Logistic Regression

Giovanni Cassani

July 22, 2019

1 Introduction

Logistic regression is a specific case of Generalized Linear Models (GLMs) where the dependent variable being modeled is binary, i.e. it is distributed according to a Bernoulli distribution with unknown parameter π governing the shape of the distribution. As every other GLM, the formula of a logistic regression can be instantiated in equation (1):

$$Y \sim \beta_0 + \beta_1 x_1 = \beta_2 x_2 + ... + \beta_k x_k \tag{1}$$

where $Y \sim Bernoulli(\pi)$. The formula in (1) consists of different components:

Random component : it is the distribution of the dependent variable Y, Bernoulli distribution in the case of logistic regression

Systematic component : the linear predictor X_k, i.e. the set of k independent variables whose effect on Y we want to assess (these variables can be both continuous and categorical)

Link function : a function *g()* that relates the expected value of Y, or the mean of Y, to the linear predictor. In the case of logistic regression, the link function is called **logit** and is specified by the function in (2)

$$g(Y) = logit(\pi(X_k)) = log\left(\frac{\pi(X_k)}{1-\pi(X_k)}\right) \tag{2}$$

From (2), it follows that

$$log\left(\frac{\pi(X_k)}{1-\pi(X_k)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_3 x_3 + \epsilon \tag{3}$$

The logit function has an inverse, the **expit** function. We can thus formulate the equation in (3) as follows:

$$\begin{aligned} Y &= \pi(X_k) \\ &= \mathbf{expit}(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \epsilon) \\ &= \frac{1}{1 + exp(-\beta_0 - \beta_1 x_1 - \beta_2 x_2 - ... - \beta_k x_k - \epsilon)} \end{aligned} \tag{4}$$

In (4), $\pi(X_k)$ is our target estimate, i.e. the probability of success, or the probability of a response variable of value 1 across all our response variables.

With binary outcomes, the expected value of Y conditional on the data is the probability that Y = 1 (or the risk): $E(Y|X) = \pi(X)$. This is the conditional mean of the Bernoulli distribution of our dependent variable Y. The variance of this distribution is $Var(Y|X) = \pi(X)(1 - pi(X))$.

When we model binary data, we inevitably violate some crucial assumptions of linear models and using such models would result in biased estimates of the effect of our linear predictor on the variable of interest. These violations include:

Bounded values : the expected value of a binary outcome variable is bounded between 0 and 1, while the expected value of a continuous outcome variable can take any value between $-\infty$ and $+\infty$

Non-linear increase/decrease : the cumulative density function of a Bernoulli distribution is S-shaped, while the identity function used as a link function in linear models changes linearly with the outcome variable

Non-normal errors : since the interval of our outcome variable is bounded, at the edge of the distribution, we cannot overestimate or underestimate the probability of success. We cannot make an estimate higher than 1 or lower than 0, which makes the distribution of errors non-normal

Non-constant variance : the variability at the edge values of a binary outcome is much lower than the variability at central values, which violates the assumption of linear models that errors between observed and predicted values of Y must be normally distributed around the predicted values with constant variance

The random component of the model, i.e. the distribution of the outcome variable takes care of the bounded values and the non-linear increase and decrease. As far as the distribution and variance of the error terms are concerned, logistic regression ties the error to the value of X and assumes that errors are Bernoulli-distributed with conditional mean $E(\epsilon|X) = 0$ and variance $Var(\epsilon|X) = \pi(X)(1 - \pi(X))$. If we input the extreme values 0 and 1 in the variance equation we see that this function entails that no variance exists at the edges of the distribution (which makes sense, as we saw before).

2 Risk and Odds

From (2) and (3), we see that our logistic model predicts the probability that our binary dependent variable is 1, i.e. the probability of success, using the linear predictor we specified. In other words, we are estimating the **log odds of Y=1**. Think of gambling: chances of winning (a success, Y=1) are usually expressed in the form of 'the odds are 5 to 1', where the first number indicates the expected wins for every loss (the second number is always one): this means that for every time you lose you are expected to win 5 times, meaning that $\pi(Y = 1) = wins/(losses + win) = 5/6$. When we express chances as odds it is straightforward to see that game B where the odds are 10 to 1 gives double the chances of winning than game A where the odds are 5 to 1. If you look carefully, though, you see that $2 \cdot \pi(Y_A = 1) \neq \pi(Y_B = 1)$, since the first is 10/11 and the second is 5/6. This makes sense because probabilities are bounded, and we cannot keep doubling: if we were to double the probability of success in game A, $\pi(Y_A = 1) = 0.83$, we would go over 1, which doesn't make sense. With odds, however, we can keep doubling and express the fact that game B gives twice the chance of winning than game B. When, in a game, we are expected to lose more often than we win, we express odds with numbers between 0 and 1, as in 'the odds are 0.5 to 1': it is more difficult to understand, but it means that for every win, we are expected to lose twice, or that for every two losses, we are expected to win once (be careful, by taking the inverse and considering the odds of losing, the above reasoning flips entirely). Odds can be converted to probabilities (also known as **risk**): the relation between odds and risk is detailed in equations (5) and (6):

