Diabetes Percentage Per U.S. State Based on Socioeconomic, Geographic, and Environmental Factors

SGK

Introduction

According to the CDC statistics, "38.4 million of Americans, which was about 11.6% of the population, had diabetes in 2021"(Centers for Disease Control and Prevention). In addition to its high prevalence, diabetes is the eighth leading cause of death in the United States, and its significant economic burden on the US. caught our team's attention. Indeed, "the total estimated cost of diagnosed diabetes in the U.S. in 2022 is $412.9 billion" (Parker, E. D.). To have a better understanding of what predicts the high prevalence of diabetes, we posed the following questions:

1. Do states with higher average income have a lower prevalence of diabetes within each state than those with lower average income?
2. Do states with a greater average amount of physicians have a lower prevalence of diabetes within each state than those with less?
3. Do environmental and locational factors, specifically such as higher walkability, have a lower prevalence of diabetes within each state compared to states with little walkability?

We obtained our data of response variable, Diabetes Prevalence in US adults since 2021, and explanatory variables, such as average income, population density, physician rate, etc., from the Social Explorer, CDC, and UVA Policy Map. To match for other explanatory variables, we manipulated the SVI data by adding each county's value together and then averaged. We also changed the data of gender, smoking, and education to a qualitative variable by classifying them by if they met a certain threshold.

Methods and Analysis

After compiling the data, we performed exploratory data analysis and found that the response level (diabetes level) was roughly unimodal and symmetric, making it suitable for regression. We then created scatterplots and determined the correlation coefficients for the quantitative variables as well as created boxplots and looked at their summaries for the qualitative variables. Satisfied with our findings, we moved on to begin with the initial hypothesis for our model. Our first model was built solely from the quantitative variables. We performed stepwise regression and ended up with four variables out of the original six. Next we added qualitative variables to our model. In order to test for their significance, we created dummy variables and conducted a nested F test to determine if each of the three variables were significant. We also double-checked our results by conducting an individual T-test of the qualitative variables. Smoking was still significant and other qualitative variables were insignificant. Based on this we added only one variable to our model from step one. Lastly, we checked for interactions in our model. In the graphs produced, no lines crossed, indicating there were no significant interactions in our model, leaving us with a final model of five predictors.

Next we checked for multicollinearity in our model. Since no individual VIF was greater than 10 and the average VIF was less than 3, we concluded that there was no concern for severe multicollinearity. From there we checked our model assumptions. Our model met the assumptions of homoscedasticity due to the absence of a fanning pattern. It also met the normality assumption since the qqPlot lined up and the histogram was roughly symmetric and unimodal. It also was not time series data so it was independent. However, the mean zero assumption was violated due to the quadratic pattern displayed for the population density. To correct for this we added a higher order term for population density so our final model now also included population density squared.

Based on this model we then checked for influential observations and outliers. To begin, we found the Cook's distance for each of our observations. Points 11, 30, and 31 were considered influential due to their value being above the 0.08 threshold. Next we plotted the studentized residuals and hat values and found observations 11 and 31 to be outliers in the y direction and 30 and 39 to be outliers in the x direction. Due to 11 and 31 being both influential points and significant outliers in the y direction, we decided to remove them from our model.

The last measure of adequacy for our model was external model validation through jackknifing. We found the adjusted R^2 and MSE from jackknifing was 0.6609 and 1.1424, which are slightly higher than those from our original model. While it is typical for the adjusted R^2 to be smaller and the MSE to be larger, these slight variations are likely due to either the original model overfitting the data or containing influential observations that are skewing the data. However, the values are very close to one another, suggesting stability and reliability of the model.

Results

After checking the interactions, variable screening, and model assumptions, we have our final model as the following: DiabetesLevel = 17.34 - (8.368e-5) Average_Income + 0.006783 Population_Density -(3.530e-6) Population_Density ^2 - 0.3809 Walkability + 2.636 SVI + 0.7856 Dummy_Smokers, where dummysmoking = 1 if yes, 0 if no. Our final model has a higher adjusted R^2, with a value of 0.7604, than the initial one, indicating that it accounts for more of the variance. Additionally, the lower MSE and p-value, with a value of 0.899 and 4.919e-19 respectively, shows that the final model is more accurate at predicting the relationship between diabetes prevalence and the explanatory variables.

