Department: Statistics

Project Title: Investigating Causes for Diabetes within the United States

Project Objective:

The objective of this research is to answer the question of which factors cause diabetes to occur amongst Americans within the United States and also determine if there are certain factors that are deemed more statistically significant than others in causing diabetes to occur within individuals. To answer the question, a dataset that originated from the CDC and then uploaded onto Kaggle will be utilized to create a regression model to determine if the predictors are significant in causing diabetes.

Project Background and Significance:

According to the CDC, around 11.6% of the entire US population has diabetes and 22.8% Americans are undiagnosed with diabetes (Centers for Disease Control and Prevention). As a result, there are many Americans who are living their everyday lives unaware of the fact that they themselves might potentially have a disability that could fatally injure them. This lack of awareness as well many lower class Americans being not as properly educated about the causes and dangers of diabetes has led to a drastic increase in cases in the 21st century (Menke et al.). This research focuses on understanding the key factors that contribute to diabetes and figuring out which ones matter most. Using statistical analysis, specifically multiple linear regression, this study will examine how things like age, body weight, physical activity, diet, income, and access to healthcare are linked to diabetes. By identifying the strongest risk factors, we can better understand what's driving the rise in diabetes cases.

The research is based on public health and statistical modeling theories. One important framework is the Health Belief Model (HBM), which suggests that people are more likely to take action against health risks if they see them as serious and believe they can do something to prevent them. By using statistical methods to find the biggest risk factors, this study can help improve health awareness and encourage preventative actions.

This research is important because it can help guide healthcare policies and community health programs. If we know what factors contribute most to diabetes, public health campaigns can focus on the right areas and reach the people who need help the most. The results could also help doctors develop better tools to predict who is at risk, leading to earlier diagnoses and better health outcomes. By combining data analysis with real-world health concerns, this research aims to make a difference. The goal is to provide clear, useful information that can help individuals, doctors, and policymakers fight diabetes more effectively.

Research Methods:

I will begin my project by uploading the diabetes dataset into R. The first step is to clean the data, which means I will remove any errors or missing values to ensure that the analysis is accurate and that the final model will be reliable. Once the dataset is finally ready to be used, I will check for any assumption violations such as linearity and independence.. To test this, I will have to view each linear model's normal probability plots and their residual plots as well to determine if any violations are present. If the data does violate the majority of the assumptions, I will perform a Box-Cox transformation to potentially create an alternative model that will be more reliable in showing the relationship between diabetes and the predictors.

After attempting to fix most of the violations, I will also view the datasets residual through jackknife residuals, Cook's distance, and leverage to determine if there are any violations within the dataset as this could affect the final outcome of the model and make it inaccurate. If there are any violations, especially using Cook's distance, those data points will have to be removed to make the model more accurate as those specific points are heavily influencing the model.

Once the data assumptions and potential issues with values are addressed, I will apply backward selection, which removes predictors that do not add significant value to the model and are not as important for causing diabetes. Backward selection helps in narrowing down the list of predictors to only those that have a strong impact on causing an individual to have diabetes. Finally, after the final model is established, the eigenvalues of the factors will be calculated to see if collinearity is present, which is causing one factor to influence another of the factors instead of directly towards diabetes.

I plan to start this project in mid-October. The first few weeks will be dedicated to data cleaning and testing assumptions. By the middle of November, I will focus on checking residuals, collinearity, and applying backward selection to finalize the model. I aim to complete the final model and begin preparing my research report, PowerPoint presentation, and poster board by early December. This timeline ensures that every step is done carefully and the project is completed in a clear, organized manner. To ensure that everything is done on time, I will be checking in with Professor Simone regularly, to ensure that is accurate and presentable.

Expected Outcome:

After completing my research, I plan to produce several simple and clear presentations and reports to share my findings about what factors may lead to diabetes in Americans. Mainly, I will produce a research paper that explains how I built a regression model to study the links between different factors and diabetes and its applications. This paper will cover the project's goal, describe how I collected and analyzed the data, and explain what the important results were. Within the report, different figures and linear models will be shown to explain the connection between the various factors and how the final model was chosen. This way, readers can easily see which factors seem to have a stronger link to the disease.

I will also prepare a PowerPoint presentation for research seminars officially by UCF faculty and also through clubs and organizations such as the collegiate math society. The presentation will give a straightforward overview of the project, including its purpose, the methods I used, and the key findings. My aim is to make the presentation easy to understand so that everyone, no matter their background, can follow along. This should help spark conversation among students and faculty from different areas and majors so that they can relate to their field of work.

In addition, I will create a poster board with clear visuals like charts and graphs. This poster will highlight the main results of my study in a simple and attractive format. It will be available for display at the conference and online via UCF's research channels/archives so that many people can see it.

Together, these resources will share the new insights I gain from studying the factors that may cause diabetes. The findings are expected to offer useful information for both the field of statistics and public health. By sharing the results through a written report, a presentation, and a poster, I hope to reach a wide audience and make it easy for anyone to understand the impact of these factors. Ultimately, this project will help educate the UCF community and the public about diabetes, and it may guide future research and health initiatives.

Literature Review:

Centers for Disease Control and Prevention. "Diabetes in Young People Is on the Rise." *Centers for Disease Control and Prevention*, https://www.cdc.gov/diabetes/data-research/research/young-people-diabetes-on-rise.html.

Centers for Disease Control and Prevention. "National Diabetes Statistics Report website." https://www.cdc.gov/diabetes/php/data-research/index.html.

International Diabetes Federation. "Diabetes Facts & Figures." *International Diabetes Federation*, https://idf.org/about-diabetes/diabetes-facts-figures/.

Klonoff, David C. "The Increasing Incidence of Diabetes in the 21st Century." *National Library of Medicine*, 2009, https://pmc.ncbi.nlm.nih.gov/articles/PMC2769839/.

Menke, Andy, et al. "Factors Associated With Being Unaware of Having Diabetes." *National Library of Medicine*, 2017, https://pmc.ncbi.nlm.nih.gov/articles/PMC5399654/.

Tahmasbi, Siamak. *Diabetes Health Indicators*. Dataset. 7 March 2025. *Kaggle*, Diabetes Health Indicators.

Preliminary Work and Experience:

With my previous experience in creating multiple regression models for my statistics classes, I have gained knowledge and experience on how to correctly calculate and form a regression model that can accurately predict and interpret a given dataset. Additionally, I have done a report in Fall 2024 that handled over 5000 subjects and was able to create a multiple linear model that could predict a Nigerian student's outcome on an exam for their education system. I believe that these experiences will allow me to demonstrate my ability to complete this research project as I am already highly knowledgeable on how I would conduct my research so that it is done efficiently and on time so that it can be readily shared and accessed for those who are interested in reading it.

IRB/IACUC Statement: Since this research requires no contact with human nor animal subjects, it will not require IRB or IACUC approval.

Budget: $0

As this research is done completely online using free and accessible resources, there is no further need to purchase anything else in order to complete this research.