1
Introduction

1.1 The role of the logit model

Logit analysis is in many ways the natural complement of ordinary linear regression whenever the regressand is not a continuous variable but a state which may or may not hold, or a category in a given classification. When such discrete variables occur among the independent variables or regressors of a regression equation, they are dealt with by the introduction of one or several $(0, 1)$ dummy variables; but when the *dependent* variable belongs to this type, the regression model breaks down. Logit analysis or logistic regression (which are two names for the same method) provides a ready alternative. At first sight it is quite different from the familiar linear regression model, and slightly frightening by its apparent complexity; yet the two models have much in common.

First, both models belong to the realm of causal relations, as opposed to statistical association; there is a clear *a priori* asymmetry between the oddly named independent variables, the regressors or covariates, which are the explanatory variables or determinants, and the dependent variable or outcome. Both models were initially designed for the analysis of experimental data, or at least for data where the direction of causation is not in doubt. In interpreting empirical applications it is often helpful to bear these origins in mind.

Within this causal context, the ordinary linear regression model offers a crude but almost universal framework for empirical analysis. Admittedly it is often no more than a simplified approximation to something else that would presumably be better; but it does serve, within its limitations, for empirical screening of the evidence. Logistic regression can be used in quite the same way for categorical phenomena.

There are of course also differences. Unlike regression, the logit model

permits of a specific economic interpretation in terms of utility maximization in situations of discrete choice. Among economists this confers a higher status on the model than that of a convenient empirical device. And there is a subtle distinction in that the ordinary regression model requires a disturbance term which is stuck on to the systematic part as a necessary nuisance, while in the logit model the random character of the outcome is an integral part of the initial specification. Together with the probit model, the logit model belongs to the class of *probability models* that determine discrete probabilities over a limited number of possible outcomes.

Finally, like the regression model, the logit model permits of all sorts of extensions and of quite sophisticated variants. Some of these are touched upon in the later chapters, but the present text is mainly concerned with plain logistic regression as a convenient vehicle for studying the determination of categorical variables.

A survey of the literature will show that a number of different varieties of the same model have been developed in almost perfect isolation in various disciplines such as biology (toxicology), medicine (epidemiology) and economics (econometrics). This has given rise to separate and distinct paradigms which share a common statistical core, but which employ different approaches, terminologies and interpretations, since they deal with different types of data and pursue different ends. Even when the application of the technique has become a mechanical routine, the original justificatory arguments still linger at the back of the practitioners' minds and condition their views. In the present text we follow the approach which originated in the bio-assay of toxicology and was later adopted (and developed further) in certain branches of economics and econometrics. We shall however try to provide links to the work in other fields wherever this is appropriate, and the reader is encouraged to follow these up.

1.2 Plan of the book and further reading

The book consists of nine chapters. Chapter 2 presents the central model that sets the course. Consideration of a single attribute gives rise to the binary model, and this simple vehicle carries the overwhelming majority of practical applications. Chapter 3 deals at some length with its estimation by what is now the standard method of maximum likelihood; while most readers will rely on program packages for their calculations, it is important that they understand the method since it determines the

properties of the resulting estimates. Chapter 3 also provides a practical illustration. Chapter 4 deals with some statistical tests and the assessment of fit, and Chapter 5 with defects like outliers, misclassified outcomes and omitted variables. At this stage just over half of the book has dealt exclusively with the simple case of a binary analysis in a random sample. In Chapter 6 we consider the analysis of separate samples, but still only of a pair. It is only in the next two chapters that the treatment is widened to more than two possible outcomes. Chapter 7 is devoted to the standard multinomial model, which is a fairly straightforward generalization of the binary model; the particular variety of multinomial models known as random utility models, which is an economic specialty, is the subject of Chapter 8. Some but not all of the embellishments of the binary model in Chapters 4, 5 and 6 carry easily over to the multinomial case. Finally Chapter 9 gives a brief account of the history of the subject, with special reference to the approach adopted here.

