Journal of Statistical Research *https://doi.org/10.3329/jsr.v58i2.80614*
2024, Vol. 58, No. 2, pp. 317-333 *ISSN 0256 - 422 X*

THE IMPACT OF MATERNAL EMPLOYMENT ON CHILD NUTRITIONAL DIVERSITY IN BANGLADESH: A CAUSAL FOREST ANALYSIS WITH CLUSTERED DATA

SAMIUL EHSAN

Institute of Statistical Research and Training (ISRT), University of Dhaka, Dhaka 1000, Bangladesh

Email: sehsan@isrt.ac.bd

MD RASEL BISWAS*

Institute of Statistical Research and Training (ISRT), University of Dhaka, Dhaka 1000, Bangladesh

Email: rasel@du.ac.bd

SUMMARY

Understanding the impact of women's employment on children's nutrition is crucial for informing effective public health policies. This study examines the relationship between mothers' employment status and the dietary diversity of their children, aged 6 months to 5 years, in Bangladesh. The Nutritional Variety Score (NVS) is used as a measure of dietary diversity, capturing the consumption of various food groups, including eggs, meat, bread, potatoes, vegetables, fruits, fish, beans, and dairy products. To explore this relationship, advanced statistical methods were employed, including causal forest models with cluster identifiers and mixed-effects multilevel logistic regression for propensity scores. The analysis utilized data from the 2022 Bangladesh Demographic and Health Survey (BDHS), a comprehensive dataset encompassing information on women's employment, household characteristics, and children's dietary intake. The models controlled for several confounding variables, including the number of children, partner's education and employment status, type of residence, wealth index, and mother's education level. The results reveal that children of employed mothers have a higher NVS than those of non-employed mothers, with an estimated average treatment effect (ATE) of 0.532 (95% CI: 0.365-0.699). This finding suggests that working mothers may have better access to resources or opportunities to provide a more diverse diet for their children. The statistically significant ATE confirms a positive causal relationship between women's employment and children's nutritional variety. This study contributes to the literature by offering robust evidence on how maternal employment affects child nutrition in Bangladesh.

Keywords and phrases: Women's Employment, Child Nutrition, Causal Forest, Bangladesh

* Corresponding author

1 Introduction

The relationship between women's employment status and their children's nutritional outcomes is a critical focus of public health research. Evidence suggests that women's empowerment through employment can improve economic conditions and increase autonomy, which may, in turn, lead to improved child nutrition (Galiè et al., 2019; Onah, 2021; Shroff et al., 2009). However, this relationship is complex and influenced by factors such as socioeconomic status, cultural norms, and household dynamics. For example, women's control over household income and decision-making has been positively associated with improved child nutrition in some contexts, although poverty can moderate these effects (Bernal Rivas and Lorenzana Albert, 2003; Onah, 2021). Additionally, time constraints resulting from employment may reduce caregiving time, potentially compromising dietary diversity and health outcomes in children (Galiè et al., 2019; Shroff et al., 2009).

This study investigates the causal effect of maternal employment on children's nutrition in Bangladesh. Understanding this relationship is essential for policymakers and health practitioners designing interventions to support both women's employment and children's nutritional well-being. The primary independent variable is the mother's employment status, categorized as working or not working, while the dependent variable is the child's Nutritional Variety Score (NVS), which captures dietary diversity. Several confounding variables, including the number of children, partner's education and employment status, type of residence (urban/rural), wealth index, and the mother's education level, are considered to ensure the accuracy of the causal inference.

To estimate the causal effect, this study employs a *causal forest* approach combined with a mixed-effects multilevel logistic regression model for propensity score estimation. The causal forest method is well-suited to this analysis as it accommodates high-dimensional data and complex interactions. The dataset for this study comes from the 2022 Bangladesh Demographic and Health Survey (BDHS), which provides detailed information on women's employment, household characteristics, and children's nutritional intake.

The primary objective of this research is to estimate the Average Treatment Effect (ATE) of maternal employment on children's NVS. The findings will offer insights into whether employment positively or negatively influences children's dietary diversity, accounting for socio-economic and demographic factors. This study contributes to the existing literature on women's empowerment and child nutrition by providing robust causal evidence. It also offers a methodological template by combining causal forest modeling with cluster-specific propensity scores, which may be valuable for future research in similar contexts.

1.1 Literature review

An increasing presence of women in the workforce is associated with improved dietary options for families (Malapit et al., 2013). Employed mothers often have greater financial capacity to provide their children with nutritious food, underscoring the significant role mothers play in managing their children's health and nutrition. However, evidence from South Asia indicates that low levels of women's empowerment can negatively affect children's dietary diversity, leading to high rates of undernutrition (Cunningham et al., 2015). Furthermore, household wealth has been found to have a strong influence on children's growth and development, with wealthier households being more likely to provide adequate nutrition (Hong et al., 2006). Thus, reducing poverty and ensuring access to essential services for low-income households is important for improving child nutrition in Bangladesh.

Dietary diversity is essential for meeting daily nutritional needs, yet among low-income populations in developing countries, diets are often dominated by starchy staples with limited consumption of animal products, fruits, and vegetables (Arimond and Ruel, 2004). Greater dietary diversity is positively correlated with improved nutritional status (Arimond and Ruel, 2002). Different foods provide distinct nutrients that are essential for children's growth. According to the 2022 BDHS, children under five years old in Bangladesh consume various types of food such as eggs, liver, red meat, chicken, and fruits. This study aggregates data on these individual food items to construct the NVS, capturing a comprehensive measure of dietary variety.

The concept of women's empowerment has been extensively studied. Empowerment refers to an individual's ability to make life choices that were previously unavailable to them (Schuler and Rottach, 2010). It is a complex social process that enables individuals to gain greater control over their lives (Mandal, 2013). One study using DHS data identified five key dimensions of women's empowerment: (i) working status, (ii) apprehension, (iii) self-respect, (iv) self-confidence, and (v) participation in decision-making (Soharwardi and Ahmad, 2020). Empowerment is thus characterized by the capacity or freedom to act as they wish, control over resources, involvement in the workplace, and self-respect. A child's development is heavily influenced by the mother's well-being and behavior. Therefore, factors that empower women are also likely to impact key aspects of child development, including nutrition.

Brand et al. (2023) highlight how machine learning enhances causal inference in sociological research by improving effect estimation, understanding group variations, analyzing causal pathways, and accounting for social interactions over time and space. Integrating machine learning with traditional methods reduces bias and enhances accuracy. Machine learning techniques like deep representation learning, adversarial networks, and balancing methods help manage high-dimensional data, while stable learning improves model robustness under changing data distributions. These advancements have practical applications, particularly in healthcare, aiding better decision-making (Cui et al., 2020)

In this study, the application of causal forests enables the estimation of treatment effects while accounting for the hierarchical cluster structure of the data. Incorporating cluster-specific propensity scores helps address clustering effects and ensures more reliable estimates of the causal relationship between maternal employment and child nutrition. The use of mixed-effects multilevel logistic regression models ensures that both individual-level and cluster-level covariates are controlled for, yielding more robust and unbiased treatment effect estimates (Suk et al., 2021). This approach underscores the importance of considering both individual and contextual factors when investigating the effects of maternal employment on child outcomes.

2 Data Source and Variables

2.1 Data

The 2022 Bangladesh Demographic and Health Survey (BDHS) provides comprehensive data on household members, individual households, women aged 15–49, and children under five (NIPORT and ICF, 2022). This rich dataset serves as a crucial resource for demographic and health research in Bangladesh. For more detailed analysis, two core datasets are utilized: one covering every women aged 15–49 and another focusing on children under five years old. These datasets are merged using shared identifiers such as cluster numbers, household numbers, and respondent line numbers, preserving the hierarchical cluster structure of the data and ensuring proper alignment of variables across different sources.

The BDHS organizes its data into a clustered and stratified design, following three distinct hierarchical levels: clusters, households within clusters, and individuals (represented by respondent line numbers) within households. Each level is uniquely identified to reflect the nested structure of the data. At the highest level, clusters represent geographic areas, followed by households within these clusters, and finally individual respondents. This multi-level structure allows for nuanced analysis across various levels of aggregation, facilitating the exploration of how community, household, and individual characteristics interact to influence outcomes.

The hierarchical nature of the BDHS data offers researchers the ability to investigate complex relationships between variables. By preserving this nested structure, the analysis can account for interactions across multiple levels, such as the influence of community-level factors on individual outcomes. This structure also supports more precise statistical modeling, helping researchers derive a holistic understanding of demographic and health trends in Bangladesh. Furthermore, the clustered and stratified design enhances the representativeness of the sample, accounting for the diverse socio-economic and geographical variations across the country. As a result, the findings from this study can be generalized to the broader population with greater confidence.

2.2 Variables

In merging the datasets for women and children, specific variables were selected to capture both exposures and outcomes related to children's nutritional status. Particular attention was given to identifying determinants of good growth in children, especially their nutritional intake. To measure dietary diversity, respondents (mothers) were asked whether they had given their children a variety of foods in the last 24 hours, including eggs, meat, bread or noodles, potatoes or cassava, pumpkins or carrots, leafy vegetables, mangoes or papayas, other fruits, fish or shellfish, beans or peas, and cheese, yogurt, or other milk products. For this study, only observations with binary responses (0 or 1) for these food variables were included. These binary responses were summed to create a new variable: the Nutritional Variety Score (NVS). This score, ranging from 0 to 11, reflects the number of different food categories provided to children under five. The NVS offers a straightforward measure of dietary diversity, which is crucial for assessing the nutritional intake necessary for healthy child development.

Alongside the food-related variables, several socio-economic and demographic variables related to women were included to explore their influence on children's nutrition. These variables include the woman's working status and number of living children, husband or partner's education, husband or partner's working status, type of residence, wealth index, and women's education. These variables provide a comprehensive view of the socio-economic and educational landscape that can influence children's nutritional outcomes. By incorporating these variables, this approach ensures a nuanced understanding of the complex interplay between socio-economic factors and children's health in Bangladesh.

3 Methodology

The foundational model for this study is based on Suk et al. (2021), which presents an adaptation of the random forest model into a causal forest framework. This causal forest model is specifically designed to incorporate the hierarchical structure of multilevel clustered data, making it particularly suitable for datasets like the BDHS which are inherently clustered. By accounting for this cluster structure, the causal forest model is well-suited for drawing causal inference from observational data.

Causal forests are a machine learning technique developed for causal inference, aiming to estimate treatment effects from observational data (Athey et al., 2019; Wager and Athey, 2018). While traditional random forests (Ho, 1995) are primarily used for predictive purposes, causal forests extend this approach to focus on estimating the impact of a treatment variable on an outcome, controlling for potential confounders (Suk et al., 2021).

3.1 Honest estimation approach

Causal forests use an honest estimation approach to improve the reliability of treatment effect estimates by reducing overfitting. In this approach, the data is divided into two separate parts: one for constructing the tree structure (splitting nodes) and the other for estimating the treatment effects within the leaves of the tree. The separation is random and a varying portion of the data is left out for estimating the treatment effects.

First, a portion of the data is used to build the tree. This involves splitting the data into various nodes based on different characteristics (covariates) so that each node represents a group of similar observations. The goal here is to form clusters of data points that are alike in terms of these characteristics. Once the tree structure is ready, the remaining data is used to estimate the treatment effects within the leaves of the tree. Each leaf now has a bunch of similar observations, and the treatment effect is calculated for these groups. Here the conditional exchangeability or unconfoundedness assumption is used due to the homogeneity of the observations within those leaves. The process ensures that the treatment effect estimation isn't biased by the same data used to create the tree, helping to avoid overfitting and providing a more accurate estimation. Essentially, the data is first organized into a meaningful structure, and then the effects of the treatment are analyzed within this well-organized framework, making the findings more robust and trustworthy.

The honest estimation approach offers several benefits. By separating the data used for tree construction and effect estimation, the approach avoids the overfitting that can occur when the same data is used for both purposes. It ensures that the treatment effect estimates are more reliable and valid because they are not biased by the tree construction process. Additionally, the estimates are more likely to generalize well to new data, making them more useful for real-world causal inference applications.

3.2 Conditional average treatment effect (CATE)

The Conditional Average Treatment Effect (CATE) in causal forests is calculated through a structured process involving several key steps. First, decision trees are grown using a subset of the data. The data is partitioned recursively based on the selected covariates to create homogeneous groups. Within each leaf node of these trees, which is a partition of data, the treatment effect is estimated by comparing the outcomes of treated and control units. These treatment effect estimates are then aggregated across all trees, weighted by their contribution, to derive the CATE for each of the observation.

To calculate the CATE, the process begins by estimating the conditional mean of the outcome, denoted as $\hat{m}^{-i}(x, w)$, and the propensity score, denoted as $\hat{e}^{-i}(x, w)$, for each observation. The estimate, $\hat{m}^{-i}(x, w)$ are obtained using a "honest approach" in the modified random forest algorithm known as causal forest, ensuring that the estimate for each observation does not include that observation's own data, which is known as out-of-bag estimation (OOB). The conditional mean of the outcome, denoted as $\hat{m}^{-i}(x, w)$, mathematically,

$$m^{-i}(x, w) = E[Y \mid X = x, W = w].$$

This represents the expected value of the outcome variable Y given the covariates X and W, when the ith observation is excluded from the estimation process. This can be thought of as predicting the average outcome for a given set of conditions while deliberately leaving out one specific observation to ensure the prediction is not influenced by that observation.

The estimate $\hat{e}^{-i}(x, w)$, also referred to as $P(Z \mid L)$, represents the probability of receiving the treatment Z given a set of confounding variables L, where the ith observation is excluded from the estimation process. This propensity score is calculated using a mixed-effects multilevel logistic model with a random intercept, tailored for data with a three-level hierarchical structure, such as that found in the BDHS. The hierarchical structure accounts for clustering within three distinct levels, allowing for more accurate estimation of the treatment probabilities by considering the nested nature of the data. The following mixed-effects multilevel logistic model is fitted to estimate the propensity score.

$$\log\left(\frac{\pi_1}{\pi_0}\right) = X\beta + b_1 + b_2 + b_3,$$

where the random intercepts, b_1, b_2, b_3, account for the hierarchical levels of clustering: clusters, households and respondent line numbers, respectively.

