8.3

Moshe Burnstein

`r Sys.Date()`

`{r setup, include=FALSE} knitr::opts_chunk$set(echo = TRUE)`

Diabetes Data

Type 2 Diabetes is a horrific disease which destroys countless lives throughout the world on a daily basis. We have yet to get a handle on this disease, even in the most advanced countries. Negative outcomes include neuropathy, kidney failure, blindness, amputations, and cardiovascular disease. Causes of the disease are assumed to include obesity, genetics, sedentary lifestyle, and diets emphasizing processed foods. The path to controlling the disease includes forming a strategy to address the main causes. Data Science can leverage these ideas with analysis of the data to determine with a high degree of probability what accounts for the greatest factor in the disease onset. We can then tailor both prophylactic and treatment regimens. We can accurately predict who will get this disease. Diabetes care will transform from a bludgeon to a sharp, directed scalpel.
My problem statement is that hypertension causes diabetes. Just as high blood pressure causes damage to the kidneys, so too does it damage the insulin-producing pancreas. If proven correct, we can aggressively treat hypertension to prevent Type 2 diabetes and to manage the disease in people afflicted. We can muster our resources and our patients' efforts in the singular battle against hypertension and not focus on strict glucose control, for instance. I will explore the relationship between diabetes and hypertension. Do diabetics generally fall under the category of hypertensive? Do patients who suffer longer from the disease, or with more advanced and less-controlled diabetes, present with higher blood pressure readings than the newly diagnosed and better controlled? Do genetics play a role in the onset of disease? Have peoples who have struggled with food supply developed genes which store energy more efficiently, and which cannot tolerate abundance? What role does a change in diet caused by stealing their water supply play? After years of starvation and with only knowing a home-grown healthy diet, the Pima were forced to take handouts from the federal government. These handouts included solid shortening, lard, and many forms of white flour. From this time, fried bread has become ubiquitous with the Pima. Does hypertension fall under the umbrella of cardiovascular disease which is caused and exacerbated by diabetes? Do different populations react differently to hypertension? What role does obesity play? Does obesity cause diabetes? Does diabetes cause obesity? Does a third factor cause both? Does obesity cause hypertension?
I will seek disparate data sources to study samples which represent the diabetes population at large. The samples will include both obese and non-obese subjects. There will be both closed (not intermarried) and mixed populations. The demographics will vary, both socio-economically and geographically. I will explore the relative relationships among such variables as pedigree function, BMI, Hba1c, and blood pressure readings. This will determine the relationship between hypertension and diabetes, both in correlation and in coefficient of determination. Exploration of other variables will show how much an effect they have on both diabetes and hypertension, further allowing us to see the real effect of hypertension on diabetes, and ruling out the tertium quid. If we find a strong positive correlation, then we can attempt to prove causal effect. Statistics will tell us the probabilities of confounding variables. [NCSU diabetes dataset]https://www4.stat.ncsu.edu/~boos/var.select/diabetes.tab.txt This dataset contains 442 observations of diabetics and 11 variables. The observations include age, sex, BMI, average blood pressures, and 6

blood serum measurements. These bloods track ldl, hdl, and total triglycerides, and total cholesterol.

[Pima Indians 2 is accessible in R through the 'mlbench' library.] https://archive.ics.uci.edu/ml/machine-learning-databases/pima-indians-diabetes/ This dataset originated in the National Institute of Diabetes and Digestive and Kidney Disease. The Pima Indians were presenting with an inordinate number of cases of Type 2 diabetes, even amongst children where Type 1 diabetes is normally so prevalent. Whether because the U.S.A. stole their water and gave them processed foods and/or because of a genetic predisposition, the Pima Indians' rate of Type 2 diabetes has averaged 50%, as opposed to 5% to 13% in the population at large. This dataset has become standard for ml prediction modeling. It contains eight predictor number variables and one outcome factor variable with two levels, 1 and 0 represent diabetic and not diabetic. All 768 observations are on females older than 20 years and of Pima Indian descent. Pima Indians have generally married their own, so their lineage is pure. The PimaIndians2 dataset replaced all zero values in the original dataset with 'NA'. There are 374 NAs in the glucose column, 227 under triceps, and much less in pressure, glucose, and mass. The pressure variable indicates the diastolic blood pressure(mm hg).

The heplots package Diabetes dataset comes from Reaven and Miller's study of 145 adults who were not obese. They used the PRIM9 system to visualize the data in 3D. They found heterogeneity in several of the variables. The data in these variables told a different story from the prevailing understanding of the progression of diabetes from the chemical stage to the overt stage. Their work is seminal in the field of statistics in the study of homogeneity of variance. It is also a groundbreaking work in in the field of diabetes care. It helped describe the stages and progression of Type 2 diabetes. The dataset has five num,int variables and one factor variable of three levels- normal, chemical, and overt.

[Syrian Refugees in Jordan]https://data.humdata.org/dataset/cct-chv-mpc-syrian-refugees-jordan? The Jordan dataset studied Syrian refugees in Jordan with Type 2 diabetes. It is a longitudinal cohort study assessing the effectiveness of the multi-layered interventions of cash investment, health education, and conditional cash transfers. There are 1042 observations and 69 variables. The variables run the gamut from age and sex, to blood pressure and BMI, to household size an economic status. This data was last updated on August 24, 2020. Several columns contain NA values. Although there are 482 NA values in 5 of the columns, there are 1042 rows.

Tinytex is required to properly knit the RMarkdown file. Plots will be presented with ggplot2. Dplyr will parse the data. Pastecs will enable statistics description. Hmisc is necessary for some correlation tests. Mlbench and heplots house two of the datasets. Scatterplots will show the relationship between 2 continuous variables and histograms will be used for categorical variables. Boxplots will help visualize outliers. Ggpairs may prove useful to compare the different variables. It produces a scatterplot matrix and Pearson correlation value and significance. Histograms and boxplots will point to outliers.

There are many different variables in the amalgamation of these datasets. Machine learning may be leveraged to create a specific description of the ideal diabetes candidate. It can predict who will develop diabetes with great probability. One must experiment to find the learning model with the greatest efficacy. It may be appropriate to calculate the biserial correlation coefficient because diabetes is a continuous dichotomy. We would need the table(), prop.table(), and polyserial() functions.