Learning the Null Hypothesis in Logistic Regression: A Beginner’s Guide

Authored by

Mohammed looti

November 2, 2025

Introduction to Logistic Regression and Binary Outcomes

Logistic Regression is an essential statistical modeling tool designed specifically for analyzing the relationship between various predictor variables and a categorical response. It is most commonly applied when the outcome variable is binary, meaning it can only assume one of two possible states, such as success/failure, presence/absence, or 0/1. Understanding this framework is crucial because it governs how we determine significance in hypothesis testing.

A key distinction separates logistic regression from standard linear regression: instead of predicting a continuous numerical value, logistic regression estimates the probability of a specific event occurring. To achieve this, it utilizes the logit function, which mathematically transforms the probability scale (0 to 1) into the scale of log odds (negative infinity to positive infinity). This transformation allows the use of a linear combination of predictors for estimation, providing a solid mathematical foundation for modeling probabilities.

The fundamental objective when fitting any logistic model is to rigorously assess whether the introduced predictors genuinely influence the outcome. This assessment relies entirely upon statistical hypothesis testing. At the heart of this process lies the Null Hypothesis (H0), which serves as the default assumption that no relationship or effect exists between the independent variables and the categorical response.

Modeling Relationships with Simple Logistic Regression

When our statistical inquiry is focused on the impact of only one predictor variable (X) on the binary response (Y), we employ **simple logistic regression**. This streamlined approach is invaluable for initial investigations or when researchers possess strong theoretical justification for isolating a single, primary factor of interest. The precise mathematical formula underlying this model is the foundation upon which we define and test our statistical hypotheses.

The core structure of simple logistic regression involves expressing the log odds of a successful outcome, p(X), as a linear function of the predictor variable X. This transformation ensures that the predictions remain statistically sound while maintaining interpretability based on the linear model form:

log = β0 + β1X

In this equation, β0 represents the intercept of the model, and β1 is the regression coefficient specifically tied to the predictor X. Crucially, the magnitude and sign of the β1 coefficient quantify the change in the log odds of the response variable (Y=1) for every one-unit increase in X. It is this coefficient that becomes the subject of our hypothesis testing.

Formulating the Null Hypothesis (H0) for Single Predictors

Hypothesis testing in the context of simple logistic regression centers entirely on evaluating whether the estimated coefficient β1 deviates significantly from zero. If the true value of β1 were zero, then changes in the predictor X would produce absolutely no change in the log odds of the outcome, effectively rendering X useless as a predictor.

The formal Null and Alternative Hypotheses utilized to assess the significance of the relationship between the single predictor and the binary response are defined as follows:

H0 (Null Hypothesis): β1 = 0
HA (Alternative Hypothesis): β1 ? 0

The Null Hypothesis (H0) states explicitly that the regression coefficient β1 is equal to zero. In substantive terms, this is the claim that there is no statistically detectable or statistically significant relationship between the predictor variable (X) and the response variable (Y). The goal of most statistical studies is typically to gather sufficient evidence to reject this conservative statement, thereby validating the relevance of the predictor.

Conversely, the **Alternative Hypothesis** (HA) asserts that β1 is not equal to zero. If the statistical evidence--usually in the form of a small p-value--allows us to reject H0 in favor of HA, we confidently conclude that X makes a significant contribution to predicting the log odds of the outcome, confirming a meaningful association between X and Y.

Expanding the Scope: Multiple Logistic Regression

When researchers need to account for the combined or controlled effects of several factors, they incorporate two or more predictor variables (x1, x2, ..., xk) to model the single binary response. This advanced technique is known as **multiple logistic regression**. It is indispensable for complex predictive modeling, as it allows for the simultaneous evaluation of each variable's unique influence while statistically controlling for all other factors included in the model.

The mathematical representation of multiple logistic regression is a direct extension of the simple model. It maintains the linear structure on the log odds scale by simply adding terms for every additional predictor introduced:

log = β0 + β1x1 + β2x2 + ... + βkxk

In this multivariate framework, each coefficient βi is interpreted as the change in the log odds resulting from a one-unit increase in the corresponding predictor xi, assuming all other predictors are held constant. Therefore, a positive coefficient indicates that increasing xi increases the likelihood (log odds) of the event (Y=1), isolated from the effects of the other variables.

