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 widely used as an indicator of nutritional adequacy and a proxy for household socio-economic status. Using dietary diversity and socio-economic data from households in Tanzania (*Dietary diversity and socio-economic data of households in Tanzania*, Ochieng et al., 2017), this study investigates the factors associated with household dietary diversity through Bayesian statistical modeling, specifically Poisson and Negative Binomial regression approaches. The Poisson model was selected over the Negative Binomial model due to its higher expected log predictive density under Leave-One-Out cross-validation (*elpd_loo*) and its lower Leave-One-Out information criterion (*looic*), indicating superior predictive performance.

Subsequent model fitting employed additional evaluation metrics, including LOOIC and WAIC, across a range of prior specifications such as Normal, Cauchy, non-informative, and Lasso priors, to ensure robustness and reliability. The Lasso prior yielded the lowest *elpd_diff* and *se_diff* values (both equal to zero) relative to the alternative priors, providing strong evidence in favor of this specification.

The final model indicates that village location, engagement in vegetable cultivation, participation in nutritional training, and household size are all associated with household dietary diversity. These findings highlight the influence of geographic and socio-economic factors on dietary outcomes and carry important policy implications for improving nutritional quality and general well-being not only in Tanzania but also in similar contexts across sub-Saharan Africa.

Introduction

Household dietary diversity serves as a key indicator of nutritional adequacy and relative food security. In countries such as Tanzania, where agriculture constitutes the primary source of income for many households, identifying the determinants of dietary diversity is essential for informing policy interventions aimed at improving nutritional outcomes and promoting the physical well-being of the population. This is particularly important given that food insecurity and malnutrition remain persistent challenges across many regions of the African continent. In addition, dietary diversity offers valuable insight into broader patterns of economic development: limited dietary diversity may signal deeper structural inequalities in the distribution of resources and wealth, making it an important dimension for social scientists seeking to understand a nation's social functioning and developmental trajectory.

This study investigates the socio-economic and demographic factors associated with household dietary diversity in Tanzania. In particular, it examines the influence of variables such as education, household size, land size, and geographic location on dietary diversity outcomes. To address these questions, the analysis employs Bayesian statistical models to generate robust, data-driven insights into the relationship between socio-economic conditions and household dietary diversity. The central research question guiding this study is: *How are socio-economic and demographic factors associated with household dietary diversity in Tanzania?*

Analysis and Results

Exploratory Data Analysis (EDA)

To develop a clearer understanding of the socio-economic factors under consideration, an exploratory data analysis was conducted on the relevant variables to identify underlying patterns and data structures. The primary variables examined include geographic location, household size, education level, age and gender of the household head, land size, participation in agricultural activities, and the presence of off-farm income sources.

The EDA process consisted of reviewing summary statistics, generating visualizations to examine the distributions of key variables, and assessing their relationships with the target variable, the Household Dietary Diversity Score (HDDS). Summary statistics offer insight into the central tendencies and variability of the data, while distributional visualizations provide an initial perspective on data structure and potential irregularities. In addition, a correlation matrix was used to evaluate multicollinearity among predictors, ensuring that the subsequent regression coefficients remain stable and interpretable in the final model.

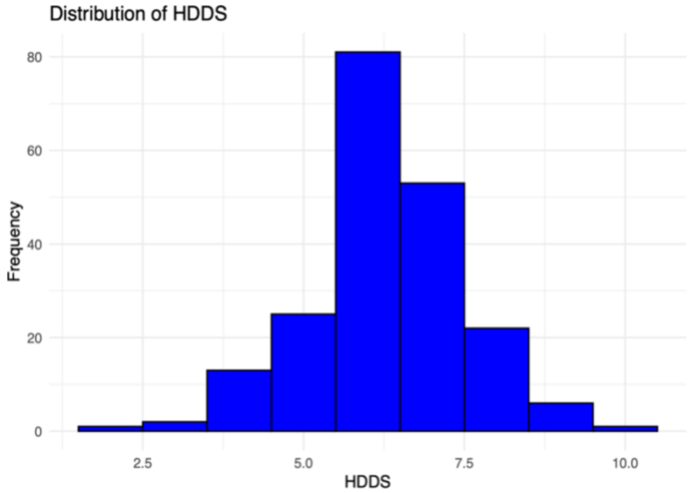


Figure 1. Distribution of HDDS, illustrating an approximately normal shape with most observations between 5 and 7.5.

The distribution of HDDS appears approximately bell-shaped, with most households reporting scores between 5 and 7.5. This centered and relatively symmetric distribution suggests that a Poisson model may be more appropriate than a Negative Binomial model, as there is no clear evidence of overdispersion in the outcome variable.

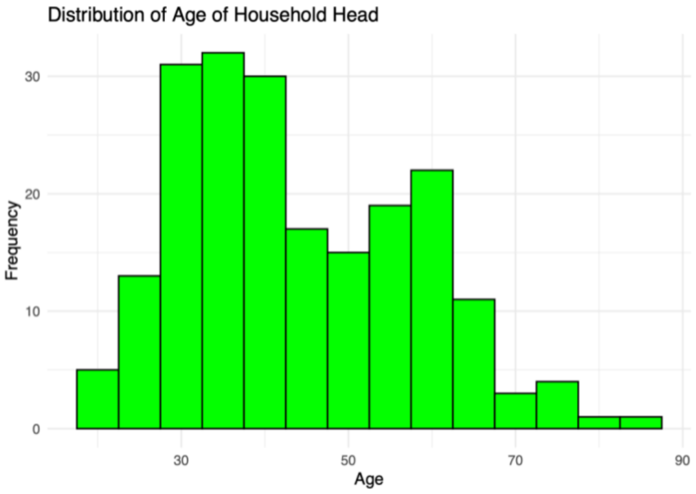


Figure 2. Distribution of Age of Household Head,
showing slight left skewness with most observations between 30 and 50 years.

The distribution of the household head's age exhibits slight left skewness, with most ages concentrated between 30 and 50 years, which is consistent with expectations for household leadership demographics. Given that the degree of skewness is minimal, transformation of the age variable does not appear necessary for model fitting. However, recoding the variable into broader age categories may be beneficial for interpretability and for capturing potential non-linear relationships.

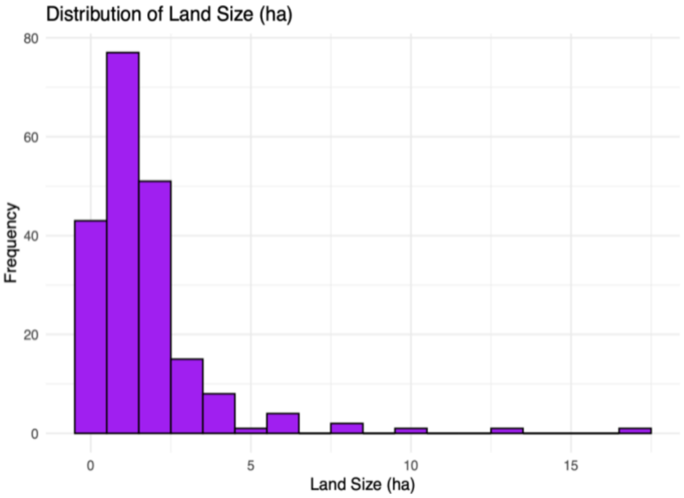


Figure 3. Distribution of household land size, showing strong
left skewness with most values clustered between 0 and 5 hectares.

Based on the visualization, the distribution of land size is highly left-skewed, with most households reporting between 0 and 5 hectares. Given this pronounced skewness, applying a logarithmic transformation is appropriate prior to including the land size variable in the regression models.

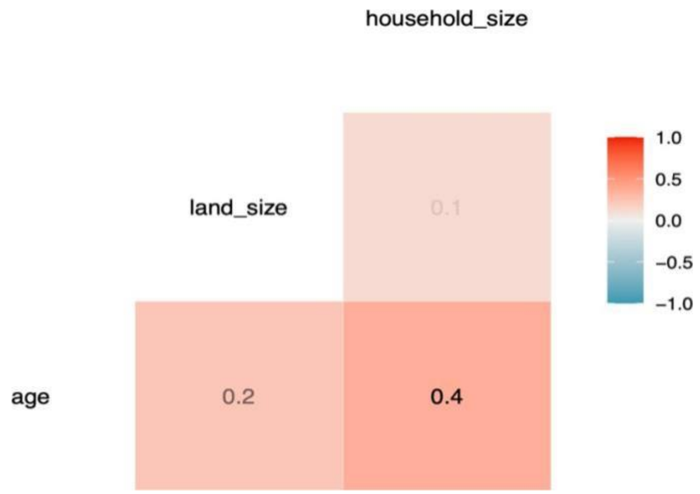
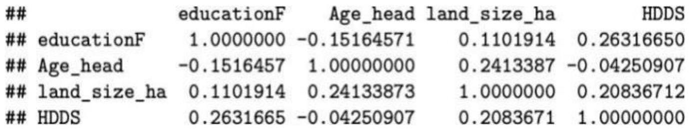


Figure 4. Correlation matrix of numerical variables, showing weak pairwise correlations across all predictors.

