Final Project Report:
An Analysis of Homeownership in the U.S. Using Bayesian Statistical Methods

Saron Tefera, Gathungu Ndirangu, Cristian Mojica, Brian Dong

May 12, 2022

I. Application

All questions posed for the analysis of homeownership in the United States come from two sources created in 2019, namely the Homeowner Assistance Fund (HAF) and the Federal Reserve. This latter source was reduced down to data for the fourth fiscal quarter of 2019 since it contains county-level debt-to-income (DTI) ratio data from each fiscal quarter since 1999. Initial exploration of the HAF data revealed that most variables were dependent on others, so in studying variables, there was the potential to create custom variables such as ratios of two variables that represented counts of homeowners that satisfied certain criteria such as earning less than 100% median area income, or being White or non-White. The Federal Reserve data is a simple dataset that identifies counties and gives their lowest and highest DTI observations for the fourth fiscal quarter of 2019. From this, we realized the potential for posing questions on such phenomena as spending power, loan qualification and cost burden, and effect of race on foreclosure rate, as these are real forces felt and considered by agents, realtors, buyers, and sellers in the real estate market. We filtered down research ideas to the following three questions.

Question 1 explores the relationship between foreclosure rates predicted by the Urban Institute, proportion of homeowners earning less than 100% median income for their area, and DTI ratio (a typical metric for loan qualification and underwriting). The method utilized to answer this question was Bayesian linear regression, primarily implemented by Stan via the rstanarm package. We hoped to answer questions involving low and high DTI ratios, not only high, to look at the least aversive conditions that are conducive to homeownership and to the conditions that make homeownership more taxing, perhaps from institutions such as the Federal Reserve and banks that adjust interest rates and allow homeowners to carry more debt. However, imputation of missing data in the low and high DTI data did not make sense considering that imputed median values in high DTI would often be less than the low DTI in those same counties, which does not make sense. It was also considered that low DTI and high DTI might be closely related and introduce multicollinearity in the regression models we would create. We ultimately choose to study only high DTI in each county and to scale this predictor for ease of interpretation and convergence in our implementation of Bayesian regression.

The findings of this research can be the basis for advice in the process of purchasing real estate. Commonly, DTI ratios are common metrics utilized to assess one’s ability to make mortgage payments, and such ratios will depend on the type of loan, as well as on its principal amount. If the DTI ratio can predict foreclosure in a given county in a given region of the United States, this information can inform loan officers and prospective homeowners as to what is an appropriate amount of debt to carry per unit of income. Also, if it is the case that the proportion of a county's homeowners that have less spending power than the top 50% of income-earning homeowners can predict foreclosure rates, there is likely some systemic factor in a county that is conducive to foreclosure, and homeowners considering moving to a new county may need to be aware of such effects of county and its wealth (or lack thereof) on the probability of foreclosure. One future question that stems from our findings (developed in Methods) include whether or not region alone has an impact on DTI ratio, or on predicted foreclosure rate alone. Another question reserved for the region of the U.S.A. known as the West asks why the proportion of homeowners earning less than 100% AMI is significant in predicting foreclosure.

Question 2 asks how the fifty states of the United States of America would rank according to the expected number of cost-burdened homeowners. We used a Poisson hierarchical model to help rank each state by county and see the expected number of cost-burdened homes at the state level. We wanted to do this because we wanted to see if population played a factor in having the highest expected cost burden on homeowners per state. We thought this may be the case because the population is the most significant factor for cost-burdened homeowners. However, we wanted to see if this is still relevant at the state level. In a future study, we may run a hierarchical beta model so that each county gets its own mean and variance. Also, by running a hierarchical beta model, we can account for the total population of homeowners per state. As for the present study, we corrected this by taking the expected count of cost-burdened homeowners and dividing it by the total population of each state to see which state had the highest percentage of cost-burdened homeowners.

In order to understand the results for this question, note two groups of the population: the total population of cost-burdened homeowners is called Group 1, and the total population of homeowners in the state is called Group 2. In studying the map of the United States in Figure 1 and the data in Figure 2, we can see that in Group 1, the top states for cost-burdened homeowners are not the states with the highest populations. This is interesting because it shows that population is not that much of a factor when just looking at the population of cost-burdened homeowners per state.

