USE AND INTERPRETATION OF LOGISTIC REGRESSION IN HABITAT-SELECTION STUDIES

KIM A. KEATING,[1] U.S. Geological Survey, Northern Rocky Mountain Science Center, Montana State University, Bozeman, MT 59717, USA
STEVE CHERRY, Department of Mathematical Sciences, Montana State University, Bozeman, MT 59717, USA

Abstract: Logistic regression is an important tool for wildlife habitat-selection studies, but the method frequently has been misapplied due to an inadequate understanding of the logistic model, its interpretation, and the influence of sampling design. To promote better use of this method, we review its application and interpretation under 3 sampling designs: random, case–control, and use–availability. Logistic regression is appropriate for habitat use–nonuse studies employing random sampling and can be used to directly model the conditional probability of use in such cases. Logistic regression also is appropriate for studies employing case–control sampling designs, but careful attention is required to interpret results correctly. Unless bias can be estimated or probability of use is small for all habitats, results of case–control studies should be interpreted as odds ratios, rather than probability of use or relative probability of use. When data are gathered under a use–availability design, logistic regression can be used to estimate approximate odds ratios if probability of use is small, at least on average. More generally, however, logistic regression is inappropriate for modeling habitat selection in use–availability studies. In particular, using logistic regression to fit the exponential model of Manly et al. (2002:100) does not guarantee maximum-likelihood estimates, valid probabilities, or valid likelihoods. We show that the resource selection function (RSF) commonly used for the exponential model is proportional to a logistic discriminant function. Thus, it may be used to rank habitats with respect to probability of use and to identify important habitat characteristics or their surrogates, but it is not guaranteed to be proportional to probability of use. Other problems associated with the exponential model also are discussed. We describe an alternative model based on Lancaster and Imbens (1996) that offers a method for estimating conditional probability of use in use–availability studies. Although promising, this model fails to converge to a unique solution in some important situations. Further work is needed to obtain a robust method that is broadly applicable to use–availability studies.

JOURNAL OF WILDLIFE MANAGEMENT 68(4):774–789

Key words: bias, case–control, contaminated control, exponential model, habitat modeling, log-binomial model, logistic model, resource selection function, resource selection probability function, sampling design, use–availability.

Logistic regression has become increasingly popular for modeling wildlife habitat selection but often is used incorrectly. Misapplications reflect an inadequate understanding among wildlife researchers concerning the logistic model, its interpretation, and especially the influence of sampling design. Design effects have been well studied by epidemiologists and economists (e.g., Prentice and Pyke 1979, Steinberg and Cardell 1992, Lancaster and Imbens 1996), but the range of designs and their influence on perceived probability of habitat use have not been clearly and accurately articulated in the wildlife literature. We address use and interpretation of logistic regression in habitat-selection studies.

We distinguish among 3 sampling designs—random, case–control, and use–availability—whose key characteristics are illustrated in the following hypothetical example. Imagine that we are designing a study of nest-site selection by the Hungarian horntail dragon (*Flammasaurus cerocaudus*; Rowling 2000:327), which nests in east European old-growth forests. If nests are common and easily seen, we might choose a random sampling design, whereby a number of trees are selected randomly from throughout the forest, and characteristics of each are measured and recorded along with information about whether the tree contains a nest. If nests are easily seen but uncommon, we might use a case–control design to ensure that our final sample contains an adequate number of nest trees. With this design, we draw 2 distinct random samples: 1 from the pool of all trees containing a horntail nest, and a second from the pool of all trees lacking a nest. Again, relevant characteristics of each sampled tree are recorded, together with information about whether the tree contains a nest. In both the random and case–control designs, we assume that nests are easily seen so that both presence and absence of a nest are determined without error. If only presence can be determined reliably, then we might employ a use–avail-

[1] E-mail: kkeating@montana.edu

ability design. For example, we might identify a random sample of nest trees by tracking radiomarked females to their nests, then measure habitat availability by randomly sampling from all trees in the forest. We record relevant characteristics of each tree sampled, but because horntail nests are notoriously cryptic, we do not know whether the trees in our available sample contained a nest. Some key differences among these designs are: (1) the random design yields a sample that contains nest and non-nest trees in approximate proportion to their occurrence in the forest; (2) the case–control design yields a sample of nest and non-nest trees, but relative proportions of the 2 are determined by the researcher and may not be representative of the underlying population of trees; and (3) the use–availability design yields a random sample of nest trees and a second random sample drawn from all trees in the forest, but we do not know whether trees in the second sample contain nests. These differences in sampling design translate into profound differences in the way that logistic regression can be applied and interpreted.

We reexamine use of logistic regression for wildlife habitat modeling under each of these sampling designs. We especially consider the role of logistic regression in estimating resource selection probability functions (RSPFs) and RSFs. Manly et al. (2002:27) defined an RSPF as "a function which gives probabilities of use for resource units of different types." An RSF is any function proportional to the RSPF (Manly et al. 2002:29); that is, RSF = kRSPF for some positive constant k. Of the various statistical methods for estimating RSPFs or RSFs (Alldredge et al. 1998, Manly et al. 2002), logistic regression is most widely used. For each sampling design, we present the formal probability model and identify relationships to commonly used forms of the RSPF and RSF. Quantitative examples are used to illustrate the different sampling designs and effects of different estimation methods. For each design, we discuss implications for modeling habitat selection.

RANDOM SAMPLING

Sampling Model

When using logistic regression, the probability that a particular habitat will be used by the species or individual of interest is assumed to take the form of a logistic model, parameterized as follows. Imagine an area comprised of multiple locations that are sampled with replacement by observing whether the sampled location was used. A binary response variable (y) is defined for each observation, such that $y = 1$ if use was observed and $y = 0$ if it was not. We assume that y is recorded without error (for discussions related to this assumption, see MacKenzie et al. 2002). Also, p covariates are measured at each location as $\mathbf{x}' = (1, x_1, \ldots, x_p)$. The logistic model describing probability of use conditioned on habitat (i.e., the RSPF) is

$$P(y = 1 \mid \mathbf{x}) = \frac{\exp(\boldsymbol{\beta}'\mathbf{x})}{1 + \exp(\boldsymbol{\beta}'\mathbf{x})}, \quad (1)$$

where $\boldsymbol{\beta}' = (\beta_0, \beta_1, \ldots, \beta_p)$ is a vector of coefficients relating probability of use to the habitat covariates via the relationship $\boldsymbol{\beta}'\mathbf{x} = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$. Model (1) is intrinsically bounded within the interval [0,1]. In this model, the sampling unit is an individual observation and $P(y = 1 \mid \mathbf{x})$ is independent of sample size. This differs from other formulations (e.g., Manly et al. 2002:83) in which the sampling unit is the physical location, so that $P(y = 1 \mid \mathbf{x})$ increases with time and hence with sample size. However, because the 2 formulations are effectively the same when the number of available locations is large relative to the sample size, we use them interchangeably.

The simplest sampling design is one in which n locations are drawn randomly with replacement from the N available locations, and y and $\mathbf{x}$ are observed and recorded for each. Model parameters are then estimated by maximizing the log-likelihood (Hosmer and Lemeshow 2000:9). With random sampling, this yields approximately unbiased estimates of the coefficients and in turn the conditional probability of use, as illustrated in the following example.

Example 1: Random Sampling

Let the true conditional probability of use be,

$$P(y = 1 \mid \mathbf{x}) = \frac{\exp(5.673 - 3.000\,ELEV)}{1 + \exp(5.673 - 3.000\,ELEV)}, \quad (2)$$

where $ELEV$ is elevation in km. Using ArcView 3.2 and ArcView Spatial Analyst (Environmental Systems Research Institute 1999), we projected Eq. (2) over a 30-m resolution grid covering a model study area of about 2,500 km^2 in the upper Yellowstone River Valley, Montana, USA ($N \approx 2.8 \times 10^6$ pixels). We then sampled n = 2,000 pixels randomly with

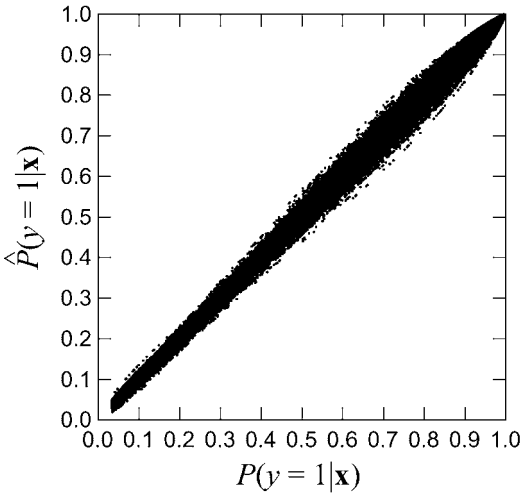

Fig. 1. Relationship between the estimated and true conditional probability of use [$\hat{P}(y = 1|\mathbf{x})$ and $P(y = 1|\mathbf{x})$, respectively] for the 1,000 models fit to data gathered according to a random sampling design.

replacement, recording *ELEV* for each and calculating $P(y = 1|\mathbf{x})$ as per Eq. (2). Whether a pixel was used was stochastically determined as $y = I[U \leq P(y = 1|\mathbf{x})]$, where U is a uniform random variable on the interval [0,1], and $I[\cdot]$ is the indicator function (i.e., $I[U \leq P(y = 1|\mathbf{x})] = 1$ if $U \leq P(y = 1|\mathbf{x})$ is true and $I[U \leq P(y = 1|\mathbf{x})] = 0$ otherwise). This sampling process was replicated 1,000 times. For each replicate, we used the LOGIT module in SYSTAT (Systat Software 2000) to fit the data to a logistic model (Eq. [1]) in which $\boldsymbol{\beta}'\mathbf{x} = \beta_0 + \beta_1 ELEV$. The resulting mean (SE) estimates of $\hat{\beta}_0 = 5.672$ (0.388) and $\hat{\beta}_1 = -3.000$ (0.175) were essentially unbiased. Estimates of β_0 and β_1 also were substituted into Eq. (2) and, for each replicate, $P(y = 1|\mathbf{x})$ was estimated for 100 randomly selected pixels. The resulting estimates ($\hat{P}(y = 1|\mathbf{x})$) were similarly unbiased (Fig. 1).