Conclusions

Our research has determined the most significant factors of predicting the prevalence of diabetes across states in the US: Average Income, Population Density, Walkability, SVI, and Smoking, which are three quantitative and one qualitative variables. Based on the correlation coefficient, increased population density, SVI, and smoking will increase the prevalence of diabetes while increasing average income and walkability decreases diabetes prevalence, which makes sense.

To predict diabetes prevalence in Virginia, the model would look like: DiabetesLevel = 17.34 - (8.368e-5)* 80615 + 0.006783* (217.4) -(3.530e-6)* $(217.4)^2$ - 0.3809*8.976217+ 2.636*0.3757 + 0.7856*0, which is approximately 9.5. This projected diabetes prevalence is just under its actual prevalence rate of 10.7, with a slight residual of 1.2. Therefore, our model is useful at predicting diabetes prevalence rate in the US based on the factors like average income, population density, walkability, SVI, and smoking or not. This answers our research question that states with higher average income and higher walkability have a lower prevalence diabetes rate than those with lower average income and walkability. However, regarding the diabetes prevalence and physician rate, there is no significant relationship between these two variables.

One limitation of the model is that diabetes prevalence can be influenced by various factors apart from the variables we have analyzed, including genetics which can be difficult to measure. Thus, the diabetes we currently focus on are only Type II, because Type I is more genetically predisposed. To improve upon the model, we may use more qualitative data, such as age, ethnicity, etc. to ensure that the collecting data captures a wide range of demographic variables, so that the model may provide a more comprehensive understanding of diabetes prevalence across different population groups.

Appendix A: Data Dictionary

Variable Name	Abbreviated Name	Description
Diabetes Percentage	Diabetes_Level	Measures the percent of the population in the state that have Diabetes. Units are measured in percentages.
Average Income	Average_Income	Measures the average income in the state. Units are in dollars.
Social Vulnerability Index	SVI	Measures the negative effects on communities caused by external stressors (disease, war, natural disasters) on human health. Calculated on a scale from 0-1, with 0 being low vulnerability and 1 being high vulnerability.
Population Density	Population_Density	Measures the average number of people per square mile in each state. Units are in mi^2
Average National Walkability Index	Walkability	Ranks the census block groups on whether primary mode of transportation is walking based on various environmental factors. Measured on a scale of 1-20.
Fast Food Restaurants	Fast_Food	Measures the average of the average number of fast food restaurants in each county of a U.S. state. No specific units, just the number of fast-food restaurants.
Average Physician Rate	Physician_Rate	Measures the number of primary care physicians per 100,000 people in each county. Rates have been averaged and grouped by state. Units are measure in percentage.
Gender Proportion	Gender	Measures the “majority” gender proportion within the state population that have diabetes. There are two levels: female and male.