The results of the correlation matrix indicate no evidence of substantial multicollinearity among the predictors.



```
##              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 showing weak associations among education, age, land size, and HDDS.

Moreover, the numerical values in the correlation matrix corroborate the patterns observed in the corresponding visualization. All pairwise correlations fall within the range of approximately ±0.3, indicating that multicollinearity is unlikely to pose a concern during model fitting. It is also informative to visualize the relationships between the socio-economic predictors and the HDDS outcome variable to obtain a preliminary sense of their potential associations and to establish baseline expectations for the subsequent modeling process.

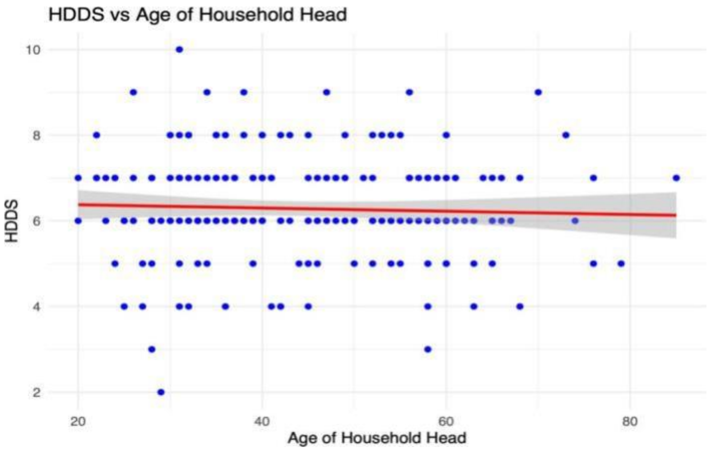


Figure 6. Relationship between HDDS and age of the household head, showing no clear association and an essentially flat trend line.

The plot suggests that there is no clear association between HDDS and the age of the household head, which is consistent with the relatively uniform distribution of the age variable shown in Figure 2. Accordingly, it is unlikely that age will emerge as a significant predictor in the modeling results.

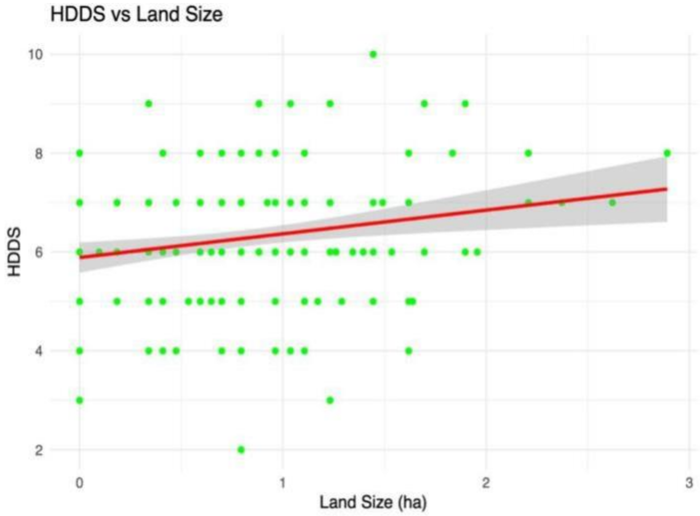


Figure 7. Relationship between HDDS and transformed land size, showing a modest positive association with a slight upward trend.

The plot depicting the relationship between HDDS and land size shows a slight upward trend, with HDDS values increasing as land size increases. This pattern suggests that land size may exhibit a positive association with dietary diversity and could emerge as a meaningful predictor in the final modeling results.

Model Fitting and Feature Selection

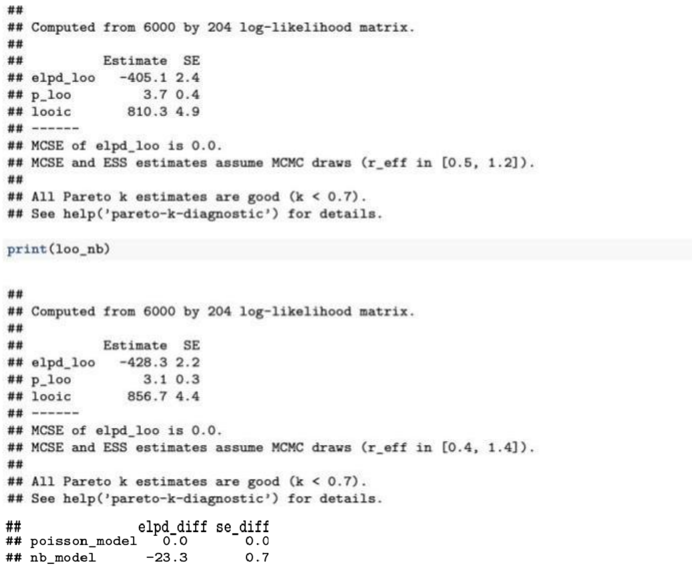
The exploratory data analysis provides an initial overview of the dataset and guides the subsequent preprocessing steps required for model fitting and feature selection. For the basic data transformations, all categorical socio-economic variables are converted into factor variables, and the numerical predictor *land_size_ha* is log-transformed to stabilize variance and reduce skewness. Variables with a large number of distinct levels, such as age, household size, and village, are grouped into broader categories to facilitate more stable estimation. The specific grouping procedures are as follows:

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

Household Size: "1-2", "3-4", "5-6", "7+" (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).

Given that the HDDS variable represents the number of dietary categories consumed by a household within the past 24 hours, it is appropriately treated as a count variable. Accordingly, Poisson regression and Negative Binomial regression are suitable modeling frameworks for this analysis. The models are estimated using the *stan_glm* function from the *rstanarm* package. Model performance is evaluated using the expected log predictive density from Leave-One-Out cross-validation (*elpd_loo*) and the Leave-One-Out information criterion (*looic*), both of which provide robust measures of predictive accuracy. The resulting performance metrics are presented below.

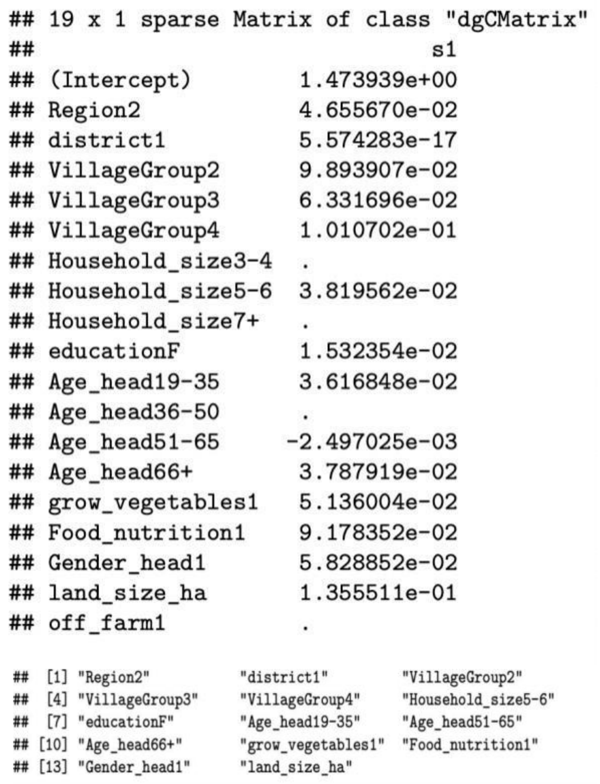


