A Statistical Tool For The Prediction Of Diabetes Mellitus

Priya.P.Menon[1], Amritha Varshini P[2]
[1, 2] Dept of Statistics
[1, 2] Maharaja's College (Autonomous), Kochi, Kerala, India

Abstract- *The data consists of 392 subjects from a larger study on diabetes of a population. The study uses logistic regression to predict whether a person is diabetic or not which can be used as early prevention before complications occur. The analysis is done using the statistical software R.*
We arrive at a conclusion that the diabetes of a person is affected by the factors like Glucose, Body Mass Index, Diabetes Pedigree Function and Age. Thereby the probability of a person being diabetic can be computed.

Keywords- Diabetes Mellitus, Logistic Regression, R software, Stepwise regression

I. INTRODUCTION

Logistic Regression is a mathematical modeling approach that can be used to describe the relationship of several predictor variables to a dichotomous dependent variable. Logistic regression is perfect for situations where you are trying to predict whether something "happens" or not. The logistic Model is a mathematical model that describes the relationship of a dichotomous (binary) dependent variable with several independent variables which may be quantitative or categorical or mixture of these. Due to this advantage it is widely used in epidemiology and survival analysis where the data is often dichotomous/categorical. Also the sigmoidal shape of the logistic function, which describes the mathematical model, makes it easier to predict. Logistic regression is a way to draw conclusions of response variable when it is binary. As in linear regression we are looking for a relationship between our response variable and a set of independent variables.

In the present study we arrive at a reduced logistic model by eliminating the insignificant variables from the actual model and identify the variables that contribute more towards the incidence of Diabetes Mellitus are . Hence the probability of a person being diabetic can be computed.

II. METHODS AND MATERIALS

The data consists of 392 subjects from a larger study on diabetes of a population. The Population for this study is the Pima Indian population near Phoenix, Arizona. That population has been under continuous study by the National Institute of Diabetes and Digestive and Kidney Diseases because of the high incidence rate of diabetes. The dependent variable is the binary variable, diabetic or not and 8 independent variables are non-categorical variables. The study uses logistic regression to predict whether a person is diabetic or not which can be used as early prevention before complications occur. The analysis is done using the statistical software R.

Separating data into training and testing sets is an important part of evaluating data mining models. Typically, when we separate a data set into a training set and testing set, most of the data is used for training, and a smaller portion of the data is used for testing. Analysis Services randomly samples the data to help ensure that the testing and training sets are similar. By using similar data for training and testing, we can minimize the effects of data discrepancies and better understand the characteristics of the model.

After a model has been processed by using the training set, we test the model by making predictions against the test set. Because the data in the testing set already contains known values for the attribute that we want to predict, it is easy to determine whether the model's guesses are correct.

In the analysis process that we carry out, 75% of the data is included in the training set and the rest 25% is taken as testing set.

Variable Names

Here is a summary of the data:

- Pregnancies – Number of times pregnant
- Glucose – Plasma glucose concentration a 2 hours in an oral glucose tolerance test
- BloodPressure – Diastolic blood pressure (mm Hg)
- SkinThickness – Triceps skin fold thickness (mm)
- Insulin – 2 Hour serum insulin (mu U/ml)
- BMI – Body Mass Index (height to weight in kg/m)
- DiabetesPedigreeFunction – Diabetes Pedigree Function (DPF)

- Age – Age (years)

The variable names are

Outcome – Class Variable (0 – healthy or 1- diabetic)

- **y : OUT :** Outcome
- **x_1: PRE :**Pregnancies
- **x_2: GLU :** Glucose
- **x_3: BP :**BloodPressure
- **x_4: ST :**SkinThickness
- **x_5:INS :**Insulin
- **x_6:BMI :**BodyMassIndex
- **x_7:DPF :**DiabetesPedigreeFunction
- **x_8:AGE :**Age

III. RESULTS AND DISCUSSION

The general model is

$$Y = \alpha + \sum_{i=1}^{8} \beta_i x_i$$

We use the backward stepwise regression which appears to be the preferred method of explanatory analysis, where the analysis begins with a full or saturated model and variables are eliminated from the model in an iterative process. The fit of the model is tested after the elimination of each variable to ensure that the model still adequately fits the data. When no more variables can be eliminated from the model, the analysis has been completed.

The above mentioned assumptions are satisfied by the data chosen for analysis. In the first step, the 8 independent variables together with constant α are included in the model. The analysis shows that independent variable X_3 (BP) is the least significant. Hence the variable x_3 (BP) is eliminated in step2. In Step3, the independent variable x_5 (INS) is the least significant. Thus the variable x_5 (INS) is eliminated in step3. In a similar way, variable x_4(ST) is eliminated from step4 and variable x_1(PRE) is eliminated from step5.The rest of the independent variables are found significant with constant α.

The reduced model is

$$Y = \alpha + \beta_2 x_2 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8$$

The values of the regression coefficients α and β_i are obtained from the table 1

The reduced model can be written as

$$Y = -10.05970 + 0.03723x_2 + 0.07525x_6 + 1.06693x_7 + 0.04885x_8$$

The accuracy obtained using this fitted model is **84.69%** which is inferred from the confusion matrix of the testing set.

Corresponding probability of being diabetic,

$$f(y) = \frac{1}{1+e^{-y}}$$

Table1. Estimates of the significant variables

	Estimate	Std. Error	z value	Pr(>|z|)
Intercept)	-10.059701	1.238917	-8.120	4.67e-16(***)
GLU	0.037233	0.006034	6.170	6.81e-10(***)
BMI	0.075251	0.024008	3.134	0.00172(**)
DPF	1.066931	0.489490	2.180	0.02928(*)
AGE	0.048853	0.015550	3.142	0.00168(**)

Signif.codes: ***** - Significant at 0.1%**
**** - Significant at 1%**
*** -Significant at 5%**

The significant variables obtained are x_2(GLU), x_6(BMI), x_7(DPF), x_8(AGE). Thus we arrive at a conclusion that the diabetes of a person is affected by the factors like Glucose, BMI, Diabetes Pedigree Function, Age. Thus probability of a person being diabetic can be computed.

IV. CONCLUSION

We arrived at a logistic model which can be utilized to compute the probability of an individual to be diabetic . The study can be used as an early prevention strategy before complications occur. Moreover, the variables that contribute towards the incidence of diabetes are identified to be **Glucose, Body Mass Index, Diabetes Pedigree Function and Age.**

REFERENCES

[1] V Altman D G.”Practical Statistics For Medical Research” , Chapman and Hall,1991
[2] Darlington R.B “Regression And Linear Models”, New York: McGraw-Hill, chapter 18. 1990
[3] David G.K, Mitchel Klein “Logistic Regression: A Self-Learning Text”, Second Edition,
Springer, New York.2002
[4] Hosmer D.W. and Lameshow. S, “Applied Logistic Regression”, Second Edition, John
Wiley & Sons, Inc, New York 2000

[5] Press S.J. and Wilson, S "Choosing Between Logistic Regression and Discriminant
Analysis" in Journal of the American Statistical Association, 73,pp. 699-705. 1978
[6] Anderson. T.W " An Introduction to Multivariate Statistical Analysis", Second Edition,
John Wiley & Sons.
[7] Rao. C.R, "Linear Statistical Inference and its Applications", John Wiley.