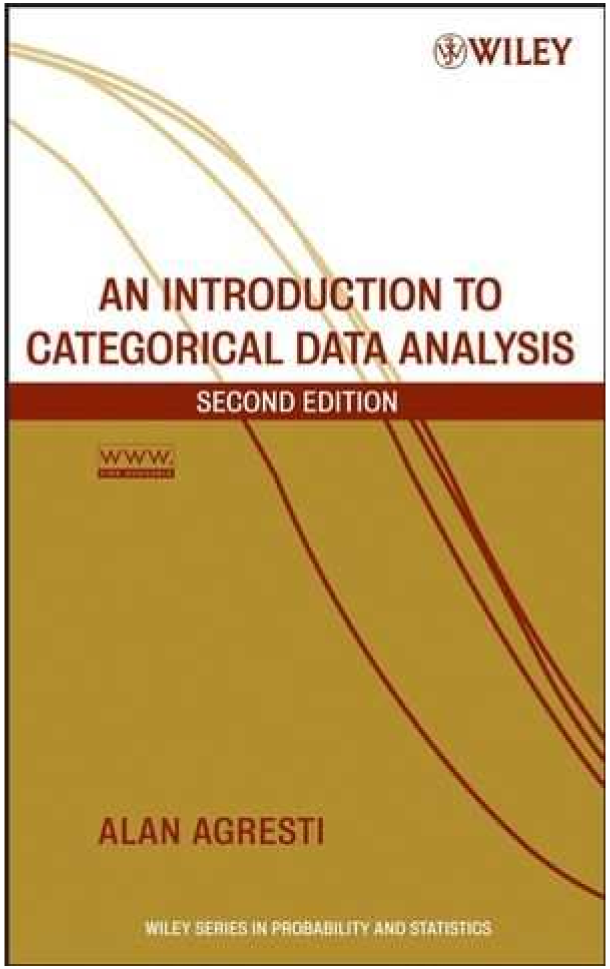
Analisi Multivariata dei Dati Sperimentali

Massimo Borelli

i modelli lineari generalizzati

- intro teorica



WILEY

AN INTRODUCTION TO CATEGORICAL DATA ANALYSIS

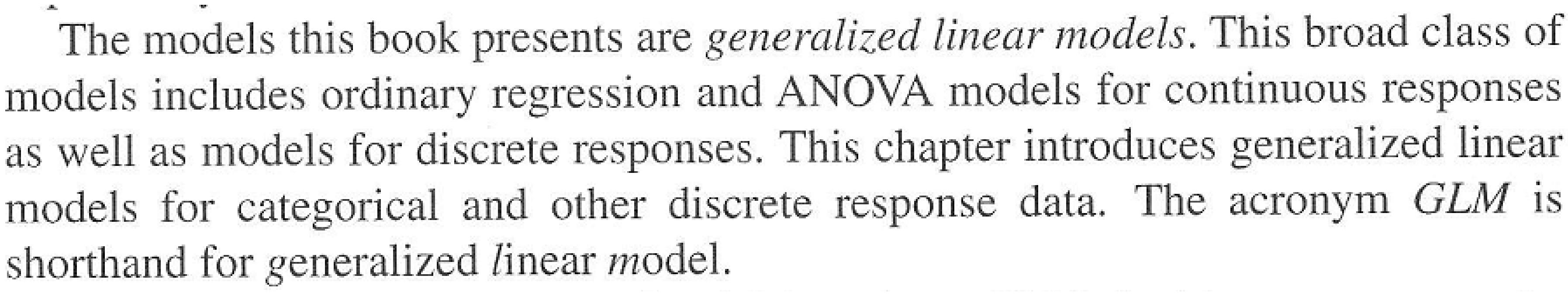
SECOND EDITION

WWW.

ALAN AGRESTI

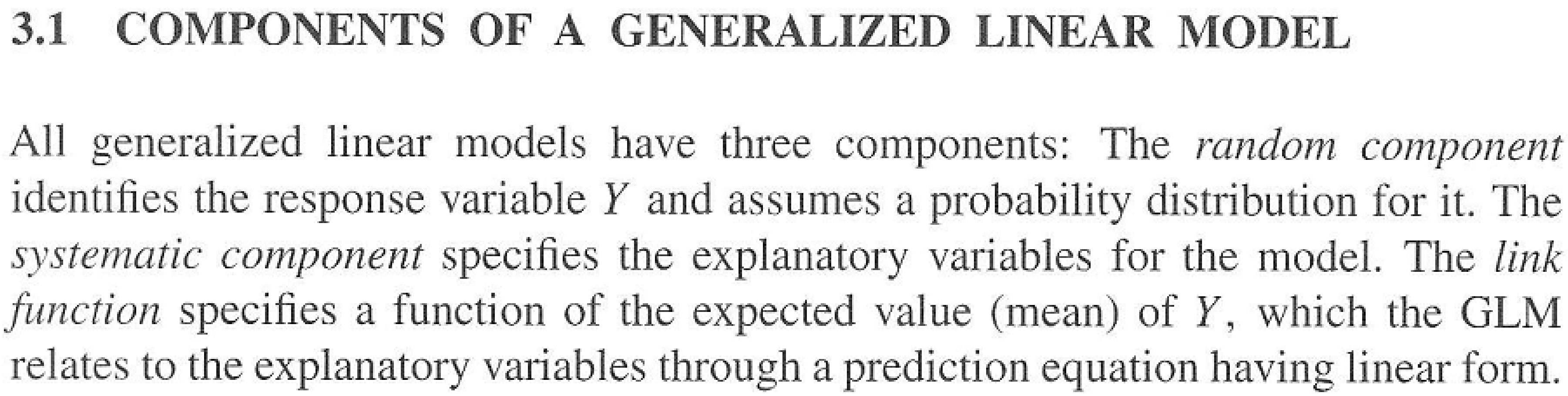
WILEY SERIES IN PROBABILITY AND STATISTICS

spiegazione



The models this book presents are *generalized linear models*. This broad class of models includes ordinary regression and ANOVA models for continuous responses as well as models for discrete responses. This chapter introduces generalized linear models for categorical and other discrete response data. The acronym *GLM* is shorthand for generalized *l*inear *m*odel.

