Poverty Levels in Costa Rica

Luke Spellman, John Oliver & Drew Kairis

November 26, 2018

Introduction

Background Information

"10 percent of the world's population lived on less than US$1.90 a day" (World Bank) in 2015. Despite the progress that has been made in this area, poverty is still a main concern across the world. This topic is analyzed from country to country, as well as a whole through a global or humanistic lense. Extreme poverty is defined as living on less than 1.90 US dollars per day, however the definition for poverty in general is less concise. Our project revolves around defining poverty without using income. This shifts the discussion in a new direction and possibly a more robust and powerful direction. This restriction allows poverty to be defined in other terms, such as standards of living, which could potentially give better insight into who is affected by poverty and why especially compared to simply a monetary number. Although poverty is a very important issue in society, it is often difficult to reach those affected by poverty due to the poorest in society not being able to provide the necessary income and expense records to prove that they qualify for such programs. This curtails the effectiveness of government programs designed to give relief to those in need. This paper aims to define poverty in a new way in hopes of shedding light on a different approach to identifying and reaching those affected.

Experimental Design

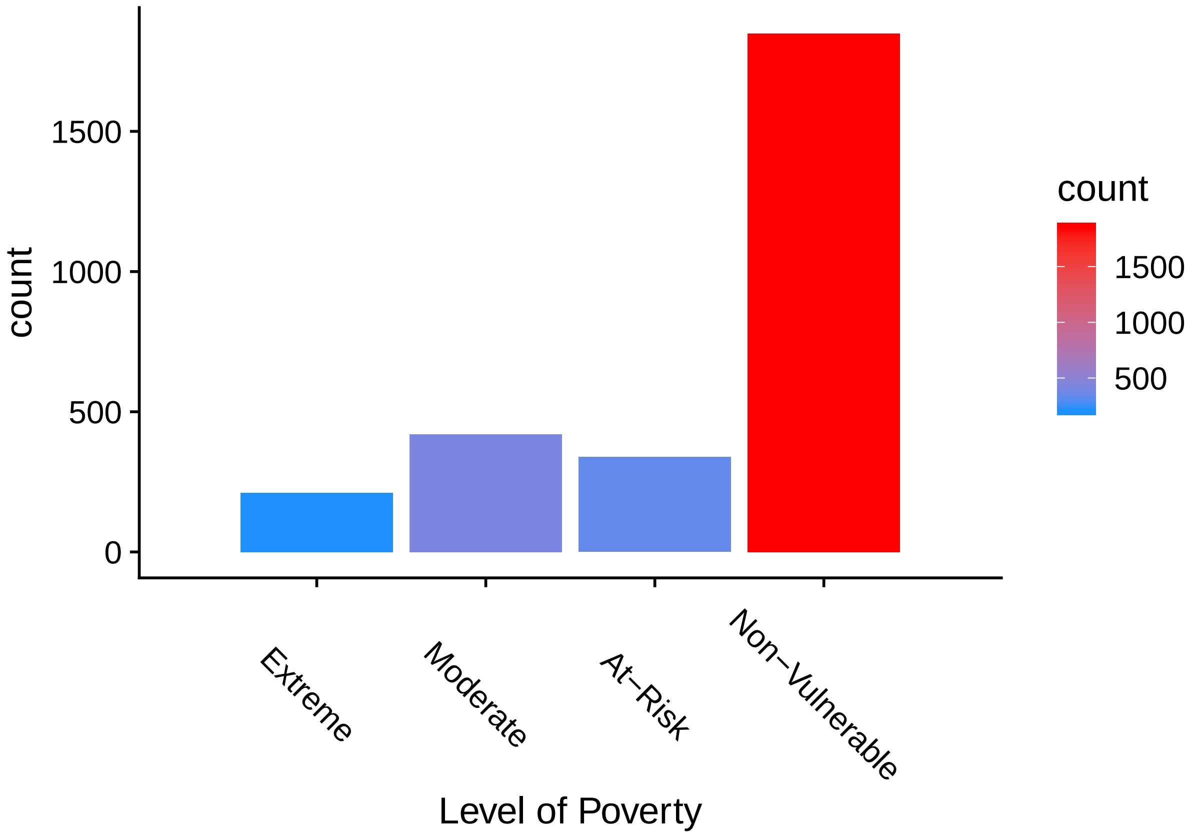
Using data containing information on individuals in Costa Rica, with their current level of poverty (extreme, moderate, at risk, non vulnerable), our goal is to build a model that is able to effectively identify and categorize the level of poverty for individuals. This can then be used to quantify household needs, and help to determine which households are in most need of assistance. The scope of this project is to predict poverty levels solely for the heads of the households, rather than the entire household. If one member of the household is in poverty we are assuming that the others who live in the same house, under the same conditions, are also in poverty. More information could be potentially gained through utuilizing everyone collected in this sample however, this information may not be beneficial in building a model to predict poverty level for the heads of the households.

Data collection

The data set we were given was collected from different households throughout Central America, specifically of Costa Rica. The data was collected by the Inter-American Development Bank, which is known for being the main source of multilateral financing in Latin America. This company is involved in multiple large scale projects, mostly focusing on the improvement of life. This data was collected at the individual level.

Data Structure

The dataset we were given initially contained 9057 people from different regions of Costa Rica, in which (143) variables were collected. Most of these variables revolved mostly around geographic and socioeconomic information, excluding income. A few examples of the information collected are: region the person is from, number of people in the household, material the house is built out of and educational background of those in the household. Despite being given 9057, we had 2819 heads of households.



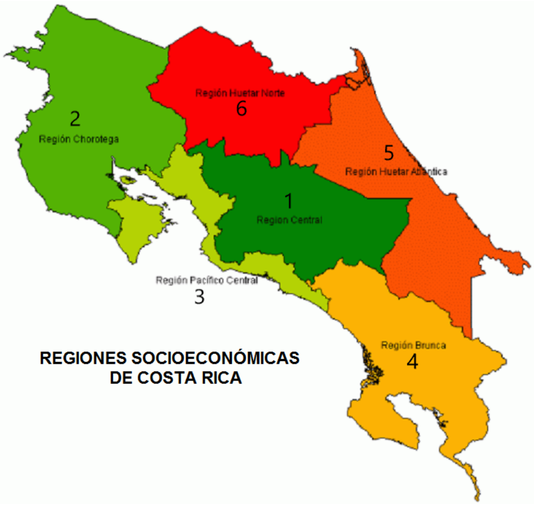
Subsetting The Data

Our task was focused on only the heads of the household, as mentioned. Therefore, any information regarding someone who is not the head may be counterproductive in regards to building a model predicting for the heads of the household. After subsetting down to just the heads, we were left with 2819 individuals.

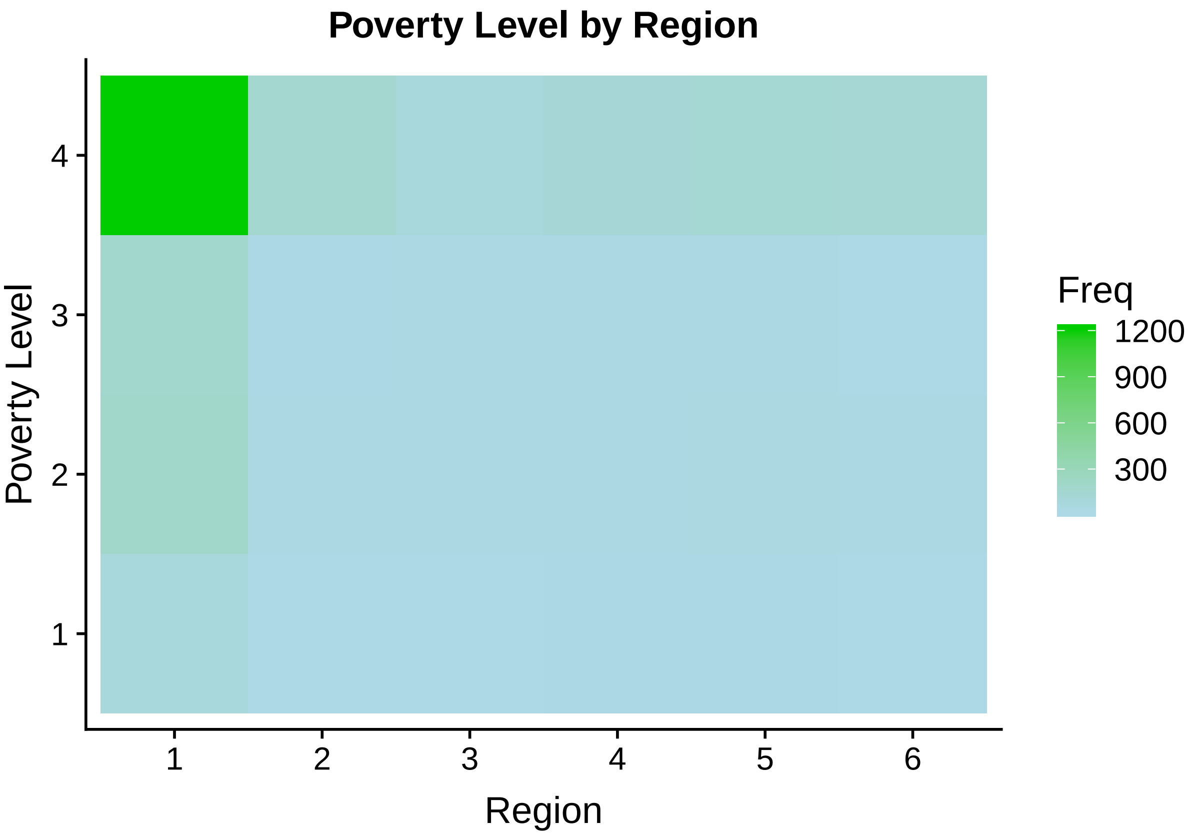
Diving into the Data

The given data set required heavy manipulation. There was redundant information within the data, as well as some variables that contained missing values and were incomplete. The information collected contained some numeric variables such as age, some count data, but mostly binary responses to questions. For instance, the region that each individual lived in was not collected in one column but six columns corresponding to a yes in one region and five no's for the other regions.

Therefore, an important part of this project will be collapsing this data appropriately, ensuring there is no unintentional data loss. Moreover, in collapsing these binary columns, count data will be created in order to represent this data more concisely. For instance, in the region area mentioned above, each region will be given a numeric value and then the six columns will then be collapsed into one variable where each observation has a numeric value corresponding to which region the person lives in. Variables with a large proportion missing were eliminated from our dataset, while the variables that only had a small proportion of values missing were kept and filled in with the average of that variable.



Collapsing and Manipulation of Data



An interesting side note is that 1207 of our observations come from the non vulnerable group in region 1. This is 42.8 of our data and it sparks curiosity of if this dataset is representative of all of Costa Rica. It would be interesting to explore possible response bias in this sample. This would also tie back to possible information in those who are not the heads of the household. Since there are more total ID's in idhogar than there are ID's for heads of households, this means that not all people who are in this dataset are in households in this dataset. Therefore, it is reasonable to subset down to only heads of households since we don't want to build a model on information from houses that are not contained in this dataset. Mostly because this could be counter productive to building a model for these heads of households.

Due to the initial form of the information collected, this data must be collapsed before analysis can be conducted. This dimensionality reduction will initially be conservative in order to preserve the information collected. Our goal is to build a simple and interpretable model, while still remaining accurate.

As mentioned above, a big part of dealing with this dataset is collapsing the initial 128 down to a workable number of variables that eliminate redundant information and are still insightful. In doing this, some feature selection is left up to our discretion, especially when it comes down to the interpretation of the variables we were given. One thing that was difficult about this project was that the data that was given had a brief description of each variable with it, which is somewhat useful but often vague. For example, in some of our variables had an ‘other’ option that was supplied. The definition of other, within the context, was unclear and made selecting how to treat that category unclear. Although the other responses were usually a small proportion of our data, getting more information on some of these responses would be helpful. As stated earlier, identifying those in poverty and getting relief to them has seemed to be tricky, therefore knowing the correct interpretation of the information collected is essential.

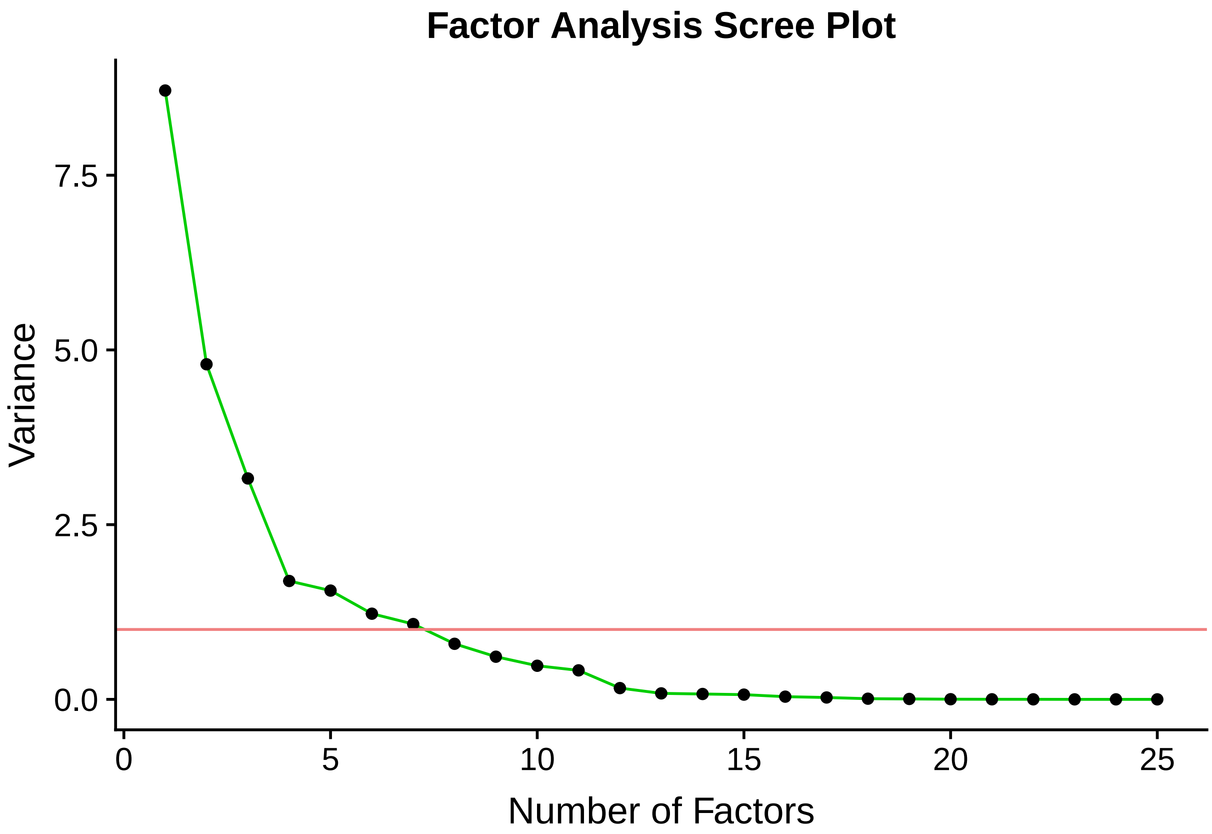
The main process’ utilized in dimensionality reduction was taking related binary variables and collapsing them down into one ordinal variable; effectively turning binary data into count data. During the process of brushing up the data, we found some repeated variables, where the same information was provided for multiple variables. These were then eliminated, as carrying multiple of the same variables is unnecessary.

Since this project revolves around predicting not only who is in poverty, but also the poverty level of individuals, some of the variables we were given were collapsed into different categories while some were directly combined. When dealing with categorical data, both nominal and ordinal, a common question is what is the distance between the categories? And are the categories equally spaced? This is simply due to the nature of categorical data but forces deeper analysis or simply thought than a quantitative variable. This may not be true in every case but one example is age. Does having the grouped age such as which decade someone was born in provided the same information as having the exact age as that same person? We collapsed a total of 16 variables that were deemed redundant, changing them from multiple binary variables to ordinal variables containing counts of each. Of these, some of the variables included: material of the walls, region which the household is in, and marital status.

Factor Analysis for Continuous Variables

After trimming down some of the binary variables above, we aimed to collapse continuous variables through a method called factor analysis. Factor analysis aims to analyze variation among the inputted variables and potentially identify unseen variables. Factor analysis recommends a number of factors to collapse down to in order to maintain ‘enough’ of the variance of the inputted variables. This is effectively performs dimensionality reduction and may give insight into latent, or unobserved, factors which are related to the observed data. This process, although slightly convoluted, is intended to be interpretable in the factors that it recommends. As a result of this, factor analysis is not an exact science and often does not converge to a single “correct” answer. Despite this, factor analysis can be very useful in identifying interesting combinations of variables.

To make sure that this method was appropriate, we ran Bartlett test which checks homogeneity of the variance across our sample. The results of said test returned that factor analysis could indeed be insightful. We then created a correlation matrix of our desired variables, and ran that through a KMO test, to validate that factor analysis could be useful. The KMO test returned an MSA of 0.71, which is a metric signifying the overall relatedness of our inputted data. Therefore, we continued with this method. We created a scree plot which allows for visual determination of approximately how many new factors should be created.

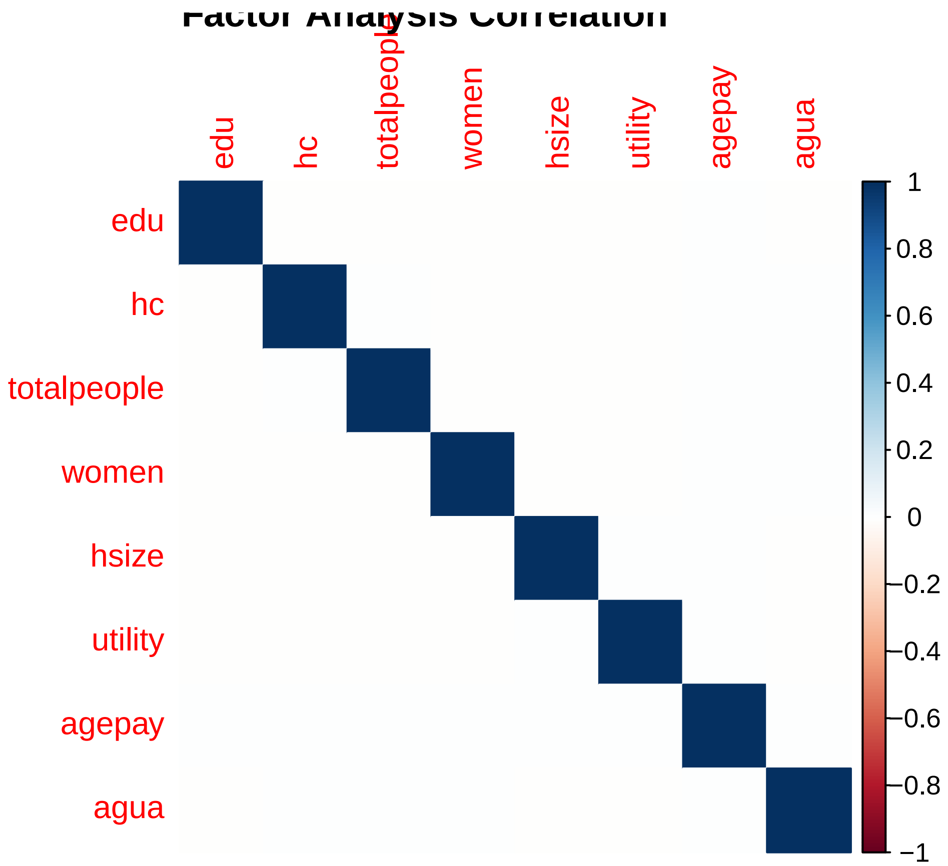


From this, it was determined that we could separate our data down to 8 separate factors.

Table 1.

Group	Variables	Number of Variables collapsed in this Factor
1	Education-related	5
2	Total person	2
3	Total women	2
4	House size and Crowding	4
5	House condition	3
6	Age	2
7	Utilities	3
8	Water	2

The way factor analysis combines these scores is somewhat involved with linear algebra however, conceptually the factors are given weights which are the importance or correlation to that latent factor. Then this weight is multiplied by the standardized version on the variable for that weight. These are called factor scores which are seen as the net effect of the variables grouped into that factor. As seen in table 1 this process definitely helped with dimensionality reduction and also with collinearity of these variables as well.