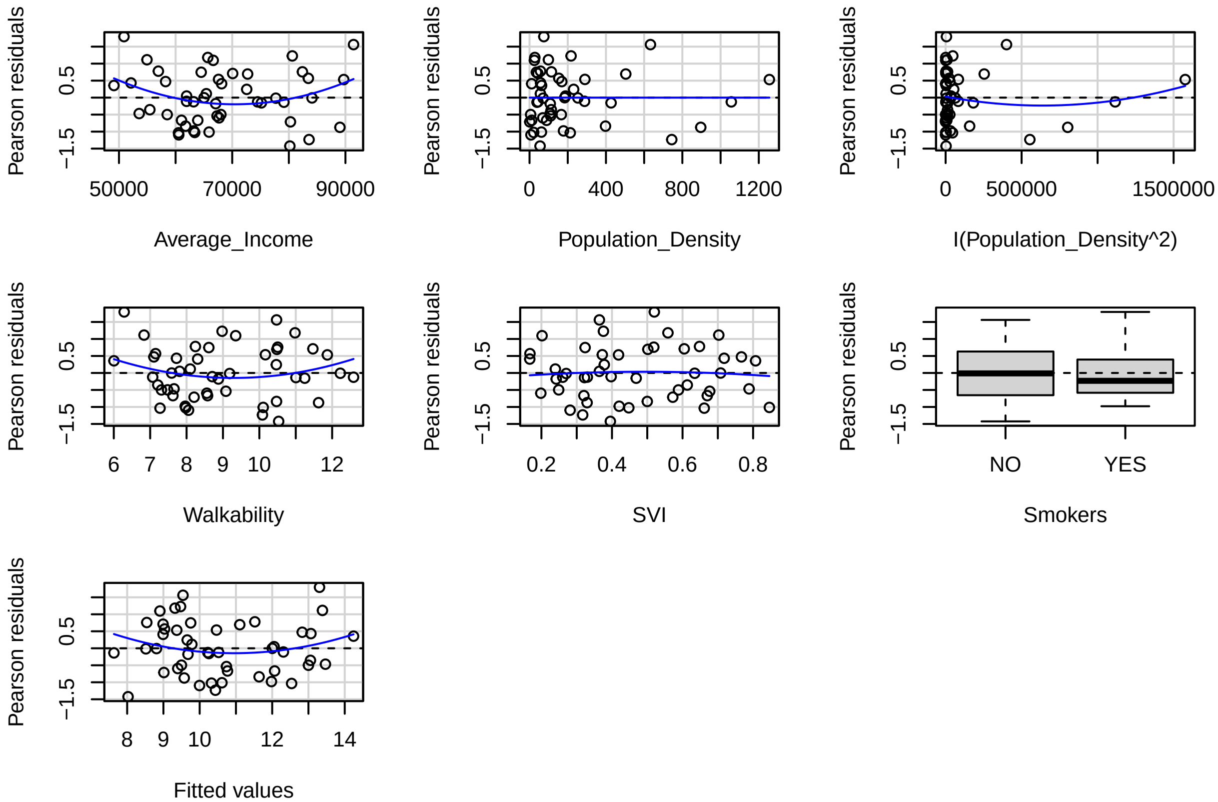
Variable Name	Abbreviated Name	Description
Smoker	Smokers	Measures if 20% of the state population over 18 years-old are smokers. There are two levels: yes and no.
Education Level	Education_Level	Measures if more than 10% of the state population have not earn high-school degree or equivalent. There are two levels: yes and no.

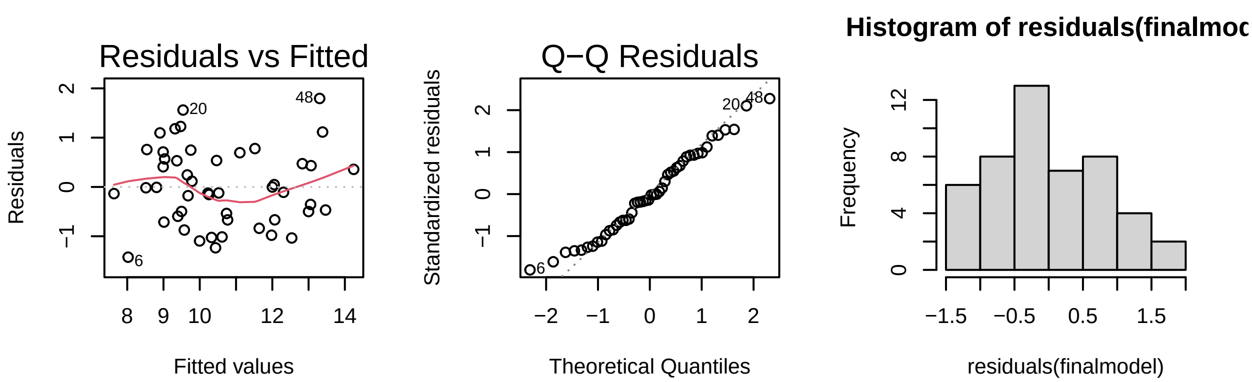
Appendix B: Data Rows

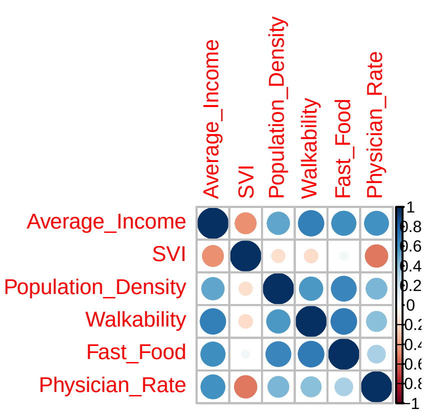
	State	Diabetes_Level	Average_Income	SVI	Population_Density
1	Alabama	14.5	54943	0.7027	98.7
2	Alaska	8.3	80287	0.5723	1.3
3	Arizona	9.6	65913	0.8461	62.3
4	Arkansas	13.5	52123	0.7180	57.8
5	California	8.8	84097	0.6344	253.1
6	Colorado	6.6	80184	0.3954	55.2

