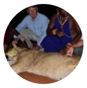
Resource Selection by Animals: Statistical Design and Analysis for Field Studies

Book *in* Journal of Animal Ecology · January 2002

DOI: 10.2307/5247

Bryan F J Manly
Manly-Biostatistics Limited



Lyman Mcdonald
Western EcoSystems Technology Inc.

Dana L. Thomas
University of Alaska System

Wallace P. Erickson
Western EcoSystems Technology Inc.

Resource Selection by Animals

Statistical Design and Analysis for Field Studies

Second Edition

Bryan F.J. Manly[1], Lyman L. McDonald[1], Dana L. Thomas[2],
Trent L. McDonald[1] and Wallace Erickson[1]

[1]Western EcoSystems Technology Inc., Cheyenne, Wyoming, USA
[2]Department of Mathematical Sciences, University of Alaska, Fairbanks, USA

TABLE OF CONTENTS

PREFACE

We have written this book as a guide to the design and analysis of field studies of resource selection, concentrating primarily on statistical aspects of the comparison of the use and availability of resources of different types. Our intended audience is field ecologists in general and, in particular, wildlife biologists who are attempting to measure the extent to which real animal populations are selective in their choice of food and habitat. As such, we have made no attempt to address those aspects of theoretical ecology that are concerned with how animals might choose their resources if they acted in an optimal manner.

The book is based on the concept of a resource selection function (RSF), where this is a function of characteristics measured on resource units such that its value for a unit is proportional to the probability of that unit being used. We argue that this concept leads to a unified theory for the analysis and interpretation of data on resource selection and can replace many ad hoc statistical methods that have been used in the past.

This second edition differs from the first edition in several ways. There is much new material including the uses of the Akaike information criterion (AIC) for model selection, a considerable rearrangement of the material on logistic regression, and completely new chapters on discrete choice models, the analysis of data from geographical information systems (GIS), approaches to studying resource selection other than using RSFs, the uses of RSFs for assessing the risks involved to animals if habitats are changed or estimating population sizes, and computer software that can be used to carry out the calculations for RSF analyses.

The book has the following structure: Chapters 1 and 2 provide a review of statistical methods that have been used in the past to study resource selection and some examples of how different types of study can be analysed in terms of RSFs; Chapter 3 gives a brief introduction to a range of statistical techniques that are used for modelling data in the later chapters; Chapter 4 covers the particularly important special case where the resources being studied are in categories of habitat or food types; Chapter 5 covers the use of logistic regression for estimating a RSF; Chapter 6 covers situations where samples of used or unused resource units are taken over a period of time; Chapter 7 covers the uses of log-linear modelling; Chapter 8 (completely new) covers the uses of discrete choice models; Chapter 9 (completely new) covers the use of GIS data; Chapter 10 covers the uses of discriminant function analysis as an alternative to logistic regression for estimating RSFs; Chapter 11 gives some suggestions on how to analyse data when the amount of use of resource units is recorded rather than simply whether or not they are used; Chapter 12 covers methods such as compositional analysis that do not use RSFs for studying resource selection; Chapter 13 covers the use of RSFs for assessing the risks involved in habitat changes, and for estimating population sizes; and finally Chapter 14 discusses the software that is available to carry out RSF calculations. The book requires much special notation, and readers may find the list that follows this preface to be useful in keeping track of this notation.

We acknowledge the help of many of our colleagues in writing this book. Particularly, we acknowledge Dan Reed who assisted with early versions of Chapter 4, Ed Arnett, C.R. Bantock, W.F. Harris, and Tom Ryder who provided us with raw data from their studies, Diana Craig who read over the final manuscript for the first edition of this book, and Ryan Nielson and Kimberley Bay who did the same for the second

edition of the book. Many users of the first edition of the book also made comments that have helped us to improve the second edition. However, any errors or omissions in this second edition are entirely our responsibility.

BFJM, LLMc, TLM, DLT and WE
Laramie, February, 2002

LIST OF SYMBOLS

I	Number of categories of resource units
m_+	Size of a sample of available resource units
m_i	Number of available units in category i in a sample of available resource units
π_i	Proportion of the population of available units that are in category i
$\hat{\pi}_i$	m_i/m_+, sample proportion of available units in category i
u_+	Size of a sample of used resource units
u_i	Number of units in category i in a sample of used units
o_i	u_i / u_+, sample proportion of used units in category i
f_i	Proportion of the available category i items that are used.
$\hat{w}_i$	$o_i / \hat{\pi}_i$, the forage ratio (also called the selection ratio and the preference index)
E_i	$(o_i - \hat{\pi}_i) / (o_i + \hat{\pi}_i)$, Ivlev's electivity index
L_i	$o_i - \hat{\pi}_i$, Strauss' linear selection index
Q_i	$[o_i (1 - \hat{\pi}_i)] / [\hat{\pi}_i (1 - o_i)]$, Jacobs' first selection index
D_i	$(o_i - \hat{\pi}_i) / (o_i + \hat{\pi}_i - 2o_i \hat{\pi}_i)$, Jacobs' second selection index
α_i	$(o_i / \hat{\pi}_i) / \Sigma(o_i / \pi_i)$, Chesson's selection index
B_{i1}	$(u_i / m_i) / \Sigma(u_i / m_i)$, Manly's standardized selection index with used resource units replenished
B_{2i}	$\log(1 - f_i) / \Sigma\log(1 - f_i)$, Manly's standardized selection index with used resource units not replenished
W_i	$f_i / \Sigma f_j$, Vanderploeg and Scavia's first selection index
E^*_i	$(W_i - 1/I)/(W_i + 1/I)$, Vanderploeg and Scavia's second selection index
SI	$\mathrm{Max}[(\Sigma m_i / m_+) - (\Sigma o_i / o_+)]$, Rondorff *et al.*'s selection intensity for continuous data

$w^*(\mathbf{x})$	Resource selection probability function for a single period of selection where resource units are characterized by their values $\mathbf{x} = (x_1,x_2,...,x_p)$ for variables X_1 to X_p
$w(\mathbf{x})$	Resource selection function, which is $w^*(\mathbf{x})$ multiplied by an unknown positive constant
$w^*(\mathbf{x},t)$	Resource selection probability function for selection from time 0 to time t
$w(\mathbf{x},t)$	Resource selection function for selection from time 0 to time t
$\phi^*(\mathbf{x},t)$	Probability that a resource unit is not used ('survives') in the time interval 0 to t
$\phi(\mathbf{x},t)$	$\phi^*(\mathbf{x},t)$ multiplied by an unknown positive constant
ß_i	Coefficient of the variable X_i in a resource selection or survival function
$\hat{\text{ß}}_i$	Estimate of ß_i
$se(\hat{\text{ß}}_i)$	Standard error of the estimator of ß_i
X_L^2	Log-likelihood chi-squared statistic for measuring the goodness of fit of a model for data
X_P^2	Pearson chi-squared statistic for measuring the goodness of fit of a model for count data
R_i	Standardized residual for the ith observation in a set of data
μ_x	Population mean of the random variable X
s_X	Sample standard deviation of the variable X
r_{XY}	Sample Pearson correlation between variables X and Y
var(X)	Variance of the random variable X
cov(X,Y)	Covariance of random variables X and Y
A_+	Size of a finite population of available resource units
A_i	Number of the A_+ units that are in category i
w^*_i	Proportion of the A_i available resource units that are used
$\hat{w}^*_i$	An estimate of w^*_i
z_α	Value that is exceeded with probability α by a standard normal random variable
E(X)	Expected (mean) value of a random variable X

u_{ij}	Number of category i resource units used by animal j
u_{i+}	Number of category i resource units used by all animals
u_{+j}	Number of resource units used by animal j
u_{++}	Total number of units used by all animals
$\hat{w}_{ij}$	Selection ratio for the jth animal and the ith resource category
π_{ij}	Proportion of the resources available to animal j that are in category i
$\hat{\pi}_{ij}$	Sample estimate of π_{ij}
S	Number of selection episodes
Θ_{ij}	Probability that a type i resource unit will be used in the time interval from t_{j-1} to t_j
P_a	Probability of sampling an available resource unit
$P_u(t)$	Probability of sampling a used resource unit at time t
$P_{\bar{u}}(t)$	Probability of sampling an unused resource unit at time t
$u_i(t)$	Number of resource units of type i found in a sample of used units taken at time t
$\bar{u}_i(t)$	Number of resource units of type i found in a sample of unused units taken at time t
P_u	Probability of sampling a used resource unit when only one sample of these units is taken
$P_{\bar{u}}$	Probability of sampling an unused resource unit when only one sample of these units is taken
$\boldsymbol{\mu}$	Population mean vector
$\boldsymbol{\Sigma}$	Population covariance matrix
$\Omega(\mathbf{x})$	A function such that selecting individuals with measurements $\mathbf{x} = (x_1,x_2,...,x_p)$ with probabilities $\Omega(\mathbf{x})$ from one multivariate population will produce a second population
ρ	Proportion of available units that are used

CHAPTER 1

INTRODUCTION TO RESOURCE SELECTION STUDIES

This chapter provides the motivation for the study of resource selection, defines various terms used in the remainder of the book, discusses study designs, and gives a historical perspective on the statistical evaluation of resource selection.

1.1 Motivation and Definitions

Adequate quantities of usable resources are necessary to sustain animal populations. Therefore, biologists often identify resources used by animals and document the availability of those resources. The need for such documentation is especially critical in efforts to preserve endangered species and manage exploited populations. Determining which resources are selected more often than others is of particular interest because it provides fundamental information about the nature of animals and how they meet their requirements for survival.

Differential resource selection is one of the principal relationships which permit species to coexist (Rosenweig, 1981). It is often assumed that a species will select resources that are best able to satisfy its life requirements, and that high quality resources will be selected more than low quality ones. The availability of various resources is not generally uniform in nature, and use may change as availability changes. Therefore, used resources should be compared to available (or unused) resources in order to reach valid conclusions concerning resource selection. When resources are used disproportionately to their availability, use is said to be selective.

The usage of a resource is defined as that quantity of the resource that is utilized by an animal (or population of animals) in a fixed period of time. The availability of a resource is the quantity accessible to the animal (or population of animals) during that same period of time. It is possible to distinguish between availability and abundance by defining the latter as the quantity of the resource in the environment. Although selection and preference are often used as synonyms in the literature, here they are defined differently: selection is the process in which an animal chooses a resource, and preference is the likelihood that a resource will be selected if offered on an equal basis with others (Johnson, 1980).

Resource selection occurs in a hierarchical fashion from the geographic range of a species, to individual home range within a geographic range, to use of general features (habitats) within the home range, to the selection of particular elements (food items) within the general features (or feeding site). The criteria for selection may be different at each level (Johnson, 1980; Wiens, 1981; Orians and Wittenberger, 1991) so that, when making inferences, researchers studying selection must keep in mind the level being studied. The specification of a level at which to study selection is closely associated with the specification of availability. As there is no single correct level to study, multiscale studies of selection have become increasingly more common (Levin, 1992; Otis, 1997).

Most commonly, selection studies deal with food or habitat selection. Food selection may be among various prey species or among sizes, colors, shapes, components, etc. of the same species. Habitat selection may be among various discrete habitat categories (e.g., open field, forest, rock outcropping) or among a continuous array of habitat attributes such as shrub density, percentage cover, distance to water, canopy height, etc. Thus, the variables observed in a selection study may be discrete, or continuous or some combination of the two.

Considerable variation exists in the motivation for conducting resource selection studies. For example, there is sometimes a need to provide quantitative information that is indicative of the long-term resource requirements of a population, as for the debate concerning whether old growth forest is vital to the continued existence of the spotted owl (*Strix occidentalis*) in the Pacific Northwest of the United States (Laymon *et al*., 1985; Forsman *et al*., 1984) or for the existence of black-tailed deer (*Odocoileus hemionus*) on Admiralty Island, Alaska (Schoen and Kirchhoff, 1985). Resource selection studies are commonly carried out for this reason. However, it should be noted that a resource item may be highly favoured but if it is difficult to find then it cannot be utilized much. Conversely, if resources which are less favoured are the only ones available then they may of necessity comprise a large proportion of those used (Petrides, 1975, Chapter 8; White and Garrott, 1990). Therefore, researchers should proceed cautiously when using the results of selection studies to determine the relative importance of resources.

Resource selection studies are often used in fisheries. Gear selectivity must be accounted for in estimating abundance, describing population size structures and evaluating mortality rates (Hamley and Regier, 1973; Beamsderfer and Rieman, 1988; Millar and Walsh, 1992; Walsh et al., 1992). These studies fall under the present theory because sampling gear is sometimes 'biased' in that the items available in the population are not selected in proportion to their availability. In addition, researchers often investigate river characteristics such as water velocity, depth, or substrate selected by fish. Examples are Parsons and Hubert's (1988) study of spawning site selection by kokanees (*Onchorhynchus nerka*) and the study of Belaud *et al*. (1989) of brown trout (*Salmo trutta*) stream site selection. The latter studies provide information from in-stream flow incremental methods (Bovee,1981) to evaluate habitat use.

Another use of resource selection studies is as part of modeling and projecting the impact of habitat change. Under certain assumptions the ratio of animal densities equals the ratio of resource availabilities for any two habitats at equilibrium (Fagen, 1988). These relationships are used to calculate relative habitat values from habitat selection studies and to define hypothetical carrying capacity curves based on habitat values under various conditions. Such an approach was used by Schoen and Kirchhoff (1985) to model deer production after logging. This approach is quite controversial, as evidenced by the differences of opinion stated by Fagen (1988), van Horne (1983) and Hobbs and Hanley (1990). Other studies utilizing selectivity to evaluate the effects of human disturbance include the study of Edge *et al.* (1987) of elk (*Cervus elaphus*) habitat selection, Edwards and Collopy's (1988) study of osprey (*Pandion haliaetus*) nest trees, and Fair and Henke's (1997) study of Texas horned lizards (*Phrynosoma cornutum*). In such studies undisturbed sites often serve to provide baseline information which helps managers to evaluate the impact of man on animals.

The process of natural selection can occur when resource selectivity results in successful (e.g., breeding) and unsuccessful individuals. Early work on habitat selection was closely associated with ideas on speciation, niche theory, and range expansion. Although habitat selection in itself is no longer generally considered a major factor in speciation for birds or mammals, some entomologists have suggested that food and oviposition selection play a significant role in evolution and speciation among some insects (Feder *et al.*, 1990; Thompson, 1988). In addition, MacCallum *et al.* (1998) demonstrated an effect of habitat selection on the genetic structure of a hybrid zone between toads (*Bombina* sp.), Clark and Shutler (1999) tested for natural selection based on nest site selection by ducks (*Anas* sp.) and Danchin *et al.* (1998) showed an association between reproductive success and breeding habitat selection for Black-legged Kittiwakes (*Rissa tridactyla*). The transmission of food preference between individuals and across generations has been demonstrated for the rabbit (Bilko *et al.* 1993) and the potential of avian selection to exert natural selection pressure on color-dimorphic fruits has been studied (Gervais *et al.* 1999).

Other situations where resource selection studies have a major role include the evaluation of the effect of domestic animals on wild animal forage, such as Bowyer and Bleich's (1984) study of the effects of cattle grazing on selected habitats of southern mule deer (*Odoncoileus hemionus*); the evaluation of the Habitat Suitability Indices (HSI) that are used by the United States Fish and Wildlife Service (1981) to characterize habitat quality for selected species, such as the test devised by Thomasma *et al.* (1991) of the HSI for the fisher (*Martes pennanti*); estimating parameters in optimal foraging and diet prediction models (Pyke *et al.*, 1977; Nelson, 1978); and the determination of the specific prey cues that result in predation, such as Zaret and Kerfoot's (1975) study of fish predation on *Bosmina longirostris* to determine whether selection was related to body-size or visibility. A recent use of resource selection study was in the estimation of population size using resource selection functions (Boyce and McDonald, 1999).

Many factors contribute to resource selection (Peek, 1986). These factors include population density, competition with other species, natural selection, the chemical composition or texture of forage, heredity, predation, habitat patch size and inter-patch distances. Numerous models and theories of resource selection have been proposed that incorporate subsets of these factors. These include foraging models (Emlen, 1966; Rapport and Turner, 1970; Werner and Hall, 1974; Ellis *et al.*, 1976; Pyke *et al.*, 1977; Rapport, 1980; Nudds, 1980, 1982; Belovsky *et al.*, 1989) and habitat selection models (Bryant, 1973; Rosenzweig, 1981;Whitham, 1980). We do not delve into such models in this book. Instead, we restrict attention to statistical techniques for the detection and measurement of the degree to which a resource is selected or avoided.

The reasons why a particular resource is selected or avoided is not directly revealed by the estimation of the amount of use or avoidance. It is not possible to be certain, for example, that any food item is distasteful just because it is rarely used. It may be the case, but there is no way of knowing from availability and use data alone that this is true. However, if we learn that there is selection for or against a resource then this is a starting point for further in-depth study (Petrides, 1975).

1.2 The Data for Resource Selection Studies

Throughout this book it is assumed that the resource being studied can be considered to consist of a number of discrete resource units and that data are collected from a census or sampling of these units. The complete set of these units will be referred to as the universe or population of available resource units. This division of the resource into units will occur naturally in some cases, such as when the resource units are individual prey items. At other times the division must be imposed by the researcher. This would be the case, for example, if a study area is divided into quadrats that may or may not be used by the individuals in an animal population. Such a division of the study are into grid cells has been recommended by Porter and Church (1987).

Some resource units such as food items can be used only once, while others, such as habitat units, may be used repeatedly. In either case there may be resource units that are unused by a particular definition over the period of a study. We can therefore partition the population of resource units into two sets consisting of used units and unused units. Resource selection may be detected and measured by comparing any two of the three possible sets (used, unused, and available) of resource units.

The methods discussed in this text address an individual's or population's use or non use of discrete resource units based on some categorization or attributes of the units. However, selectivity can be assessed by examining the amount of resource use. Controlled feeding preference experiments often measure the amount eaten of each food resource made available (e.g., Edwards *et al.*, 1994; Wang and Provenza, 1996; Stepto and Cook, 1996). Study designs and subsequent analyses for such studies continue to be refined, e.g., see Rodgers (1990), Roa (1992), and Manly (1993, 1995). Habitat selection has been studied in the field by measuring the amount of time spent by individuals in each of the available habitats (Spitz and Janeau, 1995; Durbin, 1998) or through a count of the number of seasons a plot is used. In chapter 11 several special cases concerning the amount of use of resource units are discussed.

Studies sometimes include the identification of individual animals. For example, Gionfriddo and Krasman (1986) used individually identified radio-collared sheep (*Ovis canadensis mexicana*) to study their summer habitat use. However, individual moose (*Alces alces*) were not identified by Neu *et al.* (1974) when they examined the use of areas in each of four burn categories during several aerial surveys. Similarly, studies on food habits may use collected individuals, as in Hohman's (1985) study of ring-necked ducks (*Aythya collaris*), or methods that do not distinguish individuals, such as the use by Keating *et al.* (1985) of bighorn sheep (*Ovis canadensis canadensis*) pellets to study winter food habits.

Studies may involve classifying resource units into categories, or measures of specific variables characteristic of those units may be obtained. For example, Murphy *et al.* (1985) collected habitat use data for white-tailed deer (*Odoncoileus virginianus*) by classifying radio locations into one of six habitat types. The proportion of radio locations in each habitat type was then compared to the relative availability of the respective habitat type in the study area. Alternatively, Dunn and Braun (1986) examined the selection of habitat by juvenile sage grouse (*Centocerus urophasianus*) by comparing habitat attribute data such as the shrub density and the distance to another cover type for radio locations and random sites.

A wide variety of methods can be used to collect data. For example, resource availability may be evaluated from a map (Neu *et al.*, 1974), digitally assessed using a geographic information system (McClean *et al.*, 1998; Mladenoff *et al.*, 1999) or sampled by randomly selecting sites (Marcum and Loftsgaarden, 1980). In summary, then it can be said that the data used to evaluate resource selection may be collected through a census or by sampling with one or more procedures, may be categorical or continuous, and are generally multivariate.

1.3 Sampling Protocols and Study Designs

Given the decision to study resource selection for a particular animal species, the next decision is to determine the scale or scales of selection to focus upon. In making this decision the researcher must use what is known about the biology of the animal. For example, if the animal being studied is territorial then selection is commonly studied on a different scale from what would be used for a non-territorial animal (Johnson, 1980). In their study of gray wolf (*Canis lupus*) Mladenoff *et al.* (1999) noted that an area colonized by one pack is unavailable to other packs and adjusted their measure of availability by subtracting the inhabited areas from the total area available. Also, it is important to know whether animals tend to return to some central location (as in nest site or bed site selection). If they do, then this has an impact on what resources can be assumed to be available at the scale of interest and ignoring this tendency may bias analyses (Rosenberg and McKelvey, 1999).

In some cases determining the scale of selection will be a major goal of a study and, as noted before, researchers should consider studying selection at more than one scale. For example, Danell *et al.*, (1991) investigated whether moose select forage at the individual tree level or on patches of trees by providing moose with access to artificial stands of trees for which the mixture of forage species and their spacings were controlled. Field studies that cannot manipulate resources in this way may still examine scale questions by measuring availabilities at various distances from used sites (Larsen and Bock, 1986; McClean *et al.*, 1998).

Resource selection is often affected by season, sex, age class, behavioral activity, and daily activity pattern of the animal studied. For example, if radio locations are used to assess habitat use and most locations are for inactive animals, foraging habitat will likely be underestimated (Palomares and Delibes, 1992). Similarly, if resource use or availability changes across seasons, then the study should either focus on one season or be designed to address each of the seasons of interest (Schooley, 1994; McKnight and Hepp, 1998). Focusing on a single age class may significantly reduce the variability in analyzing selection (McClean *et al.,* 1998). It is important to decide upon the timing, subpopulation, and activity to address in a resource selection study. Pooling information across times, subpopulations or activities may result in erroneous inferences.

Another important decision in designing a resource selection study is the choice of the study area and its boundaries. This choice may have a significant impact on the results of subsequent data analysis, especially if resource units are arranged in an aggregated pattern (Porter and Church, 1987). Depending upon the scale of selection studied and the sampling design used, the choice of a study area is often closely associated with the specification of availability of resources (Otis, 1998). When choosing a study area the researcher must consider the distribution of resource units, the scale of selection studied, what is truly available to the animals, and manpower and budget constraints for sampling.

As noted in the previous section, resource selection may be detected and measured by comparing any two of the three possible sets (used, unused, and available) of resource units. On this basis, the following three common sampling protocols (SP-A, SP-B, and SP-C) can be identified depending on the two sets measured:

Protoco l	What is sampled or censussed
SP-A	Available units are either randomly sampled or censussed and used resource units are randomly sampled.
SP-B	Available resource units are either randomly sampled or censussed and a random sample of unused resource units is taken.
SP-C	Unused resource units and used resource units are independently sampled.

In addition to the three sampling protocols shown above there are cases where a complete census of used and unused resource units can be made. The evidence for selection can then come from considering the variation implied by a stochastic model for the selection process rather than from the consideration of random sampling variation.

Three general study designs for evaluating selection have been identified in the literature (Thomas and Taylor, 1990). These study designs differ with respect to the level at which resource use and availability are measured; at the population level or for each animal. Each of the three sampling protocols , SP-A, SP-B, or SP-C, may be implemented for each of the sampling designs and the particular combination of design and protocol used to gather data determines some of the underlying assumptions required for a subsequent analysis (e.g., whether the availability of each resource is the same for all animals; see Section 1.5). The three study designs are defined as follows:

Design I

With this design, measurements are made at the population level. Used, unused, or available resource units are sampled or censussed for the entire study area and for the collection of all animals in the study area. Individual animals are not identified. Examples of this design are:

(1) Aerial or ground surveys such as line transects are used to classify animal locations into resource categories (habitat or forage types). Maps, aerial photography, or geographic information system (GIS) information are used to census availability or random plots are sampled to estimate availability. The percentage use for each category is then compared to its respective availability to evaluate selection. For example, Stinnett and Klebenow (1986) examined cover-type selection of California quail (*Callipepla californica*) by classifying flushes observed during ground surveys into cover types. Maps and aerial photography were partitioned into the respective cover types to evaluate availability. Haney and Solow (1992) used this design to study resource selection by marine birds. They compared the relative amount of line transect sampling conducted in four habitats (availabilty) to the proportion of bird numbers in each of these habitats (use). This approach has also been used to study diet selection by comparing the percentage of food types (vegetation or prey species) that were consumed (browsed or preyed upon) to the percentage available on plots or transects randomly located in the study area (e.g., Lagory *et al.*, 1985; Karanth and Sunquist, 1995).

(2) Randomly located plots are classified as used or unused on the basis of signs such as pellets or tracks. Plot attributes such as the percentage of vegetation cover, the shrub density, the distance to water, etc. are measured for each sampled plot, and used and unused sites compared to evaluate selection. In fisheries studies, nets are set in a sample of locations and the number of fish in each location measures the use of that location. Locations are classified into categories such as shallow or deep, with a fast or slow current, or attributes are measured at each sampled location. Available or unused locations are then compared with used locations.

Design II

With this design, individual animals are identified and the use of resources is measured for each, but availability is measured at the population level. Some examples are :

(1) A sample of animals is collected or otherwise identified via neck collars, radioactive tracing, ear tags, radio transmitters, or colored leg bands, and the resource units used by each animal are recorded. Each used resource unit may be classified into a category or some attributes are measured on it. The collection of

available resource units is sampled or censussed (e.g., using aerial photography, geographic information systems, or maps) for the entire study area.

(2) Individual animals are identified and their home range determined. The proportions of resource types in home ranges (as determined by sampling or the partitioning of maps) are compared to the proportions in the entire study area. For example, Roy and Dorrance (1985) compared habitat availabilities within coyote (*Canis latrans*) home ranges with the availabilities in the entire study area. This approach has been criticized by White and Garrott (1990, p. 201) on the grounds that the home range represents a prior selection of habitat.

(3) The relative number of relocations of a marked individual in each habitat type is compared to the proportion of that habitat in the study area. These data are then replicated for each marked animal. This method has been used for example in studies addressing habitat selection by sheep (Gionfriddo and Krausman, 1986) and elk (McCorquodale *et al.*, 1986).

(4) Foraging ecology studies often compare stomach or faecal contents of identified individuals with random samples of available food from the entire study area (Prevett *et al.*, 1985).

When only a single observation of use is made for each animal there is no way of distinguishing between design I and design II. For example, this is the case in studies of nest or bed site selection (Peterson, 1990; Huegel *et al.*, 1986) when only one site is identified for each animal. Thus, design II studies with only one observation of use per animal are a special case of design I studies.

Design III

With this design, individuals are identified or collected as in design II, and at least two of the sets (used resource units, unused resource units, available resource units) are sampled or censussed for each animal. Some examples are:

(1) The animals in a sample are radio-collared, and the relocations of an animal identify used resource units for that animal. The used units are either classified into types or attributes are measured on each. The collection of available or unused resource units within an animal's home range is sampled or censussed.

(2) A sample of animals is observed feeding for some fixed time period. Individuals are then collected and stomach analysis performed on each. Prey items are sorted by type (e.g., species or color), or measurements (length, volume) are taken on each prey item. A sample of available (or unused) prey items is taken at the feeding site of each animal collected.

(3) The home range or territory of an individual animal is determined, and the use and availability of resources are compared within that area. For example, Rolley and Warde (1985) used radio-telemetry to identify home ranges and the use of various vegetation types by individual bobcats (*Flis rufus*). LANDSAT data were then used to estimate the relative proportions of each vegetation type within each home range. Foraging studies using this design typically collect individuals for gut analyses and measure food availability at each collection site (Hohman, 1985).

1.4 Comparison of Designs

The choice of study design affects the cost of the resource selection study. Studies utilizing Design I tend to be inexpensive compared to Designs II and III because individual animals are not captured, collected, or relocated. Similarly, studies which categorize resource units into types rather than measuring attributes of units are less labor intensive.

Designs II and III each involve uniquely identified individuals. Therefore, making inferences for the population of animals requires that we assume that the animals identified are a random sample from that population. The sample design then becomes a several stage process: selection of a sample of animals, selection of samples of used and available resource units for each animal, and (often) subsampling of the chosen resource units.

Designs II and III allow an analysis of resource selection for each individual animal. Hence estimates calculated from observations on individual animals may be used to estimate parameters for the population of animals and estimates of variability of these estimates. This approach, which is called first and second stage analysis by Cox and Hinkley (1974), has been recommended for analysis of resource selection data by White and Garrott (1990) and was used by Porter and Labisky (1986) in their study of foraging habitat selection of red-cockaded woodpeckers (*Picoides borealis*) with pairs and groups of birds as first stage units. Ecologists have also argued the need for an 'individual-based' view (Judson, 1994; Sherratt and MacDougall, 1995).

The advantages of this type of approach are:

- The observations on any one animal may be time-dependent. For example, the independence of several radio relocations depends on the time between these relocations and the animal's diurnal behavior pattern. Similarly, in food studies the selection of consecutive prey items may be dependent. Therefore, as a general rule it is better to estimate sampling variances and test hypotheses using variation between animals rather than the variation between observations on one animal. In effect this means that inferences become design-based (relying on random sampling of animals) rather than model-based (relying on the assumed statistical model being correct). The advantage of design-based inference over model-based inference is that the former is far less dependent on the assumed statistical model being correct.

- It allows estimation techniques that are applicable at the population level for design I to be applied to individuals with designs II and III.

- The variation among individuals may be examined to determine, for example, whether gender or age differences between animals occur, and to identify individual animals that are unusual with respect to their selectivity.

These advantages may also be obtained by treating groups of animals as first-stage units providing that the groups used in the study are a random sample from the population of groups.

In addition to the field study designs described above, controlled experimental studies have also been conducted to evaluate resource selectivity. This approach for the evaluation of forage selection is fairly common, with examples being the studies of Colgon and Smith (1985) and Hohf *et al.*, (1987). Habitat manipulation studies are less common but can be implemented (White and Trudell, 1980; Munro and Rounds, 1985;

Fair and Henke, 1997). Our emphasis in this book is on observational studies; however, we do provide several examples of experimental studies and their analysis.

In some study designs the samples of used and available resource units are not independent. For example, in Hohman's (1985) study of ring-necked ducks and McKnight and Hepp's (1998) study of gadwalls (*Myriophyllum spicatum*), samples of available food items were taken specifically at the feeding site where birds were collected after they had been observed to feed for a short period. The use and availability samples are then paired by location. Similarly, in the study of Huegel et at. (1986) of bed site selection by white-tailed deer fawns, habitat characteristics were measured at the bed site and at an adjacent plot. Again, the measurements are paired by location.

Clearly, the analysis of data should take into account whether the samples taken of different types of resource unit (used, unused and available) are independent or paired. Many statistical techniques commonly employed to evaluate resource selection require the assumption that samples being compared are independent. If in fact the samples are paired then this assumption is violated and the results of such an analysis are suspect.

1.5 Indices of Selection

Early researchers simply described their findings on resource use and availability. Initially, quantifying resource selection occurred in the analysis of food studies. For example, see Kalmback's (1934) review of stomach analysis publications. Most of these early studies only indicated the number of animals consuming each prey item and the percentage consumption. Variability among animals and locations made it difficult for researchers to compare their results because differences were assessed subjectively.

Scott (1920) is commonly cited (e.g., Cock, 1978; Pearre, 1982) as the first author to quantify selection. He divided the average number of each prey species per fish stomach per unit of time by the number found in plankton hauls per unit area. Thus, this first index used the ratio of the rate of consumption of a prey type to the density at which is was present.

The commonly employed ratio of percentage use divided by percentage available appears to have been independently proposed by Savage (1931) (Cock, 1978), three Russian researchers including A. A. Shorygin (Ivlev, 1961), and by Hess and Swartz (1940) who referred to it as the forage ratio. Ivlev (1961) found the 0 to infinity range of the forage ratio to be cumbersome and proposed an alternative electivity index with the range -1 to +1. Numerous other indices have also been proposed, as shown in Table 1.1.

Other authors have suggested various special purpose indices related to selectivity. Pearre (1982) gave two selectivity indices based on chi-squared statistics that compare one resource at a time to all other resources combined. Holmes and Robinson (1981) used a 'tree preference index' to study tree species preferences of foraging insectivorous birds, by summing percentage deviations of bird foraging frequencies in various tree species from relative importance values of the trees. The relative importance values were calculated from the densities, frequencies of occurrence and basal areas per hectare for individuals of each tree species greater than 2.5 cm in diameter at breast height. In assessing the food selection of ungulates, Owen-Smith and Cooper (1987) proposed two indices of acceptance: site-based acceptance, calculated as the number of 30-minute intervals during which the resource was used, divided by the number of 30-minute intervals during which the number of individual plants of type I eaten, divided by the number of plants of that type encountered within neck reach. In studying habitat selection of otter (*Lutra lutra*) Durbin (1998) formed two preference indices based on the time spent in resource units, the length or area of an otters range, and the length or area of sections of rivers (resource units). One index used length of river sections and otter ranges because riparian habitats used by otters were fairly linear in nature while the other index used area.

*Table 1.1 Commonly sited indices of selectivity**

References	Index
Savage (1931)	$w_i = o_i / \hat{\pi}_i$, the forage ratio
Ivlev (1961)	$E_i = (o_i - \hat{\pi}_i) / (o_i + \hat{\pi}_i)$, Ivlev's electivity index
Strauss (1979), Jolicoeur and Brunel (1966), Ready *et al.* (1985)	$L_i = o_i - \hat{\pi}_i$, Strauss' linear index
Jacobs (1974)	$Q_i = [o_i (1 - \hat{\pi}_i)] / [\hat{\pi}_i (1 - o_i)]$ and $D_i = (o_i - \hat{\pi}_i) / (o_i + \hat{\pi}_i - 2 o_i \hat{\pi}_i)$ Also $\log (Q_i)$ and $\log (D_i)$
Chesson (1978) and Paloheimo (1979)	$\alpha_i = (o_i / \hat{\pi}_i) / \Sigma(o_i / \hat{\pi}_i)$, Chesson's index
Manly *et al.* (1972) and Manly (1973, 1974)	$B_{i1} = (u_i / m_i) / \Sigma(u_i / m_i)$ and $B_{i2} = \log (1 - f_i) / \Sigma \log (1 - f_i)$
Vanderploeg and Scavia (1979a, b)	$W_i = f_i / \Sigma f_i$ and $E^4_i = (W_i - 1/I) / (W_i + 1/I)$
Bowyer and Bleich (1984)	Importance $= o_i \hat{\pi}_i$
Rondorff *et al.* (1990)	$SI = Max[(\Sigma m_i / m_+) - (\Sigma o_i / o_+)]$, selection intensity for continuous data
Durbin (1998)	$T^l = T^{tot} S^l / R^l$ and $T^a = T^{tot} \cdot S^a / R^a$

*m_+ = size of the sample of available resource units; m_i = number of available units in category i in the sample ($i = 1, 2, \ldots, I$); $\hat{\pi}_i = m_i / m_+$ = sample proportion of available units in category i; u_+ = size of the sample of used resource units; u_i = number of units in category i in the sample of used units; $o_i = u_i / u_+$ = sample proportion of used units in category i; and f_i = proportion of the initial number of category i items that are used. For B_{i1}, m_i is the maintained number of resource i items available, kept constant by replacing resource units as they are used (if necessary). If items are replaced as used then this index is equivalent to Chesson's index. If items are not replaced then this index is equivalent to Vanderploeg and Scandia's index. The index B_{i2} is intended for field data without resource replacement. Also, T^{tot} = total amount of active time that animal was tracked, S^l = length of resource (habitat) unit, R^l = length of animal's range, S^a = area of resource (habitat) unit and R^a = area of animal's range.

Indices of selection have been reviewed by Krueger (1972), Cock (1978), Strauss (1979), Loehle and Rittenhouse (1982), Pearre (1982), and Lechowicz (1982). Lechowicz evaluated indices according to the seven criteria that are defined in Table 1.2, where these pertain to his idea of an optimal index. Pearre classified the indices that he reviewed into two types: those reflecting selection for the particular circumstance observed ($\hat{w}$, E, D, Q, L) and those measuring an invariant degree of preference (W, E*, α, B_1, B_2). The former group (with the exception of $\hat{w}$) is a collection of ad hoc methods that do not estimate any biologically meaningful value, while the latter collection attempt to estimate the probability (or some multiple of the probability) that the next resource used will be of a specific type. Because of their biological interpretation we prefer the latter indices, which relate directly to the concept of a resource selection function as discussed in later chapters of this book.

Table 1.2 Lechowicz's (1982) seven criteria for his optimal index $f(o_i, \hat{\pi}_i)$

1. Random model	$f(o_i , \hat{\pi}_i) = 0$ if, and only if, $o_i = \hat{\pi}_i$.
2. Symmetry	If $o_i = \hat{\pi}_i$, then $\lvert f(o_i + c\ \hat{\pi}_i)\rvert = \lvert f(o_i - c\ \hat{\pi}_i)\rvert$, where c is any constant.
3. Range	For any number of resources, Max $f(o_i , \hat{\pi}_i) = f(1.0,\hat{\pi}_i)$ and Min $f(o_i,\hat{\pi}_i) = f(0,\hat{\pi}_i)$, i.e., the indices maximum should occur when only one resource is used, and the minimum should occur when a resource is not used.
4. Linearity	$f(o_i,\hat{\pi}_i) - f(o_i+a,\hat{\pi}_i) = b$ for any o_i and $\hat{\pi}_i$, where a and b are constants.
5. Robustness	An index should not be sensitive to sampling errors, particularly for rare or little-used resources.
6. Testability	An index should be amendable to statistical comparisons between subgroups (e.g., genders, age groups, etc.) or between samples (times, locations).
7. Stability	An index should give comparable results for samples from sites differing in type or abundance of available resources.

1.6 Hypothesis Tests and Confidence Intervals

Hypothesis tests provide a structured means of making decisions about a population using sample data for which probabilities of errors can be evaluated. In resource selection studies tests can be used to determine objectively whether resources are being used selectively and to compare the strength of selectivity among resources. Numerous statistical tests for evaluating overall resource selectivity have been employed, as shown in Table 1.3, and, in addition, several confidence interval procedures, and tests that consider one resource at a time have been used to assess selectivity. These latter tests include procedures described by Hobbs and Bowden (1982), Talent et al. (1982), and Iverson et al. (1985), as well as intervals for many of the selectivity indices listed in Table 1.1. A comparison of several simultaneous confidence interval procedures (Cherry, 1996) concluded the Neu et al. (1974) method was inferior to intervals given by Goodman (1965) and Bailey (1980) with respect to the rate of type I and II errors.

The tests and confidence intervals for single-resource types provide a means of assessing the variability due to sampling error but commonly do not make use of the natural multivariate nature of selectivity data. Furthermore, the overall type I error rate for all resource types is not controlled. Thomas and Taylor (1990), Jelinski (1990) and

Alldredge *et al.* (1998) point out that the choice of study design often restricts the type of analysis that can be conducted, and that some misuses have occurred in the past.

Comparisons of several of the tests that are listed in Table 1.3 have been carried out by Alldredge and Ratti (1986, 1992), Alldredge *et al.* (1998) and McClean *et al.* (1998). Alldredge and Ratti used simulations of data from design II studies, with availability censussed and use sampled, to compare Neu *et al.*'s (1974) use of the chi-squared goodness-fit test, Johnson's (1980) prefer method, the Friedman (1937) test, and Quade's (1979) test with respect to the null hypotheses tested, the assumptions required, and type I and II error rates. They concluded that no single method was superior for both types of error.

McClean *et al.* (1998) assessed the sensitivity of six tests to four definitions of availability using data on 15 young (< 7 weeks old) Merriam wild turkeys (*Meleagris gallopavo merriami*). They compared the results of these tests to their expectation that young turkeys would select grass-forb habitat. They found that Neu *et al.*'s method identified selection in agreement with their expectation over all definitions of availability and the Friedman and Quade methods agreed with their expectation at the study level definition of availability. Johnson's method, MRBP and compositional analysis did not suggest selection of grass-forb habitat at any definition of availability.

Alldredge *et al.* (1998) compared the chi-squared goodness-fit test, Johnson's (1980) prefer method, the Friedman (1937) test, compositional analysis and logistic regression with respect to the observational units of study, the populations being compared, the hypotheses being tested, and the validity of the resource selection inference considering the data available for analysis. They point out that the statistical methods have various assumptions associated with them. They suggested that consideration of the validity of assumptions should help researchers in their selection of a method of analysis. As noted above, studies may involve classifying resource units into categories, or measures of specific attributes characteristic of those units may be obtained. Below we present the assumptions required for valid application of resource selection methods separately for these two types of studies. Not all methods require every assumption. In fact, consideration of the validity of assumptions should guide the selection of a method of analysis.

A. Assumptions for Studies Categorizing Resource Units

(1) A random sample of animals is generally assumed through few researchers address the problem specifically. In reality, animals available from captures of surveys are used and assumed to be representative of the population studied. The assumption that the sample is representative of the population studied may be violated when locations are pooled across animals that differ in behaviour.

(2) When individual animals are relocated over time it is assumed that relocations are independent of one another, that is, not spatially or temporally correlated. This assumption may be violated, for example, when relocations are close to one another in time. Swihart and Slade (1985, 1997) provided a test for assessing the independence of observations in animal movements.

(3) An animal's selection of a resource is assumed to be independent of selections made by all other animals. This assumption may be violated when animals are gregarious or territorial in habitat studies or when competition for food occurs in foraging studies.

Table 1.3 Hypothesis tests used to evaluate resource selection

Test	Example references
Categorical Data	
Chi-square goodness-of-fit test[1]	Neu *et al.* (1974), Byers *et al.* (1984)
Log-likelihood chi-square based on landscape architecture features	Otis (1997, 1998)
Johnson's prefer method	Johnson (1980)
Friedman's test	Pietz and Tester (1982, 1983)
Multivariate chi-square	Dasgupta and Alldredge (1998)
Chi-squared test of homogeneity	Marcum and Loftsgaarden (1980)
Quade's test	Alldredge and Ratti (1986, 1992)
Log-linear models	Heisey (1985)
Wilcoxon's signed rank test	Kohler and Ney (1982), Talent *et al.* (1982)
Compositional analysis[2]	Aebischer *et al.* (1993), Elston *et al.* (1996), Mladenoff *et al.* (1999)
Discrete choice models	McCracken *et al.* (1998)
Continuous Data	
Kolmogorov-Smirnov two sample test	Raley and Anderson (1990), Peterson (1990)
Multiple regression	Lagory *et al.* (1985), Grover and Thompson (1986), Giroux and Bedard (1988), Porter and Church (1987)
Logistic Regression	Thomasma *et al.* (1991), Hudgins *et al.* (1985), Mysterud and Ims (1998), North and Reynolds (1996)[3]
Discriminant function analysis	Dunn and Braun (1986), Edge *et al.* (1987), Rich (1986), Dubuc *et al.* (1990), Bergin (1992)
Mahalanobis distance mapping	Knick and Rottenberry (1998), Clark *et al.* (1993)
Multivariate analysis of variance	Stauffer and Peterson (1985), Bergin (1992)
Principal components	Edwards and Collopy (1988)
Geometric method	Kincaid and Bryant (1983)
Multiple response permutation procedures	Alldredge *et al.* (1991) and Mielke (1986)

[1] Link and Karanth (1994) proposed a parametric bootstrap method to improve a chi-square goodness-of-fit test based on Manly's prey selectivity index (Manly *et al.,* 1972; Chesson, 1978).
[2] Compositional analysis can also be used with continuous variable covariates.
[3] North and Reynolds (1996) used polytomous logistic regression.

(4) In many studies, all animals are assumed to have the same available resources. For example, when habitat availability is measured or estimated for the entire study area then availability is assumed to be the same for all animals.

(5) Availability is assumed known. When the availability of habitat categories is determined by partitioning a mapped area into habitat types, availability is commonly treated as known rather than estimated. This assumption pertains to controlled feeding experiments but we have not seen it in observational field studies.

(6) Availability is constant over the period of study. This assumption may be violated, for example, by changes in resource availability over seasons or through the depletion or enhancement of resources during the study period. Studies that pool selection data across seasons or years may be violating this assumption if availability changes over these periods (Schooley 1994). Arthur *et al.* (1996) developed a method to address habitat selection when the proportions of habitat types available changes during the course of the study and Jaenike (1980) proposed a method to assess spatial or temporal changes in preference without reference to availability.

(7) Used resources are classified correctly. Errors in identifying usage locations may occur, for example, when relocation positions from radio locations are not precise (Nams 1989). Kenward (1987) and Pietz and Tester (1983) suggested that telemetry locations whose error areas cover more than one resource be discarded to meet this assumption.

(8) Studies that use surveys of animals (Design I) require the assumption of equal detectability, that is, equal probability of observation in all habitats. In forage studies, this implies that each animal has the same likelihood of being collected.

B. Assumptions for Studies Using Measurable Attributes of Resource Units

(1) The distributions of the variables that characterize locations do not change during the study period.

(2) The locations available to the animals have been correctly identified. If some locations are not available because of physical barriers or distance considerations, this assumption will be violated.

(3) The variables that influence selection have been correctly identified and measured.

(4) Animals have free and equal access to all available locations. This assumption may be violated when animals are territorial. To make inferences about a population of animals, we assume that a common function of habitat variables is adequate for modeling use by all members of the population.

(5) Locations are correctly classified into those that have been used and those that have not been used. If animals use an area and are unobserved or sign of their use is not observed, then this assumption will be violated.

(6) When studies involve sampling locations, the locations are sampled randomly and independently.

The choice of a method of statistical analysis is complex and sometimes controversial (Alldredge et al., 1998; McLean *et al.*, 1998; Aebischer and Robertson 1994; Haney 1994). The purpose of this book is to help researchers make appropriate connections between their study design and the analysis to be used, and to offer methods which result in comparable inferences for a wide variety of designs and analyses.

1.7 Discussion

The above review indicates that a unified statistical theory is needed for the analysis of resource selection studies. This is provided by this book, which is based on the concept of a resource selection function, which is a function such that its value for a resource unit is proportional to the probability of that unit being used. The resource selection function will usually be dependent on several characteristics measured on the resource units. If it is specialized to a single categorical variable (e.g., habitat or food category) then the indices of Manly *et al.* (1972), Manly (1973, 1973) and Chesson (1978) are obtained. This case is considered in Chapter 4. Selection functions based on characteristics of resources include logistic regression, discriminant analysis, log-linear models, discrete multiple choice models, and proportional hazard models. These are considered in Chapters 5 to 11. Statistical inferences are made by design-based methods from the random sampling variation, or by model-based methods where evidence for selection comes from variation about a stochastic model.

Chapter Summary

- When resources are used disproportionately to their availability, use is said to be selective.
- Resource selection occurs in a hierarchical fashion.
- Resource selection may be detected and measured by comparing any two of the three possible sets (used, unused, and available) of resource units. Three common sampling protocols are identified:

  SP-A, available units are either randomly sampled or censussed and used resource units are randomly sampled.

  SP-B, available resource units are either randomly sampled or censussed and a random sample of unused resource units is taken.

  SP-C, unused resource units and used resource units are independently sampled.

- Three general study designs for evaluating selection are identified:

  Design I, measurements of use and availability are made for the collection of all animals in the study area; individual animals are not identified.

  Design II, individual animals are identified and the use of resources is measured for each, but availability is measured at the population level.

  Design III, individuals are identified and availability is measured for each animal.

- Selection indices, hypothesis tests and confidence interval procedures are reviewed
- Lists of common assumptions for categorizing resource units or for measuring attributes characteristic of resource units are provided.

CHAPTER 2

STATISTICAL MODELLING PROCEDURES

The statistical procedures that will be used in the following chapters often involve fitting several models to each set of data and deciding which is the simplest model that accounts adequately for the observed variation. The chosen model is then used to assess the amount of resource selection. Within this framework there are several models that are particularly important, and a variety of inference procedures that are often useful. These models and procedures are reviewed in this chapter.

2.1 Simple Sample Comparisons

Resource selection studies involve the comparison between samples or censuses of used, unused and available resource units. Therefore, at an early stage the analysis of data should involve the comparison of the distributions of X variables for the samples being compared. This will provide an indication of the differences, if any, that are present, and highlight any unusual aspects of the distributions. Graphical comparisons are always useful, and these can be supplemented if desired by parametric or non-parametric tests to compare means and variances. Univariate tests can be carried out on individual variables, and multivariate tests can be carried out on several variables simultaneously. Chi-squared tests can be used to compare samples in terms of the proportions of units in different categories.

If formal tests are carried out then some consideration should be given to the assumptions that are required to make these tests valid. In particular, if the entire population of available resource units is being considered then it is not valid to regard the used units as being a random sample from a population of used units and the unused units as a random sample from a population of unused units. Hence, for example, the use of a t-test to compare a mean for used units with a mean for unused units is questionable.

In fact, the question of interest is whether the division of resource units into the two groups of used and unused ones has been made at random with respect to the X variables that are used to characterize the units. This suggests that randomization tests as discussed by Manly (1997) should be considered instead of tests that involve the assumption of random sampling from populations of resource units. However, because tests for significant differences between used and unused resource units can be carried out as part of the process for estimating the resource selection probability function it is sometimes simpler to avoid making any formal tests during the initial comparisons between distributions for used and unused units.

2.2 Linear Regression

Just about the simplest statistical model that can be considered is the linear regression model. With this model there is a dependent variable Y, which is related to one or more other variables X_1, X_2, ..., X_p through the equation

$$Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + ... + \beta_pX_p + \epsilon, \qquad (2.1)$$

where β_0 to β_p are constants to be estimated from data, and ϵ represents the part of the variation in Y that is not related to the X variables. Usually ϵ is assumed to be a random variable from a normal distribution with a mean of zero and a constant variance.

The uses of the linear regression are covered in many text on statistics (e.g., Manly, 1992). Here it suffices to say that the data available for fitting the model usually consist of a number of values of Y, with the corresponding values for each of the X variables. The β values are then estimated by the method of least-squares, which means that they are chosen to make the sum of squares of the differences between the observed Y values and the corresponding values from the fitted equation as small as possible. These are also the maximum likelihood estimates of the β values if the errors denoted by ϵ in the model are normally distributed (i.e., they are the values that makes the probability of observing the data as large as possible.

The fitting of the linear regression model can be done using most statistical packages, or using the appropriate option in a spreadsheet.

2.3 Logistic Regression

The logistic regression model is used frequently in resource selection studies, and Chapter 5 of this book considers these applications in detail. With this model the observed data consist of counts of the number of 'successes' in a certain number of trials. For example, in a simple situation an animal may be presented with two types of food, A and B, on n separate occasions. Choosing food A might then be considered to be a 'success', and the observed outcome is the number of successes in n trials. If this procedure is repeated m times under different conditions (various amounts of the two foods, different light conditions, etc.) then there is likely to be interest in how the probability of choosing A is related to the various conditions applied. This could be studied using a logistic regression model.

An assumption for this type of model is that the probability of a success is given by the equation

$$\pi = \frac{\exp(\beta_0 + \beta_1X_1 + \beta_2X_2 + ... + \beta_pX_p)}{1 + \exp(\beta_0 + \beta_1X_1 + \beta_2X_2 + ... + \beta_pX_p)}, \qquad (2.2)$$

where β_0 to β_p are constants to be estimated from the available data, and X_1 to X_p are the variables that the probability of a success is to be related to. It is also usually assumed that the number of successes observed in n trials follows a binomial distribution with mean nπ and variance nπ(1 - π), which implies that the outcome for each trial is independent of the outcome of any of the other trials.

The data for fitting the logistic regression model usually consists of a list of the number of successes observed under m different conditions, the number of trials for each of the conditions, and the values for the explanatory variables X_1 to X_p, with one set of values for each of the m conditions. The principle of maximum likelihood is usually used to obtain estimates of the values β_0 to β_p. This involves some complicated iterative calculations, but can be done in many statistical packages.

2.4 Log-Linear Models

Log-linear models also have many applications with resource selection studies, as described in Chapter 7, and elsewhere in this book. The situation with this model is that there are observations Y which are counts of the number of occurrences of a certain event under different conditions. The assumption is made that the counts follow Poisson distributions, and that the expected (mean) value of one of these counts is given by the equation

$$E(Y) = \mu = \exp(\beta_0 + \beta_1X_1 + \beta_2X_2 + ... + \beta_pX_p). \qquad (2.3)$$

As for the ordinary regression model of equation (2.1) and the logistic regression model of equation (2.2), the variables X_1 to X_p are assumed here to account for the different conditions for observations. The type of situation where this model may be reasonable would be where the counts Y are numbers of sightings of elk in blocks of land, and the X variables describe differences between the blocks (slope, aspect, etc.).

Estimates of the parameters β_0 to β_p are usually obtained using the principle of maximum likelihood, using iterative calculations. Some statistical packages have this as an option.

2.5 Proportional Hazards Models

The final particular type of model that deserves a special mention because it will be used several times in the remainder of this book is the proportional hazards model. This may be applicable in situations where there is a certain population of resource units that are selected over a period of time, with the numbers remaining either being counted at times $t_1, t_2, ..., t_S$, or with samples of either used or unused units taken at these times. Such situations are discussed in detail in Chapter 6.

This model arises by supposing that the resource units initially available before any are selected can be described by the variables $X_1, X_2, ..., X_p$ that are measured on each unit. The probability of a particular unit not being selected by time t (i.e., surviving) is then assumed to take the form

$$\phi^* = \exp\{-\exp(\beta_0 + \beta_1X_1 + \beta_2X_2 + ... + \beta_pX_p)t\}, \qquad (2.4)$$

as a function of the X values on the unit and the selection time. This function is realistic in the sense that setting $t = 0$ gives $\phi^* = 1$, which must be the case, and ϕ^* tends to zero as t increases, which will also usually be expected.

A slight generalization of equation of this model has the effect of allowing the rate at which units are used to vary with time. This replaces t in equation (2.4) with some non-decreasing function g(t) so that

$$\phi^* = \exp\{-\exp(\beta_0 + \beta_1X_1 + \beta_2X_2 + ... + \beta_pX_p)g(t)\}, \qquad (2.5)$$

with $g(0) = 0$.

Equations (2.4) and (2.5) are called proportional hazards functions because if the survival function is written as $\phi^* = \exp\{-R(t)\}$ then $R(t) = -\log_e(\phi^*)$ is what is called the hazard function, which corresponds to the instantaneous death rate at time t. For equation (2.5) the hazard function is therefore

$$R(t) = \exp(\beta_0 + \beta_1X_1 + \beta_2X_2 + ... + \beta_pX_p)g(t).$$

Hence, the ratio of the hazard functions for two resource units with different values for the X variables, say

$$R_2(t)/R_1(t) = [\exp(\beta_0 + \beta_1x_{21} + ... + \beta_px_{2p})g(t)]/[\exp(\beta_0 + \beta_1x_{11} + ... + \beta_px_{1p})g(t)]$$

$$=\exp\{\beta_0 + \beta_1(x_{21} - x_{11}) + ... + \beta_p(x_{2p} - x_{1p})\},$$

is the same for all selection times. Or, in other words, the two units always have the same proportional relationship in terms of hazard functions.

With resource selection there is interest in the units that are selected rather than those that survive selection. Thus the function which gives the probability of being selected by time t is what is relevant. This is 1 - ϕ*, or

$$w^* = 1 - \exp\{-\exp(\beta_0 + \beta_1X_1 + \beta_2X_2 + ... + \beta_pX_p)g(t)\} \qquad (2.6)$$

when the survival function is given by equation (2.5). This is then a resource selection probability function.

2.6 Generalized Linear Models

All of the models described above are special cases of a general class of generalized linear models. These have the characteristic that the expected value of an observation Y is given by some function of a linear combination of explanatory variables X_1, X_2, ... X_p, so that

$$E(Y) = f(\beta_0 + \beta_1X_1 + \beta_2X_2 + ... + \beta_pX_p), \qquad (2.7)$$

with the distribution of Y being suitably defined (McCullagh and Nelder, 1989). Thus:

- Setting f(z) = z and assuming a normal distribution for Y gives the ordinary linear regression model

- Setting f(z) = exp(z)/{1 + exp(z)} and assuming that the Y values are binomial proportions gives the logistic regression model

- Setting f(z) = exp(z) and assuming a Poisson distribution for Y gives the log-linear model

- Setting f(z) = 1 - exp{-exp(z)g(t)} and assuming that the Y values are binomial proportions gives the proportional hazards model.

There are a number of other standard types of generalized linear model as well, but they will not be used in this book. These models are usually fitted to data by maximum likelihood, using the generalized linear models option that is available in some statistical packages.

2.7 Tests Used in Modelling

The estimation of resource selection probability functions and related functions amounts to estimating the coefficients of the X variables that describe units, together with approximate values for the standard errors associated with the estimated coefficients. An obvious step in the analysis therefore involves checking to see whether the estimate $\hat{\beta}_i$ of β_i, the coefficient of x_i, is significantly different from zero. To this end, the hypothesis that $\beta_i = 0$ can be tested by comparing

$$z = \hat{\beta}_i / \mathrm{se}(\hat{\beta}_i)$$

with critical values from the standard normal distribution, where $\mathrm{se}(\hat{\beta}_i)$ is the standard error of the estimator. For example $|z| > 1.96$ implies that the estimate of the coefficient is significantly different from zero at about the 5% level on a two-sided test. Also, an approximate 95% confidence interval for β_i is

$$\hat{\beta}_i - 1.96.\mathrm{se}(\hat{\beta}_i) \text{ to } \hat{\beta}_i + 1.96.\mathrm{se}(\hat{\beta}_i).$$

The theoretical justification for using these tests and confidence intervals comes from the fact that all the parametric models that are considered in this book for resource selection functions can be estimated by maximum likelihood, and it is a standard result that maximum likelihood estimators of parameters are normally distributed for large sets of data. Precisely what constitutes a 'large' set of data depends on the circumstances, and in practice the normal approximation is best regarded with some reservations unless the data being analysed are from 100 or more resource units.

When a parametric model for a resource selection probability or related function is correct, and the parameters of the model are estimated by maximum likelihood then the quantity

$$D = -2\{\log_e(L_M) - \log_e(L_F)\}, \tag{2.8}$$

called the deviance, can be used as a measure of the goodness of fit of the model. Here L_M is the maximized likelihood for the fitted model, and L_F ($\geq L_M$) is the likelihood for the model that fits the data perfectly. Under certain conditions the deviance statistic approximately follows a chi-squared distribution with the degrees of freedom (df) being the number of observations minus the number of estimated parameters, providing that the model being considered is actually correct. The deviance may then be used for testing the goodness of fit of an estimated model by comparing its value with the percentage points of the chi-squared distribution. Basically, if the deviance is not significantly large then a model can be considered to be reasonable.

The precise conditions for using this chi-squared test depend on the circumstances. However, in many cases most of the expected counts of resource units in different categories are large, where the meaning of 'large' is the same as for other chi-squared goodness-of-fit tests. Hence the comparison of the deviance with critical values from the chi-squared distribution is reasonable if most of these counts are five or more.

Another useful application of the chi-squared distribution comes from the result that if the p_1 parameters of one model are a subset of the p_2 parameters of a second model, and the first model is in fact correct, then the deviance for the first model minus the deviance for the second model will approximately be a random value from the chi-squared distribution with $p_2 - p_1$ df. Therefore, if the deviance difference is significantly large when compared with critical values of the chi-squared distribution then this indicates that the second model provides a significant improvement over the first model in fitting the data.

Like the goodness of fit test based on the deviance, the use of deviance differences to compare models relies on the data set being 'large'. However, there is some evidence that the chi-square approximation deviance differences is more robust than the approximation for the deviance itself (McCullagh and Nelder, 1989, p. 119). Therefore tests for seeing whether one model gives a significantly better fit than a simpler model can be used with some confidence even with quite small sets of data.

Another use of a deviance difference statistic that is important is the comparison of the fit of the model for which all resource units are used with equal probability (the 'no selection' model), with the fit of a model where the probability of a unit being used is a function of the X variables that are measured on the unit. If this comparison yields a significantly high deviance difference then there is evidence of selection related to at least one of the X variables being considered.

The deviance is analogous to the residual sum of squares for regression and analysis of variance models. This leads to the idea of presenting the comparison between the fits of different models in terms of an analysis of deviance table, which is similar in format to an analysis of variance table. Several of the examples in later chapters include the use of this type of table.

2.8 Model Selection Using AIC

The final application of the deviance that must be mentioned is model selection using Akaike's information criterion (AIC) and related quantities. The AIC for a model with deviance D is defined to be

$$\mathrm{AIC} = \mathrm{D} + 2\mathrm{p}, \tag{2.9}$$

where p is the number of unknown parameters in the model that must be estimated (Akaike, 1973). The "best" model from those being considered is then the model with the smallest value for AIC. In general, increasing the number of parameters in a model reduces the deviance to some extent even when the parameters are not really necessary. The use of AIC for model selection is intended to give a good compromise between reducing the deviance and increasing the number of parameters in the model.

Alternative criteria that might be used instead of AIC include the corrected AIC

$$\mathrm{AIC_c} = \mathrm{D} + 2\mathrm{p}\ \{\mathrm{n}/(\mathrm{n} - \mathrm{p} - 1)\}, \tag{2.10}$$

where n is the total number of observations, and the Bayesian information criterion (BIC) of Schwarz (1978)

$$\mathrm{BIC} = \mathrm{D} + \mathrm{p}\ \log_e(\mathrm{n}). \tag{2.11}$$

In each case, the "best" model is assumed to be the one with the smallest value of the criterion being used.

For a thorough justification of these methods of model selection, see the book by Burnham and Anderson (1998). In most cases the use of AIC is reasonable unless the number of observations (n) is not that large, in which case the corrected value $\mathrm{AIC_c}$ should be used instead.

2.9 Analysis of Residuals

The analysis of residuals (differences between observed and expected data values) is useful for highlighting any anomalous observations in a set of data, and seeing whether there are any patterns in the discrepancies between the model and the data that require further study. Hence, whenever possible this type of analysis should be carried out before accepting that a particular model is reasonable. Typically, it involves plotting the residuals against the fitted values from the model, the variables that describe resource units, etc.

The residuals are easiest to interpret when they have been standardized so that, for the correct model, the mean should be approximately zero and the variance approximately one. If, in addition, they are approximately normally distributed then it is desirable to find most standardized residuals in the range from -2 to +2, and almost all of them within the range from -3 to +3.

The most obvious standardization involves dividing differences between observed and expected data values by the standard deviations of these differences. The nature of this standardization depends on the circumstances, but in the cases considered in the following chapters there are only two important cases, which are:

(i) If the ith data value O_i follows a binomial distribution with mean $n_i\Theta_i$ and variance $n_i \Theta_i (1 - \Theta_i)$ then standardized residuals are

$$R_i = (O_i - A_i \hat{\Theta}_i) / \surd\{A_i \hat{\Theta}_i (1 - \hat{\Theta}_i)\},$$

where $\hat{\Theta}_i$ is the value of Θ_i according to the fitted model.

(ii) If the ith data value follows a Poisson distribution with mean and variance of E_i then standardized residuals are

$$R_i = (O_i - \hat{E}_i) / \surd(\hat{E}_i),$$

where $\hat{E}_i$ is the estimated value of E_i according to the fitted model.

These two types of residuals are often called Pearson residuals, because of their relationship with the Pearson chi-squared statistic.

Although Pearson residuals are often used, there is another variety of residuals called standardized deviance residuals that is recommended by McCullagh and Nelder (1989, p. 396) for general use because they are likely to have distributions that are more normal than Pearson residuals. The calculations for obtaining standard deviance residuals are not covered here, but their values can be provided by some of the programs that are available for fitting generalized linear models.

2.10 Multiple tests and confidence intervals

Often the researcher finds the need to perform several significance tests, or construct several confidence intervals at the same time using the same data. A problem then arises because error probabilities mount up. For example, if ten independent significance tests are carried out at the 5% level of significance, with null hypotheses being true, then the probability of one or more results being significant is $1 - 0.95^{10} = 0.40$. The researcher cannot therefore be sure how to react to one or two significant results out of the ten tests.

If the ten tests are not independent then the probability of getting at least one significant result will not be 0.40. Nevertheless, the principle applies that the more tests that are carried out, the more likely it is that there will be at least one significant result by chance when the null hypothesis is true for all the tests. Similarly, the probability of one or more confidence intervals not including population values increases with the number of these intervals that are constructed.

The simplest way to overcome these problems involves making use of the Bonferroni inequality, which says that if n hypothesis tests are carried out simultaneously, each at the $100(\alpha/n)\%$ level, then the probability of declaring any result significant is α or less when the null hypotheses are all true. Likewise, the inequality says that if n confidence intervals are constructed, each with confidence level $100(1-\alpha/n)\%$, then the probability that all the intervals will include the true value of the population parameter is $1-\alpha$ or more. Thus if ten tests are carried out using the $(5/10)\% = 0.5\%$ level of significance, then the overall probability of declaring a result significant in error is 5% or less. Also, if confidence intervals for ten population parameters are constructed using a $(100 - 5/10)\% = 99.5\%$ confidence level for each one, then the probability that all the intervals will contain the true population parameter values will be 0.95 or more.

A problem with the Bonferroni procedure is that if the number of comparisons being made is large then it requires very small significance levels to be applied. This has led to the development of a number of improvements that are designed to result in more power to detect effects when they do really exist. Of these, the method of Holm's (1979) appears to be the one which is easiest to apply (Peres-Neto, 1999). This uses the following algorithm:

(a) Decide on the overall level of significance α to be used (the probability of declaring anything significant when the null hypotheses are all true).

(b) Calculate the p-value for the m tests being carried out.

(c) Sort the p-values into the ascending order, to give p_1, p_2, ..., p_m, with any tied values being put in a random order.

(d) See if $p_1 \leq \alpha/k$, and if so declare the corresponding test to give a significant result, otherwise stop. Next see if $p_2 \leq \alpha/(k - 1)$, and if so declare the corresponding test to give a significant result, otherwise stop. Next see if $p_3 \leq \alpha/(k - 2)$, and if so declare the corresponding test to give a significant result, otherwise stop. Continue this process until an insignificant result is obtained, or until it is seen whether $p_k \leq \alpha$, in which case the corresponding test is declared to give a significant result. Once an insignificant result is obtained, all the remaining tests are also insignificant, because their p-values are at least as large as the insignificant one.

2.11 Bootstrap Methods

Bootstrapping was introduced as a general tool for analysing data by Efron (1979). Initially the main concern was with producing confidence intervals for population parameters, with the minimum of assumptions being made. However, more recently bootstrap tests of significance have attracted interest as well (Efron and Tibshirani, 1993; Hall and Wilson, 1991; Manly, 1997).

The essential idea behind bootstrapping is that when the only information available about a statistical population consists of a random sample from that population, then the best guide to what might be obtained by resampling the population is provided by resampling the sample.

One of the simplest and most useful applications of this idea is for the estimation of the variance of a complicated sample statistic. For example, suppose that logistic regression (Section 2.3) is used to estimate the probabilities of different types of resource units being used. The data used for estimating the logistic regression function might then consist of measurements of variables X_1 to X_p on n resource units, each one of which is either used or not. Suppose that this is the situation, and that there is interest in the probability π of use of a unit with certain specific values for the measured variables. Then a relatively simple (although computationally expensive) way to estimate the variance of the estimator of π consists of:

(1) setting up the original sample of n resource units as a bootstrap 'population', which approximates the population of resource units from which the sample came;

(2) resampling the bootstrap population with replacement;

(3) analysing the bootstrap sample in the same way as the original sample, and producing an estimate $\hat{\pi}_B$ of π;

(4) repeating steps (b) and (c) many times to generate a bootstrap distribution for the estimate; and

(5) approximating the variance of the estimator of π by the variance of the bootstrap distribution.

There are other applications of bootstrapping that are useful in analysing data from resource selection studies. These are introduced as necessary in the following chapters.

