A PRELIMINARY STUDY OF THE BIAS AND VARIANCE WHEN ESTIMATING A RESOURCE SELECTION FUNCTION WITH SEPARATE SAMPLES OF USED AND AVAILABLE RESOURCE UNITS

RYAN NIELSON[1], Western EcoSystems Technology Inc., 2003 Central Avenue, WY 82001, USA
BRYAN F.J. MANLY, Western EcoSystems Technology Inc., 2003 Central Avenue, WY 82001, USA
LYMAN L. MCDONALD, Western EcoSystems Technology Inc., 2003 Central Avenue, WY 82001, USA

Abstract: A standard method for estimating a resource selection function (RSF) involves collecting a sample of available resource units, and a separate sample of used resource units. The RSF can then be estimated using a standard logistic regression program, but with a constant term that is a function of the sampling fractions for available and used resource units. Using logistic regression for estimation under these conditions is justified by Manly *et al*. (2002) using a model that assumes that samples of used and available resource units are taken in such a way that there is a certain probability of a unit being sampled, and that the sample of available units is collected first without replacement. In this paper we examine the performance of logistic regression estimation both with this sampling scheme and others. In particular, we consider the question of when estimation is unbiased or nearly so, and how well variances are estimated using the usual logistic regression methods.

Key words: logistic regression, properties of estimators, simulation

---

We briefly review the estimation of a resource selection function (RSF) under four different sampling methods. First, a sample of resource units can be taken and it can be seen which of these are used and which are not used. Second, a sample of available resource units can be taken, and a separate sample of used units also taken. Third, a sample of available units can be taken, and a separate sample of unused units also taken. Finally, a sample of used units can be taken and a separate sample of unused units.

These sampling methods and the required estimation procedures are discussed by Manly *et al*. (2002, Chapter 5). In brief, the first sampling method involves using ordinary logistic regression, while the other methods use logistic regression but the RSF is not just equal to the logistic regression function. Rather, with sampling methods two to four the results from logistic regression have to be modified in special ways in order to obtain the RSF.

The most common sampling method in practice appears to be the one where a sample of available units is taken, and also a separate sample of used units. For example, if the resource units being considered are blocks of land that might be used by an animal then the sample of available units might be obtained from a geographical information system (GIS) , with field sampling undertaken to find blocks used by the animal. It is this sampling method that we consider in the present paper. The questions that we examine are:

- Is there any bias in the estimation of the RSF and variances, with sample sizes from small to large?
- If the sampling scheme described by Manly *et al*. (2002, p. 99), with random sample sizes, is changed to one with fixed sample sizes, then how does this effect estimation of the RSF?
- If the size of the sample of available units is effectively unlimited because it can be drawn from a GIS, then how large should the sample size be to insure that there will be no improvement in estimation by making it larger?
- The sampling scheme described by Manly *et al*. (2002, p. 99) involves selecting the sample of available units first without replacement, so that none of the units in this sample can also appear in the sample of used units. How is estimation of the RSF effected if the used units are chosen first without replacement, or if the problem of overlapping samples is just ignored?
- For some specific cases where the RSF is not an exponential function of the variables that describe resource units, how well can the function be approximated by an exponential function?

[1]Email: rnielson@west-inc.com

These questions are now considered in turn, following a brief description of the justification for using logistic regression with separate samples of available and used units.

THEORY OF ESTIMATION

The sampling scheme described by Manly *et al*. (2002, p. 99) involves taking a sample of available units in such a way that every one of these units is selected with probability P_a, independently of the selection of any other unit, and taking a separate sample of used units in such a way that each of these units is selected with probability P_u independently of the selection of any other unit. In addition, the sample of available units is taken first, without replacement, so that no unit in this sample can also appear in the sample of used units. Based on this sampling procedure, Manly *et al*. argued that the probability that the ith unit is in the used sample, given that it is in one of the samples is

$$\text{Prob(ith unit used | unit is sampled)} = \frac{(1 - P_a)\, w^*_i\, P_u}{P_a + (1 - P_a)\, w^*_i\, P_u}, \tag{1}$$

where w^*_i is the resource selection probability function (RSPF) for the ith unit, i.e. the probability that the unit is used as a function of variables $\mathbf{x} = (x_1, x_2, ..., x_p)$ measured on the unit to describe it. Assuming that the RSPF takes the particular form

$$w^*(\mathbf{x}) = \exp(ß_0 + ß_1 x_1 + ... + ß_p x_p), \tag{2}$$

then leads to the logistic regression equation

$$t(\mathbf{x}_i) = \frac{\exp\{\log_e[(1 - P_a)\, P_u / P_a] + ß_0 + ß_1 x_{i1} + ... + ß_p x_{ip}\}}{1 + \exp\{\log_e[(1 - P_a)\, P_u / P_a] + ß_0 + ß_1 x_{i1} + ... + ß_p x_{ip}\}}, \tag{3}$$

for the probability that the ith unit is in the used sample given that it is in one of the samples.

