Improving multilevel regression and poststratification with structured priors

Alex Yuxiang Gao Lauren Kennedy Daniel Simpson

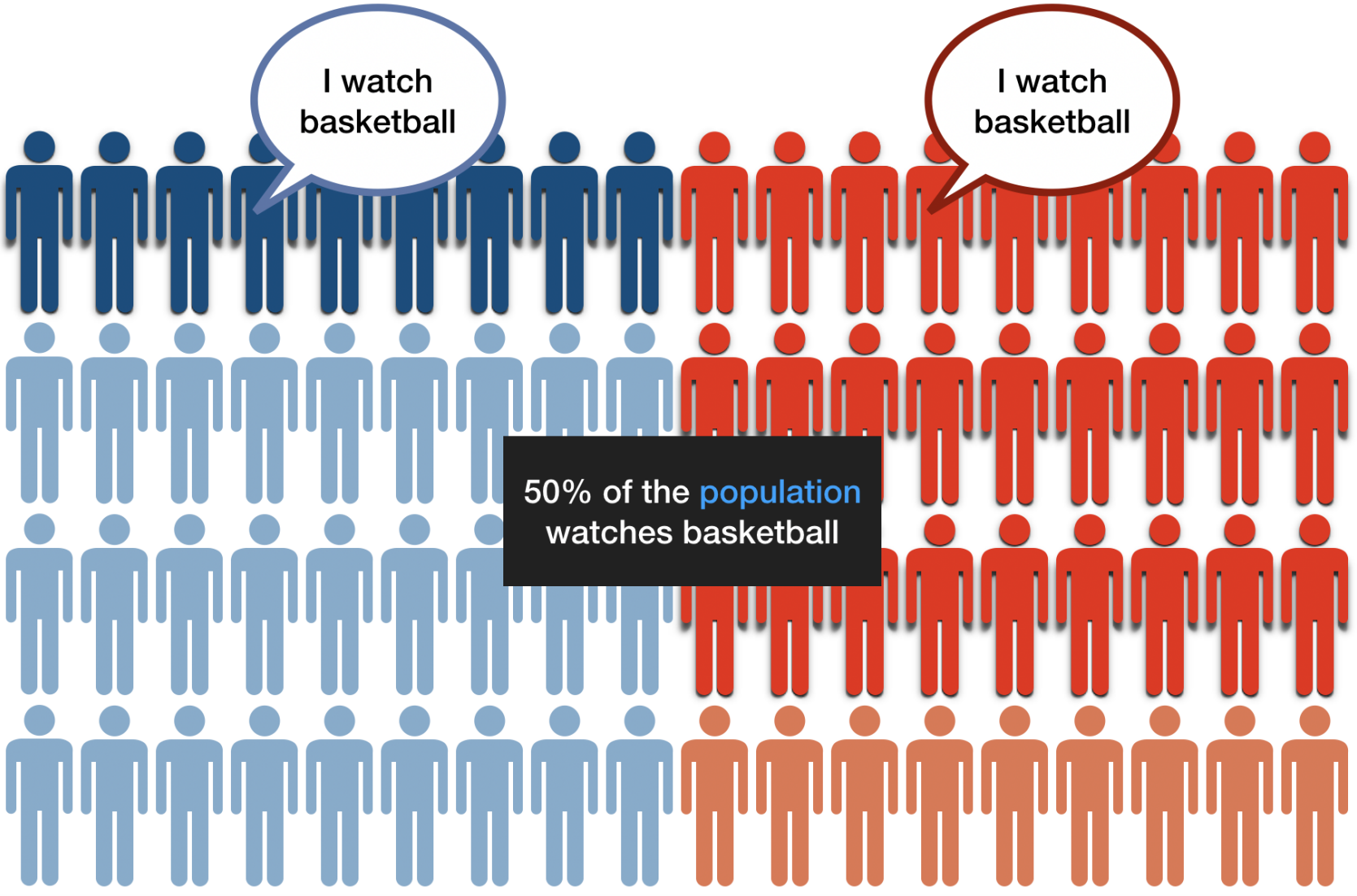
Department of Statistical Sciences, University of Toronto

