# CHAPTER 5

# RESOURCE SELECTION FUNCTIONS FROM 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

When using logistic regression with census data the assumption made is that:

> *There are N available resource units and it is known which of these have been used and which have not been used after a single period of selection.*

In this case 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. The RSPF is simply assumed to take the form

$$w^*(\mathbf{x}) = \frac{\exp(\beta_0 + \beta_1 x_1 + \ ... \ + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + ... + \beta_p x_p)}, \tag{5.1}$$

where $\mathbf{x} = (x_1, x_2, ..., x_p)$ 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

$$w^*(\mathbf{x}) = \Phi(\beta_0 + \beta_1 x_1 + ... + \beta_p x_p), \tag{5.2}$$

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

$$w^*(\mathbf{x}) = 1 - \exp\{-\exp(\beta_0 + \beta_1 x_1 + ... + \beta_p x_p)\} \tag{5.3}$$

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 $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ 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 $ß_1 = ß_2 = \ldots = ß_p = 0$, so that

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

and the deviance for a 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 as a function of the variables X variables 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 $\mathbf{x} = (x_1, x_2, \ldots, x_p)$ 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. In general, it is preferable that the logistic regression be justified by the design of the study rather than by a model based assumption.

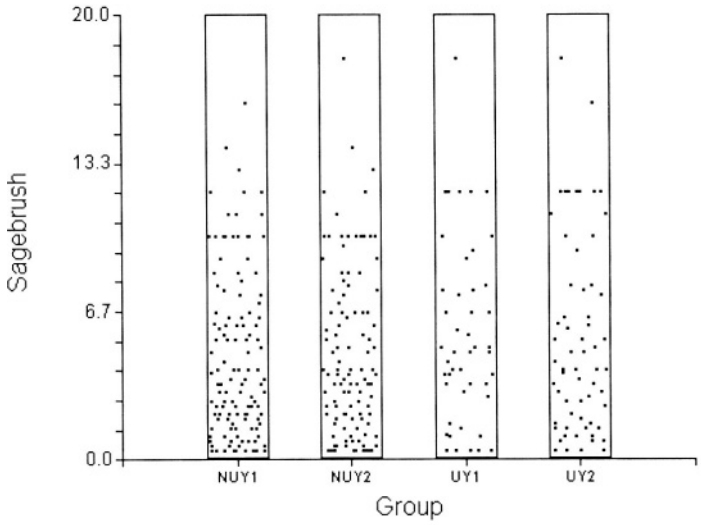
### Example 5.1 Habitat Selection by Pronghorn Antelope

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 pronghorn 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.

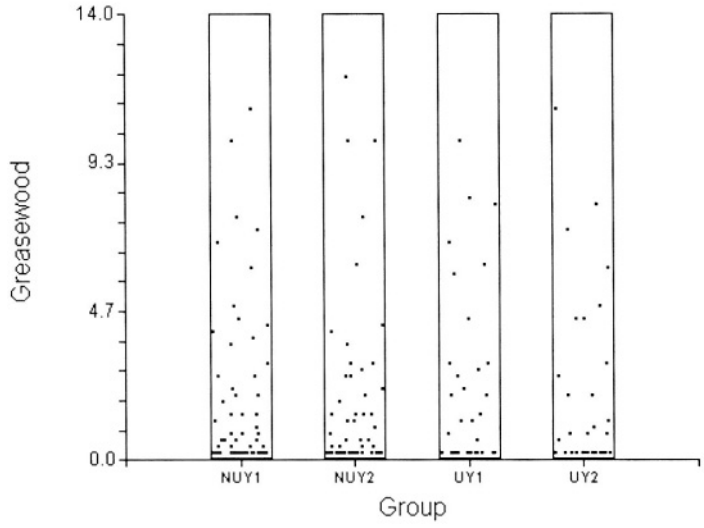
The area considered by Ryder consists of alternating blocks of public and private land, and his study plots are a systematic sample of 10% of the public land. There are therefore three possibilities in terms of the population of resource units that an estimated resource selection function applies to. 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 RSPF applies to all public land. In fact the analysis given in Sections 5.1 and 5.2 under the assumption that the 256 plots are a random sample of plots from the public land is likely conservative, because a systematic sample is often more representative than a simple random sample.

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.

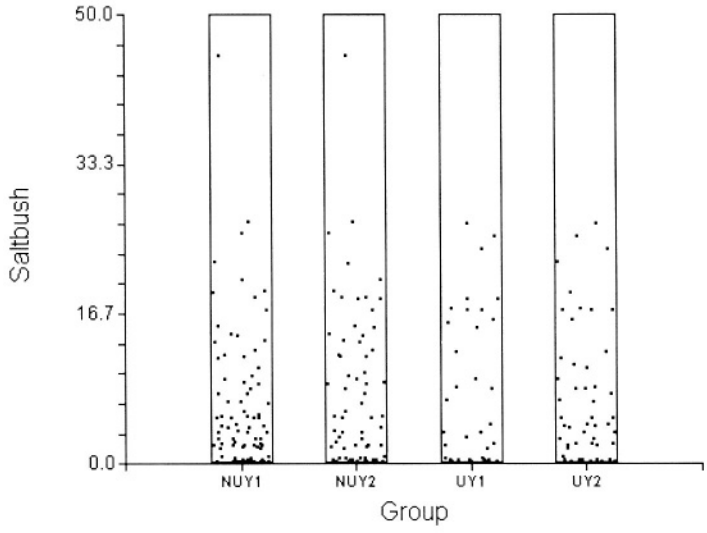
*(a) Density of Sagebrush (thousands/ha)*



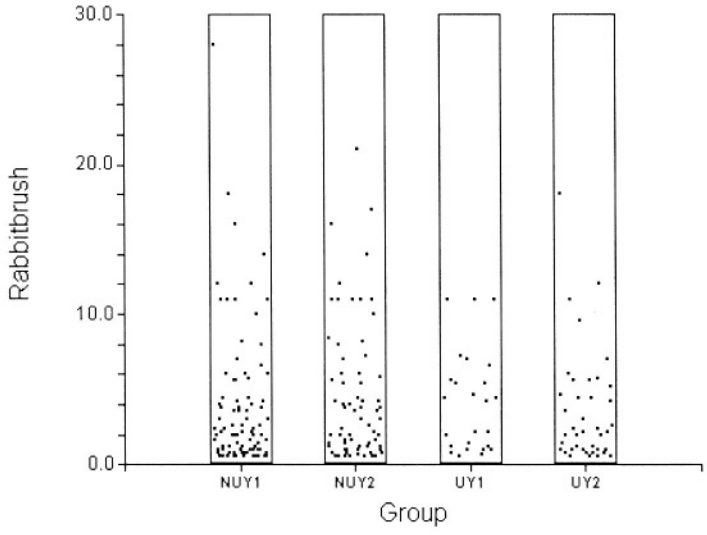

*(b) Density of Greasewood (thousands/ha)*



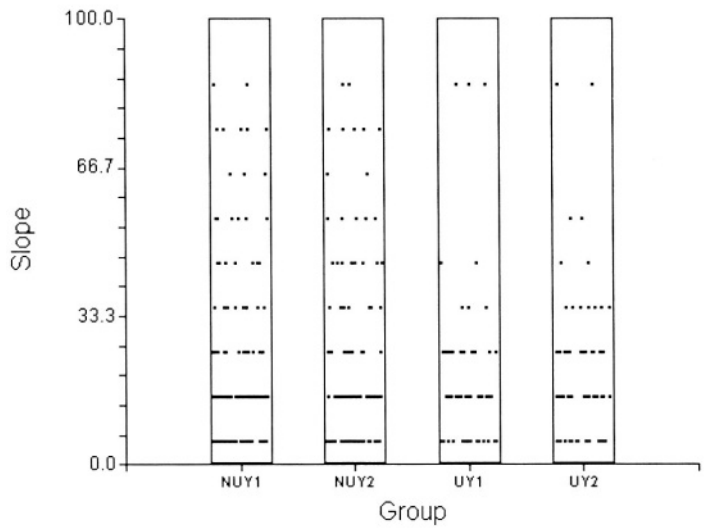
*(c) Density of Saltbrush (thousands/ha)*



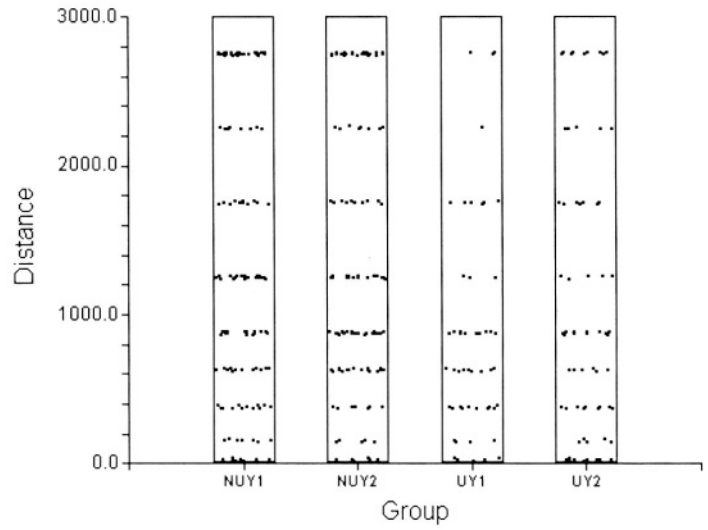

*(d) Density of Rabbitbrush (thousands/ha)*



*(e) Slope of Plot (Degrees)*




*(f) Distance to Water (m)*



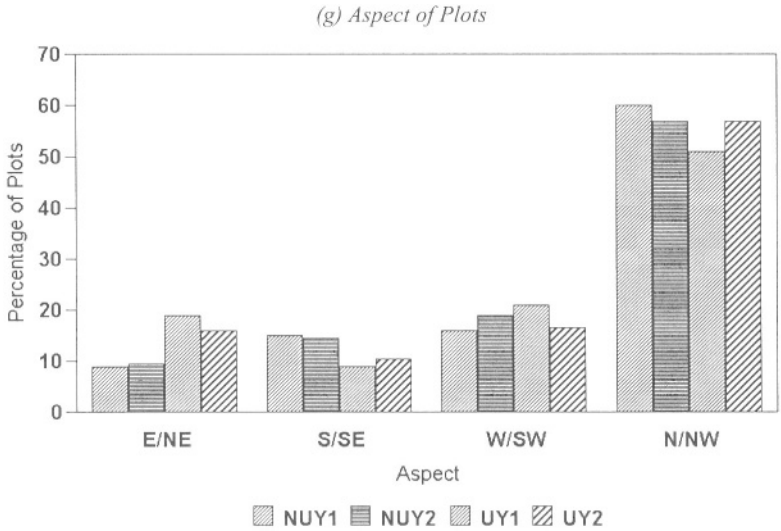



*Figure 5.1 Distributions of the variables shown in Table 3.2 for unused plots in 1980-81 (NUY1), unused plots in 1981-82 (NUY1), used plots in 1980-81 (UY1) and used plots in 1981-82 (UY2). Dotplots are used to represent the distribution of the variables other than aspect, with a dot for each data point. 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 a more complete analysis 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 | 0.41 | 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 | 0.00018 | 0.051 | 0 | 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 | 0.49 |
| Distance to water | 0 | 0.0002 | 0.01 | 0 | 0.0002 | 0.03 |
| E/NE aspect | 1.036 | 0.432 | 0.017 | 0.561 | 0.418 | 0.18 |
| S/SE aspect | -0.086 | 0.504 | 0.865 | -0.043 | 0.43 | 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 | 0.41 | 0.194 | 0.035 |
| Distance to water | 0 | 0.0001 | 0.001 |
| E/NE aspect | 0.786 | 0.301 | 0.009 |
| S/SE aspect | -0.059 | 0.325 | 0.86 |
| W/SW aspect | 0.18 | 0.26 | 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 $A_i w^*(x_i)\{1 - w^*(x_i)\}$ being five or more, is certainly not met. In fact, in this example, $A_i = 1$ for each observation. 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.*

| Model | Deviance | df | Change in Deviance | df | AIC |
|---|---|---|---|---|---|
| No selection and different probabilities of use for each winter | 640.81 | 510 | | | 644.8 |
| | | | 19.52 | 4 | |
| Model 3: selection on distance to water and aspect, plus a year difference | 621.31 | 506 | | | 633.3 |
| | | | 3.2 | 4 | |
| Model 2: selection on distance to water and aspect, with effects varying with the year | 618.11 | 502 | | | 638.1 |
| | | | 8.1 | 10 | |
| Model 1: selection on all variables, with effects varying with the year | 610.01 | 492 | | | 650 |

[1]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 for differences of deviances).

