Final Project

Earth Girls: Amanda Huang, Diyang Lu, Nan Nie, Siqi Ma

McKelvey School of Engineering, Washington University in St. Louis

INFO 574: Foundations of Analytics

Prof. Greene

December 13, 2023

1. Background and Hypothesis

1.1. Background

In recent years, global trends have shown a complex interplay between economic indicators, environmental concerns, and health issues. Economically, many countries have been navigating the aftermath of the COVID-19 pandemic, characterized by recovery efforts and adjustments to new norms. This period has witnessed fluctuations in economic output, employment rates, and government debt levels.

Environmentally, the focus has been on carbon emissions and their impact on climate change. A significant trend observed has been the gradual rebound of carbon emissions following a temporary decline due to the pandemic-induced slowdown (Liu, 2022). This rebound highlights the ongoing challenge of balancing economic recovery with environmental sustainability. In terms of public health, obesity rates among adults have been a growing concern, often linked to lifestyle changes and socioeconomic factors. There's been an increasing awareness of the impact of obesity on individual health and its broader implications on healthcare systems.

Several studies have delved into this complex relationship. For example, research has noted that there is no direct cost estimate assigned to greenhouse gas emissions due to obesity, but it acknowledges the intertwining of obesity, economic factors, and health risks (Hammond, 2010). Also, studies have found that economic growth in the United States was correlated with increasing CO2 emissions, which have indirect implications for obesity through lifestyle changes (Flechtner-Mors et al., 2015). Tomiyama (2019) also explored the relationship between economic growth and obesity and highlighting implications for both health and environmental impacts such as carbon emissions. Additionally, there is evidence to suggest that rising obesity rates may contribute to greenhouse gas emissions both directly through increased food production and indirectly through lifestyle changes (Bryant, 2022).

By synthesizing findings from existing literature, our research can build on the understanding of how economic and emission factors interact with obesity prevalence.

1.2. Hypothesis

Given these trends, our study aims to explore the potential influence of various economic and environmental factors on obesity rates among adults. We hypothesize that there is a significant relationship between these variables, where changes in economic indicators and carbon emissions may correlate with variations in obesity prevalence.

1.3. Formal Hypotheses

- Null Hypothesis (H_0): There is no significant impact of the ten variables (MLN_USD, USD_CAP, AVWAGE, FERTILITY, EMP, UR, HRWKD, GGDEBT, YNGPOP, MtCO2) on adult obesity rates.
- Alternative Hypothesis (H_1): At least one of the ten variables (MLN_USD, USD_CAP, AVWAGE, FERTILITY, EMP, UR, HRWKD, GGDEBT, YNGPOP, MtCO2) significantly impacts adult obesity rates.

2. Datasets

2.1.Data Selection

We examine the interplay between economic factors, environmental impact, and health indicators across various countries. To this end, we utilize three primary datasets, complemented by an auxiliary dataset for standardizing country codes.

1. **Country Economic Indicators:** This dataset offers a comprehensive view of the economic health and characteristics of countries. Sourced from the Organization for Economic Co-operation and Development (OECD), it includes indicators like gross economic output (MLN_USD), per capita income (USD_CAP), average wages (AVWAGE), fertility rates (FERTILITY), employment (EMP), unemployment rates (UR), average hours worked (HRWKD), government debt (GGDEBT), and the proportion of young population (YNGPOP). These indicators are pivotal in understanding the economic landscape and demographic dynamics of each country.
2. **Emissions by Country:** Addressing the environmental aspect, this dataset, obtained from The Global Carbon Project (GCP) quantifies carbon emissions (MtCO2) for each country. Carbon emission levels are a crucial measure of a country's environmental impact and contribute significantly to global climate change discussions.
3. **Obesity among Adults by Country:** From the World Health Organization, this dataset provides statistics on the prevalence of obesity (Obesity) among adults in different countries. It's a vital health indicator, reflecting the nutritional and lifestyle aspects impacting the population.

In our comprehensive analysis, we delve into the relationship between economic, environmental, and health indicators by meticulously combining data from multiple sources. Below is an elaboration on how we curated and merged these datasets to create a rich, multi-faceted dataset for our analysis.

2.2 Data Preparation and Cleaning

Obesity Data Preparation: Initially, we imported the 'Obesity among adults by country' dataset. We refined this dataset by removing gender-specific data to focus only on overall obesity rates ('both sexes'). Unnecessary columns, like an internal index and the 'Sex' column, were removed for clarity. We cleaned and reformatted the 'Obesity (%)' column to represent clear, numeric values and renamed it to 'Obesity'. A reset of the dataframe index was done to ensure data integrity.

Country Code Standardization: The auxiliary dataset 'Wikipedia ISO Country Codes' was imported to facilitate the standardization of country names into their respective ISO codes. We retained only the necessary columns, primarily the country name and its corresponding three-letter country code. The obesity dataset was then merged with this country code dataset to replace country names with standardized country codes.