Implications for Modeling Habitat Selection

With random sampling designs, logistic regression is straightforward and yields models of the probability of use conditioned on habitat ($P(y = 1|\mathbf{x})$). In the terminology of Manly et al. (2002:27), a direct estimate of the RSPF is obtained. Examples include species occurrence models constructed from grid-based samples of kangaroos (*Macropus* spp.; Walker 1990) and grizzly bears (*Ursus arctos*; Apps et al. 2004). These studies implicitly assumed that, with respect to habitat, grids were randomly located and sampled. Apps et al. (2004) violated this assumption by excluding some low-use areas from sampling during some years, thereby biasing their sample in favor of used sites. Random sampling also has been assumed in some transect-based studies. For example, Fleishman et al. (2001) applied logistic regression to model probability of occurrence of Great Basin butterflies, using data gathered along trails and roads. They implicitly assumed that trails and roads traversed a random sample of available habitats—an assumption we question, given the area's rugged terrain. Overall, habitat-selection studies using random or approximately random sampling designs are relatively uncommon in the wildlife literature. When such a design is adopted, however, associated assumptions should be clearly articulated, and their validity, if not self-evident, should be discussed.

CASE–CONTROL SAMPLING

Sampling Model

When use is rare, a prohibitively large random sample would be required to detect enough instances of use for meaningful analysis. In such cases, sampling can be stratified by y, drawing with replacement a random sample of n_1 used locations and a second random sample of n_0 unused locations. This is a case–control design (Hosmer and Lemeshow 2000:205) and is equivalent to sampling protocol C of Manly et al. (2002:5). The resulting sample is no longer described by Eq. (1) because the probability of observing an instance of use in our sample is now different than the probability of use in the population. To devise an appropriate model, a variable indicating whether a location appears in the sample is needed (Hosmer and Lemeshow 2000:206). Therefore, let $\eta = 1$ for each location that was selected as part of the sample, and $\eta = 0$ otherwise. Also, let $P_1 = P(\eta = 1|y = 1)$ be the probability that any particular used location was included in the sample, and let $P_0 = P(\eta = 1|y = 0)$ be the probability that any particular unused location was included. When sampling from a finite population where sample size is small relative to population size, these are equivalent to $P_1 = n_1/N_1$ and $P_0 = n_0/N_0$, where N_1 and N_0 are the respective numbers of used and unused locations in the population of $N = N_0 + N_1$ total locations. The probability model describing our case–control sample is then (Hosmer and Lemeshow 2000:207)

$$P(y=1\mid \mathbf{x},\eta=1)=\frac{\exp\left[\boldsymbol{\beta}'\mathbf{x}+\ln\left(\frac{P_1}{P_0}\right)\right]}{1+\exp\left[\boldsymbol{\beta}'\mathbf{x}+\ln\left(\frac{P_1}{P_0}\right)\right]}=\frac{\exp\left[\boldsymbol{\beta}^{*\prime}\mathbf{x}\right]}{1+\exp\left[\boldsymbol{\beta}^{*\prime}\mathbf{x}\right]}, \quad (3)$$

where $\boldsymbol{\beta}^{*\prime} = (\beta_0^*, \beta_1, \ldots, \beta_p)$ and $\beta_0^* = \beta_0 + \ln(P_1/P_0)$. This is a logistic model with intercept term β_0^*. Using logistic regression to fit case–control data yields estimates of $\boldsymbol{\beta}^*$ rather than $\boldsymbol{\beta}$. Note that the RSPF, $P(y = 1\mid\mathbf{x})$, is still given by Eq. (1) and can be calculated from Eq. (3) if estimates of P_0 and P_1 are available, since

$$P(y=1\mid\mathbf{x})=\frac{\exp\left[\boldsymbol{\beta}^{*\prime}\mathbf{x}-\ln\left(\frac{P_1}{P_0}\right)\right]}{1+\exp\left[\boldsymbol{\beta}^{*\prime}\mathbf{x}-\ln\left(\frac{P_1}{P_0}\right)\right]}. \quad (4)$$

In the notation of our paper, Manly et al. (2002:104, Eq. 5.19) give the RSPF for the case–control setting as

$$P(y=1\mid\mathbf{x})=\frac{\exp\left[\boldsymbol{\beta}^{*\prime}\mathbf{x}+\ln\left(\frac{P_0}{P_1}\right)\right]}{1+\exp\left[\boldsymbol{\beta}^{*\prime}\mathbf{x}+\ln\left(\frac{P_0}{P_1}\right)\right]}. \quad (5)$$

Equations (4) and (5) are equivalent because $\ln(P_0/P_1) = -\ln(P_1/P_0)$.

In most case–control studies, probability of use cannot be estimated because P_0 and P_1 are unknown. However, logistic regression still provides useful information, if interpreted carefully. Manly et al. (2002:104) suggested setting $P_0 = P_1$ in Eqs. (4) and (5), then using the resulting equation to index selectivity. This approach allows habitats to be ranked qualitatively. Quantitative comparisons of habitats are possible by examining odds ratios. Substituting from Eq. (1) and rearranging terms, we can show that

$$\frac{P(y=1\mid\mathbf{x})}{P(y=0\mid\mathbf{x})}=\exp(\boldsymbol{\beta}'\mathbf{x}). \quad (6)$$

This is the odds that a location will be used given the covariate pattern $\mathbf{x}$. In case–control studies, parameter estimates are commonly interpreted in terms of odds ratios, but interpretation depends on whether the variable is categorical or continuous. Consider a model with a single categorical predictor (x_1) with 2 levels. For example, let $x_1 = 1$ if a location was recently burned, and $x_1 = 0$ otherwise. The odds ratio (ψ) is

$$\Psi=\frac{\left[\frac{P(y=1\mid x_1=1)}{P(y=0\mid x_1=1)}\right]}{\left[\frac{P(y=1\mid x_1=0)}{P(y=0\mid x_1=0)}\right]}=\frac{\exp(\beta_0+\beta_1)}{\exp(\beta_0)}=\exp(\beta_1).$$

Thus, the odds a burned location is used is equal to $\exp(\beta_1)$ times the odds an unburned location is used. For a continuous variable (x_1), we can show that

$$\Psi=\frac{\left[\frac{P(y=1\mid x_1+1)}{P(y=0\mid x_1+1)}\right]}{\left[\frac{P(y=1\mid x_1)}{P(y=0\mid x_1)}\right]}=\exp(\beta_1).$$

Thus, for every 1-unit change in x_1, a change of $\exp(\beta_1)$ units occurs in the odds ratio. In general, for both categorical and continuous variables, if we denote a reference habitat type by $\mathbf{x}_R = (1, x_{1,R}, \ldots, x_{p,R})$, then

$$\Psi(\mathbf{x}\mid\mathbf{x}_R)=\frac{\left[\frac{P(y=1\mid\mathbf{x})}{P(y=0\mid\mathbf{x})}\right]}{\left[\frac{P(y=1\mid\mathbf{x}_R)}{P(y=0\mid\mathbf{x}_R)}\right]}=\exp\left[\beta_1(x_1-x_{1,R})+\cdots+\beta_p(x_p-x_{p,R})\right].$$

Often, it is mathematically convenient to define the reference habitat so that $\mathbf{x}_R = (1, 0, \ldots, 0)$, in which case

$$\Psi(\mathbf{x}\mid\mathbf{x}_R)=\exp(\boldsymbol{\beta}'\mathbf{x}-\beta_0)=\exp(\beta_1x_1+\cdots+\beta_px_p). \quad (7)$$

Although superficially identical to the RSF that Manly et al. (2002:100) proposed for the use–availability setting, this form of $\psi(\mathbf{x}\mid\mathbf{x}_R)$ derives from a different model and generally cannot be interpreted the same. Approximately unbiased estimates of odds ratios can be obtained from either random or case–control samples because odds ratios are unaffected by the model constant, β_0 or β_0^*. Hosmer and Lemeshow (2000)

provide further discussion with examples of interpretation in more complex settings.

Under narrow conditions, case–control results also are interpretable in terms of relative risk. Relative risk is the probability of use given **x** relative to the probability of use given a reference type, $\mathbf{x}_R$; that is,

$$\Re(\mathbf{x} \mid \mathbf{x}_R) = \frac{P(y=1 \mid \mathbf{x})}{P(y=1 \mid \mathbf{x}_R)}. \quad (8)$$

The odds ratio is related to relative risk as

$$\Psi(\mathbf{x} \mid \mathbf{x}_R) = \Re(\mathbf{x} \mid \mathbf{x}_R)\left[\frac{1-P(y=1 \mid \mathbf{x}_R)}{1-P(y=1 \mid \mathbf{x})}\right]. \quad (9)$$

Thus, if use is rare everywhere (i.e., $P(y = 1|\mathbf{x}) \approx 0$ for all **x**, including $\mathbf{x}_R$), then $\psi(\mathbf{x}|\mathbf{x}_R) \approx \Re(\mathbf{x}|\mathbf{x}_R)$, and the odds ratio can then be used to approximate relative risk. Compton et al. (2002:836) explicitly used this approximation in their study of wood turtle (*Glyptemys insculpta*, formerly *Clemmys insculpta*) habitat selection. However, this approximation is increasingly biased as $P(y = 1|\mathbf{x})$ increases (Fig. 2). Thus, using the odds ratio to approximate relative risk implicitly assumes that $P(y = 1|\mathbf{x})$ is small not just on average, but for all **x**, including $\mathbf{x}_R$.