```
##
## Computed from 6000 by 204 log-likelihood matrix.
##
##          Estimate  SE
## elpd_loo   -405.1 2.4
## p_loo         3.7 0.4
## looic       810.3 4.9
## ------
## MCSE of elpd_loo is 0.0.
## MCSE and ESS estimates assume MCMC draws (r_eff in [0.5, 1.2]).
##
## All Pareto k estimates are good (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 comparison showing superior predictive accuracy for the Poisson model, reflected in higher *elpd_loo* and lower LOOIC values relative to the Negative Binomial model.

As shown above, the Poisson model demonstrates superior overall performance, evidenced by a higher expected log predictive density in the Leave-One-Out cross-validation evaluation (−405.1 compared to −428.3) and a lower Leave-One-Out information criterion score (810.3 compared to 856.7). These results indicate more accurate predictive performance and a better fit to the data relative to the alternative model.

Given the Poisson model’s stronger performance, feature selection was subsequently conducted within this framework to remove irrelevant predictors and to ensure that the final model includes an optimally selected set of covariates informed by 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. Final set of predictors retained through cross-validated feature selection, identifying variables with nonzero coefficients in the sparse model.

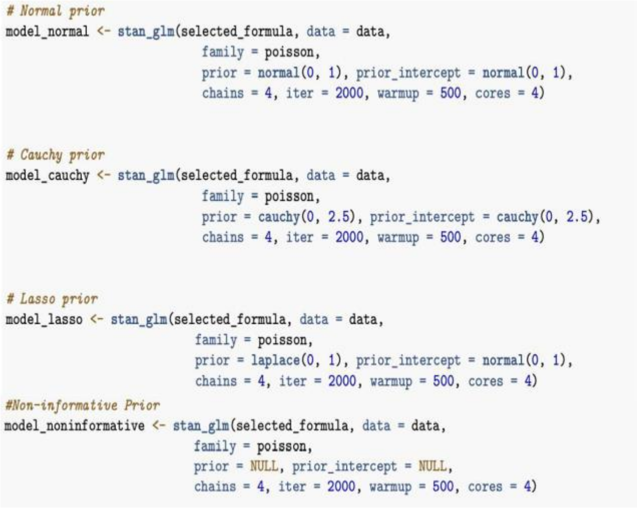
Using feature selection informed by cross-validation, the following predictors were retained for inclusion in the final model specification: *Region*, *District*, *Village*, *Household_size*, *EducationF*, *Age_head*, *Grow_vegetables*, *Food_nutrition*, *Gender_head*, and *Land_size_ha*. A supplementary visualization of the selection process is provided in the Tables and Figures section.

Prior Selection and Performance Evaluation

With the model selection criteria and the modeling framework established, we proceed to fit the Poisson model under several prior specifications in order to identify the configuration that yields the strongest performance. The priors considered include the Normal, Cauchy, and Lasso (sparsity-inducing) priors. A Horseshoe prior was also evaluated, but it proved theoretically and empirically misaligned with the structure of this dataset. Horseshoe priors are most effective in high-dimensional settings with a mix of strong and weak predictors, yet the exploratory data analysis and feature selection results indicate that the associations between HDDS and the socio-economic predictors are uniformly weak. This makes the Horseshoe prior ill-suited for the current context.

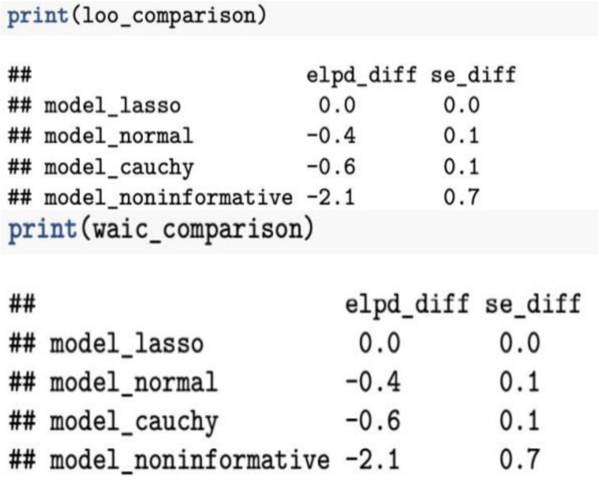
The Cauchy prior is similarly suboptimal. It is typically preferred when the outcome variable exhibits extreme values or heavy-tailed behavior, conditions not present here given the approximately Gaussian distribution of HDDS. The Normal prior represents a conventional choice for Bayesian regression; however, because it is relatively informative, it may impose structure not justified by the data. In applications such as this, where the goal is to allow the data to drive inference, weakly informative or non-informative priors are generally preferred to mitigate the risk of prior-induced bias.