Since this is after all a slim book, designed for newcomers to the subject, readers are expected to skim through the entire text, and then to return when the need arises to the bits they can use, or – even better – to continue at once with further reading. We have already noted the diversity of parallel but quite separate developments of essentially the same subject in a number of disciplines. Apart from natural differences in the type of data under consideration and in the ends that are pursued, further differences of style have arisen in the course of this development. No discipline is satisfied with establishing empirical regularities. In epidemiology and medicine the tendency is towards simplicity of technique, but statistical associations, even if they are as strong as between cigarette smoking and lung cancer, must be complemented by a reconstruction of the physiological process before they are fully accepted. Economists and econometricians on the other hand never leave well alone, favour all sorts of complications of the statistical model, and believe that the validity of empirical results is enhanced if they can be understood in terms of optimizing behaviour. Each discipline thus has a paradigm of its own that supports the same statistical technique, with huge differences in approach, terminology and interpretation. These differences are at times exaggerated to almost ideological dimensions and they have become an obstacle to open communication between scholars from different disciplines. The reader's outlook will be considerably widened by looking with an open mind at one or two texts from an alien discipline. For this purpose we recommend such varied sources as the classic book

on bio-assay of Finney (1971) (first published in 1947), or the up-to-date monograph on logistic regression from the epidemiological perspective by Hosmer and Lemeshow (2000). For the purely statistical view the reader can turn to Cox and Snell (1989), to the much more general text of McCullagh and Nelder (1989), or to the treatise on categorical variables by Agresti (1996). The survey article on case–control studies by Breslow (1996) reflects the practice of medical and epidemiological research. For the econometric approach one should read the early survey article of Amemiya (1981) or, for a rigorous treatment, Chapter 9 of his textbook of 1985. Another text from econometrics is Maddala's wide ranging survey of a whole menagerie of related models (1983), or, with a much more theoretical slant, the book of Gourieroux (2000). An introductory text that conveys the flavour of the use of the logit model in the social sciences is Menard (1995). In the book by Franses and Paap (2001) these techniques are presented as part of a much broader range with a view to marketing applications. For early economic applications we refer to the handbook of discrete choice in transportation studies of Ben-Akiva and Lerman (1987); a recent record of the achievements in this tradition is McFadden's address of acceptance of the Nobel prize (2001). Pudney (1989) gives a rigorous survey of advanced micro-economics with equal attention to the theory and to empirical issues.

This list is by no means complete, and there are also further specializations within each field: the current literature in learned journals shows a separate development of econometrics in marketing and in finance. Readers should browse for themselves to keep abreast of these advances.

1.3 Program packages and a data set

Maximum likelihood estimation is by now the accepted standard method of estimation, and for the simple binary model and the standard multinomial model this is included as a simple routine in many program packages. One of the first to do so was the BMDP package (for BIOMEDICAL DATA PROCESSING), in the late 1970s, but by now logit and probit routines are a common part of general statistical packages like SAS, SPSS and STATA. They are also found in programs of econometric inspiration, like TSP (Time Series Processing), LIMDEP (specifically aimed at the wider class of Limited Dependent Variables) and E-VIEWS (so far binary models only). All these routines will provide coefficient estimates with their standard errors and a varying assortment of diagnostic statistics. Most

illustrations in this book have been produced by LOGITJD, an early forerunner of the logit module that is now part of the econometric program package PCGIVE. It is useless to give further technical details of this and other packages as they are continually being revised and updated.

Many programs, however, do not permit the immediate calculation of specific magnitudes like the Hosmer–Lemeshow goodness-of-fit test statistic, nor do they all readily permit the estimation of even quite mild variations of the standard model, like the logit model with allowance for misclassification or the nested logit model. While the standard routines can sometimes with great ingenuity be adapted to produce nonstandard results, analysts who wish to explore new avenues are much better off writing specific programs of their own in programming languages like GAUSS or OX. Suitable short-cut procedures for maximum likelihood estimation are available in either language and these can be embedded in programs that suit the particular wishes of the analyst. By now these programming languages are quite user-friendly, and the effort of learning to use them is amply rewarded by the freedom to trim the statistical analysis to one's personal tastes.

Many program packages and some textbooks come with sample data sets that readers can use for exercises. In this book we make repeated use of a data set on private car ownership of Dutch households in 1980, and since 2000 this has been made available to users of the PCGIVE program package. It is now available to all readers of this text, who can obtain it from the Cambridge University Press website. The address is http://publishing.cambridge.org/resources/0521815886/.

The data come from a household budget survey among Dutch household held in 1980 by the then Dutch Central Bureau of Statistics (now Statistics Netherlands). This survey recorded extensive and detailed information about income and expenditure of over 2800 households. The survey is unusual in that it contains a great deal of information about the cars at the disposal of the households, with a distinction between private cars and business cars that are primarily used for business and professional purposes. Although the data are by now out of date as a source of information about car ownership, they are well suited to demonstrate various models (as in this book) and for trying out new statistical techniques (as in a number of other studies). As a rule, Statistics Netherlands, like all statistical agencies, is very reluctant to release individual survey records to third parties, in view of the disclosure risk, that is the risk that individual respondents can be identified. In the present case

a welcome exception was made since the information, which is anyhow severely limited, is over 20 years old.

