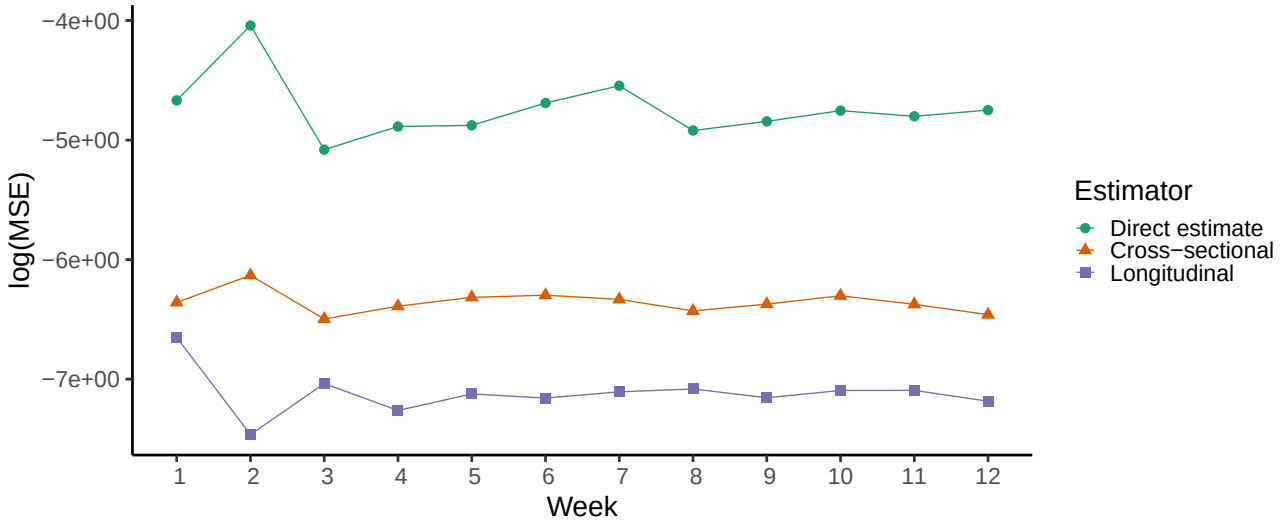
Empirical simulation

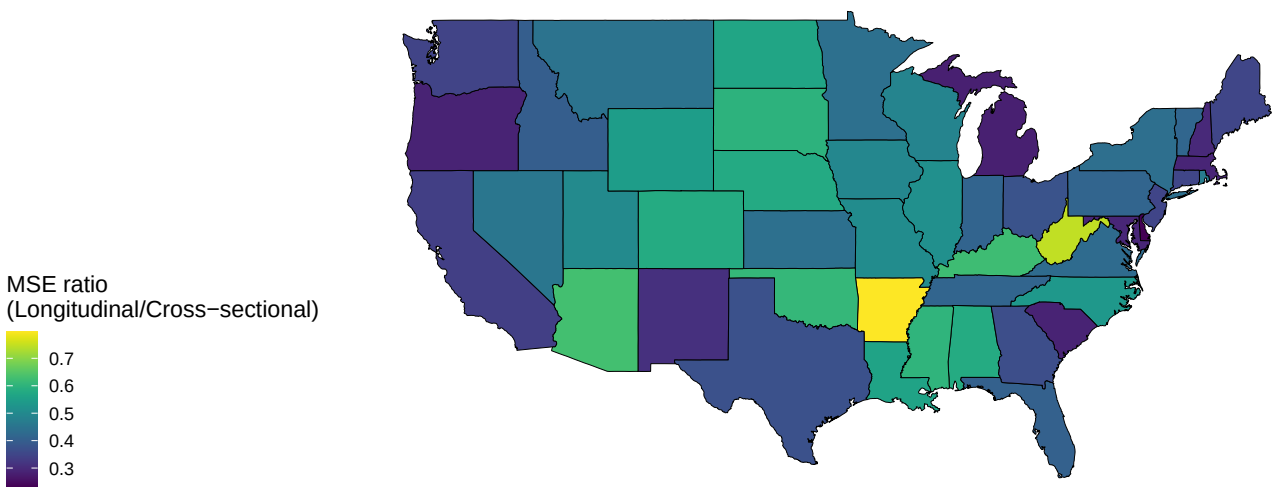
Setup

- Nominal response: "Is your house or apartment..."
  1. Owned free and clear?
  2. Owned with a mortgage or loan?
  3. Rented?
  4. Occupied without rent?
- Ordinal Response: Frequency of anxiety over previous 7 days...
  1. Not at all
  2. Several days
  3. More than half the days
  4. Nearly every day
- Covariates: sex, race, age, prev. response
- Take *probability proportional to size* sample with expected size 2% of the population ($N \approx 10^6$, both cases)

