Data Analysis 2 and Coding 1 Final_Project

Mahrukh Khan

12/21/2021

Child Height and Toilets: Can Poor Sanitation Explain Stunting Across Punjab?

I decided to used the data provided by Multiple Indicator Cluster Survey (MICS) for the year 2017-18. The survey gave information regarding children, women and men well-being and households characteristics, ranging from health and education to domestic abuse and child protection to water and sanitation in Punjab. One of the data sets used included 42,408 children in Punjab, under the age of five, for whom the survey had been filled for information regarding their age, prenatal care, and anthropometry, etc. The other data set had household characteristics for 53,840 households in Punjab. The survey provided weights for each child as there could have been under representation due to limited number of surveys coming from a single locale, ensuring that the entire population was correctly represented. Even though stunting is a prevalent problem in Pakistan, I felt that few policies considered the extreme lack of sanitation facilities as a reason. My limited exposure of visiting the rural areas of Pakistan opened my eyes to the common practice of open defecation. It inspired me to explore whether stunting in children in Pakistan could have a relationship to the common practice of open defecation in their communities.

Data Selection and Cleaning

From the children data set, I selected variables representing height-for-age z score, age, mother's education, etc. From the household data set I extracted data on open defecation, if the household is in urban or rural area,wealth index, etc. The height-for-age z-score (HAZ) is used as a measure of child height outcomes to determine the impact of open defecation in explaining the child height differential across Punjab. A child is considered stunted if the height of the child falls two standard deviations below the median of the reference population (which is set at a z-score of 0), as determined by WHO. This translates into a z-score of less than -2. As a data cleaning measure, observation for HAZ>6 or <-6 are excluded. Similarly, observations pertaining to children under 6 months are also dropped because stunting does not show significant presence in these early months. For the causal variable the PSU-level(primary sampling unit) open defecation is taken into account, in recognition of the negative externalities associated with an unsanitary environment for the entire community. Each household in a particular PSU is assigned the value representing the fraction of total households within that PSU that reported defecating openly. For the understanding of other variables please refer to Figure 1 in the appendix.

Data Descriptives

For data descriptive, distributions were made for the dependent, causal and conditioning variables. The variable, open defecation, was right skewed with mean value of 0.16 units. I decided not to manipulate the data to take logs, as the zero values in this data set are interpreted as no account of open defecation in the primary sampling unit. Hence, I decided to move forward with an imbalanced data set. Height for Age Z score followed a normal distribution with an average of -1.33 standard deviations which is below the median

HAZ Score for children under the age of 5. The households on an average belong from a middle class wealth index. On average, both mother and household head, a high majority being men, have attained primary education. Figure 1 and Figure 2 from the appendix can be used for detail.

The limitations of this data set was a high number of missing values for certain variables I wanted to explore such as dietary diversity score or immunization score for children which could have been significant for height for age z score. After creating those variables I realized the data set reduce to 1500 observations whilst excluding them shifted the data set to 18,500 observations with a better analysis of our variable of interest, open defecation.Hence I made the decision of not including them. The binary variables in the data set were labelled as 1 or 2 and sometimes 9 as 'I don't know'. I labelled 1 as being completely sure of knowing if the variable existed and 2 and 9 as zero.

Patterns of Association

After studying the dataset thoroughly, I decided to conduct a non parametric regression with a loess curve between the dependent variable and the causal variable followed by the same regressions with some confounding variables. This would also help identify the functional form for the causal and confounding variables in the regression. I decided to conduct it for open defecation, age, mother and household head's education and wealth index.Figure 3 in the appendix shows the non parametric regressions.I decided to explore this by taking two approaches, First approach would be running a regression with anlspline at knot of 0.65 units of open defecation. Second approach would be adding a polynomial of a square term of open defecation in a regression. It would add to robustness. For the age variable, two knots were added, one being at 20 months and the other being at 26 months of age. Since education and wealth index will be treated as binary variables I simply checked it for association.

Comparing Explanatory Variables

A correlation plot, which can be seen in figure 4 in the appendix, aided in gaging the multicolinearity in the variables.It makes logical sense for some variables to have a certain degree of multicolinearity as economic indicators can be dependent on each other.

A highlight was that open defecation had a strong correlation with wealth quintiles. This uncovered a socioeconomic issue of there being lack of sanitation facilities in the lower quintiles. To further understand this, I regressed open defecation on the factors of wealth index, can be seen in figure 5 of appendix. They were all statistically significant at p value less than 0.001. This shows that the poorer households are suffering as a result of open defecation higher than households on the other end of the spectrum. It's a signal for the government to do target-based policy making for ensuring there are basic sanitation facilities in such communities. I have explained later on the topic of robustness.

Main model and Reasoning

My preferred model is as follows:

$HAZ2 = \beta_0 + \beta_1 \times opendef_{(opendef<0.65)} + \beta_3 \times opendef_{(opendef>=0.65)} + \beta_6 \times age_{(age<20)} + \beta_7 \times age_{(20>=age<26)}\ \beta_8 \times age_{(age>26)} + \beta_9 \times as.factor(windex5) + allotherconfoundingvariables$

All other confounding includes the variables listed in figure 1.

