What Makes Diabetes

Statistics 218 Final Project

Hyeyeon Hannah Kim

Abstract

Diabetes is the disease in which the body's ability to produce or respond to the hormone insulin is impaired, resulting in abnormal metabolism of carbohydrates and elevated levels of glucose in the blood and urine. Based on this definition, the data collected glycosolated hemoglobin to clarify whether the diagnosis of diabetes is positive or negative. To find out whether the diagnosis of diabetes is positive or not, we need to see the value of glycosolated hemoglobin. If the value is more than 7.0, it is the positive of diabetes. Otherwise, smaller than 7.0, it is the negative of diabetes.

Background

In our lives, many symptoms are illustrated to let the people know whether they have diabetes or not. For example, you always feel hungry even if you ate before because you cannot use the glucose normally, always feel tired, feel increased thirst, frequent urination because of the high rate of glucose in kidney, nausea, blurry vision, sudden weight loss, sexual problems, slow healing of wounds, and etc. We cannot perfectly sure that you have diabetes if you have any those symptoms, but those are the symptoms that we can clarify ourselves. As we can see above, the reasons of those symptoms are all based on rate of glucose.

This data consists of 19 variables on 403 subjects from 1046 subjects who were interviewed in a study to understand the prevalence of obesity, diabetes, and other cardiovascular risk factors in central Virginia for African Americans. The researchers predicted that the waist/hip ratio may be a predictor in diabetes and heart disease. However, based on the coding, those were not the main factors for the diagnosis of diabetes. The 403 subjects were the one who were actually screened for diabetes who had more than 7.0 for glycosolated hemoglobin. From lots of variables that are given, I especially focused on the gender and age variables with the diagnosis of diabetes. The hypothesis that I made is that the gender and age variables will strongly effect the outcome of diagnosis of diabetes. I made this hypothesis because lots of people including me think the diabetes are coming to seniors with just biased data.

Methods

The 'diabetes' data that I found at the Vanderbilt University Department of Biostatistics' website, there were 1046 people who were interviewed for the data. However, I eliminated some test subjects which have no values for one of the variables. So, it ended up to 132 test subjects in total. Following lists are the variables that I used on the data.

id: subject ID (test subjects' number)
chol: Total Cholesterol
stab.glu: Stabilized Glucose
hdl: high Density Lipoprotein
ratio: Cholesterol/HDL Ratio
glyhb: Glycosolated Hemoglobin
bp.1s: First Systolic Blood Pressure
bp.1d: First Diastolic Blood Pressure
bp.2s: Second Systolic Blood Pressure
bp.2d: Second Diastolic Blood Pressure

location: Buckingham, Louisa
gender: Female, Male
frame: Small, Medium, Large

The test subjects were collected with lots of variables which have possibilities for diabetes occur. Especially the data shows if the value of

glycosolated hemoglobin is higher than 7.0, the test subject has the positive diagnosis of the diabetes.

Among the various variables, I picked age and gender first. Before all of the tests, I made glycosolated hemoglobin (glyhb) variables as a number 1 and 0 to use the variable as the number. If the glyhb is larger than 7.0, it was clarified to 1, otherwise it was 0 and the variable was called 'glyhbCAT.' I compared gender and the glyhbCAT variables to see whether diabetes have more positive diagnosis in female or male. Then I cut age variables to 4 different ranges and called that variable as 'agenew.' It was from 0 years to 100 years, so I divided them to (0,25), (25,50), (50,75), and (75,100). Similar with comparing the gender with glyhbCAT, I compared agenew and glyhbCAT to see the relationship between the age and diagnosis of diabetes. Finally, to analyze and compare three variables, agenew, gender, and glyhb, I used three-way table and calculate odd ratios of those variables to compare age and gender together with glycosolated hemoglobin. However, to make a three-way table I need to cut the age variables only in two variables so I cut it 0 years to 50 years, and 50 years to 100 years.

To find out which variables are effective to diabetes, I used stepwise algorithms. I used backward elimination, forward elimination, and stepwise selection. I found out each Akaike information criterion (AIC) and chose variables which are truly effecting the outcome of diabetes.

