What is the concept of the null hypothesis in logistic regression?

Authored by

stats writer

May 12, 2024

The null hypothesis in logistic regression is a statistical concept that assumes there is no relationship between the independent and dependent variables in a dataset. It is the default assumption that there is no significant effect of the independent variables on the outcome variable. The goal of logistic regression is to test this null hypothesis and determine whether there is a significant relationship between the variables. If the null hypothesis is rejected, it means that the independent variables do have a significant effect on the outcome variable, and a different model needs to be used to make predictions. However, if the null hypothesis is accepted, it suggests that the independent variables do not have a significant impact on the outcome variable, and the model can be used for making accurate predictions. In summary, the concept of the null hypothesis in logistic regression is crucial in determining the validity and usefulness of the model in predicting outcomes.

Understanding the Null Hypothesis for Logistic Regression

is a type of regression model we can use to understand the relationship between one or more predictor variables and a when the response variable is binary.

If we only have one predictor variable and one response variable, we can use simple logistic regression, which uses the following formula to estimate the relationship between the variables:

$$\log = \beta 0 + \beta 1X$$

The formula on the right side of the equation predicts the log odds of the response variable taking on a value of 1.

Simple logistic regression uses the following null and alternative hypotheses:

H0: β1 = 0HA: β1 ? 0

The null hypothesis states that the coefficient β1 is equal to zero. In other words, there is no statistically significant relationship between the predictor variable, x, and the response variable, y.

The alternative hypothesis states that β1 is *not* equal to zero. In other words, there *is* a statistically significant relationship between x and y.

If we have multiple predictor variables and one response variable, we can use multiple logistic regression, which uses the following formula to estimate the relationship between the variables:

log = β0 + β1x1 + β2x2 + ... + βkxk

Multiple logistic regression uses the following null and alternative hypotheses:

H0: β1 = β2 = ... = βk = 0HA: β1 = β2 = ... = βk ? 0

The null hypothesis states that all coefficients in the model are equal to zero. In other words, none of the predictor variables have a statistically significant relationship with the response variable, y.

The alternative hypothesis states that not every coefficient is simultaneously equal to zero.

The following examples show how to decide to reject or fail to reject the null hypothesis in both simple logistic regression and multiple logistic regression models.

Example 1: Simple Logistic Regression

Suppose a professor would like to use the number of hours studied to predict the exam score that students will receive in his class. He collects data for 20 students and fits a simple logistic regression model.

```
#create data
df <- data.frame(result=c(0, 0, 1, 0, 0, 0, 1, 0, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1),
hours=c(1, 5, 5, 1, 2, 1, 3, 2, 2, 1, 2, 1, 3, 4, 4, 2, 1, 1, 4, 3))

#fit simple logistic regression model
model <- glm(result~hours, family='binomial', data=df)
```

#view summary of model fit
summary(model)

Call:
glm(formula = result ~ hours, family = "binomial", data = df)

Deviance Residuals:
Min 1Q Median 3Q Max
-1.8244 -1.1738 0.7701 0.9460 1.2236

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.4987 0.9490 -0.526 0.599
hours 0.3906 0.3714 1.052 0.293

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 26.920 on 19 degrees of freedom
Residual deviance: 25.712 on 18 degrees of freedom
AIC: 29.712

Number of Fisher Scoring iterations: 4

#calculate p-value of overall Chi-Square statistic
1-pchisq(26.920-25.712, 19-18)

0.2717286

To determine if there is a statistically significant relationship between hours studied and exam score, we need to analyze the overall Chi-Square value of the model and the corresponding p-value.

We can use the following formula to calculate the overall Chi-Square value of the model:

X2 = (Null deviance - Residual deviance) / (Null df - Residual df)

The p-value turns out to be 0.2717286.

Since this p-value is not less than .05, we fail to reject the null hypothesis. In other words, there is not a statistically significant relationship between hours studied and exam score received.

Example 2: Multiple Logistic Regression

Suppose a professor would like to use the number of hours studied and the number of prep exams taken to predict the exam score that students will receive in his class. He collects data for 20 students and fits a

multiple logistic regression model.

We can use the following code in R to fit a multiple logistic regression model:

```
#create data
df <- data.frame(result=c(0, 0, 1, 0, 0, 0, 1, 0, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1),
hours=c(1, 5, 5, 1, 2, 1, 3, 2, 2, 1, 2, 1, 3, 4, 4, 2, 1, 1, 4, 3),
exams=c(1, 2, 2, 1, 2, 1, 1, 3, 2, 4, 3, 2, 2, 4, 4, 5, 4, 4, 3, 5))

#fit simple logistic regression model
model <- glm(result~hours+exams, family='binomial', data=df)

#view summary of model fit
summary(model)
```

Call:

```
glm(formula = result ~ hours + exams, family = "binomial", data = df)
```

Deviance Residuals:

Min 1Q Median 3Q Max

-1.5061 -0.6395 0.3347 0.6300 1.7014

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.4873 1.8557 -1.879 0.0602 .
hours 0.3844 0.4145 0.927 0.3538
exams 1.1549 0.5493 2.103 0.0355 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 26.920 on 19 degrees of freedom
Residual deviance: 19.067 on 17 degrees of freedom
AIC: 25.067

Number of Fisher Scoring iterations: 5

#calculate p-value of overall Chi-Square statistic
1-pchisq(26.920-19.067, 19-17)

0.01971255

The p-value for the overall Chi-Square statistic of the model turns out to be 0.01971255.

Since this p-value is less than .05, we reject the null hypothesis. In other words, there is a statistically significant relationship between the combination of hours studied and prep exams taken and final exam score received.

The following tutorials offer additional information about logistic regression: