Modeling Individual Poverty Classification

Bayesian Statistics - MATH 7393

Brady Lamson - Spring 2025

Repository Link

Background

- The Census Bureau has long tried to quantify and measure poverty.

- The first official poverty measurement came out in 1960, but opponents of it (even within the census bureau) argued that it was extremely limited.

  - " When they were developed, the official thresholds represented the cost of a minimum food diet multiplied by 3 (to allow for expenditures on other goods and services). The thresholds have been kept constant in purchasing power over time by increasing their money values to keep pace with increases in the general price level." (ssa.gov)

- Thus, to overcome these limitations, the "**Supplemental Poverty Measure** (SPM)" was created, with its first version coming out in 2011.

Data Source

1. **Supplemental Poverty Measure (SPM) Research Files** – The primary dataset is the 2023 SPM research file from the U.S. Census Bureau. These annual tables, available from 2009 to 2023, provide detailed information for poverty measurement. This analysis uses the dataset specific to 2023, so all analysis is restricted to that year.

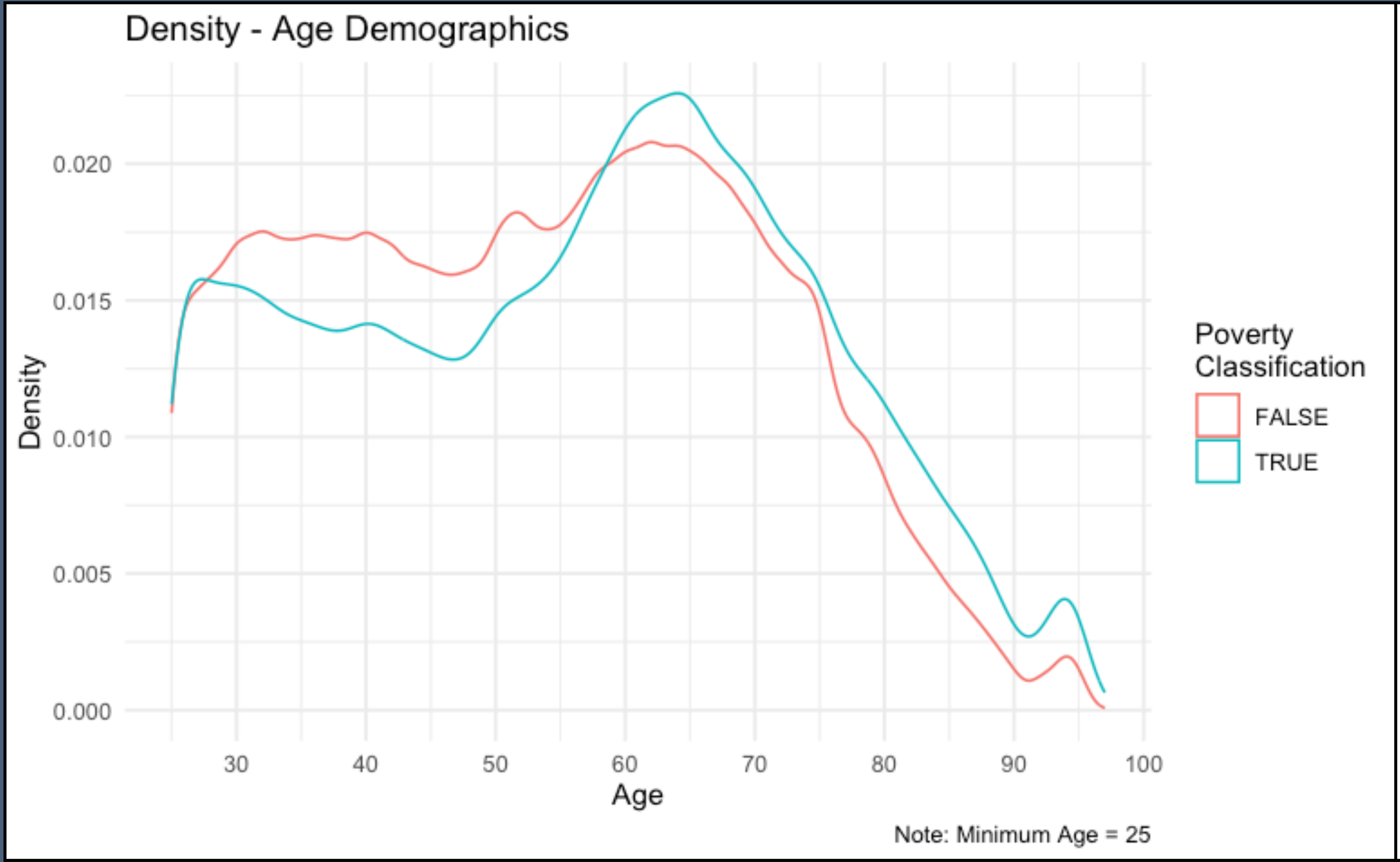
2. **FIPS Code Reference Data** – I join to the `**fips_codes**` dataset from the tidycensus R package to translate FIPS codes into state abbreviations.

