Determinants of Household Dietary Diversity in Tanzania: A Bayesian Analysis with Lasso Prior

Tiffany Hu

Table of Contents

Abstract

Dietary diversity, defined as “a qualitative measure of food consumption that reflects household access to a variety of foods” (*Guidelines for measuring household and individual dietary diversity*, Food and Agriculture Organization of the United Nations), is a well-known metric for providing a straightforward and effective measurement of nutrition adequacy and indication of a household’s socio-economic status. By adopting the dietary diversity and socio-economic data of households in Tanzania (*Dietary diversity and socio-economic data of households in Tanzania*, Ochieng, J. *et al,* 2017), this study investigates the factors influencing household dietary diversity in Tanzania using Bayesian statistical models, including Poisson and Negative Binomial regressions. The Poisson Model was selected over the Negative binomial considering its higher value in the expected log predictive density in Leave-One-Out Cross Validation(elpd_loo) and lower value in Leave-One-Out information criterion(looic). And further model fitting and comparison using metrics like LOOIC and WAIC with different priors, including normal, Cauchy, non-informative, and Lasso, was employed to enhance the robustness and reliability of the analysis. The Lasso prior produced the lowest elpd_diff and se_diff (both are 0) when compared with models with other priors, suggesting the best performance. The results in the final selected model show that factors like village location, whether the household grows vegetables or not, whether the household has participated in nutritional training, and household size are associated with the household’s dietary diversity, signifying the impact of geographical and socio-economic factors in dietary diversity, with significant policy implications for improving nutritional quality as well as general well-being for not only households in Tanzania but also other countries in Africa.

Introduction

Household dietary diversity is a valuable measure of a household's nutritional sufficiency and relative food security. In African countries like Tanzania where agriculture is the primary source of income for many households, understanding the determinants of dietary diversity can help adjust existing policies and improve nutritional outcomes and physical well-being of the citizens, especially when the challenges of food shortage and malnutrition are still unsolved in many regions in the continent. Moreover, dietary diversity provides insights into a country's general economic development. A deficiency in dietary diversity may indicate more deeply rooted inequalities in the distributions of resources and social wealth, making it a worth-while factor for social scientist to consider if they want to gain a more thorough understanding of a nation's social functioning and development status.

This study aims to explore the socio-economic and demographic factors affecting household dietary diversity in Tanzania. Specifically, it examines the impact of variables such as education, household size, land size, as well as geolocations on dietary diversity. The analysis employs Bayesian statistical models to provide meaningful insights into these relationships with the proposed study question: how are socio-economic and demographic factors associated with households' dietary diversity in Tanzania?

Analysis and Results

Exploratory Data Analysis (EDA)

To enhance our understanding of the socio-economics factors, we perform Exploratory Data Analysis on the socio-economics variables to help us discover the underlying patterns and structures of the data. Main variables analyzed include geographical locations of the household, household size, education level, age and gender of the household head, land size, engagement in agricultural activities, and whether the household receives off-farm incomes.

The main steps in the exploration of the data involve examining summary statistics and visualizations for distributions of key variables as well as their relationships between the target variable HDDS (Household Dietary Diversity Score). The summary statistics provide insights into the central tendencies and dispersion of the data, while the visualizations of their distributions enable us to gain a basic understanding of the data structures. Finally, a correlation matrix is included to check for multicollinearity to ensure the regression coefficients are robust and interpretable in the final modeling results.

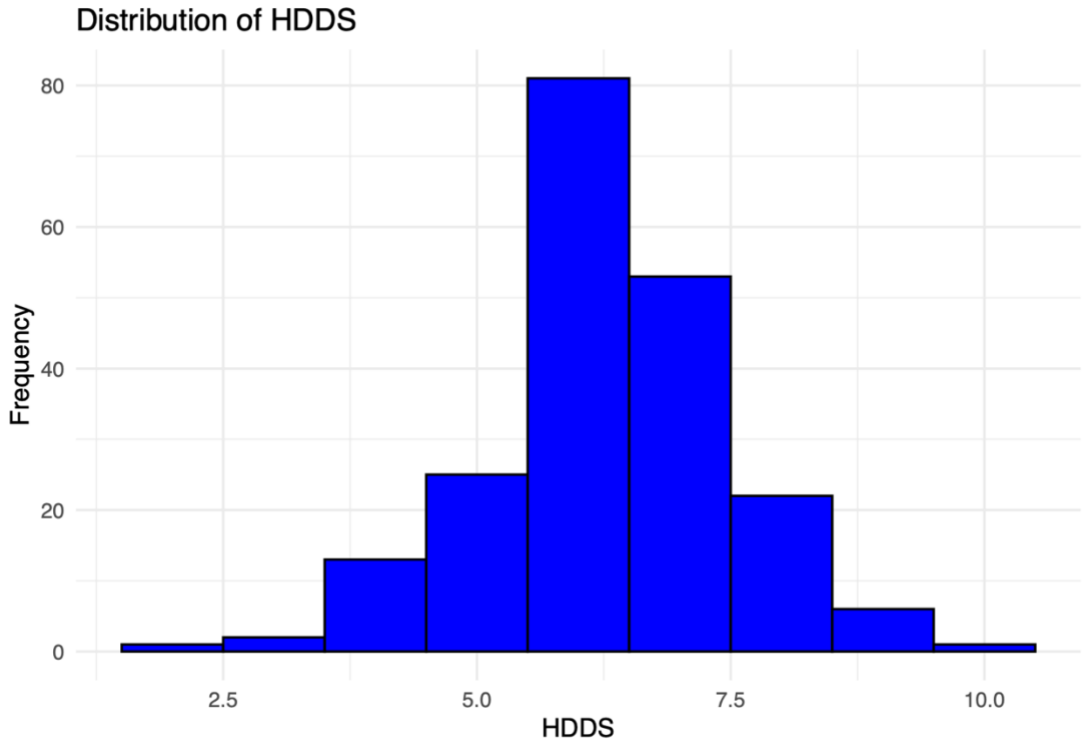


Figure 1: Distribution of HDDS (Household Dietary Diversity Score)

The distribution of HDDS seems to have a relatively normal bell-shape distribution, with the majority of the recorded HDDS scores to be around 5 to 7.5, the centered distribution of the target variable indicates that a Poisson Model might be better suited for the dataset than the Negative Binomial Model since there doesn't seem to be a visible overdispersion issue.

