Predicting Diabetes Diagnoses

Sarah Netchert

MISERICORDIA UNIVERSITY

INTRODUCTION

"Diabetes is a disease that occurs when your blood glucose, also called blood sugar, is too high."(NIH) In the last two decades, the percentage of the United States population who were diagnosed with diabetes has skyrocketed over 5%(America's Health Rankings). There have been many articles that explain ways to prevent diabetes like eating healthy, working out, getting medication, etc. I want to determine which trait in adults put them at a higher risk of getting diagnosed with diabetes. By determining which traits relate the most the diagnoses of diabetes, we can create a model that can predict which groups of people should actively seek out these diabetes prevention methods. The traits that will be used for predicting a person's chance of being diagnosed with diabetes are Cholesterol, Glucose, High Density Lipoprotein, Age, Gender, Height, Weight, Systolic Blood Pressure, Diastolic Blood Pressure, Waist Size, and Hip Size.

BACKGROUND

One dataset was used for these analyses. It is comprised of information collected during the screening process for interviewees participating in a study by Dr. John Schorling from the University of Virginia. There were over one thousand African Americans in central Virginia who were screened but for this analysis, we are only focusing on 403 interviewees who had their Hemoglobin A1C levels tested. After getting this data, it was cleaned to remove personal information. Patient id's were removed as well as city the patient lived in. Other changes to the data was changing the gender column to 1s and 0s. The column where hemoglobin was recorded as also altered. It was changed to 1s and 0s to depict if the interviewee had diabetes or not. If their hemoglobin level was equal to or greater than 6.5, then the interviewee had diabetes and was changed to a 1. If their level was below 6.5, then they did not have diabetes and was changed to a 0. Changing this column to a binary group allows to predict our probabilities.

Proposed Analysis

Since we want to determine which traits relate the most to the diagnoses of diabetes, multiple linear regression will be used. Specifically, the backwards method will be used to test all traits at once and eliminating the traits that fail multicollinearity. It will give a generalized model that is better fitted for our second analysis. After finding the model that will be used for our predictions, we will continue the backwards method until three traits remain and trait with the lowest variance inflation will be the trait that relates the most. Once we have our new model of traits that relate the most to the diagnoses of diabetes, we will then separate the traits to make up smaller models. They will be used to make multiple prediction tables consisting of two traits and specified for each gender.

Analysis Results

Full Model

Total Cholesterol	Stabilized Glucose	High Density Lipoprotein	Age	Gender	Height	Weight	Systolic Blood Pressure	Diastolic Blood Pressure	Waist	Hip
1.165	1.185	1.212	1.690	2.411	2.302	7.688	2.163	1.812	5.386	6.575

Model with weight removed

Total Cholesterol	Stabilized Glucose	High Density Lipoprotein	Age	Gender	Height	Systolic Blood Pressure	Diastolic Blood Pressure	Waist	Hip
1.161	1.178	1.206	1.602	2.280	2.043	2.144	1.78	4.19	4.282

Model with hip removed

Total Cholesterol	Stabilized Glucose	High Density Lipoprotein	Age	Gender	Height	Systolic Blood Pressure	Diastolic Blood Pressure	Waist
1.16	1.178	1.205	1.553	2.055	2.043	2.144	1.795	1.230

Model with systolic blood pressure removed

Total Cholesterol	Stabilized Glucose	High Density Lipoprotein	Age	Gender	Height	Diastolic Blood Pressure	Waist
1.16	1.176	1.204	1.214	2.054	2.041	1.076	1.229

Model with gender removed

Total Cholesterol	Stabilized Glucose	High Density Lipoprotein	Age	Height	Diastolic Blood Pressure	Waist
1.159	1.176	1.184	1.163	1.035	1.071	1.19

Model with waist removed

Total Cholesterol	Stabilized Glucose	High Density Lipoprotein	Age	Height	Diastolic Blood Pressure
1.145	1.160	1.086	1.154	1.035	1.039

Model with stabilized glucose removed

Total Cholesterol	High Density Lipoprotein	Age	Height	Diastolic Blood Pressure
1.126	1.044	1.068	1.023	1.039

Model with total cholesterol removed

High Density Lipoprotein	Age	Height	Diastolic Blood Pressure
1.016	1.014	1.022	1.01

Model with height removed

High Density Lipoprotein	Age	Diastolic Blood Pressure
1.005	1.004	1.008

Female

Total Cholesterol	High Density Lipoprotein				
	12	38	46	59	120
78	0.248	0.119	0.095	0.065	0.011
179	0.777	0.416	0.338	0.238	0.041
204	0.986	0.552	0.453	0.323	0.058
230	1.229	0.727	0.604	0.438	0.081
443	2.772	2.546	2.446	2.250	0.983

Male

Total Cholesterol	High Density Lipoprotein				
	12	38	46	59	120
78	0.169	0.081	0.065	0.044	0.007
179	0.527	0.283	0.230	0.162	0.028
204	0.667	0.375	0.308	0.220	0.039
230	0.830	0.493	0.410	0.298	0.055
443	1.851	1.703	1.638	1.508	0.665

Female

Stabilized Glucose	Age				
	19	34	45	60	92
48	0.012	0.021	0.032	0.058	0.195
81	0.041	0.073	0.111	0.196	0.595
89	0.055	0.099	0.150	0.260	0.755
106	0.105	0.184	0.275	0.463	1.178
385	2.998	2.999	2.999	3.000	3.000

Male

Stabilized Glucose	Age				
	19	34	45	60	92
48	0.005	0.010	0.015	0.027	0.093
81	0.019	0.034	0.053	0.093	0.295
89	0.026	0.047	0.071	0.125	0.382
106	0.049	0.088	0.132	0.227	0.624
385	1.998	1.999	1.999	2.000	2.00

Female

Height	Waist				
	26	33	37	41	56
52	0.190	0.368	0.524	0.727	1.806
63	0.133	0.263	0.381	0.541	1.529
66	0.121	0.239	0.348	0.497	1.453
69	0.110	0.218	0.318	0.456	1.376
76	0.087	0.174	0.256	0.371	1.201

Male

Height	Waist				
	26	33	37	41	56
52	0.186	0.349	0.484	0.652	1.391
63	0.131	0.253	0.360	0.499	1.222
66	0.119	0.232	0.331	0.461	1.173
69	0.108	0.211	0.304	0.426	1.123
Height	0.086	0.170	0.247	0.352	1.004

Female

Systolic Blood Pressure	Diastolic Blood Pressure				
	48	75	82	90	124
90	0.239	0.127	0.107	0.088	0.038
121	0.569	0.319	0.272	0.227	0.101
136	0.823	0.484	0.417	0.350	0.160
147	1.049	0.644	0.560	0.474	0.223
250	2.807	2.642	2.583	2.506	2.052

Male

Systolic Blood Pressure	Diastolic Blood Pressure				
	48	75	82	90	124
90	0.191	0.102	0.086	0.071	0.031
121	0.442	0.252	0.216	0.180	0.081
136	0.629	0.378	0.328	0.276	0.128
147	0.790	0.498	0.436	0.371	0.177
250	1.893	1.799	1.765	1.721	1.449

Conclusion

After using the backwards method, it was determined that the best fitted model includes the traits cholesterol, glucose, high density lipoprotein, age, gender, height, systolic blood pressure, diastolic blood pressure, and waist. It is the first model that passes multicollinearity with values under 4. Continuing with the backwards method, we see that when three variables are left, the age trait has the lowest variance (Var(age) = 1.004) making it the strongest factor in predicting diabetes. We should expect the tables including these two traits should have higher probability percentages.

Looking at the prediction tables, we immediately see a difference in the probability percentages between males and females. The female tables all have much higher probability of being diagnosed in diabetes. Looking at the first pair of tables on cholesterol and high density lipoprotein, the group that would have the highest probability of being diagnosed with diabetes are those with high cholesterol and low levels of high density lipoproteins. The second pair of tables on glucose levels and age result in the group with the highest probability of being diagnosed are the older people with high glucose levels. The third pair of tables on height and waist size resulted in higher probabilities for those who are shorter with larger waists. The final pair of tables on systolic and diastolic blood pressure resulted in higher probabilities for those with higher systolic blood pressure and higher diastolic blood pressure. Our expectation from checking multicollinearity using the backwards method was correct. The tables including age and high density lipoprotein had the highest probabilities with age reaching as high as 300% for women. When taking into account the results of the eight prediction tables, we can conclude the group of people who have the greatest probability of being diagnosed with diabetes are older women with high cholesterol, low levels of high density lipoprotein, high levels of glucose, high systolic and diastolic blood pressure, have bigger waists, and who are short. People with these characteristics are at the highest rick of being diagnosed with diabetes and should make large lifestyle changes to reduce their risk.

REFERENCES

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.Rproject.org/.

America's Health Rankings. Annual Report: Diabetes. 2019. plot. April 2020. <https://www.americashealthrankings.org/explore/annual/measure/Diabetes/state/ALL?edition-year=2019 >.

Centers for Disease Control and Prevention. LDL and HDL Cholesterol: "Bad" and "Good" Cholesterol. 31 January 2020. article. April 2020. <https://www.cdc.gov/cholesterol/ldl_hdl.htm>.

National Institute of Diabetes and Digestive and Kidney Diseases. What is Diabetes? December 2016. document. April 2020. <https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes>.

Schorling, Dr John. Datasets. 1997. dataset. April 2020. <http://biostat.mc.vanderbilt.edu/wiki/Main/DataSets>.

Thank you to Dr. Michael Floren for everything he has done since coming to Misericordia.