$$odds = \frac{risk}{1 - risk} \tag{5}$$

$$risk = \frac{odds}{1 + odds} \tag{6}$$

Let's consider some extreme cases to better explain the meaning of this. If we play a game where the odds are 1 to 1, we have a risk of $1/(1+1) = 0.5$, meaning that half of the time we win and half of the time we lose: 0.5 is the 'risk' of winning or the risk of losing depending on how we express the odds: if we express them as 'wins to losses', the risk is $\pi(Y = 1)$; on the contrary, if we express the odds as 'losses to wins', then the risk is $1 - \pi(Y = 1)$.

3 Interpreting coefficients

We know that (3) estimates the log-odds of Y = 1. So what do $\beta_0, \beta_1, ..., \beta_k$ represent? As in every GLM, β_0 is the average value of Y when all predictors are zeroed: in the case of logistic regression, it is the value of the logarithm of the odds of Y=1 when all xs are 0, as specified in equation (7):

$$\beta_0 = logit(\pi(X_k = 0)) = log\left(\frac{\pi(X_k = 0)}{1 - \pi(X_k = 0)}\right) = log(odds(x = 0) \tag{7}$$

From (7), it follows that $e^{\beta_0} = odds(x = 0)$.

What do other βs tell then? Suppose x_1 is a continuous variable: then β_1 is the effect that a 1-unit increase in x_1 has on the odds of Y = 1, as specified in equation (8):

$$\begin{aligned}
\beta_1 &= (\beta_0 + \beta_1(x_1 + 1)) - (\beta_0 + \beta_1 x_1) \\
&= logit(\pi(x + 1)) - logit(\pi(x)) \\
&= log\left(\frac{\pi(x + 1)}{1 - \pi(x + 1)}\right) - log\left(\frac{\pi(x)}{1 - \pi(x)}\right) \\
&= log\left(\frac{\pi(x + 1)}{1 - \pi(x + 1)} \Big/ \frac{\pi(x)}{1 - \pi(x)}\right) \\
&= log\left(\frac{odds(x + 1)}{odds(x)}\right)
\end{aligned} \tag{8}$$

β_1 is thus defined as the log(**Odds Ratio (OR)**) for a 1-unit increase in the continuous independent variable x_1. If we exponentiate β_1, we obtain the true Odds Ratio. In a logistic model, all βs estimated for the independent variables have a multiplicative effect of e^{β_1} on the odds of Y=1, not on the probability (as we saw before). When interpreting logistic regression coefficients, we must think of odds as expressed in the form 'wins to losses', which entails that the risk is to be intended as $\pi(Y = 1)$: thus, a positive β indicates an increase in the risk of a success, i.e. Y = 1.

3.1 Continuous predictors

Let's consider a simple study where height is used to predict whether somebody is or not a basketball player. The model is straightforward and is specified in equation (9):

$$logit(\pi(playsBasketball = 1) = \beta_0 + \beta_1 height \tag{9}$$

Let's assume we have centered height around the sample mean, so that height = 0 doesn't mean that somebody is actually 0 cm tall, but that it as tall as the average person in the sample. Let's assume we collected a sample and estimated the model: β_0 is the log odds that somebody of average height is a basketball player. Let's assume $\beta_0 = -5$: we know that we can get the odds of somebody being a basketball player when x = 0 by exponentiating β_0, which gives 0.0067. We can then compute the risk, or $\pi(playsBasketball|x = 0)$, using the formula in (6), which gives again 0.0067. Considering more decimals, the numbers actually differ; however, it shouldn't be a surprise that the two numbers are very close. When the odds are very small, they don't affect the denominator in (6) which stays close to 1. And dividing a number by 1 doesn't change that number.

