**Exploring Body Mass Index: Unveiling key predictors and sociodemographic sensitivity**

**Doctor Rui. Sun**

**Students: San Nguyen, Mazen Alhaffar, and Rohitha Sai Kothapalli**

Table of Contents

[I. Introduction: 3](#_Toc183611538)

[II. Literature Review: 3](#_Toc183611539)

[III. Data, Data Sources, and Data Characteristics: 6](#_Toc183611540)

[1. Obtaining data: 6](#_Toc183611541)

[2. Data cleaning & Preprocessing: 6](#_Toc183611542)

[3. Summary Statistics 7](#_Toc183611543)

[4. Data Visualization: 8](#_Toc183611544)

[5. Data Dimension Reduction: 10](#_Toc183611545)

[IV. Methodology: 10](#_Toc183611546)

[1. Presenting Model: 10](#_Toc183611547)

[2. Practical Considerations: 18](#_Toc183611548)

[3. Approach and Methodology 18](#_Toc183611549)

[a) Justify Model selection: 18](#_Toc183611550)

[b) Preprocessing: 19](#_Toc183611551)

[V. Empirical Results: 20](#_Toc183611552)

[1. Key results: 20](#_Toc183611553)

[a) Intercept: 20](#_Toc183611554)

[b) Year Variable: 20](#_Toc183611555)

[c) Real GDP: 20](#_Toc183611556)

[d) Population: 21](#_Toc183611557)

[e) Urbanization Rates: 21](#_Toc183611558)

[f) Overall\_Globalization\_Index: 21](#_Toc183611559)

[g) Human\_Capital\_Index: 21](#_Toc183611560)

[2. Conclusion: 22](#_Toc183611561)

[VI. Evaluating the Model’s Performance and Discussing Its Advantages and Weaknesses: 23](#_Toc183611562)

[VII. Reference: 23](#_Toc183611563)

# Introduction:

The population of adults living with obesity has doubled since 1990 (WHO, 2024) underscoring an urgent need to understand factors that determine our health conditions, particularly Body Mass Index (BMI). BMI is an accessible measurement that offers an overall health status based on height and weight figures. In this study, we seek to analyze the variation of BMI across 200 countries by country-level economic and socio-demographic factors inclduing GDP, Urbanization Rate, Human Capital Index, Population, Year and Globalization Index. By understanding BMI predictors, we can evaluate opportunities for patients, healthcare providers, and businesses, including:

* **1. Patients:** BMIoffers a low-cost assessment to quickly update patients’ health status. Patients can utilize this model to adjust nutrients, meal plans, and exercise routines at different life stages.
* **2. Insurance Industry:** Insurance companies can gain insight from customers regarding health risks associated with obesity to offer more tailored insurance plans and develop incentive programs.
* **3. Healthcare providers:** Healthcare providers can use this approach to evaluate health risks and tailor affordable treatments, enhancing healthcare management.
* **4. Public health initiatives:** Public health organizations can rely on this model to target at-risk communities and create health strategies to prevent and reduce obesity.

**Research questions:** The research question centered on this project is examining the impact of GDP, Urbanization Rate, Human Capital Index, Population, Year and Globalization Index on BMI and determining whether BMI is more sensitive to change in GDP or HCI.

# Literature Review:

In this literature review, we seeks to gather insightful articles to gain a comprehensive understanding of BMI measurement and the complex relationship between these factors and BMI. Valuable perspectives were gained through articles along with various variables which analyze how demographic factors interact with BMI, contributing to a comprehensive analysis of BMI’s demographic sensitivity.

**Physiology, Body Mass Index:** Asia Zierle-Ghosh and Arif Jan’s article summarizes the BMI benchmark for different classes of weight determined by the World Health Organization (WHO) and the Centers for Disease Control and Prevention (CDC). Using this benchmark as their basis, they elaborated on the issues surrounding BMI as a reliable measure, as it fails to calculate body fat percentage. This leads to miscalculating BMI of people with lean or bulky body mass. Moreover, it neglects to measure the abdominal obesity in the body, a major factor when it comes to high morbidity and mortality rates. More importantly, BMI fails to take into account the difference in body size between men and women, meaning BMI is lacking when it comes to different gender groups. Despite these facts, it has been proven that an increased BMI leads to a higher risk of developing several diseases, namely cancer, diabetes, and hypertension. This article helps us understand the shortcomings of BMI as a valid measure to predict diseases, laying the groundwork to answer our research question, and providing insight on the practical uses of the measure.

**Factors affecting BMI; Assessment of the effect of sociodemographic factors on BMI In the population of Ghulam Mohammad Abad Faisalabad:** The study carried out by Dr. Abdul Sattar, Dr. Shahbaz Baig, Dr. Naveed ur Rehman, Dr. Badar Bashir provided a comprehensive understanding about the impact of sociodemographic variables, including age, marital status, income, location on BMI. This study indicated that income and age have a positive correlation with BMI while the figure for females tends to be higher than that of males. This result aligns closely with our central research question to investigate the sensitivity of demographic factors on BMI. Additionally, this study explored further lifestyle factors on BMI such as working hours, sleeping patterns, physical activity, smoking status which were significantly associated with BMI, whereupon suggesting tailored health interventions to manage and reduce obesity among different populations. These findings highlight interaction of demographic factors with BMI and emphasize the need to address lifestyle and additional predictors to enhance our analysis.

**What use is the BMI? Archives of Disease in Childhood:** This article, written by D.M.B Hall and T.J. Cole, delves into the use of BMI to reduce obesity among children, exploring the methodologies that are used to identify early obesity signs and help communities measure and control increasing rates. Furthermore, the author uses centile-based definitions to illustrate the metric different countries use to measure overweight cases. Two approaches were presented to determine the validity of body fat measures, the first is physicochemical methods, which checks the amount of body water, and the second is imaging which measures the fat mass in the body. Nevertheless, the article also explores the pitfalls in the BMI metric, where various factors could affect the results, especially because the measure is correlated with body fat. This article enriched our understanding of BMI use when it comes to the specific age group of children, a group where accurately applying BMI principles is particularly challenging.