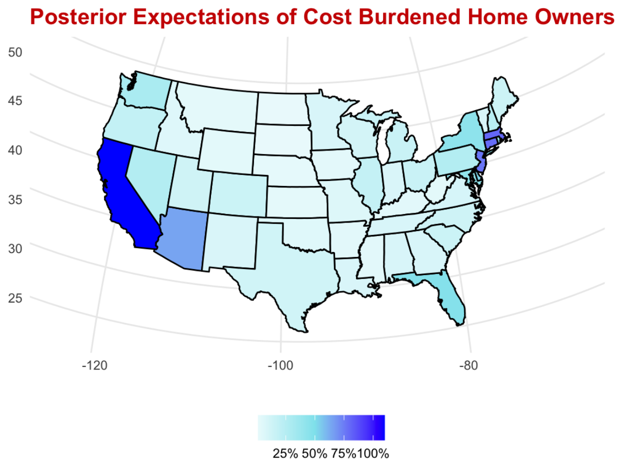
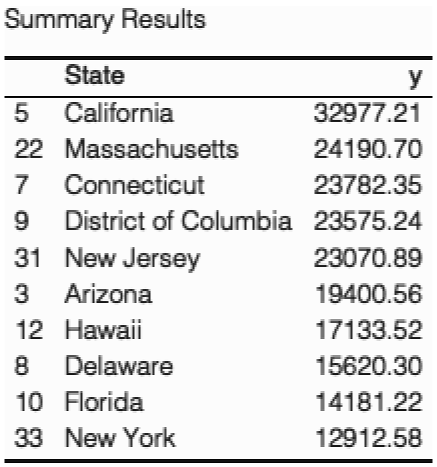


Fig. 1. Map showing states with most populated expected counts for Group 1. Light blue is a low expected count, and dark blue is a high expected count.



Summary Results

	State	y
5	California	32977.21
22	Massachusetts	24190.70
7	Connecticut	23782.35
9	District of Columbia	23575.24
31	New Jersey	23070.89
3	Arizona	19400.56
12	Hawaii	17133.52
8	Delaware	15620.30
10	Florida	14181.22
33	New York	12912.58

Fig. 2. Top 10 states with the highest expected count for cost burden homeowners for Group 1.

In studying Group 2, we look at the percentage of cost-burdened homeowners based on the expected count of cost-burdened homeowners. We get very different results. In the map of the United States in Figure 3, we see that the less-populated states are more prevalent in showing a higher percentage of cost burden homeowners than the higher-populated states. Even when we look at the table in Figure 4, we can see that none of the top populated states have the highest percentages for cost-burdened homeowners for Group 2. This can be because it is easier to see cost-burdened homeowners better in less populated states than in higher-populated states. After all, in the high-populated states, many counties might balance out counties with high counts of cost-burdened homeowners. This also touches on our method’s limitations: since we have an equal mean and variance, the bigger states might not show the proper count of cost-burdened homeowners in high-populated states.

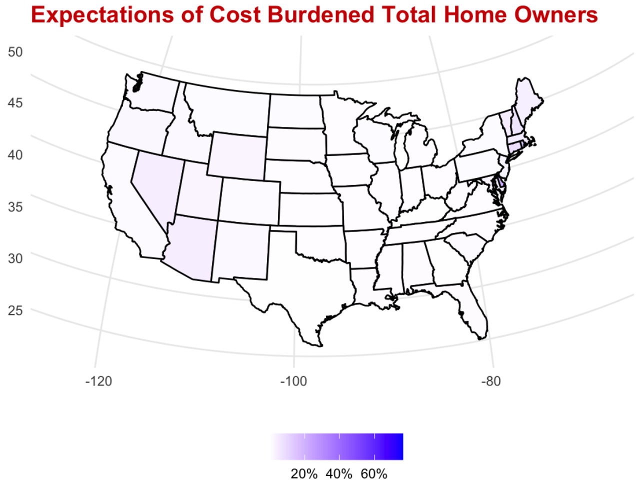
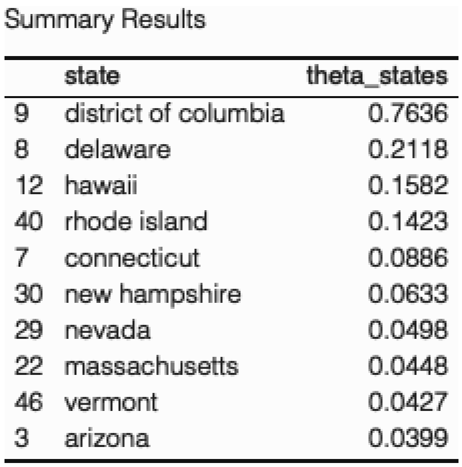


Fig. 3. Percentage of the cost burden on homeowners in every state based on the total population of homeowners per state. White means the low percentage of cost-burdened homeowners, and purple means the high percentage of cost burden homeowners.


Summary Results

	state	theta_states
9	district of columbia	0.7636
8	delaware	0.2118
12	hawaii	0.1582
40	rhode island	0.1423
7	connecticut	0.0886
30	new hampshire	0.0633
29	nevada	0.0498
22	massachusetts	0.0448
46	vermont	0.0427
3	arizona	0.0399

Fig. 4. Top 10 percentages of cost burden homeowners based on total population per state.

Although our results show that population is not a significant factor for cost-burdened homeowners at a state level, many other economic factors are at play here. Based on our study, if a new homeowner is looking for a state to live in and to not worry about becoming a cost burden to homeowners, we would suggest not residing in Delaware, Connecticut, or Hawaii, and we recommend that new homeowners live in North Dakota, Nebraska, or Virginia. If we did a similar project, we would use a hierarchical beta model to account for all the different mean and variances for each county. Some questions to answer in the future entail looking at big states and seeing how many counties actually have a high percentage of being cost-burdened homeowners. Also, looking at the life satisfaction for cost-burdened homeowners in different states, is being a cost-burdened homeowner in a given state a better experience than in another state? These would be interesting questions to answer in future studies.

Question 3 explores the relationship between foreclosure rates predicted by the Urban Institute and race (defined as White or non-White), in particular those who are making between 100% and 150% of area median income. We wanted to see whether or not race had significance in predicting foreclosure rate given the fact that the homeowners are in the same median income bracket. In addition, we utilized median owner cost as a predictor. This question was answered by regression using a comparison of ordinary least squares (frequentist) regression and Bayesian multiple regression. We used a hierarchical model based on a Gamma inverse normal distribution. The reason for doing this is because we accounted for most of the variables to follow a level of normality with the errors being independent, and the variance would be constant. Through this application we were able to answer the question of how much White ownership really had an impact on predicted foreclosure rate, but we could not infer statistically significant effects when we included the median ownership cost. One of the reasons we were not able to answer this was because median cost of ownership had missing values or lack of imputed values that affected the way in which we set up our prior. Our answer to this question has potential implications with respect to being an evidentiary point for redlining, a practice used by certain banks and financial groups within the United States and the utilization of race as a form of initial belief (or prejudice) considered to be possibly unfair and non-impactful with respect to businesses' and banks' expectations for turnout. In light of our findings we could, in the future, use a model similar to the one developed here to examine credit default, which is somewhat similar in methodology since it demands a form of (logistic) regression that can also have Bayesian priors adjusted for it.

II. Methods

Question 1

The two methods primarily considered from this question were Bayesian regression and Bayesian ANOVA. If this question were to explore the effect of region of the United States on predicted foreclosure rate, and if we were interested in comparing various models to determine which is more likely to explain observed data, then this would have been an appropriate method. Also studying DTI ratio and proportion of homeowners earning less than 100% area median income, however, would have called for some version of ANOVA that exceeds the two-way ANOVA. This question has no explicit interest in the effect of region on foreclosure rate, and we seek to find regression parameters informed by observed data and prior beliefs rather than compare models, so a more reasonable method would be Bayesian linear regression, which seeks to create posterior distributions for the regression parameters we are interested in estimating, and which works well for continuous predictors and responses.

To appreciate the phenomena of foreclosure, high debt-to-income ratios, and lower earning power, consider these visualizations (Figures 5, 6, and 7) of the dependent variable and independent variables of interest, prior to transformation.

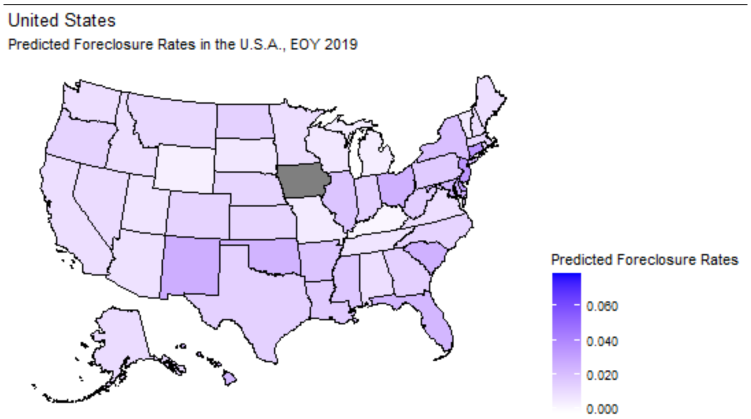

United States
Predicted Foreclosure Rates in the U.S.A., EOY 2019
Predicted Foreclosure Rates
0.060
0.040
0.020
0.000

Fig. 5. Urban Institute-predicted foreclosure rates in the United States, end of 2019.

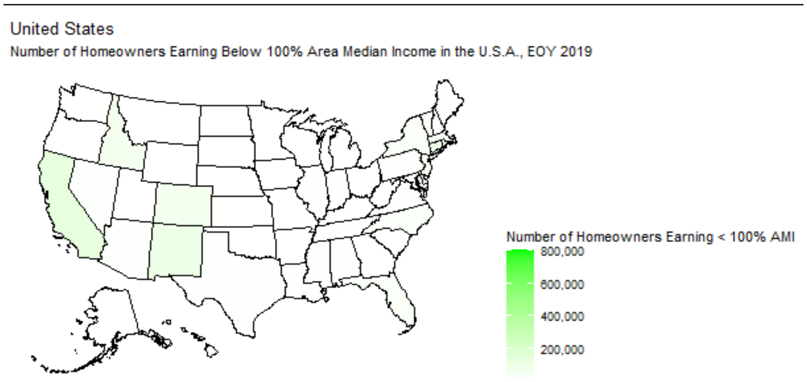

United States
Number of Homeowners Earning Below 100% Area Median Income in the U.S.A., EOY 2019
Number of Homeowners Earning < 100% AMI
800,000
600,000
400,000
200,000

Fig. 6. Number of homeowners earning below 100% of area median income in the United States, end of 2019.

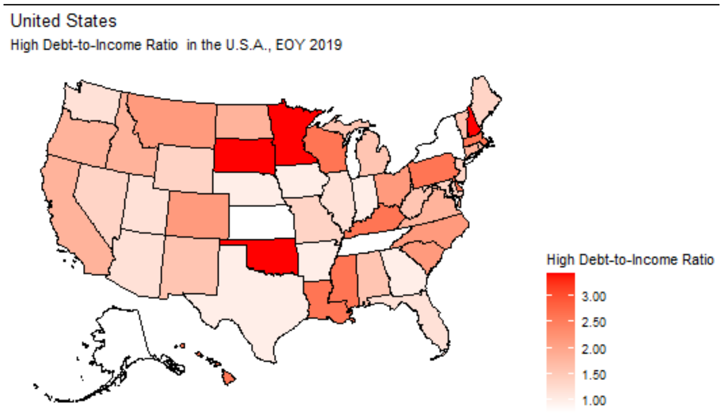

United States
High Debt-to-Income Ratio in the U.S.A., EOY 2019
High Debt-to-Income Ratio
3.00
2.50
2.00
1.50
1.00

Fig. 7. High Debt-to-Income Ratio in the United States, end of 2019.

We created five models, one for the entire United States and one for each of the four regions of the United States recognized by the U.S. Census Bureau; each model regressed Urban Institute-predicted foreclosure rates onto scaled proportion of homeowners earning less than 100% of area median income and scaled high DTI ratio. We obtained OLS estimates for each of these predictors to help us understand what to expect in Bayesian analysis, and that preliminary analysis demonstrates the statistical insignificance of scaled proportion of homeowners earning less than 100% AMI in the whole U.S.A. ($t(3133) = 0.926$, $p = 0.3547$), the Northeast ($t(214) = -1.174$, $p = 0.2418$), and the South ($t(1417) = 0.751$, $p = 0.4529$). In the Midwest model, the only significant predictor was the intercept ($t(1051) = 45.887$, $p < 0.001$), and the only insignificant predictor in the West model was scaled high DTI ratio ($t(442) = 4.473$, $p < 0.001$); all other predictors were significant. With uninformative priors on these variables, and with only some data-dependent information imposed on the prior for intercept, we should expect inferences from Bayesian analysis to be similar to those inferences made in the frequentist paradigm.

We interfaced with Stan using the rstanarm package, which has tools to create Bayesian generalized linear models; to specify linear regression, we set the “family” argument to its default value of ‘gaussian’. We specified a standard normal prior on our variance term and non-intercept regression parameters, and a prior normal distribution for the intercept, with the mean being the midpoint of minimum and maximum predicted foreclosure rates for each region, and the variance being 0.01. We ran four chains running 10,000 iterations of our models. After that, we assessed posterior distributions for variance and for each predictor for each model, as well as other diagnostics including traceplots (all indicating convergence), effective sample size (all exceeding 1,000), and Rhat values (all around 0.99 or 1). We looked at posterior means and credible intervals as our quantities of interest. We found the posterior mean for predicted foreclosure rate for the entire U.S.A. ($M = 0.013$, $SD = 0.0002$), the Northeast ($M = 0.019$, $SD = 0.0007$), the Midwest ($M = 0.012$, $SD = 0.0004$), the South ($M = 0.014$, $SD = 0.0003$), and the West ($M = 0.013$, $SD = 0.0005$).

The results from this first question indicate that with little to no prior information, the OLS and Bayesian estimates of regression parameters are highly similar to each other. The (scaled) proportion of homeowners earning less than 100% of a given area’s median income is not a significant factor in the predictor foreclosure rate that a given area is assigned, national or regional. Only in the West does the number 0 escape the 95% credible interval for scaled proportion. For regions that are not the West, this suggests that given our prior beliefs, it is highly likely that this scaled proportion has zero or negligible effect on predicted foreclosure rate. Across regions, there is considerable variability in the signs of the endpoints of 95% credible intervals for scaled high DTI ratio. This behavior suggests that posterior distributions for scaled DTI ratio are likely to depend on region, suggesting an effect of region on DTI ratio, which in turn may have an effect on predicted foreclosure rate.

This method has its advantages and limitations. Missing data were imputed by median imputation, so it is possible for certain effects of the predictors on foreclosure to inflate or deflate because of this imputation. The primary advantages to this method, however, are its ease and familiarity, and the fact that we are able to draw from the posterior distribution of each predictor, and this gives readily interpretable regression coefficients that are informed by observed data and prior beliefs.

Question 2

Hierarchical modeling works by parameterizing across multiple groups and describing the average group mean and the differences across group means by a sampling model. It helps to build reliable models that account for the variation of the influence of the predictors across clusters or groups that may form in a dataset due to some shared properties. Simply put, we used hierarchical modeling because hierarchical modeling allowed us to compare across multiple groups within a multilevel dataset. Our data was structured in a way where we have a population of states, with numerous counties (subdividing each state) that have a certain number of cost-burdened homeowners (a two-level dataset). We also wanted to utilize hierarchical modeling because of shrinkage (i.e., it takes the sample sizes of the groups into account). The sample sizes for each state depend on the number of counties in the state. The sample sizes for each state vary a great deal considering the different sizes of each state. For example, the smallest sample size was for Washington D.C., which has one county. For this reason, shrinkage was an important factor in deciding on a hierarchical model.

For our second question, we used the Poisson hierarchical model proposed below:

$$y_i \sim Poisson(\theta_i)$$
$$\theta_i \sim Gamma(\alpha, \beta)$$
$$\alpha \sim Uniform(0,\ a_0)$$
$$\beta \sim Uniform(0,\ b_0)$$

The variable y_i represents the observed number of cost burdened homeowners in county *i*. θ_i is the expected number of cost burdened homeowners. For each county we assumed that the observed number of cost-burned homeowners had a Poisson distribution whose mean is the expected number of cost-burdened homeowners. For θ_i we used a conditionally conjugate prior, the Gamma prior, with some values α and β. For the α and β values, we assumed uniform priors. This dataset has a large number of counties (over 3,000), so α and β will be well-estimated by the model. For that reason we felt a uniform prior would be suitable. The uniform priors for α and β have the values a_0 and b_0. The values a_0 and b_0 were chosen and were set large enough so that they didn't constrain α and β. Essentially, to check that a_0 and b_0 were not constraining α and β, we ran our code and got samples from the posterior distributions for α and β. We proceeded to plot histograms of those samples and made sure the range was wide enough to not constrain the posterior of α and β. We used the posterior expectations calculated for each state to see which state has the highest number of cost-burdened homeowners, then we ranked the states with the lowest expected amounts of cost-burdened homeowners highest.

Shrinkage is a major advantage of hierarchical modeling in some special cases, but it may be regarded as “unfair.” For example, a ranking based on the posterior expectation of our model may not align with a ranking based on sample counts. Although one state may have a lower sample count of cost-burdened homeowners than some other state, the posterior expectation of our model may not show the same results due to shrinkage. The states with more evidence of a lower number of cost-burdened homeowners will rank higher than those with less evidence. It is a bit of a disadvantage to smaller states which will always have less evidence because of their smaller numbers of counties. As for other limitations, we did not consider the total population of homeowners in the United States when running the Poisson hierarchical model. Therefore, the population considered in the dataset is only of the cost-burdened homeowners per state. We also realized that since we are running a Poisson regression model, all the variances and the means are the same for every county. Since we did not account for the total population of each state, this may have skewed our results.

Question 3

The primary methods used for this question were Bayesian multiple regression and Bayesian model comparison utilizing Bayesian Information Criteria. We compare these with the frequentist OLS method. This question does not examine specific foreclosure as a value of “yes” or “no” (i.e., as a binary response variable), so logistic regression is not necessary in this situation. These methods involve using a prior to provide default or baseline analysis where we can use an identically normal intercept and independent error term.

Figure 8 shows three subfigures, one for each of three predictors which we used in the OLS method to regress predicted foreclosure rate against median income of 100 to 150 for both White and non-White homeowners, and the median home ownership cost. We see that the number of non-White homeowners earning between 100% and 150% of area median income plays a level of significance in predicting foreclosure rate along with median home owner cost.

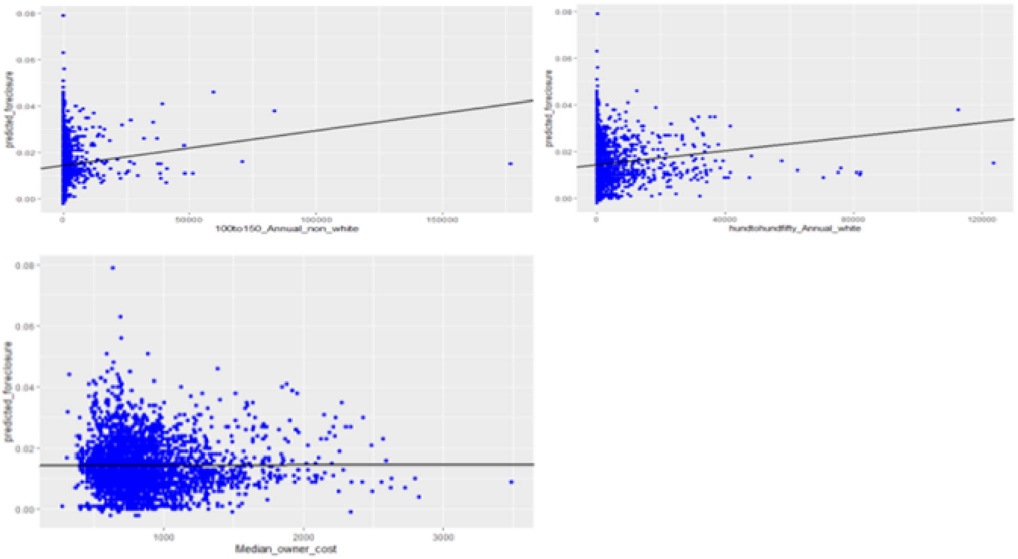


Fig. 8. Regressing predicted foreclosure rate onto various predictors. Top left: The predictor is the number of non-White homeowners earning between 100% and 150% of area median income. Top right: The predictor is the number of White homeowners earning between 100% and 150% of area median income. Bottom left: The predictor is median owner cost.

Figure 9 below proposes a model that follows an inverse Gamma distribution that makes its way into a hierarchical model because we need to specify the prior distributions for all of the regression coefficients.

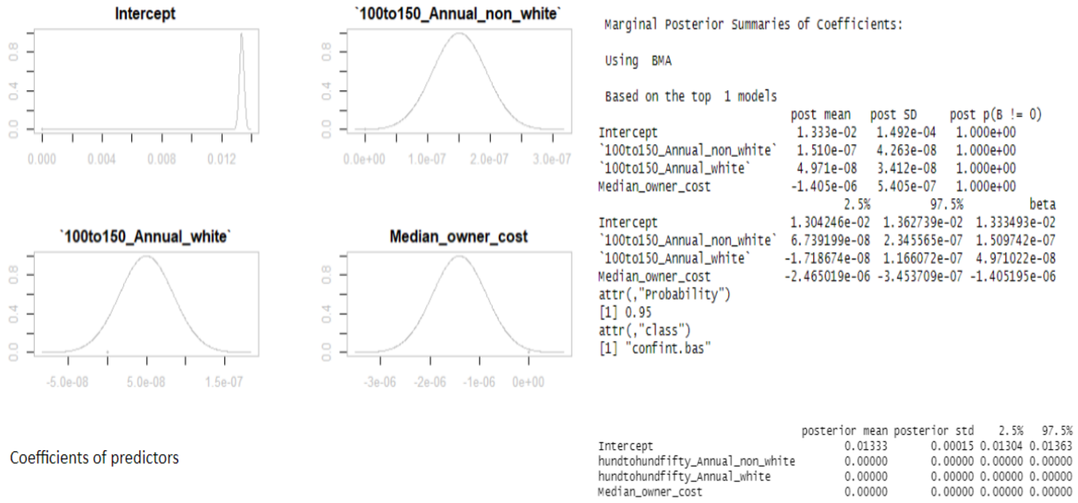
$$\beta_o\beta_1\beta_2\beta_3|\sigma^2 \sim Normal((b_0, b_1, b_2, b_3)^T, \sigma^2 \sum_0$$

$$\beta_o\beta_1\beta_2\beta_3|\sigma^2 \propto 1, \quad p(\sigma^2) \propto \frac{1}{\sigma^2}$$

Fig. 9. The hierarchical model to be incorporated into Bayesian regression for Question 3.

An informative prior which involves the median home ownership cost is further used for that. Since we do not have enough prior information about variance for median home ownership cost, we used a noninformative reference prior. Our prior distribution on all the β-values is conditioned on σ^2 associated with median home ownership cost; this is the uniform prior, and the prior of σ^2 is proportional to the reciprocal of σ^2.

The BAS library is used in this situation because it allows us to specify different model priors and coefficient priors and we can specify the response and predictor variables. We used BIC for our non-informative reference prior because the alternative (which involves AIC) would use a less conservative testing approach. We generated a summary of tables listing posterior means, standard deviations, and bounds for our intervals, and this summary is presented in Figure 10 below.



```
 Marginal Posterior Summaries of Coefficients:

 Using  BMA

 Based on the top  1 models
                            post mean   post SD     post p(B != 0)
Intercept                    1.333e-02   1.492e-04   1.000e+00
`100to150_Annual_non_white`  1.510e-07   4.263e-08   1.000e+00
`100to150_Annual_white`      4.971e-08   3.412e-08   1.000e+00
Median_owner_cost           -1.405e-06   5.405e-07   1.000e+00
                                     2.5%         97.5%          beta
Intercept                    1.304246e-02  1.362739e-02  1.333493e-02
`100to150_Annual_non_white`  6.739199e-08  2.345565e-07  1.509742e-07
`100to150_Annual_white`     -1.718674e-08  1.166072e-07  4.971022e-08
Median_owner_cost           -2.465019e-06 -3.453709e-07 -1.405195e-06
attr(,"Probability")
[1] 0.95
attr(,"class")
[1] "confint.bas"
```

```
                              posterior mean posterior std    2.5%   97.5%
Intercept                            0.01333       0.00015 0.01304 0.01363
hundtohundfifty_Annual_non_white     0.00000       0.00000 0.00000 0.00000
hundtohundfifty_Annual_white         0.00000       0.00000 0.00000 0.00000
Median_owner_cost                    0.00000       0.00000 0.00000 0.00000
```

Fig. 10. Diagnostics following Bayesian regression, Marginal Posterior Summaries of Coefficients.

The intercept posterior that we obtained from our posterior data summary is different from our OLS intercept estimate, but not by a significant amount. This is most likely from the low effects of median home ownership cost not playing a huge impact on income for both White and non-White homeowners. Looking at the posterior values for 100 to 150 annual median income of White and non-White homeowners, we can see there is no significant change and the coefficients are largely about the same. Since these distributions center the posterior distributions at the respective OLS estimates, we can fairly compare them to estimates from the frequentist method and see that they fall within the same level. Due to the effects of missing values, our posterior values were rescaled, but the outcome is interpretable. While these methods were suitable, in the future we could probably use a more ideal form of regression, one that does not rely on a normal distribution that underlies a data set with missing and imputed values.

In ultimately reflecting on the methods implemented in this research, we generally find that we have gained greater familiarity with the typical tools and methods of Bayesian inference. Questions 1 and 3 were reminiscent of simple linear regression, but in being able to think about priors on regression coefficients, we find that a regression coefficient is more than a stand-alone quantity; rather, it is a quantity that is more likely than a fixed constant to reflect our ideas of reality, and sampling for such quantities is straightforward with software such as Stan. The idea of credible intervals as being samples from posterior distributions informed by observed data and prior beliefs also became increasingly relevant over the course of answering these questions. Question 2 helped deepen our understanding of spatial statistics and how hierarchical modeling can be used to analyze large areas such as countries and states. We found that the key to using hierarchical modeling for geographic data was to analyze geographic groups such as states that most likely share common properties due to unique laws and cultural practices.

III. Balance of Effort

Saron Tefera was the timekeeper and scheduler, so she scheduled and initiated meetings for the group to meet and work on the project. She primarily answered our second research question along with Gathungu Ndirangu, and helped to facilitate recording and editing of the presentation component of this project. Gathungu was the harmonizer and communicator, so he communicated with Henry when we had any questions, issues with our analysis, and issues within the group that threatened balanced effort. In addition, he worked with Saron to develop and implement the hierarchical model of the second research question. Cristian Mojica was the deliverables manager, so he worked to the best of his ability to ensure that submitted materials met Henry's specifications. Cristian primarily formatted the final report and checked the grammar of the paper, as well as made close inspections of figures and methods before final submission of materials. He primarily answered the first question on foreclosure rates. Brian Dong focused primarily on answering the third research question, and developed and implemented the hierarchical model and marginalization necessary to do so.