Empirical simulation

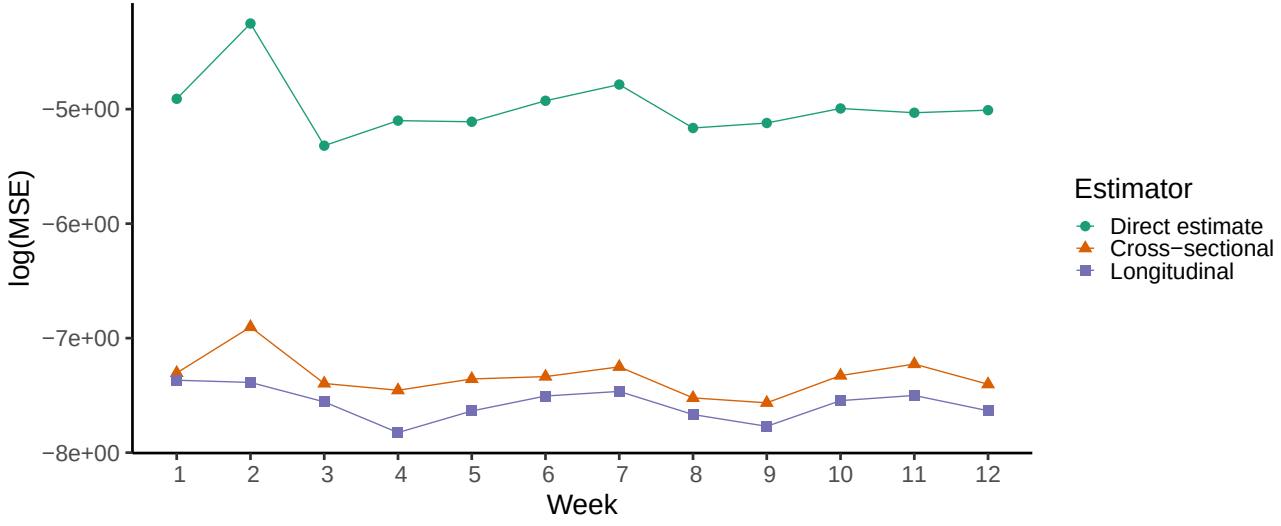
Nominal results

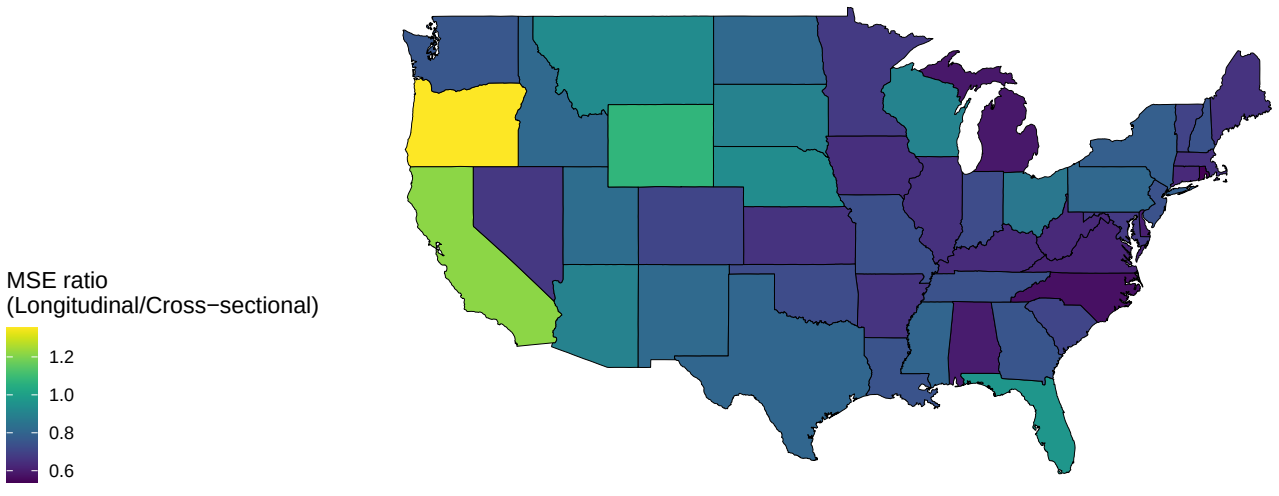




Empirical simulation

Ordinal results





Data analysis

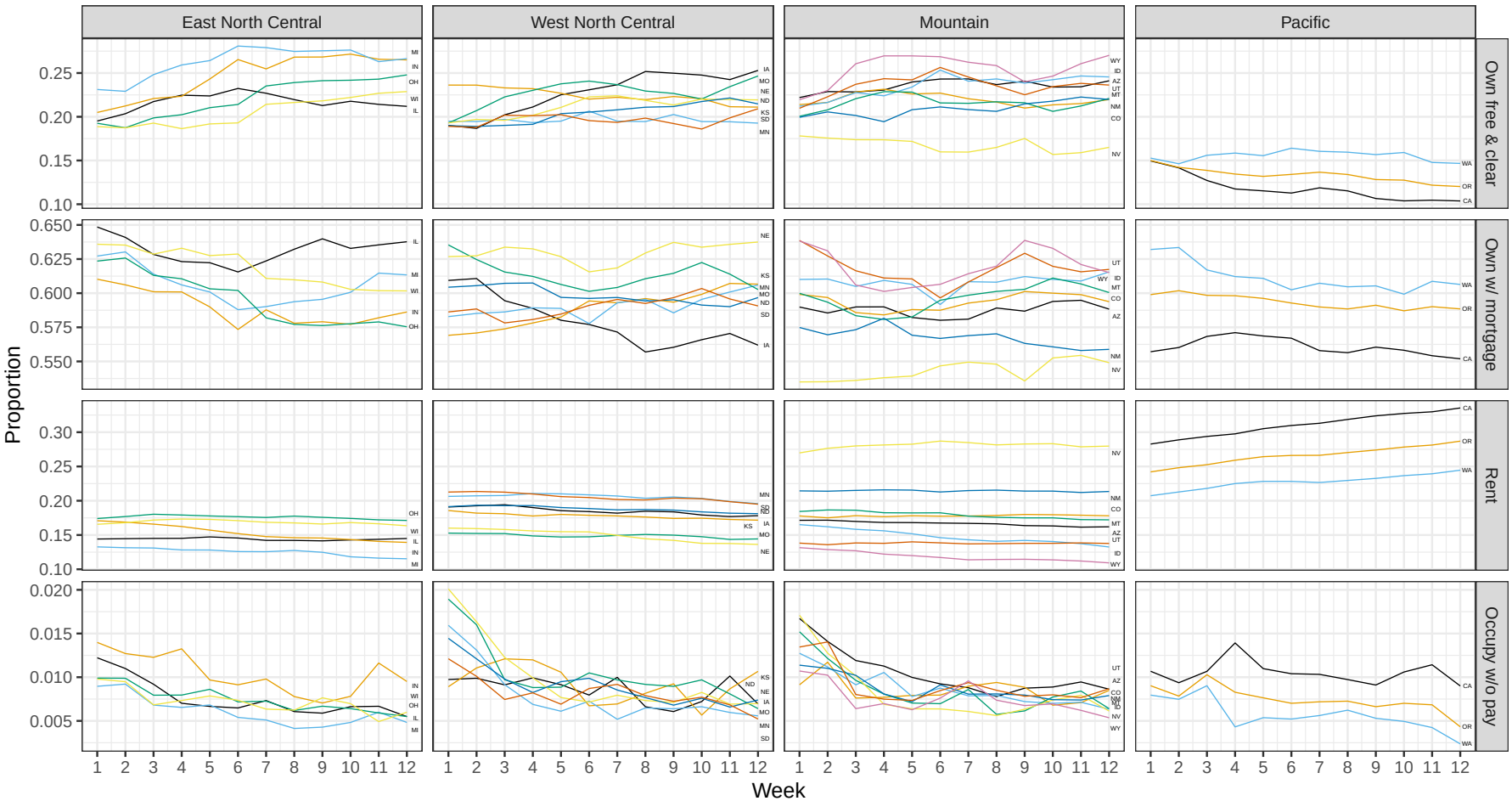
Setup

Fit our nominal model to the full HPS Phase 1 data

- Housing tenure type as response value
- Same covariates as before
- Use population-level covariates from Census Population Estimates program to generate posterior predictions

