Using Mixture Models to Assess Enumerator and Survey Quality: An Extension of Probabilistic Record Linkage

Simon Hoellerbauer

University *of* North Carolina *at* Chapel Hill

Research Objectives

Substantive problem:

- Make analysis of survey backchecking—also called field audits or re-interviews—faster, more rigorous, and more efficient

Methodological objective:

- Use mixture models and the probabilistic record linkage framework to put a probabilistic model on the backchecking process

Probabilistic Model

The model is a finite mixture model with two component distributions, each of which is a Multinomial

$$\begin{aligned}\gamma_i | M_i = m &\sim \text{Multinomial}(\pi_m) \\ M_i &\overset{\text{i.i.d}}{\sim} \text{Bernoulli}(\lambda_e) \\ \lambda_e &= \text{logit}^{-1}(\beta_0 + \beta_e) \\ \beta_e &\sim \mathcal{N}(0, \sigma_e)\end{aligned}$$

- This model combines the probabilistic record linkage model[1, 2] with the "theory selection" model[3]
  - γ_i represents the total agreement vector for the i^{th} survey-backcheck pair.
  - β_e represents random intercepts by enumerators, which contributes to the mixing parameter λ

Simulation Setup

I use actual survey data to simulate backchecks by creating artificial non-matches and disagreements, with some enumerators having more matches than others. I simulate 100 backcheck sets to create agreement matrices for each combination of the following parameters:

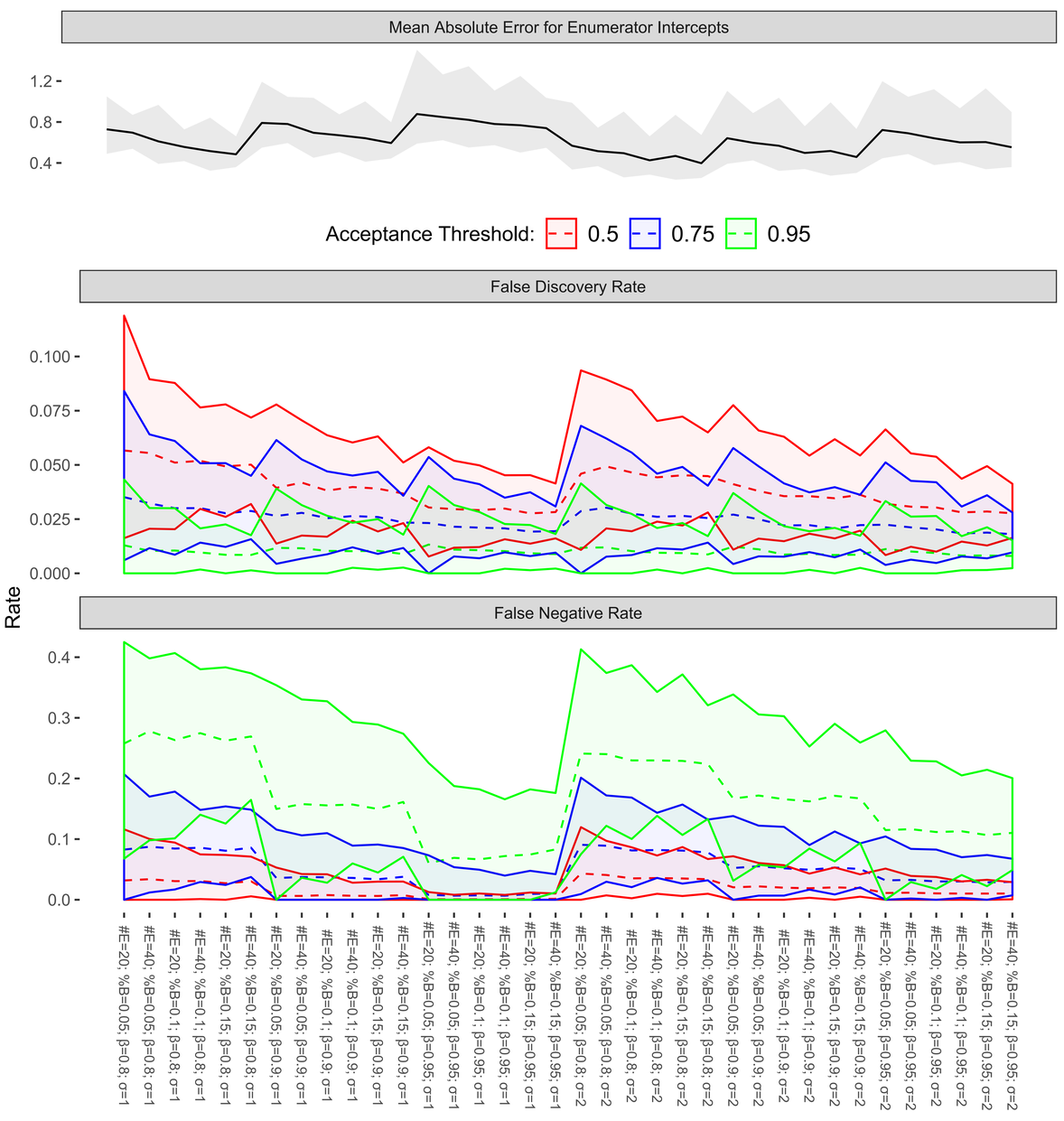
- Number of Enumerators (#E) $\in \{20, 40\}$
- Percent of Respondents Backchecked (%B) $\in \{0.05, 0.10, 0.15\}$
- Overall Match Probability (β_0) $\in \{0.8, 0.9, 0.95\}$
- Standard Deviation of β_E (σ_e) $\in \{1, 2\}$