School of Social Work, Columbia University

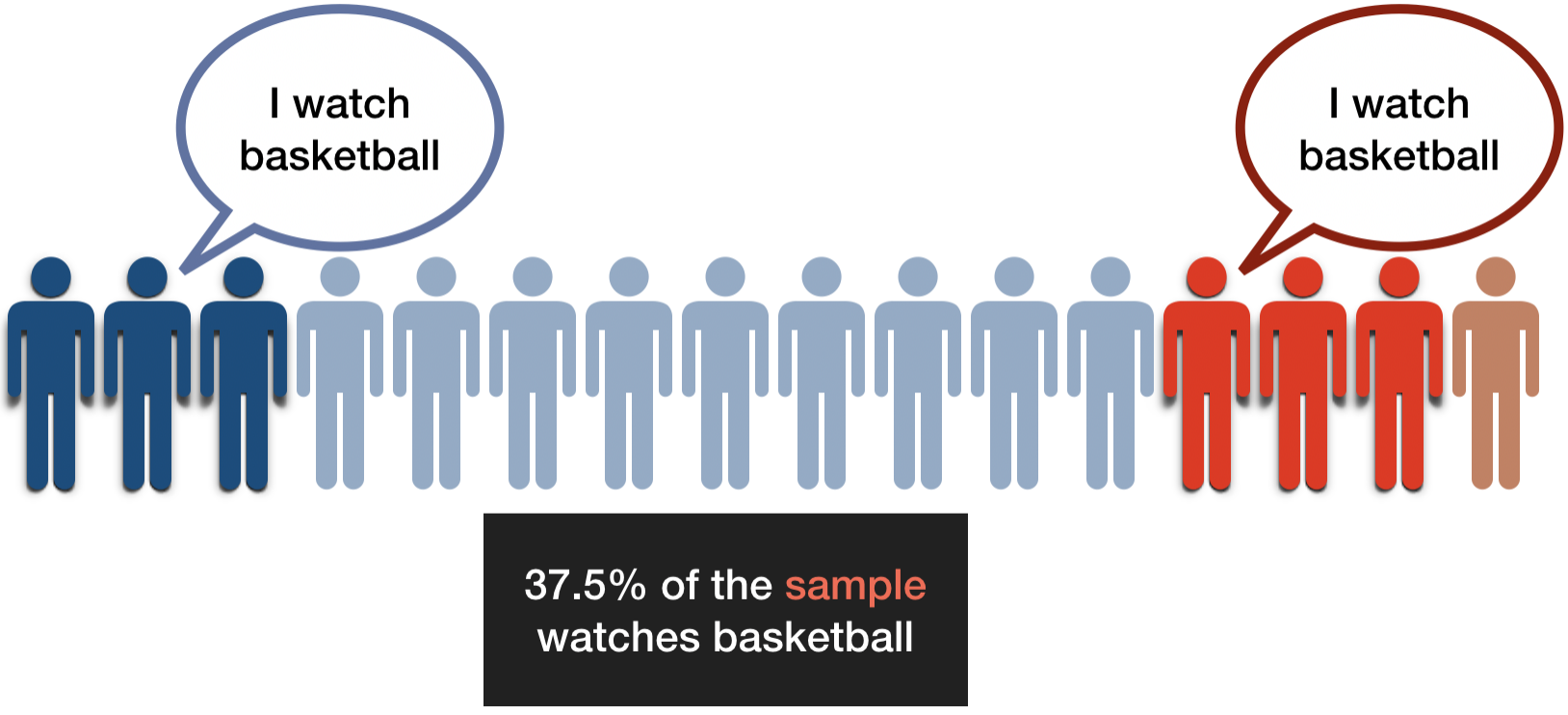
Roadmap for today's talk

1. Toy example of a non-representative survey and why it's an important issue
2. Multilevel Regression and Poststratification (MRP), a method for estimating sub-population level values based on survey data
3. Our proposed improvement for MRP reduces posterior absolute bias and variance via structured priors
4. Simulation studies based on structured priors
5. Application to real non-representative survey data

Toy example of a non-representative survey



Toy example of a non-representative survey



- When people from the orange party have a response rate lower than the blue party's response rate, you get a biased estimate of the total population preference.

Multilevel Regression and Poststratification (MRP) for sub-population level survey estimates

- MRP is a framework for estimating sub-population and the overall population level preference.
- We will assume that the population of interest is summarized by two discrete covariates, age category and income category: $\mathcal{D} := \{Y_i, X_{\text{Age},i}, X_{\text{Income},i}\}_{i=1}^{n}$ is our survey, where
  - $Y_i \in \{0, 1\}$ is the binary response for individual i
  - $X_{\text{Age},i} \in \{1, \ldots, J\}$ is the stratified age category that individual i is in
  - $X_{\text{Income},i} \in \{1, \ldots, K\}$ is the stratified income category that individual i is in
- Every category $c \in \{1, \ldots, J \times K\}$ in the population will have size N_c.

Multilevel Regression and Poststratification (MRP) for sub-population level survey estimates

Fit the hierarchical logistic regression model (MR):

$$\Pr(y_i = 1) = \text{logit}^{-1}\left(\beta_0 + \alpha_{j[i]}^{\text{Age Cat.}} + \alpha_{j[i]}^{\text{Income}}.\right), \text{ for } i = 1, \ldots, n.$$

$$\alpha_j^{\text{Age}} \mid \sigma^{\text{Age}} \overset{\text{ind.}}{\sim} \mathcal{N}(0, (\sigma_{\text{Age}})^2), \text{ for } j = 1, \ldots, J$$

$$\alpha_k^{\text{Income}} \mid \sigma^{\text{Income}} \overset{\text{ind.}}{\sim} \mathcal{N}(0, (\sigma^{\text{Income}})^2), \text{ for } k = 1, \ldots, K$$

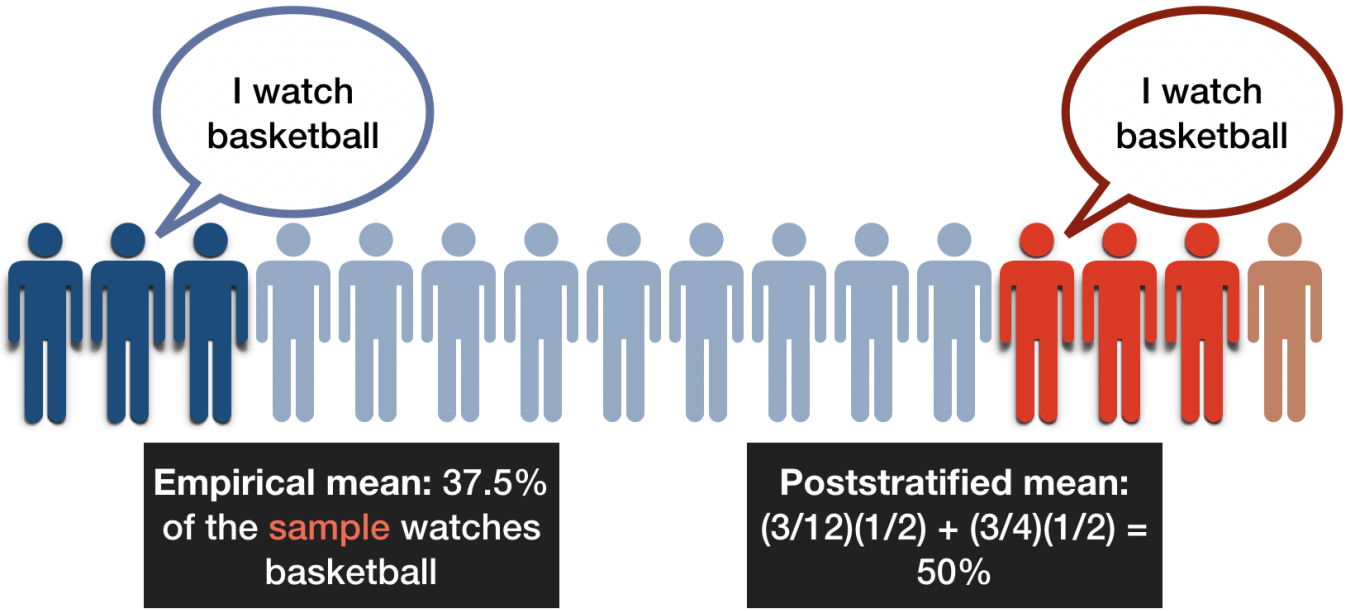
$$\sigma^{\text{Age}} \sim \text{N}_+(0, 1), \ \sigma^{\text{Income}} \sim \text{N}_+(0, 1), \ \beta^0 \sim \text{N}(0, 1)$$

Every category $c \in \{1, \ldots, J \times K\}$ will have a posterior preference probability θ_c after fitting the above model.

Multilevel Regression and Poststratification (MRP) for sub-population level survey estimates

Poststratify to get the posterior preference for the sub-population group $S \subseteq \{1, \ldots, J \times K\}$ (P):

$$\theta_O := \frac{\sum_{c \in S} \theta_c N_c}{\sum_{c \in S} N_c}$$

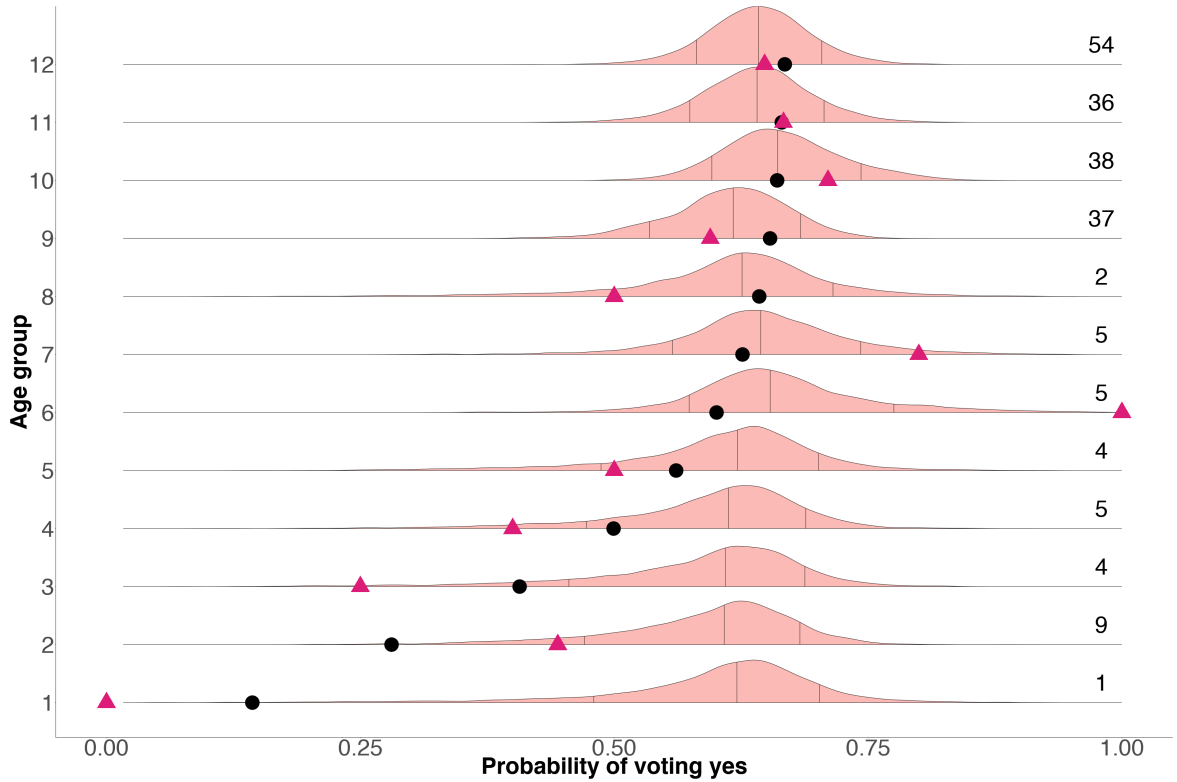


Problems with the current MRP setup

- The current setup for MRP has random effects for the age covariate and the income covariate shrinking posterior estimates towards the global mean β_0.
- Shrinking estimates toward β_0 is not ideal when one has non-representative samples.
- **Applying MRP to non-representative samples:** Scenario: True preference for voting yes in a population goes up when age of an individual goes up:
  - Case 1: Older adults in the population are undersampled
  - Case 2: Older adults in the population are oversampled

Non-representative samples: Oversampling older adults

Figure 1: Posterior preference, where probability of sampling individuals in age groups 9-12 is 0.82.