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

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

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 on public land 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

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

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 on public land 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:

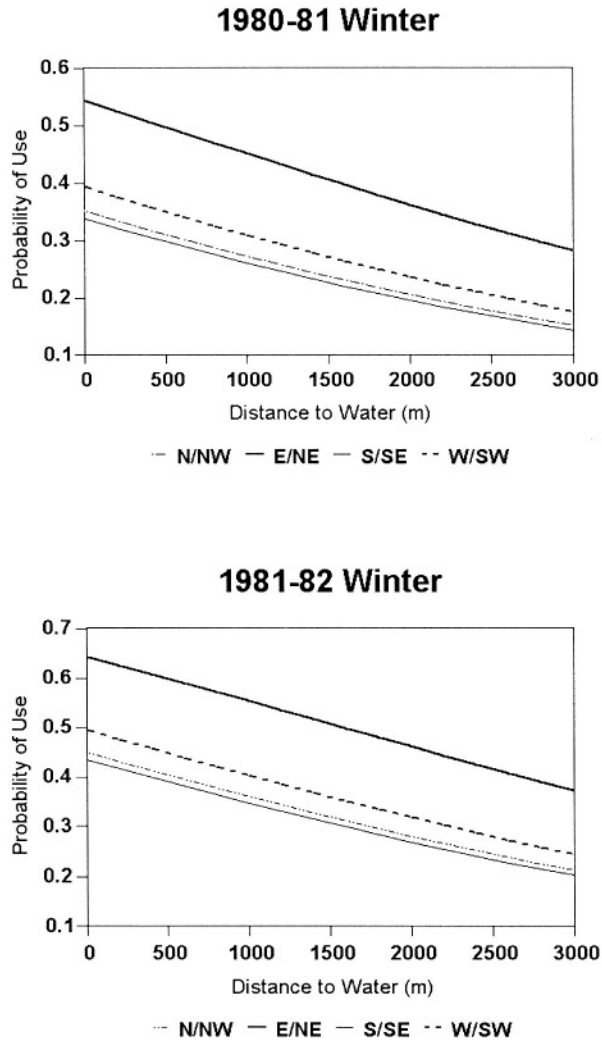
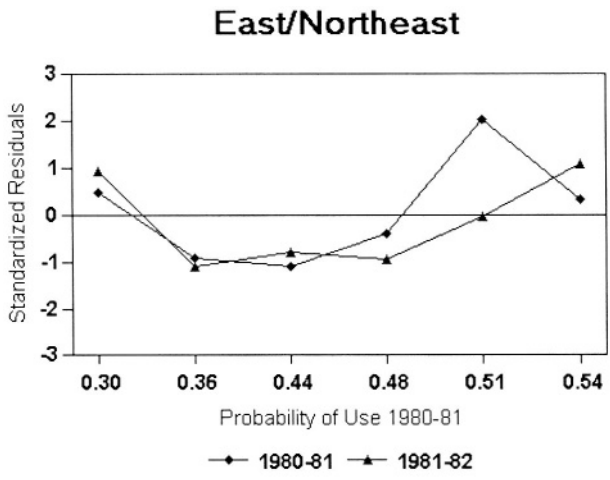
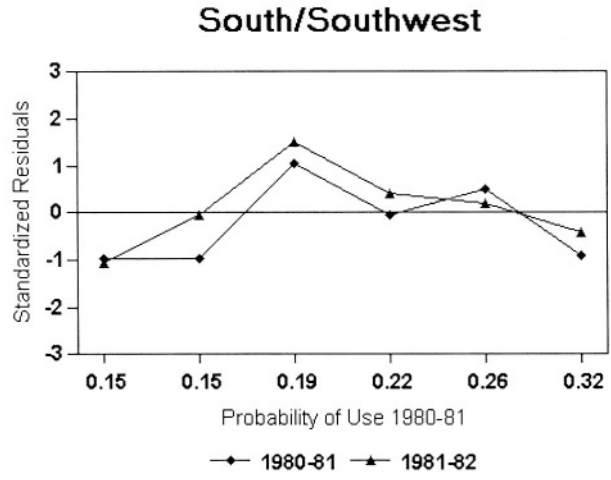



*Figure 5.2 Probabilities of study plots being used by antelope according to the resource selection probability function (5.5).*