1. Pattern of Association (Robustness)

Looking at the pattern of association between my dependent variable, height for age z score, and my causal variable, open defecation, I decided to manipulate my causal variable in two ways. First, I decided to add a piecewise linear spline at the point where my graph's slope was changing to account for the two distinct patterns.Secondly, as the pattern of association graph had a slight upwards curve in the end, I decided to test with just adding a polynomial that is the square term of the causal variable. This would allow for more robustness check.

I decided to run four regressions. The first regression captures the impact of the confounding variables on the dependent variable. The second regression adds our causal variable with a piecewise spline. The third regression simply adds the causal variable as it is. In the fourth regressions the squared term of the causal variable is added.

2. R Square

The adjusted R square and R square from all the models is slightly higher for the piecewise linear regression.

3. Observing the regressions (Robustness)

Please refer to Figure 6 for regression results With the piecewise linear regression I saw large standard errors for the observations above the knot that is more than 65 percentage points open defecation in a region. The positive association it had with height-for-age z score also rendered for more investigation. The investigation showed that there were very few observations for this segment. Whereas, this could have indicated that there are very few areas that have open defecation of more than 65 percentage points, but in this case it's not true. It's because 8 percentage points of the total population in Punjab practices open defecation, but our data set only has approximately 4.5 percentage points of households with 65 percentage points and more defecating openly. For another project, this segment could be explored to asses for why height for age z score is increasing in these areas or is it merely incorrect results.

From the quadratic regression we can see that beta 2 signals towards a convex relationship. This could be due to having very few observations for more than 65 percentage points of open defecation that is only 800 in a data set of 18500 observations. Hence, I decided to move forward with piecewise linear spline in the regression as it seemed to appropriately capture the impact of open defecation, also with the highest R square.

4. Extra Testing: Keeping Wealth Index (Robustness)

As wealth index had a multicolinearity of 0.54 with our variable of interest that is open defecation, I wanted to further explore it. Wealth quintiles are extremely important in economic analysis but for robustness checks I had to investigate keeping them in this regression model, figure 7 in the appendix, shows that the highest R squared is for the model that has both open defecation and wealth quintiles. I decided to keep wealth quintiles as dropping them brought a one percentage points decrease in the fit of the model. Also, since they are highly statistically significant in explaining the variation in height for age z score, I made the decision of keeping them.

5. Using Weights in Regression Model (Robustness)

My data set included child weights to give equal representation to children belonging from primary sampling units with lower probability of response rates. I decided to do an analysis with and without weights with the above selected model (appendix:figure 8). The result showed that keeping weights had both adjusted R2 and R2 to have a slightly higher value than the other. Hence, I decided to move forward with keeping the weighted regression. This also added to the robustness of my model.

Interpretation

The interpretation for our final regression model, figure 9 in appendix, will be based on our causal variable, open defecation. Among household observations with open defecation impact of less than 65 percentage points (0.65), height for age z score, on average, is -0.18 standard deviations lower for households with 10 percentage points (0.1) higher open defecation, ceteris paribus. We can observe from this regression that both the education levels especially for mothers is highly statistically significant. Holding all else constant, if a mother has recieved higher education(melevel4), height for age z score, on average, is 0.66 standard deviations higher than a mother who has not attended school at all (illetrate).Another very important variable is diarrhea which is highly statistically significant. Holding all else constant, for children who had suffered from diarhhea in the past two weeks (of when the survey was conducted), height for age z score, on avergae, was 0.16 standard deviations lower than a child who had not. Whether a mother had breastfed her child also held significance with 95 percentage confidence interval. A child who had ever been breastfed by the mother had height for age z score, on average, 0.01 standard deviations higher than a child who had not.

The regression output shows the t value for both piecewise linear splines is negative 2.4 and postive 2.5 respectively. Hence we can state that at a level of significance equal to 0.05 we reject the null hypothesis. With a 95% confidence level we conclude that open defecation has a significant difference on height-for-age z score.

External Validity:

From our statistical inference, we established that the patterns are very likely present in the population represented by the data: impact of open defecation on children (under the age of five) height for age z score. Multiple Cluster Survey is a good starting point in terms of its representation of the entire Province as well as a good basis for policy planning for other provinces in Pakistan. We can use data from previous years when MICS was conducted to further test for external validity. If the results are close, it might be informative about future patterns.

Causal Interpretation:

After using all our statistical tools, open defecation indicate towards a causal relationship with height for age z score but it is difficult to establish it with the length of research conducted. I tried adding as many confounding variables as I could given the limitations of the data set, but I feel that many more important confounding variables are required to establish a causal relationship. Maybe, using data from other countries where the response rate is much higher could be a good starting point.

Conclusion:

For policy makers wishing to improve the height outcomes of children in Punjab, it is crucial that they institute programs focused on reducing the rates of open defecation alongside improving the economic well-being and literacy levels. Necessary teaching of hygiene practices as part of the education curriculum should be mandated to create awareness amongst children even if households lack information. It will provide them the basic knowledge regarding the usage of installed latrine systems and maintaining them. Education also has a direct impact on increased earnings which can result in investment towards better sanitation facilities. The early years of a child's growth are crucial in determining their nutritional strength and preventing stunting. A child's required nourishment is not fulfilled if a mother, birthing at a less than two years interval, tends to shift to early weaning. Pakistan requires effective family planning programs that educate household heads on the impediment to their future welfare due to a large family size. Cultural stereotypes need to be addressed especially in rural areas. These are some of the recommendations based on my findings. Lack of sanitation facilities in Punjab should be looked over, especially for the health of children.