spiegazione

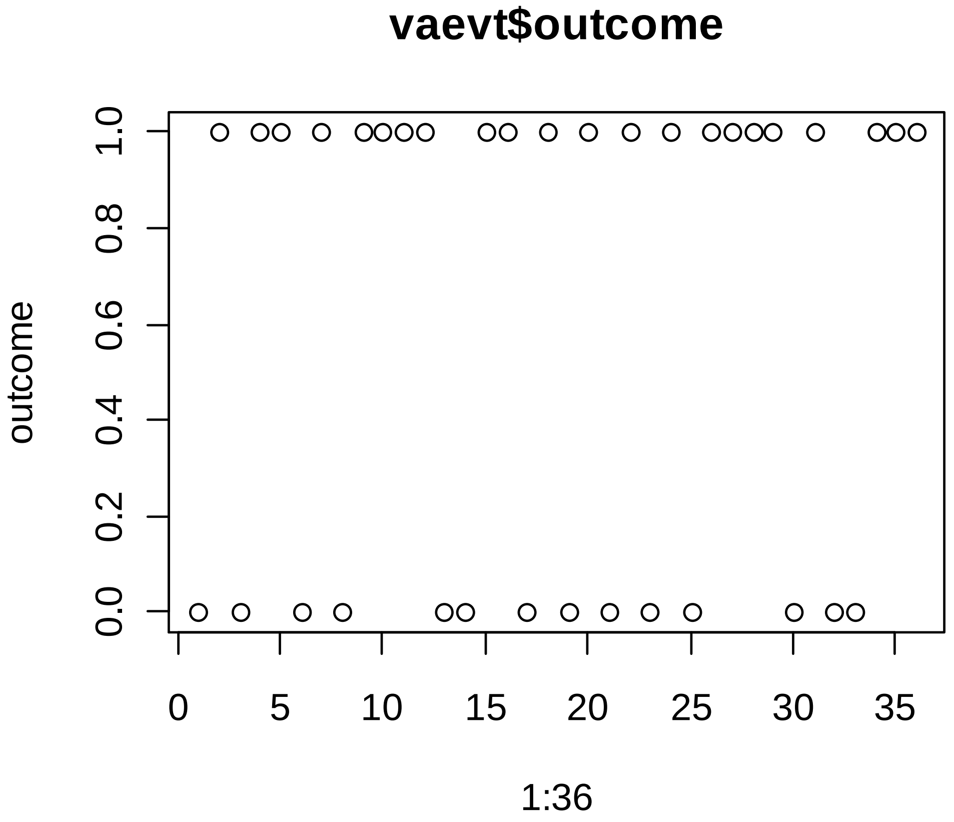


3.1 COMPONENTS OF A GENERALIZED LINEAR MODEL

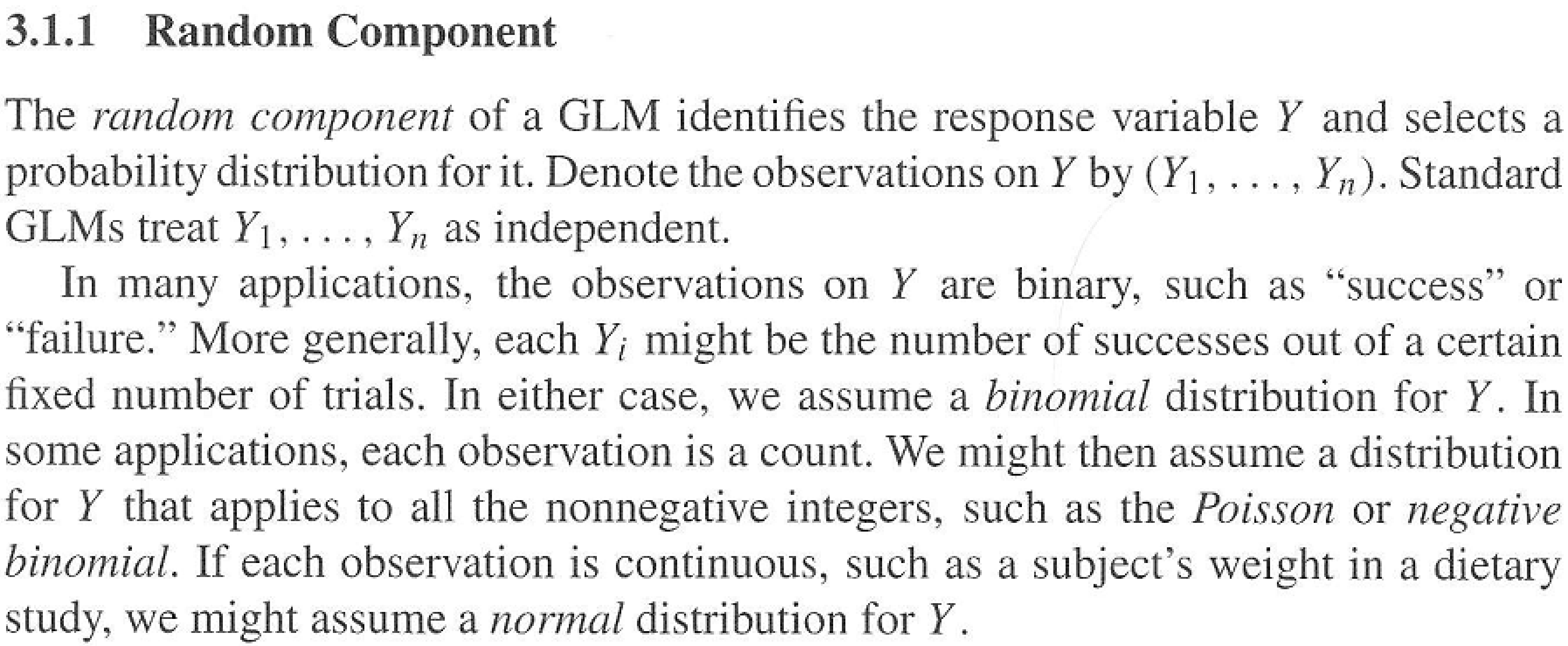
All generalized linear models have three components: The *random component* identifies the response variable Y and assumes a probability distribution for it. The *systematic component* specifies the explanatory variables for the model. The *link function* specifies a function of the expected value (mean) of Y, which the GLM relates to the explanatory variables through a prediction equation having linear form.

