A Study of Median Income across the United States

David Diaz, Noah McIntire, and Matt Rubin

ReadMe.md File and Github: https://github.com/nem2pq/DS6040-final

I. Project Description

Economic inequity is a longstanding issue in the United States that has proven difficult to solve despite decades of efforts. There is always a desire to learn and understand what factors impact various aspects of economic inequality between Americans, all of which we can measure in numerous ways. Analyzing the median household income of Americans, especially within smaller subsets, is a compelling way to understand various socioeconomic differences and demonstrate their impacts on the livelihoods of American families.

Our project goal is to help jurisdictions identify methods to reduce income gaps and eliminate poverty by modeling what impacts the median household income of Americans in each county. We can do this by creating various Bayesian models to predict the median income by county and reviewing the relationships with the predictor variables. Furthermore, this methodology will help us model the uncertainty of the relationship between these predictor variables and median income, helping to provide a range of possible relationships and metrics that smaller jurisdictions can use to assess the economic opportunity of their citizens.

II. Data Description

We utilized a Kaggle dataset containing a variety of socioeconomic and demographic factors for every county and county-equivalent in the United States. This dataset includes an aggregation of various metrics recorded by different governmental agencies. Our target variable for this study is the median household income in 2022. Some of our predictor variables included the 2016 crime rate, unemployment rate, cost of living, diversity ranking (by race), and the number of water quality violations. A complete data dictionary of the predictors is in Appendix A, Table 1.

The data cleaning process for this dataset involved several formatting changes and dropping several variables. All variables whose values were calculated by using median income, were removed. We also parsed out many numerical variables that included unnecessary symbols and had the potential to interfere with our analysis. Missing values were also present throughout the dataset, so we determined that the most effective way to manage this was to replace them with the mean value of that variable due to all missing values being in numerical columns and the assumed normal distribution of these variables (see Probability Model).

Before fitting the models on these data, we performed several data transformations. These included a logistic transformation of our predictor variable, 2022 median household income, and cost of living due to skewed distributions. Additionally, we normalized all numeric variables to allow for the use of the same prior to describe the distribution of these variables in each model. We transformed categorical variables using dummy encoding to store that information numerically.

III. Probability Model

We implemented our standard model to use all possible variables to try and develop a model that was most accurate at predicting the median income. A visual representation of the full model is present in Figure 7. We designed the model to be computationally efficient by normalizing all numerical variables, as we did for all further models. This structure also ensured that each numerical variable would have an equal impact on the final model. This understanding allowed us to use a multivariate Gaussian distribution as an informed prior for all predictors. We used a noninformative prior to enable the model to account for our intercept parameter, considering how much median income can vary over the nation. Additionally, we utilized a student-t likelihood to make the dataset more robust to extreme outliers within the dataset.

Among our probability models, the reduced model best represents the objectives of our project. We wished to fit a linear regression to predict median incomes for geographic entities across the United States. We included a visual representation of the reduced model in Figure 15, which shows how it is highly similar to our complete model but with fewer predictors. Besides the number of predictors, the probability model described for the standard model is immensely close to the reduced.

The non-linear reduced model represents our attempt at the predictive objective of our project. For this model, we made two other transformations to the predictor variables. Explicitly, we performed an inverse transformation to the unemployment and water quality columns to define the relationship between these variables and the response variable. Figure 6 shows that these variables have a right-skewed distribution, and the pairplot in Figure 1 shows that the scatterplots between these two variables and the median income response variable individually follow an inverse function pattern. Therefore, we found our transformations appropriate, scaling the transformed variables afterward. Following much of the same structure as our reduced model above for the same reasons, we fit this model intending to increase our predictive ability even further. A visual representation of this model is visible in Figure 29.

IV. Approach

By choosing several distinctive possible models to predict the median household income, we could use each to achieve different parts of our goal. Our standard model, which included many predictor variables, was used to see how accurately we could model median household income. Using almost the full dataset allowed us to model the uncertainty of the impact of various target variables on median household income within the context of many factors. This model also gave us a starting point in model building, allowing us to refine our following models more precisely.

We used a similar approach for our pseudo averaging model. We designed this model to more carefully predict the median household income while not having much interpretability. These two models could be used to predict future household incomes based on the predictor variables within the model but are not beneficial for our goal of having an interpretable model. It is important to note that due to the poor results of this model, we decided it ultimately did not represent the objective of our project. Therefore, we did not include it in the probability model section above.

Our reduced model was used to gain better insight into the impact of various target variables and focus more on the potential relationship between median household income and the most relevant variables of interest. Including at least one reduced model allows us to have a more interpretable model since the impact of our model's parameters on the response variable would be more significant without the noise of the full model. As a result, identifying the relationships between our covariates and the response variable would be more intelligible with this approach. Using a reduced model allows us to recommend metrics of interest with greater confidence to smaller counties when it comes to studying economic inequality and begin providing an understanding of what we could do to minimize it.

Our reduced non-linear model is very similar to our reduced model, but the non-linear purposefully lacks the same level of interpretability. The goal of including this model was to potentially improve the predictive ability of an individual model, build off of the reduced model, and add an extra model to our pseudo averaging model to improve predictive ability overall.

One approach we notably did not pursue when conducting this study was the construction of a hierarchical model. This choice was because our goal was to create a model that could generalize to counties all over the United States, not focus on counties in particular states. Notably, fitting a hierarchical model provides the benefit of highlighting specific findings for each state's counties. By generalizing the scope, we can determine county factors that transcend state limits and significantly impact every county in the nation.

V. Results

The full model, containing all of our predictors, did not perform as well as we had hoped, with an MSE of 3087, as seen in Figure 14. However, there are some signs that the model may be more robust than its MSE would suggest. The Bayesian p-value plot, shown in Figure 10, demonstrates that the model is a relatively good fit, with the p-value barely diverging from within the margin of error. On the other hand, only a few predictors on the forest plot (Figure 9) do not have noteworthy amounts of uncertainty. Still, it is worth noting that most numeric variables have tiny confidence intervals on this plot. The posterior predictive plot in Figure 11 also shows a strong relationship between the observed and predicted values and suggests that the sampling may not necessarily be the problem. Instead, the issue in this case is likely overfitting on the training data.

The reduced model saw the best performance of our models in fitting the data. Along with the non-linear model, it had the lowest MSE among the models at just 0.02. The beta plots in Figures 17 through 22 show that the distributions of the parameters are close to zero and nearly normally distributed, signaling effective choices of priors. Figures 26 and 27 show that our predictions for the test set were generally quite close to its actual results. Furthermore, the forest plot in Figure 23 reveals that most of the predictors, except for the ULOCALE values, have minuscule confidence intervals, indicating a high degree of certainty in their predictions. Figure 25 shows that, as with the full model, the posterior predictive mean is strikingly similar to the observed values, and the predictions are thus highly accurate. Once again, the Bayesian p-value plot, which stays within the margin of error for nearly all values, proves that the model is a good fit for predicting median income (Figure 24).

The reduced non-linear model performed just as well as the reduced model at prediction. Having an MSE of 0.02, Figures 40 and 41 show that this model's predictions were quite close to the actual observations. In addition, the uncertainty encapsulated in these predictions captures most of the observation points, with only a few points being off the mark. The forest plot in Figure 31 demonstrates that the numerical predictors, betas 0 through 6, have very narrow confidence intervals in this model, while the ULOCALE categorical variables have large confidence intervals. Furthermore, Figures 32 through 37 show the posterior distributions of the statistically significant parameters, betas 0 through 5. Figure 39 indicates that the posterior predictive mean is remarkably similar to the observed values, suggesting a highly accurate predicting ability, and the Bayesian p-value plot in Figure 38 barely diverges from within the margin of error, demonstrating a good model fit.

The Bayesian averaged model performed with an MSE of 2780.2, performing better than the full model but significantly worse than both of the reduced models. Accordingly, Figure 43 shows that this model's predictions were off by a large margin, missing the range of the actual observation values entirely. The lack of uncertainty in this figure is due to limitations we experienced with this model described in the conclusion. It is important to note that the model's weights, described in Tables 5 and 6 in Appendix D, were calculated from each of the previous model's traces using the training data and information criteria. Specifically, our model used the PSIS-LOO criteria weights, though both information criteria heavily weighed our full model. This weighting could explain the pseudo averaging model's poor MSE value compared to the reduced models. Additionally, this means that our averaged model did not give us a better way of predicting or modeling median household income and has less interpretability than our previous models. This issue goes against what we had hoped to achieve, as ensemble models are typically more accurate.

VI. Conclusions

We met almost all of our objectives by the end of this project. Based on the results, several predictors stood out as variables of interest to the median income. The variables of interest with negative relationships with median income are crime rate, unemployment, and popular vote percentage, and those with positive relationships with median income are cost of living, student-teacher ratio ranking, and racial diversity ranking. We successfully modeled a linear regression with excellent predictions for median income across the United States. These predictions have the potential to be used by poorer counties to implement policies that reduce the gap in household income with wealthier counties.

Our project had several limitations that prevented us from achieving 100% of our goals. We intended to use a Bayesian Additive Regression Tree (BART) model to use decision trees to improve our model averaging. However, due to the antiquated code of the BART module in the pmb package, we could not obtain predictions alongside the averaged model using this model's outputs and were thus forced to leave it out of the final product. Another limitation took place during our process of averaging. We wanted to carry out pseudo averaging with weights from the training data to use on our testing data, but we could not do so without manually calculating the predictions and therefore lacked any measure uncertainty in this model. This limitation was ultimately a resource limitation, as Google Colab does not utilize the latest version of Arviz, which is necessary to make precise averaged model predictions using the trace information of our individual component models. The last limitation regarded the crime rate data we utilized in our models. This data is from 2016, while the remainder of our data was from more recent years. Understanding there is a time disparity there, we assumed that the crime rate would be similar now compared to 2016 and that the overall relationships we aimed to find would not be affected. While we feel that the absence of these limitations could have slightly improved our analysis, we are satisfied with the results and conclude that our model is proficient at predicting median income.

VII. References

[1] Z. Vaughan, “City/zip/county/fips - quality of life (US),” Kaggle, https://www.kaggle.com/datasets/zacvaughan/cityzipcountyfips-quality-of-life/data (accessed Dec. 6, 2023).

VIII. Appendices

Appendix A
Exploratory Data Analysis

Table 1

Data dictionary

Column Name	**Description**
countyhelper	Unique identifier containing state and county name of row
LSTATE	2-letter abbreviation to identify the state the county is in.
NMCNTY	County or City Name.
FIPS	Federal Information Processing Standard County Code.
LZIP	Zip Code.
ULOCALE	Descriptor of urbanization level of county (urban, suburban, rural).
2022 population	Recorded County population estimate in 2022 by the US Census Bureau.
2016 Crime Rate	Reported crimes in 2016 (US Department of the Interior) over the total population of the county.
Unemployment	Percentage of the population that is unemployed.
2020PopulrVoteParty	Political party that won the popular vote in a county in the 2020 presidential election.
2020 PopulrMajor%	Percentage of the vote received by the winning party in a county in the 2020 presidential election.
AQI%Good	Air Quality Index, percentage of quality rated as good over total measured days.

WaterQualityVPV	Average number of water quality violations per visit in the county, from the Environmental Protection Agency (EPA).
NtnlPrkCnt	Count of national parks in the state that the county is in.
%CvgStatePark	Percentage of state acreage covered by national parks.
Cost of Living	Average total cost of living in the county, as reported by the Economic Policy Institute.
2022 Median Income	Average median household income in the county, as reported by the Economic Policy Institute.
AVG C2I	Average cost of living over median income.
1p0c-2p4c	Cost of living over median income for various family types.
Stu:Tea Rank	Average public school student-to-teacher ratio rank in the county.
Diversity Rank (Race)	Average public school racial diversity rank in the county.
Diversity Rank (Gender)	Average public school diversity rank in the county.

Figure 1

Pair plot of target variable median household income with various predictor variables

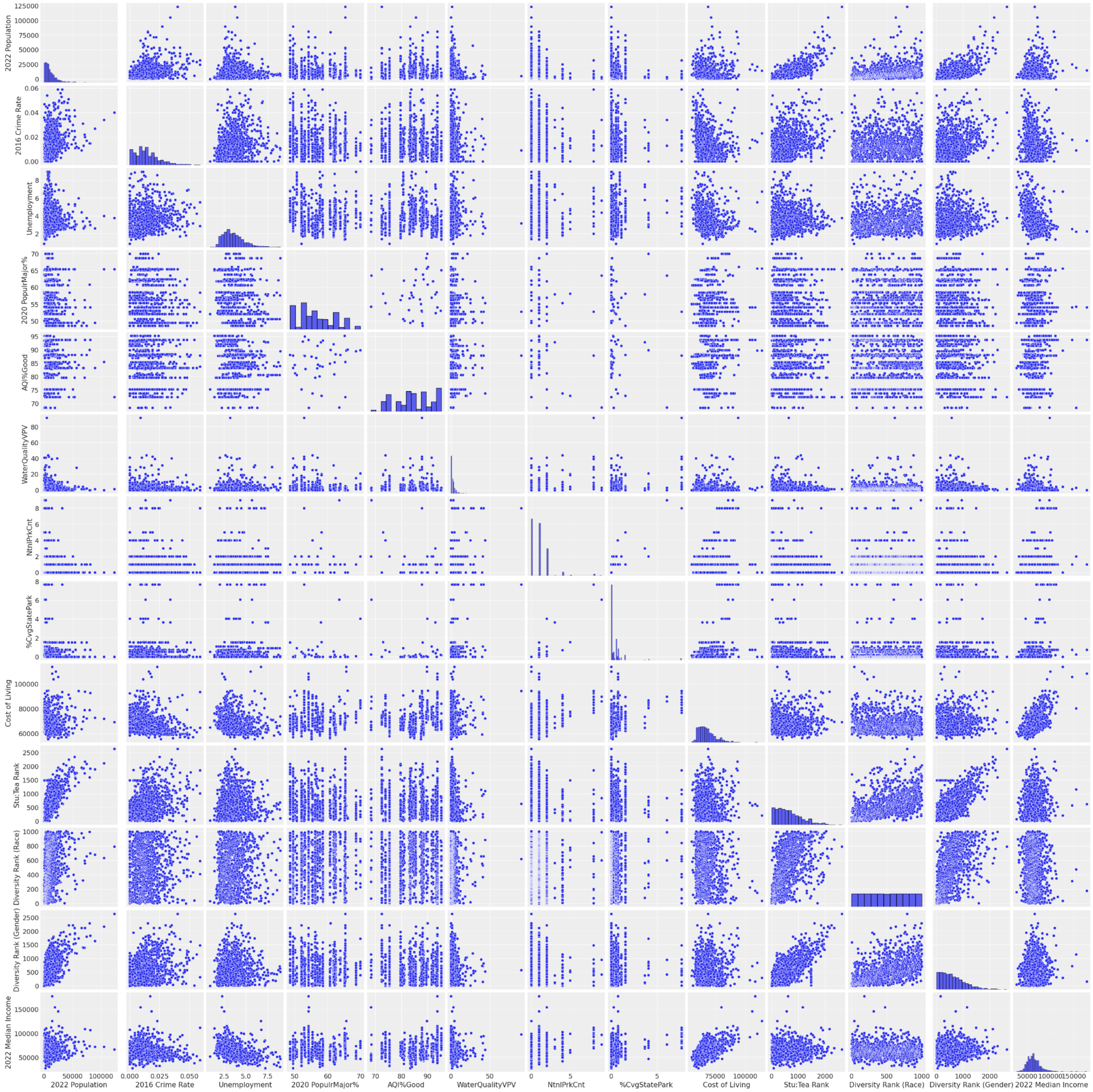


Figure 2

Average median household income by party with popular vote by county in the 2020 presidential election

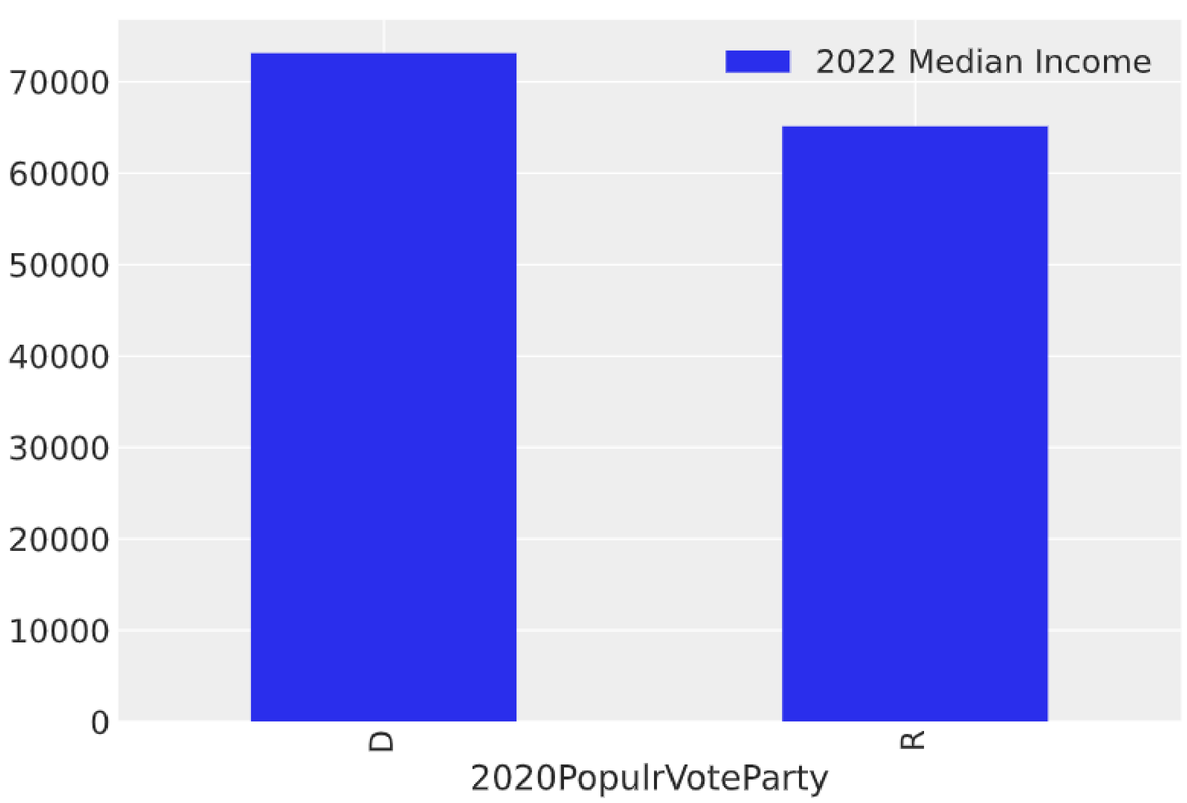


Figure 3

ULOCALE of county in each state, delineated by color

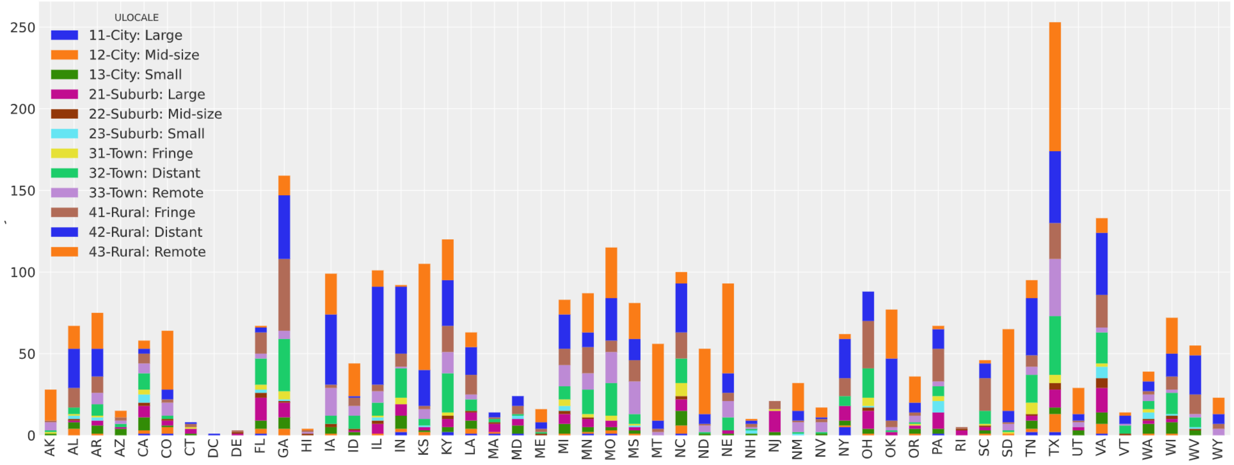


Figure 4
Average median household income by ULOCALE

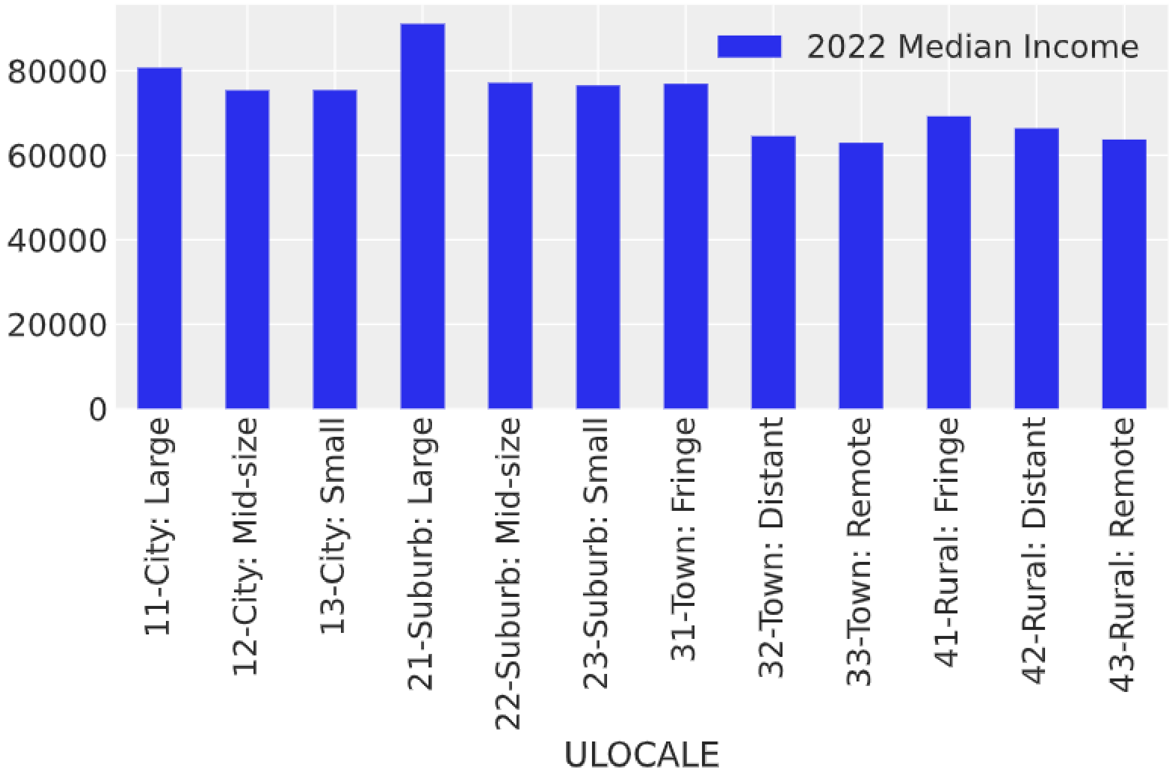


Figure 5
Average median household income by state

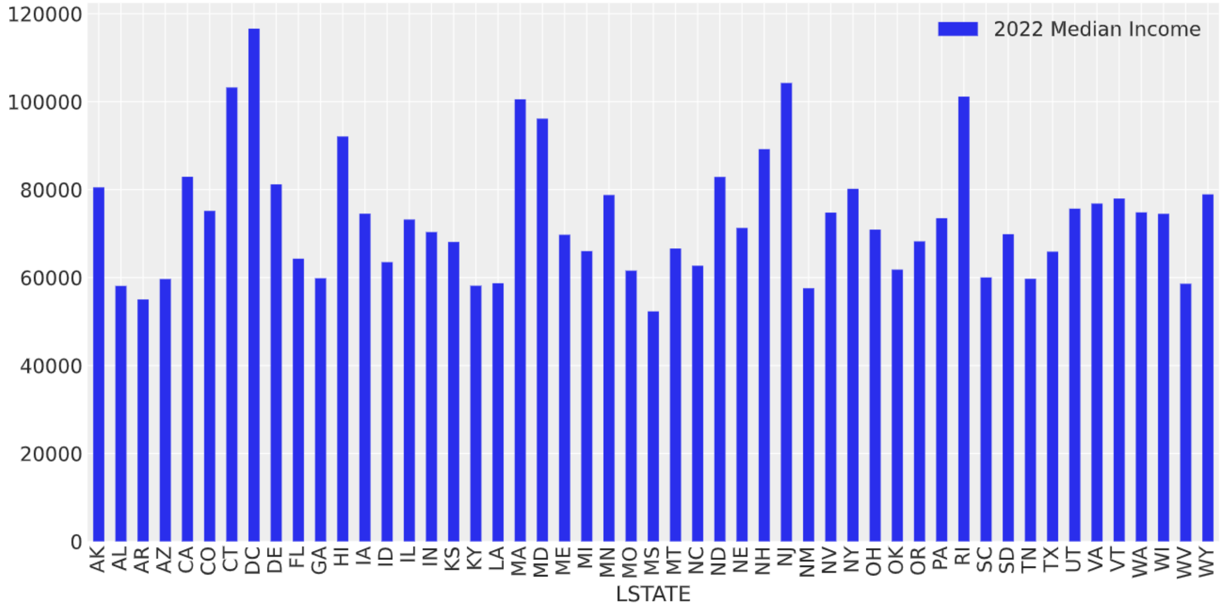
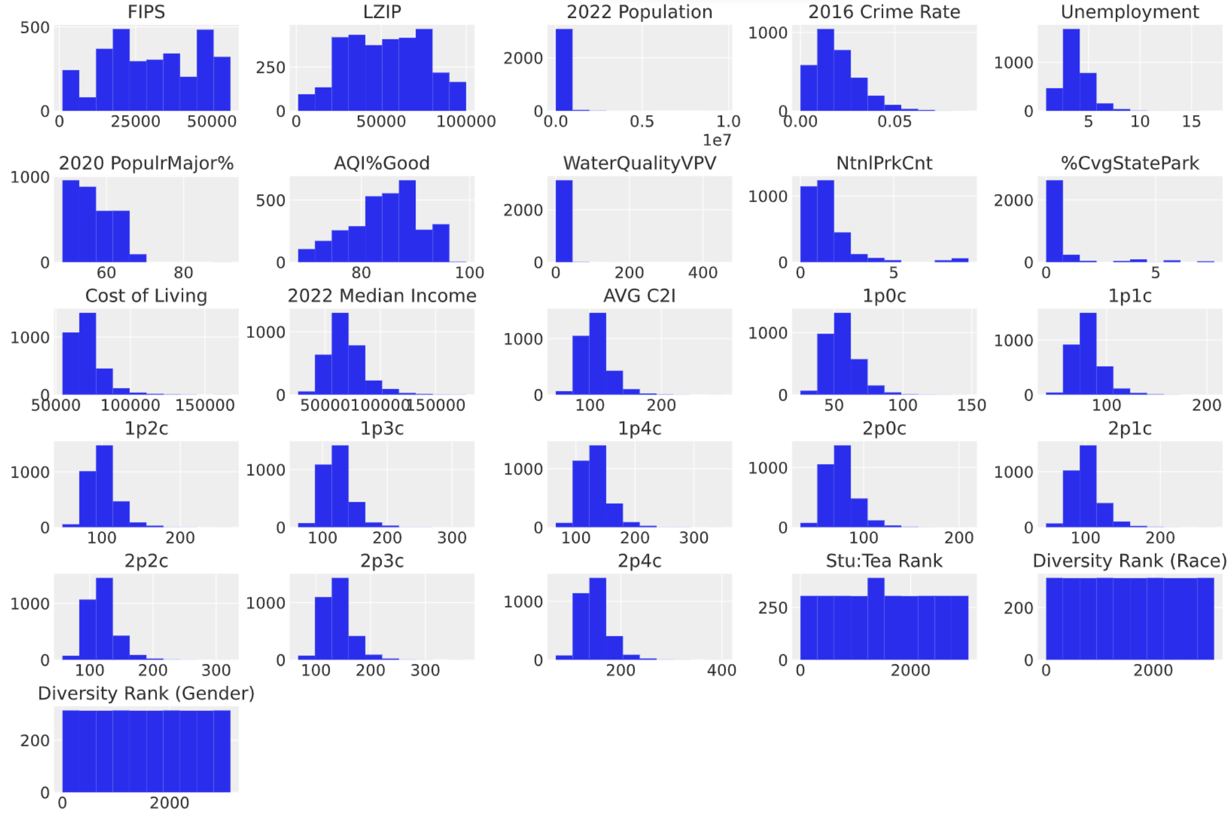


