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.

$$H_0 : \mu_1 = \mu_2$$

$$H_a : \mu_1 > \mu_2$$

where μ_1 represents the true mean volume of rainfall in the seeded group and μ_2 represents the true mean volume of rainfall in the unseeded group. This procedure relies on three assumptions: (1) independence of observations, (2) normality of errors, and (3) equality of variance between groups. In regards to independence, we do not see any egregious violations; since clouds were (presumably) only measured once, and observations were not collected sequentially over time, there is no reason to believe that rainfall across days would be related to one another. To examine the remaining assumptions, we plotted rainfall distributions for seeded and unseeded clouds, both on the raw scale and after a log transformation (Figure 1). The raw data show a strong right skew in both groups, with noticeably greater variability in the seeded group.

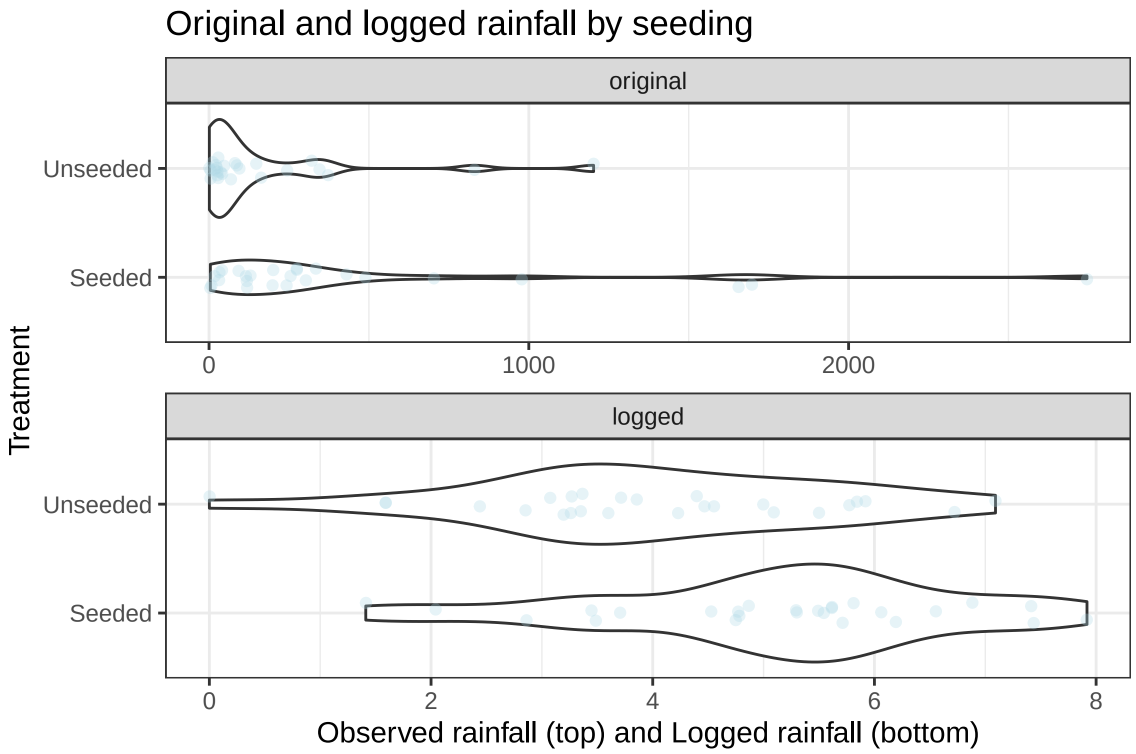


Figure 1: Plot of rainfall volume by treatment on the log (bottom) and original (top) scales.

These features suggest violations of the normality and equal-variance assumptions. Additionally, the quantile-quantile plot of the raw data (see Appendix) deviates sharply from the hypothesized normal line, consistent with a violation of normality. Similarly, the larger spread in the seeded group, visible in both violin plots and the residuals vs. fitted plot (see Appendix), indicates heterogeneity of variance. Under both of these likely violations, our t-based inference is suspect at best.

To address this issue, we applied a natural log transformation to the rainfall data. This reduced skewness and produced violin plots with more symmetric shapes and comparable spreads. The QQ plot of the transformed data aligns well with the hypothesized normal line, and the residuals vs. fitted plot shows roughly constant variance (see Appendix). Thus, the transformed data are not inconsistent with the normality and equal-variance assumptions, and we proceeded with t-based procedures on the log-transformed values.

Summary of Statistical Findings

We conducted a one-sided two sample t-test on 50 degrees of freedom, which yielded a test statistic of $t = 2.544$, with an associated p-value of 0.0141. This provides strong evidence to suggest that the mean log-rainfall for the seeded clouds is greater than that of the unseeded clouds. On the original scale, it is estimated that the median total volume of rainfall for the seeded group is 3.14 times the median volume of rainfall for the unseeded group, with a 95% confidence interval of 1.27 to 7.74.