Next, the residualized outcome, $\tilde{Y}_{ij} = Y_{ij} - \hat{m}^{-i}(X_{ij}, W_j)$ and the residualized treatment, $\tilde{Z}_{ij} = Z_{ij} - \hat{e}^{-i}(X_{ij}, W_j)$ are computed. These residualized values represent the deviations of the actual outcomes and treatments from their expected values. The residuals, $\tilde{Z}_{ij}$ and $\tilde{Y}_{ij}$ play a crucial role in isolating the part of the outcome that is solely due to the treatment effect. By subtracting the expected outcome and treatment values (which are predicted based on the covariates) from the actual values, residuals highlight the variations directly caused by the treatment. This process ensures that the estimation of the treatment effect is more precise, as it focuses on the differences attributable to the treatment itself, rather than being confounded by other variables.

The core of the CATE estimation involves performing a weighted linear regression of the residualized outcome on the residualized treatment. The formula for the CATE at a specific covariate value (x, w) is:

$$\hat{\tau}(x, w) = \frac{\sum_{ij} \alpha_{ij}(x, w)(\tilde{Y}_{ij}\tilde{Z}_{ij})}{\sum_{ij} \alpha_{ij}(x, w)(\tilde{Z}_{ij}^2)},$$

where $\alpha_{ij}(x, w)$ are weights that indicate the contribution of each observation ij to the estimation of the treatment effect at covariate value (x, w). These weights are determined using a recursive partitioning algorithm that divides the data into subsets (nodes) based on covariates. The partitioning aims to maximize the variance of pseudo-outcomes within each subset, forming a binary tree structure where each leaf represents a localized region of the covariate space.

This recursive partitioning is repeated across multiple bootstrap samples to construct a forest of trees. The estimates of the treatment effect from each tree are then aggregated to create an estimate of the CATE for every observation, and these estimates are averaged to produce the final ATE estimate. This ensemble approach helps to stabilize the estimates and reduce variance.

The causal forest approach relies on several key assumptions:

1. Unconfoundedness/Conditional Exchangeability: The treatment assignment is independent of the potential outcomes given the observed covariates:

$$Y(1), Y(0) \perp Z \mid X, W$$

2. Positivity: For all values of the covariates, there is a positive probability of receiving each treatment:

$$0 < Pr(Z = 1 \mid X = x, W = w) < 1$$

3. Stable Unit Treatment Value Assumption (SUTVA): The potential outcomes for any unit do not vary with the treatment assignments of other units, and there are no different versions of the treatment:

$$Y_i = Y_i(Z_i) \quad \text{and it does not depend on } Z_{i'},\ i' \neq i$$

4. Consistency: The observed outcome for each unit is equal to the potential outcome under the treatment that the unit actually received:

$$Y_i = Y_i(Z_i)$$

3.3 Pseudo-outcomes

To calculate the pseudo-outcomes for each observation, the algorithm adjusts the outcome Y_{ij} and the treatment Z_{ij} for their respective means and propensity scores. The pseudo-outcome q_{ij} is typically calculated as

$$q_{ij} = \frac{\tilde{Z}_{ij} \cdot \tilde{Y}_{ij}}{\hat{e}^{-i}(X_{ij}, W_j) \cdot (1 - \hat{e}^{-i}(X_{ij}, W_j))},$$

where $e^{-i}(X_{ij}, W_j)$ represents the out-of-bag estimate of the propensity score. The next step involves finding the optimal splits in the data which are called nodes. The algorithm searches for ways to divide the data so that the variance of the pseudo-outcomes within the nodes are minimized, while the variance between the nodes is maximized. This process helps in identifying splits that best capture differences in treatment effects. Maximizing the variance of pseudo-outcomes between nodes, or equivalently minimizing it within nodes, ensures that each split made by the tree is informative about the treatment effect. By focusing on splits that create subsets with different pseudo-outcomes, the algorithm effectively identifies regions of the covariate space where the treatment effect varies. This leads to a more accurate and nuanced estimate of the CATE.

3.4 Average treatment effect (ATE)

The Average Treatment Effect (ATE) is defined as the average linear contrast between the two potential outcomes $Y_{ij}(1)$ and $Y_{ij}(0)$, where $Y_{ij}(1)$ is the outcome if an individual ij receives the treatment, and $Y_{ij}(0)$ is the outcome if the same individual does not receive the treatment. ATE is calculated by taking the average of the CATE estimates taken from the large number of bootstraps made from the dataset. This averaging process aggregates the individual-specific treatment effects to provide an overall measure of the treatment effect for the population. The ATE can also be considered to be a measure of risk difference:

$$\hat{\text{ATE}} = E(\hat{\tau}(X_i, W_j)).$$

3.5 Propensity scores in causal forest

The introduction of propensity scores into causal forests significantly enhances the accuracy and reliability of treatment effect estimates. Propensity scores represent the probability of receiving a treatment given a set of covariates, and their integration helps create balanced groups. This balance ensures that comparisons between treated and control groups are fair and not skewed by underlying differences in covariates, thereby reducing bias due to confounding variables. One practical benefit of using propensity scores in causal forests is evident during the tree-building process. By incorporating these scores, the splits in decision trees are guided to create more homogeneous groups concerning treatment propensity. This leads to more accurate treatment effect estimates within each leaf of the tree.

