Impoverishment and Aid Subsidies in the United States

By: Michael Bemus

Poverty is a near constant topic of analysis in the United States as researchers attempt to assess poverty levels and determine the factors contributing to this impoverishment. Studies such as Benson (2022) will track changes in poverty levels and other factors associated with them, such as state and urban impoverishment, the population just above the poverty threshold, and total SNAP enrollment between different years. Other studies such as Romig (2020) will assess factors causing or preventing poverty, such as social security. By better understanding the current climate of poverty, legislators and other organizations are better able to address this issue and create effective solutions to keep people out of poverty.

Poverty represents more than just an economic status. According to the US Office of Disease Prevention and Health Promotion (OASH, n.d.), the lack of resources to impoverished individuals increase the risk of developing a mental illness or a chronic disease and reduces life expectancy for that individual. Poverty is typically area-based, affecting entire communities past the individual level (OASH, n.d.). This is caused by factors often outside of an individual's control, including systemic issues such as discrimination, health issues such as a lack of sustainable housing, healthy foods, or neighborhood safety, and insufficient employment and education (OASH, n.d.). Poverty is an intrinsic representation of a nation's short-comings and how it harms individuals living within that system.

According to the Center for American Progress (CAP, 2022), 11.6% of the U.S. population is impoverished, totaling to approximately 37.9 million Americans. This is determined by the poverty line, which the Office of the Assistant Secretary for Planning and Evaluation (ASPE 2023) states is an income threshold set based upon the supposed cost of living for households of variable sizes. For most of the U.S., the poverty line for a single individual is set at $16,770 (ASPE 2023). In households with more than one person, income is determined by the total income of all residents within a household (ASPE 2023). Generally, the poverty line increases by $5,910 for each individual within a given household (ASPE 2023).

Theoretically, this measurement should account for a wide range of human needs. Mowafi and Khawaja (2005), attempt to assess the features a useful poverty line should account for. They claim a poverty line should be made using measurements such as income or expenses, a set of basic needs that must be fulfilled for an individual, an identification of a person or family for which the measure will be applied, and a reference point for an average household within the economy (Mowafi & Khawaja 2005). For the United States, the poverty line is an "absolute poverty threshold," beneath which the standard of living for a given household is below what the family needs (Mowafi & Khawaja, 2005). This is different from other nations, which use "relative poverty thresholds," such as half the median household income, to determine poverty (Mowafi & Khawaja, 2005). In these cases, the poverty threshold measures only what a "low income" would be without necessarily considering whether that income represents an inability to meet a family's needs (Mowafi & Khawaja, 2005).

Regardless of measure, Mowafi and Khawaja (2005) claim that the poverty line is an ultimately disputable measure. Especially with the household model used by most nations, it is difficult to capture disparities within a household that could mean certain individuals have standards of living that qualify as impoverished, while others in the same household might not (Mowafi & Khawaja, 2005). In addition, other factors such as non-standard needs within a family—for example, a disability or other chronic conditions—could lower a household's standard of living to that of impoverishment without their income reflecting this fact (Mowafi & Khawaja, 2005). While the poverty line is important in determining what families require extra support to sustain itself, it is hard to tell whether the poverty line accurately represents all groups that could make use of this aid (Mowafi & Khawaja, 2005).

To better analyze these disparities in standard of living, the United States Census Bureau (2023b) also collects data on supplemental poverty measures. While these are not used to estimate the official poverty line, they are useful in visualizing non-income factors that could lead a household to be impoverished (US Census Bureau, 2023b). To determine an individual's supplemental poverty state, we must first compute their Supplemental Resources (SPM), which for any given individual is computed as shown in Equation 1 (US Census Bureau, 2023b).

SR = Household Income + SNAP, Housing, School Lunch, Energy, and WIC Subsidies – (FICA + Federal Tax + State Tax) – Capped Work and Childcare Expenses – Medical Expenses

Equation 1 – Equation for an individual's supplemental resources.

According to the US Bureau of Labor Statistics (2022), an individual's SR is then compared to the supplemental poverty threshold. This varies based on an individual's housing (BLS 2022). Renting households with two adults and two children have a threshold of $34,518 (BLS 2022). Those with a mortgage have a threshold of $34,235 (BLS 2022). Lastly, those without a mortgage have a threshold of $28,909 (BLS 2022).

As we can see, the supplemental poverty threshold does consider more factors than the federal measure. However, both the federal and supplemental measures are still largely arbitrary for determining the needs of a given individual. This raises the question of whether they are accurately representing the populations they are meant to distinguish.

Poverty thresholds are important because they often help determine what individuals should be eligible for subsidies and other benefits. For example, individuals can be eligible for SNAP if their monthly income is at or below 130% of the federal poverty threshold for monthly income (Cronquist & Eiffes, 2020). They are a fundamental tool allowing federal assistance to determine who is in most need of aid.

The poverty line is not necessarily fixed. To analyze its effectiveness, we must consider not only those who fall below it, but also those whose incomes fall near that line, especially those receiving federal assistance. This paper attempts to analyze the features of this non-impoverished group that is still receiving federal aid. We are primarily interested in their relation to the impoverished and non-impoverished groups. Is this new aid group's expenses, incomes, and demographic qualities more similar to that of the impoverished or non-impoverished groups? Are there demographic qualities that set them apart?

These individuals are important to examine because they could very well benefit from the label of impoverishment. If such a label were accurate for this group, federal spending could be redistributed to assist them. However, if such a label were not accurate, it could make us question why these individuals are receiving aid in the first place.

Where most poverty studies attempt to assess the performance of a predefined group, this study instead looks at those outside of the poverty group with the goal of assigning an accurate label to them, whether that be impoverished or not. We are interested in this subject because it questions the fundamental standards with which the US poverty line operates. We want to affirm that the current poverty metrics are doing enough to address this issue, and if they are not, we want to be able to analyze and address this disparity. While this study does not encompass the creation of a new poverty line, we seek to set the groundwork for such a discussion, whether a new line is necessary or not, to help assist those who need federal aid the most.

Using data collected by the US Census Bureau (2023b), this paper attempts to use statistical methods to examine and classify those individuals above the poverty line who are still receiving aid. In the following sections, we will discuss the data used, as well as a brief mention of the statistical methods applied. Then, we will provide the results of our analysis and interpret those results. Finally, we will discuss the implications of this research and provide recommendations based on our results, with some mention of the limitations latent to our research.

1. Data Set

In this study, we utilized data from the United States Census Bureau (2023b). The data was collected as part of the 2021 American Community Survey (ACS), an annual survey used to help lawmakers decide how to distribute federal funds (US Census Bureau, 2023a). Questions on the survey include topics such as employment, education, housing, veteran status, and other factors (US Census Bureau, 2023a). These questions include demographic qualities such as age, race, and sex, tax and income, medical benefits and expenses, and subsidies received by an individual (US Census Bureau, 2023b). The version of the data set we will use focuses primarily on these latter traits.

The full dataset includes 3 million observations of response from individuals who took the ACS. Individuals were selected through random sampling households across the United States (US Census Bureau, 2023c). 3.5 million addresses are selected to receive the survey each year (US Census Bureau, 2023c). Of them, a total of 3,092,744 were included in the initial data set, measured on 47 different variables. For our final data set, we reduced this to 2,137,708 observations of 34 variables.

Table 1 lists the variables used in our analysis. Many of these have been renamed and relabeled from the initial data set for our convenience. We also restricted the domains of many variables to reduce what we saw as outliers in the data set. We felt that our data set was large

Name	Definition	Data Type
age	Participant age.	Integer (26:96)
sex	Participant sex.	Character Factor - M = Male - F = Female
race	Participant race.	Character Factor - W = White - B = Black - A = Asian - O = Other
hispanic	Whether participant is Hispanic	Binary Factor - 1 = Hispanic - 0 = Not Hispanic
edu	Education attainment	Character Factor - <HS = Below High School Degree - HS = High School Degree - <C = Some College - C = College Degree
mar	Participant marital status.	Character Factor - M = Married - W = Widowed - D = Divorced - S = Separated - NM = Never Married
spm_pov_thres	Participant's poverty threshold by Supplemental Poverty Measures.	Integer (10,312:136,352)
spm_res	Resources for participant. Calculated using income, subsidies, taxes, and expenses	Decimal (-43,420:653,628)
spm_inc	Participant's income.	Decimal (0:752,000)
agi	Adjusted Gross Income	Decimal (-6,300:506,000)
fed_tax	Participant's total federal tax.	Decimal (-74,952 :99,993)
fed_tax_bc	Participant's federal tax before applying tax credits.	Decimal (0:116,102)
st_tax	Participant's state tax.	Decimal (-5,577:41,558)
eitc	Earned Income Tax Credit (EITC)	Decimal (0:19,059)
fica	Federal Insurance Contribution Act and federal retirement contributions	Decimal (0:39,824)
hi_prem	Health insurance premiums excluding Medicare Part B.	Decimal (0:28,800)
spm_hi_prem	Health insurance premiums.	Decimal (0:29,900)
med_xpen	Medical Out of Pocket expenses and Medicare Part B subsidy.	Decimal (0:49,982)
moop	Medical out of pocket expenses	Decimal (0:19,750)
mc_pb	Medicare Part B premium	Decimal (0:5,702)
cap_xpen	Capped work and child care expenses.	Decimal (0:19,889)
work_xpen	Total work expenses	Decimal (0:12,485)
cc_xpen	Childcare expenses	Decimal (0:9,999)
mortgage	Participant's tenure/mortgage status	Character Factor: - M = Has Mortgage - N = No Mortgage - R = Renting
num_adlt	Number of adults in household	Integer (1:19)
num_kid	Number of children in household	Integer (0:16)
cohabit	Whether household has cohabiting couples	Binary Factor: - 1 = Has Cohabiting Couple - 0 = No cohabiting Couple
ui_kids	Whether household has children not related to individual.	Binary Factor: - 1 = Has unrelated individual - 0 = No unrelated individual
snap	Participant's Supplemental Nutrition Assistance Program (SNAP) subsidy.	Integer (0:19,200)
slunch	Participant's school lunch subsidy	Integer (0:9,963)
energy	Participant's energy subsidy	Integer (0:6,908)
wic	Individual's Women, Infants, and Children (WIC) subsidy.	Decimal (0:2,483)
house_sub	Participant's housing subsidy.	Decimal (0:29,985)
group	Poverty and aid status.	Character Factor - poor = Impoverished - aid = Receiving aid, not poor - no = Not poor and no aid

Table 1 – Data dictionary of variables used in study.

enough that we could justify removing many outliers by condition. Table 2 contains the ranges of outliers removed from those variables which were selected. Notably, we chose to remove any observations under the age of 25. We believed that their income, aid, and poverty status would be the most likely to be determined by others in their household, namely their parents. Therefore, we did not believe they would be useful for our analysis.

Variable	**Outlier Criterion (x)**
snap	$x > 20{,}000$
house_sub	$x > 30{,}000$
slunch	$x > 10{,}000$
energy	$x > 7{,}000$
wic	$x > 2{,}500$
work_xpen	$x > 12{,}500$
med_xpen	$x > 50{,}000$
cc_xpen	$x > 10{,}000$
moop	$x > 20{,}000$
spm_hi_prem	$x > 30{,}000$
hi_prem	$x > 30{,}000$
fed_tax	$x > 100{,}000$
st_tax	$x > 70{,}000$
eitc	$x > 75{,}000$
fica	$x > 40{,}000$
age	$x < 25$
spm_inc	$0 \leq x < 1{,}000{,}000$

Table 2 – Outlier criterion for data processing.

The initial data set was very clean. We did not have to do much with data preprocessing apart from the relabeling decisions we made for many of our categorical variables. We did not need to remove any missing observations. However, for some of our models, we had to recode many of our categorical variables as binary dummy variables to use as numeric. In these cases, our baselines are those labels underlined in Table 1.

Of the variables in our data set, we created one new variable: group. This is our primary variable of interest. To create it, we combined those individuals identified as impoverished by either the federal or supplemental measure into one class, "poor." We then assigned those individuals with snap, slunch, house_sub, energy, or wic greater than 0 who were not identified as poor to the "aid" class. All other individuals were identified by the "no" class. In the total data set, 261,978 observations were of the class "poor," 427,112 observations were of class "aid," and the remaining 1,448,618 observations were of class "no." We set the baseline class for this variable to be "aid," which will have some effect in our later methods.

For our modelling step, we split our data into training, testing, and validation sets. We created representative samples of our group variable, ensuring the proportions of each class would reflect their proportions in the total population. We created a 50-30-20 split between train, test, and validation sets. Our training set had 1,068,854 observations, our test set had 641,312 observations, and our validation set had 427,543 observations.

To use this data, we treated our problem as a classification task with our group variable as the response. Using the remaining variables in our data set, we attempted to determine if there were significant differences between these three groups, specifically if we could separate the aid group from the poor and no groups. If we could not separate the aid group, we would then examine which of the other two classes it tended to be more similar to and misclassified as more frequently. To avoid biasing our results, because snap, slunch, energy, wic, and house_sub were used to define the aid group, these variables were not used in our analysis.

2. Methods

In order to address this problem, we applied three different methods to test the similarity between our groups: principal components, discriminant analysis, and logistic regression. The first is an unsupervised learning method of variance analysis typically used to simplify the number of variables in a data set. The latter two are supervised classification methods that attempt to predict the class of observations based on the features in the data.

We selected these methods because of their interpretability. Though we are generally approaching this as a classification task, we want to be able to determine what if anything is separating our variables. Principal components analysis allows us to examine which variables are most important within our classes as well as generate metrics of overall similarity between vectors. Discriminant analysis creates linear separators based on our variables to determine its classes. And logistic regression determines a probability approach which allows us to determine the odds of an individual falling into a group, as well as the risk factors associated with it. Together, these methods should give us a multifaceted lens through which we can assess the validity of our new aid group.

In the following sections, we will provide a basic summary of each methods, how it was used in our analysis, and what assumptions are implied by each method.

A. Principal Components Analysis

As we stated above, principal components analysis is an unsupervised method of examining the variance in a data set. It uses the eigenvalues and eigenvectors of a distribution's covariance or correlation matrix to create linear combinations of our initial variables which explain the most variability of the data (Richard & Wichern, 2007). These linear combinations, or components, are determined by the eigenvectors of the chosen matrix, and each component explains an amount of the variance equal to its eigenvalue (Richard & Wichern, 2007). Principal components will be orthogonal to each other, and the sum of their eigenvalues will equal the total variance of their corresponding covariance matrix (Richard & Wichern, 2007). This method will create a number of principal components equal to the number of variables in the data used to create them (Richard & Wichern, 2007).

As part of principal components analysis, the method assumes that our data is normally distributed when creating the covariance or correlation matrix. In practice, the distributions of most of our variables do not seem normal, especially within our categorical variables which are not continuous. However, we do have a very large sample size, so by the Central Limit Theorem, the distributions of our variables should be converging toward normality.

For our purposes, we created principal components based on the correlation matrix. While many of our numeric variables are measured in dollars, the ranges of these are often very different, so by using the correlation matrix, we standardize our variables and avoid these different scales from altering our results. We also had to convert our categorical variables to dummy variables as described above. Principal components can only assess the variance of numeric quantities. After making these transformations, we had a set of 36 variables not including the group variable.

We created four sets of principal components based on the training set of our data. The first set we created used the full data set to determine its components excluding the group variable. The second used only those observations in the aid group. The third was created using the poor group. The last was based upon the no group.

After creating these components, we needed to measure the similarity between them. We took two general methods to assess this similarity. First, we looked at those variables with absolute coefficients greater than $\frac{1}{\sqrt{k}}$, where k is our number of variables. This represents a significance threshold for each variable, and for each component, any variables with the absolute value of their coefficient greater than this threshold are said to be contributing significantly. Typically, this is used to determine what combination of variables are generally defining any given component.

To use these significant variables, we counted matches between corresponding components of our four component sets. We used two methods of doing this. First, we counted exact matches. Under this method, variables are ordered from greatest to least by their coefficient. We then compared these rankings and counted how often the variables in these positions matched. Second, we counted component-wise matches. In this case, we simply looked to see if the same variables were contributing to corresponding components and counted those matches as well. Regardless of method, we then divided our match counts by the average number of significant variables between the two component sets being matched. In our results section, we will refer to these metrics as Perfect Match (PM) and All Match (AM) respectively.

Our second method of comparing components was using a cosine similarity. We took three approaches here. The first used basic cosine similarity, whereas the second used the cosine similarity of the absolute values of each component. With the second approach, we wanted to look at total significance for each variable irrespective of sign. We then took the means of all cosine similarities between any two groups for these first two methods. In our results, these metrics will be labeled CS for cosine similarity and aCS for absolute cosine similarity.

Our third cosine method utilized percent variable of each component as a weighting factor. With each cosine similarity, $\theta_{i(jk)}$ for absolute value component i computed between models j and k, we created weighted similarities, Φ_{ij} and Φ_{ik}, that were multiplied by the percent variance of models j and k respectively for component i. If j and k are a perfect match, $\theta_{i(jk)}$ will be 1 for all i, so the sum of these percents, Φ_{ij} and Φ_{ik}, would each be 1. To account for slight differences in variable importance, we then took the average of the sums of Φ_{ij} and Φ_{ik} to use as our final metric, wCS.

As we can see, principal components provides us with a way of simplifying and comparing our variables among groups, which gave us preliminary estimates of how similar our distributions may be based upon their variability. To further explore this, we then moved on to our classification techniques, which were better for describing how the distributions between groups were different.

B. Discriminant Analysis

Discriminant analysis is our first classification technique, and it attempts to create decision boundaries dividing regions that can be used to predict the class of any given observation. Using an assigned cost ratio and a set of prior probabilities, discriminant analysis assigns variables based on the distributions of the different classes (Johnson & Wichern, 2007). Given g groups with prior probabilities p_i for class i, cost allocations $c(k|i)$ for the cost of misclassifying an observation of class k to class i, and distributions $f_i(x)$ class i, observations are assigned to the class for which Equation 2 is maximized (Johnson & Wichern, 2007).

$$\boxed{\sum_{i=1, i\neq k}^{g} p_i f_i(x) c(k|i)}$$

Equation 2 – Maximization criteria for discriminant analysis.

From this maximization procedure, decision rules can be created for our g groups that can be applied to any observation x_i. However, depending on the distributions of our groups, $f_i(x)$, it can be difficult to create reasonable boundaries between classes. Because of this, it is typically assumed that $f_i(x)$ is multivariate normal (Johnson & Wichern, 2007). By making this assumption, our model becomes either a linear discriminant analysis approach or a quadratic discriminant analysis approach. In the linear case, it is further assumed that the distributions of our groups share a common covariance matrix (Johnson & Wichern, 2007). In the quadratic case, this assumption is not made (Johnson & Wichern, 2007).

For our purposes, we will again rely on the Central Limit Theorem to satisfy normality. However, based upon our exploratory analysis, we do not believe we have equal covariance matrices. Despite this fact, we still tested both a linear and a quadratic discriminant analysis model to examine our groups. Though we have not met our assumptions, the linear model is more interpretable than the quadratic, so though our results will not be perfectly accurate, what we do find through our analysis can still help describe how our distributions are different.

To create our two models, we used all variables excluding those subsidy variables used to divide out aid. We first created models on the training set, then tested their performance on the test and validation sets. Next, we created models on the testing set to compare to the training model. In doing this, we planned to determine if the patterns we found in the first models were true across our distribution, even on unseen data. We also tested these testing models' performances on the training and testing sets. We did not apply them to the validation set because we did not intend to compare the testing models to our non-discriminant methods.

Our discriminant analysis models are useful for regional interpretation of our variables. The linear case helps provide thresholds between variables of our model which are easy to interpret. Our ability to classify will also help reflect whether the groups are different. By using equal costs for misclassification, we allow the model to naturally show which class the aid group is more similar to between the poor and no groups. We will use misclassification rates to determine our ability to separate our classes.

C. Logistic Regression

Logistic regression is our second classification method. Rather than approach the task through regional division, logistic regression approaches the task through odds ratios. In multinomial logistic regression, we create simultaneous Bernoulli models using the logit function given in Equation 3. Using numeric methods, we maximize the likelihood of our *g*-1 logistic models, where *g* serves as our number of classes. Using these simultaneous models, we can then compute probabilities for each class and assign our final result to the class of highest probability.

$$\boxed{\pi_{ij} = \frac{e^{\beta_{0,i}+\beta_{1,i}x_1+\ldots+\beta_{k,i}x_k}}{1+e^{\beta_{0,i}+\beta_{1,i}x_1+\ldots+\beta_{k,i}x_k}}}$$

Equation 3 – Logit function for logistic regression.

An advantage of Equation 3 is its relation to the odds ratio. Dividing by $1-\pi_{ij}$, we get a product of exponentials of the form $e^{\beta_{0,i}}e^{\beta_{1,i}x}\ldots e^{\beta_{k,i}x}$. This means that our coefficients act as exponential multipliers to the odds of an observation being in any given class. This gives us direct comparisons in pairing *i* between the probabilities of class *j* and the baseline. It is this interpretation that we will utilize to determine how our aid group is different from the poor and no groups.

Logistic regression assumes a Bernoulli probability distribution between our groups. In this analysis, we have not tested whether that assumption is accurate. We plan to test the relationship between our variables under different forms. Though this may not be the best fit for our model, we can still interpret the results of the model to see whether we can separate our classes and where failings might occur.

In application, we attempted to determine a few models to optimize our results. Along with the full model, we attempted to use forward and backward subset selection, minimizing the AIC of the generated model. Through forward selection, we got a univariate model with spm_inc as the only predictor. Using backward selection, we found the full model minimized AIC within one step of selection. We then validated these two models on the train, test, and validation sets to compare to our discriminant models. We also formulated models on the test set that we validated on the train and test sets to verify that our findings from the training set were valid.

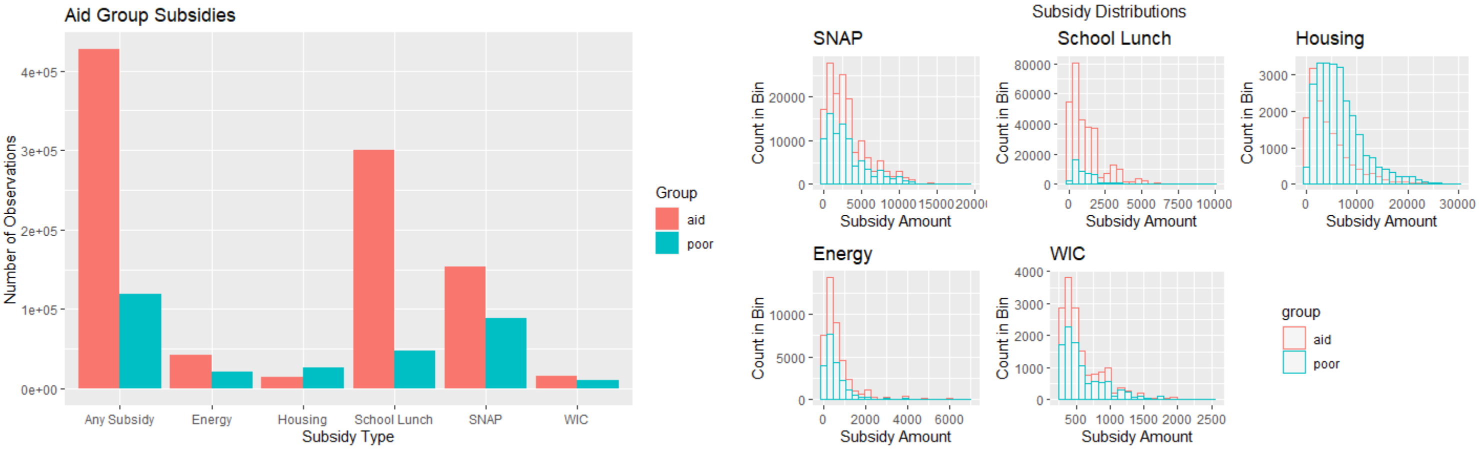


Figure 1 – (left) Counts of individuals receiving subsidies, divided by group. (right) Distributions of each subsidy variable between groups, ignoring zeroes.

3. Results

Before looking at the results of our various methods, we first consider our five subsidy variables removed from our analysis: snap, slunch, energy, wic, and house_sub. These variables define our aid group, so they are essential to understanding the relationship between it and the poverty group. Figure 1 provides plots of both the number of observations recorded as receiving each subsidy by group and the distributions of each variable by class.

As shown on the left of Figure 1, we see that we have a much greater count of individuals in the aid group who receive subsidies than those in the poor group. In the whole data set, we have about 1.63 times as many observations in the aid group as we have in the poor group. However, here we see that only 119,271 individuals in the poor group who are receiving any subsidies. That is less than half the number observed in our data. We see that about 3.58 times as many individuals who are not impoverished are receiving subsidies compared to those impoverished individuals who are receiving subsidies. The number of individuals in the aid group is about 2.99 times the number of individuals not receiving subsidies in the poor group.

We do also see that these subsidies are being distributed for different needs. In Figure 1, we see that most subsidies received by the aid group are from school lunch subsidies. As shown on the right, these do not represent as much benefit as snap or housing subsidies. The mean subsidy for slunch across all groups is approximately $1,196.71 excluding 0 observations. That is compared to snap, with zero-excluded mean around $3,124.31, and house_sub, with a non-zero mean around $5,920.84. So while we see that raw counts of individual in the aid group are receiving more subsidies, the subsidies received by those in the poor group are more essential to their needs.

The results in Figure 1 may suggest that slunch is not appropriate for defining our aid variable. However, we chose to continue utilizing it due to its role in the calculation of the supplemental poverty threshold. Under such guidelines, school lunch is an important consideration for determining an individual’s resources.

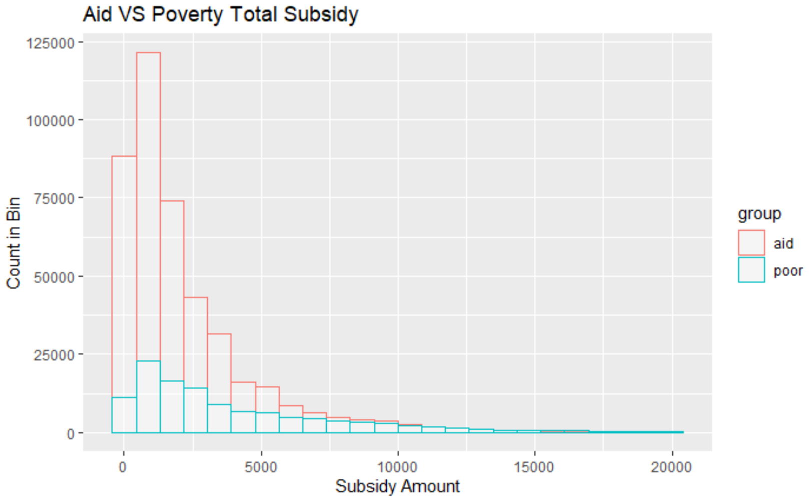


Figure 2 – Distribution of total subsidies between aid and poor group.

To further explore the relationship of our subsidies between the aid and poor groups, Figure 2 shows the distribution of total subsidies between the two groups, ignoring zero observations. We see that below $2,000, we have a much greater number of individuals who are earning subsidies, especially in the aid group. Past $2,000, we see a sharper drop in observations from the aid group compared to the drop in those from the poverty group, though this could be a matter of scale in our data set. By $5,000, we see that levels have reached approximately what we would expect given the sizes of the two populations. The number of observations in our upper bins are much closer between the groups. After $10,000, we start to see that those observations labeled as poor are earning these higher subsidy values more frequently those individuals in the aid group. Ideally, this would be desirable in the targeting of these subsidies.

What Figure 2 might contain in the left end of the distribution are all those observations which are earning slunch and other smaller subsidies in the aid group. On the right side, we see the larger subsidies such as house_sub play a greater role in the total distribution, and we know that those individuals labeled as poor more frequently receive housing subsidies than those who are not. This context helps us understand the difference between the aid and poor groups based on their subsidies earned. However, to understand our third class, the no group, we needed to use other modeling techniques to better summarize the relationships of our variables.

The first procedure we tested was principal components analysis. Figure 3 summarizes the rate at which variance was explained by each procedure we tested. As we can see, our corresponding components between groups explain approximately the same percent of the variance at every step. The most noticeable difference is found in the components for the poor group. Initially, their first component explains about 0.75 less of its group's variance than the other first components do. However, we then see that its next two components explain more of the variance to counterbalance this. Generally, our components seem to be explaining the variance at the same rate between models. There are slight fluctuations here and there depending on the model, but these do not seem significant overall.

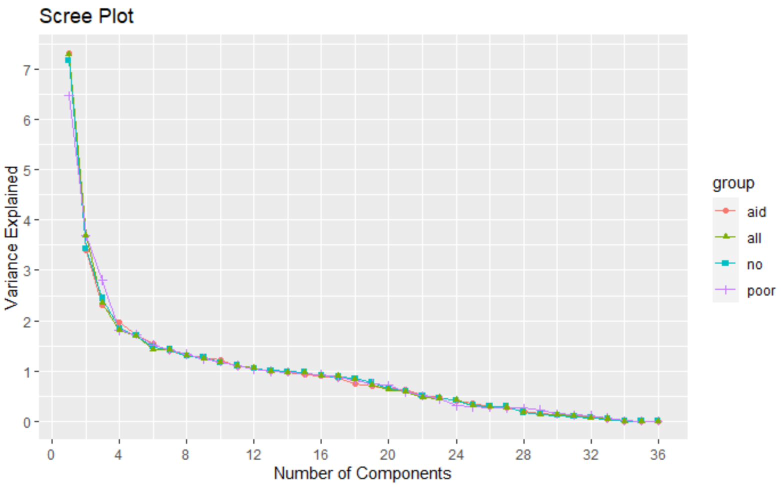


Figure 3 – Variance explained by each principle component.

For all principle components models, we find that 16 components are required to explain 80% of the distribution's variance. After the first 24 components, the remaining 12 explain less than 1% of the variance each for all models.

Variable	Full	Aid	Poor	No
cap_xpen	4	5	2	3
spm_pov_thres	4	3	3	4
spm_res	1	2	2	2
hi_prem	6	5	6	5
moop	4	4	3	5
mc_pb	6	6	4	6
agi	3	1	4	2
age	4	4	3	3
hispanic	3	3	1	3
num_adlt	5	5	6	5
num_kid	5	7	6	6
spm_inc	1	1	2	2
fica	1	1	2	1
spm_hi_prem	3	3	5	3
fed_tax	2	3	2	2
fed_tax_bc	2	3	3	2
st_tax	1	1	2	1
work_xpen	3	2	3	3
eitc	4	3	2	4
cohabit	5	5	3	3
cc_xpen	5	3	7	7
sex	6	6	6	6
ui_kids	5	5	5	5
med_xpen	2	1	2	2
divorced	7	6	5	9
married	4	2	3	4
separated	4	4	4	4
widowed	5	8	6	7
has_mortgage	5	4	4	4
renting	5	5	3	5
less_college	2	3	4	4
college	5	4	5	6
highschool	5	5	3	3
race_asian	4	6	9	6
race_black	6	6	8	5
race_other	3	3	1	4

Table 3 – Component significance counts with first 16 components.

Due to the large number of components for explanation and variables in our data set, rather than looking at individual components, we will look primarily at the number of times each variable was considered significant within the first 16 components to summarize our models. Table 3 provides these counts across models. In interpreting these, the more important variables are those which have fewer observations as significant. Those with few observations explain most of their variability in a single component compared to those with many observations that explain their variability across many components. That means that variables observed infrequently hold more individual weight within the total distribution.

We see that most of our categorical variables do not seem to hold much weight in explaining the variance of our model. This is primarily a product of our principal components. With only a handful of possible categories, these variables have little variance, so they are less able to explain the numeric trends within our data. It appears that the most important categorical variables in these models are hispanic, race_other, and less_college. This is importance in respect to our baselines: non-Hispanic, white, and less than high school education. This may suggest there is larger variability between these groups compared to differences in our other groups.

Looking at our continuous variables, we see that our income variables are generally more important. spm_res, spm_inc, and fica are all seen only one or two times within our components. In fact, if we look at our first components across models, we see spm_res and fica are significant in all four, and spm_inc is significant in all but the poor group's components. spm_inc is the largest contributing variable in the three components that use it in the first. spm_res and fica are then second and third, with the aid and full components placing spm_res second, and no placing fica second instead. In the poor group, the first two components are spm_res, then fica.

Among our other variables, state tax is rather important and used in the first components of the full, aid, and no component sets. It is used in the second and third components of the poor group. However, we also see that our other tax variables, such as fed_tax and fed_tax_bc are somewhat important. These two variables are used frequently within the first two components across models.

Looking at expenses, we see that med_xpen followed by work_xpen contribute the most within our starting components. med_xpen is the most important variable to the third component across all models. work_xpen appears in the first component of all models and in the second components of the full, aid, and no component sets.

Interestingly, cc_xpen is much more important to the aid group than our other classes. This could reflect the school lunch subsidies received by individuals in that group. However, at the same time, num_kids is contributing the least to the aid components, which could suggest there is little variance in the number of children these individuals have. Also of not about the aid components, agi contributes the most within that group, which suggests they benefit the most from income adjustments across groups.

In this preliminary analysis, it appears that the poor group is the most different among our variables in terms of significantly contributing variables. While we are seeing differences in the aid group compared to the others, the aid group appears more similar to the full and no groups than it does to the poor group.

Match	PM	AM	CS	aCS	wCS
Aid-No	0.14634	0.54409	-0.12623	0.72648	0.78652
Aid-Poor	0.06130	0.35632	0.02001	0.55025	0.70013
No-Poor	0.06654	0.38817	-0.02923	0.58295	0.75263
Aid-Full	0.21033	0.58891	-0.09884	0.79099	0.83420
No-Full	0.25092	0.70111	0.35528	0.86687	0.91591
Poor-Full	0.14634	0.54409	-0.08336	0.59498	0.76297

Table 4 – Similarity scores for principal components. (metrics from left to right: perfect component matches, any component matches, cosine similarity, absolute cosine similarity, weighted cosine similarity)

To explore this similarity further, Table 4 provides the results of our metrics applied to all of our principal components for each model pairing. Looking first at our matching metrics, we see that the most frequent matches occur between the no and full component sets. This was to be expected. The no group contains a majority of our data compared to the other groups, so it is reasonable that its components are similar to those generated by the full data set.

The second most perfect matches occurred between the aid and full group. This is again reasonable because the full group contains the data from the aid group. However, third, we see a tie between matches from the aid group with the no group and matches from the poor group and the full group. Immediately, this suggests that the poor group is the most different of our three classes since it is most different to the total distribution. The fact that it has the same number of perfect matches with components made from data containing it as the components of two disparate populations, the aid and no groups, suggests that the current poverty metrics are successfully separating a significant subset of surveyed individuals.

Interestingly, we find that the components of the poor group sees more perfect matches with the no group than it does with the aid group. We did not expect this result in our analysis. It could suggest that the different importance related to childcare variables for the aid group is significant when comparing it to the poor group, causing the relationship seen in Table 4.

We see that our AM metric provides the same general results as PM, though it does act on a different scale. We see that about 54% of the time, the significant variables in a component from the aid group matches the significant variables in the matching component of the no group. This rate is about 36% between the aid and poor group, versus a rate of approximately 39% between the no and poor group. This continues to suggest that the aid group is more similar to the no group than it is to the poor group.

Looking at our cosine metrics in Table 4, we first see that the scores for CS are generally low except in the case comparing the components of the no group with those of the full group. Because we did not take the absolute value of our similarities while computing our average, it caused issues with conflicting signs that caused cancellations in values. Therefore, we will prefer the absolute component similarity average for our analysis. Sign changes do not necessarily reflect the magnitude of importance of a given variable in a principal component. We seek to compare significance, not specific value, making this absolute value preferable for our work.

We see very similar trends in aCS as we did with our matching methods, though the metric once again represents an increase in similarity scale. This is at least in part due to the fact that our cosine similarity algorithm could see the full component vector rather than just those variables significant to it. However, we do see that our aCS has much lower similarity values for the poor group compared to the other matchings than our other methods. It is still least similar to the aid group, though this is now above a 50% cosine similarity.

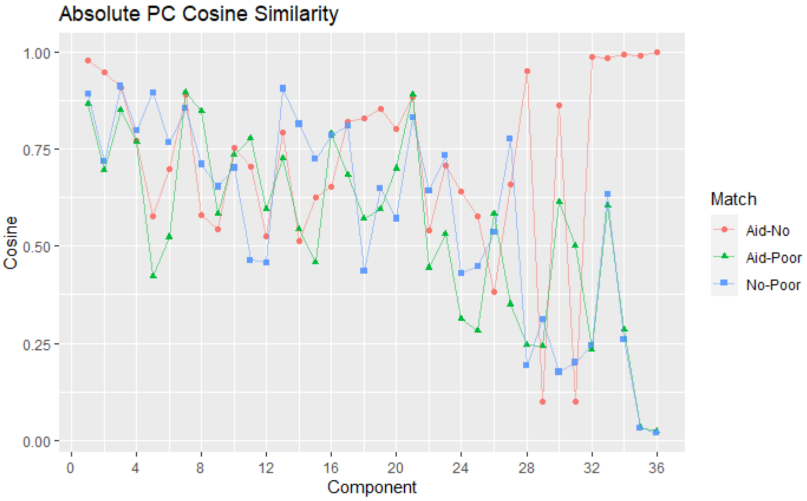


Figure 4 – aCS scores by component.

To get a better understanding of our aCS, Figure 4 shows the similarity values scored at each component for our groups of interest. We see that the similarity between the aid group and the no group is often larger than that similarity between the aid and poor groups. The aid and poor groups are occasionally more similar within small ranges of components, but the aid with no group comparison is often not too far below the measure presented between the aid and poor groups. We also see that from 4 to 7 and 12 to 16 components, the no and poor groups are most similar of our groups. These spans are generally longer than those lengths where aid and poor are the most similar.

Within the first three components of Figure 4, we see that the aid and no group are very similar. The poor group is similar to the other two in the first and third components, but not the second. Examining our second component, we see that the no and aid groups feature some of our more important categorical variables such as hispanic and race_other. They also seem concerned with expense variables and some amount of taxes. In comparison, the poor group relies almost exclusively on income and tax variables. The second component is not very clean for interpretation, but it appears that the poor group relies on those variables it did not explore in the first component that many of the other distributions had already used.

Near the end of our distribution in Figure 4, we see that the aid and no group are almost perfect matches. At components 28, 30, and all those after 32, their similarities are above 70%.

Especially within the last few components, this is much greater than the similarities of our other two matchings. However, these matches should not be as important to our general scoring as our other components because they contribute less to the variance.

To account for variance explanation, we adopted wCS, which has its results shown in Table 4. We see that when we account for component significance as well as variable similarity, the poor group is much more similar to the other two. Under wCS, the aid group is about 70% similar to the poor group, compared to 79% with the no group. While we are still seeing the aid group is more similar to the no group, this difference is much smaller under wCS. Our scores are bunched much closer together on the wCS scale than with our other metrics.

Generally, it appears that the aid group is more similar to the no group than it is to the poor group by examining their principal components. To further explore what differences are present in these groups, we move onto our classification methods, starting with discriminant analysis.