es: outcome

3.1.1 Random Component



spiegazione

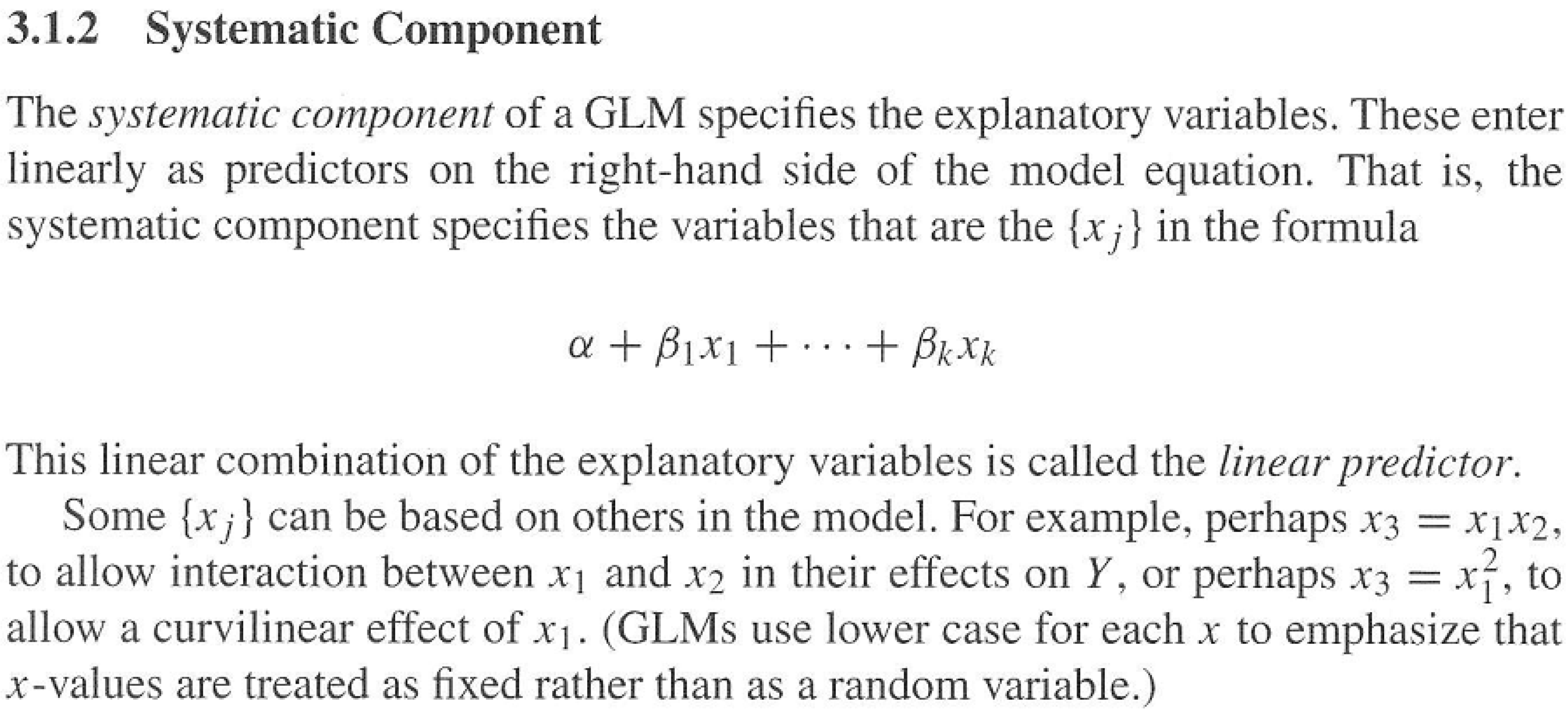


3.1.1 Random Component

The *random component* of a GLM identifies the response variable Y and selects a probability distribution for it. Denote the observations on Y by $(Y_1, \ldots, Y_n)$. Standard GLMs treat $Y_1, \ldots, Y_n$ as independent.

In many applications, the observations on Y are binary, such as "success" or "failure." More generally, each Y_i might be the number of successes out of a certain fixed number of trials. In either case, we assume a *binomial* distribution for Y. In some applications, each observation is a count. We might then assume a distribution for Y that applies to all the nonnegative integers, such as the *Poisson* or *negative binomial*. If each observation is continuous, such as a subject's weight in a dietary study, we might assume a *normal* distribution for Y.

spiegazione



3.1.2 Systematic Component

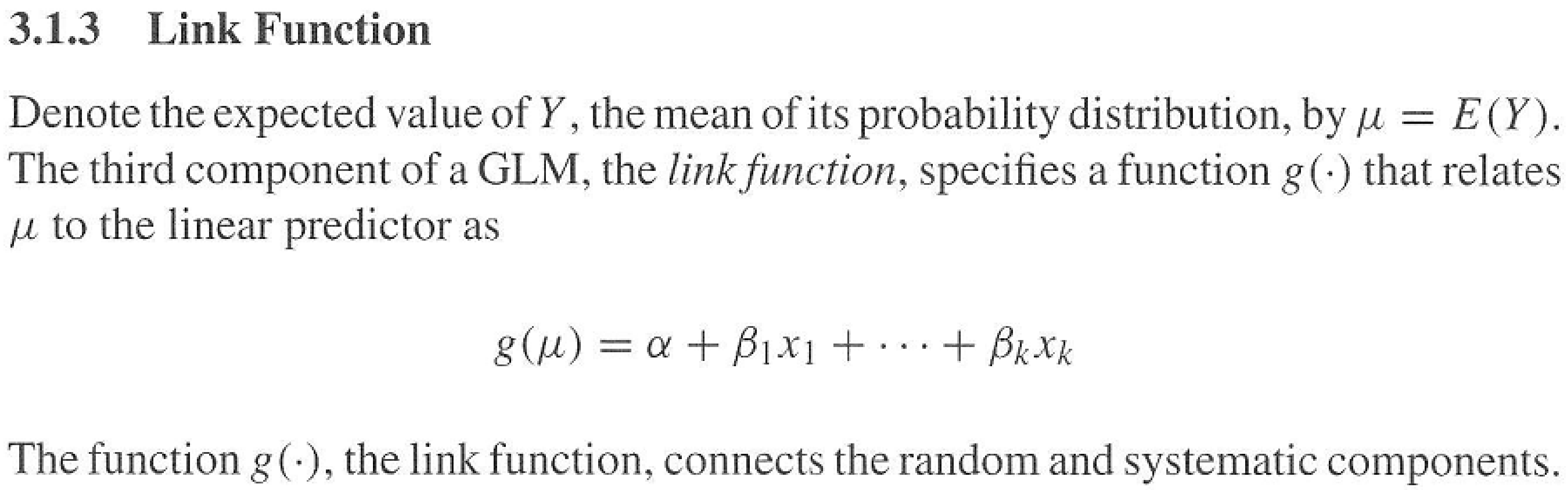
The *systematic component* of a GLM specifies the explanatory variables. These enter linearly as predictors on the right-hand side of the model equation. That is, the systematic component specifies the variables that are the $\{x_j\}$ in the formula

$$\alpha + \beta_1 x_1 + \cdots + \beta_k x_k$$

This linear combination of the explanatory variables is called the *linear predictor*.

