Commentary

A Primer on Interpreting Regression Models

FRED S. GUTHERY,[1] *Department of Natural Resource Ecology and Management, Oklahoma State University, Stillwater, OK 74078, USA*
RALPH L. BINGHAM, *Caesar Kleberg Wildlife Research Institute, Texas A&M University-Kingsville, Kingsville, TX 78363, USA*

ABSTRACT We perceive a need for more complete interpretation of regression models published in the wildlife literature to minimize the appearance of poor models and to maximize the extraction of information from good models. Accordingly, we offer this primer on interpretation of parameters in single- and multi-variable regression models. Using examples from the wildlife literature, we illustrate how to interpret linear zero-intercept, simple linear, semi-log, log-log, and polynomial models based on intercepts, coefficients, and shapes of relationships. We show how intercepts and coefficients have biological and management interpretations. We examine multiple linear regression models and show how to use the signs (+, −) of coefficients to assess the merit and meaning of a derived model. We discuss 3 methods of viewing the output of 3-dimensional models (y, x_1, x_2) in 2-dimensional space (sheet of paper) and illustrate graphical model interpretation with a 4-dimensional logistic regression model. Statistical significance or Akaike best-ness does not prevent the appearance of implausible regression models. We recommend that members of the peer review process be sensitive to full interpretation of regression models to forestall bad models and maximize information retrieval from good models (JOURNAL OF WILDLIFE MANAGEMENT 71(3):684–692; 2007)

DOI: 10.2193/2006-285

KEY WORDS graphical interpretation, linear model, logistic regression, model interpretation, models, multiple linear regression, polynomials, regression.

We believe that too many implausible regression models are appearing in print because wildlife scientists do not fully interpret the meaning of intercepts, coefficients, and graphical relationships. The advent of Akaike's Information Criterion (AIC) model selection (Burnham and Anderson 2002) seems to have exacerbated the problem because models are appearing more frequently in the literature. We have seen an Akaike best model that essentially predicted constants and, therefore, was of little predictive value, another that predicted the presence of aquatic obligates in the absence of water, and another that predicted reduced total harvest with the harvest of more animals. These types of scientific faux pas would occur less frequently if the biological and management meaning of model parameters and graphical relationships were interpreted more completely.

Another important problem with incomplete model interpretation is that useful information in a model remains unextracted. Why do authors often not interpret the biological implications of models? We suppose the finding of a statistically significant or Akaike best model is sometimes viewed as the endpoint of analysis. However, the parameters and relationships in models provide additional information. If they are not interpreted, readily available knowledge is lost.

Our goal is to give a very basic introduction to the interpretation of parameters in regression models and to methods of graphically interpreting multivariable models. To provide relevance, we base the presentation on models from the wildlife literature. We start with single-independent-variable models including zero-intercept linear, simple linear, semi-log, log-log, and polynomials (the latter use transformations of a single variable). We then turn to multivariable models (multiple linear and logistic regression models) with >1 independent variable. Discussion of multivariable models entails discussion on dimensionality (usually 1 + the no. of independent variables in a model) and its implications for graphical interpretation of model output.

SINGLE VARIABLE MODELS

Linear

Simple linear regression models receive widespread use in natural resource science. They are useful when a dependent variable (y) is proportionally related to an independent variable (x).

The graph of a zero-intercept linear model passes through the origin ($y = 0$ when $x = 0$). These models are of the form

$$y = bx.$$

Statistical packages provide the option of estimating the coefficient (b) for no-intercept regression models. The coefficient is the change in y for a 1-unit increase in x. Zero-intercept models may be appropriate whenever y must = 0 when $x = 0$. For example: 1) In predicting absolute abundance based on an index of abundance, one assumes the index would be zero if true abundance was zero. 2) In estimating the relation between age estimated from dental annuli and true age, one assumes the number of annuli would be zero if the true age was zero.

An interesting application of the zero-intercept model appeared as an early mark–recapture estimator of population size (Hayne 1949). The dependent variable was the proportion marked in the population based on a sample, and the independent variable was the number previously marked and released into the population. Recall that the regression coefficient, b, gives the increase (if $b > 0$) in y for a unit increase in x. Therefore, in the mark–recapture application, b gives the increase in the proportion marked by the mark and release of one animal. In other words, it estimates the proportion that one animal represents in the population (N) so $1/N = b$ and therefore $N = 1/b$ (an estimate of population size).