Chapter Summary

- Resource selection studies involve a comparison between samples or censuses of used, unused and available resource units. Therefore an important part of any study is the comparison between these samples in terms of the variables that describe the units. Such comparisons may be graphical, or tests for significant differences.

- Linear regression is a basic statistical tool for describing the values of a dependent variable Y as a function of other variables, through the equation

$$Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + ... + \beta_pX_p + \epsilon,$$

where ϵ is a random error term.

- Logistic regression is a generalization of linear regression where the probability of the occurrence of an event (a 'success') is assumed to be given by an equation of the form

$$\pi = \exp(\beta_0 + \beta_1X_1 + \beta_2X_2 + ... + \beta_pX_p)/[1 + \exp(\beta_0 + \beta_1X_1 + \beta_2X_2 + ... + \beta_pX_p)].$$

This model is frequently used in resource selection studies to describe the probability that a particular type of resource unit is used.

- The log-linear model is another one which is frequently used in resource selection studies. In this case it is assumed that the counts observed in a sample for different types of resource units have Poisson distributions, with mean values described by an equation of the form

$$\mu = \exp(\beta_0 + \beta_1X_1 + \beta_2X_2 + ... + \beta_pX_p).$$

- When a resource selection process is observed for several different selection periods, the proportional hazards function may be useful for describing the probabilities that resource units of different types are not used (survive) by time t. With this model the probability of survival is assumed to take the form

$$\phi^* = \exp\{-\exp(\beta_0 + \beta_1X_1 + \beta_2X_2 + ... + \beta_pX_p)\}.$$

- The linear regression, logistic regression, log-linear and proportional hazards models are all special cases of what are called generalized linear models. With this class of models, the expected value of an observation Y is assumed to take the form

$$E(Y) = f(\beta_0 + \beta_1X_1 + \beta_2X_2 + ... + \beta_pX_p),$$

  where f(z) is one of various functions, and the distribution of Y is one of several distributions.

- When generalized linear models are fitted to data the estimated coefficients of the X variables (the β values) are approximately normally distributed. This allows standard methods to be used for assessing whether the coefficients of variables are significantly different from zero, and the accuracy of the estimates. In addition, under certain conditions the deviance (minus twice the maximized log-likelihood) can be used as a measure of the goodness of fit of a model, while differences in deviance can be used to compare the fit of alternative models.

- When there are several alternative models for a set of data, the selection of an appropriate model can be made using Akaike's information criterion (AIC), a corrected version of this for small data sets (AIC_c), or the Bayesian information criterion (BIC).

- The analysis of residuals (differences between observed and expected data values) is useful for assessing anomalous observations and for seeing generally whether a model is reasonable. There are various types of residuals that can be used with generalized linear models, including simple differences between observed and expected values divided by their standard deviations, and more complicated ones like deviance residuals.

- When several related tests of significance are carried out at the same time it may be desirable to adjust the significance level used so that the probability of obtaining any significant results by chance is not large. Similarly, if several related confidence intervals are constructed then it may be desirable to ensure that the probability of all the intervals containing their true values is large. There are a variety of procedures for achieving these aims, of which the use of the Bonferroni inequality is the simplest. However, Holm’s method is also quite easy to apply with multiple tests of significance.

- Bootstrapping is mentioned as a general method for data analysis that is sometimes useful with resource selection studies. The use of bootstrapping for estimating the variance of an estimated probability is mentioned as an example.

CHAPTER 3

EXAMPLES OF THE USE OF RESOURCE SELECTION FUNCTIONS

This chapter provides an overview of the use of resource selection functions. Examples are presented to illustrate (a) the commonly occurring case where resources are considered to be in several categories, (b) the difference between census and sample data, (c) the difference between studies that involve one period of selection time and studies that involve several periods of selection time, and (d) the application of resource selection functions with data from a geographical information system (GIS). In addition, the assumptions that are required for the estimation of resource selection functions are discussed.

3.1 Introduction

A general approach to the study of resource selection can be based on the concept of a resource selection probability function, where this is a function which gives probabilities of use for resource units of different types. This approach can be used whenever the resource being considered can be thought of as consisting of a population of N available units, some of which are used and the remainder unused, and where every unit can be characterized by the values that it possesses for certain variables $\mathbf{X} = (X_1,X_2,...,X_p)$.

Situations that can be thought of in this way are very common. For example, the resource units can be items of prey, some of which are selected by predators. In that case appropriate X variables might indicate the size, colour and species of the units. Alternatively, the resource units might be plots of land, some of which are used by an organism. Appropriate X variables might then be percentages of different types of vegetation, the distance to water and the altitude.

In terms of estimation, there is a variety of situations that need to be considered within this general framework. Some important distinctions are:

(1) In some studies all the available resource units can be censused and classified as used or unused, but in other studies it is possible only to sample resource units. Different statistical models are required for these two types of study because in the first case errors in estimating the resource selection probability function only come about because resource selection is a stochastic process, but in the second case sampling errors are also involved. As will be seen later, it turns out that, with sample data, it is only possible to estimate the resource selection probability function multiplied by an arbitrary constant, unless information about sampling fractions is available.

(2) Some studies involve observing a single episode of selection but other studies involve several periods of selection, with more and more units being used as time accumulates. It is more straightforward to analyse data from the first type of study since there is no need to model the effect of increasing selection time.

(3) In some studies the resource units are characterized by the particular categories into which they fall, such as the type of habitat that they represent. In other studies each unit is either categorized in several ways or has several quantitative variables measured on it. The first type of study occurs commonly and it is therefore worthwhile to discuss it as a special case of particular interest.

3.2 Examples

It is useful at this point to discuss some examples of situations where the estimation of a resource selection probability function is a plausible approach to data analysis. All of these examples are considered more fully in later chapters. Here the intention is just to give a broad outline of the approaches that will be used in later chapters.

Example 3.1 Habitat Selection by Moose in Minnesota

Neu *et al*. (1974) considered selection of habitat by moose (*Alces alces*) on a 33,200 acre site surrounding Little Sioux Burn in northeast Minnesota during the winter of 1971-72. They determined the proportion of the study area in four habitat categories (in burn, interior; in burn, edge; out of burn, edge; and out of burn, further) using a planimeter, and during aerial surveys classified 117 observations of groups of moose or moose tracks using the same categories. The results obtained are shown in Table 3.1.

Table 3.1 The occurrence of groups of moose or moose tracks on burned, unburned and peripheral portions of a study area surrounding Little Sioux Burn.

Location	Proportion of total acreage[1]	Moose observations[2]
In burn, interior	0.340	25
In burn, edge	0.101	22
Out of burn, edge	0.104	30
Out of burn, further	0.455	40
Total	1.000	117

[1] From a map.
[2] From aerial surveys.

The resource units are not as clearly defined as is desirable, but are basically the 'points' in the study area where moose or moose tracks can potentially be observed. The question of interest is whether the comparison between the proportions of the study area in the four habitat categories and the corresponding proportions for the 117 observations indicates that the moose used the different types of habitat according to their availability, or selected in favour of one or more types. Because habitat availability was censussed for the whole region, and use was sampled for the whole population of animals, this is an example of a design I study with sampling protocol A, as defined in Sections 1.3 and 1.4.

In this example the resource selection probability function gives the probability that a resource unit in a particular category is used by the moose. However, it is clear that absolute probabilities of use cannot be estimated from the data in Table 3.1 as there is no way of knowing what proportion of resource units were used either overall, or for any of the four categories. Indeed, given the nature of the situation, it is difficult to say how these proportions should be defined.

What can be estimated is a resource selection function, which is the resource selection probability function multiplied by an arbitrary positive constant. In fact, in Chapter 4 it will be shown that the resource selection function is given by ratios of the observed to expected sample counts in different categories, or these ratios after they have been standardized in some way. Basically, this means that the function is given by the forage ratios that are defined in Table 1.1.

Thus in the situation where observations on resource use are considered to fall into several non-overlapping categories, the estimation of a resource selection function reduces to standard procedures. Or, to put this another way, the standard procedures can be justified on the basis of a general theory of resource selection functions.

Example 3.2: Habitat Selection by Antelopes

As a second example, consider a study carried out by Ryder (1983) on winter habitat selection by antelopes (*Antilocapra americana*) in the Red Rim area in south-central Wyoming. The study area consisted of blocks of alternating public and private land, and Ryder systematically sampled the public land to obtain 256 study plots of 4 ha each, covering 10% of the total public area. On each study plot Ryder recorded the presence or absence of antelope in the winters of 1980-81 and 1981-82, the density and average height of big sagebrush (*Artemisia tridentata*), black greasewood (*Sarcobatus vermiculatus*), Nuttall's saltbush (*Atriplex nuttalli*) and Douglas rabbitbrush (*Chrysothamnus viscidiflorus*), the slope, the distance to water, and the aspect. The data for the variables other than the vegetation heights are shown in Table 3.2. The vegetation height variables have been omitted here on the grounds that they are missing for plots on which the vegetation in question does not exist.

The aspect variable with values 1 to 4 cannot be used, as it stands, in a resource selection probability function as it implies, for example, that North/Northwest has four times as much 'aspect' as East/Northeast. However, as will be more fully discussed in Example 5.1, this problem is easily overcome by converting the single variable shown in Table 3.2 into three 0-1 indicator variables. With this modification, Ryder's study plots can be thought of as a population of N = 256 available resource units, each of which has values for four vegetation density variables (X_1 to X_4), the slope (X_5), the distance to water (X_6), and three indicator variables for the aspect (X_7 to X_9). Because habitat availability and use are considered for the whole population of antelopes in the study region, this is an example of a design I study with census data in the terminology of Sections 1.3 and 1.4.

Table 3.2 The presence and absence of pronghorn on 256 plots of public land in 1980/81 and 1981/82. The variables in order from left to right are: the plot number; use in 1980/81 (1 for use, 0 for no use); use in 1981/82; sagebrush density (thousands/ha); greasewood density (thousands/ha); saltbush density (thousands/ha); rabbitbrush density (thousands/ha); mean slope of plot (degrees below horizontal); distance from centre of plot to water (m); and the aspect (1 for East/Northeast, 2 for South/Southeast, 3 for West/Southwest, and 4 for North/Northwest).

	Use variables		Density of vegetation						
			Sage	Grease	Salt	Rabbit		Distance	
Plot	1980-81	1981-82	brush	wood	bush	bush	Slope	to water	Aspect
1	0	0	9.0	6.8	2.0	0.0	0	25	4
2	0	1	18.0	0.6	1.6	0.6	5	150	3
3	0	0	8.4	0.8	0.0	12.0	45	150	4
4	0	0	3.2	0.0	0.0	4.2	65	375	2
5	0	1	12.0	0.2	0.0	0.6	5	375	3
6	1	1	7.8	2.6	10.4	0.0	5	150	3
7	0	0	5.4	0.0	0.0	7.8	55	625	1
8	1	0	10.0	3.0	0.2	0.0	5	150	1
9	1	0	12.0	0.2	0.6	2.0	5	875	3
10	1	1	12.0	0.2	4.6	0.0	15	375	3
11	0	0	0.6	0.0	3.4	0.0	75	625	2
12	1	1	7.6	0.0	3.4	4.4	5	25	1
13	0	0	4.2	0.0	0.2	0.4	45	1250	3
14	1	1	12.0	0.2	4.4	0.2	5	375	3
15	0	0	8.2	0.0	0.0	5.6	25	1250	3
16	0	1	4.0	0.0	0.0	0.4	25	1250	2
17	0	0	10.0	0.0	0.0	21.0	35	1250	4
18	1	0	4.0	0.0	0.0	11.0	15	1750	4
19	0	0	3.4	0.0	0.0	11.0	75	1250	2
20	0	0	6.4	0.0	0.0	6.4	15	875	4
21	1	0	4.0	0.0	0.0	5.4	5	1250	4
22	0	1	7.8	0.0	0.0	0.8	45	875	4
23	0	0	10.0	0.0	0.0	0.2	5	875	4
24	0	0	10.0	3.6	0.2	0.8	15	1250	4
25	1	0	3.8	0.0	0.0	0.4	5	375	1
26	0	0	1.2	0.0	3.0	0.0	45	875	3
27	0	0	2.0	0.0	1.2	0.0	5	625	4
28	1	0	5.8	0.2	9.4	0.2	5	875	3
29	1	1	7.4	0.0	0.0	4.6	5	375	1
30	1	0	7.2	0.0	0.4	0.2	5	875	3
31	0	0	4.0	0.0	0.0	6.0	55	875	3
32	1	0	18.0	0.0	0.0	17.0	5	375	4
33	0	1	1.6	0.2	7.8	0.6	5	875	3
34	0	1	3.4	4.8	4.2	2.0	0	25	4

	Use variables		Density of vegetation						
Plot	1980-81	1981-82	Sage brush	Grease wood	Salt bush	Rabbit bush	Slope	Distance to water	Aspect
35	0	1	0.8	0.0	0.0	0.6	35	875	1
36	0	1	5.9	0.0	8.2	0.2	5	875	1
37	1	0	0.2	0.2	2.9	5.4	15	375	3
38	1	0	12.0	0.0	0.0	1.2	5	625	3
39	0	0	9.0	0.0	0.0	7.0	65	625	3
40	0	0	2.0	0.0	2.4	4.2	15	625	4
41	1	1	9.2	0.0	0.0	5.6	85	375	1
42	0	1	6.2	0.0	1.2	0.3	25	875	4
43	1	1	0.2	3.0	11.0	0.0	0	25	4
44	0	0	7.4	0.0	0.0	3.8	15	625	3
45	0	1	3.2	0.0	4.0	0.2	25	150	2
46	0	0	2.2	3.8	19.2	0.0	45	150	4
47	0	0	9.6	0.0	1.6	0.2	15	25	2
48	0	0	13.0	0.0	0.0	28.0	15	625	4
49	0	1	11.0	2.0	3.4	0.4	15	625	4
50	0	1	2.6	0.0	0.0	1.0	35	875	4
51	1	1	4.8	6.8	8.4	0.0	0	25	4
52	0	0	6.6	0.6	2.8	1.0	0	375	4
53	0	1	16.0	0.2	0.0	12.0	15	875	4
54	0	1	12.0	0.0	0.0	18.0	55	1750	4
55	0	0	3.0	0.0	0.0	0.6	15	1750	3
56	0	0	0.7	0.0	0.0	0.0	45	1250	4
57	0	0	3.0	0.2	3.6	0.2	75	875	4
58	0	0	2.6	0.0	0.0	2.0	0	1250	4
59	0	1	5.4	1.0	4.2	1.0	55	1250	2
60	1	0	5.6	1.0	11.4	0.0	5	875	3
61	1	0	3.8	10.0	18.2	0.0	0	150	4
62	1	1	0.0	0.0	0.0	0.0	25	1250	4
63	0	0	0.4	0.0	0.0	0.0	15	1250	2
64	0	1	1.4	0.0	0.0	0.2	15	1250	3
65	1	0	0.4	0.0	8.3	0.0	5	375	2
66	0	1	4.8	0.0	0.0	0.0	5	875	4
67	0	0	5.6	0.0	2.8	0.2	25	1250	3
68	0	0	0.2	0.2	18.4	0.0	5	625	3
69	0	1	3.9	0.0	0.0	0.0	0	25	4
70	0	1	10.0	0.0	0.0	5.6	25	875	4
71	0	0	4.2	0.0	0.0	0.2	25	875	4
72	1	0	0.2	0.0	15.6	0.0	5	875	3
73	0	1	0.2	0.2	0.0	0.0	0	150	4
74	1	0	5.0	0.0	0.0	0.0	15	625	3

	Use variables		Density of vegetation						
Plot	1980-81	1981-82	Sage brush	Grease wood	Salt bush	Rabbit bush	Slope	Distance to water	Aspect
75	0	0	5.4	0.2	0.0	1.0	55	1250	4
76	1	1	3.4	1.2	31.4	0.0	25	625	1
77	0	1	10.0	0.0	0.2	0.8	15	375	4
78	0	0	0.4	0.0	11.2	3.6	5	625	4
79	1	1	4.8	0.2	0.2	4.2	25	625	3
80	0	1	7.2	4.4	6.8	2.4	0	25	1
81	1	0	6.6	1.2	17.0	1.0	85	150	4
82	1	1	1.6	0.0	0.0	1.0	45	875	4
83	1	1	4.2	4.4	12.4	0.0	5	150	1
84	0	0	5.0	0.0	0.0	2.2	25	1750	3
85	1	1	3.4	0.0	17.0	0.0	15	875	3
86	0	0	0.8	0.0	11.8	0.0	85	1250	4
87	0	0	0.8	0.0	0.0	1.0	15	875	3
88	0	0	8.0	0.0	0.0	3.0	35	2250	1
89	0	0	0.2	0.0	5.2	0.2	5	1250	1
90	0	0	10.0	0.0	0.0	4.0	15	1250	4
91	0	0	0.2	0.0	45.4	0.2	15	375	4
92	0	0	4.2	0.0	0.0	1.4	15	1750	1
93	0	0	7.0	0.0	0.0	1.6	15	875	1
94	0	0	2.4	0.4	0.2	6.6	35	1250	1
95	0	0	10.0	0.0	0.6	6.4	55	875	4
96	0	0	7.8	0.0	0.0	0.6	15	2250	4
97	1	1	12.0	0.0	0.4	4.4	25	1250	1
98	0	0	3.2	0.0	0.0	0.8	25	1250	4
99	1	1	1.0	0.2	1.6	0.8	15	375	3
100	0	1	5.4	0.0	22.4	1.0	5	1750	4
101	0	1	11.0	0.0	2.2	2.4	0	1250	4
102	0	1	0.8	0.0	0.0	4.4	15	875	4
103	0	0	3.6	0.0	6.2	3.2	5	625	2
104	0	1	5.4	0.0	0.0	2.2	25	2250	3
105	1	1	9.0	0.0	0.0	1.8	5	1750	4
106	0	0	2.0	0.2	26.4	0.0	5	875	4
107	0	0	0.6	0.0	0.0	10.0	0	150	4
108	0	1	4.2	0.0	2.2	1.2	15	1750	4
109	0	1	1.8	0.0	2.0	0.2	5	1250	4
110	1	1	4.4	2.0	1.4	9.6	25	375	4
111	1	0	0.2	0.0	0.0	0.6	15	375	1
112	1	1	0.2	0.0	7.0	0.0	15	2250	3
113	1	0	3.8	2.8	0.6	1.6	15	1750	4
114	1	1	0.2	0.2	26.6	0.0	0	625	4

	Use variables		Density of vegetation						
Plot	1980-81	1981-82	Sage brush	Grease wood	Salt bush	Rabbit bush	Slope	Distance to water	Aspect
115	1	1	0.4	0.0	0.0	0.0	0	25	4
116	0	0	7.2	0.0	0.0	0.4	35	2750	4
117	1	1	9.4	0.0	0.0	0.2	15	1750	4
118	0	0	0.2	0.0	24.6	0.2	15	875	4
119	0	1	4.0	0.0	0.0	11.0	35	375	4
120	0	0	3.6	0.0	0.0	1.4	45	2750	2
121	0	0	2.4	0.0	0.0	7.6	15	2250	4
122	1	1	5.0	0.0	0.4	1.4	25	875	4
123	0	1	1.4	0.2	11.4	0.0	0	375	4
124	0	0	8.2	0.0	0.0	16.0	5	2250	4
125	0	0	7.4	0.0	0.0	3.8	15	1750	2
126	0	0	2.0	1.4	18.4	0.2	0	375	4
127	0	0	0.4	0.0	9.0	1.0	15	625	4
128	0	0	4.4	0.0	0.0	2.6	5	2750	4
129	1	1	0.2	0.0	1.2	0.2	25	2250	1
130	0	0	3.6	0.6	5.4	0.2	5	625	4
131	0	0	2.4	0.2	5.2	0.0	0	875	4
132	1	0	10.0	0.0	0.0	11.0	15	2250	4
133	0	0	8.4	0.0	5.8	0.2	25	1750	4
134	0	0	0.0	0.0	0.4	0.0	45	25	3
135	0	1	4.0	1.2	17.0	0.0	0	625	4
136	1	0	4.8	0.0	0.0	7.2	25	2750	3
137	0	0	1.4	0.0	1.2	1.4	15	2250	4
138	1	1	0.0	0.0	0.0	0.0	35	375	4
139	0	0	2.4	0.4	20.4	0.2	5	625	1
140	0	0	3.4	0.0	13.4	0.2	15	2750	4
141	0	0	3.0	0.2	0.0	0.4	15	1750	4
142	1	0	5.0	0.0	0.2	6.6	15	375	4
143	0	1	0.0	0.0	5.0	0.4	5	150	2
144	0	0	8.4	0.0	0.0	3.0	5	2750	4
145	1	0	7.6	1.4	0.4	0.0	25	2750	4
146	0	1	0.2	0.2	11.6	0.0	15	375	4
147	0	0	2.8	0.2	0.2	11.0	0	625	4
148	1	0	3.4	0.0	0.0	2.2	0	2750	4
149	0	1	2.4	0.0	0.0	2.2	35	2750	3
150	1	1	1.0	0.8	16.0	0.2	15	1750	2
151	1	1	2.2	2.0	25.2	0.0	0	625	4
152	1	1	0.4	8.0	2.2	0.0	0	375	4
153	0	0	1.6	0.0	0.0	4.4	45	2750	2
154	0	0	7.6	0.0	0.0	1.2	15	2750	4

	Use variables		Density of vegetation						
			Sage	Grease	Salt	Rabbit		Distance	
Plot	1980-81	1981-82	brush	wood	bush	bush	Slope	to water	Aspect
155	0	1	0.4	0.0	5.0	5.7	15	875	4
156	0	0	0.6	1.4	21.2	0.0	5	625	4
157	0	0	5.6	7.6	14.2	0.0	0	25	4
158	0	0	14.0	0.0	0.0	0.2	15	2750	4
159	0	1	3.6	0.0	2.8	3.4	15	2750	4
160	1	1	2.2	5.8	8.2	0.0	0	625	4
161	0	0	6.6	0.0	1.8	2.4	45	2750	4
162	0	0	11.0	0.0	0.4	3.6	15	2750	3
163	1	1	3.0	0.2	17.2	0.2	15	875	2
164	0	0	3.4	0.2	3.4	5.8	35	375	4
165	0	1	12.0	7.2	0.8	0.6	0	25	4
166	0	0	10.0	0.0	0.0	1.6	25	2750	3
167	1	0	3.8	0.2	1.4	1.0	15	2750	1
168	0	0	0.4	1.8	18.0	0.0	0	150	4
169	1	1	12.0	0.0	0.0	7.0	35	2750	1
170	0	1	1.0	0.0	0.0	0.0	15	2750	4
171	1	0	0.4	6.1	0.4	0.0	0	875	4
172	1	0	1.2	8.2	18.2	0.0	0	625	4
173	0	1	6.0	0.0	4.4	2.0	5	2750	4
174	0	0	6.2	0.0	0.2	1.8	25	2250	1
175	0	0	3.0	0.2	1.4	14.0	15	1750	3
176	1	1	0.4	0.6	19.6	0.0	0	625	4
177	0	1	2.0	0.2	10.6	0.2	0	25	4
178	0	0	1.2	0.0	8.8	0.2	5	2250	2
179	0	1	0.0	0.0	0.4	0.0	45	1750	4
180	0	0	0.4	0.0	6.8	0.0	0	1250	4
181	1	1	0.2	0.2	24.6	0.0	0	875	4
182	0	0	10.0	0.0	0.0	0.6	15	2750	3
183	0	0	7.2	0.2	0.6	0.4	25	1250	4
184	0	1	6.2	0.8	1.6	3.0	25	1750	4
185	1	1	0.2	0.2	0.2	0.2	0	875	4
186	0	0	1.8	1.2	1.8	0.0	0	25	4
187	0	0	7.2	0.0	0.0	8.4	55	2750	3
188	0	0	4.8	0.0	0.0	1.0	15	2750	2
189	0	1	3.2	0.0	0.4	3.6	15	2750	1
190	1	0	3.4	0.2	3.0	1.2	5	1750	4
191	0	0	0.2	0.0	8.8	2.6	5	1750	2
192	0	1	4.0	6.0	2.0	0.2	0	875	4
193	0	0	4.0	2.0	2.0	0.2	0	150	4
194	0	0	6.0	0.0	0.0	11.0	15	2750	3

	Use variables		Density of vegetation						
Plot	1980-81	1981-82	Sage brush	Grease wood	Salt bush	Rabbit bush	Slope	Distance to water	Aspect
195	0	0	0.6	0.2	26.8	0.0	5	1250	4
196	0	0	1.0	0.0	6.2	3.8	0	1250	4
197	0	0	2.0	4.0	0.2	3.8	0	375	4
198	0	1	0.2	0.0	6.4	2.0	0	150	4
199	0	1	2.6	0.0	0.0	2.6	0	2750	4
200	1	1	2.8	0.0	0.0	2.6	25	2750	4
201	0	1	6.1	0.0	1.8	0.4	15	2250	2
202	0	1	3.0	0.8	5.0	1.2	15	875	4
203	0	0	6.2	0.6	0.0	0.0	15	1250	4
204	0	0	4.0	2.6	8.0	4.0	0	875	4
205	0	0	6.0	3.0	2.0	6.0	0	25	4
206	0	1	0.8	0.0	9.4	6.0	15	2750	1
207	0	1	0.2	0.2	19.0	0.2	5	875	4
208	0	0	1.8	0.2	13.6	0.8	35	1250	4
209	0	0	8.0	12.0	5.0	8.0	0	375	4
210	0	0	0.2	0.2	25.6	0.0	0	25	4
211	0	0	2.2	0.0	0.0	2.0	15	2750	2
212	1	0	6.6	0.0	0.0	11.0	5	2750	4
213	0	0	2.6	2.2	6.6	0.6	35	1750	3
214	1	0	6.6	2.2	3.4	0.4	25	875	2
215	1	0	0.0	0.0	6.4	0.0	15	875	4
216	0	0	1.0	3.0	0.2	0.6	0	625	4
217	0	0	0.4	0.4	4.8	0.0	0	375	4
218	0	0	9.2	0.2	0.8	7.8	5	2750	1
219	0	0	3.2	0.4	4.0	2.0	5	625	4
220	0	0	1.4	0.2	8.0	0.2	15	1250	4
221	0	0	6.2	0.2	0.2	0.2	0	375	4
222	0	0	6.0	0.0	0.0	3.2	15	2750	2
223	0	1	6.4	0.0	1.4	3.4	5	2750	4
224	0	0	1.0	4.2	15.2	0.0	5	1750	4
225	1	1	0.2	0.2	17.0	0.0	15	375	1
226	1	0	1.0	0.0	0.4	0.7	15	625	4
227	0	0	0.0	0.0	0.0	0.0	0	875	4
228	1	0	3.2	0.0	0.0	2.2	0	25	4
229	0	0	0.2	1.4	1.4	0.2	25	1750	2
230	0	0	2.0	2.6	12.0	0.2	5	1250	1
231	1	1	1.4	0.0	4.8	1.0	45	2250	2
232	0	0	0.0	0.0	1.4	0.0	35	1750	4
233	0	1	0.0	0.0	0.4	0.0	0	1750	4
234	0	1	0.0	11.0	14.6	0.0	0	1250	4

	Use variables		Density of vegetation						
Plot	1980-81	1981-82	Sage brush	Grease wood	Salt bush	Rabbit bush	Slope	Distance to water	Aspect
235	0	1	1.4	0.0	0.2	0.8	85	2250	3
236	1	1	0.2	0.2	23.8	0.2	35	1750	2
237	0	0	0.0	0.0	8.0	0.0	15	2250	4
238	0	0	0.6	0.2	11.4	1.2	15	1250	3
239	0	0	0.0	0.0	0.0	0.0	5	2750	3
240	0	0	2.4	0.2	22.2	0.6	5	1750	4
241	0	0	1.6	0.0	0.2	0.8	15	2750	2
242	0	0	2.2	0.0	3.0	2.2	15	2250	3
243	1	0	1.1	0.0	15.1	0.0	25	1750	4
244	0	0	3.2	0.0	7.8	3.4	5	1750	4
245	0	0	0.0	0.0	0.0	0.0	15	2750	4
246	0	1	1.8	0.2	0.0	1.6	5	2750	2
247	0	0	4.2	0.0	10.0	0.2	65	2750	2
248	0	0	0.8	0.0	12.6	0.2	5	2750	4
249	0	0	5.2	0.0	0.2	4.2	5	2750	2
250	0	0	0.2	0.0	11.8	0.0	15	2750	2
251	1	0	0.0	0.0	2.0	0.0	85	2750	4
252	0	1	6.4	0.0	2.8	5.2	5	2750	4
253	0	0	3.4	0.0	0.0	8.2	75	2750	4
254	0	0	1.8	0.0	14.4	0.0	5	2750	4
255	0	0	0.4	0.0	9.7	0.4	5	2750	4
256	0	0	0.4	10.0	5.0	0.0	75	2750	4

In this example the resource selection probability function gives the probability of a study plot being used as a function of X_1 to X_9. However, as the presence and absence of antelopes is recorded in 1980-81 and 1981-82 there are several different ways to define 'use', with correspondingly different resource selection probability functions. Thus, plots can be considered to be 'used' if antelopes are recorded in either winter, or, alternatively, if antelopes are recorded in both winters. Other possibilities are to think of the two winters as replicates of the selection process, or to think of each plot as being observed after one and two years of selection, with a plot being considered as 'used' when antelopes are first observed.

One of the most straightforward ways of estimating a resource selection probability function involves assuming that the probability of observing antelopes in one winter, on a plot with values $\mathbf{x} = (x_1,x_2,...,x_9)$ for the nine X variables, is given by the logistic regression equation

$$w^*(\mathbf{x}) = \{\exp(ß_0 + ß_1x_1 + ...+ ß_9x_9)\}/\{1 + \exp(ß_0 + ß_1x_1 + ...+ ß_9x_9)\}$$

where $ß_0$ to $ß_9$ are unknown parameters to be estimated. This is the approach adopted in Example 5.1, with the estimation of the ß parameters being carried out using the logistic regression option of a standard computer package.

Example 3.3 Selection of Snails by Birds

An experiment described by Bantock *et al*. (1976) on the selection of *Cepaea nemoralis* and *C. hortensis* snails by the song thrush (*Turdus ericetorum*) is an example of a situation where it is necessary to consider the effect of different periods of selection time. The experimental site was a disused camellia-house covering 39 m² with ground vegetation of nettles that was almost empty of snails before the experiment began on 29 June, 1972. On that date, 498 yellow five-banded (Y5H) *C. hortensis*, 499 yellow five-banded (Y5N) *C. nemoralis*, and 877 yellow mid-banded (Y3N) *C. nemoralis* snails were released into the area. Shells were uniquely marked so that the survivors could be determined from censuses taken at various times after the population had been set up. The nearest natural population of *Cepaea* was 100 m from the experimental site. Thrush predation occurred there, but stopped when the experimental snails were released.

Simplified results from the experiment are shown in Table 3.3. Only five-banded snails are considered because extra mid-banded *C. nemoralis* were added to the population after 29 June. The methods that we propose for estimating a resource selection probability function using the simplified data can be extended to take into account this type of experimental procedure by thinking of each augmentation of the prey population as the start of a 'new' experiment, but we consider that it is better not to introduce this complication into an illustrative example.

In Table 3.3 different types of snail are defined in terms of two X variables, where X_1 is a species indicator which is 1 for *C. nemoralis* and 0 for *C. hortensis*, and X_2 is the maximum shell diameter in units of 0.3 mm over 14.3 mm. The table shows that there was a population of N = 997 available resource units (snails) at the start of the experiment, each with values for p = 2 characterizing variables, and the survivors are known after 6, 12 and 22 days of selection.

Clearly the time element cannot be ignored in this example, so that the resource selection probability function has to give the probability of a snail being eaten after t days of selection. There are various ways of estimating such a function, one of which is discussed in Example 6.1, where it is assumed that the probability of use by day t is given by

$$w^*(\mathbf{x},t) = 1 - \exp\{-\exp(ß_0 + ß_1x_1 + ß_2x_2)t\}.$$

This corresponds to assuming that the probability of not being used is given by the proportional hazards function

$$\phi^*(\mathbf{x},t) = \exp\{-\exp(ß_0 + ß_1x_1 + ß_2x_2)t\},$$

which is often used for analysing other forms of survival data.

Because the available snails are assumed to be the same for all of the birds selecting them, and no information is available on the choices of individual birds, this is another example of a design I study in the terminology of Section 1.3, with census data.

Table 3.3 Eaten and uneaten snails in an experimental population of yellow five-banded Cepaea nemoralis and C. hortensis. The species indicator variable X_1 is 0 for C. hortensis and 1 for C. nemoralis and the coded size variable is the shell diameter in units of 0.3 mm over 14.3 mm, rounded to the nearest unit.

Shell diameter	Species	Coded size	Snails eaten between days			Left on day
	X_1	X_2	0-6	6-12	12-22	22
14.6	0	1	0	0	0	2
14.9	0	2	1	0	0	2
15.2	0	3	0	0	0	2
15.5	0	4	0	1	1	3
15.8	0	5	0	1	0	4
16.1	0	6	0	3	3	6
16.4	0	7	2	2	3	12
16.7	0	8	3	1	9	23
17.0	0	9	7	6	9	20
17.3	0	10	5	8	8	35
17.6	0	11	6	9	13	29
17.9	0	12	6	10	11	32
18.2	0	13	8	9	10	35
18.5	0	14	7	12	11	26
18.8	0	15	4	9	6	17
19.1	0	16	1	3	1	11
19.4	0	17	5	1	1	7
19.7	0	18	1	0	2	5
20.0	0	19	0	1	1	1
20.3	0	20	0	0	0	4
20.9	0	22	0	0	0	1
16.7	1	8	0	0	0	1
17.3	1	10	0	0	1	0
17.6	1	11	0	0	0	1
17.9	1	12	2	1	0	2
18.2	1	13	4	2	3	0
18.5	1	14	2	2	1	9
18.8	1	15	5	5	5	7
19.1	1	16	3	8	8	11
19.4	1	17	10	14	15	30
19.7	1	18	19	18	12	19
20.0	1	19	21	17	4	16
20.3	1	20	18	17	11	13
20.6	1	21	13	8	10	21
20.9	1	22	13	10	8	13
21.2	1	23	5	7	2	12
21.5	1	24	7	1	3	4
21.8	1	25	3	2	2	5
22.1	1	26	1	2	0	3
22.4	1	27	0	1	2	0
22.7	1	28	2	0	0	0
23.0	1	29	1	0	0	0

Example 3.4 Nest Site Selection by Fernbirds

An example where only samples of available and used resource units were taken is provided by a study of nest site selection by fernbirds (*Bowdleria puncta*) in Otago, New Zealand, that is described by Harris (1986). Harris measured nine variables on 24 nest sites found during the 1982-83 and 1983-84 seasons and found comparative available sites by choosing 25 random points in the study area and locating a polystyrene model nest at the centre of the nearest clump of vegetation. Harris concluded that only three of the nine variables that he measured showed important differences between the two samples. The data for these variables (the canopy height, the distance from the outer edge of the nest to the nearest outer surface of the clump of vegetation in which the nest is situated, and the perimeter of the clump of vegetation) are shown in Table 3.4.

In this example the total number of available resource units (potential nest sites) is unknown, but clearly very large, and there are p = 3 variables measured on each sampled resource unit. From the data available there is no way of estimating the absolute probability of a potential nest site being used. However, it is possible to estimate a resource selection function which is this probability multiplied by an arbitrary constant.

One approach to estimating the resource selection function involves assuming that it takes the form

$$w(\mathbf{x}) = \exp(ß_0 + ß_1x_1 + ß_2x_2 + ß_3x_3).$$

This then leads to a logistic regression model for the sample data, as is discussed further in Example 5.2.

Harris' study has design I with sampling protocol A in the terminology of Sections 1.3 and 1.4 because availability was measured by sampling potential nest sites over the entire study region, and there is the implicit assumption that the probability of a potential site being used was approximately the same for all fernbirds.

Example 3.5 Selection of Corixids by Minnows

Although most resource selection studies based on sample data consider only a single selection period, there are cases where a population of resource units has been sampled at several times while selection is proceeding. The situation is then similar to that in Example 3.3 but with samples taken instead of censuses.

An example is a study of the use of corixids as food by minnows (*Phoxinus phoxinus*). Popham (1944) sampled the corixids in a pond every day from 13 to 19 September, 1942, introduced 50 minnows on the evening of 19 September, and then sampled again every day from 22 to 28 September. This resulted in the sample counts shown in Table 3.5 for three corixid species, with shades of grey classified as light, medium, and dark. The samples taken before the introduction of minnows have been lumped together to form a single 'available sample' as they have similar proportions for the species and colours of corixids.

Popham argued that any changes in the relative proportions of the nine types of corixid after 19 September were largely due to predation by the minnows because the effects of immigration and emigration of corixids were negligible, and newly formed adults were entering the population at a low rate. Therefore this can be thought of as a situation where there are eight samples of unused resource units (corixids), taken after selection times of 0, 3, 4, ..., 9 days.

Table 3.4 Comparison of variables measured on fernbird nest sites and randomly located sites in the same area.

	Nest sites			Random available sites		
	Canopy height	Distance to	Perimeter of	Canopy height	Distance to	Perimeter of
	1.20	14.0	8.90	0.47	13.5	3.17
	0.58	25.0	4.34	0.62	8.0	3.23
	0.74	14.0	2.30	0.75	19.0	2.44
	0.70	12.0	5.16	0.52	5.0	1.56
	1.36	14.5	2.92	0.73	8.0	2.28
	0.78	17.0	3.30	0.62	16.0	3.16
	0.45	15.0	3.17	0.60	17.0	2.78
	0.78	15.0	4.81	0.26	4.5	3.07
	0.63	16.0	2.40	0.46	15.0	3.84
	0.75	12.0	3.74	0.28	12.0	3.33
	0.55	12.0	4.86	0.53	11.0	2.80
	0.45	20.0	2.88	0.42	17.0	2.92
	1.56	16.0	4.90	0.47	20.0	4.40
	0.85	23.0	4.65	0.50	13.0	3.86
	0.58	12.0	4.02	0.54	16.0	3.48
	0.75	13.0	4.54	0.56	18.0	2.36
	0.55	18.0	3.22	0.32	7.0	3.08
	0.56	18.0	3.08	0.62	16.0	5.07
	0.57	19.5	4.43	0.39	15.0	2.02
	0.41	16.0	3.48	0.56	8.0	1.81
	0.65	18.0	4.50	0.27	9.0	2.05
	0.78	14.0	2.96	0.42	11.0	1.74
	0.64	18.0	5.25	0.70	13.0	2.85
	0.71	18.0	3.07	0.26	14.0	3.64
				0.34	9.5	2.40
Mean	0.733	16.25	4.037	0.488	12.62	2.934
Std. Dev.	0.278	3.41	1.370	0.147	4.37	0.844

There are nine types of resource unit in this example, which differ because of their species and their colour. Although these are qualitative rather than quantitative differences it is still possible to use X variables to describe the units, and hence fit this example within the framework of the present chapter. Essentially, all that is necessary is to set up appropriate 0-1 indicator variables, as discussed more fully in Example 7.3.

Because of the nature of the data, with only samples of unused resource units being available, it turns out that it is not possible to estimate a resource selection probability function, or even this function multiplied by an arbitrary constant. However, what can be done is to assume that the probability of a unit not being used by time t takes the form

$$\phi^*(\mathbf{x},t) = \exp\{(ß_0 + ß_1x_1 + ... + ß_px_p)t\},$$

where the exponential parameter is necessarily negative. It then becomes possible to describe the sample counts of different types of corixid using a log-linear model, and hence estimate the parameters $ß_1$ to $ß_p$ using a suitable computer program.

Although $ß_0$ cannot be estimated, it is possible to estimate

$$\phi(\mathbf{x},t) = \exp\{(ß_1x_1 + ß_2x_2... + ß_px_p)t\},$$

which gives the probability of not being used by time t multiplied by an arbitrary constant. The estimated function can then be used to rank the types of corixid in order of the probability of not being used, which is the reverse order to that for the probability of use. In this way, the selection of corixids can be studied to some extent at least.

Because the availability of corixids is assumed to be the same for all minnows, and there is no information about the choice of corixids by individual minnows, this is another example of a design I study in as defined in Sections 1.3 and 1.4, because unused resource units are compared with those available.

Table 3.5 Results from sampling a corixid population before and after minnows were introduced into a pond on 19 September, 1942.

			Samples of live corixids on September						
Species	Shade of grey	Available samples	22	23	24	25	26	27	28
Sigara venusta	Light	120	5	5	5	6	2	6	3
	Medium	726	102	110	131	120	105	157	134
	Dark	225	25	58	22	22	16	20	17
Sigara praeusta	Light	25	2	0	2	1	0	1	0
	Medium	33	6	2	6	3	5	4	5
	Dark	6	1	1	0	0	0	0	1
Sigara distincta	Light	15	0	0	1	0	0	0	0
	Medium	39	4	3	3	4	3	4	5
	Dark	6	0	0	0	0	0	0	0

Example 3.6 Habitat Selection by Moose in the Innoko National Wildlife Refuge

The last example that will be considered is, like the first example, about habitat selection by moose (*Alces alces*). However, the availability of a GIS allowed a more precise assessment of the habitat than was possible with the first example.

In brief, aerial line transect surveys for moose were conducted by helicopter in the Innoko National Wildlife Refuge in Alaska in March 1994 and 1996. The survey was in the north of the refuge in 1994 and in the south in 1995. The location for each moose group seen was recorded using an onboard global positioning system (GPS) connected to a laptop computer (Anthony and Stehn, 1994), with an error that was judged to be less than 100 m. Altogether, 208 moose in 135 groups were seen in 1994, and 311 moose in 123 groups were seen in 1996. Moose groups were considered to provide independent data. Erickson *et al*. (1998) describe the field methods in more detail.

The GIS system for the refuge had the total area of over 6000 square miles divided up into 30m by 30m pixels, each of which was classified as being in one of 22 land cover classes based on the type of vegetation. However, these pixels were not considered to be the large enough to represent units that might be selected by moose groups. Therefore, the habitat units used in analyses were defined to be circles 400 metres in radius, consisting of approximately 561 pixels. In this way it was possible to compare the characteristics of the units used by moose groups (circles centred on the groups) with a sample of available units. Erickson *et al*. (1998) used logistic regression for this purpose, to see how the probability of a moose group being located in a habitat unit is a function of the proportions of the 22 land cover classes in the unit.

This study is typical of many resource selection studies that have been conducted since GIS systems have become available. It is considered further in Example 9.1, which covers some of the special issues that arise from such studies including:

- How should resource units be defined in terms of size and shape?
- How should the set of available units be defined, taking into account the distance that animals can move?
- How should the set of available units be compared with the used units when there may be information on each available unit, but the number of these units is astronomically large?

This particular example is a design 1 study because there is only one observation on each moose group and selection must therefore be considered at the population level. It is sampling protocol A or B, depending on whether the GIS system is used to provide a sample of available units or a sample of unused units. Typically, with these types of study there are many available units and rather few are used so that sampling protocols A and B are more or less equivalent.

3.3 Sample Designs

Although all of the examples just considered have designs that are classified as being of design I in the terminology of Section 1.3, this should not be taken to mean that this is inevitable when a resource selection function is estimated. The methods proposed in this and the following chapters can be used with design II studies where the resources used by individual animals are measured but the availability of resources is measured at the population level. All that needs to be done is to estimate a separate selection function for each animal if this is necessary. Similarly, with design III studies, where resource use and availability is measured for several animals, it is possible to estimate a separate selection function for each animal if necessary. The choice of examples in this chapter just reflects the fact that design I studies have been most common in the past.

3.4 Assumptions

As discussed in Section 1.6, there are some general assumptions that are involved in the estimation of resource selection functions. These are worth reconsidering at this point as they are implicitly assumed to hold for all of the above examples. These assumptions are that:

(a) the distributions of the measured X variables for the available resource units and the resource selection probability function do not change during the study period;

(b) the population of resource units available to the organisms has been correctly identified;

(c) the subpopulations of used and unused resource units have been correctly identified;

(d) the X variables which actually influence the probability of selection have been correctly identified and measured;

(e) organisms have free and equal access to all available resource units; and

(f) when studies involve the sampling of resource units, these units are sampled randomly and independently.

The requirement that the distributions of the X variables for the available resource units and the resource selection probability function do not change during the study period is difficult to satisfy with many studies. With sample data the population of available resource units may change without the investigator realizing that this has happened. The resource selection probability function may change with the season, the weather, the distribution of remaining resources, etc. In all cases, inferences are made with respect to 'averages' over the period when used and unused resource units are collected. If populations are not changing rapidly then this should not be a problem but generally the selection times should be kept as short as possible. In rapidly changing populations, an attempt should be made to obtain several 'snapshots' of resource selection functions at different times.

Identification of the population of resource units that are available to the organisms, and the variables which influence the probability of selection, are probably the most crucial and most difficult aspects of the study design. Not much advice can be given here to help the researcher in specific cases. As noted by Manly (1985, p. 172) and Rexstad *et al*. (1988), the use of stepwise procedures to identify important variables from a larger set of potentially important variables is liable to give misleading results and should be used with caution.

The assumption that organisms have unrestricted and equal access to all the available resource units is most easily justified when the subpopulation of used units is small relative to the population of available units. Of course, changes in the density of animals or in the availability of resource units may change the underlying selection strategies and the selection function. Thus, statistical inferences are made with respect to the specific densities present in the study area.

If resource units are not sampled randomly and independently then the estimates of the coefficients of a resource selection probability function may still be meaningful but standard errors may not reflect the true variation in the populations. In many situations it is difficult to design studies so that the individual resource units are the basic sampling units and hence avoid pseudoreplication (Hurlbert, 1984). For example, relocations of radio-tagged animals may be recorded at a series of points in time. Care must then be taken to ensure that the time interval between recordings is sufficient to assume that identifications of used habitat points are independent events if standard model based analyses of the data are to be conducted. Another complicating factor is that different collection methods are often required to obtain used and unused units.

Given these problems, one approach is to estimate a separate resource selection function for several independent replications of batches of dependent units. Alternatively, one might consider the selection of individual animals as independent events, and estimate a separate selection function for each animal by randomly sampling the units available and the units used by each animal. This may be the only reasonable approach for study of food and habitat use by highly territorial animals.

Another (less satisfactory) way of handling pseudoreplication that can be used with some of the models discussed in the following chapters involves estimating a heterogeneity factor as the ratio of an observed chi-squared goodness of fit statistic to the value that is expected from random sampling. The variances of all parameter estimates can then be multiplied by the heterogeneity factor to adjust them for the 'extraneous' variance. A problem with this approach is that it depends on the assumption that the variances of the observed data values are all inflated by the same amount, which may well not be true. Also, the method can be used only in cases where the counts of different types of resource unit are large enough to permit a valid chi-squared goodness of fit statistic to be calculated.

Chapter Summary

- The resource selection probability function is a function that gives the probability of a resource unit being used as a function of the values that the unit possesses for certain variables X_1, X_2, ..., X_p. There are a variety of different situations that can occur, with census or sample data, one or more periods of selection, with the X variables either being continuous, discrete, or representing different categories of unit.

- Six examples are provided to illustrate the type of situations that can occur, involving habitat selection by moose in Minnesota, habitat selection by antelopes, food selection by birds, nest site selection by birds, food selection by fish, and habitat selection by moose in Alaska..

- The examples are of design 1 (use and availability compared at the population level) according to the classification proposed in Chapter 1. However, similar examples can be carried out in principle using designs 2 (use measured for individual animals) and 3 (use and availability measured for individual animals).

- The assumptions involved in estimating and using resource selection functions are discussed (the X values do not change during the study, available units are correctly identified, used and unused units are correctly identified, the right X variables are measured, there is free and equal access to all available resource units, random sampling is used).

CHAPTER 4

STUDIES WITH RESOURCES DEFINED BY SEVERAL CATEGORIES

In this chapter designs and data analysis procedures are reviewed for studies of resource selection where each resource unit is classified into one of several categories (e.g., habitat types). These studies are the simplest that can be carried out, and are also the ones that have been used most often in the past. They are therefore worth considering in their own right, although they can be thought of as special cases of the more general types of study design that are discussed in later chapters.

4.1 Introduction

This chapter deals with the estimation of selection ratios as a means of studying selection in situations where each resource unit can be classified into one of several distinct categories. These selection ratios, which are equivalent to the forage ratios of Table 1.1, are defined such that for each resource unit the value of the ratio is proportional to the probability of that unit being utilized, given that the selecting organism has unrestricted access to the entire distribution of available units. The ratios can therefore be thought of as special cases of resource selection functions. Selection ratios are not the only way to analyse the results from selection experiments with resources in several categories. Logistic regression and log-linear modelling can also be used, as discussed in later chapters.

4.2 Sampling Designs and Protocols

Three general designs and three sampling protocols have been discussed in Chapter 1. The type of study design depends on whether or not results are recorded for individual animals. With design I, data on the availability of resource units and their use are recorded at the population level, i.e., animals are not uniquely identified and use of resources is recorded for the population of animals under study. Hence any inferences inferences can concern only the overall use of resources by all animals. With design II, data on the use of resources is recorded for each of several animals but availability is measured at the population level, and implicitly assumed to be approximately the same for each animal. With design III, data on both use and availability is collected for each of several animals. With designs II and III we assume that the animals under study are selected independently with equal probability from a single population so that the animals can provide the replication that is needed in order to make inferences concerning the use of resources by the population of animals.

With sampling protocol A (SPA), selection is studied by comparing a sample of used resource units with a sample or census of available resource units. With sampling protocol B (SPB), the comparison is between a sample of unused resource units and a sample or census of available units. With sampling protocol C (SPC) the comparison is between a sample of used resource units and a sample of unused resource units.

This chapter concentrates on the consideration of studies using SPA because these are by far the most common when resource units are considered to be divided into several categories on the basis of one factor such as habitat type. However there are some important considerations with the other two types of sampling protocol.

First, if SPB is employed then the selection indices that can be estimated give the relative probabilities of different types of unit being unused, instead of relative probabilities of them being used. These indices may not be as satisfactory as indices for use, but useful interpretations are still possible, for example, to order resource units in terms of their estimated probability of use.

When SPC is employed it may well be possible in some cases to argue that taking a sample or census of unused resource units is for all intents and purposes the same as taking a sample or census of available units because the proportion of available units that are actually used may be very small. In that case, a study using SPC can be treated as one using SPA with a negligible error.

4.3 Choice of Estimators for Selection Ratios

Selection ratios are ratios of random variables. It is therefore useful to review some results concerning the estimation of ratios and differences between these ratios, because these results are not ordinarily available in introductory textbooks on statistics. A general notation that will be used often is that a capital letter such as X refers to a random variable, while a small letter such as x refers to a specific value of the random variable.

Let (Y,X) denote a pair of random variables measured on the units in a population of interest, such as the animals in a population. Suppose that it is desirable to estimate $R = \mu_Y / \mu_X$, the ratio of the population mean of Y to the population mean of X, using data $(y_1,x_1), ..., (y_n,x_n)$ from a random sample of n units. One obvious estimator of R is then

$$\hat{R} = \bar{y} / \bar{x}, \tag{4.1}$$

where $\bar{x}$ and $\bar{y}$ are the sample means for X and Y. This estimator is often recommended in statistics texts in preference to

$$\hat{R}' = \sum_{j=1}^{n} (y_j / x_j) / n,$$

the sample mean of (y_j/x_j), because $\hat{R}$ will usually have less bias and a smaller variance than $\hat{R}'$. However, the choice between $\hat{R}$ and $\hat{R}'$ is not so clear in practice for estimating selection ratios because variation in these ratios from animal to animal may be just as interesting as the mean value. The choice therefore depends on whether the researcher is just interested in the selection ratio for the entire population of animals under study, or in the selection ratios of individual animals. There is no single choice that is always best. It depends on the circumstances.

4.4 Variance of a Ratio and the Difference of Ratios

The variance of $\hat{R}$ can be estimated by

$$\mathrm{var}(\hat{R}) = \hat{R}^2 \{(s_Y / \bar{y})^2 + (s_X / \bar{x})^2 - 2r_{XY}s_Ys_X /(\bar{y}\ \bar{x})\}/n, \tag{4.2}$$

where s_Y and s_X are the sample standard deviations for Y and X and r_{XY} is the sample correlation coefficient. Alternatively, it is sometimes convenient to estimate the variance using the equation

$$\mathrm{var}(\hat{R}) = \{ \sum_{j=1}^{n} (y_j - \hat{R}x_j)^2/(n - 1)\}\{1/(n\ \bar{x}^2)\}, \tag{4.3}$$

(Cochran, 1977, p. 153).

In many applications of ratios there is no random sample of (x_i,y_i) pairs from n units. Rather, two summary statistics X and Y are available, with known or estimated variances and covariance. In that case the ratio of the means can be estimated by

$$\hat{R} = y/x$$

and its variance can be approximated by

$$\mathrm{var}(\hat{R}) = \hat{R}^2 [\mathrm{var}(Y) / y^2 + \mathrm{var}(X) / x^2 - 2\mathrm{cov}(X,Y) / (y\ x)], \tag{4.4}$$

where var(X) and var(Y) are the variances for X and Y, respectively, and cov(X,Y) is the covariance between Y and X. If X and Y are calculated from independent data then cov(X,Y) = 0.

Another situation that occurs in this chapter is where the difference between two ratios of the form of equation (4.1) is considered. Thus consider the random variable

$$\hat{D} = \hat{R}_1 - \hat{R}_2 = (y_{1+} / x_{1+}) - (y_{2+} / x_{2+}), \tag{4.5}$$

where

$$y_{1+} = y_{11}+y_{12}+...+y_{1n},$$

$$x_{1+} = x_{11}+x_{12}+...+x_{1n},$$

$$y_{2+} = y_{21}+y_{22}+...+y_{2n},$$

and

$$x_{2+} = x_{21}+x_{22}+...+x_{2n}),$$

and where x_{1j}, y_{1j}, x_{2j} and y_{2j} are observations made on the jth unit in a random sample of n units. Here it can be shown using a standard Taylor series method (Manly, 1985, p. 408) that the variance can be approximated by

$$\mathrm{var}(\hat{D}) = \{n/(n - 1)\} \sum_{j=1}^{n} \{(y_{1j} - \hat{R}_1 x_{1j}) / x_{1+} + (y_{2j} - \hat{R}_2 x_{2j})/ x_{2+} \}^2. \qquad (4.6)$$

A further possibility here is that the difference being considered has the simpler form

$$\hat{D} = (x_1 / y_1) - (x_2 / y_2),$$

where X_1, Y_1, X_2 and Y_2 are random variables with known or estimated variances and covariances. The Taylor series approximation method then yields

$$\begin{aligned}\mathrm{var}(\hat{D}) = {} & \{1/y_1^2\}\mathrm{var}(X_1) + \{x_1^2/y_1^4\}\mathrm{var}(Y_1) + \{1/y_2^2\}\mathrm{var}(X_2) \\ & + \{x_2^2/y_2^4\}\mathrm{var}(Y_2) - 2\{x_1/y_1^3\}\mathrm{cov}(X_1,Y_1) \\ & - 2\{1/(y_1y_2)\}\mathrm{cov}(X_1,X_2) + 2\{x_2/(y_1y_2^2)\}\mathrm{cov}(X_1,Y_2) \\ & + 2\{x_1/(y_1^2y_2)\}\mathrm{cov}(Y_1,X_2) - 2\{(x_1x_2)/(y_1^2y_2^2)\}\mathrm{cov}(Y_1,Y_2) \\ & - 2\ \{x_2/(y_2^3)\}\mathrm{cov}(X_2,Y_2). \qquad (4.7)\end{aligned}$$

This is a complicated equation, but in practice some of the covariances may be zero, leading to some simplification.

With a small population it is desirable to apply a finite population correction to equations (4.2), (4.3) and (4.6). This involves multiplying the right-hand sides of these equations by the finite population correction (1 - n/N), where N is the population size.

4.5 Chi-Squared Tests

Use is made of chi-squared goodness of fit tests at various places in this chapter to test for significant selection, or to test for whether different animals are using resources differently. The form of test statistic that is most commonly used for this purpose is the Pearson statistic which takes the form

$$X_P^2 = \Sigma(O_i - E_i)^2/E_i,$$

where O_i is an observed sample frequency, E_i is the expected value of O_i according to the hypothesis being considered, and the summation is over all the data frequencies. However, for most purposes we choose instead to use the log-likelihood statistic

$$X_L^2 = 2\Sigma O_i \log_e(O_i/E_i),$$

where again the summation is over all the data frequencies.

Both statistics have the same number of degrees of freedom (df) and both have chi-squared distributions for large samples if the null hypothesis being tested is correct. Furthermore, in practice X_P^2 and X_L^2 will give very similar values unless either the expected frequencies are very small or the differences between the observed and expected frequencies are very large. Therefore, it will usually not make much difference which statistic is used in terms of whether results are significant or not.

In this book the preference for using X_L^2 rather than X_P^2 is based on the fact that this is justified by the general theory of log-likelihood tests as reviewed in Chapter 2, and this theory is used extensively in the chapters that follow. Indeed, all the chi-squared statistics that are used in the present chapter can be thought of as differences between maximized log-likelihoods for fitted models based on different assumptions.

4.6 Design I with Known Proportions of Available Resource Units

The most common sampling plan used to study resource selection is the special case where (i) there is no unique identification of data collected from different animals, (ii) the proportions of available units in different resource categories are known, and (iii) a random sample of used resource units is taken. This is then design I, with sampling protocol A.

Suppose that this is the situation, with the following notation:

Symbol	**Definition**
A_i	the number of available resource units in category i, for i = 1, 2, ..., I
A_+	the size of the total population of available resource units
π_i	A_i/A_+, the proportion of available resource units that are in category i
U_i	the number of used resource units in category i in the population
U_+	the total number of used resource units in the population
u_i	the number of resource units in category i in the sample of used units
u_+	the total number of used resource units sampled
o_i	u_i/u_+, the proportion of the sample of used resource units that are in category i
w_i*	the proportion of the population of available resource units in category i that are used (the resource selection probability function)

Then the total number of used resource units in category i is

$$U_i = A_+ \pi_i w_i^*, \tag{4.8}$$

so that the selection probability for category i is

$$w_i^* = U_i / (A_+ \pi_i).$$

The proportion of used resource units in the population that are in category i is U_i/U_+, where this can be estimated by the proportion of resource units in this category in the sample of used units (o_i). Hence, an estimate of U_i is o_iU_+, and an estimate of the resource selection probability function is

$$\hat{w}_i^* = (o_i U_+)/(A_+ \pi_i).$$

In practice, it is likely that neither of the population totals U_+ nor A_+ will be known. However, the selection ratio

$$\hat{w}_i = o_i / \pi_i \tag{4.9}$$

can still be calculated, where this is $\hat{w}_i^*$ multiplied by the unknown constant A_+/U_+. This selection ratio therefore gives the resource selection function (the relative probability of selection for category i.

Selection ratios are an old intuitive approach to analysis of resource selection related to a single categorical variable but generally have not been recognized as giving rise to relative probabilities of selection. The names used for o_i/π_i have varied from forage ratios (Hess and Swartz, 1940), and selectivity indices (Manly *et al.*, 1972), to preference indices (Hobbs and Bowden, 1982).

A useful way of presenting selection ratios is with them standardized so that they add to 1. This leads to Manly's standardized selection ratio

$$B_i = \hat{w}_i / (\sum_{i=1}^{I} \hat{w}_j), \tag{4.10}$$

which has the interpretation of being the estimated probability that a category i resource unit would be the next one selected if it was possible to make each of the types of resource unit equally available. To understand this interpretation, suppose that a resource unit is randomly chosen from the population of used units. Then from equation (4.8), the probability of this unit being in resource category i is

$$\begin{aligned} U_i / U_+ &= A_+ \pi_i w_i^* / (\Sigma A_+ \pi_j w_j) \\ &= \pi_i w_i^* / (\Sigma \pi_j w_j^*) \\ &= \pi_i w_i / (\Sigma \pi_j w_j), \end{aligned}$$

where the last line follows because the value on the right-hand side of the equation is not affected by multiplying all of the w_i^* values by a constant. Hence, if all categories are equally available, so that $\pi_1 = \pi_2 = ... = \pi_I$, then

$$U_i / U_+ = w_i / (\Sigma w_j).$$

It follows that B_i of equation (4.10) gives the estimated probability that a randomly selected used resource unit would be in category i if all categories were equally frequent in the original population of available resource units (Manly *et al.*, 1972).

Example 4.1 Habitat Selection by Moose

The situation considered in Example 3.1 is an example of the type of design just described. It may be recalled that Neu *et al.* (1974) considered selection of habitat by Moose (*Alces alces*) on a site surrounding the Little Sioux Burn in northeastern Minnesota during the Winter of 1971-72, and collected the data shown in Table 3.1. The proportion of acreage in the ith habitat type, π_i, was determined by planimeter and is assumed to be known accurately. Locations of moose and moose tracks were plotted on a map of the study area during aerial surveys. Each group of moose or moose tracks was considered as an independent observation in the sample of 'used' points. This study is an example of design I (used points are not identifiable to unique animals) and sampling protocol A (a sample of used points is contrasted to the defined universe of available points in the study area).

Table 4.1 shows the calculations for the selection ratios of equations (4.9) and (4.10) for this example. It can be seen that the edge habitat in the burn appears to have been selected with about three times the probability of the interior habitat in the burn ($B_2 = 0.326$, $B_1 = 0.110$). Similarly, the edge habitat out of the burn is about four times as likely to be selected as is the habitat further from the burn ($B_3 = 0.433$, $B_4 = 0.131$). However, edge habitat in the burn and edge habitat out of the burn are selected with approximately equal probability ($B_2 = 0.326$, $B_3 = 0.433$). The same relationships hold between the unstandardized $\hat{w}_i$ values. Procedures to test whether there are statistically significant differences between these pairs of values are discussed below.

Table 4.1 Estimation of selection indices for the occurrence of moose tracks on burned, unburned, and peripheral portions of a 33,200 acre area surrounding the Little Sioux Burn in northeastern Minnesota in the winter of 1971-72.

	Population proportion	Sample count	Used sample proportion	Selection index	Standardized index
Habitat	π	u	o	$\hat{w}$	B
In burn, interior	0.340	25	0.214	0.629	0.110
In burn, edge	0.101	22	0.188	1.866	0.326
Out of burn, edge	0.104	30	0.256	2.473	0.433
Out of burn, further	0.455	40	0.342	0.750	0.131
Total	1.000	117	1.000	5.718	1.000

Example 4.2 Selection of Escape Cover by Quail

For a second example, consider the selection ratios in Table 4.2; these are computed for a subset of data on selection of escape cover by California quail (*Callipepla californica*) from a study by Stinnett and Klebenow (1986). It can be seen that the shrubland habitat was estimated to be selected with about twice the probability of riparian habitat ($B_4 = 0.034$, $B_5 = 0.061$), and field border was approximately 30 times more likely to be selected than was riparian ($B_6 = 0.889$).

4.7 Tests on Proportions of Used Units

The standard analysis in much of the recent literature for the situation where population proportions of resource categories are known and a sample of used units is taken has been to conduct a chi-squared test of the null hypothesis that animals are randomly selecting habitat in proportion to availability. If the chi-squared test is significant then it is followed by the computation of simultaneous confidence intervals for the population proportions of used resources of different types and the comparison of these with the available proportions (Neu *et al*., 1974; Byers *et al*., 1984).

Table 4.2 Relative probabilities of selection of escape cover by quail.

	Sample	Expected	Population	Selection	Standardized
	Count	count	proportion	ratio	index
Escape cover	u	πu_+	π	$\hat{w}$	B
Pasture	0	23.5	0.362	0.000	0.000
Disturbed	0	4.3	0.066	0.000	0.000
Farmstead	2	3.7	0.057	0.540	0.016
Riparian	19	16.2	0.249	1.174	0.034
Shrubland	36	17.0	0.262	2.114	0.061
Field border	8	0.3	0.004	30.769	0.889
	65	65.0	1.000	35.597	1.000

The chi-squared statistic usually used takes the form

$$X_P^2 = \sum_{i=1}^{I} (u_i - u_+ \pi_i)^2 / u_+ \pi_i,$$

with I - 1 df, where I is the number of resource categories. Thus the observed number of used resource units of type i (u_i) is compared with the expected number ($u_+\pi_i$) under the hypothesis of no selection. If X_P^2 is significantly large when compared to the percentage points of the chi-squared distribution then this indicates that there is a departure from the null hypothesis that selection is random.

As discussed in Section 4.5, the alternative statistic

$$X_L^2 = 2 \sum_{i=1}^{I} u_i \log_e\{u_i / (u_+\pi_i)\} \tag{4.11}$$

may be preferred on the grounds that this is consistent with the use of log-likelihood tests in the chapters that follow. In practice X_P^2 and X_L^2 will usually give almost thesame numerical results.

The condition for the chi-squared test to be valid is the usual one that the expected frequencies should be five or more. If this condition is not met then the test may still be valid, but obviously the outcome of the test should be treated with a certain amount of reservation. In practice, categories are often combined or dropped so that the expected frequencies are five or more in the resulting combinations.

The sample proportion o_i will have a standard error that is approximately given by

$$se(o_i) = \sqrt{\{o_i (1 - o_i) / u_+\}}. \tag{4.12}$$

Hence approximate 100(1-α)% confidence intervals for the population proportions of used resource units of different types can be taken as

$$o_i \pm z_{\alpha/2} \sqrt{\{o_i (1 - o_i) / u_+\}}, \tag{4.13}$$

where $z_{\alpha/2}$ is the percentage point of the standard normal distribution that is exceeded with probability α/2. Because there are I types of resource unit, a Bonferroni adjustment to the confidence level may be desirable, as discussed in Chapter 2. In particular, if the value of α is set at 5%/I then there will be a probability of about 0.95 that all I confidence intervals will include their respective population ratios. The Bonferroni adjustment results in very conservative (wide) confidence intervals. Whether or not this is appropriate depends on the application.

The confidence intervals defined by (4.13) will have the required level of significance providing that all of the sample proportions o_i are approximately normally distributed, and the standard errors calculated using equation (4.12) are close to the true standard errors. Because u_i should have a binomial distribution with mean estimated by u_+o_i and variance estimated by $u_+o_i(1 - o_i)$, standard statistical theory suggests that these conditions will be met providing that $u_+o_i(1 - o_i)$ is greater than about 5. This will occur if the observed number of used units exceeds five for each resource category. Basically, then, this is the same requirement as was mentioned earlier for the chi-squared test. It may of course happen that there are fewer than five resource units in some categories. In that case the corresponding confidence intervals must be regarded as merely indicative of the level of sampling errors to be expected or categories must be dropped or combined.

So far in this discussion on variances, tests of significance and confidence intervals it has been implicitly assumed that u_+, the sample size for the used resource units, is fixed in advance of sampling, so that u_i has a binomial distribution. However, in practice it will often be the case that researchers take whatever sample size they can get. This raises the question about the validity of inferences when sample sizes are random variables.

There are two facts here that justify the use of the equations given above. First, it can be argued that the inferences made are conditional on the observed sample size, and therefore rely on probability statements that apply for the restricted set of possible sets of data that can be obtained subject to the sample size observed. Second, it can be argued that if the total sample sizes is not fixed then the u_i values can be expected to have independent distributions that are approximately Poisson. It can then be shown that the statistic X_L^2 of equation (4.11) will still approximately have a chi-squared distribution with I - 1 df, and equation (4.12) will still be a valid estimate of the standard error of o_i. Consequently, all the inference procedures will still be valid. Furthermore, a reasonable condition for the chi-squared and normal distribution approximations to hold will still be that the number of used units sampled is five or more in each resource category.

