# TNSA Household Wealth Index

*Ömer Şahin*

## Introduction

2013 Turkey Demographic and Health Survey (TDHS-2013), fertility levels and trends, infant and child mortality, family planning and maternal and child health issues a sample survey at the national level designed to provide information on. And also, gives information about the wealth index of the household. In this work, the wealth index analysis and prediction according to other information about a household are evaluated.

## Data Analyze

### Libraries

Required libraries for data preparation and analyze:

```
library(readr)
library(tidyr)
library(ggplot2)
```

### Data Preperation

#### Load Household Dataset

Simplified household data of the TNSA dataset is loaded.

```
household <- read.table(file="household.csv", sep=";", header = TRUE, na.strings = c("" , " ","\t ", "M
```

Each household row has "case_id" field. These values are unique for each sample. The ID column is dropped due to has no contribution to the training.

```
household <- household[, -1]  # drop case id
```

In house ownership column has only one "Other" feature. Therefore, this row is evaluated as outliers and dropped.

```
household <- household[-which(household$house_ownership == "Other"), ]
```

In this stage, the number of samples:

```
nrow(household)
```

```
## [1] 11793
```

The dataset has some rows with unknowns attributes. The number of rows with missing values is:

```
sum(!complete.cases(household))
```

```
## [1] 178
```

The number of rows that are not complete can be ignored compared to the total number of samples, and these samples are dropped.

```
household <- na.omit(household)
```

At the end of the data clearing, the number of samples:

```r
nrow(household)
```

```
## [1] 11615
```

### Combined Region

The household data contain the combined region field that consists of combination of a cardinal direction, region, and settlement. These fields are separated into three.

```r
head(household$region_combined)
```

```
## [1] West - Istanbul - Urban/Metropol West - Istanbul - Urban/Metropol
## [3] West - Istanbul - Urban/Metropol West - Istanbul - Urban/Metropol
## [5] West - Istanbul - Urban/Metropol West - Istanbul - Urban/Metropol
## 35 Levels: Central - Aegean - Rural ... West - West Marmara - Urban
```

```r
household <- separate(household, region_combined, c("cardinal_direction", "region", "settlement"), sep =
household$cardinal_direction <- as.factor(household$cardinal_direction)
household$region <- as.factor(household$region)
household$settlement <- as.factor(household$settlement)

head(household[, 2:4])
```

```
##   cardinal_direction     region      settlement
## 1               West   Istanbul   Urban/Metropol
## 2               West   Istanbul   Urban/Metropol
## 3               West   Istanbul   Urban/Metropol
## 4               West   Istanbul   Urban/Metropol
## 5               West   Istanbul   Urban/Metropol
## 6               West   Istanbul   Urban/Metropol
```

### Wealth Index

The aim of this project is predicting the wealth index of the household. There is an order between wealth index values. Therefore, wealth index factor is releveled.

```r
# Refactor levels of wealth index
household$wealth_index <- factor(household$wealth_index,
                                 levels = c("Poorest", "Poorer", "Middle", "Richer", "Richest"))
levels(household$wealth_index)
```

```
## [1] "Poorest" "Poorer"  "Middle"  "Richer"  "Richest"
```

### Attribute Relation

Some attributes have a relation to each other. These relations can be represented as a ratio between them. In this way, all household samples will have attributes that in the same range even they are in the natural number range. After extending columns with the rate of related ones, duplicate columns are dropped.