The honest estimation approach in causal forests further benefits from the use of propensity scores. By separating the data used for constructing tree splits from the data used for estimating treatment effects, the influence of covariate imbalances is minimized. This separation results in less biased and more reliable estimates of treatment effects. Overall, the integration of propensity scores into causal forests makes the method more robust and reliable for observational studies.

3.6 Accounting for cluster effects

In a multi-stage stratified cluster sampling design, the population is divided into distinct strata based on characteristics like region and within each stratum, smaller groups (clusters) are randomly selected for data collection. This existence of clustering requires careful consideration of what the statistical approach should be, as it is known that the observation within a cluster are correlated. Mixed-effects multilevel logistic regression for propensity score, with random intercepts for clusters and including the clusters into the causal forest model help incorporate the cluster structure of the complex survey data in the proposed statistical approach. During recursive partitioning, cluster labels guide the splitting process, ensuring that the model accounts for variations across clusters. This improves the interpretability of CATE estimates, revealing how treatment effects vary across different contexts. Adjusting for cluster effects also enhances the generalizability and robustness of policy recommendations. Failing to account for cluster effects can introduce bias, as treatment effects may be incorrectly attributed to differences across clusters or due to the correlation within the clusters. Thus, including these effects ensures that the estimated treatment impact accurately reflects the true causal relationship.

4 Analysis and Results

4.1 Exploratory analysis

This study offers an overview of the variables using exploratory data analysis methods. Frequency and percentage distributions are provided for categorical variables, and for one continuous variable, the median and interquartile range are reported. Table 1 presents the list of food variables along with their frequency distributions. It shows that certain food items are less frequently given to children, such as cheese, with only 4.6% of children receiving it, and meat, consumed by 12% of children. In contrast, 61% of children had bread, noodles, or other grain-based foods, suggesting a reliance on grains in children's diets.

Table 2 provides the distribution of the Nutritional Variety Score (NVS) for children under 6 months and those aged 6 months to 5 years. According to the World Health Organization, children under six months old should be exclusively breastfed, without any other foods, (World Health Organization and UNICEF, 2003). The data reveals that 87.7% of children under 6 months did not consume any of the specified food items in the past 24 hours. To avoid potential bias from this zero inflation, the analysis will focus exclusively on children, aged 6 months to 5 years. This subgroup includes 3,206 children, with a median age of 27 months. Notably, 15.63% of children in this age range reported no intake of the listed food items in the past 24 hours, potentially reflecting issues such as poverty, food insecurity, or restricted access to diverse dietary options.

Table 3 summarizes the distribution of maternal employment status and other confounding variables among children aged 6 months to 5 years. In this table, it is seen that only 26% of mothers

Table 1: Food variables given to children aged 6 months to 5 years

Serial No.	Food Category	No (Percent)	Yes (Percent)
1	Eggs	2,205 (67%)	1,076 (33%)
2	Bread, noodles, other made from grains	761 (23%)	2,524 (77%)
3	Potatoes, cassava, or other tubers	1,591 (48%)	1,689 (51%)
4	Meat (beef, pork, lamb, chicken, etc)	2,764 (84%)	515 (16%)
5	Pumpkin, carrots, squash (yellow or orange inside)	2,954 (90%)	323 (9.8%)
6	Any dark green leafy vegetables	2,435 (74%)	844 (26%)
7	Mangoes, papayas, other vitamin A fruits	2,928 (89%)	349 (11%)
8	Any other fruits	2,449 (75%)	824 (25%)
9	Fish or shellfish	2,243 (68%)	1,035 (31%)
10	Food made from beans, peas, lentils	2,403 (73%)	876 (27%)
11	Hard or Soft Cheese	3,102 (94%)	173 (5.3%)

Table 2: Distribution of Nutritional Variety Score (NVS) for Children Under 6 Months and Those Aged 6 Months to 5 Years

NVS	Children Under 6 Months	Children Aged 6 Months to 5 Years
0	841 (87.70%)	501 (15.63%)
1	61 (6.36%)	306 (9.54%)
2	24 (2.50%)	471 (14.69%)
3	16 (1.67%)	557 (17.37%)
4	6 (0.63%)	526 (16.41%)
5	5 (0.52%)	415 (12.94%)
6	1 (0.10%)	229 (7.14%)
7	1 (0.10%)	119 (3.71%)
8	2 (0.21%)	60 (1.87%)
9	0 (0.00%)	16 (0.50%)
10	1 (0.10%)	6 (0.18%)
11	1 (0.10%)	0 (0.00%)
Total (n)	959	3206

are employed, and the majority of families reside in rural areas (67%). Household wealth is split with 44% of families classified as poor, 19% as middle income, and 37% as rich. Additionally, 34% of husbands/partners have secondary education, while a substantial 98% are employed, potentially highlighting an economic dependency on male income in these households.