Global Hypothesis Testing in Multivariate Models

In multiple logistic regression, before evaluating individual predictors, a crucial preliminary step is the global test of significance. This test determines if the full model--the model containing all predictors--is statistically superior to the null model, which contains only the intercept (β0). This overall assessment addresses the collective predictive capability of the entire set of variables.

The global Null Hypothesis (H0) and Alternative Hypothesis (HA) address whether all regression coefficients (excluding the intercept) are simultaneously equal to zero:

H0: β1 = β2 = ... = βk = 0
HA: At least one βi is not equal to zero.

If the global Null Hypothesis (H0) is true, it implies that every predictor variable included in the model is statistically irrelevant; collectively, they have zero impact on the log odds of the outcome. Failing to reject this hypothesis means the model is insignificant and provides no meaningful predictive insight beyond simply using the overall probability (the intercept).

The global test of significance often relies on comparing the Null Deviance to the Residual Deviance, yielding a Chi-Square statistic. If H0 is rejected in favor of HA, we gain statistical confidence that the overall model significantly improves the fit to the data compared to a model without predictors. This validates the combination of the chosen variables as predictors for the binary outcome.

Case Study 1: Analyzing Simple Regression Results

Consider a scenario where a university professor wishes to predict student success on an exam (Result: 1=Pass, 0=Fail) based solely on the number of hours studied (Hours). After collecting data from 20 students, the professor fits a simple logistic regression model. The central goal is to test the specific null hypothesis: H0: βhours = 0, meaning hours studied has no effect on passing the exam.

The statistical output below, generated using R, summarizes the initial model fitting and provides the necessary values for hypothesis testing:

#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),
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
```

The global significance of the model is tested using the difference between the Null Deviance (26.920) and the Residual Deviance (25.712), which results in a Chi-Square test statistic. This calculation yields an overall model p-value of **0.2717286**.

Given that the calculated p-value (0.2717) is substantially greater than the conventional significance threshold (alpha = 0.05), we must **fail to reject the null hypothesis** (H0: βhours = 0). Consequently, we conclude that, based on this specific dataset, there is insufficient statistical evidence to assert a significant relationship between the number of hours studied and the likelihood of a student passing the exam.

Case Study 2: Interpreting Multivariate Outcomes

Recognizing the limitations of the simple model, the professor attempts to enhance predictive accuracy by including a second predictor: the number of preparatory exams taken (Exams). This requires fitting a multiple logistic regression model, which must satisfy the expanded global null hypothesis: H0: βhours = βexams = 0.

The following R code executes the fitting of the multiple logistic regression model, incorporating both Hours and Exams as independent variables:

```
#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),
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
```

Upon calculating the change in deviance between the null model and the full model (26.920 - 19.067), the resulting overall model significance test yields a p-value of **0.01971255**.

Since this p-value (0.0197) is clearly less than the standard 0.05 significance threshold, we have gathered sufficient evidence to **reject the global null hypothesis** (H0: βhours = βexams = 0). The conclusion is that the model incorporating both hours studied and preparatory exams taken provides a statistically significant fit to the data. Moreover, examining the individual coefficients shows that 'Exams' is individually significant (p = 0.0355), while 'Hours' remains insignificant (p = 0.3538), suggesting that preparatory exams are the primary effective predictor variable in this specific multivariate model.

Conclusion and Further Reading

Mastering the formulation and testing of the Null Hypothesis is central to correctly interpreting any

Logistic Regression analysis. Whether dealing with a simple predictor or a complex multivariate model, the Null Hypothesis always provides the crucial benchmark: the assumption of no effect. Rejecting it allows researchers to move forward with confidence in their predictive models.

To further consolidate your understanding of hypothesis testing, coefficient interpretation, and model assessment, the following resources offer invaluable technical detail:

A detailed breakdown of the interpretation of the Log Odds and their relationship to probability ratios.
Guides explaining how the Chi-Square statistic relates to deviance in generalized linear models, particularly for model comparison.
In-depth tutorials on using the P-value to establish statistical significance and making appropriate decisions regarding H0 and HA.