On the Statistical Analysis of the Poverty Index

K. Mitka, M. Vanoušek, H. Stokman, Y. Mash'Al

November 2022

1 Introduction

In this paper, the data set utilized comes from the United Nations Data website, which provides international statistical databases. People from about 100 of the poorest countries were surveyed between 2007 and 2018. From this, various metrics about each country are combined into a Multidimensional Poverty Index (MPI). This metric tries to give a measure of poverty that is not monetary-based. Instead, other dimensions such as education, health and standard of living are measured in order to compute the MPI. For example, Health is a combined score of the indicators *quality of nutrition* and *child mortality rate*. Besides MPI, the dataset also includes other statistics related to poverty, including national and international monetary measures of poverty, being the National Poverty Line and the International Poverty Line.

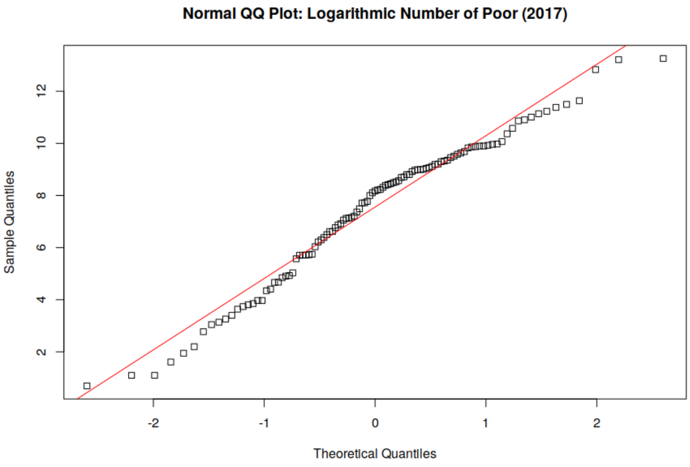
2 Quality

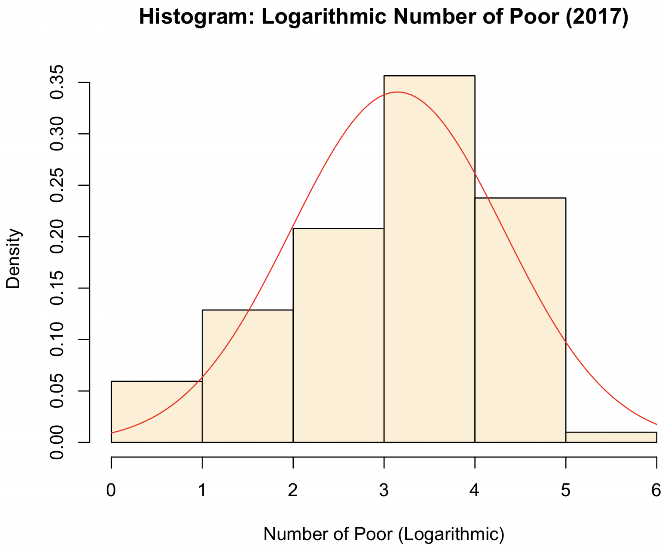
Overall, the the dataset is of high quality. It is only available for download as an XLS, where we removed some metadata and converted it to CSV using a spreadsheet program. While some underlying data used to calculate a particular dimension of the index were missing for certain countries, the authors of the dataset used various methods to calculate the MPI accurately.

We dropped columns which were not used and/or had missing values. The data set also included rows for regions (e.g."Europe and Central Asia"), which we filtered out.

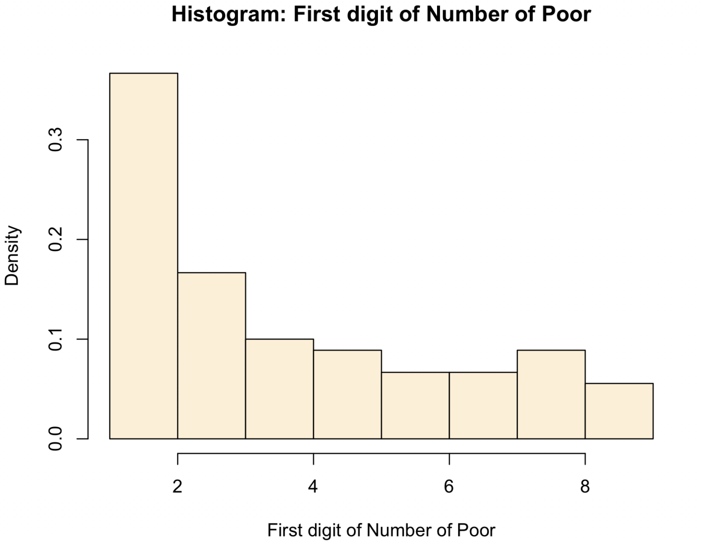
3 Descriptive statistics.