Appendix

Fig 1: Variable Description

ï..Variable	Definition
Height-for-Age z-score (HAZ)	Measures the height of the child as a number of standard deviations below or above the mean value of a reference population, according to WHO standards.
Open Defecation (open_defecation)	A fraction of the number of households in the PSU that defecate openly and the total number of households in that PSU, assigned to each household within that PSU.
Breastfeed (breastfed)	Takes the value of one if the child was ever breastfed.
Child Age (cage)	A variable that assumes a value between 6-60 months, depicting the age of the child. All observations for children below 6 months of age were dropped owing to dietary variations and their impact on child growth.
Drinking Water	Takes the value of one if the household has a facility for drinking water
Handwash (handwash)	Take the value of one if the household has a handwashing facility with running water
Siblings (siblings)	Number of siblings that the child has who are also between the ages of 6-60 months.
Urban (urban)	Takes the value of one if child resides in an Urban area and zero if rural.
Diarrhea in Last 2 Weeks (diarrhea)	A variable that takes a value of one if the child has had diarrhea in the last two weeks.
Head of Household (hhsex)	Takes the value of one if the household head is a male.
Household Head Education - None/Preschool (helevel 0)	Takes the value of one if the household head is illiterate or has received preschool education at most.
Household Head Education - Primary (helevel 1)	Takes the value of one if the household head has received primary education at most.
Household Head Education - Secondary (helevel 2)	Takes the value of one if the household head has received secondary education at most.
Household Head Education - Middle (helevel 3)	Takes the value of one if the household head has received middle school education at most.
Household Head Education - Higher (helevel 4)	Takes the value of one if the household head has received higher education.
Mother Education - None/Preschool (melevel 0)	Takes the value of one if the mother is illiterate or has received preschool education at most.
Mother Education - Primary (melevel 1)	Takes the value of one if the mother has received primary education at most.
Mother Education - Secondary (melevel 2)	Takes the value of one if the mother has received secondary education at most.
Mother Education - Middle (melevel 3)	Takes the value of one if the mother has received middle school education at most.
Mother Education - Higher (melevel 4)	Takes the value of one if the mother has received higher education.
Wealth Quintile 1 (wealth_q1)	Takes a value of one if the child lives in a household that belongs to the poorest 20% of the population.
Wealth Quintile 2 (wealth_q2)	Takes a value of one if the child lives in a household that belongs to the second poorest 20% of the population.
Wealth Quintile 3 (wealth_q3)	Takes a value of one if the child lives in a household that belongs to the middle 20% of the population.
Wealth Quintile 4 (wealth_q4)	Takes a value of one if the child lives in a household that belongs to the second richest 20% of the population.
Wealth Quintile 5 (wealth_q5)	Takes a value of one if the child lives in a household that belongs to the richest 20% of the population.

	Min	Max	P25	Median	P75	Mean	SD	P95	N
HAZ2	−5.99	5.96	−2.26	−1.38	−0.48	−1.33	1.53	1.15	18500
opendef	0.00	1.00	0.00	0.04	0.26	0.16	0.22	0.62	18500
cage	6	36	13.00	20.00	28.00	20.45	8.49	34.00	18500
breastfed	0.00	1.00	1.00	1.00	1.00	0.93	0.26	1.00	18500
siblings	0.00	4.00	0.00	1.00	1.00	0.66	0.68	2.00	18500
drinkingwater	0.00	1.00	1.00	1.00	1.00	0.92	0.28	1.00	18500
handwash	0.00	1.00	1.00	1.00	1.00	0.98	0.13	1.00	18500
hhsex	0.00	1.00	1.00	1.00	1.00	0.92	0.26	1.00	18500
urban	0.00	1.00	0.00	0.00	1.00	0.27	0.44	1.00	18500
diarrhea	0.00	1.00	0.00	0.00	0.00	0.17	0.38	1.00	18500
windex5	1	5	2.00	3.00	4.00	2.84	1.38	5.00	18500
melevel	0	4	0.00	1.00	3.00	1.37	1.47	4.00	18500
helevel	0	4	0.00	1.00	3.00	1.40	1.41	4.00	18500

