Biostat 625 Final Report

Salina Moon, Xiaochun Shao, Cheukying (Rosita) Szeto

2022-12-16

Github Link: https://github.com/yay10053/Biostat625FinalProj

Social Factor Data set

Data Processing The code used is in a separate R script named: Date_Processing

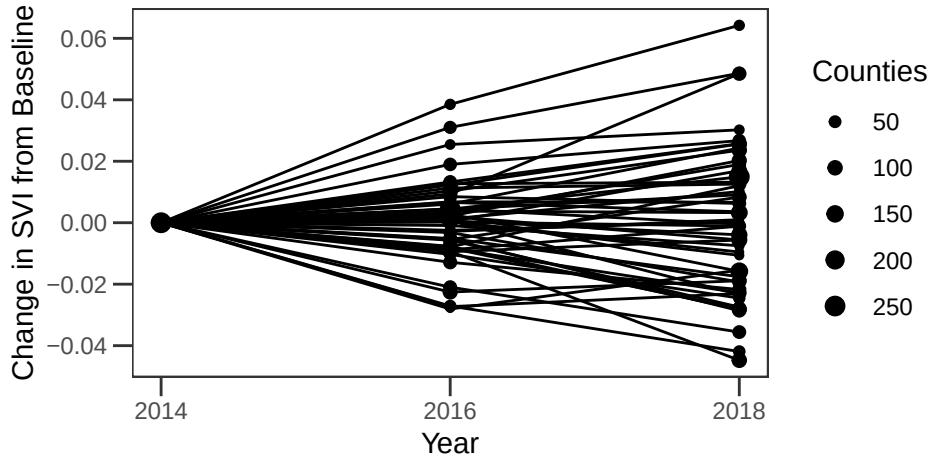
Since the data is stored in separate files, we need to consolidate the total of 9 files concerning 3 different variables and 3 different year before moving forward. We first bind the data while loading it to r and remove the "No Data" to be NA value so it would to easier to work with later. Then we merged the 3 different into one big data frame.

Exploratory Analysis A first look at the dataset reveals that it contains data for 3,141 counties across 51 states, including Washington DC. Diabetes prevalence, SVI, Food insecurity, and No health insurance data were collected at three time points: 2014, 2016, and 2018.

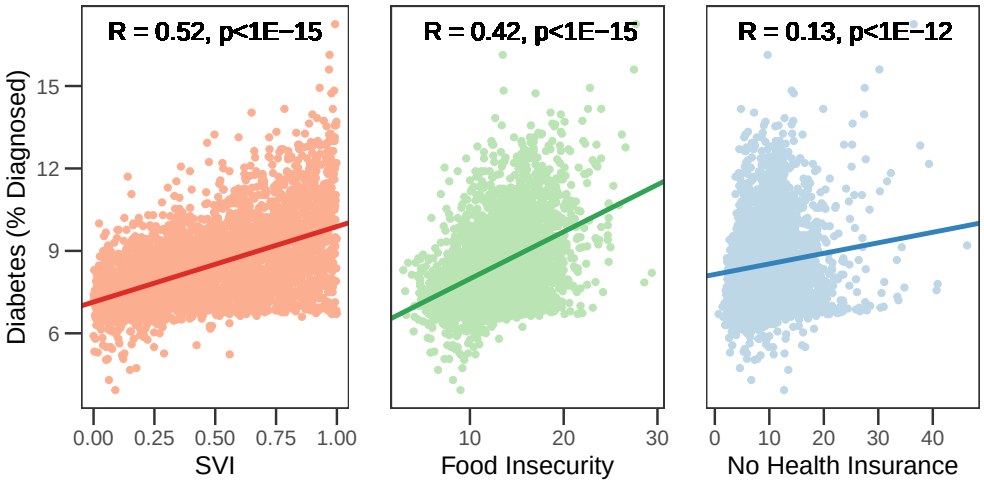
The number of counties within each state remains mostly consistent over time and may be considered constant. Because only 8 observations were incomplete out of a total 9,428, seemingly randomly, we excluded them from all further analyses.

Change over time This dataset contains both spatial and longitudinal information, and attempting to decipher a pattern directly was difficult. The dplyr package's grouping and piping functionalities were incredibly helpful at both summarizing the information and organizing the many steps we took in data manipulation.

First, we discovered that SD of Food insecurity and No health insurance at all time points were 0 in every county. In other words, Food insecurity and No health insurance were constant over time, and Diabetes and SVI were the only time-variant variables. Diabetes was relatively consistent from 2014 to 2018 (8% CV, on average). Furthermore, it resembled a normal distribution, so fitting a linear model would be appropriate. SVI also did not change much (13% CV, on average). After visualizing the mean change in SVI over time by state, we observed that SVI does not have a uniform trend across all states.



After examining each variable in isolation, we visualized their relationships with each other. We obtained lines of best fit from simple linear regression models of Diabetes with each of the 3 social covariates, as well as Pearson correlation coefficients. Combining these pieces of information shows consistently that our covariates were all positively correlated with Diabetes. Therefore, we felt confident adding SVI, Food insecurity, and No health insurance to our model for in depth analysis.



Ultimately, to study the effects of both social factors and time on the prevalence of diabetes in the US, we need to build a formal model.

Modeling In our first stage model, to detect the effect of social determinants on the diagnosed diabetes percentage, we construct a linear mixed model including Year, Overall.SVI, Food.Insecurity, No.Health.Insurance and all the interaction term with Year. For random effect, we included random slope, term of Overall.SVI, as we found high subject deviance adjusted by Year in our exploratory analysis, and the subject unit here is State.Based on this full model, we did stepwise backwards selection with threshold 0.1 and 0.05 for random and fixed effect terms.

After we compared AIC of the models by different grouping variables, County and State, our first final model is resulted to be by State and is shown as follow.

$$E[D_i|b_i] = (\beta_0 + b_0) + \beta_1 Year_i + (\beta_2 + b_1)SVI_i + \beta_3 FI_i + \beta_4 NHI_i + \beta_5 Year_i * SVI$$

From R output below, the variance between States is 1.74, and variance of error is $1.44 < 1.74$, it implies high variability between States for the percentage of diabetes. All of the point estimates for the remaining terms are significant at level of 0.05, and only No Health Insurance index negatively associated with Diabetes rate whose effect magnitude is not significant(<0.05). The fixed effect of Year is 0.11, and each increase in year is significantly associated with 0.11 increase in percentage of diabetes on average adjusting for other variables. Similarly, we found every unit change in SVI being significantly associated with 2.51 increase in percentage of diabetes,adjusting for all other variables. Similar interpretations for Food Insecurity and interaction term between Year and SVI, while every unit increase of FI and interaction terms correspond to 1.90 and 1.74 increase in diabetes rate when they are adjusting for all other terms.

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	6.8641529	0.1409989	104.9153	48.682309	0.0000000
Year	0.1099318	0.0302874	9314.6785	3.629617	0.0002854
Overall.SVI	2.5080883	0.2353972	82.0829	10.654708	0.0000000
Food.Insecurity	0.0190289	0.0063544	9287.6434	2.994610	0.0027552
No.Health.Insurance	-0.0186919	0.0042336	9119.6667	-4.415168	0.0000102
Year:Overall.SVI	0.1738681	0.0524658	9313.3924	3.313935	0.0009234

Interactive Dashboard The script that create the dashboard is in a separate R script named: rshiny

An introductory video can also be found here: https://drive.google.com/file/d/1oSjxAIbkWHpA3Hy0JCxaY3whz6GJqeFN/view?usp=sharing

The purpose of this dashboard is to educate the public using the data set we obtained from the CDC website. In the first tab, the data table was presented and audience has the option to filter the table by state and year, so they would be able to learn relevant information such as diabetes prevalence, % of food insecurity, SVI, and % people with no health insurance concerning their community. On the second tab, the same information is presented as a map, so audience can look at the information by variable comparatively across the United State. On the last tab is a self-assessment center where audience can input some personal measure such as height, weight, and A1c %. Once the information is submitted, it would output a table with calculated BMI and A1c and give suggestion according to the calculated value.

BioFactor Dataset

Introduction In order to further investigate how Obesity relates to some physical factors such as Inactivities and Obesity. We look into a new dataset from the CDC site that provide as data from 2004 to 2019.