Example 4.2 (Continued) Selection of Escape Cover by Quail

Consider again the data in Table 4.2 on the escape cover used by quail. Here the X_L^2 statistic of equation (4.11) is 112.35, with five df, which is highly significant when compared with the percentage points of the chi-squared distribution. As three of the expected frequencies are less than five, the interpretation of this result requires some caution, although there does seem to be very clear evidence of selection because X_L^2 is very large.

If $\alpha = 0.05/6 = 0.0083$ is used with the six confidence intervals (4.13) then it is possible to be approximately 95% confident that all the limits contain the population values. The z_α value is then 2.64, and the limits are as shown in Table 4.3. It can be seen that shrubland and field border were apparently selected significantly more often than is expected from the population proportions of these habitats. For example, for shrubland, $\pi_5 = 0.262$ is below the lower limit of the confidence interval 0.391 to 0.717. However, the fact that three observed frequencies are less than five must lead us to treat the significance of this result with some reservations.

This example provides a good illustration of one problem that can occur when the proportions of used resource units in different categories are compared with the proportions available. Thus suppose that the researchers consider dropping pasture and disturbed habitat from the study because these types are known to be rarely selected. As shown in Table 4.3, the observed proportion of times that the riparian habitat is chosen (0.292) exceeds the expected proportion (0.249) when six habitat types are considered. Thus this habitat seems to be favoured to some extent. However, as shown in Table 4.4, if pasture and disturbed habitats are dropped from the analysis then the observed proportion of riparian choices stays the same at 0.292 but the expected proportion increases to 0.435. Now the riparian habitat seems to be avoided. This effect of a resource category switching from preferred to avoided, or vice-versa, is not uncommon when a large but seldom used resource category is removed from the analysis. It means, of course, that the interpretation of data may depend rather crucially on decisions that are made concerning what types of habitat are available to an animal.

One of the advantages of working with selection ratios is that this effect is largely avoided when decisions must be made on whether or not to include rarely used categories in the analysis. Thus the selection ratios for farmstead, riparian, shrubland and field border that are shown in Table 4.2 (with pasture and disturbed habitats included) are proportional to the selection ratios that are shown in Table 4.4 (with pasture and disturbed habitats excluded). This is seen by the fact that the standardized ratios B_i are identical in Tables 4.2 and 4.4.

4.8 Inferences Concerning Selection Ratios

The standard error of $\hat{w}_i$ can be approximated by

$$se(\hat{w}_i) = se(o_i / \pi_i) = \sqrt{}\{o_i (1 - o_i)/(u_+ \pi_i^2)\}. \qquad (4.14)$$

Hence an approximate 100(1-α)% confidence interval for a single selection ratio is of the form

$$\hat{w}_i \pm z_{\alpha/2}\, se(\hat{w}_i). \qquad (4.15)$$

The condition for these limits to have about the right level of confidence is the same as the condition for the binomial proportion o_i to be approximately normally distributed, because if that is the case then o_i/π_i will also be approximately normally distributed. As discussed earlier, a reasonable requirement is therefore that the number of used resource units is five or more in all resource categories. If this condition does not hold for some categories then the intervals are somewhat suspect for these categories, but will be reliable for the others.

Table 4.3 Bonferroni confidence intervals for population proportions of resource units used by quail.

	Population proportion	Used proportion	Bonferroni confidence limits	
Escape cover	π	o	Lower	Upper
Pasture	0.362	0.000	-	-
Disturbed	0.066	0.000	-	-
Farmstead	0.057	0.031	0.000[1]	0.088
Riparian	0.249	0.292	0.143	0.441
Shrubland	0.262	0.554	0.391	0.717
Field border	0.004	0.123	0.015	0.231

[1] For farmstead a negative lower limit has been changed to 0.000. The limits for this habitat are unreliable because of the low sample count of used resource units.

Table 4.4 Selection of escape cover by quail with the pasture and disturbed habitat types removed.

Escape cover	u	π	o	$\hat{w}$	B
Farmstead	2	0.100	0.031	0.308	0.016
Riparian	19	0.435	0.292	0.677	0.034
Shrubland	36	0.458	0.554	1.209	0.061
Field border	8	0.007	0.123	17.582	0.889
Total	65	1.000	1.000	19.766	1.000

The selection coefficient $\hat{w}_i$ is significantly different from 1 if the confidence interval for w_i does not contain the value 1. Alternatively, an approximate test for the significance of the estimate involves comparing

$$(\hat{w}_i - 1) / \mathrm{se}(\hat{w}_i)$$

with critical values for the standard normal distribution, or

$$\{(\hat{w}_i - 1) / \mathrm{se}(\hat{w}_i)\}^2$$

with critical values for the chi-squared distribution with one df.

There is the possibility for confusion here because se($\hat{w}_i$) can be estimated either on the assumption that selection may occur, or on the assumption that there is no selection. In the first case, equation (4.14) is the appropriate one to use, with the proportion of category i resource units in the population of used units being approximated by the sample proportion o_i. In the second case, the expected value of o_i becomes equal to π_i, and equation (4.14) can be replaced by

$$\mathrm{se}(\hat{w}_i) = \surd\{(1 - \pi_i)/(u_+ \pi_i)\}. \qquad (4.16)$$

It can be argued that this last equation is the one to use when testing the hypothesis that there is no selection because it gives the standard error exactly if this hypothesis is true.

In practice, the researcher may be interested in the entire set of I selection ratios, w_i, for i from 1 to I. When this is the case, approximate simultaneous confidence intervals or tests can be constructed by use of the Bonferroni inequality that has been discussed in Section 4.7. Thus, subject to the conditions for the normal approximation to be valid, it is possible to be 100(1 - α)% confident that the intervals for all I selection ratios contain their respective true ratios if $z_{\alpha/2}$ is replaced in (4.15) by $z_{\alpha/(2I)}$. Similarly, if the significance levels used for testing

$$(\hat{w}_i - 1)/se(\hat{w}_i)$$

against the standard normal distribution is α/I then there will be a probability of only approximately α of getting any result significant when there is no selection. The decision to use the more conservative Bonferroni adjustment in the analysis depends on how conservative the researcher wishes to be in a particular application.

The discussion at the end of Section 4.7 concerning the validity of inferences when sample sizes are not fixed in advance carries over to inferences on selection ratios. Hence all the results that have been given in the present section apply equally well if u_+, the total sample size of used units, is not fixed in advance.

Example 4.2 (Continued) Selection of Escape Cover by Quail

Consider again Stinnett and Klebenow's (1986) study of the escape cover selected by quail with the data shown in Table 4.2. As there are I = 6 habitat types, it may be appropriate to test each selection ratio for significance using the (5/6)% = 0.8% level of significance, the Bonferroni adjustment. For example, the selection ratio for shrubland as escape cover is $\hat{w}_5 = 2.114$. Assuming that there is no selection, the standard error associated with this ratio is given by equation (4.14) to be

$$se(\hat{w}_5) = \sqrt{\{(1 - 0.262)/(65 \text{ x } 0.262)\}} = 0.208.$$

Hence the chi-squared statistic with one df to test for selection is

$$(\hat{w}_5 - 1)^2/se\ (\hat{w}_5)^2 = (2.114 - 1)^2 / 0.208^2 = 28.64.$$

As this is significantly large at the 0.8% level, this indicates that shrubland is used more than is expected from the availability of this habitat. Similar calculations using the conservative Bonferroni adjustment suggest that there is selection against pasture, and selection for shrubland and field border.

4.9 Comparison of Selection Ratios

Assuming that the sample of used units is a random sample from the population of used units, the observed counts for resource categories u_1 to u_I will follow a multinomial distribution, conditional on the total sample size u_+ being regarded as being fixed in advance. This means that the sample proportions o_1 to o_I will be multinomial proportions with estimated variances

$$var(o_i) = o_i\,(1-o_i)/u_+$$

and estimated covariances

$$\mathrm{cov}(o_i,o_j) = -o_io_j.$$

The variance of the difference between two selection ratios can therefore be estimated by

$$\mathrm{var}(\hat{w}_i - \hat{w}_j) = \mathrm{var}(o_i/\pi_i) - 2\mathrm{cov}(o_i/\pi_i,o_j/\pi_j) + \mathrm{var}(o_j/\pi_j),$$

$$= o_i(1 - o_i)/(u_+ \pi_i^2) - 2o_io_j/(u_+ \pi_i \pi_j) + o_j(1 - o_j)/(u_+ \pi_j^2). \qquad (4.17)$$

Using this equation, the null hypothesis of no difference in the probabilities of selection of the ith and the jth categories, i.e., that $w_i = w_j$ for $i \neq j$, can be tested by comparing the statistic

$$(\hat{w}_i - \hat{w}_j)^2 / \mathrm{var}(\hat{w}_i - \hat{w}_j)$$

with the critical values of the chi-squared distribution with one df. Also, an approximate 100(1-α)% confidence interval for the difference $w_i - w_j$ is given by

$$(\hat{w}_i - \hat{w}_j) \pm z_{\alpha/2}\mathrm{se}(\hat{w}_i - \hat{w}_j), \qquad (4.18)$$

where $\mathrm{se}(\hat{w}_i - \hat{w}_j)$ is the square root of the variance given in equation (4.17).

The validity of these confidence intervals depends on the assumption that the estimators $\hat{w}_i$ are normally distributed, which in turn depends on the sample proportions o_i being normally distributed. As discussed before, a reasonable requirement for this to hold is that the number of used units should be five or more for each resource category. Hence if this condition is not met for some of the resource categories then any of the confidence intervals (4.18) involving these categories must obviously be treated with some reservations or categories must be dropped or combined.

Using Bonferroni's inequality, a procedure very similar to that used to compare means in analysis of variance can be suggested for comparing the selection ratios. This involves ranking the selection ratios from the smallest to the largest and comparing them two at a time by use of the confidence intervals (4.18), replacing $z_{\alpha/2}$ by $z_{\alpha/(2I')}$ where

$$I' = I(I - 1)/2$$

is the total number of comparisons being made (the number of combinations of I categories taken two at a time). The selection ratios w_i and w_j are then declared significantly different if the confidence interval for $w_i - w_j$ does not contain zero.

As was the case with inferences concerning sample proportions of used resource units of different types (Section 4.7) and inferences concerning individual selection ratios (Section 4.8), all the results in the present section can be justified for use in situations where u_+, the total number of used units sampled, is not fixed in advance. In particular, the variance equation (4.17) can be shown to give a valid estimate of variance if the u_i values follow independent Poisson distributions.

Example 4.2 (Continued) Selection of Escape Cover by Quail

The results of the tests just described are shown in Table 4.5 for the selection ratios obtained from Stinnett and Klebenow's (1986) study of selection of escape cover by quail. The differences between selection ratios that are connected with a vertical bar have an approximate (100 - 5/15)% = 99.67% confidence interval that includes zero. Since there are 15 pairwise comparisons, this procedure gives about a 0.95 probability that all the confidence limits include their respective population differences between selection ratios.

Table 4.5 Comparison between selection ratios using confidence limits chosen so that the probability of all the pairwise intervals including the population difference is 0.95. Comparisons involving pasture, disturbed and farmstead are unreliable because of low or zero counts of used units.

Escape	$\hat{w}$				
Pasture	0.000	│			
Disturbed	0.000	│			
Farmland	0.540	│	│		
Riparian	1.174		│		
Shrubland	2.114			│	
Field border	30.769				│

From Table 4.5 it can be seen that:

(a) the probabilities of selection for shrubland and field border are significantly different, but both are significantly larger than the probability of selection for the other habitat types;

(b) the probability of selection of riparian is not significantly different from the probability of selection of farmland; and

(c) there are no significant differences between the probabilities of selection for pasture, disturbed habitats and farmstead.

Because the sample counts for pasture, disturbed habitats and farmstead are less than five, the comparisons involving these three categories must be regarded as indicative only.

4.10 Design I with Estimated Proportions of Available Resource Units

In the past, studies using design I and sampling protocol A often involved estimating the proportions π_i of available units using a random sample of available resource units, or some other sampling plan. For example, Marcum and Loftsgaarden (1980) estimated the proportion of different types of habitat in a study area by locating a number of random points on a map of the study area and counting the number of points hitting each type. At the present time, the proportion of different types of habitat in a study area would most likely be determined using a geographical information system (GIS). The data would then be analysed as recommended above, under the assumption that the proportions of available units are known. In effect, any measurement errors of proportions in the GIS data layers would be ignored under the assumption that they are small relative to the sampling errors involved in the estimation of the proportions of used units in each category.

Assume that the proportions of different types of available units are estimated from simple sample proportions. The usual two way chi-squared test can then be used to see whether the sample proportions for used units are significantly different from the sample proportions for available units. If a random sample of u_+ used units yields u_i units in category i, and a random sample of m_+ available units yields m_i in the same category, then the usual Pearson chi-squared statistic can be written as

$$X_P{}^2 = \sum_{i=1}^{I} [\{u_i - E(u_i)\}^2/E(u_i) + \{m_i - E(m_i)\}^2/E(m_i)],$$

where

$$E(u_i) = (m_i + u_i)\, u_+/(u_+ + m_+)$$

is the expected value of u_i, and

$$E(m_i) = (m_i + u_i)m_+/(u_+ + m_+)$$

is the expected value of m_i, on the hypothesis of no selection. If this statistic is significantly large when compared with the chi-squared distribution with I - 1 df then there is evidence that selection is occurring. The usual conditions for the validity of the chi-squared test apply, so that it should be reliable if all the expected frequencies are five or more. For the reasons discussed in Section 4.5, it is reasonable to replace $X_P{}^2$ with the log-likelihood statistic

$$X_L{}^2 = 2 \sum_{i=1}^{I} \{u_i \log_e\{u_i/E(u_i)\} + m_i \log_e\{m_i/E(m_i)\}, \qquad (4.19)$$

which has the same df and validity conditions as $X_P{}^2$.

The sample of available resource units gives the estimator

$$\hat{\pi}_i = m_i / m_+, \qquad (4.20)$$

of π_i, with estimated standard error

$$se(\hat{\pi}_i) = \hat{\pi}_i (1 - \hat{\pi}_i) / m_+. \qquad (4.21)$$

An estimator of w_i is therefore the ratio of random variables,

$$\hat{w}_i = o_i / \hat{\pi}_i, \qquad (4.22)$$

where, as before, $o_i = u_i / u_+$ is the proportion of category i resource units in the random sample of u_+ used units. The standard error of the estimated selection ratio can be approximated from the general formula (4.4) for a ratio, which provides

$$\text{se}(\hat{w}_i) = \hat{w}_i \surd\{(1 - o_i)/(o_i\, u_+) + (1 - \hat{\pi}_i\,)\,/\,(\hat{\pi}_i\, m_+)\},$$

$$= \hat{w}_i \surd\{1/u_i - 1/u_+ + 1/m_i - 1/m_+\}. \qquad (4.23)$$

There will zero covariance between o_i and $\hat{\pi}_i$ if they are estimated from independent samples.

Equation (4.23) can also be shown to be a valid estimate of the standard error of the estimated selection ratio for cases where the counts of used resource units u_1 to u_I and the counts of available resource units m_1 to m_I have independent Poisson distributions. Therefore, situations are covered where the total sample sizes u_+ and m_+ of used and available resource unit are not fixed in advance. The situation in this respect is therefore the same as for cases where population proportions of available resource units are known accurately.

Approximate simultaneous confidence intervals on the selection ratios can be constructed following the procedures suggested before. Thus the confidence intervals

$$\hat{w}_i \pm z_{\alpha/(2I)}\text{se}(\hat{w}_i) \qquad (4.24)$$

can be considered, for i from 1 to I. The selection coefficient $\hat{w}_i$ is then declared significantly different from 1 if the confidence interval on w_i does not contain the value 1. The Bonferroni inequality suggests that this procedure will give a probability of approximately 1 - α that the population selection ratios will all be within their respective intervals, and that the probability of declaring any result significant will be approximately α if there is no selection.

The validity of the confidence intervals (4.24) depends on the standard errors from equation (4.23) being accurate, and the normality of the distributions of the estimators. A minimum condition is that there are at least five resource units in each category both in the sample of used units and in the sample of available units, because this will at least ensure that the o_i and $\hat{\pi}_i$ values are approximately normally distributed, with reasonable estimates of their standard errors. However, the fact that $\hat{w}_i$ is a ratio of random variables suggests that some more stringent conditions may be needed. This matter is discussed further in the example that follows.

Estimation of the availabilities π_i by sampling protocols such as quadrat or transect sampling which do not use random points as the basic sampling unit will require a different equation for the estimation of the standard error of $\hat{w}_i$. A discussion of all possible cases is not realistic here. However, if an estimate of $\text{se}(\hat{\pi}_i)$ is available from the implemented sampling design then it can be substituted into equation (4.4) to obtain

$$\text{se}(\hat{w}_i) = \hat{w}_i \surd\{(1 - o_i\,)/\,(o_i\, u_+) + \text{se}(\hat{\pi}_i)^2\,/\,\hat{\pi}_i^{\,2}\}. \qquad (4.25)$$

This equation can then be used to obtain confidence intervals for population selection ratios in the usual way. In some cases where the estimation of the available proportions is by some special method it may be best to consider estimating the variances of the estimates by bootstrapping, as discussed in Section 2.11.

The standard error of the difference between two selection ratios can be calculated using the general equation (4.7) by setting $x_1 = o_i$, $y_1 = \hat{\pi}_i$, $x_2 = o_j$ and $y_2 = \hat{\pi}_j$. Then assuming that o_i, o_j, $\hat{\pi}_i$ and $\hat{\pi}_j$ are simple proportions, with the first two of these quantities being independent of the second two, there are the following estimates of variances and covariances:

$$\text{var}(o_i) = o_i\,(1 - o_i)\,/\,u_+,$$

$$\mathrm{var}(\hat{\pi}_i) = \hat{\pi}_i\,(1 - \hat{\pi}_i) / m_+,$$

$$\mathrm{cov}(o_i,o_j) = -o_io_j/u_+,$$

and

$$\mathrm{cov}(\hat{\pi}_i,\hat{\pi}_j) = -\,\hat{\pi}_i\,\hat{\pi}_j / m_+.$$

In addition, the covariances between o_i and $\hat{\pi}_j$ terms are zero, for all i and j. Using these results equation (4.7) produces the estimated variance

$$\mathrm{var}(\hat{w}_i - \hat{w}_j) = \{\hat{w}_i / \hat{\pi}_i + \hat{w}_j / \hat{\pi}_j - (\hat{w}_i - \hat{w}_j)^2\}/u_+ + \{\hat{w}_i^2 / \hat{\pi}_i + \hat{w}_j^2 / \hat{\pi}_j - (\hat{w}_i - \hat{w}_j)^2\}/m_+.$$

Bonferroni confidence intervals for the whole set of possible differences between selection ratios can be calculated as

$$(\hat{w}_i - \hat{w}_j) \pm z_{\alpha/(2I')}\mathrm{se}(\hat{w}_i - \hat{w}_j),$$

where I' is the number of such differences. The procedure is exactly the same as described in section 4.9.

Example 4.3 Selection of Forest Canopy Cover by Elk

Marcum and Loftsgaarden (1980) used data from Marcum (1975) as an example where 200 random (available) points were located on a map of the study area which contained a complex mosaic of forest-canopy cover, and compared these with 325 elk (*Cervus elaphus*) locations (used points) in the same region. Four categories of habitat were considered, corresponding to canopy coverages of 0%, 1-25%, 26-75% and 75-100%. The data are shown in Table 4.6, together with the estimated selection ratios, standard errors and confidence intervals that are obtained using equations (4.22) to (4.24). From equation (4.19), $X_L^2 = 21.9$, which is very highly significant when compared with percentage points of the chi-squared distribution with three df. There is therefore clear evidence of selection relative to available habitat.

From the results in the table it appears that elk used the 0% canopy cover class significantly less than in proportion to its availability because the upper limit of the confidence interval for w_1 is below the value one. Similarly, elk used the 26-75% canopy cover class significantly more than in proportion to availability because the lower limit of the confidence interval for w_3 is above one. There appears to be no significant selection either for or against the remaining two classes. These conclusions agree with the analysis of Marcum and Loftsgaarden (1980) based on simultaneous confidence intervals for the differences $(\pi_i - o_i)$ rather than on the ratios $w_i = o_i/\pi_i$.

Because there were only three used points in the 0% canopy cover class, there is some question about the validity of the confidence interval for the selection ratio in this class. A small simulation experiment was therefore carried out to investigate this matter. What was done was to generate 500 sets of data similar to the observed data, with the expected sample frequencies of available and used resource units set equal to the observed frequencies, and the actual counts following independent Poisson distributions. Selection ratios and their standard errors were then estimated for each set of artificial data, together with the z-scores

$$z_i = (\hat{w}_i - w_i)/\mathrm{se}(\hat{w}_i),$$

where the 'true' selection ratios w_i were the values obtained from the original data.

The idea behind calculating the z-scores was that if these scores have a standard normal distribution then the confidence intervals (4.24) are valid. In fact, it was found that the z-scores had distributions that were very close to normal for the estimates of all four selection ratios, although there was a slight excess of values less than -3, counterbalanced by some lack of values of +3 or more. This is shown in Figure 4.1, which compares the distribution obtained for all the observed z-scores with the standard normal distribution. The only real problem that was observed in the simulation was an occasional zero value of u_1. In these cases $\hat{w}_1 = 0$, se($\hat{w}_1$) = 0, and z_1 becomes undefined. This occurred 12 times for the 500 sets of simulated data.

Table 4.6 Estimated percentages of different habitat types, and selection indices for forest-overstory canopy coverage classes available to elk. The percentages available were estimated by a random sample of m = 200 points on a map of the study area. A sample of u_+ = 325 points selected by elk was used to estimate the proportions of elk observations in each class. The confidence limits for each selection ratio have confidence level (100-10/4)% = 97.5% in order that there is a 0.9 probability that all four limits include the population selection ratio.

Canopy cover class	m	$\hat{\pi}$	u	$\hat{o}$	$\hat{w}$	B	se($\hat{w}$)	Confidence limits Lower	Confidence limits Upper
0%	15	0.075	3	0.009	0.120	0.038	0.077	0.000[1]	0.296
1-25%	61	0.305	90	0.277	0.908	0.289	0.127	0.624	1.191
26-75%	84	0.420	181	0.557	1.326	0.422	0.128	1.039	1.613
> 75%	40	0.200	51	0.157	0.785	0.250	0.150	0.449	1.121
Total	200	1.000	325	1.000	3.139	1.000			

[1] A negative lower limit for the confidence interval for 0% has been replaced by 0.000 since negative values for the selection indices are impossible.

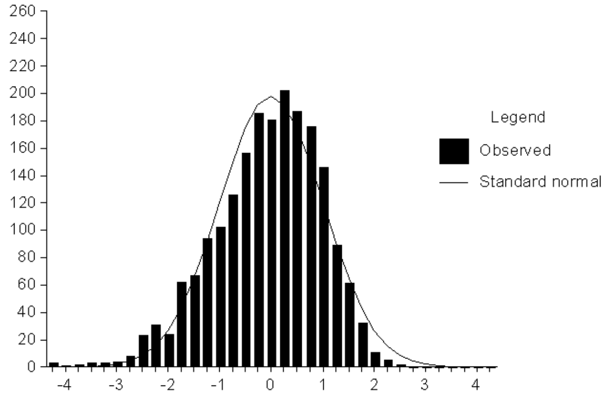


Figure 4.1 Comparison of the distribution of values of $z_i = (\hat{w}_i - w_i)/se(\hat{w}_i)$ with the standard normal distribution. If z_i follows a standard normal distribution then confidence limits (4.24) for population selection ratios will be reliable.

Overall, it seems that the confidence limits are reasonably reliable even though there is a sample count of less than five used units in the 0% canopy cover class.

4.11 Design II with Sampling Protocol A

With sample design II data are available on the selection of resource units by individual animals. For example, the proportions of resource categories might be estimated in the home ranges of individual animals or in 2 km circles centred at nest or feeding sites. Thus, assume a random sample of n animals is obtained from the population and the resource units used by the jth animal are determined to estimate the proportions for different types of resource categories. For example, the proportions of each habitat type might be measured in each home range or in each circle centred at a nest or feeding site. The population of resource units available to the population is also sampled or censussed to estimate the proportions of units in each of several categories.

With the advance of GIS we expect that most studies in the future will measure (i.e., census) the proportions of habitat types in both the areas used by the animals and in the population of available habitat. The measurement error should then be small compared to the variation in resource use among animals so that it can be ignored.

Under Design II, the animals are the primary sampling units and sampling variance arises because not all animals have exactly the same selection ratios for the resource categories. In other words, statistical inferences are based on the use of animals as replicates. Hence sampling the used units within home ranges (or circles around nest or feeding sites) for the jth animal is viewed as 'subsampling' the primary sampling unit. We assume at all times that samples of animals or resource units are randomly selected.

4.12 Census of Available Resource Units

Suppose that the proportions of resources in categories 1 to I are known to be π_1 to π_I for the entire study area defined to be available to the population of animals. Let u_{ij} be the number of type i resource units used by animal j (e.g., the number of type i resource units in the home range of animal j), u_{i+} be the number of type i resource units used by all animals, u_{+j} be the total number of units used by animal j, and u_{++} be the total number of units used by all animals.

In this situation there are two chi-squared tests that provide useful information about selection. First, a test can be carried out to test the null hypothesis that animals are using resource categories in the same proportions, irrespective of whether this is selective or not. This is the usual test for independence of the row and column categories in a two-way table, for which the log-likelihood test statistic is

$$X_{L1}^2 = 2 \sum_{j=1}^{n} \sum_{i=1}^{I} u_{ij} \log_e\{u_{ij} / E(u_{ij})\}, \tag{4.26}$$

where $E(u_{ij}) = u_{i+}u_{+j}/u_{++}$ is the expected number of units of type i used by the jth animal if that animal uses the resources in the same way as the other animals. If X_{L1}^2 is significantly large in comparison with the chi-squared distribution with (I - 1)(n - 1) df then there is evidence that the hypothesis is not correct and animals are using resources differently, which implies that at least some of them are not selecting resources in proportion to availability. As usual with a chi-squared test, it is desirable that all the expected frequencies should be five or more.

The second chi-squared test that can be considered is an overall test of the null hypothesis of selection in proportion to availability. This test involves comparing the observed frequencies with which different resource categories are used by different animals with expected frequencies calculated from the resources available. The log-likelihood test statistic is then

$$X_{L2}^2 = 2 \sum_{j=1}^{n} \sum_{i=1}^{I} u_{ij} \log_e\{u_{ij} / E(u_{ij})\}, \tag{4.27}$$

where $E(u_{ij}) = \pi_i u_{+j}$ is the expected number of resource type i units used by the jth animal if use is proportional to availability. If this statistic is significantly large in comparison with the chi-squared distribution with n(I - 1) df then there is evidence of non-random selection by at least some of the animals.

Because of the way that they are calculated, X_{L1}^2 must be less than or equal to X_{L2}^2, and the difference $X_{L1}^2 - X_{L2}^2$ with I - 1 df is a test of the null hypothesis that animals are on average using resources in proportion to availability, irrespective of whether they are selecting the same or not. If this test statistics is significantly large when compared to the chi-squared distribution with I - 1 df then there is evidence that the average selection is not in proportion to availability of resources.

The selection ratio for the jth animal and the ith type of resource is estimated by

$$\hat{w}_{ij} = (u_{ij} / u_{+j}) / \pi_i, \tag{4.28}$$

the ratio of the observed proportion of type i resource used by the jth animal to the known proportion of type i resource available to the population.

There are two plausible estimators of the ‘average’ selection ratio w_i for the whole population of animals. First, the ratios for the n sampled animals can be averaged, to give

$$\hat{w}_i' = \sum_{j=1}^{n} \hat{w}_{ij} / n. \tag{4.29}$$

This estimator should be used if interest is in the selection ratios of individual animals and the variation of the ratios within the population. These estimates of individual selection ratios are independent under the assumption that the animals are randomly selected. Hence the standard error of the average selection ratio of the ith category can be computed as the square root of the variance

$$\mathrm{var}(\hat{w}_i') = \{ \sum_{j=1}^{n} (\hat{w}_{ij} - \hat{w}_i')^2 / (n - 1)\, n\}. \tag{4.30}$$

Second, if the interest is primarily in the selection ratios for resource categories regardless of which animals in the population are doing the selection, then the ratio of totals over all animals should be used to estimate the ratio of the observed proportion of type i resource used by the population, i.e.,

$$o_i = u_{i+} / u_{++}.$$

Then the recommended estimator of the overall selection ratio of the ith category for the population is,

$$\hat{w}_i = (u_{i+} / u_{++}) / (\pi_i). \tag{4.31}$$

As discussed in Section 4.3, in most statistics texts on sampling theory the estimator (4.31) is recommended rather than the estimator (4.30) because ratios of means or totals generally have less bias and variance than means of ratios. Also, the interest is usually in population statistics, rather than on the data for the individual sampling units, which are the animals for the applications of interest in this book.

An noted in Section 4.3, the choice between the two estimators depends in the circumstances. Researchers are usually interested in the selection ratios of individual animals, and the variation in these ratios from animal to animal. However average selection ratios for the population are also needed for making predictions about the expected outcomes of management actions, in which case it seems reasonable to use estimator (4.30).

Conditional on known values for π_i, the variance of $\hat{w}_i$ is estimated by the special case of equation (4.3) with $y_j = u_{ij} / \pi_i$ and $x_j = u_{+j}$. That is,

$$\mathrm{var}(\hat{w}_i) = \{ \sum_{j=1}^{n} (u_{ij} / \pi_i - \hat{w}_i \, u_{+j})^2 / (n - 1)\} \{n / u_{++}^2\}. \tag{4.32}$$

The estimates, $\hat{w}_i$ of the selection ratios are computed by pooling observations across all animals in the sample. However, the equation (4.32) takes the variation in resource selection from animal to animal into account because the expression

$$\Sigma(u_{ij} / \pi_i - \hat{w}_i \, u_{+j})^2 / (n - 1)$$

is an estimate of the variance of $u_{ij} / \pi_i - w_i \, u_{+j}$ in the population of animals.

Simultaneous Bonferroni confidence intervals for population selection ratios can be constructed with an overall confidence level of approximately 100(1-α)%, so that the probability of all the intervals containing the true value is approximately 1 - α. These intervals are of the form

$$\hat{w}_i' \pm z_{\alpha/(2I)} \, \mathrm{se}(\hat{w}_i'), \text{ or } \hat{w}_i \pm z_{\alpha/(2I)} \, \mathrm{se}(\hat{w}_i), \tag{4.33}$$

where I is the number of habitat types.

The approximate confidence interval for w_i' is justified because the estimator is a sample mean. In fact the t-distribution for small sample sizes could be used in place of the standard normal distribution to slightly improve the accuracy of the intervals or tests of hypothesis. The validity of the confidence interval for w_i will depend be on the standard error of $\hat{w}_i$ being well estimated, and on $\hat{w}_i$ being approximately normally distributed. Because of the complex nature of the estimation it is difficult to know precisely when this will occur. This question is addressed further in the example that follows.

The difference (w_i - w_j) between the selection ratios for resource units in categories i and j can be estimated by either ($\hat{w}_i'$ - $\hat{w}_j'$) or ($\hat{w}_i$ - $\hat{w}_j$). The first of these two estimators is just the difference between two independent sample means. The standard error of the difference is therefore just

$$\mathrm{se}(\hat{w}_i' - \hat{w}_j') = \surd\{\mathrm{var}(\hat{w}_i') + \mathrm{var}(\hat{w}_j')\}. \tag{4.34}$$

The second estimator is the difference of ratios of totals, but is itself a ratio estimator of the general form discussed in Section 4.3. In this case the variance can be estimated using equation (4.3) by setting $y_j = u_{ij} / \pi_i - u_{kj} / \pi_k$ and $x_j = u_{+j}$. This gives the result

$$\text{var}(\hat{w}_i - \hat{w}_j) = [\{n / (n - 1)\}/u_{++}^2] \sum_{k=1}^{n} (u_{ik} / \pi_i - u_{jk} / \pi_j - \hat{w}_i u_{+j} + \hat{w}_j u_{+j})^2. \quad (4.35)$$

Again, confidence intervals for population differences can be constructed based on the Bonferroni inequality. These take the form

$$(\hat{w}_i' - \hat{w}_j') \pm z_{\alpha/(2I')} \text{se}(\hat{w}_i' - \hat{w}_j'), \text{ or } (\hat{w}_i - \hat{w}_j) \pm z_{\alpha/(2I')} \text{se}(\hat{w}_i - \hat{w}_j), \quad (4.36)$$

where $I' = I(I - 1)/2$ is the number of differences that can be constructed between selection ratios. In this way, the probability that all the intervals will include their respective population ratios will be approximately $1 - \alpha$.

Example 4.4 Habitat Selection by Bighorn Sheep

Arnett *et al*. (1989) studied the selection of habitat types using a sample of six radio-tagged adult female bighorn sheep (*Ovis canadensis*) in the Encampment River drainage of southeast Wyoming, with the proportions of ten habitat types available in the study area being measured from maps. A subset of their data covering the period from August to December 1988 is shown in Table 4.7. Each animal was relocated approximately 75 times during this period to obtain a sample of the habitat points used.

Table 4.7 Habitat type, proportion of study area in each type, and number of occasions a given bighorn sheep was observed in each type.

	Available	Use of habitat by sheep number						
Habitat	proportion	1	2	3	4	5	6	Total
Riparian	0.060	0	0	0	0	0	0	0
Conifer	0.130	0	2	1	1	0	2	6
Mt. shrub I	0.160	0	1	2	3	2	1	9
Aspen	0.150	2	2	1	7	2	4	18
Rock outcrop	0.060	0	2	0	5	5	2	14
Sage/bitterbrush	0.170	16	5	14	3	18	7	63
Windblown ridges	0.120	5	10	9	6	10	6	46
Mt. shrub II	0.040	14	10	8	9	6	15	62
Prescribed burns	0.090	28	35	40	31	25	19	178
Clearcut	0.020	8	9	4	9	0	19	49
Total	1.000	73	76	79	74	68	75	445

The test statistic of equation (4.26) is $X_{L1}^2 = 99.20$, with 40 df if the zero counts of riparian are ignored, or 45 df if these are included in the calculation with expected frequencies of zero. With either choice of df the statistic is very significantly large, lending evidence to the conclusion that the sheep were not using resources in the same way.

The test statistic of equation (4.27) was much larger at $X_{L2}^2 = 785.54$ with 54 df, indicating very strong evidence indeed that the sheep were selective in their use of habitat. The difference $X_{L2}^2 - X_{L1}^2 = 686.34$ with 9 df (counting riparian frequencies) is an indication that the ‘average’ level of selection for the population is very strong.

The interpretation of these test results must be tempered by the fact that many of the observed frequencies are less than five. However, the extreme levels of significance for the chi-squared statistics leave little room to doubt that selection took place, and varied from animal to animal.

There are now two approaches that can be adopted for estimating selection ratios, depending upon whether the primary interest is in selection by the population of sheep in general, or on the selection ratios for the individual sheep. These will now be considered in turn.

First, then, it is assumed that the overall objective is estimation of the population selection ratios for the habitat types regardless of differences in selection among individuals. In this case, the computations are based on equation (4.31) and (4.32), with the results that are shown in Table 4.8. For the estimation of selection ratios this is basically the same analysis as was done in Example 4.1 for Nue *et al.*’s (1974) moose data. However for Table 4.8 the standard errors were computed using the individual sheep as the unit of replication, with a sample size of six.

Table 4.8 Estimated relative probabilities of selection for different habitats by bighorn sheep with lower and upper simultaneous 90% confidence limits computed using the Bonferroni inequality with a confidence level of (100-10/10)% = 99% for the ten individual intervals.

					Bonferroni Confidence limits	
Habitat	u_{i+}	$\pi_i u_{++}$	$\hat{w}_i$	$se(\hat{w}_i)$	Lower	Upper
Riparian	0	26.70	0.000	0.000	-	-
Conifer	6	57.85	0.104	0.037	0.009	0.198
Mt. shrub I	9	71.20	0.126	0.036	0.033	0.220
Aspen	18	66.75	0.270	0.081	0.061	0.479
Rock outcrop	14	26.70	0.524	0.213	0.000[1]	1.074
Sage/bitterbrush	63	75.65	0.833	0.211	0.289	1.376
Windblown ridges	46	53.40	0.861	0.105	0.590	1.133
Mt. shrub II	62	17.80	3.483	0.471	2.270	4.696
Prescribed burns	178	40.05	4.444	0.407	3.397	5.492
Clearcut	49	8.90	5.506	1.717	1.082	9.929

[1] An impossible negative confidence limit for rock outcrop has been replaced by 0.000.

A simulation study to investigate the robustness of the selection ratios, $\hat{w}_i$, in the face of the low frequencies in Table 4.7 was carried out. For this study, 215 sets of data with ten habitats and six sheep were generated in such a way that the number of habitat i resource units used by animal j was a Poisson random variable with a mean value given by the observed frequency in Table 4.7. For each set of data, selection ratios and their estimated standard errors were calculated, and hence the z-scores

$$z_i = (\hat{w}_i - w_i)/se(\hat{w}_i),$$

for i from 1 to 10. The true selection ratios w_i were set equal to the estimates obtained from the data in Table 4.7 on the grounds that these were the population values for the simulations.

The riparian habitat was not chosen by any of the bighorn sheep for which the data are shown in Table 4.7. Consequently, this habitat was not chosen in the simulations either. As a result, $w_1 = \hat{w}_1 = se(\hat{w}_1) = 0$, and z_1 is always undefined. The following comments therefore relate to the other z-scores only. The reason for expressing the results in terms of z-scores is that if the z_i values have approximately standard normal distributions then the confidence intervals (4.33) will be valid.

Although only 215 sets of data were simulated, this produced 1,935 z-scores because there were nine values for each set of data. The nine values for one set of data are not independent. Nevertheless, a very clear pattern emerged suggesting that most z-scores are less variable than can be expected from the standard normal distribution, but there are occasional very extreme values that should not occur with the standard normal distribution. This is illustrated by Figure 4.2, which shows how the observed distribution for all 1935 z-scores compares with the standard normal distribution.

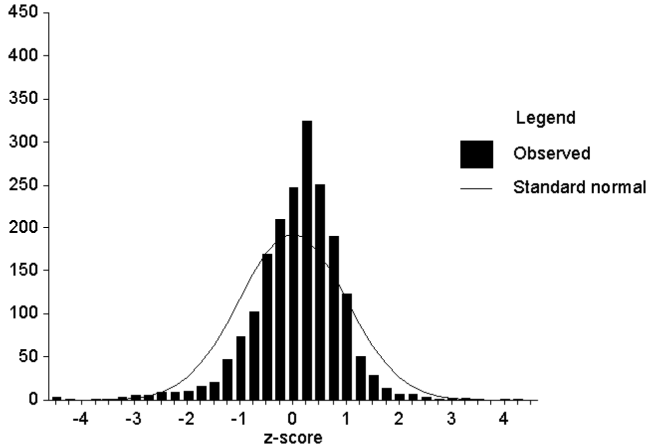


Figure 4.2 The distribution of values of $z_i = (\hat{w}_i - w_i)/se(\hat{w}_i)$ obtained by simulating 215 sets of data similar to the observed data that are shown in Table 4.7. It is required that z_i has a standard normal distribution for the confidence limits (4.33) to be valid.

This simulation study indicates that, on the whole, the confidence limits (4.33) are reliable, and, if anything, tend to be a little too narrow. Furthermore, the occasional very extreme z-scores were associated with situations where only one animal used a resource category once. Therefore, these situations can readily be identified.

Having established the validity of the ratio of totals method for determining population average confidence intervals, these intervals can now be considered for the real data. Table 4.8 contains the selection indices for the ten habitat types, together with Bonferroni confidence intervals computed from (4.36) using $\alpha = 0.1$, so that there is a probability of about 0.9 that all I = 10 intervals will contain their respective population selection ratios.

It is seen, for example, that the selection index for the prescribed burn is $\hat{w}_9 = 4.44$, with a confidence interval of 4.40 to 5.49 for the population value. Because the interval does not include the value one, there is significant selection for this habitat above what would be expected by chance. Also, mountain shrub II and clearcut have relative probabilities of selection which are above that expected under the hypotheses of no selection. Similarly, there is significant selection against conifer, mountain shrub I, and aspen relative to the amount of habitat available. There is no apparent selection for or against rock outcrop, sage/bitterbrush, and windblown ridges. Since riparian was not used at all it seems clear that this was selected against.

Consider next the comparison of pairs of selection ratios using the confidence limits (4.36). There are 45 possible comparisons between pairs of these ratios, which suggests that a reasonable α value for the limits is $0.1/45 = 0.002$. In other words, the individual confidence limits should have 99.8% confidence in order that there is a probability of approximately 0.9 that all of the intervals will contain the population values. This is achieved by using $z_{\alpha/(2I)} = z_{0.001} = 3.09$ in the limits (4.36). For example, consider the comparison of the selection ratios for clearcut ($\hat{w}_{10} = 5.506$) and prescribed burns ($\hat{w}_9 = 4.444$). From equation (4.35), $se(\hat{w}_{10} - \hat{w}_9) = 2.01$, so that the confidence limits for $w_{10} - w_9$ are

$$1.062 - 3.09(2.01) \text{ to } 1.062 + 3.09(2.01),$$

which is -5.149 to 7.273. Because this includes zero, the difference between the use of clearcut and prescribed burns is not significant.

Table 4.9 shows which selection ratios are significantly different on this basis, when the habitats are listed in order of their estimated selection ratio. It can be seen that mountain shrub II and prescribed burns are selected with significantly higher probability than are lower ranking habitat types. The use of clearcut has a high variance because the fifth sheep never used this habitat but the sixth sheep used it extensively. Thus, even though clearcut has the largest estimated selection index of $\hat{w}_{10} = 5.506$, it is not significantly larger than the selection indices for some of the lower ranking habitats. If there had been no replication between sheep then the high variance between sheep would be hidden and incorrect conclusions might be reached concerning the use of the clearcut habitat.

Although the simulation study mentioned earlier did not address the question of the validity of confidence intervals for differences between selection ratios, it is possible to say something about this. Some limited simulations indicated that the statistics

$$z_{ik} = \{(\hat{w}_i - \hat{w}_k) - (w_i - w_k)\}/se(\hat{w}_i - \hat{w}_k)$$

have distributions that are rather similar to the distributions of the statistics

$$z_i = (\hat{w}_i - w_i)/se(\hat{w}_i)$$

as shown in Figure 4.2. Thus the tendency for confidence intervals for selection ratios to be conservative seems to be shared by confidence intervals for differences between selection ratios.

Table 4.9 Significant differences between estimated selection ratios calculated from the data in Table 4.7. The entry '+' indicates a significant difference between the row habitat and the column habitat, with column habitats having obvious abbreviations for their names. The entry '-' indicates no significant difference.

Habitat	Rip	Con	MsI	Asp	Roc	S/b	Wbr	MsII	Prb
Conifer	-								
Mt. shrub I	+	-							
Aspen	+	-	-						
Rock outcrop	-	-	-	-					
Sage/bitterbrush	+	-	+	-	-				
Windblown ridges	+	+	+	+	-	-			
Mt. shrub II	+	+	+	+	+	+	+		
Prescribed burns	+	+	+	+	+	+	+	-	
Clearcut	+	+	+	+	-	-	-	-	-

For the second approach to analysing the data it is assumed that the objectives include the estimation of the selection ratios for the individual sheep, the study of the variation in selection among sheep, and the calculation of the average of the selection ratios using equation (4.29). Table 4.10 shows the results obtained. It turns out that for estimating the population selection ratios the results are very similar using the ratios of totals (Table 4.8) and simple averages (Table 4.10). This is because the numbers of relocations do not vary much among the six individuals. However, in other applications with very unequal sample sizes from individual animals, this will not necessarily be the case so that serious consideration will need to be given to whether population selection ratios are estimated using equation (4.28) or equation (4.29).

Table 4.10 Estimated relative probabilities of selection for different habitats by individual bighorn sheep, $\hat{w}_{ij}$, the arithmetic mean and the standard error (se) of the mean.

	Available	Values of $\hat{w}$ for individual sheep							
Habitat	proportion	1	2	3	4	5	6	Mean	se
Riparian	0.060	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.000
Conifer	0.130	0.00	0.20	0.10	0.10	0.00	0.21	0.10	0.034
Mt. shrub I	0.160	0.00	0.08	0.16	0.25	0.18	0.08	0.13	0.033
Aspen	0.150	0.18	0.17	0.08	0.63	0.20	0.36	0.27	0.073
Rock outcrop	0.060	0.00	0.44	0.00	1.13	1.23	0.45	0.54	0.198
Sage/bitterbrush	0.170	1.29	0.39	1.04	0.24	1.56	0.55	0.84	0.198
Windblown ridges	0.120	0.57	1.10	0.95	0.68	1.23	0.67	0.86	0.099
Mt. shrub II	0.040	4.80	3.29	2.53	3.04	2.21	5.00	3.48	0.434
Prescribed burns	0.090	4.26	5.12	5.63	4.65	4.08	2.81	4.43	0.362
Clearcut	0.020	5.48	5.92	2.53	6.08	0.00	12.67	5.45	1.569

4.13 Sample of Available Resource Units

If the proportion π_i of the ith resource category available has to be estimated then an additional source of variation is introduced into estimates of selection ratios. Thus, the selection ratio for the ith resource category by the jth animal is estimated by

$$\hat{w}_{ij} = u_{ij} / (\hat{\pi}_i u_{+j}), \tag{4.37}$$

where $\hat{\pi}_i$ is the estimated proportion of resource category i available. This is equation (4.28) but with π_i estimated. One estimator for the population selection ratio for resource units in category i is then

$$\hat{w}_i = (u_{i+} / u_{++}) / (\hat{\pi}_i), \tag{4.38}$$

where the numerator is the ratio of totals obtained by pooling data from all animals. This is equation (4.30) but with π_i estimated.

The variance of the estimator (4.38) can be determined in two stages. First, the variance of the ratio $V_i = u_{i+}/u_{++}$ can be estimated using equation (4.3), taking $y_j = u_{ij}$ and $x_j = u_{+j}$. Next, assuming that $\hat{\pi}_i$ is estimated from an independent source of data with standard error $se(\hat{\pi}_i)$, equation (4.2) can be used to estimate the variance of the final ratio $\hat{w}_i = V_i / \hat{\pi}_i$, setting $y = V_i$, $x = \hat{\pi}_i$, and $r_{xy} = 0$. Bonferroni simultaneous confidence intervals for a set of selection ratios can then be calculated as discussed in Section 4.8.

To compare selection ratios the differences

$$\hat{w}_i - \hat{w}_k = u_{i+} / (\hat{\pi}_i u_{++}) - u_{k+} / (\hat{\pi}_k u_{++}) \tag{4.39}$$

can be calculated together with their standard errors, for all possible values of i and k. Bonferroni simultaneous confidence limits can then be determined as discussed in Section 4.9.

Equation (4.39) has the same form as equation (4.5), with identification being established by setting $y_{1j} = u_{ij}$, $x_{1j} = \hat{\pi}_j u_{+j}$, $y_{2j} = u_{kj}$ and $x_{2j} = \hat{\pi}_k u_{+k}$. The variance of the difference $\hat{w}_i - \hat{w}_k$ can therefore be estimated by equation (4.6), although a modified version of this equation taking into account how the available resource proportions are estimated may be more convenient to use.

Probably the most commonly used procedure involves choosing m_+ random resource units from the available population. For example, random points in a study area might be chosen and the habitat type encountered at each point recorded. The proportion of points encountering habitat type i, $\hat{\pi}_i$, is then taken as the estimate of π_i. With this method of estimation the numbers of resource units in different categories follow a multinomial distribution so that

$$\text{var}(\hat{\pi}_i) = \pi_i (1 - \pi_i)/m_+,$$

and

$$\text{cov}(\hat{\pi}_i, \hat{\pi}_k) = -\pi_i \pi_k / m_+.$$

Using these variances and covariances, it can be shown that equation (4.6) yields the result

$$\mathrm{var}(\hat{w}_i - \hat{w}_k) = \sum_{j=1}^{n} \{u_{ij} / \hat{\pi}_i - u_{kj} / \hat{\pi}_k - (\hat{w}_i - \hat{w}_k)\, u_{+j}\}^2/(n - 1)\{n/u_{++}^2\}$$

$$+ \{\hat{w}_i^2 / \hat{\pi}_i + \hat{w}_k^2 / \hat{\pi}_k - (\hat{w}_i - \hat{w}_k)^2\}/m_+. \qquad (4.40)$$

Because of the nature of the equations that have been provided in this section for standard errors, it is difficult to be sure how accurate they will be in practice. This matter is considered in the following example, although at this point it can be said that it seems clear that it is particularly important to get good estimates of population proportions of available resource units in different categories since these play such a key role in the estimation procedures.

An alternative estimator for the population selection ratios is the average of the selection ratios for individual animals given by equation (4.29), with a variance that can be estimated using equation (4.30). However, a complication arises when comparing $\hat{w}_i'$ with $\hat{w}_j'$ by finding confidence intervals for the differences ($\hat{w}_i'$ - $\hat{w}_j'$), because the random variable $\hat{\pi}_i$ is in the first term and the random variable $\hat{\pi}_j$ is in the second. Unfortunately, these random variables will usually be correlated, which will need to be taken into account when estimating var($\hat{w}_i'$ - $\hat{w}_j'$). In practice the estimation of variances with these types of complication is often best done using a bootstrapping approach as discussed in Section 2.11.

Chi-squared tests can be used to test for whether animals are using resources in a similar way, and whether they are being selective. Thus the statistic X_{L1}^2 of equation (4.26) provides a test for the consistent use of resources by different animals because it does not take availability into account. Also, if the availability of different resource categories is determined by a random sample of m_+ available resource units, then X_{L1}^2 can be calculated treating this sample as an (n + 1)th animal that uses resources in proportion to their availability. This then gives a test for consistent proportions both for animals and the availability sample. A significant result indicates that selection occurs since this is the case either if the animals use resources categories with different probabilities or if their use is the same but differs from what is expected from the available sample.

Example 4.6 Habitat Selection by Bighorn Sheep (Partly Artificial Data)

For the sake of an example, suppose that Arnett *et al.* (1989) had not been able to determine the proportions of different habitat available for bighorn sheep exactly, but had instead estimated these proportions from a sample of 250 random points in the study region. They might then have obtained the data shown in Table 4.11.

Because the counts of used units for the six animals are the same here as in Table 4.7, the chi-squared statistic of equation (4.26) has the same value as it had with Example 4.4, where available proportions were assumed known. This is $X_{L1}^2 = 99.2$, with 45 df. It tests for a consistent choice of resources from animal to animal, irrespective of whether this is selective or not. As noted in Example 4.4, this is highly significant and indicates a lack of consistency.

If the sample of available resources is included in the chi-squared calculation as if it came from a non-selective seventh animal, then the chi-squared value increases to $X_{L2}^2 = 373.4$, with 54 df. The difference X_{L2}^2-X_{L1}^2 = 274.2 with nine df is then a measure of the amount of selectivity, irrespective of whether there are differences between animals or not. Again the result is highly significant, indicating very strong evidence of selection.

As mentioned in Example 4.4, the small frequencies for some of the categories unused by the animals, and particularly the zero frequencies for riparian, mean that the accuracy of the chi-squared approximations is questionable for these data. However, the very large chi-squared values means that the evidence for selectivity and differences between animals is clearly established.

Estimates of selection ratios and their standard errors are shown in the final two columns of Table 4.11. These can be used to produce Bonferroni confidence intervals of the form

$$\hat{w}_i \pm z_{\alpha/(2x10)} \, se(\hat{w}_i)$$

for the population selection ratios. The justification for these limits has been discussed before in this chapter, and it should be noted thatthat all of the ten possible intervals are expected to contain their respective population values with probability 1 - α. Of course, the situation with the riparian habitat which was not used at all by the sheep is unsatisfactory because the estimated selection ratio is zero, with an estimated standard error of zero. This habitat should therefore be excluded from the confidence interval calculations.

Bonferroni confidence intervals of the form

$$\hat{w}_i - \hat{w}_j \pm z_{\alpha/(2x45)} \, se(\hat{w}_i - \hat{w}_j)$$

can also be constructed for differences between population selection ratios using the variance from equation (4.40). Again, the probability that all of the 45 possible intervals contain their population values should be approximately 1 - α.

Table 4.11 The bighorn sheep data with an artificial sample of 250 available resource units assumed to be available instead of known population proportions in different resource categories.

	Available sample		Use of habitat by sheep number							
Habitat	Count	Proportion	1	2	3	4	5	6	$\hat{w}$	se($\hat{w}$)
Riparian	21	0.084	0	0	0	0	0	0	0.00	0.00
Conifer	26	0.104	0	2	1	1	0	2	0.12	0.05
Mt. shrub I	41	0.164	0	1	2	3	2	1	0.12	0.03
Aspen	40	0.160	2	2	1	7	2	4	0.25	0.08
Rock outcrop	14	0.056	0	2	0	5	5	2	0.56	0.27
Sage/bitterbrush	46	0.184	16	5	14	3	18	7	0.76	0.21
Windblown ridges	34	0.136	5	10	9	6	10	6	0.76	0.15
Mt. shrub II	8	0.032	14	10	8	9	6	15	4.35	0.16
Prescribed burns	14	0.056	28	35	40	31	25	19	7.14	1.96
Clearcut	6	0.024	8	9	4	9	0	19	4.58	2.34
Total	250	1.000	73	76	79	74	68	75		

The procedure for constructing and interpreting these limits is the same as was used for Example 4.4, except for the calculation of standard errors. Details will therefore not be provided here. However, it is useful to summarize the results of a simulation study that was designed to assess the validity of the proposed confidence intervals for population selection ratios.

As for previous examples, it can be argued that the validity of confidence intervals depends on

$$z_i = (\hat{w}_i - w_i)/se(\hat{w}_i)$$

having a standard normal distribution. A reasonable way to assess the limits therefore involves generating artificial data of a similar nature to real data, calculating z_i values, and comparing the distribution obtained for these with the standard normal distribution.

This exercise has been carried out using the data in Table 4.11 to provide expected frequencies for a simulated sample of available resource units and simulated samples from six sheep. The actual data frequencies generated had Poisson distributions about the expected frequencies. A total of 225 sets of data were generated, each providing values for z_2 to z_{10}. Riparian was never used for the simulated sets of data because the Poisson expected frequency was set to zero for all six sheep for this habitat type. Hence z_1 was always undefined.

The distribution obtained for the 225 x 9 = 2025 generated z values is shown in Figure 4.3 together with the standard normal distribution. This figure shows a great similarity to Figure 4.2 since in both cases the distribution of z values is generally less variable than the standard normal. This then suggests that the confidence intervals of the form

$$\hat{w}_i \pm z_\alpha \, se(\hat{w}_i)$$

will generally contain their population parameters with a higher probability than 1 - α.

Although the validity of confidence limits for differences between selection ratios was not investigated in the simulations, it does seem likely that the conservative nature of the confidence limits for individual selection ratios will carry over because (as discussed in Example 4.4) this is what seems to happen if the population proportions of available resource units of different types are known.

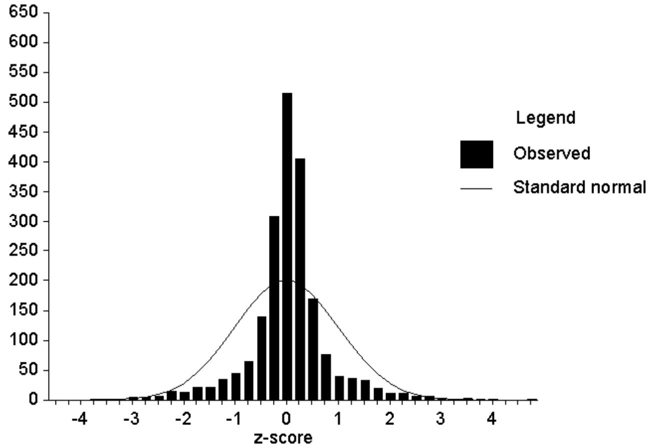


Figure 4.3 The distribution of values of $z_i = (\hat{w}_i - w_i)/se(\hat{w}_i)$ obtained by simulating 225 sets of data similar to the observed data that are shown in Table 4.10. It is required that z_i has a standard normal distribution for confidence limits for population selection ratios to be valid with data of the form shown in Table 4.10.

4.14 Design III with Sampling Protocol A

With design III studies the use and availability of resource units is measured separately for each animal. As for the other designs, there are then two cases to consider, depending on whether the proportions of units of different types that are available to each animal are known accurately, or are estimated from a random sample. We consider first the case where the available proportions are known accurately.

4.15 Census of Available Resource Units

We assume as before that there are I resource categories, the use of resources is measured for n animals, and that a random sample of u_{+j} resource units used by the jth animal is obtained and found to contain u_{ij} units in resource category i. We assume also that π_{ij} is the known proportion of resources available to animal j that are in category i.

An estimate of the selection ratio for the ith resource category by the jth animal is then

$$\hat{w}_{ij} = (u_{ij} / u_{+j})/\pi_{ij}, \tag{4.41}$$

and one estimator to use for the selection ratio for the population of animals for the ith resource category is

$$\hat{w}_i = u_{i+} / \sum_{j=1}^{n} \pi_{ij} u_{+j}. \tag{4.42}$$

Equation (4.3) can be used to estimate the variance of this estimator by setting $y_j = u_{ij}$ and $x_j = \pi_{ij}u_{+j}$, so that $\hat{R} = \hat{w}_i$. Bonferroni confidence intervals for population selection ratios can then be obtained in the same way as for design I and design II studies.

To compare two estimated selection ratios the difference $\hat{w}_i - \hat{w}_j$ can be considered with an estimate of the standard error of this difference. To this end it can be noted that the difference takes the form of equation (4.5) and that the use of equation (4.6) yields

$$\begin{aligned} \mathrm{var}(\hat{w}_i - \hat{w}_j) = \{n/(n - 1)\} \sum_{k=1}^{n} \{&(u_{ik} - \hat{w}_i\, u_{+k})/(\pi_{i1}u_{+1} + ... + \pi_{in}u_{+n}) \\ &+ (u_{jk} - \hat{w}_j\, u_{+k})/(\pi_{j1}u_{+1} + ... + \pi_{jn}u_{+n})\}^2. \end{aligned} \tag{4.43}$$

This variance is the same as what is given by equation (4.35) if all animals have the same available proportions for resource categories i and j.

Once the differences $\hat{w}_i - \hat{w}_j$ have been calculated together with their standard errors, a set of Bonferroni confidence limits can be calculated for the true population differences. The procedure is exactly the same as for design I and design II studies.

A test for whether the jth animal sampled is selective is provided by calculating

$$X_{Lj}^2 = \sum_{i=1}^{I} u_{ij} \log_e\{u_{ij} / (u_{+j}\, \pi_{ij})\},$$

with I - 1 df, where this compares the observed use of category i resource units (u_{ij}) with the expected use based on availability ($u_{+j}\pi_{ij}$). Adding up the n statistics obtained in this way then provides an overall test for selection with n (I - 1) df. As usual, these chi-squared tests require the expected frequencies $u_{+j}\pi_{ij}$ to be 'large', which means that most if not all of them should be five or more.

The alternative to estimating population selection ratios with equation (4.42) is to use the average over animals of $\hat{w}_i$ from equation (4.29), with a variance estimated from equation (4.30). Confidence intervals for the true population values and differences between these population values are then obtained using the methods that were described in Section 4.12 for the case when all animals have the same available proportions.

4.16 Sample of Available Resource Units

The final type of study design that we will consider is design III with the resources available to individual animals being estimated independently, for example by taking random samples of habitat points within home ranges.

Under these conditions equation (4.41) which applies when the resources available to individual animals are known, changes in the obvious way to

$$\hat{w}_{ij} = u_{ij} / \hat{\pi}_{ij}\, u_{+j}, \qquad (4.44)$$

where $\hat{\pi}_{ij}$ is the estimated proportion of the resource units that are available to the jth animal that are in category i. Similarly in equation (4.42) the known available resource proportions are replaced by their estimates to give the estimated population selection ratios

$$\hat{w}_i = u_{i+} / \sum_{j=1}^{n} \hat{\pi}_{ij} u_{+j}. \qquad (4.45)$$

To estimate the standard error of $\hat{w}_i$, equation (4.3) can be used yet again. It is just necessary to set $y_j = u_{ij}$ and $x_j = (\hat{\pi}_{ij}\, u_{+j})$, so that $\hat{R} = \hat{w}_i$. Bonferroni confidence intervals can then be constructed in the usual way.

To compare selection ratios, all possible differences $\hat{w}_i - \hat{w}_j$ can be calculated together with their estimated standard errors, and Bonferroni simultaneous confidence intervals constructed as for the other designs that have been considered. To this end, equation (4.7) can be used to provide the variance, taking $y_{1j} = u_{ij}$, $x_{1j} = \hat{\pi}_{ij}\, u_{+j}$, $y_{2j} = u_{kj}$, and $x_{2j} = \hat{\pi}_{kj}\, u_{+j}$.

If the population selection ratios are just estimated by the means $\hat{w}_i'$ of the values from individual animals using equation (4.29), with variances from equation (4.30) then confidence intervals for the true population values can be calculated in the usual way, with or without a Bonferroni correction. However, the estimation of the variances of differences between selection ratios will be complicated because of the correlations between estimates of available proportions that will usually exist. This is another situation where bootstrapping (Section 2.11) may be the easiest way to estimate these variances, particularly if some special purpose method has been used to estimate the available proportions in the first place.

4.17 Discussion

In Section 3.4 a number of general assumptions were mentioned as being required in order to estimate resource selection functions. In the context of the present chapter these assumptions are what is required for the valid estimation of selection ratios, and it is appropriate at this point to review the assumptions in this light.

Assumption (a) requires that the proportions of different categories of resource units that are available do not change during the sampling period. For example, this assumption might be violated if animals eat most of the food in their 'preferred' habitat during the first two weeks of a four-week study. This requirement is difficult to satisfy with many studies unless they are carried out in a short period of time.

Assumption (b) is that the population of available resource units is correctly identified. This may be particularly difficult with design III studies because of the need to identify what is available to individual animals.

Assumption (c) is that the universe of used resource units is correctly identified and sampled. For example, this requirement may be violated if animals are eating food items which are not detectable in faecal samples. This is one of the most crucial and most difficult assumptions of the study design. Specific applications of the theory must be addressed separately and a general discussion would be unduly long. We note, however, that methods such as that developed by Nams (1989) for adjusting for radio telemetry errors may need to be used in some applications.

Assumption (d) is that the variables which actually influence the probability of selection are correctly identified. For example, this assumption is violated if percentage cover by vegetation is measured, but animals are actually selecting plots on the basis of the height of the plants. Usually it is hoped that the variables used in a study are highly correlated with the variables that actually influence the probability of selection.

Assumption (e) is that animals have unrestricted access to the entire distribution of available resource units. If animals are territorial then a few aggressive individuals may control all of the 'preferred' habitat, so that this assumption is violated. The assumption is most easily justified when the subpopulation of used units is small relative to the population of available units.

Assumption (f) is that resource units are sampled randomly and independently. This requirement might be violated if sampled animals are in the same herd or if the visibility of animals varies with the habitat type. Estimates of selection indices may still be meaningful if this assumption is not satisfied, but standard errors may not reflect the true variation in the populations. For the sake of illustration, our example analyses were made on the assumption of random independent samples of resource units. It will be difficult to ensure this, especially in cases when animals occur in herds or when resource units are collected in batches. For example, consider the collection of stomach samples of animals in a design I food selection study. In this case the food items are obtained in batches and the selections of individual food items may not have been independent events because of different food preferences by different animals.

Another common but difficult situation is the analysis of relocations of radio-tagged animals. Relocations often come in a batch recorded at a series of points in time. Care must then be taken to ensure that the time interval between recordings is sufficient to assume that observations of used habitat points are independent events if the relocation points are to be considered the units of replication.

In the presence of these problems, one approach is to estimate separate resource selection indices for several independent replications of batches of dependent units. Thus, one might estimate the selection indices for each of several randomly selected sites in a large study region. Inference toward mean values of the selection ratios over the entire study region can then proceed by standard statistical procedures, using replicates to determine standard errors. Alternatively, one might consider the selection of individual animals as independent events, and estimate separate selection indices for each animal by randomly sampling the units available and the units used by each animal to give what we have called a design II or a design III study. This may be the only reasonable approach for the study of food and habitat use by highly territorial animals or for the study of selection by radio-tagged animals.

In addition to the assumptions (a) to (f), we note that with design III studies estimates of the proportions of different types of resource units available (π_{ij}) may not be truly independent among animals. For example, a sample survey of habitat available in the overall study area may be conducted. Then the observations falling into an individual animal's home range might be used to estimate the habitat available to that specific animal. In that case, if there is considerable overlap of home ranges then some data points will influence the estimate of habitat availabilities for several different animals. At this time, the procedure described in Section 4.15 is recommended for the estimation of variances of selection ratios. However it should be noted that the true sampling variance of $\hat{w}_i$ may be underestimated. Of course, independent estimates of π_{ij} should be obtained for each animal if possible, so that estimates of sampling variances are approximately unbiased.

Chapter Summary

- The chapter concerns a variety of situations where the resources available to animals are in I categories, and selection is inferred by differences between the available proportions of resources in the categories and the proportions used by the animals. These situations are considered as special cases because of the popularity of this approach.

- It is noted that in estimating the population mean value for any ratio of random variables X and Y there are two alternative approaches using the ratio of the mean of Y to the mean of X, or the mean value of Y/X. These two approaches can be applied with selection ratios when these are known for the individual animals in a sample from a population of animals.

- Equations for the variances of ratios and differences between ratios are provided because of their use with the estimation of selection ratios.

- The estimation of selection ratios is discussed for design I studies where individual animals not identified, all animals are assumed to have the same selection, and the proportions of available resources in different categories are known. Chi-squared tests for significant selection and confidence limits for population proportions of selected resource units are described. Confidence limits for selection ratios and differences between selection ratios are also covered. Examples are provided using real resource selection data.

- The changes needed for analysis when the available proportions of resources in different categories are estimated rather than known are discussed. An Example is provided using real data.

- Design II studies are considered, where the resources selected by n individual animals are known and the available resources are assumed to be the same for each of these animals. First cases are considered where the proportions of available units in different categories are known. Chi-squared tests for differences in selection between different animals, and selection overall are described. The two alternative methods for estimating population selection ratios (ratios of averages or averages of ratios) are discussed, with variances and confidence limits for population ratios and differences between these ratios. An example is provided using real data.

- Design II studies with estimated proportions of available resources are considered, with an example using an artificial data set.
- Design III studies are considered where the proportions of different categories of resource units varies between animals and information about the use of these resources is available for n animals. The situation (a) where the resources available for each animal are known, and (b) where the resources available for each animal are estimated are treated separately.
- The assumptions involved in analysing data on resource selection are discussed for the special cases considered in the chapter where resources are in several categories.

Exercise

This exercise concerns the use of five habitat types by grey partridges (*Perdix perdix*), as recorded by Smith *et al*. (1982). Ten grey partridges were radio-tagged and radio locations were classified in one of five habitats: small grain fields, row crop fields, hay fields, pasture or idle. Availability was censused by partitioning a map of the study area into five habitat types. A subset of the location data for one of six time periods is shown in Table 4.12, together with the percentages of different habitats in the study area. Noting that this is a design II study with known proportions of different habitats available,

(a) test for evidence of selection using chi-squared tests and by constructing Bonferroni simultaneous confidence limits for population selection ratios; and

(b) construct Bonferroni simultaneous confidence limits for differences between the population selection ratios for different habitats.

Table 4.12 Numbers of radio locations in different habitat types for ten grey partridges, with the percentage of the available area in each habitat for the study region as a whole.