Simple linear models with nonzero y-intercepts are of the

[1] *E-mail: fred.guthery@okstate.edu*

form

$$y = a + bx,$$

where

y = the predicted value of the dependent variable,
a = the y-intercept (value of y at $x = 0$), and
b = the slope or rate of change in y as x increases.

As above, the graph is a line with slope b interpreted as the change in y for a unit increase in x. The sign (+, −) of the slope indicates whether y increases (+) or decreases (−) with x. The sign of the coefficient of an independent variable is an important interpretive device, whether it is for a simple linear regression model (one independent variable) or for a multiple linear regression model (>1 independent variable; see later).

The following example predicts total hunter-days (y) as a function of an abundance index (x = no./km) of northern bobwhites (*Colinus virginianus*) in the Rolling Plains of Texas (Guthery et al. 2004):

$$y = 44,212 + 2,538x.$$

The intercept (44,212) indicates the number of hunter-days if $x = 0$ (no quail). In other words, the equation predicts that 44,212 hunter-days will accumulate if there are no quail in the population. This illogical outcome illustrates the danger of extrapolating predictions beyond the range of x variables in a data set (the lowest observed x value was about 5/km). The slope (2,538) indicates that each unit increase in the population index adds 2,538 hunter-days to the total hunting effort. The positive slope is reasonable because we would expect hunting pressure to increase with the abundance of quarry.

The above model also contains hidden information on the relation between bobwhite abundance and harvest pressure. Indeed, the model indicates harvest pressure (hunter-d/index quail; y/x) increases at an accelerating rate as the population declines. This is evident by dividing both sides of the equation by x to obtain

$$y/x = 44,212/x + 2,538.$$

You can see that as abundance (x) becomes smaller, harvest pressure (y/x) becomes larger. Here is an empirical counter example to the dogma that "harvest of quail is self limiting." This example fits Silver's (1998:95) observation that "sometimes when you have a set of equations and you sit down with a pencil and paper you find that they contain more than you thought they did."

Patterson and Messier (2000) estimated the following relationship between coyote (*Canis latrans*) predation of white-tailed deer (*Odocoileus virginianus*) and the abundance of snowshoe hares (*Lepus americana*):

$$y = 3.0 - 0.06x$$

where

y = deer killed per coyote per 100 days and
x = abundance of snowshoe hares (no./km^2).

From this equation we can estimate that 1) a coyote would kill about 3 deer in 100 days in the absence of snowshoe hares (y-intercept), and 2) each additional snowshoe hare on 1 km^2 reduces the kill of one coyote by 0.06 deer per 100 days (interpretation of the negative slope). By setting $y = 0$ and solving for x to obtain the x-intercept, we can estimate that a density of 50 snowshoe hares/km^2 might eliminate coyote predation on deer. That is what the model predicts, although the prediction is subject to uncertainty. If hare density exceeds 50/km^2, the kill per coyote rate (y) becomes negative, which implies that coyotes spontaneously vomit live deer. The model also indicates, in a management sense, that we could reduce coyote predation on deer by increasing the abundance of snowshoe hares. Similarly, the model indicates we might be able to increase predation on deer by managing against hares; this might be relevant to deer damage control. Finally, in an ecological sense, the model shows the effects of buffer prey (hares) on the predation rate experienced by a possible target of management (white-tailed deer). Of course, whether hares or deer are a target of management is a human value.

Semi-Log

Semi-log models also have wide application in natural resource science. These models are appropriate when 1) the dependent variable increases at an increasing rate with the independent variable (e.g., population growth in an unlimited environment) or 2) the dependent variable decreases at a decreasing rate with the independent variable (e.g., population decline in the absence of density-dependent effects and production). The typical procedure is to express the natural logarithm of y as a simple linear function of x in the form

$$\ln(y) = a + bx.$$

This equation predicts the natural logarithm (ln) of y. To obtain y we have to detransform the logarithm, which we do by exponentiating. The operator "exp" means to raise the base of a natural logarithm (e = 2.7182) to a power. For example, $\exp(2) = e^2 = 2.7182^2 = 7.38$. In exponentiating a semi-log model we have:

$$\exp[\ln(y)] = \exp(a + bx)$$

or

$$y = \exp(a + bx)$$

because $\exp[\ln(y)] = y$. We can rewrite the equation as

$$y = k[\exp(bx)]$$

where $k = \exp(a)$ since

$$\exp(a + bx) = [\exp(a)][\exp(bx)].$$

Now $\exp(a)$ is just a number, or constant, k. For example, $\exp(-0.5) = 0.6$. So, substituting,

$$y = k[\exp(bx)].$$

Here is an example (Merrill et al. 2005; Fig. 1). Their

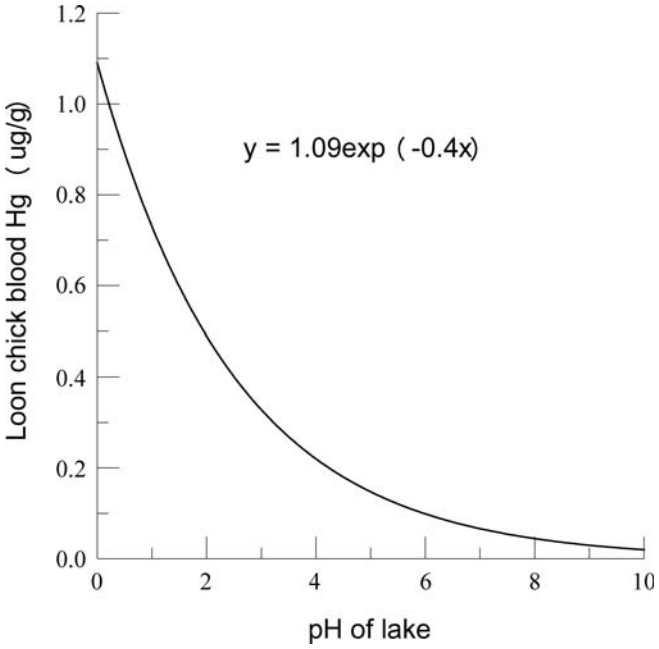

Figure 1. Mercury concentration in the blood of common loon chicks as a function of lake pH in Wisconsin, USA (Merrill et al. 2005).

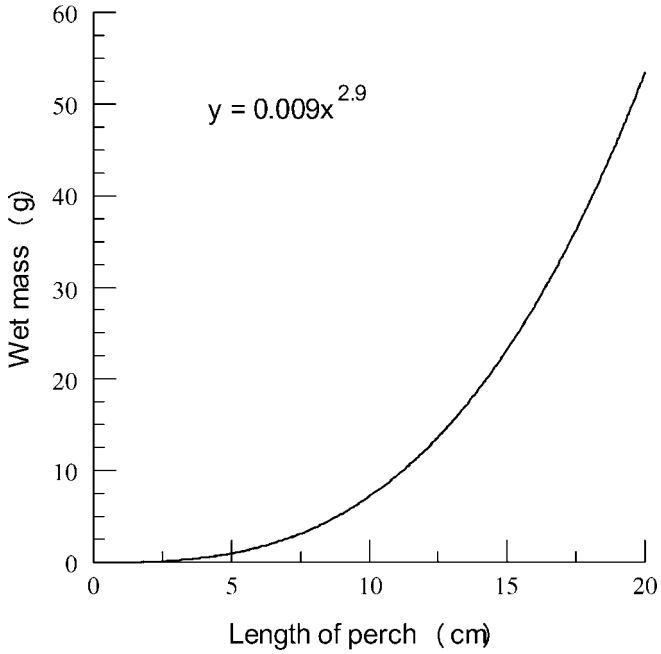

Figure 2. Wet mass of perch as a function of their length (Merrill et al. 2005).

semi-log equation was exponentiated to obtain

$$y = 1.09[\exp(-0.4x)]$$

where

y = blood mercury levels in common loon (*Gavia immer*) chicks (μg/g) and
x = the pH of lakes in Wisconsin (observed range = 5–10).

The y-intercept is k; that is, if $x = 0$, $y = 1.09$ is the estimated mercury level at a pH of 0 since $\exp(0) = 1$.

We can express the Merrill et al. (2005) equation in different form to add an interpretive angle:

$$y = 1.09(0.67^x)$$

because

$$\exp(-0.4x) = [\exp(-0.4)]^x = 0.67^x.$$

Here we discover that each 1-unit increase in pH is associated with a 0.67, or 67%, retention of contamination levels, and this value holds through all pH levels measured, at least under the model. This is a property of the exponential decay model ($y = k[\exp(bx)]$ with $b < 0$).

Log-Log

Another widely encountered model is the log-log model, with the natural logarithm of y expressed as a simple linear function of the natural logarithm of x in the form

$$\ln(y) = a + b\ln(x).$$

These models apply if the dependent variable may be expressed as some power of the independent variable (e.g., $y = x^b$; we provide examples below). Because the equation predicts the natural logarithm of y, to obtain y we again have to detransform the logarithm by exponentiating each side of the equation to obtain

$$\exp[ln(y)] = \exp\{a + b[\ln(x)]\}$$

or

$$y = [\exp(a)]\{\exp[b\ln(x)]\}$$

or

$$y = [\exp(a)]\{\exp[\ln(x)b]\}$$

or

$$y = kx^b.$$

where $k = \exp(a)$ and $x = \exp[\ln(x)]$.

Merrill et al. (2005) reported that the logarithm of wet mass (y, g) of yellow perch (*Perca flavescens*) could be modeled as a function of the logarithm of their length (x, cm; Fig. 2) according to

$$\ln(y) = -4.7 + 2.9[\ln(x)].$$

Upon solving for y (exponentiating each side of equation), we obtain

$$y = 0.009x^{2.9}.$$

Because $0^{2.9} = 0$, the y-intercept for this equation is 0.0. The value of the exponent of x (2.9) makes sense when you think about it. Merrill et al. (2005) used a linear measurement (length) to predict the mass of a volume of perch (a cubic measure). In a world not variable, the value of the exponent might have been 3.0 instead of 2.9. We can say the wet mass of a perch is approximately proportional to the cube of its length.

Whenever the value of the x exponent is between 0 and 1, we have a model wherein the dependent variable increases

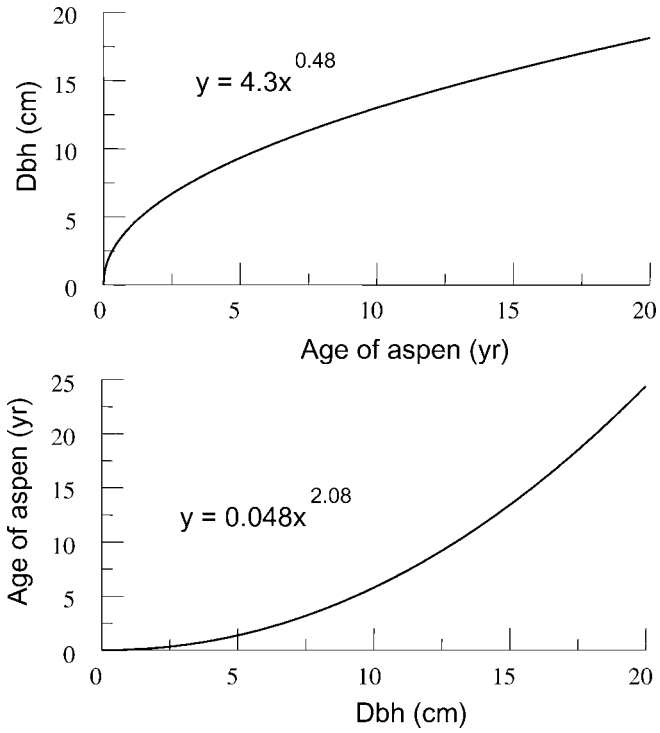

Figure 3. Top: diameter breast height of aspen as a function of tree age (Rolstad et al. 2000). Bottom: age of aspen as a function of diameter at breast height (determined by solving for x from the eq in the top graph).

with some simple positive fractional power of x. For example, $x^{0.5}$ is the square root of x. Perhaps the most widely known example of this function is the species–area curve from island biogeography (Shafer 1990),

$$S = kA^z,$$

where

S = the number of species on an island,
k = a constant
A = the size of an island (e.g., ha), and
z = some simple positive fractional power <1 of A.

Rolstad et al. (2000) used an exponentiated log-log model to predict the stem diameter at breast height (y = dbh [cm]) of quaking aspen (*Populus tremuloides*) based on the age (x = yr) of a tree (Fig. 3):

$$y = 4.3x^{0.48}.$$

This model indicates diameter at breast height = 0 cm when age = 0. Then, diameter at breast height increases approximately in proportion to the square root of age because the exponent is approximately 0.5.