Now that we know the probability that somebody of average height is a basketball player we can look at the effect of height: when somebody is taller, is it more or less likely that he/she is a professional basketball player? Let's assume our $\beta_1 = 0.25$: it's positive, meaning that the probability of playing basketball increase with height. But how much? We already know $p(playsBasketball|x = 0) = 0.0067$. We know need to compute the same probability when x = 1: we can use the expit function from (4), plugging in the right numbers:

$$p(playsBasketball|x = 1) = \frac{1}{1 + exp(-(-5) - 0.25)} = 0.0086 \tag{10}$$

0.0086 is the risk of being a basketball player when being one centimeter taller than the average. From this, we derive the odds using (5), which gives 0.0087. We can now compute the Odds Ratio for a 1-unit increase in height, 0.0087 / 0.0067 = 1.284. This means that the odds of somebody being a basketball player increase by 28% with every 1-unit increase in height. The increase in probability (or risk) is not constant, as we have explained at the beginning of this section comparing risk and odds.

3.2 Categorical predictors

When our x is categorical, a different β is estimated for each level of x, and each β gives the increase/decrease in the odds of Y = 1 as compared to the reference category, given by the estimate for β_0 - categorical predictors are coded with dummy variables, and it takes one dummy variable less than the number of levels of the predictor (if x has 4 levels, 3 dummy variables and corresponding βs are needed, since one of the level is estimated by setting all βs to 0). Consider an experiment where student's native language is used as predictor to determine the probability of success at the TOEFL exam: our predictor consists of 4 levels (Italian, French, Spanish, German). The model is outlined in equation (11):

$$\begin{aligned} logit(\pi(Pass)) &= \beta_0 + \beta x_{Lang} + \epsilon \\ &= \beta_0 + \beta_1 x_I + \beta_2 x_F + \beta_3 x_S + \epsilon \end{aligned} \tag{11}$$

From (9) we can estimate our outcome variable for all four languages, even though there is no β for German: when the student is Italian, x_I will equal 1, and all other xs will be 0; with French students, x_F will be 1 and all other xs will be 0; lastly, with Spanish students, x_S will be 1, and all other xs will be 0. For German students, it's enough to set all xs to 0. Thus, the estimated probability of success for German students will equal β_0. To know the probability of success for Italian students, we will simply sum β_0 and β_1. Thus, β_1 is the odds ratio between the success rate of Italian and German students: if β_1 is negative, Italians do worse than Germans.

Let's assume again we fitted the model, getting the following estimates for the βs in the model, as shown in (12):

$$logit(\pi(Pass)) = 0.45 + 0.001x_I + 0.7x_F - 0.56x_S \tag{12}$$

Simply looking at the estimates we see that Italians do slightly better than Germans, French students do best and Spanish students do worse. But how likely is each group to pass the TOEFL? The way to estimate β_0 is the same as for the previous model: we exponentiate the estimate to get the odds, 1.568, and derive the risk from it, 0.611. This is the probability that a German student passes the TOEFL exam in our toy example. To derive the odds, risk, and percent increase/decrease from one group to the other, we plug the right values in (4), (5), and (8).

$$Risk \begin{cases} (x_I = 1) = \frac{1}{1+exp(-0.45-0.001)} = 0.611 \\ (x_F = 1) = \frac{1}{1+exp(-0.45-0.7)} = 0.76 \\ (x_S = 1) = \frac{1}{1+exp(-0.45-(-0.56))} = 0.473 \end{cases} \tag{13}$$

$$Odds \begin{cases} (x_I = 1) = \frac{risk_I}{1-Risk_I} = 1.57 \\ (x_F = 1) = \frac{risk_F}{1-Risk_F} = 3.158 \\ (x_S = 1) = \frac{risk_S}{1-Risk_S} = 0.896 \end{cases} \tag{14}$$

Unlike the case of a continuous predictor where it is clear what goes on the numerator and what goes on the denominator, with categorical predictors we can put levels where we want: however, we must be careful in interpreting the OR.

$$OR\begin{cases}(I,G) = 1.57/1.568 = 1.001\\(I,F) = 1.57/3.158 = 0.497\\(I,S) = 1.57/0.896 = 1.752\\(G,F) = 1.568/3.158 = 0.496\\(G,S) = 1.568/0.896 = 1.75\\(F,S) = 3.158/0.896 = 3.525\end{cases}\tag{15}$$

From these ORs, we see that there is a 0.1% increase in the probability of passing the exam for Italians wrt Germans; a 75.2% increase for Italians wrt Spanish; a 75% increase for Germans wrt Spanish; a 252% increase for French wrt Spanish. To interpret OR lower than 1, it is more convenient to take the inverse and flip the order: instead of having the odds ratio for Italians wrt French, or Germans wrt French, we take the OR for French wrt Italians and Germans, which amounts to $1/OR(I,F) = 2.012$ and $1/OR(G,F) = 2.014$, i.e. a 101.2% increase for French wrt Italians, and a 101.4% increase for French wrt Germans.

4 Confidence Intervals

For every estimate, it is possible *and advisable* to compute a **confidence interval**, which can be easily done with the formula in equation (16):

$$[\hat{\beta_k} - z_\alpha \cdot \hat{se}(\hat{\beta_k}); \hat{\beta_k} + z_\alpha \cdot \hat{se}(\hat{\beta_k})]\tag{16}$$

where z_α is the critical value on the z-distribution for the chosen significance level α, and $\hat{\beta_k}$ and $\hat{se}(\hat{\beta_k})$ are the estimated coefficient and standard error for the kth predictor.

However, we know that the variance of an estimate is not constant in a logistic regression: the variance depends on the estimated probability, thus we need to compute it at different x values. In order to compute the Confidence Interval for the log(odds) and the risk, we need to compute the estimated variance in the outcome variable for the desired value of x. In our example, we need to know how the probability of being a basketball player varies for specific height values different from the average. We show how to compute it in equation (17).

$$\hat{Var}(\hat{\beta_0} + \hat{\beta_1}x) = \hat{Var}(\hat{\beta_0}) + x^2 \cdot \hat{Var}(\hat{\beta_1}) + 2x \cdot \hat{Cov}(\hat{\beta_0}, \hat{\beta_1})\tag{17}$$

Every software that allows to fit GLMs outputs variance and co-variance estimates together with estimates for βs. Let's assume $\hat{Var}(\hat{\beta_0}) = 0.64$, $\hat{Var}(\hat{\beta_1}) = 0.48$, and $\hat{Cov}(\hat{\beta_0}, \hat{\beta_1}) = -0.021$. Let's assume we want to know the variance at x = 15, i.e. for people that are 15 centimeters taller than average. We plug in the right values in equation (17), obtaining the following calculation:

$$\hat{Var}(\hat{\beta_0} + \hat{\beta_1}15) = 0.064 + 15^2 \cdot 0.048 + 2 \cdot 15 \cdot (-0.051) = 9.334\tag{18}$$

We also compute the estimate for $logit(\pi(x = 15)) = 0.223$, and the 95% Confidence Interval for it (19) - be careful, in (17) we computed the variance, so we need to take its square root to get the standard error, i.e. the standard deviation of the estimate. Finally, we can compute the estimated probability that somebody that is 15 centimeters taller than average is a basketball player (20).

$$[0.223 - 1.96 \cdot \sqrt{9.334}; 0.223 + 1.96 \cdot \sqrt{9.334}] = [-5.765; 6.211] \tag{19}$$

$$[expit(-5.765); expit(6.211)] = [0.003; 0.998] \tag{20}$$

Turns out the numbers I made up for the variances were really large, and the probability that somebody who's 15 cm taller than average is a basketball player lies somewhere between 0.003 and 0.998. But you get the idea.

5 Wrap-up

We have seen that logistic regression is an effective tool to model binary dependent variables using a set of continuous and categorical independent variables: the model assumes the dependent variable is distributed according to a Bernoulli distribution with probability π, uses the linear predictor, i.e. the set of independent variables, to estimate the expected value of the dependent variable, and finally uses the logit link function to transform the expected value into the estimate for π. We have seen that this model is capable of handling the fact that π can only fall in a bounded range, that the increase of π is not linear, that the residual errors are not normally distributed around the predicted values, and that the variance of the observed value is a function of the linear predictor.

Moreover, we have discussed the concepts of odds and risk and related them to the interpretation of a logistic regression. We have discussed how to interpret regression coefficients estimated for a continuous as well as for a categorical predictor, showing how risk, odds, and Odds Ratio are related and should be consistently used to interpret a model. Finally, we have discussed how to compute Confidence Intervals, stressing the fact that, as variance depends on the linear predictor, CIs need to be computed at different values of X.