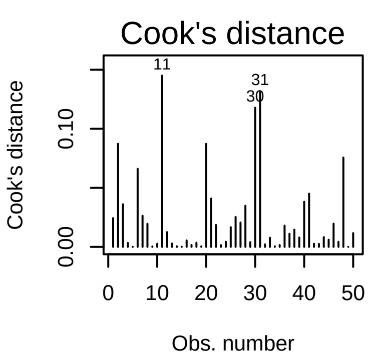
	Walkability	Fast_Food	Physician_Rate	Education_Level	Gender	Smokers
1	6.831251	57	44.8027	YES	F	YES
2	8.206548	15	94.5632	NO	M	NO
3	10.105780	306	51.8455	YES	M	NO
4	7.722636	27	46.9361	YES	M	YES
5	12.225010	517	68.9324	YES	M	NO
6	10.530890	64	67.3694	NO	M	NO

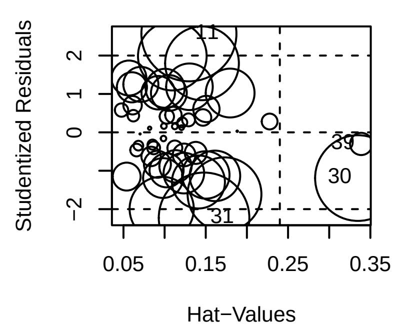
Appendix C: Tables and Figures



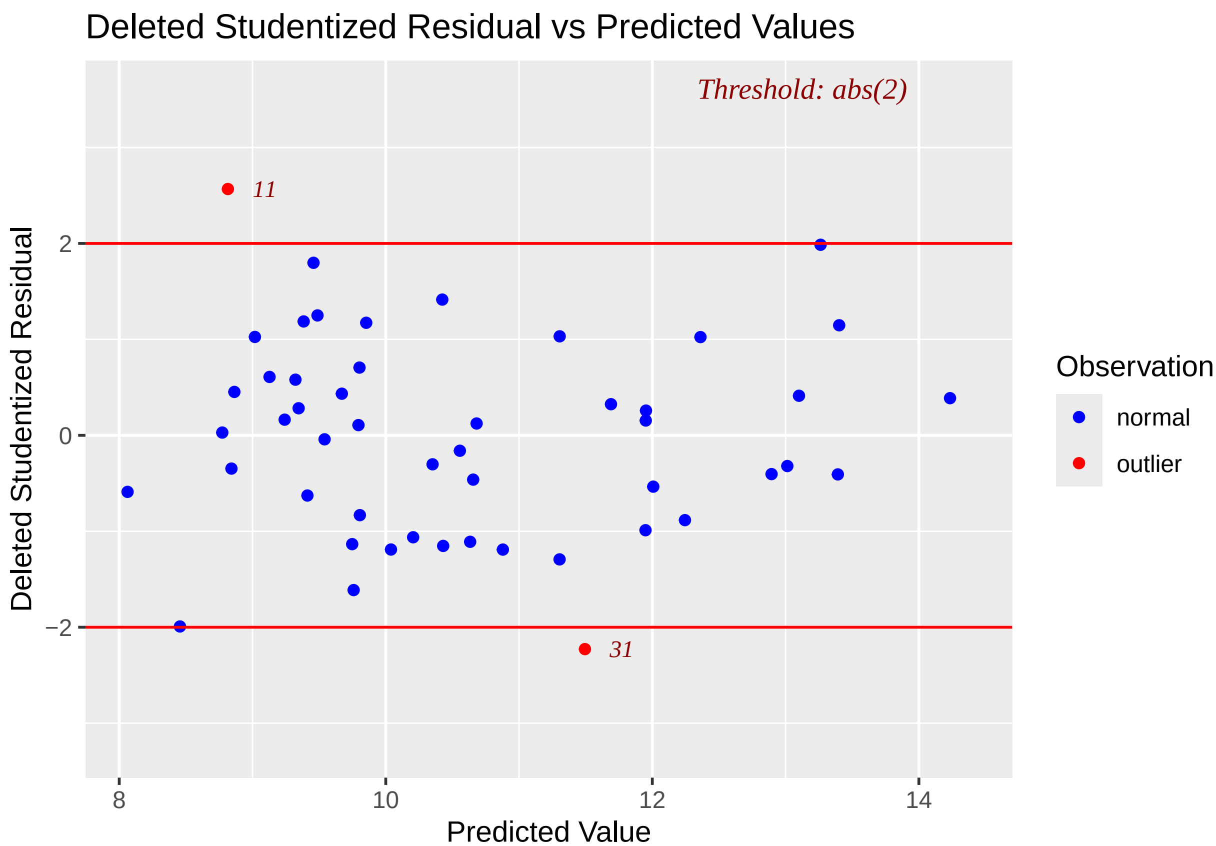








```
      StudRes       Hat       CookD
11  2.5674070 0.1295170 0.145027140
30 -1.1907946 0.3349374 0.117900959
31 -2.2279370 0.1479356 0.131763417
39 -0.3019042 0.3388558 0.007950057
```