Table 3: Distribution of Maternal Employment Status and Other Potential Confounding Variables Among Children Aged 6 Months to 5 Years

Characteristic	Frequency (Percent)
Working Status	2,032 (26%)
Number of Living Children	2.00 (1.00, 3.00)[1]
Husband/Partner Education	
Higher	1,448 (19%)
No Education	1,226 (16%)
Primary	2,409 (31%)
Secondary	2,570 (34%)
Husband/Partner Working Status	
Not Working	138 (1.8%)
Working	7,521 (98%)
Residence Type	
Rural	5,220 (67%)
Urban	2,564 (33%)
Wealth Index	
Middle	1,512 (19%)
Poor	3,429 (44%)
Rich	2,843 (37%)
Education	
Higher	1,362 (17%)
No Education	514 (6.6%)
Primary	1,887 (24%)
Secondary	4,021 (52%)
Total	7,784 (100%)

[1] The number of living children is presented as a median with the interquartile range.

4.2 Results from causal forest modelling

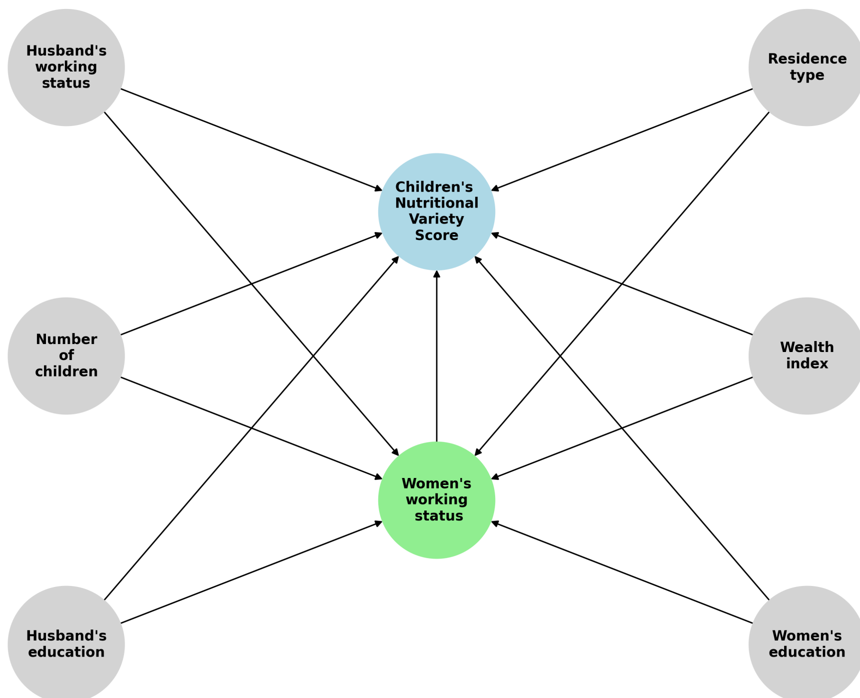

Figure 1: Directed Acyclic Graph (DAG) illustrating factors influencing children's Nutritional Variety Score

In this study, the outcome variable is the children's Nutritional Variety Score (NVS), and the treatment variable is the women's working status. The analysis focuses on controlling for several potential confounding variables that are not directly in the causal pathway of interest but have an impact on both the treatment and the outcome variables. Figure 1 shows the Directed Acyclic Graph (DAG), illustrating the confounding variables that influence the NVS, including the number of children, husband's education, husband's employment status, residence type, wealth index, and women's education level. While these variables are not the focus of the causal analysis, their influence on both the treatment and the outcome makes it important to account for them to avoid biased estimates of the effect of women's working status on children's nutritional outcomes. By doing so, the study aims to isolate the true causal effect of the treatment variable while minimizing the impact of confounding influences.

The analysis begins by constructing the outcome variable, the NVS, which is computed by summing 11 binary indicators representing whether a child consumed specific food items in the last 24 hours or not. A generalized linear mixed-effects model (GLMM) is then employed to estimate the propensity scores for women's employment status. This model incorporates random effects to capture the hierarchical structure inherent in the data, which is nested by geographical region (cluster), household, and individual respondent (identified by respondent line number).

A causal forest model is subsequently applied to estimate the Average Treatment Effect (ATE) of women's employment status on children's NVS. The ATE is estimated by taking an average of the CATE across all the observation. Table 4 shows the heterogeneity of the CATE across the observation. The covariates used in the causal forest model align with those employed in the propensity score estimation model. The causal forest model accommodates clustering by integrating cluster IDs. The causal forest has all of the tunable parameters optimized by cross-validation and Table 5 shows the estimated ATE, its standard error and the 95% confidence interval.

Following the approach of Venkatasubramaniam et al. (2023), Table 6 presents a robustness check of the causal forest model. This analysis examines the stability of the estimated ATE and its associated 95% confidence interval by varying the number of trees in the causal forest. The results demonstrate that the ATE estimates remain largely stable despite changes in the number of trees, with only minor fluctuations observed.