Some $\{x_j\}$ can be based on others in the model. For example, perhaps $x_3 = x_1 x_2$, to allow interaction between x_1 and x_2 in their effects on Y, or perhaps $x_3 = x_1^2$, to allow a curvilinear effect of x_1. (GLMs use lower case for each x to emphasize that x-values are treated as fixed rather than as a random variable.)

spiegazione



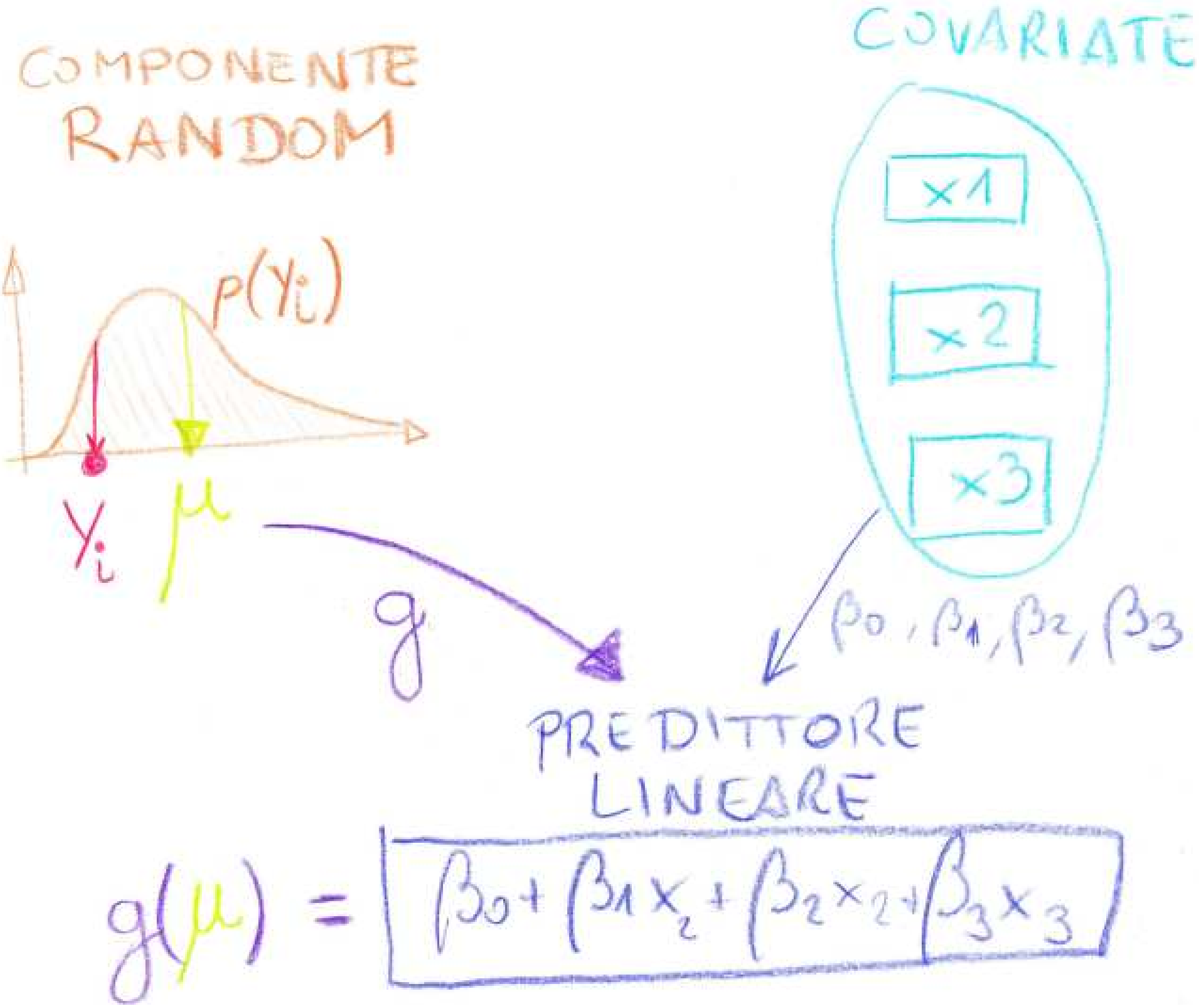
3.1.3 Link Function

Denote the expected value of Y, the mean of its probability distribution, by $\mu = E(Y)$. The third component of a GLM, the *link function*, specifies a function $g(\cdot)$ that relates μ to the linear predictor as

$$g(\mu) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$$

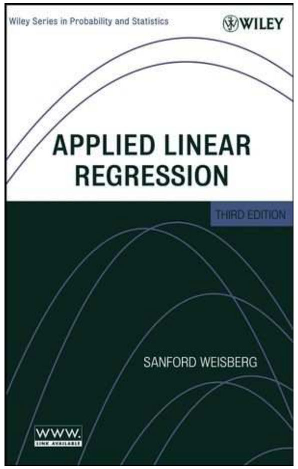
The function $g(\cdot)$, the link function, connects the random and systematic components.

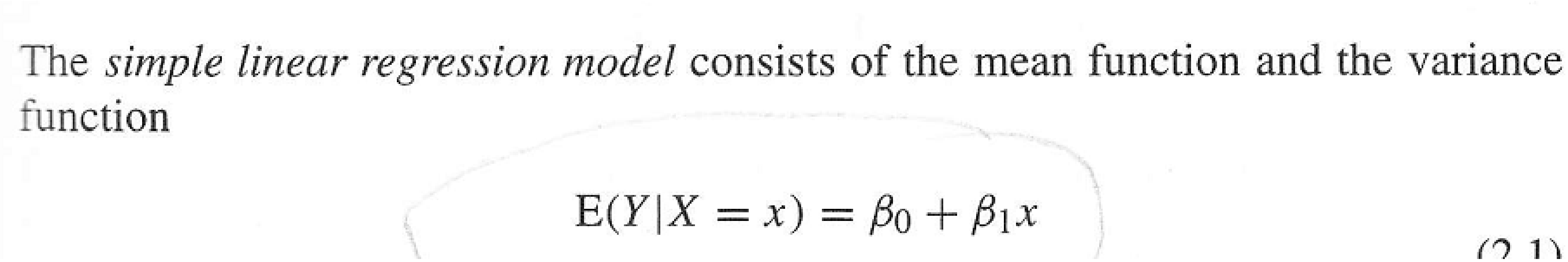
pupoletto


COMPONENTE RANDOM
$p(Y_i)$
Y_i
μ
g
COVARIATE
x1
x2
x3
$\beta_0, \beta_1, \beta_2, \beta_3$
PREDITTORE LINEARE
$g(\mu) = \boxed{\beta_0 + \beta_1 x_2 + \beta_2 x_2 + \beta_3 x_3}$

spiegazione

- ricordate?

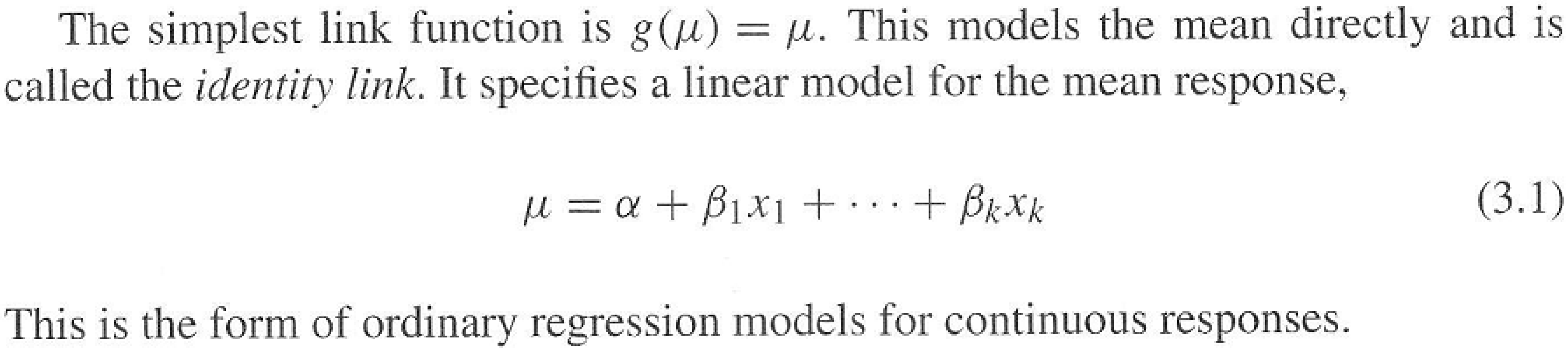

Wiley Series in Probability and Statistics
WILEY
APPLIED LINEAR REGRESSION
THIRD EDITION
SANFORD WEISBERG
WWW.



