ANALYZING FACTORS INFLUENCING URBAN POPULATION VARIATION IN SRI LANKA.

Methodology

&

Descriptive Analysis

STAT 31631- Statistical Modeling

Department of Statistics and Computer Science

University of Kelaniya.

Academic Year 2022/2023

By
Group 04

Table of Contents

1. Methodology

By following the complete methodology, we ensured that our data was precisely produced and thoroughly evaluated, establishing a foundation for subsequent inferential statistical analysis and model development to better understand the factors impacting urban population variance in Sri Lanka. This method will help us create a more accurate predictive model for future use.

1.1 Data Collection

Our study aims to identify the factors influencing urban population variation in Sri Lanka using a comprehensive dataset taken from different authoritative sources. Our primary sources of data are as follows:

1. **Department of Census and Statistics**
   This is the main source of demographic and economic data in Sri Lanka. We collected historical data on urban population growth, gross capital development, and the employment-to-population ratio.

2. **Central Bank of Sri Lanka**
   This source provided us with significant economic indicators including gross domestic income, GDP per capita, and tax revenue.

3. **World Bank**
   We utilized the World Bank database to enhance our urban population data, ensuring consistency and completeness, especially for the global comparison metrics.

From these sources, we gathered data on the following one dependent and seven explanatory variables:

1. Urban population growth (annual %) [Dependent Variable]
2. Urban population
3. Gross capital formation
4. Gross domestic income
5. GDP per capita
6. Tax revenue
7. Employment-to-population ratio (15+)
8. Literacy rate (*Note: This variable was later eliminated due to insufficient data*).

Our data-collecting timeline runs from 1960 to 2022. However, due to inconsistencies and missing data for several variables, we limited our analysis to the period 1990–2022.

1.2 Data Preprocessing

After gathering the raw data, we meticulously preprocessed it to guarantee that the dataset was usable and reliable. This phase included numerous essential steps:

1.2.1 Filtering Timeframe

We initially planned to cover the years 1960 to 2023, but due to severe data gaps in previous decades, we reduced our attention to the years 1990 to 2023, when the data was more consistent and comprehensive.

1.2.2 Data Completeness Check

We examined each variable's percentage of available data within the specified timeframe. Variables with less than 60% of the data available between 1990 and 2023 were omitted from further study. As a result of the large number of missing data points, the literacy rate variable was deleted.

1.2.3 Handling Missing Data

For the remaining variables, we addressed missing values by employing a mean imputation technique. Each missing data point was replaced with the mean of the corresponding variable derived from the available data. This method helps to keep the overall trend and variety in the data while avoiding severe bias.

1.3 Descriptive Analysis

In our descriptive analysis of the cleaned dataset, we computed essential statistical summaries to obtain insight into each variable's central tendencies and dispersion. We computed the mean and median to determine the average and midpoint values, respectively, while the standard deviation measured the variability around the mean. In addition, we examined the minimum and maximum numbers to better understand the data range. These statistical summaries provide a thorough summary of the dataset's key properties and variability, establishing the framework for future study.

1.4 Future Analytical Methods

To acquire deeper insights and ensure robust analysis, we will perform the following procedures.

1.4.1 Univariate Analysis

This stage will require analyzing each variable individually to better understand its distribution, central tendency, and dispersion. It provides the foundation for understanding the behavior of each variable in isolation.

1. **Distribution Analysis**

   We created histograms and density plots for each variable to visualize their distributions. This helps to discover any skewness, kurtosis, or outliers in the data.

2. **Correlation Analysis**

   To investigate the relationships between the dependent variable (urban population growth) and the explanatory variables, we calculated Pearson correlation coefficients. In addition, we created scatter plots to visualize the linear relationships between variables. This approach was useful in detecting potential linear correlations and guiding subsequent regression analyses.

3. **Outlier Detection**

   We created box plots to identify outliers in the data. Outliers were investigated to see whether they were genuine observations or data input errors, and necessary measures were implemented.

1.4.2 Multiple Linear Regression Analysis

We will create a full regression model that includes all explanatory variables to estimate their combined impact on urban population growth. This will include:

1. **Model Fitting:** Estimating the coefficients for each predictor.
2. **Residual Analysis:** Checking the residuals to ensure linearity, homoscedasticity, and normality.
3. **Variable Selection:** Using approaches such as stepwise selection (forward and backward) to determine the most important predictors. This will include Model Comparison and Residual Analysis for Selected Models

1.4.3 Predictive Model Enhancement

We aim to develop a predictive model with higher accuracy by validation and refrainment.

2. Descriptive Analysis

2.1 Statistical Summaries

Variable	Mean	Standard Deviation	Minimum	Q1	Median	Q3
Employment to pop ratio 15+	50.08	1.3	47.55	49.38	50.27	50.75
GDP per capita	2084.04	1435.47	466.9	832.43	1517.18	3643.83
Gross capital formation	28.24	4.81	22	24.43	27.75	28.24
Gross domestic income	7.045E+12	3.4076E+12	2.91E+12	4.49E+12	6.565E+12	7.65E+12
Tax revenue	6.14294E+11	5.39319E+11	61206000000	1.47E+11	4.685E+11	1.01E+12
Urban population	3654652.27	285740.84	3188779	3417229	3644303.64	3869404
Urban population growth (annual %)	0.88	0.27	0.48	0.68	0.88	0.99
Year	2006	9.52	1990	1998	2006	2014

Variable	Maximum	IQR	Skewness	Kurtosis
Employment to pop ratio 15+	53.43	1.35	0.07	-0.1
GDP per capita	4388.2	2736.89	0.37	-1.62
Gross capital formation	39.73	3.63	0.81	-0.38
Gross domestic income	1.33E+13	2.965E+12	0.71	-0.91
Tax revenue	1.73E+12	8.33E+11	0.75	-0.82
Urban population	4220157	441868.25	0.22	-1.06
Urban population growth (annual %)	1.86	0.29	1.17	2.56
Year	2022	15.5	0	-1.26

The descriptive statistics for the selected variables from 1990 to 2023 show important information about their central tendencies and variability. The employment-to-population ratio for people aged 15 and up has a mean of 50.08 and a standard deviation of 1.30, with values ranging from 47.55 to 53.43, and a median of 50.27. GDP per capita has a mean of 2084.04 and a standard deviation of 1435.47; the minimum is 466.90, the maximum is 4388.20, and the median is 1517.18. Gross capital formation averages 28.24 with a standard deviation of 4.81, with values ranging from 22.00 to 39.73 and a median of 27.75. Gross domestic income has a significant mean of 7045000000000.00, a standard deviation of 3407603352219.94, values ranging from 2910000000000.00 to 13300000000000.00, and a median of 6565000000000.00. Tax revenue has a mean of 614293968750.00, a standard deviation of 539318544801.23, a range of 61206000000.00 to 1730000000000.00, and a median of 468500000000.00. The average urban population is 3654652.27, with a standard deviation of 285740.84, ranging from 3188779.00 to 4220157.00, and a median of 3644303.64. Finally, urban population growth (annual%) has a mean of 0.88 and a standard deviation of 0.27, with a minimum of 0.48, a maximum of 1.86, and an average of 0.88.

3. Source Code

3.1 Preprocessing

```
#install.packages("readr")

#install.packages("dplyr")

#install.packages("summarytools")

#install.packages("tidyr")

# Load necessary libraries

library(readr)

library(dplyr)

library(summarytools)

library(tidyr)

# Load the CSV file

data <- read_csv("C:/Users/HP/Desktop/kaleniya/3rd year/statistical_modeling/Urban_Population_Data.csv")

# View the first few rows of the data

head(data)

# Calculate the percentage of missing values in each column

missing_percentage <- colSums(is.na(data)) / nrow(data) * 100

# Identify columns with more than 60% missing values

cols_to_remove <- names(missing_percentage[missing_percentage > 60])

# Remove columns with more than 60% missing values

data_cleaned <- data %>% select(-all_of(cols_to_remove))

# Replace missing values with mean values for the remaining columns

# Compute means for columns that are numeric

mean_values <- sapply(data_cleaned, function(col) {

  if (is.numeric(col)) {

    mean(col, na.rm = TRUE)

  } else {

    NA
  }
})
# Replace NA values with computed means
data_cleaned <- data_cleaned %>%
  mutate(across(where(is.numeric), ~ replace_na(., mean_values[cur_column()])))
# View the cleaned data
head(data_cleaned)
# save the cleaned dataset to a new CSV file
#write_csv(data_cleaned,"C:/Users/HP/Desktop/kaleniya/3rd year/statistical_modeling/Cleaned_Dataset.csv")
```

3.2 Descriptive Statistics

```
# Summary statistics for the entire dataset
summary(data_cleaned)

# Detailed descriptive statistics
# Descriptive statistics for all columns
descr(data_cleaned)

# summary statistics for specific columns
for (col_name in colnames(data_cleaned)) {
  cat("\nSummary for column:", col_name, "\n")
  print(summary(data_cleaned[[col_name]]))
}
```