Data Aggregation and Covariate Overview

- Original dataset is approximately 3.2 million rows. Each row is one person.
- Removal of rows with important missing info brought dataset to approximately 2.3 million rows.
- Stratified sampling brought row count down to 11000ish rows.

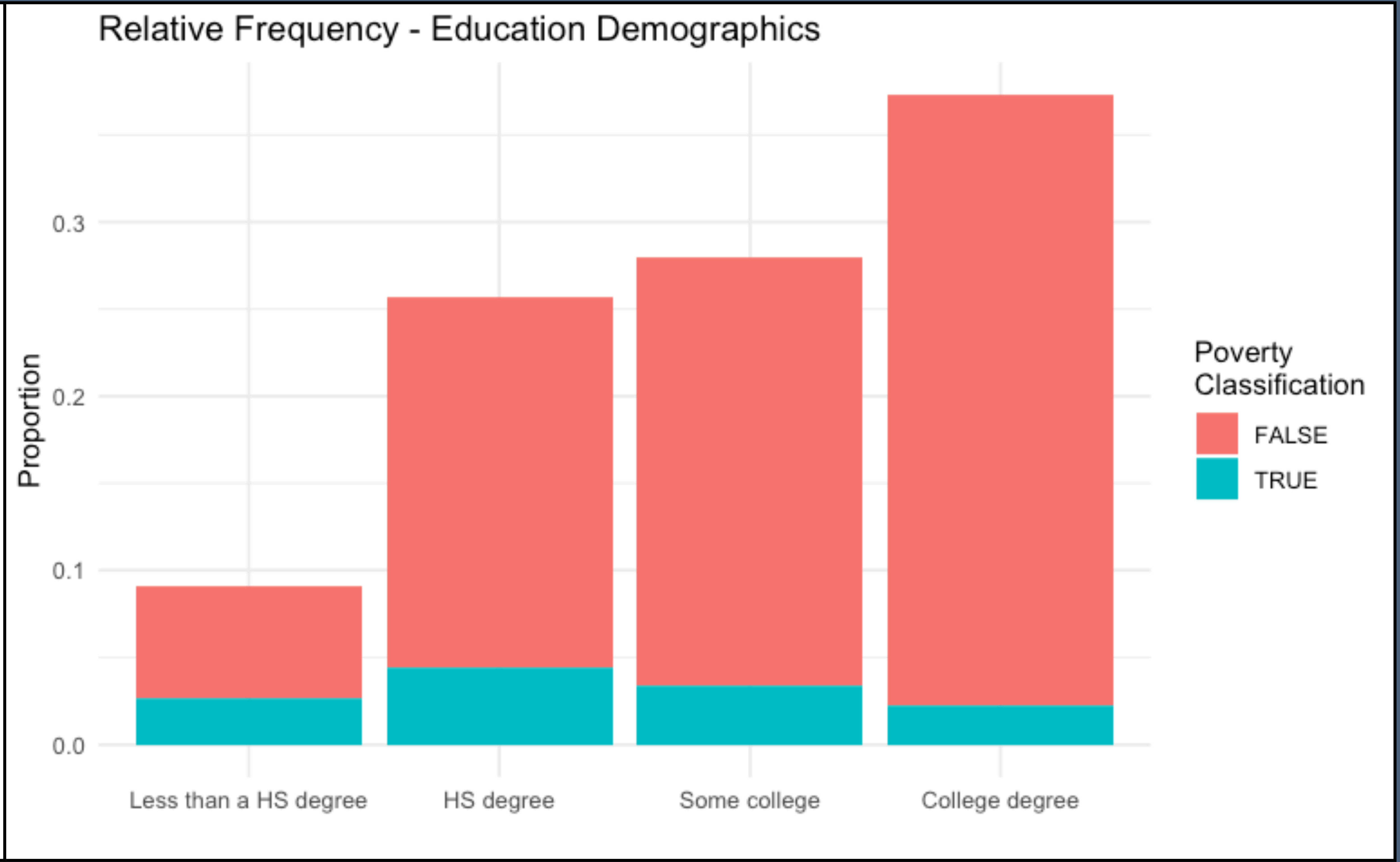
Variable	Data Type	Description
spm_poor	Boolean	Poverty classification according to the spm
spm_povthreshold	Numeric	Poverty Threshold - adjusted by a variety of factors
Age	Integer	Age
agi	Numeric	Adjusted Gross Income
Education	Factor	Less than HS degree, HS degree, some college, college degree
Region	Factor	Midwest, Northeast, South, West
Race	Factor	White, Asian, Black, Other
moop_other	numeric	Other medical out of pocket expenses
Other covariates	Boolean	sex_female, hispanic, married,
hi_premium	numeric	Health Insurance Premium