Example 2: Case–control Sampling

Using the same true model as in Example 1, we again sampled pixels randomly with replacement, stochastically determining use as $y = I[U \leq P(y = 1|\mathbf{x})]$. Sampling continued until we had drawn n_0 = 1,000 pixels for which $y = 0$, and n_1 = 1,000 pixels for which $y = 1$. We repeated this process 1,000 times. For each replicate, the data were fit to Eq. (3) using the LOGIT module in SYSTAT and letting $\boldsymbol{\beta}^{*\prime}\mathbf{x} = \beta_0^* + \beta_1 ELEV$. Comparing results with known values, the mean (SE) of $\bar{\hat{\beta}}_1 = -3.008$ (0.157) was a nearly unbiased estimate of $\beta_1 = -3.000$, but $\bar{\hat{\beta}}_0^* = 6.741$ (0.353) greatly overestimated the true model constant, $\beta_0 = 5.673$. In this example, $n_1 = n_0$, n was small relative to N, and mean (unconditional) probability of use for the study area was

$$P(y=1) = \frac{\sum_1^N P(y=1 \mid \mathbf{x})}{N} = 0.259.$$

Therefore, as per Eq. (3), the expected bias of $\hat{\beta}_0^*$ was

$$\ln\left(\frac{P_1}{P_0}\right) = \ln\left(\frac{n_1/N_1}{n_0/N_0}\right) = \ln\left(\frac{N_0/N}{N_1/N}\right) = \ln\left(\frac{1-0.259}{0.259}\right) = 1.051,$$

and the observed bias was $\bar{\hat{\beta}}_0^* - \beta_0 = 1.045$. Using the odds ratio to approximate relative risk under the (erroneous) assumption that $P(y = 1|\mathbf{x})$ was small for all **x**, the model accurately indexed probability of use. However, because the relative importance of habitats with a high probability of use was overstated, the model predicted that use would be more concentrated than it really was (Fig. 3). Overall, we cannot recommend this approximation unless the rare use assumption is justified.

Implications for Modeling Habitat Selection

Using case–control sampling, logistic regression cannot be used to model a RSPF unless one can estimate the proportions of used and unused locations sampled and thereby correct the bias in the model constant. Case–control results typically must be evaluated in terms of odds ratios, which although easily obtained, can be difficult to interpret in the context of a given problem. Under narrow conditions, odds ratios are a good estimate of relative risk and can easily be interpreted as a RSF, but this interpretation is only an approximation whose validity rests on the assumption that probability of use is small for all

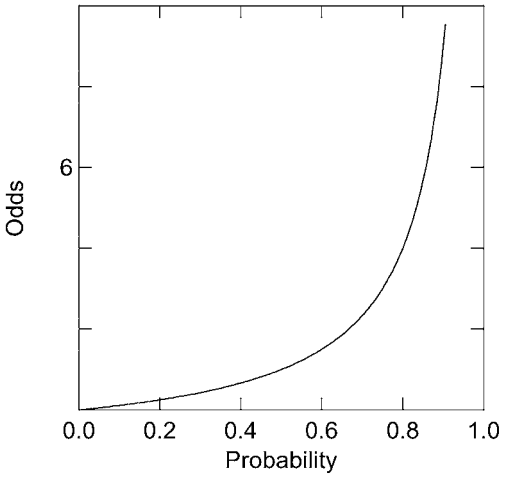

Fig. 2. Relationship between the odds of use ($P(y = 1|\mathbf{x}) / [1 - P(y = 1|\mathbf{x})]$) and probability of use ($P(y = 1|\mathbf{x})$). A curve with the same general form describes the relationship between the odds ratio and relative risk, but the axes will be scaled differently depending on the value of the reference habitat.

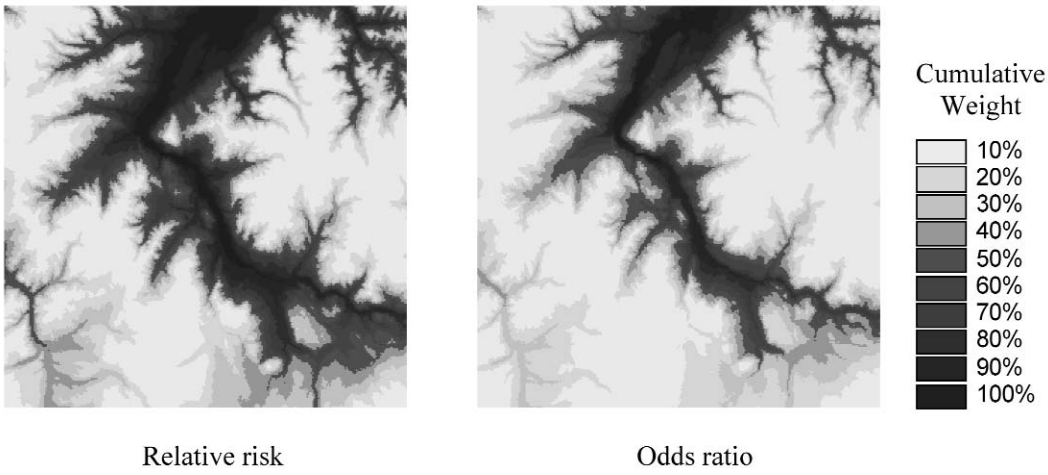

Fig. 3. Relative risk versus odds ratio for the study area and model of Example 2. Values are colored to enable comparison of distributions of total weights over the landscape. Comparison shows that odds ratios place proportionately greater weight on habitats with a high probability of use, yielding a map in which habitat values are indexed correctly but where the indices generally are not proportional to probability of use.

habitats. Simulations by Zhang and Yu (1998) suggest this approximation is unacceptable if, for any habitat, $P(y = 1|\mathbf{x}) > 0.10$. Where it is violated, relative probability of use may be greatly overestimated for high-probability locations, as illustrated in Example 2. When interpreting logistic regression in terms of odds ratios or relative risk, the reference habitat and relevant assumptions should be clearly identified.

True case–control designs are uncommon in wildlife studies, being strictly applicable only when used and unused habitats are (or are assumed to be) distinguishable. For example, in his study of the greater prairie chicken (*Tympanuchus cupido*), Niemuth (2003) used logistic regression to compare habitats used as leks versus those not used as leks, implicitly assuming that use and nonuse were detectable without error. His study illustrates, however, the need for caution when interpreting results. Because Niemuth (2003) interpreted his logistic regression results in terms of Eq. (1), he implicitly and inappropriately assumed that his data were gathered according to a random rather than a case–control design.

USE–AVAILABILITY SAMPLING

General Sampling Model

With a use–availability design, we randomly sample, with replacement, n_1 locations from the subpopulation of N_1 used locations and n_0 locations from all N available sites. If use is rare, at least on average (i.e., $P(y = 1) \approx 0$) then the use–availability and case–control designs are approximately equivalent because the sample of available sites will consist almost entirely of unused sites. In general, however, use–availability differs from the previous sampling designs because the sample of available locations can contain observations of both used and unused sites. If q is the unconditional probability of use, $P(y = 1)$, then we expect that our sample of available habitats will be comprised, on average, of $(1 - q)n_0$ unused and qn_0 used locations. From a case–control perspective, q is the expected contamination rate of the control sample, leading Lancaster and Imbens (1996) to refer to this design as case–control sampling with contaminated controls. Cosslett (1981) and Steinberg and Cardell (1992) labeled it a supplementary sampling design.

To deal with contaminated controls, we expand the sampling model following Lancaster and Imbens (1996). Let $h = n_1 / n$ (where $n = n_0 + n_1$) be the proportion of observations for which we observe $y = 1$. We make no assumption about the value of y for the n_0 observations of available locations because this sample is contaminated with some unknown proportion of used locations. Also, let s indicate sampling stratum, so that the n_1 observations of used sites are assigned the value $s = 1$, and the n_0 observations of available sites are assigned $s = 0$. As before, let $\eta = 1$ if a location appears in our sample, and $\eta = 0$ otherwise. Now, define $P(s = 1|\mathbf{x}, \eta = 1)$ as the probability that a

location will be among the n_1 locations for which use is actually observed, conditioned on the habitat and the location being among the samples drawn. The distinction between $P(y = 1|\mathbf{x})$ and $P(s = 1|\mathbf{x}, \eta = 1)$ is critical; the former is the probability of use conditioned solely on habitat (i.e., the RSPF), whereas the latter is the conditional probability that a sampled site will be among the locations for which use is observed.

Lancaster and Imbens (1996) derived the general model

$$P(s=1|\mathbf{x},\eta=1)=\frac{\frac{h}{q}P(y=1|\mathbf{x})}{\frac{h}{q}P(y=1|\mathbf{x})+1-h}, \quad (10)$$

where $P(y = 1|\mathbf{x})$ can take the form of any valid probability model. Their derivation assumes that $P(\mathbf{x}|s = 1) = P(\mathbf{x}|y = 1)$, making Eq. (10) a large-sample approximation. Dividing numerator and denominator by $(1 - h)$, Eq. (10) can be rewritten as

$$P(s=1|\mathbf{x},\eta=1)=\frac{\frac{h}{q(1-h)}P(y=1|\mathbf{x})}{1+\frac{h}{q(1-h)}P(y=1|\mathbf{x})}. \quad (11)$$

Defining $P_1 = n_1/N_1$ and $P_A = n_0/N$ as the respective proportions of used and available locations included in our sample, it follows that under finite sampling $h/[q(1 - h)] = P_1/P_A$. Therefore, under the assumption that n_0/N is quite small, Eq. (11) is approximately equivalent to Eq. (5.8) of Manly et al. (2002:99), which was derived independently. Two specific formulations of Eq. (10) have been proposed, whereby $P(y = 1|\mathbf{x})$ is assumed to conform to either the exponential (Manly et al. 2002:100) or the logistic model (Lancaster and Imbens 1996). We discuss both, but neither can be fit using logistic regression. Only the exponential form previously has been used in habitat modeling.

