Logistic

Shubham Kotal

August 2024

Logistic regression is a statistical method used to model the relationship between a binary dependent variable (let's call it Y) and one or more independent variables (let's call it X). The goal is to estimate the probability that Y equals 1 (e.g., "Yes" or "Default") given the value of X.

Key Concepts

1. Log Odds (Logit): The logistic regression model relates the probability of $Y = 1$ (denoted as $p(X)$) to X through the log odds, also known as the logit. The logit is given by:

$$logit(p(X)) = \log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X$$

This equation shows that the log odds of Y being 1 are a linear function of X.

2. Interpretation of Coefficients: - β_1 represents the change in the log odds of $Y = 1$ for a one-unit increase in X. - If $\beta_1 > 0$, an increase in X increases the probability $p(X)$. - If $\beta_1 < 0$, an increase in X decreases the probability $p(X)$.

3. Non-linear Relationship: Although the log odds are linear in X, the probability $p(X)$ itself is non-linear. This means that the effect of X on $p(X)$ changes depending on the current value of X.

Hypothesis Testing in Logistic Regression

In logistic regression, we often want to test whether there is a significant relationship between X and Y. We do this by testing the null hypothesis H_0, which states:

$$H_0 : \beta_1 = 0$$

This hypothesis suggests that X has no effect on the probability $p(X)$. If H_0 is true, the probability of $Y = 1$ does not depend on X, and the logistic regression model simplifies to a constant probability.

To test H_0, we calculate the z-statistic:

$$z = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$$

- $\hat{\beta}_1$ is the estimated coefficient, and $SE(\hat{\beta}_1)$ is its standard error. - A large absolute value of z provides evidence against H_0. - The p-value associated with z indicates the probability of observing such a z-value if H_0 were true. A small p-value (typically less than 0.05) leads us to reject H_0, concluding that X is significantly associated with Y.

Example with Default Data

Suppose we're predicting whether a customer will default on a loan based on their account balance. From a logistic regression model, we get:

$$\hat{\beta}_1 = 0.0055$$

This means that for every one-unit increase in balance, the log odds of default increase by 0.0055. In practical terms, a higher balance slightly increases the probability of default.

If the z-statistic for $\hat{\beta}_1$ is large and the p-value is small, we reject the null hypothesis $H_0 : \beta_1 = 0$. This rejection suggests a significant relationship between balance and the probability of default.

Simplified Formula to Compute Probabilities

To find the actual probability $p(X)$, we use:

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

Here, $p(X)$ gives us the probability that $Y = 1$ for a given X. The relationship between X and $p(X)$ is S-shaped, reflecting the non-linear influence of X on $p(X)$.

This is a simplified overview of logistic regression, hypothesis testing, and how changes in X influence the probability of Y.

]Introduction Summary of Logistic Regression

Logistic regression is a statistical method used to model the relationship between a binary dependent variable (let's call it Y) and one or more independent variables (let's call it X). The goal is to estimate the probability that Y equals 1 (e.g., "Yes" or "Default") given the value of X.

Key Concepts

1. Log Odds (Logit): The logistic regression model relates the probability of $Y = 1$ (denoted as $p(X)$) to X through the log odds, also known as the logit. The logit is given by:

$$logit(p(X)) = \log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X$$

This equation shows that the log odds of Y being 1 are a linear function of X.

2. Interpretation of Coefficients: - β_1 represents the change in the log odds of $Y = 1$ for a one-unit increase in X. - If $\beta_1 > 0$, an increase in X increases the probability $p(X)$. - If $\beta_1 < 0$, an increase in X decreases the probability $p(X)$.

3. Non-linear Relationship: Although the log odds are linear in X, the probability $p(X)$ itself is non-linear. This means that the effect of X on $p(X)$ changes depending on the current value of X.

Hypothesis Testing in Logistic Regression

In logistic regression, we often want to test whether there is a significant relationship between X and Y. We do this by testing the null hypothesis H_0, which states:

$$H_0 : \beta_1 = 0$$

This hypothesis suggests that X has no effect on the probability $p(X)$. If H_0 is true, the probability of $Y = 1$ does not depend on X, and the logistic regression model simplifies to a constant probability.

To test H_0, we calculate the z-statistic:

$$z = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$$

- $\hat{\beta}_1$ is the estimated coefficient, and $SE(\hat{\beta}_1)$ is its standard error. - A large absolute value of z provides evidence against H_0. - The p-value associated with z indicates the probability of observing such a z-value if H_0 were true. A small p-value (typically less than 0.05) leads us to reject H_0, concluding that X is significantly associated with Y.

Example with Default Data

Suppose we're predicting whether a customer will default on a loan based on their account balance. From a logistic regression model, we get:

$$\hat{\beta}_1 = 0.0055$$

This means that for every one-unit increase in balance, the log odds of default increase by 0.0055. In practical terms, a higher balance slightly increases the probability of default.

If the z-statistic for $\hat{\beta}_1$ is large and the p-value is small, we reject the null hypothesis $H_0 : \beta_1 = 0$. This rejection suggests a significant relationship between balance and the probability of default.

Simplified Formula to Compute Probabilities

To find the actual probability $p(X)$, we use:

$$p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

Here, $p(X)$ gives us the probability that $Y = 1$ for a given X. The relationship between X and $p(X)$ is S-shaped, reflecting the non-linear influence of X on $p(X)$.

This is a simplified overview of logistic regression, hypothesis testing, and how changes in X influence the probability of Y.