Variable	Train “No” Coeffs	Train “Poor” Coeffs	Test “No” Coeffs	Test “Poor” Coeffs
cap_xpen	$-2.632770e^{-4}$	$-1.586559e^{-4}$	$-2.527927e^{-4}$	$-1.471402e^{-4}$
spm_pov_thres	$-2.801912e^{-5}$	$1.465557e^{-5}$	$-2.824271e^{-5}$	$1.306376e^{-5}$
spm_res	$-1.992231e^{-4}$	$-1.489317e^{-4}$	$-1.987443e^{-4}$	$-1.506920e^{-4}$
hi_prem	$1.147004e^{-5}$	$-1.205710e^{-5}$	$1.220928e^{-5}$	$-1.287478e^{-5}$
moop	$3.651030e^{-6}$	$-5.157829e^{-6}$	$2.787011e^{-6}$	$-6.812311e^{-6}$
mc_pb	$3.748747e^{-4}$	$-1.031465e^{-4}$	$3.710107e^{-4}$	$-1.052106e^{-4}$
agi	$1.670870e^{-6}$	$-1.690801e^{-6}$	$1.697887e^{-6}$	$-1.633183e^{-6}$
age	$-7.751629e^{-3}$	$3.678823e^{-3}$	$-7.342002e^{-3}$	$3.937391e^{-3}$
hispanic	$2.269818e^{-2}$	$-6.941076e^{-2}$	$1.334307e^{-2}$	$-8.452688e^{-2}$
num_adlt	$-4.045802e^{-1}$	$-4.083453e^{-2}$	$-4.053110e^{-1}$	$-3.288481e^{-2}$
num_kid	$-6.668125e^{-1}$	-1.022959	$-6.657006e^{-1}$	-1.018423
spm_inc	$2.380847e^{-4}$	$1.193028e^{-4}$	$2.373710e^{-4}$	$1.212142e^{-4}$
fica	$-2.054864e^{-4}$	$-1.329545e^{-4}$	$-2.059809e^{-4}$	$-1.343780e^{-4}$
spm_hi_prem	$4.008587e^{-6}$	$-3.441157e^{-6}$	$1.962400e^{-6}$	$-3.727175e^{-6}$
fed_tax	$-2.809661e^{-4}$	$-2.045309e^{-4}$	$-2.816061e^{-4}$	$-2.069395e^{-4}$
fed_tax_bc	$-2.809661e^{-4}$	$1.950299e^{-4}$	$-8.086609e^{-5}$	$1.949720e^{-4}$
st_tax	$-1.970527e^{-4}$	$-1.489956e^{-4}$	$-1.960134e^{-4}$	$-1.505739e^{-4}$
work_xpen	$1.568886e^{-4}$	$-1.277164e^{-4}$	$1.522860e^{-4}$	$-1.399925e^{-4}$
eitc	$-1.744272e^{-4}$	$1.006995e^{-4}$	$-1.753185e^{-4}$	$1.087691e^{-4}$
cohabit	$-3.826973e^{-1}$	$5.245452e^{-1}$	$-4.090130e^{-1}$	$5.442550e^{-1}$
cc_xpen	$4.586532e^{-5}$	$1.422255e^{-5}$	$3.525193e^{-5}$	$8.966756e^{-7}$
sex	$-3.854166e^{-2}$	$2.914802e^{-2}$	$-3.717016e^{-2}$	$2.281351e^{-2}$
ui_kids	$-1.048669e^{-1}$	$-2.502076e^{-1}$	$-5.455864e^{-2}$	$-2.058667e^{-1}$
med_xpen	$-2.242616e^{-4}$	$-1.210291e^{-4}$	$-2.236566e^{-4}$	$-1.214316e^{-4}$
divorced	$-3.996551e^{-2}$	$-2.548591e^{-1}$	$-5.406888e^{-2}$	$-2.662065e^{-1}$
married	$5.702616e^{-2}$	$1.394907e^{-1}$	$4.640067e^{-2}$	$1.235831e^{-1}$
separated	$-1.765388e^{-1}$	$-4.977366e^{-2}$	$-1.813979e^{-1}$	$-4.287091e^{-2}$
widowed	$-3.621635e^{-3}$	$-2.146441e^{-1}$	$-1.100689e^{-3}$	$-2.268019e^{-1}$
has_mortgage	$1.197231e^{-2}$	$-7.637666e^{-2}$	$5.320256e^{-3}$	$-7.385760e^{-2}$
renting	$-7.916929e^{-2}$	$-6.962358e^{-2}$	$-8.664342e^{-2}$	$-7.278135e^{-2}$
less_college	$1.791166e^{-1}$	$-3.439952e^{-2}$	$1.949798e^{-1}$	$-6.926111e^{-2}$
college	$2.241659e^{-1}$	$1.075574e^{-1}$	$2.342119e^{-1}$	$8.206025e^{-2}$
highschool	$1.599717e^{-1}$	$-5.102250e^{-2}$	$1.749760e^{-1}$	$-7.075147e^{-2}$
race_asian	$-2.212521e^{-2}$	$-5.552081e^{-3}$	$-4.796911e^{-2}$	$-3.637270e^{-3}$
race_black	$-1.447209e^{-1}$	$-1.077481e^{-1}$	$-1.481489e^{-1}$	$-1.042565e^{-1}$
race_other	$-4.661825e^{-2}$	$-1.473822e^{-2}$	$-3.388270e^{-2}$	$-2.024365e^{-3}$

Table 5 – Coefficients of Linear Discrimination

First of our discriminant methods, we will look at linear discrimination (LD). Table 5 summarizes the coefficients determined by our LD model. All of these coefficients are determined relative to the aid group, so the coefficients represent how the aid group is different from the given group of the coefficients. We created two LD models, one on our training set and one on our testing set. We did this to validate whether our assumptions were met of normality and equal covariance between groups.

First looking at the difference between the train and test models, we see that our coefficients are rather similar. We do not see any sign changes between comparison. None of the estimates are perfect matches, but they are generally very close. The most difference we see are in our categorical variables. It seems that the significance of those are variable within our data set. Linear discrimination is not designed to handle categorical variables, so this is likely contributing to the issues seen here. For our purposes, we will comment only on the sign of any categorical variable in our analysis, not relying upon its magnitude.

Examining the coefficients given in Table 5, we see that some of our coefficients seem unreasonable. First, we see that our model expects the aid group to have less income than both of the other groups. This may be reasonable in the case of the no group because we would assume that those who receive subsidies have low enough income to need it. However, it would not make sense for the aid group to have income lower than the poor group. Income is a major determinant of poverty, so we would expect the poverty group to have the lowest income.

Looking at some of our other economic variables, we see that the aid group is predicted to have fewer childcare expenses than the other groups. This seems strange because the model is also predicting the aid group to have more children. The latter observation is supported by our exploratory analysis, as shown in Appendix 1. However, we did see in our exploratory analysis that most observations had childcare expenses near zero. This feature of the data does not seem important for most observations in our groups.

For our other expenses, Table 5 shows that the aid group is expected to have fewer medical. From our exploratory analysis, we know that the aid group is a younger population, as shown in Appendix 2, it would follow that their medical expenses, at least, might be lower. They would not require many of the treatments and surgeries often associated with old age. However, in Table 5, our model does not actually predict the aid group as being the youngest group. While it implies the aid group is younger than the no group, it also predicts the aid group will be younger than the poor group. The effect comparing the aid and poor group is about half the effect between the aid and no groups, so we would expect this age gap to be smaller. Still, by our exploratory analysis, this result seems somewhat unreasonable.

The LD model predicts the aid group will have work expenses between the no and poor groups, with the no group having the greatest work_xpen and the poor group having the least. The model also implies that the aid group will have the greatest total expense, with negative coefficients related to both discriminators for cap_xpen. This seems strange because cap_xpen is a sum of work_xpen and cc_xpen, and the aid group is not expected to have the of those two expenses among our groups. This coefficient could interact a multicollinearity issue present in our model.

In terms of taxes, the model appears to predict the aid group will have the greatest taxes. All coefficients are negative for both st_tax and fed_tax. Interestingly, the model predicts the poor group will have the greatest fed_tax_bc, which would imply their income is the most taxable. However, as we have already mentioned, the aid group is expected to have the greatest income. This interaction is likely the effect of multicollinearity in our model and may not be representative of the actual data.

Examining some of our categorical variables, we see that the LD model predicts the aid group to be the most diverse group. For all of race_asian, race_black, and race_other, our coefficients in Table 5 are negative. In Appendix 3, we see that this comparison may be reasonable relative to the no group. That group had the largest percent of individuals identifying as white in the total population. However, we see that our distributions are more balanced between the poor and aid groups. Though our coefficients are smaller in the latter comparison, we might expect them to be closer to zero in the absence of other variables.

Another strange observation of our model is that it predicts the aid group as least likely to be married. Table 5 shows that predicting the poor group has the greatest positive effect from the marriage variable. This result does not appear to match up with our exploratory analysis, as shown in Appendix 4. There, we found that the poor group was generally least likely to be currently married. Both the aid group and no group had majorities of observations that were married. In comparison, while those that were married were the largest of the relevant classes within the poor group, it was almost slightly greater than our baseline, which is never married.

Generally, the results of our multi-class variables are frequently different from what we would expect. However, linear discriminant analysis is not explicitly meant to handle categorical variables. Our results for these variables may not be as valid as those for our continuous variables.

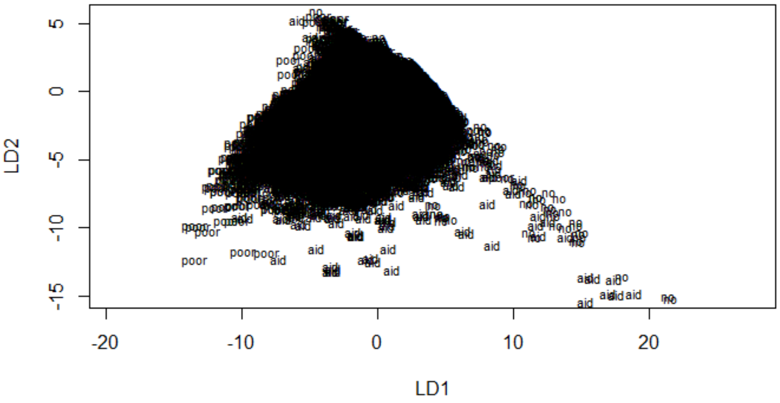


Figure 5 – Linear Discriminants on training set.

To look at the overall performance of our LD model, Figure 5 provides the distribution of the training set according to our two discriminants. We can quite clearly se that there is not much separation being done by this model. Our data is largely clouded between -10 and 5 on our first discriminant and between 5 and -10 on our second. We seem to be having observations of poor within the right side of our distribution, and asome aid observations seem to be trending toward the bottom. Interestingly, our distribution appears to have a tail or stem that is a cluster of no and aid observations. We can assume that the no group should be separated to the top right of this plot.

Along with our LD model, we also computed quadratic discriminant (QD) models from our training and testing sets. We are not able to provide coefficients for these models. However, the training and testing models appeared to have similar accuracy on the train and test sets, so we believe they were performing similarly.

LD		Predicted		
		Aid	No	Poor
Observed	Aid	38142	39523	7757
	No	10699	275555	3470
	Poor	3193	11143	38060

QD		Predicted		
		Aid	No	Poor
Observed	Aid	47582	20367	17473
	No	18797	253435	17492
	Poor	1593	3307	47496

Table 6 – (left) Confusion matrix for LD model on validation. (right) Confusion matrix for QD model on validation.

Table 6 provides the confusion matrices for our two discriminant models. Looking first on the left at the linear discriminant model, we see that we are doing fairly poorly at correctly predicting the aid group. We know that in the full data set, the poor group is about 61.34% of the size of the aid group. However, our LD model is predicting the poor group correctly almost as often as it is predicting the aid group correctly. We also know that the no group is approximately 3.39 times the size of the aid group. However, we are correctly classifying the no group about 7.22 times more often than we are correctly predicting the aid group. Together, these two observations suggest that our classifier is not properly dividing the aid group from the other classes. Our classifier is unbalanced and underpredicting aid.

We see that our LD model is primarily misclassifying the aid group within the no group. We have more observations of the aid group misclassified under the no group than we have correctly classified observations. We are also seeing some misclassified as poor, but the LD model does not appear to be misclassifying in that direction as frequently. Of our 85,422 observations within the aid group of the validation set, we have only correctly classified 44.65% of them.

Also in Table 6, we see our QD model's performance. We see that we are doing much better at classifying both the poor and aid group. Both identified about 9,500 more observations correctly compared to the LD model. However, this improvement has come at the expense of the no group. The QD model misclassified approximately 20,000 more observations than the LD model did on the validation set.

In the QD model, we see that misclassification of the aid group is slightly more balanced than it had in the LD model. Our recall for the aid group is now 55.70%. Of our observed values in the validation set, about 23.84% were misclassified under the no group, and about 20.45% were misclassified under the poor group. This still suggests that the aid group is more similar to the no group. However, both the QD and LD model rely on prior probabilities from our data. These misclassification under the no group could be a product of that class's high prior probability.

LD-B		Predicted		
		Aid	No	Poor
Observed	Aid	52616	18007	14799
	No	23738	247281	18705
	Poor	2488	1869	48039

QD-B		Predicted		
		Aid	No	Poor
Observed	Aid	37986	28127	19309
	No	30841	226660	32223
	Poor	1947	2861	47588

Table 7 – Confusion matrices on validation set from balanced prior probability discriminant models. (left) balanced LD model. (right) balanced QD model.

To isolate the effects of prior probabilities in our estimations, we created two additional models, LD-B and QD-B, which are linear and quadratic discriminant models respectively with balanced assumed prior probabilities. We relied on these solely to examine their accuracies. The confusion matrices of these models on the validation set is given in Table 7.

Looking first at LD-B, we see that by balancing our priors, we have vastly improved our accuracy at identifying the aid group. Our recall is now just below 61.60%, which is the best performance of the discriminant models we have tried. We still see that we are misclassifying under the no group more often than the poor group. Misclassifications into the no group represent 21.08% of the total aid group in the validation set. Misclassifications into the poor group make up the remaining 17.32% of aid observations.

However, by balancing the prior probabilities, we have caused more misclassifications in the no group itself. In the first LD model, we saw about 3.69% of no group observations were misclassified as aid, and about 1.20% were misclassified as poor. Initially, this continued to provide evidence that the aid group was more similar to the no group. But by balancing our prior probabilities, we now see about 8.19% of no group observations are misclassified as in the aid group, and about 6.46% of no group observations are misclassified in the poor group. While this the misclassification rate into the aid group is still greater, this reveals that the initial model's skewing toward the aid group was due in part to the higher prior probability of that group compared to the poor group.

For the poor group of LD-B, we see that we have reached accuracy levels slightly above the base QD model. However, with balanced prior probabilities, we are now tending to misclassify the poor group more frequently into the aid group rather than the no group. Before, we were misclassifying into the no group much more often due to its prior probability. In this balanced model, we start to see how the aid group may occupy a role in the middle of the other two classes.

Referring to Table 6 again, we see QD-B is performing worse than the base quadratic model at classifying the aid group. It actually has the lowest recall score for the aid group of all our models, correctly identifying only about 44.47% of observations in the validation set. Of the total aid group, approximately 32.93% were misclassified into the no group, and about 22.60% percent were misclassified into the poor group. This model has the most observations we have seen misclassified into the poor group, but the number of misclassifications into the no group is still about 45.67% larger than it, continuing to suggest that the no group is more similar to the aid group overall.

Generally, the discriminant models seem to point toward the aid group being more similar to the no group than it is to the poor group. However, in our best performing models for aid prediction, we see that our misclassification rates for the aid group into the two other classes are closer in size. Because of this, we would prefer to have a second model to further explore these differences.

For this second model, we used a multinomial logistic regression model to explore the odds structure of our data. We used two logistic regression models for our data: the full model and a univariate model chosen through forward selection. However, the deviance of the univariate model was much greater than that of the full model, so we will relay only the results of the full model here.

Variable	**No Coeff Train**	**Poor Coeff Train**	**No Coeff Test**	**Poor Coeff Test**
(Intercept)	2.876864	5.144877	2.769077*	5.116698*
cap_xpen	-0.0043603220	-0.0005401739	-0.0042982203	-0.0004937644
spm_pov_thres	$-2.992651e^{-5}$	$2.282082e^{-4}$	$-3.244926e^{-5}$*	$2.238852e^{-4}$*
spm_res	-0.0042919954	-0.0002774697	-0.0042144425*	-0.0002731917*
hi_prem	$-3.705362e^{-6}$	$2.529338e^{-6\wedge}$	$-2.988045e^{-6\wedge}$	$-3.243589e^{-8\wedge}$
moop	$-1.748360e^{-6\wedge}$	$1.761656e^{-5}$	$-3.712130e^{-6\wedge}$	$7.084925e^{-6\wedge}$*
mc_pb	0.0003868421	-0.0001568359	0.0003826727	-0.0001811422
agi	$1.960393e^{-7}$	$-7.354294e^{-5}$	$2.966080e^{-7}$	$-7.235995e^{-5}$*
age	-0.004656341	-0.004615106	-0.003638256*	-0.003329779*
hispanic1	0.006381321	-0.069005212	-0.03977009*	-0.09665433*
raceB	-0.2189606	-0.4478417	-0.1762414*	-0.5348453*
raceO	-0.06926238	-0.29069513	0.01736359*	-0.34185308*
raceW	-0.001381003	-0.206787059	0.02935365*	-0.31770956*
num_adlt	-0.2185471	0.3478506	-0.2099781*	0.3863429*
num_kid	-1.1507596	0.2760036	-1.1510268*	0.4356478*
spm_inc	0.0043062996	-0.0001203223	0.0042284391*	-0.0001188375
fica	-4.303693e-03	$5.057513e^{-5}$	$-4.226657e^{-3}$*	$5.078707e^{-5}$
spm_hi_prem	$-5.248126e^{-7}$	$-2.836738e^{-5}$	$1.826235e^{-6\wedge}$*	$-1.887842e^{-5}$
fed_tax	-0.0043815556	-0.0001353057	$-4.308265e^{-3}$*	$-9.643362e^{-5}$*
fed_tax_bc	$4.194196e^{-5}$	$1.367192e^{-3}$	0.0000482511*	0.0012962194*
st_tax	-0.0042925395	-0.0002092647	-0.004213350*	-0.000174398*
work_xpen	$7.564312e^{-5}$	$3.539322e^{-4}$	0.0000968396	0.0002991656
eitc	-0.0001211993	0.0001600206	-0.0001247841	0.0002059546*
cohabit	-0.1331431	4.5301794	-0.1311765*	4.5617854*
cc_xpen	$7.602677e^{-5}$	$1.214061e^{-4\wedge}$	$8.734934e^{-5}$	$6.894488e^{-5\wedge}$
mortageN	0.08538693	0.01753178	0.08801964*	-0.03297490*
mortgageR	-0.1570726	-0.4078408	-0.1525394*	-0.4665876*
marM	0.3324057	0.1765232	0.3453553*	0.1572672*
marNM	0.2215830	0.2939324	0.2688440*	0.3214502*
marS	-0.03877536	0.06729918	-0.001287895*	0.133200627*
marW	0.03283922	-0.18086493	0.06081249*	-0.21347240*
edu<HS	-0.1647042	-0.1695488	-0.1631308*	-0.1144703*
eduC	0.1360675	0.3014302	0.1340644*	0.2679297*
eduHS	-0.04045203	-0.10598443	-0.03021954*	-0.04751511*
sexM	0.02628307	-0.13194236	0.04235782*	-0.09149449*
ui_kids1	-0.1363533	-0.9777019	-0.1521886*	-1.2472156*
med_xpen	$-4.295296e^{-3}$	$5.759174e^{-5}$	$-4.218728e^{-3}$*	$6.814854e^{-5}$*

Table 8 – Coefficients of multinomial logistic regression. (^ indicates variable not significant to α-level of 0.05. * indicates test value outside of confidence from training set.)