It is very important to appreciate that the RSPF is given by equation (2) and not by the logistic regression equation (3). Nevertheless, once the logistic regression equation is fitted the values for $ß_1$ to $ß_p$ are estimated, so that the RSF

$$w(\mathbf{x}) = \exp(ß_1 x_1 + ... + ß_p x_p) \tag{4}$$

can be estimated, where this is proportional to the RSPF. In addition, if it happens that the values of P_a and P_u are known, then $ß_0$ can be estimated as well, so that the RSPF of equation (2) can be estimated.

Obviously, the RSPF of equation (2) will not be valid if it returns probabilities greater than one. This can be checked for if $ß_0$ can be estimated. Otherwise, a potential problem in this respect is indicated by a large proportion of units being used, with certain types of units apparently being always used. In such cases it will be necessary to include higher order polynomial terms in equation (2), or even modify the functional form of the RSPF and estimate it by maximizing the likelihood of the data without using logistic regression.

BIAS AND VARIANCE ESTIMATION

For inferences concerning regression coefficients based on maximum likelihood theory it is necessary to assume that the estimated coefficients are normally distributed, with variances that are well approximated by the values that this theory provides (Collett, 1991, p. 49). This will be the case providing that the statistics

$$t_0 = \{b_0 - \log_e[(1 - P_a)\, P_u / P_a] - ß_0\}/SE(b_0), \tag{5}$$

and

$$t_j = (b_j - ß_j)/SE(b_j) \tag{6}$$

have distributions that are close to the standard normal, where b_0 is the estimated constant term and b_j is the estimated value of $ß_j$ from the fitted logistic regression, with the estimated standard error $SE(b_j)$. Here the term $\log_e[(1 - P_a)\, P_u / P_a]$

is involved with the constant term because the constant in the logistic regression estimates $\log_e[(1-P_a)P_u/P_a] + ß_0$ rather than just $ß_0$.

To examine how this works in practice, Simulation Experiment 1 was conducted. This involved setting up a model population of 10,000 resource units described by two variables, X_1 and X_2. The values for both of these variables were generated as independent random values from the normal distribution with a mean of 1.0 and a standard deviation of 0.135, and the RSF was set equal to

$$w(\mathbf{x}) = \exp(x_1), \tag{7}$$

i.e. the coefficient of X_1 is one, and the coefficient of X_2 is zero.

The model population was used to simulate situations with 5%, 10%, 20%, 50% and 80% of the units expected to be used, with expected sample sizes of 1,000 available units and 300 used units. For example, with 5% of units expected to be used the RSF of equation (7) was scaled to the RSPF

$$w^*(\mathbf{x}) = \exp(ß_0 + x_1), \tag{8}$$

with $ß_0$ chosen to make the mean value of $w^*(\mathbf{x})$ equal to $\bar{w}^* = 0.05$ for all 10,000 units in the model population. Each unit was then determined to be used with a probability equal to its value from equation (8), resulting in about 5% of the units being considered to be used. Each unit in the model population (used or unused) was then included in the sample of available units with probability $P_a = 0.1$, resulting in a sample size of about 1,000 units. Each of the remaining used units (about 450 of them) was then included in the sample of used units with a probability of 0.667, to give a used sample size of about 300. The data were then used to fit a logistic regression model, and the statistics t_0, t_1 and t_2 calculated. This simulation process was then repeated 4,999 times to obtain a total of 5,000 t-statistics for the situation with 5% of units expected to be used. The same then done for 10%, 20%, 50% and 80% of units expected to be used, with 5,000 t-statistics being obtained for each of these scenarios.

The results of this simulation study are summarized in Table 1, in terms of the mean, standard deviation, skewness and kurtosis of the distributions of the t-statistics. The means, standard deviations, skewness values and kurtosis values are all close to what is required for a standard normal distribution (0, 1, 0 and 3, respectively), except when 80% of the units are expected to be used. With 80% expected use the mean values for t_0 and t_1 are significantly different from zero and t_0 has significant skewness.

The reason for the non-normal distribution of t_0 and t_1 with 80% expected use is apparently the result of the exponential model being an inadequate approximation for the RSPF in this case because equation (8) then allows probabilities of use greater than one. These probabilities were reduced to one in the simulations when they occurred, and this modification has presumably introduced the bias. The last column in Table 1 shows the percentage of units in the model population for which equation (8) gives a probability of more than one. This is zero for 5% to 50% expected use, but is 4.3% for 80% expected use.

A Simulation Experiment 2 was run with the two X variables used to describe the 10,000 resource units in the population randomly selected from the exponential distribution with a mean of 1.0 and a standard deviation of 0.135. The RSPF was still given by equation (8), but in this case the highly skewed distribution of X_1 means that even with 50% of units expected to be used there are some units in the population for which equation (8) gives values greater than one. Table 2 shows a summary of the results obtained from this simulation, in the same format as for Table 1. The bias is very large when 80% of units are expected to be used, in which case 6.7% of the units in the population had values from equation (8) above one.

Although the two simulation experiments only cover some very specific situations, with moderately large sample sizes, they do suggest that a bias in the estimation of the coefficients in a RSF (showing up as t-statistics having means different from zero) will not be a problem unless the proportion of used units is so high that the exponential model implies that some units have probabilities of use of more than one. In such cases it is not really appropriate to use the exponential model, and it would be better to replace the exponential model of equation (2) with another model for use in equation (1).

Some simulations were also carried out with the small expected sample sizes of 50 available units and 25 used units. These indicated that both for normally distributed and exponentially distributed X variables, with the same conditions as for Simulation Experiments 1 and 2, the t-statistics still had distributions quite close to standard normal, but the results

were if anything less affected by the RSPF giving values greater than one for some units in the population. Basically, therefore, it appears that the logistic estimation method works well even with small expected sample sizes.

In all the simulation results the coefficient of X_2 was significantly different from zero at about the correct 5% of times for a test using a 5% level of significance. Consequently, there was never any indication that a variable that should not be in the RSF will be found to be significant more or less often than expected based on statistical theory.

USING FIXED SAMPLE SIZES

Justifying the use of logistic regression with fixed sample sizes for available and used units is difficult theoretically because the probability of a unit being in the sample of used units, given that it is sampled, is not independent of which other units are used. Nevertheless, sample sizes are often fixed in practice, particularly if the sample of available units is drawn using a GIS. This raises the question of whether using fixed sample sizes matters in practice.

This was checked by running Simulation Experiment 3, which was similar to Simulation Experiment 1, but with the sample sizes fixed at 1,000 available units and 300 used units. Overall, the results were very similar for the random sample size design of Manly *et al*. (2002) with the same expected sample sizes. Apparently, therefore, using fixed sample sizes is a satisfactory procedure, at least for moderately large sample sizes.

Table 1. Results from Simulation Experiment 1 on a model population with 10,000 resource units, a RSF that is a function of one of the two variables measured for which the values are normally distributed with a mean of 1.0 and a standard deviation of 0.135. The expected size of the sample of available units was 1,000 units, and the expected size of the sample of used units was 300 units.

			t_0				t_1				t_2				
$\bar{w}^*$	P_a	P_u	Mean	SD	Skew	Kurt	Mean	SD	Skew	Kurt	Mean	SD	Skew	Kurt	%>1[b]
0.05	0.10	0.667	0.01	1.00	0.02	3.03	-0.01	0.99	-0.01	2.99	-0.01	1.01	-0.05	3.05	0.0
0.10	0.10	0.333	-0.01	1.00	-0.02	2.95	0.00	1.02	-0.02	2.96	0.02	0.99	-0.03	3.07	0.0
0.20	0.10	0.167	0.01	0.99	0.02	3.07	-0.01	1.00	-0.01	2.95	0.00	0.99	-0.06	2.94	0.0
0.50	0.10	0.067	-0.01	0.98	0.02	3.07	0.02	0.98	0.01	3.08	0.00	0.99	0.00	3.06	0.0
0.80	0.10	0.042	0.05	0.99	0.07	2.95	-0.10	0.98	-0.02	3.22	0.02	1.01	-0.01	3.01	4.3

[a]The mean, standard deviation (SD), skewness (Skew) and kurtosis (Kurt) values are for 5,000 values of the t-statistics defined by equations (5) and (6). Underlined values in the table indicate values that are significantly different at the 5% level from what is expected from a standard normal distribution, based on the standard error estimated using the 5000 replicate values for the mean and standard deviation, and Table 34 of Pearson and Hartley (1970) for the skewness and kurtosis.

[b]The percentage of units in the model population where the probability of use had to be reduced to one.

Table 2. Results from Simulation Experiment 2 on a model population with 10,000 resource units, a RSF that is a function of one of the two variables measured for which the values are exponentially distributed with a mean and standard deviation of one. The expected size of the sample of available units was 1,000 units, and the expected size of the sample of used units was 300 units. See the notes with Table 1.

			t_0				t_1				t_2				
$\bar{w}^*$	P_a	P_u	Mean	SD	Skew	Kurt	Mean	SD	Skew	Kurt	Mean	SD	Skew	Kurt	%>1
0.05	0.10	0.667	0.03	1.00	0.01	2.97	-0.01	0.99	0.03	2.99	-0.02	1.01	0.06	2.97	0.0
0.10	0.10	0.333	0.00	1.00	-0.07	2.96	0.03	0.99	0.04	3.03	0.00	1.00	0.10	2.91	0.0
0.20	0.10	0.167	0.02	0.99	-0.08	3.05	0.00	0.99	0.02	2.95	-0.01	0.99	0.13	2.97	0.0
0.50	0.10	0.067	0.05	0.99	-0.03	2.92	-0.05	0.99	0.10	3.00	0.01	1.00	0.03	2.94	0.2
0.80	0.10	0.042	0.39	1.01	0.00	2.96	-0.58	0.99	0.10	2.88	0.02	0.99	0.05	3.02	6.7

EFFECT OF CHANGING THE AVAILABLE SAMPLE SIZE

Intuitively it seems likely that if the sample of available units size is increased more and more then there will come a point where the available population of resource units is defined so well that there is little point in increasing the sample size any further. This was examined by Simulation Experiment 4, in which a population of 100,000 resource units was generated, with an average probability of use of 0.05 from equation (8). The random sample size strategy of Manly *et al*. (2000, p. 99) was then used to sample the population, with expected sizes of 500, 1000, 10000, 20000, 40000, 60000, and 80000 for the sample of available units, and an expected size of 300 units for the sample of used units. Table 3 shows the mean of the estimated standard error of the coefficient of X_1 for each of the available sample sizes. It appears that increasing the expected size of the sample of available units beyond 10,000 leads to little improvement in the accuracy of estimation. This is consistent with what was found by Erickson *et al*. (1998) for a GIS study on moose, and may be a reasonable rule of thumb for studies where the available sample size can be made as large as necessary.

ALTERNATIVE WAYS TO HANDLE SAMPLE OVERLAP

The sampling scheme proposed by Manly *et al*. (2002, p. 99) requires that the sample of available units is first selected, without replacement, followed by the selection of the sample of used units. This approach, which will be called Plan 1 here, assures that no overlap exists in the used and available samples, and gives every resource unit (used or unused) an equal chance to be in the sample of available units. Two alternative methods for sampling have been examined to see how they effect the estimation of the RSF. Plan 2 reverses the procedure of Plan 1, and calls for the used units to be sampled first, without replacement, followed by the sampling of available units. This sampling plan also ensures no overlap between available and used samples. Finally, Plan 3 ignores any potential overlap in the available and used samples, and simply calls for two independent samples of the available units and used units. As this strategy involves taking the first sample with replacement, it makes no difference which sample is drawn first.

Table 3. How the standard error of the estimated coefficient of X_1 changed as the expected size of the sample of available units increased from 300 to 80,000 resource units, in a population of 100,000 resource units.