Appendix D: References

Background

1. Centers for Disease Control and Prevention. (2023, November 29). National Diabetes Statistics Report. Centers for Disease Control and Prevention. https://www.cdc.gov/diabetes/data/statistics-report/index.html
2. Parker, E. D., Lin, J., Mahoney, T., Ume, N., Yang, G., Gabbay, R. A., ElSayed, N. A., & Bannuru, R. R. (2023, November 1). Economic costs of diabetes in the U.S. in 2022. American Diabetes Association. https://diabetesjournals.org/care/article/47/1/26/153797/Economic-Costs-of-Diabetes-in-the-U-S-in-2022
3. Bureau, U. C. (2023, November 2). National Diabetes Month: November 2023. Census.gov. https://www.census.gov/newsroom/stories/diabetes-month.html#:~:text=%E2%80%9CDiabetes%20is%20a%20disease%20that,to%20some%20types%20of%20cancer
4. Hill, J., Nielsen, M., & Fox, M. H. (2013). Understanding the social factors that contribute to diabetes: A means to informing health care and social policies for the chronically ill. The Permanente journal. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3662286/#:~:text=The%20sociobiologic%20cycle%20of%20diabetes.&text=Social%20determinants%20of%20health%20encompass,and%20access%20to%20nutritious%20food.&text=Lifestyle%20factors%20incorporate%20dietary%20choices,to%20primary%20health%20care%20services
5. Silva-Tinoco, R., Cuatecontzi-Xochitiotzi, T., De la Torre-Saldaña, V., Leon-Garcia, E., Serna-Alvarado, J., Orea-Tejeda, A., Castillo-Martínez, L., Gay, J. G., Cantu-de-Leon, D., & Prada, D. (2020, August 26). Influence of social determinants, diabetes knowledge, health behaviors, and glycemic control in type 2 diabetes: An analysis from real-world evidence - BMC endocrine disorders. BioMed Central.

Data

1. Population Density (Per Sq. Mile) [Map]. In SocialExplorer.com. ACS 2016 (5-Year Estimates) Retrieved 24 March 2024, from https://www.socialexplorer.com/8e62e93dee/view
2. Average Household Income (In 2021 Inflation Adjusted Dollars) [Map]. In SocialExplorer.com. ACS 2021 (5-Year Estimates) Retrieved 24 March 2024, from https://www.socialexplorer.com/a9676d974c/view

3. Total Population: [Map]. In SocialExplorer.com. ACS 2021 (5-Year Estimates) Retrieved 24 March 2024, from https://www.socialexplorer.com/a9676d974c/view
4. Less than High School [Map]. In SocialExplorer.com. ACS 2021 (5-Year Estimates) Retrieved 24 March 2024, from https://www.socialexplorer.com/a9676d974c/view
5. Percent of Current Smokers (Persons 18 Years and Over) [Map]. In SocialExplorer.com. Health Data 2023 Release Retrieved 24 March 2024, from https://www.socialexplorer.com/a9676d974c/view
6. https://bmcendocrdisord.biomedcentral.com/articles/10.1186/s12902-020-00604-6
7. Centers for Disease Control and Prevention. (2021). Social Determinants of health - United States Diabetes Surveillance System. Centers for Disease Control and Prevention. https://gis.cdc.gov/grasp/diabetes/diabetesatlas-sdoh.html#
8. PolicyMap. (n.d.). National walkability index in 2021 [Map based on data from EPA Smart Location Database: Data downloaded from https://www.epa.gov/smartgrowth/smart-location-mapping#SLD, January 2022]. Retrieved March 24, 2024, from http://www.policymap.com