Fig 2a: Data Summary

Fig 2b: Distribution Histograms - Binary

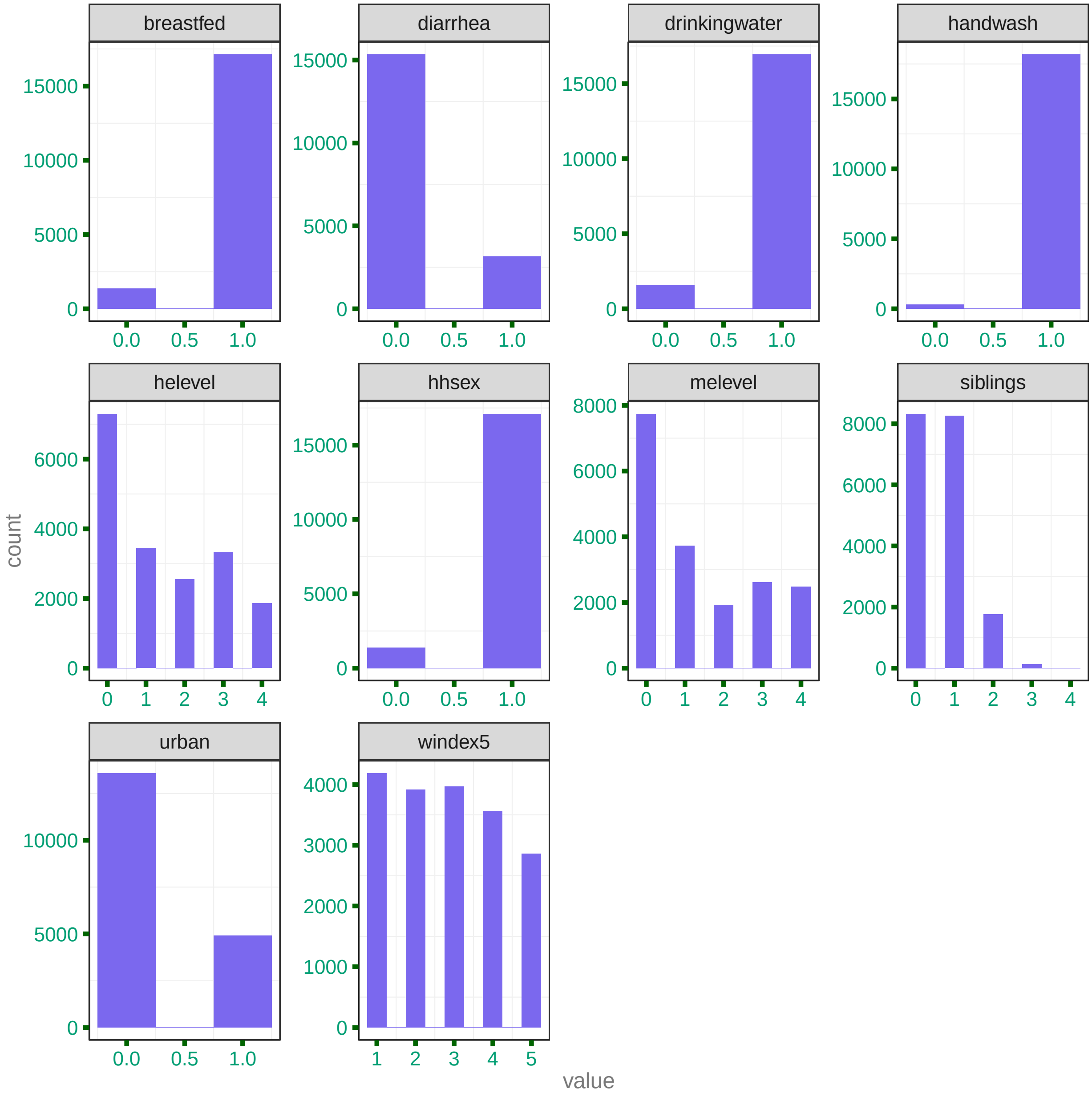

Fig 2c: Distribution Histograms - Continuous

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

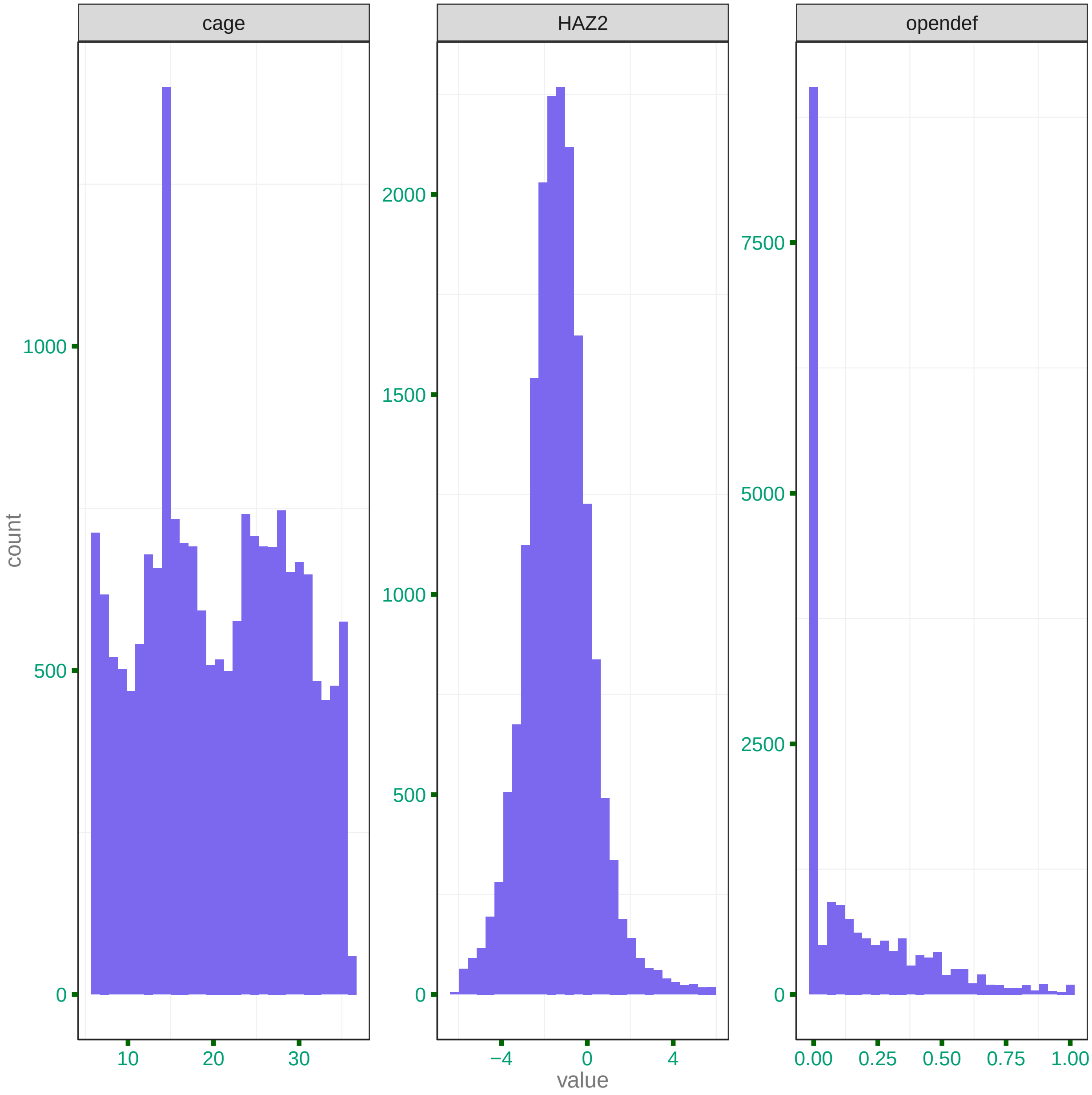

Fig 3: Patterns of Association

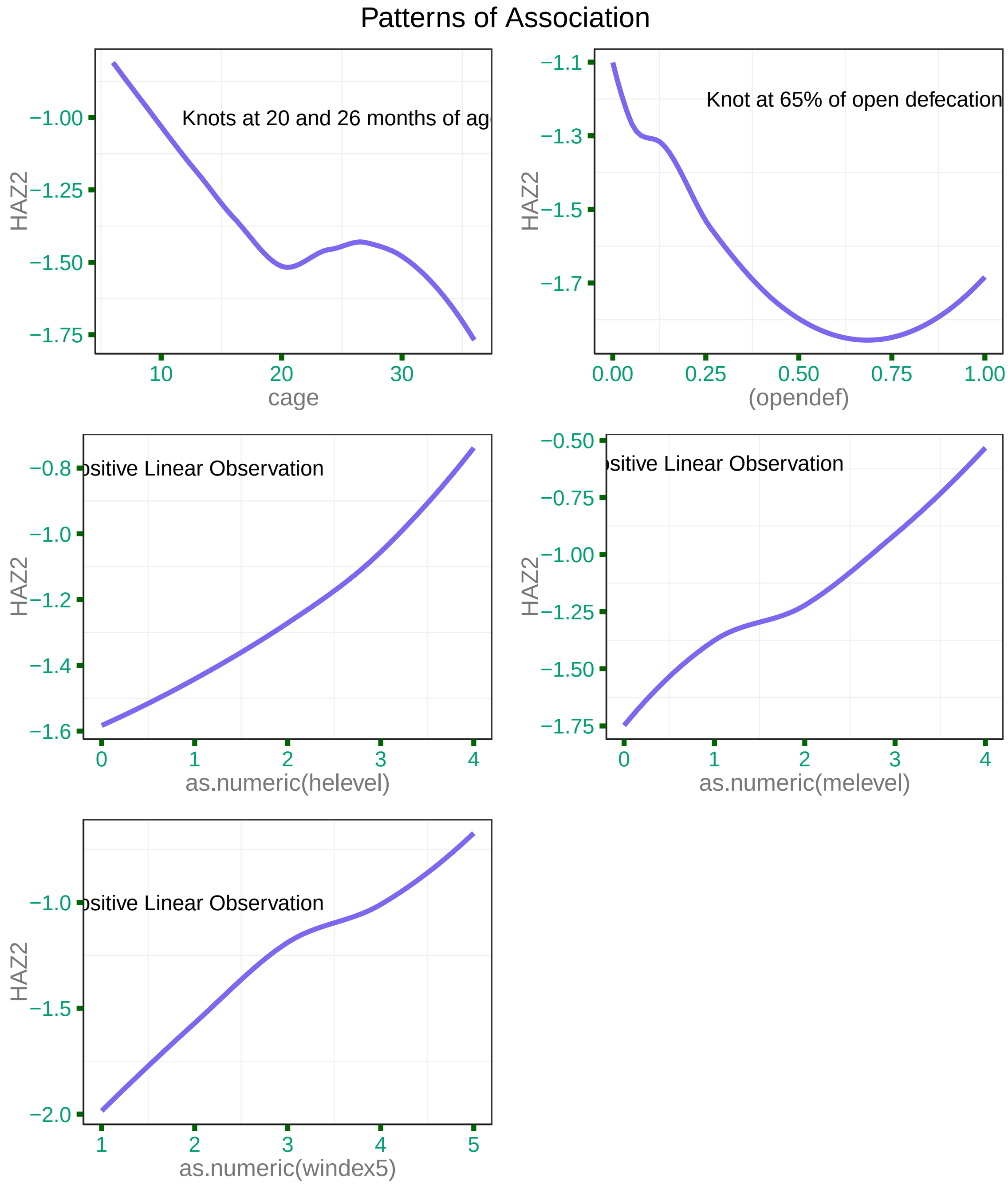

