Capstone 1 Statistical Data Analysis

Predicting Prediabetes with Machine Learning

NHANES Data 2007-2016

Goal of Statistical Analysis

I will be testing the null hypotheses that the means of several of the laboratory test results and physical exam results are the same regardless of diabetes diagnosis category:

- Meets diabetes diagnosis criteria,
- Meets prediabetes diagnosis criteria, or
- Does not meet either diagnosis criteria.

I am looking for features that help predict which people meet the diagnosis criteria for prediabetes and diabetes. Therefore, I will look at several features that have been addressed in medical literature as having a relationship with diabetes. I expect most of these features to be different for each diagnosis category and for my null hypotheses to be rejected.

Process

Several of the following features are known risk factors, so I would expect them to have different means for diabetic, prediabetic and non-diabetic samples. Some of these are known complications of diabetes, but may not necessarily be risk factors. I am not sure whether these features will have different means as it could depend on how long each person has had elevated blood glucose.

The features I chose to look at are:

1. Gender: Several studies show men older than 40 are at higher risk for developing diabetes.
2. Hypertension Diagnosis: Hypertension is a risk factor for diabetes.
3. Moderate Activity: Increased activity is known to reverse prediabetes and lessen the severity of diabetes.
4. Marital Status: Studies have shown that being widowed is associated with a higher risk of type 2 diabetes in men.
5. Triglycerides: High triglycerides are a known risk factor for diabetes.
6. Systolic Blood Pressure: Systolic Blood Pressure is the top number of a blood pressure measurement and tends to increase with age. Increased blood pressure, or hypertension, is also a known risk factor for diabetes.
7. Body Mass Index: Being overweight, or having a Body Mass Index of 25 or greater, is a significant risk factor for diabetes.
8. Serum Iron: Elevated iron level is a risk factor for diabetes.
9. Aspartate Aminotransferase (AST): AST is a measure of liver function. Liver problems are a known side effect of diabetes. Mild chronic elevations of AST (and ALT, another liver enzyme) are often due to insulin resistance.

I divided these features into two groups:

- Categorical variables, and
- Continuous variables.

I tested the difference of the means of the categorical variables with the chi-square test. The data meets many of the criteria for goodness of fit.

1. The data are categorical.
2. The categories are mutually exclusive. By the diagnostic criteria, a participant is limited to one group: no diabetes, prediabetes, or diabetes.
3. Our data is frequency data, with counts all well above 5.

Unfortunately, the data are not from a simple, random sample. They are from a complex, clustered sample. This will increase the risk for Type 1 error. Knowing this, I will go ahead with the chi-square test as the results will still be interesting.

I tested the continuous variables with the student's t test for independent samples. Some of the requirement for a t test are met in this data.

1. The data are continuous.
2. The groups are independent.

I checked the variances of the two groups and corrected each test for unequal variances.

Once again, the data are not from a simple random sample, so there could be increased Type 1 error.

The data are not normally distributed, but the sample sizes are large, so the results should still be meaningful.

Results

The results of the chi-square tests were that all null hypotheses of equal means among the groups were discarded and alternate hypotheses of unequal means were accepted. This was expected with the chosen features, and the visualizations supported this conclusion. What was not expected was the extreme values of the chi-square statistic and p-value. This could be due to the non-random sample.

The results of the t tests were also that all the null hypotheses were rejected and alternate hypothesis of different means were accepted. This was expected with most variables and is again mostly supported by the visualizations. It would be interesting to re-run the tests for serum iron and AST after accounting for the non-random sample and see if the results remain the same.