

POINTS OF SIGNIFICANCE

Logistic regression

Regression can be used on categorical responses to estimate probabilities and to classify.

In recent columns we showed how linear regression can be used to predict a continuous dependent variable given other independent variables^{1,2}. When the dependent variable is categorical, a common approach is to use logistic regression, a method that takes its name from the type of curve it uses to fit data. Categorical variables are commonly used in biomedical data to encode a set of discrete states, such as whether a drug was administered or whether a patient has survived. Categorical variables may have more than two values, which may have an implicit order, such as whether a patient never, occasionally or frequently smokes. In addition to predicting the value of a variable (e.g., a patient will survive), logistic regression can also predict the associated probability (e.g., the patient has a 75% chance of survival).

There are many reasons to assess the probability of a state of a categorical variable, and a common application is classification—predicting the class of a new data point. Many methods are available, but regression has the advantage of being relatively simple to perform and interpret. First a training set is used to develop a prediction equation, and then the predicted membership probability is thresholded to predict the class membership for new observations, with the point classified to the most probable class. If the costs of misclassification differ between the two classes, alternative thresholds may be chosen to minimize misclassification costs estimated from the training sample (Fig. 1). For example, in the diagnosis of a deadly but readily treated disease, it is less costly to falsely assign a patient to the treatment group than to the no-treatment group.

In our example of simple linear regression¹, we saw how one continuous variable (weight) could be predicted on the basis of another continuous variable (height). To illustrate classification, here we extend that example to use height to predict the probability that an individual plays professional basketball. Let us assume that professional basketball players have a mean height of 200 cm and that those who do not play professionally have a mean height of 170 cm, with both populations being normal and having an s.d. of 15 cm. First, we create a training data set by randomly sampling the

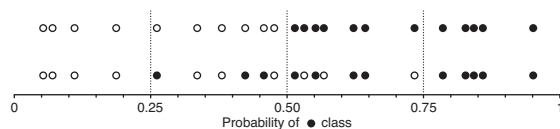


Figure 1 | Classification of data requires thresholding, which defines probability intervals for each class. Shown are observations of a categorical variable positioned using the predicted probability of being in one of two classes, encoded by open and solid circles, respectively. Top row: when class membership is perfectly separable, a threshold (e.g., 0.5) can be chosen to make classification perfectly accurate. Bottom row: when separation between classes is ambiguous, as shown here with the same predictor values as for the row above, perfect classification accuracy with a single-value threshold is not possible. The threshold is tuned to control false positives (e.g., 0.75) or false negatives (e.g., 0.25).

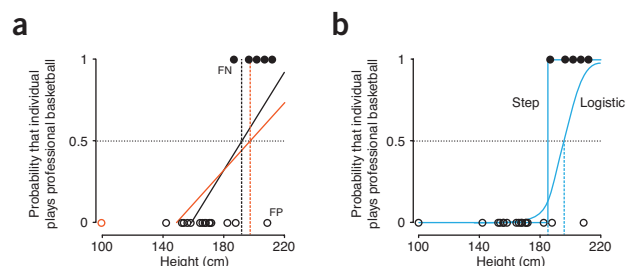


Figure 2 | Robustness of classification to outliers depends on the type of regression used to establish thresholds. (a) The effect of outliers on classification based on linear regression. The plot shows classification using linear regression fit (solid black line) to the training set of those who play professional basketball (solid circles; classification of 1) and those who do not (open circles; classification of 0). When a probability cutoff of 0.5 is used (horizontal dotted line), the fit yields a threshold of 192 cm (dashed black line) as well as one false negative (FN) and one false positive (FP). Including the outlier at $H = 100$ cm (orange circle) in the fit (solid orange line) increases the threshold to 197 cm (dashed orange line). (b) The effect of outliers on classification based on step and logistic regression. Regression using step and logistic models yields thresholds of 185 cm (solid vertical blue line) and 194 cm (dashed blue line), respectively. The outlier from (a) does not substantially affect either fit.

heights of 5 individuals who play professional basketball and 15 who do not (Fig. 2a). We then assign categorical classifications of 1 (plays professional basketball) and 0 (does not play professional basketball). For simplicity, our example is limited to two classes, but more are possible.