Fig 4: Correlation Plot

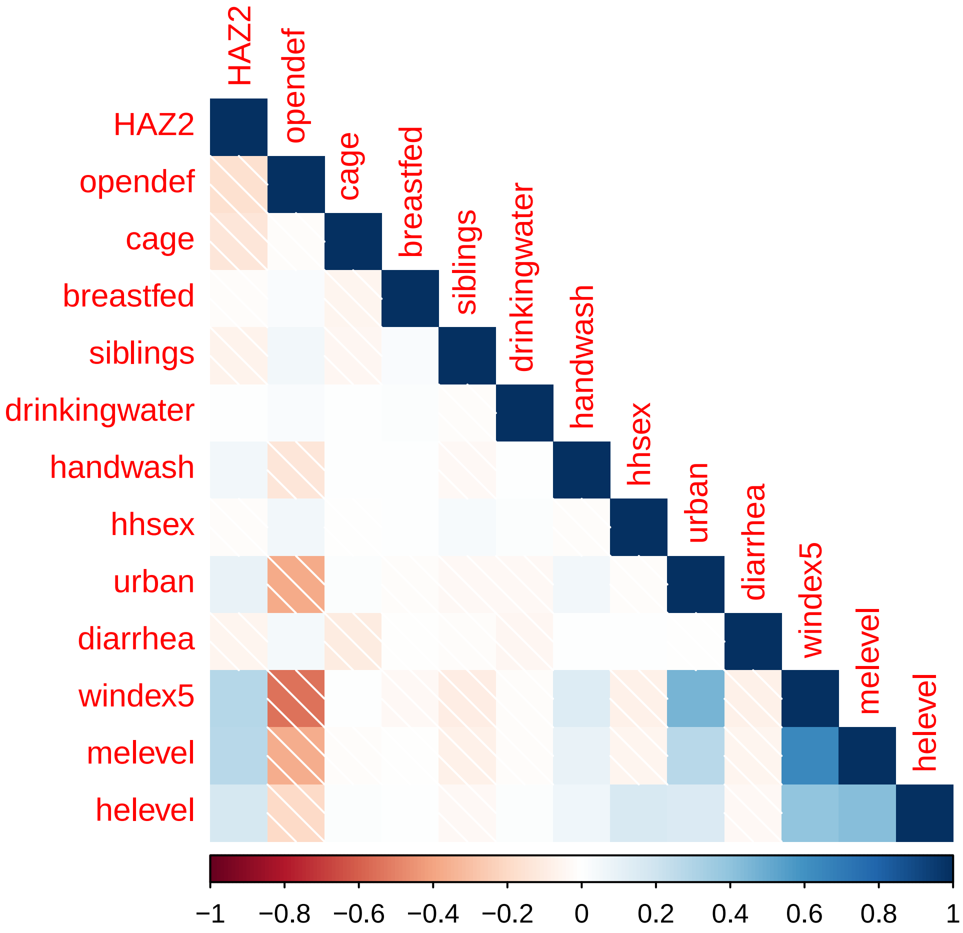

Fig 5: Explanatory Variables Regression

```
## OLS estimation, Dep. Var.: opendef
## Observations: 18,500
## Standard-errors: IID
##                      Estimate Std. Error  t value  Pr(>|t|)
## (Intercept)          0.372542   0.002804 132.8628 < 2.2e-16 ***
## as.factor(windex5)2 -0.188137   0.004034 -46.6416 < 2.2e-16 ***
## as.factor(windex5)3 -0.269662   0.004020 -67.0847 < 2.2e-16 ***
## as.factor(windex5)4 -0.323561   0.004136 -78.2372 < 2.2e-16 ***
## as.factor(windex5)5 -0.354046   0.004400 -80.4670 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.181433   Adj. R2: 0.334025
```

Fig 6: Regression Results

	reg1	reg2	reg3	reg4
Dependent Var.:	HAZ2	HAZ2	HAZ2	HAZ2
(Intercept)	-1.146*** (0.1224)	-1.100*** (0.1248)	-1.124*** (0.1245)	-1.100*** (0.1251)
handwash	0.1530. (0.0871)	0.1448. (0.0872)	0.1499. (0.0871)	0.1470. (0.0872)
breastfed	-0.0991* (0.0457)	-0.0989* (0.0457)	-0.0989* (0.0457)	-0.0985* (0.0457)
lspline(cage,c(20,26))1	-0.0560*** (0.0035)	-0.0560*** (0.0035)	-0.0560*** (0.0035)	-0.0560*** (0.0035)
lspline(cage,c(20,26))2	0.0360*** (0.0070)	0.0360*** (0.0070)	0.0360*** (0.0070)	0.0360*** (0.0070)
lspline(cage,c(20,26))3	-0.0389*** (0.0057)	-0.0387*** (0.0057)	-0.0389*** (0.0057)	-0.0388*** (0.0057)
drinkingwater	0.0382 (0.0388)	0.0414 (0.0388)	0.0387 (0.0388)	0.0414 (0.0389)
siblings	-0.1193*** (0.0164)	-0.1198*** (0.0164)	-0.1193*** (0.0164)	-0.1198*** (0.0164)
hhsex	0.0269 (0.0458)	0.0258 (0.0459)	0.0274 (0.0459)	0.0260 (0.0459)
diarrhea	-0.1640*** (0.0289)	-0.1641*** (0.0289)	-0.1638*** (0.0289)	-0.1639*** (0.0289)
urban	-0.1152*** (0.0280)	-0.1306*** (0.0290)	-0.1206*** (0.0289)	-0.1315*** (0.0293)
as.factor(melevel)1	0.1549*** (0.0308)	0.1536*** (0.0309)	0.1535*** (0.0309)	0.1540*** (0.0309)
as.factor(melevel)2	0.1777*** (0.0418)	0.1744*** (0.0419)	0.1761*** (0.0419)	0.1750*** (0.0419)
as.factor(melevel)3	0.4066*** (0.0393)	0.4040*** (0.0393)	0.4052*** (0.0393)	0.4043*** (0.0393)
as.factor(melevel)4	0.6659*** (0.0464)	0.6629*** (0.0465)	0.6644*** (0.0464)	0.6637*** (0.0465)
as.factor(helevel)1	0.0186 (0.0307)	0.0199 (0.0307)	0.0186 (0.0307)	0.0193 (0.0307)
as.factor(helevel)2	0.0803* (0.0355)	0.0819* (0.0355)	0.0803* (0.0356)	0.0813* (0.0355)
as.factor(helevel)3	0.1118** (0.0344)	0.1143*** (0.0344)	0.1122** (0.0344)	0.1140*** (0.0344)
as.factor(helevel)4	0.2304*** (0.0452)	0.2343*** (0.0452)	0.2313*** (0.0452)	0.2336*** (0.0453)
as.factor(windex5)2	0.3328*** (0.0339)	0.3216*** (0.0354)	0.3233*** (0.0354)	0.3213*** (0.0355)
as.factor(windex5)3	0.5673*** (0.0367)	0.5466*** (0.0392)	0.5545*** (0.0391)	0.5471*** (0.0393)
as.factor(windex5)4	0.6458*** (0.0412)	0.6204*** (0.0440)	0.6318*** (0.0439)	0.6200*** (0.0442)
as.factor(windex5)5	0.8478*** (0.0506)	0.8218*** (0.0529)	0.8338*** (0.0528)	0.8206*** (0.0532)
lspline(opendef,0.65)1		-0.1813* (0.0741)		
lspline(opendef,0.65)2		0.8901* (0.3495)		
opendef			-0.0532 (0.0641)	-0.3310* (0.1536)
opensq				0.3922. (0.2077)
S.E. type	Heteroskedast.-rob.	Heteroskedast.-rob.	Heteroskedast.-rob.	Heteroskedast.-rob.
Observations	18,500	18,500	18,500	18,500
R2	0.13591	0.13638	0.13594	0.13613
Adj. R2	0.13488	0.13526	0.13487	0.13500

Fig 7: Multicolinearity Testing Regression Results

	reg1	reg5	reg2
Dependent Var.:	HAZ2	HAZ2	HAZ2
(Intercept)	-1.146*** (0.1224)	-0.8087*** (0.1239)	-1.100*** (0.1248)
handwash	0.1530. (0.0871)	0.2625** (0.0863)	0.1448. (0.0872)
breastfed	-0.0991* (0.0457)	-0.1225** (0.0460)	-0.0989* (0.0457)
lspline(cage,c(20,26))1	-0.0560*** (0.0035)	-0.0560*** (0.0035)	-0.0560*** (0.0035)
lspline(cage,c(20,26))2	0.0360*** (0.0070)	0.0356*** (0.0070)	0.0360*** (0.0070)
lspline(cage,c(20,26))3	-0.0389*** (0.0057)	-0.0391*** (0.0058)	-0.0387*** (0.0057)
drinkingwater	0.0382 (0.0388)	0.0455 (0.0392)	0.0414 (0.0388)
siblings	-0.1193*** (0.0164)	-0.1338*** (0.0165)	-0.1198*** (0.0164)
hhsex	0.0269 (0.0458)	-0.0327 (0.0460)	0.0258 (0.0459)
diarrhea	-0.1640*** (0.0289)	-0.1932*** (0.0291)	-0.1641*** (0.0289)
urban	-0.1152*** (0.0280)	-0.0142 (0.0272)	-0.1306*** (0.0290)
as.factor(melevel)1	0.1549*** (0.0308)	0.2670*** (0.0300)	0.1536*** (0.0309)
as.factor(melevel)2	0.1777*** (0.0418)	0.3646*** (0.0405)	0.1744*** (0.0419)
as.factor(melevel)3	0.4066*** (0.0393)	0.6401*** (0.0362)	0.4040*** (0.0393)
as.factor(melevel)4	0.6659*** (0.0464)	0.9387*** (0.0419)	0.6629*** (0.0465)
as.factor(helevel)1	0.0186 (0.0307)	0.0500 (0.0309)	0.0199 (0.0307)
as.factor(helevel)2	0.0803* (0.0355)	0.1259*** (0.0358)	0.0819* (0.0355)
as.factor(helevel)3	0.1118** (0.0344)	0.1949*** (0.0341)	0.1143*** (0.0344)
as.factor(helevel)4	0.2304*** (0.0452)	0.3361*** (0.0450)	0.2343*** (0.0452)
as.factor(windex5)2	0.3328*** (0.0339)		0.3216*** (0.0354)
as.factor(windex5)3	0.5673*** (0.0367)		0.5466*** (0.0392)
as.factor(windex5)4	0.6458*** (0.0412)		0.6204*** (0.0440)
as.factor(windex5)5	0.8478*** (0.0506)		0.8218*** (0.0529)
lspline(opendef,0.65)1		-0.6421*** (0.0693)	-0.1813* (0.0741)
lspline(opendef,0.65)2		0.9440** (0.3486)	0.8901* (0.3495)
S.E. type	Heteroskedast.-rob.	Heteroskedast.-rob.	Heteroskedast.-rob.
Observations	18,500	18,500	18,500
R2	0.13591	0.12291	0.13638
Adj. R2	0.13488	0.12196	0.13526

