Advanced Mathematical Statistics MTH - 522 Project 1

Exploring the CDC 2018 Diabetes Data with Single & Multiple Linear Regression Models

Authors

Bhanu Prasad Thota
Naga Venkata Lokeswarao Maturi
Mantena Harsha Vardhan Varma
Lakkamraju Hitesh Kashyap Varma

1. The issues:

Our big aim in this study is to figure out how three important health issues—obesity, not moving enough, and diabetes—are all connected. These are urgent problems in our society, making the overall burden of diseases worse. We want to really understand how these factors work together and, most importantly, how they affect the health of the public. We're not just looking for connections; we want to give smart suggestions based on evidence to handle and lessen the growing health problems linked to obesity, lack of activity, and diabetes.

Issues:

Data Consistency Concerns:

One major worry we had is that the data we're using is not consistent. It's like trying to solve a puzzle with some pieces that don't quite fit. This inconsistency in data could cause problems in understanding how these health issues are connected.

Data Cleaning Challenges:

Imagine cleaning up a messy room. That's what we had to do with our data by combining different sets of information and giving them new names. But, during this cleaning process, we found that sometimes the names we gave to categories or the ways we combined data didn't match up. It's like organizing a messy closet but realizing that the labels don't match, making it easy to put things in the wrong place. This mix-up could lead to mistakes or errors in our analysis later.

2. Findings

- Diabetes and obesity rates exhibit a positive correlation, with areas experiencing rising obesity prevalence also showing increased diabetes rates. This statistically significant association explains approximately 14.8% of the variance in diabetes rates, indicating a meaningful relationship.
- Conversely, obesity rates demonstrate an inverse correlation with levels of physical activity. Regions with lower physical activity tend to have higher obesity rates, a statistically significant link that accounts for roughly 22.3% of the obesity rate differences.
- Physical inactivity is also positively correlated with diabetes rates, explaining about 19.5% of the variance. In essence, when people are less physically active, there is a higher likelihood of both obesity and diabetes spreading.
- In conclusion, these findings emphasize the importance of addressing both obesity and physical inactivity to combat the diabetes epidemic effectively. Encouraging healthier lifestyles and promoting physical activity can significantly contribute to improving public health and reducing the burden of these conditions
- Overall, these insights underscore the critical need for comprehensive public health strategies to tackle the interconnected issues of obesity, physical inactivity, and diabetes.

3. Discussion

Several significant implications for public health and policy are drawn from our examination of the 2018 CDC data on diabetes, obesity, and physical inactivity. The following conclusions can help direct efforts to address the interrelated health problems of diabetes, obesity, and inactivity:

- Obesity Management Is Essential: The association between obesity and diabetes rates is in favor, which emphasizes the need of tackling obesity as the main diabetes preventative approach. Public health initiatives should place a high priority on methods to lower obesity rates through measures including encouraging a nutritious diet and frequent exercise.

- The importance of encouraging physical exercise is highlighted by the negative relationship between obesity rates and physical inactivity. Promoting regular physical exercise among people can help reduce obesity, which in turn can reduce the risk of developing diabetes.
- Comprehensive Approach: Given the intricate connections between these health issues, it is necessary to take a comprehensive approach. Interventions in public health shouldn't concentrate solely on one element. Instead, initiatives should be combined to treat both obesity and inactivity at once.
- Interventions that are specifically targeted may be required since some areas or populations may be more affected by these health problems. To address the unique requirements of communities, tailored programs and policies might take regional variances and demography into account.

4. Appendix A: Method

Here is a detailed description of the methodology used in the analysis of data related to diabetes rates, obesity rates, and rates of physical inactivity across different regions in the United States. The following sections break down each aspect of the methodology:

1. Data Collection

- Data Source: The data used for this analysis was obtained from the Centers for Disease Control and Prevention (CDC) 2018 Diabetes dataset. This dataset is a reliable source of information related to health metrics in the United States.
- Dataset Structure: The dataset was organized into three separate sheets: "Diabetes," "Obesity," and "Inactivity." Each sheet contains data specific to its respective health metric.

2. Variable Creation

Understanding the Columns: Each row represents a combination of specific details for a particular region each year. Let's break down what each column represents:

- YEAR: This shows the specific year for the data entry.
- FIPS: FIPS stands for Federal Information Processing Standards. It's a code that uniquely identifies a county or region. So, each row corresponds to a specific region identified by its FIPS code.
- COUNTY: This column represents the name of the county or region.
- STATE: The abbreviation or full name of the state to which the region belongs.
- % DIABETIC: This column indicates the percentage of people in that region who are diabetic. It gives an idea of the prevalence of diabetes in that specific area.
- % OBESE: This column shows the percentage of individuals in that region who are considered obese. It gives insight into the obesity rates for that particular area.
- % INACTIVE: This column reflects the percentage of people in the region who are physically inactive. It provides information on the levels of physical inactivity within that specific area.

- Data Cleaning: The process of data cleaning is essential to ensure the quality and accuracy of the data used in the analysis. The data was loaded from Excel files using the Pandas library in Python, a popular data manipulation and analysis tool.

- Data Cleaning Procedures: Standard data cleaning procedures were applied, which typically include handling missing values (e.g., filling in missing data or removing incomplete records), renaming columns for clarity, and merging datasets if necessary. In this case, the datasets were merged based on a common identifier, the FIPS (Federal Information Processing Standards) code for regions. This merging likely allowed for the combination of data from different sheets into a single dataset for analysis.

3. Analytic Methods

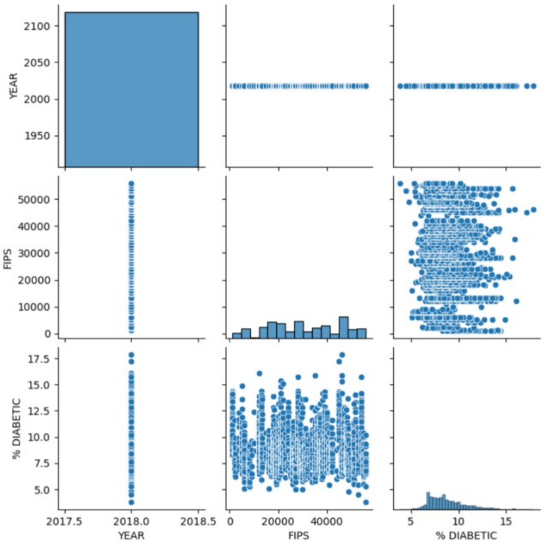
- Descriptive Statistics: Descriptive statistics were calculated to provide an overview of the dataset. These statistics include measures of central tendency (e.g., mean), measures of dispersion (e.g., standard deviation), and summary statistics like minimum and maximum values. Descriptive statistics help in understanding the basic characteristics of the data.
- Correlation Analysis: Correlation matrices were computed to assess the relationships between variables. Specifically, the analysis focused on the correlation between diabetes rates, obesity rates, and rates of physical inactivity. Correlation analysis helps identify potential associations between variables.
- Data Visualization: Data distributions were visualized using histograms with kernel density estimates (KDE). This visualization technique provides insights into the shape and characteristics of data distributions, helping to identify patterns and potential outliers.
- Linear Regression: Simple and multiple linear regression models were used for analysis. Linear regression is a statistical method used to analyze the relationships between variables and make predictions. It can help determine how one or more independent variables (such as obesity rates and physical inactivity rates) relate to a dependent variable (diabetes rates) and can be used to make predictions based on these relationships.

6. Appendix B: Results

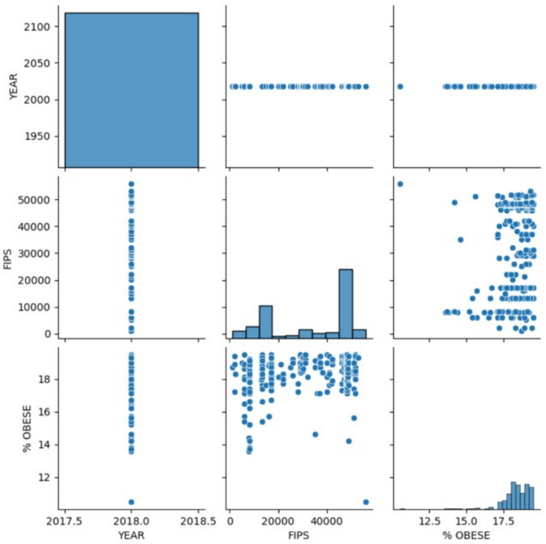
In this section, we are presenting the results of the data analysis and regression models for the three key health factors: diabetes (% DIABETIC), obesity (% OBESE), and physical inactivity (% INACTIVE). We will begin with descriptive statistics, followed by the results of the individual linear regression models, and finally, the results of the multiple linear regression model.

Key Descriptive Statistics:

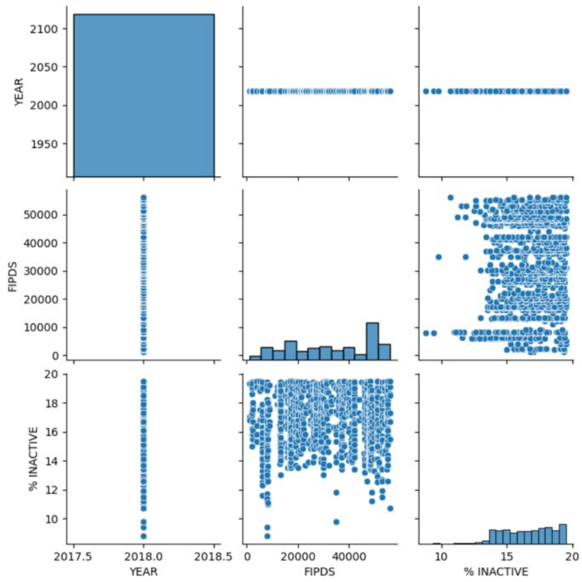
DIABETIC: The average diabetes prevalence in the studied areas is approximately 7.14%, with a moderate level of variability, ranging from 3.80% to 10.00%. This information provides us with a baseline understanding of diabetes rates in the dataset.



OBESE: On average, the obesity prevalence in these areas stands at about 18.25%, with some variation, ranging from 10.50% to 19.50%. This data helps us gauge the level of obesity in the regions under consideration.



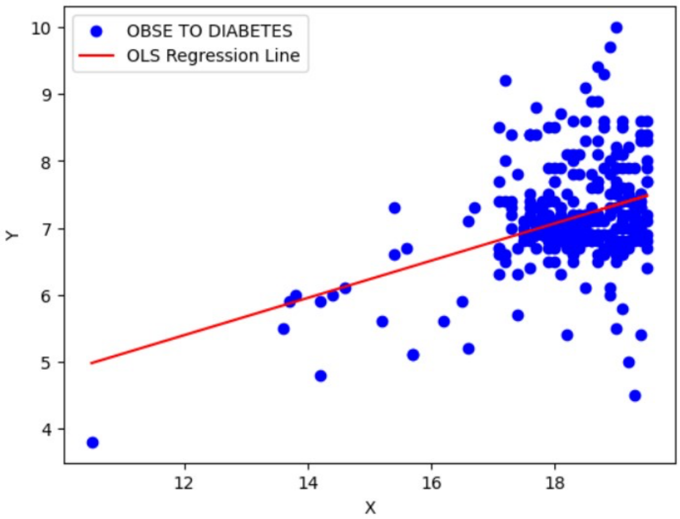
INACTIVE: The average level of physical inactivity is approximately 16.54%, with variability from 8.80% to 12.10%. This statistic informs us about the prevalence of physical inactivity in the studied areas.



Key Observations from Linear Regression:

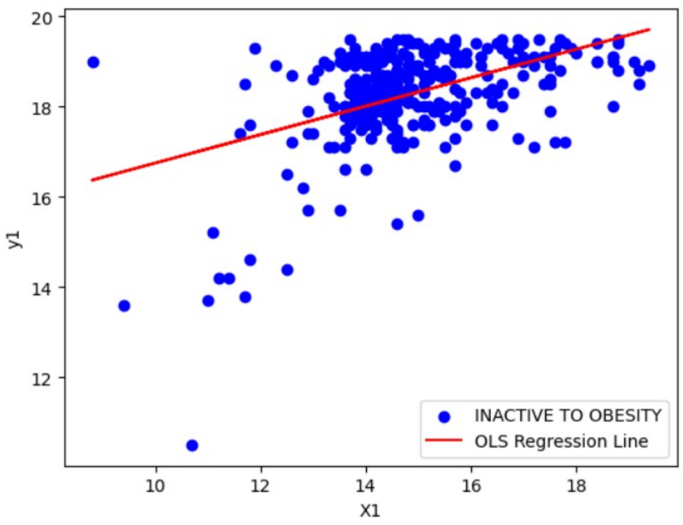
1. Obesity and Diabetes Relationship:

Our analysis revealed a notable positive correlation between obesity (% OBESE) and diabetes (% DIABETIC). In simpler terms, as the percentage of obese individuals in a region increases, so does the percentage of individuals with diabetes. This finding suggests a direct influence of obesity on diabetes prevalence.