Let us first approach this classification using linear regression, which minimizes least-squares¹, and fit a line to the data (Fig. 2a). Each data point has one of two distinct y -values (0 and 1), which correspond to the probability of playing professional basketball, and the fit represents the predicted probability as a function of height, increasing from 0 at 159 cm to 1 at 225 cm. The fit line is truncated outside the [0, 1] range because it cannot be interpreted as a probability. Using a probability threshold of 0.5 for classification, we find that 192 cm should be the decision boundary for predicting whether an individual plays professional basketball. It gives reasonable classification performance—only one point is misclassified as false positive, and one point as false negative (Fig. 2a).

Unfortunately, our linear regression fit is not robust. Consider a child of height $H = 100$ cm who does not play professional basketball (Fig. 2a). This height is below the threshold of 192 cm and would be classified correctly. However, if this data point is part of the training set, it will greatly influence the fit³ and increase the classification threshold to 197 cm, which would result in an additional false negative.

To improve the robustness and general performance of this classifier, we could fit the data to a curve other than a straight line. One very simple option is the step function (Fig. 2b), which is 1 when greater than a certain value and 0 otherwise. An advantage of the step function is that it defines a decision boundary (185 cm) that is not affected by the outlier ($H = 100$ cm), but it cannot provide class probabilities other than 0 and 1. This turns out to be sufficient for the purpose of classification—many classification algorithms do not provide probabilities. However, the step function also does not differentiate between the more extreme observations, which are far from the decision boundary and more likely to be correctly assigned, and those near the decision boundary for which membership in

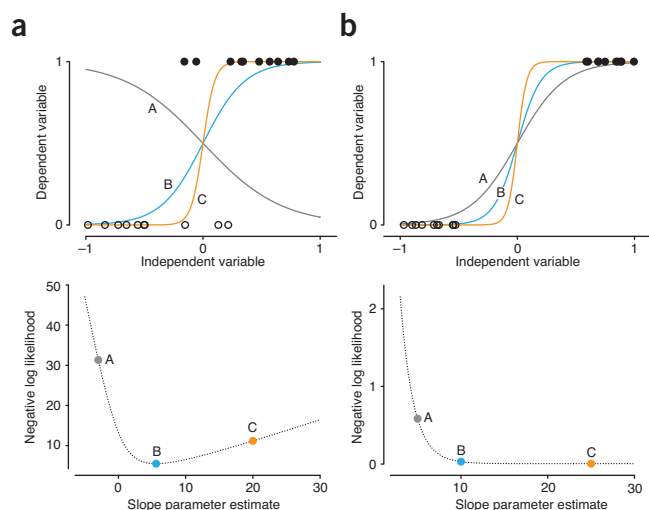


Figure 3 | Optimal estimates in logistic regression are found iteratively via minimization of the negative log likelihood. The slope parameter for each logistic curve (upper plot) is indicated by a correspondingly colored point in the lower plot, shown with its associated negative log likelihood. (a) A non-separable data set with different logistic curves using a single slope parameter. A minimum is found for the ideal curve (blue). (b) A perfectly separable data set for which no minimum exists. Attempts at a solution create increasingly steeper curves—the negative log likelihood asymptotically decreases toward zero, and the estimated slope tends toward infinity.

either group is plausible. In addition, the step function is not differentiable at the step, and regression generally requires a function that is differentiable everywhere. To mitigate this issue, smooth sigmoid curves are used. One used commonly in the natural sciences is the logistic curve (Fig. 2b), which readily relates to the odds ratio.