Table 8 summarizes the resulting coefficients of our logistic regression model run on both the training and testing set. We created models for each to verify that our resulting coefficients were reasonable within the context of our study. As shown in Table 8, when trained on these different samples, we see some changes in our variables, specifically in their significance. In the original model, we see only hi_prem for the poor sub-model, moop for the no sub-model, and cc_xpen for the poor sub-model as being insignificant. Each of these are only insignificant to one of the two logistic sub-models, justifying their place in the overall model. However, in the test sample model, hi_prem and moop are insignificant to both sub-models, spm_hi_prem is insignificant to the no sub-model, and cc_xpen is still insignificant to the poor sub_model. This suggests that the larger sample size of the training set contains more relationships between our variables than are found in the test set.

Apart from these changes in significance, Table 8 also shows that many of our test coefficients are outside of the confidence ranges of our training model. This is partially due to the size of the training set. With so many observations, the standard errors for our estimates decrease greatly, which causes these intervals to shrink. For many of our categorical variables, these intervals were almost closed, meaning that even relatively small changes in coefficient could leave an estimate outside of the confidence range. This seems to be a reflection of our categorical variables not acting normally due to their discrete nature.

Of the changes in coefficient, the change in age between models seems most interesting. Both coefficients increased by 0.001, approaching closer to 0 in effect. As mentioned earlier with discriminant analysis, we know that the age distribution of the aid group seems centered at a younger age than our other groups. It is somewhat intriguing to see this effect decrease from a change in sample. Also notably, the effect of num_kid for the poor group increases substantially in the test model. As seen with the school lunch subsidies from our exploratory analysis and the importance of cc_xpen in our principal components to the aid group, it is interesting to see this large fluctuation in a child-related variable. We do see, however, that cc_xpen is not significantly different between the poor and aid groups. This observation in and of itself is interesting for similar reasons. Together, these observations could imply that the childcare variables are not what discern the aid group from our other classes, though in the absence of slunch, we cannot fully verify this.

Comparing our other classes to the aid group through the training model, we see that the coefficients of this model appear to be more reasonable than the LD model. First, the coefficient for spm_inc is positively in the no sub-model and negatively in the poor sub-model, which implies the aid group has income below the no group and above the poor group. This seems reasonable. We would expect subsidies to be given to lower income households, and we would also expect those who are not classified as poor to have higher income than those who are. Interestingly, the coefficient for the no sub-model is larger than the coefficient for the poor sub-model. This implies a large income is more important for identifying the no group from the aid group than for identifying the poor group from the aid group.

Interestingly in relation to spm_inc, spm_res is negative in both sub-models. This implies that the aid group would be expected to have the greatest resource according to our supplemental metrics. Given the context of the data, this appears to be a multicollinearity issue, with spm_res likely acting as a counterbalance to spm_inc or other financial variables.

By our tax variables, it appears that the aid group is expected to have greater federal and state tax, but lower tax before credits are applied. For the relation between the aid and poor groups, this makes sense. However, this seems somewhat unreasonable compared to the no group, which is expected to have higher income. Again, this seems to be a multicollinearity issue in the creation of our models. Though all variables are significant, what they are measuring is similar in nature and can therefore cause odd interactions between the coefficients of our models.

Moving on to some of our categorical variables, we see that our coefficients for our race variable are all negative. This would imply that the aid group is more likely to be non-Asian than our other groups. Based on our findings in Appendix 1, this is the opposite of what we would have expected. We would expect the coefficients of the poor group to be nearer to zero compared to the aid group, and the coefficient raceW to be positive for the no group. It seems strange that the magnitude of these negative coefficients are larger in the poor sub-model than the no sub-model because we know that the no group is percentage-wise less diverse. However, this could just mean that the no group has more similar racial distributions than the aid group, which would suggest the aid group is more similar to the no group.

We see that many of our medical and insurance variables are small or insignificant to the model, especially comparing the poor and aid groups. This could suggest that these medical variables are not useful for separating the aid group from our other two classes.

To test the logistic model, we used it to create predictions on the validation set, then compared that to our true labels to create a confusion matrix. The log-model's confusion matrix is given by Table 9. Immediately, we see that this model is performing much closer to what we would expect compared to our discriminant models. We have about 37.43% more correct classifications of the larger aid group than we have of the smaller poor group. While this is not the 61.34% size difference we would expect based on our population size, this model is achieving much closer to the true ratio than what our other models have done so far.

Logistic		Predicted		
		Aid	No	Poor
Observed	Aid	61975	18064	5383
	No	2236	285668	1820
	Poor	1943	5357	45096

Table 9 – Confusion matrix on validation set for logistic regression model.

In Table 9, we see that the logistic regression model also has the greatest recall for the aid variable. Of all the observations in the aid group, approximately 72.55% have been correctly classified by this model. Of the remaining observations, about 21.14% were misclassified into the no group, and about 6.30% were misclassified into the poor group. As we have been seeing across all models, misclassifications are occurring most frequently into the no group. Here, we see more than three times as many observations from the aid group were misclassified as no compared to those misclassified as poor. This continues to suggest that the aid group is most similar to the no group compared to the poor group.

	LD	LD-B	QD	QD-B	Logistic
Accuracy	82.2743%	81.3805%	81.5155%	73.0300%	91.8597%
Recall	44.6513%	61.5954%	55.7023%	44.4686%	72.5516%
Precision	73.3021%	66.7360%	70.0024%	53.6723%	93.6829%
F1	55.4970%	64.0627%	62.0389%	48.6389%	81.7742%

Table 10 – Accuracy, Recall, Precision, and F1 of all supervised models.

Finally, to summarize the results of our group division methods, Table 10 provides confusion scores for each of our models computed on the validation set. While model selection is not the purpose of our research, these scores are helpful, allowing us to compare results and gauge which model best meets its assumptions and is therefore best for our analysis.

Across all models, it appears that the logistic model is the best overall. It has the highest scores of all five models. This would suggest that our results from that model may be most representative of the actual data itself, which would imply that the aid group is most similar to the no group for our analysis.

Interestingly, of the discriminant models, the original LD is performing the best in terms of accuracy. This is strange because we believed that the assumption of an equal covariance matrix was not met. However, looking at the F1 score, we see that it is doing worse than all models except QD-B at predicting the aid group. We know from Table 6 that the LD model was relying heavily upon predicting the no group. It is its tendency to predict the largest group that likely makes the LD model's overall accuracy higher than the rest.

In terms of predicting the aid group, we see that the LD-B model has the greatest F1 score of the discriminant models. We saw in Table 7 that the LD-B model achieved the most correct classifications of our discriminant methods. Where it misclassified, we saw that it was close to balanced misidentifying aid observations into the no and poor groups. It did show slight favoritism toward the no group, though, even with balanced prior probabilities. The base QD model saw similar results both in terms of score and in terms of misclassification rates, as seen in Table 6. The QD-B model also preferred no group misclassification for the aid group, though its overall performance favored the poor group, as shown in Table 7.

A feature we have not examined too deeply so far has been our precision scores. We see across models that our precision scores for the aid group are larger than the aid scores. This would imply that we are misclassifying fewer observations into the aid group than we are misclassifying aid group observations themselves. This trend could imply the aid group is a "weaker class" than at least one other class in our data set. Based on our confusion matrices, it would seem that the "stronger class" that is classified into most frequently is the no group.

Of note, examining the LD-B confusion matrix in Table 7, it appears that the aid group is receives more no and poor misclassifications than the other groups. This could be an indication that the aid group is a "middle group" between the states of the poor and no groups. However, this does not hold for the poor group of our logistic model in Table 9. The no group received more poor group misclassifications than the aid group, though the aid group received more no group misclassifications than the poor group. The results from the logistic model also held true in our Table 6 discriminant models. And in Table 7's QD-B model, the poor group received the fewest misclassifications from the other groups. By majority vote, it would seem that the aid group is largely more similar to the no group as it receives more misclassifications from that group overall compared to the poor group.

4. Conclusion

Given all our results, it appears that the aid group is more similar to the no group than it is to the poor group in the absence of the subsidy variables used to define it. Our principal components yielded similarity scores that all preferred the comparison between the aid group and the poor group of our three main groups. The confusion matrices of all of our discriminant and logistic models also saw more frequent misclassification of the aid group into the no group. Across all our division methods, the aid group saw the lowest F1 scores, implying it to be the weakest of our three groups and the hardest to classify into. Together, these features lead us to believe that the aid group as a whole does not need to be classified as poor because its observations appear more different to that group than they would be to the general population classified under the no group. Through our exploratory analysis, it appears that the aid group most generally represents families in our data set, specifically those whose children receive prepaid school lunches. It seems that this majority of the aid population should not be classified as impoverished and instead remain classified within the non-impoverished population.

5. Limitations

While our results suggest that the aid group should not be classified as poor, our results are by no means definitive. First, many of our methods required our variables to be normally distributed. Though our sample size should be large enough for the central limit theorem to apply, our exploratory analysis showed that most of our variables had very skewed distributions. We also utilized many categorical variables which would be difficult to generalize to a normal distribution. Our discriminant models could see improved performance if many of our continuous variables were transformed, which could potentially change the results given by those models. We do still have our logistic regression model, which does not require normality. That model would imply our results would hold, but we did not explore that possibility here.

In the coefficients of our models, we also saw many potential multicollinearity problems. In both our discriminant and logistic regression models, we saw coefficients that seemed contradicted by our exploratory analysis. Using a linear model predicting age, we find VIF statistics suggesting high multicollinearities within cap_xpen, spm_pov_thres, spm_res, num_adlt, num_kid, spm_inc, fica, fed_tax, fed_tax_bc, st_tax, work_xpen, cc_xpen, and med_xpen. We know that spm_res at least is a combination of many of these variables. However, even when it is removed, only fica, fed_tax, and med_xpen no longer show high multicollinearity. Many of the remaining variables still show multicollinearity then, so a more thorough analysis of which should be removed could benefit our results as well.