Sampling Model – Exponential Form

With use–availability sampling, Manly et al. (2002:100) assumed that the RSPF could be approximated by the exponential function

$$P(y=1|\mathbf{x})=\exp(\boldsymbol{\beta}'\mathbf{x}), \quad (12)$$

where $\boldsymbol{\beta}'\mathbf{x} \le 0$ for all $\mathbf{x}$. The constraint $\boldsymbol{\beta}'\mathbf{x} \le 0$ ensures $P(y = 1|\mathbf{x}) \le 1$. Substituting into Eq. (11) yields

$$P(s=1|\mathbf{x},\eta=1)=\frac{\exp(\boldsymbol{\beta}^{*\prime}\mathbf{x})}{1+\exp(\boldsymbol{\beta}^{*\prime}\mathbf{x})}, \quad (13)$$

where $\boldsymbol{\beta}^{*\prime} = (\beta_0^*, \beta_1, ..., \beta_p)$, $\beta_0^* = \beta_0 + 1n(P_1/P_A)$, and $\boldsymbol{\beta}'\mathbf{x} \le 0$. This is Eq. (8.6) of Manly et al. (1993:127) and, under the assumption that n_0/N is small, also is approximately equivalent to Eq. (5.10) of Manly et al. (2002:100). The log-likelihood is

$$L(\boldsymbol{\beta}^*)=\sum_{i=1}^{n}\{s_i\ln[P(s_i=1|\mathbf{x}_i,\eta_i=1)] + (1-s_i)\ln[1-P(s_i=1|\mathbf{x}_i,\eta_i=1)]\}. \quad (14)$$

At first glance, Eqs. (13) and (14) appear to specify a logistic model, leading Manly et al. (2002:100) to recommend that logistic regression be used to fit model (13) and thereby estimate the parameters of model (12). This approach relies on at least 2 critical assumptions. First, the constraint $\boldsymbol{\beta}'\mathbf{x} \le 0$ is assumed to be either optional or somehow satisfied by the logistic regression procedure. If true, then parameter estimates should always translate into valid probability estimates. Second, RSFs calculated from the resulting parameter estimates are assumed to be proportional to the true probability of use; that is, we should observe $\exp(\hat{\boldsymbol{\beta}}'\mathbf{x} - \hat{\beta}_0)/P(y = 1|\mathbf{x}) = k$, for some positive constant k. Examples 3 and 4 show that neither assumption is necessarily true. In epidemiological studies, where Eq. (12) is known as the log-binomial model (Schouten et al. 1993, Skov et al. 1998), use of this approach has been similarly criticized (Edwardes 1995, Ma and Wong 1999).

Example 3: Use–Availability, Exponential Form I

In this example, we show that using logistic regression to fit model (13) cannot guarantee that the resulting probability model will be valid, even when the true underlying model is exponential and the resulting parameter estimates are unbiased. Let x_1 be a continuous positive covariate, distributed in the populations of used and available locations as $f_1(x_1) = 2\exp(-2x_1)$ and $f(x_1) = \exp(-x_1)$, respectively. Also, let $q = 0.5$. From Bayes' Rule we get

$$P(y=1|x_1)=\frac{f_1(x_1)q}{f(x_1)}=\exp(-x_1). \quad (15)$$

This is an RSPF of exponential form, where $\beta_0 = 0$ and $\beta_1 = -1$. Substituting into Eq. (11) and specifying that samples will be drawn so that $h = 0.5$, we get

$$P(s=1\,|\,x_1,\eta=1)=\frac{\exp[\ln(2)-x_1]}{1+\exp[\ln(2)-x_1]},$$

where $x_1 > 0$. This is model (13), with

$$\beta_0^* = \beta_0 + \ln\left(\frac{P_1}{P_A}\right) = \beta_0 + \ln\left(\frac{h}{q(1-h)}\right) = \beta_0 + \ln(2).$$

Thus, an approximately unbiased estimate of β_0 is $\hat{\beta}_0 = \hat{\beta}_0^* - \ln(2)$. Using S-PLUS 2000 (MathSoft 1999), we drew 1,000 random samples each from $f_1(x_1)$ and $f(x_1)$ then estimated the relevant coefficients using the glm function with a binomial family argument. This process was repeated 1,000 times. Mean estimates (SE) were $\bar{\hat{\beta}}_0^*$ = 0.696 (0.051), $\bar{\hat{\beta}}_0$ = 0.003 (0.051), and $\bar{\hat{\beta}}_1$ = –1.005 (0.079). Estimates were essentially unbiased. However, because the logistic regression procedure did not impose the required constraint, $\boldsymbol{\beta}'\mathbf{x} \le 0$, the estimated maximum probability of use was >1 in 505 of the 1,000 simulations. Of those, an average of 6% of the locations sampled had fitted probabilities >1. We also evaluated whether the estimated RSFs were proportional to probability of use (i.e., whether RSF = *k*RSPF for some positive constant *k*) as required by definition. For this example, we know that $k = 1$. On average, observed values of

$$\hat{k} = \frac{\exp(\hat{\boldsymbol{\beta}}'\mathbf{x} - \hat{\beta}_0)}{P(y=1\,|\,\mathbf{x})}$$

were clustered around 1, but values associated with any particular replicate varied systematically and were not constant (Fig. 4). Thus, the statistical expectation of proportionality does not guarantee that the estimated RSF will, in fact, be even approximately proportional to probability of use in a given study.

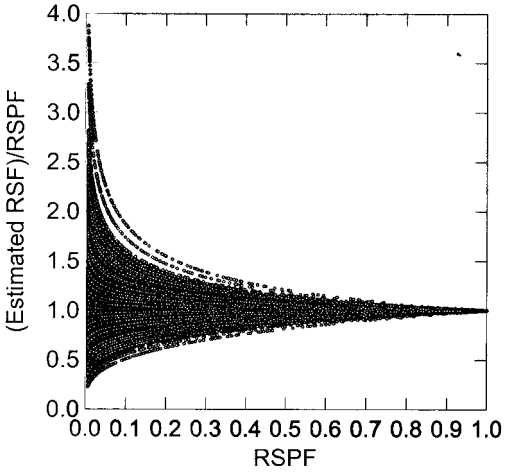

Fig. 4. Observed ratios of the resource selection function (RSF), estimated according to the method of Manly et al. (2002:100), and known resource selection probability function (RSPF) plotted on the RSPF, for 100 randomly selected locations for each of the 1,000 RSF models fit in Example 4. The RSF must be proportional to the RSPF, by definition. Therefore, if the method of Manly et al. (2002:100) yielded valid RSFs, this graph should have been comprised of 1,000 approximately horizontal lines.

Example 4: Use–Availability, Exponential Form II

Next, we illustrate the confounding effects of using common model-selection procedures together with logistic regression to fit use–availability data to model (13). Using the same true model as in Example 1, we sampled randomly with replacement, drawing n_1 = 1,000 pixels for which $y = 1$. Each was labeled as belonging to sampling stratum $s = 1$. Without regard to *y*, we then randomly drew n_0 = 1,000 pixels with replacement from our model study area and assigned each to stratum $s = 0$. This procedure was repeated 1,000 times. For each replicate, the data were fit to model (13) using the LOGIT module in SYSTAT. Preliminary analyses suggested that a polynomial of order $m \ge 4$ was needed to approximate the logistic form of the true model; therefore, we fit the data from each replicate to models of order $m = 1$ to 5 and used Akaike's Information Criterion (AIC; Burnham and Anderson 2002:61) to select the most parsimonious model.

To evaluate whether the AIC-selected models violated the assumption that $\boldsymbol{\beta}'\mathbf{x} \le 0$, we calculated probabilities of use implied by each model. We first corrected for sampling bias using known values of P_1 and P_A to calculate $\hat{\beta}_0 = \hat{\beta}_0^* - \ln(P_1/P_A)$, where

$$\ln\left(\frac{P_1}{P_A}\right) = \ln\left(\frac{N}{N_1}\right) = \ln\left(\frac{1}{q}\right) = \ln\left(\frac{1}{0.259}\right) = 1.351. \quad (16)$$

The value $\hat{\beta}_0$ was then substituted for $\hat{\beta}_0^*$ to obtain $\hat{\boldsymbol{\beta}}$. We then applied each of our 1,000 bias-corrected models to estimate probability of use for every pixel in our study area and recorded the maximum estimated probability for each model. Probability estimates ranged as high as $\hat{P}(y=1\,|\,\mathbf{x})$ = 5.71 (Fig. 5), and estimates >1 were observed for 75% of the models, indicating that the assumption $\boldsymbol{\beta}'\mathbf{x} \le 0$ was consistently violated.