The *simple linear regression model* consists of the mean function and the variance function

$$E(Y|X = x) = \beta_0 + \beta_1 x \quad (2.1)$$

esempio 1: GLM = regressione

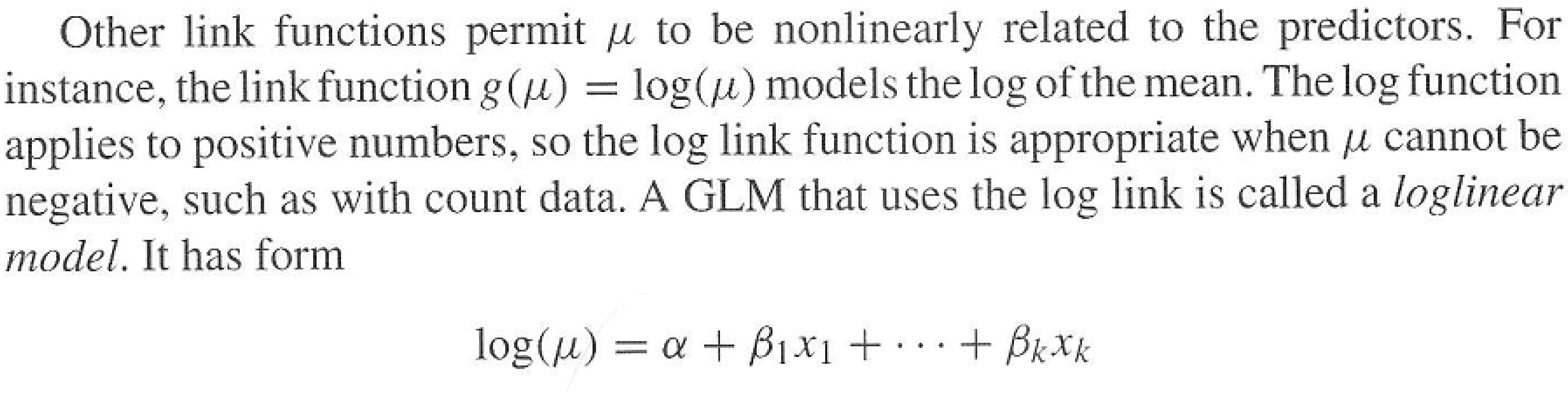


The simplest link function is $g(\mu) = \mu$. This models the mean directly and is called the *identity link*. It specifies a linear model for the mean response,

$$\mu = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k \tag{3.1}$$

This is the form of ordinary regression models for continuous responses.

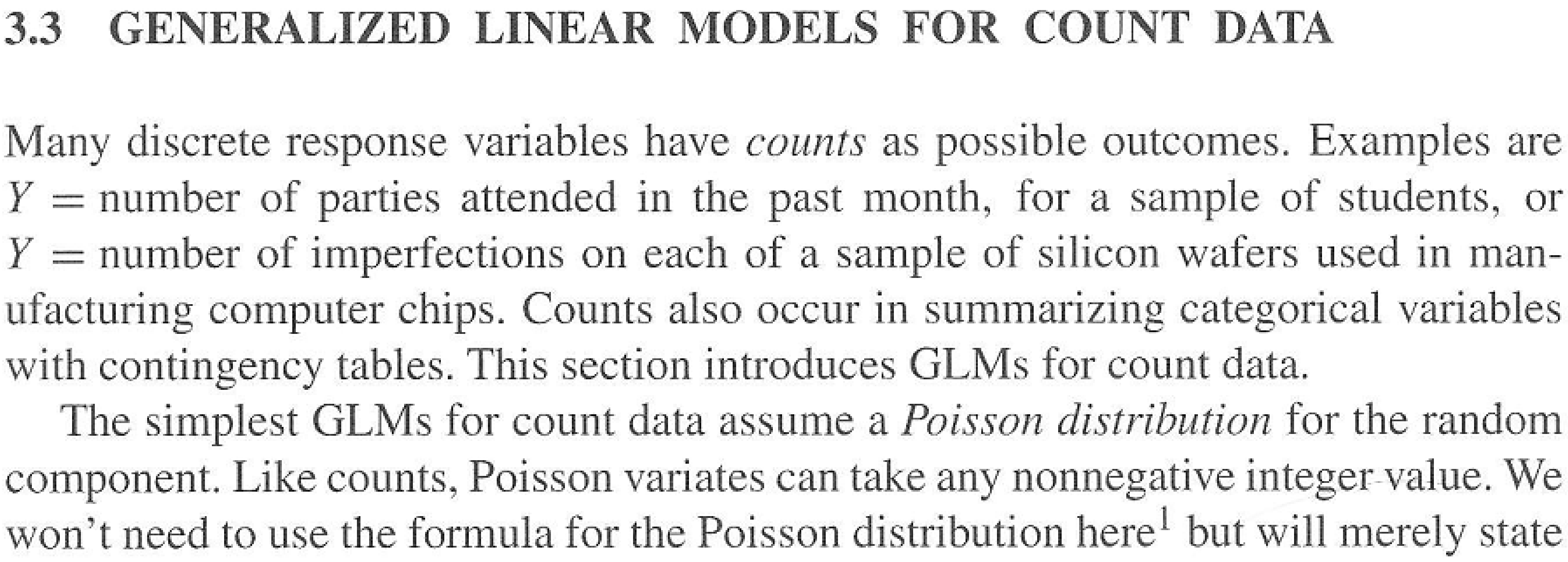
esempio 2: Poisson



Other link functions permit μ to be nonlinearly related to the predictors. For instance, the link function $g(\mu) = \log(\mu)$ models the log of the mean. The log function applies to positive numbers, so the log link function is appropriate when μ cannot be negative, such as with count data. A GLM that uses the log link is called a *loglinear model*. It has form

$$\log(\mu) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$$

esempio 2: Poisson



3.3 GENERALIZED LINEAR MODELS FOR COUNT DATA

Many discrete response variables have *counts* as possible outcomes. Examples are Y = number of parties attended in the past month, for a sample of students, or Y = number of imperfections on each of a sample of silicon wafers used in manufacturing computer chips. Counts also occur in summarizing categorical variables with contingency tables. This section introduces GLMs for count data.

The simplest GLMs for count data assume a *Poisson distribution* for the random component. Like counts, Poisson variates can take any nonnegative integer value. We won't need to use the formula for the Poisson distribution here[1] but will merely state

esempio 2: Poisson

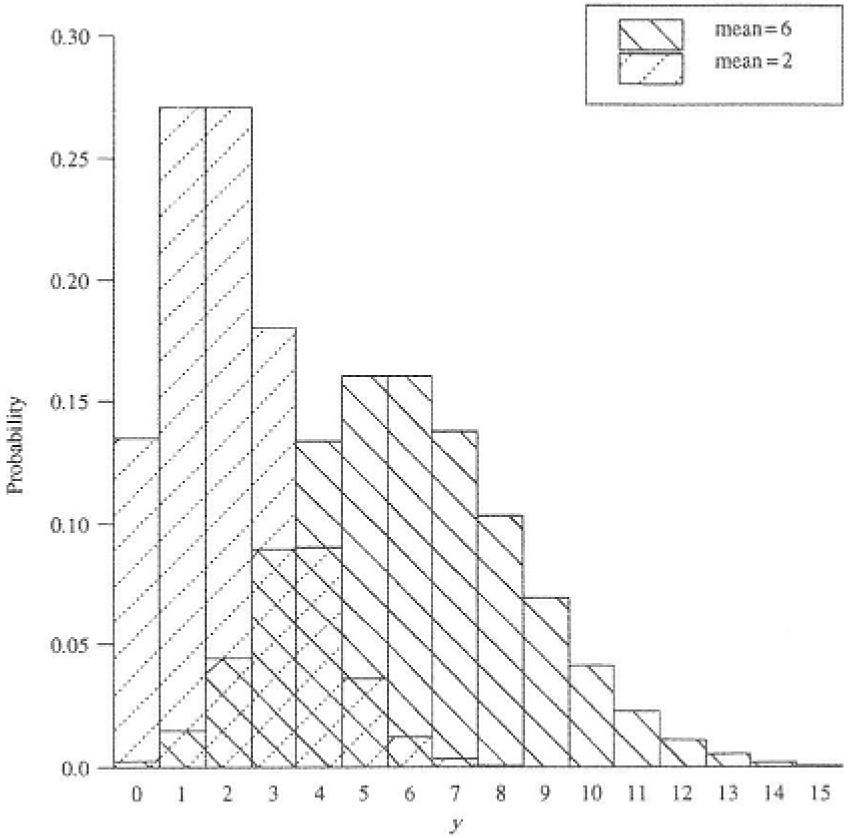
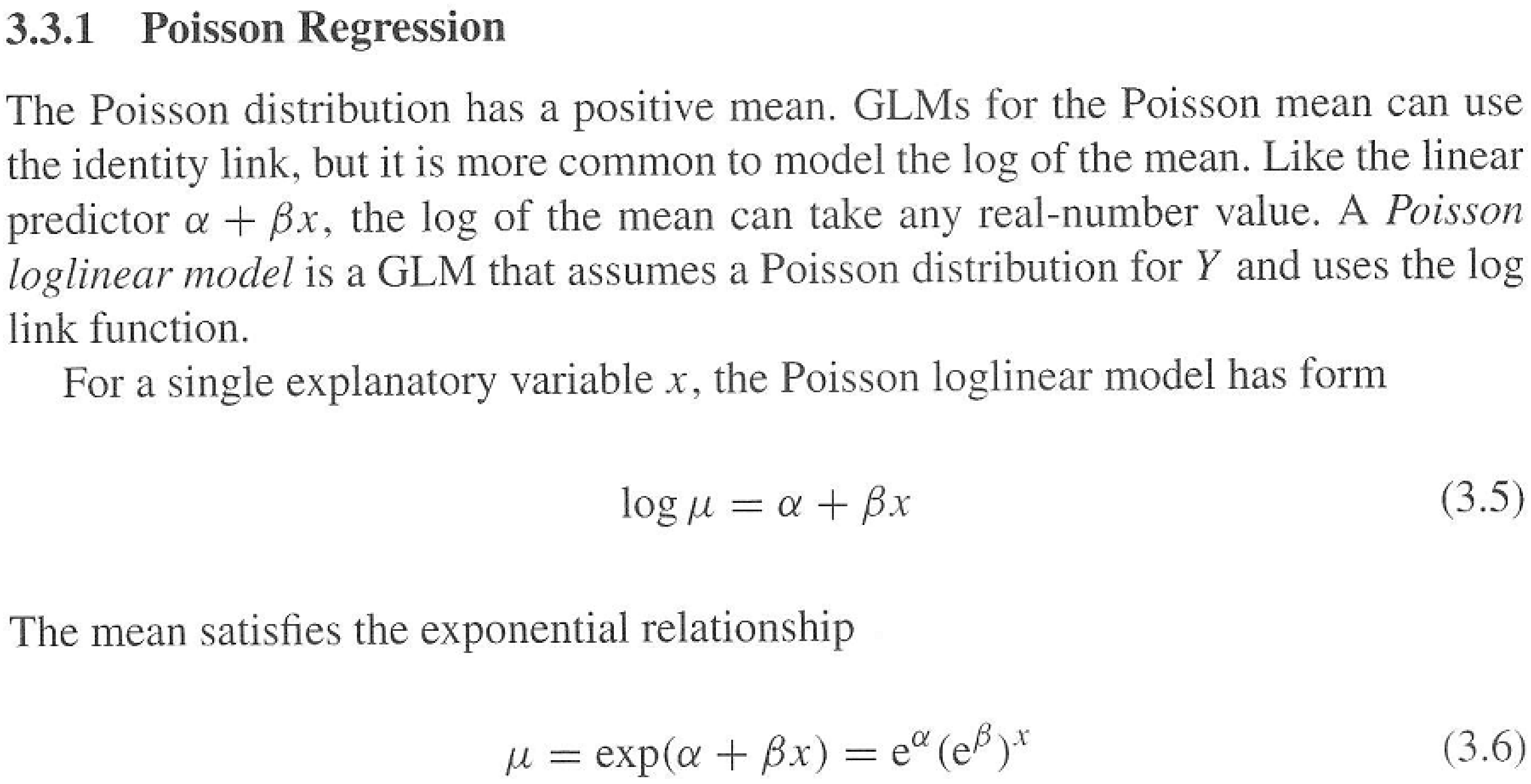


Figure 3.3. Poisson distributions having means 2 and 6.

esempio 2: Poisson



3.3.1 Poisson Regression

The Poisson distribution has a positive mean. GLMs for the Poisson mean can use the identity link, but it is more common to model the log of the mean. Like the linear predictor $\alpha + \beta x$, the log of the mean can take any real-number value. A *Poisson loglinear model* is a GLM that assumes a Poisson distribution for Y and uses the log link function.