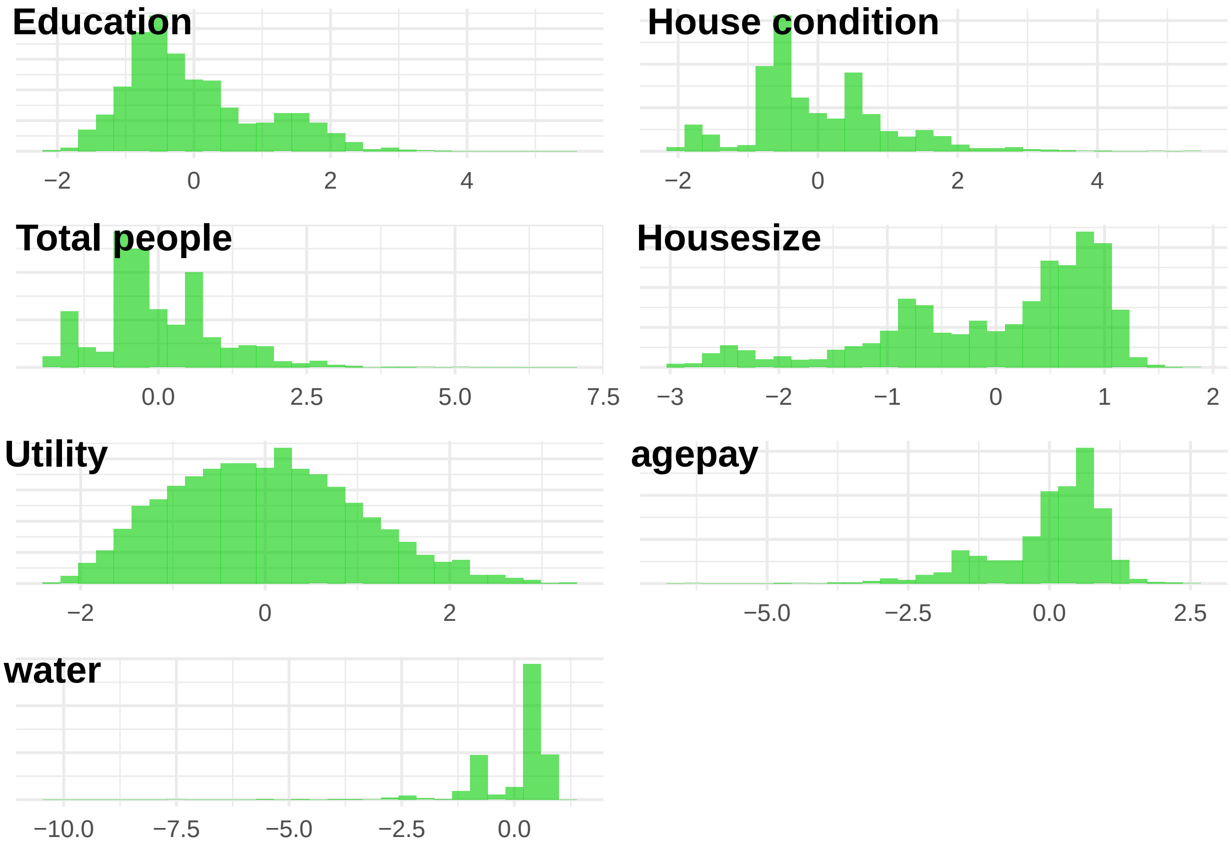


One thing to note, is that factor analysis was only applied to ordinal and continuous variables. We tried to implement factor analysis for categorical (nominal) variables but we ran out of time to do so. We also explored packages which supported both continuous and categorical variables however, time was an issue here as well. Since factor analysis was only applied to ordinal and continuous variables, many of the count data created earlier in this project were reordered to maximize the ordinal variables to take advantage and utilize the natural ordering of the data.

