Analysis between Diabetes and BMI using Logistic Model

Jingyao Wang

December 2020

Abstract

Studies have shown that diabetes is one of the most serious diseases affecting people's lives. We used the Pima Indians Diabetes Database to observe various factors that affect diabetes. Among them, we found that obesity is closely related to diabetes. A higher BMI indicates a higher chance of getting diabetes.

Introduction

Diabetes is currently one of the most common metabolic diseases, and its main feature is hyperglycemia. If blood sugar levels cannot be controlled, long-term high blood sugar can cause serious damage to organs and nerves. For example, blindness and kidney failure are common complications of diabetes. According to the research of WHO, in 2014, over 8.5% people who are older than 18 had diabetes. Moreover, 1.6 million people died due to high blood sugar in 2016. WHO estimates that one of the major causes of death in 2016 is diabetes. Between 2000 and 2016, the number of premature deaths due to diabetes showed an upward trend. Due to the number of people suffering from diabetes has continued to rise in recent years, they have also paid more and more attention to it. For this reason, many preventive measures have been summarized, and the hospital will also check related factors carefully during the physical examination. Since diabetes is not an infectious disease, people can control or prevent it by adjusting their living habits. Fitness, healthy eating and avoiding staying up late are highly recommended for maintaining body in a healthy status.

Type 2 Diabetes is highly related to obesity. Research demonstrates that 85% of people with type 2 diabetes are overweight. Body mass index (BMI) is an index that measure body fat. By calculating BMI, we can intuitively know whether a person is obese or not. We define obesity as a BMI over 25.0.

Data

We used a set of data that is released by the National Institute of Diabetes and Digestive and Kidney Diseases. It contains eight diagnostic measurements (variables) related to diabetes. Through these variables, we can judge the connection with the diagnosis result. 769 observations were collected with different factors that related to diabetes. Since obesity is one of the most common problems at present, we want to study the relationship between BMI and other variables first. Propensity matching score was applied for matching the treated and controlled observations. The data of BMI were separated into two different levels,smaller than 25.0 and lager or equal to 25.0. We created a new binary variable named *BMI_num* in order to set up the logstic model. *Pregnancies*. *Glucose*, *BloodPressure*, *SkinThickness*, *Insulin*, *DiabetesPedigreeFunction* and *Age* are all considered as variables of this model.

Table 1: Meaning of Variables

Variables	Meanings
Pregnancies	times of pregnancy
Glucose	concentration of plasma glucose in an oral glucose tolerance test
BloodPressure	blood pressure
SkinThickness	thickness of triceps skin fold
Insulin	serum insulin in 2 hrs
DiabetesPedigreeFunction	a function that score the probability of having diabetes based on family history
Age	The age of observations

Model

Now we need to build a logistic model to predict the result of *Outcome*, which is whether an observation is having diabetes or not.

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 \tag{1}$$

$$\frac{p}{1-p} = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8} \tag{2}$$

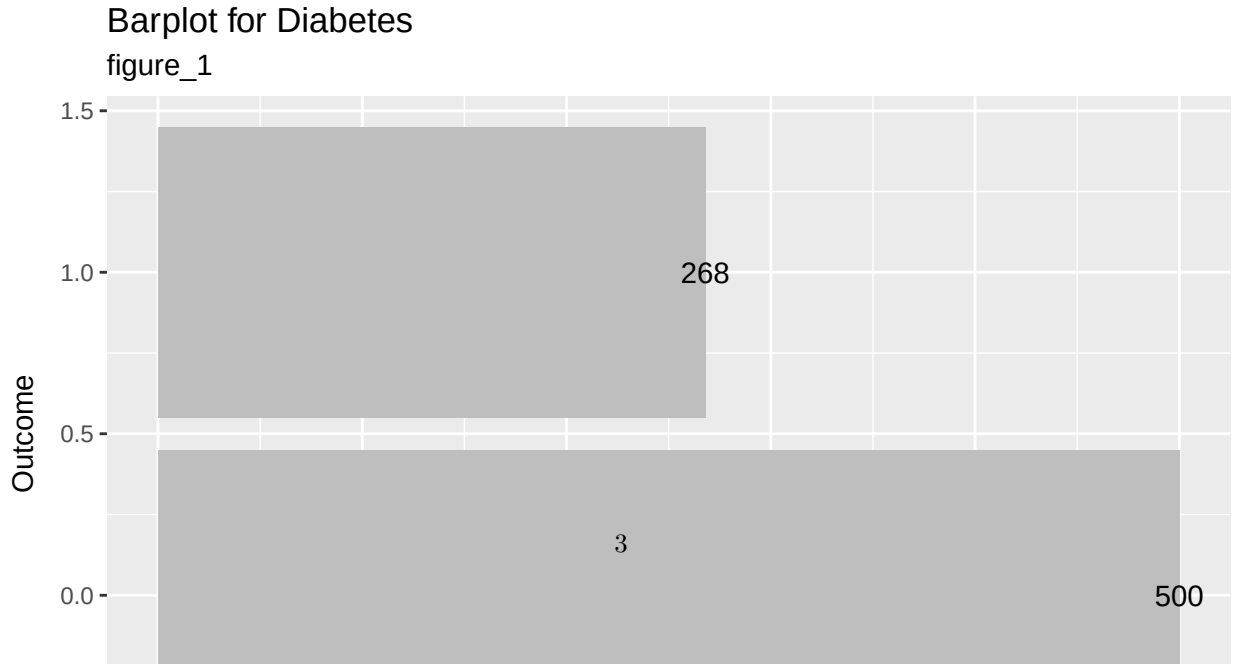
Result

We draw a bar plot to see the Outcome of the dataset.

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

	(1)
(Intercept)	-0.193
	(0.212)
Pregnancies	0.058 *
	(0.021)
Glucose	0.001
	(0.002)
Insulin	0.006
	(0.006)
SkinThickness	0.006
	(0.009)
DiabetesPedigreeFunction	-0.403
	(0.331)
Age	0.005
	(0.004)
BMI_num	-0.063
	(0.137)
N	22
logLik	5.258
AIC	7.484

*** p < 0.001; ** p < 0.01; * p < 0.05.



Result of Logistic Model

Table 2: Coefficients of Variables

Variables	Coefficients
(Intercept)	-0.192866
Pregnancies	0.058395
Glucose	0.001185
BloodPressure	0.005582
SkinThickness	0.005710
Insulin	0.005582
DiabetesPedigreeFunction	-0.402789
Age	0.005358
BMI_num	-0.063085

Discussion

The coefficient of BMI_num is -0.063085 which implies that a person with lower BMI is less likelihood have diabetes. It is essential to keep body weight within a normal range. According to Table 2 above, we noticed that coefficients of *Pregnancies*, *BloodPressure*, *SkinThickness*, *Insulin* and *Age* are positive, which means that the rise of these variables shows a positive correlation with diabetes. People who are experiencing a higher insulin level often associated with Hyperinsulinemia, which is related to diabetes.

People who experienced more times of pregnancies are likely to have a higher chance of getting diabetes than others. Move on to the concentration of plasma glucose in an oral glucose tolerance test, a higher index indicates a higher probability of having diabetes. And as the age increases, the risk of diabetes is also increasing.

Weakness

We are very satisfied that the result of β is in line with our expectation. Besides, all the variables are independent which makes the bias smaller. However, there are several weakness of the model. The sample size of this dataset is not that large. It is hard for us to make a percise conclusion. Also, even though all of the variables are highly related to our interest outcome, it would be better to include the gender, Country of Citizenship or A1C result in order to compare the difference among countries, female or male.

Next step

Next we want to use a more complex model such as Bayer's model to make predictions in order to get more accurate results. Also, since obesity is not the only reason for diabetes, we would like to find out the connections between the disease and other factors.

Reference

Pima Indians Diabetes Database UCI Learning - https://www.kaggle.com/uciml/pima-indians-diabetes-database

Diabetes https://www.who.int/news-room/fact-sheets/detail/diabetes

Body Mass Index (bmi) Calculator https://www.diabetes.ca/managing-my-diabetes/tools---resources/body-mass-index-(bmi)-calculator

Hyperinsulinemia: Is It Diabetes? https://www.mayoclinic.org/diseases-conditions/type-2-diabetes/expert-answers/hyperinsulinemia/faq-20058488

The Uk Is the Fattest Country in Europe. The Number Of Obese Adults Is Forecast To Rise By 73% Over the Next 20 Years from To 26 Million People, Resulting in More Than a Million Extra Cases Of Type 2 Diabetes, Heart Disease and Cancer. 15th January 2019 By Editor- Editor - https://www.diabetes.co.uk/diabetes-and-obesity.html

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

David Robinson, Alex Hayes and Simon Couch (2020). broom: Convert Statistical Objects into Tidy Tibbles. R package version 0.7.3. https://CRAN.R-project.org/package=broom

Andrew Gelman and Yu-Sung Su (2020). arm: Data Analysis Using Regression and Multilevel/Hierarchical Models. R package version 1.11-2. https://CRAN.R-project.org/package=arm

Baptiste Auguie (2017). gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.3. https://CRAN.R-project.org/package=gridExtra

Hao Zhu (2020). kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. R package version 1.2.1. https://CRAN.R-project.org/package=kableExtra

Appendix

Code and data supporting this anaylysis is available at: https://github.com/Wang-Lucy107/Analysis-of-Diabetes