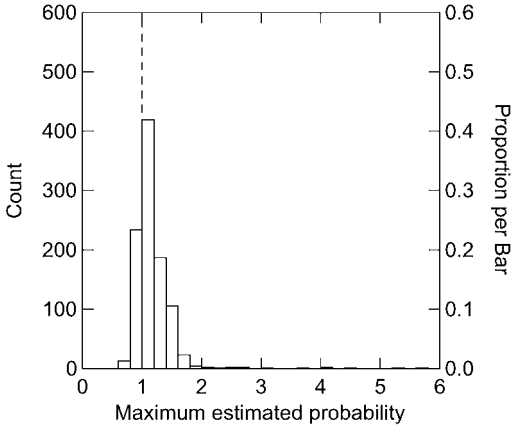

Fig. 5. Distribution of maximum estimates of probability of use, calculated by applying the 1,000 best models of Example 3, ranked by Akaike's Information Criterion, to the elevations in the model study area. The dashed line indicates the upper bound of 1 that should have been observed if the procedure of Manly et al. (2002:100) yielded valid probability models.

We also evaluated whether RSFs were proportional to probability of use. For each AIC-selected model, a RSF was estimated as $\exp(\hat{\boldsymbol{\beta}}^{*\prime}\mathbf{x} - \hat{\beta}_0^*)$, as per Manly et al. (2002:100). For each model, we then randomly sampled 100 locations with replacement from our study area, calculated

$$\hat{k} = \frac{\exp(\hat{\boldsymbol{\beta}}'\mathbf{x} - \hat{\beta}_0)}{P(y = 1 \mid \mathbf{x})}$$

for each location, and plotted those values against $P(y = 1 \mid \mathbf{x})$. Results showed that RSFs were not proportional to probability of use (Fig. 6). Although RSFs usefully indexed probability of use in many cases, this was not guaranteed. For models of polynomial order $m > 1$, locations with very different probabilities of use often were indexed by identical or nearly identical RSF values.

Sampling Model – Logistic Form

Lancaster and Imbens (1996) presented an alternative form of model (10). In the notation of our paper, they assumed

$$P(y = 1 \mid \mathbf{x}) = \frac{1}{1 + \exp(-\boldsymbol{\beta}'\mathbf{x})}. \qquad (17)$$

This is the logistic model of Eq. (1) with terms rearranged and requires no constraint on $\boldsymbol{\beta}'\mathbf{x}$. Substituting into Eq. (10) yields

$$P(s = 1 \mid \mathbf{x}, \eta = 1) = \frac{1}{1 + \frac{q(1-h)}{h} + \exp\left[-\boldsymbol{\beta}'\mathbf{x} + \ln\left(\frac{q(1-h)}{h}\right)\right]}. \qquad (18)$$

Substituting Eq. (1) into Eq. (11), the model also can be written as

$$P(s = 1 \mid \mathbf{x}, \eta = 1) = \frac{\exp\left[\boldsymbol{\beta}'\mathbf{x} + \ln\left(\frac{h}{q(1-h)}\right)\right]}{1 + \exp\left[\boldsymbol{\beta}'\mathbf{x} + \ln\left(1 + \frac{h}{q(1-h)}\right)\right]}. \qquad (19)$$

The log-likelihood is given by Eq. (14). This model is not logistic and cannot be fit using logistic regression (Lancaster and Imbens 1996). Lancaster and Imbens (1996) provided 2 methods (maximum likelihood and generalized method of moments) for estimating the model parameters $\boldsymbol{\beta}$ and the nuisance parameter q, and showed that the 2 are essentially equivalent for estimating $\boldsymbol{\beta}$. In Example 5, we use maximum likelihood to estimate $\boldsymbol{\beta}$ and q and compare estimates with known values.

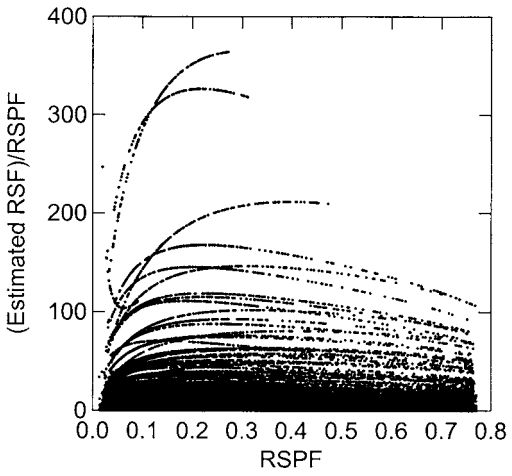

Fig. 6. Observed ratios of the resource selection function (RSF), estimated according to the method of Manly et al. (2002:100), and known resource selection probability function (RSPF) plotted on the RSPF, for 100 randomly selected locations for each of the 1,000 RSF models fit in Example 5. The RSF must be proportional to the RSPF, by definition. Therefore, if the method of Manly et al. (2002:100) yielded valid RSFs, this graph should have been comprised of 1,000 approximately horizontal lines.

Example 5: Use–Availability, Logistic Form

Using the same 1,000 samples as in Example 4, we fit the data to Eq. (18) using the nonlinear regression module (NONLIN) in SYSTAT. To obtain maximum-likelihood estimates, we set the loss function equal to the negative of the log-likelihood (Eq. [14]). Because $P(s = 1|\mathbf{x}, \eta = 1)$ is binomial, variance was a function of $\mathbf{x}$. We therefore used iterative reweighting (Cox and Snell 1989:19), whereby the weight for each observation was recalculated as the inverse of the variance,

$$\frac{1}{\hat{P}(s=1|\mathbf{x},\eta=1)\,[1-\hat{P}(s=1|\mathbf{x},\eta=1)]},$$

following each iteration of the least-squares nonlinear regression procedure. The nonlinear regression procedure was sensitive to starting values and often failed to converge if starting values were very far from actual values. To resolve this problem in a manner suited to actual field studies, we first fit each of the 1,000 data sets to the simple logistic model of Example 1 using standard logistic regression. The resulting estimates ($\hat{\beta}_0$ and $\hat{\beta}_1$) were then used as starting values in our nonlinear regressions. To calculate starting values for q, we used elevation data for our study area together with $\hat{\beta}_0$ and $\hat{\beta}_1$ to estimate probability of use as

$$\hat{P}(y=1|\mathbf{x}) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 ELEV)}{1+\exp(\hat{\beta}_0 + \hat{\beta}_1 ELEV)}, \qquad (20)$$

then averaged those estimates over the study area to obtain $\hat{q} = \bar{\hat{P}}(y = 1|\mathbf{x}) = \hat{P}(y = 1)$. The final mean (SE) estimates of $\bar{\hat{\beta}}_0 = 5.594$ (1.528), $\bar{\hat{\beta}}_1 = -3.015$ (0.444), and $\bar{\hat{q}} = 0.250$ (0.082) obtained via nonlinear regression were nearly unbiased. However, uncertainties associated with $\hat{\beta}_0$ and $\hat{\beta}_1$ were much greater than we observed using random sampling (cf. Example 1). Estimates of $P(y = 1|\mathbf{x})$ were similarly uncertain. Final estimates of β_0 and β_1 were substituted into Eq. (20) and, for each replicate, $P(y = 1|\mathbf{x})$ was estimated for 100 randomly selected pixels. The resulting estimates, $\hat{P}(y = 1|\mathbf{x})$, reflected the considerable uncertainty in $\hat{\beta}_0$ and $\hat{\beta}_1$ (Fig. 7).

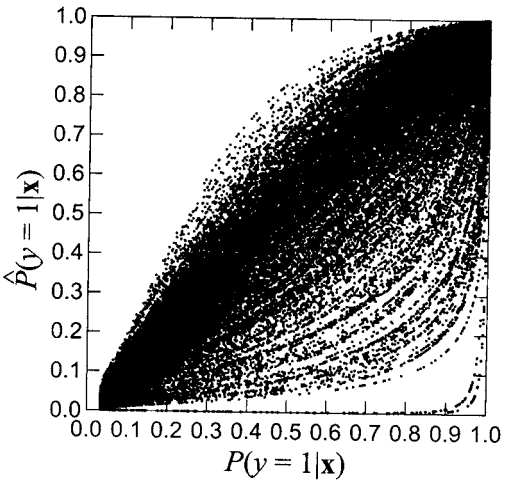

Fig. 7. Relationship between the estimated and true conditional probability of use [$\hat{P}(y = 1|\mathbf{x})$and $P(y = 1|\mathbf{x})$, respectively] for the 1,000 models fit to data gathered according to a use–availability sampling design.

Implications for Modeling Habitat Selection

Use–availability is perhaps the most common sampling design in habitat-selection studies. Unfortunately, logistic regression is commonly misused in this setting. For example, applying logistic regression to use–availability data for mountain quail (*Oreortyx pictus*), Wyoming toads (*Bufo baxteri*), and southwestern myotis (*Myotis auriculus*), respectively, Brennan et al. (1986), Parker and Anderson (2003), and Bernardos et al. (2004) incorrectly interpreted their results in terms of Eq. (1)—as if the data had been gathered according to a random sampling design. To avoid these and other types of errors requires an understanding of the many nuances and assumptions that accompany the use–availability design. In the following sections, we discuss some of the more important ones.