2.3 Variable Extraction and Integration

Economic Indicators Extraction: Multiple economic datasets were imported from the 'Country Economic Indicators' package, each focusing on a specific indicator like GDP, average wage, fertility rates, etc. For each dataset, relevant columns were selected, and any unnecessary ones were removed to streamline the data. We used the pivot table method to reshape certain datasets, ensuring that different measures (like GDP in millions of USD and GDP per capita) were represented as separate columns. Each of these datasets was then cleaned and standardized, focusing on removing non-numeric or null values.

Emissions Data Integration: We imported the 'Emissions by Country' dataset, focusing on the metric tons of CO2 emissions. Country names in this dataset were also replaced with standardized ISO country codes for uniformity.

2.4 Data Merging and Refinement

To create a cohesive dataset, we merged all the cleaned and standardized datasets based on the common keys of 'Country_Code' and 'Year'. Each merging operation was performed using an inner join to ensure that only records with complete data across all datasets were included. The final merged dataset provides a comprehensive view of each country's obesity rates, economic indicators, and CO2 emissions across different years.

2.5 Key Variables Overview

Table 1 provides a detailed overview of the variables extracted from our primary and auxiliary datasets.

Table 1： Key Variables and Their Descriptions in the Data Analysis

	NO.	**Variable Name**	**Description**
Country Economic Indicators Dataset	1	MLN_USD	Total economic output of the country in millions of U.S. dollars
	2	USD_CAP	Economic output per capita in U.S. dollars
	3	AVWAGE	Average annual wages in the country
	4	FERTILITY	Fertility rate, representing the average number of children a woman will have
	5	EMP	Employment rate, the percentage of the working-age population that is employed
	6	UR	Unemployment rate, the percentage of the labor force that is unemployed
	7	HRWKD	Average hours worked per week

	8	GGDEBT	Gross government debt as a percentage of GDP
	9	YNGPOP	Percentage of the population that is young (usually defined as under 15 or 18 years)
Emissions by Country Dataset	10	MtCO2	Emissions measured in metric tons of CO2
Obesity among Adults by Country Dataset	11	Obesity	Prevalence of obesity among adults, often expressed as a percentage of the population

The datasets are intertwined by the key identifiers of ‘Year’ and ‘Country_Code,’ ensuring a cohesive and comparative analysis across different dimensions. Additionally, the 'Wikipedia ISO Country Codes' dataset serves as an essential tool for standardizing country names into universally recognized ISO codes, facilitating accurate data merging and comparison.

We concentrated specifically on the data from the year 2016. This decision was driven by several considerations:

a. Relevance and Comparability: The year 2016 represents a period prior to major global disruptions such as the COVID-19 pandemic. This allows for a more consistent comparison across countries without the confounding effects of the pandemic on economic, environmental, and health indicators.
b. Data Completeness and Reliability: The datasets for 2016 were chosen due to their completeness and reliability. We ensured that the data for this year was robust, encompassing a wide range of indicators without significant gaps or inconsistencies.

c. Representation: Despite focusing on a single year, the datasets from 2016 provide a rich and representative snapshot of global trends in economic indicators, CO2 emissions, and obesity rates.

We extract the 2016 data with a data frame called df_2016, which consists of 30 observations, each representing a unique combination of country-specific indicators for that year.

This concentrated approach allows us to perform a detailed and focused analysis of the relationships between economic output, environmental impact, and health parameters in a pre-pandemic context. The findings from this year serve as a valuable benchmark for understanding subsequent global trends and shifts.

3. Methodology

After the preprocessing, we chose a fixed year (2016) for the data, which results in our having 30 distinct countries with features (Year, Obesity, Country_Code, MLN_USD, USD_CAP, AVWAGE, FERTILITY, EMP, UR, HRWKD, GGDEBT, YNGPOP, MtCO2).

3.1 Correlation

The utilization of heatmaps serves as a pivotal tool in data analysis, primarily due to their capacity to distill complex correlation data into a visually accessible and intuitive format. A heatmap represents correlations between variables as a grid of colored squares, where each cell delineates the correlation coefficient between two distinct variables.

In this context, the color assigned to each cell symbolizes the magnitude and nature of the correlation: hues of red typically denote positive correlations, whereas blue hues suggest negative correlations. The intensity of the color correlates with the strength of this relationship.

The values within these cells range from -1 to 1, reflective of the potential spectrum of the Pearson correlation coefficient. Here, a value of 1 signifies a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 implies the absence of any linear relationship.