We use only a small fraction of the rich material of the budget survey, namely the information about car ownership, income, family size, urbanization and age. The data set consists of 2820 records, one for each household, with six variables. In the order of the dataset (which differs from the order of the analyses in this book) these are:

- PRIVATE CAR OWNERSHIP status in four categories, numbered from 0 to 3, namely *none, used, new* (for one used or new car respectively) and *more*. Private cars are all cars at the disposal of the households that are not business cars (see below).
- INC, income per equivalent adult in Dutch guilders per annum.
- SIZE, household size, measured by the number of equivalent adults. This is calculated by counting the first adult as 1, other adults as 0.7, and children as 0.5.
- AGE, the age of the head of household, measured by five-year classes, starting with the class 'below 20'.
- URBA, the degree of urbanization, measured on a six-point scale from countryside (1) to city (6).
- BUSCAR, a $(0, 1)$ dummy variable for the presence of a business car in the household. A business car is a car that is primarily used for business or professional purposes, regardless of whether it is paid for wholly or in part by the employer or whether its costs are tax-deductible.

In all analyses in this book we follow the common usage of taking the logarithm of income and, since it is closely related to this, of size as well, denoting the transformed variables by LINC and LSIZE.

When Windmeijer (1992) used this data set he identified one outlier: this is a household which owns a new private car while it has a very low income and disposes of a business car. In the dataset this is observation 817. In the calculations reported in this book it has not been removed from the sample.

1.4 Notation

I have aimed at a consistent use of various scripts and fonts while respecting established usage, but the resulting notation is not altogether uniform. It is also often incomplete in the sense that it can only be understood in the context in which it is used. A full classification with a

distinct notation for each type of expression and variable is so cumbersome that it would hinder understanding: as in all writing, completeness does not ensure clarity. I must therefore put my trust in the good sense of the reader. My main misgiving is that I found no room for a separate typographical distinction between random variables and their realization; in the end the distinction of boldface type was awarded to vectors and matrices as opposed to scalars.

The first broad distinction is that, as a rule, the Greek alphabet is used for unknown parameters and for other unobservables, such as disturbances, and Roman letters for everything else. The parameter β (and the vector $\boldsymbol{\beta}$) has the same connotation throughout, but λ and to a lesser extent α are used as needed and their meaning varies from one section to another. Greek letters are also occasionally employed for specific functions, like the normal distribution function.

In either alphabet there is a distinction between scalars and vectors or matrices. Scalars are usually designated by capital letters, without further distinction, but vectors and matrices are set in boldface, with lower-case letters for (column) vectors and capitals for matrices. I use a superscript ^{T} for transposition, the dash being exclusively reserved for differentiation.

The differentiation of vector functions of a scalar and of scalar functions of a vector habitually causes notational problems. In an expression like

$$\mathbf{y} = \mathbf{f}(X),$$

$\mathbf{y}$ is a column vector with elements that are functions of a scalar X. Differentiation will yield $\mathbf{f}'$, which is again a column vector. But in

$$Y = f(\mathbf{x}),$$

the scalar Y is a function of several arguments that have been arranged in the column vector $\mathbf{x}$. Differentiation with respect to (the elements of) $\mathbf{x}$ will yield a number of partial derivatives, which we arrange, by convention, in a row vector $\mathbf{f}'$. By the same logic, if $\mathbf{y}$ is an $r \times 1$ vector and $\mathbf{x}$ an $s \times 1$ vector, and if $\mathbf{y} = \mathbf{f}(\mathbf{x})$, $\mathbf{f}'$ is an $r \times s$ matrix of partial derivatives, and it should be named by a capital letter.

We use the standard terms of estimation theory and statistics, such as the expectation operator E, the variance of a scalar var and the variance-covariance matrix $\mathbf{V}$, applying them directly to the random variable to which they refer, as in EZ, var$Z$, and $\mathbf{Vz}$. The arguments on which these

(and other) expressions depend are indicated in parentheses. Thus

$$\mathbf{V}\mathbf{z}\,(\boldsymbol{\theta})$$

indicates that the variance matrix of $\mathbf{z}$ is a function of the parameter vector $\boldsymbol{\theta}$, while

$$\mathbf{V}\hat{\boldsymbol{\theta}}$$

is the variance matrix of $\hat{\boldsymbol{\theta}}$. This is an estimate of $\boldsymbol{\theta}$ as indicated by the circumflex or hat above it. Again,

$$\hat{\mathbf{V}}\mathbf{z} = \mathbf{V}\mathbf{z}(\hat{\boldsymbol{\theta}})$$

indicates how an estimated variance matrix is obtained.

