Logistic regression introduction

Biometry 755

Spring 2009

Odds and odds ratios

The *odds* of an event, A, is the ratio of the probability in favor of event A to the probability against event A. That is,

$$\text{odds } A = \frac{\text{Prob}(A)}{1 - \text{Prob}(A)} = \frac{\text{Prob}(A)}{\text{Prob}(\overline{A})}.$$

Odds and odds ratios (cont.)

Consider the following 2×2 table.

		Disease $+$	Disease $-$	
Exposure	$+$	a	b	$a+b$
	$-$	c	d	$c+d$
		$a+c$	$b+d$	$a+b+c+d$

1. Odds of disease given exposed = _____
2. Odds of disease given unexposed = _____
3. Odds ratio comparing exposed to unexposed = _____

Introduction to logistic regression

Logistic regression is akin to linear regression in that its goals are quite similar: to find the best fitting, plausible model to describe the relationship between an outcome variable and a set of independent variables. What distinguishes the logistic from the linear regression model is that the response variable in logistic regression is *binary* or *dichotomous*. In linear regression, the response variable is continuous. This feature is reflected in both the choice of the logistic regression model as well as the modelling assumptions, but the guiding principles of linear regression continue to serve us in this new paradigm.

Motivation

- Y is a binary (0/1) response variable. Therefore, Y is a Bernoulli random variable (in linear regression the response was assumed to come from a Normal distribution).
- The expected (average) value of a Bernoulli random variable is π which is equal to the probability of success (i.e. $E(Y) = \pi$)
- $X_1, \ldots, X_k$ are k predictor variables (covariates). They can be any combination of continuous, ordinal or nominal variables.

Goal of logistic regression

Recall that in linear regression, we fit the expected (average) response $E(Y)$ as a linear function of covariates. Our goal is similar in logistic regression except that now $E(Y) = \pi$. Therefore, our goal is to model the probability that $Y = 1$ given a particular set of covariate values. For example, we might be interested in estimating a patient's probability of disease given certain risk factors such as the patient's age, race, sex and other comorbid conditions.

Modelling a probability

Therefore, we want to do something like this ...

$$\text{Prob}(Y = 1|X_1, \ldots, X_k) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k.$$

But there is one big problem with this ...

A probability must be constrained to fall between 0 and 1, so the left hand side of the equation (i.e. $\text{Prob}(Y = 1|X_1, \ldots, X_k)$) can't be any smaller than 0 or any larger than 1. But $\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$ has the potential to take on any positive or negative number!

A model like this could lead to some ridiculous values for probabilities. Clearly, this model will NOT work.

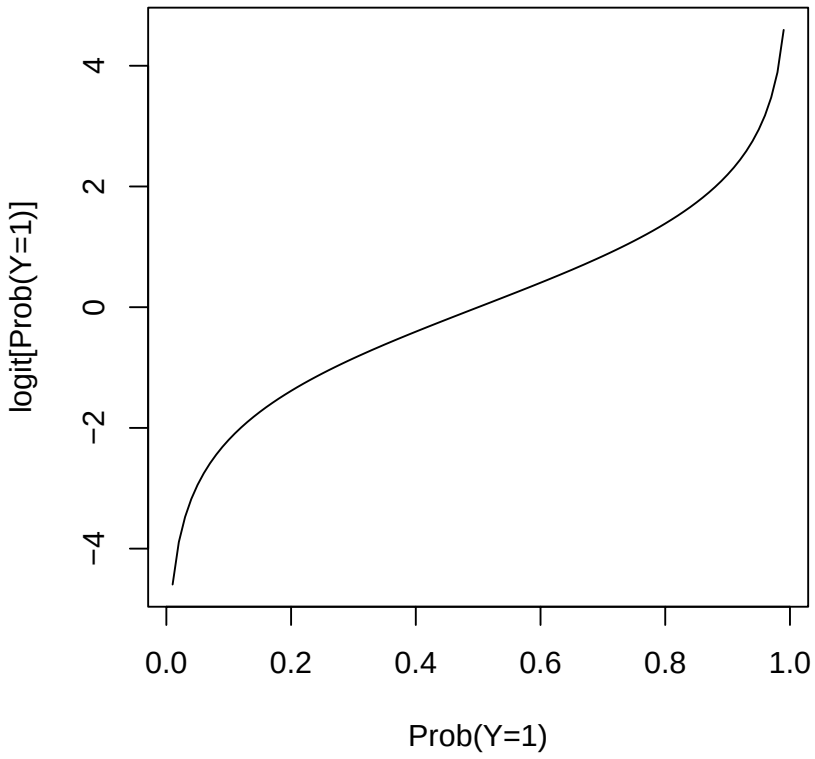
The logit function

So what should we do? The solution to the problem is to use a function that transforms the unit interval, $[0, 1]$ (the space where probabilities live), to the entire real line, $(-\infty, \infty)$, (the space where a general class of covariates live). That transformation is the logit function. It is defined as follows.

$$\text{logit}[\text{Prob}(Y = 1|X_1, \ldots, X_k)] \quad = \quad \ln\left[\frac{\text{Prob}(Y = 1|X_1, \ldots, X_k)}{1 - \text{Prob}(Y = 1|X_1, \ldots, X_k)}\right]$$

In words, the logit model expresses the (natural) log odds of the event $Y = 1$ given covariate values $X_1, \ldots, X_k$.

Graph of the logit function



The logit function (cont.)

By taking the logit transformation of Prob$(Y = 1|\mathbf{X})$, we are able to explore the relationship between the probability of an outcome (which must be between 0 and 1) with one or more covariates that can take on values anywhere on the real line. To get back to the scale of a probability, all we need to do at the end is transform back using the inverse function. The logistic regression model is written as

$$\begin{aligned} \text{logit}[\text{Prob}(Y = 1|\mathbf{X})] &= \ln\left[\frac{\text{Prob}(Y = 1|\mathbf{X})}{1 - \text{Prob}(Y = 1|\mathbf{X})}\right] \\ &= \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k. \end{aligned}$$

Digression into some convenient notation

In manipulating the logistic model, it becomes cumbersome to write

$$\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k.$$

It is much simpler to write this as the inner product of two vectors. Let $\mathbf{X} = \begin{pmatrix} 1 \\ X_1 \\ X_2 \\ \vdots \\ X_k \end{pmatrix}$ and $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}$. Let $\mathbf{X}'$ be the transpose of $\mathbf{X}$, i.e. $\mathbf{X}' = (1 \; X_1 \; X_2 \; \ldots \; X_k)$.

Convenient notation (cont.)

Then

$$\begin{aligned} \mathbf{X}'\boldsymbol{\beta} &= (1 \; X_1 \; X_2 \; \ldots \; X_k) \times \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} \\ &= \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k. \end{aligned}$$

What is the inverse logit function?

For simplicity, write $\pi = \text{Prob}(Y = 1|\mathbf{X})$. If

$$\ln\left[\frac{\pi}{1-\pi}\right] = \mathbf{X}'\boldsymbol{\beta},$$

then $\pi =$?.

The IMPACT study

The IMPACT data set is a subset of data collected for the University of Massachusetts Aids Research Unit IMPACT Study. This was a 5-year (1989-1994) project comprising two randomized trials of residential treatment for drug abuse. The purpose of the study was to compare treatment programs of different planned durations designed to reduce drug abuse and to prevent high-risk HIV behavior. The data are described in the handout.