Accordingly, Poisson models with Normal, Cauchy, Lasso, and non-informative priors are fitted and evaluated separately in the analysis presented below.



```r
# 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 used to fit Poisson regression models under Normal,
Cauchy, Lasso, and non-informative prior specifications.



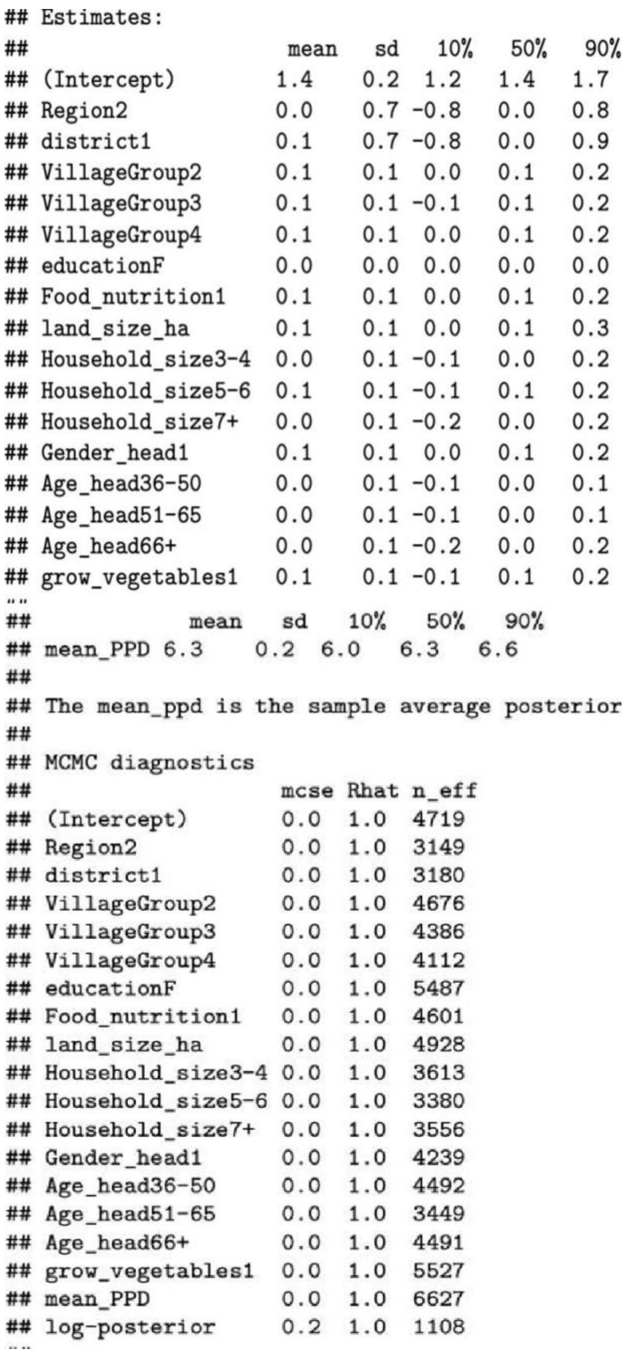
```
print(loo_comparison)

##                      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 comparison of prior specifications, showing that
the Lasso prior achieves the best predictive accuracy based on *elpd_diff* and *se_diff*.

According to the LOO and WAIC criteria, as well as the associated Pareto k diagnostics, the model employing a Lasso (sparsity-inducing) prior demonstrates the strongest performance. This conclusion is supported by its *elpd_diff* and *se_diff* values, both of which are zero under the LOO and WAIC evaluations, indicating no evidence of inferior predictive accuracy relative to competing models. Based on these results, we conclude that the Poisson model with a Lasso prior is the optimal specification for this analysis.

Final Model and Its Interpretation

With the final model established, we now examine it more closely to interpret the insights it provides in relation to the research 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. Summary of the final Poisson model with Lasso prior,
including posterior estimates and MCMC diagnostics demonstrating good convergence.

As shown in the summary tables, the predictors *district*, *village*, *Food_nutrition*, *land_size_ha*, *Household_size*, *Gender_head*, and *grow_vegetables* exhibit relatively weak associations with household dietary diversity. Nonetheless, several modest patterns emerge. Dietary diversity tends to be slightly higher in certain districts and villages, as well as among households with larger land holdings and those engaged in vegetable cultivation. Households that have participated in nutritional training also demonstrate a marginally greater likelihood of achieving a more diverse diet.

The finding that households with 5–6 members exhibit slightly higher dietary diversity is less straightforward to interpret. However, when considered jointly with land size and other indicators of socio-economic position, this pattern may reflect that relatively wealthier households possess both the resources and the awareness necessary to prioritize nutritional adequacy. The positive association for *Gender_head* (with '1' denoting female) is also plausible, as women often hold primary responsibility for food preparation and dietary decisions in many cultural contexts.

Finally, the MCMC diagnostics, specifically the near-zero MCSE values, $\hat{R}$ values of 1.0 across all parameters, and large effective sample sizes, indicate that the model exhibits good convergence and low sampling variability, supporting the reliability of the parameter estimates.

Discussion

The analysis highlights the role of socio-economic factors in shaping household dietary diversity in Tanzania. Variables such as land size, gender of the household head, geographic location, prior nutritional training, and household size all exhibit measurable, though relatively weak, associations with Household Dietary Diversity Scores (HDDS). The modest effect sizes suggest that these predictors alone may not fully capture the socio-economic determinants of nutritional quality, and that additional contextual or behavioral variables may be needed to better explain variation in dietary diversity.

A significant limitation of the study is the restricted level of detail in the dataset, which constrains the depth of analysis. Nutritional categories for key subpopulations, such as women and children, are not explicitly defined, limiting the ability to assess subgroup-specific drivers of dietary diversity. Existing literature, including *Socio-Economic Inequalities in the Double Burden of Malnutrition among Under-Five Children: Evidence from 10 Selected Sub-Saharan African Countries*, underscores the persistence of malnutrition linked to socio-economic inequalities across African contexts, particularly among vulnerable groups. More granular data would allow for disentangling how socio-economic indicators differentially affect these subgroups and would provide policy-relevant evidence for interventions aimed at improving access to nutritionally adequate diets.

The dataset's temporal and spatial limitations further reduce the robustness of the findings. With only approximately 200 observations collected over a single month, the dataset may not capture seasonal or regional variability in dietary patterns. Expanding the sample size across broader geographic regions and longer time frames would significantly strengthen the external validity of the results.

Methodologically, the study would also benefit from incorporating additional model selection and validation approaches. While cross-validation was used, complementary techniques, such as stepwise selection based on the Akaike Information Criterion (AIC), could offer further insight into predictor relevance. In addition, the modeling process was constrained by limited computing resources, resulting in fewer iterations and lower sampling precision. Although this may have minimal impact given the small dataset, such constraints would be problematic in larger-scale analyses where finer convergence diagnostics and higher-precision estimates are required.

Future extensions of this work should involve identifying larger or more specialized datasets and replicating the analysis to improve generalizability and subgroup inference. Engaging with social scientists and researchers specializing in socio-economic and nutritional inequality would further strengthen the study's theoretical grounding and ensure that methodological decisions align with substantive knowledge in the field.

Conclusion

This study examines the potential determinants of household dietary diversity in Tanzania, focusing on commonly cited socio-economic indicators such as age, education level, and geographic location. Using Bayesian statistical modeling and associated diagnostic techniques, which enhance the reliability and rigor of the analytical results, the analysis identifies land size, gender of the household head, and access to nutritional training as key predictors associated with dietary diversity scores. The findings are presented in a manner that remains accessible to readers without a formal statistical background while still maintaining analytical depth. The discussion section also addresses the methodological limitations of the study and outlines possible avenues for refinement. Additional sources are provided in the references for readers who wish to explore the broader literature on dietary diversity and socio-economic determinants. More broadly, this work highlights the value of Bayesian modeling for social and nutritional research in data-limited environments, demonstrating its capacity to extract reliable insights even when analytic conditions are constrained.

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.

Appendix: Figures

Supplementary EDA Visualizations

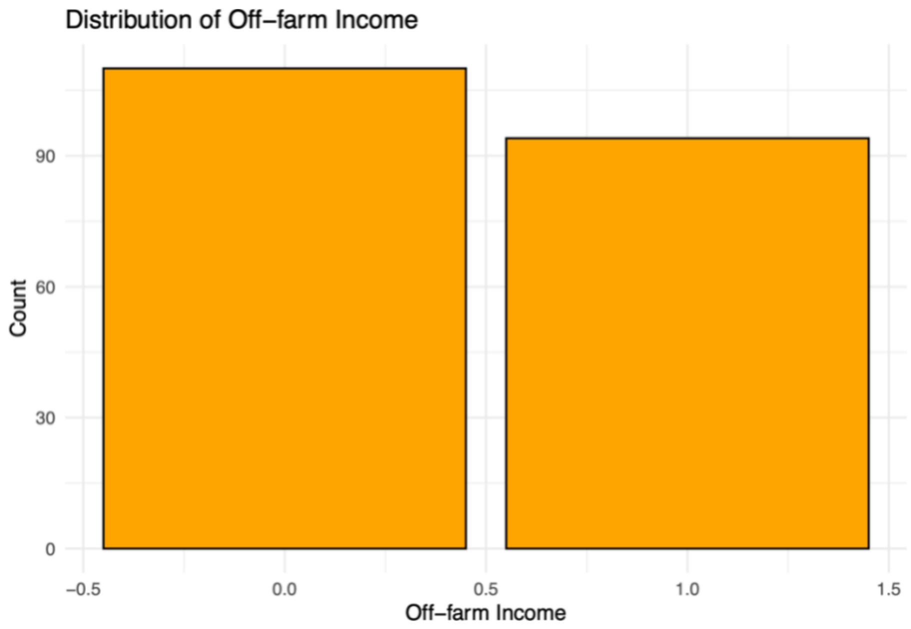


Figure A1. Distribution of off-farm income, showing a roughly even split between households with and without off-farm income sources.