Scope of Inference

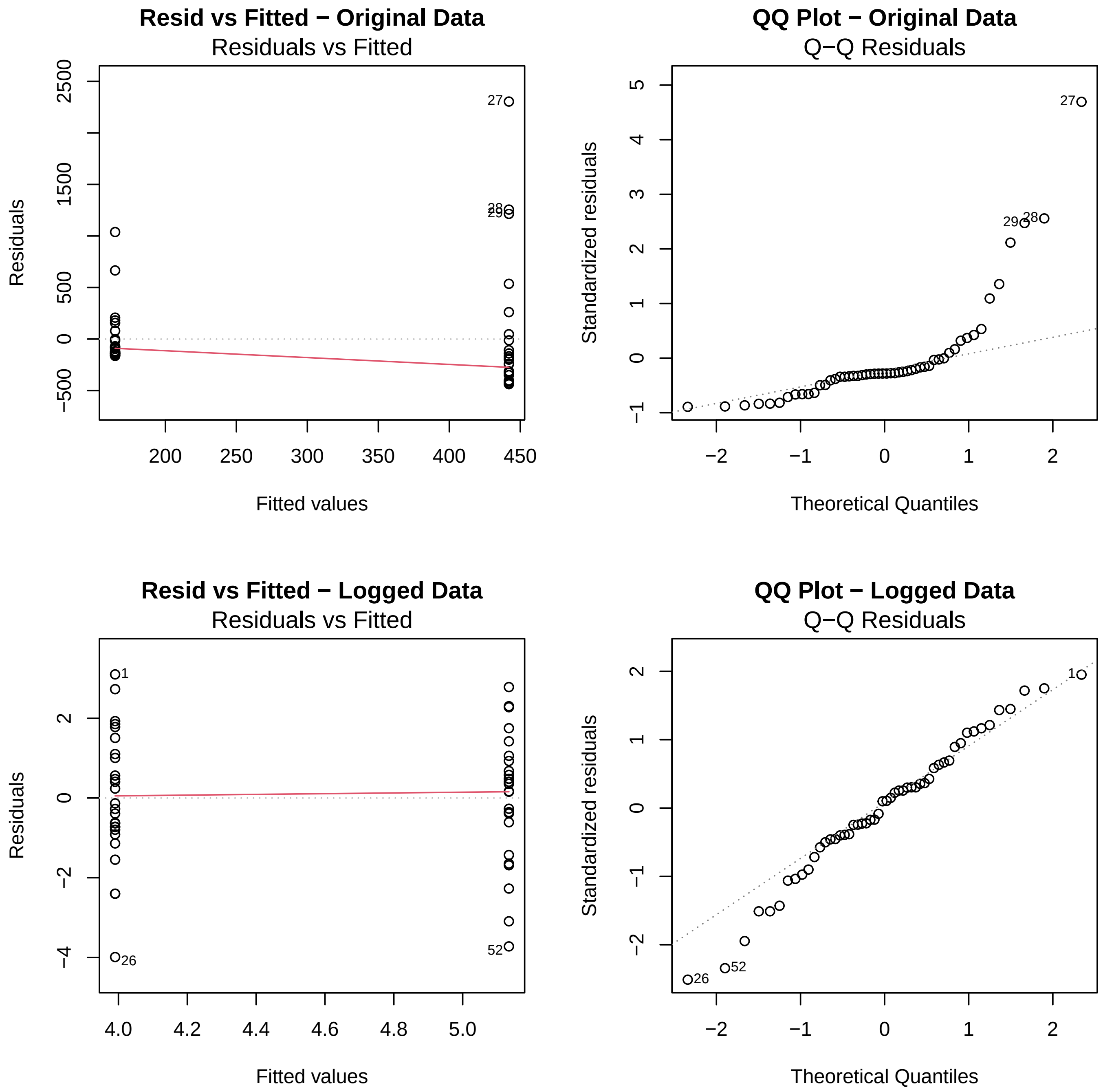
Since a random mechanism was used to determine whether clouds were seeded, we can conclude that the differences in rainfall that we observed were caused by the seeding. However, there is no reason to believe that the days on which clouds were seeded were randomly selected over the course of the four 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

Meaghan and Christian both wrote the report together. Meaghan made the exploratory graphics while Christian did the analysis with the help of ChatGPT 5.0, which was used to troubleshoot code to run the t-test.

Appendix

Plots



Code

```r
# create dataframe for ggplot
rainfall <- c(cloud$Rainfall, log(cloud$Rainfall))
treatment <- rep(cloud$Treatment, 2)
resp_type <- factor(c(rep('original', nrow(cloud)), rep('logged', nrow(cloud))))

plot_df <- tibble(rainfall, treatment, resp_type)
plot_df$resp_type<- relevel(plot_df$resp_type, ref = "original")

# plot
ggplot(plot_df, aes(x = treatment, y = rainfall)) +
  geom_violin() +
  geom_jitter(width = 0.1, alpha = 0.3, color = "lightblue") +
  facet_wrap(~ resp_type, scales = "free", nrow = 2) +
  coord_flip() +
  labs(title = 'Original and logged rainfall by seeding',
       x = 'Treatment',
       y = 'Observed rainfall (top) and Logged rainfall (bottom)') +
  theme_bw()
```

```r
cloud$Treatment <- relevel(cloud$Treatment, ref = 'Unseeded') #relevel to match book
cloud.mod <- lm(log(Rainfall) ~ Treatment, cloud) #fit model
summary(cloud.mod) #summary
exp(1.1438) #estimate on original scale
exp(confint(cloud.mod)) #95% ci on original scale
```

```r
#create diagnostic plots for original and transformed data
par(mfrow = c(2,2))
plot(lm(Rainfall ~ Treatment, data = cloud), which = 1,
  main = 'Resid vs Fitted - Original Data')
plot(lm(Rainfall ~ Treatment, data = cloud), which = 2,
  main = 'QQ Plot - Original Data')
plot(
  lm(log(Rainfall) ~ Treatment, data = cloud), which = 1,
  main = 'Resid vs Fitted - Logged Data'
)
plot(
  lm(log(Rainfall) ~ Treatment, data = cloud), which = 2,
  main = 'QQ Plot - Logged Data'
)
```