

How to Test the Null Hypothesis in Logistic Regression

Authored by
stats writer

December 3, 2025

RECOMMENDED CITATION

stats writer (2025). *How to Test the Null Hypothesis in Logistic Regression*.

PSYCHOLOGICAL SCALES. Retrieved from <https://scales.arabpsychology.com/?p=104378>

The core concept of the Null Hypothesis (H_0) in statistical modeling, including Logistic Regression, posits that there is absolutely no statistically significant relationship or effect present between the variables under study. Specifically, for a Logistic Regression model, the Null Hypothesis states that the log odds of the outcome occurring are independent of the predictor variables included in the model.

This translates mathematically to the statement that all regression coefficients (the Beta values, β) associated with the independent variables are zero. If H_0 holds true, it implies that the presence or absence of the predictor variables does not influence the likelihood of the binary outcome. Effectively, the odds of the event happening are statistically indistinguishable from the odds of it not happening, regardless of the input data.

Understanding the Fundamentals of Logistic Regression

The Logistic Regression model is a fundamental statistical tool used when the response or dependent variable is binary, meaning it can only take on two possible values (e.g., 0 or 1, Yes or No, Success or Failure). Unlike linear regression, which predicts a continuous numerical outcome, logistic regression estimates the probability that an event belongs to a particular category. It achieves this by utilizing the logistic function (or sigmoid function) to map the linear combination of the input variables to a probability value between zero and one.

The central output of the model is not the probability itself, but the transformation of probability known as the log odds, also called the logit. This transformation ensures that the model's prediction remains within the bounds of a probability distribution. Understanding this relationship is critical because the coefficients we test under the Null Hypothesis are interpreted in terms of changes in these log odds, not changes in probability directly. This allows us to quantify how strongly the **predictor variables** influence the likelihood of the outcome occurring.

The Essential Role of Hypothesis Testing in Statistical Modeling

Hypothesis testing serves as the backbone of statistical inference. It provides a formal framework for scientists and analysts to make data-driven decisions about a population based on sample data. In the context of regression models, we use hypothesis testing to ascertain whether the relationship observed between the variables is a genuine effect or merely the result of random chance. This structured approach prevents erroneous conclusions and ensures that any claimed relationships are statistically reliable.

Every hypothesis test involves stating two opposing claims: the Null Hypothesis (H_0) and the Alternative Hypothesis (H_1). The H_0 represents the status quo--the assumption of no effect or no difference. The H_1 , conversely, suggests that there is a significant relationship or effect. The goal

of the analysis is to gather enough evidence from the data to potentially reject H_0 in favor of H_1 . If the evidence is insufficient, we simply fail to reject H_0 .

Defining the Null Hypothesis (H_0) for Logistic Models

For Logistic Regression, the Null Hypothesis is precisely defined based on the regression coefficients (β values). These coefficients quantify the expected change in the log odds for a one-unit increase in the corresponding predictor variable. If a coefficient (β_j) is zero, it means that changing the associated predictor variable (X_j) has no effect whatsoever on the outcome's log odds.

The general formulation of the Null Hypothesis for a model containing several predictor variables is that all regression coefficients are simultaneously equal to zero. This comprehensive test assesses the overall usefulness of the model. If H_0 is rejected, it indicates that at least one predictor variable significantly contributes to explaining the variation in the response variable. If we fail to reject H_0 , it suggests the model, as a whole, is no better than simply predicting the outcome based on the overall mean probability.

Simple Logistic Regression: Formulation and Hypotheses

When we focus on the relationship between only one predictor variable and a binary response variable, we employ **simple logistic regression**. This model is characterized by two coefficients: the intercept (β_0) and the slope coefficient (β_1) corresponding to the single predictor. The model structure is designed to estimate the relationship using the log odds scale:

The formula used to estimate the relationship between the variables in simple logistic regression is:

$$\log = \beta_0 + \beta_1 X$$

The left side of the equation represents the log odds of the event occurring ($p(X)$) versus not occurring ($1-p(X)$). The term on the right, $\beta_0 + \beta_1 X$, is the linear predictor. For this setup, we primarily focus on testing the significance of the slope coefficient, β_1 .

The hypotheses for simple logistic regression are structured as follows:

H_0 : $\beta_1 = 0$ (The predictor variable X has no statistically significant impact on the log odds of the outcome.)

H_1 : $\beta_1 \neq 0$ (The predictor variable X has a statistically significant impact on the log odds of the outcome.)

If we find sufficient evidence to reject H_0 , we conclude that there is a statistically significant relationship between the predictor variable, X , and the response variable, Y . This rejection is

typically based on comparing the calculated Z-statistic (or Wald statistic) and its corresponding p-value against a predetermined significance level (alpha, usually 0.05).

Multiple Logistic Regression: Extended Scope and Hypotheses

When the analysis involves two or more **predictor variables** ($x_1, x_2, \dots, x_k$), we use **multiple logistic regression**. This extension allows us to model the combined and individual effects of several factors on the binary outcome, holding other variables constant. The complexity increases as we now have k slope coefficients, one for each predictor.

The formula for estimating the relationship in a multiple logistic regression model is expanded to include all predictors:

$$\log = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$$

Testing hypotheses in multiple logistic regression involves two primary approaches: testing the overall model fit and testing individual coefficient significance. The overall test determines if the group of predictors, taken together, improves the model fit compared to a null model (a model with only the intercept).

