

Physiological Factors Affecting Diabetes in Women using Binary Logit Regression Model

Yiming Cao^{1, *}

¹ School of Mathematics, Nanjing University, Nanjing, Jiangsu 210033, China

* Corresponding Author Email: 201840047@smail.nju.edu.cn

Abstract. The purpose is to use statistical methods to analysis some female data, trying to figure out what physiological factors will affect the occurrence of diabetes on female and how. Method: using dataset from “AKSHAY DATTATRAY KHARE” in Kaggle, which comes from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) in original. In particular, all patients here are females of Pima Indian descent who are at least 21 years old. Use binary logit regression model to analysis. Result: Pregnancies, skin thickness, insulin and blood pressure, as long as age can't be taken into account ($p > 0.05$). The remaining three elements, which is glucose (OR=1.039), BMI(OR=1.073), diabetes pedigree function (OR=3.130), are all risk factors. The overall correct rate is 78.3%. The conclusion is using binary logit regression model can analysis factors related diabetes well. The logistic regression model among the influencing factors of diabetes shows that glucose, BMI, and diabetes pedigree function are likely to cause diabetes, and diabetes pedigree function has a greater impact on the psychological function of patients.

Keywords: Diabetes; logit regression model; affecting factors.

1. Introduction

Diabetes mellitus is one of the most common diseases in the world, characterized by the patient's blood sugar being higher than normal for a long time. Moreover, if not treated in time, diabetes can cause many complications, influencing patients' health. In 2019, World Health Organization (WTO) released top 10 global causes of death, where diabetes is ninth on the chart. Since 2000, diabetes has increased by as much as 70%, which leads to it became one of the 10 leading causes of death. Among these 10 causes, diabetes is also the cause of the largest increase in deaths among men, with an 80% increase since 2000. In 2021, the number of people with diabetes worldwide has reached 537 million. Among these people, China has over 140 million adults with diabetes, maintaining the country with most diabetes patients in the world. As a result, studying on factors affecting diabetes is of vital importance, which helps us to learn more about diabetes patients and thereby improve their quality of life.

Former researchers have done a lot of work in diabetes. Ambady and Ronald studied Several distinct features of factors of diabetes mellitus and Asian populations' thresholds, in order to aware the public that the economic burden due to diabetes at personal, societal, and national levels is huge [1]. They emphasize the necessity for the government to take steps to raise public awareness about the disease and to improve standard of care and implementation of programs for primary prevention are urgently needed. Gale and Gillespie reviewed what kinds of role gender acts in diabetes [2], then came to the conclusion that although gender has hardly effects on Type I diabetes. It is witnessed that populations with the highest incidence all show male excess. Furthermore, fathers with Type I diabetes are more likely than affected mothers to transmit the condition to their offspring. Besides, men appear to be more susceptible to laziness and obesity than women, possibly due to differences in regional fat deposition and insulin sensitivity. However, women are more likely to pass type 2 diabetes on to their offspring. There also exist global report on diabetes by WHO, which exhaustively introduced global burden of diabetes, preventing diabetes, managing diabetes and national capacity for prevention and control of diabetes [3]. Besides, Willian and Erin described the method and data sources for these diabetes predictions, with key implications discussed [4]. They find that despite medical advances and prevention efforts, diabetes will remain a major health crisis in the United

States. Finally, they built a model to predict future trends in diabetes in the U.S. Moreover, Cathy studied the influence of stress on diabetes, whose research has indicated that stressful experiences have an impact on diabetes, leading to bad effects on the onset of diabetes [5]. As a result, it's important to have greater understanding of stress, as well as awareness of the need of psychosocial support for people in this predicament.

