MODELING THE EFFECTS OF SOCIOECONOMIC AND ENVIRONMENTAL FACTORS ON CHILD MALNUTRITION AND CONTAMINATION RISK USING GENERALIZED ESTIMATING EQUATIONS

by

Samuel Amoateng

B.S. University of Ghana, 2020

A Project Submitted in Partial Fulfillment of the Requirements

for the Degree of

Master of Science

in

Statistics and Data Science

University of Alaska Fairbanks

May 2025

APPROVED:

Julie McIntyre, Committee Chair

Scott Goddard, Committee Member

Margaret Short, Committee Member

Leah Berman, Chair

Department of Mathematics and Statistics

Dedication

To My family:

This project is dedicated to my beloved family, whose constant love, support, and encouragement have been my greatest source of strength throughout this academic journey. Their patience and belief in me have strengthened my will and perseverance. I am deeply grateful for their sacrifices, which have made the completion of this work possible.

Abstract

Malnutrition and environmental contamination remain critical public health problems among children under five years old, particularly in developing environments. This paper investigates the sociodemographic, environmental, and behavioral factors that influence the risk of contamination and child malnutrition using Generalized Estimating Equations (GEE) in order to adjust correlations across repeated observations. Two multivariate binary response variables were taken into consideration: SWUbinary comprised of three binary indicators of whether the child was stunted, wasted, or underweight; and RTbinary comprised of two binary indicators of contamination exposure based on the presence of Relative Light Unit (RLU) and Total Coliform Analysis (TCA) contamination in a household. Using Point-Biserial Correlation and Cramér's V Statistic, relevant predictor variables were screened and backward stepwise selection was used to determine the best set of predictors from those remaining for each of three correlation structures: unstructured, exchangeable and independent. The best model for each of the two response variables was chosen using the Akaike Information Criterion (AIC). For SWUbinary, the exchangeable correlation model was selected, and for RTbinary, the independent correlation model was the selected model. The results show how important it is to look at both nutrition and environmental factors together when trying to improve the health and well-being of children.

Acknowledgment

I would like to express my deepest gratitude to the Almighty God for His guidance and strength throughout the course of this project. I also want to thank Professor Andrea Bersamin from the Department of Biology and Wildlife at the University of Alaska Fairbanks for providing the dataset for this research. My sincere appreciation goes to my supervisor, Professor Julie McIntyre for her valuable guidance and constructive feedback throughout the duration of my research.

Table of Contents

List of Tables

Chapter 1: Introduction

Child malnutrition is still an important public health problem worldwide, with the biggest burden in low- and middle-income countries. Over 200 million children under five years are affected by malnutrition (undernutrition and overnutrition), and undernutrition is responsible for approximately 45% of all deaths of children under age five globally. In sub-Saharan Africa and South Asia, prevalence of chronic undernutrition (stunting) and acute undernutrition (wasting) in children continue to be high (Obasohan *et al.*, 2024). Nigeria is a perfect example of this, ranking among the countries with the highest incidence of child malnutrition. In a survey conducted in 2018, it was reported that 37% of Nigerian children under the age of five were stunted, 7% wasted, and 22% underweight (Raiwe, 2020). Despite slight improvements over the past years, Nigeria's child malnutrition levels are still poor, calling for more actions and interventions (Raiwe, 2020).

Domestic contamination, which include unsafe water for drinking, bad sanitation and hygiene, and bad food handling or safety practices at home, is another serious risk to children's health. Contaminated water and unhygienic living conditions increase the number of certain infections, leading to an increase in undernutrition due to loss of nutrients (UNICEF Nigeria, 2025). Only 26.5% of Nigerians live in areas with better access to water and sanitation, while about 22% still defecate in the open (UNICEF Nigeria, 2025). As a result, water-borne infections are common; over 70,000 Nigerian children under five die from diarrhea each year, and the lack of access to clean water, sanitation and hygiene (WASH) is responsible for over 73% of the nation's diarrheal sickness (UNICEF Nigeria, 2025). Children in low-resource areas are likely to have continuous issues of poor WASH and malnutrition, resulting in infections and stunting of growth. For example, a study on Ibadan, Nigeria reported extremely high rates of malnutrition (44% stunting, 34% underweight and 38% wasting) with high diarrheal illness rates, and found that limited access to safe water was strongly associated with child malnutrition in that environment (Aleru *et al.*, 2023). Results like these show how home and environmental contamination, like unsafe water and absence of good sanitation can cause malnutrition in children.

The main objective of this project is to identify the key household, behavioral, and environmental

predictors of malnutrition indicators among children under five years old and of household contamination risk. We will use the Generalized Estimating Equations (GEE) method to do this. GEE is frequently used to evaluate longitudinal or clustered data. The findings will help us understand the burden of malnutrition and contamination risks and promote more strategies to improve child health.

Chapter 2: Methodology

2.1 Overview

This study adopts an analytical approach to investigate child nutrition in Nigeria, using an epidemiological dataset. Malnutrition in early childhood can have serious effects both in the short and long term. In the short term, children who are malnourished are more likely to get sick because their immune systems are weaker. Over time, malnutrition can affect them, making it harder for children to learn and grow properly. The primary objective is to explore the associations between nutritional outcomes and various socioeconomic and environmental factors.

Given the nature of the dataset, where multiple observations are clustered within individuals and households, the Generalized Estimating Equations (GEE) framework was chosen for statistical modeling (Liang & Zeger, 1986). GEEs are particularly useful for handling correlated data and produces robust estimates. This method takes care of within-cluster dependencies, ensuring that the standard errors are correctly estimated. Unlike other methods, such as Generalized Linear Mixed Models (GLMM), which account for subject-specific variability using random effects, GEE estimates population-averaged effects, focusing on the general trend across all individuals rather than making inferences about specific subjects (Venables *et al.*, 2002).

2.2 Dataset

The dataset used was from a study that investigates food safety and nutrition status of children under five years of age in Ibadan, Oyo state, Nigeria (Atoloye *et al.*, 2024). It includes observations of children within households that record socioeconomic and environmental factors that affect the outcomes of malnutrition. The data was carefully collected to capture a wide range of factors, including nutritional outcomes, socio-economic conditions, and environmental factors. The indicators of malnutrition in children were measured on a total of 625 households. In addition, for a subset of 236 households, researchers measured the presence of Relative Light Unit (RLU) and Total Coliform Analysis (TCA) contamination from swabs of surfaces in each household. RLU contamination is the type which detects organic material, like bacteria, on surfaces, whereas TCA contamination refers to the presence of coliform bacteria, which are indicators of microbial con-

tamination.

2.3 Study Location

This study looks at Nigerian children's nutrition, specifically, children in Ibadan metropolis in Oyo state. Ibadan is the third most populous city in Nigeria, being home to about 3 million people (Carrell, 2024). Ibadan Metropolis consists of five local government areas (Ibadan North, Ibadan North-West (NW), Ibadan North-East (NE), Ibadan South-West (SW), and Ibadan South-East (SE)), which were all considered (Atoloye *et al.*, 2024). Climate, agriculture, socioeconomic conditions, and access to healthcare vary greatly throughout all areas of the city. These differences help explain why children's nutritional outcomes vary. To successfully address malnutrition, specific interventions must be designed with an understanding of these variations.

2.4 Variables in the Study Dataset

This study uses a subset of variables from the larger dataset mentioned earlier. These variables were selected based on their relevance to understanding the factors contributing to malnutrition. We considered two sets of clustered response variables. The first consisted of three health indicators of malnutrition (stunting, wasting, underweight) measured on children in each household. The second consisted of two indicators of bacterial contamination (RLU and TCA) measured on the subset of households. Table 2.1 provides a summary of the key variables analyzed in this study.

Table 2.1: Key Variables In The Dataset

Variable	Description
Stunting	Binary indicator for stunting status (1 = stunted, 0 = not stunted)
Wasting	Binary indicator for wasting status (1 = wasted, 0 = not wasted)
Underweight	Binary indicator for underweight status (1 = underweight, 0 = not underweight)
RLUcontam	Indicator of contamination based on Relative Light Units (pass, fail)
TCAcontam	Indicator of contamination based on Total Coliform Analysis (none, some, very)
Age	Child's age in months
hc4_gender	Gender of child
hc2a_familysize, hc2c_familysize	Total number of household members
Nation_Windex	Household wealth index, a composite measure based on asset ownership and living conditions.
hc9a_eduhhdhead_recode	Education of household head, categorized as primary, secondary, or tertiary
hc9b_eduhhdcook_recode	Education of household cook
hc10_occuphhdcook_recode	Occupation of household cook (self-employed or paid-employed)
hc8a_nativelang_recode, hc8b_ethnicgrp_binary	Indicators for native language spoken and ethnic group classification in a household
hs_toilet_Binary	Binary variable indicating whether the household has access to a sanitary toilet (1 = Yes, 0 = No)
hs_shared	Binary indicator of whether the household shares living spaces with others (1 = Yes, 0 = No)
hc12_ownbuildg_binary	Binary indicator of whether the household owns its residence
DD_Scores	Dietary diversity score, with higher scores indicating better nutrition
fsa_control	An indicator of household food security
fss_runout, fss_money	Indicators for food shortages, including running out of food or lacking money to buy food
hc3b_hhdcookage	Age of the person responsible for cooking in a household
hc13_cookwhere_recode	Type of kitchen setup used in the household (Private or shared kitchen)
hc5_mstatus_binary	Marital status, single vs. non-single household heads
hc7_religion	Household's religious affiliation
hws_Binary	Indicator of household access to safe drinking water
swm_kitchenwaste, swm_disposed, swm_animals	Household waste disposal practices, important for hygiene and sanitation
SDG_Classes	Categorical variable aligning households with Sustainable Development Goals (SDG) indicators.

2.4.1 Variable Screening

Selecting the appropriate predictor variables is a crucial step in ensuring the effectiveness and precision of statistical modeling. Given that the dataset for this study contains a large number of continuous and categorical predictors, as a first step, we eliminated variables that had little association with the responses. To achieve this, Point-Biserial Correlation ($|r_{pb}|$) and Cramér's V statistic were used to identify the most relevant continuous and categorical variables, respectively. This approach helped to reduce the number of predictors considered, by screening out those with little evidence of association with the response variable.

2.4.2 Point-Biserial Correlation for Continuous Variables

Point-Biserial Correlation ($|r_{pb}|$) is a specialized form of Pearson's correlation used when one variable is continuous and the other is binary. This correlation quantifies the strength and direction of the relationship between the continuous predictor and the binary outcome (Kornbrot, 2014). It is computed as:

$$r_{pb} = \frac{\bar{X}_1 - \bar{X}_0}{s} \times \sqrt{\frac{p_1 p_0}{n}} \tag{2.1}$$

where;

- $\bar{X}_1$ and $\bar{X}_0$ are the means of the continuous variable for the two binary groups,
- s is the standard deviation of the continuous variable,
- p_1 and p_0 are the proportions of observations in each group,
- n is the total sample size.

A higher absolute value of r_{pb} indicates a stronger relationship between the continuous predictor and the binary outcome. For this study, Point-Biserial Correlation was computed between each binary response and all continuous predictor variables in the dataset. Variables that exhibited a weak association ($|r_{pb}| < 0.1$) for all components in each multivariate response were excluded from further modeling to improve efficiency.

2.4.3 Cramér's V Statistic for Categorical Variables

Cramér's V Statistic is a measure of association between two categorical variables based on the chi-square test. It quantifies the strength of the relationship without assuming a linear correlation. It is given by:

$$V = \sqrt{\frac{\chi^2}{n(k-1)}} \tag{2.2}$$

where;

- χ^2 is the test statistic for the chi-square test of association,
- n is the total sample size,
- k is the smaller of the number of rows or columns in the contingency table.

Cramér's V ranges from 0 to 1, where values closer to 1 indicate a stronger association between the categorical predictor and the binary outcome. In this study, Cramér's V will be calculated for all categorical predictor variables and each binary response. Variables with weak association ($V < 0.1$) for all components in each multivariate response were excluded from further analysis to retain only meaningful predictors.

2.4.4 Implementation in Variable Screening

To systematically refine the selection of predictor variables, the following approach was taken:

1. Compute Point-Biserial Correlation for all continuous variables against the binary outcomes.
2. Compute Cramér's V Statistic for all categorical variables against the binary outcomes.
3. Set predefined threshold values to retain only significant predictors ($|r_{pb}| > 0.1$ for continuous variables and $V > 0.1$ for categorical variables).
4. Exclude variables with low associations from further modeling steps.

Thus, for both multivariate binary responses, a predictor was retained if it met the threshold criterion for any one of the individual components that contributed to the response variable.

2.5 Statistical Methods

2.5.1 Generalized Estimating Equations (GEE)

GEEs were used in this study to analyze the relationship between nutritional outcomes (stunting, wasting, underweight) and various socio-demographic, environmental, and behavioral predictors. GEEs are particularly suited for data where repeated measurements or observations within clusters are correlated, as it accounts for within-cluster dependencies while providing robust, population-averaged estimates. This makes it ideal for public health studies that aim to generate findings that are applicable to entire populations rather than individual-level predictions (Liang & Zeger, 1986). One advantage of GEEs is that they produce consistent parameter estimates even when the correlation structure is misspecified, making it a flexible and reliable choice for analyzing correlated data (Hardin & Hilbe, 2002). This ensures that even if the true nature of within-cluster correlation is unknown, the estimated effects of predictors remain valid for informing public health interventions.

2.5.2 Fundamentals of GEEs

The GEE method, a quasi-likelihood extension, is becoming more popular for the analysis of longitudinal and other correlated responses, especially if they are binary or counts (Venables *et al.*, 2002). The primary goal of GEE is to estimate the regression parameters in the presence of correlated responses without making strong assumptions about the distribution of random effects.

GEEs are designed to provide reliable estimates of model parameters along with standard errors, without requiring the full specification of a likelihood function. This is particularly advantageous when dealing with multivariate categorical responses, where constructing a complete likelihood can be challenging. A key aspect of GEEs is their ability to account for the correlation among repeated observations of the same subject while maintaining robustness in estimation. Rather than needing an exact specification of the correlation structure, GEEs allow these correlations to be estimated directly from the data. This flexibility ensures that the model remains effective even if the working correlation structure is not perfectly specified, making GEEs a practical choice for analyzing clustered or longitudinal data.

2.5.3 Comparison with GLM and GLMM

GEEs, Generalized Linear Models (GLMs), and Generalized Linear Mixed Models (GLMMs) are widely used statistical methods for analyzing non-normal response data. Even though they share similarities in handling various types of outcome distributions, they differ significantly in how they account for dependencies within the data. Understanding these differences is important for selecting the most appropriate modeling approach based on the research objectives and data structure.

GLMs extend ordinary regression models by allowing the response variable to follow distributions from the exponential family, such as binomial, poisson, or gamma. They establish a relationship between the response and predictor variables through a link function, which transforms the mean response as follows:

$$g(\mu) = \mathbf{X}\beta \tag{2.3}$$

where $g(\cdot)$ is the link function, μ represents the expected response, and β is the vector of regression coefficients. However, GLMs assume that all observations are independent, which can lead to biased estimates when analyzing repeated measures or clustered data, as these settings naturally involve within-group correlations (Fitzmaurice *et al.*, 2012).

To address this limitation, GEEs introduce a working correlation structure that models the dependencies among repeated observations within the same subject or cluster. Unlike GLMs, GEEs estimate marginal effects, capturing the overall impact of predictors on the entire population rather than focusing on individual differences (Liang & Zeger, 1986). One of the key advantages of GEEs is their response to misspecification of the correlation structure, as the estimates of β remain consistent due to the use of strong variance estimators.

