Journal of Data Science 22 (4), 605–620
October 2024
DOI: 10.6339/23-JDS1113
Data Science in Action

Exploring Racial and Ethnic Differences in US Home Ownership with Bayesian Beta-Binomial Regression

Jhonatan Medri[1,*], Tejasvi Channagiri[1,*], and Lu Lu[1]

[1]*Department of Mathematics & Statistics, University of South Florida, USA*

Abstract

Racial and ethnic representation in home ownership rates is an important public policy topic for addressing inequality within society. Although more than half of the households in the US are owned, rather than rented, the representation of home ownership is unequal among different racial and ethnic groups. Here we analyze the US Census Bureau's American Community Survey data to conduct an exploratory and statistical analysis of home ownership in the US, and find sociodemographic factors that are associated with differences in home ownership rates. We use binomial and beta-binomial generalized linear models (GLMs) with 2020 county-level data to model the home ownership rate, and fit the beta-binomial models with Bayesian estimation. We determine that race/ethnic group, geographic region, and income all have significant associations with the home ownership rate. To make the data and results accessible to the public, we develop an Shiny web application in R with exploratory plots and model predictions.

Keywords *census; exploratory analysis; generalized linear models; GLM; housing; sociodemographic factors; statistical analysis*

1 Introduction

Home ownership is popularly considered a core element of the American dream. Owning a home provides social status, shelter, and financial stability; a starting point for families to start accumulating wealth (Austin and Felicity, 1999). Moreover, the US has consistently reported rates above 60% since 1960, meaning US citizens have been more likely to own rather than rent a home (Robb, 2021). However, current data suggests there have been considerable differences between the home ownership rates of different racial/ethnic groups throughout that time.

The 2022 Data Challenge Expo of the Sections on Statistical Computing, Statistical Graphics, and Government Statistics of the American Statistical Association (ASA) provided an opportunity to explore the relationship between home ownership rates and other sociodemographic variables in the US. Our analyses focused on US Census data from the year 2020 and tried to answer the following primary research question: Are US home ownership rates significantly different among different racial/ethnic groups? In addition, our analyses addressed the following secondary question: What variables and estimation methods help explain differences in home ownership rates among different racial groups?

In this article, we primarily investigate the home ownership gaps between the WhiteNH (White Non-Hispanic), Hispanic, Black, and Asian groups. The existing literature on Hispanic home ownership and its disparity in comparison to other racial/ethnic groups has identified

*Corresponding and joint first authors. Email: jm192@usf.edu or tchannagiri@gmail.com.

Received January 16, 2023; Accepted July 31, 2023

various drivers, including financial, demographic, and assimilative factors. For instance, studies have examined the influence of householder age, household income, and other assimilative drivers on home ownership rates (Sanchez-Moyano, 2021). Additionally, the geographic concentration of the Hispanic population in ethnic enclaves has been recognized as a potentially influential factor affecting the likelihood of owning a home (Flippen, 2010).

Some authors have explored the Black/White gap and attribute this difference to sociodemographic factors such as household income and educational attainment (Choi et al., 2019), but also to financial factors such as mortgage access (Aronowitz et al., 2020) and tax exemptions (Thomas, 2021). The Asian group, in contrast, exhibits different patterns of home ownership, since the group's immigration background and high diversity influence home ownership rates (Kuebler, 2013). While the Black group is significantly less likely than the White group to earn home equity, the Asian group is not that disadvantaged and has shown more improvement than other minority groups, even since 2000 (Krivo and Kaufman, 2004).

Regarding estimation methods, previous research has made use of Bayesian statistical models in public policy topics, such as neighborhood quality (Mast, 2010), local government land use decisions (Deslatte et al., 2018), and even environmental regulation violations (Paleologos et al., 2018). Moreover, some authors have developed Bayesian models, such as hierarchical (Hui et al., 2010) or averaging methods (Erdoğdu et al., 2021), to explain residential property valuation or housing prices. In the study of home ownership, methods using ordinary least squares and logistic regression (binomial family) have been used (Moore, 1991; Goodman and Mayer, 2018; Jones, 1989; Delgadillo, 2009). In this article, we used Bayesian beta-binomial regression to help explain racial/ethnic differences in home ownership rates. To the best of our knowledge, this is the first application of beta-binomial regression in the literature to study home ownership rates.

This article is an extension of our preliminary work submitted to JSM Proccedings 2022 (Medri and Channagiri, 2022). Our current study includes the Hispanic ethnicity in the analysis and focuses only on 2020 county data, since our previous results showed no significant change of home ownership rates in the last six years. In our previous study, we considered additional US Census tables and variables, such as housing tax, housing cost, etc. We omitted these variables from this study because though they were available per county, they were not broken by racial/ethnic group. Additional variables would also have added computational expense to fitting the models. Although we found evidence that spatial regression techniques may be useful, we did not pursue such techniques here and left that for future work. We also decided not to use imputation for missing values because counties with missing values had very small populations (see Section S.2 (Supplementary)).

This article is structured as follows. Section 2 provides an overview of the data and outlines the selected variables for our analyses. In Section 3, we present the binomial and beta-binomial Bayesian statistical models used in this study. Our main results are discussed in Section 4. Section 5 offers conclusions drawn from our study and proposes avenues for future research. Finally, Section S (Supplementary) contains additional data exploration and analyses, along with a description of a Shiny app developed for this research.

2 Data

The data used in this article is from the American Community Survey (ACS) 2020 5-year estimate subject tables (US Census Bureau, 2015–2020) obtained at the county level. In our previous work, we aggregated this data into state-level data by ignoring missing values and

Table 1: Housing and sociodemographic variables by racial/ethnic group in 2020. Notable findings include higher home ownership rates among WhiteNH (White Non-Hispanic) and Asian groups, income disparities with the highest annual income among Asian households, and population distribution showing the largest share among the WhiteNH group. "Home Own." = "Home Ownership Rate", "H. S. Edu." = "High School Education Rate", "Unemp." = "Unemployment Rate", "Pop. Share" = "Population Share", "Income" = "Anual Income".