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


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

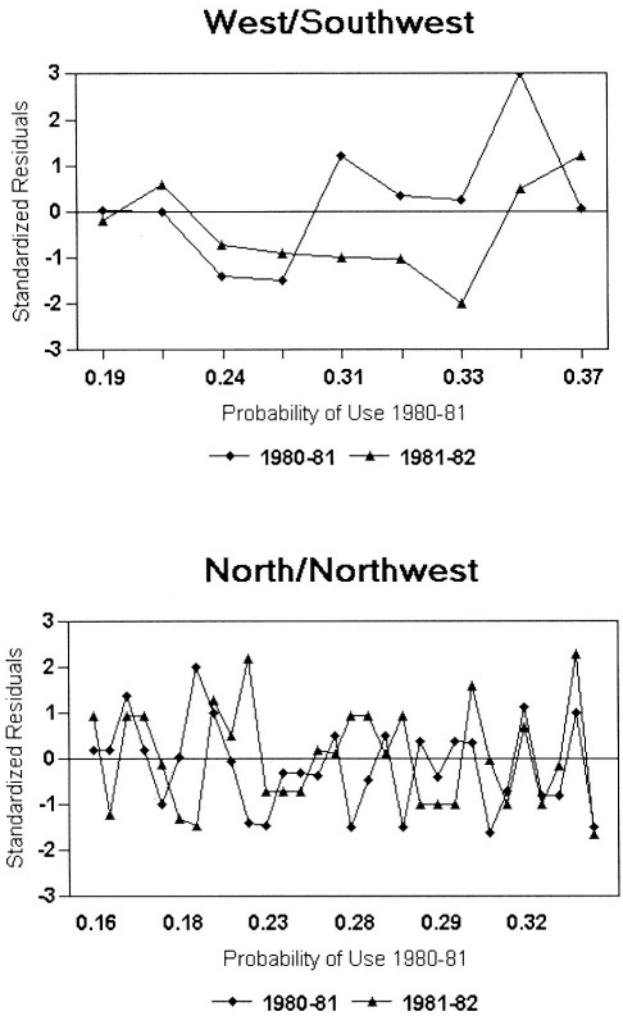



*Figure 5.3 Standardized residuals plotted against the estimated probability of use in 1980-81, separately for each of the four aspects.*

(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: available, used, and unused. 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 Units

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

- 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. The assumption will also be made initially 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, with 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

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

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

$$\text{Prob(ith unit used|sampled)} = \text{Prob(used and sampled)/Prob(sampled)}$$

$$= (1 - P_a)w^*(\mathbf{x}_i)P_u/\{P_a + (1 - P_a)w^*(\mathbf{x}_i)P_u\}. \qquad (5.8)$$

It is convenient at this point to also assume that the resource selection probability function takes a particular exponential form of function which is

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

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

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

This is a logistic regression equation in which the parameter $\beta_0$ 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 units.

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,

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

This log likelihood is identical to a binomial log likelihood with the number of trials set to 1. Consequently, standard logistic regression computer programs can be used to estimate the coefficients $\beta_0, \beta_1, \ldots, \beta_p$ of the log-linear function (5.9), with variances.

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 $\beta_0$ in the resource selection probability function (5.9) can be estimated by subtracting the quantity $\log_e[(1 - P_a)P_u/P_a]$ from the estimated constant in the logistic regression equation. If the sampling fractions are not known then $\beta_0$ cannot be estimated, but it is still possible to estimate the resource selection function (RSF)

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

and use this to compare resource units.

It is critical to note that the correct probabilities of use or relative probabilities of use are given by substituting the estimates of $\beta_0, \beta_1, \ldots, \beta_p$ into the log-linear functions (5.9) or (5.11). The probabilities obtained from the logistic regression computer program, $\tau(\mathbf{x}_i)$ in (5.10), are not correct estimates for the resource probability selection, $w^*(\mathbf{x}_i)$, or resource selection function, $w(\mathbf{x}_i)$, although in the past this has sometimes been assumed to be the case.

The sampling scheme used to derive equation (5.11) 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 variables 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 aprobability $P_a$ 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, ifthe systematic sample is large enough then it should represent the population ofavailable 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 make sure that for the final size used the results obtained are very similar with alternative samples.