Fig 8: Weighted Analysis

	weight_with	weight_without
Dependent Var.:	HAZ2	HAZ2
(Intercept)	-1.100*** (0.1248)	-1.037*** (0.1183)
lspline(opendef,0.65)1	-0.1813* (0.0741)	-0.1833* (0.0722)
lspline(opendef,0.65)2	0.8901* (0.3495)	0.8096* (0.3314)
handwash	0.1448. (0.0872)	0.1126 (0.0839)
breastfed	-0.0989* (0.0457)	-0.0862* (0.0424)
lspline(cage,c(20,26))1	-0.0560*** (0.0035)	-0.0569*** (0.0033)
lspline(cage,c(20,26))2	0.0360*** (0.0070)	0.0361*** (0.0066)
lspline(cage,c(20,26))3	-0.0387*** (0.0057)	-0.0363*** (0.0055)
drinkingwater	0.0414 (0.0388)	0.0483 (0.0375)
siblings	-0.1198*** (0.0164)	-0.1167*** (0.0158)
hhsex	0.0258 (0.0459)	0.0100 (0.0417)
diarrhea	-0.1641*** (0.0289)	-0.1782*** (0.0273)
urban	-0.1306*** (0.0290)	-0.1407*** (0.0284)
as.factor(melevel)1	0.1536*** (0.0309)	0.1342*** (0.0300)
as.factor(melevel)2	0.1744*** (0.0419)	0.1594*** (0.0392)
as.factor(melevel)3	0.4040*** (0.0393)	0.3782*** (0.0382)
as.factor(melevel)4	0.6629*** (0.0465)	0.6227*** (0.0443)
as.factor(helevel)1	0.0199 (0.0307)	0.0344 (0.0293)
as.factor(helevel)2	0.0819* (0.0355)	0.0993** (0.0340)
as.factor(helevel)3	0.1143*** (0.0344)	0.1350*** (0.0325)
as.factor(helevel)4	0.2343*** (0.0452)	0.2319*** (0.0436)
as.factor(windex5)2	0.3216*** (0.0354)	0.3084*** (0.0343)
as.factor(windex5)3	0.5466*** (0.0392)	0.5576*** (0.0382)
as.factor(windex5)4	0.6204*** (0.0440)	0.6390*** (0.0430)
as.factor(windex5)5	0.8218*** (0.0529)	0.8340*** (0.0510)
S.E. type	Heteroskedast.-rob.	Heteroskedast.-rob.
Observations	18,500	18,500
R2	0.13638	0.13188
Adj. R2	0.13526	0.13075

Fig 9: Final Regression Model

```
## OLS estimation, Dep. Var.: HAZ2
## Observations: 18,500
## Standard-errors: Heteroskedasticity-robust
##                            Estimate Std. Error   t value   Pr(>|t|)
## (Intercept)               -1.099723   0.124751  -8.81533  < 2.2e-16 ***
## lspline(opendef, 0.65)1   -0.181268   0.074142  -2.44487 1.4500e-02 *
## lspline(opendef, 0.65)2    0.890132   0.349509   2.54680 1.0879e-02 *
## handwash                   0.144831   0.087207   1.66077 9.6777e-02 .
## breastfed                 -0.098876   0.045662  -2.16539 3.0370e-02 *
## lspline(cage, c(20, 26))1 -0.055996   0.003472 -16.12552  < 2.2e-16 ***
## lspline(cage, c(20, 26))2  0.035994   0.006994   5.14625 2.6847e-07 ***
## lspline(cage, c(20, 26))3 -0.038738   0.005722  -6.77036 1.3233e-11 ***
## ... 17 coefficients remaining (display them with summary() or use argument n)
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 1.42572   Adj. R2: 0.13526
```