	Habitat types					
Bird	Small grain fields	Row crop	Hay	Pasture	Idle	Total
1	0	8	0	20	2	30
2	25	21	0	0	1	47
3	17	11	0	0	2	30
4	4	0	0	0	2	6
5	20	0	0	9	0	29
6	22	0	0	2	0	24
7	0	7	6	0	1	14
8	10	26	2	8	0	46
9	21	0	4	0	3	28
10	44	1	0	0	5	50
Total	163	74	12	39	16	304
Availability (%)	28.2	41.7	10.2	13.5	6.3	100.0

CHAPTER 5

LOGISTIC REGRESSION

One of the simplest ways of estimating a resource selection probability function (RSPF) involves taking a census of the used and unused units in a population of resource units, and fitting a logistic regression function for the probability of use as a function of variables that are measured on the units. Logistic regression can also be used with samples of resource units, although this is complicated by the need to vary the estimation procedure according to the sampling protocol that is used. These uses of logistic regression are discussed in this chapter, and illustrated using data on the selection of winter habitat by antelopes and nest site selection by fernbirds.

5.1 Census Data

Suppose that there are N available resource units and that it is known which of these have been used and which have not been used after a single period of selection. Then logistic regression, as discussed in Section 2.3, can be used to relate the probability of use to variables X_1 to X_p that are measured on the resource units. In this case, the RSPF is assumed to take the form

$$w^*(\mathbf{x}) = \frac{\exp(ß_0 + ß_1x_1 + \ldots + ß_px_p)}{1 + \exp(ß_0 + ß_1x_1 + \ldots + ß_px_p)}, \qquad (5.1)$$

where $\mathbf{x} = (x_1,x_2,\ldots,x_p)$ holds the values for the X variables that are measured on a unit.

This logistic function has the desirable property of restricting values of $w^*(\mathbf{x})$ to the range 0 to 1, but is otherwise arbitrary. Other functions that could be used include the probit

$$w^*(\mathbf{x}) = \Phi(ß_0 + ß_1x_1 + \ldots + ß_px_p), \qquad (5.2)$$

where $\Phi(z)$ is the integral from $-\infty$ to z for the standard normal distribution, and the proportional hazards function

$$w^*(\mathbf{x}) = 1 - \exp\{-\exp(ß_0 + ß_1x_1 + \ldots + ß_px_p)\} \qquad (5.3)$$

of Section 2.5. The main justification for using the logistic function rather than any other to approximate the RSPF is the fact that it is widely used for other statistical analyses in biology, and computer programs for estimating the function are readily available.

Suppose that the N available resource units can be divided into I groups so that within the ith group the units have the same values $\mathbf{x}_i = (x_{i1},x_{i2},...,x_{ip})$ for the X variables. The number of resource units used in group i, u_i, can then be assumed to be a random value from the binomial distribution with parameters A_i and $w^*(\mathbf{x}_i)$, where A_i is the number of available resource units in the group. Maximum likelihood estimates of the ß parameters in equation (5.1) can then be calculated using any of the standard computer programs for logistic regression. The input that is required for estimation are the group sizes (A_1 to A_I), the vectors of X values ($\mathbf{x}_1$ to $\mathbf{x}_I$), and the numbers of used units for each group (u_1 to u_I).

Often all of the available resource units will have different values for the X variables, so that each of the I groups consists of just one resource unit. This causes no difficulties as far as estimation is concerned, and in fact some computer programs are specifically designed to handle this case only.

As explained in Section 2.7, the deviance can under certain conditions be used as a statistic indicating the goodness of fit of the model. In the present context this statistic is

$$D = 2 \sum_{k=1}^{I} [u_i \log_e\{u_i /(A_i \hat{w}^*(\mathbf{x}_i))\} + (A_i - u_i) \log_e\{(A_i - u_i)/(A_i - A_i \hat{w}^*(\mathbf{x}_i))\}], \quad (5.4)$$

where the degrees of freedom (df) are I - p - 1. The condition for this to have an approximately chi-squared distribution is that most values of $A_i\, w^*(\mathbf{x})\{1 - w^*(\mathbf{x})\}$ are 'large', which in practice means that they are five or more. However, differences between the deviances for different models can reliably be tested against the chi-squared distribution even when this condition does not hold (McCullagh and Nelder, 1989, p. 119).

Some computer programs for logistic regression output the difference between the deviance for the no selection model with

$$w^*(\mathbf{x}) = \exp(ß_0)/\{1 + \exp(ß_0)\}$$

and the deviance for particular model being fitted with one or more X variables included, but do not output the deviances themselves. It is therefore useful to note that the deviance for the no selection model can be found by substituting

$$\hat{w}^*(\mathbf{x}) = u_+/N$$

in equation (5.4), where u_+ is the total number of used units out of the N available. The reason for this is that in the absence of selection the maximum likelihood estimate of the probability of use for all units is the observed proportion of units used. The 'no selection' deviance has I - 1 df.

5.2 Use With a Random Sample of Resource Units

Suppose that the units for which information is available are not all of the resource units in the population. Instead, a random sample of units is selected from the full population and it is observed whether each of these is used or not. Then equation (5.1) can still be used to approximate the probability of use for the ith unit, and logistic regression can be applied to the sample just as well as if there had been a full census.

The situation is slightly different if results are available for a sample of units that was not randomly selected from the population of all resource units. For example, the units selected might come from only a small part of the area covered by the full population. In that case logistic regression can still be used, by taking one of two points of view. First, the population of interest can be redefined to consist only of those in the smaller area. This then changes the sample to a census from the smaller area, and logistic regression can be used to estimate the RSPF for this area only. Nothing can then be said about the resource selection function in other areas.

Alternatively, it can be assumed that the RSPF is the same everywhere, and can therefore be estimated by a sample from just a part of the total area. An important consideration if this view is taken is that the use of logistic regression to estimate a RSPF does not require that the units analysed are a random sample from the population of interest. Instead, it is a model-based approach that draws its validity from the assumption that if a unit has values $\mathbf{x} = (x_1,x_2,...,x_p)$ then the probability of it being recorded as used is given by equation (5.1), independent of the use or otherwise of any other unit.

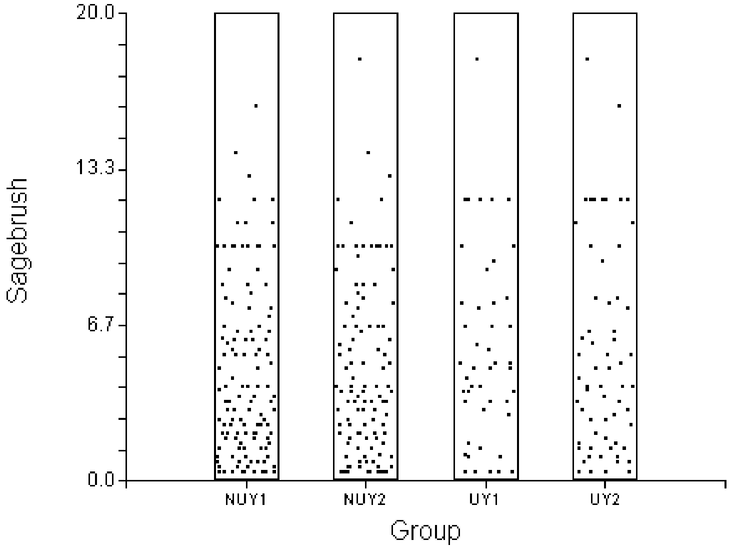
Example 5.1 Habitat Selection by Antelope

As an example of the use of logistic regression to assess resource selection, consider the study carried out by Ryder (1983) on winter habitat selection by antelope (*Antilocapra americana*) in the Red Rim area in south-central Wyoming that has already been described in Example 3.2. Recall that Ryder set up 256 study plots and recorded the presence or absence of antelope in the winters of 1980-81 and 1981-82, together with a number of characteristics of each plot.

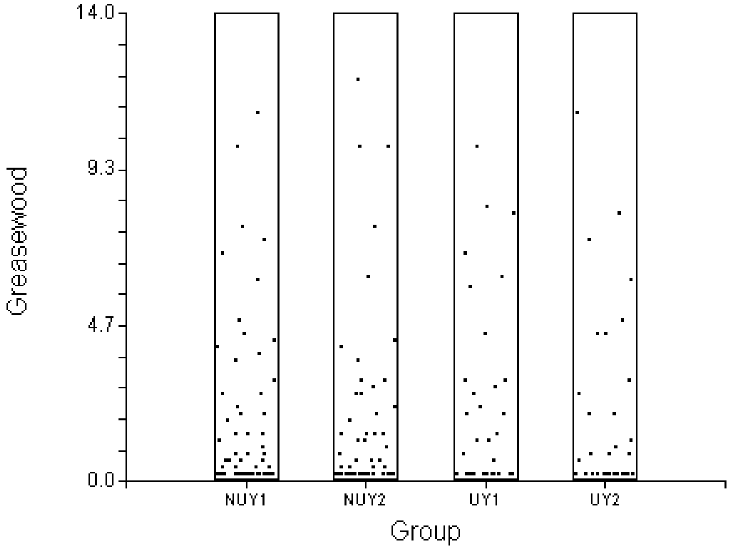
The area considered by Ryder consists of alternating blocks of public and private land, and his study plots are a systematic sample of 10% of the public land. There are therefore three possibilities in terms of the population of resource units that an estimated resource selection function applies to. First, the 256 sampled plots can be regarded as the population of interest. Second, it can be assumed that the resource selection function is the same on all public land. In that case the estimated function applies to all plots in this population. Third, it can be assumed that the resource selection function is the same on all public and private land. In that case, the estimated function applies to the whole of the Red Rim area. It is completely a matter of judgement as to which of these populations is relevant. As no private land was sampled, the third population does not seem reasonable. However, the sampled plots were systematically laid out on the public land so it will be assumed here that the estimated function applies to all public land.

Ryder's data are shown in Table 3.2 but with a number of vegetation height variables omitted because these are not defined on some study plots. Figure 5.1 gives a comparison between the distributions of the variables for the unused plots and the plots used at least once, separately for each year. It can be seen from this figure that the distribution of the distance to water and the use of the East/Northeast aspect are somewhat different for these four groups.

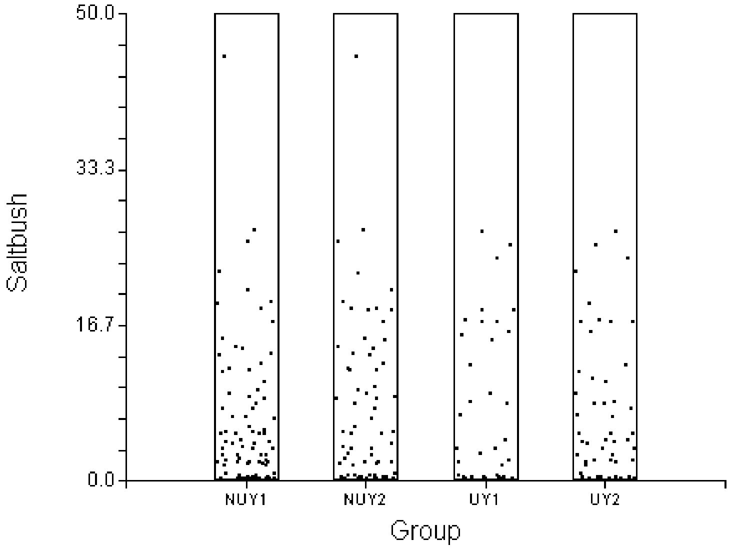
(a) Density of Sagebrush (thousands/ha)



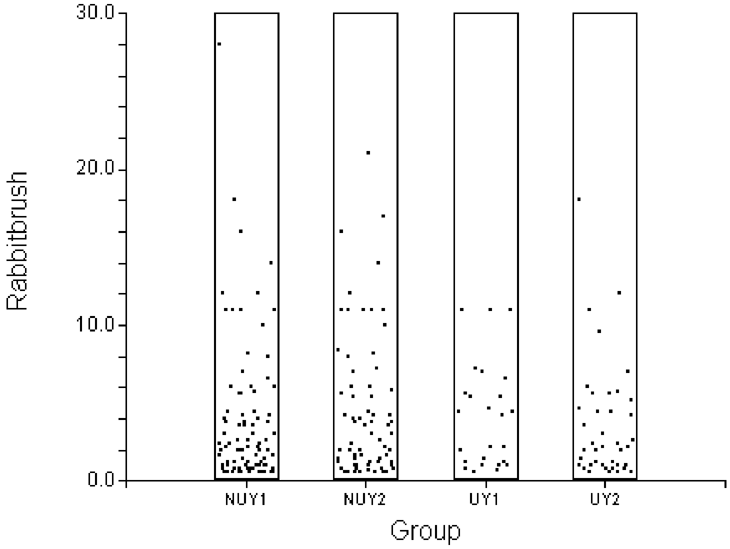
(b) Density of Greasewood (thousands/ha)



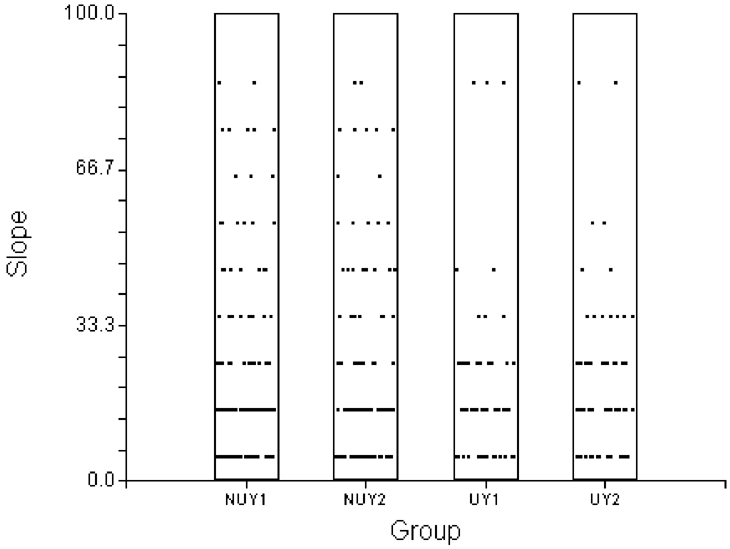
(c) Density of Saltbrush (thousands/ha)



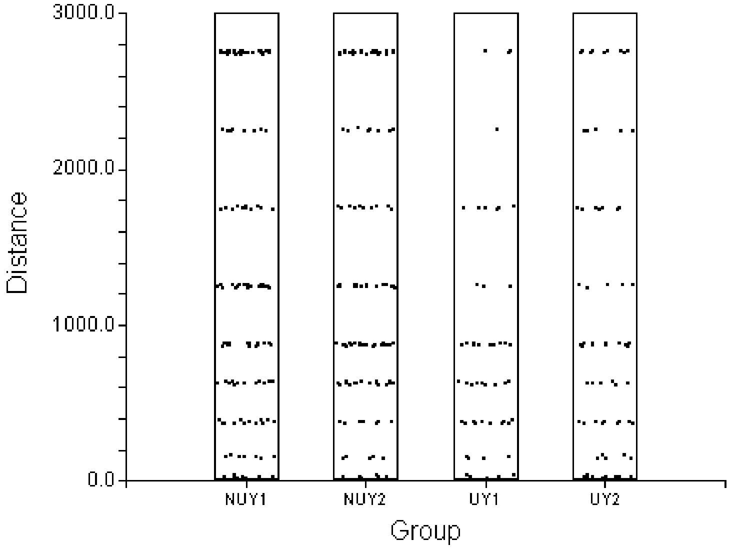
(d) Density of Rabbitbrush (thousands/ha)



(e) Slope of Plot (Degrees)



(f) Distance to Water (m)



(g) Aspect of Plots

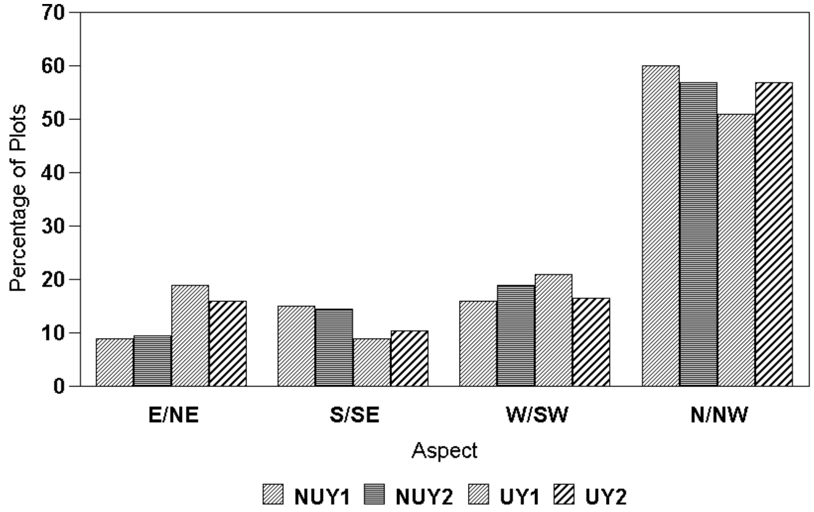


Figure 5.1 Distributions of the variables shown in Table 3.2 for unused plots in 1980-81 (NUY1), unused plots in 1981-82 (NUY1), used plots in 1980-81 (UY1) and used plots in 1981-82 (UY2). Dotplots are used to represent the distribution of the variables other than aspect, with a dot for each data point. For aspect, the percentage frequency of the four aspects is shown separately for NUY1, NUY2, UY1 and UY2.

To allow the estimation of a resource selection probability function where each of the four aspects (East/Northeast, South/Southeast, West/Southwest and North/Northwest) has a different probability of use, three 0-1 indicator variables can be used to replace the single aspect number shown in the last column of Table 3.2. The first of these indicator variables can be set equal to 1 for an East/Northeast plot or otherwise 0, the second indicator variable can be set equal to 1 for a South/Southeast plot or otherwise 0, and the third indicator variable can be set equal to 1 for a West/Southwest plot or otherwise 0. For example, the values of these dummy variables for the first plot are 0 0 0 because this has aspect 4 (North/Northwest), while for the second plot the values are 0 0 1 because this has aspect 3 (West/Southwest). Only three indicator variables are needed to allow for differences between four aspects because the North/Northwest aspect can be considered as the 'standard' aspect, and it is only necessary to allow the three other aspects to differ from this.

With the introduction of the indicator variables for aspect there are nine variables available to characterize each of the 256 study plots: X_1 = density (thousands/ha) of big sagebrush (*Artemisia tridentata*); X_2 = density (thousands/ha) of black greasewood (*Sarcobatus vermiculatus*); X_3 = density (thousands/ha) of Nuttall's saltbush (*Atriplex nuttalli*); X_4 = density (thousands/ha) of Douglas rabbitbrush (*Chrysothamnus viscidiflorus*); X_5 = slope (degrees); X_6 = distance to water(m); X_7 = East/Northeast indicator variable; X_8 = South/Southeast indicator variable; and X_9 = West/Southwest indicator variable.

As noted in Example 3.2, the fact that the study plots could be used once or twice during the study period means there are several approaches that can be used for analysing the data by logistic regression, depending on what definition of use is applied. Here the obvious possibilities are:

(a) A study plot can be considered to be used if antelopes are recorded in either the first or the second winter (as in the comparisons made in Figure 5.1). On this basis the application of logistic regression to approximate the probability of a plot being used is straightforward.

(b) A study plot can be considered to be used if antelopes are recorded in both years. This leads to probabilities of use that are smaller than for definition (a), but an analysis of the data using logistic regression is still straightforward.

(c) A study plot can be considered to be used when antelopes are recorded for the first time. This turns the situation into one where units are used up because the pool of unused study plots decreases with time. This approach requires that the effect of time be modelled, which therefore means that logistic regression is not a convenient approach for data analysis. A method of analysis that allows for the effect of time is discussed in Chapter 6.

(d) The two years can be considered to be replicates, in which case it is interesting to know whether the nature of the resource selection (if any) was different for the two years. In this case, logistic regression can be used separately in each year, or one equation can be fitted to both years of data. This leads to a more complicated analysis than is needed if one of the definitions (a) and (b) is used but better use is made of the data.

It is approach (d) that will be used for this example. Thus the observational unit will be taken to be a study plot in one year. There are 512 such units, each of which is recorded as either being used or not used. The question of whether it is reasonable to regard the two years as providing independent data is discussed further below. Here it will merely be noted that an examination of this assumption is required in order to establish the validity of the analysis being used.

Initially, three logistic regression models were fitted to the data. For model 1 it was assumed that the RSPF was different for the two winters 1980-81 and 1981-82. Hence the logistic equation (5.1) was fitted separately to the data for each of the two years, with all the variables X_1 to X_9 included. This produced the estimates with standard errors that are shown in Table 5.1.

Chi-squared tests on deviances can be used to assess whether there is any evidence that the probability of use of a study plot was related to one or more of the variables being considered (Section 2.7). This is done by seeing whether the deviance obtained from fitting model 1 is significantly less than the deviance for the 'no selection' model, in comparison with critical values from the chi-squared distribution, with the df being equal to the number of variables in the model.

Table 5.1 Results from fitting model 1 separately to the data on habitat selection by antelopes in 1980-81 and 1981-82.

	1980-81			1981-82		
Variable	Coefficient	Std. err.[1]	P-value[2]	Coefficient	Std. err.	P-value
Constant	-0.896	0.410	0.029	-0.056	0.376	0.882
Sagebrush	0.015	0.044	0.727	0.045	0.041	0.267
Greasewood	0.057	0.073	0.433	-0.038	0.073	0.607
Saltbush	0.02	0.021	0.343	-0.001	0.02	0.967
Rabbitbrush	-0.021	0.046	0.642	-0.086	0.045	0.058
Slope	-0.0043	0.0082	0.603	-0.0043	0.0075	0.565
Distance to water	-0.00035	0.00018	0.051	-0.00031	0.00016	0.054
E/NE aspect	1.003	0.443	0.013	0.534	0.427	0.211
S/SE aspect	0.007	0.519	0.989	-0.068	0.443	0.878
W/SW aspect	0.714	0.393	0.069	-0.332	0.387	0.392

[1] Estimated standard errors output from the fitting process.
[2] The p-values shown are obtained by calculating the ratios of estimates to their standard errors and finding the probability of a value that far from zero for a standard normal variable.

For 1980-81 the null model deviance is 304.2 with 255 df, which is reduced to 286.3 with 246 df for the nine variable model. The reduction in the deviance is 17.9 with nine df, which is significantly large at the 5% level. The equivalent statistic for 1981-82 is 13.0 with nine df, which is not significantly large at the 5% level. There is therefore some evidence of selection in 1980-81 but not in 1981-82. The sum of the two deviance reductions is 30.9 with 18 df. This is a measures of the evidence of selection for both years combined, which is significantly large at the 5% level.

Inspection of the coefficients in Table 5.1 indicates that there is not much evidence that habitat selection was related to the vegetation variables or the slope in either 1980-81 or 1981-82. However, the coefficient for the distance to water is nearly significant at the 5% level in both years, and the East/Northeast dummy variable for aspect is significantly large at about the 1% level for the first year. It was therefore decided to refit the logistic regression equations, again separately for each year, with the vegetation variables and the slope omitted. This resulted in the estimates shown in Table 5.2 for what will be called model 2.

The deviance for model 2 is 289.3 with 251 df for 1980-81, an increase of 3.0 over the deviance for model 1, with an increase of 5 df. This is not at all significantly large, verifying that the decision to drop some of the variables is reasonable. For 1981-82 the deviance for model 2 is 328.8 with 251 df, an increase of 5.2 over model 1, with 5 df. Again, this is not at all significant, indicating that the simpler model is appropriate.

To assess the evidence for selection, the deviances for model 2 in the two years can be compared with the corresponding deviances for the no-selection model. For 1980-81 the reduction in deviance by fitting model 2 instead of the no-selection model is 14.9 with four df, which is significantly large at the 1% level. The same statistic is 7.8 with four df for 1981-82, which is significantly large at the 10% level. The total of 22.7 with eight df is significantly large at the 1% level, giving strong evidence of selection overall.

Table 5.2 Results from fitting model 2 separately to the data on habitat selection by antelopes in 1980-81 and 1981-82.

	1980-81			1981-82		
Variable	Coefficient	Std. err.	P-value	Coefficient	Std. err.	P-value
Constant	-0.655	0.256	0.011	-0.164	0.238	0.490
Distance to water	-0.00044	0.00017	0.010	-0.00033	0.00015	0.030
E/NE aspect	1.036	0.432	0.017	0.561	0.418	0.180
S/SE aspect	-0.086	0.504	0.865	-0.043	0.430	0.921
W/SW aspect	0.613	0.371	0.099	-0.216	0.367	0.555

The somewhat similar coefficients for the two years that are shown in Table 5.2 suggest that it may be possible to get about as good a result by fitting all the data together with a dummy variable introduced to allow for a difference between the years. This produces what will be called model 3. The results of fitting this model are shown in Table 5.3. The dummy variable 'Year' was set equal to 0 for all the 1980-81 results and 1 for all the 1981-82 results.

Table 5.3 Results from fitting model 3 fitted to the combined data for winters 1980-81 and 1981-82.

Variable	Coefficient	Std. err.	P-value
Constant	-0.613	0.199	0.002
Year	0.410	0.194	0.035
Distance to water	-0.00037	0.00010	0.001
E/NE aspect	0.786	0.301	0.009
S/SE aspect	-0.059	0.325	0.860
W/SW aspect	0.180	0.260	0.489

The total deviance for model 2 is 618.1 with 502 df, while the total deviance for model 3 is 621.3 with 506 df. The difference is 3.2 with 4 df, which is not at all significant. Consequently, the simpler model 3 seems better for describing the data.

Looking at the results in Table 5.3, it can be seen that the coefficients of year, distance to water, and the dummy variable for the East/Northeast aspect are all significantly different from zero at the 5% level. The non-significant coefficients for the other two dummy variables for aspect merely indicate that the probabilities of use for the South/Southeast and West/Southwest aspects are about the same as the probabilities for the standard North/Northwest aspect.

Table 5.4 is an analysis of deviance table which summarises the results of comparing the models. The models are listed from the simplest (no selection) to the most complicated (model 1, with all nine variables used and different coefficients estimated for each year). The Akaike information criteria (AIC) values are also shown in this table. The 'best' model with respect to AIC is the one for which the AIC values is lowest (Section 2.8). This is again model 3.

The fact that all the model deviances shown in Table 5.4 are significantly large might be thought to show that none of the models is a satisfactory fit to the data. However, this is not the case because the condition for these statistics to have approximately chi-squared distributions, most values of $A_iw^*(\underline{x})\{1 - w^*(\underline{x}_i)\}$ being five or more, is certainly not met. Hence the chi-square approximation is not reliable for testing the goodness of fit statistics, although it can be used for testing differences between these statistics for different models.

Table 5.4 Analysis of deviance table for assessing models fitted to the data on habitat selection by antelopes.

			Change in		
Model	Deviance	df	Deviance	df	AIC
No selection and different probabilities of use for each winter	640.8[1]	510			644.8
			19.5[2]	4	
Model 3: selection on distance to water and aspect, plus a year difference	621.3[1]	506			633.3
			3.2	4	
Model 2: selection on distance to water and aspect, with effects varying with the year	618.1[1]	502			638.1
			8.1	10	
Model 1: selection on all variables, with effects varying with the year	610.0[1]	492			650.0

[1] Significantly large at the 0.1% level when compared with critical values of the chi-squared distribution (chi-squared approximation is not reliable).
[2] Significantly large at the 0.1% level when compared with critical values of the chi-squared distribution (chi-squared approximation is reliable).

The amount of selection is indicated by Figure 5.2, which shows values of the estimated RSPF

$$\hat{w}^*(\mathbf{x}) = \exp(V)/\{1 + \exp(V)\}, \tag{5.5}$$

where

$$V = \exp\{-0.613 + 0.410(\text{YEAR}) - 0.00037(\text{DW}) + 0.786(\text{E/NE}) \\ - 0.059(\text{S/SE}) + 0.180(\text{W/SW})\},$$

and where YEAR indicates the 0-1 variable for the year, DW indicates the distance to water, and E/NE, S/SE and W/SW are the dummy variables for aspect. The probabilities of use calculated from this function are plotted against the distance from water, separately for the 1980-81 and 1981-82 winters, and the four aspects. There was apparently a maximum probability of use of about 0.65 for East/Northeast study plots close to water in 1981-82, and a minimum probability of use of about 0.20 for South/Southeast plots far from water in 1980-81.

Residual plots can be examined to see whether there are any systematic deviations between the data and model 3. However, the standardized residuals

$$R_i = \{u_i - A_i\,\hat{w}^*(\mathbf{x}_i)\} \,/\, \surd[A_i\,\hat{w}^*(\mathbf{x}_i)\{1 - \hat{w}^*(\mathbf{x}_i)\}] \qquad (5.6)$$

are not very informative because all the group sizes A_i are one. It is therefore unreasonable to expect these residuals to be approximately standard normally distributed. This problem can be overcome by grouping observations, and plotting standardized residuals for groups.

According to model 3, the probability of a study plot being used depended on the aspect, the distance to water, and the year. These are therefore the factors that the grouping of observations should depend on. Having decided this, it must be admitted that any basis for grouping has to be somewhat arbitrary. The method that was used here involved first dividing the 256 study plots into four groups on the basis of their aspect, and then ordering them within each group from those most distant from water to those closest to water. In effect, this meant that within each of the four aspect groups the plots were ordered according to their estimated probabilities of use for model 3. The study plots were then divided into sets of five within each aspect group, so that the first set consisted of the five plots estimated to be least likely to be used, followed by a set of five plots with higher estimated probabilities of use, and so on, with the last set allowed to have more or less than five plots in order to make use of all the plots available.

At this point, equation (5.6) was used to calculate two standardized residuals for each set of study plots within each aspect group, using average values for the estimated probabilities of use. The first of the standardized residuals was based on the plots used in 1980-81, and the second one was based on the plots used in 1981-82. In this way, six standardized residuals were obtained for the East/Northeast aspect in 1981-82 and another six for this aspect in 1981-82. Similarly, six standardized residuals were calculated for the South/Southeast aspect for each of the two years, nine standardized residuals for the West/Southwest aspect for each of the two years, and 30 standardized residuals for the North/Northwest aspect for each of the two years.

According to model 3, the ordering of the 256 study plots by their probabilities of use was the same in 1980-81 and 1981-82, although the probabilities were slightly higher in the second year. One interesting residual graph is therefore of the two standardized residuals for each plot against the estimated probability of use in 1980-81. If model 3 is correct then this graph is expected to show no patterns at all, with most of the standardized residuals within a range from -2 to +2, and almost all of them within a range from -3 to +3.

Figure 5.3 shows graphs of this type, separately for each of the four aspects. It can be seen from this figure that all of the standardized residuals are within a reasonable range, but there are some disturbing patterns in the graphs for two of the aspects. In particular:

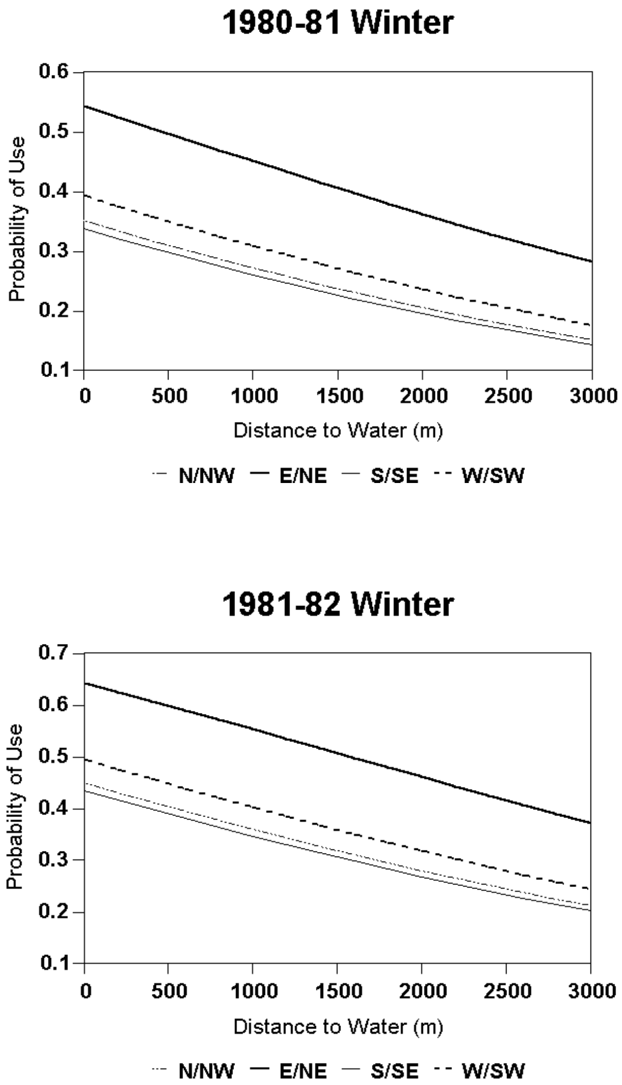
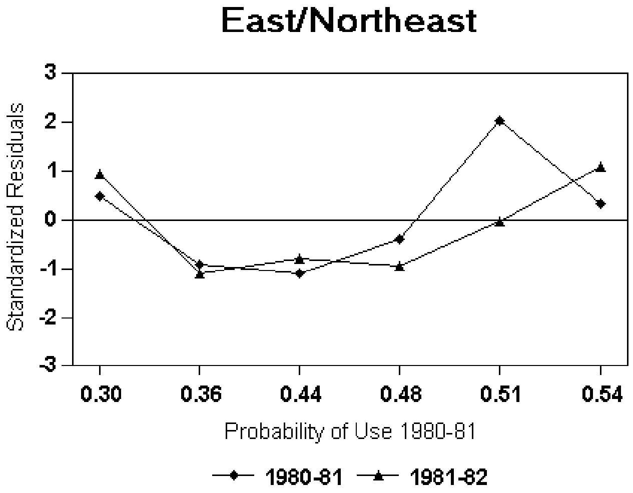
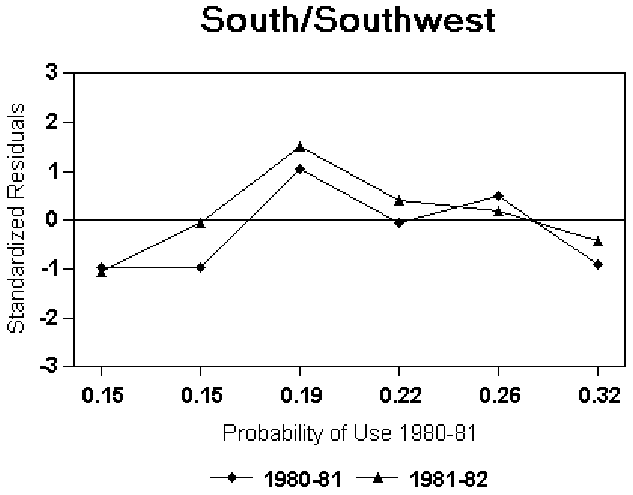


Figure 5.2 Probabilities of study plots being used by antelopes according to the resource selection probability function (5.5).


East/Northeast
3
2
1
0
-1
-2
-3
Standardized Residuals
0.30
0.36
0.44
0.48
0.51
0.54
Probability of Use 1980-81
1980-81
1981-82


South/Southwest
3
2
1
0
-1
-2
-3
Standardized Residuals
0.15
0.15
0.19
0.22
0.26
0.32
Probability of Use 1980-81
1980-81
1981-82

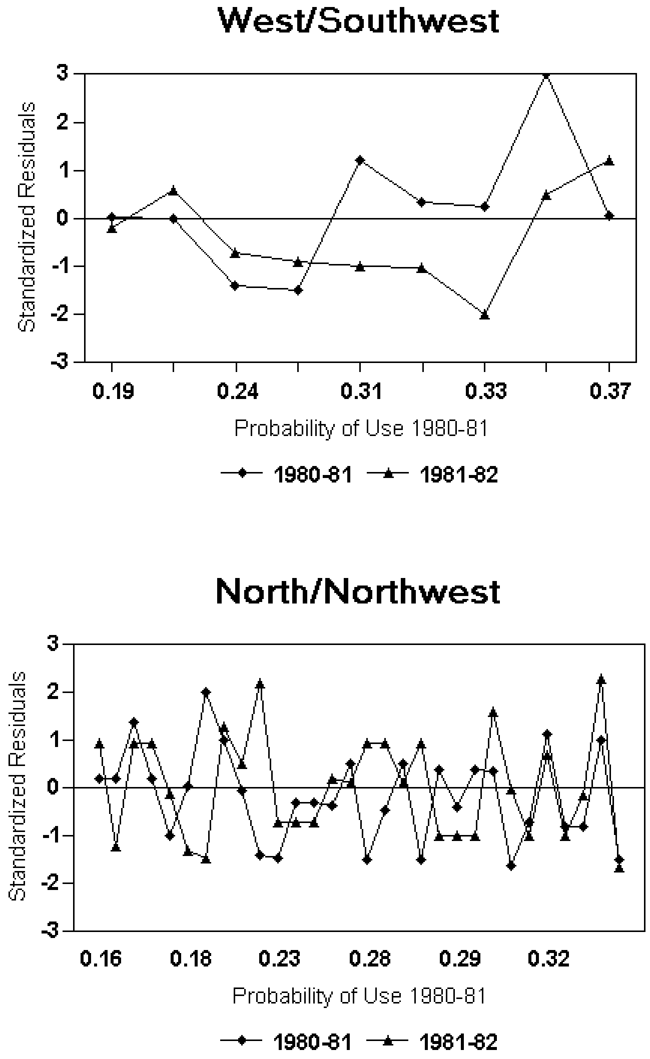


Figure 5.3 Standardized residuals plotted against the estimated probability of use in 1980-81, separately for each of the four aspects.

(a) For the East/Northeast aspect there are only six sets of study plots but the residuals are so similar for 1980-81 and 1981-82 that the assumption of independent data in the two years looks suspect. The Pearson correlation between the residuals for the two years is quite high at 0.48, although this is not significantly different from zero at the 5% level.

(b) The graph for the South/Southeast plots also indicates that the data are not independent for the two years. In this case the Pearson correlation is 0.85, which is significantly different from zero even with only six pairs of standardized residuals.

If the patterns for the first two aspects were repeated for the other two aspects as well then there would be little doubt that the assumption of independent data for the two years is untenable. However, the graphs for West/Southwest and North/Northwest study plots show little indication of dependence between the data for the two years, with the Pearson correlation coefficients being 0.22 based on nine pairs of standardized residuals, and 0.09 based on 30 pairs of standardized residuals, respectively.

Taken overall the residual plots do not show clear evidence of dependence between the data for 1980-81 and 1981-82 because the correlation between the standardized residuals for East/Northeast and South/Southeast plots is obscured by the low correlations for the other two aspects. In fact the Pearson correlation for all 51 pairs of standardized residuals is 0.20, which is not significantly different from zero at the 5% level. Still, there is cause for some concern and it is appropriate to conclude this example with a brief discussion of alternative explanations for the correlations indicated for the East/Northeast and South/Southeast residuals.

One explanation is, of course, that the behaviour of antelope is consistent from year to year so that an individual animal tends to be seen in the same study plots every year. If this is true then the data obtained for different years will be correlated, with the result that residuals will also be correlated. If this is the situation then the estimated RSPF may still provide good estimates of probabilities of use for different plots. However, the calculated standard errors of ß estimates will be too small because the effective amount of data available is less than the apparent amount. Also, the significance of differences in chi-squared values for the fits of different models will be exaggerated.

An alternative explanation for residuals from different years being correlated is that one or more important variables is missing from the resource selection probability function. In this case, the probability of use will be underestimated for some study plots and overestimated in others, and this bias will be present in both years. Consequently, there may be some plots with a high probability of use that are estimated to have a low probability of use. These plots will tend to be used in both years and hence provide positive residuals in both years. On the other hand, study plots with a low probability of use that are estimated to have a high probability of use will tend to give negative residuals in both years.

If this second explanation for correlated residuals is correct then there is a problem because estimated probabilities of use may be seriously biased. In the case of the present example there is nothing that can be done about this without taking further measurements on the plots of land. However, the graphs in Figure 5.3 suggest that if there is a missing variable then the values of this variable are related to the distance to water in the East/Northeast study plots and in the South/Southeast study plots in much the same way as the standardized residuals, but show little relationship to the distance to water for study plots with the other two aspects.

5.3 Separate Sampling

The situation is more complicated if independent separate random samples are taken of different types of unit. Logistic regression can then still be used, but it needs a special justification, which depends on the types of samples involved. Three situations are considered here: (a) there is a sample of the available units and a sample of the used units in the population, (b) there is a sample of the available units and a sample of the unused units in the population, and (c) there is a sample of unused units and a sample of used units. These cases will now be considered in turn.

5.4 Separate Samples of Available and Used Resource Units

The types of situation that we envisage for the separate sampling of available and used units are:

- A random sample is taken of the trees in an area that might be used for nests of a species of bird, and a random sample of trees with nests is taken in the same area. Characteristics of the trees in both samples are measured to determine which of these seems important for the selection of nesting sites by the birds.

- The locations of groups of moose is recorded from aerial surveys in a national park, and a sample of available locations is selected from a geographical information system (GIS). The characteristics of the locations in both samples are determined from the GIS, possibly in terms of the pixels where they occur, to see what type of location is selected by the moose.

- A random sample is taken of the prey available to the predators in an area, and stomach samples of the predators are taken to see which types of prey they are selecting for food.

For these situations, suppose that the population of available units is of size N, with the ith unit having the values $\mathbf{x}_i = (x_{i1},x_{i2},...,x_{ip})$ for the variables X_1 to X_p, and a corresponding probability of $w^*(\mathbf{x}_i)$ of being used after a certain amount of time. Suppose also that the sampling scheme is such that every available unit has a probability P_a of being sampled, and every used unit has a probability P_u of being sampled, withe the available sample being selected first without replacement so that units in this sample cannot also appear in the used sample. In that case, the probability of a unit being used and sampled is $(1 - P_a)w^*(\mathbf{x}_i)P_u$ and the probability of a unit being in either the available or the used sample is

$$\text{Prob(ith unit sampled)} = P_a + (1 - P_a)w^*(\mathbf{x}_i)P_u. \tag{5.7}$$

It then follows that the probability that the ith unit is in the used sample, given that it is in one of the samples is

$$\begin{aligned}\text{Prob(ith unit used|sampled)} &= \text{Prob(used and sampled)/Prob(sampled)}\\ &= (1 - P_a)w^*(\mathbf{x}_i)P_u/\{P_a + (1 - P_a)w^*(\mathbf{x}_i)P_u\}. \qquad (5.8)\end{aligned}$$

It is convenient at this point to assume that the resource selection probability function takes the particular form

$$w^*(\mathbf{x}_i) = \exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}), \qquad (5.9)$$

where the argument of the exponential function should be negative. Then, letting $\tau(\mathbf{x}_i)$ = Prob(ith unit used|sampled), equation (5.8) can be written

$$\tau(\mathbf{x}_i) = \frac{\exp\{\log_e[(1 - P_a)P_u/P_a] + \beta_0 + \beta_1 x_{i1} +...+ \beta_p x_{ip}\}}{1 + \exp\{\log_e[(1 - P_a)P_u/P_a] + \beta_0 + \beta_1 x_{i1} +...+ \beta_p x_{ip}\}} \qquad (5.10)$$

This is a logistic regression equation in which the parameter β_0 is modified to

$$\beta_0' = \log_e[(1 - P_a)P_u/P_a] + \beta_0$$

to allow for the sampling probabilities of available and used.

Assuming independence of observations, the probability of observing resource unit i as used is $\tau(\mathbf{x}_i)$ and the probability of observing it as available is $1-\tau(\mathbf{x}_i)$. Let y_i be an indicator of whether a sampled unit was used. That is, $y_i = 0$ if sampled unit i came from the available sample, $y_i = 1$ if unit i came from the sample of used units. The probability of observing unit i can then be written as

$$L_i = \tau(\mathbf{x}_i)^{y_i}\{1 - \tau(\mathbf{x}_i)\}^{1-y_i}$$

and the log-likelihood of observing the entire sample is,

$$\log_e\{ L(\beta_0,\beta_1,...,\beta_p)\} = \sum_{i=1}^{n} \log L_i = \sum_{i=1}^{n} [y_i \log_e\{\tau(\mathbf{x}_i)\} + (1 - y_i)\log_e\{1 - \tau(\mathbf{x}_i)\}].$$

This log likelihood is identical to a binomial log likelihood with the number of trials set to 1. Consequently, standard logistic regression programs can be used to estimate the coefficients β_0, β_1, ..., β_p. Variance estimates for the coefficients computed by standard logistic regression estimation routines are based on the second derivative of the log likelihood surface. Variance estimates computed by the logistic regression routine are correct for this situation regardless of the size of either the sample of used or available units and provided that the probability of a unit being in both samples is negligible.

The fact that the constant in the logistic regression is $\log_e[(1 - P_a)P_u/P_a] + \beta_0$ means that if the sampling probabilities P_u and P_a are known then the parameter β_0 in the resource selection probability function can be estimated by subtracting $\log_e[(1 - P_a)P_u/P_a]$ from the estimated constant in the logistic regression equation. If the sampling fractions are not known then β_0 cannot be estimated, but it is still possible to estimate the resource selection function (RSF)

$$w(\mathbf{x}) = \exp(\beta_1 x_1 + ... + \beta_p x_p) \qquad (5.11)$$

and use this to compare resource units.

The sampling scheme used to derive equation (5.10) will often be reasonable with field data, particularly for the sample of used units, which are often found by searching the area where selection is taking place. In some cases, however, sample sizes will be fixed in advance rather than being decided by giving each resource unit a probability of inclusion. For example, the sample of available units may be obtained from a GIS, in which case it would be common to just take a simple random sample of n from the population of N units.

Equations (5.7) to (5.10) still hold with one or both of the samples having a size fixed in advance, but with P_a and P_u defined as sampling fractions rather than sampling probabilities, if necessary. The use of logistic regression for estimation is therefore still justified. But there is the complication that fixing sample sizes introduces some dependency in the dependent variable for the logistic regression, which may affect the properties of estimators. For instance, suppose that the sample sizes are 100 for both available and used units. Then the data for the logistic regression will consist of 200 observations, of which exactly 100 are used. The constraint that exactly 100 units are used means that the 200 observations are not completely independent, which is the usual assumption for logistic regression. To avoid this type of complication, it is best to use the sampling scheme whereby available and used resource units have probabilities P_a and P_u of being included in their respective samples so that logistic regression likelihood, that assumes independence applies.

Of course, if sample sizes are fixed in advance then it is always possible to use bootstrapping to assess variances, with the bootstrap sampling designed to mimic the sampling used to collect the real data.

When using data from a GIS system, information is recorded on all the available units but the number of these may be astronomical, making the use of all the data difficult or impossible even with modern computers. This leads to the idea of taking a large systematic sample of the available units to get a good 'representative' sample, which will represent the full population of available units with negligible error for a logistic regression.

The question then arises as to whether it is valid to use the systematic sample as if it is effectively the same as a sample drawn such that each available unit has a probability P_a of selection for a logistic regression analysis. This question can be answered in two ways. First, it can be argued that the systematic sample should represent the population of available units better than the random sample obtained by giving each unit a probability P_a of selection. This suggests that, if anything, treating the systematic sample as a random sample will mean that the level of sampling errors indicated by variances will be overestimated. The analysis should therefore be conservative in this respect. On the other hand, if the systematic sample is large enough then it should represent the population of available units with negligible sampling errors, as would a random sample of the same size. On this basis, the systematic sample is effectively equivalent to a random sample of the same size.

What does seem important under these circumstances is to ensure that the sample of available units is large enough so that it leads to negligible sampling errors, whether it is systematic or random. To investigate this it is sensible to take several samples at each of several sizes and making sure that for the final size used the results obtained are very similar with alternative samples.

5.5 Separate Samples of Available and Unused Resource Units

Separate samples of available and unused resource units occur with situations such as:

- The prey items available in an area are sampled before and after a predator is introduced, to determine what type of prey the predator chooses.

- Plots of land not used by an animal are sampled, and compared with a sample of all plots in the study area. Characteristics of the plots are measured to determine which are related to the probability of use by the animal.

It is assumed here that samples are obtained in such a way that every available unit has a probability P_a of being included in the sample of available units, and every unused unit has a probability $P_{\bar{u}}$ of being included in the sample of used units, with the available sample taken first without replacement, so that units cannot appear in both samples. As before, the population of available units is of size N, with the ith unit being described by values $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{ip})$ for the p variables $(X_1, X_2, ..., X_p)$.

With these assumptions, the probability of the ith resource unit in the population being in one of the samples is

$$\text{Prob(ith unit sampled)} = P_a + (1 - P_a)\{1 - w^*(\mathbf{x}_i)\}P_{\bar{u}}. \qquad (5.12)$$

The conditional probability of the ith unit being unused, given that it is sampled is therefore

$$\text{Prob(ith unit unused|sampled)} = \tau_i = \frac{(1 - P_a)\{1 - w^*(\mathbf{x}_i)\}P_{\bar{u}}}{P_a + (1 - P_a)\{1 - w^*(\mathbf{x}_i)\}P_{\bar{u}}].} \qquad (5.13)$$

If it is then further assumed that the RSPF is well approximated by

$$w^*(\mathbf{x}) = 1 - \exp(\beta_0 + \beta_1 x_1 +...+ \beta_p x_p), \qquad (5.14)$$

where the argument of the exponential function should be negative, then equation (5.13) can be written as

$$\tau_i = \frac{\exp\{\log_e[(1 - P_a)P_{\bar{u}} /P_a] + \beta_0 + \beta_1 x_{i1} +...+ \beta_{ip}x_{ip}\}}{1 + \exp\{\log_{e[}(1 - P_a)P_{\bar{u}} /P_a) + \beta_0 + \beta_1 x_{i1} +...+ \beta_{ip}x_{ip}\}} \qquad (5.15)$$

This is a model that can be fitted using logistic regression, with the dependent variable being an indicator variable that is one if a sampled unit is in the unused sample, or otherwise zero.

The constant term in the fitted logistic regression equation equates to

$$\beta_0' = \log_e[(1 - P_a)P_{\bar{u}} /P_a] + \beta_0.$$

It follows that β_0 can be estimated only if the sampling probabilities are known. If this is not the case, then the best that can be done is to note that

$$1 - w^*(\mathbf{x}) = \exp(\beta_0 + \beta_1 x_1 +...+ \beta_p x_p)$$

is the probability of the ith resource unit surviving (i.e. being unused). Therefore the function

$$\exp(ß_1x_1 +...+ ß_px_p)$$

which can be estimated, gives relative probabilities of different types of unit not being used.

The comments at the end of Section 5.5 regarding sampling schemes also apply here. Logistic regression can still be justified if one or both of the available and unused samples have sizes fixed in advance, but this is best avoided because it means that the observations for the logistic regression are no longer strictly speaking independent. It may be desirable to use bootstrapping to assess variances for sampling schemes that are different from what is assumed in this section.

5.6 Separate Samples of Used and Unused Resource Units

The types of situation now considered are ones like the following:

- Samples of the prey items in an area are sampled after predation by animals, and stomach samples are taked of used prey items. These samples are compared to see which types of prey are selected by the animals.

- A study area is divided into plots where it is easy to see which plots have been used by animals, but recording information on the plots is a time consuming process. This information is obtained for a sample of the used plots and a separate sample of the unused plots to determine which characteristics of the plots are related to the probability of use.

Here the samples are assumed to be obtained in such a way that every used unit has a probability P_u of being included in the sample of used units, and every unused unit has a probability $P_{\bar{u}}$ of being included in the sample of unused units. As before, the population of available units is of size N, with the ith unit being described by values $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{ip})$ for the p variables $(X_1, X_2, ..., X_p)$.

The probability of the ith unit being in one of the two samples is now

$$\text{Prob(ith unit sampled)} = w^*(\mathbf{x}_i)P_u + \{1 - w^*(\mathbf{x}_i)\}P_{\bar{u}}, \qquad (5.16)$$

and the conditional probability of the ith unit being used, given that it is sampled is

$$\text{Prob(ith unit used|sampled)} = \tau_i = w^*(\mathbf{x}_i)P_u/[w^*(\mathbf{x}_i)P_u + \{1 - w^*(\mathbf{x}_i)\}P_{\bar{u}}]. \quad (5.17)$$

This can be rewritten as

$$\tau_i = \frac{(P_u/P_{\bar{u}})w^*(\mathbf{x}_i)/\{1 - w^*(\mathbf{x}_i)\}}{1 + (P_u/P_{\bar{u}})w^*(\mathbf{x}_i)/\{1 - w^*(\mathbf{x}_i)\}}$$

which defines a logistic regression function by setting

$$(P_u/P_{\bar{u}})w^*(\mathbf{x}_i)/\{1 - w^*(\mathbf{x}_i)\} = \exp(ß_0 + ß_1x_{i1} + ... + ß_px_{ip}),$$

in which case

$$\tau_i = \frac{\exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_{ip} x_{ip})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_{ip} x_{ip})}, \tag{5.18}$$

and

$$w^*(\mathbf{x}_i) = \frac{\exp\{\log_e(P_{\bar{u}} / P_u) + \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}\}}{1 + \exp\{\log_e(P_{\bar{u}} / P_u) + \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}\}}. \tag{5.19}$$

The parameters in equation (5.18) may be estimated by carrying out a logistic regression using the resource units in the used and unused samples as the observations, with the dependent variable being an indicator of whether a unit is used. An estimated RSPF can then be obtained by substituting these estimates into equation (5.19), providing that the ratio of sampling probabilities $P_{\bar{u}}/P_u$ is known.

If the ratio of sampling probabilities is not known and cannot be estimated then it is not possible to estimate the RSPF, or even this function multiplied by some unknown constant. The best that can be done is to arbitrarily set $P_{\bar{u}}/P_u = 1$ in equation (5.19), and recognize that the estimated function thereby obtained is an index of selectivity in the sense that if resource units are ranked in order using this function then they are being placed in the same order as they would be if the ratio of sampling probabilities were known.

As before, the comments at the end of Section 5.5 regarding sampling schemes apply. Logistic regression can still be justified if one or both of the available and unused samples have sizes fixed in advance, but this is best avoided because it means that the observations for the logistic regression are no longer strictly speaking independent. It may be desirable to use bootstrapping to assess variances for sampling schemes that are different from what is assumed in this section.

Example 5.2 Nest Selection by Fernbirds

Harris' (1986) study on nest selection by fernbirds (*Bowdleria puncta*) that produced the data shown in Table 3.4 has been discussed in Example 3.4. There is a sample of available resource units (random points in the study region) and a sample of used resource units (nest sites), so that the method described in Section 5.4 applies. Sampling fractions are unknown, but are clearly very small.

A logistic regression was carried out, with the dependent variable being 0 for available sites and 1 for nest sites, and the three variables canopy height, distance to edge, and perimeter of clump used as predictor variables. This produced the fitted equation

$$\hat{\tau} = \frac{\exp\{-10.73 + 7.80(\text{CANOPY}) + 0.21(\text{EDGE}) + 0.88(\text{PERIM})\}}{1 + \exp\{-10.73 + 7.80(\text{CANOPY}) + 0.21(\text{EDGE}) + 0.88(\text{PERIM})\}}$$

with obvious abbreviations for the variables. The deviance for this model is 40.48 with 45 df, compared to the deviance of 67.91 with 48 df for the no selection model. The difference in these deviances is 27.43 with 3 df. This is very highly significant in comparison with the chi-squared distribution ($p < 0.001$), giving strong evidence for selection.

The standard errors for the coefficients of CANOPY, EDGE and PERIM are 3.25, 0.12, and 0.48, respectively. Using the ratios of the estimated coefficients to their standard errors to test for the significance of these estimates against the standard normal distribution (Section 2.7) gives 7.80/3.25 = 2.40 ($p = 0.016$) for CANOPY, 0.21/0.12 = 1.73 ($p = 0.083$) for EDGE, and 0.88/0.48 = 1.84 ($p = 0.066$) for PERIM. The significance is therefore borderline for EDGE and PERIM. However, if EDGE is removed from the equation then the deviance for the model increases by 3.61 with 1 df, which is nearly a significant change at the 5% level ($p = 0.057$), while if PERIM is dropped then the deviance increases by 3.98 with 1 df, which is significant at the 5% level ($p = 0.046$). It therefore seems reasonable to accept the model as it stands, with all three variables included.

The fitted logistic regression equation corresponds to equation (5.10). The RSPF would therefore be estimated by

$$\hat{w}^*(\mathbf{x}) = \exp\{-10.73 - \log_e[(1-P_a)P_u/P_a] + 7.80(\text{CANOPY}) + 0.21(\text{EDGE})$$
$$+ 0.88(\text{PERIM})\}$$

if the sampling probabilities P_u and P_a were known. Because these are not known, all that can be estimated is the RSF that is obtained by omitting the constant terms from the last equation, i.e.,

$$\hat{w}(\mathbf{x}) = \exp\{7.80(\text{CANOPY}) + 0.21(\text{EDGE}) + 0.88(\text{PERIM})\}.$$

Figure 5.4 shows how the values from this function compare when it is evaluated at the nest sites and the random sites. To keep the values within a reasonable range, the function has been scaled so that it takes the value 1.0 when CANOPY, EDGE and PERIM are equal to the mean values for these variables at the random sites. This is done by evaluating the equation for each of the sites and then dividing by the value of the function when CANOPY = 0.49, EDGE = 12.6, and PERIM = 2.93 (Table 3.4). Even with this scaling, the range of values is very large, requiring a logarithmic scale for the plot. It appears, therefore that there was very considerable selection in the choice of sites by the fernbirds.

5.7 Variances for Estimators and Their Differences

With logistic regression and other models for RSPFs the amount of selection for or against a particular type of resource unit, or the comparison between the selection for two types of resource units involves the consideration of exponential functions of estimated parameters. It is therefore useful at this point to review of statistical methods that can be used to assess the accuracy of estimates of exponential functions and their differences.

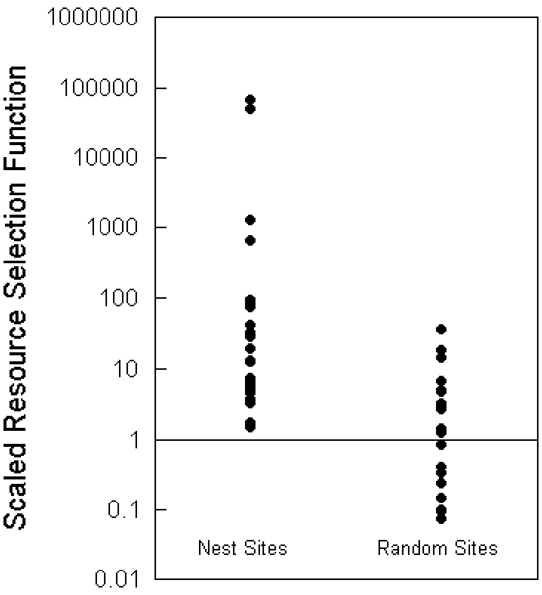


Figure 5.4 Values of the resource selection function estimated for fernbirds, with the values scaled so that the value is 1.0 for a site with the average values of the predictor variable at the randomly located sites.

First, suppose that a logistic function for the probability that a resource unit with measurement x_1 to x_p has been estimated. This then takes the form

$$\hat{w}^*(\mathbf{x}) = \frac{\exp(\hat{ß}_0 + \hat{ß}_1 x_1 + ... + \hat{ß}_p x_p)}{1 + \exp(\hat{ß}_0 + \hat{ß}_1 x_1 + ... + \hat{ß}_p x_p)}. \qquad (5.20)$$

The variance of this function can be determined using the Taylor series method (Manly, 1985, p. 408) to be approximately

$$\mathrm{var}\{\hat{w}^*(\mathbf{x})\} = w^*(\mathbf{x})^2\{1 - w^*(\mathbf{x})\}^2 \sum_{i=0}^{p} \sum_{j=0}^{p} x_i x_j \mathrm{cov}(\hat{ß}_i,\hat{ß}_j), \qquad (5.21)$$

taking $x_0 = 1$. Here $\mathrm{cov}(\hat{ß}_i,\hat{ß}_j)$ is the variance of $\hat{ß}_i$ if i=j, or is otherwise the covariance between $\hat{ß}_i$ and $\hat{ß}_j$, where these variances and covariances should be available as part of output from the computer program used to fit the logistic function. To use the equation the true value of the RSPF will need to be replaced by the estimate from equation (5.20).

The Taylor series method also shows that if the difference between the probability of use for a resource unit with X values $\mathbf{x}_1 = (x_{11},...,x_{1p})$ and the probability of use for a resource unit with $\mathbf{x}_2 = (x_{21},...,x_{2p})$ is estimated by $\hat{w}^*(\mathbf{x}_1) - \hat{w}^*(\mathbf{x}_2)$, then this estimator has the approximate variance

$$\mathrm{var}\{\hat{w}^*(\mathbf{x}_1) - \hat{w}^*(\mathbf{x}_2)\} = w^*(\mathbf{x}_1)\{1 - w^*(\mathbf{x}_1)\}w^*(\mathbf{x}_2)\{1 - w^*(\mathbf{x}_2)\}$$

$$\times [\sum_{i=0}^{p} \sum_{j=0}^{p} x_{1i}\, x_{2j}\, \mathrm{cov}(\hat{\beta}_i,\hat{\beta}_j)]\ , \qquad (5.22)$$

taking $x_{10} = x_{20} = 1$. Again the estimated values of the RSPF will have to replace the true values in order to apply the equation.

Often the estimated RSPF takes the form

$$\hat{w}(\mathbf{x}) = \exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_p x_p). \qquad (5.23)$$

Then the Taylor series method gives the approximation

$$\mathrm{var}\{\hat{w}(\mathbf{x})\} = w(\mathbf{x})^2 \sum_{i=0}^{p} \sum_{j=0}^{p} x_i\, x_j\, \mathrm{cov}(\hat{\beta}_i,\hat{\beta}_j), \qquad (5.24)$$

and

$$\mathrm{var}\{\hat{w}(\mathbf{x}_1) - \hat{w}(\mathbf{x}_2)\} = w\,(\mathbf{x}_1)\, w(\mathbf{x}_2) \sum_{i=0}^{p} \sum_{j=0}^{p} x_{1i}\, x_{2j}\, \mathrm{cov}(\hat{\beta}_i,\hat{\beta}_j)], \qquad (5.25)$$

taking $x_{10} = x_{20} = 1$.

It may be desirable to compare the ratio of two function values rather than the difference. To this end it can be noted that

$$\hat{w}(\mathbf{x}_1)/\hat{w}(\mathbf{x}_2) = \exp\{\hat{\beta}_1(x_{11} - x_{21}) + ... + \hat{\beta}_p(x_{1p} - x_{2p})\},$$

so that equation (5.24) provides the result

$$\mathrm{var}\{\hat{w}(\mathbf{x}_1)/\hat{w}(\mathbf{x}_2)\} = \{w(\mathbf{x}_1)/w(\mathbf{x}_2)\}^2 \sum_{i=1}^{p} \sum_{j=1}^{p} (x_{1i} - x_{2i})\,(x_{1j} - x_{2j})\, \mathrm{cov}(\hat{\beta}_i,\hat{\beta}_j). \quad (5.26)$$

In the next chapter the estimated RSPF

$$\hat{w}^*(\mathbf{x}) = 1 - \exp\{ -\exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_p x_p)\} \qquad (5.27)$$

is used. Here the Taylor series method gives the approximate variances

$$\mathrm{var}\{\hat{w}^*(\mathbf{x})\} = w^*(\mathbf{x})^2\, [\log_e\{1 - w^*(\mathbf{x})\}]^2 \sum_{i=0}^{p} \sum_{j=0}^{p} x_i\, x_j\, \mathrm{cov}(\hat{\beta}_i,\hat{\beta}_{\Sigma j}), \qquad (5.28)$$

and

$$\mathrm{var}\{\hat{w}^*(\mathbf{x}_1)-\hat{w}^*(\mathbf{x}_2)\} = w^*(\mathbf{x}_1)\log_e\{1-w^*(\mathbf{x}_1)\}w^*(\mathbf{x}_2)\log_e\{1 - w^*(\mathbf{x}_2)\}$$

$$\times [\sum_{j=0}^{p} \sum_{j=0}^{p} x_{1i}\, x_{2j} \,\mathrm{cov}(\hat{\beta}_i,\hat{\beta}_j), \qquad (5.29)$$

taking $x_{10} = x_{20} = 1$.

Example 5.3 Logistic Selection Function for Habitat Selection by Antelopes

Example 5.1 was concerned with the selection of winter habitat by antelopes in the Red Rim area of south-central Wyoming. The analysis of the data in this case led to the estimated RSPF that is given by equation (5.5).

The matrix of variances and covariances for the estimated constant term -0.613 and the coefficients of YEAR, DW, E/NE, S/SE and W/SW are shown in Table 5.5, where this was output from the computer program used to carry out the estimation. The elements of this matrix are the covariance values that are needed for evaluating equations (5.21) and (5.22).

Table 5.5 The covariance matrix obtained from the computer program used to estimate a resource selection function for habitat selection by antelopes, where the value in a cell of the table is the covariance between the estimated coefficients for the variables shown in the row and column labels.

	Constant	Year	DW	E/NE	S/SE	W/SW
Constant	4.0401E-02	-1.9575E-02	-1.3057E-05	-1.4944E-02	-9.6384E-03	-1.4058E-02
Year	-1.9575E-02	3.7636E-02	-5.8666E-07	1.3431E-03	-6.2856E-05	3.0264E-04
DW	-1.3057E-05	-5.8666E-07	1.2544E-08	-1.5508E-06	-5.8424E-06	-1.8637E-06
E/NE	-1.4944E-02	1.3431E-03	-1.5508E-06	9.0601E-02	1.6579E-02	1.6122E-02
S/SE	-9.6384E-03	-6.2856E-05	-5.8424E-06	1.6579E-02	2.1625E-02	1.6764E-02
W/SW	-1.4058E-02	3.0264E-04	-1.8637E-06	1.6122E-02	1.6764E-02	6.7600E-02

One application of equation (5.21) is to find confidence intervals for true probabilities of use. For example, consider just the East/Northeast study plots in 1980-81. For these plots, the variables YEAR, S/SE and W/SW are always zero, which means that the variances and covariances associated with these terms do not contribute to the sum on the right-hand side of equation (5.21). A further simplification is that the E/NE variable is always equal to one. The result is that the equation for the variance of $\hat{w}^*$ is fairly straightforward to apply.

Figure 5.5 shows the estimated RSPF for study plots at different distances to water, with approximate 95% confidence limits that are the estimated probabilities plus and minus 1.96 estimated standard errors. It is apparent that the function is not well estimated in this case.

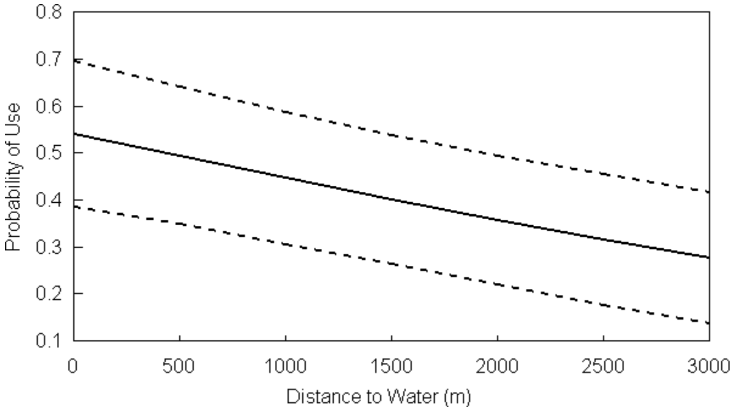


Figure 5.5 Probabilities of use by pronghorn for East/Northeast study plots in 1980/81, as a function of the distance to water. The estimated RSPF $\hat{w}^*$ is the continuous line and the broken lines are approximate 95% confidence intervals given by $\hat{w}^* \pm 1.96se(\hat{w}^*)$.

5.8 Discussion

The logistic regression approaches described in this chapter has much to recommend them when there is no need to take into account varying amounts of selection time. This type of model is used widely for other biological applications, and many computer programs for estimation exist.

The two examples that have been presented are design I studies in the terminology of Chapter 1, with resource availability and use being measured at the population level. However, this does not mean that logistic regression cannot be used with other designs. The antelope study would have had design II if the use of study plots by individual animals had been recorded. In principle, it would then have been possible to estimate a RSPF for each animal using logistic regression. An interesting question would then be whether a model that allows each animal to have a different RSPF gives a significantly better fit to the data than a model that assumes all animals have the same function. In a similar way, differences between sexes, age groups, etc. could be studied.

With a design III study, availability is measured for each animal as well as use. This again would permit a separate RSPF to be estimated for each animal, and tests for differences between animals or groups of animals would be possible.

With either a design II or design III study it would be desirable to have enough animals to use differences between them to assess the accuracy of estimated RSPFs, rather than relying on the standard errors produced by computer programs for logistic regression. Thus, if equation (5.1) is estimated using separate data for n animals, then the standard error of $\hat{\beta}$ can be estimated with n - 1 df using the observed standard deviation of the n individual estimates. In this way, the estimation of a RSPF for each animal gives a first stage analysis, and inferences concerning the population of animals can be carried out using a second stage analyses.

As noted in Chapter 1, the advantage of this approach is that it is still valid even if the observations on each animal are not independent, providing that different animals do give independent observations. If different animals do not give independent observations then it may still be possible to isolate independent groups of animals and conduct a first stage analysis on each of these groups. In that case, second stage analyses can be based on regarding the groups as providing replicates.

Chapter Summary

- Logistic regression is a useful way to model the probability that a resource unit described by certain variables X_1 to X_p is used during a period of selection, given information on which units were used in the population. The calculations can be carried out by many standard statistical packages.

- If the use or non-use is only known for a random sample of resource units from a population then the logistic regression function can still be estimated in the usual way. However, if the units for which information is available are not a random sample then it may be necessary to redefine the population of interest.

- An example is provided where the use of 256 study plots in the Red Rim area of Wyoming, USA, by antelopes in two years is related to vegetation densities, slope, distance to water, and the aspect of the plots.

- Situations are considered where used, unused or available resource units are sampled separately. The estimation of resource selection functions is discussed for the three cases where there are two samples of units, consisting of (a) available and used, (b) available and unused, and (c) unused and used.

- An example involving the selection of nest sites by fernbirds in Otago, New Zealand, is provided where logistic regression is used to estimate a resource selection function from a sample of available sites and a sample of nest sites.

- Equations are provided for the variances of estimates from resource selection functions, and differences between such estimates. The results from these calculations are illustrated on the resource selection probability function estimated for antelopes in Wyoming.

- The uses of logistic regression with design II (availability measured at the population level and use measured for individual animals) and design III (availability and use measured for individual animals) studies is discussed.

Exercises

1. Many of the problems facing the biologist in studying resource selection by animals are also found by the archaeologist studying the use of resources by human societies. One such study concerns the location of prehistoric Maya sites within the Corozal District of Belize in Central America. The investigator was Green (1973) who discusses the proposition that "sites were located so as to minimize the effort expended in acquiring scarce resources". The resource units being considered are plots of land. The whole study area was divided into 151 of these, each being a square with 2.5 km sides. Thirteen variables were measured on each square, related to soil types, vegetation types, distance to navigable water, the distance to Santa Rita (a possible prehistoric commercial and political centre), and the number of sites in neighbouring squares. One or two sites were known to exist on 29 of the squares, giving 34 sites in total.

The data for a selection of the variables measured by Green are shown in Table 5.6. Use logistic regression to see whether the presence of one or more sites on a square can be related to the measured characteristics. A point to note with this example is that the existence of some misclassification has to be accepted because Maya sites that have not been found may well exist on some of the squares that are recorded as being unused. Thus the estimated probability of a square being used will in fact be an estimate of the probability of use multiplied by the probability of a site being discovered. This need be of no concern providing that the probability of a site being discovered is approximately constant for all the existing sites. Exactly the same problem occurs with Ryder's study of habitat selection by antelopes where the classification of a plot of land as used depends on an antelope being sighted at least once on that land.

(2) Table 5.7 shows plankton and yellow perch (*Perca flavescens*) stomach samples of *Daphnia publicaria* taken by Wong and Ward (1972) on five different days in 1969 from West Blue Lake, Manitoba, Canada. The investigators recorded the lengths of the *D. publicaria* in both samples, and considered the question of whether the predators were selective, and whether the selection changed with time. Note that this is a situation where there is a (plankton) sample of available resource units and a (stomach) sample of used resource units, for each of the five sample times. Use logistic regression to estimate a resource selection function of the form

$$w^*(x,t) = \exp(ß_0 + ß_1x + ß_2x^2 + ß_3x^3),$$

for each of the sample days. Use appropriate tests to compare the fit of the linear, quadratic and cubic models. Discuss the nature of the changes in the resource selection function over time.

Table 5.6 Data on the presence of prehistoric Maya sites in the Corozal District of Belize in Central America.*

	Number	Soil percentages				Vegetation percentages				Other variables			
Plot	of sites	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
1	0	40	30	0	30	0	25	0	0	1	0.5	30	15.0
2	0	20	0	0	10	10	90	0	0	2	0.5	50	13.0
3	0	5	0	0	50	20	50	0	0	2	0.5	40	12.5
4	0	30	0	0	30	0	60	0	0	1	0.0	40	10.0
5	0	40	20	0	20	0	95	0	0	3	1.3	30	13.8
6	0	60	20	0	5	0	100	0	0	4	2.8	0	11.5
7	0	90	0	0	10	0	100	0	0	3	2.5	0	9.0
8	0	100	0	0	0	20	80	0	0	3	2.5	0	7.5
9	0	0	0	0	10	40	60	0	0	2	1.3	50	8.8
10	2	15	0	0	20	25	10	0	0	0	0.0	50	9.0
11	0	20	0	0	10	5	50	0	0	1	0.5	40	10.0
12	0	0	0	0	50	5	60	0	0	1	0.5	50	11.0
13	0	10	0	0	30	30	60	0	0	2	3.8	20	7.0
14	0	40	0	0	20	50	10	0	0	1	2.3	50	7.0
15	0	10	0	0	40	80	20	0	0	1	3.0	0	7.5
16	0	60	0	0	0	100	0	0	0	0	3.0	0	8.8
17	0	45	0	0	0	5	60	0	0	0	0.3	45	12.5
18	0	100	0	0	0	100	0	0	0	0	2.0	45	10.3
19	1	20	0	0	0	20	0	0	0	0	0.0	100	12.5
20	0	0	0	0	60	0	50	0	0	0	0.3	50	15.0
21	0	0	0	0	80	0	75	0	0	0	0.5	50	14.8
22	0	0	0	0	50	0	50	0	0	0	0.0	50	16.3
23	0	30	10	0	60	0	100	0	0	2	2.5	20	14.8
24	0	0	0	0	50	0	50	0	0	0	0.0	50	16.5
25	0	50	20	0	30	0	100	0	0	3	2.5	0	15.0
26	0	5	15	0	80	0	100	0	0	1	2.5	0	12.5
27	0	60	40	0	0	10	90	0	0	2	4.0	0	10.0
28	0	60	40	0	0	50	50	0	0	2	7.8	0	7.5
29	0	94	5	0	0	90	10	0	0	2	10.0	0	6.3
30	0	80	0	0	20	0	100	0	0	1	3.0	0	11.0
31	0	50	50	0	0	25	75	0	0	3	5.2	0	9.8
32	0	10	40	50	0	75	25	0	0	3	7.5	0	6.5
33	0	12	12	75	0	10	90	0	0	2	5.3	0	4.0
34	0	50	50	0	0	15	85	0	0	2	5.0	0	11.3

	Number	Soil percentages				Vegetation percentages				Other variables			
Plot	of sites	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
35	1	50	40	10	0	80	20	0	0	3	7.3	0	9.8
36	0	0	0	100	0	100	0	0	0	0	7.0	0	6.3
37	0	0	0	100	0	100	0	0	0	0	3.8	0	4.8
38	0	70	30	0	0	50	50	0	0	2	4.5	0	11.5
39	0	40	40	20	0	50	50	0	0	2	8.8	0	10.0
40	0	0	0	100	0	100	0	0	0	0	6.3	0	7.5
41	1	25	25	50	0	100	0	0	0	1	3.8	0	5.2
42	0	40	40	0	20	80	20	0	0	3	2.0	0	4.0
43	0	90	0	0	10	100	0	0	0	1	5.0	0	3.8
44	0	100	0	0	0	100	0	0	0	0	3.8	0	5.0
45	0	100	0	0	0	90	10	0	0	0	2.5	25	7.6
46	1	10	0	0	90	100	0	0	0	2	3.5	0	2.5
47	1	80	0	0	20	100	0	0	0	1	2.8	5	0.0
48	0	60	0	0	30	80	0	0	0	1	1.3	50	3.0
49	0	40	0	0	0	0	30	0	0	0	0.0	100	5.3
50	2	50	0	0	50	100	0	0	0	1	2.0	50	2.0
51	2	50	0	0	0	40	0	0	0	0	0.0	100	1.3
52	1	30	30	0	20	30	60	0	0	2	1.3	50	4.0
53	0	20	20	0	40	0	100	0	0	2	1.0	50	17.6
54	0	20	80	0	0	0	100	0	0	1	3.0	0	15.2
55	0	0	10	0	60	0	75	0	0	1	0.3	50	21.3
56	0	0	50	0	30	0	75	0	0	2	2.8	20	18.8
57	0	50	50	0	0	30	70	0	0	2	5.5	80	16.3
58	0	0	0	0	60	0	60	0	0	0	0.0	50	24.0
59	0	20	20	0	60	0	100	0	0	2	2.5	20	21.5
60	1	90	10	0	0	70	30	0	0	1	5.0	0	20.0
61	0	100	0	0	0	100	0	0	0	0	6.3	0	17.6
62	0	15	15	0	30	0	40	0	0	2	1.0	50	25.2
63	1	100	0	0	0	25	75	0	0	0	3.0	0	23.8
64	1	95	0	0	5	90	10	0	0	0	5.5	0	21.4
65	0	95	0	0	5	90	10	0	0	0	8.0	0	20.0
66	1	60	40	0	0	50	50	0	0	1	6.0	0	12.6
67	0	30	60	10	10	50	40	0	0	3	8.5	0	11.0
68	1	50	0	50	50	100	0	0	0	1	3.0	0	9.0
69	1	60	30	0	10	60	40	0	0	1	1.3	25	7.5
70	1	90	8	0	2	80	20	0	0	1	7.5	0	14.8

	Number	Soil percentages				Vegetation percentages				Other variables			
Plot	of sites	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
71	1	30	30	30	40	60	40	0	0	4	4.8	0	11.5
72	1	33	33	33	33	75	25	0	0	3	1.8	40	11.0
73	0	20	10	0	40	0	100	0	0	2	0.0	100	9.8
74	0	50	0	0	50	40	60	0	0	1	5.3	0	16.0
75	0	75	12	0	12	50	50	0	0	2	2.5	0	14.8
76	0	75	0	0	25	40	60	0	0	1	0.5	100	13.0
77	0	30	0	0	50	0	100	0	0	2	0.0	100	11.5
78	0	50	10	0	30	5	95	0	0	3	5.0	0	17.5
79	0	100	0	0	0	60	40	0	0	1	2.5	0	17.3
80	0	50	0	0	50	20	80	0	0	2	0.0	100	15.0
81	0	10	0	0	90	0	100	0	0	1	0.3	100	14.9
82	0	30	30	0	20	0	85	0	0	3	0.8	80	6.3
83	0	20	20	0	20	0	75	0	0	3	0.0	100	6.3
84	1	90	0	0	0	50	25	0	0	0	0.5	100	7.5
85	0	30	0	0	0	30	5	0	0	0	0.0	100	8.7
86	2	20	30	0	50	20	80	0	0	4	1.0	100	8.8
87	0	50	30	0	10	50	50	0	0	1	0.0	100	8.8
88	0	80	0	0	0	70	10	0	0	0	1.8	100	8.9
89	1	80	0	0	0	50	0	0	0	0	0.8	100	10.0
90	0	60	10	0	25	80	15	0	0	3	1.3	50	11.3
91	0	50	0	0	0	75	0	0	0	0	0.0	100	11.3
92	0	70	0	0	0	75	0	0	0	0	0.0	100	11.5
93	0	100	0	0	0	85	15	0	0	0	2.5	0	13.3
94	0	60	30	0	0	40	60	0	0	3	2.5	25	13.3
95	0	80	20	0	0	50	50	0	0	1	0.0	100	13.8
96	0	100	0	0	0	100	0	0	0	0	2.5	40	14.5
97	0	100	0	0	0	95	5	0	0	0	5.0	0	15.0
98	0	0	0	0	60	0	50	0	0	2	0.3	45	34.0
99	0	30	20	0	30	0	60	0	40	3	1.3	45	32.5
100	0	15	0	0	35	20	30	0	0	0	0.0	50	40.0
101	1	40	0	0	45	70	20	0	0	2	1.3	50	37.8
102	0	30	0	0	45	20	40	0	20	3	0.0	100	35.2
103	0	60	10	0	30	10	65	5	20	3	1.3	20	33.8
104	0	40	20	0	40	0	25	0	75	3	1.0	60	27.0
105	1	100	0	0	0	70	0	0	30	0	3.0	0	25.0
106	1	100	0	0	0	40	60	0	0	2	6.0	0	23.5

	Number	Soil percentages				Vegetation percentages				Other variables			
Plot	of sites	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
107	0	80	10	0	10	40	60	0	0	2	8.0	0	21.4
108	1	90	0	0	10	10	0	0	90	0	1.3	75	28.8
109	1	100	0	0	0	20	10	0	70	0	3.0	0	26.5
110	0	30	50	0	20	10	90	0	0	2	6.0	0	25.0
111	0	60	40	0	0	50	50	0	0	1	5.3	0	23.3
112	0	100	0	0	0	80	10	0	10	0	2.5	0	33.0
113	1	60	0	0	40	60	10	30	0	1	4.8	0	28.4
114	0	50	50	0	0	0	100	0	0	2	7.0	0	27.0
115	0	60	30	0	10	25	75	0	0	3	4.5	0	25.5
116	0	40	0	0	60	30	20	50	0	1	5.0	0	31.5
117	0	30	0	0	70	0	50	50	0	2	7.5	0	30.3
118	0	50	20	0	30	0	100	0	0	3	6.0	0	29.0
119	0	50	50	0	0	25	75	0	0	1	6.5	0	27.5
120	0	90	10	0	0	50	50	0	0	1	5.5	0	20.2
121	0	100	0	0	0	60	40	0	0	0	3.0	0	18.5
122	0	50	0	0	50	70	30	0	0	1	0.0	100	17.5
123	0	10	10	0	80	0	100	0	0	2	0.3	100	17.4
124	0	50	50	0	0	30	70	0	0	2	3.8	0	22.0
125	1	75	0	0	25	80	20	0	0	1	1.3	90	20.5
126	0	40	0	0	60	0	100	0	0	2	0.3	90	20.0
127	0	90	10	0	10	75	25	0	0	2	3.5	20	19.0
128	0	45	45	0	55	30	70	0	0	2	2.3	30	23.8
129	0	20	35	0	80	10	90	0	0	2	0.3	100	22.8
130	0	80	0	0	20	70	30	0	0	2	2.8	10	22.3
131	0	100	0	0	0	90	0	0	0	0	5.0	0	21.3
132	0	75	0	0	25	50	50	0	0	2	1.0	60	26.3
133	0	60	5	0	40	50	50	0	0	2	0.3	100	25.0
134	0	40	0	0	60	60	40	0	0	1	2.8	0	24.0
135	0	60	0	0	40	70	15	0	0	1	5.0	0	23.8
136	0	90	10	0	10	75	25	0	0	1	2.0	30	16.3
137	0	50	0	5	0	30	20	0	0	0	0.0	100	16.3
138	0	70	0	30	0	70	30	0	0	1	2.0	20	17.0
139	0	60	0	40	0	100	0	0	0	1	4.8	0	17.5
140	2	50	0	0	0	50	0	0	0	0	0.0	100	19.0
141	0	30	0	50	0	60	40	0	0	1	1.3	60	19.0
142	0	5	0	95	0	80	20	0	0	1	3.8	0	19.0

	Number	Soil percentages				Vegetation percentages				Other variables			
Plot	of sites	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
143	0	10	0	90	0	70	30	0	0	1	6.3	0	19.5
144	0	50	0	0	0	15	30	0	0	0	0.0	100	21.3
145	0	20	0	80	0	50	50	0	0	1	2.8	0	21.3
146	0	0	0	100	0	90	10	0	0	0	5.3	0	22.0
147	0	0	0	100	0	75	25	0	0	0	7.5	0	22.0
148	0	90	0	10	0	60	30	10	0	1	1.3	20	23.8
149	0	0	0	100	0	80	10	10	0	0	3.8	0	23.8
150	0	0	0	100	0	60	40	0	0	0	6.3	0	23.8
151	0	0	40	60	40	50	50	0	0	1	8.3	0	23.9

*Variables are: X_1 = percentage of soils with constant lime enrichment; X_2 = percentage meadow soil with calcium groundwater; X_3 = percentage soils formed from coral bedrock under conditions of constant lime enrichment; X_4 = percentage alluvial and organic soils adjacent to rivers and saline organic soil at the coast; X_5 = percentage deciduous seasonal broadleaf forest; X_6 = percentage high and low marsh forest, herbaceous marsh and swamp; X_7 = percentage cohune palm forest; X_8 = percentage mixed forest composed of types listed for X_5 and X_7; X_9 = number of soil boundaries in square; X_{10} = distance to navigable water (km); X_{11} = percentage of square within 1 km of navigable water; X_{12} = distance from the site of Santa Rita (km).

Table 5.7 Distributions of the lengths of Daphnia publicaria in plankton (P) and in the stomachs (S) of yellow perch fry in five samples taken on different days in 1969 from West Blue Lake, Manitoba. This table was constructed from Figure 1 of Wong and Ward (1972).

	1 July		15 July		29 July		12 August		25 August	
Length (mm)	P	S	P	S	P	S	P	S	P	S
0.5 -	20	59	28	20	2	0	1	0	6	27
0.7	22	84	49	40	11	12	2	0	2	42
0.9	20	154	59	101	21	61	7	34	2	124
1.1	18	138	62	126	33	95	9	127	0	138
1.3	26	44	46	146	59	172	17	230	3	261
1.5	24	10	33	60	31	233	28	241	12	303
1.7	22	5	28	2	24	168	14	218	35	604
1.9	24	0	33	5	22	78	12	218	63	606
2.1	26	0	13	2	16	21	4	92	36	289
2.3	16	0	13	2	11	9	6	34	15	193
2.5	11	0	7	0	7	1	4	11	5	55
2.7	7	0	7	0	2	0	1	5	0	58
2.9	1	0	2	0	1	0	0	6	0	0

CHAPTER 6

RESOURCE SELECTION OVER SEVERAL TIME PERIODS

When resource units are censussed or sampled after several periods of selection time then it may be necessary for an analysis to take into account the increasing proportion of used units as time increases. With census data the situation is straightforward because a resource selection study is like many other studies on the relative survival rate of different types of individual. Here the use of the proportional hazards model in particular is suggested, and its use is illustrated using data on the selection of snails by birds. With sample data the analysis becomes non-standard and complicated, unless the population of resource units being considered is large and either (a) only a small proportion of resource units are used, or (b) only samples of unused resource units are available. In these special cases the resource selection function (RSF) or the resource selection probability function (RSPF) can be estimated using generalizations of the logistic regression methods that have been proposed for sample data in Chapter 5. These methods are illustrated using data from a studies of habitat selection by the northern spotted owl and the predation of corixids by minnows.

6.1 Census Data

In Section 5.1 logistic regression was suggested as a method for estimating a RSPF for census data when there is no need to take into account variable amounts of selection time. Here an alternative model is proposed for data involving two or more periods of selection, possibly of different durations.

The situation envisaged is that a population of resource units is subjected to S selection episodes following one after another, with a census of resource units at the end of each episode. The structure of the selection process is then as indicated in Figure 6.1. Thus the available resource units at time 0 are divided at time t_1 into a group of unused units and a group of used units. Then, during the second selection episode, from time t_1 to time t_2, some of the unused units at time t_1 become used. The same process continues until time t_S and, in general, as the selection time increases there are fewer and fewer unused units, and more and more used units. Note that this description of a selection process holds for resource units that can be used more than once if it is the first use that changes the status of a unit.

As in the previous chapter, it is assumed that each resource unit is characterized by the values that it possesses for p variables $\mathbf{X} = (X_1, X_2, ..., X_p)$, and that the resource units initially available can be divided into I classes so that all the A_i units within the ith class have the X values $\mathbf{x}_i = (x_{i1},x_{i2},...,x_{ip})$. The question at issue then becomes how the probability of use by time t for units in the ith class depends on $\mathbf{x}_i$.

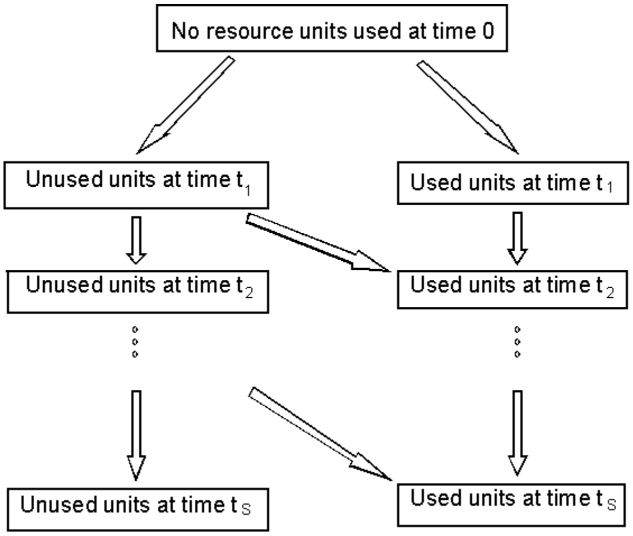

No resource units used at time 0
Unused units at time t_1
Used units at time t_1
Unused units at time t_2
Used units at time t_2
Unused units at time t_S
Used units at time t_S

Figure 6.1 The structure of the resource selection process when resource units can only be used once and censuses of units are made after selection times of t_1, t_2, ..., t_S.

From censuses of resource units taken at times t_1 to t_S, and a knowledge of what resource units were available at time 0, it will be possible to determine U_{ij}, the number of type i units used between times t_{j-1} and time t_j, and $\bar{U}_{ij}$, the number of type i units that are unused at time t_S. Hence the data resulting from the study can be set out in the form shown in Table 6.1.

Clearly, the selection of resource units is a type of survival process so that it is appropriate to model the RSPF by

$$w^*(\mathbf{x},t) = 1 - \phi^*(\mathbf{x},t), \tag{6.1}$$

where $\phi^*(\mathbf{x},t)$ is a standard survival function, giving the probability that a unit with $\mathbf{X} = \mathbf{x}$ survives until time t. There are many possible choices for $\phi^*(\mathbf{x},t)$, of which one is the proportional hazards model (Section 2.5). This takes the form

$$\phi^*(\mathbf{x},t) = \exp\{-\exp(\beta_0 + \beta_1 x_1 + ... + \beta_p x_p)t\},$$

so that

$$w^*(\mathbf{x},t) = 1 - \exp\{-\exp(\beta_0 + \beta_1 x_1 + ... + \beta_p x_p)t\}, \qquad (6.2)$$

which is realistic in the present context because it implies that $w^*(\mathbf{x},0) = 0$ (no resource units are used when $t=0$) and $w^*(\mathbf{x},\infty) = 1$ (all resource units are eventually used).

Table 6.1 Form of the data arising from taking censuses of I types of resource units before selection, and after S periods of selection.

Type of unit	X values				Used during the time period				Unused at time			
	X_1	X_2	...	X_p	$0\text{-}t_1$	$t_1\text{-}t_2$	...	$t_{S-1}\text{-}t_S$	t_0	t_1	...	t_S
1	x_{11}	x_{12}	...	x_{1p}	U_{11}	U_{12}	...	U_{1S}	A_1	$\bar{U}_{11}$	...	$\bar{U}_{1S}$
2	x_{21}	x_{22}	...	x_{2p}	U_{21}	U_{22}	...	U_{2S}	A_2	$\bar{U}_{21}$	...	$\bar{U}_{2S}$
.												
.												
.												
I	x_{I1}	x_{I2}	...	x_{Ip}	U_{I1}	U_{I2}	...	U_{IS}	A_I	$\bar{U}_{I1}$	...	$\bar{U}_{IS}$

If each resource unit is used or not used independently of the other units, then the distribution of the numbers of type i resource units used in the S intervals $(0, t_1)$, (t_1, t_2), ..., (t_{S-1}, t_S), and the number of this type of unit that are unused at time t_S, will follow a multinomial distribution such that the probability of use in the time interval (t_{j-1},t_j) is

$$\Theta_{ij} = \phi(\mathbf{x}_i,t_{j-1}) - \phi(\mathbf{x}_i,t_j)$$

$$= \exp\{-\exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip})t_{j-1}\} - \exp\{-\exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip})t_j\}, \qquad (6.3)$$

and the probability of not being used by time t_S is

$$\Theta_{i\,S+1} = \exp\{-\exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip})t_S\}. \qquad (6.4)$$

This model can be fitted to data using the principle of maximum likelihood. One way to do this is to convert it to a generalized linear model using an approach that is justified by Manly (1985, p. 419). Then any computer program for fitting these models can be used to estimate the parameters of the model.

What is done is to recognize that U_{i1}, the number of type i units used in the selection period 0 to t_1 is a binomial random variable with A_i individuals at risk, and a probability

$$p_{i1} = 1 - \phi^*(\mathbf{x}_i,t_1) = 1 - \exp\{-\exp(\beta_0 + \beta_1 x_1 + ... + \beta_p x_p)t_1\} \qquad (6.5)$$

of being used. If any survive this first period then U_{i2}, the number used in the second period is a binomial random variable with $\bar{U}_{i1}$ at risk, and a probability

$$p_{i2} = 1 - \phi^*(\mathbf{x}_i,t_2)/\phi^*(\mathbf{x}_i,t_1) = 1 - \exp\{-\exp(\beta_0 + \beta_1 x_1 + ... + \beta_p x_p)(t_2 - t_1)\}, \qquad (6.6)$$

of being used. Continuing in this way, if there are any survivors to time t_{j-1} then U_{ij}, the number used in the period from t_{j-1} to t_j, is a binomial random variable with $\bar{U}_{ij-1}$ at risk, and a probability

$$p_{ij} = 1 - \phi^*(\mathbf{x}_i,t_2)/\phi^*(\mathbf{x}_i,t_1) = 1 - \exp\{-\exp(\beta_0 + \beta_1x_1 + ... + \beta_px_p)(t_j - t_{j-1})\}, \quad (6.7)$$

of being used. As shown by Manly (1985, p. 419), it is perfectly valid to analyse the data with the numbers used in the different periods of time being treated as having independent binomial distributions. The model parameters can then be estimated with what is called the complementary log-log link for the generalized linear model (McCullagh and Nelder, 1989, p. 31).

The general form of equations (6.5) to (6.7) is

$$p_{ij} = 1 - \exp\{-\exp(\beta_0 + \beta_1x_1 + ... + \beta_px_p)D_i\}, \quad (6.8)$$

where p_{ij} is the probability that an item that is unused at time t_{i-1} is used in the time interval t_{i-1} to ti, and $D_i = t_i - t_{i-1}$ is the duration of this interval, with $t_0 = 0$. This can also be written as

$$p_{ij} = 1 - \exp\{-\exp(\beta_0 + \beta_1x_1 + ... + \beta_px_p + \gamma_i\}, \quad (6.9)$$

where $\gamma_i = \log_e(D_i)$. The reason for doing this is that the parameters $\gamma_1, \gamma_2, ..., \gamma_S$ can either be treated as being known, and allowed for in the fitting process using what is sometimes called an offset, or γ_1 alone can be treated as known, and γ_2 to γ_S estimated, to allow the effective selection time to differ from the chronological time.

As discussed in Section 2.7, the deviance will follow a distribution that is approximately chi-squared if the model being fitted is correct, with the degrees of freedom (df) being the number of data frequencies minus the number of estimated parameters. This deviance should be output by the computer programme used to fit the model to data. It can be used for a goodness of fit test providing that the expected frequencies are 'large', and differences between deviances can be used to compare the fit of different models even if this is not the case.

If the modelling assumptions are correct then the counts $U_{i1}, U_{i2}, ..., U_{iS}$ of the number of units used in the different selection periods will follow binomial distributions, with the mean and variance of U_{ij} being $\bar{U}_{ij-1}p_{ij}$ and $\bar{U}_{ij-1}p_{ij}(1 - p_{ij})$, respectively. The assumptions can therefore be checked by residual plots, as discussed in Section 2.9.

Example 6.1 Selection of Snails by Birds

As an example, consider the experiment of Bantock *et al.* (1976) on the selection of *Cepaea nemoralis* and *C. hortensis* snails by the song thrush *Turdus ericetorum*, that was described in Example 3.3. Recall that an experimental population of 498 yellow five-banded (Y5H) *C. hortensis*, 499 yellow five-banded (Y5N) *C. nemoralis* and 877 yellow mid-banded (Y3N) *C. nemoralis* snails was set up on 29 June, 1973, with the shells being uniquely marked so that the survivors could be determined from censuses taken at various times after the population was set up. Extra Y3H were added to the population on 5 July and on 7 July. This complicates any analysis that includes this morph and therefore for the purposes of this example this morph will be ignored and only the results for used and unused Y5N and Y5H will be considered.

Simplified results from the experiment are shown in Table 3.3, with different types of snail defined in terms of two variables, X_1, a species indicator which is 1 for *C. nemoralis* and 0 for *C. hortensis*, and X_2 is the maximum shell diameter in units of 0.3 mm over 14.3 mm.

Various versions of the proportional hazards model defined above were fitted to the data. First, the no selection model, with only the $ß_0$ term in equation (6.9) was fitted using the known survival time durations of 6, 6 and 10 days to fix the values of γ_1, γ_2 and γ_3. This model, which will be referred to as model 0A, gave a deviance of D_{0A} = 207.7 with 121 df. The no selection model with the effective durations of the last two selection times determined by setting $\gamma_1 = 0$ and estimating γ_2 and γ_3 was also fitted. This model, which will be called model 0B, gave a deviance of D_{0B} = 196.3 with 123 df. The difference in deviance is 11.4 with 2 df. As this is significantly large at the 1% level, it seems that model 0B is more satisfactory than model 0A.

Next, models allowing for selection related to species and size was fitted by including these variables in the model. As for the no selection case, two models were considered. For model 1A the duration variables γ_1, γ_2 and γ_3 were fixed based in chronological time. This then resulted in a deviance of D_{1A} = 145.0 with 119 df. For model 1B, γ_1 was set at 0, and γ_2 and γ_3 were estimated, with the result that the deviance dropped to D_{1B} = 133.5, with 117 df. The difference in deviance between model 1A and model 1B is 12.5 with 2 df, which is significantly large at the 1% level. Hence the model with estimated sample times is a distinctly better fit than the model using chronological times. In addition, both of the models that allow for selection have much lower deviances than the no selection models, providing very strong evidence that selection occurred.

Because many of the expected sample frequencies are very small, it is questionable whether it is valid to compare D_{1B} = 133.5 with 117 df with the chi-squared distribution as a test for the absolute goodness of fit. However, if this is done then the statistic is found to be not at all significantly large.

Models involving effects that are quadratic in size were also fitted by including SIZE^2 in the model. As the addition of this extra variable led to very little reduction in deviance, this modification will not be considered further here. Models that allowed the coefficient of the size variable to vary with the species were also fitted, but again this led to little reduction in deviance.

It seems from this analysis that model 1B is realistic for the data. The estimated parameters found for this model are shown in Table 6.2, and it can be seen by considering the ratios of ß estimates to their standard errors that there is clear evidence of a species effect but little evidence of a size effect.

The estimate of the duration parameter for the second selection period is $\hat{\gamma}_2$ = 0.300. This estimates the natural logarithm of the effective duration relative to the duration of the first selection period, so that the estimate of the effective duration itself is D_2 = 6exp(0.300) = 8.1 days. In a similar way, the effective duration of the third selection period is estimated to be $D_3 = 6\exp(\hat{\gamma}_3)$ = 10.3 days. The effective duration of the second selection period is therefore estimated to be somewhat more than the chronological time of 6 days, while the effective duration of the third selection period is estimated to be close to the actual 10 days.

The RSPF for model 1B is

$$w^*(\mathbf{x},t^*) = 1 - \exp[-\exp\{-2.233 + 0.0238(\text{SIZE}) - 0.487(\text{SPECIES})\}t^*], \quad (6.10)$$

where SIZE denotes the coded size of the snails, SPECIES denotes the dummy variable for the species (0 for *C. hortensis*, 1 for *C. nemoralis*), and t* denotes the effective selection time. Because the second and third selection times were estimated, probabilities can only be determined from this function for the census times of six days (using t* = 6), 12 days (using t* = 14.1) and 22 days (using t* = 24.4).

Table 6.2 Estimated parameters for the proportional hazards resource selection function fitted to data on the selection of snails by birds.

Parameter	Estimate	Standard error	Ratio	p-value[1]
Constant, β_0	-2.233	0.184	-12.14	
Species effect, β_1	0.487	0.133	3.66	0.000
Size effect, β_2	0.024	0.013	1.77	0.076
Period 2 duration, γ_2	0.300	0.103	2.90	0.004
Period 3 duration, γ_3	0.545	0.106	5.13	0.000

[1] Determined by comparing the ratio of the estimate to the standard error with the standard normal distribution.

Figure 6.2 shows the level of selection that is suggested by model 1B. There is apparently a rather higher probability of use for *C. nemoralis* than for *C. hortensis*, and for large snails rather than small ones. However, it must be kept in mind that the size effect is not significant at the 5% level. Figure 6.3 shows standardized deviance residuals (McCullagh and Nelder, 1989, p. 396). A separate graph is provided for the residuals based on the numbers used up to day 6, the numbers used between day 6 and day 12, and the number used between day 12 and day 22. The species involved with a residual is indicated by 'H' or 'N'. All the plots indicate the type of distribution expected from standard normal variables so that it seems that the model gives an adequate description of the data.

6.2 Sample Data

The situation that will now be considered is like that considered in the previous section, where a population of N available resource units exists at time 0, and these are then gradually used as time goes on. However, it will now be assumed that instead of the numbers of used and unused units being counted at times t_1, t_2, ..., $t_{S,}$ only samples of these units are available for the estimation of a RSPF. If there is only one sample time then the logistic regression methods of Chapter 5 can be used. Hence only multiple sample time situations are considered here.

One such situation was discussed in Example 3.5, which was concerned with Popham's (1944) study of the use of corixids as food by minnows (*Phoxinus phoxinus*), and selection related to the species and colour of the corixids. What Popham did was to sample the corixid in a pond every day for seven days, introduce 50 minnows to the pond on the evening of the seventh day, and then sample again every day from the third to ninth days after this change to the pond. The samples taken before the introduction of minnows can be lumped together to form a single 'available sample', because they have similar proportions for the species and colours of corixids, while the samples taken after the introduction of minnows are samples of unused prey. Therefore this can be thought of as a situation where there are eight samples of unused resource units (corixids), taken after selection times of 0, 3, 4, ..., 9 days.

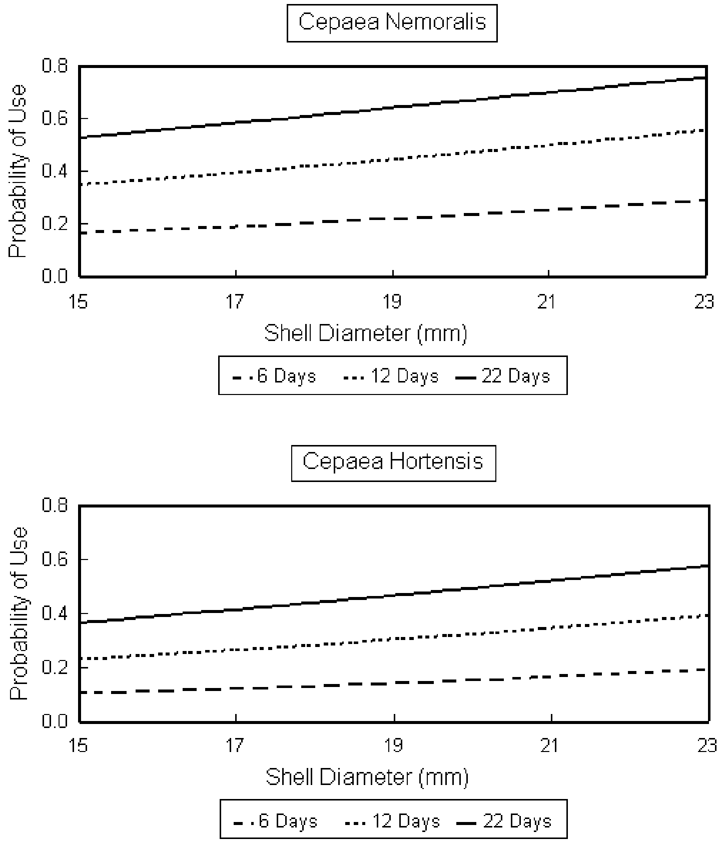


Figure 6.2 Estimated probabilities of use after 6, 12 and 22 days for Cepaea nemoralis and C. hortensis.

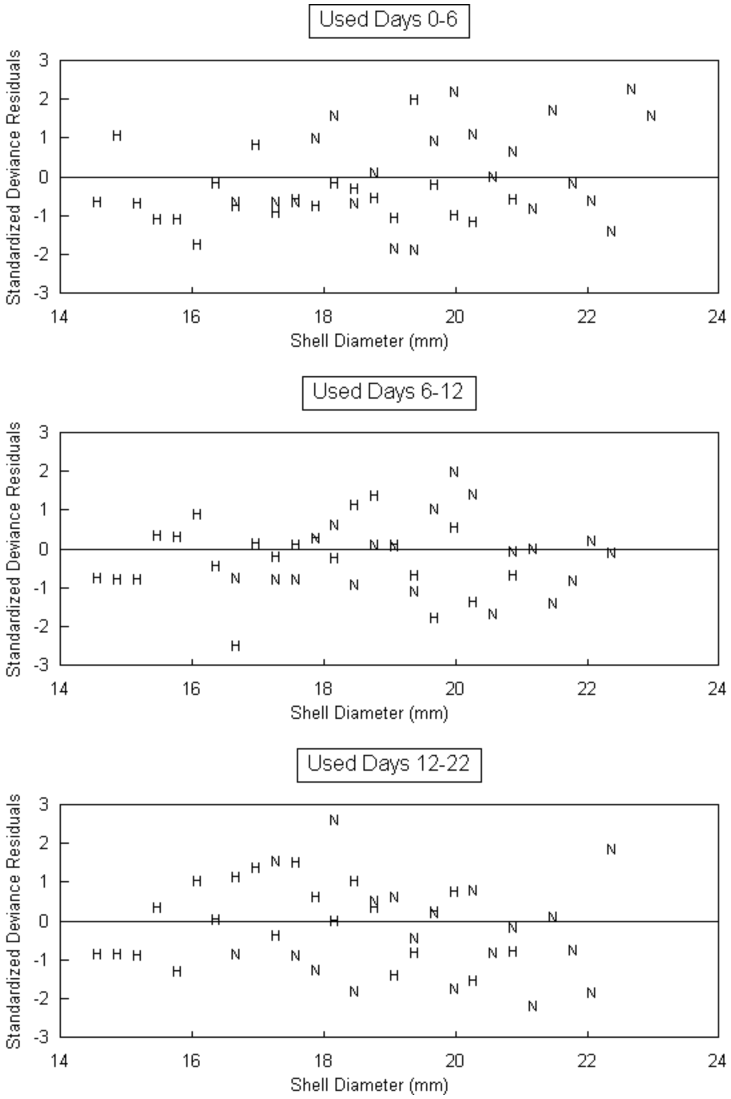


Figure 6.3 Standardized deviance residuals plotted against the maximum shell diameters of snails, with a separate plot for the numbers used for each of the time intervals 0-6, 6-12 and 12-22 days (H is a residual for *Cepaea hortensis*, N for *C. nemoralis*).

To understand the nature of sample data, as before it will be assumed that all resource units fall within one of I classes such that all the units within class i have the vector of values $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ for the variables X_1 to X_p. There are then three types of samples that need to be considered: (a) available resource units before selection begins, (b) used units at time t_j, and (c) unused units at time t_j. To find expected sample frequencies for the first type of sample, suppose that at time 0, before selection begins, there are A_i available units in the ith class, of which a_i are seen in a random sample. Then the expected value of a_i is

$$E(a_i) = P_a A_i, \tag{6.11}$$

where P_a is the sampling fraction.

As the selection time increases, more and more of the unused units in the ith class will become used units, until eventually no more unused units are left. To allow for this, suppose that a random sample is taken from all the units used up to time t. Then the expected value of the number of type i individuals in the sample, is

$$E_{ui}(t) = P_u(t)\, A_i\, w^*(\mathbf{x}_i, t), \tag{6.12}$$

where $P_u(t)$ is the sampling fraction and $w^*(\mathbf{x}_i, t)$ is the resource selection probability function, which gives the probability that a type i unit will be used by time t. Equation (6.12) still applies in situations where units can be used more than once if 'use' is defined to be use at least once. Also, because no units are used at time t = 0, it follows that $w^*(\mathbf{x}_i, 0) = 0$, and hence that $E_{ui}(0) = 0$.

An equation for the expected number of type i individuals in a random sample of unused units taken at time t is

$$E_{\bar{u}i}(t) = P_{\bar{u}}(t)\, A_i\, \{1 - w^*(\mathbf{x}_i, t)\}, \tag{6.13}$$

where $P_{\bar{u}}(t)$ is the sampling fraction. As the unused units at time 0 are the available units, equation (6.13) with t = 0 is the same as equation (6.11). Formally this identification is achieved by recognizing that $P_a = P_{\bar{u}}(0)$ and $w^*(\mathbf{x}_i, 0) = 0$.

At this point the function $w^*(\mathbf{x}_i, t)$ in equations (6.12) and (6.13) can take any form as long as the condition $0 \le w^*(\mathbf{x}_i, t) \le 1$ is satisfied. However, in order to estimate the function it is usual to assume a specific parametric form. One possibility is the proportional hazards survival function, with $w^*(\mathbf{x}, t)$ given by equations (6.1) and (6.2), in which case the expected values for used and unused units become

$$E_{ui}(t) = P_u(t)\, A_i\, [1 - \exp\{-\exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)t\}], \tag{6.14}$$

and

$$E_{\bar{u}i}(t)\} = P_{\bar{u}}(t)\, A_i \exp\{-\exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)t\}, \tag{6.15}$$

with equation (6.15) reducing to equation (6.11) when t = 0.

Between them, these equations describe the expected values of sample frequencies in terms of the parameters A_1 to A_I, P_a, $P_u(t)$, $P_{\bar{u}}(t)$, and β_0 to β_p. In principle it is therefore possible to estimate some or all of these parameters if observed sample frequencies are available for two or more samples either of different types of units or of the same type of units taken at different times. However, these equations for expected sample frequencies are not by themselves sufficient to estimate the parameters in RSPFs by the method of maximum likelihood. It is also necessary to make assumptions about the distributions of sample frequencies. The simplest possibility here involves assuming that sample counts have independent Poisson distributions, which is reasonable if samples are taken independently, are random, and sample sizes are much smaller than the populations of resource units being sampled.

Given the Poisson assumption, estimates of unknown parameters can be obtained using a general computer program for maximum likelihood estimation, because the model is not standard. The usual procedure is to maximize the log-likelihood function, which takes the form

$$\sum \{-E_k + O_k\log_e(E_k) - \log_e(O_k!)\},$$

where O_k is an observed sample frequency, which has the expected value E_k for the model being fitted. The E_k values are then functions of the parameters to be estimated, given by equations (6.14) and (6.15), and the summation is for all of the observed data frequencies.

It is important to remember that equation (6.14) applies to a sample from all of the units used by time t. If a sample is taken just from the units used from time t_{j-1} to time t_j then the equation must be modified accordingly.

6.3 Small Proportions of Used Resource Units Sampled

Although the estimation of a RSPF is complicated for general situations where samples of used and unused units are available, there are two special cases that are relatively straightforward. The first of these is where only a small proportion of available resource units are used during the study period. What happens under these circumstances is that because relatively few units are selected for use, the exponential term $\exp(\beta_0 + \beta_1x_1 + ... + \beta_px_p)t$ must be small in equation (6.2) for the resource selection probability function based on the proportional hazards survival function. As a result, the approximation

$$\begin{aligned} w^*(\mathbf{x},t) &= 1 - \exp\{-\exp(\beta_0 + \beta_1x_1 + ... + \beta_px_p)t\} \\ &\approx \exp(\beta_0 + \beta_1x_1 + ... + \beta_px_p)t \qquad (6.16) \end{aligned}$$

applies. In addition, the probability of a unit with $\mathbf{X} = \mathbf{x}$ being used in the time interval from t_{j-1} to t_j becomes

$$w_j^*(\mathbf{x}) \approx \exp(\beta_0 + \beta_1x_1 + ... + \beta_px_p)\, D_j,$$

where $D_j = t_j - t_{j-1}$.

It is the last equation that is the important one for the development of a method of estimation. However, before using this it will be generalized by letting the β parameters vary with j, which means that the RSPF may vary with time. The RSPF for the period t_{j-1} to t_j then becomes

$$w_j^*(\mathbf{x}) \approx \exp(\beta_{0j} + \beta_{1j}x_1 + ... + \beta_{pj}x_p)\, D_j. \qquad (6.17)$$

Using this equation it is possible to generalize the logistic regression method that was described in Section 5.4 to the situation where there is one sample of available units and two or more samples of used units collected at different times.

Assume that the sample of available units is collected in such a way that each of these units has the same probability P_a of being included in the sample, independent of the other units in the sample. Also assume that the samples of used units are collected at successive times t_1, t_2, ..., t_S, where the jth of these samples is taken from the units used in the interval from time t_{j-1} to time t_j, in such a way that each unit used in this interval has the same probability P_j of being included in the sample, independently of the other units in the sample. For the first interval, t_0 is defined to be the time when selection starts, possible, but not necessarily time, zero. Further assume that the population of available units is very large, with the ith of these units described by values $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{ip})$.

With this formulation, the probability that the ith unit in the population of available units is included in the available sample is P_a, and the probability that it is included in the jth used sample is $P_j w_j^*(\mathbf{x}_i)$. Therefore, the probability of this unit appearing in any of the samples is

$$\text{P(unit sampled)} = 1 - \{1 - P_a\}\{1 - P_1\, w_1^*(\mathbf{x}_i)\}\{1 - P_2\, w_2^*(\mathbf{x}_i)\} \ldots \{1 - P_S\, w_S^*(\mathbf{x}_i)\}.$$

We now use the assumption that all of the samples are small in comparison with the total number of units, so that the probability of a unit appearing in more than one sample is negligible. In this case the last equation simplifies to

$$\text{P(unit sampled)} \approx P_a + P_1 w_1^*(\mathbf{x}_i) + P_2 w_2^*(\mathbf{x}_i) + \ldots + P_S w_S^*(\mathbf{x}_i).$$

It then follows that the conditional probability that the ith unit appears in the available sample, given that it appears in one of the samples, is

$$\tau_a(\mathbf{x}_i) = P_a / \{P_a + P_1 w_1^*(\mathbf{x}_i) + P_2 w_2^*(\mathbf{x}_i) + \ldots + P_S w_S^*(\mathbf{x}_i)\},$$

$$= P_a / \{P_a + \sum_{k=1}^{S} P_k \exp(\beta_{0k} + \beta_{1k} x_{i1} + \ldots + \beta_{pk} x_{ip})\, D_k\},$$

while the probability that it appears in the jth used sample, given that it appears in one of the samples is

$$\tau_j(\mathbf{x}_i) = P_j\, w_j^*(\mathbf{x}_i) / \{P_a + P_1 w_1^*(\mathbf{x}_i) + P_2 w_2^*(\mathbf{x}_i) + \ldots + P_S w_S^*(\mathbf{x}_i)\},$$

$$= P_j \exp(\beta_{0j} + \beta_{1j} x_{i1} +\ldots+ \beta_{pj} x_{ip})\, D_j / \{P_a + \sum_{k=1}^{S} P_k \exp(\beta_{0k} + \beta_{1k} x_{i1} +.. + \beta_{pk} x_{ip})\, D_k\}.$$

At this point it needs to be recognized that all of the parameters in these equations cannot be estimated. To account for this, the equation can be rewritten as

$$\tau_a(\mathbf{x}_i) = 1 / \{1 + \sum_{k=1}^{S} \exp(\beta'_{0k} + \beta_{1k} x_{i1} +\ldots+ \beta_{pk} x_{ip})\}, \tag{6.18}$$

and

$$\tau_j(\mathbf{x}_i) = \exp(\beta'_{0j} + \beta_{1j} x_{i1} +\ldots+ \beta_{pj} x_{ip}) / \{1 + \sum_{k=1}^{S} \exp(\beta'_{0k} + \beta_{1k} x_{i1} +\ldots+ \beta_{pk} x_{ip})\}, \tag{6.19}$$

where

$$\beta'_{0j} = \beta_{oj} + \log_e(P_j/P_a) + \log_e(D_j).$$

All of the parameters in the equation can then be estimated, with the modified constant terms allowing for different sampling probabilities, different time duration for selection, and also for the RSPFs to have different constant terms for different time intervals.

Equations (6.18) and (6.19) describe what is sometimes called a polytomous regression model. The likelihood function is derived by defining a response vector $\mathbf{y}_i = (y_{i0}, y_{i1}, ..., y_{iS})$ for the ith sampled unit, where this indicates which sample the unit is in. For an available unit, $y_{i0} = 1$ and $y_{i1} = y_{i2} = ... = y_{iS} = 0$, whereas for a unit in the jth used sample all of the values $y_{i0}, y_{i1}, ..., y_{iS}$ are 0, except y_{ij}, which is 1. For any unit, $\mathbf{y}_i$ is then a random observation from a multinomial distribution, for which the probability of the ith unit appearing in the sample where it does is

$$L_i = \tau_a(\mathbf{x}_i)^{y_{i0}}\, \tau_1(\mathbf{x}_i)^{y_{i1}} \ldots \tau_S(\mathbf{x}_i)^{y_{iS}},$$

and the log-likelihood function for the entire data set is

$$l(\beta'_{01}, \beta_{11}, ..., \beta_{p1}, \beta'_{02}, \beta_{12}, ..., \beta_{p2}, ..., \beta_{pS}) = \sum \log_e(L_i), \qquad (6.20)$$

where the sum is over all sampled units.

Maximum likelihood estimators of the unknown parameters are obtained by maximizing the likelihood function (6.20) with respect to these parameters. Generally, there will be three versions of the model that will be of interest. The first is the no selection model that is obtained by setting all of the parameters except β'_{01}, β'_{02} ..., β'_{0S} equal to zero. The second is the constant selection model that is obtained by setting $\beta_{j1} = \beta_{j2} = ... = \beta_{jS}$, for all j (i.e, making the coefficient of X_j the same for all the used samples). The third is the variable selection model for which all of the used samples have different coefficients for each of the X-variables.

The deviance for a fitted model is minus twice the difference between the log-likelihood for the model and the log-likelihood for the model that fits the data perfectly. In the present case the model that fits perfectly will give a probability of exactly 1 for observing the ith unit in the sample where it occurs and a probability of exactly 0 of observing it in any other sample. Thus $L_i = 1$ and $\log_e(L_i) = 0$ for the perfectly fitting model, for all i. Consequently, the perfectly fitting model gives a total log-likelihood of 0, and the deviance is just minus twice the log-likelihood for this model.

Maximum likelihood estimation may possibly be performed by some standard statistical packages. For example, MINITAB can fit the variable selection model using the nominal regression option (Minitab Inc., 1997). However, MINITAB does not allow the fitting of the model where the coefficients of the X-variables are the same for all used samples. Fitting of this model may therefore require the use of a special computer program, as discussed in Chapter 14.

The important outcome of the fitting process is estimates of the coefficients of the X-variables for the jth used sample. These allow the estimation of the RSF for the sample period that this sample relates to. This function is

$$w_j(\mathbf{x}_i) = \exp(\beta_{1j}x_{i1} + \beta_{2j}x_{i2} ... + \beta_{pj}x_{ip})t, \qquad (6.21)$$

which is proportional to the RSPF $w_j^*(\mathbf{x}_i)$ of equation (6.17).

Example 6.2 Habitat Selection by the Northern Spotted Owl

This example comes from a study of habitat selection by a male and female pair of the northern spotted owl (*Strix occidentalis caurina*) in a forest near Eugene, in Oregon, USA. The sample units are blocks of land in the forest, for which information is available in a geographical information system (GIS) on the elevation (km), the distance from the nest (km), and whether the block is old growth, young growth, or clearcut. A primary interest in the study was to determine the extent to which the owls are selective in the use of the old, young and clearcut blocks.

To represent availability, a sample of 1000 units was selected at random from the GIS. For the purpose of this example this will be treated as being equivalent to a sample selected in such a way that each unit had a small probability P_a of selection. The owls had radio collars attached, and the location of the owls was also determined by radio telemetry at various times during the period from 25 August 1997 to 13 August 1998. These locations are used here to provide three samples of used units, for August to December 1997 (128 observations), January to March 1998 (90 observations), and April to August 1998 (91 observations).

Three models were fitted to the data. The first was the no selection model for which only the constant terms β'_{01}, β'_{02} and β'_{03} were included in equations (6.18) and (6.19). For this model, the deviance is 2100.91, with 1306 df.

The second model allowed the resource selection function to vary with the elevation (X_1), and the distance to the nest (X_2). In addition, it was allowed to vary with the type of habitat by including two indicator variables in the model to reflect this. The first of these indicator variables (X_3) was 1 for a unit with young growth, or otherwise 0, while the second indicator variables (X_4) was 1 for a clearcut unit, or otherwise 0. This then made the 'standard' type of unit one with old growth. For this model the deviance is 1927.47, with 1302 df. The reduction in deviance in moving from model 1 to model 2 is therefore 2100.91 - 1927.47 = 173.44, with 1306 - 1302 = 4 df. This difference is very highly significant when compared with the chi-squared distribution with 4 df, giving very strong evidence that the owls were selective in their choice of habitat.

The third model allowed the coefficients of the X-variables to vary with the used samples. For this model the deviance is 1896.42, with 1294 df. The reduction in deviance in moving from model 2 to model 3 is therefore 1927.47 - 1896.42 = 31.05, with 1302 - 1294 = 8 df. This difference is very highly significant in comparison with the chi-squared distribution with 8 df, giving very clear evidence that the selection of habitat by the owls varied to some extent over the study period.

Table 6.3 shows the estimated selection functions that are obtained from the variable selection model, with the approximate standard errors of the estimated parameters that are provided by the standard theory of maximum likelihood. The effect of elevation is significant for the first and third used samples, with the negative coefficient indicating less use of units as they become higher. The distance from the nest is always very significant, with less use of units as they become further from the nest. The coefficient for young growth is always negative, and very significantly so for the second and third used sample. This indicates that old growth (the standard habitat) is selected more than young growth, particularly later in the study period. Similarly, the coefficient of clearcut is always negative, and it is very highly significant for the first and third samples. No clearcut at all appeared in the second used sample, so that the selection probability for this is estimated to be zero.

The estimated RSFs give an immediate assessment of the relative probabilities of use for old growth, young growth, and clearcut resource units. For example, considering August to December 1997, the RSF is estimated to be

$$w_1(\mathbf{x}) = \exp\{-0.798(\text{Elevation}) - 0.464(\text{Distance}) - 0.338(\text{Young}) -2.778(\text{Clearcut})\},$$

with obvious descriptions of the four variables being considered. This means that if the probability of an owl selecting old growth is θ, then the probability of it selecting young growth is $\exp(-0.338)\theta = 0.713\theta$, and the probability of selecting clearcut is $\exp(-2.778)\theta = 0.062\theta$. In this way, the estimated relative probabilities of use old growth, young growth, and clearcut for the three sampling periods are found to be: August to December, 1997, 1.000, 0.713 and 0.062; January to March, 1998, 1.000, 0.535 and 0.000; and April to December, 1998, 1.000, 0.363 and 0.075. It appears that, relative to old growth units, young growth units became selected less over the study period, and clearcut units always had a low probability of selection.

Table 6.3 Estimated parameters for resource selection functions estimated for a pair of northern spotted owls in three time periods.

Coefficient	Estimate	Std. err.	Ratio	P-value[1]
	August - December 1997			
X_1, Elevation	-0.798	0.383	-2.08	0.037
X_2, Distance to nest	-0.464	0.164	-2.84	0.005
X_3, Young growth	-0.338	0.226	-1.50	0.134
X_4, Clearcut	-2.778	0.614	-4.52	0.000
	January - March 1998			
X_1, Elevation	0.145	0.461	0.31	0.753
X_2, Distance to nest	-1.498	0.210	-7.14	0.000
X_3, Young growth	-0.625	0.253	-2.48	0.013
X_4, Clearcut[2]	-13.141			
	March - August 1998			
X_1, Elevation	-1.031	0.477	-2.16	0.031
X_2, Distance to nest	-0.919	0.191	-4.81	0.000
X_3, Young growth	-1.015	0.243	-4.18	0.000
X_4, Clearcut	-2.596	0.542	-4.79	0.000

[1] Determined by comparing the ratio of the estimate to the standard error with the standard normal distribution

[2] Clearcut did not appear in the second used sample. The large negative estimate in this case just gives an estimated probability of use very close to zero. The true maximum likelihood estimate is minus infinity, with no approximate standard error available.

6.4 Samples of Unused Resource Units Only

Another important special case is where a sample of available units is taken, and also S samples of unused resource units are taken at times t_1, t_2, ..., t_S. Then the approach used in Section 6.3 can again be used, providing that all the samples are small fractions of their respective populations of resource units.

There are, however, changes in interpretation required. Now, the probability of the ith unit surviving unselected to time t_j must be assumed to be well approximated by

$$\phi_j^*(\mathbf{x}_i) = \exp\{(\beta_{0j} + \beta_{1j}x_{i1} +...+ \beta_{pj}x_{ip})t_j\}, \tag{6.22}$$

where the argument of the exponential function is negative. Following the same derivation used for equations (6.18) and (6.19), it then follows that the probability that the ith unit is observed in the available sample, conditional upon it appearing in one of the samples is

$$\tau_a(\mathbf{x}_i) = 1 / \{1 + \sum_{k=1}^{S} \exp\{(\beta'_{0k} + \beta_{1k}x_{i1} +...+ \beta_{pk}x_{ip})t_k\}, \tag{6.23}$$

where $\beta'_{0k} = \beta_{0k} + \log_e(P_k/P_a)$, with P_a being the probability that an available unit is included in the available sample, and P_k is the probability that a unit which is still unused at time t_k is included in the sample of unused units taken at that time. Similarly, the probability of the ith unit appearing in the jth sample of unused units, conditional on it appearing in one of the samples is

$$\tau_j(\mathbf{x}_i) = \exp\{(\beta'_{0j} + \beta_{1j}x_{i1} +...+ \beta_{pj}x_{ip})t_j\}/\{1 + \sum_{k=1}^{S} \exp\{(\beta'_{0k} + \beta_{1k}x_{i1} +...+ \beta_{pk}x_{ip})t_k\}. \tag{6.24}$$

Equations (6.23) and (6.24) are similar to equations (6.18) and (6.19), but there are important differences. To begin with, in equations (6.18) and (6.19) the effects of time are incorporated into the parameters β'_{0k}, along with the sampling probabilities, and the original constant β_{0k} in the RSPF. This is not possible with equations (6.23) and (6.24), where the selection times multiply the linear combinations of the exponential functions.

Secondly, when the parameters in equations (6.23) and (6.24) are estimated by maximum likelihood, they do not allow the resource selection function to be estimated. Instead, what can be estimated is

$$\phi_j(\mathbf{x}_i) = \exp\{(\beta_{1j}x_{i1} + \beta_{2j}x_{i2} + ... + \beta_{pj}x_{ip})t_j\}, \tag{6.25}$$

which gives the probability that a unit is unused at time t, multiplied by an unknown constant, i.e., it is a relative survival rate function.

From an estimate of $\phi_j(\mathbf{x}_i)$ it is possible to order resource units on the basis of their estimated probabilities of not being used by time t_j, which will be the opposite to the order for their probability of being used. This is not as satisfactory as estimating the resource selection probability function, but may give sufficient information to make a study worthwhile. Exactly the same situation was found when logistic regression is used with one sample of used units (Section 5.5).

A likelihood function can be derived from equations (6.23) and (6.24) as explained for equation (6.20). Maximum likelihood estimation of the unknown parameters can then be carried out using either a standard statistical package, or more likely using a special purpose program. As in the previous section, there will generally be three models of particular interest. The first is the no selection model that is obtained by setting all of the parameters except β'_{01}, β'_{02}, ..., β'_{0S} equal to zero. The second is the constant selection model that is obtained by setting $\beta_{j1} = \beta_{j2} = ... = \beta_{jS}$, for all j (i.e, making the coefficient of X_j the same for all the unused samples). The third is the variable selection model for which all of the unused samples have different coefficients for each of the X-variables.

Example 6.3 Selection of Corixids by Minnows

For an example of a situation with samples of unused resource units, consider Popham's (1944) study of the use of corixids as food by minnows (*Phoxinus phoxinus*) for which the data are shown in Table 3.5. This was the subject of Example 3.5, and was also mentioned in Section 6.2. Popham sampled the corixid in a pond before introducing minnows, and then sampled daily after from three to nine days of predation to see whether the distribution was changing with regard to nine types of corixid defined by the species and shade of grey. No information is available about sampling fractions for this example because Popham endeavoured to obtain about 150 corixids in each of his daily samples, irrespective of the number present in the population.

Altogether, there were 2333 corixids collected by Popham. These therefore provide the data for the model described by equations (6.23) and (6.24). A special purpose program was used for fitting the various version of these models that are considered (Chapter 13).

Because the counts of different types of corixid are reasonably large, this set of data can be analysed as a log-linear model, and is used as an example of this type in the next chapter. Here, however, the data are analysed using the method just described. Both analyses give essentially the same results. However, the log-linear model approach does allow an easier assessment of the absolute goodness of fit of models through the comparison of observed and expected counts of corixids in different categories. In general, the advantage that the method based on equations (6.23) and (6.24) has over log-linear modelling is that it can be used even when all resource units are unique so that counting numbers in different categories of resource units is not feasible.

As noted above, there are three obvious models that can be entertained to account for the sample data. The simplest of these (model 1) is the no selection model, which says that the population proportions of the nine types of corixid were the same at all the sample times, and the sample differences in proportions were just due to random sampling effects. When fitted to the data this model gives a deviance of 7645.84 with 2326 df.

The second model that can be entertained (model 2) says that the resource selection probability function is dependent on a species effect and a colour effect, with no interaction between these, and with the effects the same for all samples of unused units. This can be achieved by creating appropriate indicator variables for use in the model. Specifically, four variables were constructed with values for each of the sampled corixids, with $X_1 = 1$ for *Sigara venusta*, or otherwise 0, $X_2 = 1$ for *S. praeusta*, or otherwise 0, $X_3 = 1$ for light corixids, or otherwise 0, and $X_4 = 1$ for medium corixids, otherwise 0. Use of these variables means that the 'standard' corixid is a dark *S. distincta*, for which all X values are zero.

Fitting equations (6.23) and (6.24) with constant coefficients for the four X-variables resulted in a deviance of 7530.39 with 2322 df. Comparing this to the no selection model, it is seen that the deviance is reduced by 115.45, with 4 df. This is very significantly large in comparison with the chi-squared distribution with 4 df, giving strong evidence of selection.

The model with the coefficients of X_1 to X_4 allowed to vary with the sample time (model 3) was considered next. This gave a deviance of 7487.02 with 2298 df. The reduction in deviance obtained by allowing for variable selection is 43.37, with 24 df. This is significantly large at the 5% level, giving some evidence of variation.

Two other models were also fitted. The first of these (model 4) allowed the level of selection to vary with all nine different types of corixid, but assumed that this was constant over time. This was achieved by defining indicator variables to be $X_1 = 1$ for medium *S. venusta*, or otherwise 0, $X_2 = 1$ for dark *S. venusta*, or otherwise 0, and so on up to $X_8 = 1$ for dark *S. distincta*, or otherwise 0. The 'standard' corixid is then light *S. venusta*, for which all X values are zero, although another of the species-colour combinations would have served just as well as the standard. Fitting this model gave a deviance of 7524.99 with 2318 df.

Finally, the model was fitted where the coefficients of X_1 to X_8 varied with the sample of unused corixids taken (model 5). This gave a deviance of 7471.24, with 2270 df. Compared with model 4, there is a deviance reduction of 53.75, with 48 df. This is not at all significant in comparison with the chi-squared distribution, therefore providing no evidence of changing selection.

There is a problem in comparing these models because they are not all hierarchical. In particular, model 3 and model 4 are both generalizations of model 2, but models 3 and 4 are not in order. This is a situation where Akaike's information criterion (AIC, Section 2.8) is useful. Based on AIC, which is the deviance plus twice the number of estimated parameters, the best model is the one with the smallest value. The AIC values for models 1 to 5, respectively, are 7659.84, 7552.39, 7557.02, 7554.99, and 7597.24. This therefore suggests that model 2 (constant selection related to species and colour) is best for the data, even though there is some evidence for changing selection.

Parameter estimates from model 2 are shown in Table 6.4. According to this model, the probability of survival to time t is proportional to

$$\hat{\phi}_t(\mathbf{x}) = \exp\{(0.106\text{Venusta} + 0.063\text{Praeusta} - 0.158\text{Light} + 0.092\text{Medium})t\},$$

with obvious names for the indicator variables that allow for species and colour effects. Setting t = 1 gives relative survival rates for one day, as shown in Table 6.5. It seems that *S. venusta* survived better than *S. praeusta*, which in turn survived better than *S. distincta*. Also, medium coloured corixids survived better than dark corixids, which in turn survived better than light coloured ones. In other words, the corixids that were favoured by the minnows tended to be light coloured *S. distincta*.

Table 6.4 Estimates of parameters for the model of equation (6.26) fitted to the data from Popham's experiment on the predation of corixids by minnows.

Coefficient	Estimate	Standard error	Ratio	p-value[1]
S. venusta, β_1	0.106	0.038	2.77	0.006
S. praeusta, β_2	0.063	0.050	1.26	0.209
Light grey, β_3	-0.158	0.035	-4.49	0.000
Medium grey, β_4	0.092	0.017	5.32	0.000

[1] Determined by comparing the ratio of the estimate to the standard error with the standard normal distribution.

Table 6.5 Estimated relative daily survival rates for different types of corixid.

Species	Light	Medium	Dark
S. venusta	0.95	1.22	1.11
S. praeusta	0.91	1.17	1.07
S. distincta	0.85	1.10	1.00

Chapter Summary

- When there are several periods of selection, with more and more resource units being used as time goes on, the reduction in unused units can be modelled as a survival process, possibly using the proportional hazards survival function. If counts of the number of units used in each survival period are known then the survival function, and hence the resource selection probability function can be estimated by setting up a generalized linear model for the data, with the complementary log-log link function.