Use–Availability as an Approximation to Case–Control.—Use–availability is approximately equivalent to a case–control design if the unconditional probability of use (q) is small because we then expect that the control sample will be composed almost exclusively of unused sites. Thus, use–availability studies that treat the data as if they had been collected using a case–control design implicitly assume that use is rare, at least on average. Examples of use–availability studies treated as case–control designs include those for the northern spotted owl (*Strix occidentalis caurina*), sage grouse (*Centrocercus urophasianus*), and flammulated owl (*Otus flammeolus*; Ramsey et al. 1994); grizzly bear (Mace et al. 1996); and wood turtle (Compton et al. 2002). Of these, only Compton et al. (2002) explicitly discussed the underlying rare use assumption. Particularly for telemetry-based studies (e.g., the sage grouse and

grizzly bear studies of Ramsey et al. [1994)] and Mace et al. [1996]), it often is not self-evident that randomly selected available habitats were in fact unused or that q was necessarily small. When analyzing use–availability data as if they had been gathered according to a case–control design, the assumption that q is small merits explicit and careful consideration. Also, when using this approximation, results generally must be interpreted as odds ratios (as per Eq. [7]) rather than RSFs—the latter being, by definition, proportional to $P(y = 1|\mathbf{x})$, while the former are not. Only when $P(y = 1|\mathbf{x})$ is small for all habitats (i.e., all $\mathbf{x}$ values, including $\mathbf{x}_R$) can odds ratios be interpreted as approximate RSFs. Although reasonable in some cases, we believe the assumption that $P(y = 1|\mathbf{x})$ is small for all habitats is dangerous in general. Knowing when use is rare enough for this approximation to hold will be strictly possible only when P_1 and P_A are both known (i.e., when q is known and model [12] is a good approximation to the true probability of use). Unfortunately, determining when these conditions are even approximately satisfied is often difficult. The assumption that q is small is far less stringent and thus more likely to prove useful in wildlife studies.

Use–Availability and the Exponential Model.—In general, logistic regression cannot be applied to analyze use–availability data because the underlying probability model (Eq. [10]) is not logistic. A common misconception (e.g., Manly et al. 2002:100) is that standard logistic regression, when applied to use–availability data, yields maximum-likelihood estimates of the parameters of the exponential model of Eq. (12). Maximizing the likelihood of the logistic model and maximizing the likelihood subject to the constraint $\boldsymbol{\beta}'\mathbf{x} \le 0$ are not equivalent problems and, in general, will not yield the same solutions. Because standard logistic regression does not impose the requisite constraint, using it to fit model (13) and thereby estimate the parameters of model (12) does not guarantee maximum-likelihood solutions, valid standard errors, or even valid probabilities, as demonstrated in Examples 3 and 4. Furthermore, invalid probabilities imply invalid log-likelihoods—a fact with serious implications for studies using this approach together with likelihood-based model-selection methods like AIC.

Perceptions regarding the exponential model are confounded by the fact that logistic regression can yield mathematically acceptable results in some cases. Nielsen et al. (2004) reported that logistic regression yielded approximately unbiased parameter estimates and valid probability estimates (i.e., $0 \le \hat{P}(y = 1|\mathbf{x}) \le 1$) in use–availability simulations with contamination rates as high as $q = 0.5$. They concluded that logistic regression should yield valid results, unless q is quite large. In Examples 3 and 4, however, we observed invalid probability estimates when $q = 0.259$ and $q = 0.5$, respectively. This illustrates that the validity of the estimated model is not determined solely by q. The same conclusion can be reached more formally by considering the inequality $\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p) \le 1$, which follows from model (12) and the constraint $\boldsymbol{\beta}'\mathbf{x} \le 0$. Multiplying both sides by $n_1/(n_0 q)$ and remembering that

$$\beta_0^* = \beta_0 + \ln\left(\frac{n_1}{n_0 q}\right)$$

for the exponential model, this inequality can be rewritten as

$$\exp(\beta_0^* + \beta_1 x_1 + \cdots + \beta_p x_p) \le \frac{n_1}{n_0 q}.$$

Rearranging terms yields

$$q \le \frac{n_1}{n_0 \exp\left(\beta_0^* + \beta_1 x_1 + \cdots + \beta_p x_p\right)}, \qquad (21)$$

which must hold for all $\mathbf{x}$ to ensure that estimated probabilities are ≤1. One implication of Eq. (21) is that no single threshold value exists for q that will ensure a valid probability model. Instead, the permissible upper bound for q is a function of the particular exponential model ($\hat{\boldsymbol{\beta}}^*$), study area ($\mathbf{x}$ values), and the sampling proportion n_1/n_0. Indeed, for every exponential model, study area, and sampling proportion, some upper bound for q is implicitly assumed. This assumed upper bound can be surprisingly small. Consider the fernbird (*Bowdleria punctata*) example of Manly et al. (2002:105), in which $n_1 = 24$, $n_0 = 25$, and

$$\exp(\hat{\boldsymbol{\beta}}^{*\prime}\mathbf{x}) = \exp(-10.73 + 7.79x_1 + 0.21x_2 + 0.88x_3).$$

Examining only the 49 values of $\mathbf{x}$ actually observed in that study (Manly et al. 2002:40, Table 3.4), a maximum value of $\exp(\hat{\boldsymbol{\beta}}^{*\prime}\mathbf{x}) = 11{,}968$ was obtained for the observation $\mathbf{x}' = (1, 1.2, 14, 8.9)$. For the estimated model to yield valid probabilities, we must therefore assume that $q \le 24/(25 \times$

11,968) ≈ 0.00008. This bound would undoubtedly be lower if covariate values from all locations in the study area had been examined. Even so, the implicit threshold of $q \leq 0.00008$ contrasts sharply with the threshold of $q \leq 0.5$ suggested by Nielson et al. (2004) and further illustrates that no single threshold value for q can ensure a valid probability model in all cases. Knowing whether logistic regression results are acceptable in any particular instance is likely to be study and model dependent and hence difficult in practice. However, given information about $\mathbf{x}$ for all locations in a study area, users of the exponential model can apply Eq. (21) to calculate the assumed upper bound on q and at least consider whether that assumption might be realistic.

Despite the problems discussed above, one may reasonably question whether RSFs obtained for the exponential model might be proportional to probability of use. Examples 3 and 4 (Figs. 4, 6) showed that proportionality is not assured. Observed relationships between fitted RSFs and the true probability of use typically were nonlinear and, in some cases, maximum RSF values were associated with intermediate rather than maximum probabilities. Thus, observed RSFs were rarely proportional to probability of use and sometimes were unreliable even as an index of that probability. In Example 4, in which we approximated a logistic model using an exponential model with polynomial terms, lack of proportionality probably was due, in part, to model misspecification. In Example 3, however, the model was correctly specified, yet calculated RSFs still were not proportional to the RSPF. Instead, relationships between the RSFs and the RSPF were a function of $\mathbf{x}$. This can be understood as follows. Recall that by definition RSF/RSPF = k, where k is a positive constant. Estimates of k for Example 3 are therefore given by

$$\begin{aligned}\hat{k} &= \frac{\text{Estimated RSF}}{\text{RSPF}} \\ &= \frac{\exp(\hat{\beta}_1 x_1)}{\exp(\beta_0 + \beta_1 x_1)} \\ &= \frac{1}{\exp(\beta_0)} \exp[(\hat{\beta}_1 - \beta_1) x_1].\end{aligned} \tag{22}$$

The expected proportionality constant in Example 3 was $k = 1/\exp(\beta_0) = 1$. From Eq. (22), we see that this expectation will be realized only if $x_1 = 0$ or β_1 is estimated without error (i.e., $\hat{\beta}_1 - \beta_1 = 0$). In all other cases, $\hat{k}$ is an exponential function of x_1 and the estimation error. The observed lack of proportionality between the RSF and RSPF raises serious questions about applications that rely critically on the assumption of proportionality. For example, RSF-based assessments of risk (McDonald and McDonald 2002), carrying capacity (Boyce and Waller 2003), or population size (Boyce and McDonald 1999) seem especially difficult to justify.

Beyond these issues, the general form of the exponential model exhibits several undesirable properties. First, because modeled probabilities of use are bounded on the upper end only via the constraint $\boldsymbol{\beta}'\mathbf{x} \leq 0$, the validity of any particular model depends on $\mathbf{x}$. Consequently, even if the underlying constraint were satisfied for a model constructed for 1 area or time period, it may not be satisfied if that model is extrapolated to areas or periods where the range of $\mathbf{x}$ values is different. Extrapolations of exponential models (e.g., Boyce and Waller 2003) are therefore suspect on purely mathematical grounds. Second, using the exponential model to approximate non-exponential relationships requires higher-order polynomials. For example, the most commonly assumed relationship between probability of use and a continuous covariate (x_1) is

$$P(y = 1 \mid x_1) = \frac{\exp(\beta_0 + \beta_1 x_1)}{1 + \exp(\beta_0 + \beta_1 x_1)},$$

which yields the classic sigmoidal curve assumed for random and case–control sampling. To approximate this form using model (12) would require a fourth- or fifth-order polynomial; that is, it would require that for $m = 4$ or 5,

$$P(y = 1 \mid x_1) = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_1^{\,2} + \cdots + \beta_m x_1^{\,m}). \tag{23}$$

Such models are rare in the literature. Although use of quadratic terms has become more common, we know of only 2 studies (McDonald and McDonald 2002, Manly et al. 1993:112) that included cubic terms. The fact that higher-order polynomials are rarely even considered suggests that few researchers appreciate this property of the exponential model. More importantly, polynomial models become increasingly unstable as higher-order terms are added, confound model interpretation, and impose an unnecessarily high cost in terms of commonly used model-selection criteria like AIC. Such problems are exacerbated when multiple predictors or interaction terms are included. Third, the exponential model implies that use and

nonuse are influenced by habitat in fundamentally different ways. For a continuous covariate, the model implies that relative probabilities of use are independent of the covariate value because

$$\frac{P(y=1\mid x_1+a)}{P(y=1\mid x_1)}=\frac{\exp[\beta_0+\beta_1(x_1+a)]}{\exp(\beta_0+\beta_1x_1)}=\exp(a\beta_1)$$

for any arbitrary constant (a). Thus, an increase of a units in the covariate x_1 has the same effect on relative probability of use regardless of the actual value of x_1. In contrast, relative probabilities of nonuse are given by

$$\frac{P(y=0\mid x_1+a)}{p(y=0\mid x)}=\frac{1-\exp[\beta_0+\beta_1(x_1+a)]}{1-\exp(\beta_0+\beta_1x_1)},$$

which depends on x_1. The biological justification for this lack of symmetry in the habitat-selection process is unclear.