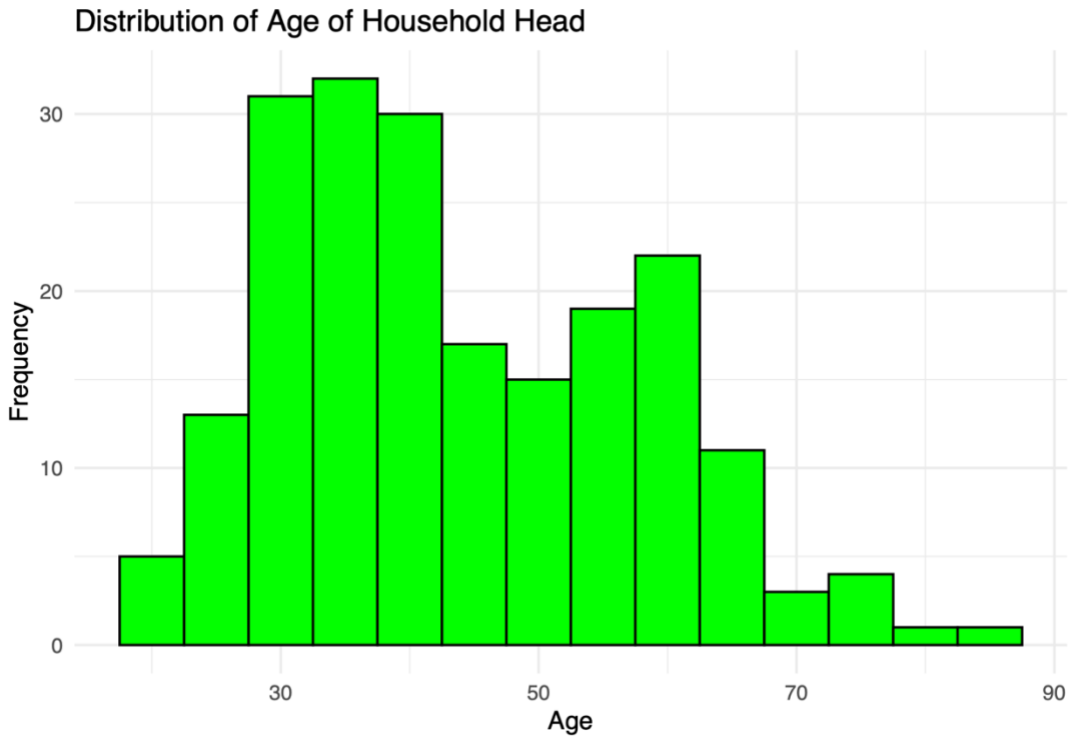


Figure 2: Distribution of Age of Household Head

The distribution of the Age of the Household Head seems to exhibit a slight left-skewness, with the age centered more around 30 to 50 years so, which corresponds to the common sense. Since the skewness is rather minor, I don't think there's a need for transforming the Age variable prior to model fitting; however, it is worth considering transforming it into a categorical variable.

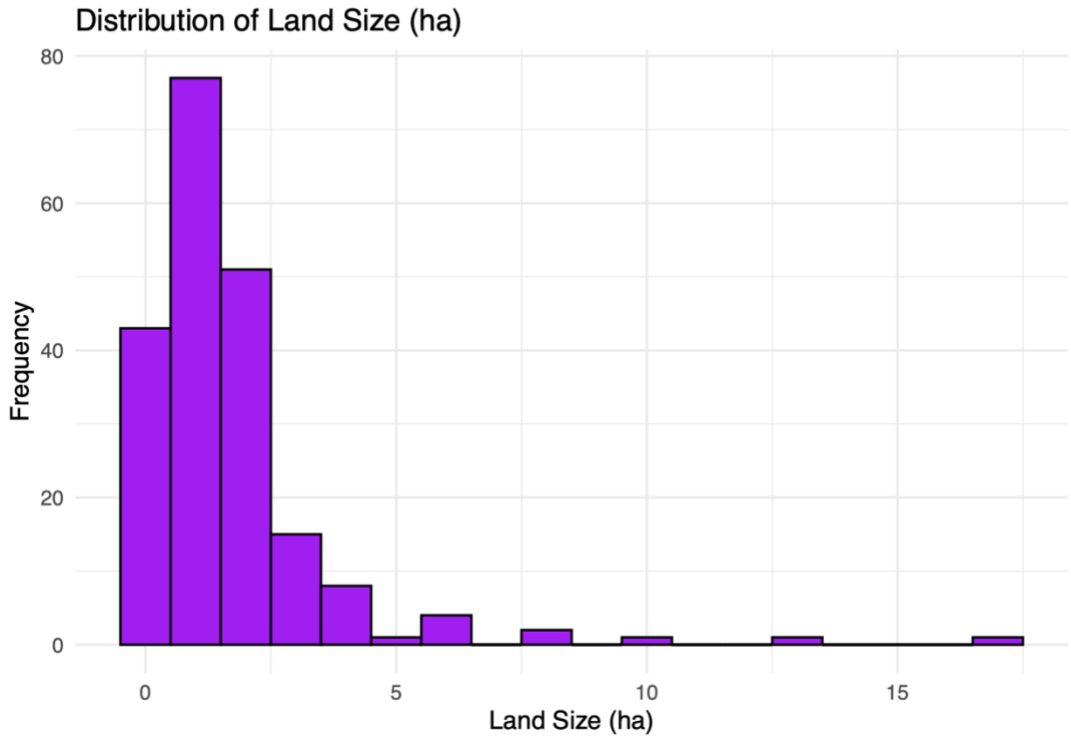


Figure 3: Distribution of Land Size

Based on the visualization, the distribution of Land Size exhibits a strong left-skewness, with almost all observed household land sizes to be around 0 and 5 Ha. Due to the nature of skewness, a logarithmic transformation is reasonable before fitting the land size variable into the regression models.

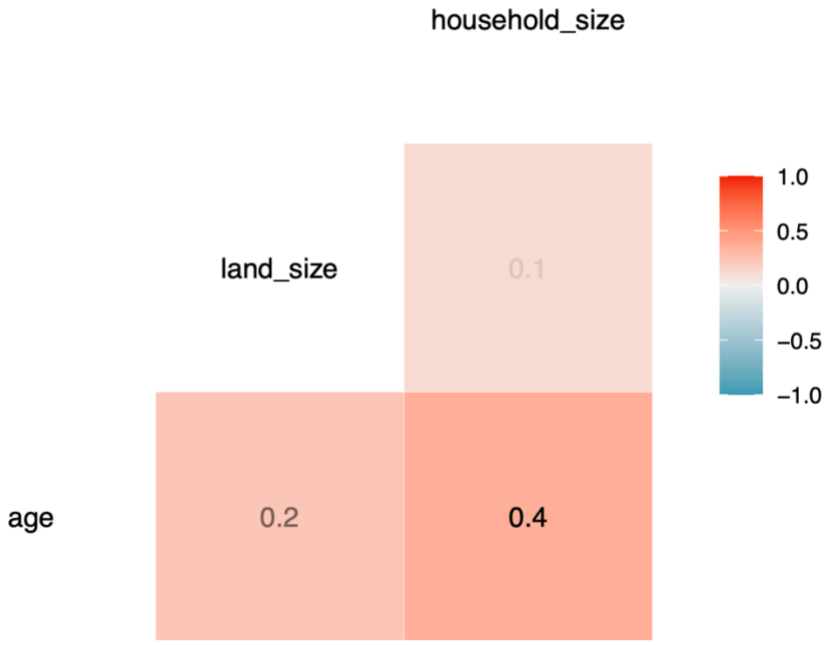


Figure 4: Correlation Matrix Visualization for numerical variables in the dataset

Based on the result of the correlation matrix, there doesn't seem to be a noticeable multicollinearity issue in the dataset.

```
##              educationF    Age_head land_size_ha        HDDS
## educationF    1.0000000 -0.15164571    0.1101914  0.26316650
## Age_head     -0.1516457  1.00000000    0.2413387 -0.04250907
## land_size_ha  0.1101914  0.24133873    1.0000000  0.20836712
## HDDS          0.2631665 -0.04250907    0.2083671  1.00000000
```

Figure 5: Correlation Matrix

Moreover, the calculated correlation matrix values confirm the results shown in correlation matrix visualization. All values are within +- 0.3, showing that we don't have to worry about potential multicollinearity issue before model fitting.

It is also useful to visualize the relationships between the socio-economic predictors and the target variable HDDS to gain a general understanding of their potential associations and set up a basic expectation for what to expect from the modeling result.

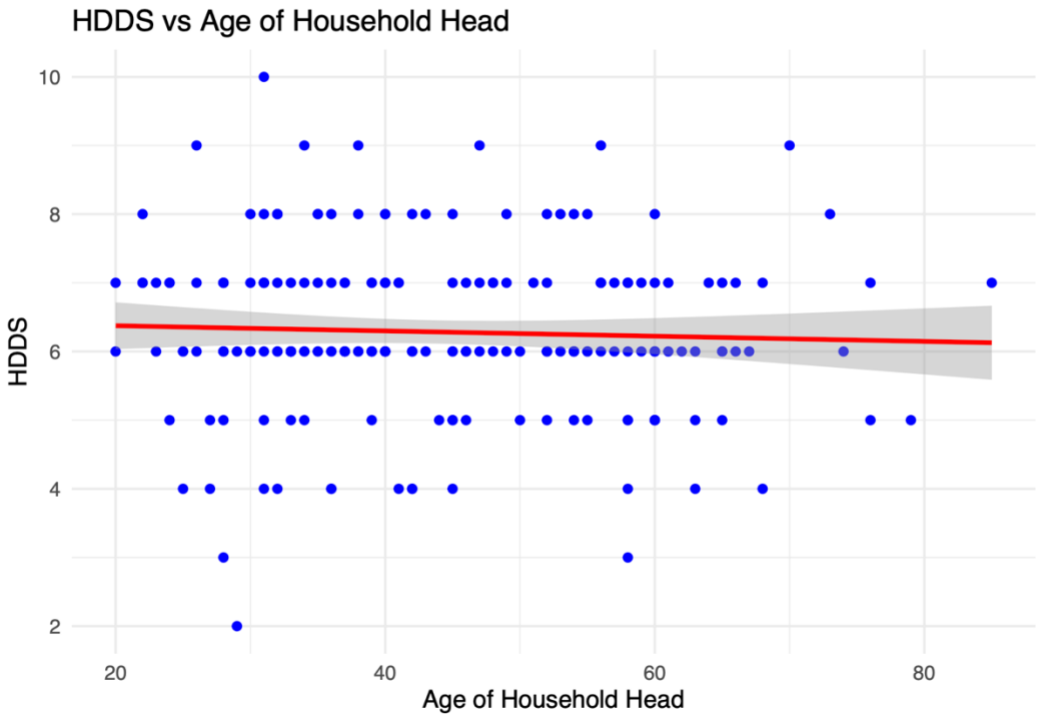


Figure 6: Relationship between HDDS and Age of the Household Head

The plot indicates that there is not a noticeable association between HDDS and Age of the Household Head variable, which is expected based on the rather evenly distributed nature of the Age variable shown in Figure 2. Therefore, it is likely that Age of Household Head will not be significant in the modeling result.

