**Title: Logistic Regression: An Overview and Application Report**

**Introduction**

Logistic regression is a statistical modeling technique used to analyze the relationship between a binary or categorical dependent variable and one or more independent variables. It is widely employed in various fields, including biomedical research, social sciences, finance, marketing, and more. This report provides an overview of logistic regression, explaining its principles, assumptions, advantages, limitations, and practical applications.

**Principles of Logistic Regression**

- Definition and Purpose

Logistic regression models the probability of an event occurring by estimating the odds of the event happening relative to the odds of it not happening. It is used when the dependent variable is categorical, such as yes/no or presence/absence, and cannot be directly predicted using linear regression.

- Logistic Function

Logistic regression employs the logistic function (also known as the sigmoid function) to transform the linear combination of independent variables into a probability value between 0 and 1. The logistic function is defined as:

P(Y=1|X) = 1 / (1 + e^(-z))

Where P(Y=1|X) represents the probability of the dependent variable being 1 (success) given the values of the independent variables X, and 'z' represents the linear combination of the independent variables.

**Assumptions of Logistic Regression**

- Binary or Categorical Dependent Variable

Logistic regression assumes that the dependent variable is binary or categorical in nature.

- Independence of Observations

The observations used in logistic regression should be independent of each other to avoid violating the assumption.

- Linearity of Independent Variables and Log Odds

Logistic regression assumes a linear relationship between the independent variables and the logarithm of the odds ratio.

- No Multicollinearity

Logistic regression requires that the independent variables are not highly correlated with each other to avoid multicollinearity issues.

**Advantages of Logistic Regression**

-Suitable for modeling binary or categorical outcomes.

-Provides interpretable coefficients that represent the impact of independent variables on the log odds or odds ratio.

-Can handle both continuous and categorical independent variables.

-Allows for the inclusion of interaction terms to capture complex relationships.

-Robust to outliers.

**Limitations of Logistic Regression**

-Assumes a linear relationship between independent variables and log odds, which may not hold in all cases.

-Requires a sufficient sample size to ensure stable parameter estimates.

-Cannot handle missing data effectively.

-Assumes independence of observations.

-Prone to overfitting if too many independent variables are included compared to the sample size.

**Practical Applications of Logistic Regression**

Logistic regression finds applications in various fields, including:

-Predicting the likelihood of disease occurrence based on risk factors.

-Assessing the impact of marketing campaigns on customer response (e.g., purchase or not).

-Analyzing factors influencing customer churn in telecommunications or subscription-based services.

-Evaluating the effectiveness of interventions or treatments on binary outcomes.

-Modeling credit risk to predict loan default probabilities.

**Conclusion**

Logistic regression is a powerful statistical technique used to model the relationship between a binary or categorical dependent variable and independent variables. It offers valuable insights into the probability of an event occurring and allows for the interpretation of the impact of variables on the outcome. By understanding the principles, assumptions, advantages, and limitations of logistic regression, researchers can apply this method effectively in their respective fields to gain valuable insights and make informed decisions.

***A logistic regression on Diabetes Healthcare***

***Introduction***

NIDDK (National Institute of Diabetes and Digestive and Kidney Diseases) research creates knowledge about and treatments for the most chronic, costly, and consequential diseases. The dataset used in this project is originally from NIDDK. The objective is to predict whether a patient has diabetes, based on certain diagnostic measurements included in the dataset.  
we will try to Build a model to accurately predict whether the patients in the dataset have diabetes or not.

Dataset Description: The datasets consist of several medical predictor variables and one target variable (Outcome). Predictor variables includes the number of pregnancies the patient has had, their BMI, insulin level, age, and more.

Variables Description:  
-Pregnancies Number of times pregnant  
-Glucose Plasma glucose concentration in an oral glucose tolerance test  
-Blood Pressure Diastolic blood pressure (mm Hg)  
-BMI Body Mass Index  
-DiabetesPedigreeFunction Diabetes pedigree function  
-Age in years  
-Outcome Class variable (either 0 or 1).

***First: Descriptive analysis***

**For the discrete variables:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | | | |
| Outcome table 1 | | | | | | |
|  | | | Frequency | Percent | | Valid Percent | Cumulative Percent |
| Valid | 0 | | 475 | 65.6 | | 65.6 | 65.6 |
| 1 | | 249 | 34.4 | | 34.4 | 100.0 |
| Total | | 724 | 100.0 | | 100.0 |  |

* We can see that there are 475 persons in our data set doesn’t suffer from diabetes with 65.6% , and there are 249 persons from 729 suffer from diabetes with 34.4%.

**For the continuous variables:**

Table 2

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Descriptive Statistics** | | | | | |
|  | N | Minimum | Maximum | Mean | Std. Deviation |
| Pregnancies | 724 | 0 | 17 | 3.87 | 3.363 |
| Glucose | 724 | 44 | 199 | 121.88 | 30.750 |
| BloodPressure | 724 | 24 | 122 | 72.40 | 12.380 |
| BMI | 724 | 18.2 | 67.1 | 32.467 | 6.8889 |
| DiabetesPedigreeFunction | 724 | .078 | 2.420 | .47477 | .332315 |
| Age | 724 | 21 | 81 | 33.35 | 11.765 |
| Valid N (listwise) | 724 |  |  |  |  |