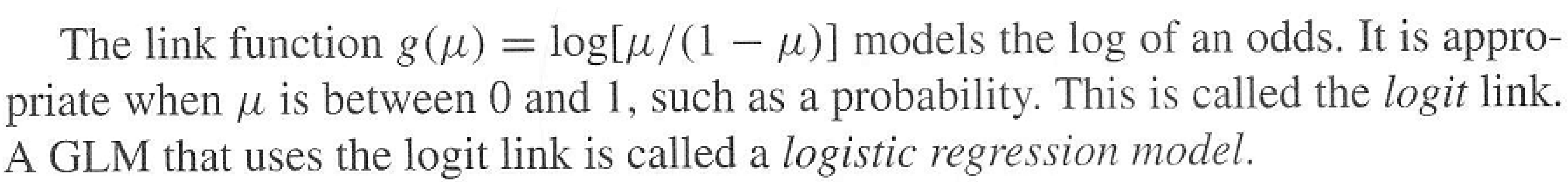
For a single explanatory variable x, the Poisson loglinear model has form

$$\log \mu = \alpha + \beta x \tag{3.5}$$

The mean satisfies the exponential relationship

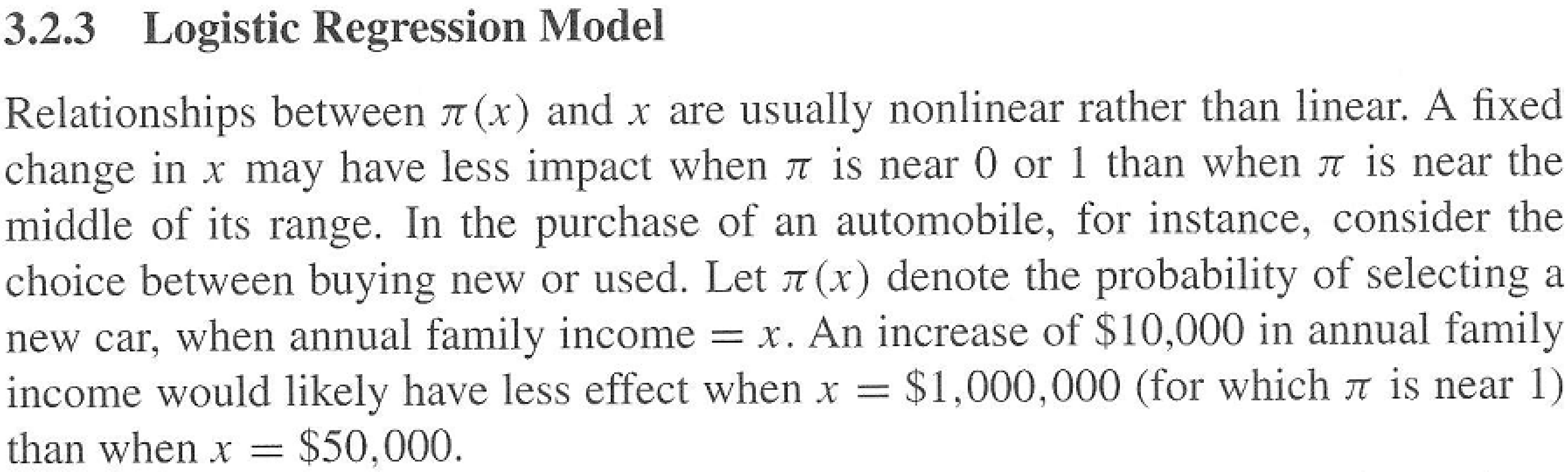
$$\mu = \exp(\alpha + \beta x) = e^{\alpha}(e^{\beta})^{x} \tag{3.6}$$

esempio 3: Binomiale



The link function $g(\mu) = \log[\mu/(1-\mu)]$ models the log of an odds. It is appropriate when μ is between 0 and 1, such as a probability. This is called the *logit* link. A GLM that uses the logit link is called a *logistic regression model*.

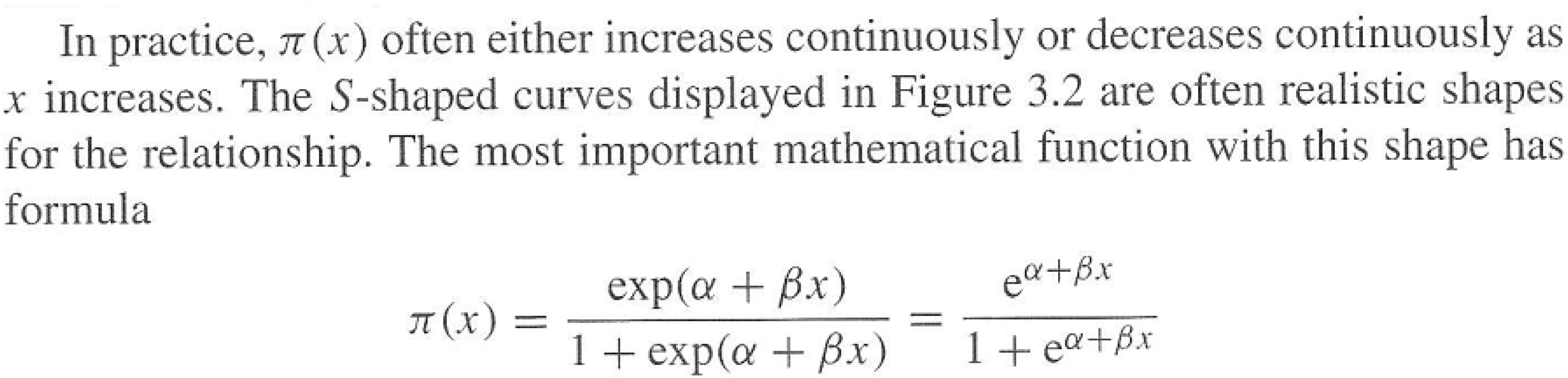
esempio 3: Binomiale



3.2.3 Logistic Regression Model

Relationships between $\pi(x)$ and x are usually nonlinear rather than linear. A fixed change in x may have less impact when π is near 0 or 1 than when π is near the middle of its range. In the purchase of an automobile, for instance, consider the choice between buying new or used. Let $\pi(x)$ denote the probability of selecting a new car, when annual family income $= x$. An increase of \$10,000 in annual family income would likely have less effect when $x = \$1{,}000{,}000$ (for which π is near 1) than when $x = \$50{,}000$.