Model Parameters

$$\begin{aligned}\gamma_i &\sim \text{Multi}(\pi_1) & \beta_e &\sim \mathcal{N}(0, \sigma_e) & \pi_0 &\sim \text{Dir}(2, 1) \\ \gamma_i &\sim \text{Multi}(\pi_0) & \sigma_e &\sim \text{Gamma}(1, 1) & \mu_{\beta_0} &\sim \mathcal{N}(0, .1) \\ \beta_0 &\sim \mathcal{N}(\mu_{\beta_0}, \sigma_{beta_0}) & \pi_1 &\sim \text{Dir}(1, 2) & \sigma_{beta_0} &\sim \text{Gamma}(1, 1)\end{aligned}$$

Models were estimated using `Stan`'s `R` interface `rstan`.

Model Assessment



Survey Evaluation Measures

Respondent Match Probability:

$$\lambda_{e_i} = \text{Logit}^{-1}(\beta_0 + \beta_{e_i})$$

$$\xi_i = \frac{\lambda_{e_i} \prod_{k=1}^{K} \pi_k^{\gamma_{ik}}}{\sum_{m=0}^{1} \lambda_{e_i}^m (1 - \lambda_{e_i})^m \prod_{k=1}^{K} \pi_{km}^{\gamma_{ik}}}$$

Enumerator Quality:

$$Q_e = \frac{\sum_{i_e}^{N_e} \xi_{i_e}}{N_e}$$

Survey Quality:

$$Q_S = \frac{\sum_{i}^{N} \xi_i}{N}$$

Conclusion

- Mixture models and agreement vectors can be used to facilitate backchecking
- However, when there are a large number of errors *and* a small proportion of an enumerator's respondents are chosen, the model does not perform as well

Next Steps

- Transition to treating agreement vector as a series of categorical variables, which would make it possible to identify variables that are consistently incorrect across backchecks
- Investigate use of responsibilities as survey weights to avoid discarding data

Email: hoellers@unc.edu

References

[1] Ivan P. Fellegi and Allan B. Sunter.
A theory for record linkage.
Journal of the American Statistical Association, 64(328):1183–1210, 1969.

[2] Ted Enamorado, Benjamin Fifield, and Kosuke Imai.
Using a probabilistic model to assist merging of large-scale administrative records.
American Political Science Review, pages 1–19, 2018.

[3] Kosuke Imai and Dustin Tingley.
A statistical method for empirical testing of competing theories.
American Journal of Political Science, 56(1):218–236, 2012.