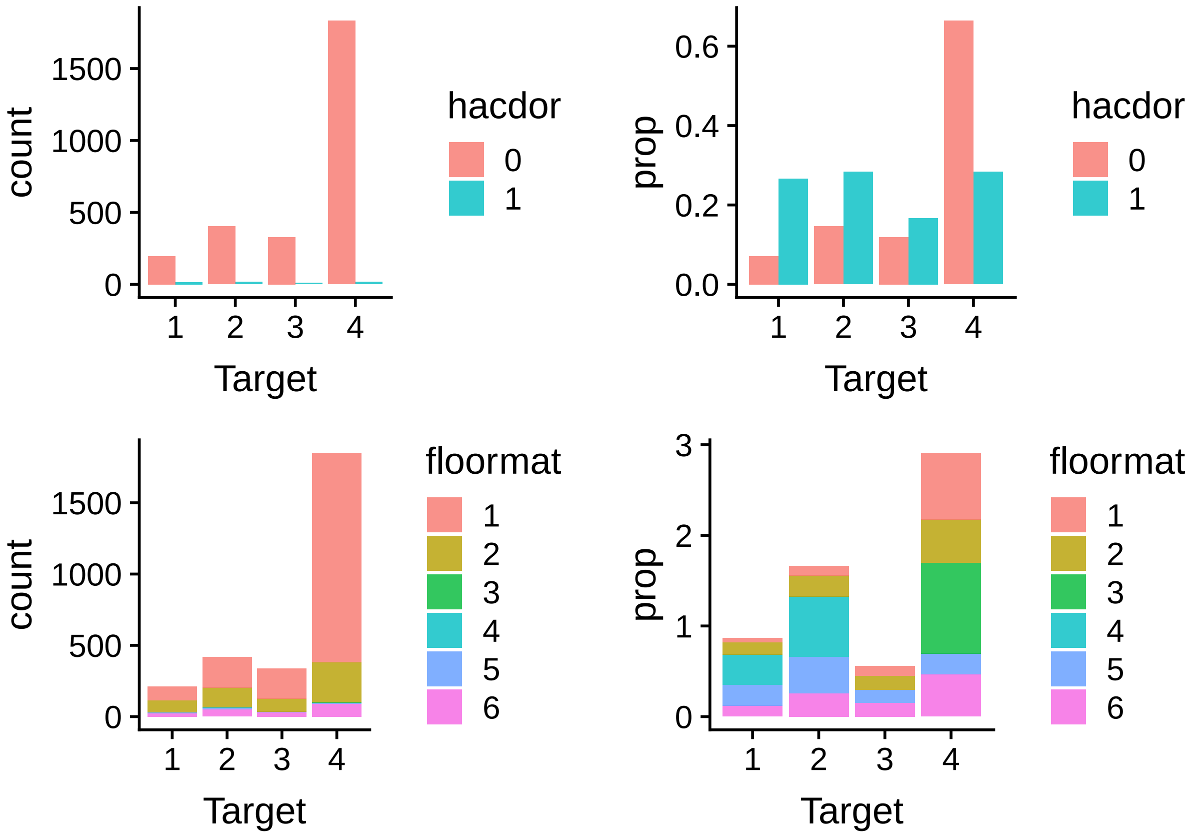
In addition, standardizing the variables was done before performing factor analysis in order to account for the different scales of all of the variables. Differing scales could allow a variables with a large influence in factor analysis simply because tat variable takes large values. To account for this all the variables used in factor analysis were normalized. In other words, transformations were applied on our continuous and ordinal variables in order to minimize the effect of skewness as well as equally weight the variables being inputted into factor analysis. The specific transformation was simply standardization of every variable, which is subtracting the mean from every observation then dividing by the standard deviation of that variable. This effectively normalized our continuous data.

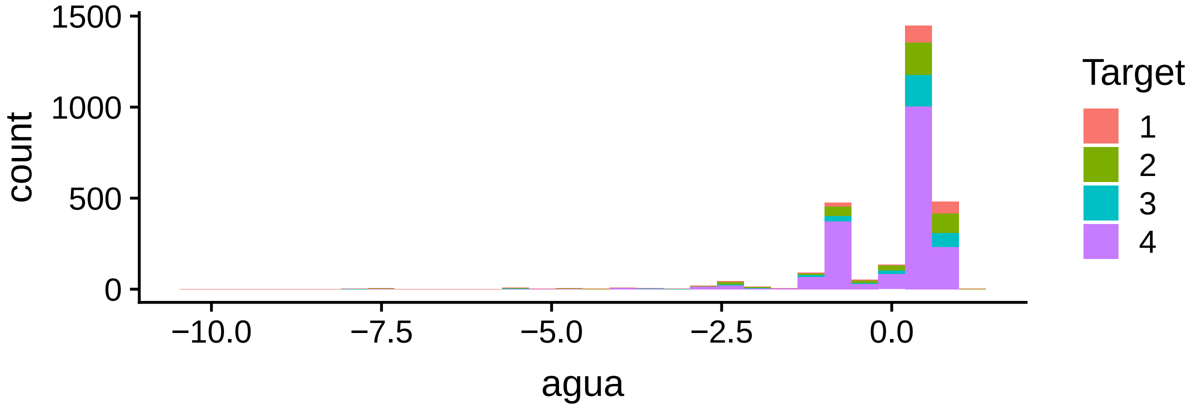
Variable Exploration

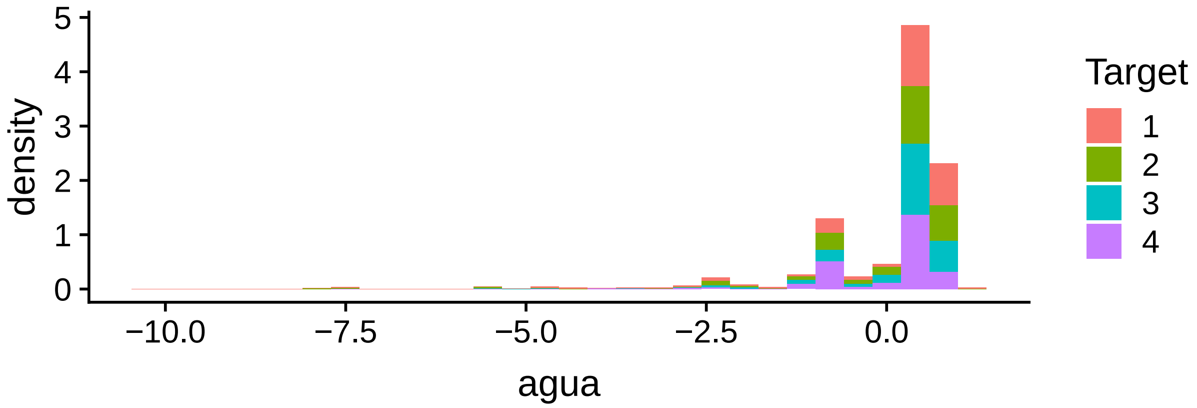
After performing factor analysis on the continuous and ordinal variables, the new factors and the nominal variables were bound into a workable dataset. This dataset was then analyzed using various plotting techniques to explore distributions of variables as well as to look into correlations between variables. Analysis of the ordinal and continuous variables was more straightforward since quantitative data is more commonly dealt with and more types of plots are available for data visualization. Specifically, density plots and histograms were utilized:



These histograms above convey that the new factors from factor analysis are normal (or as normal as they can be) and this was also verified by running the transformTukey function from the rcompanion package. This function recommends a transformation to apply in order to transform inputted data to normal. When applying this function to our new factor variables, this function recommended raising all the variables to the power of 1. This clearly shows that the normalization applied before factor analysis was maintained during factor analysis (due to linear combinations of normal variables maintaining the normal distribution) and that there is no further need for transformations on these variables.


count
1500
1000
500
0
1
2
3
4
Target
hacdor
0
1
prop
0.6
0.4
0.2
0.0
1
2
3
4
Target
hacdor
0
1
count
1500
1000
500
0
1
2
3
4
Target
floormat
1
2
3
4
5
6
prop
3
2
1
0
1
2
3
4
Target
floormat
1
2
3
4
5
6





According to the plots above it appears that floor material 3 and 4 may add some information in predicting one's poverty level however, analyzing the counts of this variables show that there are only 2 observations in category 3 and 3 people in category 4. Similarly, the hacdor variable seems to be interesting as well until the counts of this variable is analyzed and a similar scenario is discovered. As other variables were explored this became a common theme, specifically that many of the binary data were clear signals of poverty however, the number of people blatantly in poverty was very small in comparison to the rest of our data. This makes analyzing proportions deceptive, however, this is still good to do but being aware of this is extremely important.

Building a Model

The methods carried out through data exploration carry over into building an initial model. Using the selected transformations, we remove all highly correlated and redundant variables since. The goal of this section is to build an interpretable model that is able to accurately classify poverty level of the heads of the households.

Splitting Data Into Test and Training Set

Upon implementation, our model will be used to categorize poverty level on a held-out test set of samples we have not seen before in order to validate that the model constructed is able to generalize to new data. To ensure confidence prior to implementation, the data we have available to us should be partitioned such that we can both train a model and check the fit on data the model hasn't used. To do this, we split our transformed data into 75% training, and 25% cross-validation. The process is done through random sampling based on a set seed (for replicable results), and results in a training set size of 2114 and cross validated set size of 705.

Model selection

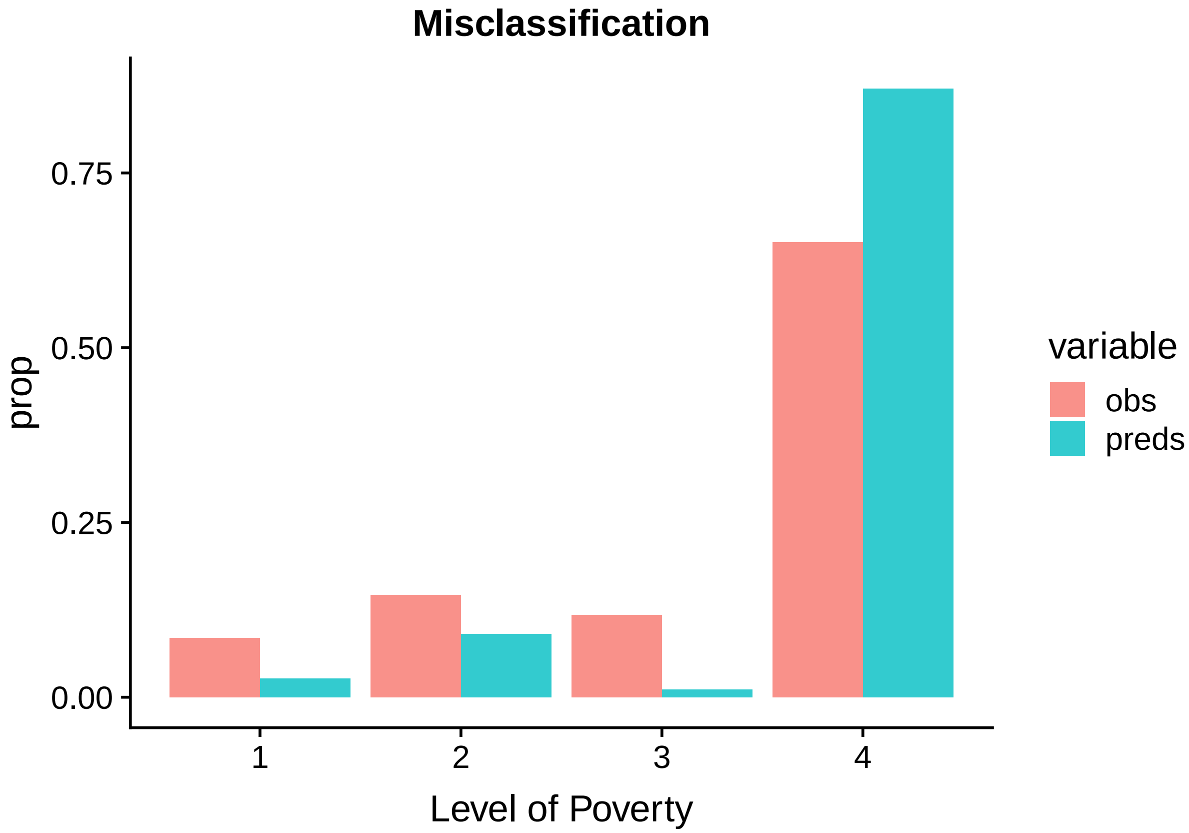
Since our objective is to classify poverty levels (1-4) there are a few different types of models that can potentially be applied to this type of problem. The different types of models that were considered were poisson or negative binomial regression, multinomial regression and cumulative logistic regression.