There are also a lot of works that use statistical methods to study on different aspects of diabetes. Gao and Hu reviewed the epidemiological characteristics of diabetes mortality and the main causes of deaths due to diabetes in China, finding that the mortality caused by diabetes is on the rise [6]. Then they provide a basis for controlling the diabetes epidemic and reducing the risk of death, suggesting that government should strengthen early intervention to control the diabetes epidemic. Zhang used decision tree CHAID model and logistic regression model to analyze the influencing elements of patients' life quality who have type 2 diabetes [7], to understand the current situation of the quality of life of patients with type 2 diabetes, and then drew the conclusion that type 2 diabetes The quality of life of diabetic patients needs to be improved, and the disease has a greater impact on the psychological function of patients. Chen discussed the relationship between diabetes epidemiological investigation and cancer risk, using logistic regression to analysis the relationship between diabetes and tumors [8]. They found that diabetes is a risk factor for the occurrence of malignant tumors and age, gender and so on are risk factors for diabetes. While coffee, fruit, dancing can reduce the risk of diabetes. Besides, there also exist researches (Dai,2022) that analyzed of the changing trend of diabetes-related disease mortality and the probability of premature death among residents in Yangpu District, Shanghai [9]. They found that the mortality caused by diabetes-related disease is on the rise. Besides, male mortality and premature death rates are increasing too for all three diseases. Also, Zhang analyzed the domestic and foreign research and development trends in the field of diabetes from the aspects of papers, patents, drugs, and industrial development [10], coming to the conclusion that it is necessary to take measures to strengthen the transformation of research results and product development to meet the huge demand for diabetes diagnosis and management in China in the future.

However, none of them have particularly analyzed diabetes when associated with females. Although male is easier to be affected by diabetes, we cannot ignore the existence of female diabetes patients too. As a result, what will do in this paper is using statistical methods to analysis some female data, trying to figure out what physiological factors will affect the occurrence of diabetes on female and how.

2. Methodology

This dataset is from AKSHAY DATTATRAY KHARE in Kaggle, which comes from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) in original. There are several limitations to selecting these instances from larger databases. In particular, all patients here are females of Pima Indian descent who are at least 21 years old. Considering that in this data analysis, there are only two values of the dependent variable(outcome), binary logit regression analysis can be chosen as the analysis method of this data. There are 3 steps in this method.

2.1. Data Processing

In order to analysis data, first data needs to be processed. First and foremost, transform categorical data into quantitative data by setting dummy variables. In addition, binary logistic regression requires that the dependent variable can only be 2 items, and the numbers must be 0 and 1. The number 1 means YES, willing, buying, sick, etc., and the number 0 means no, unwilling, not buying, not sick etc. If this is not the case, then the data needs to be coded against the dependent variable Y.

2.2. Chi-Squared Test or Analysis of Variance

Before using the model to analysis, it's important to first know the difference relationship in the independent variables and then it is possible to further have an influence relationship. Therefore, before performing binary logistic regression analysis, you can first perform a difference analysis on X to screen out X that is different from Y.

If X is categorical data, then to analyze differences, using chi-square analysis is better. On the contrary, use analysis of variance to study the difference between X and Y if X is quantitative data.

After analyzing the difference of X and Y, filter out X which is different, then put it into the model for binary logistic regression, which has two purposes. One is to simplify the model, because the simpler the model, the easier it is to be fitted with a even better result. The other is to be aware of the general relationship between the data in advance.

2.3. Binary Logit Regression Analysis

The fitting of the binary Logistic regression analysis model or the judgment of the model effect will involve three indicators, namely the Hosmer and Lemeshow test, the R² value and the model prediction accuracy table. The Hosmer and Lemeshow test are used to test whether the factual data is consistent with the model fitting results. If the P value is greater than 0.05 during the Hosmer and Lemeshow test, it means that the factual data is consistent with the model fitting results, which means that the model fits in good condition.

3. Results and Discussion

3.1. Descriptive Statistical Analysis

Before starting modeling analysis, first some descriptive statistical analyses are provided. The result is as follows.

Table 1. Overall description of all elements

	Number	Minimum	Maximum	Average	S. D.
Pregnancies	392	0	17	3.3	3.211
Glucose	392	56	198	122.63	30.861
BloodPressure	392	24	110	70.66	12.496
SkinThickness	392	7	63	29.15	10.516
Insulin	392	14	846	156.06	118.842
BMI	392	18.2	67.1	33.086	7.027
DiabetesPedigreeFunction	392	0.085	2.420	0.523	0.345
Age	392	21	81	30.86	10.201
Outcome	392	0	1	0.33	0.471

From the table 1 it can be verified that all the dependent variables satisfy the normal distribution.