Probabilities abound. We write

$$\Pr(Y_i = 1)$$

for the probability of an event, described within brackets; the suffix i denotes a particular trial or observation. We will then continue as in

$$\Pr(Y_i = 1) = P_i = P(X_i)$$

where P_i is a number between 0 and 1 and $P(X_i)$ the same probability as a function of X_i. Its complement is denoted by

$$Q_i = 1 - P_i, \quad Q(X_i) = 1 - P(X_i).$$

The vector $\mathbf{p}$ consists of a number of probabilities that usually sum to 1. At times we shall also make use of a different notation and use $\Pr(Y_i)$ for the probability of the observed value Y_i; in the binary case this is P_i if $Y_i = 1$ and Q_i if $Y_i = 0$, and it can be written as

$$\Pr(Y_i) = P_i^{Y_i} Q_i^{1-Y_i}.$$

Equations are numbered by chapter, but sparingly, and only if they are referred to elsewhere in the text.

2

The binary model

Binary discrete probability models describe the relation between one or more continuous determining variables and a single attribute. These simple models, probit and logit alike, account for a very large number of practical applications in a wide variety of disciplines, from the life sciences to marketing. In this chapter we discuss their background, their main properties, their justification and their use. Section 2.3 presents the latent variable regression model that is used as the standard derivation throughout this book. Although the emphasis is on the logit model, much of the discussion applies to the probit model as well.

2.1 The logit model for a single attribute

The logit model has evolved independently in various disciplines. One of its roots lies in the analysis of biological experiments, where it came in as an alternative to the probit model. If samples of insects are exposed to an insecticide at various levels of concentration, the proportion killed varies with the dosage. For a single animal this is an experiment with a determinate, continuously variable stimulus and an uncertain or random discrete response, viz. survival or death. The same scheme applies to patients who are given the same treatment with varying intensity, and who do or do not recover, or to consumer households with different income levels who respond to this incentive by owning or not owning a car, or by adopting or eschewing some other expensive habit. Married women may or may not take up paid employment and their choice is influenced by family circumstances and potential earnings; students' choices among options of further education are affected by their earlier performance. The class of phenomena or models thus loosely defined is variously referred to in the biological literature as *quantal variables* or as

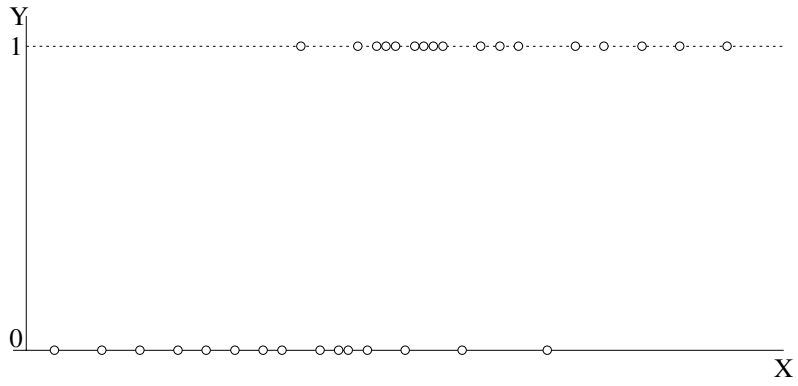

Fig. 2.1. Car ownership as a function of income in a sample of households.

stimulus and response models, in psychology and economics as *discrete choice*, and in econometrics as *qualitative* or *limited dependent variables*.

We examine the car ownership example more closely. The relation of household car ownership to household income can be observed in a household survey. The independent or determining variable is household income, which is continuous, and the dependent variable or *outcome* is ownership status, which is a discrete variable. For a single attribute (like car ownership as such) the outcome Y is a scalar which can take only two values, conventionally assigned the values 0 and 1. The event $Y = 1$ is habitually designated as a *success* of the experiment, and $Y = 0$ as a *failure*, regardless of their nature. In the present case we have

$$Y_i = 1 \text{ if household } i \text{ owns a car,}$$
$$Y_i = 0 \text{ otherwise.}$$

When these values are plotted against income X_i for a sample of households we obtain the scatter diagram of Figure 2.1.

A regression line could be fitted to these data by the usual Ordinary Least Squares (OLS) technique, but the underlying model that makes sense of this exercise does not apply.† One may of course still *define* a linear relationship, and make it hold identically by the introduction of an additive disturbance ε_i, as in

$$Y_i = \alpha + \beta X_i + \varepsilon_i.$$

† There is no short-cut formula for the OLS regression of Y on X. If X were regressed on Y, however, the regression line would pass through the mean incomes of car-owners and of non-car-owners.