What are the 6 assumptions of logistic regression and can you provide examples of each?

Authored by

stats writer

April 20, 2024

Logistic regression is a statistical technique used for predicting the probability of a categorical outcome based on one or more independent variables. In order to use this method effectively, there are six key assumptions that need to be met.

1. Binary Outcome: The dependent variable in logistic regression should be binary in nature, meaning it can take only two distinct values. For example, predicting whether a person is likely to buy a product (yes or no) based on their demographic information.

2. Independence of Observations: The observations should be independent of each other, meaning the value of one observation should not influence the value of another. For example, a study on the effectiveness of a new drug should not include data from the same person multiple times.

3. Linearity of Independent Variables: The relationship between the independent variables and the logit of the dependent variable should be linear. This means that the impact of a change in the independent variable on the logit of the dependent variable should be consistent. For example, the relationship between age and the likelihood of purchasing a product should be consistent across different age groups.

4. Absence of Multicollinearity: The independent variables should not be highly correlated with each other. This can lead to unstable and unreliable estimates of the regression coefficients. For example, using both height and weight as independent variables in predicting a person's fitness level would likely result in multicollinearity.

5. Adequate Sample Size: Logistic regression requires a sufficient number of observations to produce reliable results. As a general rule, there should be at least 10-15 observations for each independent variable included in the model.

6. No Outliers: Outliers are extreme values that can significantly influence the results of the regression analysis. It is important to identify and remove any outliers before conducting a logistic regression. For example, in a study on the factors affecting students' grades, a student who earned a perfect score on every exam may be considered an outlier and should be removed from the analysis.

In summary, the six assumptions of logistic regression are binary outcome, independence of observations, linearity of independent variables, absence of multicollinearity, adequate sample size, and no outliers. It is important to check these assumptions before conducting a logistic regression analysis to ensure accurate and reliable results.

The 6 Assumptions of Logistic Regression (With Examples)

Logistic regression is a method that we can use to fit a regression model when the response variable is binary.

Before fitting a model to a dataset, logistic regression makes the following assumptions:

Assumption #1: The Response Variable is Binary

Logistic regression assumes that the response variable only takes on two possible outcomes. Some examples include:

Yes or NoMale or FemalePass or FailDrafted or Not DraftedMalignant or Benign

How to check this assumption: Simply count how many unique outcomes occur in the response variable. If there are more than two possible outcomes, you will need to perform ordinal regression instead.

Assumption #2: The Observations are Independent

Logistic regression assumes that the observations in the dataset are independent of each other. That is, the

observations should not come from repeated measurements of the same individual or be related to each other in any way.

How to check this assumption: The easiest way to check this assumption is to create a plot of residuals against time (i.e. the order of the observations) and observe whether or not there is a random pattern. If there is *not* a random pattern, then this assumption may be violated.

Assumption #3: There is No Multicollinearity Among Explanatory Variables

Logistic regression assumes that there is no severe multicollinearity among the explanatory variables.

Multicollinearity occurs when two or more explanatory variables are highly correlated to each other, such that they do not provide unique or independent information in the regression model. If the degree of correlation is high enough between variables, it can cause problems when fitting and interpreting the model.

For example, suppose you want to perform logistic regression using max vertical jump as the response variable and the following variables as explanatory

variables:

Player heightPlayer shoe sizeHours spent practicing per day

In this case, height and shoe size are likely to be highly correlated since taller people tend to have larger shoe sizes. This means that multicollinearity is likely to be a problem if we use both of these variables in the regression.

How to check this assumption: The most common way to detect multicollinearity is by using the variance inflation factor (VIF), which measures the correlation and strength of correlation between the predictor variables in a regression model. Check out this tutorial for an in-depth explanation of how to calculate and interpret VIF values.

Assumption #4: There are No Extreme Outliers

How to check this assumption: The most common way to test for extreme outliers and influential observations in a dataset is to calculate Cook's distance for each observation. If there are indeed outliers, you can choose to (1) remove them, (2) replace them with a

value like the mean or median, or (3) simply keep them in the model but make a note about this when reporting the regression results.

Assumption #5: There is a Linear Relationship Between Explanatory Variables and the Logit of the Response Variable

Logistic regression assumes that there exists a linear relationship between each explanatory variable and the logit of the response variable. Recall that the logit is defined as:

Logit(p) = log(p / (1-p)) where p is the probability of a positive outcome.

How to check this assumption: The easiest way to see if this assumption is met is to use a Box-Tidwell test.

Assumption #6: The Sample Size is Sufficiently Large

Logistic regression assumes that the sample size of the dataset if large enough to draw valid conclusions from the fitted logistic regression model.

How to check this assumption: As a rule of thumb, you should have a minimum of 10 cases with the least frequent outcome for each explanatory variable. For

example, if you have 3 explanatory variables and the expected probability of the least frequent outcome is 0.20, then you should have a sample size of at least (10*3) / 0.20 = 150.

Assumptions of Logistic Regression vs. Linear Regression

In contrast to linear regression, logistic regression does not require:

A linear relationship between the explanatory variable(s) and the response variable.The residuals of the model to be normally distributed.The residuals to have constant variance, also known as homoscedasticity.

The Four Assumptions of Linear Regression

4 Examples of Using Logistic Regression in Real Life

How to Perform Logistic Regression in SPSS

How to Perform Logistic Regression in Excel

How to Perform Logistic Regression in Stata