**Gender Differences in the BMI-Income Relationship:** (Instrumental Variable Strategy) Edwards et al used genetic variants from the HUNT study as instruments to address endogeneity. The study finds a significant relationship between BMI and income for women but inconclusive results for men. For women, each 1-unit increase in BMI is associated with a 2% reduction in income. Specifically, the IV model using a weighted genetic risk score (GRS) shows a coefficient of-0.02 for females. The IV models used GRS based on 97 SNPs associated with BMI. The F-statistics for the instruments were well above 10, indicating strong relevance. For women, the weighted GRS IV model produced a coefficient of-0.020, suggesting a strong negative effect of BMI on income. For men, however, the IV models showed no significant effect on income, with results varying based on the instruments used. The study included 50,910 participants under the age of 67. For men, the highest mean income was found among those classified as overweight, while for women, those with normal weight had the highest mean income. Obese women earned significantly less.

**The Association Between Physical Activity and Body Mass Index:** The study highlights that obesity significantly influences the PA-BMI relationship. In non-obese individuals, PA’s impact on BMI is minimal, while in obese individuals, the relationship is much stronger, particularly for vigorous PA. Hemmingsson and Ekelund conducted a study to explore this relationship using objective PA measurements in a diverse sample, including severely obese individuals and non-obese controls. PA was measured using accelerometers over seven days, capturing six categories: sedentary time, light PA, moderate PA, vigorous PA, total activity counts per day, and steps per day. In non-obese participants, the association between PA and BMI was weak. This suggests that obesity may act as a barrier to PA, creating a cycle where lower PA leads to higher BMI, which further reduces PA.

In conclusion, the literature review establishes a fundamental foundation for BMI metrics and analyzes important factors that correlate with BMI. With the integration of lifestyle factors, we aim to seek additional data to support our multiple regression analysis, allowing us to examine these diverse predictors comprehensively and accurately.

# Data, Data Sources, and Data Characteristics:

## Obtaining data:

Our data were retrieved from Mendeley which was collected by NCD Risk Factor Collaboration. It provided detailed information on GDP, Urbanization rate, Human Capital Index, Globalization Index, and Population of countries to figure out the relationship between these variables and identify the impact of those macroeconomic and societal factors on the average Body Mass Index.

## Data cleaning & Preprocessing:

We got the data from Mendeley Data and it was based on the database collected by the NCD Risk Factor Collaboration. The data was gathered from 200 different countries between 1975 and 2014 and had around 19.2 million adult participants, including 9.9 million men and 9.3 women. The data-cleaning process was simple and consisted of removing the data guide and codes along with the link to the database.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Description** | **Type** | **Unit** |
| Average BMI | The average Body Mass Index (BMI) of the population in each country | Numerical (continuous) | BMI (kg/m²) |
| GDP | The Gross Domestic Product (GDP) of the country | Numerical (continuous) | US Dollars |
| Globalization Index | A country’s global integration based on economic, social, and political fields | Numerical (continuous) | Index score |
| Urbanization rate | The percentage of the population living in urban areas | Numerical (continuous) | Percentage (%) |
| Year | The year when Body Mass Index (BMI) data was collected for various countries | Numerical (discrete) | Year |
| Human Capital Index | Measure a child’s potential productivity based on education and health outcomes | Numerical (continuous) | Range between 0 and 1 |
| Population | The population of each country between 1975 and 2014 | Numerical (discrete) | Person/km² |

## Summary Statistics

In this project, we use multiple regression to investigate impact of societal and economic factors over BMI by following variables: Average BMI, Globalization Index, Population, Human Capital Index, Urbanization Rate, GDP and Country.

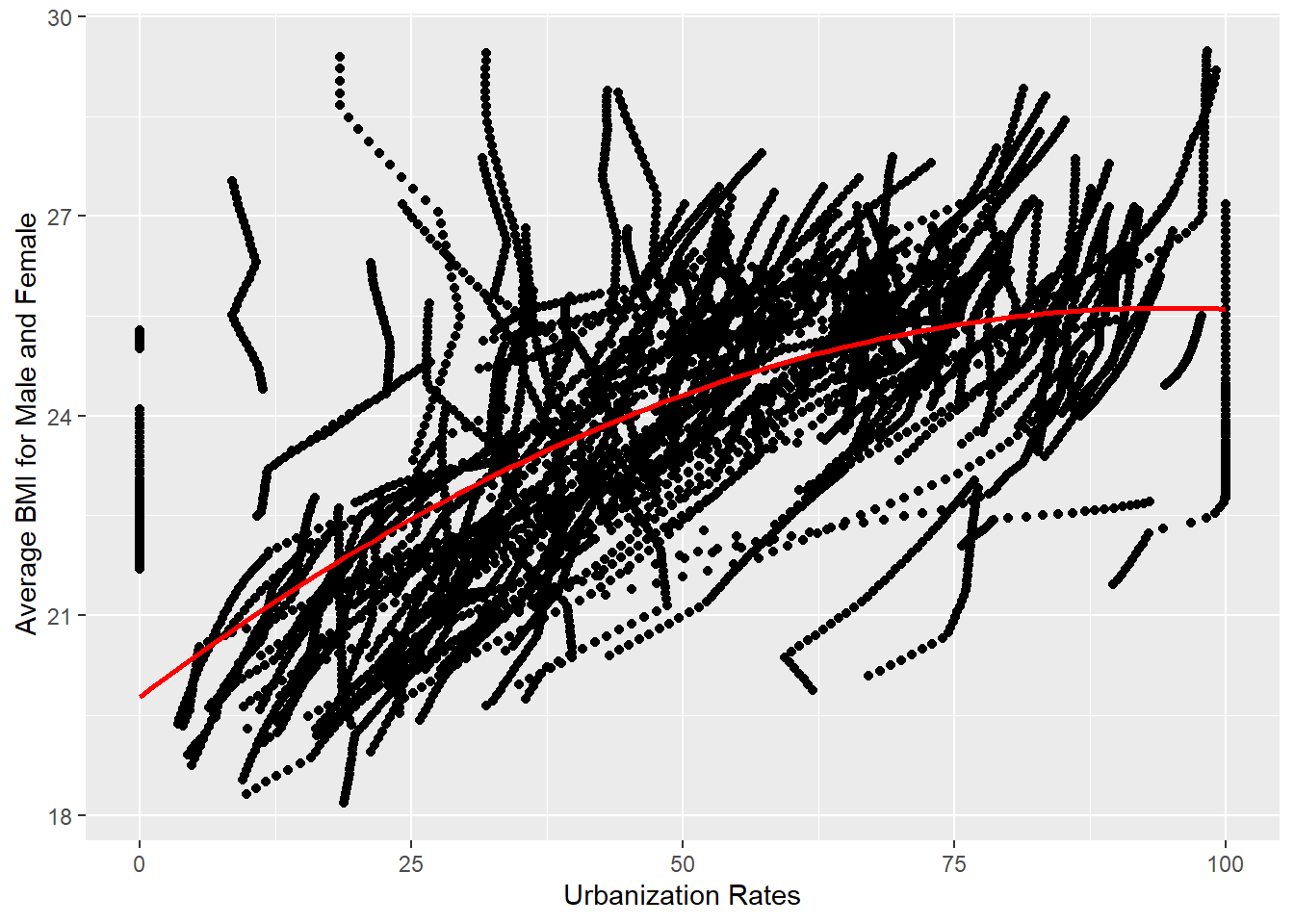
By running this multiple regression, we have the following equation:

Y=β0​+β1​X1​+β2​X2​+β3​X3​+β4​X4​+β5​X5​+ei​

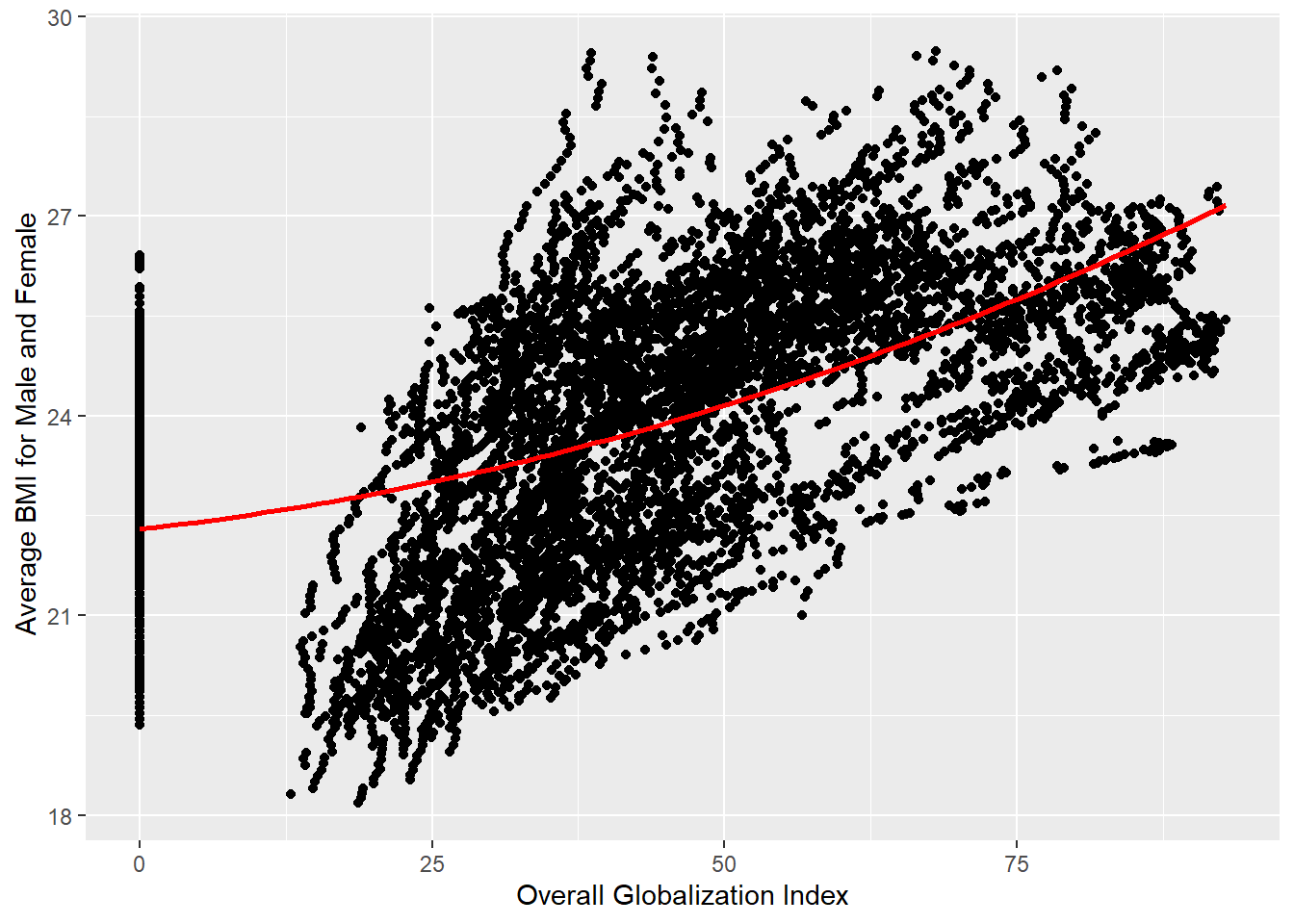
In this scenario, Y represents the average BMI for both females and males within a country. The variables GDP, Globalization Index, and Urbanization Rate are denoted as beta 1, beta 2 and beta 3 , with X1, X2 and X3 indicating the variations in these independent variables including GDP, Globalization Index, Urbanization rate and country. To examine the central tendencies and distributions of these variables, we have provided a table summarizing key statistics.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Min.** | **1st Qu.** | **Median** | **Mean** | **3rd Qu.** | **Max.** |
| **Country** | - | - | - | - | - | - |
| **Year** | 1975 | 1985 | 1994 | 1994 | 2004 | 2014 |
| **Ctry\_dum** | 1 | 43.75 | 86.5 | 86.5 | 129.25 | 172 |
| **BMI\_Male** | 18.37 | 21.95 | 24.06 | 23.74 | 25.44 | 29 |
| **BMI\_Female** | 17.6 | 22.68 | 24.68 | 24.32 | 25.88 | 30.57 |
| **Average\_BMI** | 18.18 | 22.41 | 24.44 | 24.03 | 25.61 | 29.48 |
| **Real\_GDP** | 0 | 6427 | 28112 | 301830 | 166489 | 17080304 |
| **Population** | 0 | 1.34 | 6.239 | 31.741 | 20.203 | 1369.436 |
| **Human\_Capital\_Index** | 0 | 1.171 | 1.824 | 1.744 | 2.62 | 3.734 |
| **Urbanization\_Rates** | 0 | 32.91 | 51.3 | 51.63 | 70.37 | 100 |
| **Overall\_Globalization\_Index** | 0 | 31.12 | 43.12 | 44.45 | 58.45 | 92.84 |