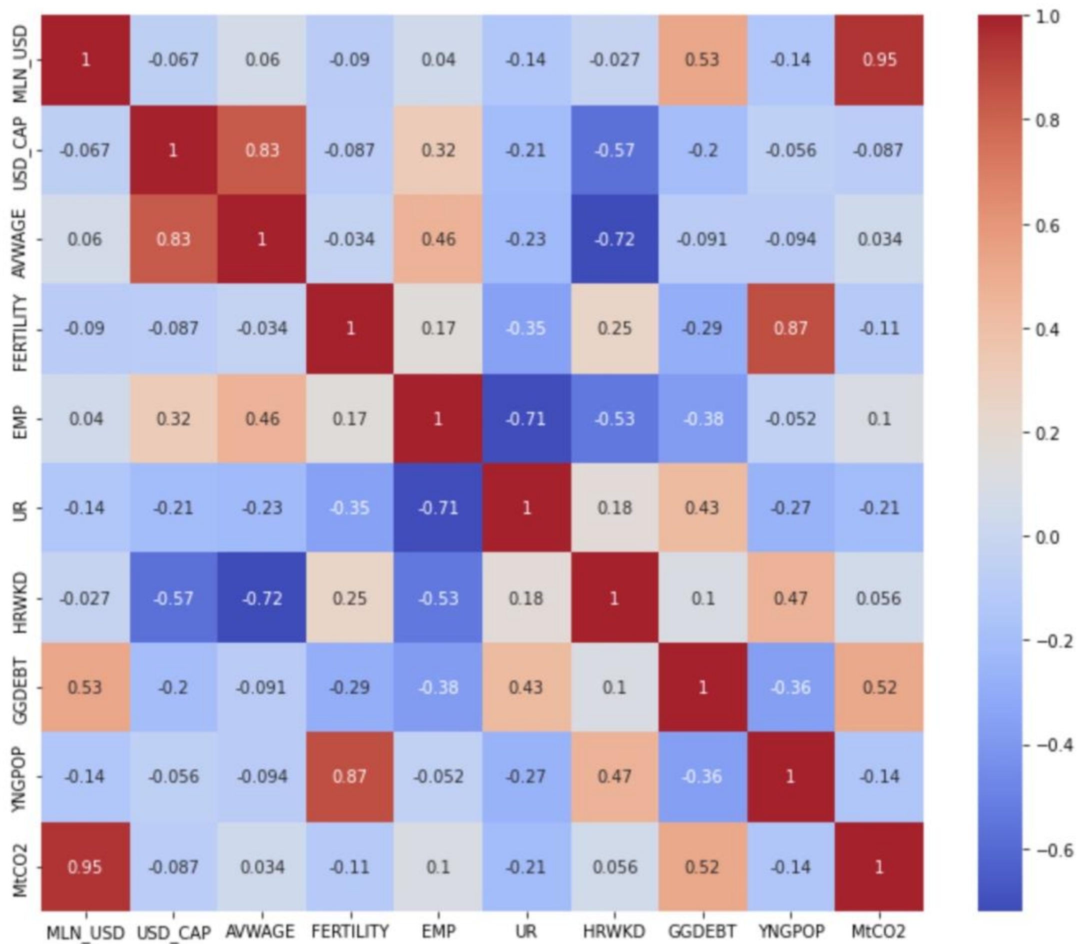


Figure 1. Heatmap of Correlation

The cell intersecting between MLN_USD (Millions of USD) and MtCO2 (Metric tons of CO2 emissions) has a value of 0.95, which indicates a very strong positive correlation, suggesting that as MLN_USD increases, MtCO2 tends to increase as well.

EMP (employment rate) and UR (unemployment rate) show a strong negative correlation of -0.71, implying that higher employment rates are associated with lower unemployment rates.

GGDEBT (government debt) and YNGPOP (young population) have a negative correlation of -0.36, indicating that higher government debt is associated with a lower proportion of young population.

It is imperative to acknowledge that correlation does not equate to causation. These observed relationships indicate trends rather than direct causal links. The interpretation of these correlations necessitates a thorough consideration of contextual factors and other potential influences. For instance, the strong correlation between GDP and CO2 emissions does not definitively establish causality; other factors might concurrently affect these variables. Additionally, a variable exhibiting high correlations with multiple variables concurrently may suggest the presence of multicollinearity, which warrants further investigation.

For initial investigations of the impact of economic and environmental factors on obesity rates, the linear assumption is a reasonable starting point so we start with Ordinary Least Squares (OLS) regression since it is not a classification problem and is simple. Then, before doing regression analysis, we checked 5 assumptions, which are “Collinearity”, “Linearity”, “Independent Residuals”, “Homoscedasticity”, and “Residuals Normality” respectively.