3.1.1 Dependent variable

First, analyze the outcome and get the result in Figure 1. It can be seen that in all 392 data, the number of people without diabetes (0 value) is 262, and the number of people with diabetes (1 value) is 130. In both cases the numbers of people are all large, as a result the data is of statistical value. At the same time, diabetic patients account for about 33%, while the global proportion of diabetic patients is 10.5%. Although the data is not in line with the actual situation, larger percentage of diabetic patients can help us to better study the data and obtain more accurate results.

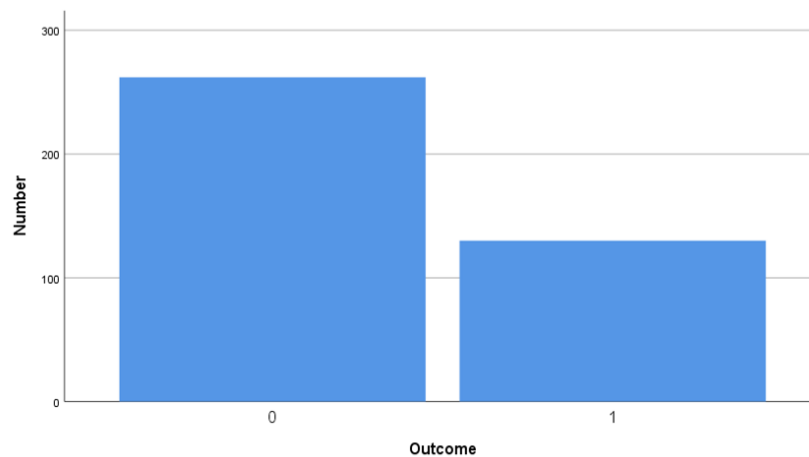


Fig. 1 Description of outcome

3.1.2 Independent variable

When it comes to studying the factors that affect diabetes among women, one thing that differs a lot from studying men is women's pregnancy. Therefore, a boxplot for pregnancies and outcomes is made, and the chart and figure are as follow. As is shown in Figure 2 and table 2, women with diabetes, on average, have more children than women without diabetes. Also, the difference in the number of children born among women with diabetes was greater than among women without diabetes, with a variance of 6.853 for an outcome of 0 and a variance of 15.336 for an outcome of 1.

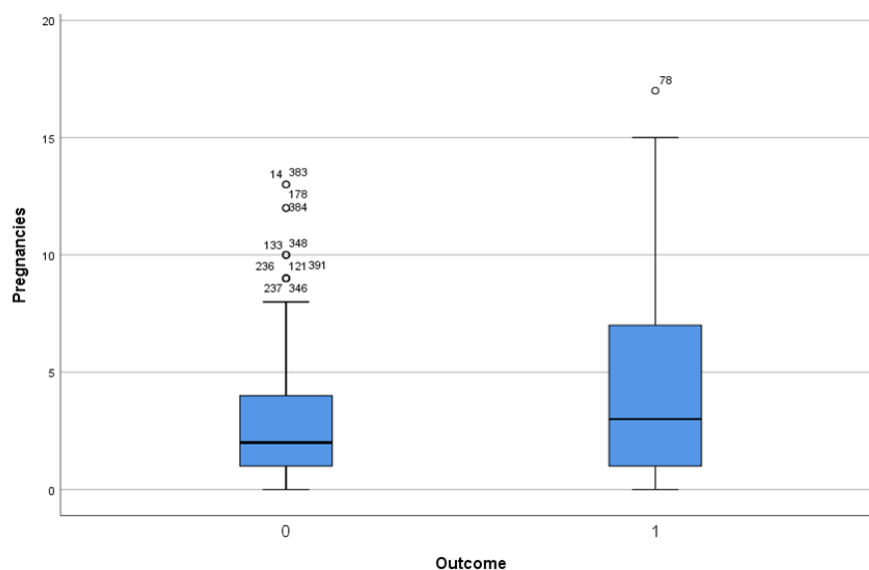


Fig. 2 Boxplot of pregnancies and outcome

Table 2. Data of pregnancies related with outcome

Outcome		statistic	S. D.
0	Average	2.72	0.162
	Median	2.00	-
	Variance	6.853	-
	Interquartile range	3	-
1	Average	4.47	0.343
	Median	3.00	-
	Variance	15.336	-
	Interquartile range	6	-

