Analysing of Diabetes using Logistic Regression model.

AIMS | African Institute for Mathematical Sciences
RWANDA

Albert NDENGEYINTWALI: AIMS232401904

Malaria modelling,
African Institute for Mathematical Sciences.

Supervisor: Prof. Evans Gouno

Kigali – October 30, 2023

Contents

1. Introduction
2. Problem statement
3. Objectives
4. Methodology
5. Results
6. Conclusion

Introduction: Table showing detailed information

Table: Diabetes csv

ni	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	Y
1	6	148	72	35	0	33.6	0.627	50	1
2	1	85	66	29	0	26.6	0.351	31	0
3	8	183	64	0	0	23.3	0.672	32	1
4	1	89	66	23	94	28.1	0.167	21	0
5	0	137	40	35	168	43.1	2.288	33	1
6	5	116	74	0	0	25.6	0.201	30	0
7	3	78	50	32	88	31	0.248	26	1
8	10	115	0	0	0	35.3	0.134	29	0
9	2	197	70	45	543	30.5	0.158	53	1
10	8	125	96	0	0	0	0.232	54	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
762	9	170	74	31	0	44	0.403	43	1
763	9	89	62	0	0	22.5	0.142	33	0
764	10	101	76	48	180	32.9	0.171	63	0
765	2	122	70	27	0	36.8	0.34	27	0
766	5	121	72	23	112	26.2	0.245	30	0
767	1	126	60	0	0	30.1	0.349	47	1
768	1	93	70	31	0	30.4	0.315	23	0

You can click (here) to see a whole dataset.

Explanation of variables

Dependent variable

'Outcome' takes on values of 0 and 1, where 0 typically represents the absence of diabetes (no) and 1 represents the presence of diabetes (yes).

Independent variables

1. 'Pregnancies': The number of times a person has been pregnant(Unit: Count).
2. 'Glucose': Blood sugar level (Unit: mg/dL).
3. 'BloodPressure': Blood pressure measurement (Unit: mm Hg).
4. 'SkinThickness': Skinfold thickness measurement (Unit: mm).
5. 'Insulin': Insulin level (Unit: mu U/ml).
6. 'BMI': Body Mass Index, a measure of body fat based on height and weight (Unit: kg/m^2).
7. 'DiabetesPedigreeFunction': A function that represents the diabetes history in relatives and the genetic influence (Unit: Dimensionless)
8. 'Age': Age of the individual (Unit: Years).

.

About Logistic Regression that is going to be used

- Logistic regression is a statistical analysis method to predict a binary outcome, such as yes or no, based on prior observations of a data set.
- Taking Y as dependent variable , $X_i's$ are independent variables and $\beta_i's$ as regression coefficients. The modal is written as follow:

Formula

$$P(Y=1) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n}} \quad \text{In case outcome is yes.}$$

$$P(Y=0) = 1 - P(Y=1) \quad \text{In case outcome is no.}$$

$$\text{Odds of outcome} = \frac{P(Y=1)}{P(Y=0)}$$

$$\text{Log-Odds (Logit)} = \ln\left(\frac{P(Y=1)}{P(Y=0)}\right)$$

Why Logistic Regression ?

- Logistic regression streamlines the mathematics for measuring the impact of multiple variables $(X_i's)$ with a given outcome(Y).
- It can also estimate the probabilities of events, including determining a relationship between features and the probabilities of outcomes.

Problem statement

- In project, we consider diabetes csv, as represented by the 'Outcome' variable, in a dataset containing information about individuals.
- The dataset includes the following independent variables: 'Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI', 'DiabetesPedigreeFunction' and 'Age'.
- We intend to use Logistic Regression (LR) to understand how these factors collectively influence the likelihood of diabetes.