We will analyze *Number of poor*, but as it is a fundamentally multiplicative phenomena, we will not study the variable itself, but rather its order of magnitude. Looking at the histogram and QQ-plot, we see the distribution is not quite normal. The QQ-plot isn't straight, it has a wavy pattern. If there was any doubt left, the Shapiro-Wilk's test rejects normality at a 99% confidence level.





As *Number of Poor* spans multiple orders of magnitude, Benford's law should apply. We can verify this with a histogram of the first digit. The distribution isn't perfect, but as you can see, 1 is by far the most common first digit and the frequencies roughly decrease, as predicted by the law.



4 Analysis

4.1 Are small countries more likely to be outliers?

For a good estimator, variance decreases with sample size. Extreme values are often caused by unusual means, as well as by a large measurement error of the underlying parameter. This gives rise to following research question: Are outliers (below 10th or above 90th percentile of the *Index*) more likely to come from small countries (below 20th percentile of *Number of poor*)? Note that both of these groups (small countries and outliers) are by construction equally large.

We will not assume the underlying distribution. Instead, we shall perform a non-parametric test on a contingency table. As this is a one sided test, we can't use Chi-squared test and must use Fisher's exact test. We test the null hypothesis 'Outliers are less likely or equally likely to be small countries.' against the alternative 'Outliers are more likely to be small countries'. The resulting p-value of 0.74% is very small. This leads us to reject the null hypothesis for all common values of α. Therefore, outliers are more likely to be small countries with a high level of statistical certainty.

```
FALSE TRUE                Fisher's Exact Test
FALSE    62   10        data:  contingency_table
TRUE     10   8       p-value = 0.007435
```

4.2 Is the National Poverty Line higher than the International Poverty Line?

The data set provides two data points per each country that measure the proportion of population living under a certain point of wealth. The first observation being the %NPL *(% of population living under the national poverty line, which is decided by authorities in each of the countries)* and the second one being the %IPL *(population living under the international poverty line, which is standardized to be PPP of 1.90 dollars a day)*. It is in our interest to see whether the NPL is on average drawn higher than the IPL in the developing countries, meaning that the international poverty line is set much lower than what the authorities of developing countries would consider to be adequate.

Since both observations measure the population proportion under a certain line of wealth, we cannot assume their independence, and instead we will treat the observations as paired samples. For each country we compute the difference X = (%NPL - %IPL) and treat all the differences $X_1, X_2..X_n$ as independent and $N(\mu, \sigma^2)$-distributed with the unknown expected difference μ and unknown variance σ^2. The assumption of normality should be taken with a grain of salt as we are dealing with percentage values, which is a bounded statistic, and it is unlikely that %NPL and %IPL follow a normal distribution. You can imagine %NPL is a combination of what the cutoff of the NPL is and how wealth is distributed in that country and for the %IPL the cutoff is fixed (PPP 1.9 dolars a day), so the %IPL depends entirely on the distribution of wealth in a country. This makes it unlikely that the %NPL - %IPL is normally distributed. Assuming normality should still result in accurate results.

The question of whether NPL is on average higher than IPL can be conceived as a test on H_0: $\mu = 0$ against $H_1 : \mu > 0$ with a significance level $\alpha = 0.01$.

Our test statistic will be $T = \frac{\bar{X}}{S/\sqrt{n}}$ where n is the number of rows considered from our data set. The R script gives values $\bar{X} \approx 12.904$ and $S \approx 16.049$. Therefore the mentioned test statistics $T \approx 7.628$. Since this is an upper-tailed test we compute the quantile function of the T-distribution with $df = n - 1$ and $\alpha = 0.01$, which yields the rejection region to be: $T > c = 2.369$.