3.2 Assumption 1: Collinearity

Checking the assumption of collinearity in regression analysis is an important step to ensure the reliability and interpretability of the model. Collinearity refers to the situation where two or more predictor variables in a regression model are highly correlated, and it can lead to various issues. Variance Inflation Factor (VIF) is a quantitative measure of the extent of multicollinearity in a regression model. Specifically, the VIF for each predictor variable is calculated based on how much the variance of that variable's estimate is increased due to collinearity with other variables. A high VIF (usually greater than 5) is indicative of a problematic level of collinearity. We tested the VIF score for both attempts.

feature	VIF
const	1422.263135
MLN_USD	15.752480
USD_CAP	3.915619
AVWAGE	6.461392
FERTILITY	7.719707
EMP	4.432994
UR	3.277995
HRWKD	6.475678
YNGPOP	11.715303
GGDEBT	3.124238
MtCO2	18.685135

Figure 1. VIF after deleting Obesity

After deleting our y (Obesity) from the data frame (shown in Figure 1), the VIF for MLN_USD, AVWAGE, FERTILITY, HRWKD, YNGPOP, and MtCO2 are greater than 5. Although the VIF for MtCO2 is the biggest, it is our important independent variable. Therefore, we determined to delete the MLN_USD since MLN_USD and USD_CAP are similar indicators, and the VIF for MLN_USD is the second biggest.

	feature	VIF
0	const	1147.247128
1	USD_CAP	3.915079
2	AVWAGE	6.305317
3	FERTILITY	7.692634
4	EMP	4.065157
5	UR	3.241069
6	HRWKD	4.983642
7	YNGPOP	11.544099
8	GGDEBT	3.121381
9	MtCO2	2.194338

Figure 2. VIF after deleting MLN_USD

After deleting MLN_USD, the VIF for AVWAGE, FERTILITY (shown in Figure 2), and YNGPOP are still higher than 5. We tried two attempts; one was to delete features directly and another attempt was to combine certain features.

i. Attempt 1: Deleting Features

Aiming to obtain a VIF score smaller than 5, we deleted some features one by one. After deleting YNGPOP, and AVWAGE in order, the final VIF table is shown in Figure 3. We can see that the features remaining have reasonable VIF scores. This result is ideal since in the correlation section, we can see that MLN_USD and MtCO2, FERTILITY and YNGPOP, and AVWAGE and USD_CAP have a high correlation. Therefore, dropping these can also address the correlation problem.

feature	VIF
const	1132.203242
USD_CAP	1.573431
FERTILITY	1.383803
EMP	3.251171
UR	3.076783
HRWKD	2.466367
GGDEBT	2.424678
MtCO2	2.114772

Figure 3. VIF after deleting YNGPOP, and AVWAGE

ii. Attempt 2: Combining features

With the aim of retaining more information and enhancing interpretability, we attempted to combine features. Based on the correlation matrix and intuition, we chose to combine the young population (YNG) and fertility (Fer) to represent the extent to which this group contributes to the total fertility rate, and the combination of wage (AVWAGE) and employment (EMP) indicates the overall wage income level. The VIF score after combination is shown in Figure 4.

	feature	VIF
0	const	358.839870
1	USD_CAP	2.781762
2	YNGxFer	1.674867
3	WagexEmployment	4.786740
4	UR	2.108346
5	HRWKD	3.197445
6	GGDEBT	2.353892
7	MtCO2	2.159999

Figure 4. VIF after combining features

3.3 Assumption 2: Linearity

Checking the assumption of linearity in regression analysis is crucial to ensure that the relationship between the predictor variables (X) and the response variable (y) is adequately modeled. The assumption of linearity in regression refers to the assumption that the relationship between the independent variables (predictors) and the dependent variable is linear.

i. Attempt 1: Deleting Features

After deleting certain features, we used the remaining features to test the linearity assumption. As shown in Figure 5, we can see that the remaining features have some degree of linearity, although pretty weak.

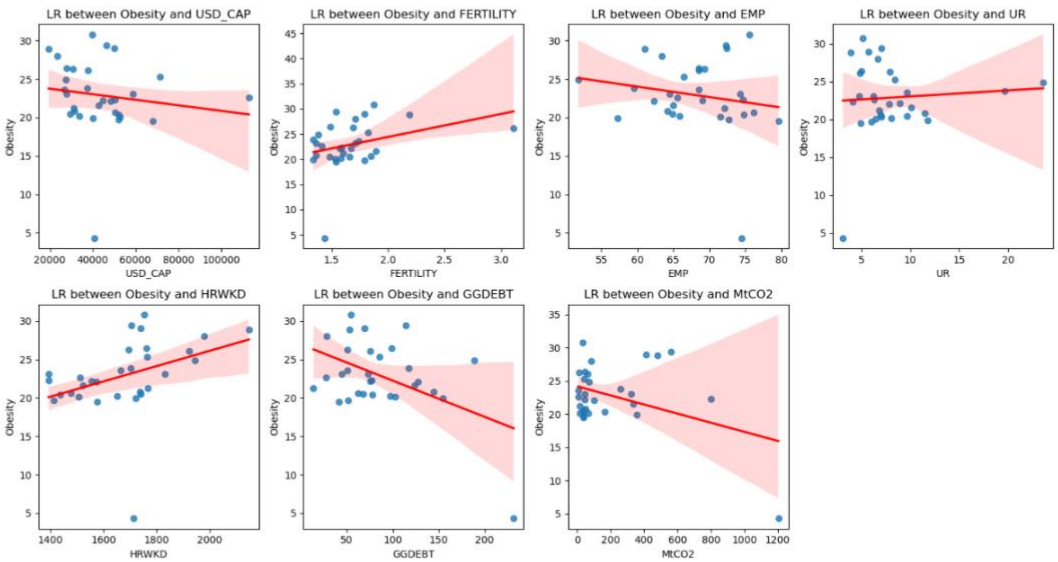


Figure 5. Linear Relationships between features and Obesity

ii. Attempt 2: Combining features

We use the combined features and the remaining features to test the linearity assumption.

