Logistic Regression

([gustavoc@msu.edu)](mailto:gustavoc@msu.edu))

**Binary Outcomes**

Many outcomes of interest are binary, implying that they can take two values (say, 0/1). Disease is a typical example of this. Binary random variables follow Bernoulli distributions: or .

**Logistic Regression**

We are often interested on learning the effects of some factors (e.g., sex) and covariates (e.g., age) on the probability of a binary outcome (e.g., disease). In logistic regression, we make a function of covariates. Since we cannot model directly using linear regression because a linear function can take any value in the real line. To deal with this problem we introduce a “link” function (e.g., probit, logit). A link function maps from the real line onto the . The most commonly used link is the logit which is the logarithm of the odds of success, that is:. This function can take values in the real line, thus, we can model the logit using linear methods

. [1]

Note that the above regression is a regression for the probability, not for the data, thus, it typically does not include an error term (in some over-dispersed models it may contain an error).

**From regression to probabilities**

Solving [1] for gives

. [2]

Letting the right-hand side of [1], i.e., the regression function, be then we have: .

**Likelihood function**

The likelihood function is the probability of the data given the parameters. As before, we will assume conditional independence, meaning that

The probability of the ith data-point is:

;

or,

Therefore, assuming conditional independence, the joint likelihood becomes

[3]

Note that above: (i) is a realized value of the corresponding Bernoulli random variable (, therefore, can take values either 0 or 1. (ii) is a function of both covariates and parameters.

**Maximum Likelihood estimation**

Maximum likelihood estimates are obtained by maximizing [3] with respect to the parameters (). The function glm in R fits logistic regression via maximum likelihood. We can also fit a logistic regression using a general-purpose optimization algorithm (e.g., optim in R). The entry logisticRegression.md in our gitHub repository shows how to fit logistic regression using glm and optim.