The Simple Logistic Model

Let's consider a scenario with a binary response with two possible response categories, Pass and Fail. We are interested in the probability that the response category, or outcome, is Fail.

How should we model the relationship between this probability and the potential predictors? We might consider using a linear regression model. Let's say we have k predictors. The linear model would take this form.

However, the predictions from a linear model are not constrained to fall between 0 and 1, and we know the probabilities can only fall within this interval. Even if the relationship is nearly linear over the range of predictor values, the predictions can fall outside the (0,1) interval. So, the probabilities cannot be directly modeled as a linear function of the predictors.

Rather than model the probability that Y = Fail, we model a function of this probability, called the logit. Let me explain this. Consider an event with the probability of occurrence, p. For a binary response, the probability of the event not occurring is 1-p. We calculate the odds for the event occurring as the ratio of these two probabilities.

For example, consider rolling a fair die. There are six possible outcomes. The odds of obtaining a 3 or higher when rolling a fair die are 4 to 2 or 2 to 1. This quantity, the odds, is the foundation of logistic regression.

It turns out that we can model the log of the odds as a linear function of the predictors, with the predicted values constrained to fall between zero and one. This is our logistic model. It relates the log of the odds, or simply the log odds, to a linear model in the predictors.

Here, ln denotes the natural logarithm, and the predictors can be continuous or categorical. This is called the log-odds of p, or the logit. The logistic model predicts the log odds of an event as a linear function of the predictors. But we can reorganize this formula to calculate the predicted probability of the event.

Let's take the case of one binary response, Y, and one continuous predictor, X. The probability of a particular event, or outcome, is shown here. In this equation, $\beta0$ is the intercept, and $\beta1$ is the slope. We can graph the logistic model to visualize the probabilities for different values of the intercept and the slope.

In this graph, the predicted probability of the event is plotted on the Y axis and the value of the predictor, X, is plotted on the X axis. The s-shaped or sigmoidal curve in the graph, which we call the logistic curve, shows how the predicted probability changes as we increase the value of the predictor.

In this example, the intercept is zero, and the slope is 0.3. This logistic curve shows that an increase in the value of the predictor, X, results in an increase in the probability of the event. This curve provides a graphical measure of the strength of the relationship between the predictor and the categorical response variable. The stronger the relationship, the steeper the logistic curve.

Here, we see the logistic curve when the intercept is zero and the slope is increased from 0.3 to one. Notice that the curve is much steeper, so the probability changes much more quickly as the value of the predictor increases.

In this example, the probability is 0.5 when X is zero. Changing the intercept from zero to two, with the slope of one, simply shifts the logistic curve along the X axis. Now, the probability is 0.5 when X is -2.0.

Here, we see the logistic curve with an intercept of zero and a slope of -0.5. With a negative slope, as the value of the predictor increases, the probability of the event decreases.

When both the intercept and the slope are zero, the logistic curve is a horizontal line. The predicted probability is 0.5 for all values of the predictor.

Statistical Thinking for Industrial Problem Solving

Close