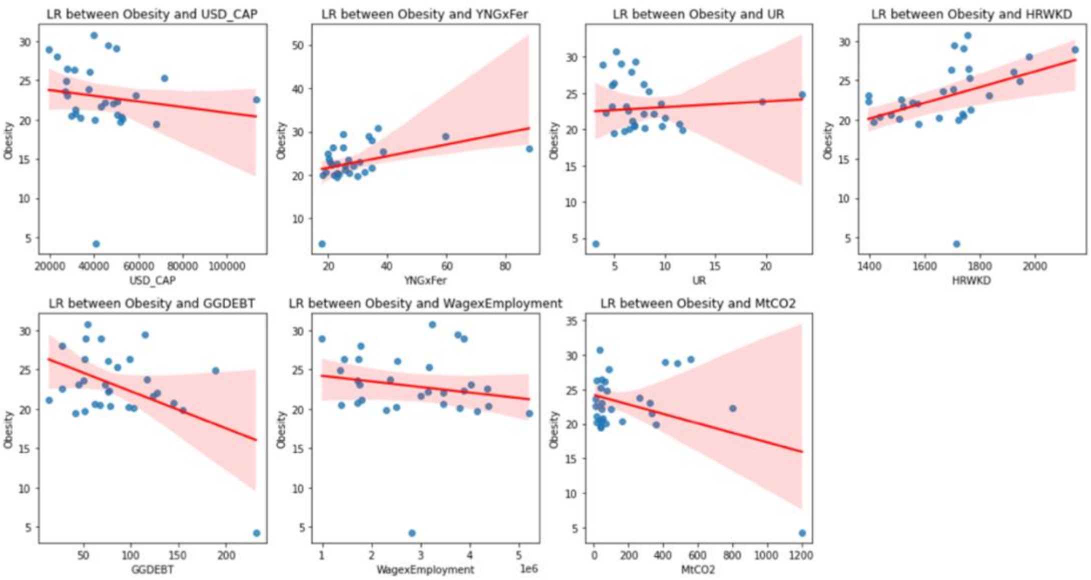


Figure 6. Linear Relationships between features and Obesity

According to Figure 6, we can see there exists linearity between features and the dependent variable, although some are weak.

3.4 Assumption 3: Independent Residuals

Checking the assumption of independent residuals in regression analysis is essential to ensure the validity of statistical inferences and the reliability of the model. The assumption of independent residuals, also known as the assumption of independence or the assumption of independence of errors, implies that the errors (residuals) from the regression model are not correlated with each other. The Durbin-Watson Test is a widely used method for detecting autocorrelation in the residuals from a statistical regression analysis. It helps us determine whether the past values of a variable have an impact on its current values.

i. Attempt 1: Deleting Features

The score of the Durbin-Waston Test turned out to be 1.42, which is some what close to 1.5 indicating no autocorrelation. Hence, satisfied this assumption.

```
from statsmodels.stats.stattools import durbin_watson

#perform Durbin-Watson test
durbin_watson(model.resid)
```

```
1.4177984017671879
```

Figure 7: Durbin-Watson statistic

ii. Attempt 2: Combine features

The Durbin-Watson statistic is 1.47, which is close to 1.5 implying no autocorrelation. Thus, it barely satisfied this assumption, and we would consider autocorrelation not to be problematic.

```
from statsmodels.stats.stattools import durbin_watson

#perform Durbin-Watson test
durbin_watson(model.resid)
```

```
1.4655253404708712
```

Figure 8: Durbin-Watson statistic

3.5 Assumption 4: Homoscedasticity

Homoscedasticity is a key assumption of linear regression, and it refers to the assumption that the variance of the residuals (the differences between observed (X) and predicted values (y)) is constant across all levels of the independent variables. In simpler terms, it means that the spread of the residuals should be roughly the same for all values of the independent variable(s).

i. Attempt 1: Deleting Features

The residuals in Figure 9 show that they form a horizontal band with no discernible pattern. The spread of residuals is relatively constant, so homoscedasticity is met.