Group	Home Own.	H. S. Edu.	Unemp.	Pop. Share	Income
WhiteNH	72.3%	93.2%	4.4%	60.1%	$74,014
Hispanic	48.6%	70.3%	6.3%	18.2%	$55,791
Black	42.5%	86.7%	9.3%	12.6%	$45,880
Asian	60.1%	87.3%	4.3%	5.6%	$95,301
All	64.5%	88.4%	5.4%	100.0%	$67,619

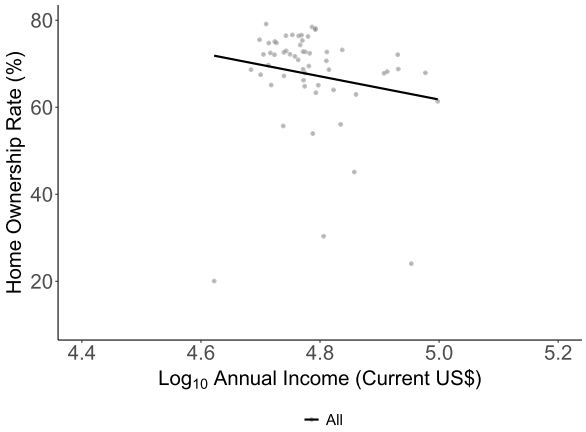
taking weighted averages based on the population of each county. However, in this article we extend our analysis to the county-level data. We consider occupied housing units (both owned and rented) as our observational units, though some tables included information from only owner-occupied housing units.

Table 1 indicates the variables considered in our analyses. We decided to focus our analyses on the variables from the survey that were related to housing and sociodemographics. Among housing variables, our main variable of interest was the home ownership rate, the percentage of owned housing units out of occupied housing units in a certain county. Regarding sociodemographics, we considered education, income, employment, and population variables. The educational attainment was defined as the percentage of people 18 years and over that are high school graduates or higher. We also include median income in the past 12 months expressed in current 2020 dollars, and unemployment rate, the number of unemployed as a percentage of the labor force. We finally considered the total population in each county.

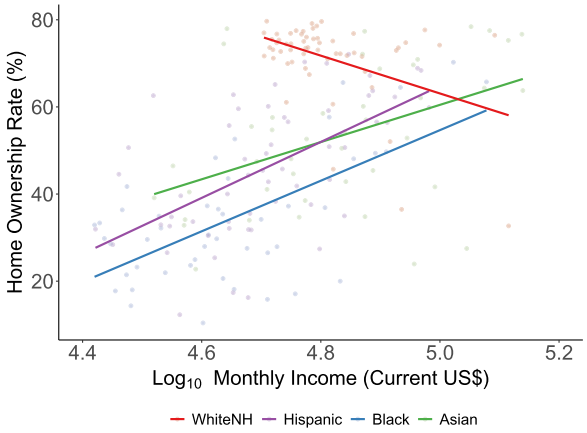
We can make some general observations regarding home ownership in the US in the year 2020. According to US Census Bureau (2015–2020), approximately 64.5% of housing units are owned, indicating a preference for home owning over renting among US residents. The WhiteNH group reports, on average, the highest home ownership rate (72.3%), high school attainment (93.2%), and population share (60.1%). For annual income and unemployment rate, the WhiteNH group has the second highest and lowest value ($74,014 and 4.4%, respectively). The Hispanic group, which is the second largest group, represents 18.2% of the population. They have the second lowest home ownership rate (48.6%), annual income ($55,791), and unemployment rate (6.3%). They also report the lowest high school attainment rate (70.3%).

The Black group reports the lowest home ownership rate (42.5%) and annual income ($45,880), but highest unemployment rate (9.3%) and second lowest high school attainment rate (86.7%). They are also the third biggest racial/ethnic group in term of population share (12.6%). Finally, the Asian group reports the highest annual income ($95,301) but smallest population share (5.6%) and smallest unemployment rate (4.3%). For both home ownership rate (60.1%) and high school attainment rate (87.3%), the Asian group has the second highest value.

The values in Table 1 suggest that an analysis of home ownership in the US may require controlling for racial/ethnic groups to properly estimate relationships with other variables. For example, when looking at the association between home ownership rate and monthly income in

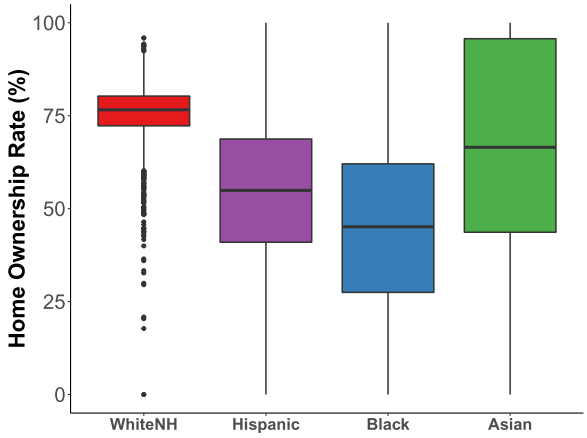


(a) Household income and home ownership rate show a negative association in all counties in New York state.

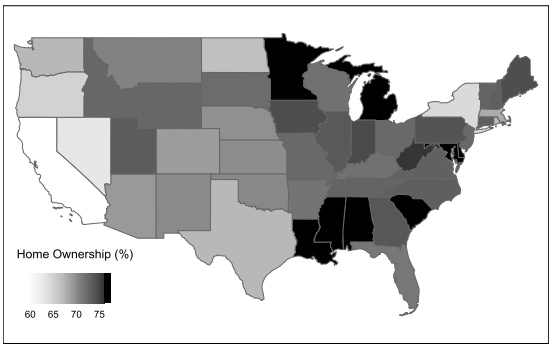


(b) Household income and home ownership rate show a different association in each racial/ethnic group in New York state.

Figure 1: New York household annual income and home ownership rate scatter plot for 2020. The income variable, on the x-axis, has a $\log_{10}$ scale. Each dot represents a county.



(a) US home ownership rates box plot for 2020 by racial/ethnic group. Different racial/ethnic groups show significantly dissimilar distributions.



(b) US home ownership rate choropleth map for 2020. We see significantly different home ownership rates in different states. The states Alaska and Hawaii were omitted from the plot, but reported rates of 65.5% and 61.2% respectively.

Figure 2: Variation between racial/ethnic groups (a) and states (b).

the state of New York for the year 2020 in Figure 1a, we notice a negative association between both variables. However, when separating the relationship by race, we notice in Figure 1b a different association in each group, indicating that interaction effects may be significant. Moreover, the box plot in Figure 2a suggests that each racial/ethnic group may have a different distribution. In the choropleth map in Figure 2b, we see that there are significant differences in the home ownership of different states. Thus, we may need to control for the effect of state when developing statistical models.

Data manipulations and visualizations in this section were done in R (R Core Team, 2021) using the R package suite tidyverse (including dplyr (Wickham et al., 2022) and ggplot2 (Wick-

ham, 2016)). Choropleth maps were made using the R packages sf (Pebesma, 2018) and tmap (Tennekes, 2018). Please see Section S.1 (Supplementary) for further descriptions of the data set.

3 Methods

3.1 Modeling Variables

We transformed and renamed several of the raw US Census data variables to make them more amenable to modeling. The variables *hsedu* (high-school education proportion) and *unemp* (unemployment proportion), were proportions in the range [0, 1]. The monetary variable, *income* (annual income), was originally in units of 2020 US dollars ($) but was transformed to units of $100,000 so that all variables would be in similar ranges. The variables *state* (US geographic state) and *group* (racial/ethnic group) were categorical variables with fifty and four levels, respectively, and were expanded into dummy variables in the modeling stage. The *state* categorical variable was used to control for variation between states but was not of intrinsic interest. The levels for *group* were WhiteNH (White, Non-Hispanic), Hispanic, Black, and Asian. The baseline group for determining the coefficients was WhiteNH, since it was the most populous. Non-Hispanic tabulations for the Asian and Black groups were not available in all the US Census tables, and so the data from these groups may overlap with the Hispanic group. According to a 2010 US Census brief, the estimated population that was both Hispanic and Black was 1,243,471 (2.5% of the Hispanic population or 3.2% of the Black population), and the estimated population that was both Hispanic and Asian was 209,128 (0.4% of the Hispanic population or 1.4% of the Asian population) (Bureau, 2011). While this is not insignificant, we feel the relatively small degree of overlap of the Hispanic with the Black and Asian groups will not be problematic for our inferences. The *unemp* variable was available in US Census table as a percentage with one decimal, so its values have only two to three significant digits.

The *sampsize* variable represents the number of households interviewed for a given county and group. This variable was used as the size parameter for the binomial and beta-binomial distributions, and for weighting histogram plots. We obtained the number of interviewed households in each county from a US Census table. However, this data was not broken by group, and so we estimated the group-level values by multiplying the county-level value by the population share of each group in that county and rounding to the nearest integer. For example, if the county had 1,000 interviews and the WhiteNH group was .7312 of the county's population, we would assign WhiteNH's *sampsize* in that county to 731. Similarly, the number of owned households, *owners*, in each county and group was estimated from the tables by multiplying the owned household proportion by *sampsize* and rounding to the nearest integer. We omitted all observations that had $sampsize = 0$. These approximations likely lead to rounding errors and we discuss a possible fix in Section 5.

The original data tables from the US Census had data arranged in one row per county and separate columns for the variables of different racial/ethnic groups. We formatted each table so that each row represented a unique county and group. Then we joined the separate tables into a single table using the county and group as keys. Some variables had missing values, with a maximum of 1.99% of the sample size and 56.84% of the observations missing from the income variable of the Asian group (see Section S.2 (Supplementary)). While some of the percentages of missing observations were high, the corresponding sample sizes were relatively small. This indicates that the observations with missing values had small sample sizes, and may have had

their data suppressed (US Census Bureau, 2021a). Thus, we decided to omit these observations from the final modeling. The final data set had 3,142 counties, 12,568 observations (3,142 counties $\times$ 4 groups), and 8,820 non-missing observations with *sampsize* > 0.

The data manipulations were performed using several tidyverse (Wickham et al., 2019) R packages (including dplyr (Wickham et al., 2022), tidyr (Wickham et al., 2023b), (Wickham et al., 2023a), tibble (Müller and Wickham, 2023), stringr (Wickham, 2022), and forcats (Wickham, 2023)). Table output was formatted using the xtable (Dahl et al., 2019) R package. Please see Section S.2 (Supplementary) for additional tables and figures summarizing the modeling variables.

3.2 Model Formulation

To study the relationship between home ownership and the other sociodemographic variables in the US Census data, we formulated several binomial and beta-binomial generalized linear models (GLMs). In our models, the binomial density was parametrized as

$$p(x;\theta,N) \propto \binom{N}{x}\theta^x(1-\theta)^{N-x} \tag{1}$$

while the beta-binomial density was parametrized as

$$p(x;\theta,\phi,N) \propto \binom{N}{x}\int_0^1 (\theta')^{x+\theta\phi-1}(1-\theta')^{N-x+(1-\theta)\phi-1}\,d\theta'. \tag{2}$$

Both distributions have mean $N\theta$. However, the variance of the binomial distribution is $N\theta(1-\theta)$, while the variance of the beta-binomial is $N\theta(1-\theta)(\phi+N)/(\phi+1)$. Thus, the precision, ϕ, controls the amount of over-dispersion of the beta-binomial response relative to the binomial response, with lower precision resulting in higher over-dispersion. At the extremes, $\phi \to \infty$ makes the variance approach the binomial variance, and $\phi \to 0$ makes the variance approach N times the binomial variance. Formally, the beta-binomial is a mixture of binomial distributions with the proportion parameter (θ' in (2)) having a prior beta distribution with shape parameters $\alpha = \theta\phi$ and $\beta = (1-\theta)\phi$ (Prentice, 1986; Ferrari and Cribari-Neto, 2004).

One motivation for using count models was due to the large number of 0 and 1 home ownership rates in observations with small populations (i.e., all interviewed households were owned or all were rented). Large numbers of 0 and 1 proportions may not fit well with a continuous distributions such as the beta (Ospina and Ferrari, 2012). Additionally, the observations had a large range of sample sizes, and the binomial and beta-binomial distributions naturally weight observations by their sample sizes. The motivation for selecting the beta-binomial over the binomial was due to the beta-binomial's greater flexibility. Particularly, as discussed in Section S.3 (Supplementary), the binomial was a poor fit for the data.

The models, named Model 0–8, are stated in Table 2, and were chosen by a model selection process described in Section 3.4. Model 0 was a preliminary binomial model used for selecting Bayesian priors and is analyzed in Section S.3 (Supplementary). In all models the response (x in (1) and (2)) was the number of owned households, *owners*, and the size parameter (N in (1) and (2)) was the estimated sample size of the group, *sampsize*. The logit home ownership rate (θ in (1) and (2)) was a linear function of the predictors, where we use the standard definition $\text{logit}(\theta) = \log(\theta/(1-\theta))$. In Models 1–7, the log precision (ϕ in (2)) was a linear function of *logpop*, the log of the number of households in the county and group. We used only *logpop* as

Table 2: GLM formulations. The response distribution is binomial for Model 0 and beta-binomial for all others. We use the standard Wilkinson-Rogers notation for model formulation (Wilkinson and Rogers, 1973).

Model	Proportion (logit(θ))	Precision (log(ϕ))
0	$state + group + hsedu + unemp + income$	N/A
1	$state + group + hsedu + unemp + income$	$logpop$
2	$state + group + hsedu : group + unemp : group + income : group$	$logpop$
3	$state + group + hsedu : group + income : group$	$logpop$
4	$state + group + income : group$	$logpop$
5	$state + group$	$logpop$
6	$group + income : group$	$logpop$
7	$state + group + income$	$logpop$
8	$state + group + income : group$	1

a covariate for the precision in the interest of parsimony and due to the natural interpretation of precision being positively associated with larger population (and, hence, larger sample size). Model 8 had a constant precision parameter.

The beta-binomial model has been used previously in applications in healthcare data (Najera-Zuloaga et al., 2018), microbiome data (Hu et al., 2018; Martin et al., 2020), and many others (Ascari and Migliorati, 2021). As in our study, a key motivation for using the beta-binomial model is that it provides a better fit than the binomial model for some count data (Haseman and Kupper, 1979) and allows us to preserve the sample size information in the data (Martin et al., 2020). In addition to its flexibility, the beta-binomial model can be interpreted as modeling binary outcomes when the outcomes within each observation are correlated (Prentice, 1986). The correlation between outcomes is related to the precision parameter, with higher precision associated with lower correlation. In our case, this would naturally be interpreted as the correlation of ownership status within households of the same county and group. Following this reasoning, our models predict that higher population is associated with higher precision and lower correlation. A possible interpretation is that in regions with more households, each household will on average have less correlation with other households in the same region (e.g., because they have less contact with each other on average).

3.3 Model Fitting

Models 1–8 were fit using using the brms (Bürkner, 2017) R package. This package is an interface for the probabilistic programming language Stan (Carpenter et al., 2017), which fits models using Markov chain Monte Carlo (MCMC) simulation. All exploratory models were fit using four MCMC chains with 1,000 warmup iterations (we use the terminology of (Gelman et al., 2013, p. 282)) and 1,000 post-warmup iterations, resulting in 4,000 post-warmup posterior draws. The final model was fit using four MCMC chains with four times as many iterations, resulting in 16,000 posterior draws. No thinning of the draws was performed. To check for chain convergence, we examined the $\hat{R}$, the convergence criteria described in (Gelman et al., 2013, p. 285), which summarizes the stationarity and well-mixing of the chains. We had $\hat{R} < 1.002$ for all parameters

in the final model. Vehtari et al. (2021) recommend $\hat{R} < 1.01$ (closer to 1 is better). The bulk effective sample size (bulk-ESS), which summarizes the stability of estimates of central tendencies (e.g., mean and median), was $\geqslant$ 1,000 for all parameters in the final model. The tail-ESS (Vehtari et al., 2021), which summarizes the stability of estimates for extreme tendencies (e.g., extreme quantiles), was $\geqslant$ 2,000 for all parameters in the final model. Vehtari et al. (2021) recommend bulk- and tail-ESS > 400 (larger is better). Model 0 was fit using maximum likelihood estimation (MLE) with the base R function `glm` (R Core Team, 2021).

3.4 Model Selection

The models in Table 2 were specified by hand in a step-wise fashion to determine the best model. The main model selection criteria used was the LOOIC (leave-one-out information criteria) described in (Gelman et al., 2013, p. 175), which quantifies how well the model would perform on out-of-sample data (lower values indicate better predictive performance). The LOOIC and its standard error were computed using the loo (Vehtari et al., 2022) R package and are described in Vehtari et al. (2017). In addition to the LOOIC, we used our subjective interpretation of parameters to select a reasonable model.

Model 1 was specified as the base model with all numerical and categorical predictors with no interactions. Model 2 extended Model 1 by adding interactions of the numerical predictors with the group, and significantly improved the LOOIC. The posterior credible intervals in Model 2 showed that the unemployment variable was not significant. Thus, we removed the unemployment variable to obtain Model 3, which again did not significantly worsen the LOOIC. However, Model 3 predicted a negative association of high-school education with the home ownership rate in most groups, which we could not explain. Thus, the high-school education was removed to obtain Model 4, which did not significantly worsen the LOOIC. To check whether the remaining variables were significant in Model 4, we dropped each variable and checked whether the LOOIC significantly worsened. Only Model 7, the model with the income-group interaction removed, had a similar LOOIC. However, because exploratory plots (Figure 1) showed evidence of income-group interaction and the credible intervals of these interaction terms in Model 4 were significant, we decided to keep Model 4 as the final model. The LOOIC values for the fitted models are shown in Section S.4 (Supplementary).

3.5 Bayesian Priors and Sensitivity Analysis

We used weakly informative prior distributions for our model parameters. The prior distribution of all parameters of the logit home ownership rate, $\text{logit}(\theta)$, was $\text{N}(0, 0.25)$ (where $\text{N}(\mu, \sigma)$ is the normal distribution with mean μ and standard deviation σ), reflecting the prior assumption that the variables are not associated with the home ownership rate. However, we set the prior of the logit home ownership rate Intercept to $\text{N}(\text{logit}(0.64), 0.25)$ to reflect the national home ownership rate of approximately 64% in 2020. The prior distribution of the log population for the log precision, $\log(\phi)$, was also $\text{N}(0, 0.25)$, again reflecting our prior ambiguity about the association of population. However, we set the prior of the log precision Intercept to $\text{N}(3.25, 0.25)$ using an analysis of the binomial model residuals described in Section S.3 (Supplementary). Prior predictive checks were used to confirm that the priors reasonably mimicked the data (see Section S.5 (Supplementary)). The value of the standard deviation, $\sigma = 0.25$, was chosen heuristically based on these checks.

We also performed a sensitivity analysis of the priors by fitting Model 1 with the following

alternative choices of priors.

- Lower precision alternative: replace the log precision Intercept prior with $\mathrm{N}(1, 0.25)$.
- Higher precision alternative: replace the log precision Intercept prior with $\mathrm{N}(5, 0.25)$.
- Non-informative alternative: replace all priors for both the logit home ownership rate and the log precision with $\mathrm{N}(0, 100)$.

To compare the resulting posteriors we used the following method. Let $\{\hat{\theta}_k\}_k$ and $\{\hat{\theta}'_k\}_k$ be the posterior predictive means of the home ownership rates for the primary and alternative priors, respectively. We then compute the root-mean-square-error ($RMSE$) and Pearson correlation (r) as

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{\theta}_i - \hat{\theta}'_i\right)^2},$$

$$r = \frac{\sum_{i=1}^{N}\left(\hat{\theta}_i - \bar{\hat{\theta}}\right)\left(\hat{\theta}'_i - \bar{\hat{\theta}}'\right)}{\sqrt{\sum_{i=1}^{N}\left(\hat{\theta}_i - \bar{\hat{\theta}}\right)^2 \cdot \sum_{i=1}^{N}\left(\hat{\theta}'_i - \bar{\hat{\theta}}'\right)^2}},$$

where $N = 8{,}820$ is the number of observations in the model. We had $RMSE \leqslant 0.01$ and $r \geqslant 0.998$ for all such pairings of the primary with an alternative prior, confirming that the posterior was not highly sensitive to the choice of prior distribution. This is likely due to the large sample sizes in the US Census data causing the likelihood to dominate the posterior.

3.6 Model Diagnostics

We used posterior (prior) predictive checks for model diagnostics. Posterior (prior) predictive checks make use of replicate data sets, which are obtained by using posterior (prior) parameter draws and simulating new sets of observations using the likelihood model, as described in (Gelman et al., 2013, ch. 6.3) and Gabry et al. (2019). Using several replicate data sets, we plotted the distribution of the home ownership rates or a summary statistic thereof. These plots were produced using the ggplot2 (Wickham, 2016) R package, but were based on the functionality of the bayesplot (Gabry and Mahr, 2022) R package. The summary statistics used were the weighted mean and variance, weighted by the sample sizes, and were computed using the matrixStats (Bengtsson, 2017) R package. Manipulations of the posterior (prior) draws were done using the posterior (Bürkner et al., 2023) R package. Please see Section S.5 (Supplementary) for the model diagnostic results.

4 Results

The fitted coefficient means, standard deviations, and 95% credible intervals for Model 4 are shown in Table 3. The baseline category for the racial/ethnic group was WhiteNH, so each of the group coefficients, Hispanic, Black, and Asian, can be interpreted as a difference from the WhiteNH group. The state categorical variables are omitted for brevity and because they were mainly used for controlling for differences between states, but are included in Section S.6 (Supplementary).

As suggested by our exploratory analyses in Section 2, the fitted coefficients predict that the order of the groups from highest to lowest home ownership rate is WhiteNH, Asian, Hispanic, and Black. The model predicts that this association holds even when controlling for income.

Table 3: Posterior coefficient estimates for Model 4 (omitting the state indicator variables). "Mean" is the posterior mean, "SD" is the standard deviation, and "2.5%" and "97.5%" are the corresponding quantiles. The notation "A:B" is the interaction term between the variables "A" and "B". The parameter θ is the home ownership rate and ϕ is the precision.

(a) Coefficients at log-odds (for θ) and log (for ϕ) scale.

Component	Name	Mean	SD	2.5%	97.5%
logit(θ)	*Intercept*	1.03	0.04	0.96	1.11
logit(θ)	*Hispanic*	−1.28	0.05	−1.37	−1.18
logit(θ)	*Black*	−1.47	0.04	−1.56	−1.39
logit(θ)	*Asian*	−1.20	0.06	−1.31	−1.08
logit(θ)	*income*:*WhiteNH*	0.26	0.05	0.16	0.35
logit(θ)	*income*:*Hispanic*	0.91	0.07	0.78	1.05
logit(θ)	*income*:*Black*	0.76	0.07	0.62	0.89
logit(θ)	*income*:*Asian*	0.97	0.06	0.86	1.09
log(ϕ)	*Intercept*	0.10	0.06	−0.02	0.21
log(ϕ)	*logpop*	0.34	0.01	0.32	0.35

(b) Coefficients at odds-ratio (for θ) and ratio (for ϕ) scale.

Component	Name	Mean	SD	2.5%	97.5%
$\theta/(1-\theta)$	*Intercept*	2.81	0.11	2.60	3.04
$\theta/(1-\theta)$	*Hispanic*	0.28	0.01	0.25	0.31
$\theta/(1-\theta)$	*Black*	0.23	0.01	0.21	0.25
$\theta/(1-\theta)$	*Asian*	0.30	0.02	0.27	0.34
$\theta/(1-\theta)$	*income*:*WhiteNH*	1.29	0.06	1.18	1.42
$\theta/(1-\theta)$	*income*:*Hispanic*	2.50	0.17	2.18	2.86
$\theta/(1-\theta)$	*income*:*Black*	2.14	0.15	1.87	2.44
$\theta/(1-\theta)$	*income*:*Asian*	2.65	0.16	2.35	2.97
ϕ	*Intercept*	1.10	0.07	0.98	1.24
ϕ	*logpop*	1.40	0.01	1.38	1.42

As expected, income is positively associated with home ownership rate. However, we see from the interaction terms that this association is significantly stronger for non-White groups, with the WhiteNH groups having a log-odds effect size in $[0.16, 0.35]$ and non-White groups having log-odds effect sizes in $[0.62, 1.09]$. We also see that, as expected, the precision is positively associated with the log-population. Since the precision, ϕ, has a log-link, this can be interpreted as $\phi \propto pop^{0.34}$, where *pop* is the population in number of households.

Table 4 shows the variables' predicted effects at the proportion scale for two counties. The estimated standard deviation of the posterior predictive home ownership rate is relatively large (4.72%-8.41%), so that a large change in income such as $+\$50,000$ is needed to significantly shift the posterior distribution. This ability to more accurately estimate the posterior predictive variance is a key feature of the beta-binomial model. While the predicted effects for both counties are similar, the fitted home ownership rates are different, reflecting the differences in income and

Table 4: Model 4 (final model) effects at the proportion scale for two counties. "HOwnR" is the home ownership rate. The "Empirical" columns are empirical values from the data. "Fitted" and "SD" are the posterior predictive home ownership rate mean and standard deviation, respectively. The remaining columns show the predicted effect on the home ownership rate for different changes in the income.

(a) Hillsborough County, Florida.

Group	HOwnR Empirical	HOwnR Fitted	HOwnR SD	Income Empirical	Income +$10k	Income +$25k	Income +$50k
WhiteNH	68.9%	78.6%	4.72%	$72k	0.43%	1.11%	2.07%
Hispanic	50.4%	57.6%	6.50%	$50k	2.25%	5.48%	10.72%
Black	40.0%	49.5%	7.04%	$44k	1.94%	4.66%	9.43%
Asian	59.5%	67.6%	8.41%	$84k	2.02%	5.08%	9.63%

(b) New York County, New York.

Group	HOwnR Empirical	HOwnR Fitted	HOwnR SD	Income Empirical	Income +$10k	Income +$25k	Income +$50k
WhiteNH	32.7%	71.1%	4.99%	$130k	0.51%	1.34%	2.57%
Hispanic	8.5%	42.3%	6.38%	$43k	2.29%	5.69%	11.29%
Black	10.4%	35.4%	6.62%	$40k	1.83%	4.50%	8.97%
Asian	23.9%	56.3%	6.96%	$91k	2.25%	5.84%	11.49%

state. However, when comparing the fitted and empirical home ownership rates for New York County, we see a shortcoming of the model. It appears to significantly overestimate the home ownership rate in urban counties with high population densities, which tend to have relatively low home ownership rates (Delgadillo, 2009; Moore, 1991). We discuss a possible remedy to this shortcoming in Section 5.

5 Conclusion

In this study, we collected data from the US Census American Community Survey (ACS) and performed an observational analysis of home ownership in the US. In our exploratory analyses (Section 2), we observed noteworthy differences in the home ownership rates across different racial/ethnic groups and states. Our goal was to model the association of home ownership rates with other sociodemographic factors in the ACS data. To do so, we formulated several beta-binomial generalized linear models (Section 3). The beta-binomial model was chosen in order to account for over-dispersion and heteroscedasticity in the response compared to a binomial model. Our models were fit with Bayesian estimation, which offers flexibility in fitting distributional models and estimating posterior intervals for effect sizes at transformed scales. We performed model selection and found that the state, group, and income variables appeared to be significant. In the final fitted model, we interpreted the predictions and found that a relatively large increase in income (about $50,000) is predicted to be associated with a significant increase in the home ownership rate (about 11%) in some counties.

The main thrust of this article was to apply the Bayesian beta-binomial model to study home ownership in the US. In a previous article (Medri and Channagiri, 2022), we applied several different response distributions when modeling state-level data, such as a normal response for the logit-transformed home ownership rates, beta response for the home ownership rates, and binomial response for the count of home owners. For several reasons, we prefer the beta-binomial. The logit transform cannot be used for county-level data, where many observations with small sample sizes have 0% or 100% for the home ownership rate. The beta model allows for a flexible variance structure through its precision parameter, but does not inherently account for sample sizes in its formulation. The binomial model accounts for sample sizes, but has an inflexible variance structure, and was shown to be a poor fit in Section S.3 (Supplementary). Because the beta-binomial distribution is a mixture of the binomial and beta distributions, it both includes the sample sizes in its formulation and has a precision parameter. We accounted for heteroscedasticity by modeling both the proportion and precision parameters of the distribution. A shortcoming of this analysis was that the sample sizes and home owner counts for each group were not directly provided by the ACS summary tables, but were estimated using the sample sizes, group proportions, and group home ownership rates for each county (see Section 3.1). However, for future work, the individual survey responses are available from the American Community Survey Microdata (US Census Bureau, 2021b), from which true count data may be obtained. The MicroData may also be used to get more precise data, such as the amount of overlap of the Hispanic group with the Black and Asian group, or tabulations for non-Hispanic Black and non-Hispanic Asian populations.

In this study, we explored several sociodemographic factors and found that state, racial/ethnic group, and income had the strongest association with the home ownership rate. Notably, we also found evidence of an interaction effect between group and income, with the income having a stronger effect for non-White groups. This may indicate that as income increases, the White/non-White gap of home ownership rate decreases, and that lower incomes may be a cause of the lower home ownership rates in minority groups. We also found indication that additional predictors may be needed to form a better fitting model. In particular, we observed that our model poorly predicts the home ownership rate in highly urban counties, such as New York County, New York (see Section 4). The model tends to overestimate the home ownership rate in such counties, because they have a relatively high income but a relatively low home ownership rate. Thus, further predictors, such as the population density (Delgadillo, 2009), may be useful to better control for these differences between counties.

Supplementary Material

We have included a separate Supplementary section with additional discussion of the data, modeling analyses and results, and a description of a Shiny app developed in R for data exploration. The app features a user-friendly web interface created with the R packages Shiny (Chang et al., 2022) and shinyWidgets (Perrier et al., 2023), enabling users to perform customized and interactive explorations of the data and models presented in this work. The current version of the interface was initially showcased at the American Statistical Association (ASA) Data Challenge Expo 2022 (in the Joint Statistical Meeting (JSM) 2022), and was subsequently refined and expanded in this article. The code for the Shiny app and all our analyses may also be found at https://github.com/jmedri/JSM2022_HomeOwnership.

Acknowledgement

We thank our faculty advisor Dr. Lu Lu for generously giving her time to provide valuable feedback throughout this project. We thank the administrative staff and faculty in the USF Department of Mathematics & Statistics and the USF Graduate School for providing funding and making the travel arrangements to attend JSM 2022. We thank the Sections on Statistical Computing, Statistical Graphics, and Government Statistics of the ASA for creating this event and giving us the opportunity to participate.

Funding

The USF Department of Mathematics & Statistics Department and the USF Graduate School provided the travel funding to attend JSM 2022.

References

Aronowitz M, Golding EL, Choi JH (2020). *The Unequal Costs of Black Homeownership*. Massachusetts Institute of Technology, Golub Center for Finance and Policy, Cambridge, MA.

Ascari R, Migliorati S (2021). A new regression model for overdispersed binomial data accounting for outliers and an excess of zeros. *Statistics in Medicine*, 40(17): 3895–3914. https://doi.org/10.1002/sim.9005

Austin TM, Felicity S (1999). *Mortgage Lending Discrimination: A Review of Existing Evidence*. The Urban Institute.

Bureau UC (2011). Overview of Race and Hispanic Origin: 2010. Technical report. US Census Bureau.

Bürkner PC (2017). brms: An R package for Bayesian multilevel models using stan. *Journal of Statistical Software*, 80(1): 1–28.

Bürkner PC, Gabry J, Kay M, Vehtari A (2023). posterior: Tools for working with posterior distributions. R package version 1.4.1.

Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, et al. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76: 1–32.

Chang W, Cheng J, Allaire J, Sievert C, Schloerke B, Xie Y, et al. (2022). shiny: Web Application Framework for R. R package version 1.7.2.

Choi JH, McCargo A, Neal M, Goodman L, Young C (2019). Explaining the Black-White Homeownership Gap, Washington, DC: Urban Institute. Retrieved March, 25: 2021.

Dahl DB, Scott D, Roosen C, Magnusson A, Swinton J (2019). xtable: Export Tables to LaTeX or HTML. R package version 1.8-4.

Delgadillo L (2009). A model of factors correlated to homeownership: The case of Utah. *Family and Consumer Sciences Research Journal*, 30: 3–36. https://doi.org/10.1177/1077727X01301001

Deslatte A, Tavares A, Feiock RC (2018). Policy of delay: Evidence from a Bayesian analysis of metropolitan land-use choices. *Policy Studies Journal*, 46(3): 674–699. https://doi.org/10.1111/psj.12188

Erdoğdu H, Erdem N, Nacar F (2021). Housing appraisal under model uncertainty: Bayesian model averaging method. *Advanced Engineering Journal*, 1(1): 26–34.

Ferrari S, Cribari-Neto F (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 31(7): 799–815. https://doi.org/10.1080/0266476042000214501

Flippen CA (2010). The spatial dynamics of stratification: Metropolitan context, population redistribution, and black and Hispanic homeownership. *Demography*, 47(4): 845–868. https://doi.org/10.1007/BF03214588

Gabry J, Mahr T (2022). bayesplot: Plotting for Bayesian Models. R package version 1.9.0.

Gabry J, Simpson D, Vehtari A, Betancourt M, Gelman A (2019). Visualization in Bayesian workflow. *Journal of the Royal Statistical Society. Series A. Statistics in Society*, 182(2): 389–402. https://doi.org/10.1111/rssa.12378

Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB (2013). *Bayesian Data Analysis*, Third edition. *Chapman & Hall/CRC Texts in Statistical Science*. Chapman & Hall/CRC, Philadelphia, PA,

Goodman LS, Mayer C (2018). Homeownership and the American dream. *The Journal of Economic Perspectives*, 32(1): 31–58. https://doi.org/10.1257/jep.32.1.31

Haseman JK, Kupper LL (1979). Analysis of dichotomous response data from certain toxicological experiments. *Biometrics*, 35(1): 281–293. https://doi.org/10.2307/2529950

Bengtsson H (2017). matrixStats: Functions that Apply to Rows and Columns of Matrices (and to Vectors). R package version 0.52.2. Available at https://github.com/HenrikBengtsson/matrixStats.

Hu T, Gallins P, Zhou YH (2018). A zero-inflated beta-binomial model for microbiome data analysis. *Stat*, 7(1): e185.

Hui SK, Cheung A, Pang J, et al. (2010). A hierarchical Bayesian approach for residential property valuation: Application to Hong Kong housing market. *International Real Estate Review*, 13(1): 1–29. https://doi.org/10.53383/100117

Jones LD (1989). Current wealth and tenure choice. *Real Estate Economics*, 17(1): 17–40. https://doi.org/10.1111/1540-6229.00471

Krivo LJ, Kaufman RL (2004). Housing and wealth inequality: Racial-ethnic differences in home equity in the United States. *Demography*, 41(3): 585–605. https://doi.org/10.1353/dem.2004.0023

Kuebler M (2013). Closing the wealth gap: A review of racial and ethnic inequalities in homeownership. *Sociology Compass*, 7(8): 670–685. https://doi.org/10.1111/soc4.12056

Martin BD, Witten D, Willis AD (2020). Modeling microbial abundances and dysbiosis with beta-binomial regression. *Annals of Applied Statistics*, 14(1): 94–115.

Mast BD (2010). Measuring neighborhood quality with survey data: A Bayesian approach. *Cityscape*, 12(3): 123–142.

Medri J, Channagiri T (2022). Exploratory Analysis of Racial Representation in American Home Ownership. In: *JSM Proceedings. Statistical Computing Section*, 983–1013. American Statistical Association, Alexandria, VA.

Moore DJ (1991). Homeownership affordability series forecasting the probability of homeownership: A cross-sectional regression analysis. *Journal of Housing Research*, 2(2): 125–143. Publisher: American Real Estate Society.

Müller K, Wickham H (2023). tibble: Simple Data Frames. R package version 3.2.1.

Najera-Zuloaga J, Lee DJ, Arostegui I (2018). Comparison of beta-binomial regression model approaches to analyze health-related quality of life data. *Statistical Methods in Medical Research*, 27(10): 2989–3009. https://doi.org/10.1177/0962280217690413

Ospina R, Ferrari SLP (2012). A general class of zero-or-one inflated beta regression mod-

els. *Computational Statistics & Data Analysis*, 56(6): 1609–1623. https://doi.org/10.1016/j.csda.2011.10.005

Paleologos EK, Elhakeem M, Amrousi ME (2018). Bayesian analysis of air emission violations from waste incineration and coincineration plants. *Risk Analysis*, 38(11): 2368–2378. https://doi.org/10.1111/risa.13130

Pebesma E (2018). Simple features for R: Standardized support for spatial vector data. *The R Journal*, 10(1): 439–446. https://doi.org/10.32614/RJ-2018-009

Perrier V, Meyer F, Granjon D (2023). shinyWidgets: Custom Inputs Widgets for Shiny. https://github.com/dreamRs/shinyWidgets, https://dreamrs.github.io/shinyWidgets/.

Prentice RL (1986). Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors. *Journal of the American Statistical Association*, 81(394): 321. https://doi.org/10.1080/01621459.1986.10478275

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Robb BH (2021). Council Post: Homeownership and the American Dream. https://www.forbes.com/sites/forbesrealestatecouncil/2021/09/28/homeownership-and-the-american-dream/.

Sanchez-Moyano R (2021). Geography and Hispanic homeownership: A review of the literature. *Journal of Housing and the Built Environment*, 36(1): 215–240. https://doi.org/10.1007/s10901-020-09745-5

Tennekes M (2018). tmap: Thematic maps in R. *Journal of Statistical Software*, 84(6): 1–39. https://doi.org/10.18637/jss.v084.i06

Thomas AF (2021). The racial wealth gap and the tax benefits of homeownership. *The New York Law School Law Review*, 66: 247.

US Census Bureau (2015-2020). American Community Survey. Tables S1501, S1903, S2301, S2501, S2502, S2503, S2506, S2507, DP05. Technical report, US Census Bureau.

US Census Bureau (2021a). Data Suppression. Technical report, US Census Bureau. Available at https://www.census.gov/programs-surveys/acs/technical-documentation/data-suppression.html.

US Census Bureau (2021b). Understanding and Using the American Community Survey Public Use Microdata Sample Files: What Data Users Need To Know. Technical report, US Census Bureau. Available at https://www.census.gov/content/dam/Census/library/publications/2021/acs/acs_pums_handbook_2021.pdf.

Vehtari A, Gabry J, Magnusson M, Yao Y, Bürkner PC, Paananen T, et al. (2022). loo: Efficient Leave-One-Out Cross-Validation and WAIC for Bayesian Models. R package version 2.5.1.

Vehtari A, Gelman A, Gabry J (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27: 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Vehtari A, Gelman A, Simpson D, Carpenter B, Bürkner PC (2021). Rank-normalization, folding, and localization: An improved R^ for assessing convergence of MCMC (with discussion). *Bayesian Analysis*, 16(2): 667–718. Publisher: International Society for Bayesian Analysis. https://doi.org/10.1214/20-BA1221

Wickham H (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, New York.

Wickham H (2022). stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.5.0.

Wickham H (2023). forcats: Tools for Working with Categorical Variables (Factors). R package version 1.0.0.

Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, et al. (2019). Welcome to the tidyverse. *The Journal of Open Source Software*, 4(43): 1686. https://doi.org/10.21105/joss.01686

Wickham H, François R, Henry L, Müller K (2022). dplyr: A Grammar of Data Manipulation. R package version 1.0.10.

Wickham H, Hester J, Bryan J (2023a). readr: Read Rectangular Text Data. R package version 2.1.4.

Wickham H, Vaughan D, Girlich M (2023b). tidyr: Tidy Messy Data. R package version 1.3.0.

Wilkinson GN, Rogers CE (1973). Symbolic description of factorial models for analysis of variance. *Journal of the Royal Statistical Society. Series C. Applied Statistics*, 22(3): 392–399.