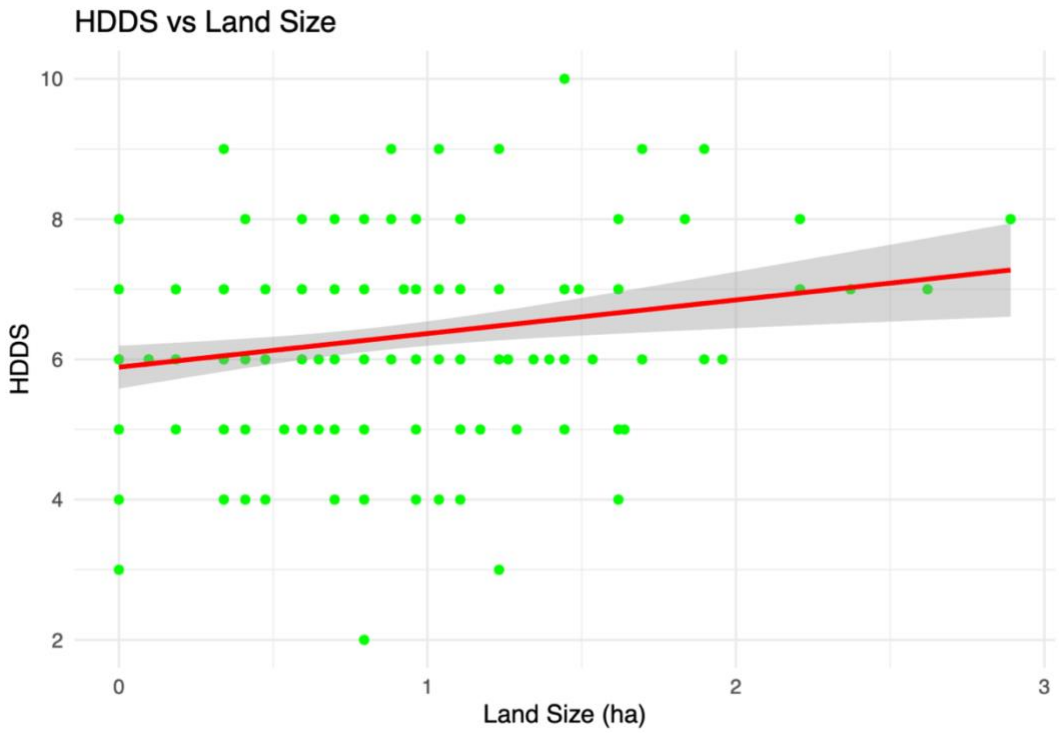


Figure 7: Relationship between HDDS and Land Size (Land Size is transformed)

By observing the plot about the relationship between HDDS and Land Size, there is a slight increase in the values of HDDS with the increase of Land Size, so we might expect to see an association between Land Size and HDDS in the final modeling result.

Model Fitting and Feature Selection

After the EDA, we have gained a thorough understanding about the dataset, and we are ready to proceed to preprocessing of the predictors so that we are prepared for the model fitting and feature selection.

For basic data transformations, I factorized all the categorial socio-economic indicators and performed log transformation for the numerical predictor Land Size. For those with large numbers of levels, such as Age, Household Size, and Village, I also regrouped them by assigning a wider range for each group. To be more specific, the grouping details are as the following:

Age: "0-18", "19-35", "36-50", "51-65", "66+"

Household Size: "1-2", "3-4", "5-6", "7+" (this is the number of people in a household)

Village: Group1: c(6, 4, 3, 8, 11), Group2: c(2, 1, 12, 9, 14), Group3: c(13, 19, 7, 16, 17), Group4: c(20, 10, 15, 5, 18). (there are originally 20 levels).

Considering that the HDDS variable represents the number of dietary categories the household consumed in the last 24 hours, it can be thought of as a count variable for the dietary categories. Therefore, the Poisson Regression Model and Negative Binomial Model are suited for model fitting. The models are fit using “stan_glm” function from the “rstanarm” package, and the expected logarithmic predicted density for Leave-One-Out cross validation (elpd_loo) and Leave-One-Out information criterion (looic) are calculated and used as metrics for performance assessment. The results are shown below:

```
##
## Computed from 6000 by 204 log-likelihood matrix.
##
##          Estimate  SE
## elpd_loo   -405.1 2.4
## p_loo         3.7 0.4
## looic       810.3 4.9
## ------
## MCSE of elpd_loo is 0.0.
## MCSE and ESS estimates assume MCMC draws (r_eff in [0.5, 1.2]).
##
## All Pareto k estimates are good (k < 0.7).
## See help('pareto-k-diagnostic') for details.
```

```
print(loo_nb)
```

```
##
## Computed from 6000 by 204 log-likelihood matrix.
##
##          Estimate  SE
## elpd_loo   -428.3 2.2
## p_loo         3.1 0.3
## looic       856.7 4.4
## ------
## MCSE of elpd_loo is 0.0.
## MCSE and ESS estimates assume MCMC draws (r_eff in [0.4, 1.4]).
##
## All Pareto k estimates are good (k < 0.7).
## See help('pareto-k-diagnostic') for details.
```

```
##               elpd_diff se_diff
```

```
## poisson_model   0.0       0.0
## nb_model      -23.3       0.7
```

Figure 8: Model Performance Results Table

As we can see the Poisson Model exhibits a better overall performance by having both a higher expected logarithmic predicted density for Leave-One-Out cross validation (-405.1 vs -428.3) and a lower Leave-One-Out information criterion score (810.3 vs 856.7), signifying that it has a more accurate and reliable prediction result and that is a better suit for model fitting for our dataset.

We then perform feature selection with the better candidate Poisson Model to remove irrelevant predictors and ensure the set of predictors used for the final model is optimized using cross validation.

```
## 19 x 1 sparse Matrix of class "dgCMatrix"
##                               s1
## (Intercept)         1.473939e+00
## Region2             4.655670e-02
## district1           5.574283e-17
## VillageGroup2       9.893907e-02
## VillageGroup3       6.331696e-02
## VillageGroup4       1.010702e-01
## Household_size3-4   .
## Household_size5-6   3.819562e-02
## Household_size7+    .
## educationF          1.532354e-02
## Age_head19-35       3.616848e-02
## Age_head36-50       .
## Age_head51-65      -2.497025e-03
## Age_head66+         3.787919e-02
## grow_vegetables1    5.136004e-02
## Food_nutrition1     9.178352e-02
## Gender_head1        5.828852e-02
## land_size_ha        1.355511e-01
## off_farm1           .
```

```
##  [1] "Region2"           "district1"         "VillageGroup2"
##  [4] "VillageGroup3"     "VillageGroup4"     "Household_size5-6"
##  [7] "educationF"        "Age_head19-35"     "Age_head51-65"
## [10] "Age_head66+"       "grow_vegetables1"  "Food_nutrition1"
## [13] "Gender_head1"      "land_size_ha"
```

Figure 9: Feature Selection Final Results

By performing feature selection with cross validation, the following features (predictors) will be included in the final model formula: Region, district, Village, Household_size, educationF, Age_head, grow_vagetables, Food_nutrition, Gender_head, and land_size_ha (supplementary plot in Tables and Figures section).

Prior Selection and Performance Evaluation

Now that we have settled on the model metric and modeling formula, we are ready to fit the Poisson model on different prior metrics and arrive at our final model that yields the best performance. The priors I included for model fitting are Normal, Cauchy, and Lasso (Sparsity) (Professor Handcock suggested that I use Sparsity prior in the model fitting, which makes sense. But I also wanted to do an actual evaluation on different priors to achieve a more robust analysis and evaluation for my model). I also encountered Horseshoe prior when I was doing my research, but it seemed suboptimal: it is usually used when there are several strong predictors and associations among a high-dimensional dataset. As we have observed from the previous Exploratory Data Analysis visualization and feature selection results, the association between HDDS and all socio-economic predictors are rather weak, making horseshoe prior unsuited for this dataset. Therefore, I did not consider horseshoe prior. Theoretically, Cauchy prior would also be suboptimal because it is more robust for cases when there are visible outliers in the target variable. Again, since we have seen previously that the distribution for HDDS is rather normal, I would not expect a good performance on the model fitting with Cauchy prior. Normal prior is a standard and common choice when doing model fitting. However, since it is an informative prior, I would prefer a non-informative or weakly informative prior over an informative prior as a Bayesian statistician to ensure that my model is data-driven, and my fitting results are not bias-induced. In my analysis, I tested models with Normal, Cauchy, Lasso, and non-informative priors.

```
# Normal prior
model_normal <- stan_glm(selected_formula, data = data,
                         family = poisson,
                         prior = normal(0, 1), prior_intercept = normal(0, 1),
                         chains = 4, iter = 2000, warmup = 500, cores = 4)


# Cauchy prior
model_cauchy <- stan_glm(selected_formula, data = data,
                         family = poisson,
                         prior = cauchy(0, 2.5), prior_intercept = cauchy(0, 2.5),
                         chains = 4, iter = 2000, warmup = 500, cores = 4)


# Lasso prior
model_lasso <- stan_glm(selected_formula, data = data,
                        family = poisson,
                        prior = laplace(0, 1), prior_intercept = normal(0, 1),
                        chains = 4, iter = 2000, warmup = 500, cores = 4)
```

```
#Non-informative Prior
model_noninformative <- stan_glm(selected_formula, data = data,
                        family = poisson,
                        prior = NULL, prior_intercept = NULL,
                        chains = 4, iter = 2000, warmup = 500, cores = 4)
```

Figure 10: Code for model fitting

```
print(loo_comparison)
```

```
##                       elpd_diff se_diff
## model_lasso            0.0       0.0
## model_normal          -0.4       0.1
## model_cauchy          -0.6       0.1
## model_noninformative -2.1       0.7
```

```
print(waic_comparison)
```

```
##                       elpd_diff se_diff
## model_lasso            0.0       0.0
## model_normal          -0.4       0.1
## model_cauchy          -0.6       0.1
## model_noninformative -2.1       0.7
```

Figure 11: Performance Evaluation

As shown above by using LOO and WAIC (Widely Applicable Information Criterion) as well as checking pareto k values, the model with a Lasso (Sparsity) prior yields the best performance validated by having 0.0 elpd_diff and se_diff in both LOO and WAIC metrics. It is obvious that we should select Poisson model with Lasso prior as our final model given the values in LOO and WAIC.

Final Model and Its Interpretation

Now that we have identified the final model, let's take a closer look at it and interpret what insight it gives regarding our proposed question.

```
## Estimates:
##                       mean   sd   10%   50%   90%
## (Intercept)          1.4    0.2  1.2   1.4   1.7
## Region2              0.0    0.7 -0.8   0.0   0.8
## district1            0.1    0.7 -0.8   0.0   0.9
## VillageGroup2        0.1    0.1  0.0   0.1   0.2
## VillageGroup3        0.1    0.1 -0.1   0.1   0.2
## VillageGroup4        0.1    0.1  0.0   0.1   0.2
## educationF           0.0    0.0  0.0   0.0   0.0
## Food_nutrition1      0.1    0.1  0.0   0.1   0.2
## land_size_ha         0.1    0.1  0.0   0.1   0.3
## Household_size3-4    0.0    0.1 -0.1   0.0   0.2
## Household_size5-6    0.1    0.1 -0.1   0.1   0.2
## Household_size7+     0.0    0.1 -0.2   0.0   0.2
## Gender_head1         0.1    0.1  0.0   0.1   0.2
## Age_head36-50        0.0    0.1 -0.1   0.0   0.1
## Age_head51-65        0.0    0.1 -0.1   0.0   0.1
## Age_head66+          0.0    0.1 -0.2   0.0   0.2
## grow_vegetables1     0.1    0.1 -0.1   0.1   0.2
##
```

```
##            mean   sd   10%   50%   90%
## mean_PPD 6.3    0.2  6.0   6.3   6.6
##
## The mean_ppd is the sample average posterior
##
## MCMC diagnostics
##                     mcse Rhat n_eff
## (Intercept)         0.0  1.0  4719
## Region2             0.0  1.0  3149
## district1           0.0  1.0  3180
## VillageGroup2       0.0  1.0  4676
## VillageGroup3       0.0  1.0  4386
## VillageGroup4       0.0  1.0  4112
## educationF          0.0  1.0  5487
## Food_nutrition1     0.0  1.0  4601
## land_size_ha        0.0  1.0  4928
## Household_size3-4   0.0  1.0  3613
## Household_size5-6   0.0  1.0  3380
## Household_size7+    0.0  1.0  3556
## Gender_head1        0.0  1.0  4239
## Age_head36-50       0.0  1.0  4492
## Age_head51-65       0.0  1.0  3449
## Age_head66+         0.0  1.0  4491
## grow_vegetables1    0.0  1.0  5527
## mean_PPD            0.0  1.0  6627
## log-posterior       0.2  1.0  1108
##
```