Use of the exponential model also undermines comparability among studies employing different sampling designs. Imagine a situation where use, nonuse, and habitat data are collected for a single continuous covariate (x_1) according to each of the 3 sampling designs we outlined above. Using these data, 3 different analyses are conducted, assuming as appropriate (Eqs. [1], [4], [12])

$$P(y=1\mid\mathbf{x})=\frac{\exp(\beta_0+\beta_1x_1)}{1+\exp(\beta_0+\beta_1x_1)},$$

$$P(y=1\mid\mathbf{x})=\frac{\exp\left[\boldsymbol{\beta}_0^*-\ln\left(\frac{P_1}{P_0}\right)+\beta_1x_1\right]}{1+\exp\left[\boldsymbol{\beta}_0^*-\ln\left(\frac{P_1}{P_0}\right)+\beta_1x_1\right]},\ \text{or}$$

$$P(y=1\mid\mathbf{x})=\exp(\beta_0+\beta_1x_1+\beta_2x_1^2+\beta_3x_1^3+\beta_4x_1^4+\beta_5x_1^5),\qquad \boldsymbol{\beta}'\mathbf{x}\le 0.$$

Ideally, these approaches should yield comparable results because they all describe the same underlying reality. However, the coefficients in the third model are not comparable to those in the first 2, nor are the covariates themselves treated the same because polynomial transformations are required to ensure that the exponential model yields approximately the same values for probability of use. This lack of comparability undermines efforts to develop a truly unified framework for habitat-selection modeling.

The Exponential Model as a Logistic Discriminant Function.—Despite serious theoretical and practical concerns regarding the exponential model, it has been widely applied, with most researchers reporting that the resulting RSFs appear reasonable. If truly flawed, why should this model seemingly perform well? One explanation is that an RSF of the form $\exp(\beta_1x_1+\cdots+\beta_px_p)$ is proportional to a logistic discriminant function. To see this, we reformulate the problem as follows. Let $f_1(\mathbf{x})$ be the distribution of covariates in a pool consisting of all habitat choices made by the species of interest, and let $f_2(\mathbf{x})$ be the distribution in a pool of all choices made randomly. The sampling universe is hypothetical, containing all possible realizations of the 2 different sampling methods; that is, we imagine a hypothetical population $g(\mathbf{x})$ such that $g(\mathbf{x})=\pi f_1(\mathbf{x})+(1-\pi)f_2(\mathbf{x})$, where π is the probability of drawing from $f_1(\mathbf{x})$, and $1-\pi$ is the probability of drawing from $f_2(\mathbf{x})$. In this hypothetical formulation, $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ are mathematically distinct populations. We want to determine which population, $f_1(\mathbf{x})$ or $f_2(\mathbf{x})$, a particular observation $\mathbf{x}$ belongs to. This is a classification problem, and 1 way of handling the problem is to assume (Seber 1984:308)

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})}=\exp(\beta_0+\beta_1x_1+\cdots+\beta_px_p).$$

This logistic discriminant function is used to classify $\mathbf{x}$ as belonging to $f_1(\mathbf{x})$ if $\beta_0+\beta_1x_1+\cdots+\beta_px_p>\ln[(1-\pi)/\pi]$. Of course, π is rarely known, and a common practice is to classify based on sampling proportions; that is, classify $\mathbf{x}$ as belonging to $f_1(\mathbf{x})$ if

$$\beta_0+\beta_1x_1+\cdots+\beta_px_p>\ln\left(\frac{n_1}{n_2}\right),$$

where n_1 and n_2 are the sample sizes drawn from $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$, respectively. Thus, using logistic regression to fit use–availability data to the exponential model yields a logistic discriminant function (Seber 1984:308–317, Manly 1994:118–125), where the nuisance parameter q and the troublesome constraint $\boldsymbol{\beta}'\mathbf{x}\le 0$ are no longer considerations. Being proportional to this discriminant function, $\exp(\beta_1x_1+\cdots+\beta_px_p)$ should allow meaningful ranking of habitats—although, as seen in Example 3, use of polynomial regression can undermine the accuracy of those rankings. This approach also should allow meaningful identification of those habitat characteristics (or their surrogates) most strongly associated with habitat selection.

When using logistic regression to obtain a logistic discriminant function, the result must be interpreted differently than in RSF analysis. To see this, let $f_0(\mathbf{x})$ denote the population of unused habitats. Use of logistic regression in RSF analysis assumes that the available population is a mixture of used and unused habitats (i.e., $f_2(\mathbf{x}) = qf_1(\mathbf{x}) + (1 - q)f_0(\mathbf{x})$). This implies that the result is a logistic discriminant function that distinguishes between $f_1(\mathbf{x})$ and $f_0(\mathbf{x})$ (use and nonuse), rather than $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ (observed use and random use). The latter interpretation is always valid under use–availability sampling, but the former is not unless one is willing to make strong assumptions. The difference is subtle, and confusion exists in the literature about which data-generating model applies. The use–availability sampling plan described by Manly et al. (2002:99) clearly assumes that the available population is a mixture of used and unused units (Manly et al. 2002:27), but in other discussions of resource selection, some of these same authors describe a sampling plan more like the logistic discriminant setting (e.g., the pie example in McDonald and McDonald [2002], where no population of unused pies exists). The study of northern raccoon (*Procyon lotor*) den-site selection by Henner et al. (2004) provides an example of a use–availability design where logistic regression was explicitly used to discriminate between observed use and random use, rather than use and nonuse.

Use–Availability and the Logistic Model.—Lancaster and Imbens (1996) provide a framework (Eq. [10]) for developing alternative models applicable to use–availability data. In doing so, they demonstrate the theoretical feasibility of estimating $P(y = 1|\mathbf{x})$ in this setting, while retaining the logistic model assumed in the random and case–control settings. This introduces a consistent mathematical framework, whereby researchers can estimate habitat-specific probabilities of use from the same underlying logistic model, regardless of sampling design. Moreover, in contrast to the exponential model, the logistic model is intrinsically bounded so that $0 \leq P(y = 1|\mathbf{x}) \leq 1$ regardless of $\mathbf{x}$, avoids use of high-order polynomials, and assumes symmetry with respect to habitat use–nonuse decisions. Nontrivial problems we encountered in applying the Lancaster-Imbens logistic model include the lack of commercially available software to implement their generalized method of moments solution. More seriously, their maximum-likelihood method failed to converge to a unique solution when using categorical covariates or some transformations of continuous covariates. Reasons for this are unclear, but alternative approaches apparently are needed to provide a generally robust method for estimating RSPFs from use–availability data. This problem is the subject of ongoing study.

MANAGEMENT IMPLICATIONS

The validity of habitat models is affected by many factors, including choice of statistical method and interpretation of results. Logistic regression is among the most popular methods for constructing habitat models but is easily misapplied. Because misapplications most often have been due to lack of understanding about the assumptions and constraints applicable under different sampling designs, we offer the following summary guidelines (Table 1).

With random sampling of use–nonuse data, logistic regression can be applied to estimate the conditional probability of use ($P(y = 1|\mathbf{x})$), which is the RSPF of Manly et al. (2002:27). With case–control sampling, logistic regression generally cannot be used to estimate $P(y = 1|\mathbf{x})$, but it

Table 1. Summary of methods and interpretations appropriate to particular sampling designs under a range of assumptions or conditions relevant to wildlife habitat-selection studies. $P(y = 1)$ is the unconditional probability of use; $P(y = 1|\mathbf{x})$ is the probability of use conditioned on the habitat ($\mathbf{x}$); RSPF is the resource selection probability function; and "≃" indicates approximation.

Sampling design	Special assumptions or conditions	Method	Interpretation
Random	None	Logistic regression	RSPF
Case–control	$P_0 = P_1$	Logistic regression	Habitat ranking
	None	Logistic regression	Odds ratio
	Use rare everywhere; $P(y = 1|\mathbf{x}) \simeq 0$ for all $\mathbf{x}$	Logistic regression	≃ Relative risk
Use–availability	None	Lancaster-Imbens[a]	RSPF
	None	Logistic regression[b]	Habitat ranking
	Use rare, on average; $P(y = 1) \simeq 0$	Logistic regression	≃ Odds ratio
	Use rare everywhere; $P(y = 1|\mathbf{x}) \simeq 0$ for all $\mathbf{x}$	Logistic regression	≃ Relative risk

[a] Lancaster and Imbens (1996) generalized method of moments or maximum-likelihood methods.
[b] Used here to obtain a logistic discriminant function.

does yield valid estimates of odds ratios. In general, odds ratios are not proportional to the RSPF and thus cannot be interpreted as an RSF. However, a special case exists if $P(y = 1 | \mathbf{x})$ is small for all habitats (i.e., all $\mathbf{x}$ values) because the odds ratio is then approximately equal to relative risk and thus can be treated as a good approximation to an RSF. With use–availability sampling, logistic regression will yield a logistic discriminant function that can be used to rank habitats based on a comparison of observed versus random use and identify those habitat characteristics most strongly correlated with habitat use. The reliability of habitat rankings may be undermined, however, if models are poorly specified. If use is rare, then use–availability data also may be treated as being approximately equivalent to a case–control sample. Logistic regression results may then be interpreted as either odds ratios (if $P(y = 1)$ is small) or RSFs (if $P(y = 1 | \mathbf{x})$ is small for all $\mathbf{x}$).