A common situation occurs when the proportion of used units is ‘small’ compared to the overall population of N available units. In this case, the likelihood of overlap of the available used samples is small and the results will be approximately correct even if there is minor overlap of the samples. If there is concern about not being able to count some units as used because they appear in the available sample first, then the sampling scheme discussed in Section 5.6 might be preferred to the one considered in this section, with separate samples being taken of used and unused units. Unfortunately,

if the used sample is taken first without replacement then a conditional probability argument similar to that leading to equation (5.8) does not lead to a standard logistic regression equation unless $P_a$ is very small, in which case equation (5.10) can still be applied with $1 - P_a$ set equal to 1.

## 5.5 Separate Samples of Available and Unused 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.

In such situations the assumption will be made 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 $P_u$ 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 $(X_1, X_2, ..., X_p)$.

Then the probability of the ith resource unit in the population being in one of the samples is

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

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

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

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

$$\tau(x_i) = \frac{\exp\{\log_e[(1 - P_a)P_u / P_a] + \beta_0 + \beta_1 x_{i1} + ... + \beta_{ip} x_{ip}\}}{1 + \exp\{\log_e[(1 - P_a)P_u / P_a) + \beta_0 + \beta_1 x_{i1} + ... + \beta_{ip} x_{ip}\}} \qquad (5.15)$$

This is a model that can be fitted using standard logistic regression programs, 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

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

It follows that $\beta_0$ 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

$$\exp(\beta_1 x_1 + ... + \beta_p x_p)$$

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

The comments at the end of Section 5.4 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. If the sampling probabilities are small then there is little likelihood of overlap of the samples and the results are approximately correct.

## 5.6 Separate Samples of Used and Unused 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 taken of used prey items. These samples are compared to see which types of prey are selected by the animals. If stomach samples are taken then the study will necessarily have design II or III.

- A study area is divided into plots where it is easy to see which plots have been used or unused 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.

In these situations the following assumption will be made:

*Samples are 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 $P_u$ 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

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

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

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

This can be rewritten as

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

which defines a logistic regression function by setting

$$(P_u/P_{\bar{u}})w^*(\mathbf{x}_i)/\{1 - w^*(\mathbf{x}_i)\} = \exp(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}),$$

in which case

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

and

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

The parameters in equation (5.18) may be estimated using a logistic regression program with 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 or not. 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 5.4 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 and there is very little chance of overlap in the two samples. We can regard the sample of available points as being sampled with replacement.

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 as a function of CANOPY, EDGE and PERIM.

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(5.9) or (5.11), with all three variables included.

The fitted logistic regression equation corresponds to equation (5.10) and is only a convenient 'trick' to obtain the coefficients of (5.9) or (5.11) and their standard errors by use of handy computer software packages. The RSPF would therefore be estimated by equation (5.9)

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

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, (5.11),

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

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 for nest sites that have larger values for CANOPY, EDGE and PERIM.

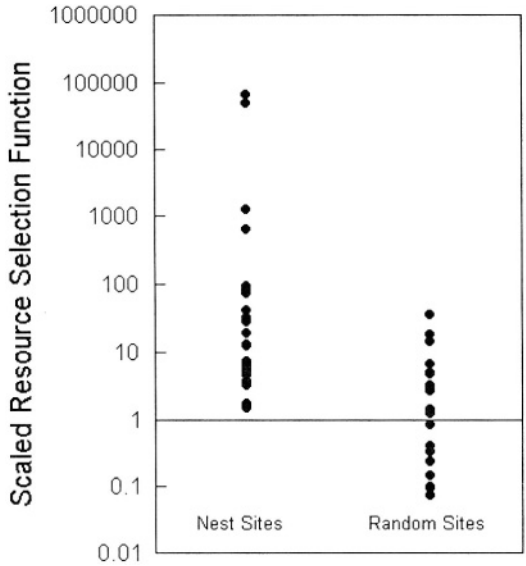



*Figure 5.4 Values of the resource selection function estimated for fernbirds, with the values scaled so that the value is 1.0 for a site with the average values of the predictor variable at the randomly located sites.*

## 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 statistical methods that can be used to assess the accuracy of estimates of exponential functions and their differences.

