Reall Household Income and Affordability Calculator

Methodology Paper: Pakistan 2019

This paper lays out the methodology used to develop the 2019 Pakistan data for Reall's Household Income and Affordability Calculator, which can be found at www.reall.net/calculator.

The Household Integrated Economic Survey (HIES) is produced by the Pakistan Bureau of Statistics (PBS). HIES is a nationally-representative household survey focused on presenting income and consumption expenditure data, with the current round covering 24,809 households. The full published report and tables[1] provide a broad overview of the landscape, with the raw data freely available for more detailed analysis[2].

HIES uses robust processes for calculating both income and consumption expenditure, with both containing both cash and in-kind receipts. Section 12CE contains a verification process, checking whether income is at least 85% of expenditure, and if not, requesting additional data from the respondent. This sheet uses this income data, rather than expenditure, using the 'best case' of the first figure where it is at least 85% of expenditure, or the second figure where this was not the case.

Data is split into separate data files for different sections of the survey. Key sections of the HIES survey used for this work are 'sec 1a' providing number of people per household; 'sec 12a' providing number of income earners per household and income per person; 'sec 12ce', which provides summarised income and expenditure for each household; 'code' which provides names for province, division and region codes; and 'weight', which provides weightings for each household.

Methodology

The following actions were undertaken on these datasets:

Creating Dataset

- weight
    - Rename [psu] to [PSU]
- sec 1a
    - Grouping data by [hhcode] (household code) to create [hhsize] (household size) column
- sec 12a
    - Filter data to exclude cases where [idc] (household member ID code) is 99 (denoting total row for each household)

[1] https://www.pbs.gov.pk/publication/household-integrated-economic-survey-hies-2018-19
[2] https://www.pbs.gov.pk/content/pslm-hies-2018-19-microdata

- Creation of new [is_earner] column in which value is 1 for each household member where [s12aq08] (total sum of all income columns, covering occupations, wages in-kind and pensions etc) is greater than 0
- Group all data by [hhcode] (household code), summing number of [is_earner] household members in a new [no_of_earners] column

- sec 12ce
  - Creation of new [monthly_income] which uses the larger of [t_income] and [t_income1] and then divides this by 12
  - Merger of [hhsize] from 'sec 1a' by [hhcode]
  - Merger of [no_of_earners] from 'sec 12a' by [hhcode]
  - Extract the following from [PSU]:
    - [province] using 1st number in [PSU] value
    - [division] using 2nd and 3rd numbers in [PSU] value
    - [region] using 4th number in [PSU] value
  - Merger of [province_name], [region_name] and [district_name] from 'codes', by [province], [division] and [region]
  - Merger of [weight] from 'weight' by [PSU]

```
# Data processing for income and earners
sec1a_dataset <- sec1a_dataset %>%
        group_by(hhcode) %>%
        summarise(HHsize = n()) %>% #changed from mutate
        ungroup()

weight_dataset <- weight_dataset %>%
        rename(
        PSU = psu
        )

sec12A_dataset <- sec12A_dataset %>%
        filter(idc != 99) %>%
        mutate(is_earner = ifelse(s12aq08 != 0, 1, 0)) %>%
        group_by(hhcode) %>%
        summarise(no_of_earners = sum(is_earner)) %>% #changed from mutate
        ungroup()

sec12CE_dataset <- sec12CE_dataset %>%
        mutate(monthly_income = ifelse(ratio_lrg == "Yes", t_income / 12, t_income1 / 12))

# Merge all datasets for Pakistan
merged_dataset <- left_join(sec12CE_dataset, weight_dataset, by = "PSU") %>%
        left_join(sec1a_dataset[c('hhcode', 'HHsize')], by = "hhcode") %>%
        left_join(sec12A_dataset[c('hhcode', 'no_of_earners')], by = "hhcode") %>%
        left_join(sec00_dataset[c('hhcode', 'Q05')], by = "hhcode") %>%
        mutate(
                province = as.integer(substr(PSU, 1, 1)),
                division = as.integer(substr(PSU, 2, 3)),
                region = as.integer(substr(PSU, 4, 4))
                ) %>%
```

```
        left_join(codes, by = c('province', "division", "region")) %>%
        left_join(perc_income, by = c('province_name', "Urban_rural")) %>%
        group_by(hhcode) %>%
        filter (row_number() == 1)

merged_dataset <- merged_dataset %>%
        mutate(Urban_rural = case_when(
                Urban_rural == "urban" ~ "Urban",
                Urban_rural == "rural" ~ "Rural",
                TRUE ~ Urban_rural
                ))
```

Calculating Sample Sizes

- Grouping dataset by [district_name] and [province_name] to create sample sizes for each district
- Grouping dataset by [district_name], [province_name] and [urban_rural] to create separate urban and rural sample sizes for each district

```
sample_size_pakistan <- merged_dataset %>%
        group_by(province_name, Urban_rural, district_name, Year = 2019) %>%
        summarise(sample_size = n())

sample_size_pakistan_u_r <- merged_dataset %>%
        group_by(province_name, Urban_rural ="All", district_name, Year = 2019) %>%
        summarise(sample_size = n())
```

Calculating Percentiles

- Multiplying each [hhcode] record by [weight] to create full weighted dataset
- Grouping data by [district_name] and [province_name], and extracting data for records at 1% increments, creating figures for each percentile of every district
- Repeating the step above but also grouping by [urban_rural], creating separate urban and rural percentile figures for each district

```
# Function to calculate quantiles and related statistics
calculate_quantiles <- function(data, quantiles) {
        do.call(rbind, lapply(quantiles, function(q) {
                data.frame(
                        Quantile = q * 100,  # Convert quantile probability to percentage
                        HH_exp = quantile(data$HH_exp, probs = q, na.rm = TRUE)
                        )
                }))
        }

all_data_pakistan <- merged_dataset %>%
        uncount(weights = as.integer(weight / 100)) %>%
```

```
        rename(HH_exp = monthly_income, HH_size = HHsize)

# Calculate quantiles for each location
pakistan_location_quantiles <- all_data_pakistan %>%
        group_by(Country = "Pakistan", urban_rural = Urban_rural, Location = province_name,
        City =district_name, Year = 2019) %>%
        group_modify(~ calculate_quantiles(.x, quantile_probs)) %>%
        ungroup()

# Additional mixed urban/rural quantiles for each location
quantile_probs <- seq(0.01, 0.99, by = 0.01)
pakistan_location_quantiles_u_r <- all_data_pakistan %>%
        group_by(Country = "Pakistan", urban_rural = "All", Location = province_name, City =
        district_name, Year = 2019) %>%
        group_modify(~ calculate_quantiles(.x, quantile_probs)) %>%
        ungroup()
```

Creating Final Dataset

- Define and align common columns to enable merging of quantiles datasets
- Combine relevant 'quantiles' and 'quantiles_u_r' datasets into a single 'summary_dataset'
- Define and align common columns to enable merging of sample size datasets
- Combine relevant 'sample_size' datasets into a single 'combined_sample_size'
- Join 'combined_sample_size' dataset to 'summary_dataset'
- Create a 'state_aggregated' version of 'summary_dataset' by grouping all household data by state
- Create a 'national_aggregated' version of 'summary_dataset' by grouping all household data by country
- Join 'state_aggregated' and 'national_aggregated' datasets to 'summary_dataset'

```
# Define common columns for final summary
common_columns <- c("Country", "urban_rural", "Location", "City", "Year", "Quantile",
"HH_exp", "no_of_earners", "HH_size", "Percent_Income_Spent_on_Housing")

# Function to align columns across datasets
align_columns <- function(df, common_cols) {
        df %>%
        mutate(across(setdiff(common_cols, colnames(df)), ~ NA)) %>%
        select(all_of(common_cols))
        }

# Align columns and combine all datasets
Pakistan_location_quantiles <- align_columns(Pakistan_location_quantiles,
common_columns)
Pakistan_location_quantiles_u_r <- align_columns(Pakistan_location_quantiles_u_r,
common_columns)
```

```
# Combine all country datasets into a single summary dataset
summary_dataset <- bind_rows(Pakistan_location_quantiles,
Pakistan_location_quantiles_u_r)

sample_size_Pakistan <- sample_size_Pakistan %>%
        rename(Location = state, sample_size = sample_size) %>%
        mutate(Country = "Pakistan")

sample_size_Pakistan_u_r <- sample_size_Pakistan_u_r %>%
        rename(Location = state, sample_size = sample_size) %>%
        mutate(Country = "Pakistan")

# Combine the sample size tables into one
combined_sample_size <- bind_rows(
        sample_size_Pakistan,
        sample_size_Pakistan_u_r,

summary_dataset <- summary_dataset %>%
        left_join(combined_sample_size, by = c ("Country", "urban_rural", "Location", "City",
        "Year"))

# Aggregate data for state level
state_aggregated <- summary_dataset %>%
        group_by(Country, Location, Year, Quantile,urban_rural) %>%
        summarise(
                sample_size = sum(sample_size),
                HH_exp = mean(HH_exp, na.rm = TRUE),
                no_of_earners = mean(no_of_earners, na.rm = TRUE),
                HH_size = mean(HH_size, na.rm = TRUE),
                ) %>%
        ungroup()%>%
        mutate(City = "All")

# Aggregate data for the national level by combining all states within each country
national_aggregated <- summary_dataset %>%
        group_by(Country, Year, Quantile,urban_rural) %>%
        summarise(
                sample_size = sum(sample_size),
                HH_exp = mean(HH_exp, na.rm = TRUE),
                no_of_earners = mean(no_of_earners, na.rm = TRUE),
                HH_size = mean(HH_size, na.rm = TRUE)
                ) %>%
        ungroup() %>%
        mutate(Location = "All", City = "All")

# Combine both the state-level and national-level data
final_dataset <- bind_rows(state_aggregated, national_aggregated,summary_dataset)
```

Calculating Inflation

All data is inflated using median annual inflation rates from 2010-23. Median rates were used rather than actual figures to help compensate for large-scale inflation across many economies in 2022 and 2023.

Consumer Price Index inflation rates were taken from the World Bank[3] and consisted of the following figures, creating a final median rate of 9.59%.

CPI	Pakistan
2000	4.366665
2001	3.148261
2002	3.290345
2003	2.914135
2004	7.444625
2005	9.063327
2006	7.921084
2007	7.598684
2008	20.28612
2009	13.64777
2010	12.93887
2011	11.91609
2012	9.682352
2013	7.692156
2014	7.189384
2015	2.529328
2016	3.765119
2017	4.085374
2018	5.078057
2019	10.57836
2020	9.739993
2021	9.496211
2022	19.87386
2023	30.76813

[3] https://data.worldbank.org/indicator/FP.CPI.TOTL.ZG