The assumption that estimated RSFs or RSPFs are proportional to the true probability of use is critical in some applications. Our results show that, regardless of the modeling approach used, this assumption should be considered carefully. First, unless the model is correctly specified, a statistical expectation of proportionality may not exist. Example 4 (Fig. 6) illustrates problems that can arise due to model-selection uncertainty. Second, even where the true model is correctly specified and a statistical expectation of proportionality exists, stochastic variation in parameter estimates can create a situation in which proportionality is rare. In our Monte Carlo trials, this situation occurred when variability about the statistical expectation varied with $\mathbf{x}$ (Examples 3, 5; Figs. 4, 7). This underscores the need for relatively large samples in applications in which proportionality is assumed.

Overall, use of logistic regression in habitat-selection studies currently requires that researchers navigate with meticulous care through a set of choices regarding sampling design, the underlying probability model, and associated assumptions. Ultimately, these determine the appropriateness of interpretations made from their analyses. New developments hint at the possibility of simplifying this complex situation. Specifically, RSFs previously have been accepted as a necessary complication because RSPFs were viewed as unattainable, except under random sampling. The logistic-based model of Lancaster and Imbens (1996) provides a solution to this problem for studies using only simple continuous covariates but apparently fails in other situations; thus, we do not recommend its use. Nonetheless, the model suggests the feasibility of directly estimating conditional probability of use from use–availability data. We encourage additional work in this direction to enable estimation of a single underlying probability model that is comparable across sampling designs.

ACKNOWLEDGMENTS

Montana State University, the U.S. Geological Survey, and the National Park Service supported this work. For their helpful reviews of this manuscript, we thank R. Anderson-Sprecher, R. J. Boik, M. S. Boyce, M. A. Hamilton, C. Johnson, B. C. Lubow, L. L. MacDonald, E. M. Olexa, G. A. Sargeant, W. L. Thompson, and an anonymous referee. Discussions with M. S. Boyce, M. A. Hamilton, B. F. J. Manly, and L. L. McDonald helped to clarify our thinking on many of the ideas contained herein. However, acknowledgment does not imply agreement with our ideas. We thank S. J. Boccadori for assisting with simulations.

LITERATURE CITED

ALLDREDGE, J. R., D. L. THOMAS, AND L. L. MCDONALD. 1998. Survey and comparison of methods for study of resource selection. Journal of Agricultural, Biological, and Environmental Statistics 3:237–253.

APPS, C. D., B. N. MCCLELLAN, J. G. WOODS, AND M. F. PROCTOR. 2004. Estimating grizzly bear distribution and abundance relative to habitat and human influence. Journal of Wildlife Management 68:138–152.

BERNARDOS, D. A., C. L. CHAMBERS, AND M. J. RABE. 2004. Selection of Gambel oak roosts by southwestern myotis in ponderosa pine-dominated forests, northern Arizona. Journal of Wildlife Management 68:595–601.

BOYCE, M. S., AND L. L. MCDONALD. 1999. Relating populations to habitats using resource selection functions. Trends in Ecology and Evolution 14:268–272.

———, AND J. S. WALLER. 2003. Grizzly bears for the Bitterroot: predicting potential abundance and distribution. Wildlife Society Bulletin 31:670–683.

BRENNAN, L. A., W. M. BLOCK, AND R. J. GUTIÉRREZ. 1986. The use of multivariate statistics for developing habitat suitability index models. Pages 177–182 *in* J. Verner, M. L. Morrison, and C. J. Ralph, editors. Wildlife 2000: modeling habitat relationships of terrestrial vertebrates. University of Wisconsin Press, Madison, Wisconsin, USA.

BURNHAM, K. P., AND D. R. ANDERSON. 2002. Model selection and inference: a practical information–theoretic approach. Second edition. Springer-Verlag, New York, New York, USA.

COMPTON, B. W., J. M. RHYMER, AND M. MCCOLLOUGH. 2002. Habitat selection by wood turtles (*Clemmys insculpta*): an application of paired logistic regression. Ecology 83:833–843.

COSSLETT, S. R. 1981. Efficient estimation of discrete-choice models. Pages 51–111 *in* C. F. Manski and D.

McFadden, editors. Structural analysis of discrete data with econometric applications. MIT Press, Cambridge, Massachusetts, USA.

COX, D. R., AND E. J. SNELL. 1989. Analysis of binary data. Second edition. Chapman & Hall, New York, New York, USA.

ENVIRONMENTAL SYSTEMS RESEARCH INSTITUTE. 1999. ArcView. Version 3.2. Environmental Systems Research Institute, Redlands, California, USA.

EDWARDES, M. D. DEB. 1995. Letter to the editor: Risk ratio and rate ratio estimation in case–cohort designs: hypertension and cardiovascular mortality. Statistics in Medicine 14:1609.

FLEISHMAN, E., R. MACNALLY, J. P. FAY, AND D. D. MURPHY. 2001. Modeling and predicting species occurrence using broad-scale environmental variables: an example with butterflies of the Great Basin. Conservation Biology 15:1674–1685.

HENNER, C. M., M. J. CHAMBERLAIN, B. D. LEOPOLD, AND L. W. BURGER, JR. 2004. A multi-resolution assessment of raccoon den selection. Journal of Wildlife Management 68:179–187.

HOSMER, D. W., AND S. LEMESHOW. 2000. Applied logistic regression analysis. Second edition. John Wiley & Sons, New York, New York, USA.

LANCASTER, T., AND G. IMBENS. 1996. Case–control studies with contaminated controls. Journal of Econometrics 71:145–160.

MA, S., AND C.-M. WONG. 1999. Letter to the editor: Estimation of prevalence proportion rates. International Journal of Epidemiology 28:175.

MACE, R. D., J. S. WALLER, T. L. MANLEY, L. J. LYON, AND H. ZUURING. 1996. Relationships among grizzly bears, roads and habitat in the Swan Mountains, Montana. Journal of Applied Ecology 33:1395–1404.

MACKENZIE, D. I., J. D. NICHOLS, G. B. LACHMAN, S. DROEGE, J. A. ROYLE, AND C. A. LANGTIMM. 2002. Estimating site occupancy rates when detection probabilities are less than one. Ecology 83:2248–2255.

MANLY, B. F. J. 1994. Multivariate statistical methods: a primer. Second edition. Chapman & Hall, New York, New York, USA.

———, L. L. MCDONALD, AND D. L. THOMAS. 1993. Resource selection by animals: statistical design and analysis for field studies. Chapman & Hall, New York, New York, USA.

———, ———, ———, T. L. MCDONALD, AND W. ERICKSON. 2002. Resource selection by animals: statistical design and analysis for field studies. Second edition. Kluwer Press, New York, New York, USA.

MATHSOFT. 1999. S-PLUS 2000 guide to statistics. Volume 1. MathSoft, Seattle, Washington, USA.

MCDONALD, T. L., AND L. L. MCDONALD. 2002. A new ecological risk assessment procedure using resource selection models and geographic information systems. Wildlife Society Bulletin 30:1015–1021.

NIELSON, R., B. F. J. MANLY, AND L. L. MCDONALD. 2004. A preliminary study of the bias and variance when estimating a resource selection function with separate samples of used and available resource units. Pages 28–34 *in* S. Huzurbazar, editor. Resource selection methods and applications. Proceedings of the 1st International Conference on Resource Selection, Laramie, Wyoming, January 13–15, 2003. Western EcoSystems Technology, Inc., Cheyenne, Wyoming, USA.

NIEMUTH, N. D. 2003. Identifying landscapes for greater prairie chicken translocation using habitat models and GIS: a case study. Wildlife Society Bulletin 31:145–155.

PARKER, J. M., AND S. H. ANDERSON. 2003. Habitat use and movements of repatriated Wyoming toads. Journal of Wildlife Management 67:439–446.

PRENTICE, R. L., AND R. PYKE. 1979. Logistic disease incidence models and case–control studies. Biometrika 66:403–411.

RAMSEY, F. L., M. MCCRACKEN, J. A. CRAWFORD, M. S. DRUT, AND W. J. RIPPLE. 1994. Habitat association studies of the northern spotted owl, sage grouse, and flammulated owl. Pages 189–209 *in* N. Lange, L. Ryan, L. Billard, D. Brillinger, L. Conquest, and J. Greenhouse, editors. Case studies in biometry. John Wiley & Sons, New York, New York, USA.

ROWLING, J. K. 2000. Harry Potter and the goblet of fire. Scholastic Press, New York, New York, USA.

SCHOUTEN, E. G., J. M. DEKKER, F. J. KOK, S. LE CESSIE, H. C. VAN HOUWELINGEN, J. POOL, AND J. P. VANDENBROUCKE. 1993. Risk ratio and rate ratio estimation in case–cohort designs: hypertension and cardiovascular mortality. Statistics in Medicine 12:1733–1745.

SEBER, G. A. F. 1984. Multivariate observations. John Wiley and Sons, New York, New York, USA.

SKOV, T., J. DEDDENS, M. R. PETERSEN, AND L. ENDAHL. 1998. Prevalence proportion ratios: estimation and hypothesis testing. International Journal of Epidemiology 27:91–95.

STEINBERG, D., AND N. S. CARDELL. 1992. Estimating logistic regression models when the dependent variable has no variance. Communications in Statistics – Theory and Methods 21:423–450.

SYSTAT SOFTWARE. 2000. SYSTAT. Version 10. Systat Software, Inc., Point Richmond, California, USA.

WALKER, P. A. 1990. Modelling wildlife distributions using a geographic information system: kangaroos in relation to climate. Journal of Biogeography 17:279–289.

ZHANG, J., AND K. F. YU. 1998. What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. Journal of the American Medical Association 280:1690–1691.

Received 20 June 2003.
Accepted 24 August 2004.
Associate Editor: Lubow.