- An example on the selection of different species and sizes of *Cepaea nemoralis* and *C. hortensis* is used to illustrate the fitting process with census data of this type.

- If only samples are available of different types of resource units (available, used and unused) then the proportional hazards model can still be applied but the model is non-standard and requires special calculations to find maximum likelihood estimates of parameters.

- A important special case is where the proportion of resource units used during the total selection time is only a small fraction of the total population of units. In this case the resource selection function can be estimated from a sample of available units, and two or more samples of used units, using a generalization of the logistic regression method for a sample of available and a sample of used units.

- The analysis of an available sample and three samples of used units is illustrated using data on habitat selection by the northern spotted owl.

- Another important case is where the data available consists of a sample of available units and two or more samples of unused units. In this case a relative survival rate function can be estimated by a generalization of logistic regression.

- The analysis of an available sample and several unused samples is illustrated with an example involving the predation of different species and colours of corixids by minnows.

Exercises

(1) Example 5.1 was concerned with Ryder's (1983) study of winter habitat selection by antelopes (*Antilocapra americana*) in the Red Rim area of south-central Wyoming. This study involved recording a number of variables on each of 256 plots of ground in the region, and noting whether pronghorn were seen on each plot

in the winters of 1980-81 and 1981-82. From the analysis with the example it seems that the probability of a plot being used was mainly a function of the distance to water and the aspect of the plots. For the present exercise, assume that the distance to water and the aspect of a plot are the only important variables for determining probability of use. On this basis, it is possible to group the 256 study plots into I = 36 different types of unit and hence fit the proportional hazards model to the data providing that a plot is considered to be used the first time that pronghorn are recorded. In this way, the extensive set of data provided in Table 3.2 reduces to the much smaller set shown in Table 6.6.

Fit the proportional hazards model to this reduced set of data using the principle of maximum likelihood, by applying equation (6.9). Compare the estimated RSPF that is obtained in this way with the function that was estimated by logistic regression in Example 5.1. Examine residuals to see whether the model appears to give a reasonable fit to the data for all aspects and distances to water.

(2) In a study carried out in Marley Wood, near Oxford, England, Sheppard (1951) collected two samples of live snails (*Cepaea nemoralis*) and seven samples of broken shells close to a thrush 'anvil'. The results obtained are shown in Table 6.7, with the snails classified into two groups according to the colour of the shells (pink and brown or yellow), and the nature of the sample (broken shells or live snails). On Day 1 (6 April, 1950) the thrush 'anvil' was cleared of broken shells, so that the sample of broken shells collected on day 5 was of snails taken by thrushes over four days, the sample of broken shells collected on day 17 was a sample of snails taken by thrushes over 12 days, and so on. The two samples of live snails were of the available snails at the sample times. Analyse the data using the method described in Section 6.3, assuming that the two samples of live snails can be pooled to give the available sample.

Table 6.6 The data on winter habitat use by antelopes from Table 3.2 arranged as required for the estimation of a proportional hazards model of equations, when only the distance to water and the aspect of plots are considered. The three aspect variables are E/NE = 1 for East/Northeast, otherwise 0, S/SE = 1 for South/Southeast, otherwise 0, and W/SW = 1 for West/Southwest, otherwise 0. The fourth aspect, North/Northwest, is the 'standard' aspect that receives 0 for E/NE, S/SE and W/SW. The use variables are U_1 = number of plots used in winter 1, U_2 = number of plots used for the first time in winter 2, and $\bar{U}$ = number of plots not used in either winter.

Type of Plot	Distance to Water (m)	Aspect Variables			Use Variables		
		E/NE	S/SE	W/SW	U_1	U_2	$\bar{U}$
1	25	1	0	0	1	1	0
2	150	1	0	0	2	0	0
3	375	1	0	0	5	0	0
4	625	1	0	0	1	0	2
5	875	1	0	0	0	2	1
6	1250	1	0	0	1	0	3
7	1750	1	0	0	0	0	1
8	2250	1	0	0	1	0	2
9	2750	1	0	0	2	2	1
10	25	0	1	0	0	0	1
11	150	0	1	0	0	2	0
12	375	0	1	0	1	0	1
13	625	0	1	0	0	0	2
14	875	0	1	0	2	0	0
15	1250	0	1	0	0	0	4
16	1750	0	1	0	2	0	3
17	2250	0	1	0	1	1	1
18	2750	0	1	0	0	1	9
19	25	0	0	1	0	0	1
20	150	0	0	1	1	1	0
21	375	0	0	1	4	1	0
22	625	0	0	1	3	0	6
23	875	0	0	1	6	1	3
24	1250	0	0	1	0	1	4
25	1750	0	0	1	0	0	4
26	2250	0	0	1	1	2	1
27	2750	0	0	1	1	1	6
28	25	0	0	0	4	4	5
29	150	0	0	0	2	2	5
30	375	0	0	0	5	4	8
31	625	0	0	0	6	2	10
32	875	0	0	0	6	11	10
33	1250	0	0	0	2	3	15
34	1750	0	0	0	6	6	6
35	2250	0	0	0	1	0	5
36	2750	0	0	0	5	6	13

Table 6.7 Colour composition of samples of unused (live, L) and used (broken, B) Cepaea nemoralis snails from Marley Wood, Oxford, with days counted from 6 April, 1950.

	Day								
	5	8	17	24	31	43	46	50	50
Sample	B	L	B	B	B	B	B	B	L
Pink and brown	4	250	10	21	25	16	6	12	147
Yellow	3	80	7	11	9	3	1	2	57

CHAPTER 7

LOG-LINEAR MODELLING

Log-linear models are defined in Section 2.4. They provide a general purpose tool for analysing count data that can be used to relate resource selection to factors such as the individual animal involved and the time when selection is made, as well as characteristics of the resource units. In addition there is the possibility of analysing data on samples of resource units collected at different times from a population that is changing as selection takes place. In this chapter, log-linear modelling is discussed further, with its use illustrated with examples concerning habitat selection by elk and white-tailed deer, and the predation of corixids by minnows.

7.1 General Log-linear Modelling

As noted by Heisey (1985), log-linear modelling has a great deal of potential for modelling changes in selection related to factors such as the individual animal concerned, the time of day, the season of the year, etc.

With a general log-linear model it is assumed that the data available consists of m counts y_1 to y_m, where the ith count has a Poisson distribution that is independent of the other counts, with a mean value that can be expressed as

$$\mu_i = \exp(ß_0 + ß_1x_{i1} + ... + ß_px_{ip}), \qquad (7.1)$$

where $ß_0$ to $ß_p$ are constants to be estimated, and x_{i1} to x_{ip} are known values for the variables X_1 to X_p.

Sometimes it is known that μ_i should be proportional to some base rate, B_i, so that equation (7.1) is better written as

$$\mu_i = B_i\exp(ß_0 + ß_1x_{i1} + ... + ß_px_{ip}). \qquad (7.2)$$

For example, suppose that the data count y_i is the observed number of times that a certain animal is seen in the ith type of habitat, where it is known that 60% of the habitat available to the animal is of this type. Then it is appropriate to set $B_i = 0.60$ so that if there is no selection then the model $\mu_i = 0.60\exp(ß_0)$ should fit the data, with the term $\exp(ß_0)$ allowing for the total number of observations made on the animal.

Log-linear models are usually fitted by maximum likelihood, which requires that a suitable computer program is available. There are many such programs available (Chapter 14), and there is no need to discuss the use of these here, other than to mention that in many cases these programs will construct sets of X variables to allow for the effects of categorical variables. For example, consider the data in Table 6.6 on the use by antelopes of study plots with different distances from water and different aspects. Here the four aspects (East/Northeast, South/Southeast, West/Southwest and

North/Northwest) are allowed for using the three 0-1 variables E/NE, S/SE and W/SW. Some computer programs will set up variables of this type automatically if the data input indicates that the data frequencies are classified by a factor at four levels. Note, however, there are alternative ways of defining the X variables for categorical variables, so that some computer programs would produce three X variables with different values from those shown in Table 6.6.

The goodness of fit of a log-linear model can be measured by the deviance, which is minus twice the difference between the maximised log-likelihood for the fitted model and the log-likelihood for the model which fits the data perfectly (Section 2.7). This deviance takes the form

$$D = 2 \sum_{i=1}^{m} y_i \log_e(y_i/\hat{\mu}_i),$$

where $\hat{\mu}_i$ is the expected value of the ith data frequency according to the fitted model. It has m - p - 1 degrees of freedom (df), and if this statistic is significantly large in comparison with the chi-squared distribution then the model is a poor fit to the data. As usual, most of the estimated expected values $\hat{\mu}_i$ should be five or more for the test to be reliable. Also, as discussed in Section 2.7, differences between the deviances for different models can be tested against the chi-squared distribution to assess the relative goodness-of-fit of those models.

7.2 Examples

Log-linear models are so versatile that is not possible to cover all the situations where resource selection data can be analysed using these models. However, the following examples illustrate the type of approach that can be used.

Example 7.1 Selection of Forest Canopy Cover by Elk

Consider again Marcum and Loftsgaarden's (1980) study of the selection of forest over-story canopy cover by elk (*Cervus elaphus*), that was discussed in Example 4.3 in the context of estimating selection ratios. Here the data (Table 4.6) consist of counts in four canopy over-story classes for a sample of 200 random points on a map of the study area, and an independent sample of 325 points selected by the population of elk.

To analyse the data as a log-linear model, consider the data as set out in Table 7.1. Here the first column gives the observed sample counts, while columns two to eight give the following X variables: $X_1 = 0$ for a count of available points and 1 for a count of used points; $X_2 = 1$ for the 0% canopy class, otherwise 0; $X_3 = 1$ for the 1-25% canopy class, otherwise 0; $X_4 = 1$ for the 26-75% canopy class, otherwise 0; X_5 = the product of X_1 and X_2; X_6 = the product of X_1 and X_3; and X_7 = the product of X_1 and X_4. Note that the definitions of X_2 to X_7 are such that the availability of the 76-100% canopy class is being regarded as the 'standard' level of availability, and the selection for or against the 75-100% canopy class is taken as the standard amount of selection.

Using these variables, a log-linear model gives the expected value of the ith frequency as

$$\mu_i = \exp(ß_0 + ß_1 x_{i1} + ... + ß_7 x_{i7}). \tag{7.3}$$

Here X_1 allows for the fact that the sample of used points is not the same size as the sample of available points, X_2 to X_4 allow for the availability of the 0%, 1-25% and 26-75% canopy classes to be different from the availability of the 76-100% class, and X_5 to X_7 allow for the counts in the first three canopy classes to vary between the sample of available points and the sample of used points. Thus this is a log-linear model for the data which allows for the possibility of resource selection through the inclusion of variables X_5 to X_7.

Table 7.1 Data counts and X variables for a log-linear model for Marcum and Loftsgaarden's example on the selection of forest over-story canopy by elk.

Data count	X-variables						
Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7
15	0	1	0	0	0	0	0
61	0	0	1	0	0	0	0
84	0	0	0	1	0	0	0
40	0	0	0	0	0	0	0
3	1	1	0	0	1	0	0
90	1	0	1	0	0	1	0
181	1	0	0	1	0	0	1
51	1	0	0	0	0	0	0

If the no selection model is fitted, using variables X_1 to X_4 only (by setting β_5 to β_7 equal to zero), then the deviance is $D_0 = 21.96$ with 3 df. This is clearly a very poor fit.

If the model is expanded to include the variables X_5 to X_8 then there are eight data counts and eight parameters in the model, which means that the log-linear model must fit the data exactly. This means that for this model the deviance must be $D_1 = 0$, with 0 df. The difference $D_0 - D_1 = 21.96 - 0 = 21.96$, with $3 - 0 = 3$ df, is then a measure of the improvement in the fit of the model due to allowing for habitat selection. As this is very significantly large, there is very strong evidence of selection. Note that this test for selection based on the comparison of the fit of log-linear models with and without an allowance for selection is exactly the same as the test using equation (4.19) for whether the sample of used resource units comes from the same population as the sample of available units.

Because the model allowing for selection must fit the data exactly, one way to determine estimates of the parameters ß$_0$ to ß$_7$ involves just equating each of the observed counts to its expected value and solving the resulting eight equations. Alternatively, one of the standard programs for estimating the parameters of log-linear models can be used, where this has the advantage of also producing standard errors for the estimated parameters.

Estimates with their standard errors are shown in Table 7.2. Because ß$_5$ is more than two standard errors below zero, it appears that there is selection against the 0% canopy class, in comparison to the 76-100% class. On the other hand, because ß$_7$ is more than two standard errors above zero it appears that there is selection in favour of the 26-75% canopy class, in comparison to the 76-100% class.

The parameters of the log-linear model can be related to the selection ratios that have been discussed in Chapter 4. In that chapter the selection ratio for type i resource units was defined to be the ratio of the proportion of used units to the proportion of available units of that type, which is proportional to the ratio of the number of used type i units to the number of available type i units. In the context of the log-linear model this implies that the selection ratio for the 0% canopy cover is proportional to

$$\mu_5/\mu_1 = \exp(ß_1 + ß_2 + ß_5)/\exp(ß_2) = \exp(ß_1 + ß_5),$$

taking into account the 0-1 X values shown in Table 7.1. In a similar way it can be argued that the selection ratios for the 1-25%, 26-75% and 76-100% canopy classes are proportional to

$$\mu_6/\mu_2 = \exp(ß_1 + ß_6),$$
$$\mu_7/\mu_3 = \exp(ß_1 + ß_7)$$

and

$$\mu_8/\mu_4 = \exp(ß_1),$$

respectively.

Table 7.2 Estimated coefficients for the log-linear model for canopy selection by elk, with standard errors, ratios of estimates to their standard errors, and significance levels.

Parameter	Estimate	Standard error	Ratio	p-value[1]
$ß_0$, constant	3.690	-	-	-
$ß_1$, used/unused indicator	0.243	0.211	1.15	0.249
$ß_2$, 0% canopy class	-0.983	0.303	-3.24	0.001
$ß_3$, 1-25% canopy class	0.421	0.203	2.07	0.038
$ß_4$, 26-75% canopy class	0.741	0.192	3.86	0.000
$ß_5$, use of 0% canopy class	-1.854	0.667	-2.78	0.005
$ß_6$, use of 1-25% canopy class	0.146	0.268	0.54	0.586
$ß_7$, use of 26-75% canopy class	0.525	0.249	2.11	0.035

[1] Ratio tested against the standard normal distribution, to see if it is significantly different from zero.

The implication of these results is that the estimated selection ratios $\hat{w}_1$ to $\hat{w}_4$ are proportional to exp(b_5), exp($_6$), exp(b_7) and exp(0) = 1, respectively, where the b_i is the estimated value for $ß_i$ from the fitted the log-linear model. For the data being considered, the ß estimates are as shown in Table 7.2, so that the estimated selection ratios are proportional to exp(-1.854) = 0.157, exp(0.146) = 1.157, exp(0.525) = 1.690 and 1.000, respectively. These are then the selection ratios shown in Table 4.6 multiplied by 0.785.

In summary, what this means is that the selection ratios that were estimated using equation (4.22) in example 4.3 are proportional to exponential functions of the parameters of the log-linear model that relate to selection. A similar result will apply whenever a log-linear model is fitted to data from a sample of available resource units and a sample of used resource units.

Of course, the calculations used in Chapter 4 are much more straightforward to use than a log-linear model, which means that in practice we would not recommend the use of a log-linear model for a simple situation like that of the present example. Therefore the reason for presenting this example was just to illustrate the relationship between log-linear modelling and the methods that have been discussed in Chapter 4.

Example 7.2 Habitat Selection by White-Tailed Deer

A more practically useful application of log-linear modelling is described by Heisey (1985), using data reported by Nelson (1979) from a radio tracking study of habitat use by white-tailed deer (*Odoncoileus virginianus*). Here there are four habitat types, and relocations were observed for two deer at two times of the day. The counts of relocations in different habitats and the proportional availability of different habitats are shown in Table 7.3.

Table 7.3 Habitat use (HU) of two white-tailed deer in four types of habitat, with the proportional availability of those habitats (HA).

	Midday				Morning and evening			
	Deer 68		Deer 342		Deer 68		Deer 342	
Habitat	HU	HA	HU	HA	HU	HA	HU	HA
Aspen	18	0.66	29	0.65	43	0.66	46	0.65
Clearcut	2	0.20	1	0.13	33	0.20	29	0.13
Plantation	0	0.09	4	0.13	5	0.09	4	0.13
Spruce	0	0.05	0	0.09	0	0.05	2	0.09

In the terminology of Chapter 1, this is an example of a Design III study with sampling protocol A, with availability censussed and use sampled for each of two animals. Furthermore, since the resources being studied are defined by several categories, it is similar to the situation that has been discussed in Section 4.15. There is, however, the complication that the used resources were sampled at two times of day.

One method for analysing the data that might be considered involves using equation (4.41) to estimate selection ratios first using the midday results, and then separately for the morning and evening results. Equation (4.3) could then be used to estimate variances as explained in Section 4.15. However, the problem with this approach is that the variance estimates would be rather unreliable because they would be based on differences between only two animals. Hence tests for selection and for differences between selection ratios would also be unreliable.

A log-linear model analysis overcomes these problems providing that it can be assumed that the counts in Table 7.3 are values from independent Poisson distributions. This will be a reasonable providing that the individual observations on deer locations were far enough apart in time to be independent. We assume that this was the case.

Heissey used the computer program GLIM (Francis *et al*., 1993) to analyse the data. This program automatically constructs X variables for the four habitat types, two deer, and two times of day. However, these variables can be set up easily enough for use in a program that does not have this facility, as shown in Table 7.4. It will be seen from this table that base rates and 12 variables are needed to account for selection related to the time of day and the deer. The model takes the form

$$\mu_i = B_i \exp(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_{12} x_{i12}), \tag{7.4}$$

where μ_i is the expected value of the ith count in Table 7.4, B_i is the associated base rate which reflect the availability of different types of habitat, and the exponential part of the equation is the resource selection function (RSF). Equation (7.4) gives the full model considered, with other models being obtained by setting some β parameters equal to zero.

Table 7.4 Data for fitting log-linear models to account for the frequencies with which different types of habitat are used by white-tailed deer. The base rates are the proportions available of different habitats. The X variables are: X_1 = 1 for clearcut, otherwise 0; X_2 = 1 for plantation, otherwise 0; X_3 = 1 for spruce, otherwise 0; X_4 = 1 for midday, 0 for morning and afternoon; X_5 = 1 for deer 68, 0 for deer 342; $X_6 = X_1X_4$; $X_7 = X_2X_4$; $X_8 = X_3X_4$; $X_9 = X_1X_5$; $X_{10} = X_2X_5$; $X_{11} = X_3X_5$; and $X_{12} = X_4X_5$.

Sample count	Base rate	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
18	0.66	0	0	0	1	1	0	0	0	0	0	0	1
29	0.65	0	0	0	1	0	0	0	0	0	0	0	0
43	0.66	0	0	0	0	1	0	0	0	0	0	0	0
46	0.65	0	0	0	0	0	0	0	0	0	0	0	0
2	0.20	1	0	0	1	1	1	0	0	1	0	0	1
1	0.13	1	0	0	1	0	1	0	0	0	0	0	0
33	0.20	1	0	0	0	1	0	0	0	1	0	0	0
29	0.13	1	0	0	0	0	0	0	0	0	0	0	0
0	0.09	0	1	0	1	1	0	1	0	0	1	0	1
4	0.13	0	1	0	1	0	0	1	0	0	0	0	0
5	0.09	0	1	0	0	1	0	0	0	0	1	0	0
4	0.13	0	1	0	0	0	0	0	0	0	0	0	0
0	0.05	0	0	1	1	1	0	0	1	0	0	1	1
0	0.09	0	0	1	1	0	0	0	1	0	0	0	0
0	0.05	0	0	1	0	1	0	0	0	0	0	1	0
2	0.09	0	0	1	0	0	0	0	0	0	0	0	0

The first column of Table 7.4 shows the counts of habitat use by the two deer, the second column gives the proportional availabilities of habitats, where these can be thought of as base rates B_i as in equation (7.3), variables X_1 to X_3 allow for sample counts to vary with the habitat in addition to the variation that is expected from the base rates, X_4 allows sample counts to vary with the time of day, X_5 allows sample counts to vary with the deer, X_6 to X_8 allow the effects of different habitats to vary with the time of the observations, X_9 to X_{11} allow the effects of different habitats to vary with the deer, and X_{12} allows the deer effect to vary with the time of day.

On the basis of these variables, a no selection model is one that includes X_4, X_5 and X_{12}. This allows the expected counts to depend on the deer and the time of day, with the deer effect possibly varying with the time of day, but says that habitat use is proportional to the availability. The deviance is 78.26 with 12 df.

Adding X_1 to X_3 into the model allows for some selection to take place, where this is at the same level for both deer and both times of day. The resulting model has a deviance of 32.42 with 9 df.

At this stage it is possible to expand the model to either allow selection to depend on the time of day (by adding X_6 to X_8 into the model) or to allow selection to depend on the deer (by adding X_9 to X_{11} into the model). It turns out that the deviance is reduced considerably to 6.44 with 6 df if the first of these two options is taken, but reduced hardly at all to 30.88 with 6 df if the second option is chosen. The first option is therefore best.

The next stage in model building consists of allowing selection to depend on both the deer and the time of day. The model, which then includes all of the variables X_1 to X_{12}, has a deviance of 4.41 with 3 df.

The only way that the model can be expanded at this stage is by adding X variables that allow the selection of habitat by deer to vary with the time of day. However, there would then be as many parameters as sample frequencies so that the model would be saturated with parameters, and fit the data exactly.

The model building process just described is summarised in the analysis of deviance shown in Table 7.5. A reasonable conclusion from this summary is that there was selection, and that this depended on the time of day but not on the deer. Estimates for this model are shown in Table 7.6, where the large negative estimate for the coefficient of X_8 comes about because spruce was never used at midday (i.e., it makes the expected frequency of spruce at midday very close to zero).

Table 7.5 Analysis of deviance table for the log-linear model analysis of habitat selection by white-tailed deer.

			Difference	
Model	Deviance	df	Deviance	df
No selection of habitat	78.26 *	12		
			45.84 *	3
Constant selection on habitat	32.42 *	9		
			25.98 *	3
Selection varies with time	6.44	6		
			2.03	3
Selection varies with time and deer	4.41	3		
			4.41	3
Selection with the time effect varying with the deer	0.00	0		

*Significantly large at the 0.1% level.

The expected frequencies for the fitted model are given by the equation

$$\mu_i = B_i\exp(4.29 + 1.03x_{i1} - 0.52x_{i2} - 1.58x_{i3} - 0.47x_{i4} - 0.15x_{i5} - 2.35x_{i6} - 0.21x_{i7} -15.82x_{i8} - 0.40x_{i12}),$$

where the variables X_4, X_5 and X_{12} take into account differences in the amount of use at different times of day and by different deer, while X_1, X_2, X_3, X_6, X_7 and X_8 allow for selection of habitat by the deer. To estimate the magnitude of selection it is simplest to fix on the standard time of day (morning and evening) and the standard deer, for which $X_4 = X_5 = X_{12} = 0$, and assume one unit of habitat is available, so that $B_i = 1$. This then produces a RSF, which gives the estimated probability of selecting different types of habitat, multiplied by an unknown constant. Because of the unknown multiplication factor, the constant term in the exponential function can also be removed, to give the estimated RSF

$$\hat{w}_i = \exp(1.03x_{i1} - 0.52x_{i2} - 1.58x_{i3} - 2.35x_{i6} - 0.21x_{i7} -15.82x_{i8} - 0.40x_{i12}).$$

The values obtained from this function are shown in Table 7.7. It appears that at midday aspen is the preferred habitat, followed by plantation and clearcut. Spruce is estimated to have a zero probability of selection because the two deer in the study never used this at midday. However, in the morning and evening clearcut is the most preferred habitat, followed by aspen, plantation and spruce.

Table 7.6 Parameter estimates for the log-linear model for habitat selection by white-tailed deer, with selection varying with the time of day.

Coefficient of		Estimate	SE	Ratio	p-value
Constant		4.289	0.129	33.17	0.000
X_1	Clearcut	1.032	0.166	6.21	0.000
X_2	Plantation	-0.521	0.350	-1.49	0.137
X_3	Spruce	-1.581	0.715	-2.21	0.027
X_4	Midday	-0.472	0.221	-2.13	0.033
X_5	Deer 68	-0.146	0.158	-0.92	0.357
X_6	Clearcut.Midday	-2.349	0.619	-3.80	0.000
X_7	Plantation.Midday	-0.208	0.628	-0.33	0.741
X_8	Spruce.Midday	-15.820	2581.000	-0.01	0.995
X_{12}	Deer 68.Midday	-0.397	0.324	-1.23	0.220

Table 7.7 Estimated values from the resource selection function from the log-linear model for habitat selection by deer.

	Relative selection probability	
Habitat	Midday	Morning and
Aspen	1.00	1.00
Clearcut	0.27	2.81
Plantation	0.48	0.59
Spruce	0.00	0.21

Example 7.3 Selection of Corixids by Minnows

For a final example of log-linear modelling, Popham's (1944) study of the use of corixids as food by minnows (*Phoxinus phoxinus*) is revisited. The data are shown in Table 3.5. This study was the subject of Example 3.5, and the results have already been analysed one way in Example 6.3. It may be recalled that Popham sampled the corixid in a pond before introducing minnows, and then sampled daily after from three to nine days of predation to see whether the distribution was changing with regard to nine types of corixid defined by the species and shade of grey.

For the log-linear modelling approach to analysis, the first model to consider is the no selection model, which says that the population proportions of the nine types of corixid were the same at all the sample times, and the sample differences in proportions were just due to random sampling effects. For this model the expected frequencies of the nine types of corixid for each sample are just the sample total allocated out according to the proportions found with all samples lumped together, as in a standard chi-squared test for no association between row and column categories in a two-way contingency table. On this basis, the deviance is $D_1 = 174.6$, with 56 df. This is very significantly large, indicating that the model is a very poor fit to the data.

A second model that can be considered says that the resource selection probability function is dependent on a species effect and a colour effect, with no interaction between these. This can be achieved by taking $X_1 = 1$ for *Sigara venusta*, or otherwise 0, $X_2 = 1$ for *S. praeusta*, or otherwise 0, $X_3 = 1$ for light corixids, or otherwise 0, and $X_4 = 1$ for medium corixids, otherwise 0. The standard corixid is then a dark *S. distincta*, for which all X values are zero. These are the same X-variables as used with Example 6.3, where the same model was fitted a different way.

Suppose that the probability of a type i corixid surviving uneaten until time t_j takes the form

$$\phi_j(\mathbf{x}_i) = \exp\{(ß_0 + ß_1x_{i1} + ß_2x_{i2} + ß_3x_{i3} + ß_4x_{i4})\, t_j\},$$

where $\mathbf{x}_i = (x_{i1},x_{i2},x_{i3},x_{i4})$ are the variables that describe this individual. On this basis, the expected number of type i corixids in a sample taken after t_j days of predation is

$$E(\bar{u}_{ij}) = (P_jA_i)\exp\{(ß_0 + ß_1x_{i1} + ß_2x_{i2} + ß_3x_{i3} + ß_4x_{i4})\, t_j],$$

where P_j is the probability that a corixid is included in the sample on day j, and A_i is the number of type i corixids in the population before predation began. Actually, because P_j and β_0 are both unknown they cannot be estimated separately. Therefore they can be combined together into the parameter $\beta'_j = \log_e(P_j) + \beta_0t_j$, to give

$$E(\bar{u}_{ij}) = \exp\{\log_e(A_i) + ß'_j + (ß_1x_{i1} + ß_2x_{i2} + ß_3x_{i3} + ß_4x_{i4})\, t_j\}. \qquad (7.4)$$

For the available sample the expected count of type i corixids is just

$$E(a_i) = P_aA_i,$$

where Pa is the probability of a corixid being included in the available sample. To match equation (7.4), this can be written as

$$E(a_i) = \exp\{\log_e(A_i) + \beta'_0\}, \qquad (7.5)$$

where $\beta'_0 = \log_e(P_a)$.

Between them, equations (7.4) and (7.5) describe a log-linear model for the data. The values $\log_e(A_i)$ represent the effects of an abundance factor with a different level for every identifiable type of corixid. The values β'_0 to β'_j represent the effects of a second factor with a different level for every sample. Finally, the linear combination $(ß_1x_{i1} + ß_2x_{i2} + ß_3x_{i3} + ß_4x_{i4})\, t_j$ represents the effects of selection up to the time of the jth used sample. In order to include this linear combination in the model, the X-values are just multiplied by t_j when used with the observations from the jth used sample.

The model described by equations (7.4) and (7.5) was fitted to the data. The only parameters that are important are the coefficients of the X-variables, which are exactly the same as those shown in Table 6.4 where the same model was estimated a different way. The deviance is $D_2 = 59.1$ with 52 df, which is not significantly large at the 5% level in comparison with the chi-squared distribution. The model is therefore a good fit. Note that this chi-squared test for goodness of fit is valid for this example because most of the observed and expected frequencies are reasonably large.

A third model that can be entertained for the data is one which says that the level of selection varies with all nine different types of corixid. This can be fitted by defining eight dummy variables to be $X_1 = 1$ for medium *S. venusta*, or otherwise 0, $X_2 = 1$ for dark *S. venusta*, or otherwise 0, and so on up to $X_8 = 1$ for dark *S. distincta*, or otherwise 0. The standard corixid is then light *S. venusta*, for which all X values are zero. Again, this is one of the models considered in Example 6.3. Fitting by log-linear modelling proceeds as before. The deviance $D_3 = 53.7$, with 48 df, which shows that the model gives a good fit to the data. However, the reduction in deviance in moving from model 2 to model 3 is only 5.4, with 4 df, which is not at all significant. Again the estimated parameters are the same as was obtained for essentially the same model when it was fitted in Example 6.3.

Other models could be fitted to the data by log-linear modelling. However, these would just give the same estimates as the method used in Example 6.3, so this will not be done. What is useful about the log-linear modelling approach is that it produces residuals directly, and these can be used to assess the fit of a model. Here this is illustrated on model 2, which seems to be a reasonable one for the data.

Figure 7.1 shows the standardized deviance residuals plotted against the sample times. It is expected that almost all of the residuals will be in the range from -3 to +3, which they are, with three exceptions. The most notable exception is the large value of about five which occurs because an observed frequency of 58 dark *S. venusta* has an expected frequency of only 29.7 on the fourth day after predation began. Clearly the observed frequency is anomalous. However, it is a genuine value so that a fair conclusion is that the fitted model is generally reasonable but there may have been some problem with the sampling for the third sample.

Chapter Summary

- Log-linear modelling is a useful tool for analysing data consisting of counts of the numbers of different types of individuals in different types of samples. Log linear models are usually fitted by maximum likelihood, assuming that counts follow Poisson distributions. Deviances can be used to measure the goodness of fit of models providing that most counts are five or more. Differences between deviances can be used to assess the relative goodness of fit of different models.

- An example on the selection of habitat by elk illustrates how log-linear modelling relates to the selection ratio methods discussed earlier in Chapter 4.

- An example of habitat selection by white-tailed deer illustrates how a log-linear model can account for selection, and changes in selection related to the time of observations and the individual animal doing the selection.

- A final example on the predation of corixids by minnows illustrates how a log-linear model can estimate the function describing the probability of surviving selection (rather than a resource selection function) in situations where the data available comes from a sample of available resource units, and one or more samples of used units. Here the log-linear modelling approach is an alternative to a method described in the previous chapter, for which estimates are identical.

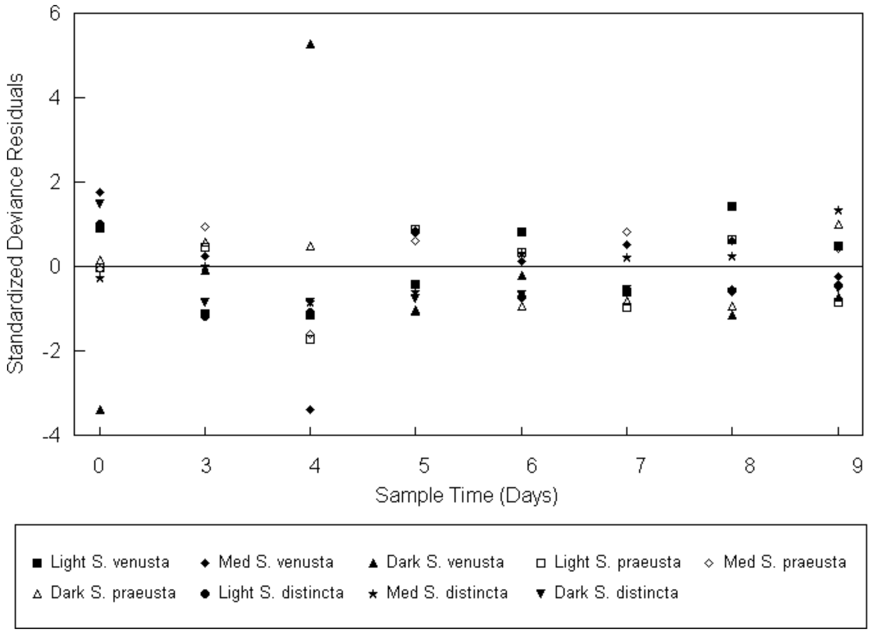


Figure 7.1 Standardized deviance residuals for a log-linear model fitted to the data from Popham's experiment on the predation of corixids by minnows.

Exercise

Sheppard's (1951) study of the selection of snails with different shell colours in Marley Wood, Oxford, England, was used for Exercise 2 in Chapter 6. Analyse the data, which are given in Table 6.7, using a log-linear model approach.

CHAPTER 8

DISCRETE CHOICE MODELS

In the analyses of previous chapters, resource availability has been modelled as constant over the study period and in most cases with equally available for individual animals. Often this is reasonable, but in some cases availability changes with time either for the individual animals or for the entire population of animals. In that case one approach for data analysis that can sometimes be used is based on discrete choice models, or what is also sometimes called polytomous regression. This approach is covered in the present chapter, with an example based on the estimation of ocean habitat by seabirds.

8.1 The Theory of Discrete Choice Models

Discrete choice models view the resource selection problem as a series of choices made by the animals under study, with the unique set of resource units available at each choice called the choice set. For example, a choice set might be defined as a number of alternative locations within a small geographic area, of which one is chosen for use by an animal. Then as the animal moves around the choice set changes. Or, an animal might be presented with food items A, B, and C, at one point in time, of which one is chosen, but the choice might be from food items A, B, and D, at another point in time.

With a discrete choice model for resource selection, the ith choice is described by (i) the choice set of n_i resource units (habitat or food) that are available to be chosen; (ii) values for variables A_1, A_2, ..., A_u that characterize the animal making the choice (e.g., age, sex, etc.); and (iii) values for variables B_1, B_2, ..., B_v that characterize the resource units (e.g., vegetation type, elevation, etc. with habitat selection).

There is an extensive literature on discrete choice models in situations where it is humans making the choice (Manski, 1981; McCracken *et al*., 1998). In fact, there are several alternative models that have been proposed, including the nested logit, multinomial probit, and generalized extreme-value. However here only the most popular model will be considered, which is variously called polytomous regression, nominal regression and the multinomial logit model.

Suppose that at the ith choice there are n_i alternative resource units available to an animal, each of which is described by variables p variables X_1, X_2, ..., X_p, such that the values of these variables for the jth unit are $\mathbf{x}_{ij} = (x_{ij1}, x_{ij2}, ..., x_{ijp})$. Suppose also that the probability of the jth unit being selected for this ith choice is proportional to

$$w(\mathbf{x}_{ij}) = \exp(\beta_1 x_{ij1} + \beta_2 x_{ij2} + ... + \beta_p x_{ijp}), \qquad (8.1)$$

where β_1, β_2, ..., β_p are unknown parameters. Then the probability that the jth unit is the one selected will be equal to

$$p_{ij} = \exp(\beta_1 x_{ij1} + \beta_2 x_{ij2} + ... + \beta_p x_{ijp}) / \sum_{k=1}^{n_i} \exp(\beta_1 x_{ik1} + \beta_2 x_{ik2} + ... + \beta_p x_{ikp}). \quad (8.2)$$

Thus if the jth unit is the one that is selected at the ith choice, then the probability of this occurring is p_{ij}. This means that given S independent choices the probability of observing a particular series of selections is equal to the probabilities for the successful choices multiplied together, where the probabilities take the form shown in equation (8.2). This probability of the observed series of selections is the likelihood function. It can be written concisely by representing the outcome of the ith choice by a series of indicator variables y_{ij}, where $y_{ij} = 1$ if resource unit j is chosen for use, or is otherwise 0, for j = 1, 2, ..., n_i, in which case the likelihood function becomes

$$L = \prod_{i=1}^{S} (p_{i1})^{y_{i1}} (p_{i2})^{y_{i2}} ... (p_{in_i})^{y_{in_i}}, \quad (8.3)$$

where the p_{ij} values are given by equation (8.2).

Maximum likelihood estimates of the β parameters are obtained by maximizing L with respect to these parameters, with the standard theory providing estimated standard errors, tests of significance, etc. The estimated function

$$\hat{w}(\mathbf{x}_{ij}) = \exp(\hat{\beta}_1 x_{ij1} + \hat{\beta}_2 x_{ij2} + ... + \hat{\beta}_p x_{ijp}), \quad (8.4)$$

is then a resource selection function (RSF) because it gives relative probabilities of use for different types of resource units. Various computer programs are available for carrying out the estimation of the β parameters, with procedures that are variously described as multinomial logit regression, polytomous regression, nominal regression, etc. See Chapter 14 for more details. One special case that can be handled easily is where there are only two resource units involved in each choice. This is considered in the next section of this chapter.

The likelihood of equation (8.3) has been justified here on purely empirical grounds. However, McFadden (1973) derived the same likelihood in a choice behaviour setting by first defining a linear model for the utility of a resource unit and then assuming that error terms follow an extreme value distribution. With this approach the utility that an animal derives by selecting the jth unit in the ith choice set is assumed to be of the form

$$U_{ij} = \beta_1 x_{ij1} + \beta_2 x_{ij2} + ... + \beta_p x_{ijp} + \epsilon_{ij},$$

where ϵ_{ij} is an error term. Assuming the ϵ_{ij} are independent and identically distributed extreme value random variables, the probability of the jth unit being selected at the ith choice can then be shown to be given by equation (8.2). According to this approach, the utility of the jth unit is $\log_e\{w(\mathbf{x}_{ij})\}$.

According to the discrete choice model considered above the relative probability of selecting a unit with attributes $\mathbf{x}_{ij}$ over a unit with attributes $\mathbf{x}^*_{ij}$ does not depend upon the other available choices. This is the so-called 'independence of irrelevant alternatives' property (IIA). If this property is not appropriate for a particular situation then the model presented here can still be used, but it must be realized that the estimated RSF will change according to what is available for selection. On the other hand, if the IIA assumption is appropriate, then the RSF can potentially be transported to other areas because the relative preference for different resource units is constant. Obviously this consideration applies with all approaches for studying resource selection.

A useful idea from the theory of discrete choice models is the concept of the marginal rate of substitution (MRS) of one resource for another (Cooper and Millspaugh, 1999; Nicholson, 1990). This is the change in one attribute on a unit that is required in order to compensate for a one unit change in another attribute on the unit in order to leave the probability of selecting the unit unchanged.

For example, assume that the selection of plots of land for nest sites by a rare bird species is of interest, where the plots contain varying amounts of young and old forest. For one plot a proposed cutting of old forest will convert one hectare of old forest into one hectare of young forest. Then the MRS of old forest to young forest measures how much young forest must be added to that unit in order for the overall probability of birds selecting that unit to remain unchanged. If the RSF of equation (8.1) is

$$w = \exp\{1.50(\text{Old hectares}) + 0.75(\text{Young hectares})\},$$

then it can be seen that reducing the old forest hectares by one will require an increase of two hectares of young forest in order to keep w unchanged. The MRS of old forest for young forest is therefore 1.50/0.75 = 2.0 and it is found that two hectares of young forest (one more than the harvested hectare) would be necessary to offset every harvested hectare of old forest in order for the overall probability of selecting the unit to remain the same. In other words, two hectares of young forest is worth one hectare of old forest for the birds.

8.2 Two Units Per Choice Set

The special case of two choices per selection ($n_i = 2$ for all i) sometimes arises. It is particularly common in medical applications where it is referred to as a matched case-control study (Collet, 1991, Section 7.7). For this situation there is a trick for analysing the data using a program for logistic regression.

Suppose that the resource units are labelled so that the first one is selected for each of the S choices of one unit from a pair of units. Then the probability of this outcome from the ith choice is

$$p_i = \frac{\exp(\beta_1 x_{i11} + \beta_2 x_{i12} + \ldots + \beta_p x_{i1p})}{\exp(\beta_1 x_{i11} + \beta_2 x_{i12} + \ldots + \beta_p x_{i1p}) + \exp(\beta_1 x_{i21} + \beta_2 x_{i22} + \ldots + \beta_p x_{i2p})},$$

Dividing all terms through by the second exponential term in the denominator then produces

$$p_i = \frac{\exp\{\beta_1 (x_{i11}-x_{i21}) + \beta_2 (x_{i12}-x_{i22}) + \ldots + \beta_p (x_{i1p}-x_{i2p})\}}{1 + \exp\{\beta_1 (x_{i11}-x_{i21}) + \beta_2 (x_{i12}-x_{i22}) + \ldots + \beta_p (x_{i1p}-x_{i2p})\}}, \qquad (8.5)$$

which is the probability of a success for a logistic regression model where all the responses are successes, the explanatory variables are differences between the values for the first and second units in each choice pair, and there is no constant term β_0 in the argument of the exponential functions. In this case, standard logistic regression software can be used to estimate the coefficients and standard errors for the discrete choice model, providing that the missing constant term can be handled.

8.3 Examples

The discrete choice model can be used whenever one or more animals selects one resource unit from a set of possible units, with this selection being repeated on a number of occasions. The following two examples illustrate how such situations occur in practice. The first example is relatively straightforward with the choices being between four types of habitat by polar bears. As a result, a special iterative method can be used to estimate the parameters of the selection model. The second example is more complicated because it concerns the choice by pigeon guillemot seabirds of habitat units that are described by several variables.

Example 8.1 Selection of Sea Ice Habitat by Polar Bears

A discrete choice model was used by used by Arthur *et al.* (1996) to study the selection of habitat polar bears (*Ursus maritimus*) in a special case where the estimation of the parameters of the model can be carried out by a relatively simple iterative procedure.

The data available to Arthur *et al.* came from five adult female bears with radio-collars from a population that ranges over the sea ice in the northern Bering Sea, the Chukchi Sea, the western Beaufort Sea and the Eastern Siberian Sea, as shown in Figure 8.1. The positions of the bears were determined either every three days or every six days, which was judged to provide independent data because of the mobility of the bears and the speed with which the ice changes. Only observations in the Spring (May and June) or late Summer (September and October) were considered, for 1990.

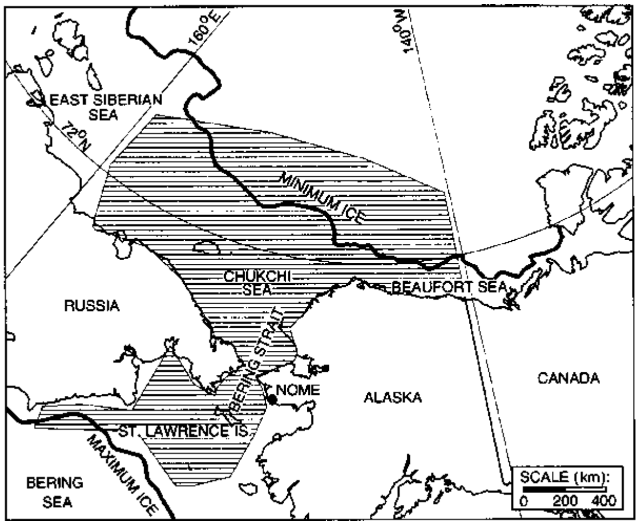


Figure 8.1 The shaded area was traversed by adult female polar bears monitored using radio-collars in the Bering and Chukchi Seas from 1986 to 1993. The minimum (October 1990) and maximum (February 1991) extents of the polar pack ice are also shown.

The basic assumption used by Arthur *et al*. was that given the position of a bear at one point in time, the habitat available to that bear for determining its position three days later was contained within a circle with radius 200 km centred at the first position, and that the habitat available to that bear for determining its position six days later was contained within a circle with radius 300 km centred at the first position. The size of these circles was a matter of judgement based on information on the movement of bears, and it was accepted that it was somewhat arbitrary.

Daily ice maps were obtained giving the percentage of ice within 25 by 25 km grid cells in the study region. It was therefore possible to calculate the areas within the available circles determined for each bear that had ice percentages in each of the four categories of 1-25%, 26-50%, 51-75%, and 76-100%, as illustrated in Figure 8.2.

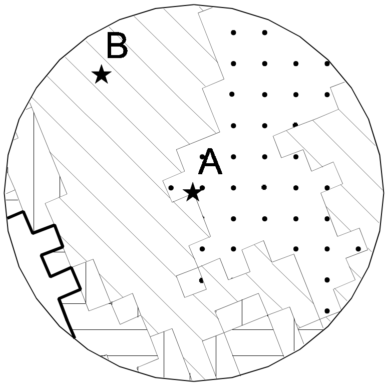


Figure 8.2 Illustration of the determination of the available habitat and the habitat chosen by a bear. The radius of the circle is the maximum distance a bear is likely to travel in three days (200 km). The patterns represent the ice categories of 1-25% (horizontal lines), 26-50% (vertical lines), 51-75% (diagonal lines), or 76-100% (dots), and the stars represent the locations of a bear on one day (A) and three days later (B). The heavy line represents the edge of the ice pack. Areas with less than 1% of ice (unshaded) and land areas were not considered to be available to the bears.

It was then assumed that the probability of a bear choosing to be in category j habitat at the ith relocation was given by

$$p_{ij} = A_{ij}w_j \Big/ \sum_{k=1}^{4} A_{ik}\, w_k, \tag{8.6}$$

where A_{ij} is the fraction of the available habitat in the jth category, and w_j is a selective value that allows for some habitats to be preferred to others. If $w_1 = w_2 = w_3 = w_4$ then

$$p_{ij} = A_{ij} / \sum A_{ik} = A_{ij},$$

because the A_{ij} values are fractions adding to one. Hence in this case the probability of choosing a particular habitat is equal to the fraction of the area that it covers, and there is no selection. Alternatively if, for example, w_j is much larger than the other selective values then habitat in category j is strongly selected for.

From equation (8.6) it can be seen that w_j is proportional to p_{ij}/A_{ij}, which is the probability that habitat j is used divided by the proportional availability of that habitat. It is therefore essentially the selection ratio of Chapter 4, which is sometimes called the forage ratio. However, estimation is more complicated when availability is changing.

Arthur *et al*. show that in general with S choices of habitat and H categories of habitat maximum likelihood estimates of the selection ratios are the values which make the total number of selections of habitat j equal to the sum of the probabilities of these selections. They also show that these estimates can be calculated by iterating on the two equations

$$\hat{w}_k = \sum_{i=1}^{S} o_{ik} / [\sum_{i=1}^{S} A_{ik} / \sum_{j=1}^{H} A_{ij} B_j], \qquad (8.7)$$

and

$$B_k = \hat{w}_k / \sum_{j=1}^{H} \hat{w}_j, \qquad (8.8)$$

where o_{ik} is one if habitat type k is chosen at the ith choice, or otherwise is zero. To begin with all the estimates are set equal to 1/H. Substituting into equation (8.8) then gives a new set of estimates for k from 1 to H. These are standardized to have a sum of one using equation (8.8). The standardized values are then substituted back into equation (8.8) to get new estimates for $\hat{w}_1$ to $\hat{w}_k$. This process continues until the estimates become constant from one iteration to the next.

Arthur *et al*. also provided equations for the variances and covariances of the estimators, and discussed a range of inference procedures for testing for selection, and for differences between the selection of different animals, and demonstrated some of their properties with a simulation study. All of these various methods were illustrated using data from 86 habitat choices in Spring and 68 habitat choices in late Summer for their five bears. They found strong evidence of selection in both seasons, but the selection was not the same in both seasons.

Although Arthur *et al*. did not use the discrete choice model to analyse their data, they could have. However, the usual form of the model given by equation (8.2) has to be modified to take into account the areas for different categories of habitat. To see this let $w_j = \exp(\beta_j)$. Then equation (8.6) becomes

$$p_{ij} = A_{ij} \exp(\beta_j) / \sum_{k=1}^{4} A_{ik} \exp(\beta_k),$$

which can also be written as

$$p_{ij} = A_{ij}\exp(\beta_1 x_{ij1} + \beta_2 x_{ij2} + \beta_3 x_{ij3} + \beta_4 x_{ij4}) / \sum_{k=1}^{4} A_{ik}\exp(\beta_1 x_{ik1} + \beta_2 x_{ik2} + \beta_3 x_{ik3} + \beta_4 x_{ik4}),$$

where $x_{ikr} = 1$ if k = r, or is otherwise 0. Apart from the weightings of the exponential terms terms by the fractions of different types of habitat, this probability takes the same form as equation (8.2), showing that the Arthur *et al*. model can be thought of as a variety of the discrete choice model. Furthermore, the likelihood function is still given by equation (8.3), and in principle maximum likelihood estimates of the β parameters (and hence the w parameters) can be obtained by maximizing this likelihood. In practice some special software would be needed to do the calculations.

There is another way that the discrete choice model could be used with this example without the complication of having to allow directly for the areas in different ice cover categories. The choice set could be thought of as the 25 by 25 km grid cells within the available circles for each bear. In a circle with radius 200 km for a three day relocation of a bear there would be 201 of these cells, less any classified as open water or land, while with a circle with radius 300 km for a six day relocation of a bear there would be 452 cells, less any classified as land or open water. The bear would then choose one of the cells for its relocation and the standard discrete choice model of equation (8.2) would apply with three indicator variables set up to describe the four habitat types. The only problem with this approach to the analysis is the handling of the large number of choices possible for each relocation. In practice, therefore, with an example like this it is easiest to use the iterative method of estimation proposed by Arthur *et al*., rather than the usual discrete choice model.

A valid criticism of the Arthur *et al*. analysis is the assumption that all parts of a circle centred at an animal's location at one point in time are equally available for the animal's next location. This issue was addressed by Hjermann (2000) who used a random walk model for the movement of animals to calculate the probability of a grid cell being used in the absence of selection, as a function of the distance of the cell from the initial location. Some approach along these lines does seem more realistic than assuming equal availability. However, the method used in the following example will be easier to use than the method proposed by Hjermann (2000).

Example 8.2 Foraging Habitat Selection by Pigeon Guillemots

The second example concerns seabird foraging site selection. The data were collected as part of a larger study conducted by Gregory Golet in conjunction with colleagues at the U.S. Fish and Wildlife Service office in Anchorage, Alaska. A fuller analysis will be published elsewhere (Gollet, personal communication).

The aim of the study was to identify marine habitat characteristics that influence foraging site selection of the pigeon guillemot (*Cepphus columba*), a semi-colonial pursuit-diving seabird. During the breeding season guillemots forage in near-shore ocean waters typically within 12 km of the colony. They typically fly over water to reach foraging locations where they dive in pursuit of small fish or crustaceans. A series of dives in one location is called a foraging bout, and guillemots often conduct several bouts before returning to the colony.

The data collected during the study consisted of the locations of the foraging bouts of three pigeon guillemots fitted with radio transmitters. The study was conducted during the chick provisioning stage, and the nests of the three birds were within 10 m of each other on Jackpot Island, located in south western Prince William Sound (Golet *et al*., 2000, 2002). Technicians in high-speed motorboats followed the radio-tagged birds on their foraging trips and recorded all locations where diving was observed. Approximately 30 foraging bouts per individual were recorded, with a total of 82 bouts (Figure 8.3). Foraging occurred primarily in bays and around islets to the north and west of the colony. Because the birds all nested at essentially the same location, each bird can be assumed to have been presented with the same foraging options. Thus data from this study may be used to test for differences in foraging patterns between the individual birds without having to control for differences in colony location. Due to the non-random nature of the choice of birds, and the small number of them, inferences from these three individuals to a larger population of pigeon guillemots is not considered to be realistic.

The study region was divided into 232 grid squares of 500m by 500m, each of which contained at least 50% of ocean water habitat. These grid squares became the resource units for the study. Foraging bout locations were plotted in a geographical information system (GIS) and moved to the centres of the grid cells that contained them. This association had the effect of moving the foraging locations a maximum of 1.414 x 250 = 354m. For each grid cell there were values for the average water depth and the forage fish density estimated from aerial surveys.

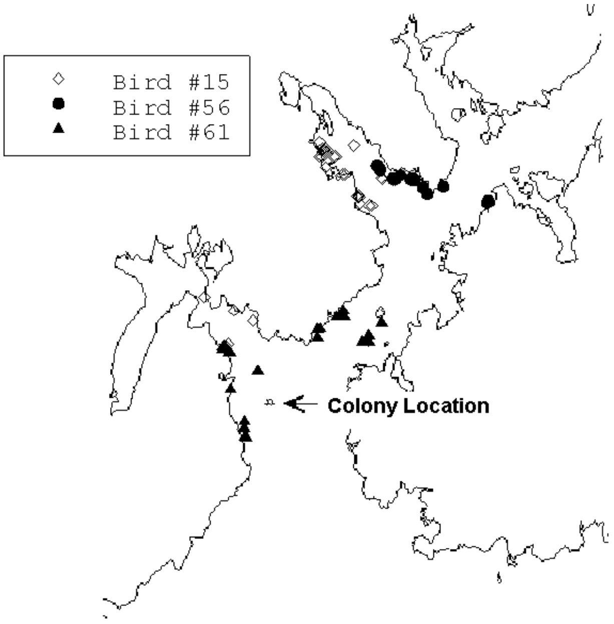


Figure 8.3 Locations of 82 foraging bouts conducted by three radio-collared pigeon guillemot birds in Prince William Sound, Alaska. Some points plot on top of one another.

Distances measured over the water from each grid cell to the observed foraging locations were also calculated to provide a distance from last foraging bout variable. For example, the shortest distances over water from the centre of the cell containing foraging bout number one to the centres of all the 232 grid cells were calculated and used as a covariate for the location of foraging bout number two. Similarly, the distance of each cell from foraging location number two was used as a variable for foraging location number three, and so on except that for the first foraging location the distance from the colony was used for this variable. This distance from the last foraging bout variable was included in the analysis because it seemed reasonable to assume that locations far away from a bird's current location would have a smaller relative probability of selection than locations nearer to a bird's current location. In effect this is an alternative to Hjermann's (2000) method for allowing the availability of resource units to decrease with distance, that is relatively easy to use.

Including the distance from the last foraging bout as a variable describing the resource units had the effect of changing the attributes of available resource units for each choice of a location for a foraging bout. For this reason, the discrete choice model was used to model the resource selection of these birds.

Each choice set for the discrete choice model consisted of all 232 grid cells with their associated values fot the forage fish density, the water depth, and the distance from the previous bout. The first 10 rows of data from the much larger data set containing 82 x 232 = 19,024 rows is shown in Table 8.1, with the full data set being available at the web site www.west-inc.com.

Table 8.1 The first ten rows of data from a data set use in the pigeon guillemot example. All distances are measured over the water (i.e., the measured path is diverted around land masses). The full data set has 19,024 rows.

Bird	Bout	Grid point	Grid column	Grid row	Albers x Coordinate	Albers y Coordinate	Distance to last bout (m)	Distance to colony (m)	Water depth (m)	Forage fish density	Used
56	2	1	27	7	320235	1171146	4536	11778	10	0.000	0
56	2	2	28	7	320735	1171146	4036	11278	10	0.000	0
56	2	3	29	7	321235	1171146	3828	11071	10	0.000	0
56	2	4	35	7	324235	1171146	3828	11899	50	2.910	0
56	2	5	36	7	324735	1171146	3621	11692	30	2.541	0
56	2	6	37	7	325235	1171146	3414	11485	10	2.541	0
56	2	7	29	8	321235	1170646	3328	10571	10	0.360	0
56	2	8	35	8	324235	1170646	3328	11399	50	0.000	0
56	2	9	36	8	324735	1170646	3121	11192	10	0.955	0
56	2	10	37	8	325235	1170646	2914	10985	10	0.955	0

From radio tracking records and direct observation of the foraging bouts, it was determined that two of the birds fed predominantly on benthic fish and the other bird fed predominantly on herring. While this difference in prey type clearly represented a difference in the selection behaviour between the two groups of birds, it was also of interest to determine whether a difference in the selection for travel distances existed that could be associated with prey type. To examine this possibility, the coefficients of the variables describing the resource units were initially allowed to vary according to the prey type in the discrete choice model through the introduction of interaction terms. If interactions were not significant, they were dropped and a common coefficient was assumed for both prey types.

The results of testing for differences in the coefficients of the distance to the last bout, the water depth, and the forage fish density between the food type groups supported the use of the full model with interactions for the forage site selection. There was a significantly non-zero interaction of the distance to last bout and the food type (p=0.023), as well as the interaction of depth and feed type (p=0.001). The test for the equality of forage fish density selection for the two food type groups was nearly significant at the 5% level (p=0.057) and this interaction was retained because the difference in selection of was deemed to be biologically important.

The form of the final model was:

$$w(x) = \exp\{ - 0.855(\mathrm{DIST}) - 0.456(\mathrm{DIST}\times I_{\mathrm{BEN}}) + 0.007(\mathrm{DEPTH})$$

$$+ 0.032(\mathrm{DEPTH}\times I_{\mathrm{BEN}}) + 0.262(\mathrm{FFD}) - 0.497(\mathrm{FFD}\times I_{\mathrm{BEN}}))$$

where DISTt is the over water distance from the previous foraging bout in km, DEPTH is the water depth at the foraging site in m, FFD is the forage fish density, and I_{BEN} is an indicator function equal to 1 if the bird making the choice fed on benthic fish, and 0 otherwise. The standard errors for the coefficients were 0.136, 0.200, 0.005, 0.010, 0.195, 0.260 for DIST, DIST×I_{BEN}, DEPTH, DEPTH×I_{BEN}, FFD, and FFD×I_{BEN}, respectively. Note that I_{BEN} does not appear in the model by itself because it was constant within each choice set.

By substituting 0's and 1's into this model in place of the prey-type indicator variable, factoring terms, and adding coefficients, the following models were found to be estimated for the two feeding groups:

benthic feeding birds

$$w(x) = \exp\{ - 1.311(\mathrm{DIST}) + 0.038(\mathrm{DEPTH}) - 0.235(\mathrm{FFD})\},$$

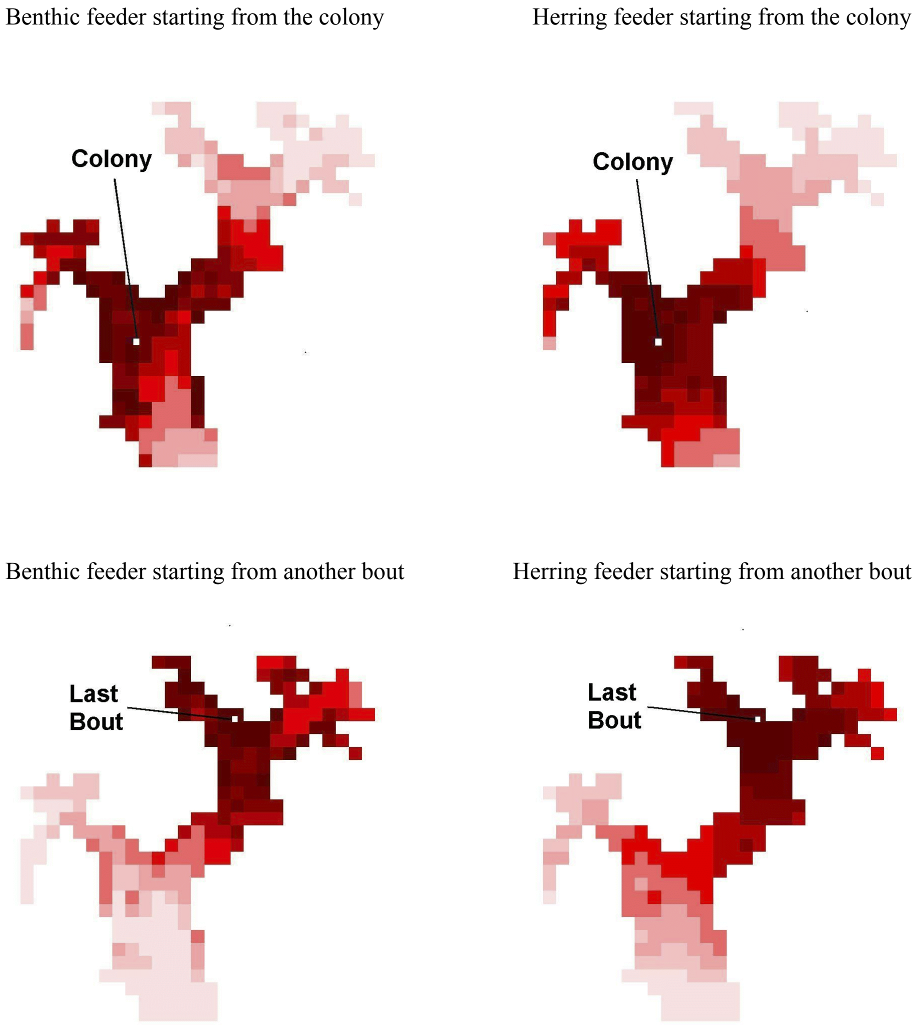
herring feeding bird

$$w(x) = \exp\{ - 0.855(\mathrm{DIST}) + 0.007(\mathrm{DEPTH}) + 0.262(\mathrm{FFD})\}.$$

The standard errors for the benthic model are 0.147 for DIST, 0.008 for DEPTH, and 0.172 for FFD, while the standard errors for the herring model are 0.136 for DIST, 0.005 for DEPTH, and 0.195 for FFD.

This final model indicates significant selection for short distances between bouts in both prey type groups. The model also indicates significant selection for shallow water depths in both prey type groups. However, benthic feeding birds displayed stronger selection for short distances between bouts and for shallow water than is the case for herring feeding birds. The selection of grid cells with high forage fish densities differed by feeding group. Benthic feeding birds displayed selection against areas of high forage fish densities, while herring feeding birds displayed selection for areas of high forage fish density.

The estimated relative probabilities of each bird selecting one of the 232 grid cells assuming the bird starts from the colony, or starts from a foraging bout in the north are displayed in Figure 8.2. This figure illustrates how the inclusion of distance in the model means that the relative probability of selection is high only for cells that are close to the starting point.


Benthic feeder starting from the colony

Herring feeder starting from the colony

Benthic feeder starting from another bout

Herring feeder starting from another bout

Figure 8.4 Estimated relative probabilities of selecting grid cells for a bird starting at the colony, either a benthic feeder or a herring feeder, and for a bird starting from a foraging bout in the north, either a benthic or herring feeder. Dark shades represent cells with relatively high probabilities of selection.

8.4 Availability Sampled

The theory provided above and the examples just considered refer to situations where the resource choices available are completely enumerated. For example with the pigeon guillemot example the choice was always one of the 232 grid cells. However, situations arise where there are so many possible choices that it is not practical to use every one in an analysis. Instead, a resource unit chosen by an animal must be compared with a sample from the population of available units. This will frequently be the case with studies based on GIS databases where there may be millions of pixels forming the set of locations where an animal might choose to be.

One approach that can be used when available units are sampled is based on a conditional probability argument similar to ones used in Chapters 5 and 6. Thus suppose that an animal makes a choice of one of N available resource units, and n other available units are randomly selected for comparison with the used unit. It is then possible to consider the probability that out of the set of n + 1 units that are observed, it is a particular one that is used.

As before, let each of the n + 1 units be described by their values for variables X_1 to X_p, and assume that the probability of the ith of these units being selected out of the full set of N units is

$$w(\mathbf{x}_i) = \exp(\beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip})\ \gamma,$$

where γ is the constant that ensures that the probabilities of selection add to exactly one for all N units. Then the probability of obtaining the set of n + 1 units analysed is the sum of the probability that the ith of these units is used, times the probability that the other n units appear in the available sample. The latter probability is just the reciprocal of the total number of samples of size n that can be selected from the N - 1 unused units. Hence the probability of obtaining the observed units for analysis is

$$\text{P(units analysed)} = \sum_{i=1}^{n+1} w(\mathbf{x}_i)\ \{n!(N - 1 - n)!/(N - 1)!\}\ .$$

The conditional probability that the jth of the n + 1 units is the used one, given that it is one of the n + 1 units is therefore

$$p_j = w(\mathbf{x}_i)\ \{n!(N - 1 - n)!/(N - 1)!\}/\text{P(units analysed)},$$

which reduces to

$$p_j = \frac{\exp(\beta_1 x_{j1} + \beta_2 x_{j2} + ... + \beta_p x_{jp})}{\sum \exp(\beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip})}, \qquad (8.9)$$

where the sum in the denominator is for i from 1 to n + 1. This is exactly the same form as equation (8.2) for discrete choice models in general, with the likelihood function for S choices taking the form of equation (8.3). Consequently, as far as estimating the RSF

$$w(\mathbf{x}) = \exp(\beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p)$$

is concerned, it is immaterial whether all possible available units are used, or just a random sample of them.. Note, however, that in order for the results of S selections to be independent it is necessary for a fresh sample of available units to be used for each of the selections.

The argument provided here is not the only one that can be used to derive a likelihood function for estimation with a sample of used units. McCracken *et al*. (1998) proposed using the likelihood function for the full set of N available units, with part of this estimated from the data, while Collett (1991, Appendix B2) uses a different approach to derive equation (8.9) in the context of case-control studies in medicine.

Chapter Summary

- The theory of discrete choice models is reviewed as an approach for estimating a resource selection function when each choice by an animal or group of animals involves a different set of available resource units.

- The special case where each selection involves choosing one of a pair of available units is considered because estimation of a resource selection function can then be reduced to fitting a logistic regression function with no constant term.

- Two examples are presented, with one involving the selection of habitat with different percentages of sea ice by polar bears in the Arctic, and the other involving the selection of foraging habitat by pigeon guillemot seabirds.

- Situations are considered where the characteristics of a selected unit are compared with the characteristics of a sample of used units. This is necessary sometimes because of the huge number of available units that exist. It is argued that the discrete choice model can be used unchanged with these types of situation.

Exercises

(1) Arthur *et al*. (1996) illustrated their iterative method for estimating selective values using the data shown in Table 8.2 for the selection of ice cover categories by an adult female polar bear. Using a spreadsheet program or otherwise, use these data to estimate the standardized selection values B_1 to B_4 for the choice of ice categories 1-25% to 76-100% by this bear. Use equations (8.7 and (8.8) for the iterations, as explained in Example 8.1.

(2) The data analysed by Arthur *et al*. (1996) involved classifying the habitat into four level of ice cover, and measuring the amount of each class available. An alternative procedure would have been to record the ice cover in the pixel where a polar bear was located and compare this with a sample of pixels chosen randomly from those available to the bear (based on a circle centred where the bear was previously located). The situation would then be as discussed in Section 8.4. Suppose that this alternative procedure was carried out and resulted in the (artificial) data shown in Table 8.3. Use the discrete choice model to estimate a RSF

$$w(x) = \exp(\beta_0 + \beta_1 x),$$

giving the relative probability of selecting a pixel with a percentage x of ice cover.

Table 8.2 Use and availability of four categories of ice cover habitat (1, 1-25%; 2, 26-50%; 3, 51-75%; and 4, 76-100%) for an adult female polar bear on seven days during September and October, 1990.

	Habitat use by bear (o_{ij})[1]				Habitat availability (A_{ij})[2]			
Date	1	2	3	4	1	2	3	4
8 Sep	0	1	0	0	0.061	0.109	0.139	0.691
14 Sep	0	1	0	0	0.064	0.138	0.477	0.321
17 Sep	1	0	0	0	0.076	0.197	0.386	0.341
5 Oct	0	0	0	1	0.091	0.101	0.114	0.694
17 Oct	0	0	0	1	0.249	0.138	0.140	0.474
20 Oct	0	0	1	0	0.044	0.188	0.160	0.608
23 Oct	0	0	1	0	0.037	0.080	0.289	0.595

[1] The value of o_{ij} is 1 if at the ith choice the ice cover class j is chosen for use, or is otherwise 0.
[2] The value of A_{ij} is the fraction of the available habitat that is in ice cover class j when the ith choice is made.

Table 8.3 Artificial data from a situation where the percentage ice cover in a pixel used by a female polar bear is compared with the ice cover in a random sample of 20 pixels selected from those in the area where the polar bear could be located, i.e. available pixels. The location of the bear is determined seven times.

	Percentage ice cover in available sample[1]																			
Used[2]	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
48	59	71	93	44	94	52	94	82	49	84	78	92	83	79	3	99	90	45	80	76
49	67	61	1	54	87	92	63	88	51	67	44	73	84	95	76	4	37	1	98	74
24	44	38	73	48	86	68	68	64	53	50	87	90	89	70	36	37	27	48	89	85
85	89	69	94	99	75	78	90	90	92	85	79	98	59	92	7	80	53	52	91	84
87	99	23	12	88	37	10	77	77	92	59	75	75	37	50	37	83	74	44	9	53
71	46	80	99	81	92	98	82	84	48	25	52	58	2	31	60	78	74	32	64	74
63	61	94	91	89	53	67	54	79	0	97	25	81	88	62	82	94	88	98	80	74

[1] For each of the seven locations of the bear (the rows in the body of the table) a new random sample of 20 available units is selected.
[2] The percentage of ice cover for the pixel used by the bear.

CHAPTER 9

APPLICATIONS WITH GEOGRAPHIC INFORMATION SYSTEMS

Geographic information systems (GIS) have become an integral part of many resource selection applications. Information from the GIS on habitat characteristics such as the topography and vegetation classifications are often utilized in the development of variables for modelling resource selection. Radio and geographical positioning system (GPS) equipment provide accurate information on the location of wildlife on the landscape. These locations are combined with the GIS for the development of habitat variables for the used resources. In this chapter the issues associated with the use of GIS in modelling resource selection are discussed. These include the definition of use and availability, the initial selection of variables for use in a resource selection function (RSF), model selection using these variables, and the mapping of relative probabilities of use. Two case studies involving habitat selection by moose and brown bears are used as illustrations.

9.1 Defining Use and Availability

Habitat use in GIS examples is typically defined by the habitat characteristics at the locations of the point locations of the animal or animals under study. The coordinates of these locations are typically imported into the GIS and values for variables are derived at these locations to give the sample values for the used resource units, which might be the pixels within which the locations occurred. The distance to landscape features such as water or vegetation cover can also be calculated from the coordinates and GIS. Area based variables are also commonly used. For example, percentages of vegetation cover classes within circular buffers centred at the used locations can be calculated and compared to a large sample of similar buffers in the area that is considered to be available for use. In some applications such as those involving nest site selection, use may be defined at a point or on an area basis. Sometimes points and areas are used in the same study (e.g., Marzluff *et al*., 1997).

When used resource units are defined on an area basis the available units should be defined in the same way. For example, Erickson *et al*. (1998) compared the proportion of resources within buffers surrounding moose locations to the proportions in similar available buffers systematically selected across the study area. This comparability of used and available units is obviously particularly crucial for variables that depend on the size of the units, such as vegetation diversity, richness and patchiness.

The appropriate size and shape of the area defining use depends on the scale of the study, patch sizes, the types of variable being used to describe habitat, the biology of the animal under study, and the magnitude of likely errors in the locations of animals.

If home range selection is of interest, then the obvious choice of the area to use is one based on a standard estimator of the home range size. With variables that depend on the area surveyed such as the vegetation diversity, samples of average home ranges may be used to define availability. When small buffers around relocations are considered, then it is best to use a buffer size greater than the average relocation error. It is also valuable to try using several different buffer sizes to see how this changes the results obtained. Defining use on an area basis may be preferable because of location errors and because selection often occurs on a mosaic of habitats (Rettie and McCloughlin, 1999).

The difficulties in defining available habitats are common to most resource selection studies. These are present with GIS studies, but the use of a GIS does allow flexibility in evaluating the robustness of the availability definition using different choices.

9.2 Variable and Model Selection

When a GIS is available it is very easy to develop many possible variables that might be useful in a RSF. Although this may seem to be an advantage over situations where all the data must be collected in the field, it does mean that model selection may become very difficult. Often ten or more variables may be considered, in which case there are 1023 possible combinations of variables giving alternative models. Therefore most GIS resource selection studies should be considered exploratory, because they are observational studies rather than manipulative experiments and a large set potential models are considered. Model selection procedures are not robust when a large number of variables are considered, and the opportunity for spurious results increases with the number of variables used. Therefore a candidate set of variables and models should be defined before the analysis begins, where these are chosen on biological grounds. Graphical displays of individual variables and tests comparing distributions for used and available resource units can help reduce the number of candidate variables.

9.3 Mapping of Resource Selection Functions

When a final estimated RSF is derived, having a GIS allows this function to be mapped for the entire study area. Examples of these types of maps have already been given in Figure 8.4, although in practice colour would be used. In most applications only relative probabilities rather than absolute probabilities can be used for this purpose. As is the case with RSFs in general, the maps that predict high probability of use for certain areas do not necessarily define optimal habitats.

9.4 Examples

The following two examples are typical of studies based on GIS. They differ because the first study involved the locations of moose being detected from aerial surveys, whereas the second involved use being detected from relocations of radio-tagged bears. Also, with the first study the habitat was just described by a set of land cover categories but with the second study more sophisticated measures of the habitat were used.

Example 9.1 Habitat Selection by Moose in the Innoko National Wildlife Refuge

The first example has already been briefly discussed in Example 3.6. It comes from a study reported by Erickson *et al*. (1998) of moose (*Alces alces*) winter habitat selection on the northern half of the Innoko National Wildlife Refuge in west central Alaska in 1994. Transects within the river corridors were flown by helicopter with a saw-tooth path. In an attempt to minimize the bias associated with varying visibility in different land cover types, the speed of the helicopter was varied depending on the habitat. The survey transect positions were recorded on a computer interfaced with a GPS unit when each moose group was perpendicular to the flight line of the helicopter, and the positions of moose groups were offset from the flight line by the estimated perpendicular distance from the flight line to the group. Errors in moose group positions were judged to be less than 100m. This example is specifically concerned with the resource selection model that was developed for the Innoko River corridor, an area of about 500 square miles where an estimated half of the north half of the refuge moose population reside during March.

A land cover base map was developed prior to the aerial surveys. The entire study area was divided up into 30m by 30m pixels, with each pixel classified into one of 22 land cover classes. Erickson *et al*. (1998) report an analysis of the results for the study based on using a circle with radius 400m around each moose group to describe a used resource unit. Each of these circles contained approximately 560 pixels, so that the resource unit could be described by the proportions of each of the 22 land cover classes within it.

The population of available resource units was considered to be circles of radius 400m around every pixel in the entire study area. As this population was very large a closely spaced systematic sample of pixels with associated 400m circles was selected and used to define the available units. Approximately every 13th pixel was selected, which resulted in overlapping 400m circles covering the entire study area. Because of the overlapping units, the proportions in the different land cover classes were practically identical for the systematic sample and the full population of units.

Standard model selection techniques were employed to select a parsimonious set of variables to adequately describe the selection of habitat by the moose. Any of the 22 land cover proportions habitat classifications were candidates for inclusion in a RSF providing that they covered 1.5% or more of the total study area. This eliminated land cover classes which were present but very rare in the study area, and also in the used sample. Quadratic terms for some of the variables were also considered based on a judgment that these may be important.

The study has design I (selection by a population of animals), with sampling protocol A (available units and used units separately sampled) according to the classifications discussed in Section 1.3. Therefore logistic regression was used to estimate a RSF using the methods described in Section 5.4. Best subset regression techniques (SAS Institute 1999) with AIC as the objective criterion (Akaike 1973) was used to determine the variables for inclusion in the RSF.

The final estimated RSF was

$$\hat{w} = \exp\{10.3(P_4) + 5.3(P_5) - 4.2(P_6) + 3.4(P_8) - 8.6(P_{10}) + 4.9(P_{17}) + 8.2(P_{22}) \\ -10.9(P_4)^2 - 9.3(P_{22})^2\}$$

where P_i is the proportion in the ith land cover class as described in Tables 1 and 2 of Erickson *et al*. (1998). Standard errors are reported in Table 9.1. The variables P_4, P_8 and P_{22} with positive coefficients are lowland categories that include willow that are important food sources for the moose.

Table 9.1. Coefficients of the estimated RSF for selection of habitat by bears in the Innoko River corridor, with estimated standard errors (Std. err.).

Variable	Coefficient	Std. err.
P4	10.32	2.77
P5	5.33	1.85
P6	-4.16	3.61
P8	3.37	1.05
P10	-8.61	3.58
P17	4.91	1.18
P22	8.24	3.17
$P4^2$	-10.94	4.73
$P22^2$	-9.33	7.18

Following model selection and estimation, maps of the study area were generated, which portrayed the predicted relative probability of selection for every possible resource unit (400m circle) the study area. For mapping, the predicted values of the RSF were classified as low selection (1st - 40th percentile of the distribution), moderate selection (41st - 70th percentile of the distribution), and high selection (71st - 100th percentile of the distribution). The result is shown in Figure 9.1. Areas of relatively high probability of use occur mainly along the river, where willow exist.

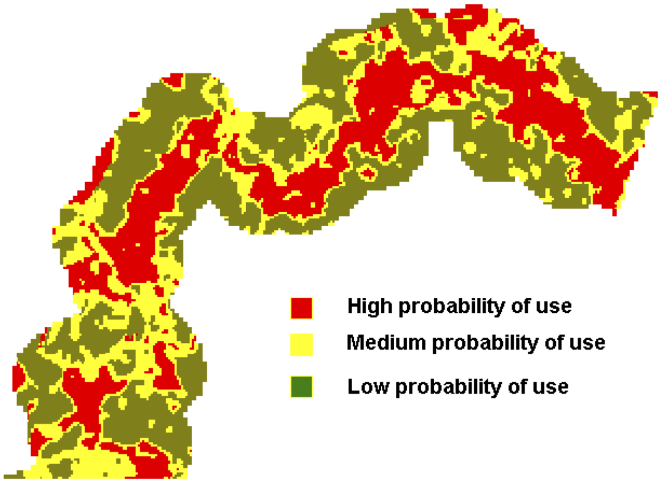


Figure 9.1 Results of the Innoko River corridor resource selection study. The map shows the relative probabilities of selection for each pixel, assuming free and equal access to all pixels by all animals.

Example 9.2 Habitat Selection by Brown Bears on the Kenai Peninsula, Alaska.

This example comes from a radiotelemetry study of 25 female brown bears (*Ursus arctos horribilis*) without cubs during the summer season from 1993-1996 on the Kenai Peninsula, Alaska. Bears with a minimum of 20 relocations during this time frame were used. Data were pooled across the years, because the variables that were considered for describing the habitat did not vary by the year.

The entire Kenai Peninsula was considered available except for permanent ice and glaciers. A systematic sample of 14,063 pixels was used to describe the available habitat and a GIS was used to calculate several continuous variables that were identified as potential predictors of selection by the bears. Initially, t-tests were carried out to compare variable means for each bear with the corresponding means for the entire study area, with the sample size for these tests being the 25 bears. Any variable found to give a result significant at the 10% level was retained at that stage. Further reductions in the variables to be considered set were made by eliminating some variables that were highly correlated with other variables. Thus if the correlation were 0.70 or more for a given pair of variables then only the one judged to be most appropriate was retained.

The study has design II (used units were determined for individual animals but availability was assumed to be the same for all animals) with sampling protocol A (used and available units sampled separately). Logistic regression was used to estimate a RSF following the method discussed in Section 5.4, with bear locations providing the used data and a random selection of points from the entire study area providing the available data.

The final RSF model was determined by a backward elimination stepwise procedure using jackknifing (Manly *et al.*, 1997), although other procedures could have been employed. Jackknifing was used because of computer memory limitations. A model was fitted with every bear in the data set, and also with each bear sequentially removed from the data set. Jackknife pseudo-values for regression coefficients were obtained for each variable as

$$b_j = n\bar{b} - (n - 1)\,\bar{b}_{-j},$$

where n is the number of bears, $\bar{b}$ is the regression coefficient calculated using the data from all of the bears, and $\bar{b}_{-j}$ is the regression coefficient calculated without using the data from bear j. The jackknife estimate of the regression coefficient was found by averaging the pseudo-values, and a t-test was carried out to see whether this was significantly different from zero at the 5% level. If one or more variables had a jackknife estimate of its regression coefficient that was not significantly different from zero then the variable with the least significant coefficient was removed from the equation. This process continued until every variable left in the equation had a regression coefficient that was significant at the 5% level.

The results of t-test are shown in Table 9.2, the final estimated RSF is summarized in Table 9.3, and a map of the values of the RSF is depicted in Figure 9.2. The final estimated RSF is

$$\hat{w} = \exp\{-0.358 - 0.483(\text{COVER}) - 0.431(\text{DEV2}) - 0.0142(\text{KESTM1}) \\ + 0.0034(\text{KESTM2}) + 0.0013(\text{SSTML2})\}.$$

In summary, the locations of the bears included in the study (females in the summer season with no cubs) were associated with short distances to cover (COVER), a low density of development (DEV2), and the presence of salmon streams (KESTM1, KESTM2 and SSTML2).

Table 9.2. Initial t-tests to screen out unimportant variables for the study of resource selection by brown bears.

Variable	Description	Mean Used	SE of mean	Mean Available	t-value	P-Value
COVER	Distance to cover[1]	0.36	0.03	0.60	-7.34	< 0.001
DEV1	Distance to human development[1]	69.44	5.11	63.55	1.15	0.260
DEN2	Density of human development per km^2	0.03	0.01	0.27	-22.10	<0.001
ELEV	Elevation (m)	315.53	34.51	373.12	-1.67	0.108
HROAD	Density of high use roads[2]	8.82	3.70	15.82	-1.89	0.070
KESTM1	Distance to potential salmon spawning streams[1]	12.22	1.52	21.83	-6.30	< 0.001
KESTM2	Density of potential salmon spawning streams[3]	153.02	16.58	28.30	7.52	< 0.001
LCOVVAR	Number of landcover classes within 100m	3.15	0.08	3.00	1.90	0.069
SLAKES	Distance to lakeshore[1]	201.72	31.24	294.00	-2.95	0.007
SSTML1	Distance to low potential salmon streams[1]	28.70	3.05	36.90	-2.69	0.013
SSTML2	Density of high potential salmon streams[2]	112.29	13.62	26.31	6.31	0.066

[1] Units are 100m .
[2] Miles per km^2.
[3] Number per hectare.

Table 9.3. Estimated coefficients of the RSF for the final brown bear model from the jackknife procedure.

Variable	Coefficient	Std. Err.	t-value	P-value
Constant	-0.3584	0.25847	-1.39	0.178
COVER	-0.4834	0.19453	-2.49	0.020
DEV2	-0.4310	0.17745	-2.43	0.023
KESTM1	-0.0142	0.00579	-2.46	0.021
KESTM2	0.0034	0.00039	8.81	< 0.001
SSTML2	0.0013	0.00026	5.22	< 0.001

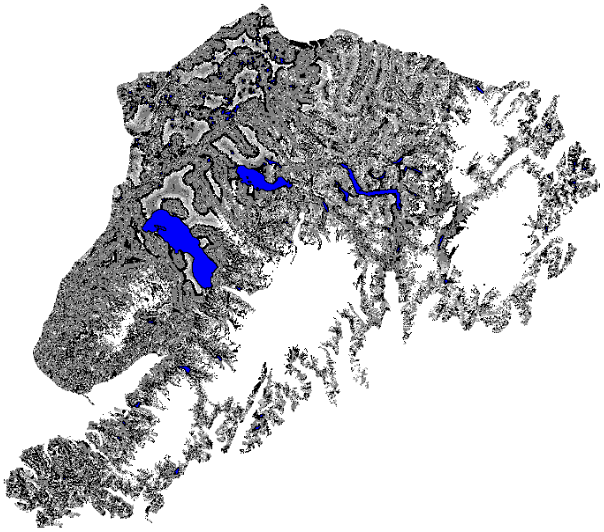


Figure 9.2 The Kenai Peninsula map of values from the estimated resource selection function, with darker areas being having a higher probability of selection than lighter areas.

Chapter Summary

- The definitions of used and available resource units are discussed for situations where these resource units are described by variables calculated using a geographical information system.

- Variable selection methods are considered, noting that the large number of variables that can be calculated using a geographical information system may make this a difficult process in which judgement should play an important part.

- The fact that the results of fitting a resource selection function can be mapped is noted as being an important advantage of using a geographical information system.

- Two examples are presented to illustrate typical applications using a geographical information system. These relate to habitat selection by moose in Innoko National Wildlife Refuge in Alaska, and by brown bears on the Kenai Peninsular, also in Alaska.

CHAPTER 10

DISCRIMINANT FUNCTION ANALYSIS

When a study of resource selection involves the collection of only two samples of normally distributed data then it is possible to estimate a resource selection function using discriminant function analysis. This approach to estimation is discussed in this chapter.

10.1 Linear Discriminant Function Analysis

Linear discriminant function analysis can be thought of as a way to estimate a resource selection function (RSF) for cases where the variables that are measured on resource units have multivariate normal distributions for the populations that are sampled. To see this, it is useful to begin by reviewing discriminant function analysis in terms of selection on a population.

Suppose that a population has a p variable multivariate normal distribution with a mean vector $\boldsymbol{\mu}_1$ and covariance matrix $\boldsymbol{\Sigma}$, and that the members of the population are selected to form a second population in such a way that the probability of an individual with the values $\mathbf{x}' = (x_1,x_2,...,x_p)$ being selected takes the form

$$\begin{aligned}\Omega(\mathbf{x}) &= \exp(\beta_0 + \beta_1 x_1 + ... + \beta_p x_p) \\ &= \exp(\beta_0 + \mathbf{x}'\,\boldsymbol{\beta}), \qquad (10.1)\end{aligned}$$

where $\boldsymbol{\beta}' = (\beta_1,\beta_2,...,\beta_p)$. Then the second population will also have a multivariate normal distribution, with the vector of means being

$$\boldsymbol{\mu}_2 = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}\ \mathbf{ß}$$

and with the covariance matrix still equal to $\boldsymbol{\Sigma}$ (Manly, 1985, p. 68). It follows that the ß values other than $ß_0$ in equation (10.1) are given by the equation

$$\mathbf{ß} = \boldsymbol{\Sigma}^{-1}\ (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1), \qquad (10.2)$$

which means that they are also the coefficients in Fisher's linear discriminant function, which is

$$L(\mathbf{x}) = ß_1 x_1 + ... + ß_p x_p \qquad (10.3)$$

(Seber, 1984, p. 109).

The value of β_0 can be related to the proportion of the first population that is selected to become part of the second population. This is the expected (mean) value of $\Omega(\mathbf{x})$ in the first population which is

$$E\{\Omega(\mathbf{x})\} = \exp(\beta_0)\, E\{\exp(\boldsymbol{\beta}'\mathbf{x})\},$$

from equation (10.1). Here $E\{\exp(\boldsymbol{\beta}'\mathbf{x})\}$ is the moment generating function of the first multivariate normal distribution with mean $\boldsymbol{\mu}_1$ and covariance matrix $\boldsymbol{\Sigma}$, which is well known to be $\exp(\boldsymbol{\beta}'\, \boldsymbol{\mu}_1 + \frac{1}{2}\, \boldsymbol{\beta}'\, \boldsymbol{\Sigma}\, \boldsymbol{\beta})$. Hence

$$E\{\Omega(\mathbf{x})\} = \exp(\beta_0).\exp(\boldsymbol{\beta}'\, \boldsymbol{\mu}_1 + \tfrac{1}{2}\, \boldsymbol{\beta}'\, \boldsymbol{\Sigma}\, \boldsymbol{\beta}),$$

from which it follows that

$$\exp(\beta_0) = E\{\Omega(\mathbf{x})\}\exp(-\boldsymbol{\beta}'\, \boldsymbol{\mu}_1 - \tfrac{1}{2}\, \boldsymbol{\beta}'\, \boldsymbol{\Sigma}\, \boldsymbol{\beta}) = E\{\Omega(\mathbf{x})\}\exp\{-\tfrac{1}{2}\, \boldsymbol{\beta}'\, (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\},$$

using equation (10.2).

Substituting into equation (10.1) then produces

$$\Omega(\mathbf{x}) = E\{\Omega(\mathbf{x})\}\exp\{\boldsymbol{\beta}'(\mathbf{x} - \tfrac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\}. \tag{10.4}$$

Equations (10.2) and (10.4) can form the basis of a method for estimating the selection function $\Omega(\mathbf{x})$. If random samples are taken from the two populations to provide sample mean vectors $\hat{\boldsymbol{\mu}}_1$ and $\hat{\boldsymbol{\mu}}_2$, and a sample pooled covariance matrix $\hat{\boldsymbol{\Sigma}}$ then the vector $\boldsymbol{\beta}$ can be estimated by

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\Sigma}}^{-1}\, (\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1). \tag{10.5}$$

and the selection function by

$$\hat{\Omega}(\mathbf{x}) = E\{\Omega(\mathbf{x})\}\, \exp\{\hat{\boldsymbol{\beta}}'\, (\mathbf{x} - \tfrac{1}{2}\, (\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2)\}. \tag{10.6}$$

Obviously, if $E\{\Omega(\mathbf{x})\}$ is not known then it will only be possible to estimate $\Omega(\mathbf{x})$ multiplied by an arbitrary multiplicative scaling.

An approximation to the covariance matrix for the estimators $\hat{\beta}_1$ to $\hat{\beta}_p$ (which assumes that $\hat{\boldsymbol{\Sigma}}$is error free) is

$$\hat{\boldsymbol{\Sigma}}^{-1}\, (1/n_1 + 1/n_2),$$

where n_i is the size of sample i (Manly, 1985, p. 68). This will tend to be an underestimate, and bootstrapping is probably a better method for determining the covariance matrix, as discussed in Section 2.11.

Equations (10.5) and (10.6) can be used with any two multivariate normal distributions without the requirement that one of the populations is obtained from the other one by some selection process if $\Omega(\mathbf{x})$ is interpreted as the selection function that is required to change the first population into the second population. More generally, $\Omega(\mathbf{x})$ can be thought of as the function which says how many individuals there are in the second population with $\mathbf{X} = \mathbf{x}$ for each individual with these measurement in the first population. From this point of view, the function does not necessarily have to produce probabilities. In fact, there is no reason why values greater than one should be prohibited.

As with logistic regression (Chapter 5), there are three situations that need to be considered in the context of estimating a RSF from two samples of resource units, corresponding to the samples being (a) a sample of available resource units and a sample of used resource units, (b) a sample of available resource units and a sample of unused resource units, and (c) a sample of unused resource units and a sample of used resource units. These three cases will now be considered in turn.

10.2 Samples of Available and Used Resource Units

Equations (10.5) and (10.6) apply immediately to the estimation of a resource selection probability function (RSPF) from a sample of available units and a sample of used units. Sample 1, with the mean vector $\hat{\boldsymbol{\mu}}_1$, should be the available sample, and sample 2, with mean vector $\hat{\boldsymbol{\mu}}_2$, should be the used sample. The selection function $\Omega(\mathbf{x})$ then has the same form as the RSPF $w^*(\mathbf{x})$ of equation (5.8) for logistic regression because it can be written as

$$w^*(\mathbf{x}) = \exp(\beta_0 + \beta_1 x_1 + ... + \beta_p x_p), \tag{10.7}$$

with the various terms in equation (10.4) that do not involve the X variables being combined into the constant β_0. What this means is that if β_1 to β_p are estimated using the discriminant function equation (10.6), then these are the same coefficients that are being estimated by logistic regression as described in Section 5.4.

10.3 Samples of Available and Unused Resource Units

If sample 1 for equations (10.5) and (10.6) is of available resource units, and sample 2 is of unused resource units then the function $\Omega(\mathbf{x})$ gives the probability that an individual with $\mathbf{X} = \mathbf{x}$ remains unused. Hence the RSPF

$$w^*(\mathbf{x}) = 1 - \Omega(\mathbf{x})$$

has the same form as equation (5.14) for logistic regression with a sample of available and a sample of used units. Thus in this situation the RSPF can be estimated by

$$\hat{w}^*(\mathbf{x}) = 1 - E\{\Omega(\mathbf{x})\}\exp\{\hat{\boldsymbol{\beta}}'(\mathbf{x} - \tfrac{1}{2}(\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2)\}. \tag{10.8}$$

Note that in this case $E\{\Omega(\mathbf{x})\}$ is the proportion of the available resource units that are unused, rather than the proportion used. If this proportion is not known then the function $\exp(\beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p)$ can be estimated. However it is proportional to the probability of not being used and is therefore not the RSF. This mirrors the situation when logistic regression is used for estimation, as described in Section 5.5.

10.4 Samples of Unused and Used Resource Units

The situation is more complicated when a resource selection probability is to be estimated from a sample of unused resource units and a sample of used resource units. In this case, for every unit with $\mathbf{X} = \mathbf{x}$ in the available population there are 1 - w*($\mathbf{x}$) units in the unused population and w*($\mathbf{x}$) units in the used population. Hence, for each unit with $\mathbf{X} = \mathbf{x}$ in the unused population there are w*($\mathbf{x}$)/{1 - w*($\mathbf{x}$)} units in the used population. In other words, the distribution of $\mathbf{X}$ in the used population can be obtained from the distribution in the unused population by selecting individuals with $\mathbf{X} = \mathbf{x}$ with a probability proportional to w*($\mathbf{x}$)/{1 - w*($\mathbf{x}$)}.

What this implies is that if equations (10.5) and (10.6) are used with sample 1 being of unused units and sample 2 being of used units, then $\Omega(\mathbf{x})$ corresponds to w*($\mathbf{x}$)/{1 - w*($\mathbf{x}$)}, and E{$\Omega(\mathbf{x})$} must be interpreted as the average value of $\Omega(\mathbf{x})$ for the units in the unused sample. This means that

$$E\{\Omega(\mathbf{x})\}\ \exp\{\boldsymbol{\beta}'(\mathbf{x} - \tfrac{1}{2}\,(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\} = w^*(\mathbf{x})/\{1 - w^*(\mathbf{x})\},$$

so that the RSPF is being assumed to have the form

$$w^*(\mathbf{x}) = E\{\Omega(\mathbf{x})\}\exp\{\boldsymbol{\beta}'\,(\mathbf{x} - \tfrac{1}{2}\,(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\}/[1 + E\{\Omega(\mathbf{x})\}\exp\{\boldsymbol{\beta}'\,(\mathbf{x} - \tfrac{1}{2}\,(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\}].$$

Another way of defining E{$\Omega(\mathbf{x})$} is the expected number of units in the used population for each unit in the unused population. Hence, if a proportion p of available units are used then

$$E\{\Omega(\mathbf{x})\} = p/(1 - p).$$

Substituting in to the last equation then gives

$$w^*(\mathbf{x}) = \frac{\{p/(1 - p)\}\exp\{\boldsymbol{\beta}'(\mathbf{x} - \tfrac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\}}{1 + \{p/(1 - p)\}\exp\{\boldsymbol{\beta}'(\mathbf{x} - \tfrac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\}}$$

which can be estimated by

$$\hat{w}^*(\mathbf{x}) = \frac{\{p/(1 - p)\}\exp\{\hat{\boldsymbol{\beta}}'\,(\mathbf{x} - \tfrac{1}{2}\,(\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2)\}}{1 + \ \ \{p/(1 - p))\}\exp\{\hat{\boldsymbol{\beta}}'\,(\mathbf{x} - \tfrac{1}{2}\,(\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2)\}} \qquad (10.9)$$

Thus, to estimate the RSPF from a sample of unused and a sample of used resource units, the ß values should be estimated with equation (10.6), and substituted into the last equation, together with the sample means and the proportion of used units.

If p is not known then the estimation of the RSPF, or even this function multiplied by an unknown scaling factor, becomes impossible. However, there are two ways to proceed that may be considered reasonable. First, setting p = ½ (or any other arbitrary value) will give an estimated function that is monotonically related to the true RSPF. Thus using this function will enable resource units to be ranked in order of their probability of use, so that it can be regarded as a type of selectivity index. Second, if p is small then w*($\mathbf{x}$) can be approximated by

$$\hat{w}^*(\mathbf{x}) = p.\exp[\hat{\boldsymbol{\beta}}'\{\mathbf{x} - \tfrac{1}{2}\,(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)\}],$$

so that an estimated RSF is

$$\hat{\mathrm{w}}(\mathbf{x}) = \exp(\hat{\boldsymbol{\text{ß}}}'\,\mathbf{x}). \tag{10.10}$$

In effect this is equivalent to treating the sample of unused resource units as a sample of available resource units.

Example 10.1 Nest Selection by Fernbirds Reconsidered

The linear discriminant function method can be used as an alternative to logistic regression to estimate a RSF from Harris' (1986) fernbird data that are given in Table 3.4. Because there is a sample of available resource units (random points in the study region) and a sample of used resource units (nest sites), equations (10.5) and (10.6) can be applied directly.

The sample mean vector and covariance matrix for the 25 available nest sites (with the variables in the order X_1 = canopy height, X_2 = distance to edge and X_3 = perimeter of clump) are found to be

$$\hat{\boldsymbol{\mu}}_1 = \begin{bmatrix} 0.4884 \\ 12.6200 \\ 2.9336 \end{bmatrix} \quad \text{and} \quad \hat{\boldsymbol{\Sigma}}_1 = \begin{bmatrix} 0.0216 & 0.1898 & -0.0007 \\ 0.1898 & 19.1308 & 1.5550 \\ -0.0007 & 1.5550 & 0.7128 \end{bmatrix}.$$

For the 24 nest sites the corresponding quantities are

$$\hat{\boldsymbol{\mu}}_2 = \begin{bmatrix} 0.7325 \\ 16.2500 \\ 4.0367 \end{bmatrix} \quad \text{and} \quad \hat{\boldsymbol{\Sigma}}_2 = \begin{bmatrix} 0.0773 & -0.1613 & 0.1352 \\ -0.1613 & 11.6087 & -0.4115 \\ 0.1352 & -0.4115 & 1.8759 \end{bmatrix}.$$

The pooled sample covariance matrix is therefore

$$\hat{\boldsymbol{\Sigma}} = \begin{bmatrix} 0.0489 & 0.0180 & 0.0568 \\ 0.0180 & 15.4498 & 0.5926 \\ 0.0658 & 0.5926 & 1.2820 \end{bmatrix},$$

with the inverse

$$\hat{\boldsymbol{\Sigma}}^{-1} = \begin{bmatrix} 21.9899 & 0.0181 & -1.1378 \\ 0.0181 & 0.0659 & -0.0314 \\ -1.1378 & -0.0314 & 0.8530 \end{bmatrix}.$$

Substituting into equation (10.6) then produces the estimates

$$\hat{\boldsymbol{\text{ß}}} = \begin{bmatrix} \hat{\text{ß}}_1 \\ \hat{\text{ß}}_2 \\ \hat{\text{ß}}_3 \end{bmatrix} = \begin{bmatrix} 4.1780 \\ 0.2090 \\ 0.5492 \end{bmatrix}.$$

The proportion p of available sites that were used by the fernbirds is not known, although it is obviously very small. Therefore, only a RSF that is proportional to the probability of selection can be estimated. Ignoring constant terms in the exponential argument this is

$$\hat{w}(\mathbf{x}) = \exp\{4.18(\text{CANOPY}) + 0.21(\text{EDGE}) + 0.55(\text{PERIM})\}. \qquad (10.11)$$

It is interesting to compare this function with the function

$$\hat{w}(\mathbf{x}) = \exp\{7.80(\text{CANOPY}) + 0.21(\text{EDGE}) + 0.88(\text{PERIM})\} \qquad (10.12)$$

that was calculated in Example 5.2 using logistic regression. On the face of it, these two functions agree only as far as the coefficient of the distance to edge is concerned. However, if the values obtained from equation (10.11) are plotted against the values obtained from equation (10.12) for all the 49 sampled resource units, as shown in Figure 10.1, then it is found that they are almost linearly related on a logarithmic scale, although there is far more variation in the estimates obtained from equation (10.12) than there is for those obtained from equation (10.13).

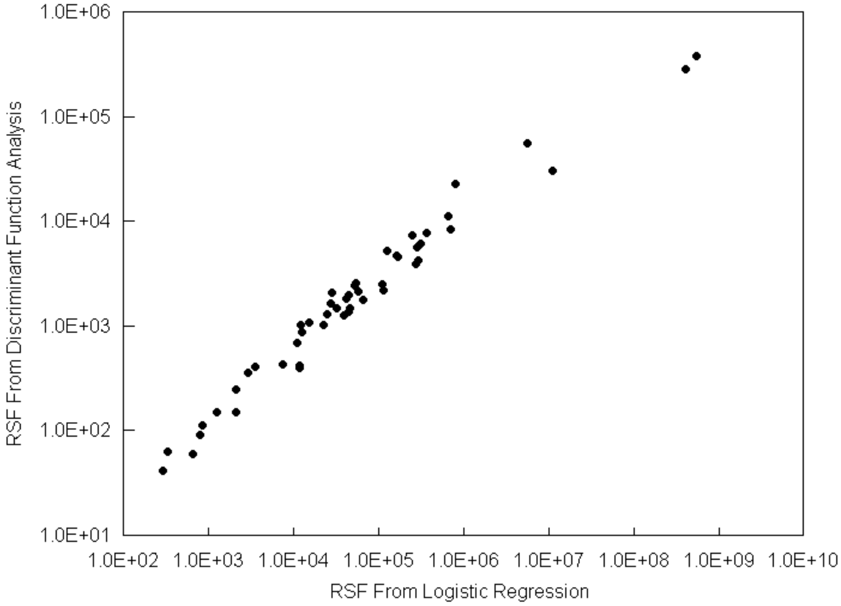


Figure 10.1 Comparison of values from RSFs estimated for nest site selection by fernbirds. The horizontal axis gives values obtained from a function estimated by logistic regression and the vertical axis gives values obtained from a function estimated using linear discriminant function equations. Logarithmic scales are used to accommodate the large range of values for the 49 sampled sites.

10.5 Quadratic Discriminant Function Analysis

The theory developed above for the estimation of a RSF in terms of Fisher's linear discriminant function can be extended to the allow of a selection functions of the form

$$w(\mathbf{x}) = \exp\{(\boldsymbol{\beta}'\ \mathbf{x} + \mathbf{x}'\ \boldsymbol{\Omega}\ \mathbf{x})\},$$

where $\underline{\Omega}$ is a symmetric p by p matrix. This means that the argument of the exponential function involves squares and products of the X variables as well as linear terms, which implies that the covariance matrices for the two populations can be different as well as the mean vectors. Using equations provided by Manly (1985, p. 66), it can be shown that $\log\{w(\mathbf{x})\}$ is in this case the usual quadratic discriminant function when the two populations have multivariate normal distributions for the X variables.

This approach for estimating a RSPF will not be pursued further here on the grounds that if it is considered that this function should include squares and products of the X variables then it is more straightforward to include these terms in a logistic regression formulation.

10.6 Discussion

In this chapter it has been shown that discriminant function analysis can be used as an alternative to logistic regression when there are two samples of resource selection units being compared. There are three different possible cases, depending on whether the samples are of available, used or unused resource units. This requires the assumption that the variables that are measured on resource units have multivariate normal distributions in the two populations sampled, which gives a more efficient method of estimation when the assumptions are valid (Efron, 1975). However, distributions are typically not normal with resource selection studies so that in practice discriminant function analysis has limited value.

Chapter Summary

- Fisher's linear discriminant function for two samples can be interpreted as the function that changes one multivariate normal distribution into a second multivariate normal distribution. It is therefore related to a resource selection function which shows how the individuals in one population must be selected in order to produce a second population.

- In terms of the estimation of a resource selection function using discriminant function analysis, three cases have to be considered: (a) when there is a sample of available resource units and a sample of used units, (b) when there is a sample of available units and a sample of unused units, and (c) when there is a sample of unused units and a sample of used units. With cases (b) and (c) there are complications if the proportion of used units is not known.

- Quadratic discriminant function analysis can also be used if selection changes the covariance matrix of a multivariate normal distribution.

- Because variables measured on resource units seldom have approximately normal distributions it is suggested that the discriminant function approach to estimating a resource selection function has limited practical use.

Exercise

Consider the data given in Table 5.7 for selection of *Daphnia publicaria* by yellow perch fry on 1 July, 1969. Note that this is a situation where there is a sample of available resource units and a sample of used resource units. Use logistic regression to estimate a resource selection function of the form $w(\mathbf{x}) = \exp(ß_1 X + ß_2 X^2 + ß_3 X^3)$, where X denotes the length of the *Daphnia* in mm. Verify that the estimated function is the same as that obtained in Exercise 2 in Chapter 5. Use chi-squared tests to compare the fit of the cubic model with quadratic and linear models in X. Compare the estimated linear function with the estimate obtained using linear discriminant function methods.

CHAPTER 11

ANALYSIS OF THE AMOUNT OF USE

In this chapter consideration is given to situations where the amount of use is recorded for resource units, rather than just whether they are used or not. Three cases are recognized: (a) where the amount of use is a count (such as the number of animals present); (b) where the amount of use is measured (such as the amount of biomass eaten); and (c) where the distribution of the amount of use is a mixture of zeros for unused units and a distribution of positive values for used units.

11.1 Introduction

Up to this point the situations that have been considered have all been in terms of whether individual resource units are used or unused. However, in some cases data are also available on the amount of use received by the used units. This raises the question of how such data can be analysed.

In this chapter three situations are considered in turn. First, it may be possible for a resource unit to be used by more than one animal. Then the amount of use can be recorded as the number of animals present, with zero indicating no use. Second, it may not be possible to know how many animals use a resource unit but the amount of the resource taken can be measured. Third, the distribution of the amount of use may consist of zeros for unused units and positive values for used units.

An example of the first case would be where quadrats are randomly placed in a study area and the number of animals of a certain species in each quadrat (U, say) is recorded, together with variables X_1, X_2, ... X_p that measure the physical characteristics of the quadrats. It would then be interesting to relate the counts U to the X variables.

An example of the second case would be where quadrats are randomly placed in a study area and the biomass of vegetation eaten (Y, say) is recorded for each quadrat, along with the variables X_1 to X_p that measure the physical characteristics of the quadrat. Again, it would be interesting to relate the Y values to the X variables. Here the point of view that is adopted is that the important difference between this case and the first case is simply that the first case involves count data but the second case does not.

The example of the second case becomes an example of the third case if the biomass eaten on a quadrat is zero for quadrats that animals did not visit and greater than zero for the quadrats that were visited.

11.2 Analysis of Counts of the Amount of Use

It is useful to begin the consideration of the analysis of counts of the amount of use by discussing the proposition that all cases where the amount of use is counted can be considered as similar. The most obvious situation of interest is, of course, where the counts U recorded on the resource units are the numbers of individual animals found on the units. However, suppose instead that U is the number of 'signs' of use such as the number of tooth marks on trees. Then the 'signs' for a resource unit may have been made by more than one animal and the relationship between the number of 'signs' and the number of animals may be unknown. Still, establishing a relationship between the number of 'signs' and the characteristics of resource units as measured by variables X_1 to X_p may be a valuable contribution to the understanding of resource selection by the animal in question. Indeed, the number of animal 'signs' may be a more relevant measure of use than the number of animals if it is the effect of the animals on the resource units that is the main concern.

In some cases it is possible to relate the counts of the number of times that resource units are used to the X variables using standard statistical methods. In particular, a log-linear model may be suitable so that it can be assumed that the expected value of the amount of use of resource unit i is a Poisson random variable with mean value

$$\mu_i = \exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}), \tag{11.1}$$

where x_{ij} is the value of X_j for this unit (Section 2.4).

Two problems are likely to occur with this approach. To begin with, it may be that the probability of including a resource unit in the sample varies according to the amount of use. This could be because used and unused resource units are sampled separately, or because there is a sampling bias caused by the fact that used units are more (or less) visible than unused units. The other problem is the natural tendency of many animal populations to be clustered, which means that the counts of the amount of use for different resource units will often show more variation than is expected from the Poisson distribution even when equation (11.1) gives the correct relationship between the expected amount of use and the X variables.

The first problem can only be overcome by explicitly recognizing that different resource units have different probabilities of being recorded. For example, suppose that the log-linear model for the counts of the amount of use of different resource units is correct, so that these counts have Poisson distributions with mean values given by equation (11.1), but the probability of recording information on an unused resource unit is $P_{\bar{u}}$ and the probability of recording information on a used resource unit is P_u. Then, if $P_{\bar{u}}$ and P_u are not equal the observed data will not follow the usual log-linear model. In fact, Bayes' theorem shows that the probability of observing zero use of a resource unit, conditional on that resource unit being sampled, is

$$\begin{aligned}\text{Prob}(U = 0 \mid \text{unit sampled}) &= \frac{\text{Prob}(\text{unit sampled} \mid U = 0).\text{Prob}(U = 0)}{\text{Prob}(\text{unit sampled})} \\ &= \frac{P_{\bar{u}}\exp(-\mu)}{P_{\bar{u}}\exp(-\mu) + P_u\{1-\exp(-\mu)\}} \\ &= \frac{\exp(-\mu)}{\exp(-\mu) + \{P_u/P_{\bar{u}}\}\{1-\exp(-\mu)\}}, \qquad (11.2)\end{aligned}$$

where μ is the unconditional expected count on the unit. Similarly, the probability of observing $U = u > 0$ is

$$\text{Prob}(U = u \mid \text{unit sampled}) = \frac{\text{Prob}(\text{unit sampled} \mid U = u).\text{Prob}(U = u)}{\text{Prob}(\text{unit sampled})}$$

$$= \frac{P_u \exp(-\mu)\mu^u/u!}{P_{\bar{u}}\exp(-\mu) + P_u\{1 - \exp(-\mu)\}}$$

$$= \frac{\{P_u/P_{\bar{u}}\}\exp(-\mu)\mu^u/u!}{\exp(-\mu) + \{P_u/P_{\bar{u}}\}\{1 - \exp(-\mu)\}} . \qquad (11.3)$$

Equations (11.2) and (11.3) define a two parameter modified Poisson distribution, and together with equation (11.1) they define a modified log-linear model. Maximum likelihood estimates for the parameters of the model can be obtained by standard iterative methods, as outlined in the Appendix to this chapter.

The second problem mentioned above with modelling data on counts of the amount of use of resource units (more variation in counts than is expected with the Poisson distribution) can be handled in various ways. For example, it might be reasonable to assume that the amount of use for the ith unit is a negative binomial random variable with the mean value μ_i given by equation (11.1). The variance of the count is then greater than the Poisson variance (which is equal to μ_i). Alternatively, it can be assumed that a log-linear model is correct except that the variances of the counts are inflated by a constant heterogeneity factor. As discussed in Section 2.4, it is then valid to estimate parameters as for a log-linear model and simply adjust the standard errors of parameter estimators by multiplying by the square root of the estimated heterogeneity factor.

Example 11.1 Habitat Selection by Galaxiids

This example concerns a study carried out by McIntosh *et al.* (1992) using two stretches of the Shag River in North Otago, New Zealand. In one stretch the native fish *Galaxias vulgaris* was present but not the introduced brown trout (*Salmo trutta*). In the other stretch both species were present. The question addressed by the study was whether the selection of habitat by galaxiids is the same irrespective of whether trout are present or not.

Within the stretch of river without trout, three 15 m long sampling sites were chosen and each site was divided into approximately 210 quadrats with a size of 50 cm by 50 cm. Forty random quadrats were then chosen from each site and, for each of these quadrats, the number of galaxiids present was recorded together with the following 14 variables: X_1 = the width of the stream; X_2 = the distance to the nearest bank; X_3 = the proportion of bedrock; X_4 = the proportion of gravel; X_5 = the proportion of cobble; $X_6 = 1 - X_3 - X_4 - X_5$ = the proportion of boulder; X_7 = the mean surface area; X_8 = the maximum surface area; X_9 = the mean interstitial space; X_{10} = the maximum interstitial space; X_{11} = the left current velocity in the quadrat; X_{12} = the middle current velocity in the quadrat; X_{13} = the right current velocity in the quadrat; and X_{14} = the mean depth of the stream.

Three sites were also chosen in the stretch of water without trout. However, it was found that for these 'trout' sites the sampling scheme used for the 'no trout' sites was not satisfactory because of the low numbers of galaxiids present. To overcome this problem, the sampling area for each 'trout' site was extended to include about 40 m of river (approximately 560 quadrats) and a large random sample of quadrats was chosen and checked for the presence of galaxiids. A random subsample of 29 or 30 quadrats was then taken from the large sample, and augmented by a random subsample of between 10 and 13 quadrats randomly chosen from those containing galaxiids. It can be calculated that this sampling scheme gave a probability of approximately $P_{\bar{u}} = 0.049$ of sampling a quadrat without galaxiids and a probability of approximately $P_u = 0.331$ of sampling a quadrat with galaxiids at each of the 'trout' sites. For each of the subsampled quadrats the 14 variables indicated above were measured.

Because the 14 variables that characterize the quadrats are in some cases highly correlated and linearly related, it was decided to make a principal component analysis (Manly, 1994, Chapter 6) the first step in the treatment of the data. There were 237 sampled quadrats in all, and the principal component analysis was applied to these without regard to any differences between the six sites that were sampled. All variables were initially scaled to have unit variances. The outcome of the principal component analysis was that five principal components were found to have variances of more than one, which is commonly taken as a rule of thumb for deciding how many components are important. Between them these five principal components accounted for 79.7% of the variation in the original data.

Table 11.1 shows the coefficients of the first five principal components after these principal components have been scaled to have a variance of unity over the 237 sampled quadrats. For example, the first principal component is

$$\begin{aligned} Z_1 = {} & 0.22x_1 + 0.21x_2 - 0.17x_3 - 0.48x_4 - 0.17x_5 + 0.77x_6 + 0.97x_7 + 0.92x_8 \\ & + 0.97x_9 + 0.89x_{10} + 0.10x_{11} + 0.00x_{12} + 0.11x_{13} - 0.07x_{14}, \end{aligned}$$

where x_i is the value of the measured variable X_i after standardization to a mean of zero and a variance of one. From the coefficients in Table 11.1 it seems reasonable to identify the components as measuring the following properties of quadrats: Z_1, surface area and interstitial space; Z_2, velocity and depth of the stream; Z_3, width of stream; Z_4, bedrock and lack of cobble; and Z_5, gravel.

Following the principal component analysis, an attempt was made to relate the number of galaxiids found in quadrats to the values of the principal components. Because of the different sampling schemes at 'no trout' and 'trout' sites, it is convenient to analyse the data separately for the two cases. These separate analyses will now be described in turn, followed by a discussion of what can be learned from the whole study.

Analysis of Results From the 'No Trout' Sites

A series of log-linear models were fitted to the data from the three 'no trout' sites to determine how well the numbers of galaxiids in quadrats can be accounted for using some or all of the first five principal components and dummy (0-1) variables allowing for site effects. The results obtained were as follows:

(1) The null model (model A) with the same expected number of galaxiids in each quadrat was fitted. This gave a deviance of 104.82 with 119 degrees of freedom (df).

Table 11.1 Coefficients of the first five principal components obtained from the data on all sampled quadrats. These are coefficients for the standardized (mean zero and variance one) values of both the principal components and the original X variables.

	Principal component				
Variable	1	2	3	4	5
1 Width of stream	0.23	-0.40	-0.60	0.10	0.39
2 Distance to nearest bank	0.21	0.17	-0.71	0.16	0.41
3 Proportion of bedrock	-0.17	-0.21	0.54	0.61	0.39
4 Proportion of gravel	-0.48	-0.14	-0.50	0.31	-0.59
5 Proportion of cobble	-0.17	0.37	-0.01	-0.79	0.01
6 Proportion of boulder	0.77	0.08	0.12	-0.19	0.31
7 Mean surface area	0.97	-0.03	-0.03	-0.03	-0.06
8 Maximum surface area	0.92	-0.05	0.07	0.13	-0.25
9 Mean interstitial space	0.97	-0.04	-0.03	-0.01	-0.09
10 Maximum interstitial space	0.89	-0.05	0.07	0.14	-0.29
11 Left current velocity	0.10	0.72	-0.21	0.11	-0.04
12 Middle current velocity	0.00	0.81	0.00	0.11	0.04
13 Right current velocity	0.11	0.79	0.05	0.13	0.11
14 Mean depth of the stream	-0.07	0.69	0.04	0.34	-0.10

(2) Next, model B included all five principal components and the dummy variables for site effects (S_1 = 1 for a quadrat from site 1, or otherwise 0; S_2 = 1 for a quadrat from site 2, or otherwise 0) was fitted. This gave a deviance of 86.19 with 112 df, which is significantly lower than the value from the null model (difference = 18.63 with 7 df, $p < 0.01$ by comparison with the chi-squared distribution).

(3) The third, fourth and fifth principal components were removed from the model one by one, without any significant increases in the deviance. This resulted in model C with a deviance of 91.29 with 115 df. The increase in the deviance over that of model B was not significant (difference = 5.10 with 3 df, $p > 0.1$).

(4) The second principal component was removed to produce model D. This resulted in a deviance of 94.45 with 116 df. The increase in the deviance over that of model C is not significant at the 5% level, although it is approaching significance (difference = 3.16 with 1 df, $0.1 > p > 0.05$).

(5) The site effect variables S_1 and S_2 were removed to produce model E. This resulted in a significant increase in the deviance above the value for model D (difference = 8.63 with 2 df, $0.05 > p > 0.01$).

As will be seen below, there is clear evidence that the number of galaxiids was related to the first two principal components at the 'trout' sites. For this reason, it seems appropriate to accept model C as the best model for the 'no trout' sites even although the effect of the second principal component is not significant at the 5% level. According to this model,

$$\hat{\mu} = \exp(-1.80 + 0.45Z_1 + 0.30Z_2 - 1.80S_1 + 0.54S_2), \qquad (11.4)$$

so that the expected number of galaxiids in a quadrat was high for quadrats with large values for Z_1 (surface area and interstitial space) and large values for Z_2 (velocity and depth of the stream). Relative to site 3, there seem to have been low numbers of galaxiids at site 1 and slightly high numbers at site 2.

Table 11.2 shows the estimates that are obtained when a model relating galaxiid counts to Z_1 and Z_2 is fitted separately for each of the three sites. The consistency of the coefficients of these variables is reassuring since it indicates that the selection of habitat by galaxiids was much the same at each site. Furthermore, without going into details it can be mentioned that estimating separate coefficients for Z_1 and Z_2 for each of the three sites gives an insignificant reduction in the deviance of 2.37 with 4 df.

Table 11.2 Results obtained from fitting log-linear models relating galaxiid counts at the 'no trout' sites to the first two principal components.

Parameter		Estimates	Std. Err.	Ratio
		Estimates from site 1		
Constant		-1.24	0.32	-
Coefficient of	Z1	0.52	0.41	1.28
	Z2	0.25	0.45	0.56
		Estimates from site 2		
Constant		-0.24	0.31	-
Coefficient of	Z1	0.72	0.34	2.10 *
	Z2	0.17	0.24	0.70
		Estimates from site 3		
Constant		-1.95	0.55	-
Coefficient of	Z1	0.27	0.30	0.90
	Z2	0.48	0.31	1.57
		Estimates from all sites		
Constant		-1.80	0.40	-
Coefficient of	Z1	0.44	0.18	2.47 *
	Z2	0.30	0.17	1.73
	S1	0.54	0.51	1.06
	S2	1.33	0.46	2.89 *

*Significant at the 5% level (outside the range ± 1.96).

The observed frequencies of galaxiids ranged from 0 to 2 and the expected frequencies were all quite small. Therefore, it is questionable to compare the deviance statistic of 91.29 with 115 df for model C with the percentage points of the chi-squared distribution in order to assess the absolute goodness-of-fit of the log-linear model. In fact, if this comparison is made then the fit seems to be unbelievably good because the probability of a goodness-of-fit statistic this low is about 0.002. A better idea of the adequacy of the model is obtained by considering Figure 11.1 which shows standardized residuals for groups of five quadrats. This figure was constructed by ordering the 40 quadrats within each site from the one with the smallest expected number of galaxiids to the one with the largest number of expected quadrats. The quadrats were then grouped into sets of five within the sites and the observed and expected number of galaxiids determined for each group. The standardized residuals (Observed - Expected)/√(Expected) were then calculated and plotted. These standardized residuals appear to have approximately standard normal distributions and the model therefore seems to be a good fit to the data.

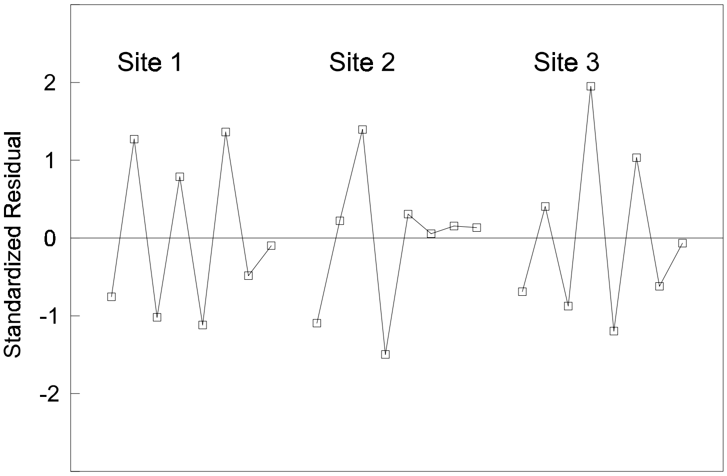


Figure 11.1 Standardized residuals calculated using equation (11.4) for sets of five quadrats from the 'no trout' sites.

Analysis of Results From the 'Trout' Sites

The analysis carried out at the 'no trout' sites was repeated on the trout sites but with the modified log-linear model that is described by equations (11.1) to (11.3) and discussed further in the Appendix to this chapter. As noted above, the sampling probability for quadrats without galaxiids was approximately 0.049 and the sampling probability for quadrats with galaxiids was approximately 0.331. Hence it was assumed for model fitting that $P_u/P_{\bar{u}} = 0.331/0.049 = 6.7$.

It was found by first fitting a model with effects for the first five principal components, and then removing nonsignificant terms one by one, that only the first two of the principal components seem to have important effects. With just these terms in the model it is estimated that the expected number of galaxiids in a quadrat is given by

$$\hat{\mu} = \exp(-2.91 + 0.86Z_1 + 0.73Z_2). \quad (11.5)$$

Table 11.3 shows that the coefficients of Z_1 and Z_2 are both quite large relative to their standard errors, and that quite similar estimates are obtained if the model is fitted separately to the data from the three sites. Furthermore, the improvement in fit that is obtained by estimating the parameters separately for each site is small and insignificant (difference in deviance = 3.67 with 6 df).

Table 11.3 Results obtained from fitting modified log-linear models relating galaxiid counts to the first two principal components for data from the 'trout' sites.

Parameter		Estimate	Std. Err.	Ratio	
		Estimates from site 1			
Constant		-2.93	0.47	-	
Coefficient of	Z1	0.94	0.32	2.96	*
	Z2	0.33	0.45	0.74	
		Estimates from site 2			
Constant		-2.69	0.40	-	
Coefficient of	Z1	0.75	0.33	2.29	*
	Z2	0.93	0.42	2.23	*
		Estimates from site 3			
Constant		-3.22	0.56	-	
Coefficient of	Z1	0.81	0.43	1.88	
	Z2	1.17	0.59	1.97	*
		Estimates from all sites			
Constant		-2.91	0.26	-	
Coefficient of	Z1	0.86	0.20	4.21	*
	Z2	0.73	0.26	2.81	*

*Significant at the 5% level (outside the range ± 1.96).

Analysis of Combined Results

The coefficients of Z_1 and Z_2 in equations (11.4) and (11.5) are quite similar. In fact, the two equations indicate virtually the same habitat selection by the galaxiids at the 'no trout' and 'trout' sites, as can be seen from Figure 11.2, which shows the logarithm of the values obtained from equation (11.5) plotted against the logarithms of the values obtained from equation (11.4) (with $S_1 = S_2 = 0$ to remove site effects). To be more precise, logarithms of the functions (11.4) and (11.5) were evaluated for the 120 sampled quadrats at the 'no trout' sites. This provided the 120 points shown in part (a) of the figure. The functions (11.4) and (11.5) were also evaluated for the 117 sampled quadrats at the 'trout' sites. This provided the 117 points shown in part (b) of the figure. Note that logarithms were plotted simply to avoid 'bunching' of the points for the relatively large number of quadrats with small estimated means.

The plotted values show an extremely high correlation (overall: $n = 237$, $r = 0.992$, $p < 0.001$). Hence, although there were far fewer galaxiids at the 'trout' sites than at the 'no trout' sites, the function estimated from the 'no trout' sites predicts very well the habitat use at the 'trout' sites, and the function estimated at the 'trout' sites predicts very well the habitat use at the 'no trout' sites.

A comparison of parts (a) and (b) of Figure 11.2 indicates that the distribution of the values of equations (11.4) and (11.5) are virtually the same for the quadrats from the 'no trout' sites and the quadrats from the 'trout' sites. Thus the 'no trout' and 'trout' sites seem to be inherently about as attractive to galaxiids. This raises the question of why there are so many less galaxiids at the 'trout' sites. The obvious answer is the presence of trout which could inhibit the galaxiid population either by predating on it or by competing for space. However, because the three 'no trout' sites and the three 'trout' sites are really pseudoreplicates (Hurlbert, 1984) taken from areas that were not chosen at random, this may or may not be true. The galaxiid numbers may be related to some entirely different factor that has not been recognized in this study.

11.3 Analysis of Continuous Measures of the Amount of Use

Suppose now that the amount of use of a resource unit is a continuous variable Y with all resource units having the same probability of inclusion in the sample to be analysed, and with no complications from a substantial part of the population consisting of unused units with zero measured use. In that case it may well be that standard methods of statistical analysis will be sufficient to determine the relationship between the amount of use and variables measured on resource units. For example, the linear regression model

$$Y = ß_0 + ß_1X_1 + ... + ß_pX_p + \epsilon$$

might be used to relate Y to the X variables that measure characteristics of the resource unit. Also, if the assumptions of the linear regression model are not valid then it may be possible to transform the data so that the model can be used. Alternatively, a non-linear regression might be considered.

There are obviously endless possibilities in this type of situation and it is difficult to make any further general comments.

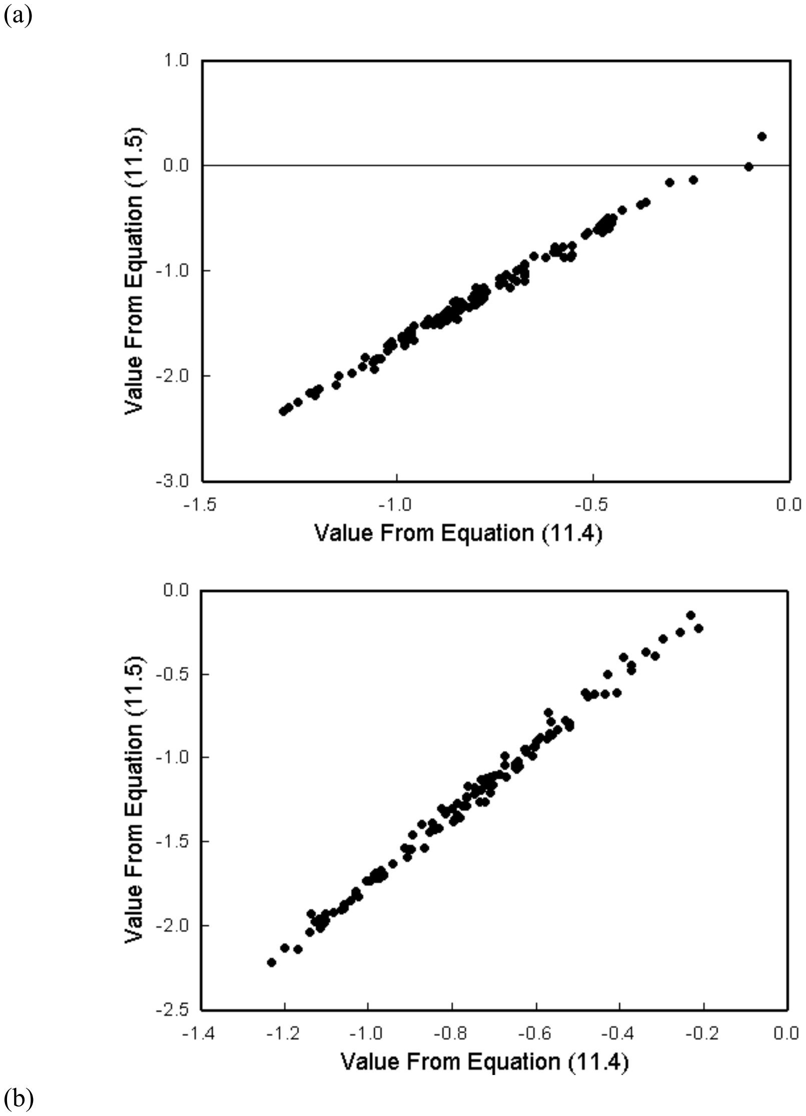


Figure 11.2 Plot of logarithms of mean numbers of galaxiids in quadrats as estimated from the data from 'trout' sites with equation (11.5) against mean numbers estimated from the data from 'no trout' sites with equation (11.4): (a) plot for the 120 quadrats from 'no trout' sites; (b) plot for the 117 quadrats from 'trout' sites.

11.4 Mixtures of Zeros and Positive Measures of the Amount of Use

The final situation that we consider is where the distribution of the amount of use Y of a resource consists of 0 with probability p and values from a positive distribution with probability 1 - p. Thus p is the probability that a resource unit is not used, so that Y is necessarily zero, and the continuous distribution is the distribution of Y conditional on use.

Models of this type have been considered by Lachenbruch (1976) for the situation where two samples need to be compared to see whether the probability p or the positive distribution differ in the populations that the samples are drawn from. Exponential and log-normal distributions were considered for the non-zero values in the samples, as well as the Poisson distribution with zero values omitted. Lachenbruch discussed maximum likelihood estimation of the parameters of the two part distribution and showed that the maximum likelihood estimate of p is the sample proportion of zeros and that the parameters of the non-zero distribution are exactly the same as is found by ignoring the zero data. Furthermore, the estimates of p and the parameters of the non-zero distribution are independent.

In the context of resource selection the situation is more complicated because the probability of a unit being used and the amount of use of a unit are likely to be functions of the characteristics of that unit as measured by the variables X_1 to X_p. However, it is still true that the estimation of the probability of use and the distribution of the amount of use can be considered separately.

To see this, suppose that the values of the X variables for the ith resource unit are x_{i1}, x_{i2}, ..., x_{ip}, and that the probability of this unit being used is given by a function $p(x_{i1},x_{i2},...,x_{ip})$. Suppose also that if this unit is used then the probability density function of the amount of use Y_i of the unit is $f(y_i;x_{i1},x_{i2},...,x_{ip})$. Then if data are collected on n resource units, of which the first n_0 are unused, the likelihood function for the data takes the form

$$L = \prod_{i=1}^{n_0} \{1 - p(x_{i1},x_{i2},...,x_{ip})\} \prod_{i=n_0+1}^{n} p(x_{i1},x_{i2},...,x_{ip})\ f(y_i;x_{i1},x_{i2},...,x_{ip}).$$

Hence the log-likelihood function is

$$\log(L) = \sum_{i=1}^{n_0} \log\{1-p(x_{i1},x_{i2},...,x_{ip})\} + \sum_{i=n_0+1}^{n} \log\{p(x_{i1},x_{i2},...,x_{ip})\}$$

$$+ \sum_{i=n_0+1}^{n} \log\{f(y_i;x_{i1},x_{i2},...,x_{ip})\}.$$

It can be seen that this log-likelihood function is in two parts. The first two terms on the right-hand side relate to the probability of use, while the last term relates to the amount of use conditional on a unit being used. This implies that the maximum likelihood estimators of any parameters of the function p are not a function of the measured amounts of use y_i on the used units, and that the maximum likelihood estimators of any parameters of the function f are not a function of the X values for unused units. In practice, therefore, the two functions p and f can be estimated quite separately.

There is a reservation here that must be made. If the functions p and f have any shared parameters then, of course, maximum likelihood estimation must be carried out using all the data simultaneously taking this into account. For example it might be reasonable to assume that p and f are both functions of the same linear combination $ß_0 + ß_1X_1 + ... + ß_pX_p$ of the X variables, where this measures the desirability of a resource unit both in terms of whether it is used at all and, if so, how much it is used. In that case estimating p and f separately will not be satisfactory.

The methods used for separate estimation of p and f will obviously depend on the nature of the data. However, a reasonable approach to try would involve estimating the probability of use function with a logistic regression as discussed in Chapter 5, and then estimating the amount of use function using multiple regression.

Chapter Summary

- Three situations are considered where the amount of use is recorded for resource units that are used. These are where there are (a) counts of the number of uses, (b) continuous measures of the amount of use, and (c) zeros for unused units and a measure of the amount of use for used units.
- A modified log-linear model is defined for counts of the number of uses of different resource units.
- For analysing measures of the amount of use, various different approaches might be used, including multiple linear regression.
- A maximum likelihood estimation approach is suggested for situations where there is a probability p of no use and a distribution of the amount of use for used units.

Appendix: Estimation of a Modified Log-Linear Model

In this Appendix the maximum likelihood estimation of the parameters Θ, $ß_0$, $ß_1$, ..., $ß_p$ of the model specified by equations (11.1) to (11.3) are considered for the case where the data available consists of the values of variables X_1 to X_p for each of n_0 unused resource units, and for each of n_1 resource units used at least once. For the used resource units the number of times used must also be known.

In this situation the likelihood function (the probability of the observed data as a function of the unknown parameters) is given by

$$L = \prod_{i=1}^{n_0} 1/[1+\Theta\{\exp(\mu_i)-1\}] \prod_{i=n_0+1}^{n_0+n_1} [\{P_u/P_{\bar{u}}\}\mu_i^{u_i}/(u_i!)]/[1+\{P_u/P_{\bar{u}}\}\{\exp(\mu_i)-1\}],$$

where the labelling of the resource units is such that the first n_0 are the unused ones and the last n_1 are the used ones,

$$\mu_i = \exp(ß_0 + ß_1x_{i1} + ... + ß_px_{ip})$$

is the expected number of times that the ith unit will be used, and u_i is the observed amount of use for the ith unit.

The likelihood function L is a function of $P_u/P_{\bar{u}}$, $ß_0$, $ß_1$, ..., and $ß_p$. Therefore, maximum likelihood estimates of some or all of these parameters can be obtained by maximizing L or its logarithm $\log_e(L)$. Equations for finding these estimates can be found in the usual way by equating the derivatives of $\log_e(L)$ to zero for each of the parameters. However, these equations do not have an explicit solution so that some numerical method for finding the estimates must be used. The Newton-Raphson method (Manly, 1985, p. 405) seems to work very well for this purpose, and was what was used for the calculations for the 'trout' site data of example 11.1.