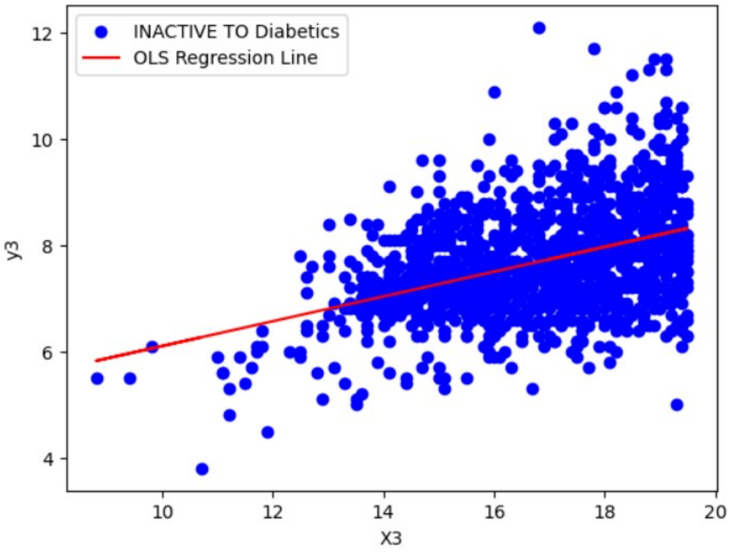
2. Physical Inactivity and Obesity Relationship:

We observed a similar positive relationship between physical inactivity (% INACTIVE) and obesity (% OBESE). When physical inactivity levels are higher in an area, it tends to coincide with increased obesity rates. This indicates that a sedentary lifestyle contributes to obesity.



3. Physical Inactivity and Diabetes Relationship:

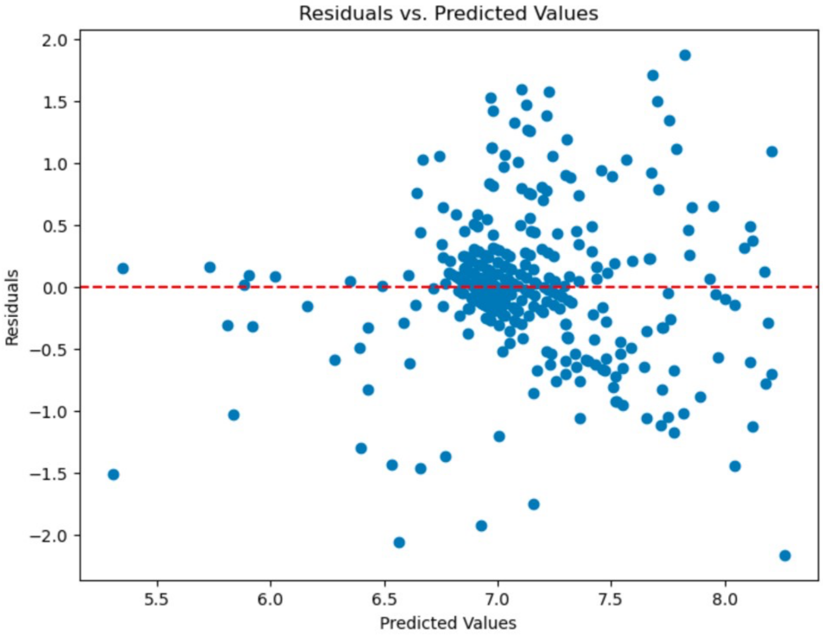
Our analysis also uncovered a positive connection between physical inactivity (% INACTIVE) and diabetes (% DIABETIC). In regions with higher physical inactivity rates, there is a corresponding increase in diabetes prevalence. This implies that reduced physical activity is linked to a higher risk of diabetes.

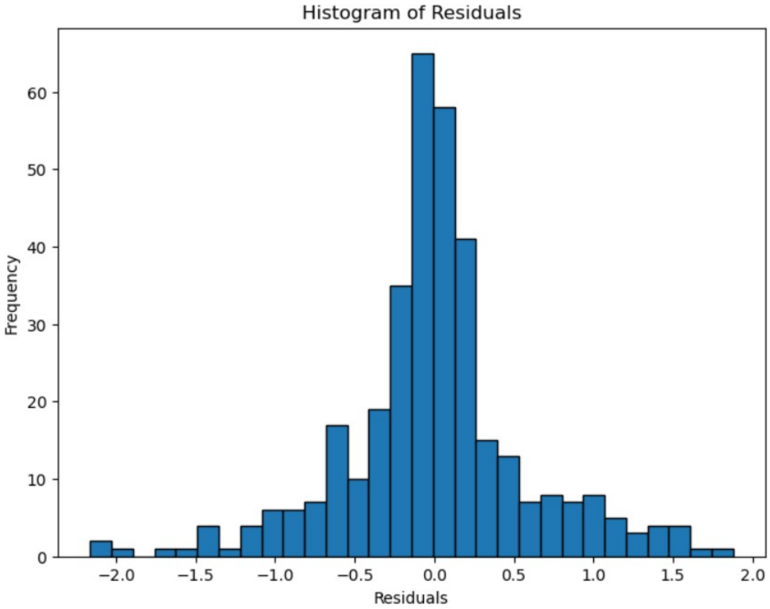


Key Insights from Multiple Linear Regression:

Predicting DIABETIC:

- In our multiple linear regression model, we considered both obesity (% OBESE) and physical inactivity (% INACTIVE) as predictors for diabetes prevalence (% DIABETIC).
- The model indicates that both obesity and physical inactivity have statistically significant impacts on diabetes rates.
- Specifically, for each unit increase in obesity (% OBESE), we expect a 0.2783 unit increase in diabetes prevalence (% DIABETIC).
- Similarly, for each unit increase in physical inactivity (% INACTIVE), we anticipate a 0.2331 unit increase in diabetes prevalence (% DIABETIC).





Conclusion:

The results of our analysis emphasize the critical importance of addressing both obesity and physical inactivity as key drivers of diabetes prevalence. These findings have significant implications for public health strategies:

- To combat the diabetes epidemic effectively, interventions should focus on promoting healthier lifestyles, including increased physical activity.
- Public health initiatives should aim to control obesity rates, as reducing obesity can lead to a decrease in diabetes prevalence.
- The insights gained from this analysis can guide targeted efforts to improve community health, reduce the burden of diabetes, and ultimately enhance the overall well-being of the population.

In summary, understanding the interplay between these health factors helps inform evidence-based strategies to prevent and manage diabetes, a major public health concern.

7. Appendix C: Data and code

Below is the link for code:

https://github.com/bhanuprasadthota/MTH-522-Project-1/blob/main/STATS.ipynb

8. References

1. Centers for Disease Control and Prevention (CDC). (2018). Diabetes dataset. Retrieved from [CDC Diabetes Dataset](https://www.cdc.gov/diabetes/data/index.html)
2. Angrist, J. D., & Pischke, J. S. (2008). Mostly harmless econometrics: An empiricist's companion. Princeton University Press.
3. Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Lawrence Erlbaum Associates.
4. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825-2830.
5. Python Software Foundation. (2021). Python Language Reference, version 3.9. Available at [Python.org](https://docs.python.org/3/)
6. OpenAI. (2021). ChatGPT. Retrieved from [OpenAI ChatGPT](https://openai.com/research/chatgpt)

9. Contributions

Bhanu Prasad Thota spearheaded the project's inception, emphasizing the importance of understanding links between obesity, inactivity, and diabetes in public health. They took charge of data cleaning, addressing consistency issues, and facilitating the seamless merging of datasets, ensuring data integrity for robust analysis.

Naga Venkata Lokeswarao Maturi led the data collection phase, showcasing expertise in sourcing data from the CDC 2018 Diabetes dataset. Their meticulous approach extended to data preparation, where they managed missing values and column renaming, optimizing the dataset for in-depth analysis.

Mantena Harsha Vardhan Varma excelled in data analysis, applying various statistical methods and delivering insights. Their role encompassed descriptive statistics, correlation analysis, data visualization, and linear regression, enhancing understanding of research findings.

Lakkamraju Hitesh Kashyap Varma provided a holistic perspective by exploring the practical implications of analysis results in public health and policy. They stressed the significance of obesity management and physical activity promotion. Additionally, Hitesh contributed to crafting the methodology and results sections, offering clarity on data processes and summarizing key findings.

Together, our team's collective efforts seamlessly blended our individual strengths and expertise to comprehensively explore complex health factor relationships. This collaboration yielded valuable insights for addressing critical public health challenges, showcasing the impact of our contributions.