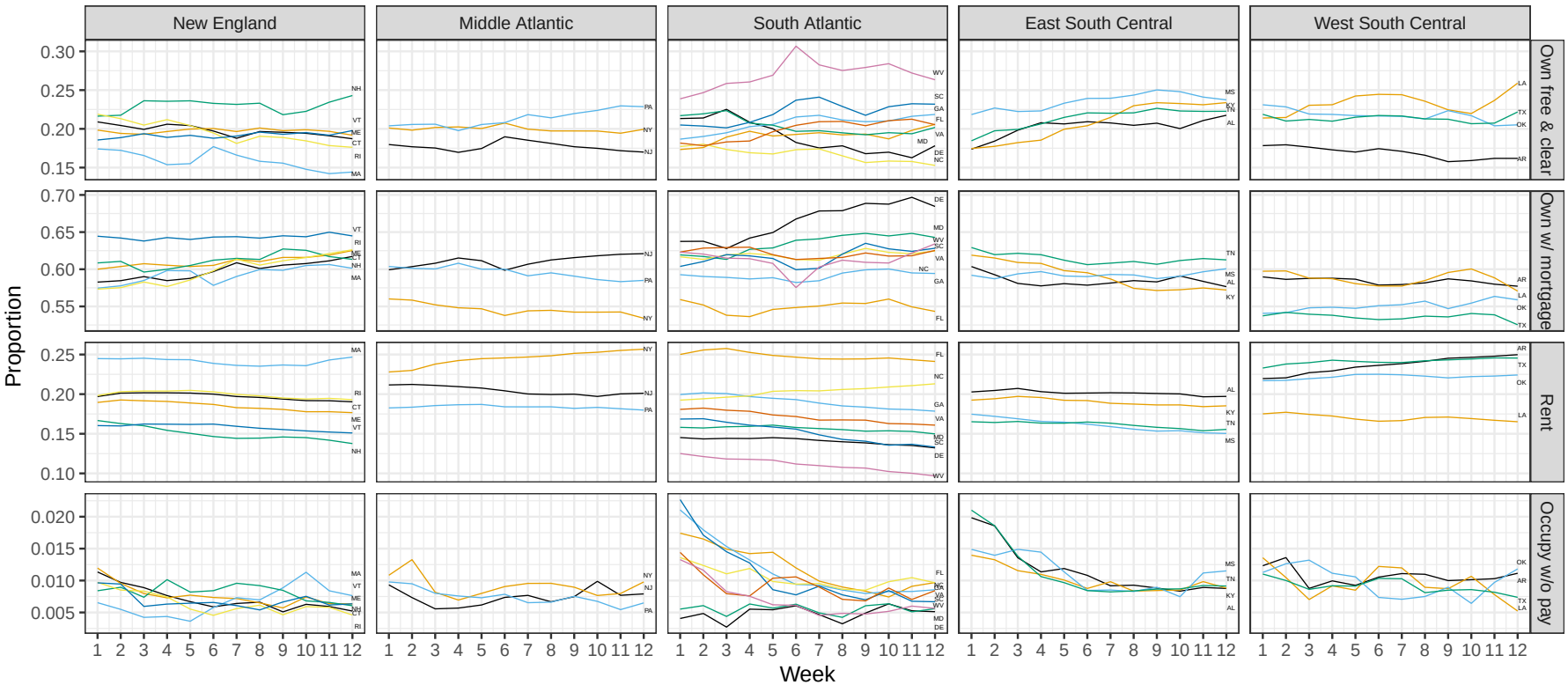
Data analysis

Estimated housing tenure, Asian males aged 45 to 50 by Census div.



Data analysis

Estimated housing tenure, Asian males aged 45 to 50 by Census div.