Although there is no simple test for the goodness-of-fit of this model (because the observed data will not have Poisson distributions), it is still possible to compare the goodness-of-fit of two models if one of the models is a special case of the other. Thus if $\log_e(L_1)$ is the maximized log-likelihood for the special case model, with p_1 estimated parameters, and $\log_e(L_2)$ is the maximized log-likelihood for the more general model, with $p_1 + p_2$ parameters (of which the first p_1 are the same as for the first model), then for large samples $-2\{\log_e(L_2) - \log_e(L_1)\}$ can be compared with the chi-squared distribution with p_2 degrees of freedom to see if the more general model is a significantly better fit than the special case model. In effect, this means that the value of $-2 \log_e(L)$ can be treated as a measure of the goodness-of-fit of a model, although this cannot be compared directly with critical values of the chi-squared distribution.

CHAPTER 12

OTHER TYPES OF ANALYSIS

There are many other approaches for analysing resource selection data apart from the use of resource selection functions. In this chapter we consider two methods that continue to be applied on a regular basis. these are compositional analysis (Aebischer *et al*., 1993) and the a Mahalonbis distance method (Clark *et al*., 1993). The recent use of neural networks as an alternative to regression methods for estimating the equivalent of resource selection functions is also briefly reviewed.

12.1 Compositional Analysis

Compositional analysis is an extension of multivariate analysis of variance (Aebischer *et al*., 1993). The technique is applicable with radiotelemetry studies in particular, can be used with classification variables, and uses individual animals as replicates.

Resource use is defined in terms of the proportion of different types of resources within the estimated home range boundary, or other used area. Using the animal as the unit of observation may avoid problems related to the sampling level (Kenward, 1992), the unit-sum constraint whereby the avoidance of one resource type leads to some selection for alternatives (Aebischer *et al*., 1993), the differential use of resources by different groups of animals (Aebischer *et al*., 1993), and pseudoreplication if animals behave independently (Hurlbert, 1984). Despite these advantages, there are important assumptions underlying compositional-based analyses, including those of independence for the data from different animals and multivariate normality. The need to add an arbitrary constant to zero data under some circumstances (Pendleton *et al*., 1998) has also lead to the method being criticized.

Compositional analysis can be conducted easily in standard statistical packages using the option for multivariate analysis of variance (MANOVA). If use is defined as the proportional occurrences of different types of resources within home ranges and availability is defined as in the same way but for a larger area, then Aebischer *et al.* (1993) recommended a two-stage analysis, corresponding to Johnson's (1980) 2nd and 3rd orders. Standard MANOVA tests such as one based on Wilk's lamda can be used to see if there is evidence for selection. When selection is indicated, tests to compare pairs of resource types can be carried out.

A compositional analysis proceeds as follows. Assume that there are D types of resource available, and that an individual animal's proportional resource use of these resources is described by the composition x_{u1}, x_{u2}, ... x_{uD} where x_i is the estimated proportion of the resources used by the individual that are of type i, and the proportions sum to one. Similarly, let the available proportions for the same animal be x_{a1}, x_{a2}, ..., x_{aD}. It can be shown that for any component x_j of a composition, the log-ratio transformation $y_i = \log_e(x_i/x_j)$ produces new variables that are linearly independent, with a specific choice for j. Based on this result, the differences

$$d_i = \log_e(x_{ui}/x_{uj}) - \log_e(x_{ai}/x_{aj})$$

are calculated for the ith animal to represent the difference between the relative use and availability of resources i and j. With no selection the mean value of d_i is expected to be zero over all animals, for all i. An overall test for selection therefore involves seeing whether the vector of mean values of d_i is significantly different from a zero vector, using Wilk's lambda test, for example. If an overall test indicates selection then t-tests or randomization tests can be used to compare pairs of resource types. If necessary, zero values vave to be replaced by a small positive number when calculating d_i values.

Compositional analysis is very similar to Johnson's ranking method (1980) which relies on the form

$$d^*_i = \{\text{rank}(x_{ui}) - \text{rank}(x_{ai})\} - \{\text{rank}(x_{uj}) - \text{rank}(x_{aj})\}.$$

Also note that the equation for d_i can be written as

$$d_i = \ln(x_{ui}/x_{ai}) - \ln(x_{uj}/x_{aj}) = \ln(w_i) - \ln(w_j),$$

the difference in logs of the selection ratios.

Example 12.1 Habitat Selection by Bears on the Kenai Peninsular

This example of compositional analysis is based on data from nine radio-tagged brown bears *(Ursus arctos horribilis*) and six habitat types. The data were collected from 1996 to 1998 on the Kenai Peninsula in Alaska and are from female bears without cubs with at least 100 relocations. Here we consider part of the full data set, which is also referred to in Example 9.2. Minimum convex polygon home ranges were used to define the availability for each bear. This is therefore an example of a design III study, with availability defined within the home ranges for each of the individuals, and use determined by the proportion of relocations of each individual within each habitat type.

The habitat types are coniferous forest (CF), mixed forest (MF), deciduous forest (DF), shrub (SH), herbaceous (HE), and other (OT). Table 12.1 summarizes the use and availability estimates for each individual bear for each of these habitats, and Table 12.2 shows the values of

$$d_i = \ln(x_{ui}/x_{ai}) - \ln(x_{uj}/x_{aj})$$

using the other (OT) landcover class as the reference category.

MANOVA was used to test for overall selection using the SAS PROC GLM (SAS Institute Inc., 1999). Wilks' Lambda statistic is 0.0985, indicating significant selection ($p = 0.038$). The ranking of the resources is in the order mixed forest, deciduous forest, coniferous forest, shrub, other, and herbaceous based on the average of their d values across the nine animals, with the reference category receiving a value of zero.

To see which resource types are significantly different from others, t-tests can be used. Thus to compare coniferous forest, mixed forest, deciduous forest, shrub and herbaceous, to the other habitat, a one-sample t-test can be conducted comparing the mean of the d_i values to zero, while to compare coniferous forest, mixed forest, deciduous forest, shrub and herbaceous habitats two at a time paired t-tests are appropriate to compare the average difference in mean d values with zero.

Table 12.1 Used and available proportions of six habitat types for nine radio-tagged bears.

	Conifer forest		Mixed forest		Deciduous		Shrub		Herbaceous		Other	
Bear	A[1]	U[2]	A	U	A	U	A	U	A	U	A	U
1	0.454	0.566	0.152	0.164	0.060	0.091	0.178	0.070	0.070	0.038	0.086	0.070
2	0.398	0.471	0.020	0.077	0.020	0.048	0.320	0.308	0.110	0.048	0.132	0.048
3	0.774	0.666	0.028	0.140	0.016	0.018	0.032	0.030	0.078	0.058	0.072	0.088
4	0.600	0.548	0.174	0.228	0.018	0.033	0.102	0.122	0.098	0.050	0.008	0.019
5	0.634	0.785	0.172	0.084	0.058	0.029	0.066	0.066	0.062	0.034	0.008	0.002
6	0.550	0.497	0.132	0.209	0.038	0.100	0.132	0.100	0.092	0.044	0.056	0.050
7	0.364	0.294	0.016	0.083	0.064	0.037	0.278	0.404	0.156	0.110	0.122	0.073
8	0.125	0.321	0.021	0.085	0.009	0.024	0.353	0.410	0.149	0.075	0.344	0.085
9	0.162	0.283	0.013	0.050	0.006	0.008	0.350	0.475	0.262	0.113	0.207	0.071

[1] Available proportion of habitat.
[2] Used proportion of habitat.

Table 12.2 Differences in log-ratios calculated from data in Table 1 comparing habitat use within home ranges to availability defined by the home range. See text for definitions of the habiata types CF, MF, etc.

	Differences in log ratios (d)				
Bear	CF/OT (d_1)	MF/OT (d_2)	DF/OT (d_3)	SH/OT (d_4)	HE/OT (d_5)
1	0.426	0.282	0.622	-0.727	-0.405
2	1.180	2.360	1.887	0.973	0.182
3	-0.351	1.409	-0.083	-0.265	-0.497
4	-0.956	-2.897	-0.259	-0.686	-1.538
5	1.600	0.670	0.693	1.386	0.786
6	0.012	0.573	1.081	-0.164	-0.624
7	0.300	2.160	-0.034	0.887	0.164
8	2.341	2.796	2.379	1.548	0.712
9	1.628	2.417	1.358	1.375	0.229

For example, the comparison of deciduous forest to the other category is conducted by a t-test comparing the mean for d_3 to zero. The mean for d_3 is 0.849 giving a p-value of 0.024 and hence indicating significantly higher selection for deciduous forest compared to the other category. On the other hand, the comparison of deciduous forest with herbaceous is conducted by comparing the mean of the differences between d_3 and d_4 across the nine animals with a paired t-test. This mean difference is 0.905 and the t-test gives a p-value of 0.005, indicating significant selection for deciduous forest compared to herbaceous. Table 12.3 summarizes the results of all the possible comparisons of this type. It can be noted that although mixed forest has the highest mean d estimate, the t-tests show no statistically significant differences between mixed forest and the other categories, apparently due to high variability in the d values for mixed forest.

Table 12.3 Means, standard deviations (SD) and t-test results for making pairwise comparisons of habitat types (one sample and paired t-tests with eight df).

Comparison	Differences d_i	Mean	SD	P-value	P-value[1]
CF versus OT	d_1	0.687	1.069	0.090	0.216
MF versus OT	d_2	1.085	1.751	0.100	0.032
DF versus OT	d_3	0.849	0.914	0.024	0.043
SH versus OT	d_4	0.481	0.933	0.161	0.384
HE versus OT	d_5	-0.110	0.733	0.664	0.304
CF versus MF	d_1- d_2	-0.399	1.241	0.363	0.048
CF versus DF	d_1- d_3	-0.163	0.613	0.449	0.232
CF versus SH	d_1- d_4	-0.196	0.523	0.271	0.250
CF versus HE	d_1- d_5	0.797	0.504	0.002	0.008
MF versus DF	d_2- d_3	0.236	1.387	0.624	0.144
MF versus SH	d_2- d_4	0.605	1.258	0.187	0.021
MF versus HE	d_2- d_5	1.196	1.245	0.021	0.008
DF versus SH	d_3- d_4	0.369	0.808	0.208	0.099
DF versus HE	d_3- d_5	0.960	0.749	0.005	0.007
SH versus HE	d_4- d_5	0.591	0.429	0.003	0.002

[1] P-value for the comparison of selection ratios (see text).

Selection ratios were also calculated for this data set by averaging individual selection ratios for each animal and paired t-tests were conducted to compare the means for the different habitat types. The final column of Table 12.3 shows the p-values for these comparisons, which are generally rather similar to those from compositional analysis.

12.2 Mahalanobis Distance

The Mahalanobis distance measure has been proposed for mapping habitat and animal relationships (Clark *et al.*, 1993), particularly with geographical information (GIS) systems. Each resource unit (e.g. the pixels in a GIS) is described by the values that it possesses for variables X_1 to X_p, and the Mahalanobis distance of an available unit to the mean for the used units is taken as a measure of the suitability for use of that unit, or its probable use. Maps of probable use can then be created by computing the Mahalanobis distance for all resource units throughout the study area (Clark *et al.*, 1993; Knick and Dyer, 1997; Knick and Rotenberry, 1998).

The Mahalanobis distance for a unit with measurements $\mathbf{x} = (x_1, x_2, ..., x_p)$ is calculated as

$$M = (\mathbf{x} - \hat{\boldsymbol{\mu}})' \hat{\boldsymbol{\Sigma}}^{-1} (\mathbf{x} - \hat{\boldsymbol{\mu}})$$

where $\hat{\boldsymbol{\mu}}$ is the mean vector of habitat characteristics estimated from the used locations and $\hat{\boldsymbol{\Sigma}}$ is the estimated covariance matrix estimated from the used locations. Assuming multivariate normality, Mahalanobis distances are approximately distributed as chi-squared with p - 1 df.

The technique has some advantages over other resource selection methods. In contrast to logistic regression and other methods for mapping resource selection, only used resources need to be identified. The method is also unique among selection analyses because it uses only mean characteristics and the variability of the use patterns. There is no need to compute resource availability, which helps circumvent an important and often difficult aspect of some resource selection techniques. For this reason, the approach may be useful compared to logistic regression and other mapping techniques if defining availability is not practical. Furthermore, the Mahalanobis distance is not prone to problems due to the correlation between the variables that describe the resource units, which may upset regression techniques (Knick and Rotenberry 1998).

Despite these advantages, Knick and Rotenberry (1998) discussed various limitations of the Mahalanobis distance technique and only recommended its use when the various landscapes are represented well in the samples and do not change during the period of study. It should also be noted that although the technique does not require the defining of availability, the mapping procedure does requires the definition of a study area boundary, which may influence the interpretation of the results.

The Mahalanobis distance analysis is not readily available in statistical or GIS software packages. The calculation of the variance-covariance matrix can be calculated using most standard statistical packages, but the actual calculation of the Mahalanobis distance for each pixel in the study area would require a special algorithm to be written in the GIS.

Example 12.2 Habitat Selection by Bears on the Kenai Peninsular Reconsidered

Like the last example, this concerns habitat selection by female brown bears *(Ursus arctos horribilis*) on the Kenai Peninsular in Alaska. The data now being considered consist of relocations of 26 bears without cubs collected during the summer season over a four-year period. Only bears with at least 20 relocations were included, but the number of relocations varied from 20 to 442 relocations per bear.

The entire Kenai Peninsula was considered available except for the permanent ice and glaciers. A systematic sample of 13,385 pixels was used to describe the study area, rather than attempting to calculate the Mahalanobis distances for all possible pixels. The GIS was used to derive several continuous covariates identified as potential predictors of selection. Here six variables are considered to describe the pixels, which are:

SSTML the density of high potential salmon spawning streams (miles within a 1 km block around the pixel)
KESTM1 density of potential salmon spawning streams (miles with a 1 km block around the pixel)
ELEV The mean elevation in a 100 m block around the pixel
KESTM2 The distance to the nearest salmon stream in 100 m units
KROAD The distance to nearest road in 100 m units
COVER The distance to cover in 100 m units
DEV The density of human developments in numbers per km^2

Means for the animal locations and the systematically sample of available points are shown in Table 12.4. Based on mean comparisons only, the bears included in this analysis were associated with high salmon stream densities (SSTML and KESTM1), short distances to salmon streams (KESTM2), long distances to roads (KROAD), short distances to cover habitats (COVER), low elevations (ELEV), and low densities of human developments (DEV).

Table 12.4 Summary statistics for the Kenai brown bear study area and use data.

	Study Area	Used by Bears	
Variable	Mean	Mean	Std. Dev.
SSTMH	26.305	112.288	69.453
KESTM1	28.302	153.022	84.565
ELEV	372.892	316.448	176.193
KESTM2	21.828	12.224	7.770
KROAD	80.169	86.508	77.360
COVER	0.597	0.360	0.165
DEV	0.272	0.027	0.057

The map based on Mahalonbis distances is shown in Figure 12.1. Some authors have plotted probabilities on maps like this, on the assumption that the distances follow a chi-squared distribution with p - 1 df, which follows if the variables measured on resource units follow a multivariate normal distribution. Plotting of probabilities is not recommended in general because the assumption of a multivariate normal distribution will usually be suspect. Rather than probabilities, Figure 12.1 shows scaled Mahalonbis distances, following the example of Knick and Rotenberry (1998). Dark shades represent low distances (habitat similar to the average used habitat), while light shades represent high distances (habitat rather different from the average used).

12.3 Estimation Using Machine Learning Methods

Two recent papers have used machine learning approaches to model resource selection. Özesmi and Özesmi (1999) consider the choice of nest locations by the red-winged blackbird (*Agelaius phoeniceus*) and the marsh wren (*Cistothorus palustris*) in two diked wetland basins on the southwest of Lake Erie, in the United States, while Kobler and Adamic (2000) considered the choice of habitat by brown bears (*Ursus arctos*) in the southwest of Slovenia.

The situation considered by Özesmi and Özesmi (1999) was where ordinary logistic regression could be used for modelling the presence of either the blackbird or the wren nests. Basically there were a number of locations where the nests could be (the available resource units), of which some were used, and each location was described by six variables (the vegetation durability, the stem density, the stem height, the distance to open water, and the water depth). Logistic regression was used to predict the location of the blackbird nests, and the results were compared with what was obtained using artificial neural networks. For the logistic regression analysis there was stepwise variable selection, while the details of the artificial neural network procedure required several pages of explanation by Özesmi and Özesmi because of the number of choices that had to be made. It was concluded that the artificial neural network procedure was more effective than logistic regression except when wrens were present. an artificial neural network model was therefore developed for the presence of either blackbird or wren nests.

The use of artificial neural networks rather than logistic regression seems at the present time to have both advantages and disadvantages. The advantages are that the artificial neural network approach is very flexible so that it is able to handle highly non-linear relationships, and can outperform logistic regression if this is carried out in a very straightforward way. The disadvantages are that there is no final explicit equation for the resource selection probability function, the software needed is not currently widely available, and considerable knowledge seems to be needed to decide how to run the artificial neural network. It seems that in time the neural network approach may receive wider use, but that at present it is not an easy approach for a novice to use.

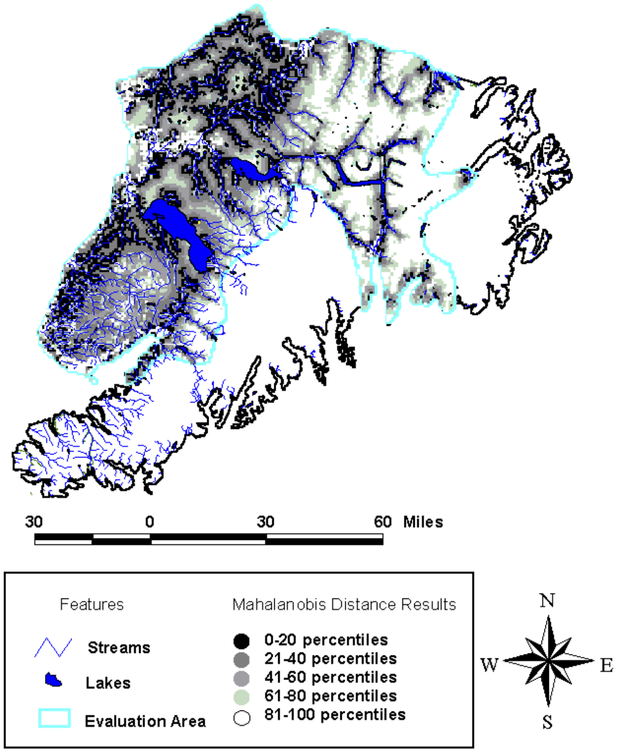


Figure 12.1 Mahalanobis distance map for the Kenai Brown bear example.

Kobler and Adamic (2000) used an inductive learning algorithm to predict the presence of the brown bears in different pixels of a GIS. The presences were determined from bear sightings between 1990 and 1998, and from radio-tracking of some bears from 1993 to 1995. Only sightings of females with cubs were considered, of which there were 830. These used pixels were compared with 825 other pixels chosen by random stratified sampling from all pixels in the study area. Each pixel used in the analysis were described by 36 attributes such at the distance to the nearest human settlement and the percentage of agricultural land.

Kobler and Adamic ran the learning algorithm with 1555 of the pixels that they had data for, and kept 100 pixels aside in order to test the accuracy of their final model. For this sample they found 87% success in correctly classifying pixels.

It is not possible to tell from this study whether there was any advantage in using the learning algorithm rather than logistic regression. Like the artificial neural network approach, special software is needed for running the learning algorithm and this would not be easy for a novice to use. Hence, this is another potentially useful approach for studying resource selection that requires more investigation before it can be recommended for general use.

12.4 Chapter Summary

- Compositional analysis is described as a general approach for comparing the proportions of different types of habitat that are available and used, with an example involving the habitat choice of radio-tagged brown bears on the Kenai Peninsular of Alaska.

- The Mahalanobis distance mapping method is considered. With this method, each resource unit is described by its values for p variables. The mean of these variables for the used units is then considered to represent the idea habitat, and the Mahalonabis distance from a resource unit to the mean is considered to measure the suitability of the unit for use, with small values indicating probable use. Brown bears on the Kenai Peninsular are again used for an example.

- Two recent studies involving artificial neural networks and an inductive machine learning algorithm are to estimate a resource selection function are described. It is concluded that these approaches are potentially useful, but at present relatively difficult to use, and not clearly better than more conventional statistical methods.

CHAPTER 13

SOME APPLICATIONS OF RESOURCE SELECTION FUNCTIONS

Previous chapters have described estimation of resource selection functions for a range of conditions, and some applications such as for mapping areas with relatively good habitat. In this chapter, two additional applications are described. The first makes use of estimated relative probabilities of selection to compute an index of risk associated with possible perturbations of a landscape, while the second uses estimated relative probabilites of selection and the size of habitat patches to the carrying capacity in an area.

13.1 Risk Assessment

Historically, risk assessment with wildlife populations has meant many things. In general the term has referred to the process of judging actions, either past or future, relative to their merit for supporting or harming wildlife. Risk assessments have been undertaken in a wide range of situations including to estimate the harm done to bird populations after the application of a herbicide or pesticide and to estimate the relative worth of clean-up actions at contaminated sites. These risk assessments have usually been undertaken with little field data and relied heavily on expert biological opinion and judgment. Moreover, even when field data were available, it has sometimes been unclear how to incorporate that field data even into a quantitative risk assessment.

In this section, a method is described by which risk assessment can be made quantitative (McDonald and McDonald, 2002). This approach makes use of an estimated resource selection function to ascribe a quantitative, relative worth or risk of future or past actions on a landscape. It is assumed that the resource units under study are plots of land, which will often be pixels of a certain size in a geographic information system (GIS). Alternatively, the plots might be circular buffers surrounding animal locations, or square quadrats laid out in a systematic grid across the landscape. Throughout the rest of this section, these basic units of selection will be called pixels.

The proposal is to assign the relative worth or risk to actions using a resource selection function (RSF) by integrating under the surface that this function defines both before and after the actions of interest take place. In practical terms this integrations is accomplished by calculating the estimated relative probability of selection for every pixel in the landscape, of interest before and after the planned actions, and summing the difference in relative probabilities for the effected area. The resulting measure is then called the relative risk for the actions, which can be positive (beneficial) or negative (harmful).

The validity of this method for assessing risk is based on the assumption that the effects of the planned actions can reliably be quantified by the changes in the variables that are included in the RSF. Clearly, if changes are taking place in other important variables that are not part of the RSF then there must be doubt about whether the risk measure is satisfactory.

Assume that the estimated RSF takes the form

$$\hat{w} = \exp(\hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_p x_p), \tag{13.1}$$

where x_1 to x_p are values for p variables, and that for the ith pixel in the landscape the value of this function is $\hat{w}$ i before the action of interest takes place, and $\hat{w}_i'$ after this action. Then the index of risk is

$$\gamma = \sum (\hat{w}_i' - \hat{w}_i), \tag{13.2}$$

where the summation is over all pixels in the region of interest, where in fact, this can be used just as well with any other form of equation for the RSF.

There is some sampling error in γ due to the errors of estimation in the β values. With the RSF of equation (13.1) a Taylor series expansion gives the approximation

$$\mathrm{var}(\gamma) \approx \sum_{j=1}^{p} \sum_{k=1}^{p} A_j A_k \,\mathrm{cov}(\hat{\beta}_j, \hat{\beta}_k) \tag{13.3}$$

(Manly, 1985, p. 408), where

$$A_j = \sum_{i=1}^{N} (\hat{w}_i' x_{ij}' - \hat{w}_i x_{ij}),$$

there are N pixels in the region of interest, x_{ij} is the value of the jth variable for the ith pixel before the planned action, and x_{ij}' is the value of the jth variable for the ith pixel after the proposed change. With another form of RSF, or if the variances and covariances of the estimated β values are not known, it may be simplest to estimate var(γ) by bootstrapping.

Example 13.1 Risk of Timber Harvest to Spotted Owls

In Example 6.2 a RSF for foraging site selection was estimated for a pair of Northern spotted owls inside their home range approximately 40 miles west of Eugene, Oregon. In the present example the relative risk of three proposed timber harvest plans is evaluated using the method just described. Each timber harvest plan entailed clear-cutting a section of forest within the owl's home range, as shown in Figure 13.1. Actually only the first plan was proposed for the area. The other two clear-cuts have been invented purely for the sake of this example. Relative risks are only considered for the early part of the breeding season, which is May to June.

The resource units were defined to be the 30 m by 30 m pixels within the study area, with each pixel defined to be within one of the four classes of older forest, sapling/broadleaf forest, pole/young forest, or clearcut. The relative risks associated with clear-cutting one of the three areas shown in Figure 13.1 is therefore calculated by (i) calculating the value of the RSF for all pixels in the area and adding these values up, (ii) changing the classification of each of the pixels in the area to clear-cut, recalculating the values of the RSF with this new classification, and adding them up, and (iii) taking the difference between the second sum and the first sum, as shown in equation (13.2).

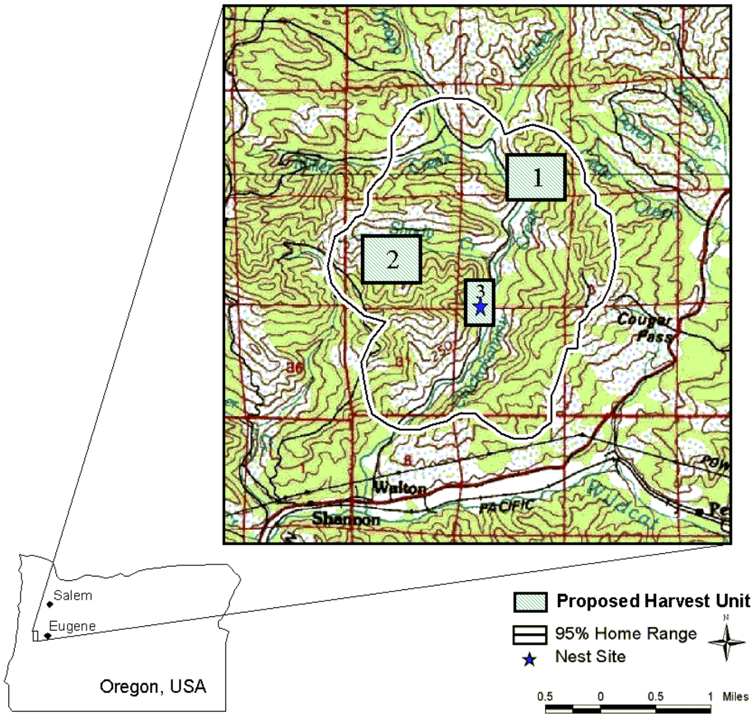


Figure 13.1 The study area showing the home range for a pair of northern spotted owls, and three hypothetical areas for clear-cutting of the forest.

The risk index is -651.8 for area 1, -883.7 for area 2, and -2382.9 for area 3, indicating that the impact of cutting area 3 is much higher than the risk of cutting areas 1 and 2, even although the actual area affected is much smaller for area 3. Another way of expressing this is to say that the estimated impacts of cutting areas 1 and 2 are only 27% and 37%, respectively of the estimated impact of cutting area 3. The high risk index associated with area 3 is not surprising because most of the high values for the RSF are in this area (Figure 13.2). However, the difference in the risk index between areas 1 and 2 is not so obvious.

13.2 Estimating Population Size

Boyce and McDonald (1999) suggest two approaches for estimating a population size, one using a RSF and the other a RSPF. It can be anticipated that the approach based on the RSF will be more widely applicable than the other due to the difficulties that often occur in estimating absolute probabilities of use for resource units.

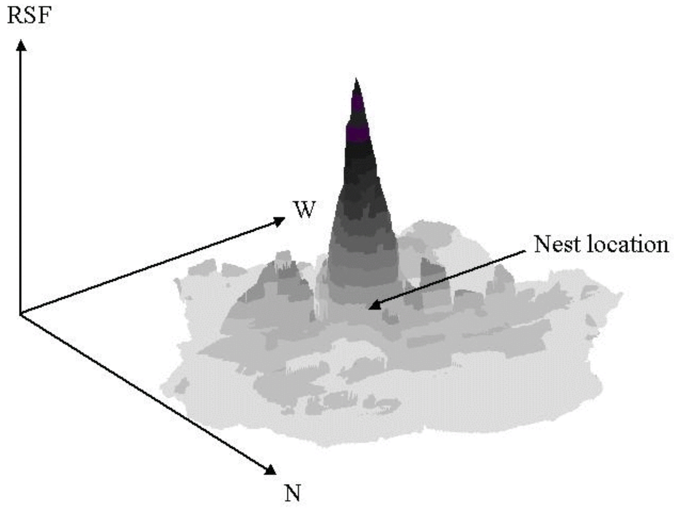


Figure 13.2 Estimated values of the RSF for all pixels in the 95% home range of a pair of northern spotted owls.

The RSF approach is considered first. This relies on a knowledge of the population size in the area where the data for estimating the RSF was obtained, which is considered to be a control area. Under some circumstances the RSF for this area can be applied to estimate the population size in another area. It is assumed that the resource units are the same size in both the control and new area.

First, note that the RSF is proportional to the probability of use of a resource unit. Suppose that N animals are known to exist in the control area where there are m available resource units, with $\hat{w}(\mathbf{x}_i)$ being the estimated value of the RSF for the ith resource unit. Also, let the estimated RSF be scaled so that the sum is one for all units in the control area, i.e.,

$$\sum_{i=1}^{m} \hat{w}(\mathbf{x}_i) = 1.$$

Then it is expected that there will be approximately N $\hat{w}(\mathbf{x}_i)$ animals in the ith resource unit in this area. Hence, if similar selection can be assumed in another area, then an estimate of population size in the new area is

$$\hat{N}' = \sum_{j=1}^{m'} N \ \hat{w}(\mathbf{x}_j), \tag{13.3}$$

where m' is the number of resource units in the new area.

The situation is even simpler if a RSPF is estimated in the control area. In that case the value of this estimated RSPF for the ith unit in the control area, $\hat{w}^*(\mathbf{x}_i)$, is the estimated probability that the unit contains an animal. Assuming that each resource unit can hold only one animal and that the RSPF holds for a new area, leads to the estimation of the population size in the new area by the sum of the values of the RSPF in that area,

$$\hat{N}' = \sum_{j=1}^{m'} \hat{w}^*(\mathbf{x}_j). \tag{13.4}$$

This was the approach used by Boyce and McDonald (1999) for an example involving the estimation of the size of a population of the northern spotted owl (*Strix occidentalis caurina*).

It must be stressed that the use of equations (13.3) and (13.4) for estimating population sizes requires caution. In particular, the availability of different types of resource units will almost certainly vary between the control area and the new area, and this may lead to the selection function changing (Mysterud and Ims, 1999; Boyce *et al.*, 1999).

Chapter Summary

- A method for assessing the risk to wildlife due to changing the habitat in an area is proposed, where the risk index is the difference between the sum of the values of the resource selection function for all resource units after the change, minus the same sum before the change. The use of this method for assessing the relative risks of different proposed changes is illustrated using an example on the effect of timber harvesting on northern spotted owls.

- Under certain circumstances a resource selection function or a resource selection probability function can be used to estimate the size of the animal population in an area. Methods for doing this are described, but the importance of the selection function remaining unchanged between the area where it is determined and the new area where size is to be estimated is stressed.

CHAPTER 14

COMPUTING

Many different analyses are described in the earlier chapters of this book, with the calculations required ranging from very simple to very complicated. In this chapter some guidance is provided concerning how to do these calculations, and, in particular how to do the calculations that are not part of standard statistical packages. Code is also provided for some procedures for the SAS package.

14.1 General Considerations

Chapters 4 to 13 are the ones where specific analyses are discussed in some detail. For convenience, the main types of analysis are listed in Table 14.1, with brief comments about computing considerations.

It is unfortunate that many of the analyses associated with the use of resource selection functions are rather special, and are not readily carried out using standard computer packages. Programs are however available for some of these calculations at the web site www.west-inc.com. In particular this is a source for one-off windows programs for logistic regression, log-linear modelling, the polytomous regression model of Section 6.3 where several samples of used units are compared with a sample of available units, the polytomous regression model of Section 6.4 where several samples of unused units are compared with a sample of available units, the general discrete choice model, and the iterative discrete choice model referred to in Example 8.1. These programs are available "as-is" without guarantees, although every effort has been made to ensure that they produce the correct results.

14.2 Some Examples of SAS Code

Table 14.1 Analyses listed by chapter with brief comments about computing methods.

Chapter	Main Analyses	Computing Methods
4	Ratios and variances of ratios	Can be done in a spreadsheet program
5	Logistic regression	A common option in general statistical packages
6	Proportional hazards models	Available in comprehensive statistical packages
	General maximum likelihood estimation	Requires special programming
	Polytomous regression	Requires special programming as commonly available programs do not do the special calculations needed
7	Log-linear modelling	A common option in general statistical packages
8	Discrete choice modelling	Available in some specialized statistical packages, and other packages can be made to do the necessary calculations
9	Geographical information system calculations	Requires special programming
10	Discriminant function analysis	A common option in general statistical packages
11	Analysis of the amount of use	Some calculations are available in standard statistical packages but others require special programming
12	Compositional analysis	Standard calculations for multivariate analysis of variance
	Mahalonobis distance analysis	Requires special programming
	Artificial neural networks and other machine learning methods	Some special purpose packages are available
13	Risk assessment calculations	Requires special programming
	Estimation of population size	Requires special programming

REFERENCES

Aebischer, M., Robertson, P.A. and Kenword, R.E. (1993). Compositional analysis of habitat use from animal radio-tracking data. *Ecology* 74: 1313-25.

Aebischer, M. and Robertson, P. (1994). Testing for resource use and selection by marine birds: A comment. *Journal of Field Ornithology* 65: 214-20.

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory* (Eds. B.N. Petrov and F. Csaki), pp. 267-81. Akademiai Kiado, Budapest.

Alldredge, A., Deblinger, R. and Peterson, J. (1991). Birth and bedsite selection by pronghorns in a sagebrush steppe community. *Journal of Wildlife Management* 55: 222-7.

Alldredge, J. R. and Ratti, J. T. (1986). Comparison of some statistical techniques for analysis of resource selection. *Journal of Wildlife Management* 50: 157-65.

Alldredge, J. R. and Ratti, J. T. (1992). Further comparison of some statistical techniques for analysis of resource selection. *Journal of Wildlife Management* 56: 1-9.

Alldredge, J. R., Thomas, D. L. and McDonald, L. L. (1998). Survey and comparison of methods for study of resource selection. *Journal of Agricultural, Biological and Environmental Statistics* 3: 237-53.

Anthony, R.M. and Stehn, R.A. (1994). Navigating aerial transects with a laptop computer map. *Wildlife Society Bulletin* 22: 674-6.

Arnett, E., Cook, J., Lindzey, F. and Irwin, L. (1989). *Encampment River bighorn sheep study, June 1987 - December 1988 Summary*. Department of Zoology and Physiology, University of Wyoming, Laramie, Wyoming.

Arthur, S.M., Manly, B.F.J., McDonald, L.L. and Garner, G. W. (1996). Assessing habitat selection when availability changes. *Ecology* 77: 215-27.

Bailey, B., (1980). Large sample simultaneous confidence intervals for the multinomial probabilities based on transformation of the cell frequencies. *Technometrics* 22: 583-9.

Bantock, C., Bayley, J. and Harvey, P. (1976). Simultaneous selective predation on two features of a mixed sibling species population. *Evolution* 29: 636-49.

Beamsderfer, R.C. and Rieman, B.E. (1988). Size selectivity and bias in estimates of population statistics of small mouth bass, walleye, and northern squawfish in a Columbia River reservoir. *North American Journal of Fisheries Management* 8: 505-10.

Belaud, A., Lim, P. and Sabaton, C. (1989). Probability-of-use curves applied to brown trout (*Salmo trutta Fairo* L.) in rivers of southern France. *Regulated Rivers: Research and Management* 3: 321-36.

Belovsky, G., Ritchie, M. and Moorehead, J. (1989). Foraging in complex environments: when prey availablity varies over time and space. *Theoretical Population Biology* 36: 144-60.

Bergin, T. (1992). Habitat selection by the western Kingbird in Western Nebraska: a hierarchical analysis. *The Condor* 94: 903-11.

Bilko, A., Altbacker, V. and Hudson, R. (1994). Transmission of food preference in the rabbit: the means of information transfer. *Physiology and Behavior* 56: 907-12.

Bovee, K.D. (1981). *A User's Guide to the Instream Flow Incremental Methodology*. United States Fish and Wildlife Service Report FWS/OBS-78/07, FWS/OBS-78/07, Fort Collins, Colorado.

Bowyer, R.T. and Bleich, V.C. (1984). Effects of cattle grazing on selected habitats of southern mule deer. *California Fish and Game* 70: 240-7.

Boyce, M.S. and McDonald, L.L., (1999). Relating populations to habitats using resource selection functions. *Trends in Ecology and Evolution* 14: 268-72.

Boyce, M.S., McDonald, L.L. and Manly, B.F.J. (1999). Reply. *Trends in Ecology and Evolution* 14: 490.

Bryant, E. (1973). Habitat selection in a variable environment. *Journal of Theoretical Biology* 41: 421-9.

Burnham, K.P. and Anderson, D.R. (1998). *Model Selection and Inference: a Practical Information-Theoretic Approach*. Springer-Verlag, New York.

Byers, C., Steinhorst, R. and Krausman, P. (1984). Clarification of a technique for analysis of utilization-availability data. *Journal of Wildlife Management* 48: 1050-3.

Cherry, S. (1996). A comparison of confidence interval methods for habitat use-availability studies. *Journal of Wildlife Management* 60: 653-8.

Chesson, J. (1978). Measuring preference in selective predation. *Ecology* 59: 211-5.

Clark, J., Dunn, J.E. and Smith, K. (1993). A multivariate model of female black bear habitat use for a geographic information system. *Journal of Wildlife Management* 57: 519-26.

Clark, R. and Shutler, D. (1999). Avian habitat selection: pattern from process in nest-site use by ducks? *Ecology* 80: 272-87.

Cochran, W.G. (1977). *Sampling Techniques*, 2nd edit. Wiley, New York.

Cock, M.J.W. (1978). The assessment of preference. *Journal of Animal Ecology* 47: 805-16.

Colgon, P.W. and Smith, J.T. (1985). Experimental analysis of food preference transitivity in fish. *Biometrics* 41: 227-36.

Collett, D. (1991). *Modelling Binary Data*. Chapman and Hall, London.

Cooper, A. and Millspaugh, J. (1999). The application of discrete choice models to wildlife resource selection studies. *Ecology* 80: 566-75.

Cox, D. and Hinkley, D. (1974). *Theoretical Statistics*. Chapman and Hall, London.

Danchin, E., Boulinier, T. and Massot, M. (1998). Conspecific reproductive success and breeding habitat selection: implications for the study of coloniality. *Ecology* 79: 2415-28.

Danell, K., Edenius, L. and Lundberg, P. (1991). Herbivory and tree stand composition: moose patch use in winter. *Ecology* 72: 1350-7.

Dasgupta, N. and Alldredge, J. (1998). A multivariate analysis of resource selection data. *Journal of Agricultural, Biological and Environmental Statistics* 3: 323-34.

Dubuc, L.J., Krohn, W. B. and Owen, R. B. (1990). Predicting occurrence of river otters by habitat on Mount Desert Island, Maine. *Journal of Wildlife Management* 54: 594-9.

Dunn, P.O. and Braun, C.E. (1986). Summer habitat use by adult female and juvenile sage grouse. *Journal of Wildlife Management* 50: 228-35.

Durbin, L. (1998). Habitat selection by five otter Lutra lutra in rivers of Northern Scotland. *Journal of Zoology* 245: 85-92.

Edge, W.D., Marcum, C.L. and Olson-Edge, S.L. (1987). Summer habitat selection by elk in Western Montana: a multivariate approach. *Journal of Wildlife Management* 51: 844-51.

Edwards, T.C. and Collopy, M.W. (1988). Nest Tree preference of osprey in north central Florida. *Journal of Wildlife Management* 52: 103-7.

Edwards, G., Newman, J., Parsons, A. and Krebs, J. (1994). Effects of scale and spatial distribution of the food resource and animal state on diet selection: an example with sheep. *Journal of Animal Ecology* 63: 816-26.

Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association* 70: 892-8.

Efron, B. (1979). Bootstrap methods - another look at the jackknife. *Annals of Statistics* 7: 1-26.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman and Hall, New York.

Ellis, J., Wiens, J., Rodell, C. and Anway, J. (1976). A conceptual model of diet selection as an ecosystem process. *Journal of Theoretical Biology* 60: 93-108.

Elston, D., Illius, A. and Gordon, I. (1996). Assessment of preference among a range of options using log ratio analysis. *Ecology* 77: 2538-48.

Emlen, J.M. (1966). The role of time and energy in food preference. *American Naturalist*:100: 611-7.

Erickson, W.P., McDonald, T.L. and Skinner, R., (1998). Habitat selection using GIS: a case study. *Journal of Agricultural, Biological and Environmental Statistics* 3: 296-310.

Fagen, R. (1988). Population effects of habitat change: a quantitative assessment. *Journal of Wildlife Management* 52: 41-6.

Fair, W. and Henke, S. (1997). Effects of habitat manipulations of Texas horned lizards and their prey. *Journal of Wildlife Management* 61: 1366-70.

Feder, J.L., Chilcote, C.A. and Bush, G.L. (1990). The geographic pattern of genetic differentiation between host associated populations of Rhagoletis pomonella (Diptera: Tephritidae) in the eastern United States and Canada. *Evolution* 44: 570-94.

Forsman, E., Meslow, E. and Wight, H.M. (1984). Distribution and biology of the spotted owl in Oregon. *Wildlife Monographs* 87: 1-64.

Francis, B., Green, M. and Payne, C., Eds. (1993). *The GLIM System, Release 4*. Clarendon Press, Oxford.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association* 32: 675-701.

Gervais, J., Noon, B. and Wilson, M. (1999). Avian selection of the colour-dimorphic fruits of salmonberry, *Rubus spectabilis*: a field experiment. *Oikos* 84: 77-86.

Gionfriddo, J.P. and Krausman, P.R., (1986). Summer habitat use by mountain sheep. *Journal of Wildlife Management* 50: 331-6.

Giroux, J.F. and Bedard, J. (1988). Use of bullrush marshes by greater snow geese during staging. *Journal of Wildlife Management* 52: 415-20.

Golet, G. H., Kuletz, K.J., Roby, D.D. and Irons, D.B. (2000). Adult prey choice affects chick growth and reproductive success of pigeon guillemots. *The Auk* 117: 82-91.

Golet, G. H., Seiser, P.E., McGuire, A.D., Rody, D.D., Fischer, J.B., Kuletz, K.J., Irons, D.B., Dean, T.A., Jewett, S.C. and Newman, S.H. (2002). Long-term direct and indirect effects of the *Exxon Valdez* oil spill on Pigeon Guillemots in Prince William Sound, Alaska. *Marine Ecology Progress Series in press*.

Goodman, L.A. (1965). On simultaneous confidence intervals for multinomial proportions. *Technometrics* 7: 247-54.

Green, E. (1973). Location analysis of prehistoric Maya sites in British Honduras. *American Antiquity* 38: 279-93.

Grover, K.E. and Thompson, M.J. (1986). Factors influencing spring feeding site selection by elk in Elkhorn mountains, Montana. *Journal of Wildlife Management* 50: 466-70.

Hall, P. and Wilson, S. (1991). Two guidelines for bootstrap hypothesis testing. *Biometrics* 47: 757-62.

Hamley, J.M. and Regier, H.A. (1973). Direct estimates of gillnet selectivity to walleye (*Stizostedion vitreum vitreum*). *Journal of the Fisheries Research Board of Canada* 30: 817-30.

Haney, J. and Solow, A. (1992). Testing for resource use and selection by marine birds. *Journal of Field Ornithology* 63: 43-52.

Haney, J. (1994). Testing for resource use and selection by marine birds: a reply to Aebischer and Robertson. *Journal of Field Ornithology* 65: 214-20.

Harris, W.F. (1986). *The Breeding Ecology of the South Island Fernbird in Otago Wetlands*. Ph.D. Thesis, University of Otago, Dunedin, New Zealand.

Heisey, D. (1985). Analysing selection experiments with log-linear models. *Ecology* 66: 1744-8.

Hess, A. and Swartz, A. (1940). The forage ratio and its use in determining the food grade of streams. *Transactions of the North American Wildlife Conference* 5: 162-4.

Hjermann, D. Ø. (2000). Analysing habitat selection in animals without well-defined home ranges. *Ecology* 81: 1462-8.

Hobbs, N.T. and Bowden, D.C. (1982). Confidence intervals on food preference indices, *Journal of Wildlife Management* 46: 505-7.

Hobbs, N.T. and Hanley, T.A. (1990). Habitat evaluation: do use/availability data reflect carrying capacity? *Journal of Wildlife Management* 54: 515-22.

Hohf, R.S., Ratti, J.T. and Croteau, R. (1987). Experimental analysis of winter food selection by spruce grouse. *Journal of Wildlife Management* 51: 159-67.

Hohman, W.L. (1985). Feeding ecology of ringnecked ducks in northwestern Minnesota. *Journal of Wildlife Management* 49: 546-57.

Holm, S. (1979). A simple sequential rejective multiple test procedure. *Scandinavian Journal of Statistics* 6: 65-70.

Holmes, R.T. and Robinson, S.K. (1981). Tree species preference of foraging insectivorous birds in a northern hardwoods forest. *Oecologia* 48: 31-5.

Hudgins, J., Storm, G. and Wakely, J. (1985). Local movements and diurnal-habitat selection by male American woodcock in Pennsylvania. *Journal of Wildlife Management* 49: 614-9.

Huegel, C., Dahlgren, R., and Gladfelter, H. (1986). Bedsite selection by white-tailed deer fawns in Iowa. *Journal of Wildlife Management* 50: 474-80.

Hurlbert, S.H. (1984). Pseudoreplication and the design of ecological field experiments. *Ecological Monographs* 54: 187-211.

Iverson, G., Vohs, P., and Tacha, T. (1985). Habitat use by sandhill cranes wintering in western Texas. *Journal of Wildlife Management* 49: 1074-83.

Ivlev, V.S. (1961). *Experimental Ecology of the Feeding of Fishes*. Yale University Press, New Haven.

Jacobs, J. (1974). Quantitative measurement of food selection: a modification of the forage ratio and Ivlev's electivity index. *Oecologia* 14: 412-7.

Jaenike, J. (1980). A Relativistic Measure of Variation in Preference. *Ecology* 61: 990-7.

Jelinski, D. (1991). On the use of chi-square analyses in studies of resource utilization. *Canadian Journal of Forestry Research* 21: 58-65.

Johnson, D. (1980). The comparison of usage and availability measurements for evaluating resource preference. *Ecology* 61: 65-71.

Jolicoeur, P. and Brunel, P. (1966). Application du diagramme hexagonal a l'etude de la selection de ses proies par la Morue. *Vie Milieu B, Oceanography* 17: 419-33.

Judson, O. (1994). The rise of the individual-based model in ecology. *Trends in Ecology and Evolution* 9: 9-14.

Kalmback, E. R. (1934). Field observations in economic ornithology. *Wilson Bulletin* 46: 73-90.

Karanth, K. and Sunquist, M. (1995). Prey selection by tiger, leopard, and dhole in tropical forests. *Journal of Animal Ecology* 64: 439-50.

Keating, K. A., Irby, L. R. and Kasworm, W. F. (1985). Mountain sheep winter food habits in the upper Yellowstone Valley. *Journal of Wildlife Management* 49: 156-61.

Kenward, R.E. (1987). *Wildlife Radio Tagging*. Academic Press, London.

Kenward, R. E. (1992). Quantity versus quality: programmed collection and analysis of radio-tracking data. In *Wildlife Telemetry: Remote Monitoring and Tracking of Animals* (Eds. G. Priede and S.M. Swift), pp. 231-46. Ellis Horwood, New York.

Kincaid, W.B. and Bryant, E.H. (1983). A geometric method for evaluating the null hypothesis of random habitat utilization. *Ecology* 64: 1463-70.

Knick, S.T. and Dyer, D.L. (1997). Distribution of black-tailed jackrabbit determined by GIS in Southwestern Idaho. *Journal of Wildlife Management* 61: 5-85.

Knick, S.T. and Rotenberry, J.T. (1998). Limitations to mapping habitat use areas in changing landscapes using the Mahalanobis distance statistic. *Journal of Agricultural, Biological, and Environmental Statistics* 3: 311-22.

Kobler, A. and Adamic, M. (2000). Identifying brown bear habitat by combining GIS and machine learning method. *Ecological Modelling* 135: 291-300.

Kohler, C. and Ney, J. (1982). A comparison of methods for quantitative analysis of feeding selection of fishes. *Environmental Biology of Fishes* 7: 363-8.

Krueger, W.C. (1972). Evaluating animal forage preference. *Journal of Range Management* 25: 471-5.

Lachenbruch, P.A. (1976). Analysis of data with clumping at zero. *Biometrical Journal* 18: 351-6.

Lagory, M.K., Lagory, K.E., and Taylor, D.H. (1985). Winter browse availability and use by white-tailed deer in southeastern Indiana. *Journal of Wildlife Management* 49: 120-4.

Larsen, D. and Bock, C. (1986). Determining avian habitat preference by bird-centered vegetation sampling. In Wildlife 2000: *Modelling Habitat Relationships of Terrestrial vertebrates* (Eds. J. Verner, M.L. Morrison and C.J. Ralph), pp. 37-43. University of Wisconsin Press, Madison.

Laymon, S.A., Salwasser, H. and Barrett, R. H. (1985). *Habitat Suitability Index Models: Spotted Owl*. United States Fish and Wildlife Service, Biological Services Program, Biological Report 82(10.113), Washington, D. C.

Lechowicz, M. J. (1982). The sampling characteristics of electivity indices. *Oecologia* 52: 22-30.

Levin, S. (1992). The problem of pattern and scale in ecology. *Ecology* 73: 1943-67.

Link, W. and Karanth, K. (1994). Correcting for overdispersion in tests of prey selectivity. *Ecology* 75: 2456-9.

Loehle, C. and Rittenhouse, L. (1982). An analysis of forage preference indices. *Journal of Range Management* 35: 316-9.

MacCallum, C., Nurnberger, B., Barton, N., and Szymura, J. (1998). Habitat preference in the Bombina hybrid zone in Croatia. *Evolution* 52: 227-39.

Manly, B.F.J., Miller, P. and Cook, L. (1972). Analysis of a selective predation experiment. *American Naturalist* 106: 719-36.

Manly, B.F.J. (1973). A linear model for frequency-dependent selection by predators, *Researches on Population Ecology* 14: 137-50.

Manly, B.F.J. (1974). A model for certain types of selection experiments. *Biometrics* 30: 281-94.

Manly, B.F.J. (1985) *The Statistics of Natural Selection on Animal Populations*. Chapman and Hall, London.

Manly, B.F.J. (1992). *The Design and Analysis of Research Studies*. Cambridge University Press, Cambridge.

Manly, B.F.J. (1993). Comments on design and analysis of multiple-choice feeding-preference experiments. *Oecologia* 93: 149-52.

Manly, B.F.J. (1994). *Multivariate Statistical Methods: a Primer*, 2nd ed. Chapman and Hall, London.

Manly, B.F.J. (1995). Measuring selectivity from multiple choice feeding-preference experiments. *Biometrics* 51: 709-15.

Manly, B.F.J. (1997). *Randomization, Bootstrap and Monte Carlo Methods in Biology*, 2nd edit. Chapman and Hall, London.

Manski, C. (1981). Structural models for discrete data: the analysis of discrete choice. In *Sociological Methodology* (Ed. S. Leinhardt), pp 58-109. Jossey-Bass, San Francisco.

Marcum, C. (1975). *Summer-Fall Habitat Selection and Use by a Western Montana Elk Herd*. Ph.D. Thesis, University of Montana, Missoula.

Marcum, C. and Loftsgaarden, D. (1980). A non-mapping technique for studying habitat preferences. *Journal of Wildlife Management* 44: 963-8.

Marzluff, J.M., Knick, S.T., Vekasy, M.S., Schueck, L.S. and Zarrielo, T.J. (1997). Spatial use and habitat selection of golden eagles in Southwestern Idaho. *The Auk* 114: 673-87.

McClean, S., Rumble, M., King, R. and Baker, W. (1998). Evaluation of resource selection methods with different definitions of availability. *Journal of Wildlife Management* 62: 793-801.

McCorquodale, S.M., Raedeke, K.J. and Taber, R. D. (1986). Elk habitat use patterns in the shrub-steppe of Washington. *Journal of Wildlife Management* 50: 664-9.

McCracken, M.L., Manly, B.F.J. and Vander-Heyden, M. (1998). The use of discrete-choice models for evaluating resource selection. *Journal of Agricultural, Biological, and Environmental Statistics* 3: 268-79.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd edit. Chapman and Hall, London.

McDonald, T.L. and McDonald, L.L. (2002). A new ecological risk assessment procedure using resource selection models and geographical information systems. *Wildlife Society Bulletin* (in press).

McFadden, D. (1973). Conditional logit analysis of qualitative choice behaviour. In *Frontiers in Econometrics* (Ed. P. Zarembka), pp. 105-42. Academic Press, New York.

McIntosh, A., Townsend, C. and Crowl, T. (1992). Competition for Space between introduced brown trout (*Salmo trutta* L.) and a native galaxiid (*Galaxias vulgaris Stokell*) in a New Zealand stream. *Journal of Fish Biology* 41: 63-81.

McKnight, S. and Hepp, G. (1998). Diet selectivity of gadwalls wintering in Alabama, *Journal of Wildlife Management* 62: 1533-43.

Mielke, P.W. (1986). Non-metric statistical analyses: some metric alternatives. *Journal of Statistical Planning and Inference* 13: 377-87.

Millar, R.B. and Walsh, S.J. (1992). Analysis of trawl selectivity studies with an application to trouser trawls. *Fisheries Research* 13: 205-20.

Minitab, Inc. (1997). *Minitab User's Guide 2: Data Analysis and Quality Tools*. Minitab Inc., State College, Pennsylvania.

Mladenoff, D., Sickley, T. and Wydeven, A. (1999). Predicting gray wolf landscape recolonization: logistic regression models vs. New field data. *Ecological Applications* 9: 37-44.

Munro, H. and Rounds, R. (1985). Selection of artificial nest sites by five sympatric passerines. *Journal of Wildlife Management* 49: 125-8.

Murphy, R.K., Payne, N.F. and Anderson, R. K. (1985). White-tailed deer use of an irrigated agricultural-grassland complex in central Wisconsin. *Journal of Wildlife Management* 49: 125-8.

Mysterud, A. and Ims, R.A. (1998). Functional responses in habitat use: availability influences relative use in trade-off situations. *Ecology* 79: 1435-41.

Mysterud, A. and Ims, R.A. (1999). Relating populations to habitat. *Trends in Ecology and Evolution* 14: 489-90.

Nams, V. O. (1989). Effects of radiotelemetry error on sample size and bias when testing for habitat selection. *Canadian Journal of Zoology* 67: 1631-6.

Nelson, J.R. (1978). Maximizing mixed animal species stocking rates under proper-use management. *Journal of Wildlife Management* 42: 172-4.

Nelson, M.E. (1979). *Home Range Location of White-Tailed Deer*. United States Forest Service Research Paper, NC-173, NC-173.

Neu, C., Byers, C. and Peek, J. (1974). A technique for analysis of utilization-availability data. *Journal of Wildlife Management* 38: 541-5.

Nicholson, W. (1990). *Intermediate Microeconomics and its Application*, 5th edit. Dryden Press, Chicago.

North, M. and Reynolds, J. (1996). Microhabitat analysis using radiotelemetry locations and polytomous logistic regression. *Journal of Wildlife Management* 60: 639-53.

Nudds, T. (1980). Forage 'preference': theoretical considerations of diet selection by deer. *Journal of Wildlife Management* 44: 735-9.

Nudds, T. (1982). Theoretical considerations of diet selection by deer: a reply. *Journal of Wildlife Management* 46: 257-8.

Orians, G. and Wittenberger, J. (1991). Spatial and temporal scales in habitat selection. *American Naturalist* 137: 29-49.

Otis, D. (1997). Analysis of habitat selection studies with multiple patches within cover types. *Journal of Wildlife Management* 61: 1016-22.

Otis, D. (1998). Analysis of the influence of spatial pattern in habitat selection studies. *Journal of Agricultural, Biological and Environmental Statistics* 3: 254-67.

Owen-Smith, N. and Cooper, S. M. (1987). Assessing food preferences of ungulates by acceptability indices. *Journal of Wildlife Management* 51: 372-8.

Özesmi, S.L. and Özesmi, U. (1999). An artificial neural network approach to spatial habitat modelling with interspecific interaction. *Ecological Modelling* 116: 15-31.

Paloheimo, J. E. (1979). Indices of food type preference by a predator. *Journal of the Fisheries Research Board of Canada* 36: 470-3.

Palomares, F. and Delibes, M. (1992). Data analysis and potential bias in radio-tracking studies of animal use. *Acta Oecologia* 13: 221-6.

Parsons, B. and Hulbert, W. (1988). Influence of habitat availability on spawning site selection by kokanees in streams. *North American Journal of Fisheries Management* 8: 426-31.

Pearre, S. (1982). Estimating prey preference by predators: uses of various indices, and a proposal of another based on χ^2. *Canadian Journal of Fisheries and Aquatic Sciences* 39: 914-23.

Peek, J. M. (1986). *A Review of Wildlife Management*. Prentice-Hall, New Jersey.

Pendleton, G., Titus, K., Lowell, R., Degayner, E. and Flatten, C. (1998). Compositional analysis and GIS for study of habitat selection by goshawks in southeast alaska. *Journal of Agricultural, Biological and Environmental Statistics* 3: 280-95.

Peres-Neto, P.R. (1999). How many tests are too many? The problem of conducting multiple ecological inferences revisited. *Marine Ecology Progress Series* 176: 303-6.

Petersen, M.R. (1990). Nest-site selection by emperor geese and cackling Canada geese. *Wilson Bulletin* 102: 413-6.

Petrides, G. (1975). Principal foods versus preferred foods and their relations to stocking rate and range condition. *Biological Conservation* 7: 161-9.

Pietz, P.J. and Tester, J.R. (1982). Habitat selection by sympatric spruce and ruffed grouse in north central Minnesota. *Journal of Wildlife Management* 46: 391-403.

Pietz, P.J. and Tester, J.R. (1983). Habitat selection by snowshoe hares in north central Minnesota. *Journal of Wildlife Management* 47: 686-96.

Pinkas, L., Oliphant, M. S. and Iverson, I. L. K. (1971). *Food Habits of Albacore, Bluefin Tuna, and Bonito in California Waters*. Fisheries Bulletin 152, California Department of Fisheries and Game.

Popham, E. (1944). A study of the changes in an aquatic insect population using minnows as the predators. *Proceedings of the Zoological Society of London* A114: 74-81.

Porter, W.F. and Labisky, R.F. (1986). Home range and foraging habitat of red-cockaded woodpeckers in northern Florida. *Journal of Wildlife Management* 50: 239-47.

Porter, W.F. and Church, K.E. (1987). Effects of environmental pattern on habitat preference analysis. *Journal of Wildlife Management* 51: 681-5.

Prevett, J.P., Marshall, I.F. and Thomas, V.G. (1985). Spring foods of snow and Canada geese at James Bay. *Journal of Wildlife Management* 49: 558-63.

Pyke, G., Pulliam, H. and Charnov, E. (1977). Optimal foraging: a selective review of theory and tests. *Quarterly Review of Biology* 52: 137-54.

Quade, D. (1979). Using weighted rankings in the analysis of complete blocks with additive block effects. *Journal of the American Statistical Association* 74: 680-3.

Rachlin, J. W., Pappantoniou, A. and Warkentine, B. E. (1987). A bias estimator of the environmental resource base in diet preference studies with fish. *Journal of Freshwater Ecology*. 4: 23-31.

Raley, C. and Anderson, S. (1990). Availability and use of arthropod food resources by Wilson's warblers and Lincoln's sparrows in southeastern Wyoming. *Condor* 92: 141-50.

Rapport, D. and Turner, J. (1970). Determination of predator food preferences. *Journal of Theoretical Biology* 26: 365-72.

Rapport, D. (1980). Optimal foraging for complementary resources. *American Naturalist* 116: 324-46.

Ready, R.C., Mills, E.L. and Confer, J.L. (1985). A new estimator of, and factors influencing, the sampling variance of the linear index of food selection. *Transactions of the American Fisheries Society* 114: 258-66.

Rettie, W.J. and McCloughlin, P.D. (1999). Overcoming radiotelemetry bias in habitat selection studies. *Canadian Journal of Zoology* 77: 1175-84.

Rexstad, E., Miller, D., Flather, C., Anderson, E., Hupp, J. and Anderson, D. (1988). Questionable multivariate statistical inference in wildlife habitat and community studies. *Journal of Wildlife Management* 52: 794-8.

Rich, T. (1986). Habitat and nest-site selection by borrowing owls in the sagebrush steppe of Idaho. *Journal of Wildlife Management* 50: 548-55.

Roa, R. (1992). Design and analysis of multiple-choice feeding preference experiments. *Oecologia* 89: 509-15.

Rodgers, A. (1990). Evaluating preference in laboratory studies of diet selection. *Canadian Journal of Zoology* 68: 188-90.

Rolley, R. E. and Warde, W. D. (1985). Bobcat habitat use in southeastern Oklahoma. *Journal of Wildlife Management* 49: 913-20.

Rondorff, D., Gray, G. and Fairley, R. (1990). Feeding ecology of subyearling Chinook Salmon in riverine and reservoir habitats of the Columbia River. *Transactions of the American Fisheries Society* 119: 16-24.

Rosenberg, D. and McKelvey, K. (1999). Estimation of habitat selection for central-place foraging animals. *Journal of Wildlife Management* 63: 1028-38.

Rosenzweig, M. L. (1981). A theory of habitat selection. *Ecology* 62: 327-335.

Roy, L. D. and Dorrance, M. J. (1985). Coyote movements, habitat use, and vulnerability in central Alberta. *Journal of Wildlife Management* 49: 307-13.

Ryder, T.J. (1983). *Winter Habitat Selection by Pronghorn in South Central Wyoming*. M.S. Thesis, Department of Zoology and Physiology, University of Wyoming, Laramie, Wyoming.

SAS Institute Inc. (2000). *SAS Technical Report TS-621, SAS/STAT Software: Changes and Enhancements, Release 8*. SAS Institute Inc., Cary, North Carolina.

Savage, R. E. (1931). The relation between the feeding of the herring off the east coast of England and the plankton of the surrounding waters. *Fishery Investigation, Ministry of Agriculture, Food and Fisheries, Series 2*, 12: 1-88.

Schoen, J. W. and Kirchhoff, M. D. (1985). Seasonal distribution and home range patterns of Sitka black-tailed deer on Admiralty Island, southeast Alaska. *Journal of Wildlife Management* 49: 96-103.

Schooley, R. (1994). Annual variation in habitat selection: patterns concealed by pooled data. *Journal of Wildlife Management* 58: 367-74.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* 6: 461-4.

Scott, A. (1920). Food of Port Erin mackerel in 1919. *Proceedings and Transactions of the Liverpool Biological Society* 34: 107-11.

Seber, G. (1984). *Multivariate Observations*. Wiley, New York.

Sheppard, P.M. (1951). Fluctuations in the selective value of certain phenotypes in the polymorphic land snail *Cepaea nemoralis* (L.). *Heredity* 5: 125-34.

Sherratt, T. and MacDougall, A. (1995). Some population consequences of variation in preference among individual predators. *Biological Journal of the Linnean Society* 55: 93-107.

Smith, R.E., Hupp, J.W. and Ratti, J.T. (1982). Habitat use and home range of grey partridge in eastern South Dakota. *Journal of Wildlife Management* 46: 580-7.

Spitz, F. and Janeau, G., (1995). Daily selection of habitat in wild boar (*Sus scrofa*). *Journal of Zoology* 237: 423-34.

Stauffer, D. F. and Peterson, S. (1985). Ruffed and blue grouse habitat use in southeastern Idaho. *Journal of Wildlife Management* 49: 459-66.

Stepto, N. and Cook, P. (1996). Feeding preferences of the juvenile South African abalone (*Haliotis midae*) (Linnaeus,1758). *Journal of Shellfish Research* 15: 653-7.

Stinnett, D. and Klebenow, D. (1986). Habitat use of irrigated lands by California quail in Nevada. *Journal of Wildlife Management* 50: 368-72.

Strauss, R. E. (1979). Reliability estimates for Ivlev's electivity index, the forage ratio, and a proposed linear index of food selection. *Transactions of the American Fisheries Society* 108: 344-52.

Swihart, R. and Slade, N. (1985). Testing for independence of observations in animal movements. *Ecology* 66: 1176-84.

Swihart, R. and Slade, N. (1997). On testing of independence of animal movements. *Journal of Agricultural, Biological and Environmental Statistics* 2: 48-63.

Talent, L. G., Krapu, G. L. and Jarvis, R. L. (1982). Habitat use by mallard broods in south central North Dakota. *Journal of Wildlife Management* 46: 629-35.

Thomas, D. and Taylor, E. (1990). Study designs and tests for comparing resource use and availability. *Journal of Wildlife Management* 54: 322-30.

Thomasma, L. E., Drummer, T. D. and Peterson, R. (1991). Testing the Habitat Suitability Index for the fisher. *Wildlife Society Bulletin* 19: 291-7.

Thompson, J. N. (1988). Evolutionary ecology of the relationship between oviposition preference and performance of offspring in phytophagus insects. *Entomologia Experimentalis et Applicata* 47: 3-14.

United States Fish and Wildlife Service (1981). *Standards for the Development of Suitability Index Models*. Ecology Service Manual 103, Division of Ecological Services, Washington, D.C.

Vanderploeg, H. and Scavia, D. (1979a). Two electivity indices for feeding with special reference to zooplankton grazing. *Journal of the Fisheries Research Board of Canada* 36: 362-5.

Vanderploeg, H. and Scavia, D. (1979b). Calculation and use of selectivity coefficients of feeding: zooplankton grazing. *Ecological Modelling* 7: 135-49.

Van Horne, B. (1983). Density as a misleading indicator of habitat quality. *Journal of Wildlife Management* 47: 893-901.

Walsh, S., Millar, R., Cooper, C. and Hickey, W. (1992). Codend selection in American plaice: diamond versus square mesh. *Fisheries Research* 13: 235-54.

Wang, J. and Provenza, F. (1996). Food preference and acceptance of novel foods by lambs depend on the composition of the basal diet. *Journal of Animal Science* 74: 2349-54.

Werner, E. and Hall, D. (1974). Optimal foraging and the size selection of prey by the bluegill sunfish (*Lepomis macrochirus*). *Ecology* 55: 1042-52.

White, R. G. and Trudell, J. (1980). Habitat preference and forage consumption by reindeer and caribou near Atkasook, Alaska. *Arctic and Alpine Research* 12: 511-29.

White, G. and Garrott, R. (1990). *Analysis of Wildlife Radio-Tracking Data*. Academic Press, New York.

Whitham, T. G. (1980). The theory of habitat selection: examined and extended using Pemphigus aphids. *American Naturalist* 115: 449-66.

Wiens, J. A. (1981). Scale problems in avian censusing. In *Estimating Numbers of Terrestrial Birds* (Eds. C.J. Ralph and J.M. Scott), pp. 513-21. Studies in Avian Biology 6, Cooper Ornithological Society.

Wong, B. and Ward, F. (1972). Size selection of Daphnia publicaria by yellow perch (*Perca flavescens*) fry in West Blue Lake, Manitoba. *Journal of the Fisheries Research Board of Canada* 29: 1761-4.

Zaret, T.M. and Kerfoot, W.C. (1975). Fish predation on Bosmina longirostris: body-size selection versus visibility selection. *Ecology* 56: 223-37.

NAME INDEX

SUBJECT INDEX