Figure 12: Final Model Summary Table

As we can observe from the summary tables above, predictors district, Village, Food_nutrition, land_size_ha, Household_size, Gender_head and grow_vegetables have weak associations with a household's dietary diversity outcome. More specifically, we learn that there's a slight trend of increase for certain district and some villages and for households with larger land sizes and engage in agricultural cultivation of vegetables. If the household has had previous nutritional training, they have a slightly higher chance of having a more diverse diet. The fact that households have a family size of 5-6 people are slightly more likely to achieve a higher dietary diversity is a bit harder to interpret, but I think it makes sense to associate it with land size and treat them as indicators of the relatively high wealth and socio-economic level of the househole, and it makes sense that more affluent families are more likely to pay attention to their nutritional adequacy and diversity. It is also reasonable that Gender_head has a positive association to the nutritional diversity (1 represents female) since female are more likely to pay attention to and be in charge of family diets in conventional sense. Additionally, the all 0 mcse (Monte Carlo Standard Error) for predictors, all 1 values for Rhat, as well as all large n_eff (effective sample size) in the MCMC diagnostic table suggest that the model is stable with low variability, and it is well converged.

Discussion

The analysis underscores the relationship of socio-economic factors to improving household dietary diversity in Tanzania. Variables like land size, gender of the household, geolocation, access to nutritional training, as well as household size are shown to have impacts on the household's dietary diversity scores. However, the association between HDDS and all predictors are shown rather weak, making the final result less indicative of the driving socio-economic factors in enhancing Tanzanian households' nutrition quality. One major flaw in my study is that I couldn't perform more in-depth analysis of the dietary diversity scores due to the lack of data specification in my dataset. The nutritional categories measured for subgroups like women and children are not specified, making it less informative and less feasible to discover the underlying associations between socio-economic factors in determining the presence of certain specific nutritional categories. I read from articles like "Socio-Economic Inequalities in the Double Burden of Malnutrition among under-Five Children: Evidence from 10 Selected Sub-Saharan African Countries" that the issue of malnutrition caused by socio-economic inequalities and the lack of resources, especially among children, persists in many countries in Africa. I think it would be meaningful if I were given more details in the dataset that enables me to conduct investigations into how socio-economic indicators affect specific subgroups like children or women, because there are clearly needs for policy amendments so that groups like children are able to gain access to a more nutritionally diverse diet. Moreover, the data is collected across a span of a month and there are only around 200 observations, it would be more ideal if more observations can be included across a larger geographical and time span.

I would also like to reflect on potential drawbacks in my research methodologies. I think as an alternative to cross-validation, stepwise selection using AIC criterion would be informative too.

Secondly, due to the constraint of computing power, my modeling was performed with a relatively low number of precision and iterations. It would have a lower impact when the dataset is small but would be insufficient in cases of large datasets and the need for fine-tuning.

I think some further steps to take from here are that I can find larger datasets or datasets that target a more specific group and conduct similar studies. I also think it would be helpful to enrich the study interpretation by consulting social scientists who are more experienced in solving socio-economical and nutritional inequality problems.

Conclusion

The study strives to investigate potential determinants of a household's dietary diversity in Tanzania among common socio-economic indicators like age, education level, and geographical location. By utilizing Bayesian statistical modeling and analysis techniques (which ensure the reliability and robustness in the final result), we have arrived at a final model with predictors such as land size, gender of the head of the household, and access to nutritional training to be associated with the dietary diversity score. The analysis and results are presented and discussed in a fashion that is easily understandable and informative to audiences without a statistical background, and potential drawbacks and applicable ways of improvement are reflected in the discussion part. Readers can find more materials in the reference part if they are interested in the topic and are willing to learn more about the field.

References

1. Weerasekara, Permani C, et al. “Understanding Dietary Diversity, Dietary Practices and Changes in Food Patterns in Marginalised Societies in Sri Lanka.” *Foods (Basel, Switzerland)*, U.S. National Library of Medicine, 13 Nov. 2020, www.ncbi.nlm.nih.gov/pmc/articles/PMC7696452/.

2. Ochieng, Justus; Afari-Sefa Victor; Philipo Lukumay; Dubois Thomas, 2017, "Dietary diversity and socio-economic data of households in Tanzania", https://doi.org/10.7910/DVN/INRWQA, Harvard Dataverse, V1,UNF:6:S9+dAUSN0sRGGv8oJkBQGw== [fileUNF].

3. Alaba, Olufunke A, et al. “Socio-Economic Inequalities in the Double Burden of Malnutrition among under-Five Children: Evidence from 10 Selected Sub-Saharan African Countries.” *International Journal of Environmental Research and Public Health*, U.S. National Library of Medicine, 12 Apr. 2023, www.ncbi.nlm.nih.gov/pmc/articles/PMC10138555/.

