Print

Modeling a Binary Response

A linear regression model assumes the data is continuous, but for logistic regression, the response is binary. So, we can't simply use regression model assumptions, fitting techniques, and procedures we've learned thus far.

Let's do a quick comparison of linear regression and logistic regression. Linear regression uses a predictor variable (X) to estimate the conditional mean of the response (Y), which is continuous. With linear regression, you assume that the expected value of the response has a linear relationship with the predictor variable. The conditional mean of the response has the linear form $\beta_0 + \beta_1 X$, and it ranges from negative infinity to positive infinity. For binary data, the mean of the response is the probability of a success. Suppose you're working with a binary response variable that has the values Yes and No. You can code the values numerically: Yes=1 and No=0. But these values are still categories and the coding is arbitrary. If the response variable has only two levels, you can't assume the constant variance and normality that are required for linear regression.

In logistic regression, we want to model the probability of both levels of the response. Although probabilities are continuous, they are bounded. So, if we were to use linear regression to estimate the probability of each level, we could predict values outside the probability range of 0 to 1. Also, the relationship between the probability of the outcome and a predictor variable is usually nonlinear rather than linear. In fact, the relationship often resembles an S-shaped, or sigmoidal, curve. Furthermore, probabilities do not have a random normal error, but rather a binomial error of p * (1 - p). The error is greatest at probabilities close to 0.5 and lowest near 0 and 1. Finally, there's no such thing as an observed probability, and therefore, least squares methods cannot be used.

A logistic regression model applies a logit transformation, or simply the log odds transformation, to the probabilities.

$$\log\left(\frac{p}{1-p}\right)$$

Here, log means natural log, rather than log with a base 10. The logit effectively avoids the boundary problem for probabilities. As p approaches its minimum value of 0, the logit approaches negative infinity, and as p approaches its maximum value of 1, the logit approaches positive infinity. So, the logit is unbounded, just like in linear regression, but the probabilities maintain the original bounds of 0 to 1.

In addition, the logit transformation enables us to move from modeling the probability with a sigmoidal nonlinear curve, to modeling the logit with a linear function of the predictors. And modeling the logit allows you to indirectly model the probability of the response. Whatever the predicted value of the logit is, we can simply back-transform to the probability scale to get a value between 0 and 1.

The logistic model now looks quite familiar, compared to linear regression.

$$\text{logit}\left(p_i\right) = \beta_0 + \beta_1 X_i$$

The logit is equal to a linear function of your parameters and predictors. Again, β_0 is the intercept of the regression equation, and the other betas are the slopes of the model predictors. And there is no error here because one, we no longer assume normally distributed errors, and two, the error is a function of p, the probability of the event.

To fit the model, logistic regression requires a more computationally complex estimation method, named the method of maximum likelihood, to estimate the parameters. This method finds the values of the parameters that make the observed data most likely. This is accomplished by maximizing the likelihood function that expresses the probability of the observed data as a function of the unknown parameters.

Statistics 1: Introduction to ANOVA, Regression, and Logistic Regression

Close