In contrast, GLMMs extend GLMs by incorporating random effects to model subject-specific deviations. This approach allows for the estimation of both fixed effects, which describe the general trends across the population, and random effects, which capture individual-level variations (Bolker *et al.*, 2009). The GLMM formulation takes the form:

$$g(\mu_{ij}) = \mathbf{X}_{ij}^T\beta + \mathbf{Z}_{ij}^T u_i \tag{2.4}$$

where $\mathbf{X_{ij}}$ is the vector of predictors with dimension $p \times 1$, $\mathbf{Z_{ij}}$ represents the random effects design matrix, and $\mathbf{u_i}$ is the vector of subject-specific deviations. Unlike GEEs, which provide marginal

estimates, GLMMs produce subject-specific effects, making them particularly useful when the objective is to model individual variations over time or within clusters. However, GLMMs require correct specification of the random effects distribution, making them computationally more complex and sensitive to model assumptions.

2.5.4 Mathematical Framework for GEE

The general form of the GEE model is given as:

$$g(\mu_{ij}) = \mathbf{X}_{ij}^T \beta, \tag{2.5}$$

where:

- $g(\cdot)$ is the link function (e.g., logit for binary outcomes),
- μ_{ij} is the expected value of the outcome for observation j in cluster i,
- $\mathbf{X_{ij}}$ is the vector of predictors with dimension $p \times 1$,
- β is the vector of regression coefficients with dimension $p \times 1$.

The working correlation matrix, $\mathbf{R}_i(\alpha)$, is a square matrix with dimension $m_i \times m_i$ where m_i is the number of observations in cluster i. This accounts for within-cluster correlations:

$$\mathbf{V}_i = \mathbf{A}_i^{1/2} \mathbf{R}_i(\alpha) \mathbf{A}_i^{1/2}, \tag{2.6}$$

where $\mathbf{V}_i$ is a matrix that represents the covariance structure of the response variable within cluster i, $\mathbf{A}_i$ is a diagonal matrix of variance functions, and α is the parameter vector for the correlation structure. For binary outcomes, such as stunting, wasting, and underweight, the logit link function is used:

$$g(\mu_{ij}) = \log\left(\frac{\mu_{ij}}{1 - \mu_{ij}}\right), \tag{2.7}$$

which transforms probabilities into the log-odds scale for linear modeling.

2.5.5 Correlation Structures Tested

Three correlation structures were tested during the modeling process to determine the best fit for the data:

- **Independent Correlation**: Assumes no correlation between observations within the same cluster. This served as a baseline for comparison and can be useful when correlations are weak.
- **Exchangeable Correlation**: Assumes that all pairs of observations within a cluster have the same correlation. This is suitable for datasets with homogeneous within-cluster relationships.
- **Unstructured Correlation**: Places no constraints on the correlation matrix, allowing correlations between all pairs of observations to vary freely. This is the most flexible structure but requires a large sample size for reliable estimation.

The choice of the working correlation structure can influence the efficiency of the parameter estimates. However, one of the strengths of GEEs is that even if the chosen working correlation structure is misspecified, the estimates of β remain consistent, although standard errors may be affected. The choice of correlation structure was guided by model fit statistics, including the Quasi-likelihood under Independence Model Criterion (QIC) (Hin & Wang, 2009).

2.5.6 Which Correlation Structure is Best?

Choosing the right correlation structure in GEEs is essential, as it influences the efficiency of parameter estimates and the accuracy of standard errors. Unlike models that require strict assumptions about random effects, GEE makes it easy to specify a working correlation structure based on the nature of the data.

As seen earlier, all three correlation structures play different roles as far as within-cluster correlation is concerned. We compared different correlation structures using model selection criteria such as the Akaike Information Criteria (AIC). A lower AIC value suggests a better fitting model, guiding the choice of the most appropriate correlation structure for the data.

2.5.7 Initial Model And Final Model

The initial models for the two multivariate response variables, that is, SWUbinary (malnutrition status) and RTbinary (presence of contamination), were specified by including all the predictor variables found during the variable screening procedure explained earlier. These models were fit with each of the three correlation structures. For achieving the best-fit models, backward selection according to the Quasi-likelihood Information Criterion (QIC) was used. In the process of this iterative approach, a few predictors were dropped, and a final model was achieved for each correlation structure. The smallest AIC value model was used as the best-fit model for each response variable and kept for interpretation.

2.5.8 Software and Packages Used

The statistical analyses were conducted using the R programming language (R Core Team, 2024), which provided a solid framework to implement advanced statistical methods. Several specialized packages were utilized to facilitate the analysis:

- **glmtoolbox**: This package (Vanegas *et al.*, 2024) was very important for fitting GEE models, testing various correlation structures, and calculating model performance metrics, including AIC. Its advanced features enabled systematic exploration of the dataset's clustered structure.

- **geepack**: This package (Højsgaard *et al.*, 2006) complemented glmtoolbox by providing additional tools for validating results and fitting GEE models. It ensured consistency in parameter estimates and allowed for cross-validation of correlation structures.

Chapter 3: Results

3.1 Overview

The results of the analysis are shown in this section, with a focus on the variables affecting health outcomes linked to malnutrition and bacterial contamination in homes. The findings are arranged to provide a clear picture of the relationship between behavioral, environmental, and sociodemographic factors and nutritional outcomes (wasting, stunting, and underweight) and household contamination (RLU and TCA contamination).

The GEE technique was employed to handle within-cluster correlations arising from repeated measures. Two multivariate binary response variables were modeled: the three-dimensional response, SWUbinary, indicated whether a child experienced any of the three outcomes indicating malnutrition, and the two-dimensional response variable, RTbinary, which indicated either of the household contamination status.

The section starts with the results of the variable screening procedure using Cramér's V for categorical variables and Point-Biserial Correlation for continuous predictors. Next is a summary of the model selection using Akaike Information Criterion (AIC) to select the best set of predictors for each correlation structure (unstructured, exchangeable, and independent). To maintain clarity and brevity, detailed results are reported only for the best-fitting model that has lowest AIC among the three correlation structures considered.

3.2 Results For SWUbinary

3.2.1 Model Screening Results

Table 3.1 presents the list of predictor variables retained after the initial screening process for the SWUbinary. The selection was based on the strength of association between each variable and the response, using Point-Biserial Correlation for continuous variables and Cramér's V for categorical variables. Only variables with a correlation or statistic value greater than 0.1 were considered for further modeling.

Table 3.1: Selected Variables After Screening Using Point-Biserial Correlation and Cramér's V (SWUbinary)

Point-Biserial Correlation (Continuous Variables)	PB Correlation
Nation_Windex	0.1482
age	0.2383
hc3b_hhdcookage	0.1114
Cramér's V (Categorical Variables)	**Cramér's V**
SDG_Classes	0.1402
swm_kitchenwaste	0.1051
hc9b_eduhhdcook_recode	0.1333
hc8a_nativelang_recode	0.1297
hc13_cookwhere_recode	0.1136
hc9a_eduhhdhead_recode	0.1016
hc12_ownbuildg_binary	0.1305
hc8b_ethnicgrp_binary	0.1106
hs_shared	0.1101
hs_disposed_binary	0.1099

3.2.2 Model Selection Results

After the variable screening process, the first set of GEE models were fit. For each of the correlation structure, a full model with the selected predictors was fitted and then reduced using backward selection to remove non-important variables.

To determine the most important correlation structure and model, the AIC values of the final GEE models for SWUbinary were compared across the three correlation structures. The results are shown in Table 3.2.

Table 3.2: AIC Comparison for Final Models (SWUbinary)

Model	AIC
final.model.unst	1581
final.model.exch	1538
final.model.indep	1592

Among the three correlation structures considered, the Exchangeable correlation structure yields the lowest AIC value. Since a lower AIC value indicates a model being a better fit, the Exchangeable correlation structure was selected as the most appropriate for modeling SWUbinary. This indicated that assuming a constant correlation between repeated observations within individuals provides a

better model fit than assuming unstructured or complete independence (zero) correlation.

3.2.3 Interpretation Of Best Model

This section presents the results of the GEE analysis, which looks at the relationship between child malnutrition and the covariates. Table 3.3 presents the final model estimates using the Exchangeable correlation structure. The model includes only the most relevant predictors after variable screening, telling us how various factors affect the nutritional health outcomes of children.

Table 3.3: Final GEE Model Results For The Response SWUbinary (Exchangeable Correlation). Variables in bold are significant at $\alpha = 0.05$

Variable	Estimate	Std. Error	z-value	p-value	Odds Ratio
Intercept	**1.725**	**0.808**	**2.135**	**0.03276**	**5.61**
Measurement_Underweight	**-1.022**	**0.102**	**-10.009**	$\mathbf{< 2 \times 10^{-16}}$	**0.36**
Measurement_Wasting	**-2.607**	**0.197**	**-13.200**	$\mathbf{< 2 \times 10^{-16}}$	**0.07**
SDG_ClassesSDG.2	0.689	0.428	1.611	0.10721	1.99
SDG_ClassesSDG.3	-0.499	0.642	-0.777	0.43697	0.61
SDG_ClassesSDG.4	0.415	0.426	0.975	0.32963	1.51
hc9b_eduhhdcook_recodeprimary	0.266	0.588	0.453	0.65086	1.30
hc9b_eduhhdcook_recodesecondary	-0.401	0.566	-0.709	0.47849	0.67
hc9b_eduhhdcook_recodetertiary	-0.871	0.578	-1.506	0.13205	0.42
swm_kitchenwastekitchen.2	**-0.625**	**0.249**	**-2.512**	**0.01201**	**0.54**
swm_kitchenwastekitchen.3	-0.347	0.336	-1.034	0.30120	0.71
swm_kitchenwastekitchen.4	-0.114	0.288	-0.396	0.69230	0.89
swm_kitchenwastekitchen.5	0.051	0.405	0.127	0.89888	0.95
age	**-0.023**	**0.005**	**-4.260**	$\mathbf{2.05 \times 10^{-05}}$	**0.98**
hc3b_hhdcookage	**-0.030**	**0.010**	**-2.940**	**0.00328**	**0.97**

Working Correlation Matrix

$$\begin{bmatrix} 1.00 & 0.22 & 0.22 \\ 0.22 & 1.00 & 0.22 \\ 0.22 & 0.22 & 1.00 \end{bmatrix}$$

From Table 3.3, we observe that, the MeasurementUnderweight and MeasurementWasting variables remained highly significant, with coefficients of -1.022 and -2.607, respectively. This suggests that, the odds of a child found to be malnourished when examined for being underweight were 64% lower than the odds of being found to be malnourished when examined for stunting, and the odds of

a child found to be malnourished when examined for wasting were 93% lower than the odds of being found to be malnourished when examined for stunting, given fixed values of the covariates. Also, an estimate of -1.022 for MeasurementUnderweight means that, the log odds of being underweight are 1.022 units lower than the log odds of being stunted. The second kitchen waste disposal variable showed a significant negative association with an estimate of -0.625, an odds ratio of 0.54 and a p-value of 0.012, indicating that children in this waste disposal category had about 46% decrease in odds of experiencing the outcome variables compared to the first level of this variable. Age of child and age of household cook variables were also retained, with age of child having a very small p-value and an odds ratio of 0.98, indicating a decreasing likelihood of any of the three outcomes as age increases.