From this we conclude that we reject H_0 as $T = 7.628 > c = 2.369$ at a significance level of $\alpha = 0.01$. This proves that the International Poverty Line is significantly lower than the poverty line drawn by the national authorities of the developing countries.

4.3 Does the size of a country influence its poverty rate?

We will split every country into two groups: small countries and big countries. First of all, we have to define what it means to be a small country or a big country. As a cutoff, the median of our data is going to be used, where we would have an equal amount of countries in both categories by construction. However, the total population of each country is not given in our data set. We will add an extra column and calculate the total population by dividing *the number of poor people* by the fraction of *people living below the national poverty line(NPL)*. Our research question will be: Is there a difference in the poverty rate(fraction of people living below NPL) between small countries and big countries?

To answer this question we will assume people in small and big countries have a p_1 and p_2 probability of being poor respectively. Therefore, the number of poor people living in small countries X is $B(n, p_1)$ distributed and the number of poor people in big countries Y is $B(m, p_2)$ distributed, where n and m is the number of people living in small and big countries respectively. To calculate n and m we can sum the population of all *small/big* countries and to calculate the sample proportions $\widehat{p_1}$ and $\widehat{p_2}$, we can sum over the amount of poor people in *small/big* countries and divide that by the n/m we just calculated.

Since we are dealing with large n and m we can say that $\frac{X}{n}$ and $\frac{Y}{m}$ are approximately normally distributed with expectation p_1 and p_2 and variance $\frac{p_1(1-p_1)}{n}$ and $\frac{p_2(1-p_2)}{m}$ respectively. Assuming X and Y are independent, we can also calculate the expectation and variance of $\frac{X}{n} - \frac{Y}{m}$, which is $(p_1 - p_2)$ and $Var(\frac{X}{n}) + Var(\frac{Y}{m})$ respectively.

To answer our research question we will test $H_0 : p_1 = p_2$ against $H_1 : p_1 \neq p_2$ at a significance level of $\alpha = 0.01$. Our test statistic under H_0 will be:

$$Z = \frac{\widehat{p_1} - \widehat{p_2}}{\sqrt{\frac{\widehat{p_1}(1-\widehat{p_1})}{n} + \frac{\widehat{p_2}(1-\widehat{p_2})}{m}}} \sim N(0,1)$$

We will reject H_0 in favor of H_1 if the $p\text{-}value = 2(1 - pnorm(Z)) \leq \alpha = 0.01$. From our R code we find $n = 6.20 \times 10^9$, $m = 1.17 \times 10^8$, $\widehat{p_1} = 0.377$ and $\widehat{p_2} = 0.202$. From this we compute $Z = 3881$ and we get a *p-value* of $0 \leq \alpha = 0.01$. So we reject H_0 in favor of H_1 at a significance level of $\alpha = 0.01$. We conclude that, at a 99% confidence level, the poverty rate in small countries and big countries is different.

This result doesn't seem entirely reflective of the real-world, where smaller countries aren't necessarily poorer than bigger countries. Our assumption of binomial distributions and independence of X and Y and independence of countries within X and Y is not completely accurate. Also, since we didn't have the population of each country, we had to estimate it with the amount of poor people and the %NPL. This estimation would have propagated the most significant error in our test.

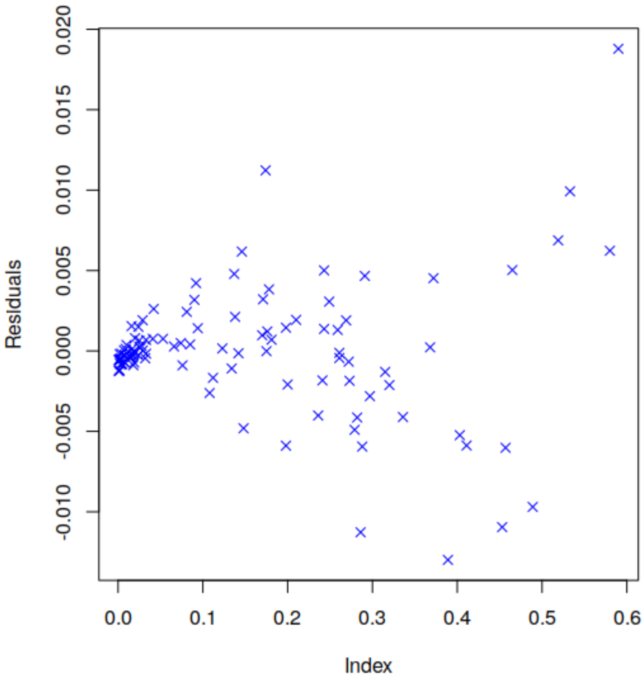
5 Regression analysis

In the data set, the Poverty Index is a function of a multitude of other variables. Thereby, the Index is the response variable, while the other variables (Health, population, etc.) are the explanatory variables. But is the index just a linear combination of the other variables? We can use linear regression to answer this question.

$$\hat{Y} = X\hat{\beta} + \epsilon, \tag{1}$$

where $\hat{Y}$ is vector composed of the response variables, X is the vector of the explanatory variables, $\hat{\beta}$ is the vector composed of the coefficients of the regression model and ϵ is the vector of the disturbances.

As the adjusted R^2 is very high (0.9993), we can conclude that the Poverty Index is indeed a linear combination of the other variables. However, the regression model can, surprisingly, be simplified. While Headcount and Population in Severe Multidimensional Poverty (PSMP) were highly significant (***), no other variable was. If we instead model index using just these two variables, the adjusted R^2 doesn't change. Plotting the residuals, we see a typical pattern – residuals are proportional to the index. As the residual plot does not appear random, we conclude that the linear regression model is not accurate.



Applying our newfound regression model, we can compute a 95% prediction interval for the Index, given the Headcount and PSMP (in %). For a Headcount = 100 and PSMP = 15%, we interpolate and get a prediction interval that is lower bounded by 0.363 and upper bounded by 0.387.

6 Conclusion

The data-set proved to be comprehensive in that it allowed us to conduct various statistical tests in order to draw conclusions.

Comparing the NPL and IPL, we found that the NPL is set significantly higher than the IPL when assuming that the %NPL - %IPL is normally distributed. It's interesting that national authorities set the cutoff for poverty a lot higher than the international cutoff. We had expected the IPL to be more like the average of all the NPLs.

We also looked at the difference of big and small countries. First we had to categorise countries into big and small countries. Then we assumed a binomial distribution for the number of poor people in small and big countries and found that, for our data set, there is a significant difference in the poverty rate in small and big countries, with the small countries having higher a higher poverty rate (we only tested for difference, but this was clear from the statistics).

On the other hand, we tested whether outliers are more likely to come from small countries. We applied the Fisher's exact test, which yielded a p-value of 0.74%. Thereby, we rejected the null hypothesis and concluded that outliers are more likely to come from small countries.

Additionally, the regression model analysis proved that the Index can be explained by Headcount and PSMP. This was further applied when we utilized the simplified regression model in order to compute a 95% prediction interval, given Headcount and PSMP.

Lastly, the data-set can be further analyzed by comparing more qualitative variables using a Chi-Square test. However, since the data-set is composed of continuous data, we would have to group the continuous variables into categorical groups, composed of more than two categories, in order to perform a chi-quare test (we do have have a test on 2 categories, so we could perform a chi-square test, but it's equivalent to a two-tailed Z test). For example, it would be interesting to create categories from a dimension (*Health, Education* and *Standard of living*) which has the greatest/lowest impact on the MPI and see whether that influences the value of the MPI or the poverty rate.

A Appendix: R code

```
the_color = "papayawhip"

data = read.csv(file.choose(), header=TRUE)

drops = c("Year.and.survey", "Inequality.among.the.poor")

data = data[ , !(names(data) %in% drops)]

data$Number.of.poor..2017. = strtoi(gsub(",", "",data$Number.of.poor..2017.))

#library for filtering

library(dplyr)

regions = c("Arab States", "East Asia and the Pacific", "Europe and Central Asia",

            "Latin America and the Caribbean","South Asia","Sub-Saharan Africa")

#filter out regions

data = filter(data, !(Country %in% regions))

data = na.omit(data)

attach(data)


#Section 3

poor_population_log = log(Number.of.poor..2017., base = 10)

qqnorm(poor_population_log, main = "Normal QQ Plot: Logarithmic Number of Poor (2017)", pch=0)

qqline(poor_population_log, col="red")

shapiro.test(poor_population_log)


#Histogram of the number of poor vs the Density.

poor_population_log = log(Number.of.poor..2017., base = 10)

hist(poor_population_log, main="Histogram: Logarithmic Number of Poor (2017)", prob=TRUE,

                      col = the_color, breaks = 6, xlab="Number of Poor (Logarithmic)")

#add a normal p.d.f.

m=mean(poor_population_log, na.rm=T)

stdev=sd(poor_population_log, na.rm=T)

curve(dnorm(x,mean=m, sd=stdev),col="red",add=TRUE)

first_digit = as.numeric(substr(Number.of.poor..2017., 1, 1));

hist(first_digit, main="Histogram: First digit of Number of Poor", prob=TRUE, col = the_color,

                                        breaks = 6, xlab="First digit of Number of Poor")
```

```
#Section 4.1
outlier_percentile = .1
data$Small.country = data$Number.of.poor..2017. < quantile(data$Number.of.poor..2017.,
                                                           2*outlier_percentile)
data$Outlier = data$Index < quantile(data$Index, outlier_percentile) |
              data$Index > quantile(data$Index, 1-outlier_percentile)
contingency_table = contingency_table = table(data$Outlier, data$Small.country,
                                              dnn = c("Outlier", "Small"))
contingency_table
fisher.test(contingency_table,alternative = "greater")

# Section 4.2
# Compute the difference between National poverty line and International poverty line
data$Difference = data$National.poverty.line - data$PPP..1.90.a.day
# Compute the mean of the difference
mean_difference = mean(data$Difference)
# Compute the standard deviation of the difference
sd_difference = sd(data$Difference)
# Hypothesis test
# H0: mean_difference = 0
# H1: mean_difference > 0
# alpha = 0.05
# t-statistic
t_statistic = mean_difference / (sd_difference / sqrt(length(data$Difference)))
# quantile of the t-distribution
quantile_t = qt(0.99, df = length(data$Difference) - 1)
# Conclusion
if (quantile_t < t_statistic) {
  print("Reject H0")
} else {
  print("Accept H0")
}
```

```
#Section 4.3: Poverty in small vs big countries
#calculate population
data$population = 1000*data$Number.of.poor..2017. / (data$National.poverty.line/100)
med = median(data$population)
#data set for small and big countries
small = filter(data, data$population <= med)
big = filter(data, data$population > med)
#calculate number of people in small countries and number of people in big countries
n = sum(small$population)
m = sum(big$population)
#sample proportion of small and big countries
sp1 = 1000*sum(small$Number.of.poor..2017.) / n
sp2 = 1000*sum(big$Number.of.poor..2017.) / m

#statistic and p_value
Z = (sp1 - sp2) / sqrt((sp1*(1 - sp1)/n + sp2*(1 - sp2)/m))
p_value = 2*(1 - pnorm(Z))

#Section 5
#building a multi-linear regression model using Index as response variable
#and the other factors as explanatory variables.
#Result shows that only Headcount and Population in severe multidimensional poverty are relevant.
model <- lm(Index ˜ Headcount + Intensity.of.deprivation + Number.of.poor..2017. +
              Population.in.severe.multidimensional.poverty +
              Population.vulnerable.to.multidimensional.poverty + Health +
              Education + Standard.of.living , data = data)
summary(model)

#Building a new regression model
regr <- lm(Index˜Headcount + Population.in.severe.multidimensional.poverty, data=data)
summary(regr)
```

```
#plot the residuals

plot(Index, residuals(regr),col="blue",pch=4)

abline(regr)


############ PREDICTION INTERVAL ############

#compute a 95% prediction interval for Index

#given that Headcount = 100 and

#Population.in.severe.multidimensional.poverty = 15.

newdata=data.frame(Headcount = 100, Population.in.severe.multidimensional.poverty =15)

predict(regr, newdata, interval="predict")


#fit       lwr       upr

#1 0.3753793 0.3645504 0.3862082


detach(data)
```