A RCT for nutritional Intervention in Bangladeshi Children

Howard Hu

August 13, 2022

1 Exclusive Summary

Dr. Jennie Z. Ma of the public health sciences department provided the data for this report. We use subgroup analysis to find which children improve under the nutritional intervention. First, we categorize children into subgroups such as highly educated parents and families with high incomes. After doing multiple t-tests, we did not find any statistically significant subgroups, but the p-value of the "dadedu+momedu==1" is 0.132, which is the closest to our significant level of 0.05. Then we use the *Pysubgroup* package and build a logistics regression model. However, we did not find any statically significant variables in the logistics regression model, but we found out the p-value of septic toilet is 0.087, which is close to the significant level of 0.05. It implies that different toilets may affect whether children improve under our nutritional intervention. We find out that children with a highly educated dad and joint family have the highest probability of gaining benefits from our nutritional intervention. We also implement a reverse solution. We use PCA and K-means to classify our treatment group. Based on our K-means model, two variables that may create a subgroup that improves under our nutritional intervention are dad's educational level and toilet-related variables.

2 Introduction

This report aims to do a subgroup analysis on the data from Dr. Jennie Z. Ma of the department of public health sciences. Firstly, we separate children into different subgroups such as highly educated parents and high expenditure families. Then we would like to do multiple t-tests to check whether there is a statistical significance of the mean difference of those subgroups' biomarkers. Secondly, we use the " Pysubgroup " package to find which subgroup has a higher probability of being determined as an improvement in a logistical regression model. Thirdly, we solve the problem in a reverse way. We use PCA and do the K-means classifier. Then we are trying to find any common characteristics in those classified groups.

2.1 General Background

Malnutrition and stunting have become severe problems, especially in developing countries. According to the background knowledge given by Dr. Jennie Z. Ma, by the age of 15 months, one-third of children are undernourished, and the number of stunted and malnourished children under the age of 5 is about 159 million in those developing countries. Research by Prendergast and Humphrey supports "With an estimated 165 million children below five years of age affected, stunting has been identified as a major public health priority, and there are ambitious targets to reduce the prevalence of stunting by 40% between 2010 and 2025" [PH14]. Therefore, public health researchers need to develop a helpful nutritional intervention to help those children.

We have three different data sets for this report. Firstly, researchers collect the baseline information of 200 children. In the baseline data set, the age of those children is recorded as days. Secondly, for each child, researchers track their individual health information such as Height for age, weight for age, and weight for Height in five visits. Those three variables are recorded as Z scores in the data set. Thirdly, researchers recorded each child's four bio-markers (MPO, Reg1b, sCD14, CRP) before and after nutritional intervention. The target of the nutritional intervention is trying to reduce those four bio-markers factors of children's blood.

2.2 Objectives

According to the discussion with our client, we know the outcome measures for the two arms don't differ significantly. Then we have the following questions. If treatment effects and clinical factors are heterogeneous, are there subsets of children who benefit from nutritional supplements? To develop targeted interventions, how would we identify those children? In conclusion, we would like to analyze our data subgroup to determine which subset of children benefited from the PTM nutritional intervention.

2.3 Exploratory Data Analysis

We receive our data from Dr. Jennie Z. Ma of the department of public health sciences. Firstly, we would like to EDA on PTM baseline. In the PTM Baseline data set, we have 200×19 samples, including 100 placebo samples and 100 treatment samples. Researchers record the total monthly income in taka and the total monthly expenditure in taka for each sample. We can create the following figures.

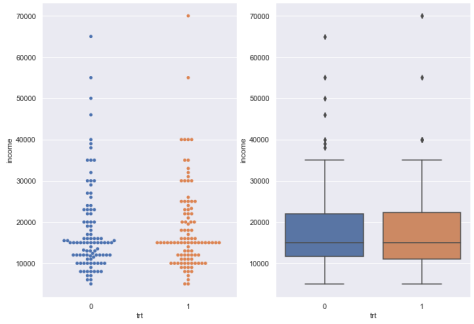

Figure 1: Total monthly income

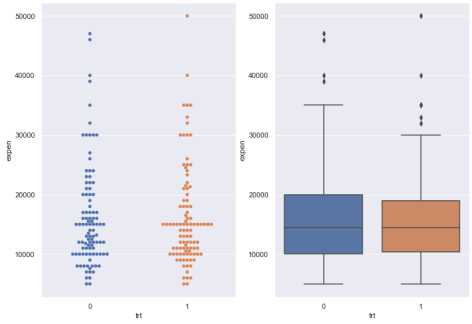

Figure 2: Monthly total expenditure

Figure 1 and figure 2 show the median monthly income and expense are around the same for the placebo and treatment groups. As we know, total monthly income and total monthly expenditure are related to socioeconomic status. People who have high income and high expenditure will have high socioeconomic status. Children from families with high socioeconomic status may have less influence under nutritional intervention because they already eat enough food. For analyzing data related to socioeconomic status, we can find the median income is 15000.0 and the median of expense is 15000.0. Then we can separate people into two different groups. We find out there are 117 children whose families have higher than the median income and 97 children whose families have a higher expense. In the approach section, we can use a two-sample t-test to check whether the mean difference of those groups' average bio-markers is the same or not.

Instead of finding the potential subgroup using socioeconomic status, we can consider parents' education level. Intuitively, parents who do not experience high education may ignore the importance of nutrition on children's foods. Therefore, their children may have malnutrition and stunting problems, and nutritional intervention may become a valuable method to improve their health. It means we may separate 200 children into two groups, highly-educated parents(if both parents have high education levels) and not high educated parents. We have 112 children with highly educated parents and 88 children who have not high educated parents.

Secondly, we focus on analyzing the PTM Anthro long data set. This data set includes children's health variables of five visit times. We can build a point graph to see general trends of HAZ, WAZ, and WHZ.

Since the time gap between visit time 1 (screening/ enrollment) and visit time 2 (start of intervention) is close, we can average those two days' data in the later approach section. From Figures 3,4, and 5, we have a big difference between visit time 4 (Follow up) and visit time 2. If we can find a subgroup with a huge difference in HAZ, we can prove this subgroup has a beneficial effect under a nutritional intervention. The detail of hypothesizing test are shown in the approach section.

Thirdly, we focus on the PTM lab data set. Observation 6038, 6103, 6155, 6196, 6200, 6202, 6263, 6294, and 6319 contain missing data on 1 month after intervention. Therefore, we would like to drop those samples. Then we have 372×7 samples in the PTM Lab data set. According to the meeting with our client, we know if children have lower biomarkers one month after the intervention.

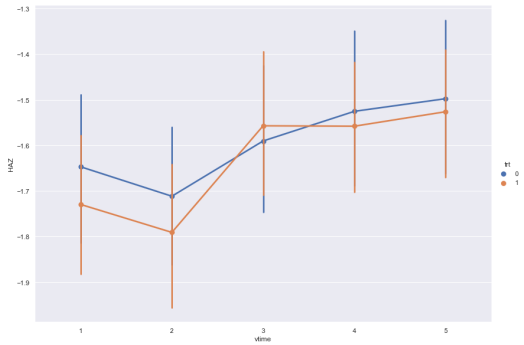

Figure 3: The mean HAZ of 200 children in 5 visit time

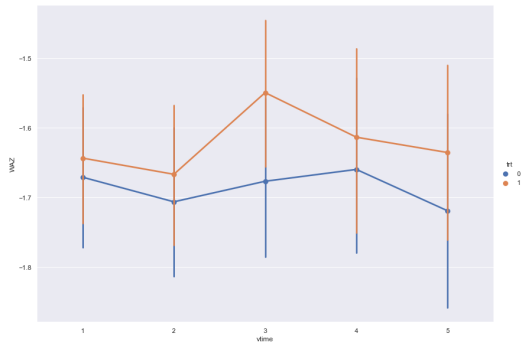

Figure 4: The mean WAZ of 200 children in 5 visit time

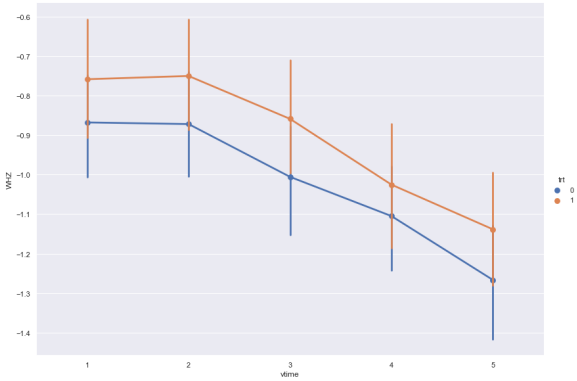

Figure 5: The mean WHZ of 200 children in 5 visit time

They can be considered as gaining improvement by this intervention. Therefore, we can calculate the difference between before and after the intervention. We can also create a new variable, "Improve," a binary variable that indicates which child gets improvement. Then we can standardize and average four biomarkers to get a new variable called "Ave." This variable shows the general trend of those biomarkers. After merging the PTM lab data set with the PTM baseline data set, we have 191×25 samples.

3 Approach

Firstly, we would like to do a two-sample T-test to check if the mean difference of averaging biomarkers between the placebo group and treatment group is the same or not. We build the following hypothesizing testing.

Null hypothesis H(0): the mean difference of averaging biomarkers between the placebo subgroup and treatment subgroup is the same. (It means those subgroup's children do not get improvement on this intervention)

Alternative hypothesis H(1): the mean difference of averaging biomarkers between the placebo and treatment groups is different. (It means the intervention works on this subgroup's children)

From the previous table, we did not find any of those subgroups is statistically significant. However, the lowest p-value is 0.132, which indicates that when only one of the parents has a high education level, our nutritional intervention may have more influence on their children than other subgroups.

For checking whether the distribution of two groups of female children's average biomarkers is similar or not, we plot this box plot.

Subgroup	P-value
Sex=2	0.935
Sex=1	0.475
momedu=1	0.438
momedu=0	0.939
dadedu=0	0.289
dadedu=1	0.831
dadedu+momedu $>$ 0	0.381
dadedu+momedu $>$ 1	0.978
dadedu+momedu $=$ 1	0.132
hhclass $=$ 1	0.266
hhclass $=$ 2	0.146
hhclass $=$ 3	0.663
hhclass $=$ 4	0.872
treatedwater $=$ 1	0.649
expen ¿ medianexpense	0.503
expen ¡ medianexpense	0.841

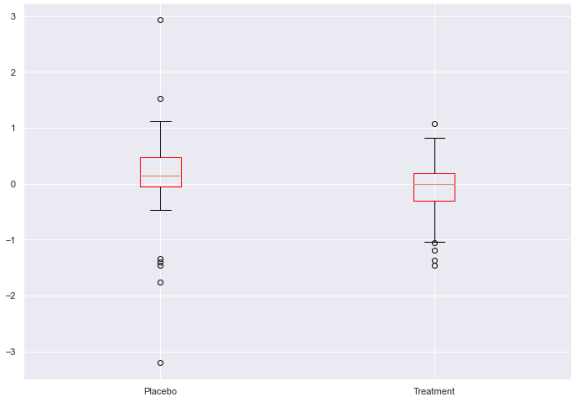

Figure 6: The average biomarkers for female children

This figure proves those two groups have a similar distribution, and the p-value should be very large and insignificant. Secondly, we would like to build logistic regression and use Z-test to see which baseline factor may influence their biomarkers' difference. As mentioned in the EDA section, we set the "Improve" variable as our response variable in our regression model. Then we create the following regression model.

$$Improve = Intercept + sex + aged + member + children + hhlive + sleep + familytp + ownhh + income + expen + hhclass + momedu + dadedu + septictoilet + treatedwater + toiletshare + opendrain$$

where: $Improve = 0$:No improvement 1:have improvement

From the previous table, we did not find any statistically significant variable. However, the smallest p-value, 0.087, is the septic toilet. It implies different kinds of toilets may affect whether children have improved or not under our nutritional intervention.

Thirdly, we would like to use the "Pysubgroup" package from the Python community to determine which potential subgroup may have successful improvement.

We know dad's education has become a popular factor in the five subgroups from the previous table. These results consist of the table given by a two-sample T-test, which indicates that only children who belong to one highly educated parent will have a more considerable influence on our nutritional intervention. Therefore, we would like an extra two-sample t-test with an only dad who

variable	coef	P value
Intercept	0.5166	0.832
sex	0.1496	0.635
aged	0.0029	0.735
member	-0.3282	0.531
children	0.1212	0.530
hhlive	-0.0227	0.467
sleep	0.2652	0.608
familytp	-0.0771	0.869
ownhh	-0.6003	0.129
income	6.748e−6	0.883
expen	2.904e−5	0.617
hhclass	-0.0632	0.731
momedu	-0.1133	0.775
dadedu	-0.4775	0.198
septictoilet	-1.4415	0.087
treatedwater	0.4382	0.237
toiletshare	0.3723	0.340
dadedu	0.3723	0.198
opendrain	0.3099	0.332

quality	subgroup
0.0436117430991475	dadedu==1 AND familytp==2
0.0413914092267208	dadedu==1 AND toiletshare==0
0.0398289520572353	dadedu==1 AND income¿=25000
0.0358542803102985	dadedu==1 AND hhclass==3
0.0353608727830926	dadedu==1 AND ownhh==2

has high education. However, the p-value of this t-test is 0.659, which shows those two groups still have a similar mean.

Fourthly, we would like to solve this problem in a reversed way. We firstly use PCA to get the first two principal components. Then we use a classifier to classify those samples. Then we are trying to find the common characteristics of those classified groups.

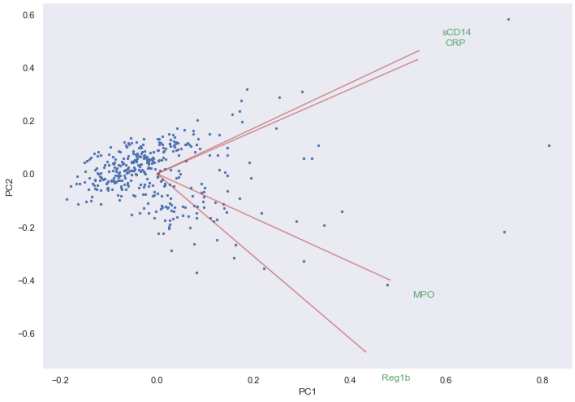

Figure 7: PCA for biomarkers

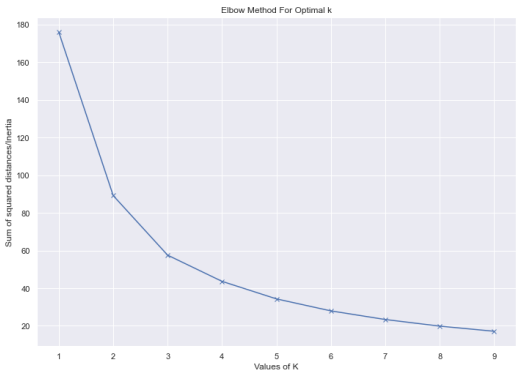

Figure 8: Elbow method for optimal k

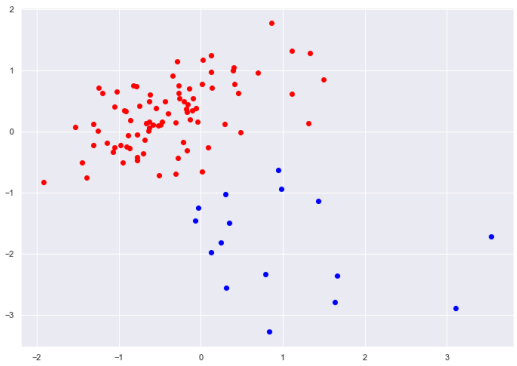

Figure 9: K means result

Figure 8 shows no elbow point for our PCA, but the most significant jump from 1 to 2. Therefore, we choose two as our k number for K-means. Moreover, we only do K-means on treatment groups because our client wants to know which treatment subgroup benefits.

For the blue dot in figure 9, we have 16 samples. 94% of samples' septic toilet is 0, and 68.8% of samples' dad education is 1. Moreover, 94% of blue dot children have lower biomarkers. Therefore, we would like to say dad's education level and toilet-related variables might be two potential factors that can create a subgroup that improves under our nutritional intervention.

4 Conclusion

The data for this report was provided by Dr. Jennie Z. Ma of the public health sciences department. We used subgroup analysis to determine which children improved with the nutritional intervention. Firstly, we classify children into subgroups such as highly educated parents and families with high incomes. After doing multiple t-tests, there are no statistically significant subgroups, but the p-value of the "dadedu+momedu==1" is 0.132, which is the closest to our significant level of 0.05. Then we use the *Pysubgroup* package and build a logistics regression model. However, we did not find any statistically significant variables in the logistics regression model, but we found out the p-value of septic toilet is 0.087, which is close to the significant level of 0.05. It shows that different toilets may affect whether children improve under our nutritional intervention. We find out that children with a highly educated dad and joint family have the highest probability of gaining benefits from our nutritional intervention. We also solve the problem reversely. We use PCA and K-means to classify our treatment group. Based on our K-means model, two most possible variables that may create a subgroup are dad's educational level and toilet-related variables.

5 Appendix

References

[PH14] Andrew J. Prendergast and Jean H. Humphrey. The stunting syndrome in developing countries. *Paediatrics and International Child Health*, 34(4):250–265, 2014.