The working correlation structure for the exchangeable model reported a moderate positive within-cluster correlation of 0.22, which means that a child who is affected by one indicator of malnutrition (i.e., stunting, wasting & underweight) is likely to be minimally affected by another malnutrition indicator.

3.3 Results for RTbinary

3.3.1 Model Screening Results

Table 3.4 presents the variables selected for the RTbinary response. The same procedure was used as described earlier in the model screening section of SWUbinary. This process ensured that only predictors with meaningful relationships to the contamination outcome were included in the GEE model for further analysis.

3.3.2 Model Selection Results

To determine the best-fitting correlation structure for RTbinary, the final models under the Unstructured, Exchangeable, and Independent structures were compared using their AICs. Table 3.5 shows the AIC values for the three models.

As seen in Table 3.5, the Independent correlation structure provides the best model fit for RTbinary. There is a very small difference in AIC values among the models which may suggest that modeling the within-cluster correlation does not highly improve the model's performance for contamination-

Table 3.4: Selected Variables After Screening Using Point-Biserial Correlation and Cramer's V (RTbinary)

Point-Biserial Correlation (Continuous Variables)	PB Correlation
Nation_Windex	0.1717
T_FSknowledge	0.1644
T_FSbehavior	0.1198
hc2c_familysize	0.1250
hc3b_hhdcookage	0.1008
Cramér's V (Categorical Variables)	**Cramér's V**
SDG_Classes	0.1271
hs_toilet_Binary	0.1725
hs_shared	0.1669
hs_disposed_binary	0.1796
swm_kitchenwaste	0.1104
hc9b_eduhhdcook_recode	0.1351
hc9a_eduhhdhead_recode	0.2295
hc10_occuphhdcook_recode	0.1376
hc13_cookwhere_recode	0.2222
fss_money	0.1090
swm_disposed	0.1400

Table 3.5: AIC Comparison for Final Models (RTbinary)

Model	AIC
final.model.unst	536
final.model.exch	536
final.model.indep	532

related outcomes. As a result, we will retain the independent model.

3.3.3 Interpretation Of Best Model

Table 3.6 similarly presents the final model estimates using the Independent correlation structure, which was selected based on the AIC. Again, the model includes only the most relevant predictors after variable screening.

From Table 3.6, we can observe that the variable MeasurementTCAcontam has a positive coefficient of 0.483 with a statistically significant p-value of 0.024, indicating that, the log odds of TCA contamination is 0.483 units higher than that of RLU contamination, the baseline contamination

Table 3.6: Final GEE Model Results For The Response RTbinary (Independent Correlation - RTbinary). Variables in bold are significant at $\alpha = 0.05$

Variable	Estimate	Std. Error	z-value	p-value	Odds Ratio
Intercept	1.781	1.993	0.893	0.372	5.935
MeasurementTCAcontam	**0.483**	**0.213**	**2.267**	**0.024**	**1.621**
SDG_ClassesSDG.2	-0.369	0.773	-0.478	0.633	0.691
SDG_ClassesSDG.3	**-2.614**	**0.936**	**-2.793**	**0.005**	**0.073**
SDG_ClassesSDG.4	-0.846	0.772	-1.095	0.273	0.429
hs_toilet_Binarytoilet.2	**-0.980**	**0.367**	**-2.675**	**0.007**	**0.375**
hs_toilet_Binarytoilet.3	0.440	0.586	0.752	0.452	1.553
hc9a_eduhhdhead_recodeprimary	0.113	0.678	0.166	0.868	1.120
hc9a_eduhhdhead_recodesecondary	0.376	0.534	0.703	0.482	1.456
hc9a_eduhhdhead_recodetertiary	-0.046	0.551	-0.083	0.934	0.955
hc10_occuphhdcook_recodepaidemploy	**1.031**	**0.457**	**2.258**	**0.023**	**2.804**
hc10_occuphhdcook_recodeselfemploy	0.503	0.424	1.184	0.236	1.653
hs_sharedshared.1	**0.586**	**0.257**	**2.284**	**0.023**	**1.797**
hc9b_eduhhdcook_recodeprimary	-0.416	1.058	-0.392	0.695	0.660
hc9b_eduhhdcook_recodesecondary	-0.407	1.010	-0.403	0.687	0.666
hc9b_eduhhdcook_recodetertiary	0.015	1.064	0.014	0.989	1.015
swm_kitchenwastekitchen.2	0.376	0.387	0.971	0.331	1.457
swm_kitchenwastekitchen.3	0.647	0.497	1.300	0.193	1.910
swm_kitchenwastekitchen.4	0.129	0.466	0.278	0.782	1.138
swm_kitchenwastekitchen.5	-0.654	0.508	-1.287	0.199	0.520
Nation_Windex	-0.484	0.386	-1.254	0.210	0.616
T_FSknowledge	-0.097	0.112	-0.862	0.389	0.908

Working Correlation Matrix

$$\begin{bmatrix} 1.00 & 0.00 \\ 0.00 & 1.00 \end{bmatrix}$$

variable. The odds ratio of 1.62 for this variable also tells us that children have about 1.6 times the odds of being found in contaminated TCA environments compared to failing the contamination threshold of RLU contamination.

Education-related variables, such as hc9a_eduhhdhead_recodeprimary and secondary, all show positive coefficients but with non-significant p-values, indicating that higher education levels of household heads may not really be linked to increased contamination risk. Looking at hc9b_eduhhdcook_recode, which links to the educational level of the household cook. Despite not strong p-values, the negative coefficients may suggest that higher education levels may be associated with reduced contamination

risk.

Finally, hs_toilet_Binarytoilet.2, we observe a negative coefficient of -0.980 and a significant p-value of 0.007, indicating that having this toilet type reduces contamination odds. The odds ratio of 0.375 shows that this group experiences about 62.5% lower odds of contamination failure relative to the reference toilet type. The variable hc10_occuphhdcook_recodepaidemploy, which indicates paid employment status of the household cook, we found a significant p-value with a positive coefficient of 1.031, and an odds ratio of 2.80. This may imply that children in households where the cook is paid-employed are more than twice as likely to experience contamination failure.

The independent structure suggested in this model aligns with the working correlation matrix, which assumes zero correlation between repeated measures. This supports the model's simpler structure by indicating that repeated contamination observations for the same child or family are not correlated.

3.4 Discussion

3.4.1 Malnutrition Model (SWUbinary)

Using SWUbinary as the response variable, the analysis of malnutrition showed that Measurement-Wasting and MeasurementUnderweight were significant variables of malnutrition. In line with other research showing the fact that stunting can often be the most chronic and persistent form of malnutrition, children identified as wasting had much lower odds of being malnourished than those classified as stunting (Islam *et al.*, 2018). Similar findings have been seen in longitudinal studies, where stunting is a sign of long-term malnutrition and recurrent infection exposure, while wasting is mostly linked to more severe nutritional deficiencies. (Ntenda & Chuang, 2018).

The model was later improved by adding contamination variables (TCAcontam and RLUcontam), which showed a possible connection between environmental contamination and malnutrition. The results showed that children who were exposed to a certain amount of TCA contamination were more than twice as likely to suffer from malnutrition as children who were not exposed to any contamination. This aligns with research that suggests that microbial and toxic contamination of food

and water raises the risk of illnesses, especially diarrhea, which hinders the absorption of nutrients and leads to poor nutritional outcomes (Andrews-Trevino *et al.*, 2021). The effect of RLUcontam was not statistically significant, which may suggest that the degree of contamination is not adequately captured in a way that is highly correlated with the risk of malnutrition.

Several studies have connected children's growth deficiencies to long-term exposure to environmental contaminants. For instance, Budge *et al.* (2019) noted that environmental enteric dysfunction (EED), a subclinical condition that affects gut function and nutrient absorption, has been linked to greater rates of poor water, sanitation, and hygiene (WASH) conditions. The necessity of integrated WASH and nutrition interventions to address both foodborne and environmental concerns is further supported by the significant association our study also revealed between TCA contamination and malnutrition.

The fact that socioeconomic factors contributed to malnutrition but were not necessarily reliable predictors was another important observation in this study. The sustainable development objectives (SDG.ClassesSDG) and other economic indicators were not statistically significant in the final model, despite the fact that some household-level variables, such as the education of the household cook and waste disposal practices, displayed weak relationships. This is in contrast to studies in Malawi and Bangladesh, where economic status has been consistently linked to malnutrition risk (Islam *et al.*, 2018); (Ntenda & Chuang, 2018). The lack of significance in this study may be due to context-specific factors, such as differences in public health interventions or food security programs, or dataset limitations.

Generally, these results demonstrate the complex and multifaceted nature of malnutrition, in which dietary determinants, illnesses, household conditions, and environmental exposures all have interrelated effects. Given the association between TCA contamination and malnutrition, it is important to address contamination risk as an important part of nutritional interventions.

3.4.2 Contamination Model (RTbinary)

The contamination risk analysis (RTbinary) identified significant economic and household exposure factors. The strongest predictor was MeasurementTCAcontam, which was statistically significant and had a very small p-value. This result is in line with previous research that tracked the concentration of fecal indicator bacteria in household environment using GEE models (Harris *et al.*, 2019). Given the high predictive power of TCA contamination, high-risk households should prioritize checking the quality of their food and water.

A key takeaway from the contamination model is that economic factors play an important role in contamination exposure. Higher contamination risk was shown to be substantially associated with financial predictors (fss_moneymoney.2), indicating that households with lower incomes might not be able to afford the right waste disposal methods, safe food storage, or water treatment. This is in alignment with previous studies showing that poverty is a major factor in household contamination since low-income families frequently use unimproved water sources and lack proper sanitary facilities (Harris *et al.*, 2019). Furthermore, a South Sudanese study discovered that children from lower-income families were more likely to eat contaminated food, most likely as a result of improper handling and storage techniques (Wells *et al.*, 2024).

Waste disposal practices (swm_disposed) were another major factor in contamination risk. Contamination was more common in households using unimproved disposal techniques, which is in line with research showing a link between an increased bacterial presence in household environments and inadequate sanitation (Budge *et al.*, 2019). This is particularly concerning because consuming unhealthy food and water increases the risk of illnesses and causes long-term nutritional deficiencies.

Interestingly, household head's education (hc9b_eduhhdcook_recode) was not a significant predictor of contamination risk in this study. This stands in contrast to research conducted in Tanzania and Ecuador, which found that lower levels of contamination in stored water and feeding utensils were associated with better maternal education (Harris *et al.*, 2019). This study's lack of significance may indicate that environmental elements, such the quality of the infrastructure and food handling

procedures, are more reliable indicators of contamination than education alone.

The model's results support the idea that economic and infrastructure factors have an influence on contamination risk. Various interventions are needed to curb contamination exposure, such as enhancing the infrastructure for sanitation, expanding access to clean water, and offering financial assistance for proper handling and storage of food.

3.5 References

Aleru, E.O., Bodundeb, I.O., Denirana, I.A., Ajania, O., Olayinkac, A.O. & Fawolea, A.O. (2023) Wash practices increased the prevalence of malnutrition among under-five children (6–59 months) in an urban slum area in Ibadan, Nigeria. *Journal of Water, Sanitation and Hygiene for Development* **13**, 910–920.

Andrews-Trevino, J.Y., Webb, P., Shively, G., Kablan, A., Baral, K., Davis, D., Paudel, K., Shrestha, R., Pokharel, A., Acharya, S. *et al.* (2021) Aflatoxin exposure and child nutrition: measuring anthropometric and long-bone growth over time in nepal. *The American journal of clinical nutrition* **113**, 874–883.

Atoloye, A.T., Samuel, F., Aluko, O.O., Torimiro, N., Bamgbade, B., Areola, A.A., Otegbayo, B., Leger, D. & Bersamin, A. (2024) Factors associated with caregivers' food safety knowledge, behavior, perception of food safety control, and the nutrition status of under-5 children in nigeria. *BMC public health* **24**, 2813.

Bolker, B.M., Brooks, M.E., Clark, C.J., Geange, S.W., Poulsen, J.R., Stevens, M.H.H. & White, J.S.S. (2009) Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in ecology & evolution* **24**, 127–135.

Budge, S., Parker, A.H., Hutchings, P.T. & Garbutt, C. (2019) Environmental enteric dysfunction and child stunting. *Nutrition reviews* **77**, 240–253.

Carrell, G. (2024) 31 facts about ibadan. *Facts.net* .

Fitzmaurice, G.M., Laird, N.M. & Ware, J.H. (2012) *Applied longitudinal analysis*. John Wiley & Sons.

Hardin, J.W. & Hilbe, J.M. (2002) *Generalized Estimating Equations*. Chapman and Hall/CRC, New York, 1st edn.

Harris, A.R., Pickering, A.J., Boehm, A.B., Mrisho, M. & Davis, J. (2019) Comparison of analytical techniques to explain variability in stored drinking water quality and microbial hand contamination of female caregivers in Tanzania. *Environmental Science: Processes & Impacts* **21**, 893–903.

Hin, L.Y. & Wang, Y.G. (2009) Working-correlation-structure identification in generalized estimating equations. *Statistics in medicine* **28**, 642–658.

Højsgaard, S., Halekoh, U. & Yan, J. (2006) The r package geepack for generalized estimating equations. *Journal of Statistical Software* **15**, 1–11.

Islam, M.M., Sanin, K.I., Mahfuz, M., Ahmed, A.S., Mondal, D., Haque, R. & Ahmed, T. (2018) Risk factors of stunting among children living in an urban slum of Bangladesh: findings of a prospective cohort study. *BMC public health* **18**, 1–13.

Kornbrot, D. (2014) Point biserial correlation. *Wiley StatsRef: Statistics Reference Online* .

Liang, K.Y. & Zeger, S.L. (1986) Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

Ntenda, P.A.M. & Chuang, Y.C. (2018) Analysis of individual-level and community-level effects on childhood undernutrition in Malawi. *Pediatrics & Neonatology* **59**, 380–389.

Obasohan, P.E., Walters, S.J., Jacques, R. & Khatab, K. (2024) Socio-economic, demographic, and contextual predictors of malnutrition among children aged 6–59 months in Nigeria. *BMC nutrition* **10**, 1.

R Core Team (2024) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raiwe, C. (2020) The nutritional status of Nigerian children: What needs to change. *The Nation* .

UNICEF Nigeria (2025) Water, sanitation and hygiene. Accessed: 2025-03-31.

Vanegas, L.H., Rondón, L.M. & Paula, G.A. (2024) *glmtoolbox: Set of Tools to Data Analysis using Generalized Linear Models*. R package version 0.1.12.

Venables, W.N., Ripley, B.D., Venables, W. & Ripley, B. (2002) Random and mixed effects. *Modern Applied Statistics with S* pp. 271–300.

Wells, J., Abugo, D.G., Angong, J., Lamwaka, N.G., Gallandat, K., Hassan, J.L., Deng, L., Save, D., Braun, L., Gose, M. *et al.* (2024) Risk factors for food contamination among children discharged from community management of acute malnutrition programmes in South Sudan: A cross-sectional study and hazard analysis critical control point approach. *Maternal & Child Nutrition* **20**, e13612.