The Null and Alternative Hypotheses for the overall model significance are:

H₀: $\beta_1 = \beta_2 = \dots = \beta_k = 0$ (All slope coefficients in the model are simultaneously equal to zero, meaning none of the predictors significantly impact the outcome.)

H_a: At least one $\beta_j \neq 0$ (Not every coefficient is simultaneously equal to zero, indicating that the combined set of predictors significantly improves the model.)

If the overall H₀ is rejected, we proceed to examine the individual coefficients to identify which specific predictors are statistically significant.

Interpreting Model Coefficients and Log Odds

The interpretation of the logistic regression coefficients is crucial for drawing meaningful conclusions from the model. Since the model output is on the log odds scale, the coefficient β_j represents the change in the log odds of the outcome occurring for a one-unit increase in the predictor x_j , assuming all other predictors are held constant (in the case of multiple regression).

While log odds are mathematically convenient for modeling, they are not intuitive for general interpretation. Therefore, analysts often exponentiate the coefficients (e^{β_j}) to obtain the **Odds Ratio**. The Odds Ratio allows for a clearer understanding of the relationship. An Odds Ratio greater than 1 suggests that an increase in the predictor variable increases the odds of the event occurring, while an Odds Ratio less than 1 suggests a decrease in the odds. An Odds Ratio

exactly equal to 1 corresponds to a β of 0, which is the exact condition specified by the Null Hypothesis.

In practical interpretation, when we conduct hypothesis testing on a coefficient, we are essentially testing whether its corresponding Odds Ratio is significantly different from 1. If the p-value associated with a coefficient is small (typically less than 0.05), we reject H_0 , concluding that the Odds Ratio is significantly different from 1, and the variable is a significant predictor.

Practical Application: Testing H_0 in Simple Logistic Regression

To illustrate how to test the Null Hypothesis in practice, let us consider a scenario involving simple logistic regression. Suppose a professor wants to use the number of hours a student studied to predict whether they will pass (1) or fail (0) an exam. Data is collected from 20 students, and a simple logistic regression model is fitted using statistical software like R.

The relevant hypotheses are $H_0: \beta = 0$ (Hours studied does not affect the outcome) and $H_a: \beta \neq 0$ (Hours studied does affect the outcome). The analysis involves calculating the overall goodness-of-fit statistic, often derived from the deviance statistics, and examining the coefficient table provided by the model summary.

We utilize the following R code block to set up the data and fit the 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 using the binomial family

```
model <- glm(result~hours, family='binomial', data=df)
```

#view summary of model fit, which includes coefficient testing

```
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 properly determine if there is a statistically significant relationship, we analyze two key metrics: the individual coefficient test and the overall model test (using the difference between the Null Deviance and Residual Deviance, which follows a Chi-Square distribution). The overall model test is calculated as: $X^2 = (\text{Null deviance} - \text{Residual deviance})$. The corresponding degree of freedom is calculated as $(\text{Null df} - \text{Residual df})$.

From the output above, the overall Chi-Square calculation yields a p-value of **0.2717286**. Since this p-value is substantially greater than the conventional significance level of $\alpha = 0.05$, we **fail to reject the Null Hypothesis** ($H_0: \beta = 0$). This means that based on this dataset and model, we do not have enough evidence to conclude that hours studied has a statistically significant relationship with the exam outcome.

Furthermore, looking at the coefficient summary table, the individual p-value for the 'hours' predictor is 0.293, which confirms the decision to fail to reject H_0 for this specific variable.

Practical Application: Testing H_0 in Multiple Logistic Regression

Next, we explore a scenario utilizing multiple logistic regression. The professor now wishes to use both the number of hours studied and the number of prep exams taken (x_1) to predict the final exam outcome. This scenario requires testing the overall Null Hypothesis ($H_0: \beta_0 = \beta_1 = \beta_2 = 0$) to see if the model as a whole is useful, and then testing the individual coefficients.

We incorporate the second predictor variable, 'exams,' into the R analysis:

```
#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
```

First, we assess the overall model significance by comparing the Null Deviance (26.920) and the Residual Deviance (19.067), which tests $H_0: \beta_0 = \beta_1 = \beta_2 = 0$. The p-value for this overall Chi-Square statistic is calculated to be **0.01971255**.

Since 0.0197 is less than the typical significance level of 0.05, we **reject the Null Hypothesis**. This rejection implies that the combination of hours studied and prep exams taken provides a

statistically significant improvement over the null model, suggesting that the model is useful for prediction.

Next, we look at the individual coefficients to understand their specific contributions. The 'hours' coefficient has a p-value of 0.3538 (fail to reject H_0), indicating that when controlling for the number of prep exams, hours studied is not a significant predictor. However, the 'exams' coefficient has a p-value of 0.0355, which is less than 0.05. We reject H_0 for this individual term, concluding that the number of prep exams taken is a statistically significant predictor of the exam outcome.

Summary of Hypothesis Decision Making

The process of testing the Null Hypothesis in Logistic Regression relies fundamentally on comparing the calculated p-value to the chosen significance level (α). The decision rule is straightforward: if the p-value is less than α (e.g., 0.05), we reject H_0 . If the p-value is greater than α , we fail to reject H_0 .

The decision to reject H_0 means that we have found statistically compelling evidence that the predictor variable(s) significantly affect the log odds of the outcome. In the context of the overall model fit (using the deviance test), rejection confirms that the proposed model offers a significant predictive advantage over a model that assumes no relationship. The ability to correctly interpret and test the Null Hypothesis ensures robust and reliable conclusions drawn from any Logistic Regression analysis.

The following tutorials offer additional information about logistic regression: