Analysis of Annual Household Income and Expenses in the Philippines

Ayesha Kumbhare, ID 604936823, Stat 102B

I. Introduction

This paper is an analysis of the annual household income and expenses in the Philippines. The Philippine Statistics Authority (PSA) is an organization that conducts a survey every three years to gather information regarding income, expenditure, consumption of item by expenditure, sources of income in cash by household in the Philippines. This data is from the most recent survey conducted by The Philippine Statistics Authority, which I retrieved from Kaggle [1]. It contains more than 40,000 observations and 60 variables describing income and expenditures of the families. For the sake of this analysis, I have taken a random subset of 400 observations to adhere to my computer's memory and processing limitations.

Using a k-means clustering algorithm, I have conducted a cluster analysis on this data. I have also conducted a Principal Components analysis to reduce dimensionality in the number of variables. Lastly, I have used the Newton Algorithm to find the MLE and asymptotic confidence intervals for the parameters of the most important variable in the dataset. The analysis is conducted on the 45 interval variables found in the data. The methods and algorithms used in this analysis are from Stat 102B at UCLA [2]. References can be found at the end.

II. (Non-Probabilistic) K-means Cluster Analysis

Non-probabilistic k-means is a method to find patterns in data. This method simplifies a large set of multivariate data by dimension reduction into two smaller groups. In this section, I search for natural groupings in the household income and expenses data.

After running the non-probabilistic k-means algorithm, I retrieved 2 clusters. Figure 1 below displays the number of observations in each cluster.

Cluster 1	**Cluster 2**
347 observations	53 observations

Figure 1

Even though I worked with a subset of the data (400 observations), it was difficult to construct a scatterplot visualization to understand the clusters. For this reason, I analyzed the summary statistics. Since the colMeans vector for each vector was quite large (45 variables), I did not report the full table here. The vector is reported in the corresponding .R file of this report. Figure 2 below describes the mean total income and total food expenditure for each cluster in the dataset. These two variables are meaningful, as they display disparities between clusters.

Means

	Total.Household.Income	Total.Food.Expenditure
Cluster 1	172434.7	72554.03
Cluster 2	697601.5	176649.1

Figure 2

From Figure 2, it is evident that cluster 1 contains families of lower total income who also have a lower food expenditure. Cluster 2 contains families of higher total income who also have a higher food expenditure. This makes sense intuitively, as there are 347 observations in cluster 1 which may encompass the lower and middle classes and 53 observations in cluster 2 which may encompass the upper class in the Philippines. The median values for total household income and total food expenditure for each cluster in Figure 3 reflect this analysis as well.

Medians

	Total.Household.Income	Total.Food.Expenditure
Cluster 1	148360	68026
Cluster 2	617320	174071

Figure 3

Standard Deviations

	Total.Household.Income	Total.Food.Expenditure
Cluster 1	100539.3	32938.83
Cluster 2	323771.0	54699.56

Figure 4

Figure 4 demonstrates that there is more variability within cluster 1 regarding total household income. This is expected because the means and medians suggest that cluster 1 encompasses the lower-middle classes which is larger (347 observations fall in this cluster) and this may contain more variability.

Through my cluster analysis, I have found two clusters within the dataset regarding Family Income and Expenditure in the Philippines. In these clusters, the intra-cluster similarity is high, but the inter-cluster similarity is low. The first cluster encompasses families of lower income and lower total food expenditure, while the second one encompasses families of higher income and total food expenditure.

III. Principal Components (PC) Analysis

The non-probabilistic k-means analysis above provides evidence for the existence of two clusters in family income and expenditure data. In this section, I explore a way to summarize the 45 variables in this dataset as concisely as possible.

Principle component analysis allows us to reduce the number of variables to a smaller number of indices (the principal components) that are linear combinations of the original variables. For principal component analysis to be beneficial, the variables must be highly correlated. Figure 5 is a table displaying some of the higher correlations between variables from the dataset.

	Income	Housing /Water	Bread /Cereals	Rice	Food	Miscellaneous Goods	Restaurants /Hotels
Income		**0.851**			**0.7604**	**0.7166**	
Housing / Water	**0.851**						
Bread / Cereals				**0.9062**			
Rice			**0.9062**				
Food	**0.7604**						
Miscellaneous Goods	**0.7166**						**0.7048**
Restaurants / Hotels						**0.7048**	

Figure 5

The principal components are the values of the indexed for each family. They are linear combinations of the original centered and scaled data and the eigenvectors. After calculating the cumulative sum of variability with each principal component, I found that the first 27 principal components (compared to 45 variables in the data set) accounted for 90% of the variance. Similarly, the first 33 principal components accounted for 95% of the variance in the data. Figure 6 describes the variance explained by each PC variable.

