CSE 780 Assignment 2 2021-02-05

Diabetes Logistic Regression Analysis

Noah Frank, 400050264 frankn1@mcmaster.ca

1 Data Description

1.1 Data Source

The data used for this assignment is the 'Pima Indians Diabetes Database' originally the National Institute of Diabetes and Digestive and Kidney Diseases, but retrieved from Kaggle.[1] The data is comprised of information from 768 individuals taken in 1990, all patients being females of at least 21 years old of Pima Indian heritage.

1.2 Variable Details

The data set is made up of 8 features and 1 outcome variable. All of the features are numeric variables, and the outcome is a binary variable denoted whether a patient has diabetes or not. For the blood pressure variable, 35 zero values were found which is biologically impossible. These values were replaced with the median blood pressure from the respective outcome of that patient. The same was done for the 11 zero values for BMI, and the 5 zero values for glucose. The variables used in the analysis are described here:

- **Pregnancies:** Number of pregnancies the patient has had (0-17)
- **Glucose:** Plasma glucose concentration after 2 hours in an oral glucose tolerance test (44-199 mg/dl)
- **Blood Pressure:** Diastolic blood pressure (24-122 mm Hg)
- **Skin Thickness:** Triceps skin fold thickness (0-99 mm)
- **Insulin:** 2-Hour serum insulin (0-846 mu/ml)
- **BMI:** Body mass index (18.2-67.1)
- **Diabetes Pedigree Function:** Function scoring family history of diabetes (0.078-2.42)
- **Age:** Age of the patient (21-81)
- **Outcome:** Whether a patient has diabetes (1) or not (0)

1.3 Descriptive Analysis

For this descriptive analysis we will focus on some biological factors that are most common when dealing with diabetes. These will include glucose, insulin, and pregnancies. The glucose measurements are taken after 2 hours of a glucose test, where a higher amount of glucose in the blood means that it is not being metabolized insulin and taken into a patient's cells. The glucose measurements range from 44 to 199 mg/dl with a mean value of 121.68. Similarly, a test is done measuring insulin levels in the blood 2 hours after receiving glucose. Around 48% of insulin levels were 0 throughout the dataset, with the highest value being 846 mu/ml and a mean of 79.8 mu/ml.

In Fig. 1 we can see a scatter plot of glucose levels against all non-zero insulin levels. We can observe that the blue points representing patients with diabetes are more prevalent in high levels of glucose, whereas those without diabetes can be seen on the lower end of the glucose scale. Insulin response seems to be proportional to glucose levels regardless of diabetes or not, however this is only looking at the patients with a non-zero insulin response.

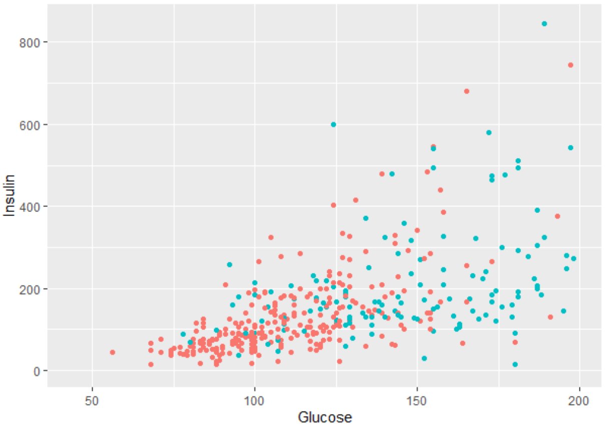

Figure 1: Glucose Levels vs. Insulin Levels (Left), Count of Glucose Levels (right)

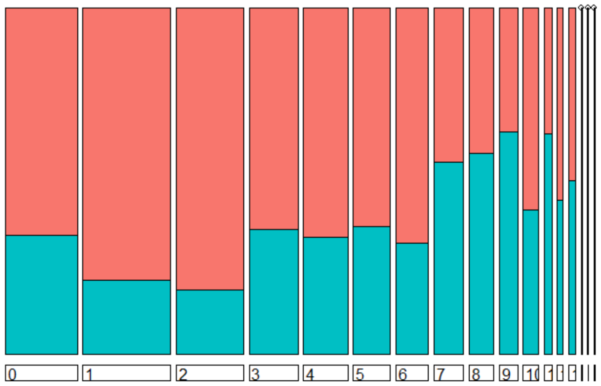

Figure 2: Representation of race in census grouped by income level

In Fig. 2 we can see a doubledecker plot of the the number of pregnancies against and the proportion of cases of diabetes. There is a clear trend showing that women who have had more pregnancies have a higher proportion of diabetes than lower amounts of pregnancies.

[1] UCI Machine Learning Kaggle. *Pima Indiands Diabetes Database*. 2016. URL: https://www.kaggle.com/uciml/pima-indians-diabetes-database.

2 Problem Description

The aim of this analysis is to determine what are the major factors which can determine whether an individual has diabetes or not. Specifically, the analysis will also look at how biological markers such as blood glucose levels, insulin levels, and blood pressure affect diabetes. Through this analysis we will find the greatest predictors for diabetes, and which factors increase the odds of an individual having diabetes.

3 Technique Description

Binary logistic regression is used to predict the relationship between independent feature variables and a dependant predicted binary variable. A multiple binary logistic regression model uses multiple independent predictor variables of many types such as continuous, binary, and categorical.[2] The model can be described as follows:

$$\log[\tfrac{\pi(x)}{1-\pi(x)}] = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p,$$
$$\pi(x) \in (0,1)$$

Deviance can be calculated in order to determine the goodness of fit of a model. The deviance can be compared between models with differing parameters to measure the effect of parameters on the goodness of fit. The deviance D and change in deviance G are given as follows:

$$D = -2\log[\mathcal{L}_{\text{fitted model}}]$$
$$G = D_{\text{without variable}} - D_{\text{with variable}}$$
$$G = -2\log[\tfrac{\mathcal{L}_{\text{without}}}{\mathcal{L}_{\text{with}}}]$$

Knowing that $G \sim \chi^2_v$, with v denoting the degrees of freedom as the difference in the number of variables, we can test to see whether the new model without variables can replace the original model with variables.

In order to see the effect of each variable on the odds of the outcome, we may look at the odds ratio. The odds ratio is defined as follows:

$$\text{Odds Ratio} = e^{\beta_i}$$

This expression will give us the change in the odds of the outcome, given a unit change in the predictor variable X_i.

4 Result Description

The multiple binary logistic regression analysis described in Section 3 was carried out using the 'glm' method in R, with the logit function. When comparing models with removed variables, a test was carried out using the 'pchisq' method against the null hypothesis that the new model is just as good as the original model. The first model was computed using all the variables from the data set. From that model, the variables which were the least significant were blood pressure, skin thickness, insulin and age. We then created a second model, by removing the least significant parameters of the first model listed previously. Based on the chi-squared test, we failed to reject the null hypothesis that the new model fits as well as the old model. Attempting to remove any more parameters lead to rejecting the null hypothesis. Thus, the final model comprised of glucose, pregnancies, BMI, and the diabetes pedigree function (DPF). The final model can be summarized in Table 1.

Table 1: Final Model Parameters

Variable	Estimate	Std. Error	p-value	Odds Ratio
Intercept	-9.222	0.706	<2e-16	-
Pregnancies	0.142	0.028	2.43e-07	1.153
Glucose	0.037	0.003	<2e-16	1.037
BMI	0.089	0.015	1.47e-09	1.093
DPF	0.883	0.296	0.003	2.418

From our final model, ranking the variables in order of significance, we have: glucose, BMI, pregnancies and the diabetes pedigree function. As all of these variables are continuous variables, the odds ratios are not very large. However, we can see that they are all leading to an increase in odds for diabetes as these measure increase.

5 Conclusions

Our results show that the major factors in our model for determining if an individual has diabetes are the blood glucose levels, number of pregnancies, body mass index, and the diabetes pedigree function. The blood glucose level and BMI were the strongest predictors of diabetes, which is to be expected as these are the current metrics used for determining if someone has diabetes. This is due to an individual with diabetes having an inadequate insulin response to elevated glucose levels. This can be expected as insulin is released as a response to elevated blood glucose levels, and we must keep in mind that diabetes can strongly effect and even eliminate the insulin response.[3] We might have expected insulin to be a strong predictor, however it was not. This can be due to the fact that almost half of the patients in the data set had an insulin level of 0 mU/ml. A level this low can be both a sign of a healthy decrease in insulin after a successful response to glucose, or also a complete lack of insulin response due to diabetes. The outcome variable does not differentiate between type-1 or type-2 diabetes, thus the insulin metric between diabetics can very greatly. Our model also showed the effects of multiple pregnancies on diabetes, increasing the odds of having diabetes by 1.153 for each pregnancy. This is likely due to gestational diabetes, a form of diabetes caused by pregnancy which is especially prevalent among American Indians, such as the Pima Indians. Our model successfully demonstrates the prevalence of certain biological factors due to diabetes.

[2] Prof. Sharon McNicholas. *CSE780: Logistic Regression*. McMaster University, School of Computational Science & Engineering. 2021.

[3] Centers for Disease Control and Prevention. *The Insulin Resistance–Diabetes Connection*. 2019. URL: https://www.cdc.gov/diabetes/basics/insulin-resistance.html.