We also do not necessarily know the efficacy of the aid group. While we have shown that it is possible to discern much of this group from the other classes, we do not know that the aid group is in and of itself the best class to begin with. As shown in Figure 1, much of the aid group receives school lunch subsidies, but not all. Further research would need to be conducted to see if this is the primary feature actually defining this group, and if so, whether observations without school lunch subsidies should be removed. Our research has not fully explored the nature of this group and whether subdivisions exist within it. Clustering methods could be applied to examine these features of the aid group and our other groups to see if different classification schemas are more accurate to represent the data.

A fourth element that is important to our results is that we removed our subsidy variables from the data set after using them to define our aid group. Because of this, we have not done much analysis using them. One potential direction of research could be attempting to divide those who are receiving subsidies between poor and not. This would avoid biasing results and provide more evidence as to whether the aid group is distinct.

Another direction of research could utilize model creation without the aid group being considered. One could train a discriminant or logistic model to discern only the poor and no groups from each other without using any observations from the aid group. Under this method, we can then attempt to classify aid group observations under either the no or poor classes. Whichever class aid group observations fall under the most would likely be the class the aid group is most similar to under such a model. We did not apply such a method here because it would require a high accuracy model for classifying between the no and poor groups. However, as we have seen, the logistic model seems to have high performance on this task, so it could be useful in such an analysis.

Generally, our findings seem to suggest that current poverty metrics are performing as desired for those individuals within our aid group. The group does not generally seem to be similar to others who are classified as impoverished. However, we do not have perfect discrimination between our groups. We do still see some observations misclassified as impoverished. It is these individuals that we would recommend further study upon in regard to any changes that could be made to U.S. poverty metrics. If there are underlying features that separate those poor misclassified individuals from the rest of the aid group, they could benefit from being identified as poor.

By our exploratory analysis, it seems that most of our research has been concerned with those individuals whose children receive school lunch subsidies. Before considering further research, we would recommend researchers familiarize themselves with these U.S. subsidy programs to better understand their goals and be able to determine whether they are achieving what they intend to through their spending. This research suggests nothing about the efficacy of any subsidy programs brought to attention for our analysis. It simply explores the relation of a subset of those receiving aid to the sample of impoverished individuals given in our data. Our current analysis suggests inaction. There does not appear to be a problem with the classification of the aid group as being non-impoverished.

Works Cited

Benson, C. (2022). *Poverty 2018 and 2021*. United States Census Bureau, 1-20, https://statistical-proquest-com.ezproxy.library.pfw.edu/statisticalinsight/result/pqpresultpage.previewtitle?docType=PQSI&titleUri=%2Fcontent%2F2023%2F2316-15.76534.xml&accountid=11649

Bureau of Labor Statistics (2022). *2022 Research Supplemental Poverty Measure Thresholds*. https://www.bls.gov/pir/spm/spm_thresholds_2022.htm#:~:text=As%20seen%20in%20Chart%201,for%20two%20adults%20two%20children).

Center for American Progress (2022). *Data on Poverty in the United States*. https://www.americanprogress.org/data-view/poverty-data/#tab1

Cronquist, K. & Eiffes, B. (2020). *Characteristics of Supplemental Nutrition Assistance Program Households: Fiscal Year 2020*. Food and Nutrition Service (USDA), xvii-xxi, https://statistical-proquest-com.ezproxy.library.pfw.edu/statisticalinsight/result/pqpresultpage.previewtitle?docType=PQSI&titleUri=%2Fcontent%2F2023%2F1364-8.xml&accountid=11649

Richard, J. & Wichern D. (2007). *Applied Multivariate Statistical Analysis* (6th Edition). Pearson Education Inc..

Mowafi, M. & Khawaja, M. (2005). *Poverty*. Journal of Epidemiology and Community Health (1979-), 59(4), 260–264. http://www.jstor.org/stable/25570683

Office of the Assistant Secretary for Planning and Evaluation (2023). *Poverty Guidelines*. ASPE. https://aspe.hhs.gov/topics/poverty-economic-mobility/poverty-guidelines

Office of Disease Prevention and Health Promotion (n.d.). *Poverty*. Healthy People 2030. https://health.gov/healthypeople/priority-areas/social-determinants-health/literature-summaries/poverty

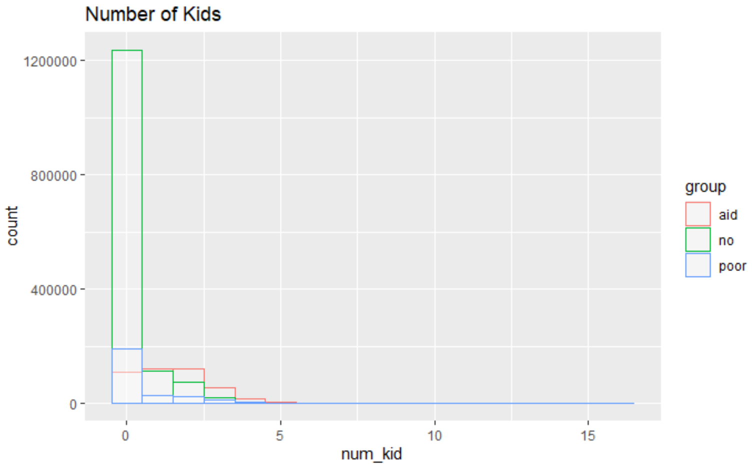
Romig, K. (2020). *Social Security Lifts More People Above the Poverty Line Than Any Other Program*. Center on Budget and Policy Priorities, 1-13, https://statistical-proquest-com.ezproxy.library.pfw.edu/statisticalinsight/result/pqpresultpage.previewtitle?docType=PQSI&titleUri=%2Fcontent%2F2022%2FR3834-37.xml&accountid=11649

United States Census Bureau (2023a). *About the American Community Survey*. https://www.census.gov/programs-surveys/acs/about.html

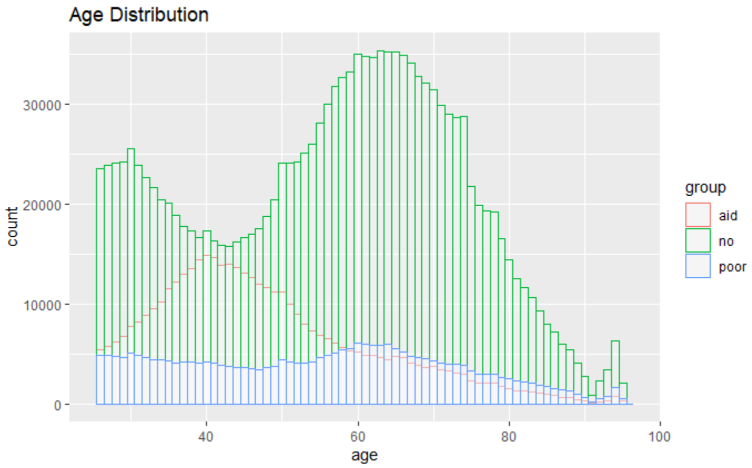
United States Census Bureau (2023b). *ACS Supplemental Poverty Measures (SPM) Research Files: 2009 to 2019, 2021*. https://www.census.gov/data/datasets/time-series/demo/supplemental-poverty-measure/acs-research-files.html

United States Census Bureau (2023c). *Top Questions About the Survey*. https://www.census.gov/programs-surveys/acs/about/top-questions-about-the-survey.html

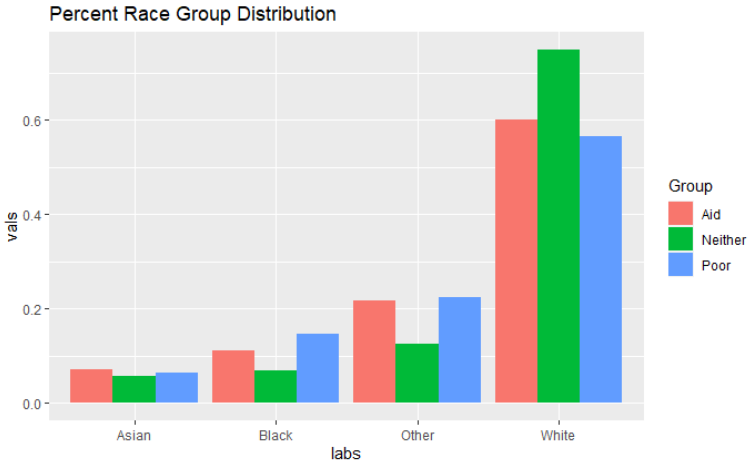
Appendix



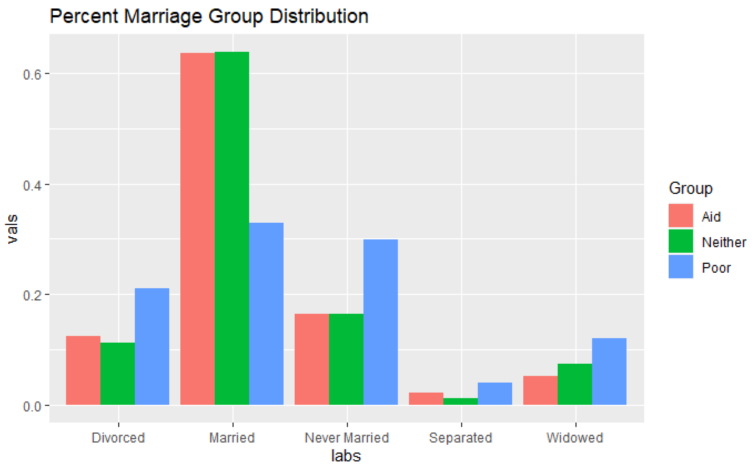
Appendix 1 – Total distribution of num_kid by group.



Appendix 2 – Age distribution by group.



Appendix 3 – Within group percent distribution of racial categories.



Appendix 4 – Within group percent distribution of marriage status.