1	2	3	4	5	6	7	8	9	10	11	12
28.064	35.601	39.962	44.277	47.796	50.939	53.849	56.589	59.152	61.607	63.988	66.280
13	14	15	16	17	18	19	20	21	22	23	24
68.387	70.488	72.469	74.375	76.168	77.870	79.515	81.052	82.544	83.945	85.281	86.561
25	26	**27**	28	29	30	31	32	**33**	34	35	36
87.804	88.930	**90.029**	91.071	92.075	92.965	93.796	94.610	**95.370**	96.096	96.725	97.348

37	38	39	40	41	42	43	44	45
97.911	98.429	98.838	99.234	99.470	99.667	99.824	99.956	100.000

Accounting for 90% of the variability: 27 principal components
Accounting for 95% of the variability: 33 principal components

Figure 6

Compared to some of the examples regarding principal components that we saw in class, my principal component cumulative sum requires more principal components to reach the target variability. This is because this dataset has 45 variables, which is much more than the examples we saw in class. Even though I could have removed some variables to make this analysis smoother, I decided to keep all the interval variables found in the original dataset, because they did not hinder my computer's performance and I wanted to incorporate and stay true to the dataset as much as possible. However, I found that conducting this principal component analysis results in a much more concise way of representing the data. For this PCA, I decided to keep the first 33 PCs to account for 95% of the variance.

The correlation between the principal component (PC) and the variables tells us how strongly the PC loads on that variable. For this reason, I constructed the correlation matrix between the original dataset and the PC matrix. By doing so, I found that most of the higher correlations occurred in the first couple principal components. I also found many high negative correlations which suggests that as the latent variable increases, the original variable decreases. The complete correlation matrix can be found in the corresponding .R script file. In Figure 7, I attempt to analyze and interpret the latent variables as shown by the some of the higher correlations found in the first couple PCs of the data.

	PC1
Total.Household.Income	-0.89462843
Total.Food.Expenditure	-0.87784592
Meat.Expenditure	-0.75348804
Restaurant.and.hotels.Expenditure	-0.72519758
Housing.and.water.Expenditure	-0.81954475
Transportation.Expenditure	-0.74867244
Clothing..Footwear.and.Other.Wear.Expenditure	-0.70324775

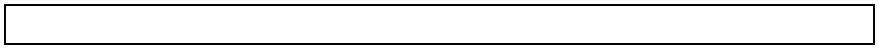


Figure 7.1

Figure 7.1 shows that the expenditures for luxury/expensive items are negatively correlated with the latent variable in PC1. For this reason, I believe that the first latent variable is frugality, or necessity to save.

	PC2
Bread.and.Cereals.Expenditure	-0.70387101
Total.Rice.Expenditure	-0.68204808
Total.Number.of.Family.members	-0.83166547
Members. With.age.5…17.years.old	-0.58028537

Figure 7.2

Figure 7.2 shows that the family size and expenditures for carbs such as bread, cereals, and rice are negatively correlated with the latent variable in PC2. Breads, cereals, and rice are more sustenance items than fruit and meat since they are cheaper and keep people full and satisfied. Additionally, bigger families tend to have more children and young children often consume foods such as rice and baby cereals. For this reason, I believe that the second latent variable is spending per family member. This makes sense intuitively, because the more family members and children there are in a family, the less money they would tend to spend on each family member.

IV. Newton Algorithm for Parameters (MLE)

From conducting a non-probabilistic k-means cluster analysis and a principal component analysis, I have found the most important variable in my to be Total.Household.Income. This is because the two clusters have very different means and medians for this variable, and this variable was the most highly correlated one with the latent variables described in part III. With this new information, I have determined a probability model that applies well to this variable and have used the Newton algorithm to find the MLE of the parameters and asymptotic confidence intervals for the parameters.

The probability model I have used is the lognormal distribution, because the log of Total.Household.Income, is distributed normally. Figure 8 displays the histograms that demonstrate we must use the lognormal distribution:

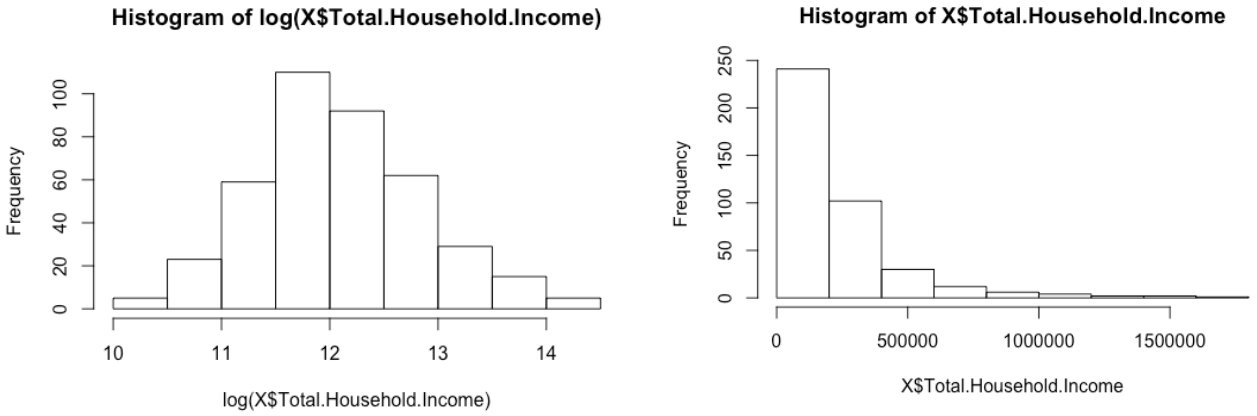
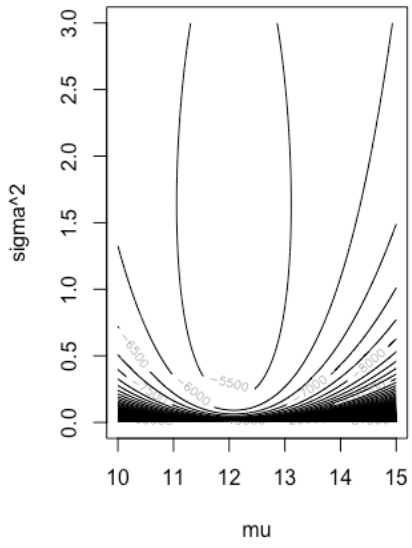


Figure 8

Because the values for Total.Household.Income in the data are all greater than 0, the objective function is log likelihood of the log normal distribution.

L(mu, sigma^2) =
-sum(log(data)) - n*log(sqrt(sigma^2)) - n/2*log(2*pi) - 1/(2* sigma^2)*sum((log(data) - mu)^2)

By optimizing this objective function using Newton's Algorithm, I found 95% confidence intervals for the lognormal parameters mu and sigma^2. Through the contour plot in Figure 9, I found initial values for the parameters.



Initial Value for Mu: 12
Initial Value for Sigma^2: 1.5

Figure 9

Then, I ran the algorithm until I received converging values for these parameters. The final results were mu = 12.08436 and sigma^2 = 1.72566. Lastly, I computed the 95% confidence errors for these parameters using the standard errors obtained from the variances. Here are the 95% confidence intervals:

95% confidence interval for mu: (11.95563, 12.21309)
95% confidence interval for sigma^2: (1.596898, 1.854365)

The interpretations for these confidence intervals is as follows:

- We are 95% confident that the Mu of the model lies between 11.95563 and 12.21309.
- We are 95% confident that the Sigma^2 of the model lies between 1.596898, 1.854365.

The results obtained in my analysis using Newton's algorithm are expected. When I calculated the mean of the log(Total.Household.Income), I found that the true mean is 12.08436. This lies in the confident interval obtained and is exactly the mu obtained after convergence.

V. Conclusions

In this research, I have discovered two clusters within the dataset regarding Family Income and Expenditure in the Philippines. One of the clusters encompasses families of lower income and lower total food expenditure. On the other hand, the other cluster encompasses families of higher income and total food expenditure. Since the first cluster is much larger in size than the second one and contains more variability, I have concluded that cluster 1 incorporates lower-middle class families while cluster 2 incorporates families of high incomes.

Through my principal component analysis on this dataset, I have detected two latent variables in the data, one being frugality and the other being the spending on each family member. I have also condensed the 45 variables in this data set to 27-33 principal components. While this may seem like a lot of principal components, I was able to reduce the number of variables to a smaller number of indices (the principal components) that are linear combinations of the original variables. Even though I could have removed some variables from the original dataset to make the PCA account for more variance with less PCs, I decided to keep all the interval variables found in the original dataset, because they did not hinder my computer's performance and I wanted to incorporate the original dataset as much as possible. Overall, this principal component analysis gave me further insight into family demographics in the Philippines.

Lastly, I obtained the MLE of the parameters using Newton algorithm as well as the asymptotic confidence intervals for them. The results were as expected and resembled the key aspects of the lognormal distribution on the dataset. After viewing the histogram for the dataset, I knew that there must be a better probability model. By inspecting the histogram of the log of the data, I saw that my original data must follow the lognormal distribution.

Overall, through this analysis I have better understood the demographics of and the expenditure/incomes of families in the Philippines. I conclude that using non-probabilistic k-mean, principal components analysis, and Newton algorithm to find MLE of parameters are great ways of inspecting and analyzing new and unfamiliar data.

References

[1] Dataset: https://www.kaggle.com/grosvenpaul/family-income-and-expenditure?fbclid=IwAR0g9pxX-D50GkRplEHQWRaR4xdotlB-caoI0GxKOJxwpRBxHf1uBmguil8

[2] Sanchez, J. Stat 102B lecture notes and R code.