First, suppose we have census data on used and unused units as in Section 5.1 and Example 5.1 and 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

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

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

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

taking $x_0 = 1$. Here $\text{cov}(\hat{\beta}_i, \hat{\beta}_j)$ is the variance of $\hat{\beta}_i$ if i=j, or is otherwise the covariance between $\hat{\beta}_i$ and $\hat{\beta}_j$, 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, $w^*(\mathbf{x})$ in (5.21) 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

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

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

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

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

as in Sections (5.3) through (5.6) and Example 5.2. Then the Taylor series method gives the approximation

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

and

$$\text{var}\{\hat{w}^*(\mathbf{x}_1) - \hat{w}^*(\mathbf{x}_2)\} = w^*(\mathbf{x}_1)\, w^*(\mathbf{x}_2) \sum_{i=0}^{p} \sum_{j=0}^{p} x_{1i} x_{2j} \text{cov}(\hat{\beta}_i, \hat{\beta}_j)], \tag{5.25}$$

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

$$\text{var}\{\hat{w}(\mathbf{x}_1)/\hat{w}(\mathbf{x}_2)\} = \{w(\mathbf{x}_1)/w(\mathbf{x}_2)\}^2 \sum_{i=1}^{p} \sum_{j=1}^{p} (x_{1i} - x_{2i})(x_{1j} - x_{2j}) \text{cov}(\hat{\beta}_i, \hat{\beta}_j). \tag{5.26}$$

In the next chapter the estimated RSPF

$$\hat{w}^*(\mathbf{x}) = 1 - \exp\{-\exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_p x_p)\} \tag{5.27}$$

is used. Here the Taylor series method gives the approximate variances

$$\text{var}\{\hat{w}^*(\mathbf{x})\} = w^*(\mathbf{x})^2 [\log_e\{1 - w^*(\mathbf{x})\}]^2 \sum_{i=0}^{p} \sum_{j=0}^{p} x_i x_j \text{cov}(\hat{\beta}_i, \hat{\beta}_j), \tag{5.28}$$

and

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

$$\times [\sum_{j=0}^{p} \sum_{j=0}^{p} x_{1i} x_{2j} \text{cov}(\hat{\beta}_i, \hat{\beta}_j), \tag{5.29}$$

taking $x_{10} = x_{20} = 1$.

**Example 5.3 Habitat Selection by Pronghorn Antelope**

Example 5.1 was concerned with the selection of winter habitat by pronghorn antelope 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.

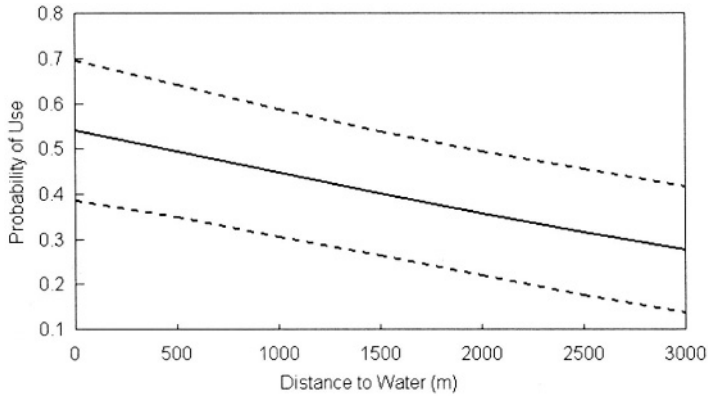



*Figure 5.5 Probabilities of use by pronghorn for East/Northeast study plots in 1980/81, as a function of the distance to water. The estimated RSPF $\hat{w}^*$ is the continuous line and the broken lines are approximate 95% confidence intervals given by $\hat{w}^* \pm 1.96se(\hat{w}^*)$.*

## 5.8 Discussion

The logistic regression methods described in this chapter have much to recommend them for estimation of a RSPF or RSF when there is no need to take into account varying amounts of selection time. These types of models are 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 of the animals involved. However, this does not mean that these models 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 variation among 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 the coefficient $\beta_i$ 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. It is, of course, critical that data be collected on each animal using the same methods regardless of whether they are independent or dependent.

## 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 census information on which units were used in the study area or the available food items. 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 limit the influences to a restricted subset of the study area or food items.

- An example is provided where the use and non-use of 256 study plots in the Red Rim area of Wyoming, USA, by pronghorn antelope 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. In these cases an exponential function is proposed for the RSPF or RSF.

- An example involving the selection of nest sites by fernbirds in Otago, New Zealand, is provided where a computer program for logistic regression is used to estimate the coefficients in an exponential 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 antelope in Wyoming.

- The uses of logistic regression and exponential resource selection functions 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

$$w^*(x,t) = \exp(\beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3),$$

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 |
| 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 |