We can use the Rolstad et al. (2000) model to predict age of aspen based on diameter at breast height. This involves solving the original equation for x (age). The resulting formula is

$$x = 0.048y^{2.08}.$$

Thus, whereas diameter at breast height increased approximately in proportion to the square root of age, age increases approximately in proportion to the square of diameter at breast height (Fig. 3).

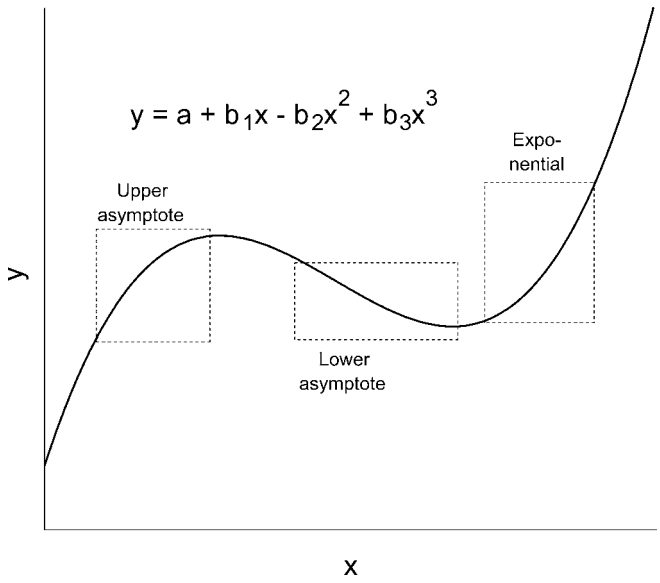

Figure 4. Various curve types along a cubic polynomial.

Polynomials

Polynomials are functions of the form

$$y = a + b_1x + b_2x^2 + \cdots + b_kx^k$$

where $b_k \neq 0$. A linear function $y = a + bx$ with $b \neq 0$ is a polynomial of degree 1 (x raised to the first power), a quadratic function $y = a + b_1x + b_2x^2$ with $b_i \neq 0$ is a polynomial of degree 2 (x raised to the second power), and so on. Polynomials are useful in that we can approximate many types of functions, including logarithmic and exponential functions, to any desired degree of accuracy with polynomials. Indeed, a polynomial of high enough order will fit almost any data set. For example, a quadratic polynomial can provide a perfect fit to exactly 3 data points with different x values, a cubic polynomial to 4 data points, a quartic polynomial to 5 data points, and so on. This is a dangerous property of polynomials because a good fit may lead to absurd predictions, especially with small sample sizes and higher-ordered polynomials. In natural resource science polynomials higher than degree 3 are uncommon.

Polynomials have 2 main benefits in modeling. First, they provide a means of invoking curvature in the relation between a dependent and an independent variable. A second benefit in simple curve-fitting is that any polynomial of order 2 or higher has a variety of shapes (curves) contained in the response. For example, sections of a cubic polynomial will appear to have upper asymptotes and lower asymptotes, to show exponential declines and increases, and depending on scaling, to appear linear (Fig. 4).

Stewart et al. (2000) used a quadratic polynomial to describe the relationship between white-tailed deer use of different plots as a function of an index of carrying capacity (amt of food) on plots (Fig. 5). Their model was

$$y = 0.01 + 0.03x - 0.003x^2$$

where

y = average number of deer per scan sample on a plot (deer use) and
x = an index to carrying capacity on a plot.

Interpretation of the relationship indicates practically no deer use at a carrying capacity index of zero, maximum use at an index of 5, and no use at indexes >10 (Fig. 5). We may estimate the carrying capacity associated with maximum use by examining the graph for the maximum value of the quadratic function or calculate it exactly using calculus or algebra.

The model would appear to have some problems. After the carrying capacity index exceeded about 5, why would deer use decline? They would have more food. It turns out that plots with the highest carrying capacity had too much woody cover (i.e., food), which impeded deer use (Stewart et al. 2000).

Look at the low values for average number of deer per plot (Fig. 5). Look at the nature of the quadratic relationship. Do you see that 1) in the range of carrying capacities of 2–8, carrying capacity made little difference in deer use and 2) in the range of 4–6 deer use was almost independent of carrying capacity (flat part of curve)? These are some biological and management interpretations we can draw from the Stewart et al. (2000) curve. The Stewart et al. (2000) example highlights the value of examining graphs of estimated relationships.