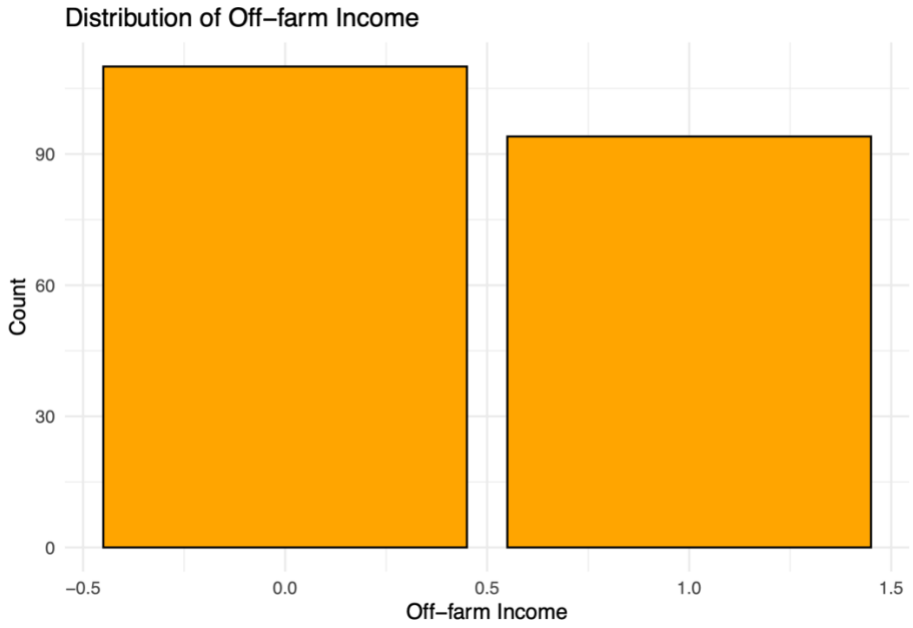
4. Alkerwi, Ala’a, et al. “Demographic and Socioeconomic Disparity in Nutrition: Application of a Novel Correlated Component Regression Approach.” *BMJ Open*, U.S. National Library of Medicine, 11 May 2015, www.ncbi.nlm.nih.gov/pmc/articles/PMC4431064/#:~:text=People%20with%20high%20socioeconomic%20status,to%20their%20poorer%20health%20status.

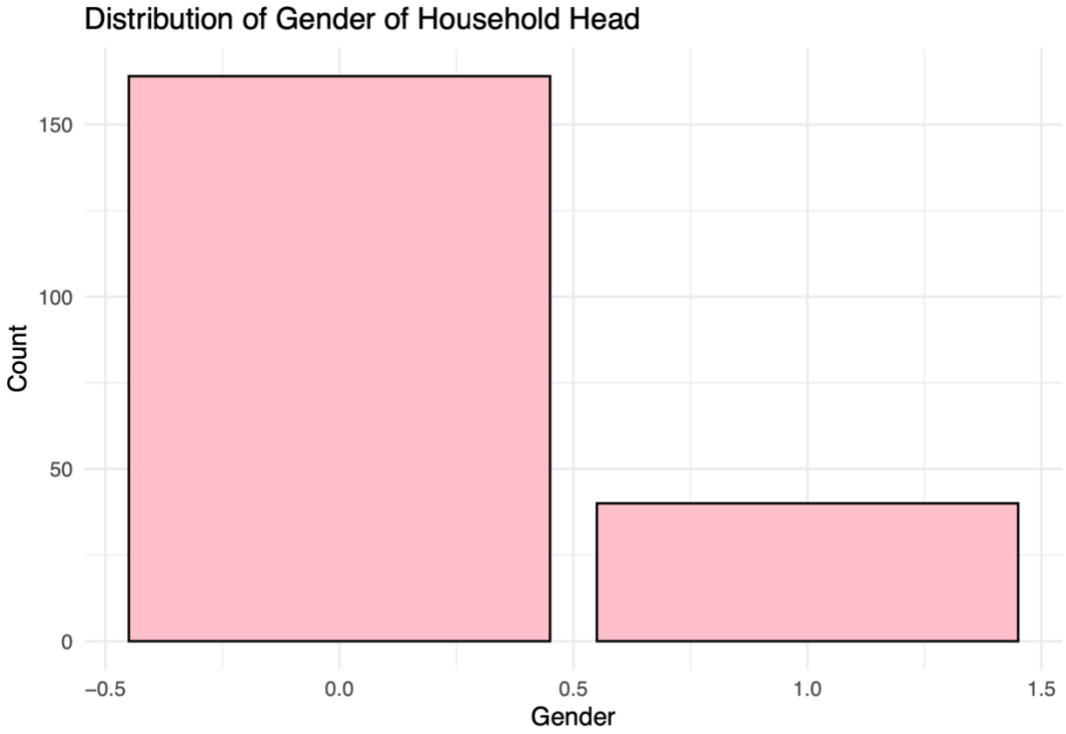
5. “Dietary Diversity.” *Wikipedia*, Wikimedia Foundation, 18 Apr. 2024, en.wikipedia.org/wiki/Dietary_diversity.

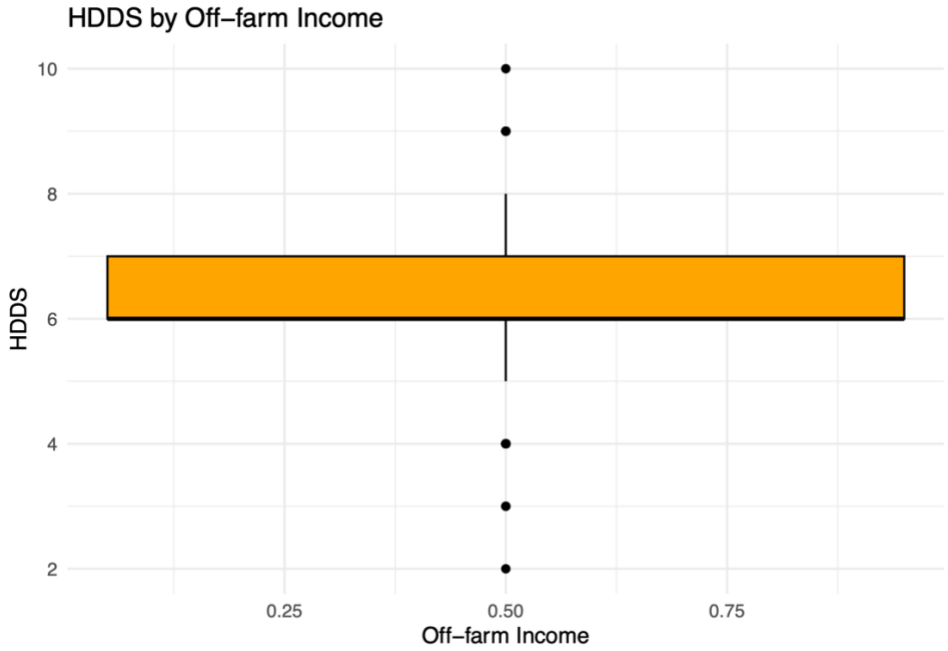
6. *Guidelines for Measuring Household and Individual Dietary Diversity*, www.fao.org/4/i1983e/i1983e00.htm#:~:text=Dietary%20diversity%20is%20a%20qualitative,administered%20low%2Dcost%20assessment%20tool. Accessed 15 June 2024.

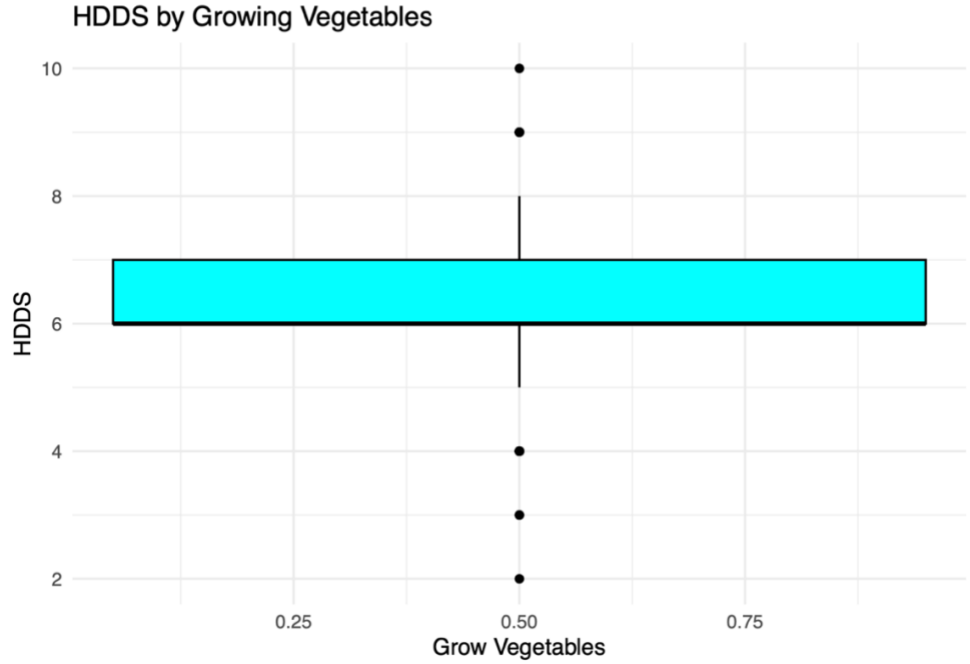
Appendix: Figures

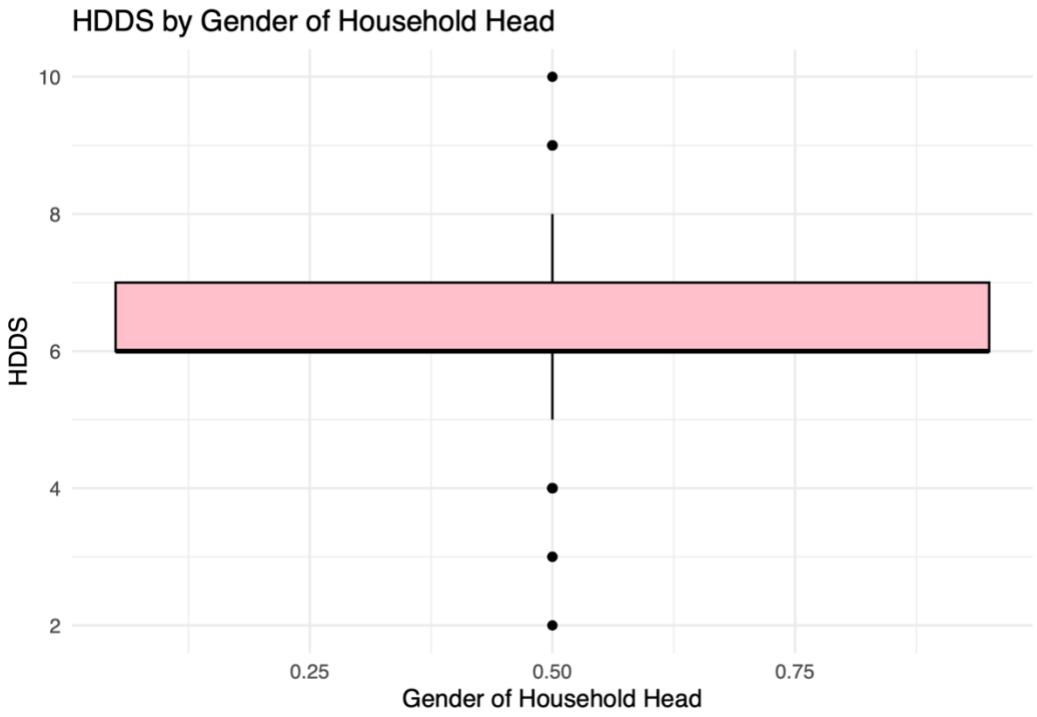
Part A: Supplementary EDA Visualizations




Distribution of Gender of Household Head
150
100
50
0
Count
−0.5
0.0
0.5
1.0
1.5
Gender


HDDS by Off−farm Income
HDDS
10
8
6
4
2
0.25
0.50
0.75
Off−farm Income


HDDS by Growing Vegetables
HDDS
10
8
6
4
2
0.25
0.50
0.75
Grow Vegetables


HDDS by Gender of Household Head
10
8
6
4
2
HDDS
0.25
0.50
0.75
Gender of Household Head