Available sample size	300	5,000	10,000	20,000	40,000	60,000	80,000
Standard error	0.450	0.440	0.434	0.432	0.430	0.428	0.428

To examine the potential effects on estimation that can result from these three competing sampling plans, Simulation Experiment 5 was conducted. A population was simulated with 100,000 resource units, with 40% of these expected to be used. Three scenarios, each representing a different degree of potential overlap in the available and used samples if sampling Plan 3 was used. In the first scenario there were an expected 8,000 units in the sample of available units and an expected 4,000 units in the sample of used units. This represented an average of 2.67% overlap in the two samples when Plan 3 was used. Scenario 2 called for 15,000 units expected in each of the two samples, and represented an expected 7.5% overlap in the two samples when Plan 3 was used to draw the samples. The third scenario had a 10% overlap in the two samples for Plan 3, with an expected available sample size of 20,000 and an expected used sample size of 20,000. The relative bias in model coefficients, along with the significance of this statistic was used to compare the three sampling plans. There were 1,000 sets of data generated for each scenario-by-sampling plan combination.

The result obtained from this simulation was that there was no significant bias for any of the three methods for sampling the data. Consequently, no evidence was found suggesting that the use of sampling Plan 2 or sampling Plan 3 introduces any major biases, although obviously they may under other circumstances.

APPROXIMATING A NON-EXPONENTIAL RESOURCE SELECTION FUNCTION

There may be some question about the ability of the RSF modeling procedure described in Manly *et al*. (2002) to estimate a RSF that has something other than an exponential form. Simulation Experiment 6 was therefore carried out to examine the ability of the procedure in this area. The simulated population consisted of 100,000 resource units, with every unit assigned values for a single variable X, which was normally distributed with a mean of 16 and a standard deviation of 5. Any values of X less than zero were set equal to zero. The relative probability of use of a unit with X = x was then set equal to

$$w(x) = x^3 \exp(-x/4) / 16, \tag{9}$$

and the relative probabilities were scaled so that the probabilities of use were all between zero and one, with the average of the probabilities giving the required expected number of used units. Situations with 1%, 5%, 10%, 20%, 50% and 80% of the population expected to be used were considered.

Figure 1 shows the relative bias in the estimated probability of selection when equation (9) is approximated by quadratic, cubic and quartic exponential resource selection functions. To produce these figures the relative probability of selection was predicted for units in the population having true probabilities of selection falling on the 5th, 10th, 20th, 30th, ..., 90th, 95th, and 99th percentiles of the distribution of true probabilities for the entire simulated population. These predicted relative probabilities and the true probabilities were scaled so the probability of selection (both true and relative) was equal to one for a unit with the mean value for X. The bias was then calculated by taking the scaled relative probability of selection minus the scaled true relative probability of selection.

It appears that a quartic function gives a reasonable fit to equation (9), although there is some overestimation of relative probabilities of use for units with moderately low probabilities of use and units with high probabilities of use.

DISCUSSION

The main points to emerge from the simulation studies considered here are that (a) care needs to be taken to ensure that the assumption of an exponential RSF combined with a high overall probability of use for resource units does not imply that there are probabilities of use greater than one for some units, (b) changing the sampling procedure to one with fixed sample sizes has had minimal effects for the situations considered, (c) in general there may be little point in having available sample sizes of more than 10,000 units, (d) changing the order of the selection of the available and used sample, or just selecting them independently ignoring any problems of overlap has had minimal effects for the situations considered, and (e) non-exponential RSF functions can be approximated using the exponential RSF, although this may involve some relatively small bias.

Clearly these are only preliminary findings, and more simulations covering a far wider range of situations are needed to confirm the extent to which they apply in general. We plan to carry out these more extensive simulations at a later date.

LITERATURE CITED

COLLETT, D. 1991. Modeling binary data. Chapman & Hall, London, U.K.

ERICKSON, W.P., T.L. MCDONALD, AND R. SKINNER. 1998. Habitat selection using GIS data: a case study. Journal of Agricultural, Biological and Environmental Statistics 3:296-310.

MANLY, B.F.J., L.L. MCDONALD, D.L. THOMAS, T.L. MCDONALD, AND W.P. ERICKSON. 2002. Resource selection by animals: statistical design and analysis for field studies. Second edition. Kluwer Academic Publishers, Dordrecht, The Netherlands.

PEARSON, E.S., AND H.O. HARTLEY. 1970. Biometrika tables for statisticians, vol. 2. Cambridge University Press, Cambridge.

Figure 1. The bias in the estimated relative probability of selection when equation (9) is approximated by exponentials of second degree (quadratic) to fourth degree (quartic) functions. Curves are only shown for 1%, 20% and 80% of units used because the curves for the other percentages considered (5%, 10% and 50%) are so similar.