Results

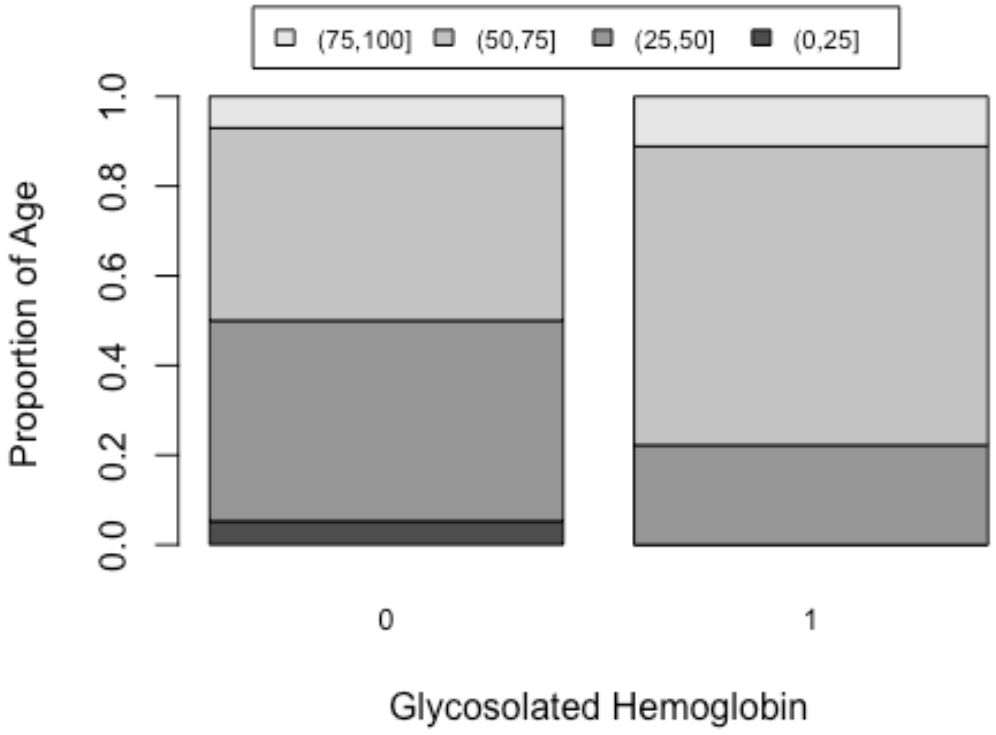
Out of 132 subjects, there were 18 people who had positive diagnosis of diabetes. We can know this by the value of glyocosolated hemoglobin. To know which gender is more occurring in diabetes, I compared glycosolated hemoglobin value and gender by proportion. I made a bar plot with the x-axis as glycosolated hemoglobin and y-axis as the proportion of the gender which makes the total 100% which is

1.00 in proportion.

	Female	*Male*	*Total*
0	0.535 (61)	0.465 (53)	1.000 (114)
1	0.667 (12)	0.333 (6)	1.000 (18)
Total	1.202 (73)	0.798 (59)	2.000 (132)

To compare age and diabetes, I used 'agenew' variables which were cut by 25 years each from the age variable. First, I made a table

with agenew variable and glycosolated hemoglobin. Similarly, 0 simply stands for negative diagnosis for diabetes and 1 stands for positive diagnosis for diabetes. Then, I made a bar plot from the proportion values of agenew variables' values.



	(0,25)	*(25,50)*	*(50,75)*	*(75,100)*	*Total*
0	0.053	0.447	0.430	0.070	1.000
1	0	0.222	0.667	0.111	1.000
Total	0.526	0.669	1.097	0.181	2.000

To compare age and gender with glyocosolated hemoglobin, I made three-way table. I cut the age variables to (0,50), (50,100) to make the same length with gender and glyhbCAT.

		(0,50)	*(50,100)*	*Total*
Female, 0		26	35	61
		2	10	12
1				
Male, 0		31	22	53
	1	2	4	6
Total 0		57	57	114
	1	4	14	18
Total		61	72	132

The conditional odds ratios of agenew2 and gender at the negative diagnosis (glyhbCAT = 0) is 0.5272 and the conditional odds ratios of agenew2 and gender at the positive diagnosis (glyhbCAT = 1) is 0.4.

To be more accurate with the relationship between the variables and glycosolated hemoglobin which is the outcome for the diagnosis of diabetes, I did stepwise algorithms to find out which variables make the data good fit. From the outcome of stepwise algorithms, lower AIC is the better outcome to use from this method. First, I did backward elimination; the AIC was equal to 107.2 and the variables were all eliminated so the logistic regression was as follow:

Logit(p)=log(p/(1-p)) =-1.846.

So, I did forward elimination; the AIC was equal to 104.9 and the remaining variable was stab.glu which stands for stabilized glucose. The logistic regression was as follow:

logit(p) = log(p/(1-p)) = -2.785+0.0075x_*stab.glu*.

Finally, I did stepwise selection; the AIC was equal to 104.9 and the remaining variable was also stab.glu. The logistic regression was as follow:

logit(p) = log(p/(1-p)) = -2.785+0.0075x_*stab.glu*

Conclusions

As the bar plot of glycosolated hemoglobin with gender showed that 66.7% of people who were positive to diabetes were female and 33.3% of subjects were male. Based on the bar plot and the table from above, we can simply conclude that female can get diabetes more than male. From the table of proportion of age and glycosolated hemoglobin tell us that the age from 0 years to 25 years have 0% of positive diagnosis of diabetes, 25 years to 50 years have 22.2%, 50 years to 75 years have 66.7%, and 75 years to 100 years have 11.1%. We can carefully conclude that 50 years to 75 years have the highest proportions of positive diagnosis to diabetes. To compare the

glycosolated hemoglobin with age and gender altogether, I made a three-way table. From the table, I calculated the odds ratios of those variables. The conditional odds ratios of age and gender for the positive diagnosis is 0.5272. This shows us that female who got negative diagnosis are less to be in 0 years to 50 years than male. Moreover, the conditional odds ratios of age and gender for the negative diagnosis is 0.4. This shows us that male who got positive diagnosis are likely to be in 0 years to 50 years more than female who got positive diagnosis.

To interpret the data set more accurately, I used stepwise algorithms for all of the variables which were used as factors for the diagnosis of diabetes. From the stepwise algorithms, there are three ways to eliminate the variables and make the data best fit for the best outcome. By three eliminations, the lower AIC is the best elimination for the data. As the results above, forward elimination and stepwise selection had the same lowest AIC which is equal to 104.9. Also, the remaining variable for the outcome was same, which was stabilized glucose. By the results of stepwise algorithms' code, the logistic regression was

$$\text{logit}(p) = \log(p/(1-p)) = -2.785+0.075x\_stab.glu$$

From this equation, we can interpret that the odds of glycosolated hemoglobin when the stabilized glucose is 0 is e^ (-2.785). As stabilized glucose increases 1 each, then the odds of glycosolated hemoglobin increases multiplicatively by
e^ (0.0075).

From the stepwise algorithms, we can clearly conclude that age and gender are not strongly related to diabetes. The variable which is the stabilized glucose is the factor that effect the diabetes the most. What I already mentioned at the background paragraph, the fact that most of the symptoms were coming from the amount of glucose, can be one of the clear evidence for this project.

Limitation

While I was working on this project, I made a hypothesis that the diagnosis of diabetes will depend on the age, gender, weight, height, etc which are all of the variables listed on this data that I used. So, I thought when I do stepwise algorithms, those factors' coefficients will come out with lowest AIC. However, the outcome from the stepwise algorithms, stabilized glucose was the only factor that effect the diagnosis of diabetes. So, I tried to freeze the variables that I want to compare with glycosolated hemoglobin which were age and gender. I simply just compare those with glyhbCAT each in number of subjects and make them as proportion to make bar plots. By bar plots, I could compare the variables easily and simply. I tried to compare age, gender, and diabetes so I made a three-way table. However, the length should be all same with two so I cut the age variables in a different way as the logistic regression.

Appendix

For R code: the RMarkdown file called “Final Project - Coding“

Work Cited

DataSets < Main < Vanderbilt Biostatistics Wiki, http://biostat.mc.vanderbilt.edu/wiki/Main/DataSets.

YouTube, YouTube, 22 Oct. 2019, https://www.youtube.com/watch?v=Kz9L0dRYbEc&feature=youtu.be.

“What Are the Symptoms for the Diabetes?” Seoul Bae Hospital: Naver Blog, 20 Oct. 2019, https://doctorbae75.blog.me/221683351055.