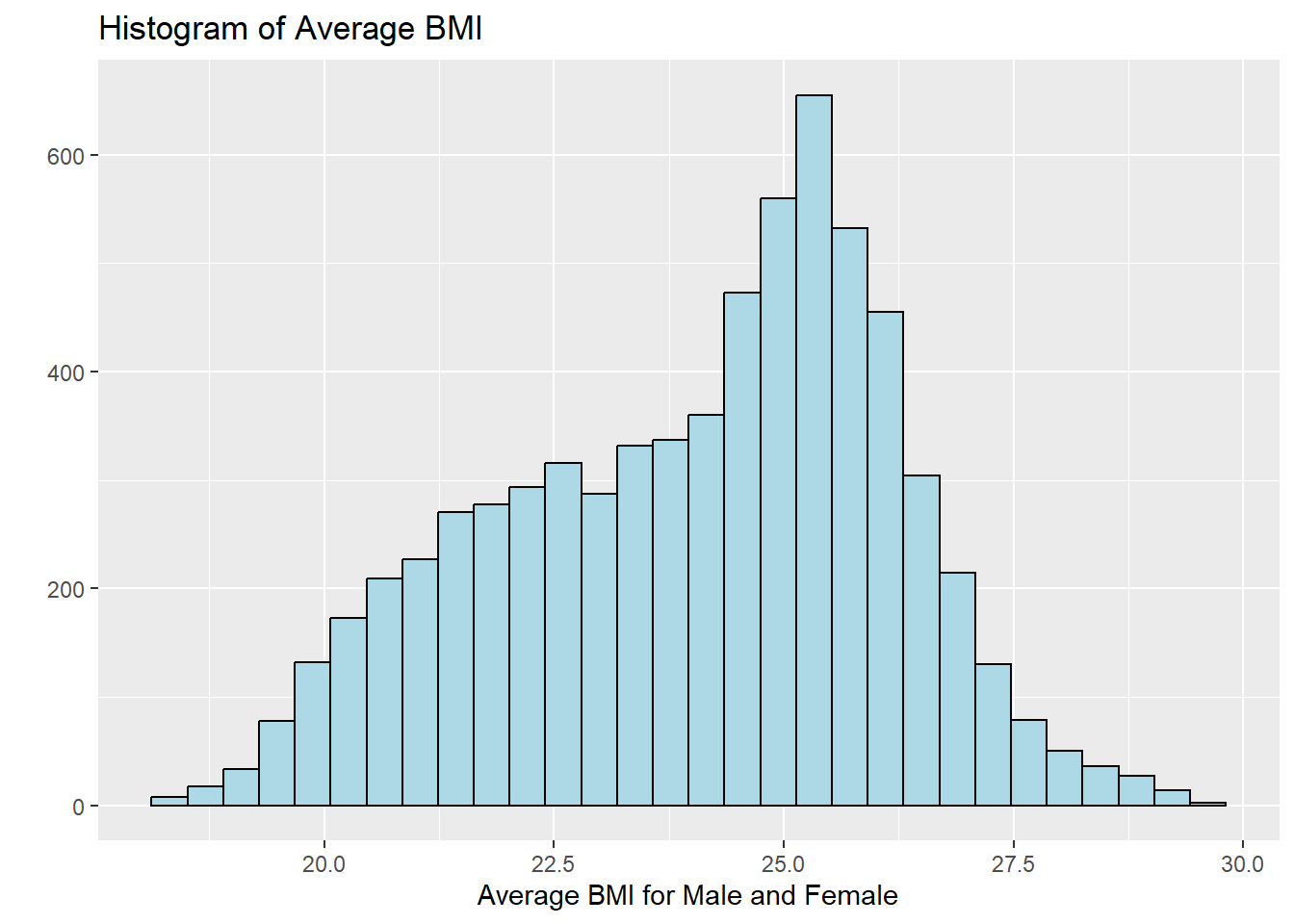
## Data Visualization:



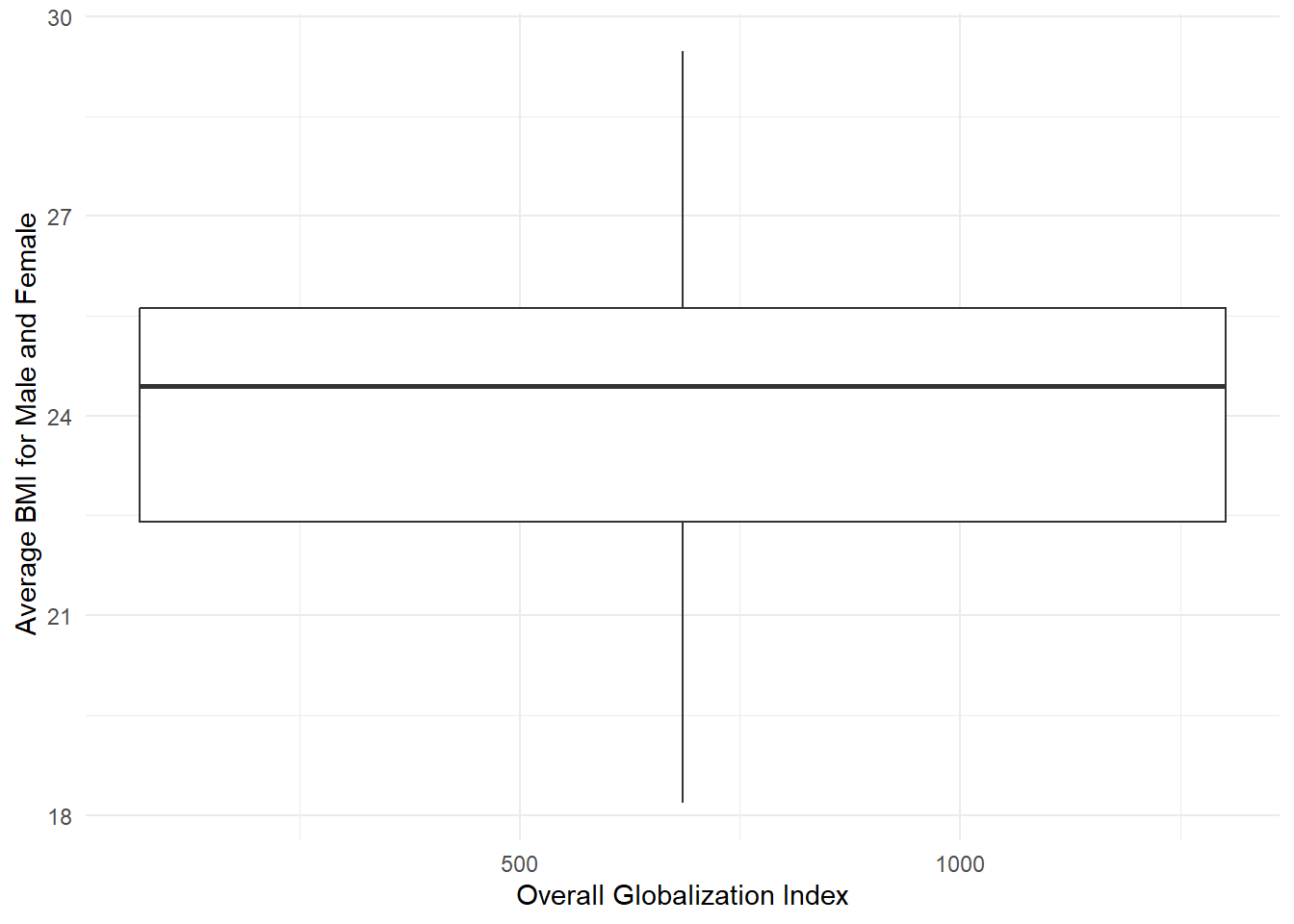
The above scatter plot shows the relationship between Average BMI and Urbanization Rates.



The above scatter plot shows the relationship between Average BMI and Overall Globalization Index.



This is the Histgoram for Average BMI for Male and Female.



The above is the box-plot for the Overall Globalization index and Average BMI.

## Data Dimension Reduction:

In our analysis, we found that Population (POP) was statistically insignificant, indicating it doesn’t significantly impact BMI. Although the Human Capital Index (HC) had a small p-value, suggesting some potential influence, we see challenges in interpreting its relationship with BMI. Consequently, we decided not to include POP or HC in our regression model.

By excluding these variables due to their limited explanatory power, we can concentrate on more relevant factors: Real GDP (RGDP), Urbanization rate (URBAN), and the Globalization index (OGA). This dimensionality reduction enhances the model’s clarity and efficiency.

# Methodology:

## Presenting Model:

In this project, we decided to use multiple linear regression to identify the association of GDP, Human Capital Index, Globalization Index, Urbanization Rate, Population on Average BMI. We run the model by following steps:

***Model 1: lm(formula = Average\_BMI ~ . - BMI\_Male - BMI\_Female, data = BMI)***

***Modification to the Data:*** *Removed unneeded columns for variables Ctry\_dum and Code, transformed Country variable data type from character to factor, and releveled the reference country for the variable Country into the United States of America*

***Key Results:***

* *All variables were statistically significant with a p-value below 0.05, except for the variable Population and the two countries Jordan and Qatar*
* *Adjusted R-squared was at* ***98.37%****, which suggests there is an over-fitting problem with our model*
* *F-statistic had a p-value of* ***< 2.2e-16*** *below 0.05, meaning that there is a statistically significant relationship between the predictors and the outcome variable*

***Validating the Model:***

***1) Testing for Correlation between Variables:***

A table of numbers and symbols

Description automatically generated with medium confidence

The correlation matrix shows that there are some variables that are somewhat correlated like Average BMI with Urbanization Rates, and Urbanization Rates with Human Capital Index.

**2) Testing for Variance Inflation Factor (VIF):**

|  |  |  |  |
| --- | --- | --- | --- |
| **Variance Inflation Factor - Model 1** | **GVIF** | **Df** | **GVIF^(1/(2\*Df))** |
| **Country** | 1.44E+05 | 171 | 1.035348 |
| **Year** | 4.02E+00 | 1 | 2.006052 |
| **Real\_GDP** | 8.87E+00 | 1 | 2.978392 |
| **Population** | 5.50E+01 | 1 | 7.41695 |
| **Human\_Capital\_Index** | 1.10E+01 | 1 | 3.31237 |
| **Urbanization\_Rates** | 3.42E+01 | 1 | 5.849136 |
| **Overall\_Globalization\_Index** | 1.05E+01 | 1 | 3.24534 |

Testing for VIF revealed that the variable Population has multicollinearity with other variables.To produce more accurate results, we decided to remove it our model. When running VIF, and after adjusting GVIF for the variable Country that has more than one degree of freedom, results show that the variable Population has multicollinearity with other variables. Thus, we decided to remove it from our model to receive more accurate results.

**Model 2: lm(formula = Average\_BMI ~ . - BMI\_Male - BMI\_Female - Population, data = BMI)**

**Modification to the Data:** None

**Key Results:**

* All variables were statistically significant with a p-value below 0.05, except for the two countries Jordan and Qatar.
* Adjusted R-squared was at **%98.37**, there was clearly an over-fitting issue with our model, which may cause the model to end up fitting random fluctuations in the data rather than genuine relationships.
* F-statistic had a p-value of **< 2.2e-16** below 0.05, meaning that there is a statistically significant relationship between the predictors and the outcome variable.

After running the model and testing for correlation, it was clear that the R-squared is still very high, being at 98.37%. This means that the model has an over-fitting issue, meaning that the number of observations are minimal compared to the number of predictors, which may cause the model to end up fitting random fluctuations in the data rather than genuine relationships. Next, we executed some plots to test for Nonlinear Association, Heteroscedasticity, Outliers, and High leverage Points.

A collage of graphs and charts

Description automatically generated

The Residuals vs. Fitted graph shows a linear relationship between the outcome variable and the predictors. Q-Q Risduals graph shows that there are outliers with significant deviations in both the extreme negative and positive tails of model, meaning further investigation is needed since these points might affect the validity of the model. Scale location graph suggests there is no heteroscedasticity since the data points are relatively evenly scattered. Risduals vs. Leverage shows that there are no high leverage points that might affect the validity of the model since no point exceeds cook’s distance. At the next stage, we decided to *remove the Outliers and Observing the Effects.*

**Model 3: lm(formula = Average\_BMI ~ . - BMI\_Male - BMI\_Female - Population, data = BMI2)**

**Modification to the Data:** Removed outliers that have been identified in the Q-Q Risduals graph

**Key Results:**

* All variables were statistically significant with a p-value below 0.05, except for the variable Urbanization Rates and three countries Egypt, Jordan, and Qatar
* Adjusted R-squared has increased to reach **98.42%**, meaning the over-fitting issue persisted and we have to start experimenting with our variables to get better results
* F-statistic had a p-value of **< 2.2e-16** below 0.05, meaning that there is a statistically significant relationship between the predictors and the outcome variable

**Validating the Model:**

**Testing for Variance Inflation Factor (VIF):**

|  |  |  |  |
| --- | --- | --- | --- |
| **Variance Inflation Factor - Model 3** | **GVIF** | **Df** | **GVIF^(1/(2\*Df))** |
| **Country** | 4034.849 | 171 | 1.024574 |
| **Year** | 4.064167 | 1 | 2.015978 |
| **Real\_GDP** | 5.706441 | 1 | 2.388816 |
| **Human\_Capital\_Index** | 10.92815 | 1 | 3.305775 |
| **Urbanization\_Rates** | 34.4907 | 1 | 5.872878 |
| **Overall\_Globalization\_Index** | 10.5171 | 1 | 3.243009 |

A collage of graphs

Description automatically generated

We tried to remove the outliers and see if it can improve the results of the model, however the R-squared remained too high and the risduals graphs didn't witness any significant changes. Therefore, we decided to remove the Country variable to test R-squared.

**Model 4: lm(formula = Average\_BMI ~ . - Country - BMI\_Male - BMI\_Female + Human\_Capital\_Index \* Real\_GDP, data = BMI)**

**Modification to the Data:** transformed data type for Year variable from number to factor, with the first year 1975 as the reference year

**Key Results:**

* All variables were statistically significant with a p-value below 0.05, except for the years between 1976 and 1982
* Adjusted R-squared finally improved to reach **49.71%**, meaning the over-fitting issue has been resolved
* F-statistic had a p-value of **< 2.2e-16** below 0.05, meaning that there is a statistically significant relationship between the predictors and the outcome variable

**Validating the Model:**

**Testing for Variance Inflation Factor (VIF):**

|  |  |  |  |
| --- | --- | --- | --- |
| **Variance Inflation Factor - Model 4** | **GVIF** | **Df** | **GVIF^(1/(2\*Df))** |
| **Year** | 1.324528 | 39 | 1.00361 |
| **Real\_GDP** | 83.36055 | 1 | 9.1302 |
| **Population** | 3.838089 | 1 | 1.959104 |
| **Human\_Capital\_Index** | 1.743137 | 1 | 1.320279 |
| **Urbanization\_Rates** | 1.566708 | 1 | 1.251682 |
| **Overall\_Globalization\_Index** | 2.556565 | 1 | 1.598926 |
| **Real\_GDP:Human\_Capital\_Index** | 69.23821 | 1 | 8.32095 |

Testing for VIF shows that no multicollinearity exists between the variables, but collinearity with the interaction term and the variables used for the interaction term is normal. After testing to remove some variables from the model, we found that the variable Country was related to the high R-squared value, and as such we decided to remove it to produce a more valid model. Then, we tested some interaction terms and found that Human\_Capital\_Index\*Real\_GDP is the interaction term that improves the R-squared the most, reaching 49.71%. However, there wasn't any reasonable explanation for the interaction term. As such, we next tested some polynomial variables to study the effects on the model.

**Model 5: lm(formula = Average\_BMI ~ . - Human\_Capital\_Index - Country - BMI\_Male - BMI\_Female + poly(Human\_Capital\_Index, degree = 4, raw = TRUE), data = BMI)**

Testing for polynomial terms landed a better R-squared, with degree 4 producing the most significant results.

**Modification to the Data:** None

**Key Results:**

* All variables were statistically significant with a p-value below 0.05, except for the years between 1976 and 1984
* Adjusted R-squared finally improved to reach **57.85%**, meaning the over-fitting issue has been resolved
* F-statistic had a p-value of **< 2.2e-16** below 0.05, meaning that there is a statistically significant relationship between the predictors and the outcome variable

**Validating the Model:**

**Testing for Variance Inflation Factor (VIF):**

|  |  |  |  |
| --- | --- | --- | --- |
| **Variance Inflation Factor - Model 5** | **GVIF** | **Df** | **GVIF^(1/(2\*Df))** |
| **Year** | 1.368692 | 39 | 1.004032 |
| **Real\_GDP** | 1.9069 | 1 | 1.380905 |
| **Population** | 1.625672 | 1 | 1.275018 |
| **Urbanization\_Rates** | 1.903981 | 1 | 1.379848 |
| **Overall\_Globalization\_Index** | 2.64387 | 1 | 1.625998 |
| **poly(Human\_Capital\_Index, degree = 4, raw = TRUE)** | 3.132391 | 4 | 1.153412 |

Checking for collinearity produced favorable results as no variable is correlated to other variables. Testing for polynomial terms landed a better R-squared, reaching 57.85%. While checking for collinearity produced favorable results as no variable is correlated to other variables.

A collage of graphs and diagrams

Description automatically generated

The graphs show that model5 retained the linear relationship between the outcome variable and the predictors as tested previously. Q-Q Risduals graph shows significant improvement after adding the polynomial terms to the model. Scale location still shows that there is no heteroscedasticity, as the data points are still relatively evenly scattered. Finally, the Residuals vs. Leverage graph still shows no high leverage points that might affect the validity of the model since no point exceeds the cook's distance. However, the model captures the non-linear relationship between the Human capital index and Average BMI, as evident by the variation in the slope amount between each degree in the polynomial term.

In the literature review, Asia Zierle-Ghosh and Arif Jan state that BMI would inappropriately serve as a measurement when it comes to different body sizes between females and males. Therefore, we create a new variable called \*Average BMI\* which average BMI of men and women as a general predictor to run the model. After running the model with Average BMI, we get an R-squared of 0.4954 which implies that 49.54% of Average BMI variation can be explained by these variables. This model aligns closely with our research question since it has a relatively high R-squared to justify the relationship and interaction of Year, Real GDP, Population, Urbanization\_Rates, Globalization Index, and BMI.

## Practical Considerations:

**Availability of data**: The data consists of both categorical time-related variables (e.g., years) and continuous variables (e.g., GDP, Population, Urbanization Rates). For handling this kind of mixed data, multiple linear regression works well.

**Linear Regression vs. Logistic Regression**: Since the outcome variable is the average BMI, which is continuous rather than binary outcome, logistic regression would not be appropriately used in this case.

**Time-series variable**: The model accounts for temporal trends in Average BMI with the year variable as a categorical predictor, contributing a significant dimension to the analysis. When analyzing a model that includes economic variables, particularly the time effect is significant, this model is a standard approach.

**Previous research**: Linear regression models are utilized in health economics and demographics studies to analyze public health data and develop healthcare interventions. For example, the Income and obesity in an urban poor community: a cross-sectional study carried out by Harvard Dataverse analyzed the relationship between low-high income countries and obesity among residents in an urban community. They also used linear regression to identify BMI predictors and investigate the association of those variables.

## Approach and Methodology

### Justify Model selection:

Multiple linear regression model is chosen in this study based on the following reasons:

**Complexity of problems**: The dataset includes 6880 observations with 10 variables which makes multiple linear regression well-suited to identify the relationship between those variables. The goal is to determine the interaction and contribution of several variables such as GDP, Human Capital Index, Globalization Index, Urbanization Rate, and Population on Average BMI at a country level.

**High compatibility**: This dataset contains both numerical and categorical variables to examine the data for country-level analysis. Multiple linear regression can effectively handle these types of variables by using dummy variables.

**Interpretation**: It’s simply easy to comprehend how each predictor affects the outcome variable - Average BMI - when using multiple linear regression models, which aligns closely with our research question in investigating the impact of those variables on BMI.

Additionally, multiple regression consists of several considerations and principles including interaction term, Hierarchical principle, Variance Inflation Factor (VIF) , and diagnostic plots which we can utilize to validate our result.

Interaction term studies any synergy effects of two predictors that might jointly affect the BMI of several countries. If any non-linear relationship exists, we can utilize polynomial regression to extend the relationship to achieve a better-fit model. In this case, our R-squared reached 57% which implies 57% of average BMI can be explained by variables.

Therefore, it is effectively utilized in this study to address the research question in investigating what economic and sociodemographic factors influence BMI on a country level. We assume that the Globalization Index, Human Capital Index, Year, and Urbanization Rate have a linear contribution to BMI, making this model an optimal selection to determine the strength and interpretation of how predictors contribute to BMI.

### Preprocessing:

**Data Transformation**: This dataset includes specific BMI for both males and females, we created a new variable called Average BMI variable which calculates the average BMI for both males and females to improve integration, insightful interpretation, and analysis.

**Dummy Encoding:** Convert the categorical **Year** variable into binary dummy variables to include in the regression model which allows us to analyze country-specific effects while maintaining the linear form of the model.

**Feature Engineering:**  
**Dummy variables**: The year variables is characterized as categorical variables and it’s will be converted into dummy variables to enhance the analysis.

**Model training/validation**:

**Variance Inflation Factor (VIF)**: We utilize VIF to check the multicollinearity of multiple variables to guarantee the accuracy of analysis. In this case, we investigate our variables are largely independent to other variables since their VIF are from 1 to 3 which are generally accepted due to their moderate correlation.

**Model Fit:** R-squared and ajusted R-squared are used to determine the fitness of model which provide 57%of Average BMI can be explained by independent variables.

**Model Analysis:** Assess the residual plots to validate model assumption including distribution, linear relationship, heteroscedasticity, outliers and high leverage points. In our study, residuals vs fitted values are perfectly linear.

**Unique Features and Advantages of the Model**:

**Flexibility - Polynomial Regression**: This model can be extended by additional features of multiple linear regression. To capture the nonlinear relationship of the human capital index, we use the polynomial term and figured out that degree 4 is significant to the model, which increased the R-squared of the model from 49% to 57%.

**Statistical Significance**: The result shows that many predictors show highly significant p-values (< 0.001), suggesting a meaningful impact on Average BMI. This prioritizes which variables we should focus on.

# Empirical Results:

Interpreting the key findings for the model, the model predicts Average BMI, and the predictors are Year, Real GDP, Human Capital Index, Urbanization rates, Population, and Overall Globalization Index.

## Key results:

### Intercept:

The Estimate is 21.87, represents the Average BMI when all other variables are zero.

### Year Variable:

Year is a dummy variable that represents the effects of each specific year with the estimated difference in Average\_BMI compared to a reference year (1975). Years 1985-2014 have a significant positive impact on BMI, indicating a trend of increasing BMI over the years.

### Real GDP:

The Estimate is 1.04e-07, positive and statistically significant coefficient for Real GDP (p < 0.001) suggests that as a country’s GDP increases, so does the Average BMI. However, since the coefficient is very small, the impact of Real\_GDP on Average\_BMI is minimal but statistically significant.

### Population:

The Estimate is -2.455e-03, population has a statistically significant negative impact on BMI. As population size increases, the average BMI slightly decreases.While there seems to be a weak inverse relationship, it’s likely influenced by population differences across countries.

### Urbanization Rates:

The Estimate is 2.994e-02, the Urbanization Rates variable is significant (p < 0.001), and its positive coefficient suggests that more urbanized regions tend to have higher BMIs.Urban areas often show higher average BMI, likely influenced by city living habits—think sedentary lifestyles and dietary shifts.

### Overall\_Globalization\_Index:

The Estimate is 1.472e-02, A positive and statistically significant coefficient for the Globalization Index suggests that as countries become more globalized, average BMI increases.

### Human\_Capital\_Index:

The inclusion of a 4th-degree polynomial for the Human Capital Index suggests a complex, nonlinear relationship between human capital and BMI.  
First degree poly (negative):The Estimate is -1.001e+01, Suggests an initial decrease in BMI with increases in human capital.  
Second degree poly(positive):The Estimate is 1.075e+01,Reverses the trend, suggesting that beyond a certain point, increases in human capital lead to higher BMI.  
Third degree poly(negative):The Estimate is -3.804e+00,Introduces further complexity, suggesting additional curvature in the relationship.  
Fourth degree poly(positive):The Estimate is 4.391e-01,Continues this pattern, confirming the model’s flexibility in capturing a nuanced relationship between human capital and BMI.

## Conclusion:

The model shows a clear trend of increasing Average BMI over time, with later years significantly contributing to the rise. Real GDP has a small positive impact, while Population has a slight negative effect on BMI. The Human Capital Index exhibits a strong nonlinear influence, generally reducing BMI as it increases, but with more complexity captured by the polynomial terms. Both Urbanization Rates and the Globalization Index have small but significant positive effects on BMI. Overall, the model provides a comprehensive view of how socio-economic and demographic factors influence BMI, highlighting both positive and negative contributors.

The research aimed to understand the socio-economic and demographic drivers of Average BMI globally. Real GDP shows a positive impact on BMI, indicating that economic growth contributes to weight gain. The nonlinear relationship of Human Capital Index suggests a more complex interaction between how well countries are investing in the health and education of their people and BMI, with overall reductions in BMI at higher levels of human capital index.

**Interpreattion for non-tech audience**: Over the years, average body mass index (BMI) has been steady. Economic development, urbanization, and globalization contribute to this trend. Interestingly, higher human capital (such as education) tends to correlate with lower BMI, emphasizing the importance of investing in education for better public health outcomes. Meanwhile, as countries become wealthier (as measured by real GDP), managing the health consequences of lifestyle changes becomes critical, especially in rapidly growing economies. Urbanization and globalization also nudge BMI upward, highlighting the need for urban health policies that promote active living and balanced diets. These findings provide valuable guidance for policymakers aiming to combat the global obesity challenge through economic policies, and thoughtful urban planning.

**Statistical Evidence and Model Metrics:** The p-values of variables like Year, Real GDP, Population, and Human Capital Index indicate high statistical significance (p < 0.001). These results provide strong evidence that these factors play important roles in influencing BMI. (R-squared)The model explains 58.14% of the variance in Average BMI (R-square = 0.5814). At 1.385, the model has a relatively low error, meaning it predicts BMI fairly accurately on average.

# Evaluating the Model’s Performance and Discussing Its Advantages and Weaknesses:

After testing several models, we landed on model5 which included the polynomial term for Human Capital Index and removing the variable Country which was causing the overfitting issue in the previous models. The model validates our study's assumption on the influence of socio-demographic factors on BMI, and results show that the factors year, population, real GDP, Human Capital Index (HCI), urbanization rates, and Overall Globalization Index have a statistically significant relationship with the outcome variable BMI, since the p-value was significant for all these variables our model.

Our model shows that the aforementioned factors has 57.85% influence on BMI's results, and although this is a significant relationship, it's still not perfect. Moreover, the added polynomial term captures the complex relationship between HCI and BMI, as they have a non-linear relationship indicated by the resulting slope for each degree in the model.

# Reference:

Hall, D. M. B., and Cole, T. J. (2006). *What use is the BMI?* Archives of Disease in Childhood. <https://adc.bmj.com/content/91/4/283.info>

Sattar, A., Baig, S., Rehman, N. U., & Bashir, B. (2013*). Factors affecting BMI: Assessment of the effect of sociodemographic factors on BMI In the population of Ghulam Mohammad Abad Faisalabad*. Professional Medical Journal. 20\*(6), 956-964. <https://applications.emro.who.int/imemrf/Professional_Med_J_Q/>

Zierle-Ghosh, A., and Jan, A. (2023, November 5). *Physiology, Body Mass Index.* <https://www.ncbi.nlm.nih.gov/books/NBK535456>

Edwards, C. H., Bjorngaard, J. H., & Kinge, J. M. (2020, June 13). *The relationship between body mass index and Income.* Health Economics. <https://onlinelibrary.wiley.com/doi/full/10.1002/hec.4285>

Hemmingsson, E., & Ekelund, U. (2007, April 30). *The association between physical activity and body mass index. Nature Publishing Group. Research Gate*. [(PDF) Is the association between physical activity and body mass index obesity dependent? (researchgate.net)](https://www.researchgate.net/publication/6837299_Is_the_association_between_physical_activity_and_body_mass_index_obesity_dependent)