If p is the probability that a person plays professional basketball, then the odds ratio is $p/(1-p)$, which is the ratio of the probability of playing to the probability of not playing. The log odds ratio is the logarithmic transform of this quantity, $\ln(p/(1-p))$. Logistic regression models the log odds ratio as a linear combination of the independent variables. For our example, height (H) is the independent variable, the logistic fit parameters are β_0 (intercept) and β_H (slope), and the equation that relates them is $\ln(p/(1-p)) = \beta_0 + \beta_H H$. In general, there may be any number of predictor variables and associated regression parameters (or slopes). Modeling the log odds ratio allows us to estimate the probability of class membership using a linear relationship, similar to linear regression. The log odds can be transformed back to a probability as $p(t) = 1/(1 + \exp(-t))$, where $t = \beta_0 + \beta_H H$. This is an S-shaped (sigmoid) curve, with steepness controlled by β_H that maps the linear function back to probabilities in $[0, 1]$.

As in linear regression, we need to estimate the regression parameters. These estimates are denoted by b_0 and b_H to distinguish them from the true but unknown intercept β_0 and slope β_H . Unlike linear regression¹, which yields an exact analytical solution for the estimated regression coefficients, logistic regression requires numerical optimization to find the optimal estimate, such as the iterative approach shown in Figure 3a. For our example, this would correspond to

finding the maximum likelihood estimates, the pair of estimates b_0 and b_H that maximize the likelihood of the observed data (or, equivalently, minimize the negative log likelihood). Once these estimates are found, we can calculate the membership probability, which is a function of these estimates as well as of our predictor H .

In most cases, the maximum-likelihood estimates are unique and optimal. However, when the classes are perfectly separable, this iterative approach fails because there is an infinite number of solutions with equivalent predictive power that can perfectly predict class membership for the training set. Here, we cannot estimate the regression parameters (Fig. 3b) or assign a probability of class membership.

The interpretation of logistic regression shares some similarities with that of linear regression; for instance, variables given the greatest importance may be reliable predictors but might not actually be causal. Logistic regression parameters can be used to understand the relative predictive power of different variables, assuming that the variables have already been normalized to have a mean of 0 and variance of 1. It is important to understand the effect that a change to an independent variable will have on the results of a regression. In linear regression the coefficients have an additive effect for the predicted value, which increases by β_i when the i th independent variable increases by one unit. In logistic regression the coefficients have an additive effect for the log odds ratio rather than for the predicted probability.

Similar to linear regression, correlation among multiple predictors is a challenge to fitting logistic regression. For instance, if we are fitting a logistic regression for professional basketball using height and weight, we must be aware that these variables are highly positively correlated. Either one of them already gives insight into the value of the other. If two variables are perfectly correlated, then there would be multiple solutions to the logistic regression that would give exactly the same fit. Correlated features also make interpretation of coefficients much more difficult. Discussion of the quality of the fit of the logistic model and of classification accuracy will be left to a later column.

Logistic regression is a powerful tool for predicting class probabilities and for classification using predictor variables. For example, one can model the lethality of a new drug protocol in mice by predicting the probability of survival or, with an appropriate probability threshold, by classifying on the basis of survival outcome. Multiple factors of an experiment can be included, such as dosing information, animal weight and diet data, but care must be taken in interpretation to include the possibility of correlation.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Jake Lever, Martin Krzywinski & Naomi Altman

- Altman, N. & Krzywinski, M. *Nat. Methods* **12**, 999–1000 (2015).
- Krzywinski, M. & Altman, N. *Nat. Methods* **12**, 1103–1104 (2015).
- Altman, N. & Krzywinski, M. *Nat. Methods* **13**, 281–282 (2016).

Jake Lever is a PhD candidate at Canada's Michael Smith Genome Sciences Centre. Martin Krzywinski is a staff scientist at Canada's Michael Smith Genome Sciences Centre. Naomi Altman is a Professor of Statistics at The Pennsylvania State University.