The first step would be data processing, since the data is stored in separate files, we need to consolidate the files before moving forward. This is challenging because the data is stored separately by variables as well as by year, meaning we will need to consolidate 45 files. Unlike the previous data set, merging them one by one is not feasible in this case.

Data processing The code used is in the separate R script named: Date_Merge

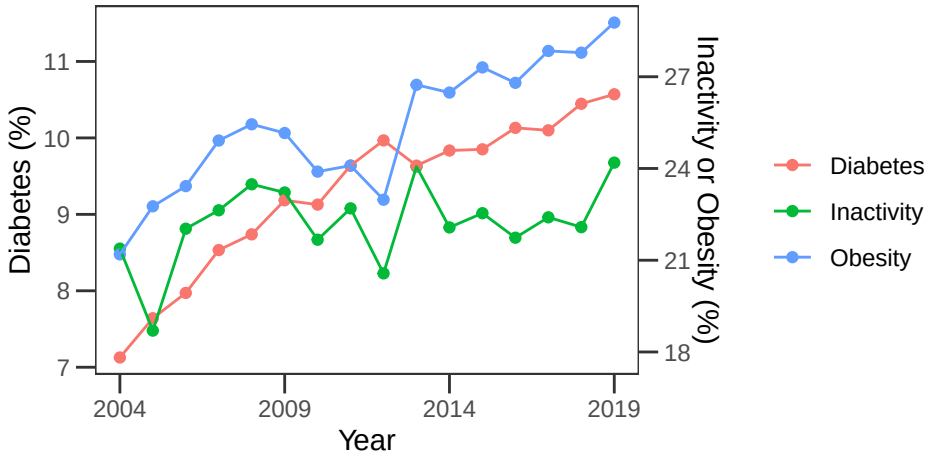
In order to consolidate the large amount of files, we first save all files in one single folder. A new function is created to get data from the same variable but different years: get_data(file_pattern, variable_name)

We will need to specifiy the file pattern so we can load the file corresponding to the desired variable, then load the dataset as list, impute the year varible, then reduce the list into a single dataframe. After we obtain the dataframe from all 3 variables, we can further reduce it by merging. The resulting dataframe contains variable “Year”, “CountyFIPS”,“County”,“State”, “Obesity”, “inactive”, and “diabetes”. The dataframe contains 51,616 observations and 7 variables.

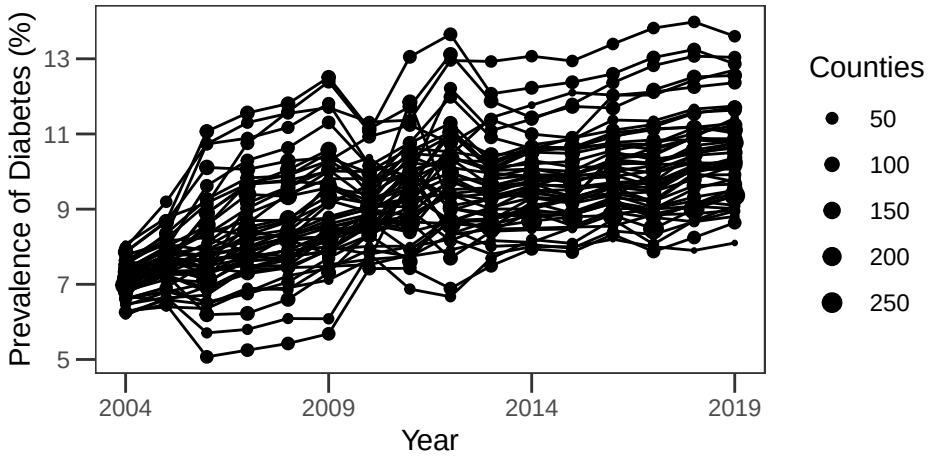
Exploratory Analysis With an even larger dataset than previously, grouping the data into categories was crucial to identify important patterns. A major example of this was that, after grouping our data into the 52 "states," we discovered that there was no data for Puerto Rico. After removing incomplete cases, we were left with 3,148 counties in 51 states at 16 time points, every year from 2004 to 2019.

Obesity, Physical Inactivity, and Diabetes each approximate normal distributions. So, Diabetes can be appropriately modeled with linear regression, and Obesity and Inactivity do not need to be transformed.

Unlike the social factors, our biofactors all changed over time. Also unlike the previous part, the span of time is much longer than four years, so it is not reasonable to assume that these variables would not change much and could be described in terms of mean and standard deviation. In order to examine the longitudinal change, we grouped and summarized our data by year. We visualized the mean values over all counties at each time point and found that they all increased over time. In addition, Obesity and Inactivity follow similar patterns - overall increasing with a drop around 2010 to 2012. This connection is to be expected because obesity and physical activity are closely linked medically. On the other hand, the national average prevalence of diabetes was steadily climbing from 2004 to 2019.



Though diabetes had a consistent, positive trend nationally, this clear line removes much of the detail and nuance available in the larger dataset. So, we further visualized our outcome on the state level. Most states did appear to have similar slopes over time to each other and the national average, though with varying intercepts, which justified our use of the linear mixed model.



GG Animated Plot Link to the animated plot: https://rpubs.com/cszeto/984619

The challenge with this biofactor dataset is organizing large amount of information into meaningful way. Since we have information across 15 span, it would be interest to visualize the data by year. We constructed an animated plot giving obesity rate vs inactivity rate, and the dot size represented by the Diabetes rate. The data is visualized on the state level and given the option to move the slide to change the year of the information illustrated. From the plot, we observed that there is a positive relationship between inactivity rate and obesity rate. Although the relationship with the Obesity rate is unclear as the rate is about the same in each state. The precise information of a state can be obtained by moving the cursor over the dot.

While putting the variables in relation of others is interesting, it might be more meaningful to visualize the nation-wise average over the year. We first summarize the average rate for the 3 variables in each Year. We noticed that while all 3 variables has increased over the year Diabetes rate steadily increases and Obesity rate increases at a higher rate than inactivity rate. This information can help us understand how diabetes relates to some physical factors.

Modeling We recollect data from 2004 to 2019 to see the relationship between diabetes percentage and risk factor including physical inactive and obesity. As there will occur singularity problem adding random slope here, to avoid overfitting and convergence problem, we only included random intercept this time. Compare the model by State to by County and AIC shows model by County performs better. We still used backwards selection and no variable should be removed, and here is our second final model:

$$E[D_i|b_i] = (\beta_0 + b_0) + \beta_1 Year_i + \beta_2 Inactive_+ \beta_3 obesity_i + \beta_4 Year_i * Inactive + \beta_5 Year_i * Obesity_i$$

Here is R output of our second stage model. Random variance is 0.9257(<1), which implies relatively low between subject variability. All the point estimates from variables of interest or interaction term are significant. Specifically, one year increase is significantly associated with 0.129 increase in diabetes percentage across the whole population adjusting for the all other variables. And one percent increase in physical inactive and obesity are respectively associated with 0.062 and 0.12 increase in diagnosed diabetes percentage on average adjusting for all other variables. For the interaction terms, one unit increase of Year is ,on average ,significantly associated with 0.008 increase in effect of inactive term on the outcome, adjusting for other variables. Similary, the Year increase is significantly associated with 0.005 decline in effect of obesity rate on average on the outcome of interest.

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	3.5097473	0.0821225	46243.88	42.73793	0
Year	0.1291733	0.0072554	48580.28	17.80383	0
inactive	0.0617637	0.0033245	49000.93	18.57816	0
obesity	0.1180803	0.0036093	49256.62	32.71534	0
Year:inactive	0.0082712	0.0003254	48610.42	25.41747	0
Year:obesity	-0.0058636	0.0003245	48627.05	-18.06764	0

In conclusion, for every point estimate, the magnitudes of the fixed effect over the diagnosed diabetes are almost not observable while they are statistically significant(at significant level of 5%). ??(I am not sure about the cutoff for the shrinkage degree, and if there is not standard answer, we can delete this one)The shrinkage degree of this model looks good as we found the variability for the random effect is smaller than variance of error.

The model diagnosis for both models are checked with QQplot and residual plot. LMM model assumption is ensured since no violation is observed suggesting our models are legit.