Basic Demographics II



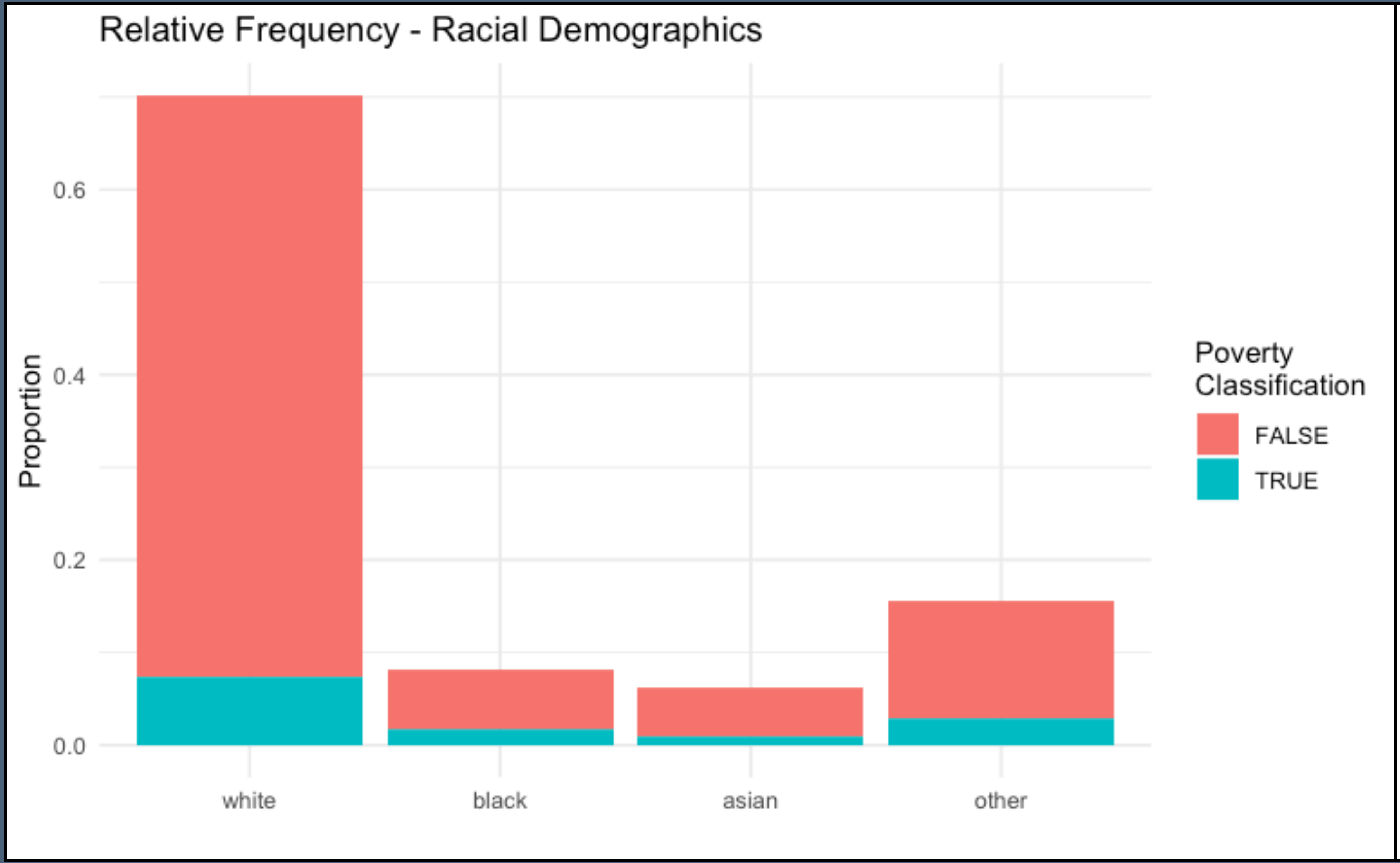
Density - Age Demographics

Note: Minimum Age = 25

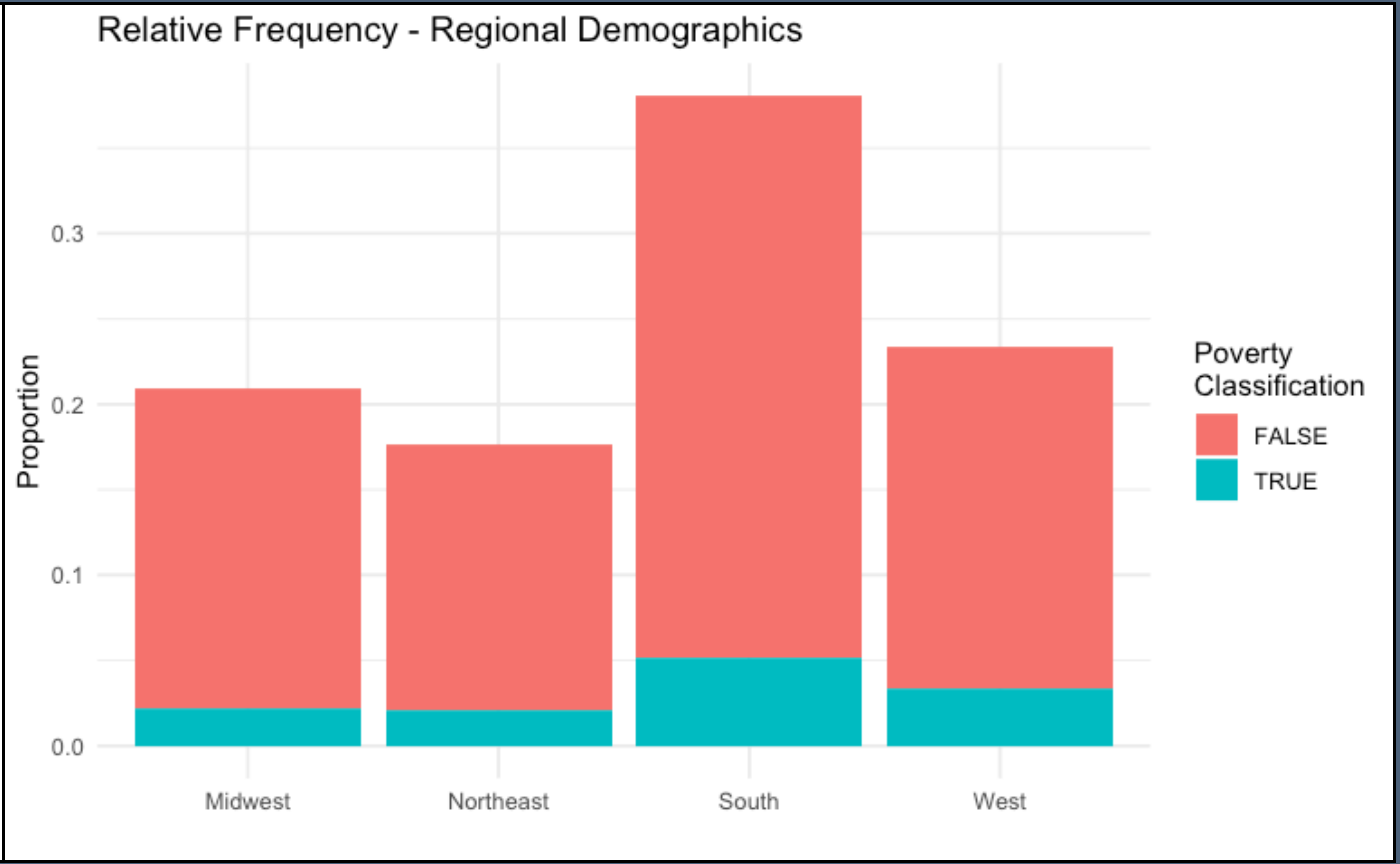


Relative Frequency - Education Demographics

Basic Demographics I



Relative Frequency - Racial Demographics



Relative Frequency - Regional Demographics

Approximately matches proportions of overall US demographics.

Variable Selection

- Handled in 3 steps.

1. Performed LASSO regression on all candidate covariates (Full Model)

   - Penalizes uninformative covariates and brings them close to 0.

2. Removed covariates straddling zero, updated the model (Model 2)

3. Examined correlated covariates, removed messy numeric candidates. (Model 3)

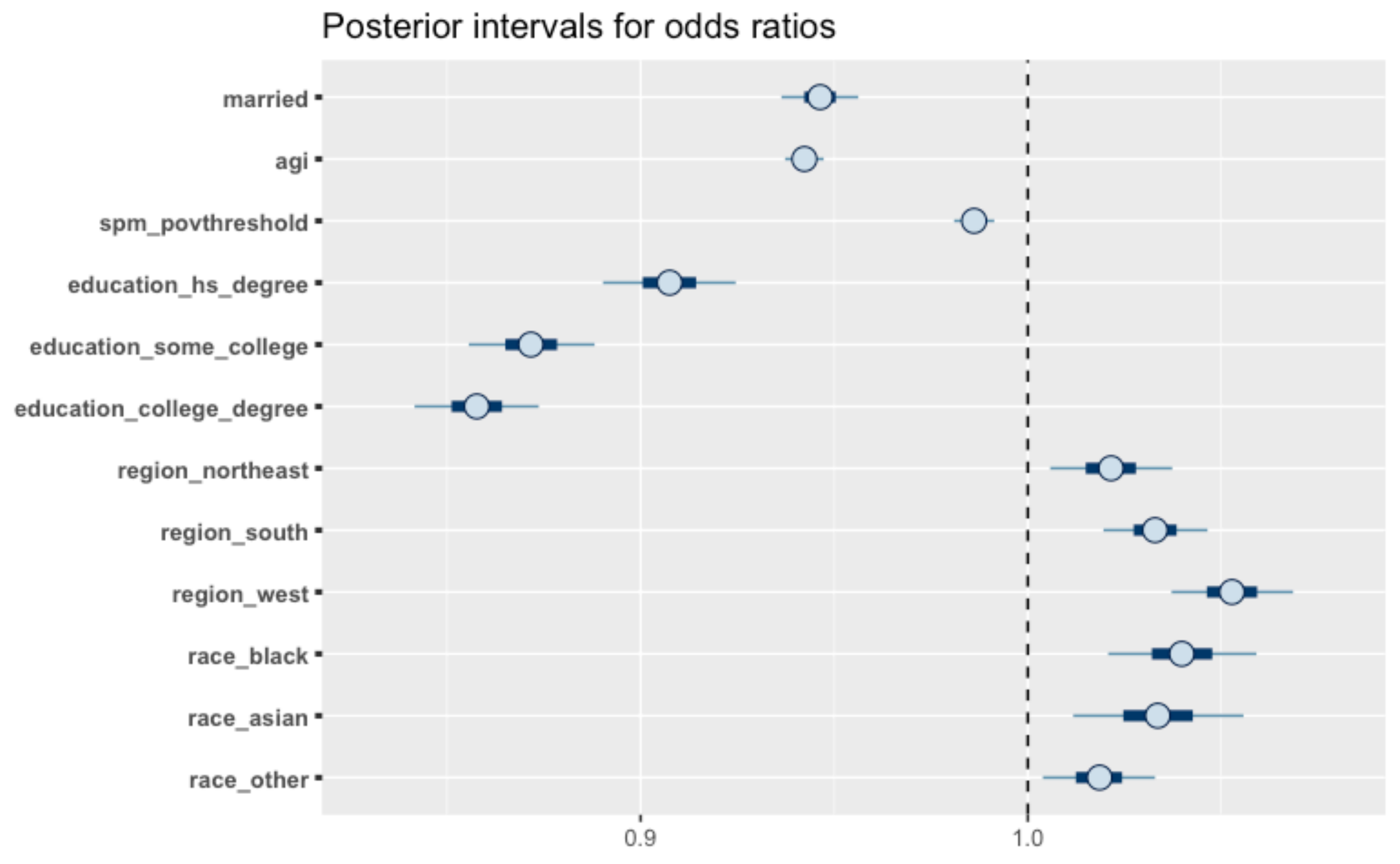
4. Compared WAIC scores for all 3 models, change in WAIC falls well