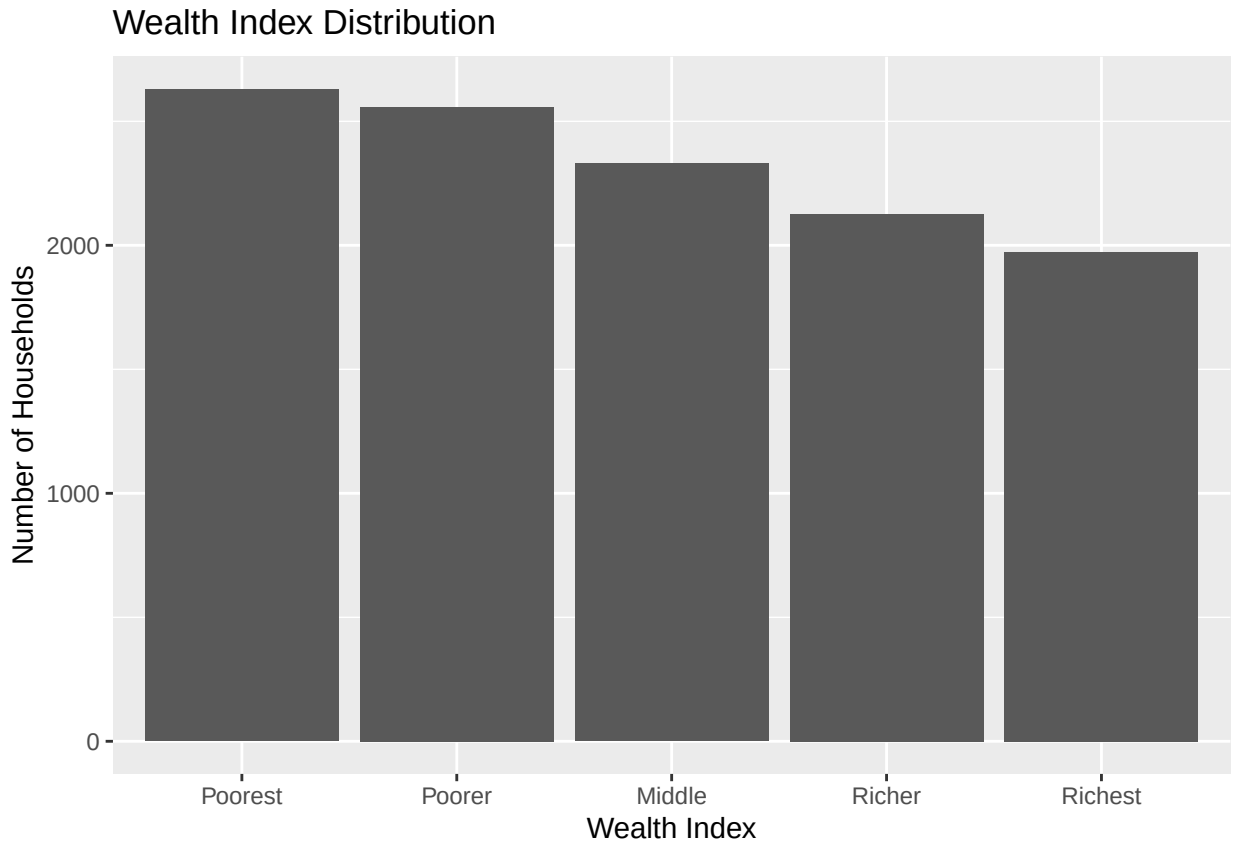
```r
# Rate of related attributes
household$man_member_rate = (household$household_member - household$woman_member) / household$household
household$woman_member_rate = household$woman_member / household$household_member
household$child_member_rate = household$children_under_5 / household$household_member
household$bedroom_rate = household$bedroom_number / household$rooms_number

# Drop duplicated columns with rate values
household <- household[ , -which(names(household) %in% c("woman_member",  # woman_member_rate
```

```
                                                    "children_under_5", # children_member_rate
                                                    "bedroom_number"))]  # bedroom_rate
```

### Data Distribution

The wealth index of a household is the target that is tried to predict. The distribution of the wealth index is nearly balanced. It helps to ensure the model does not tend to a class when the target class is balanced.

```
ggplot(household, aes(x = wealth_index)) + geom_bar() +
        xlab("Wealth Index") + ylab("Number of Households") +
        ggtitle(label = "Wealth Index Distribution")
```




Distributions of the some of binary attributes about households are too imbalanced. There are not enough counter samples for these attributes. Therefore, these attributes can misguide the model when predicting the wealth index. Imbalanced attributes:

```
summary(household[, c("refrigerator", "garbage_grinder", "washing_machine", "washer_dryer",
                      "home_theather", "mobile_phone", "taxi_minibus", "tractor", "motorcycle")])
```

```
##  refrigerator garbage_grinder washing_machine washer_dryer home_theather
##  No :  175    No :11544       No :  515       No :11411    No :11282
##  Yes:11440    Yes:   71       Yes:11100       Yes:  204    Yes:  333
##  mobile_phone taxi_minibus tractor     motorcycle
##  No :  565    No :11126    No :10642   No :10806
##  Yes:11050    Yes:  489    Yes:  973   Yes:  809
```

According to results, nearly all households have the refrigerator, washing machine, and mobile phone. On the other hand, nearly none of the households have the garbage grinder, washer dryer or home theatre.

```
# Plot distribution of data
# for(colnm in colnames(household)) {
#   print(ggplot(household, aes_string(x = colnm)) +
#           geom_bar() + ylab("Number of Households"))
# }

# for(colmn in colnames(household)) {
#   tbl <- table(household$wealth_index, household[,colmn])
#   tbl <- tbl / rowSums(tbl)
#   conf_matrix <- as.data.frame(tbl)
#   print(
#     ggplot(data =  conf_matrix, mapping = aes(x = Var1, y = Var2)) +
#       xlab("wealth_index") + ylab(colmn) +
#       geom_tile(aes(fill = Freq)) +
#       geom_text(aes(label = sprintf("%0.4f", Freq)), vjust = 1) +
#       scale_fill_gradient(low = "blue",
#                           high = "red",
#                           trans = "log")
#   )
# }
```