Summary

We propose unit-level models for nominal and ordinal data that

- account for longitudinal design
- account for informative sampling
- are computationally efficient

We illustrate their performance via

- an empirical simulation study
- and a full data analysis

Thank you!

dvedensky@mail.missouri.edu

References

Albert, J. H. and Chib, S. (1993). "Bayesian Analysis of Binary and Polychotomous Response Data." *Journal of the American Statistical Association*, 88, 422, 669–679.

— (2001). "Sequential ordinal modeling with applications to survival data." *Biometrics*, 57, 3, 829–836.

Binder, D. A. (1983). "On the Variances of Asymptotically Normal Estimators from Complex Surveys." *International Statistical Review*, 51, 3, 279.

Fienberg, S. E. (1980). "The analysis of cross-classified categorical data." *Massachusetts Institute of Technology Press, Cambridge and London*.

Linderman, S., Johnson, M. J., and Adams, R. P. (2015). "Dependent Multinomial Models Made Easy: Stick-Breaking with the Polya-gamma Augmentation." In *Advances in Neural Information Processing Systems*, eds. C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, vol. 28. Curran Associates, Inc.

McCullagh, P. (1980). "Regression models for ordinal data." *Journal of the Royal Statistical Society: Series B (Methodological)*, 42, 2, 109–127.

Parker, P. A., Holan, S. H., and Janicki, R. (2022). "Computationally Efficient Bayesian Unit-level Models for Non-Gaussian Data Under Informative Sampling with Application to Estimation of Health Insurance Coverage." *The Annals of Applied Statistics*, 16, 2, 887 – 904.

Parker, P. A., Janicki, R., and Holan, S. H. (2023). "A Comprehensive Overview of Unit-Level Modeling of Survey Data for Small Area Estimation Under Informative Sampling." *Journal of Survey Statistics and Methodology*.

Polson, N. G., Scott, J. G., and Windle, J. (2013). "Bayesian Inference for Logistic Models Using Pólya–Gamma Latent Variables." *Journal of the American Statistical Association*, 108, 504, 1339–1349.

Savitsky, T. D. and Toth, D. (2016). "Bayesian Estimation Under Informative Sampling." *Electronic Journal of Statistics*, 10, 1, 1677 – 1708.

Skinner, C. J. (1989). "Domain Means, Regression and Multivariate Analysis." In *Analysis of Complex Surveys*, eds. C. J. Skinner, D. Holt, and T. M. F. Smith, chap. 2, 59–84. Chichester: Wiley.

Tutz, G. (1990). "Sequential item response models with an ordered response." *British Journal of Mathematical and Statistical Psychology*, 43, 1, 39–55.

Vedensky, D., Parker, P. A., and Holan, S. H. (2023). "Bayesian Unit-level Models for Longitudinal Survey Data under Informative Sampling: An Analysis of Expected Job Loss Using the Household Pulse Survey."

Cumulative ordinal models

McCullagh (1980); Albert and Chib (1993)

Suppose

$$P(y_i \leq k|\boldsymbol{\gamma}, \boldsymbol{\mu}) = F(\gamma_k - \boldsymbol{\mu}_i)$$

which implies

$$P(y_i = k|\boldsymbol{\gamma}, \boldsymbol{\mu}) = F(\gamma_k - \boldsymbol{\mu}_i) - F(\gamma_{k-1} - \boldsymbol{\mu}_i)$$

or can be viewed as

$$y_i = k \text{ if } \gamma_{k-1} < \mu_i < \gamma_k.$$

Sequential ordinal models

Assuming

$$P(y_i = k | y_i \geq j, \beta, \eta, \gamma) = F(\gamma_k - \mu_i)$$

implies

$$P(y_i = k | \beta, \eta, \gamma) = F(\gamma_k - \mu_i) \prod_{j=1}^{k-1} (1 - F(\gamma_j - \mu_i)).$$

Alternatively, this can be viewed as saying

$$y_i = \begin{cases} 0 & \text{if } \mu_i \leq \gamma_1 \\ 1 & \text{if } \mu_i' > \gamma_1, \mu_i \leq \gamma_2 \\ \vdots & \vdots \\ K-1 & \text{if } \mu_i > \gamma_1, \mu_i > \gamma_2, \ldots, \mu_i > \gamma_{K-2}, \mu_i \leq \gamma_{K-1} \\ K & \text{if } \mu_i > \gamma_1, \mu_i > \gamma_2, \ldots, \mu_i > \gamma_{K-1} \end{cases}$$