Initially, saturated models of all of these types of regression were fit and accuracy of each type of model was calculated.

Table 2

Model Type	Training Accuracy	Cross Validation Accuracy
Negative Binomial	.445	.452
Multinomial	.684	.652
Cumulative Logit	.674	.658

Moreover, this table above shows straight accuracy and although that is a good metric, we were interested in where the predictions were miscategorizing people. This misclassification was analyzed through the use of a side by side bar graph for the predictions and observed counts for each poverty level.



This Graph above shows an example for the cross validation set for a multinomial regression. Once these plots were analyzed for both the training and cross validation sets across the three different types of models, it became apparent that the negative binomial model was classifying some individuals as higher than non vulnerable. Rather than further pursuing this and attempting to correct this, the negative binomial regression was dropped since the other two models seemed more promising.

Stepwise Selection procedure

Although stepwise procedures are inherently flawed due to their internal issues in handling multicollinearity, they can be useful. The main concern with stepwise selection procedures is often that they are not good at dealing with collinearity and are prone to overfitting. Moreover, many have an issue with selecting models based off of any distance metric because too much weight is then put on that singular metric which does not encompass the "goodness" of a model. Despite this, we utilized stepwise regression mostly to cut out any unimportant nominal variables that we were not able to analyze through factor analysis. Next during this stepwise regression, we narrowed our model to a multinomial model since this performed better after the backward elimination process.

LASSO Regression

One technique that can further dimensionality reduction but also balances accuracy is LASSO regression. Least Absolute Shrinkage and Selection Operator performs both feature selection by minimizing a function similar to SSE with a squared term in account for the coefficients of each variable. By setting a threshold on how large the summation of the squared coefficients can be and then minimizing this function effectively performs dimensionality reduction by shrinking the coefficients of the variables while also maximizing accuracy.

We experimented with implementing LASSO regression in place of a stepwise selection procedure but ran into the issue of having four different sets of variables selected depending on the poverty level. When analyzing this coefficient matrix, none of the variables that were eliminated by LASSO regression was conserved across the four classes. Even implementing a cutoff value, in order to perform more aggressive dimensionality reduction, few of the coefficients deemed small were not the same across the classes.

Results

After experimenting with a few different types of models and a few different techniques our final model was a multinomial regression with a few factors cut out through backward elimination. This model gave had relatively high training accuracy (.678) compared with the other models and had the best cross validation accuracy as well (.664). This model consisted of 23 variables and incorporated effects for each education of the household, the house condition, total people in the house, total women in the house, household size, quality of utilities, age and ownership of the household, overcrowding, a few technology indicators, a rural effect, different effects for marital status, and a regional affect for the different areas in Costa Rica.

Even though this was our best model, it was only slightly more accurate than naively predicting everyone is in group 4 since 65% of our household heads were in the "non vulnerable" poverty level. Despite the predictive power of this model not being very high, this does not mean this model is not useful. This model provides insight into what has an affect on the poverty level of people in Costa Rica. Even though the information gained from this is not clear or definitive, it does not mean progress wasn't made. Furthermore, our goal in this project as to be interpretable, which was maintained. There are other ways in which more accuracy could be squeezed out of this data however, the interpretability of many of those methods may be lost.

A recurring issue that was independent of the models fit and the techniques attempted was the fact that it was difficult to differentiate between extreme poverty and moderate poverty and also between at risk from non vulnerable households. Similarly, less than 8% of the heads of the households were in extreme poverty, about 12% were at risk while approximately 66% of people in Costa Rica were non vulnerable households. Going off this, in many of the basic necessity variables, such as access to water, there were often very few observations (less than 1%) that had no access to water. Similarly, there were a multiple categories given for how people disposed of their garbage, some of which were awful such as throwing garbage in an unoccupied space, however, there were very few people in these categories. More broadly, even though there were observations in extreme poverty and that didn't have basic necessities, the overwhelmingly major of people did. Extending this, the some identifiers of the poverty level of people most likely were not contained in this dataset. For instance, an interesting part of the electricity variable is that a decent amount of people relied on the government for power. This implies that at least some of those people couldn't pay for power, and the difference between the poverty levels might not be as glaring as does one have basic necessities. On the contrary, it may be a subtle distinction such as how much is a household reliant on the government for basic necessities.

Conclusion

Overall, it appeared that the data collected provided some insight into the poverty level of people in Costa Rica. Moving forward, for future exploration we would recommend gathering more detailed data on the standards of living. For instance, further depth in what is inside the house, and the condition of what is in the houses would be helpful since it became apparent that the majority of residents had "decent" houses. As mentioned earlier, this became a common theme, that the majority of residents in our sample were not in glaring extreme poverty although there were a few. Therefore, in order to more effectively distinguish poverty level, more detailed information on standards of living, work and transportation among others could possibly provide crucial insight into distinguishing poverty levels.