within standard error moving to a simpler model. Model 3 was chosen.

Model	WAIC	SE
Full	5094	117.2
2	5092	117.2
3	5107	117.1

Final Model

Covariates chosen:

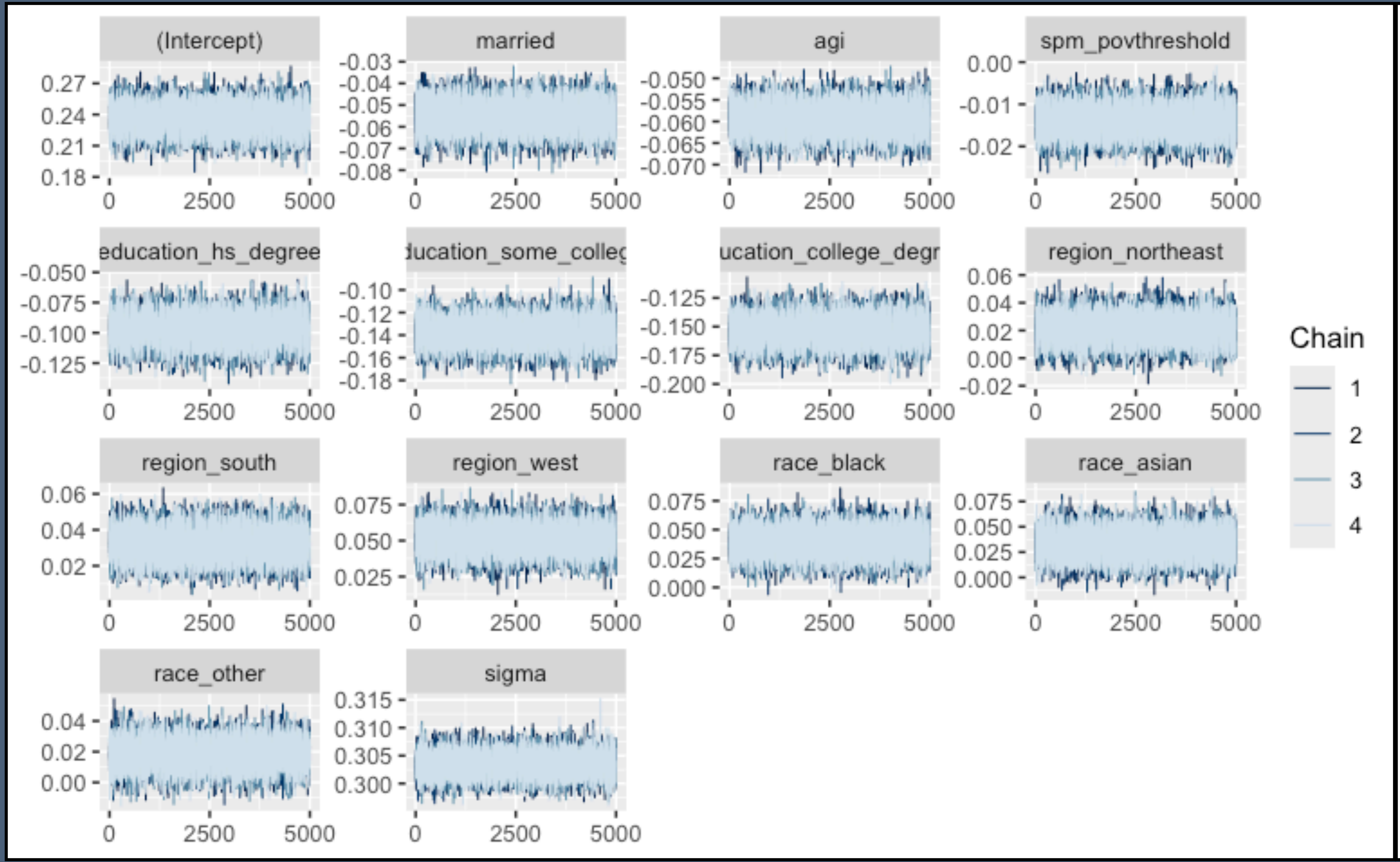
- Married
- AGI
- POV Threshold
- Education
- Region
- Race



