Threshold functions for logistic regression

Odds, mathematically speaking.

We are going to take the notion of odds, put a simple mathematical framework around it and then use our previous knowledge of linear regression to create a model that predicts binary outcomes.

We need two bits of prior knowledge to understand this.

1. Exponentials and Natural Logarithms [1]
2. A very elementary understanding of Probability. [2]

Basically all we need to know is that Probability is a number between 0 and 1 and indicates the likelihood of an event occurring. We remind ourselves that: probability = 0 is as good as the event being impossible and probability = 1 is as good as it being certain.

We should also remind ourselves that if the probability of an event happening is p then the probability of it not happening is 1-p

That's all the probability we need.

In these exercises we will have math gradually added and most of it will be at the simplest possible level needed to work with these techniques. We have a philosophy of "Just Enough Math" so if you see a math symbol in a lesson it's because we really need it.

Having said that let's start talking about odds.

Odds and Odds Ratio

When bettors say the odds of winning are 1:4 what is this in terms of probability?

It means 1 part chance of winning to 4 parts chance of losing. Note that total # of parts = 5 and odds of winning is 1 out of 5. So p is 1/5 = 0.2 and 1-p is 0.8. Here p is small and 1-p is large.

The odds might be 1:1 which means p = 1/2 and 1-p = 1/2 i.e. equal chances of an even happening or not = "even odds".

The odds might be 3:2 which means p = 0.6 and 1-p = 0.4. Here p is greater and 1-p is smaller. So depending on the ratio of p to 1-p we have more or less confidence in a bet winning.

This suggests we might want to look at:

$$OddsRatio(OR) = \frac{p}{1-p}$$

If OddsRatio is high say:

$$OR > 4$$

then the event might be considered very likely and if:

$$OR < 0.25$$

then very unlikely.

The Logit Function

Mathematicians like to work with a function derived from this called the Logit function. It's the Log of the OddsRatio

$$logit(p) = log(\frac{p}{1-p})$$

or the LogOdds function.

And here is where we wave a magic wand again and bring in our linear regression formula. Let's say we want to plot a set of events on the x-axis and the logit value for that event on the y-axis and we want to fit a linear model to this.

So we would say:

$$logit(p) = log(\frac{p}{1-p}) = b_0 + b_1X$$

where X is the "value" of the event. So here instead of $Y = b_0 + b_1X$ we want to plot $logit(p)$ on the Y axis and the event or the score on the X axis.

So this is how the linear model slips in - we want to express log odds as a linear function of score. Patience now, we are just one step away.

Finally, the Logistic Function

We don't really like plotting LogOdds as it involves awkward intermediate steps of calculating the logit() value. So we do the hard work once and for all and solve for p as a function of X and when we do that we get

$$p(X) = \frac{1}{1 + e^{b_0 + b_1X}}$$

which, voila, is what we pulled out of the hat - but now we know why we pulled that particular one and why we didn't use any other.

References

[1] Exponential and Logarithmic functions https://www.khanacademy.org/math/trigonometry/exponential_and_logarithmic_func

[2] Probability https://www.khanacademy.org/math/probability/independent-dependent-probability