Logistic Regression

Ciro S. Costa

06 Out, 2015

Contents

Intro

Logistic Regression **(logit model)** "is a regression model where the **dependent variable is categorical** and **binary**. Cases with **more than two** categories are called **multinomial logistic regression**" (Wikipedia)

It treats the dependent variable as the oucome of a Bernoulli trial rather than a continuous outcome (as it is done in the case of the traditional linear regression). By doing so, assumptions of traditional linear regression are violated (residuals won't be normally distributes, e.g.). Traditional linear regression also doesn't fit here as we predict probabilities - with lin reg we could get values below 0 or above 1, which wouldn't make sense for our model. Because of that we need to apply a transformation to the binary variable to transform it into a continuous (real, ranging from `-inf` to `+inf`) one so that we can use the tools we have to estimate the dependent variable. The mentioned transformation is the `logit` (or `log-odds`).

> "although the dependent variable in logistic regression is binomial, the logit is the continuous criterion upon which linear regression is conducted"

Probability and Odds

$$Pr(something) = \frac{outcomesOFinterest}{allPOSSIBLEoutcomes}$$

Knowing that $odds = \frac{F(x+1)}{1-F(x+1)}$, i.e, odds equals to the probability of ocurring divided by the probability of not occurring, the odds ratio corresponds to:

$$OR = \frac{odds(x+1)}{odds(x)}$$

The **odds ratio** in logistic regression represents **how the odds change with a 1 unit increase in that variable holding all other variables constants**. The interesting thing of odds ration is that it lets us know the increase in the odds of something happening independently of where in the variable spectrum we are.

> "The odds can have a large magnitude change even if the underlying probabilities are low"

Linking Bernoulli to Linear Combination

The dependent variable in the logistic regression follows the bernoulli distribution having and unknonw probability `p`. As said before, in logistic regression we're estimating an unknown `p` for any given linear combination of the independent variables. Now, to link together independent and dependent variables we use the `logit` (natural log of the odds-ratio), which maps the linear combination of variables to the domain $[0, 1]$.

$$logit(p) = ln(\frac{p}{1-p}), p \in [0, 1]$$

As $logit(p)$ gives us a sigmoid over the y axis (while we want over the x axis) we have to simply take its inverse:

$$logit^{-1}(\alpha) = \frac{e^{\alpha}}{1+e^{\alpha}}, \alpha \in \mathrm{R}$$

As α can be any number, there we can put our linear combination of variables and their corresponding coefficients estimated. The inverse-logit will then map that to the probability of being a 1 or a 0.