LOGISTIC REGRESSION

Rishitha Pallala

5 August 2023

1 Introduction

Logistic regression differs from linear because the dependent variable is in binary form like true or false. This model fits an S-curve to the data and can work with independent variables (both continuous and discrete).

2 Logistic function and Analysis

2.1 Example and Log Odds Line

To understand better lets use the example :people are tested for covid and tested if the have fever at various degree. The body temperature is independent variable and covid diagnosis are true or false is the dependent variable.

Lets take the given points and plot them on an x-y plane. the x-axis has the independent variable and the y-axis has the log(odds) ranging from negative to positive infinity.

2.2 s-shape sigmoid function and Likelihood

This can be converted into plot that gives probability of dependent variable ranging from 0 to 1. the x-axis has fever and y-axis has probabilities(p) of positive cases .the equation(1) shows how to derive p from log(odds).

$$p = \frac{e^{logodds}}{1 + e^{logodds}} \tag{1}$$

After the plotting we get , s-shape function ready to find the likelihood.
Log likelihood of a single data point is the probability of getting the observed dependent variable. example: if 100^{o}F has outcome as positive (pos), equation (2) gives log likelihood for that data point. if 95^{o}F has outcome as negative (neg) ,equation (3) gives log likelihood for that data point.

$$loglikelihood(100^oF) = ll(p(100^o)) \tag{2}$$

$$loglikelihood(95^o F) = ll(1 - p(95^o)) \tag{3}$$

The log likelihood of the function is the sum of all log likelihoods.

2.3 Maximum likelihood

The function with the maximum likelihood will be detected and used by the model by iterative methods . The maximum likelihood ranges from 0 to negative infinity as the probabilities range from 0 to 1. So, closer the likelihood of the function to 0, the better it fits the data.

3 Evalution metrics

3.1 Finding McFadden's Pseudo R^2

Lets draw a line in the x-y plane with body temperature on x-axis, positive as 1 and negative as 0 on y-axis and given by equation (4)

$$y = p(positivecases) \tag{4}$$

The loglikelihood of each data point is : $log(y)$ if result is positive and $1 - log(y)$ if result is negative.
The likelihood of the this line is sum of all likelihoods along the line and given by $ll(overall)$.

The R^2 is given by equation (5) , where ll(fit) is the maximum likelihood (likelihood of the best fit sigmoid function.) and ll(saturated) is 0 for logistic regression and can be ignored. The more close the R^2 value is to 1 , the better the variation is explained by independent variable.

$$R^2 = \frac{ll(overall) - ll(fit)}{ll(overall) - ll(saturated)} \tag{5}$$

3.2 p-values

To calculate the p-value , we should take chi-squared distribution function who's degree of freedom is the difference in the parameters of $ll(overall)$ and $ll(fit)$ model. The p-value is given by equation(6) and it is calculated from the graph.

$$p - value = Pr(x >= 2(ll(fit) - ll(overall)) \tag{6}$$

Thus ,logistic regression is closely related to linear regression and it is useful in classification as we can assign probabilities.