**Comment:**

* For the glucose variable, we can see the mean is 121.88 mg/dl which means that the average glucose measure in our data set is 121.88 mg/dl. and we have a standard deviation of 30.75 which is a large value indicating that the data are more spread out from the mean. The maximum glucose measure in the data is 199 and the minimum is 44.
* For the blood pressure variable, we can see that the mean is 72.4 mm/hg which means that the average blood pressure in our data is 72.4. and the st.d is 12.38 which is a large value indicating that the data are more spread out from the mean. The maximum blood pressure measure is 122 and the minimum is 24.
* For the BMI the mean is 32.467 which mean that the average BMI measure in our data is 32.467. and the standard deviation is 6.88 which mean that the data are more spread out. We have a maximum BMI measure which is 67.1 and a minimum value of 18.2.
* For the Diabetes Pedigree Function variable, we have a mean of 0.474 which mean that the average Diabetes Pedigree Function measure in our data is 0.474. and the standard deviation is 0.332 which mean that the data is clustered around the mean. We have a maximum Diabetes Pedigree Function measure which is 2.420 and a minimum value of 0.078.
* for the age variable, we have mean of 33.35 which means that the average age in our data is 33.35 years. and a standard deviation of 11.765 which means that the data are more spread out. the maximum age in our data is 81 and the minimum age is 21.
* For pregnancy variable , we can see the mean is 3.87 which means that the average number of pregnancies in our data is 3.87. and we have a standard deviation of 3.363 which is a large value indicating that the data are more spread out from the mean. The maximum number of pregnancies is 17 and minimum is 0 .

***Second: logistic regression***

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Variables in the Equation table 4** | | | | | | | |
|  | | B | S.E. | Wald | df | Sig. | Exp(B) |
| Step 0 | Constant | -.646 | .078 | 68.145 | 1 | .000 | .524 |

* in this table 4 we can see an equation includes only the intercept “the constant “and it reflects block 0

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Omnibus Tests of Model Coefficients table 5** | | | | |
|  | | Chi-square | df | Sig. |
| Step 1 | Step | 259.076 | 6 | .000 |
| Block | 259.076 | 6 | .000 |
| Model | 259.076 | 6 | .000 |

* In table 5 we can see that the model that contains the other variables in the data set is more significant than the model that contains only the intercept. and that we have a good fitting model.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Variables in the Equation table 6** | | | | | | | | | |
|  | | B | S.E. | Wald | df | Sig. | Exp(B) | 95% C.I.for EXP(B) | |
| Lower | Upper |
| Step 1a | Pregnancies | .118 | .033 | 12.439 | 1 | .000 | 1.125 | 1.054 | 1.201 |
| Glucose | .035 | .004 | 95.316 | 1 | .000 | 1.036 | 1.029 | 1.043 |
| BloodPressure | -.009 | .009 | 1.070 | 1 | .301 | .991 | .975 | 1.008 |
| BMI | .091 | .016 | 33.372 | 1 | .000 | 1.095 | 1.062 | 1.130 |
| DiabetesPedigreeFunction | .961 | .306 | 9.826 | 1 | .002 | 2.613 | 1.433 | 4.764 |
| Age | .017 | .010 | 2.969 | 1 | .085 | 1.017 | .998 | 1.037 |
| Constant | -8.962 | .821 | 119.193 | 1 | .000 | .000 |  |  |
| a. Variable(s) entered on step 1: Pregnancies, Glucose, BloodPressure, BMI, DiabetesPedigreeFunction, Age. | | | | | | | | | |

* We can see in table 6 that there are 2 insignificant variables which are age and blood pressure so we will remove them and re model the variables .

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Variables in the Equation table 7** | | | | | | | | | |
|  | | B | S.E. | Wald | df | Sig. | Exp(B) | 95% C.I.for EXP(B) | |
| Lower | Upper |
| Step 1a | Pregnancies | .143 | .028 | 25.513 | 1 | .000 | 1.154 | 1.092 | 1.220 |
| Glucose | .036 | .004 | 104.151 | 1 | .000 | 1.036 | 1.029 | 1.044 |
| BMI | .084 | .015 | 31.716 | 1 | .000 | 1.088 | 1.057 | 1.120 |
| DiabetesPedigreeFunction | .973 | .304 | 10.228 | 1 | .001 | 2.647 | 1.458 | 4.806 |
| Constant | -9.000 | .717 | 157.554 | 1 | .000 | .000 |  |  |
| a. Variable(s) entered on step 1: Pregnancies, Glucose, BMI, DiabetesPedigreeFunction. | | | | | | | | | |

After removing the 2 insignificant variables we computed the model above in table 7, interpretations:

* The odds ratio for pregnancies is 1.154, meaning that the for the odds of a person suffering from diabetes (y=1) change by a factor of 1.154 with every unit increase on pregnancy. [Since we are multiplying odds by 1.154 per unit increase on the predictor, this must mean our odds are increasing with each increase on the predictor.]
* The odds ratio for glucose is 1.036, meaning that the for the odds of a person who suffers from diabetes (y=1) change by a factor of 1.036 with every unit increase on glucose. [Since we are multiplying odds by 1.154 per unit increase on the predictor, this must mean our odds are increasing with each increase on the predictor.]
* The odds ratio for BMI is 1.088, meaning that the for the odds of a person suffers of diabetes (y=1) change by a factor of 1.088 with every unit increase on pregnancy. [Since we are multiplying odds by 1.088 per unit increase on the predictor, this must mean our odds are increasing with each increase on the predictor.]
* The odds ratio for Diabetes Pedigree Function is 2.647, meaning that the for the odds of a person suffering from diabetes (y=1) change by a factor of 2.647 with every unit increase on Diabetes Pedigree Function. [Since we are multiplying odds by 2.647 per unit increase on the predictor, this must mean our odds are increasing with each increase on the predictor.]