

## The Logistic Regression Equation

The form of the logistic regression equation is:

$$\text{logit}[p(x)] = \log\left(\frac{p(x)}{1-p(x)}\right) = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \dots$$

- This looks just like a linear regression and although logistic regression finds a best fitting equation, just as linear regression does, the principles on which it does so are rather different.
- Instead of using a least-squared deviations criterion for the best fit, it uses a maximum likelihood method, which maximizes the probability of getting the observed results given the fitted regression coefficients.
- A consequence of this is that the goodness of fit and overall significance statistics used in logistic regression are different from those used in linear regression.
- The probability that a case is in a particular category,  $p$ , can be calculated with the following formula.

$$p = \frac{\exp(b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \dots)}{1 + \exp(b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \dots)}$$

## 0.1 The Logistic Regression Equation

The form of the logistic regression equation is:

$$\text{logit}[p(x)] = \log\left(\frac{p(x)}{1-p(x)}\right) = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \dots$$

This looks just like a linear regression and although logistic regression finds a best fitting equation, just as linear regression does, the principles on which it does so are rather different. Instead of using a least-squared deviations criterion for the best fit, it uses a maximum likelihood method, which maximizes the probability of getting the observed results given the fitted regression coefficients. A consequence of this is that the goodness of fit and overall significance statistics used in logistic regression are different from those used in linear regression.

The probability that a case is in a particular category,  $p$ , can be calculated with the following formula (which is simply another rearrangement of the previous formula).

$$p = \frac{\exp(b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \dots)}{1 + \exp(b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \dots)}$$

## 1 Review of Logistic Regression

There is a direct relationship between the coefficients produced by **logit** and the odds ratios produced by the logistic procedure. First, let's define what is meant by a logit: A logit is defined as the log base  $e$  (log) of the odds,

$$\text{logit}(p) = \log(\text{odds}) = \log(p/q)$$

Logistic regression is in reality ordinary regression using the logit as the response variable,

$$\text{logit}(p) = a + bX$$

$$\log(p/q) = a + bX$$

This means that the coefficients in logistic regression are in terms of the log odds, that is, the coefficient 1.694596 implies that a one unit change in gender results in a 1.694596 unit change in the log of the odds.

Equation [3] can be expressed in odds by getting rid of the log. This is done by taking  $e$  to the power for both sides of the equation.

$$p/q = e^{a+bX}$$

The end result of all the mathematical manipulations is that the odds ratio can be computed by raising  $e$  to the power of the logistic coefficient,

$$OR = e^b = e^{1.694596} = 5.44$$

### 1.0.1 Log transformation

Unfortunately a further mathematical transformation a log transformation is needed to normalize the distribution. Transformations, such as log transformations and square root transformations transform non-normal/skewed distributions closer to normality.

This log transformation of the  $p$  values to a log distribution enables us to create a link with the normal regression equation. The log distribution (or logistic transformation of  $p$ ) is also called the logit of  $p$  or ***logit(p)***.

## 1.1 The Logit

The convention for binomial logistic regression is to code the dependent class of greatest interest as 1 and the other class as 0, because the coding will affect the odds ratios and slope estimates.

The  $\text{logit}(p)$  is the log (to base  $e$ ) of the odds ratio or likelihood ratio that the dependent variable is 1. In symbols it is defined as:

$$\text{logit}(p) = \ln \left( \frac{p}{(1-p)} \right)$$

Whereas  $p$  can only range from 0 to 1,  $\text{logit}(p)$  scale ranges from negative infinity to positive infinity and is symmetrical around the logit of 0.5 (which is zero)

## 1.2 Log-Odds

As an alternative to modeling the value of the outcome, logistic regression focuses instead upon the relative probability (odds) of obtaining a given result category. The natural logarithm of the odds is linear across most of its range, allowing us to continue using many of the methods developed for linear models. The result of this type of regression can be expressed as follows:

$$\text{Ln} \left[ \frac{p}{1-p} \right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 \dots b_kx_k + e$$

## Logits

In logistic regression, the logit may be computed in a manner similar to linear regression:

$$\eta_i = \beta_0 + \beta_1x_1 + \beta_2x_2 + \dots$$

## 1.3 Logistic function

The logistic function of any number is given by the inverse-logit:

$$\text{logit}^{-1}(\alpha) = \frac{1}{1 + \exp(-\alpha)} = \frac{\exp(\alpha)}{1 + \exp(\alpha)}$$

## 1.4 Logistic Regression: Odds Ratio

What are odds? The odds of outcome 1 versus outcome 2 are the probability (or frequency) of outcome 1 divided by the probability (or frequency) of outcome 2.

$$\hat{Y} = \frac{\text{Odds}}{1 + \text{Odds}}$$

# 2 Logistic Regression

Logistic regression, also called a logit model, is used to model **dichotomous outcome** variables. In the logit model the **log odds** of the outcome is modeled as a linear combination of the predictor variables.

In logistic regression theory, the predicted dependent variable is a function of the probability that a particular subject will be in one of the categories (for example, the probability that a patient has the disease, given their set of scores on the predictor variables).

## 2.1 Odds

- The odds in favor of an event or a proposition are the ratio of the probability that an event will happen to the probability that it will not happen.
- 'Odds' are an expression of relative probabilities. Often 'odds' are quoted as odds against, rather than as odds in favor of, because of the possibility of confusion of the latter with the fractional probability of an event occurring.

$$\text{Odds} = \frac{p}{1 - p}$$

$$\hat{Y} = \frac{\text{Odds}}{1 + \text{Odds}}$$