Figure 6

Histograms of variables of interest



Appendix B
Standard/Full Model

Figure 7
Visual representation of developed standard model

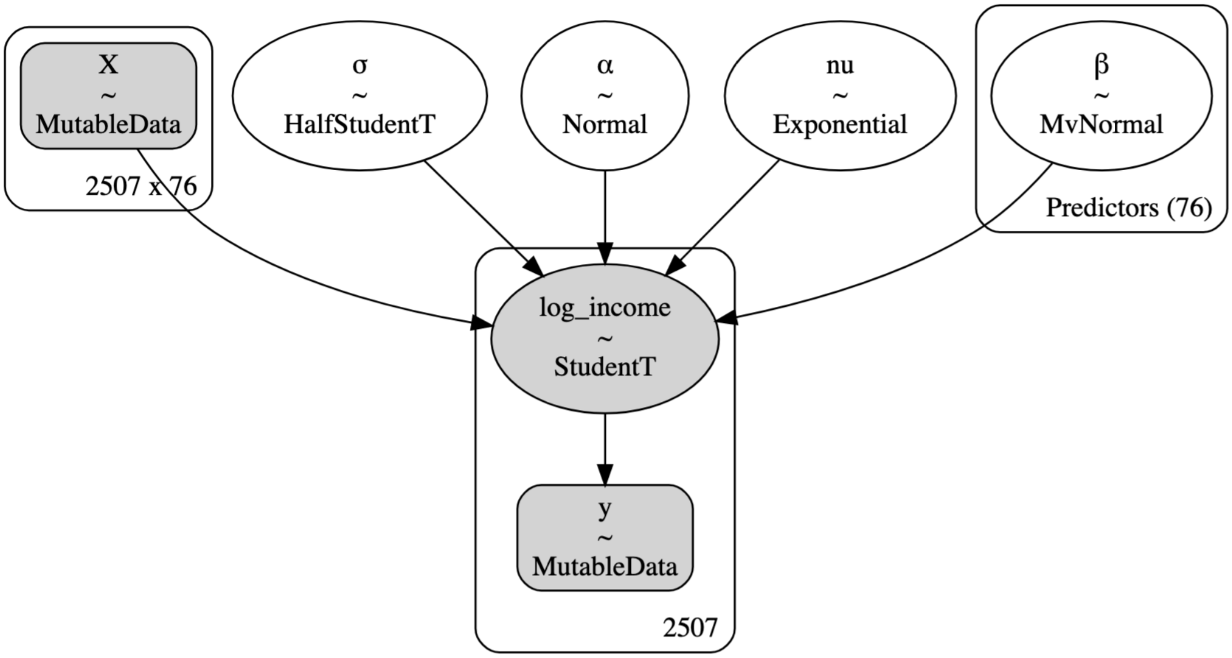


Figure 8
Trace plot of standard model to evaluation sampling

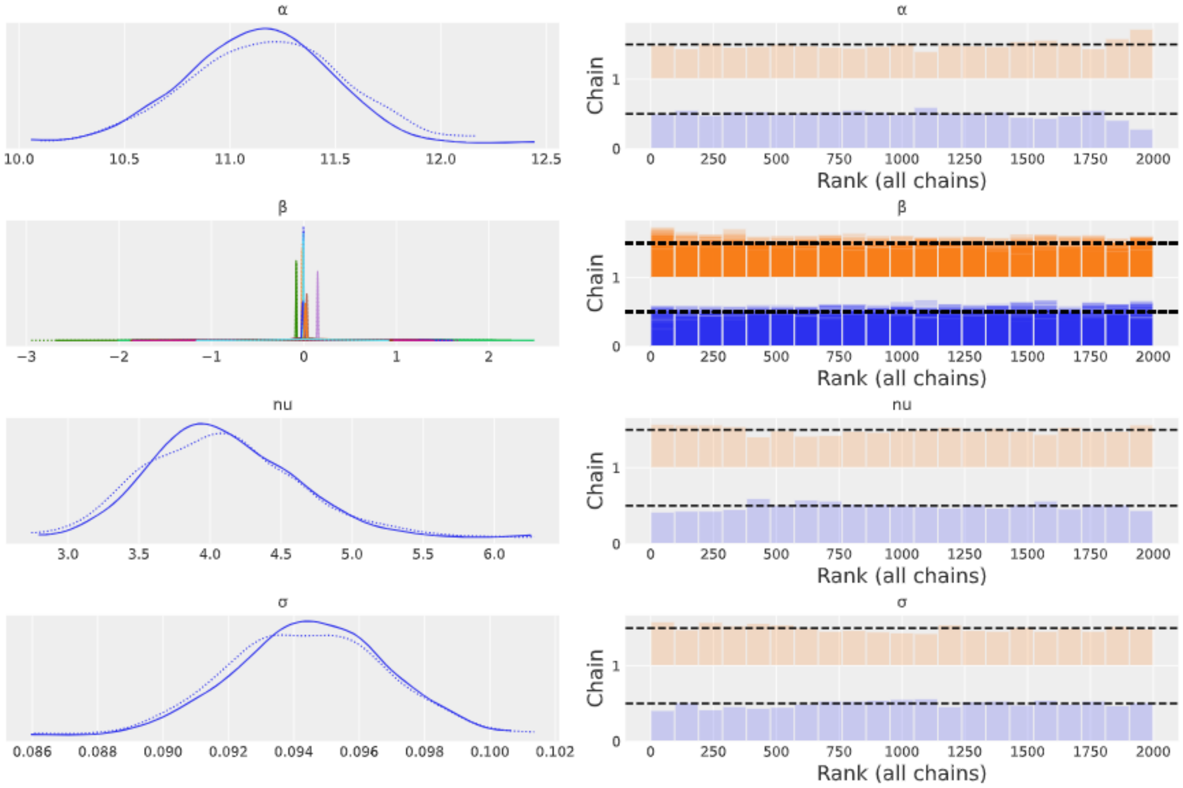
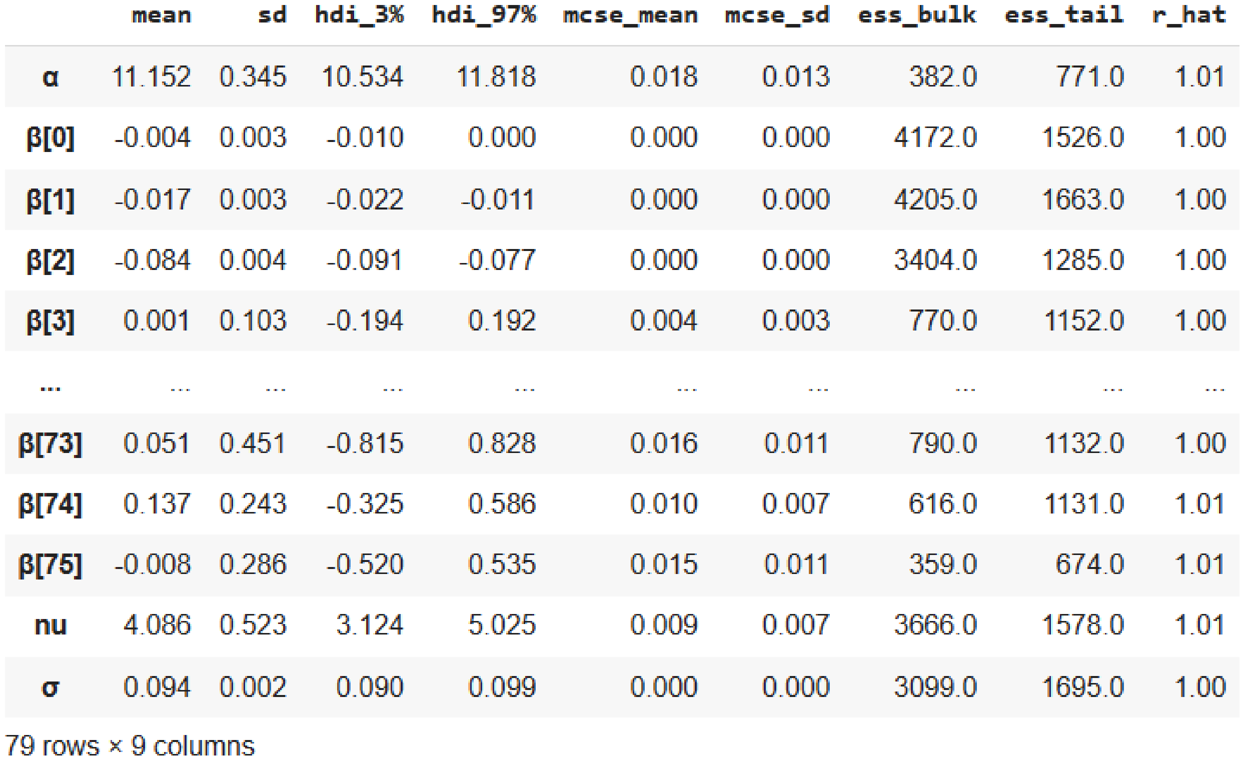


Table 2

Information on model parameters after fitting (az.summary())



	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
α	11.152	0.345	10.534	11.818	0.018	0.013	382.0	771.0	1.01
β[0]	-0.004	0.003	-0.010	0.000	0.000	0.000	4172.0	1526.0	1.00
β[1]	-0.017	0.003	-0.022	-0.011	0.000	0.000	4205.0	1663.0	1.00
β[2]	-0.084	0.004	-0.091	-0.077	0.000	0.000	3404.0	1285.0	1.00
β[3]	0.001	0.103	-0.194	0.192	0.004	0.003	770.0	1152.0	1.00
...	...	...	...	...	...	...	...	...	...
β[73]	0.051	0.451	-0.815	0.828	0.016	0.011	790.0	1132.0	1.00
β[74]	0.137	0.243	-0.325	0.586	0.010	0.007	616.0	1131.0	1.01
β[75]	-0.008	0.286	-0.520	0.535	0.015	0.011	359.0	674.0	1.01
nu	4.086	0.523	3.124	5.025	0.009	0.007	3666.0	1578.0	1.01
σ	0.094	0.002	0.090	0.099	0.000	0.000	3099.0	1695.0	1.00

79 rows × 9 columns

Figure 9

Forest of Beta Distributions

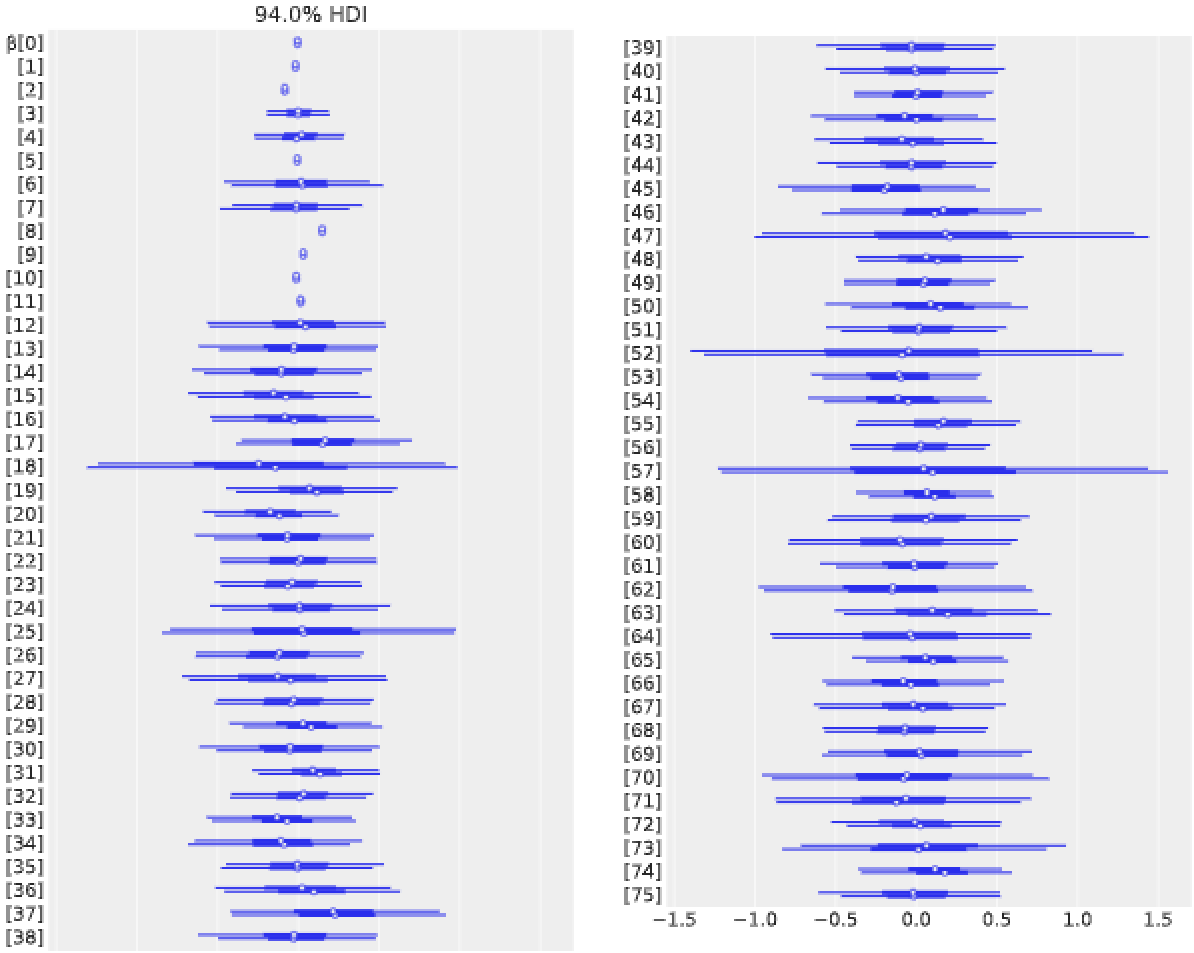


Figure 10
Bayesian P-value Plot

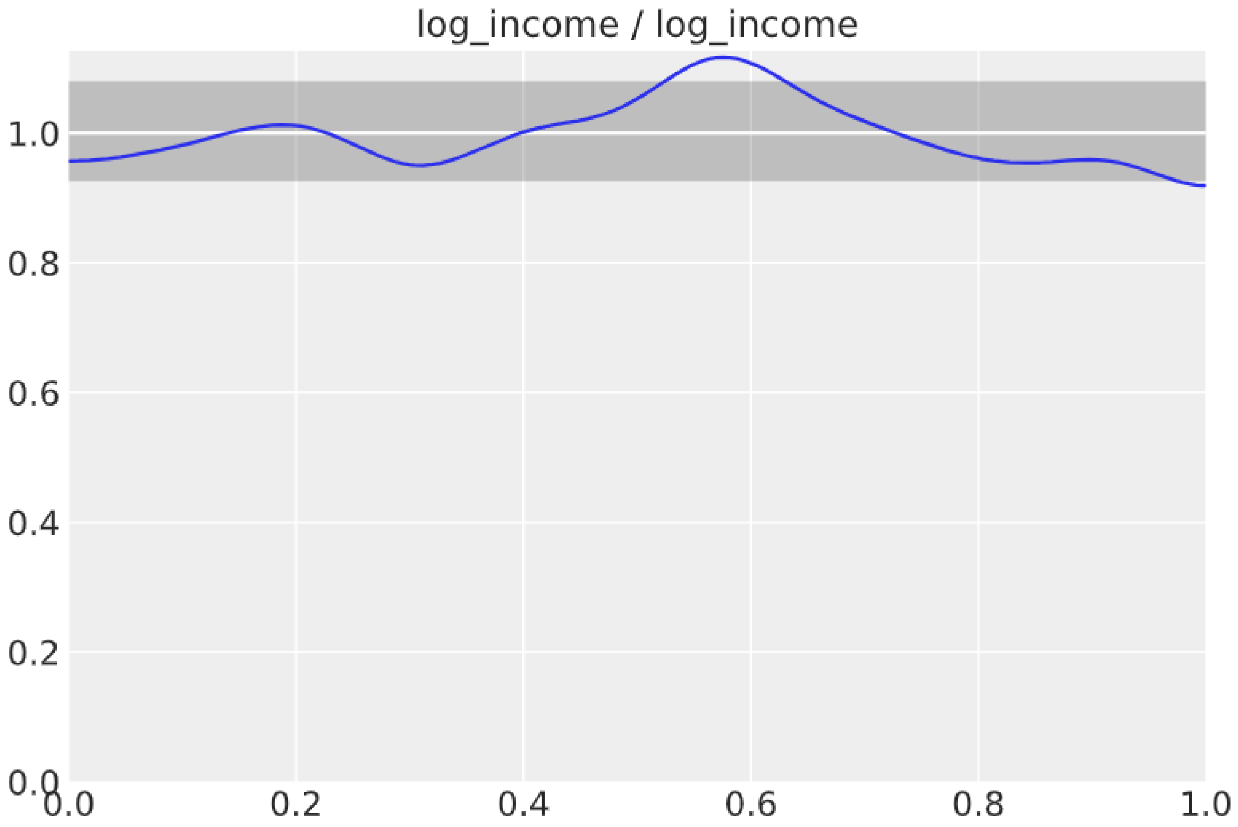
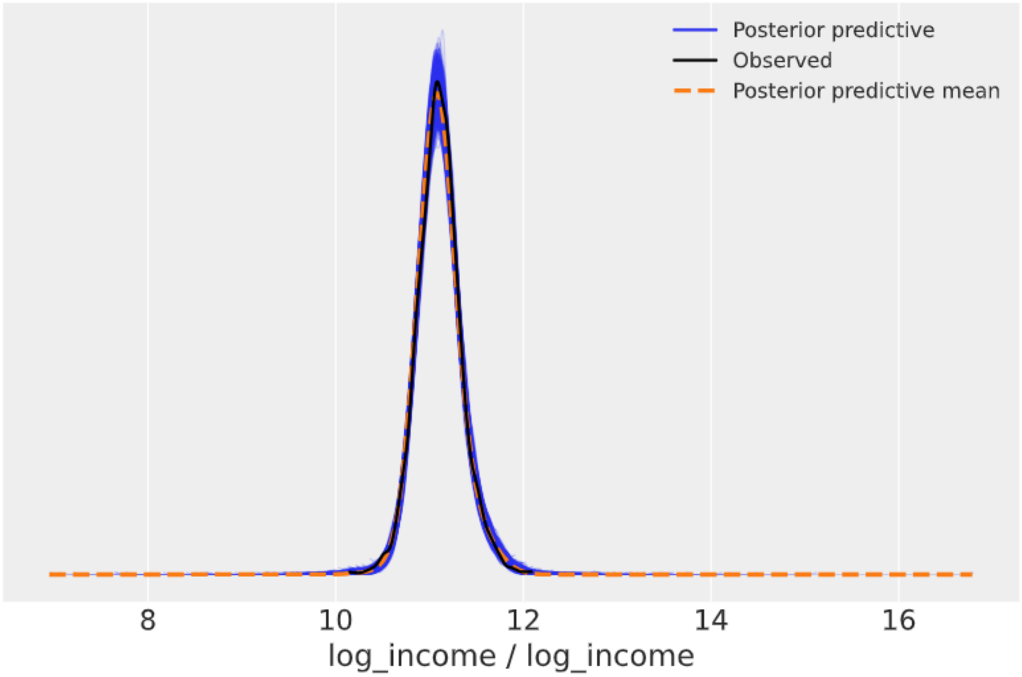


Figure 11
Posterior Predictive of Log of Median Household Income



Figures 12 and 13

Plot of Full Model Predictions showing uncertainty vs Test Set Values

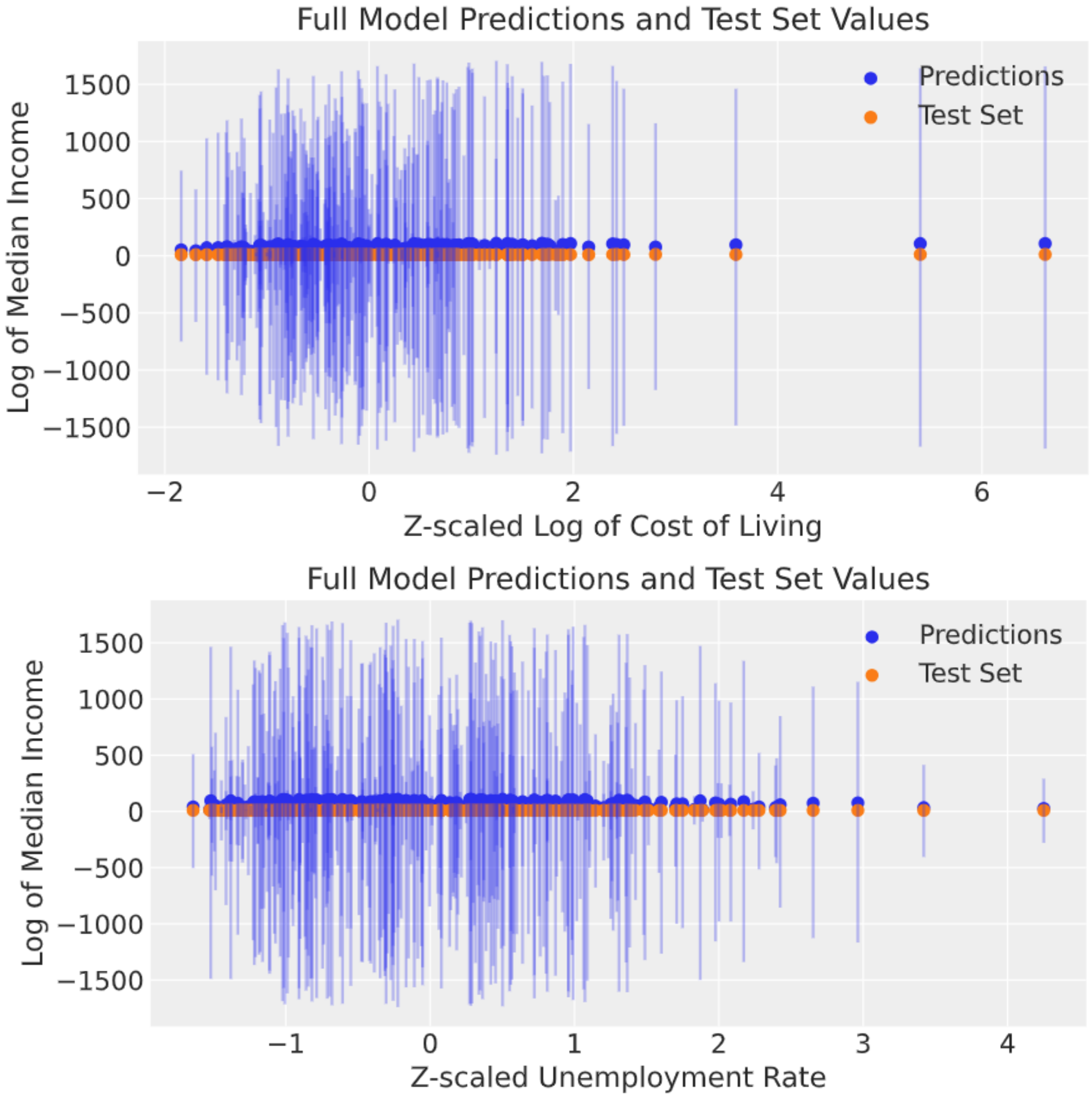
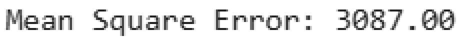


Figure 14

MSE value obtained from test data



```
Mean Square Error: 3087.00
```

Appendix C
Reduced Model

Figure 15

Visual representation of reduced model

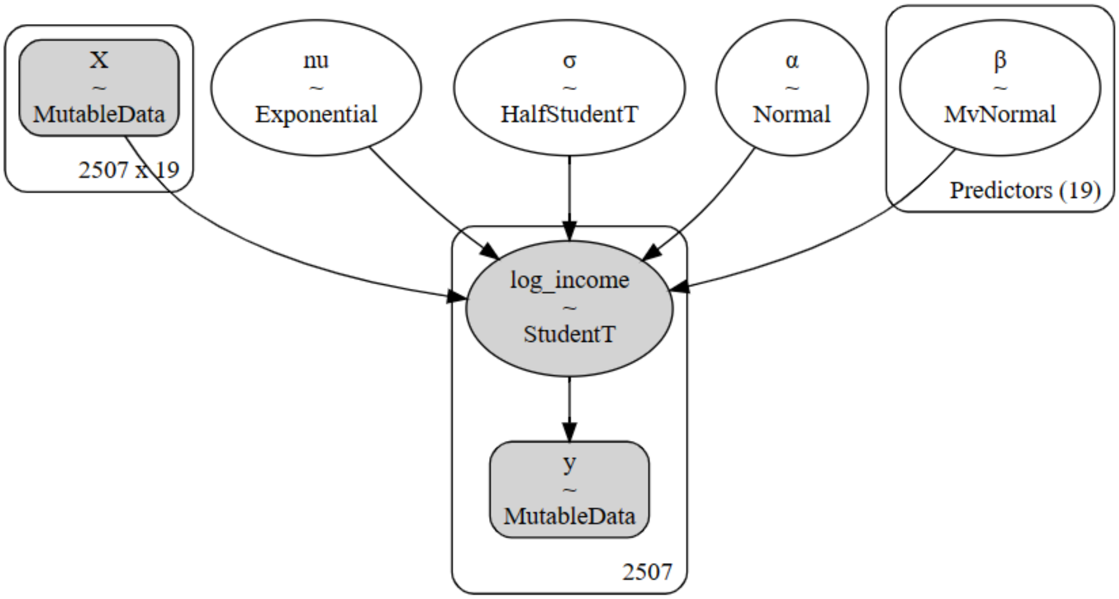


Figure 16

Trace plots to evaluate model sampling

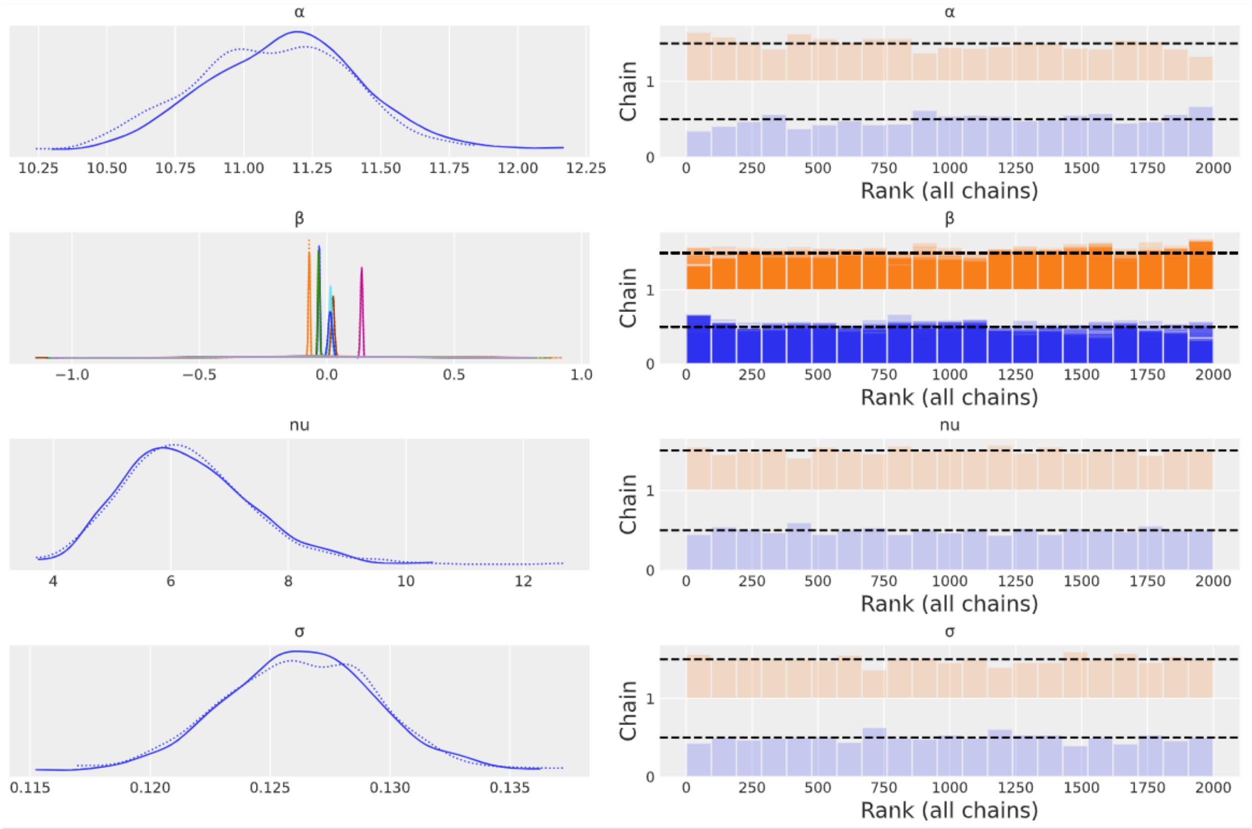
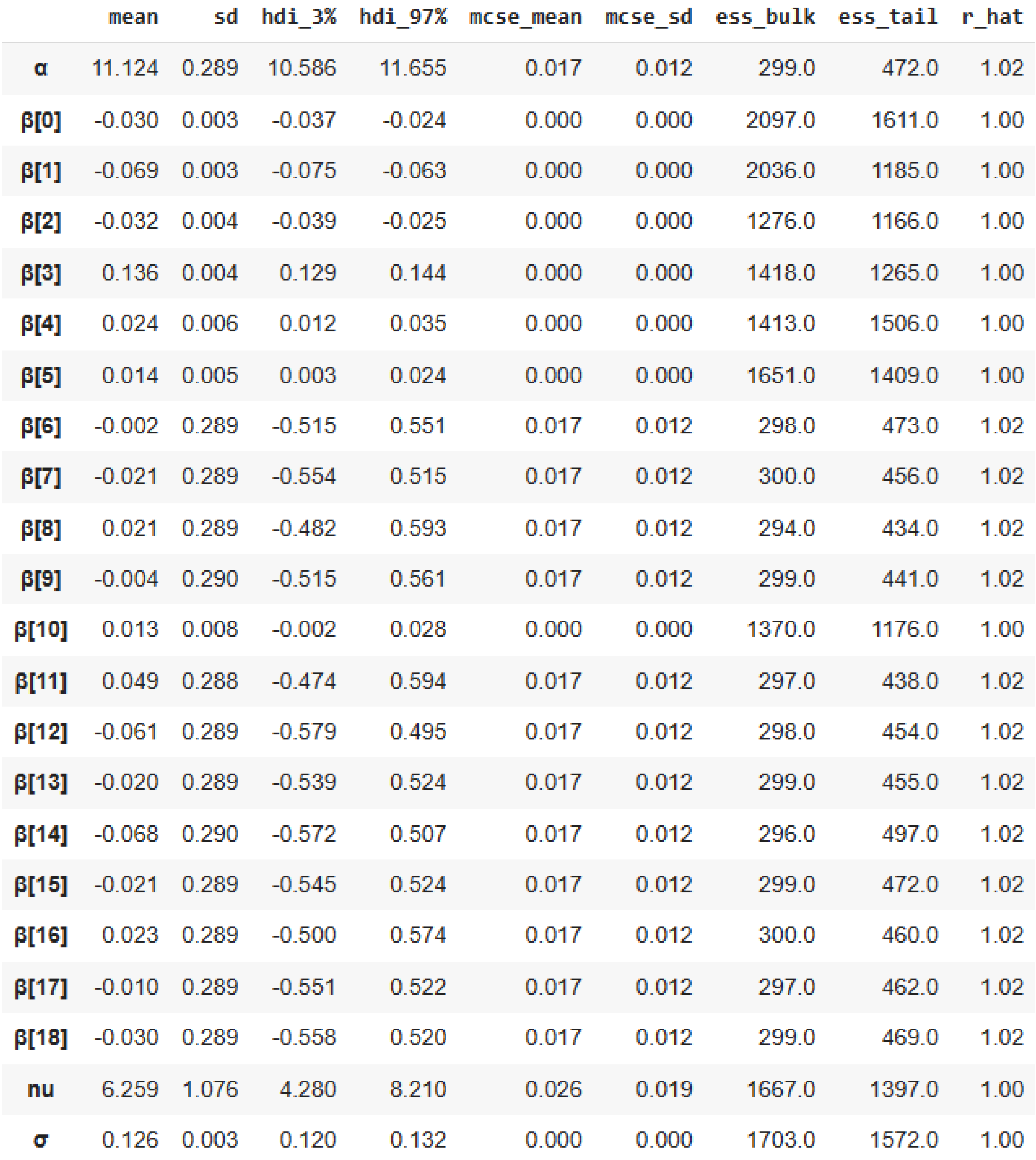


Table 3

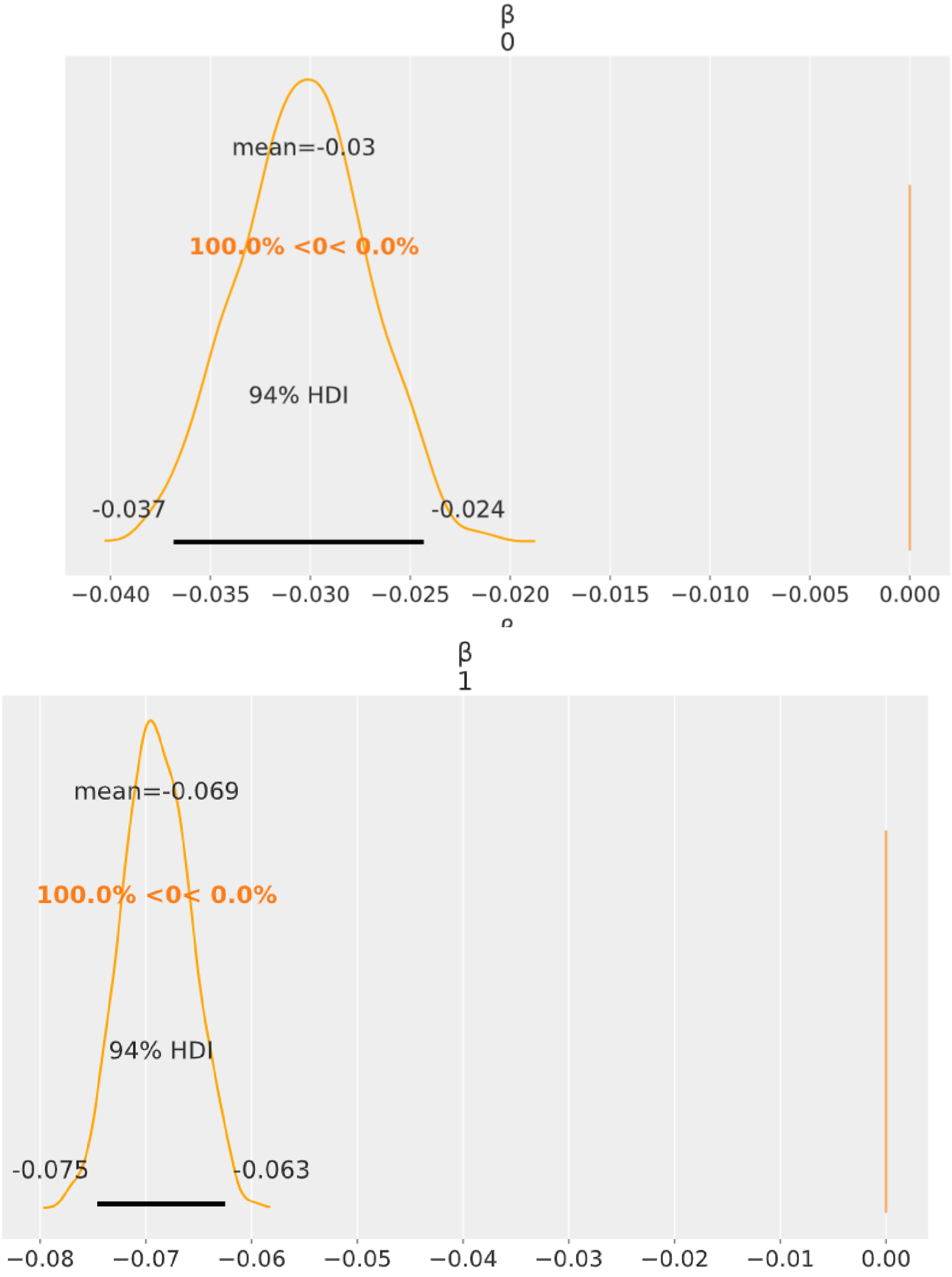
Parameter table after model fitting (az.summary())



	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
α	11.124	0.289	10.586	11.655	0.017	0.012	299.0	472.0	1.02
β[0]	-0.030	0.003	-0.037	-0.024	0.000	0.000	2097.0	1611.0	1.00
β[1]	-0.069	0.003	-0.075	-0.063	0.000	0.000	2036.0	1185.0	1.00
β[2]	-0.032	0.004	-0.039	-0.025	0.000	0.000	1276.0	1166.0	1.00
β[3]	0.136	0.004	0.129	0.144	0.000	0.000	1418.0	1265.0	1.00
β[4]	0.024	0.006	0.012	0.035	0.000	0.000	1413.0	1506.0	1.00
β[5]	0.014	0.005	0.003	0.024	0.000	0.000	1651.0	1409.0	1.00
β[6]	-0.002	0.289	-0.515	0.551	0.017	0.012	298.0	473.0	1.02
β[7]	-0.021	0.289	-0.554	0.515	0.017	0.012	300.0	456.0	1.02
β[8]	0.021	0.289	-0.482	0.593	0.017	0.012	294.0	434.0	1.02
β[9]	-0.004	0.290	-0.515	0.561	0.017	0.012	299.0	441.0	1.02
β[10]	0.013	0.008	-0.002	0.028	0.000	0.000	1370.0	1176.0	1.00
β[11]	0.049	0.288	-0.474	0.594	0.017	0.012	297.0	438.0	1.02
β[12]	-0.061	0.289	-0.579	0.495	0.017	0.012	298.0	454.0	1.02
β[13]	-0.020	0.289	-0.539	0.524	0.017	0.012	299.0	455.0	1.02
β[14]	-0.068	0.290	-0.572	0.507	0.017	0.012	296.0	497.0	1.02
β[15]	-0.021	0.289	-0.545	0.524	0.017	0.012	299.0	472.0	1.02
β[16]	0.023	0.289	-0.500	0.574	0.017	0.012	300.0	460.0	1.02
β[17]	-0.010	0.289	-0.551	0.522	0.017	0.012	297.0	462.0	1.02
β[18]	-0.030	0.289	-0.558	0.520	0.017	0.012	299.0	469.0	1.02
nu	6.259	1.076	4.280	8.210	0.026	0.019	1667.0	1397.0	1.00
σ	0.126	0.003	0.120	0.132	0.000	0.000	1703.0	1572.0	1.00

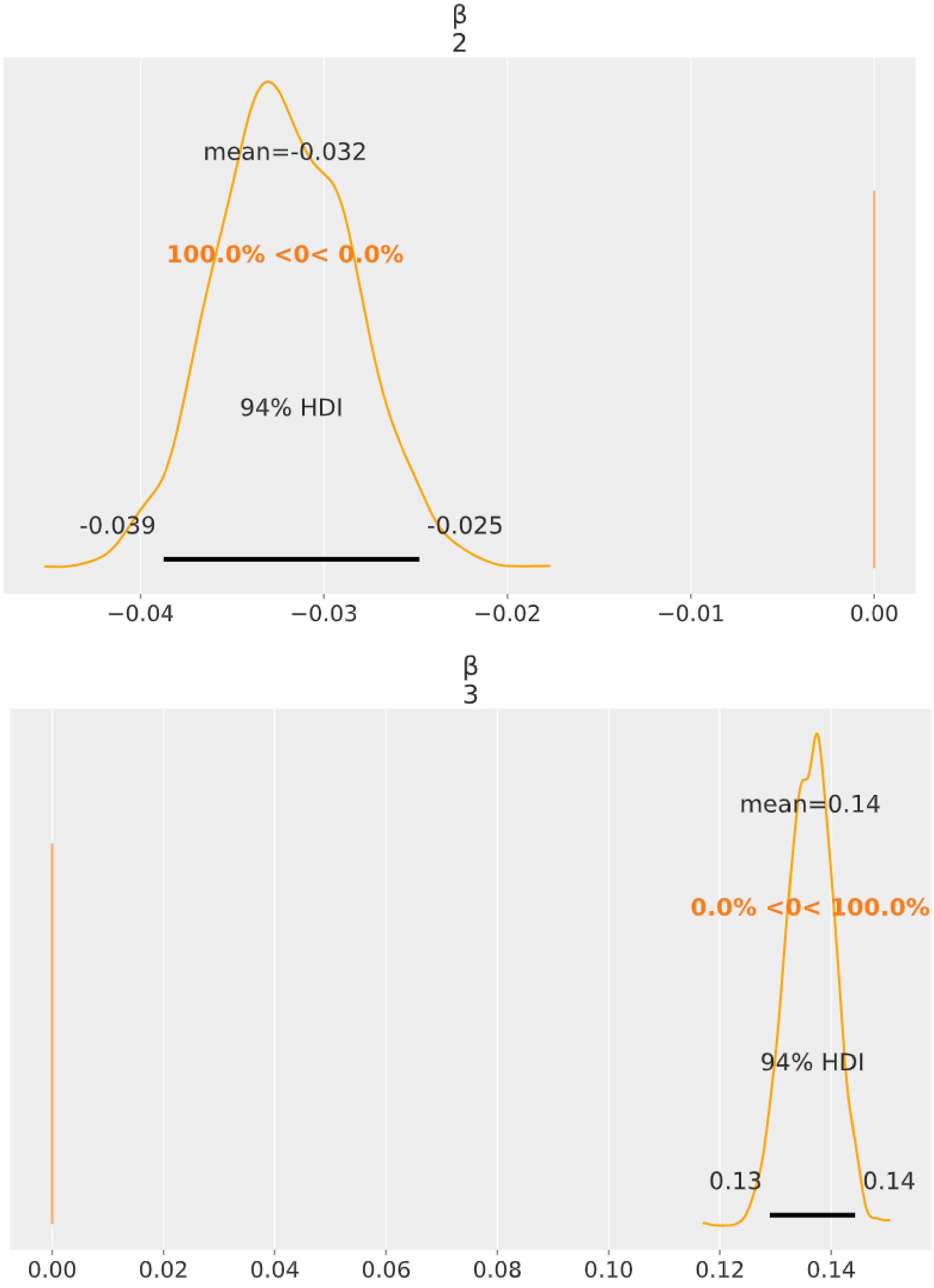
Figures 17 and 18

Beta distributions of interest in reduced model

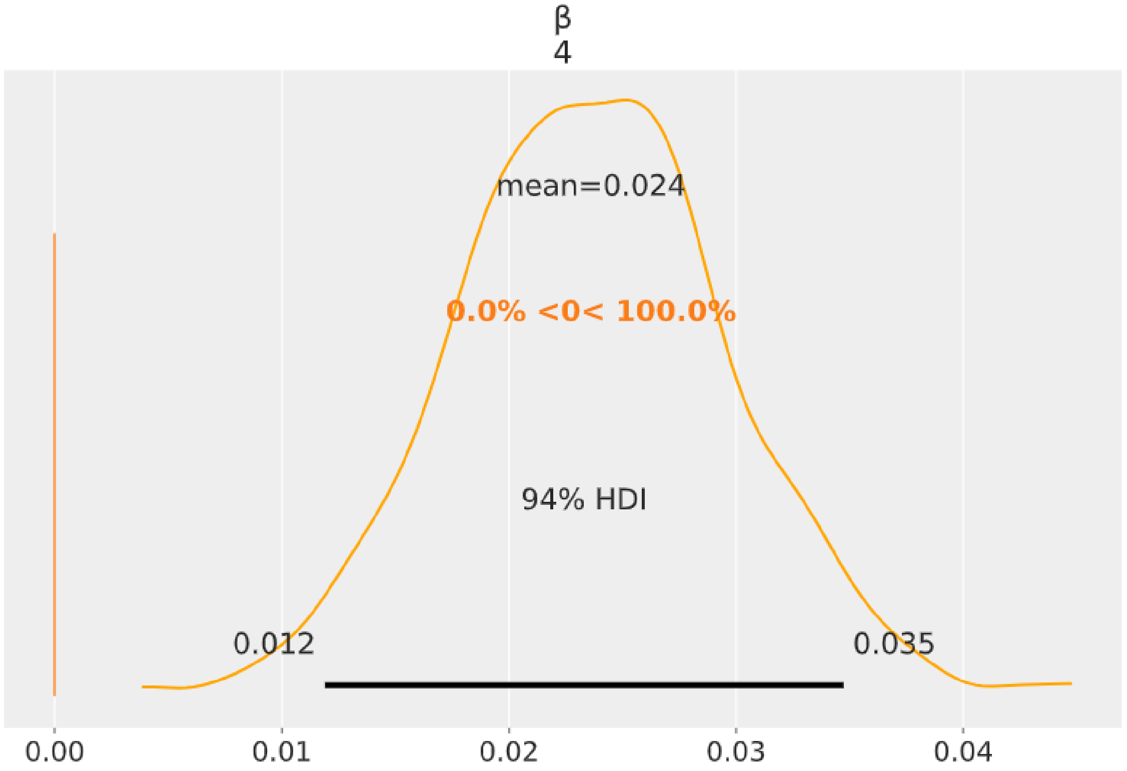


Figures 19 and 20

Beta distributions of interest in reduced model



Figures 21 and 22
Beta distributions of interest in reduced model



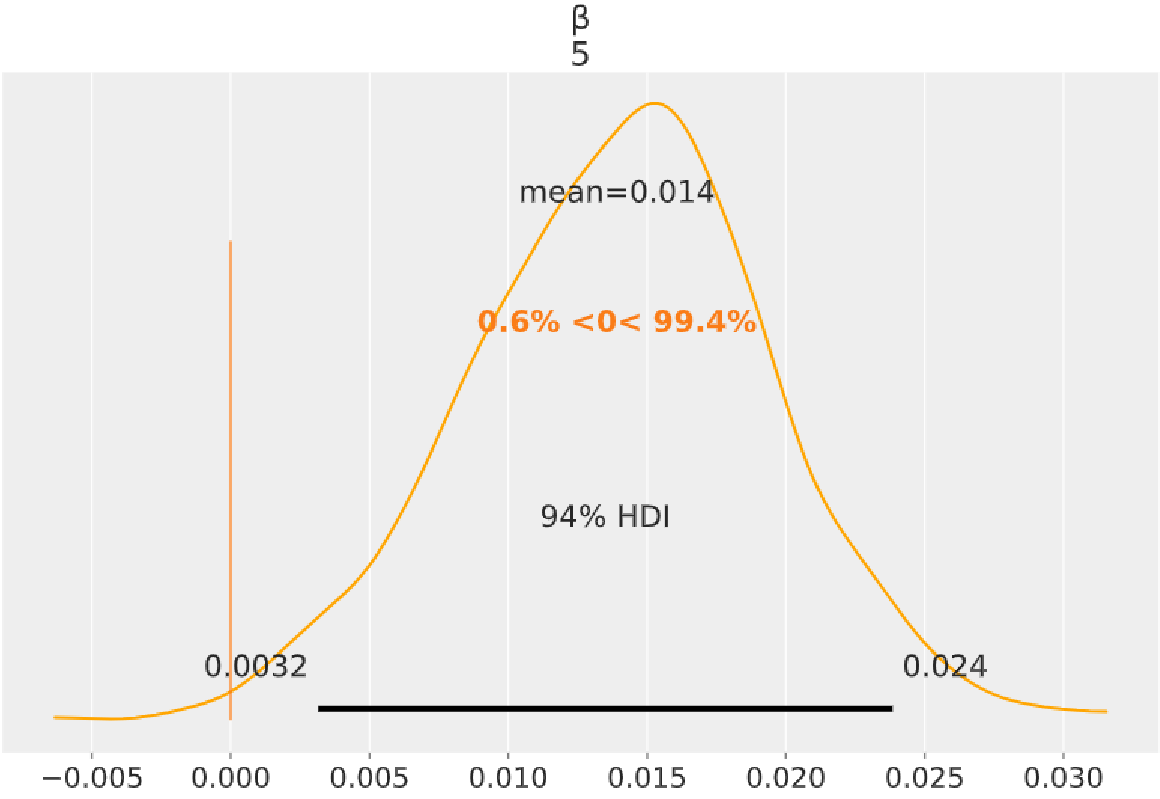


Figure 23
Forest of beta distributions

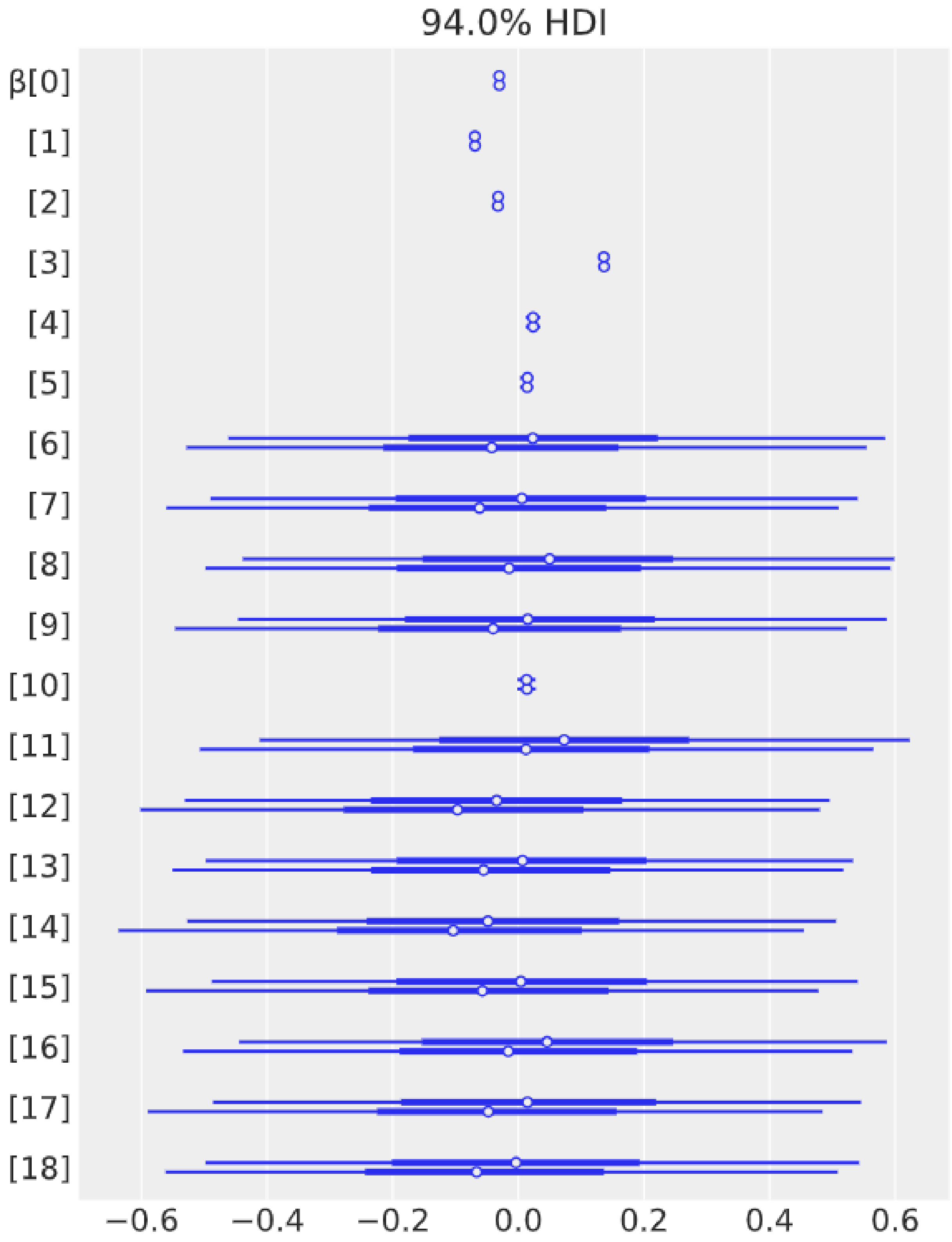


Figure 24
Bayesian p-value plot

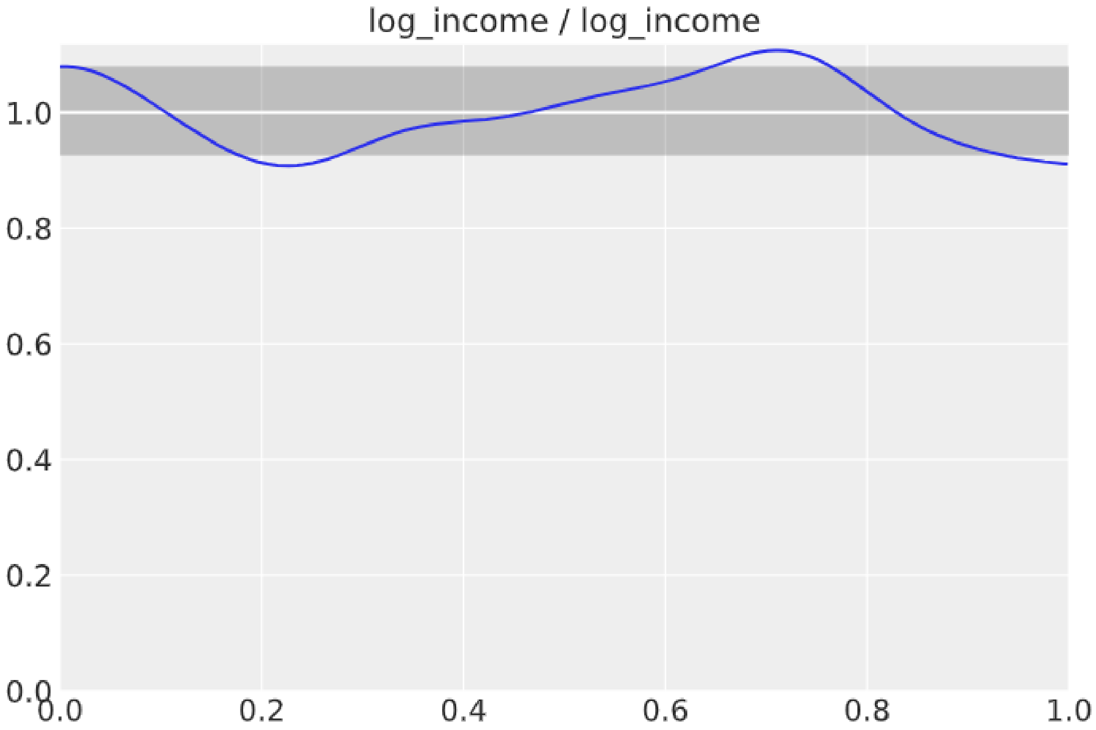
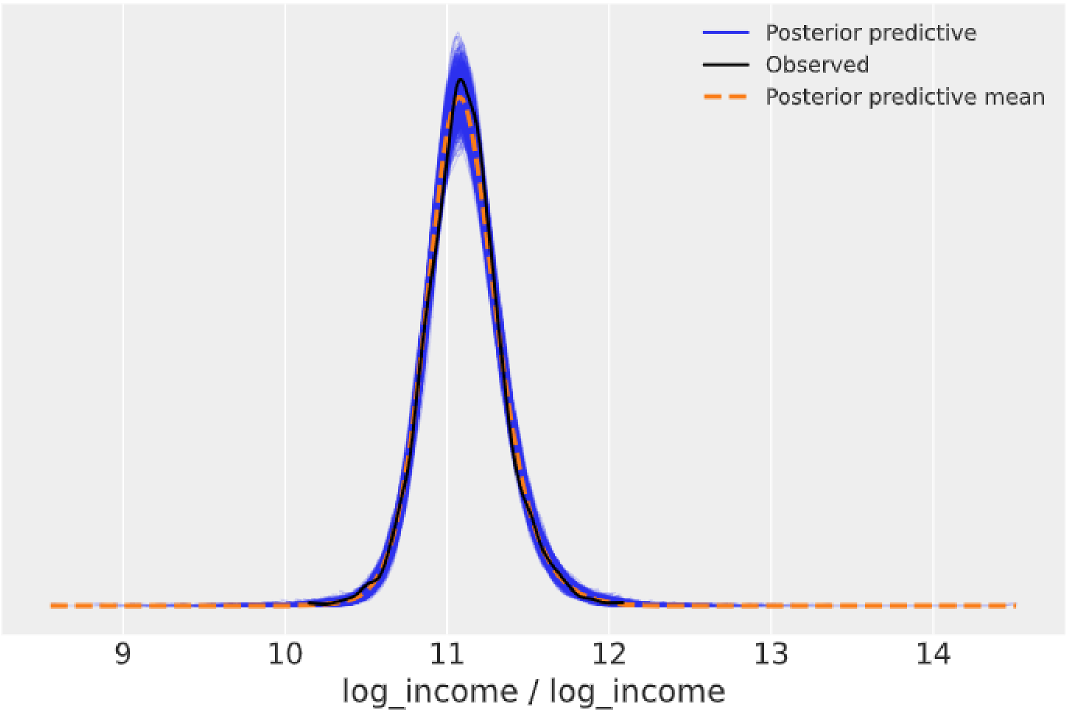


Figure 25
Posterior Predictive of Log of Median Household Income



Figures 26 and 27
Plots of Reduced Model Predictions showing Uncertainty versus Test Set Values

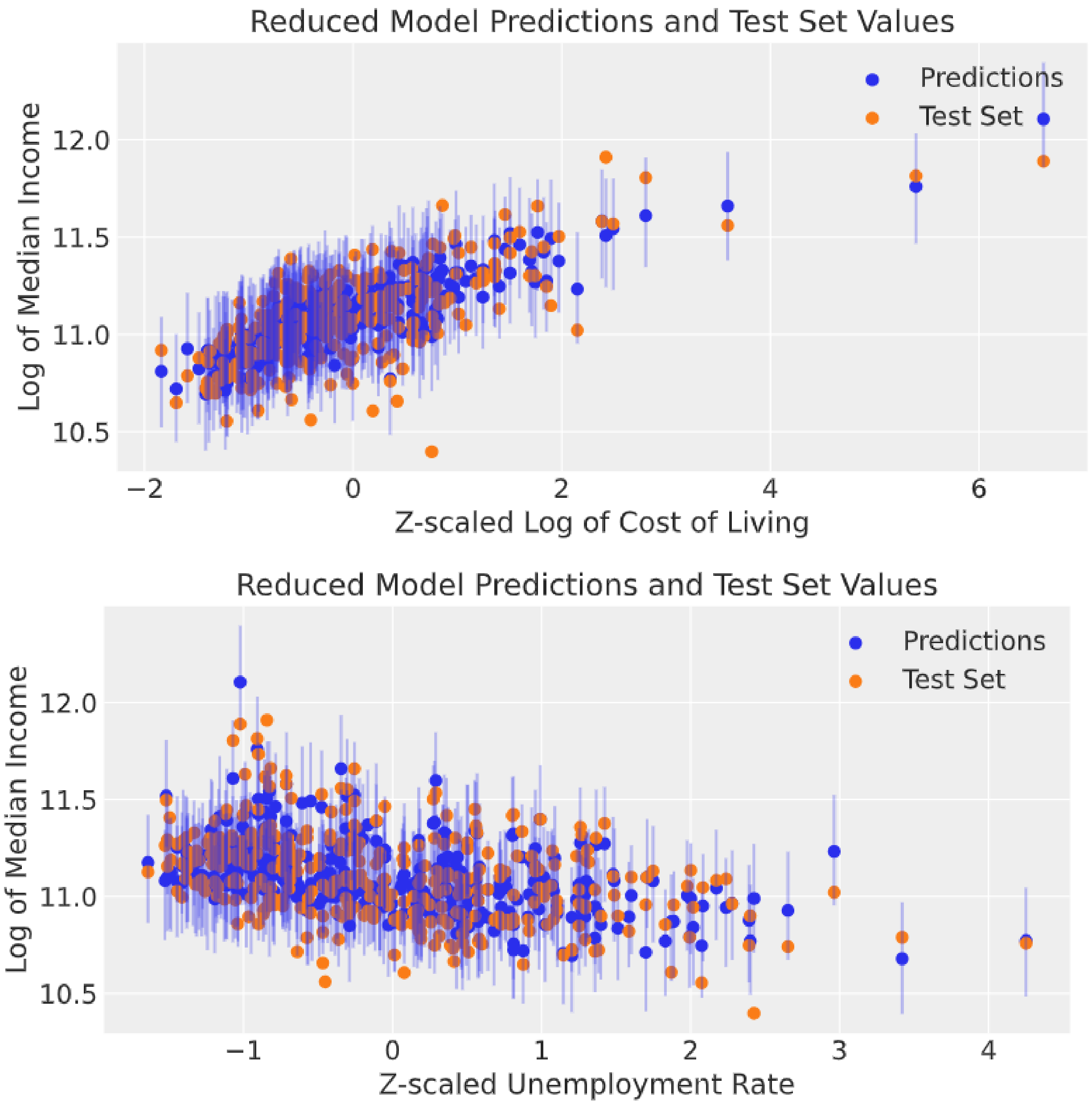


Figure 28
MSE value obtained from test data

```
Mean Square Error: 0.02
```

Appendix D
Non-Linear Reduced Model

Figure 29

Visual representation of the non-linear reduced model

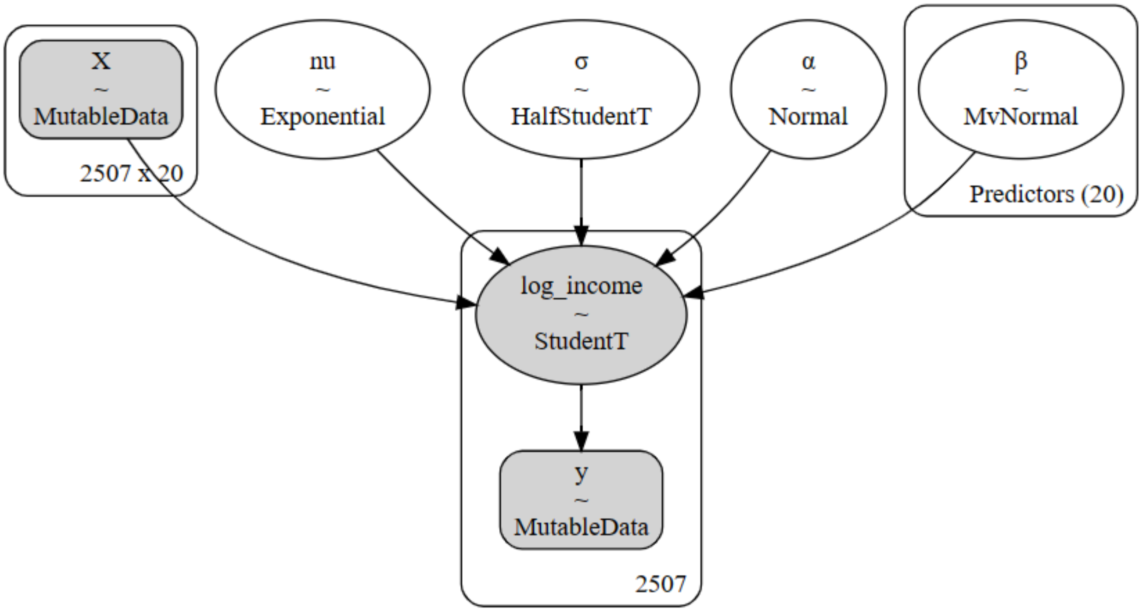


Figure 30

Trace plots to evaluate sampling

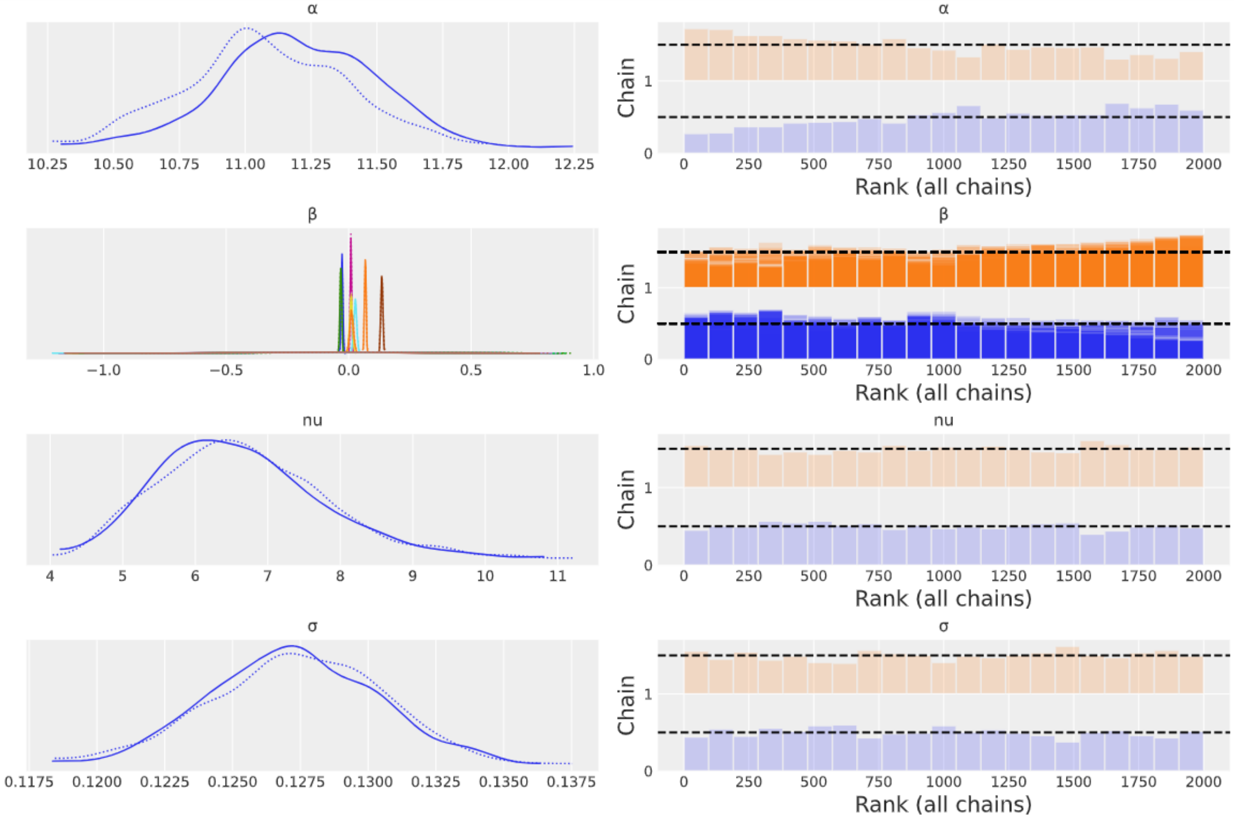
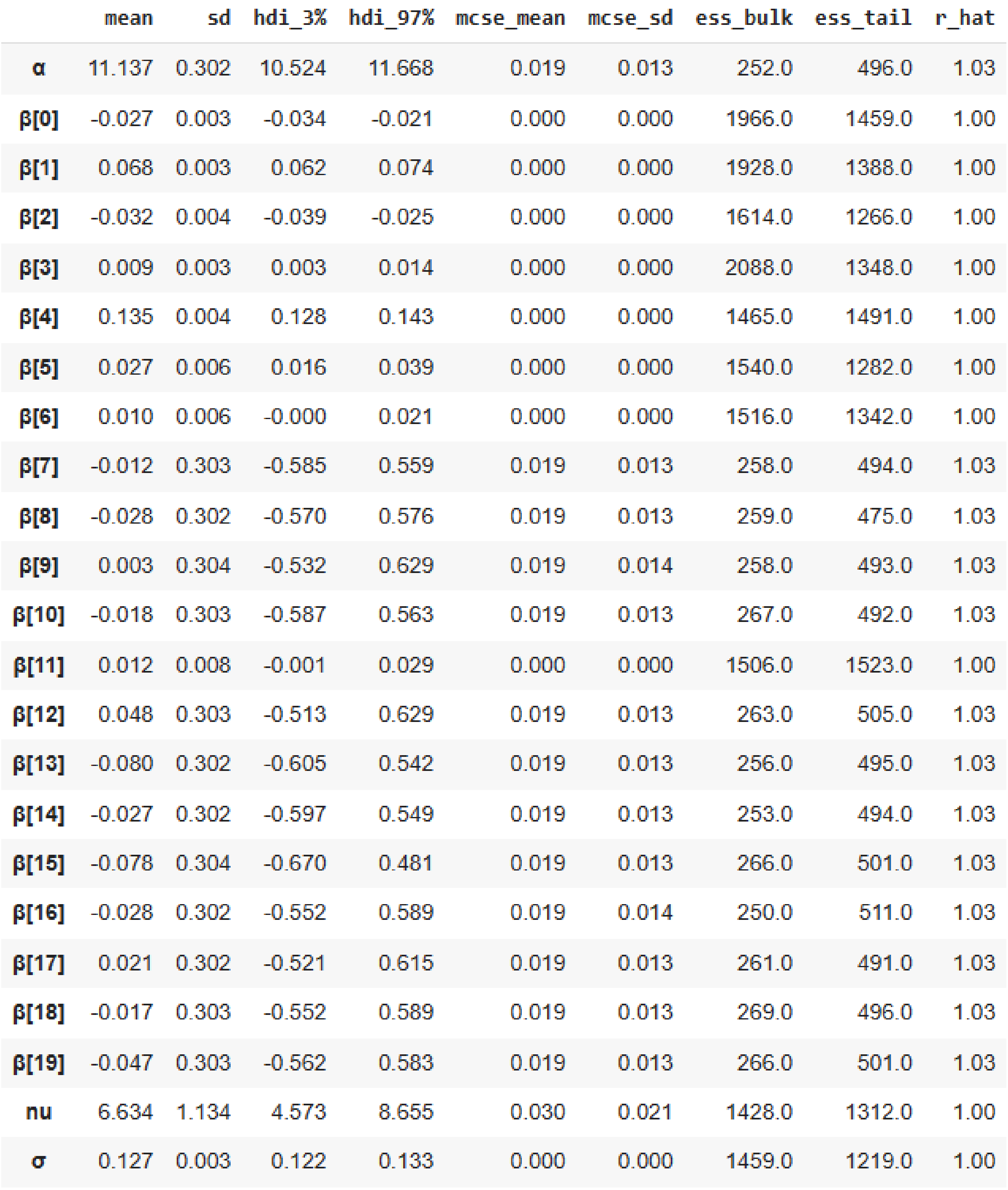


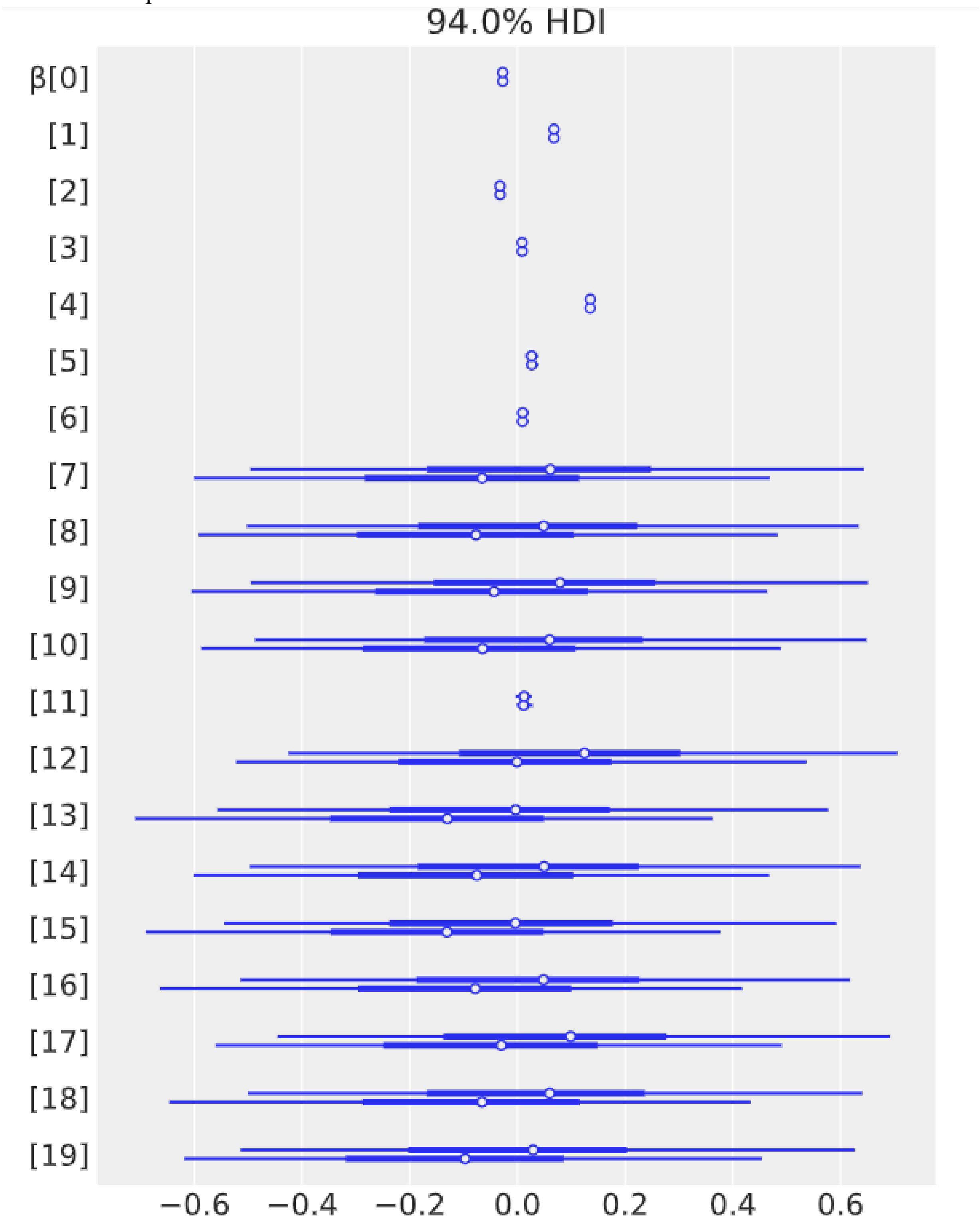
Table 4

Parameter table after model fitting (az.summary())

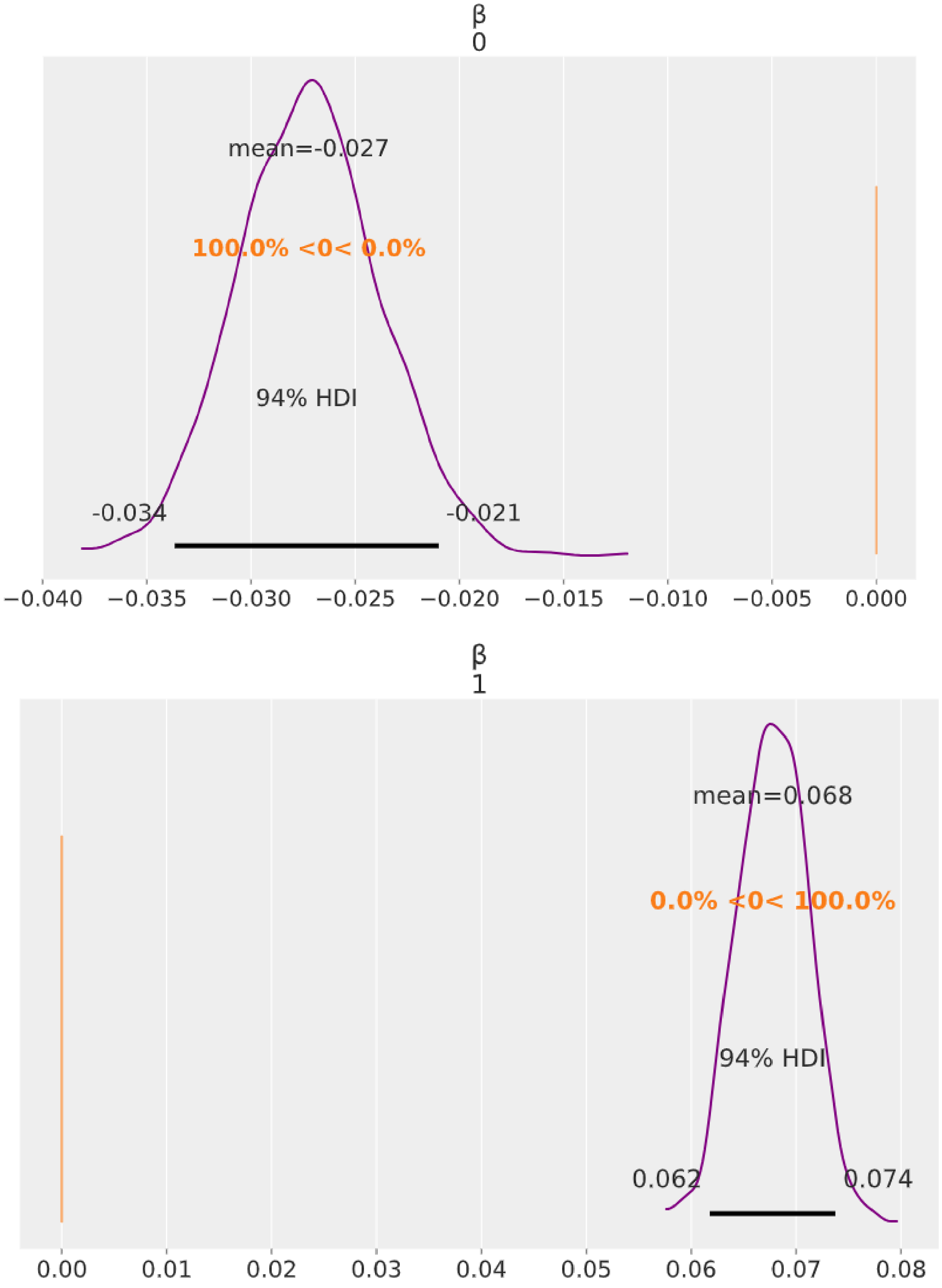


	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
α	11.137	0.302	10.524	11.668	0.019	0.013	252.0	496.0	1.03
β[0]	-0.027	0.003	-0.034	-0.021	0.000	0.000	1966.0	1459.0	1.00
β[1]	0.068	0.003	0.062	0.074	0.000	0.000	1928.0	1388.0	1.00
β[2]	-0.032	0.004	-0.039	-0.025	0.000	0.000	1614.0	1266.0	1.00
β[3]	0.009	0.003	0.003	0.014	0.000	0.000	2088.0	1348.0	1.00
β[4]	0.135	0.004	0.128	0.143	0.000	0.000	1465.0	1491.0	1.00
β[5]	0.027	0.006	0.016	0.039	0.000	0.000	1540.0	1282.0	1.00
β[6]	0.010	0.006	-0.000	0.021	0.000	0.000	1516.0	1342.0	1.00
β[7]	-0.012	0.303	-0.585	0.559	0.019	0.013	258.0	494.0	1.03
β[8]	-0.028	0.302	-0.570	0.576	0.019	0.013	259.0	475.0	1.03
β[9]	0.003	0.304	-0.532	0.629	0.019	0.014	258.0	493.0	1.03
β[10]	-0.018	0.303	-0.587	0.563	0.019	0.013	267.0	492.0	1.03
β[11]	0.012	0.008	-0.001	0.029	0.000	0.000	1506.0	1523.0	1.00
β[12]	0.048	0.303	-0.513	0.629	0.019	0.013	263.0	505.0	1.03
β[13]	-0.080	0.302	-0.605	0.542	0.019	0.013	256.0	495.0	1.03
β[14]	-0.027	0.302	-0.597	0.549	0.019	0.013	253.0	494.0	1.03
β[15]	-0.078	0.304	-0.670	0.481	0.019	0.013	266.0	501.0	1.03
β[16]	-0.028	0.302	-0.552	0.589	0.019	0.014	250.0	511.0	1.03
β[17]	0.021	0.302	-0.521	0.615	0.019	0.013	261.0	491.0	1.03
β[18]	-0.017	0.303	-0.552	0.589	0.019	0.013	269.0	496.0	1.03
β[19]	-0.047	0.303	-0.562	0.583	0.019	0.013	266.0	501.0	1.03
nu	6.634	1.134	4.573	8.655	0.030	0.021	1428.0	1312.0	1.00
σ	0.127	0.003	0.122	0.133	0.000	0.000	1459.0	1219.0	1.00

Figure 31
Forest of beta parameters in model

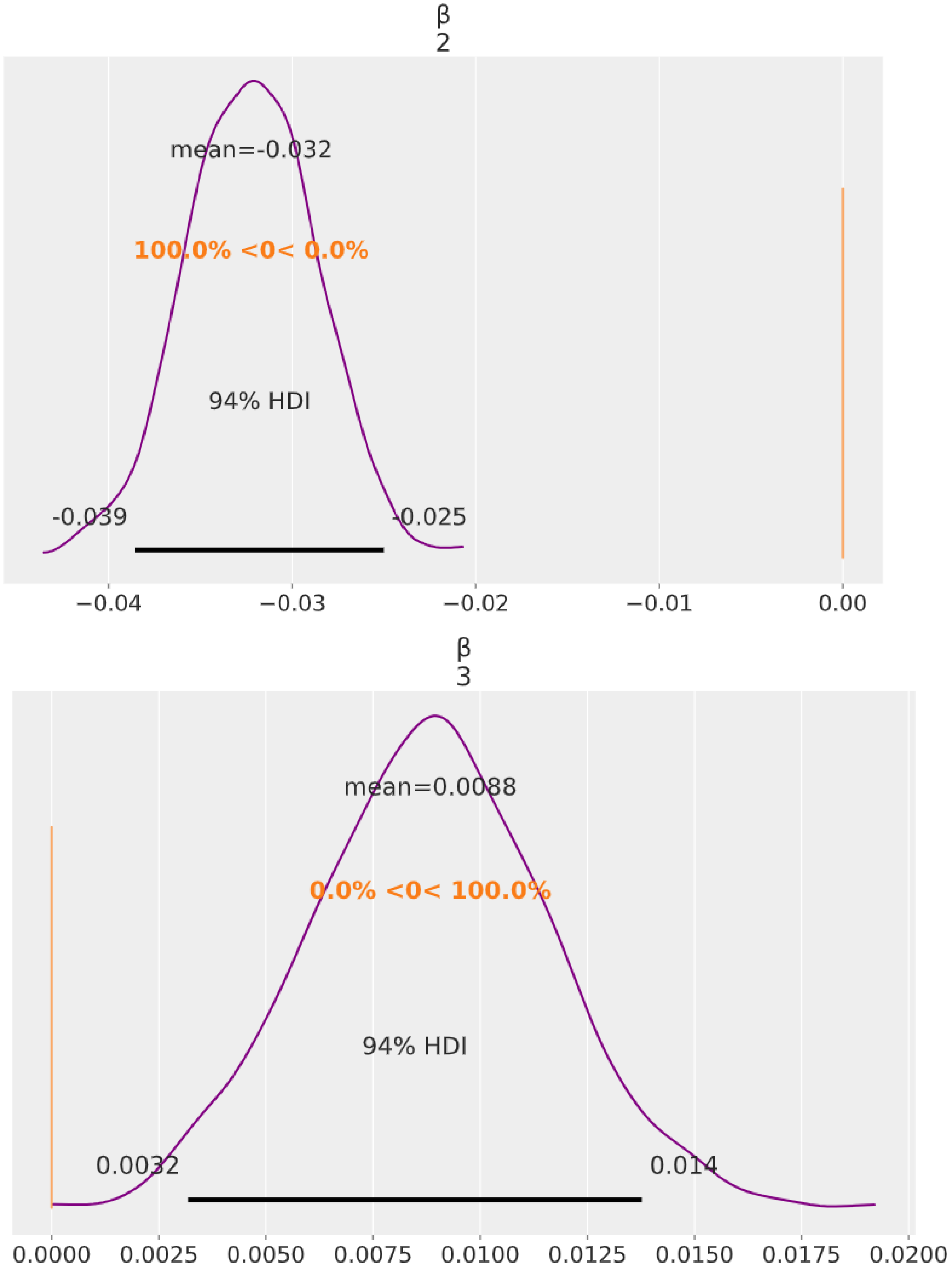


Figures 32 and 33
Beta distributions of interest



Figures 34 and 35

Beta distributions of interest



Figures 36 and 37
Beta distributions of interest

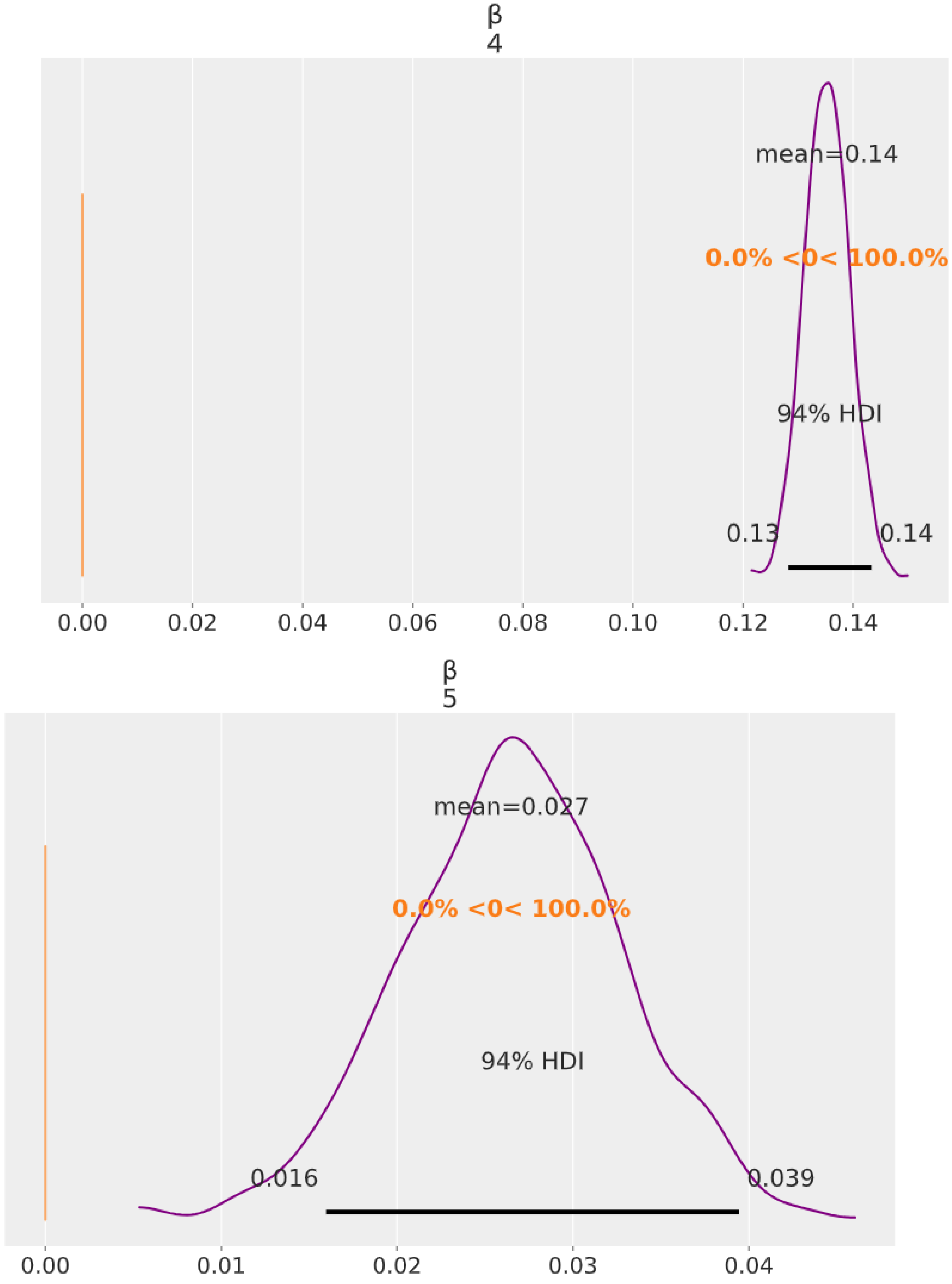


Figure 38
Bayesian p-value plot

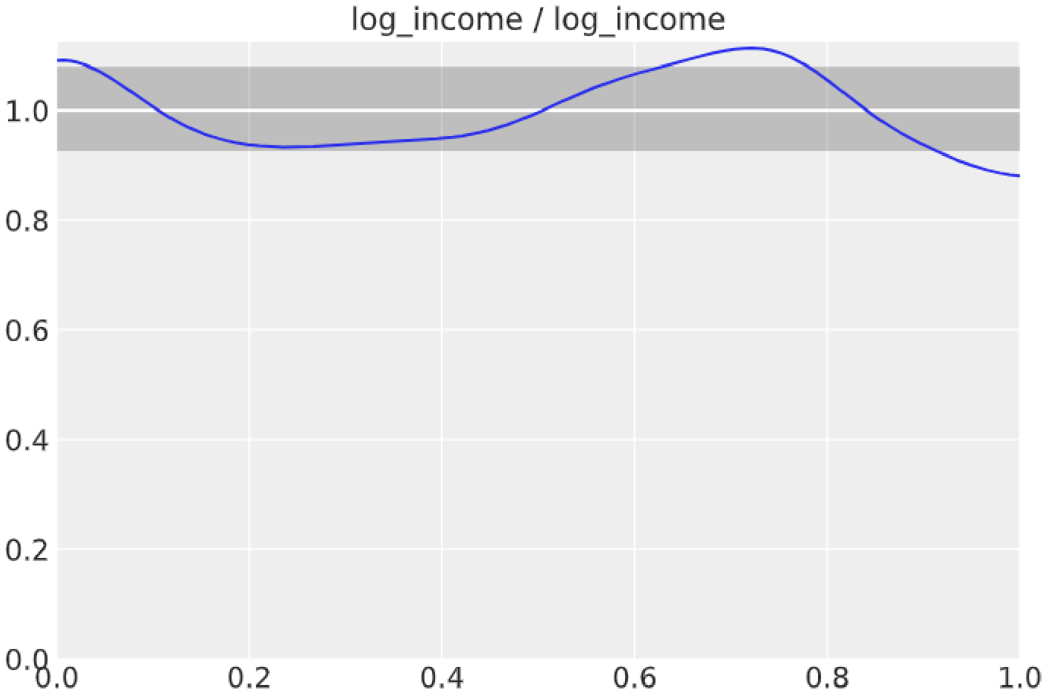
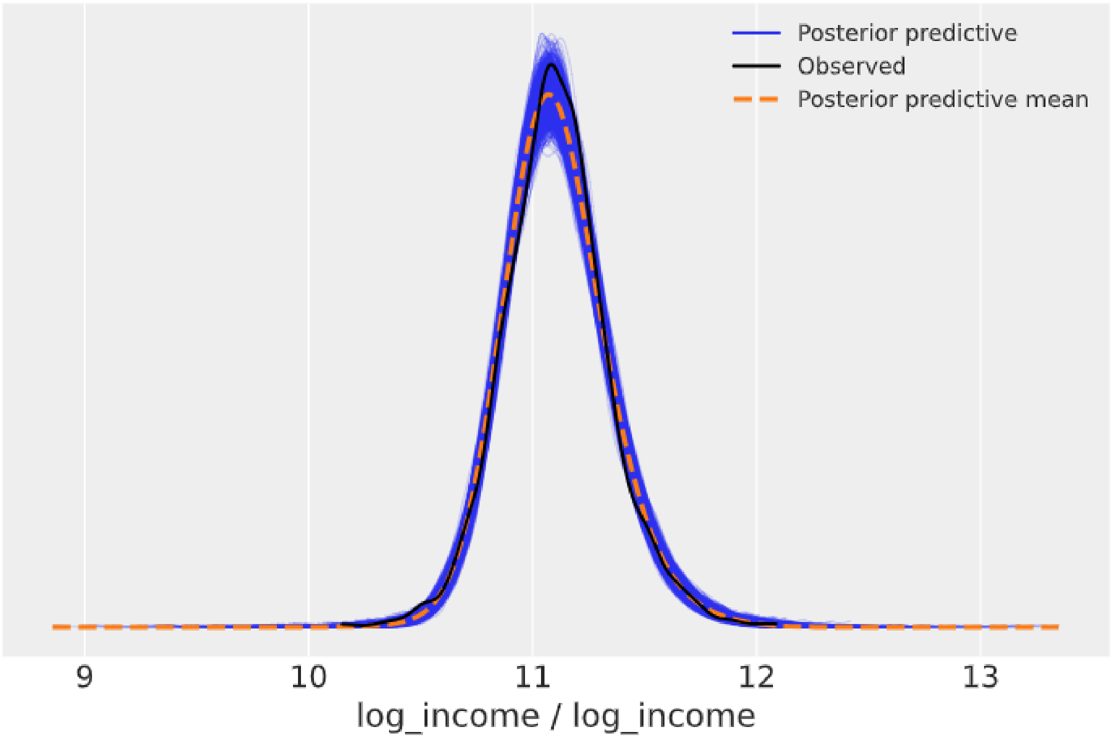


Figure 39
Posterior Predictive of Log of Median Household Income



Figures 40 and 41

Non-Linear Model Predictions Showing Uncertainty Versus Test Set Values

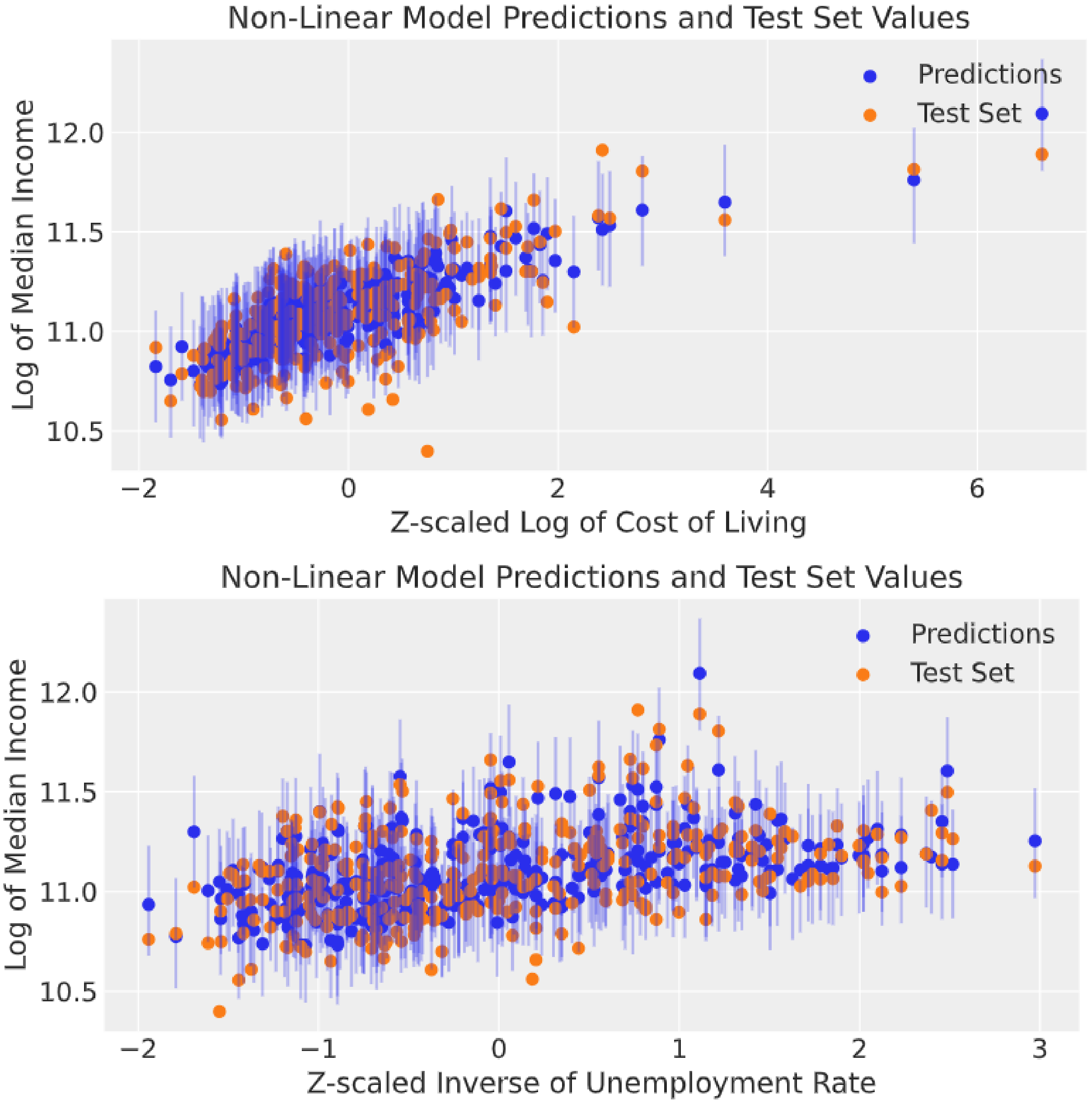
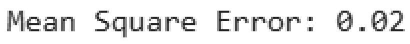


Figure 42

MSE value obtained from test data

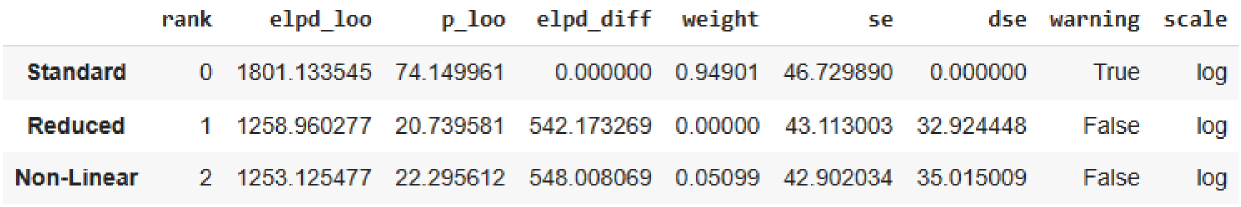


```
Mean Square Error: 0.02
```

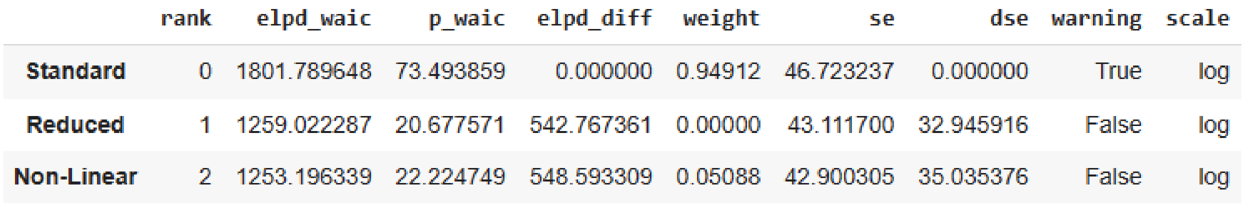
Appendix D
Bayesian Model Averaging Model

Tables 5 and 6

Model Comparisons between three previous models



	rank	elpd_loo	p_loo	elpd_diff	weight	se	dse	warning	scale
Standard	0	1801.133545	74.149961	0.000000	0.94901	46.729890	0.000000	True	log
Reduced	1	1258.960277	20.739581	542.173269	0.00000	43.113003	32.924448	False	log
Non-Linear	2	1253.125477	22.295612	548.008069	0.05099	42.902034	35.015009	False	log



	rank	elpd_waic	p_waic	elpd_diff	weight	se	dse	warning	scale
Standard	0	1801.789648	73.493859	0.000000	0.94912	46.723237	0.000000	True	log
Reduced	1	1259.022287	20.677571	542.767361	0.00000	43.111700	32.945916	False	log
Non-Linear	2	1253.196339	22.224749	548.593309	0.05088	42.900305	35.035376	False	log

Figure 43

Predictive BMA model vs. actual values

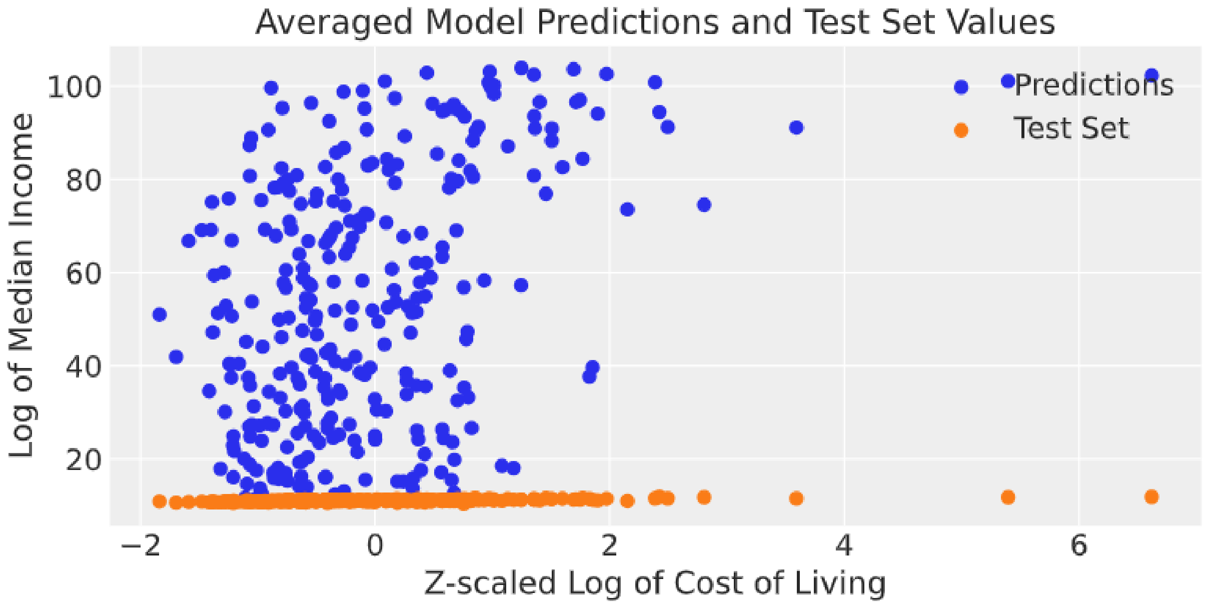
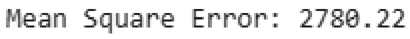


Figure 44

Reported BMA Model MSE



```
Mean Square Error: 2780.22
```