Objectives

Main objective

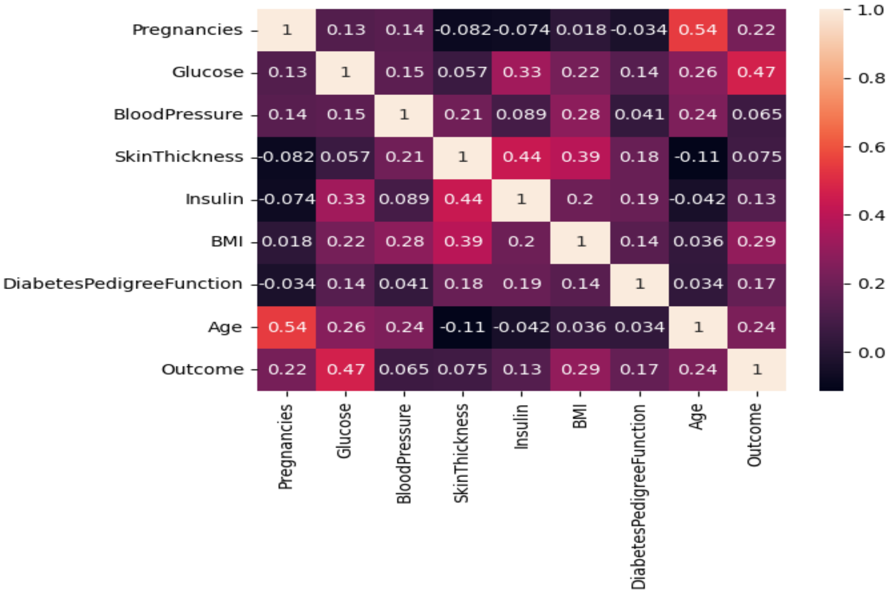
The aim of this project is to build logistic regression that would likely be to predict whether an individual is at risk for diabetes (1) or not at risk for diabetes (0).

Methodology

The general methodology we followed based on this analysis is as follows:

- Data Preparation.
- Exploratory Data Analysis.
- Logistic Regression Model Building.
- Interpretation of Coefficients and P-Values addition to finding the probabilities of having diabetes or not.

Results: General information



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Pregnancies	1	0.13	0.14	-0.082	-0.074	0.018	-0.034	0.54	0.22
Glucose	0.13	1	0.15	0.057	0.33	0.22	0.14	0.26	0.47
BloodPressure	0.14	0.15	1	0.21	0.089	0.28	0.041	0.24	0.065
SkinThickness	-0.082	0.057	0.21	1	0.44	0.39	0.18	-0.11	0.075
Insulin	-0.074	0.33	0.089	0.44	1	0.2	0.19	-0.042	0.13
BMI	0.018	0.22	0.28	0.39	0.2	1	0.14	0.036	0.29
DiabetesPedigreeFunction	-0.034	0.14	0.041	0.18	0.19	0.14	1	0.034	0.17
Age	0.54	0.26	0.24	-0.11	-0.042	0.036	0.034	1	0.24
Outcome	0.22	0.47	0.065	0.075	0.13	0.29	0.17	0.24	1

★ **Note**:Glucose seems to be most correlated with outcome because it has the highest value compared to others which is 0.47 but it is weak

Results: Output from R in computing L.R

```
data <- read.csv('/home/albert/Downloads/archive/diabetes.csv')
X <- data[, c('Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI',
              'DiabetesPedigreeFunction', 'Age')]
y <- data$Outcome
model <- glm(Outcome ~ ., data = data, family = binomial)
summary(model)
```

```
##
## Call:
## glm(formula = Outcome ~ ., family = binomial, data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5566  -0.7274  -0.4159   0.7267   2.9297
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -8.4046964  0.7166359 -11.728  < 2e-16 ***
## Pregnancies               0.1231823  0.0320776   3.840 0.000123 ***
## Glucose                   0.0351637  0.0037087   9.481  < 2e-16 ***
## BloodPressure            -0.0132955  0.0052336  -2.540 0.011072 *
## SkinThickness             0.0006190  0.0068994   0.090 0.928515
## Insulin                  -0.0011917  0.0009012  -1.322 0.186065
## BMI                       0.0897010  0.0150876   5.945 2.76e-09 ***
## DiabetesPedigreeFunction  0.9451797  0.2991475   3.160 0.001580 **
## Age                       0.0148690  0.0093348   1.593 0.111192
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

Results: Interpretations of coefficients and p-values

The column of "estimated" shows the value of all $\beta_i's$ and their related p-values in column of "$Pr(>|z|)$"

1. **const (Intercept)**: The constant term is -8.4047. This represents the estimated log-odds of the response variable when all other predictors are zero.
2. **Pregnancies**: For each additional pregnancy, the log-odds of the response variable increase by 0.1232. The p-value is significant ($p = 0.001$), indicating a statistically significant effect.
3. **Glucose**: For each additional unit of glucose, the log-odds of the response variable increase by 0.0352. The p-value is significant ($p = 0.001$), indicating a statistically significant effect.
4. **BloodPressure**: For each additional unit of blood pressure, the log-odds of the response variable decrease by -0.0133. The p-value is significant ($p = 0.011$), indicating a statistically significant effect.
5. **SkinThickness**: The log-odds of the response variable change very slightly (0.0006) for each additional unit of skin thickness. The p-value is not significant ($p = 0.929$), suggesting that skin thickness may not be a strong predictor.
6. **Insulin**: For each additional unit of insulin, the log-odds of the response variable decrease by -0.0012. The p-value is not significant ($p = 0.186$), suggesting that insulin may not be a strong predictor.
7. **BMI**: For each additional unit of BMI, the log-odds of the response variable increase by 0.0897. The p-value is significant ($p = 0.001$), indicating a statistically significant effect
8. **DiabetesPedigreeFunction**: For each additional unit of the diabetes pedigree function, the log-odds of the response variable increase by 0.9452. The p-value is significant ($p = 0.002$), indicating a statistically significant effect.variables constant.
9. **Age**: For each additional year of age, the log-odds of the response variable increase by 0.0149. The p-value is not significant ($p = 0.111$), suggesting that age may not be a strong predictor.

Result: Output from R in computing confusion matrix

```
data <- read.csv('/home/albert/Downloads/archive/diabetes.csv')
X <- data[, c('Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI',
              'DiabetesPedigreeFunction', 'Age')]
y <- data$Outcome

model <- glm(Outcome ~ ., data = data, family = binomial)
predicted_probabilities <- predict(model, newdata = data, type = "response")
tab<-table(predicted_probabilities >0.5,data$Outcome)
tab
```

```
##
##           0   1
##   FALSE 445 112
##   TRUE   55 156
```

```
sum(diag(tab))/sum(tab)*100
```

```
## [1] 78.25521
```

```
table(data$Outcome)
```

```
##
##   0   1
## 500 268
```

```
500/(500+268)*100
```

```
## [1] 65.10417
```

Result:Interpretations of confusion matrix

- **True Negatives (TN):** The model correctly identified 445 cases as not having diabetes.
- **False Positives (FP):** The model incorrectly predicted 112 cases as having diabetes when they don't.
- **False Negatives (FN):** The model incorrectly predicted 55 cases as not having diabetes when they do.
- **True Positives (TP):** The model correctly identified 156 cases as having diabetes.
- **78.25521% Accuracy:** represents the percentage of correct predictions made by the model for a binary classification task (0 for not having diabetes and 1 for having diabetes).
- **65.10417% Accuracy:** This accuracy value appears to be calculated as 500 / (500 + 268), which is the accuracy achieved when you simply predict the majority class (in this case, 0 for not having diabetes).

★ An accuracy of *78.25521%* is higher than the baseline accuracy of *65.10417%*, indicating that **the model is performing better.**

Result: Probability of having diabetes or not

- First, taking an individual with $x_i^{(0)} \in X_i$ for $i = 1, 2, \cdots 8$, Example:

$$\underbrace{x_1^{(0)} = 4, x_2^{(0)} = 121, x_3^{(0)} = 69, x_4^{(0)} = 21, x_5^{(0)} = 80, x_6^{(0)} = 32, x_7^{(0)} = 0.47, x_8^{(0)} = 33}$$

- Probability of having diabetes $P(Y = 1) = \dfrac{e^{\beta_0 + \sum_{i=1}^{8} \beta_i x_i^{(0)}}}{1 + e^{\beta_0 + \sum_{i=1}^{8} \beta_i x_i^{(0)}}} \approx 0.299$
- Probability of not having diabetes $P(Y = 0) = 1 - P(Y = 1) \approx 0.701$
- Odds of Diabetes $= \dfrac{P(Y = 1)}{P(Y = 0)} = \dfrac{0.299}{0.701} \approx 0.426$
- Log-Odds (Logit)$=$ $\ln\left(\dfrac{P(Y = 1)}{P(Y = 0)}\right) \approx -0.853$

Conclusion

- In conclusion, our logistic regression analysis indicates that the probability of having diabetes is estimated at 0.299, while the probability of not having diabetes is estimated at 0.701. Again,the model is performing better since An accuracy of 78.25521% is higher than the baseline accuracy of 65.10417%.
- This suggests that several factors, including Pregnancies, Glucose,BloodPressure, BMI and DiabetesPedigreeFunction,are significant predictors of the outcome, suggesting higher values of these variables are associated with an increased or decreased likelihood of the outcome based on their related reression coefficient signs. However, the influence of other variables, such as SkinThickness, Insulin, and Age, are not statistically significant in this analysis.

Bibliography

- Gareth James . Trevor Hastie . Daniela Witten . Robert Tibshirani, *An Introduction to Statistical Learning with Applications in R.*, Second Edition, Corrected Printing: June 21, 2023.
- LaValley, Michael P., *Logistic regression.*,volume 117. Am Heart Assoc, 2008.
- Menard, Scott., *Applied logistic regression analysis.*, Sage, 2002.

End

Thank You