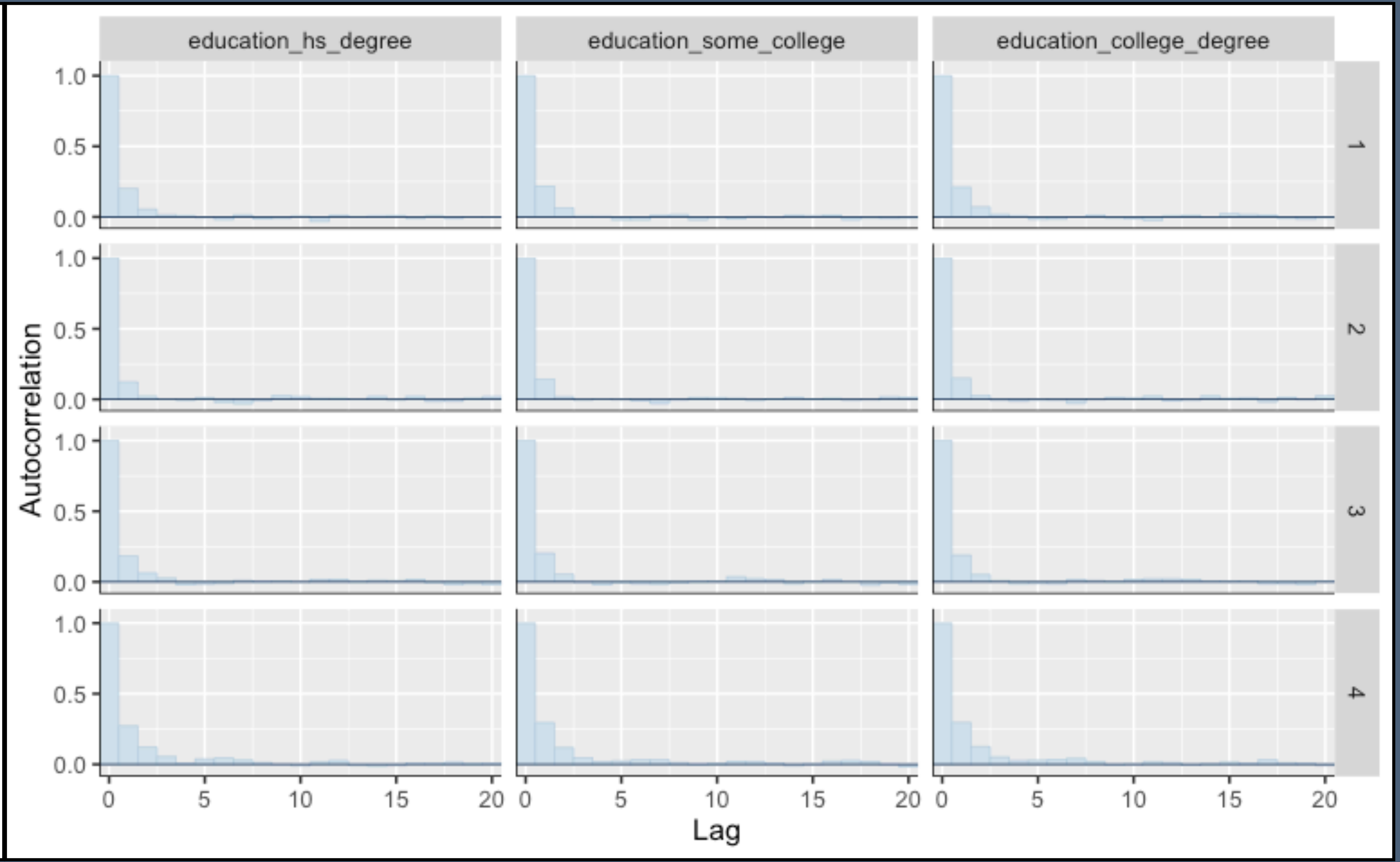
Posterior Odds Ratios - Table Version

Variable	5%	95%
(Intercept)	1.23977	1.29214
married	0.93637	0.95623
agi	0.93731	0.94728
spm_povthreshold	0.98096	0.99135
education_hs_degree	0.89027	0.92456
education_some_college	0.85563	0.88812
education_college_degree	0.84162	0.87371
region_northeast	1.00579	1.03730
region_south	1.01953	1.04644
region_west	1.03711	1.06853
race_black	1.02077	1.05906
race_asian	1.01172	1.05571
race_other	1.00387	1.03289
sigma	1.35008	1.35890