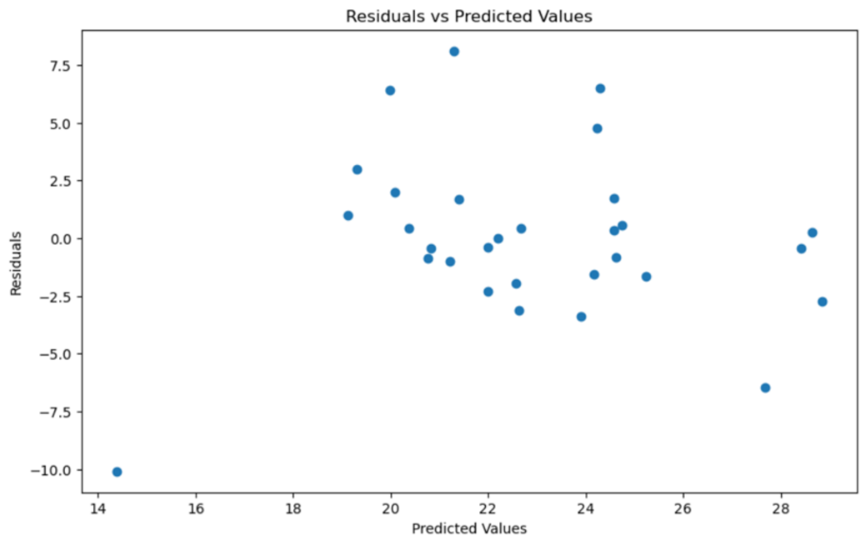


Figure 9. Scatter Plot for Homoscedasticity

ii. Attempt 2: Combining features

As shown in Figure 10, the points are randomly distributed around the horizontal axis with no clear pattern, which indicates the homoscedasticity is met.

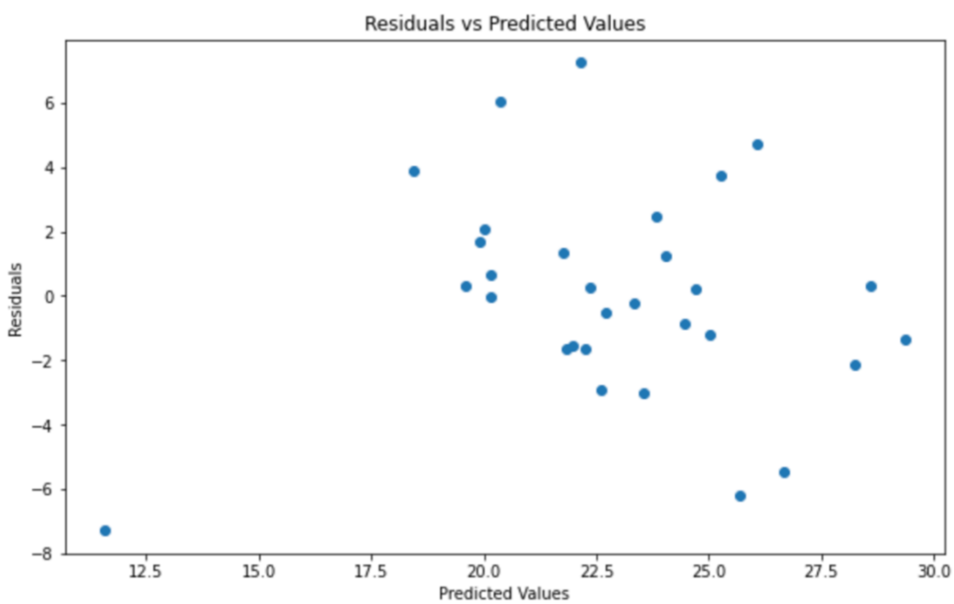


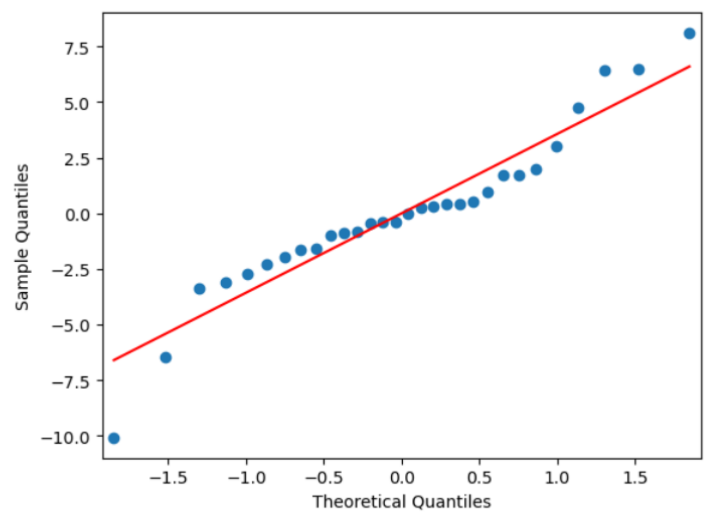
Figure 10. Scatter Plot for Homoscedasticity

3.6 Assumption 5: Residuals Normality

The assumption of normality of residuals is another important aspect of linear regression. It assumes that the residuals (the differences between the observed and predicted values) are normally distributed. We plotted a histogram or a quantile-quantile (Q-Q) plot of the residuals. In a Q-Q plot, the residuals should fall along a straight line. Departures from a straight line suggest deviations from normality. Besides the Q-Q plot, we also utilized the Shapiro-Wilk test, a statistical test to check the normality of a distribution.

i. Attempt 1: Deleting Features

In Figure 11, the data points generally follow the reference line in the plot. The null hypothesis (H0) posits that the data adheres to a normal distribution. The Shapiro-Wilk test statistic approaches 1, suggesting a strong compatibility with a normal distribution. Furthermore, with a p-value exceeding 0.05, there is inadequate evidence to reject H0, implying that the residuals can be deemed sufficiently proximate to a normal distribution.

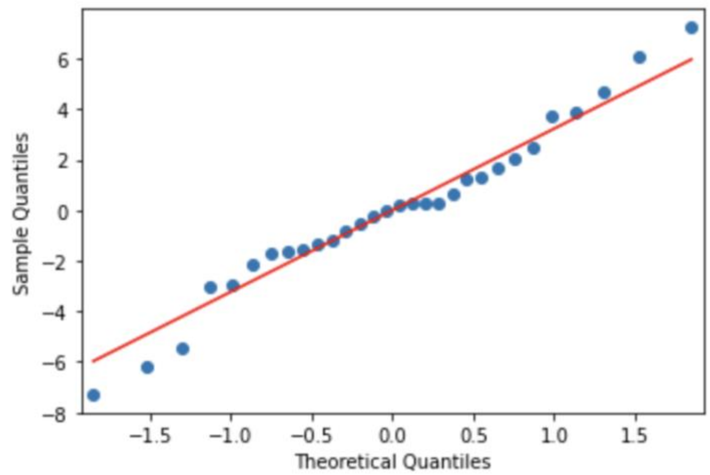



```
Statistics=0.942, p=0.105
Sample looks Gaussian (fail to reject H0)
```

Figure 11. QQ plot for Residuals Normality

ii. Attempt 2: Combining features

As shown in Figure 12, the points are aligned roughly along the reference line in the plot. H0 is that the data are normally distributed. The statistic of the Shapiro-Wilk test is close to 1, which indicates that the data are a good fit for the normal distribution. Meanwhile, the p-value is greater than 0.05, so there is insufficient evidence to reject H0 and the residuals can be considered sufficiently close to the normal distribution.




```
Statistics=0.977, p=0.750
Sample looks Gaussian (fail to reject H0)
```

Figure 12. QQ plot for Residuals Normality

4. Results

After making sure the assumptions were met, we ran a linear regression.

i. Attempt 1: Deleting Features

In Figure 13, we see the results of OLS Regression. In this case, the R-squared is 0.496, indicating that about 49.6% of the variance in Obesity is explained by the model. Adjusted R-squared takes into account the number of predictors, and in this case, it's 33.5%. This might suggest that there is some overfitting or that not all included predictors significantly contribute to explaining the variability in Obesity. Only Government Debt (GGDEBT) shows some form of significance compared to other features.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                Obesity   R-squared:                       0.496
Model:                            OLS   Adj. R-squared:                  0.335
Method:                 Least Squares   F-statistic:                     3.087
Date:                Wed, 13 Dec 2023   Prob (F-statistic):             0.0200
Time:                        10:27:25   Log-Likelihood:                -78.869
No. Observations:                  30   AIC:                             173.7
Df Residuals:                      22   BIC:                             184.9
Df Model:                           7
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         22.8900      0.715     32.015      0.000      21.407      24.373
USD_CAP        0.1242      0.897      0.138      0.891      -1.736       1.984
FERTILITY      0.8733      0.841      1.038      0.310      -0.871       2.618
EMP           -0.2876      1.289     -0.223      0.826      -2.961       2.386
HRWKD          1.5192      1.123      1.353      0.190      -0.809       3.848
UR             1.3822      1.254      1.102      0.282      -1.219       3.983
GGDEBT        -2.8016      1.113     -2.516      0.020      -5.110      -0.493
MtCO2         -0.0487      1.040     -0.047      0.963      -2.205       2.108
==============================================================================
Omnibus:                        2.552   Durbin-Watson:                   1.418
Prob(Omnibus):                  0.279   Jarque-Bera (JB):                1.265
Skew:                           0.340   Prob(JB):                        0.531
Kurtosis:                       3.742   Cond. No.                         3.85
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Figure 13. OLS summary for attempt 1

ii. Attempt 2: Combining features

Before the regression analysis, we standardized the predictor variables using StandardScaler to ensure that each variable was in the same order of magnitude and to avoid disproportionate effects of certain variables on the model due to larger units of measure. According to the result summary shown in Figure 14, we can see the R-squared of the model is 0.532, which indicates that the model explains 53.2% of the total variability. Considering the p-value, only the government debt is smaller than 0.05, demonstrating a significant effect on the obesity rate.

OLS Regression Results

Dep. Variable:	Obesity	R-squared:	0.532
Model:	OLS	Adj. R-squared:	0.383
Method:	Least Squares	F-statistic:	3.567
Date:	Wed, 13 Dec 2023	Prob (F-statistic):	0.0103
Time:	02:22:11	Log-Likelihood:	-77.756
No. Observations:	30	AIC:	171.5
Df Residuals:	22	BIC:	182.7
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	22.8900	0.689	33.224	0.000	21.461	24.319
USD_CAP	-0.9413	1.149	-0.819	0.421	-3.324	1.442
YNGxFer	0.6379	0.892	0.715	0.482	-1.211	2.487
WagexEmployment	1.9268	1.507	1.278	0.214	-1.199	5.053
HRWKD	2.4029	1.232	1.950	0.064	-0.152	4.958
UR	1.6522	1.000	1.652	0.113	-0.422	3.727
GGDEBT	-2.7060	1.057	-2.560	0.018	-4.898	-0.514
MtCO2	-0.3464	1.013	-0.342	0.736	-2.446	1.754

Omnibus:	0.588	Durbin-Watson:	1.466
Prob(Omnibus):	0.745	Jarque-Bera (JB):	0.050
Skew:	-0.018	Prob(JB):	0.975
Kurtosis:	3.196	Cond. No.	4.59

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Figure 14. OLS summary for attempt 2

5. Notable Attempts

In this part, we tried to fit the data into different models besides the OLS regression model, including Adaboosting, gradient boosting, and random forest.

1. AdaBoost

AdaBoost, short for Adaptive Boosting, is a popular machine learning algorithm that belongs to the family of ensemble methods. It is specifically designed for binary classification problems, but it can be extended to multiclass classification and regression tasks as well. The primary idea behind AdaBoost is to combine the predictions of multiple weak learners to create a strong learner.

2. Gradient Boosting

Gradient Boosting is another popular ensemble learning technique, and it's known for its high predictive accuracy and flexibility. It is a machine learning algorithm that builds a series of decision trees, each correcting the errors of the previous one. The basic idea behind Gradient Boosting is to fit a new tree to the residual errors of the current ensemble.

3. Random Forest

Random Forest is another powerful ensemble learning algorithm, and it belongs to the bagging family of techniques. Random Forest builds multiple decision trees during training and merges their predictions to improve accuracy and control overfitting.

We ran through cross-validation using CVgrid to choose the best parameters and use the best parameters to calculate the R2 scores.

Table 2. Performance between AdaBoost, Gradient Boosting, and Random Forest

Models	RMSE	R2 score
AdaBoost	6.3971	0.2178
Gradient Boosting	6.2353	0.2554
Random Forest	6.3660	0.23212

Through Table 2, we can see that all of these models did not perform too well on the given data. We guess it is because we have too little data with too complicated models. These models do not have enough data to keep the loss down. Hence, they are not so suitable for our datasets.

6. Discussion

After collecting relevant materials, we found that the surprising result may reflect complex socioeconomic dynamics (Ogden, C. L. et al, 2010). High government debt may signal broader economic pressures, which may indirectly affect obesity rates in several ways, including reduced access to healthy food, increased pressure on the population, and changes in health-related policies or funding. The lower R-squared values (0.49 and 0.53) suggest that the model explains less than 60% of the variability in obesity rates. Obesity is influenced by a multitude of factors including genetic, behavioral, cultural, and environmental influences that may not be fully captured by economic and environmental indicators.

7. Conclusions

From the hypotheses and results of the study, we rejected the null hypothesis (H0). The results indicate that at least one variable (Government Debt, GGDEBT) has a statistically significant impact on adult obesity rates. Additionally, by comparing attempts 1 and 2, we discovered that combining variables retains more information in the model and improves interpretability, as opposed to simply deleting them.

Our GitHub link: https://github.com/Amanda-L/2023INFO574_Final-Project

References

Bryant, C. J. (2022). Plant-based animal product alternatives are healthier and more sustainable than animal products. Food and Nutrition Sciences, 3(1). https://www.sciencedirect.com/science/article/pii/S2666833522000612

Flechtner-Mors, M., Thamm, M., Wiegand, S., Reinehr, T., Schwab, K. O., Kiess, W., ... & Holl, R. W. (2015). Overweight and Obesity Based on Four Reference Populations in Pediatric Type 1 Diabetes Subjects. Scientifica, 2015. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4466469/

Hammond, R. A. (2010). The economic impact of obesity in the United States. Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy, 3, 285–295. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3047996/

Liu, Z., Deng, Z., Davis, S. et al. Monitoring global carbon emissions in 2022. Nat Rev Earth Environ 4, 205–206 (2023). https://doi.org/10.1038/s43017-023-00406-z

Ogden, C. L., Lamb, M. M., Carroll, M. D., & Flegal, K. M. (2010). Obesity and socioeconomic status in adults: United States 1988–1994 and 2005–2008. NCHS data brief, 50, 1-8.

Tomiyama, A. J. (2019). Stress and Obesity. Annual Review of Psychology, 70, 703–718. https://www.annualreviews.org/doi/10.1146/annurev-psych-010418-102936