esempio 3: Binomiale



In practice, $\pi(x)$ often either increases continuously or decreases continuously as x increases. The S-shaped curves displayed in Figure 3.2 are often realistic shapes for the relationship. The most important mathematical function with this shape has formula

$$\pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)} = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}}$$

esempio 3: Binomiale

3.2 GENERALIZED LINEAR MODELS FOR BINARY DATA

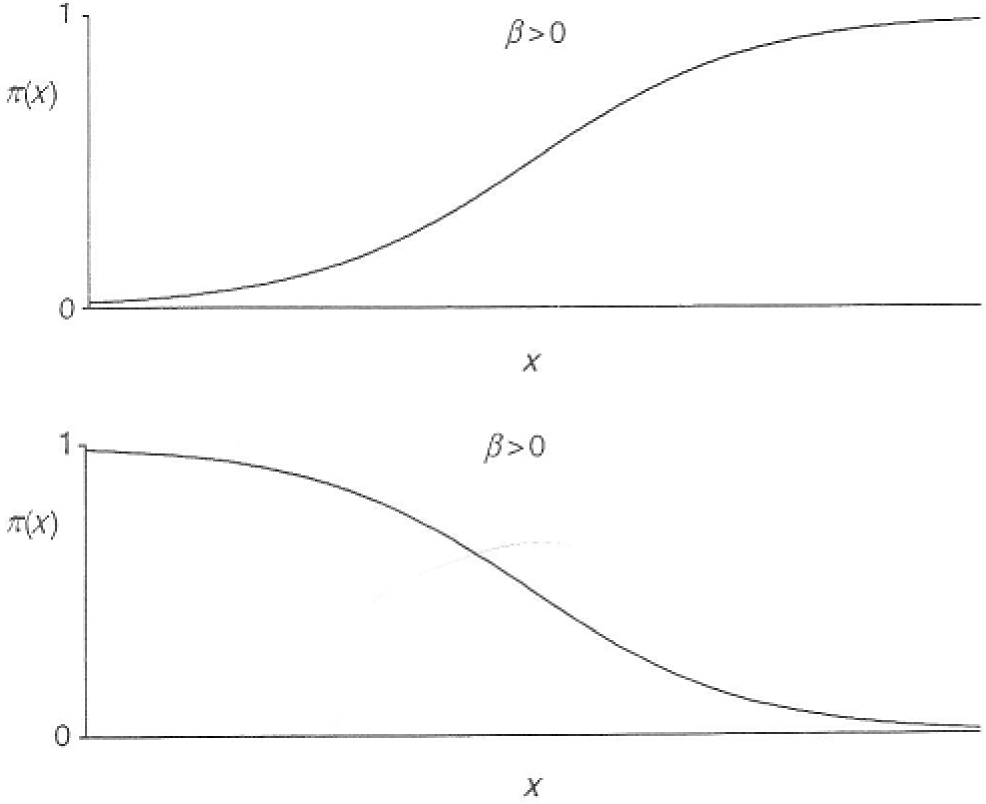
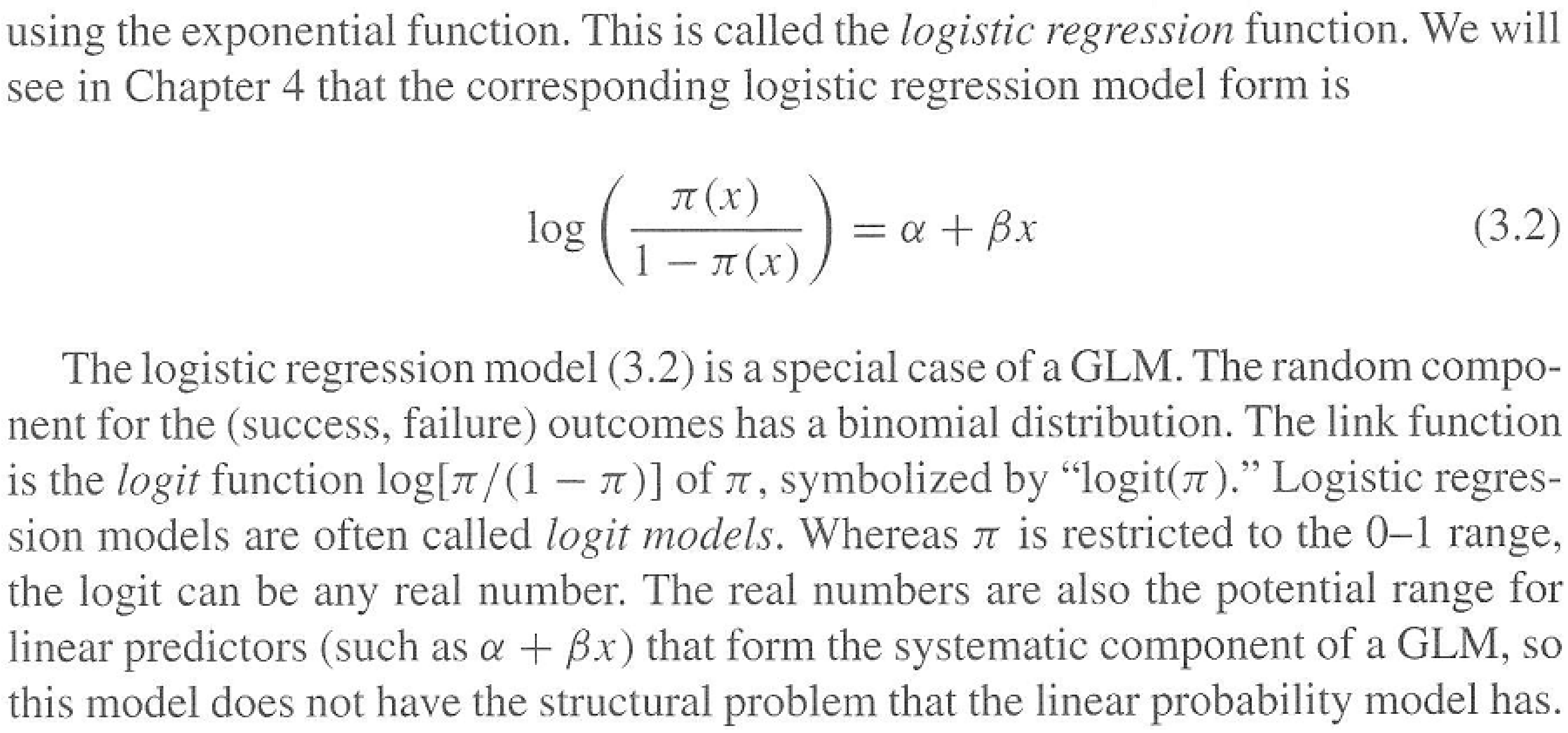


Figure 3.2. Logistic regression functions.

esempio 3: Binomiale



using the exponential function. This is called the *logistic regression* function. We will see in Chapter 4 that the corresponding logistic regression model form is

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta x \tag{3.2}$$

The logistic regression model (3.2) is a special case of a GLM. The random component for the (success, failure) outcomes has a binomial distribution. The link function is the *logit* function $\log[\pi/(1-\pi)]$ of π, symbolized by "logit(π)." Logistic regression models are often called *logit models*. Whereas π is restricted to the 0–1 range, the logit can be any real number. The real numbers are also the potential range for linear predictors (such as $\alpha + \beta x$) that form the systematic component of a GLM, so this model does not have the structural problem that the linear probability model has.