- Sample size $n = 200$.
- Black dots are the true preference.
- Red triangles are the empirical preference.

Can we make MRP better?

- So now we that shrinking Posterior estimates to the global mean β_0 is not ideal in the presence of non-representatives samples.
- A lot of bias is introduced in the Posteriors that are under/over-sampled.
- Can we do better? Yes!

- The current problem is that Posterior estimates for every age are borrowing information from the global mean.
- We actually want Posterior estimates for every age to borrow information from their neighbouring ages.
- Structured priors in multilevel regression will introduce shrinkage towards neighbours!

Structured Priors for MRP

- Covariates in more complex populations will have more complex structures (graph structures, spatial structures e.t.c)
- In our case, the Age covariate has an ordinal structure to it.
- Instead of having $\alpha_j^{\text{Age}} \overset{\text{ind.}}{\sim} \mathcal{N}(0, (\sigma^{\text{Age}})^2)$ for $j \in \{1, \ldots, J\}$, we have the following prior specification for Age:

$$\alpha_1 \sim \mathcal{N}(0, \frac{(\sigma^{\text{Age}})^2}{1-\rho^2})$$

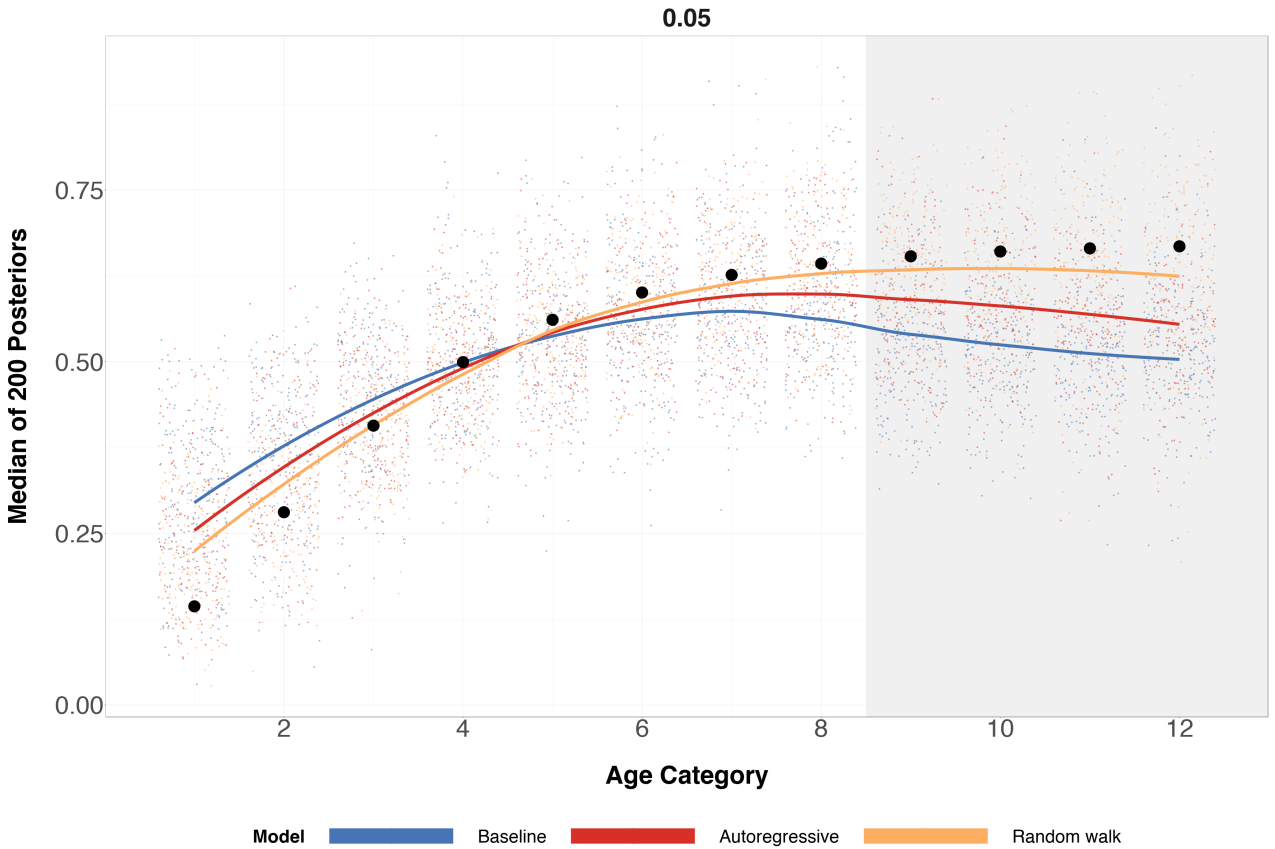
$$\alpha_j^{\text{Age}} = \rho \alpha_{j-1}^{\text{Age}} + \epsilon_j, \ \epsilon_j \overset{\text{ind.}}{\sim} \mathcal{N}(0, (\sigma^{\text{Age}})^2), \text{ for } j = 2, \ldots, J$$

$$2(\rho + 1) \sim \text{Beta}(0.5, 0.5)$$

- This is a first-order autoregressive process for the random effect on Age. In the case that $\rho = 1$, we have a first-order random walk process.
- We should expect posterior estimates of every age to shrink towards posterior estimates of the previous age from this specification, thus reducing posterior bias and variance.

Simulation results: Undersampling older adults

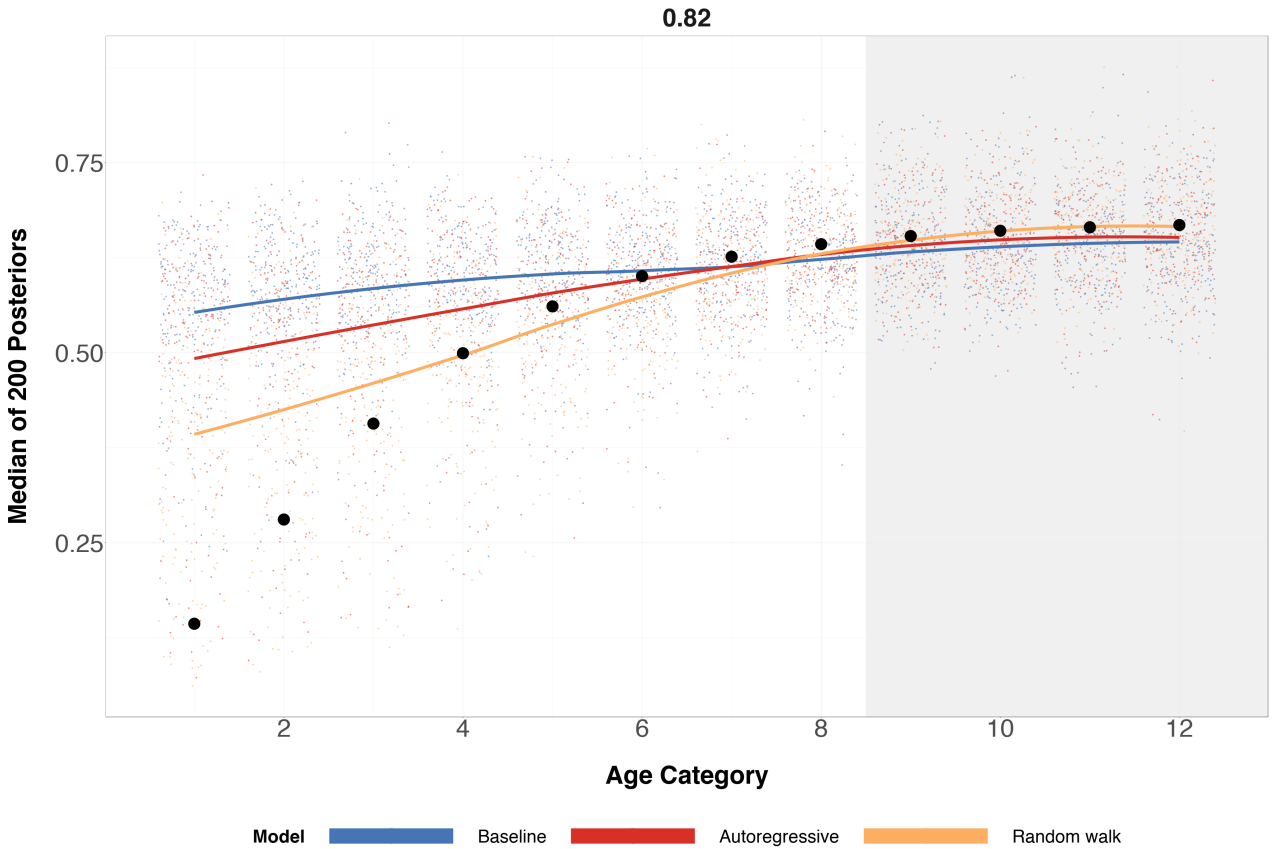
Figure 2: Posterior medians for 200 simulation runs, where the probability of sampling individuals in age groups 9-12 is 0.05.



- Sample size $n = 100$.
- Black dots are the true preference for every age.

Simulation results: Oversampling older adults

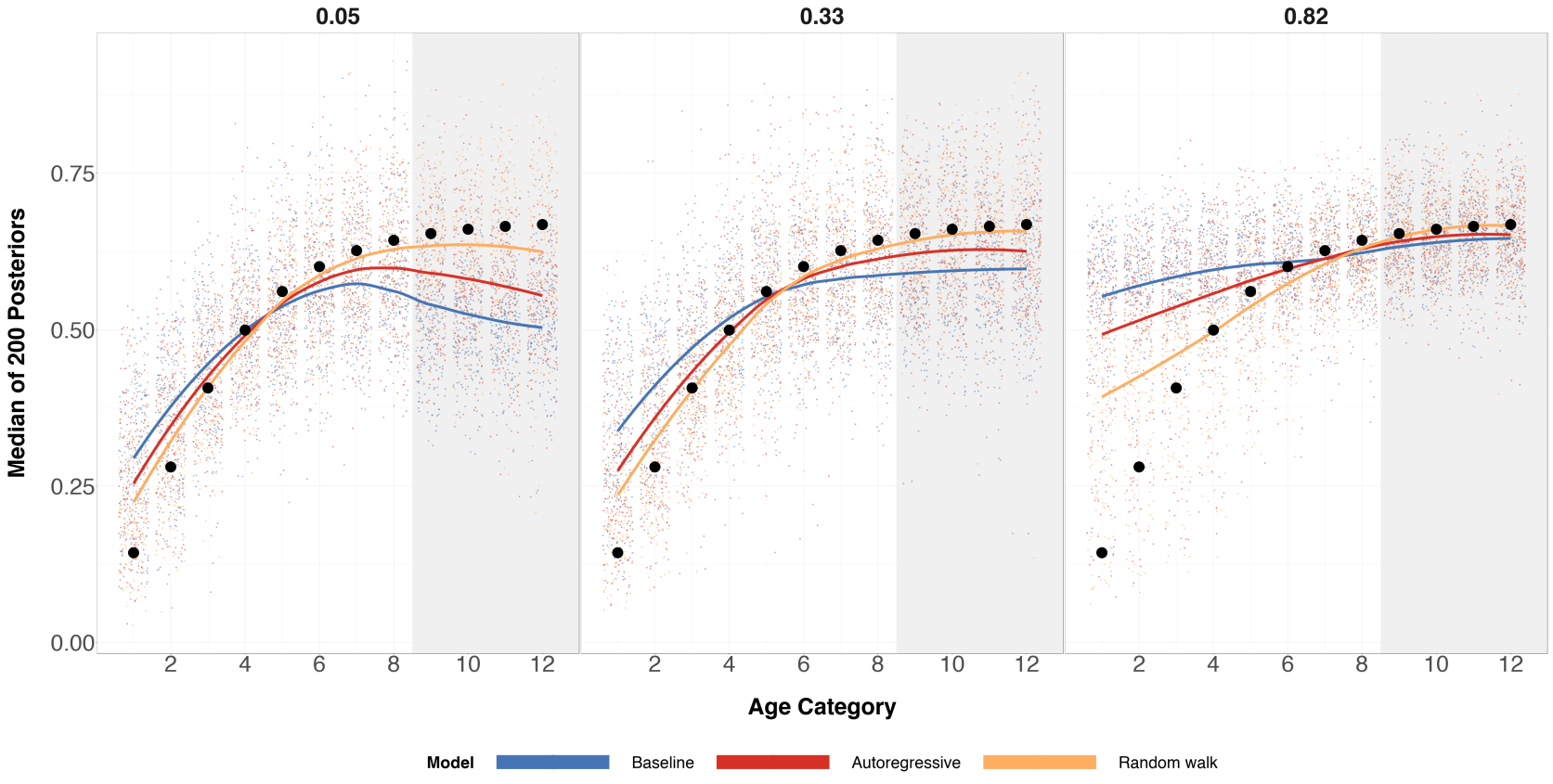
Figure 3: Posterior medians for 200 simulation runs, where the probability of sampling individuals in age groups 9-12 is 0.82.



- Sample size $n = 100$.
- Black dots are the true preference for every age.

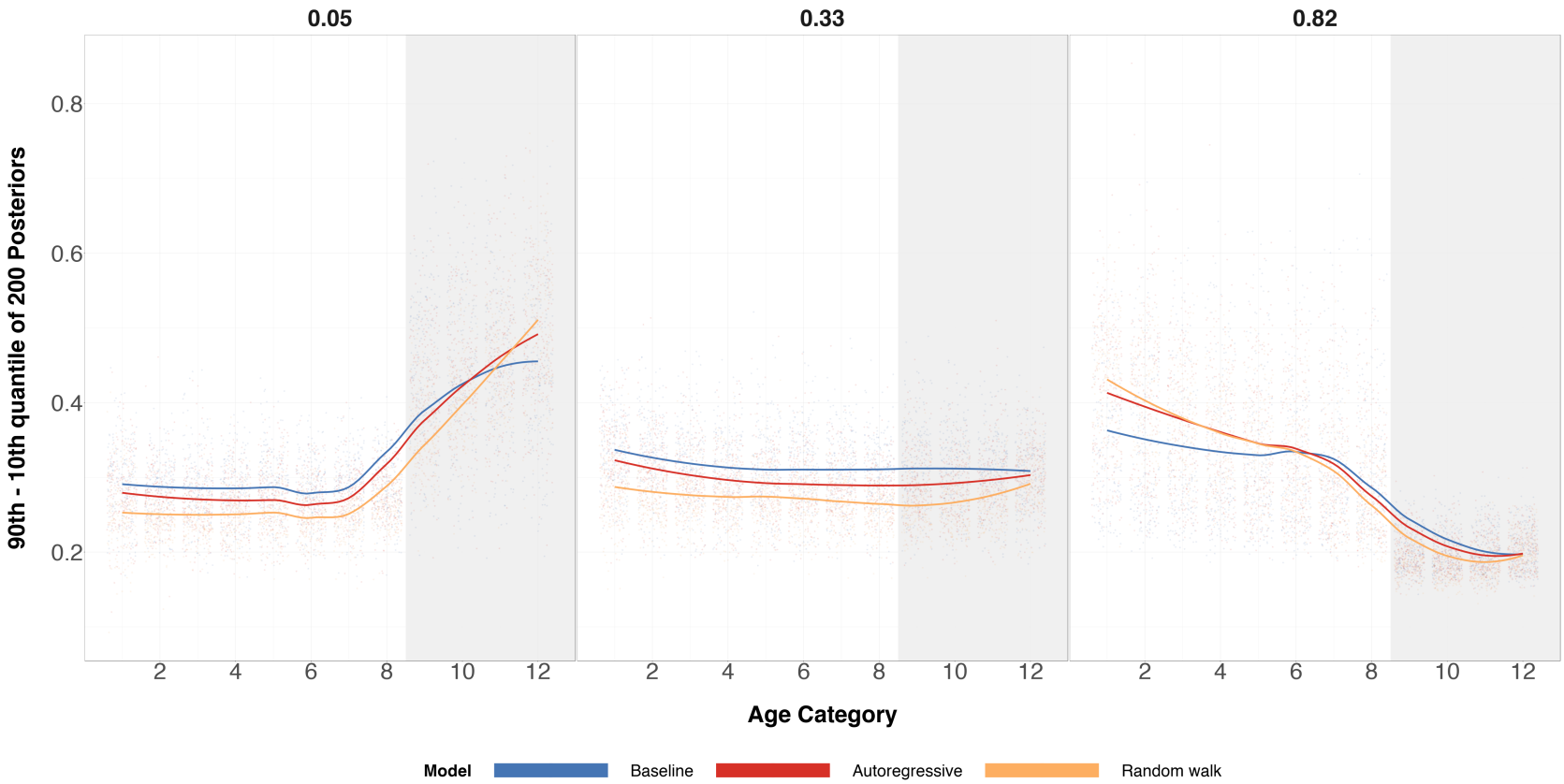
Simulation results: Summary of posterior bias

Figure 4: Posterior medians for 200 simulation runs, where the probability of sampling individuals in age groups 9-12 is: 0.05 (left), 0.33 (middle), 0.82 (right). Sample size $n = 100$. Black dots are the true preference for every age category.



Simulation results: Summary of posterior standard deviation

Figure 5: Posterior quantile differences for 200 simulation runs, where the probability of sampling individuals in age groups 9-12 is: 0.05 (left), 0.33 (middle), 0.82 (right). Sample size $n = 100$.

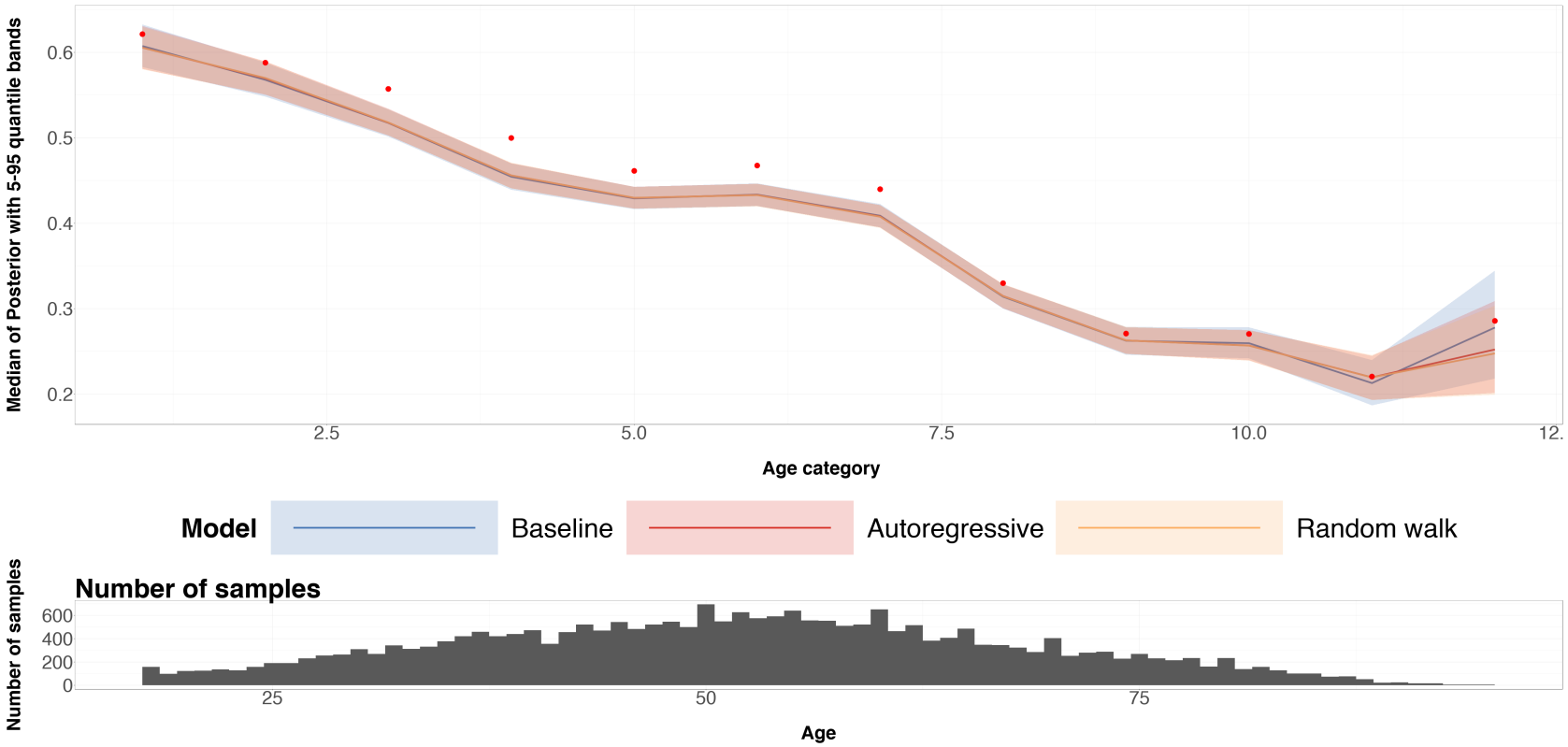


Real data analysis

- We analyze the non-representative survey data set *The Annenberg Public Policy Center's National Annenberg Election Survey 2008 Phone Edition*, originating from the Annenberg Public Policy Center of the University of Pennsylvania
- Annenberg survey sample size $n = 24,387$
- The response variable of interest is whether someone favors gay marriage or not
- The 2006 - 2010 5-year American Community Survey is used to form the poststratification matrix
- Covariates used in modelling are sex, race/ethnicity, income, state, age, education

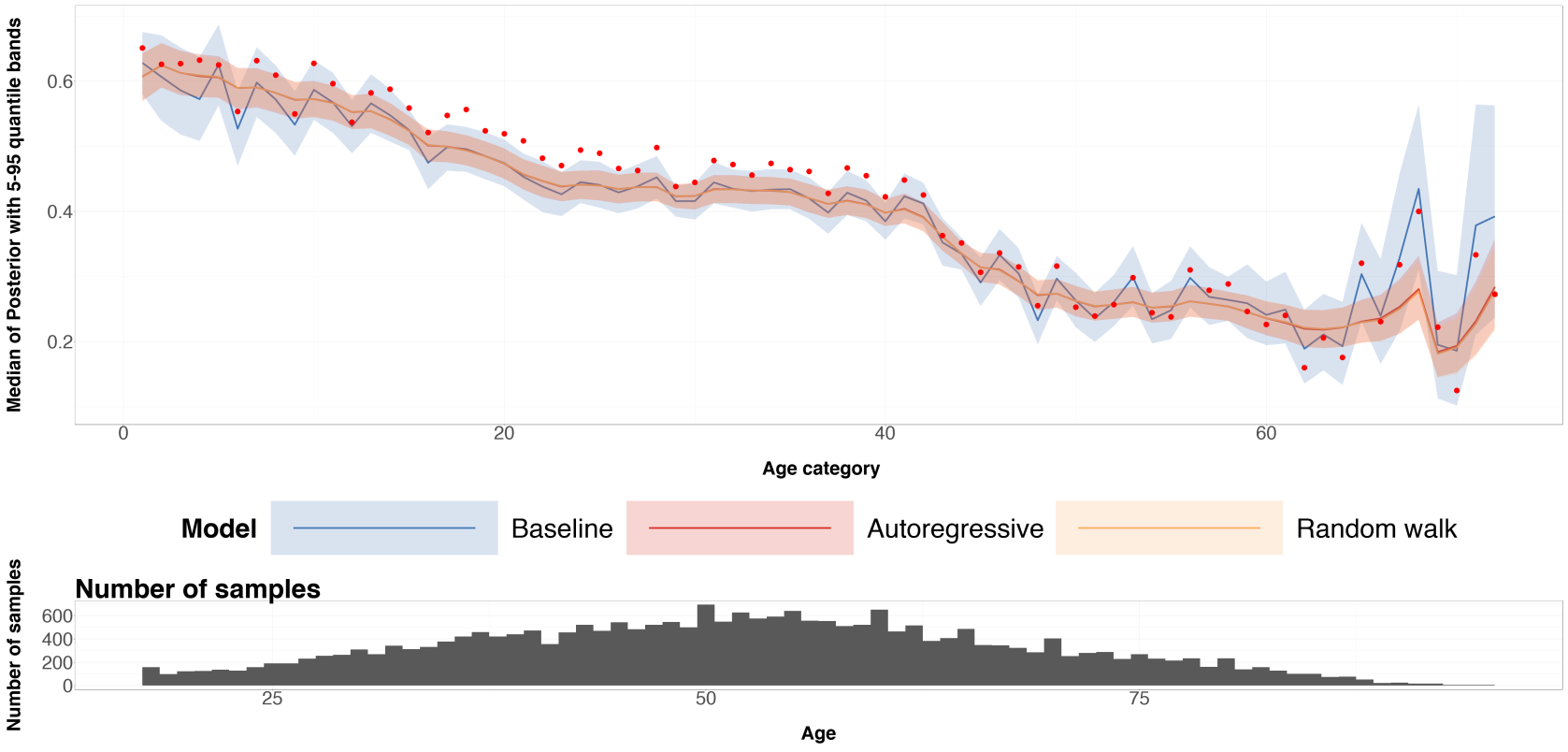
Real data analysis

Figure 6: Posterior preferences for 12 age categories



Real data analysis

Figure 7: Posterior preferences for 72 age categories



Structured priors on age has a nonparametric smoothing effect!

What did we learn?

Summary:

1. Non-representative samples in surveys introduces unnecessary bias in the classical MRP specification. We introduce structured priors to address this
2. For example, a structured prior for age would be an first-order autoregressive process
3. We show empirically that the shrinkage-to-neighbours in the autoregressive random effects reduces sub-population level posterior bias and variance

Things we did not consider:

1. Structured priors with interactions
2. Selection of what structured covariates to incorporate in MRP

References

1. Annenberg Public Policy Center of the University of Pennsylvania. (2008). *The Annenberg Public Policy Center's National Annenberg Election Survey 2008 Phone Edition (NAES08-Phone)* [Data file and code book]. Retrieved from https://www.annenbergpublicpolicycenter.org/naes-data-sets/