Model Assessment I

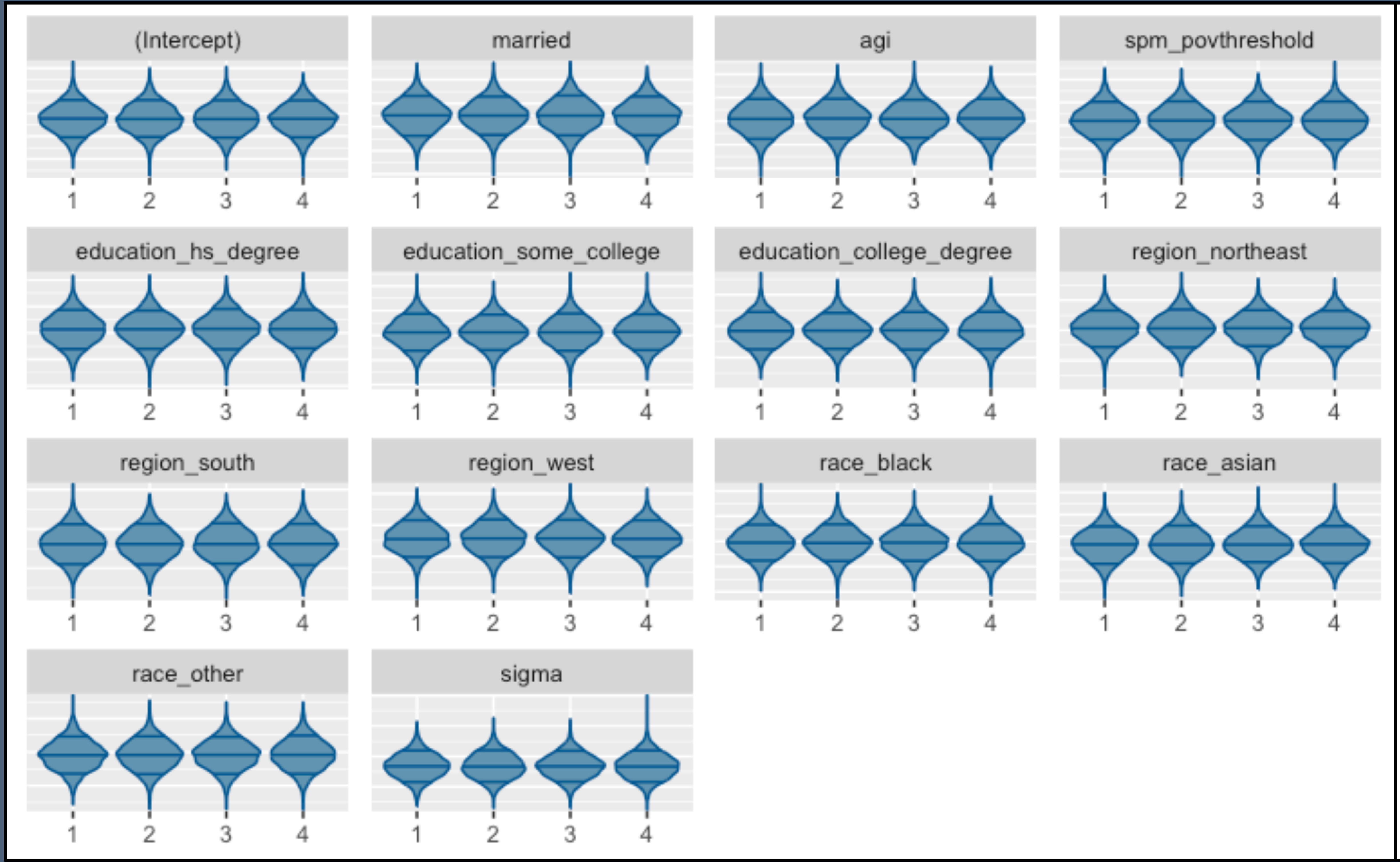


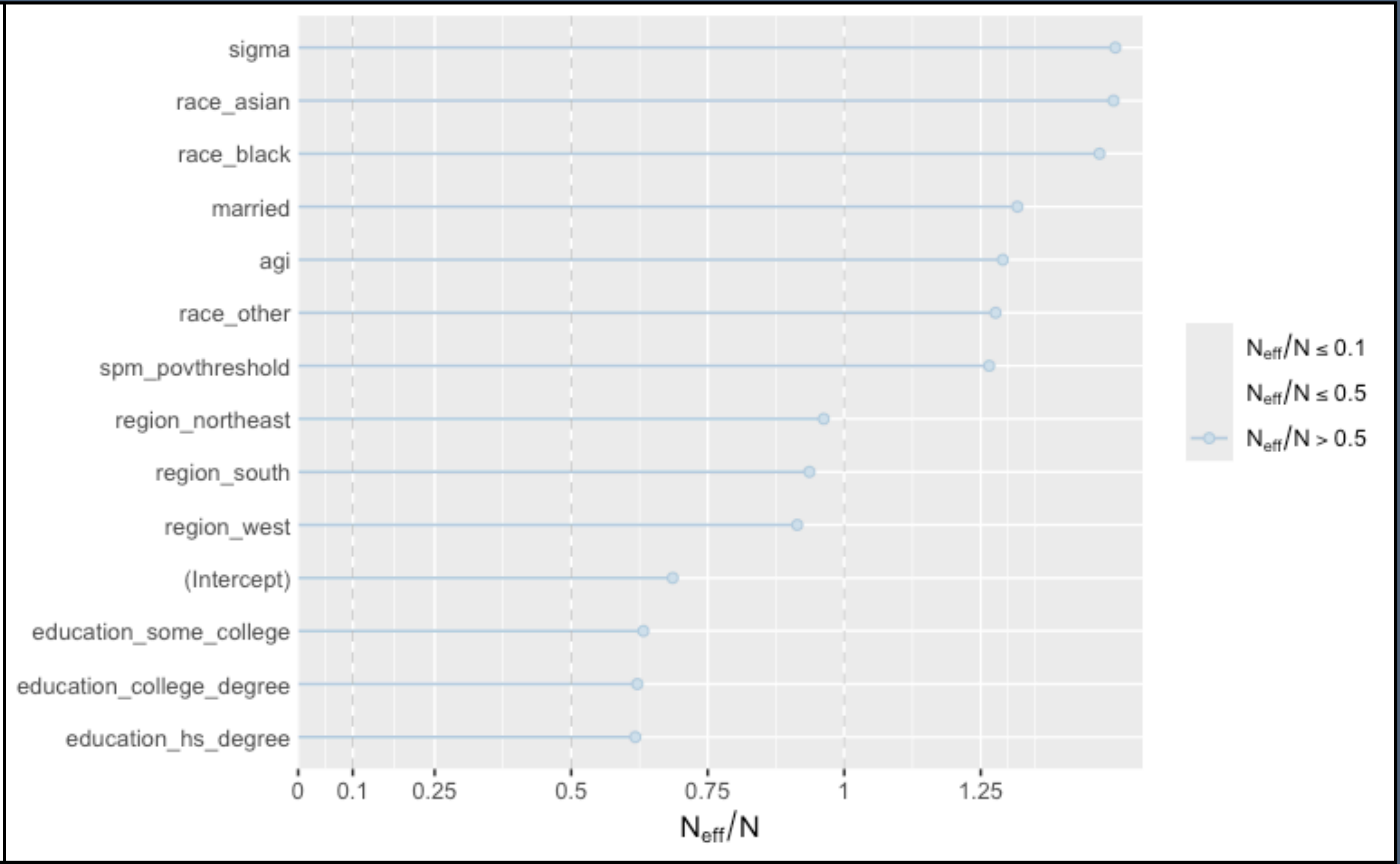
Trace Plots all look good.



Example Autocorrelation plot, all showed similar behavior.

Model Assessment II





No concerning violins or diverging chains.

None of the effective sample sizes fall into problematic ranges. (Ex: Neff/N < 0.1)

Conclusion and Limitations

Mostly Limitations

- Results here are largely unsurprising but it is fascinating to see them shown through this process.
- Ease of this modeling process is a testament to the Census Bureau’s ability to craft a good dataset.
- Not a deep enough understanding of the demographic makeup of those removed from the data. Who exactly was I removing from this analysis?
- I feel I could have handled “Age” better. The census bureau in their analyses, even the SPM, uses age buckets and makes it a factor variable. It’s possible I discarded a valuable covariate due to poor handling.
- Ran out of time to test the predictive performance of this model, need for complex stratified sampling makes test/train split conceptually tricky.