For this study, the missing data has been addressed by conducting a complete case analysis (CCA), assuming that the missing data mechanism follows a Missing Completely at Random (MCAR) pattern. All analysis were conducted using the `R` programming language (R Core Team, 2024). Key packages included `grf`, which was used to implement the causal forest model, and `lme4`, which was employed to fit the mixed-effects multilevel logistic regression model for propensity score estimation (Tibshirani et al., 2024; Bates et al., 2015).

Table 4: The frequency distribution of Conditional Average Treatment Effect (CATE) across the observations

CATE Range	**Frequency**
$[0.328, 0.375]$	54
$(0.375, 0.422]$	342
$(0.422, 0.469]$	411
$(0.469, 0.516]$	535
$(0.516, 0.563]$	506
$(0.563, 0.61]$	297
$(0.61, 0.657]$	151
$(0.657, 0.704]$	423
$(0.704, 0.751]$	410
$(0.751, 0.798]$	41

Table 5: Estimated Average Treatment Effect (ATE) on children's Nutritional Variety Score (NVS)

Variable	ATE	Standard Error	95% CI
Women's Employment	0.532	0.085	0.365 - 0.699

Table 6: Robustness checks: Estimated average treeatment effect with the change of the number of trees in the causal forest algorithm.

Number of Trees	ATE	95% Confidence Interval
10	0.507	[0.338, 0.676]
100	0.500	[0.331, 0.669]
1000	0.495	[0.326, 0.664]
2000	0.493	[0.324, 0.662]
5000	0.499	[0.330, 0.668]

4.3 Hypothesis test

The following hypothesis investigates whether a causal relationship, represented by the Average Treatment Effect (ATE), exists between women's employment status and their children's Nutritional Variety Score (NVS).

$$H_0 : ATE = 0 \quad \text{(Causal effect does not exist)}$$
$$H_1 : ATE \neq 0 \quad \text{(Causal effect does exist)}$$

The estimate of the ATE and its standard error was derived using the bootstrap method. These estimates can also be used to construct confidence intervals. Under the assumption that the ATE follows a normal distribution, which is reasonable given that the ATE is aggregated from Conditional Average Treatment Effects (CATE) across numerous trees in the causal forest. The application of the central limit theorem ensures that with a sufficiently large sample size, the sampling distribution of the ATE converges to normality. Thus, the 95% confidence interval for the ATE is [0.365, 0.699], indicating a significant causal effect of women's employment status on the nutritional variety score of children.

5 Discussion and Conclusion

The relationship between women's employment and children's nutritional outcomes is complex, but this study provides valuable insights into the impact of mothers' working status on their children's dietary diversity in Bangladesh. By employing advanced statistical techniques, such as the causal forest model with cluster identifiers and a mixed-effects multilevel logistic model for propensity score, this study has been able to rigorously analyze this relationship and offer robust conclusions.

The findings indicate that children, aged 6 months to 5 years, of employed mothers have a slightly higher Nutritional Variety Score (NVS) than those of non-employed mothers, with an estimated Average Treatment Effect (ATE) of 0.532 and a standard error of 0.085. This suggests that working mothers may have greater access to financial resources or information, enabling them to provide a more diverse diet for their children. However, the modest effect size implies that additional socio-economic factors also play a crucial role in shaping children's nutritional outcomes.

These findings carry important policy implications. They highlight the need for greater employment opportunity for women such that they can ensure greater nutritional diversity and nutritional intake for their children. They also highlight the need for supportive interventions such as access to affordable childcare, flexible work environments, and nutrition education programs. Such initiatives can simultaneously enhance women's economic participation and promote better nutrition for children. This study also illustrates the value of integrating causal inference methods with comprehensive survey data, offering a methodological framework for future research in similar settings. By addressing confounding factors more effectively, these methods yield clearer insights into the causal link between maternal employment and child nutrition.

While this study offers important insights, several limitations should be acknowledged. First, the cross-sectional nature of the Bangladesh Demographic and Health Survey (BDHS) data limits our ability to establish definitive causality. Although advanced techniques such as causal forest models and cluster-specific propensity scores help infer causal relationships, the possibility of unmeasured confounders remains. Second, the Nutritional Variety Score, although informative, is based on binary indicators that capture only the presence or absence of food items, without accounting for the quantity or quality of food consumed. Future research could incorporate these dimensions to provide a more holistic understanding of dietary diversity. Furthermore, the generalizability of these findings may be limited. While the results are relevant to Bangladesh, they may not directly apply to other regions with different socio-economic and cultural contexts. Comparative studies across multiple countries would help determine whether similar patterns exist elsewhere. Third, the reliance on self-reported data introduces potential reporting biases, as variables such as women's employment status and children's dietary intake may not be accurately reported. Future studies could benefit from validating self-reports with objective measures. Additionally, this study focuses on short-term outcomes; longitudinal research would be beneficial to explore how the relationship between women's employment and child nutrition evolves over time, capturing any delayed effects.

