Predicting Weight to Height Z-Score in the Global South

Sam Loescher and Camiel Schroeder

Introduction

Humans have struggled with malnutrition since before the dawn of civilization. The root causes of the problem have shifted, however. The issue is no longer our limited ability to grow food. Humanity now produces more than enough food to adequately feed every living human, yet the problem of malnutrition persists. Rather than true scarcity, economic inequality and poor redistribution of resources are now the key drivers of malnutrition. One way to remedy this resource allocation inequality is through direct cash transfers (DCT). Researchers in _____ province, India researched the effects of DCTs to new parents on numerous health outcomes, including the weight to height z-score (WHZ) of children.

The effects of DCTs, while important, are well documented. Giving cash to people in poverty improves their health outcomes. The researchers' data set is extensive, however, which opens the possibility to further statistical analysis beyond the temporal effects of DCTs. Rather than observing change over time, we used their data to examine the relationship between numerous variables such as parent's education levels, empowerment levels, and distance from nearest market with their children's HWZ. This allowed us to observe which of these variables are strong predictors of HWZ; these associations in turn indicate which variables the global health community should target when seeking to improve young children's WHZ in poor communities.

Statistical procedures used

In this study, the response variable is the height to weight z-score of children in _____, India, and the explanatory variables are mother's education level, father's education level, empowerment index, and distance from market. These variables were all measured on the 3,142 families in our chosen subset of the study. These families represent the observational unit in this study. To address the research question, we implemented a K-nearest neighbors test, formalized mathematically as follows:

What this means in plain English is that the algorithm calculates how different all of the points in the data set are from one unknown point, and then takes the mean of some number (K) of the most similar points. For instance, if we wanted to predict the WHZ for a child living twelve km from the market with parents who both had 10 years of education, we would find the K most similar ("nearest") points, and take their mean WHZ to create out prediction.

One key assumption in any KNN algorithm is that there is enough data for the number of dimensions (explanatory variables). If there are too many dimensions for the amount of data provided, then there is a chance that there will not be many near neighbors to the target point. This will make the predictions much less accurate, and often necessitates removing dimensions from the analysis.

There are many ways of calculating how similar points are to one another (the "distance"). For this report, we used the Gower Distance. Gower distance standardizes the variable scale among numeric variables, and is able to account for categorical variables in its distance measurement. This makes it perfect for the complexity of data set we are working with, with many types of data and different scales involved.

After exploring many variables, we chose mother's education level (measured in years in school), happiness level (a self-prescribed score from 1-5), and ration card status (an effective measure of overall socioeconomic standing) as our variables because they represented key indicators of different aspect of life which logically seem like they could explain large parts of why some children are undernourished while others are not.

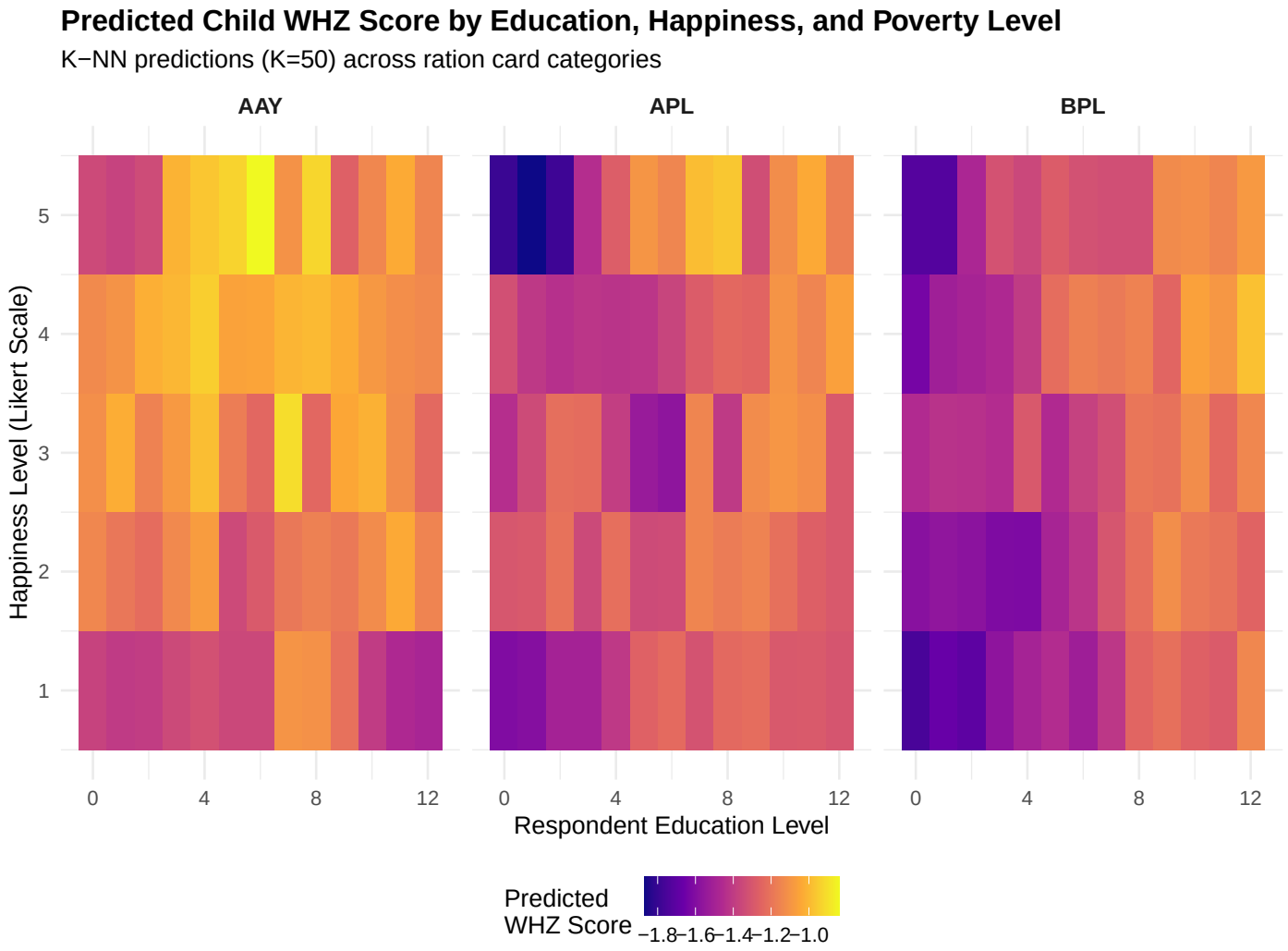

Figure caption.

Summarize whatever graph we put here

Summary of Statistical Findings

In order to determine the K-value for our algorithm, we calculated the mean squared error (MSE) for 9 k-values ranging from 3 to 100 using a subset of our data set. The prediction error (MSE) decreased as K increased, stabilizing around K=50-100. K=50 was selected as optimal, achieving an MSE of 1.03 while maintaining local prediction characteristics. This represents using approximately 2.4% of the

training data for each prediction, balancing bias-variance tradeoff effectively. The error for each K that was tested are available in the appendix.

Scope of Inference

The KNN algorithm's optimal MSE of 1.03 is higher than a naive estimator of WHZ: the variance. This means that using these three predictors to find points similar to the target point is worse than simply predicting the mean value for every data point, regardless of the participant's education level, happiness index, or ration card status. Thus we conclude that the KNN algorithm using these predictors should not be used to make predictions about the WHZ of children, nor should any correlation between the variables and malnutrition be accepted on the basis of this analysis.

Since a random mechanism was used to determine whether clouds were seeded, we can conclude that the differences in rainfall that we observed were caused by the seeding. However, there is no reason to believe that the days on which clouds were seeded were randomly selected over the course of the four year experiment. Therefore, these results apply to only the 52 days deemed suitable to cloud seed, and only in the area in which the study was conducted.

Acknowledgements and Author Contributions

Sam and Camiel worked on this report together. Sam capitalized on the rain trapping Professor Stratton in Monroe to have a very length discussion about the KNN algorithm in office hours. Sam and Camiel both used ChatGPT for debugging, and for conceptual understanding of in-class coding examples.

Appendix

Code

```
# setwd("C:/Users/Owner/STAT_LEARNING_REPORT1/report1/cash_transfers_dataset")

# ========================================
# LOAD AND PREPARE DATA
# ========================================

data_old <- haven::read_dta("cash_transfers_dataset/p1y2_endline_field_hh_PUBLIC.dta")
data_2 <- data_old %>%
  mutate(
    happy_numeric = as.numeric(happy_likert),
    edu_numeric = as.numeric(resp_edu),
    ration_status = as.numeric(rationcard)
  ) %>%
  select(uid, happy_numeric, edu_numeric, ration_status) %>%
  filter(!is.na(happy_numeric), !is.na(edu_numeric), !is.na(ration_status))

p1y1_end_anth_old <- haven::read_dta("cash_transfers_dataset/p1y1_endline_anthropometrics.dta
p1y1_end_anth <- p1y1_end_anth_old %>%
  mutate(whz_numeric = as.numeric(whz)) %>%
  select(uid, whz_numeric) %>%
  filter(!is.na(whz_numeric)) %>%
  distinct(uid, .keep_all = TRUE)

data_final <- data_2 %>%
  inner_join(p1y1_end_anth, by = "uid") %>%
  mutate(
    ration_status = factor(ration_status,
                           levels = c(1, 2, 3),
                           labels = c("APL", "BPL", "AAY"))
  ) %>%
  filter(!is.na(ration_status))

cat("Final sample size:", nrow(data_final), "\n")
cat("Ration card status distribution:\n")
print(table(data_final$ration_status))
cat("\nWHZ summary by ration status:\n")
print(data_final %>% group_by(ration_status) %>%
      summarise(mean_whz = mean(whz_numeric),
                sd_whz = sd(whz_numeric),
```

```
                n = n()))

# ========================================
# K-NN FUNCTION
# ========================================

my_knn <- function(K, target_edu, target_happy, target_ration, dat){
  tmp <- tibble(
    edu = c(target_edu, dat$edu_numeric),
    happy = c(target_happy, dat$happy_numeric),
    ration = factor(c(as.character(target_ration), as.character(dat$ration_status)))
  )

  dist_matrix <- cluster::daisy(tmp, metric = "gower") %>% as.matrix()
  dist_vector <- dist_matrix[, 1][-1]

  K_adjusted <- min(K, sum(!is.na(dist_vector)))

  dat %>%
    mutate(diff = dist_vector) %>%
    arrange(diff) %>%
    slice(1:K_adjusted) %>%
    summarise(pred = mean(whz_numeric, na.rm = TRUE)) %>%
    pull(pred)
}

# ========================================
# CREATE PREDICTION GRID
# ========================================

edu_vector <- seq(min(data_final$edu_numeric), max(data_final$edu_numeric), by = 1)
happy_vector <- seq(min(data_final$happy_numeric), max(data_final$happy_numeric), by = 1)
ration_vector <- levels(data_final$ration_status)

grid <- expand.grid(
  edu = edu_vector,
  happy = happy_vector,
  ration = ration_vector,
  KEEP.OUT.ATTRS = FALSE,
  stringsAsFactors = FALSE
)

# ========================================
# RUN PREDICTIONS
```

```r
# ======================================

pred_vec <- numeric(nrow(grid))
for(i in seq_len(nrow(grid))){
  if(i %% 50 == 0) cat("Processed", i, "of", nrow(grid), "\n")
  pred_vec[i] <- my_knn(
    K = 50,
    target_edu = grid$edu[i],
    target_happy = grid$happy[i],
    target_ration = grid$ration[i],
    dat = data_final
  )
}

p1_df <- grid %>% mutate(pred = pred_vec)
saveRDS(p1_df, "p1_df.rds")
```

```r
p1_df <- readRDS("cash_transfers_dataset/p1_df.rds")
p1 <- ggplot(p1_df, aes(x = edu, y = happy, fill = pred)) +
  geom_raster() +
  scale_fill_viridis_c(option = "plasma",
                       name = "Predicted\nWHZ Score",
                       na.value = "grey50") +
  facet_wrap(~ration, ncol = 3) +
  labs(
    title = "Predicted Child WHZ Score by Education, Happiness, and Poverty Level",
    subtitle = "K-NN predictions (K=50) across ration card categories",
    x = "Respondent Education Level",
    y = "Happiness Level (Likert Scale)"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(face = "bold", size = 13),
    strip.text = element_text(face = "bold", size = 10),
    legend.position = "bottom"
  )
p1
```