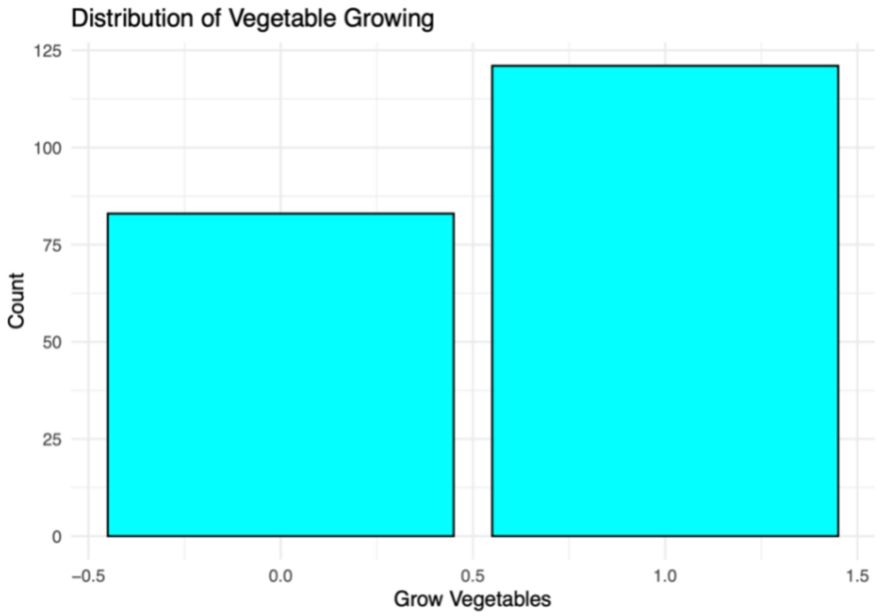


Figure A2. Distribution of household vegetable cultivation, showing that a majority of households engage in growing vegetables.

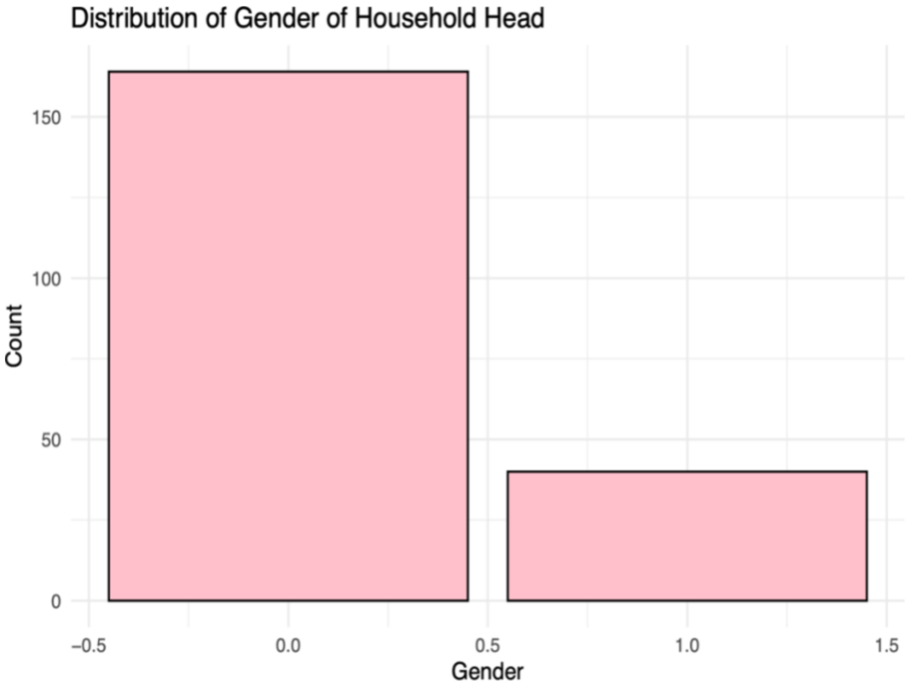


Figure A3. Distribution of household head gender, showing that most households are headed by men.