Next, since it's diabetes that is discussed, what can't be ignored is glucose. Here are the boxplot and related data of glucose and outcome. As can be seen from figure 3 and table 3, women with diabetes tend to have apparently higher glucose content than those who have no diabetes problem. This result is not surprising, as it's well known that glucose is close related with diabetes. However, here are some scatters in the boxplot when women don't have diabetes, which indicates that even some women have high glucose content, they still don't have diabetes. It can be conjectured that this result maybe be affected by other elements, which will be discussed later.

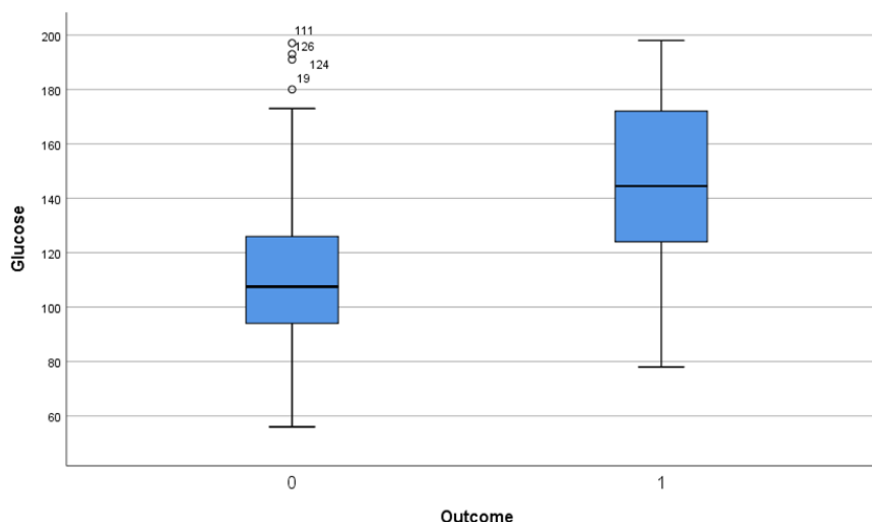


Fig. 3 Boxplot of glucose and outcome

Table 3. Data of glucose related with outcome

Outcome		Statistic	S. D.
0	Average	111.43	1.522
	Median	107.50	
	Variance	607.235	
	Interquartile range	24.642	
1	Average	145.19	2.617
	Median	144.5	
	Variance	890.389	
	Interquartile range	29.839	

Another element that deserves to be discussed is age. As can be seen from figure 4 and table 4, women who have diabetes are on average older than women without diabetes. However, there are still some women who have diabetes in a young age, which is younger than some of women who don't have diabetes. On the opposite, many old ladies are also free from diabetes. The deeper reason will be discussed below.

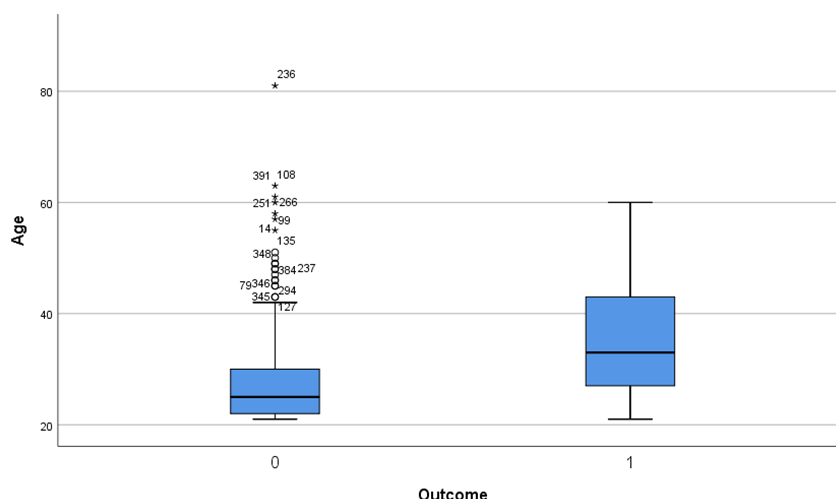


Fig. 4 Boxplot of age and outcome

Table 4. Data of age related with outcome

Outcome		Statistic	S. D.
0	Average	28.35	0.555
	Median	25.00	
	Variance	80.802	
	Interquartile range	8	
1	Average	35.94	0.933
	Median	33.00	
	Variance	113.097	
	Interquartile range	16	

Apart from these elements mentioned above, there also exist other elements that maybe related to diabetes. In this paper, the blood pressure measurement (BloodPressure), the thickness of the skin (SkinThickness), the insulin level in blood (Insulin), Body mass index (BMI) and diabetes percentage (DiabetesPedigreeFunction) are also included. The results and tables, along with figures are omitted, considering space limitations and similarity. Values of women with diabetes are much higher than those who have no diabetes. How these relates with the outcome of diabetes will be discussed later.

3.2. Logit Regression Analysis

The data has been processed. All rows with missing data have been deleted. Moreover, categorical data has been transformed into quantitative data by setting dummy variables, setting having no diabetes as 0, having diabetes as 1. As what is mentioned above, analysis of variance is required for every independent element. The result table 5 is as follow.

Table 5. Analysis of variance

	Outcome=0	Outcome=1	F	P
Pregnancies	262	130	21.196	<0.01
Glucose	262	130	124.340	<0.01
BloodPressure	262	130	15.036	<0.01
SkinThickness	262	130	27.337	<0.01
Insulin	262	130	32.875	<0.01
BMI	262	130	30.881	<0.01
DiabetesPedigreeFunction	262	130	17.873	<0.01
Age	262	130	48.902	<0.01

It can be seen from the table 5 that the significance of each element is less than 0.01, which means that all the dependent elements have differences with the independent element. As a result, all the elements can be taken into logit regression model.

Next logit regression can be proceeded. With diabetes as the dependent variable (0 for those without diabetes and 1 for those with diabetes), logistic regression analysis was performed with statistically significant variables in univariate analysis as independent variables. Among them, the respective variables are continuous variables, which can be directly substituted into the model.

Table 6. Logit regression analysis of related elements in female patients who have diabetes

	β	SE	Wald χ^2	P	OR(95%CI)
Pregnancies	0.082	0.055	2.197	0.138	1.086(0.974-1.210)
Glucose	0.038	0.06	44.025	<0.01	1.039(1.027-1.051)
BloodPressure	-0.001	0.012	0.014	0.904	0.999(0.976-1.022)
SkinThickness	0.011	0.017	0.431	0.511	1.011(0.978-1.046)
Insulin	-0.01	0.001	0.399	0.528	0.999(0.997-1.002)
BMI	0.071	0.027	6.655	0.010	1.073(1.017-1.132)
DiabetesPedigreeFunction	1.141	0.427	7.125	0.001	3.130(1.354-7.233)
Age	0.034	0.018	3.412	0.065	1.035(0.998-1.072)

As is shown in table 6, pregnancies, blood pressure, skin thickness, insulin and age can't be taken into account ($p>0.05$). The remaining three elements, which is glucose (OR=1.039), BMI(OR=1.073), diabetes pedigree function (OR=3.130), are all risk factors.

Table 7. Correct rate

	0	1	Correct rate
outcome	0 233	29	88.9%
	1 59	74	56.9%
overall			78.3%

Table 7 shows the correct rate of this model. The overall correct rate is 78.3%, which shows its reliability.

3.3. Discussion

3.3.1 Glucose

Diabetes is a chronic disease characterized by an increase in blood glucose concentration. Years of high blood sugar will cause irreversible damage to various internal organs, eventually leading to complications and death. As a result, it's not strange to find glucose is risk factor of diabetes. In the process of diabetes control, monitoring blood sugar is a very important thing, because the results of blood sugar monitoring can be used to judge blood sugar control on the one hand, and can also be used to guide treatment on the other hand.