In conclusion, although women's employment positively affects their children's nutritional variety, this is only part of a broader narrative. Encouraging maternal employment alone may not be sufficient to significantly improve child nutrition. A more holistic policy approach, addressing various aspects of the family's socio-economic context, is essential to achieve lasting improvements. Future public health interventions must be designed with a comprehensive understanding of the multifaceted realities faced by families. This study offers a foundation for such efforts, underscoring the importance of integrated policies that support both women's workforce participation and children's nutritional health.

References

Arimond, M. and Ruel, M. T. (2002), “Progress in developing an infant and a child feeding index: an example using the Ethiopia Demographic and Health Survey 2000,” Tech. rep.

— (2004), “Dietary diversity is associated with child nutritional status: evidence from 11 demographic and health surveys,” *The Journal of nutrition*, 134, 2579–2585.

Athey, S., Tibshirani, J., and Wager, S. (2019), “Generalized random forests,” *The Annals of Statistics*, 47.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015), “Fitting Linear Mixed-Effects Models Using lme4,” *Journal of Statistical Software*, 67, 1–48.

Bernal Rivas, J. and Lorenzana Albert, P. (2003), “Dietary diversity and associated factors among beneficiaries of 77 child care centers: Central Region, Venezuela,” *Archivos latinoamericanos de nutricion*, 53, 52–58.

Brand, J. E., Zhou, X., and Xie, Y. (2023), “Recent developments in causal inference and machine learning,” *Annual Review of Sociology*, 49, 81–110.

Cui, P., Shen, Z., Li, S., Yao, L., Li, Y., Chu, Z., and Gao, J. (2020), “Causal inference meets machine learning,” in *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 3527–3528.

Cunningham, K., Ruel, M., Ferguson, E., and Uauy, R. (2015), “Women’s empowerment and child nutritional status in South A sia: a synthesis of the literature,” *Maternal & child nutrition*, 11, 1–19.

Galiè, A., Teufel, N., Girard, A. W., Baltenweck, I., Dominguez-Salas, P., Price, M. J., Jones, R., Lukuyu, B., Korir, L., Raskind, I., et al. (2019), “Women’s empowerment, food security and nutrition of pastoral communities in Tanzania,” *Global Food Security*, 23, 125–134.

Ho, T. K. (1995), “Random decision forests,” *Proceedings of the Third International Conference on Document Analysis and Recognition*, 1, 278–282.

Hong, R., Banta, J. E., and Betancourt, J. A. (2006), “Relationship between household wealth inequality and chronic childhood under-nutrition in Bangladesh,” *International journal for equity in health*, 5, 1–10.

Malapit, H. J., Kadiyala, S., Quisumbing, A. R., Cunningham, K., and Tyagi, P. (2013), “Women’s empowerment in agriculture, production diversity, and nutrition: Evidence from Nepal,” .

Mandal, K. C. (2013), “Concept and Types of Women Empowerment.” in *International Forum of Teaching & Studies*, vol. 9.

NIPORT and ICF (2022), “Bangladesh Demographic and Health Survey,” Tech. rep., National Institute of Population Research and Training, Dhaka, Bangladesh.

Onah, M. N. (2021), "Women's empowerment and child nutrition in South-Central Asia; how important is socioeconomic status?" *SSM-Population Health*, 13, 100718.

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

Schuler, S. R. and Rottach, E. (2010), "Women's empowerment across generations in Bangladesh," *The journal of development studies*, 46, 379–396.

Shroff, M., Griffiths, P., Adair, L., Suchindran, C., and Bentley, M. (2009), "Maternal autonomy is inversely related to child stunting in Andhra Pradesh, India," *Maternal & child nutrition*, 5, 64–74.

Soharwardi, M. A. and Ahmad, T. I. (2020), "Dimensions and determinants of women empowerment in developing countries," *International Journal of Sustainable Development and Planning*, 15, 957–964.

Suk, Y., Kang, H., and Kim, J.-S. (2021), "Random forests approach for causal inference with clustered observational data," *Multivariate Behavioral Research*, 56, 829–852.

Tibshirani, J., Athey, S., Sverdrup, E., and Wager, S. (2024), *grf: Generalized Random Forests*, r package version 2.3.2.

Venkatasubramaniam, A., Mateen, B. A., Shields, B. M., Hattersley, A. T., Jones, A. G., Vollmer, S. J., and Dennis, J. M. (2023), "Comparison of causal forest and regression-based approaches to evaluate treatment effect heterogeneity: an application for type 2 diabetes precision medicine," *BMC Medical Informatics and Decision Making*, 23, 110.

Wager, S. and Athey, S. (2018), "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests," *Journal of the American Statistical Association*, 113, 1228–1242.

World Health Organization and UNICEF (2003), *Global strategy for infant and young child feeding*, Geneva: World Health Organization.

Received: January 8, 2025

Accepted: March 10, 2025