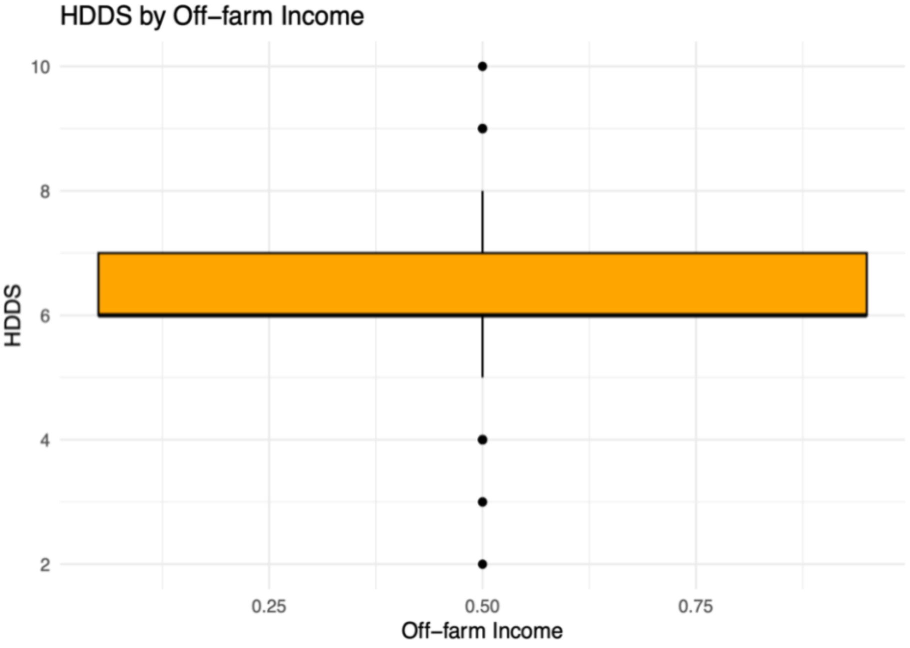


Figure A4. Distribution of HDDS by off-farm income status, showing similar dietary diversity levels between households with and without off-farm income.

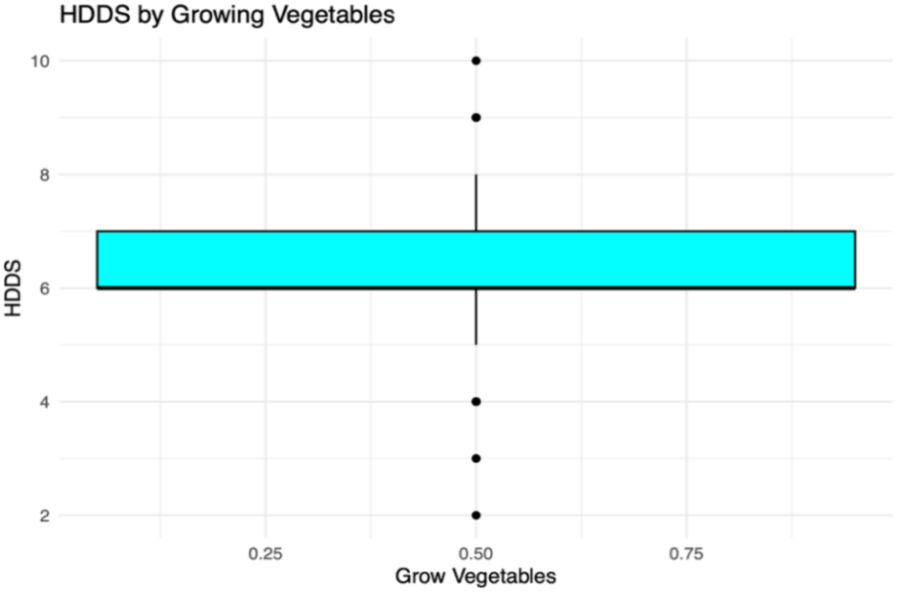


Figure A5. Distribution of HDDS by vegetable cultivation status, showing slightly higher dietary diversity among households that grow vegetables.

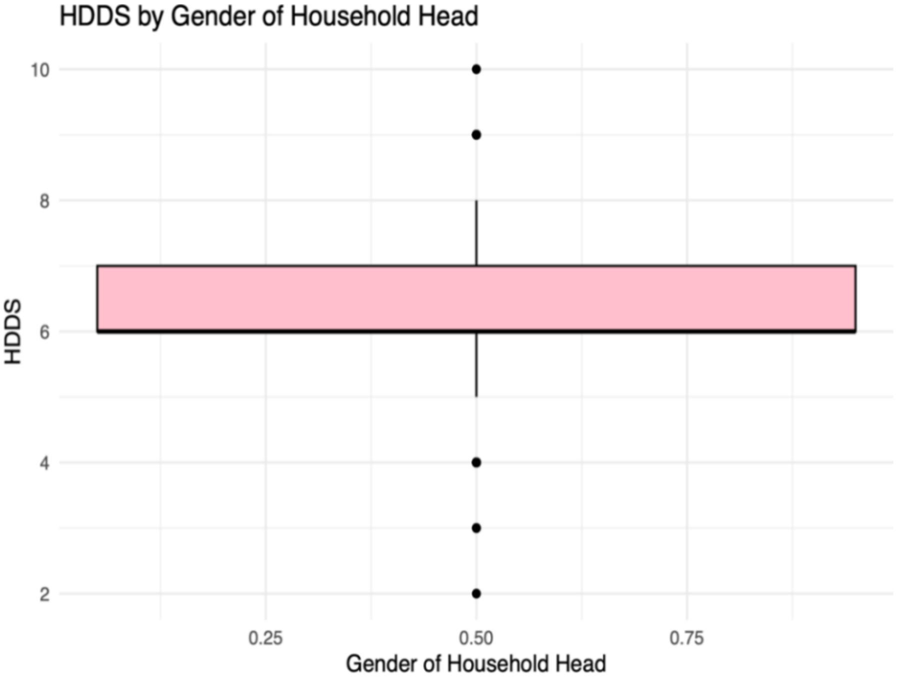


Figure A6. Distribution of HDDS by gender of household head, showing similar dietary diversity levels across male- and female-headed households.