3.3.2 BMI

Actually, many researches have proved that BMI has close relationship with diabetes. It's said that after the BMI exceeds 22-23, the risk of diabetes increases significantly. That is to say, if the BMI exceeds 22 before the age of 25, the risk of developing diabetes in the future will increase. Over time, it was found that the two groups with a BMI of 22-23 and those with a BMI around 23 had a very high risk of developing diabetes, while the group with a BMI of less than 21 had a very small risk.

Besides, research also found that even having moderately elevated BMI is also associated with increased risk of developing diabetes and its DM complications. As a result, being overweight can lead to a variety of diseases, including dyslipidemia and cardiovascular disease, in addition to diabetes risk. As a result, weight management should start in adolescence. For those older ladies, they should pay more attention to diet and exercise, so as to maintain their BMI and reduce the risk of diabetes.

3.3.3 Diabetes Pedigree Function

It can be easily inferred that diabetes percentage has effect on diabetes itself, which is also the reason why it has such a large number of OR.

4. Conclusion

Through this study, the Logistic regression model among the influencing factors of diabetes shows that glucose, BMI, and diabetes pedigree function are likely to cause diabetes, and diabetes pedigree function has a greater impact on the psychological function of patients. In the future, medical staff should focus on diabetic patients with these above factors, and actively improve the disease cognition level and psychological function of such patients. For ordinary women, they should pay more attention to their diet, control sugar intake, and maintain BMI, which can prevent diabetes well. Limitations of this study: Convenience sampling method was adopted, and the selected samples were only from India, with a slightly small sample size and selection bias. Future research should expand the sample size and use stratified sampling to investigate multiple countries, cities, hospitals, and communities.

References

- [1] Ramachandran A, Ma R C W, Snehalatha C. Diabetes in asia. *The Lancet*, 2010, 375(9712): 408-418.
- [2] Gale E A M, Gillespie K M. Diabetes and gender. *Diabetologia*, 2001, 44(1): 3-15.
- [3] World Health Organization. Global report on diabetes. World Health Organization, 2016.
- [4] Rowley W R, Bezold C, Arikan Y, et al. Diabetes 2030: insights from yesterday, today, and future trends. *Population health management*, 2017, 20(1): 6-12.
- [5] Lloyd C, Smith J, Weinger K. Stress and diabetes: a review of the links. *Diabetes spectrum*, 2005, 18(2): 121-127.
- [6] Mingfei Gao, Ruying, Chonggao Hu. Research progress on the epidemiological characteristics of diabetes death in China. *Preventive Medicine*, 2022, 34(07): 692-695.
- [7] Wenhui Zhang, Yanmei Wang, Xijuan Wei, Pei Yang. Analysis of influencing factors of quality of life in patients with type 2 diabetes based on decision tree and Logistic regression model. *Modern Clinical Nursing*, 2022, 21(05): 1-8
- [8] Jing Chen, Weiping Liu, Ru Liu. Epidemiological investigation of diabetes among residents over 40 years old and correlation study of tumor risk. *Chinese Cancer Clinic and Rehabilitation*, 2022, 29(08): 941-944. DOI: 10.13455/j.cnki.cjcor.2022.08.12.
- [9] Yin Dai, Hui Li, Jia Zhao, Zhengzheng Zhang, Xue Han, Jianhua Yin. 2002-2020 Analysis on the Mortality Trend of Diabetes-related Diseases and the Probability of Early Death among Residents in Yangpu District, Shanghai. *Shanghai Preventive Medicine*, 2022, 1-7. <http://kns.cnki.net/kcms/detail/31.1635.R.20220718.1820.022.html>.
- [10] Liwen Zhang, Meihua Ruan, Jialan Liu, Caihong He, Jianrong Yu. Analysis of research and development situation in the field of diabetes. *Heredity*, 2022, 1-26. DOI: 10.16288/j.yczz.22- 272.