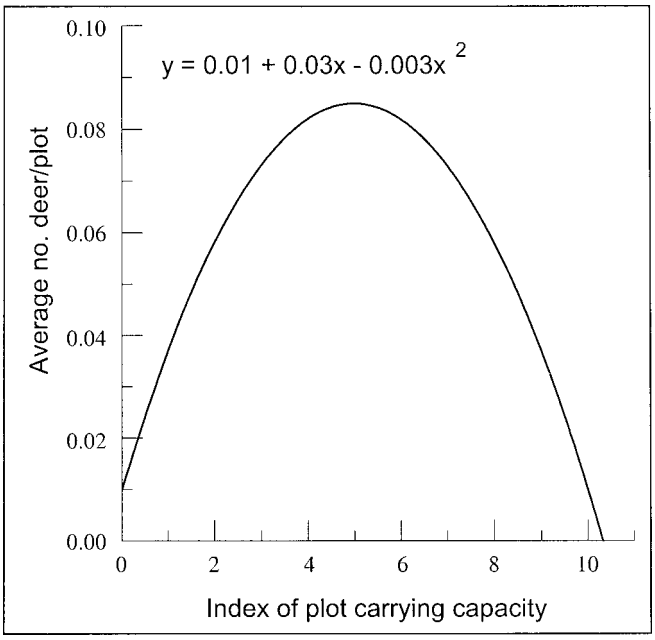

Figure 5. White-tailed deer use of plots as a function of plot carrying capacity (Stewart et al. 2000).

MULTI-VARIABLE MODELS

Multiple Linear Regression

Multiple linear regression involves ≥2 independent (x) variables in a model of the form

$$y = a + b_1x_1 + b_2x_2 + \cdots + b_kx_k.$$

where

y = the predicted value of the dependent variable,
a = the y-intercept or the value of y if all independent variables $x_i = 0$, and
b_i = the rate of change in y for a unit increase in the variable x_i if you hold all other x values constant.

If a model has the property that no independent variable is a function of the other independent variables (e.g., no interaction terms or powers of x_i), then the signs (+, −) of the coefficients (b_i) show the direction of the change in y as x_i increases, holding other x_i at fixed values. Simply by looking at the coefficients, you can do some analytical interpretation of a multiple regression model. Given the conditions specified above, does it make sense that y declines (negative coeff.) or increases (positive coeff.) with some x variable? If it does not make sense, there is something wrong with the model, the data, or your understanding of nature.

Here is an example of a multiple linear regression model from Ellis et al. (1972). Their goal was to develop equations to predict prehunt density (y; birds/40 ha) of bobwhites on 2 state parks in Illinois. Independent variables were March (prebreeding) density (x_1), the average number of bobwhite calls for 2-minute listening stops (x_2), and the average number of calling males per stop (x_3). Data were collected for 8 consecutive years on each of the 2 study areas. The resulting models were

$$y_1 = 15.9 - 0.4x_1 + 3.0x_2 - 10.3x_3 \qquad \text{for study area 1,}$$

and

$$y_2 = 4.9 - 0.8x_1 + 0.6x_2 + 0.3x_3 \qquad \text{for study area 2.}$$

How do we interpret these models?

First, Area 1 would have about 16 bobwhites per 40 ha and Area 2 about 5 per 40 ha in the prehunt population if no birds were present in the breeding population and no calling males were heard. We base this conclusion on the intercept (first term on the right side of the equal sign). So the intercepts do not make sense and possibly represent extrapolation beyond the range of the data, which is often the case where at least some of the x_i values are >0 for all data points.

Second, if we hold other independent variables constant, prehunt density declined at a rate of 0.4 birds per bird in the breeding population (x_1) on Area 1 and at 0.8 birds per bird on Area 2. This is based on the negative coefficients for x_1. This result makes no sense; in other words, we would expect prehunt density to increase with breeding density.

Third, if we hold other independent variables constant, prehunt density increased on both areas as total calls (x_2) increased based on positive regression coefficients. This result makes sense. More calls imply more males imply higher prehunt densities.

Fourth, if we hold other independent variables constant,

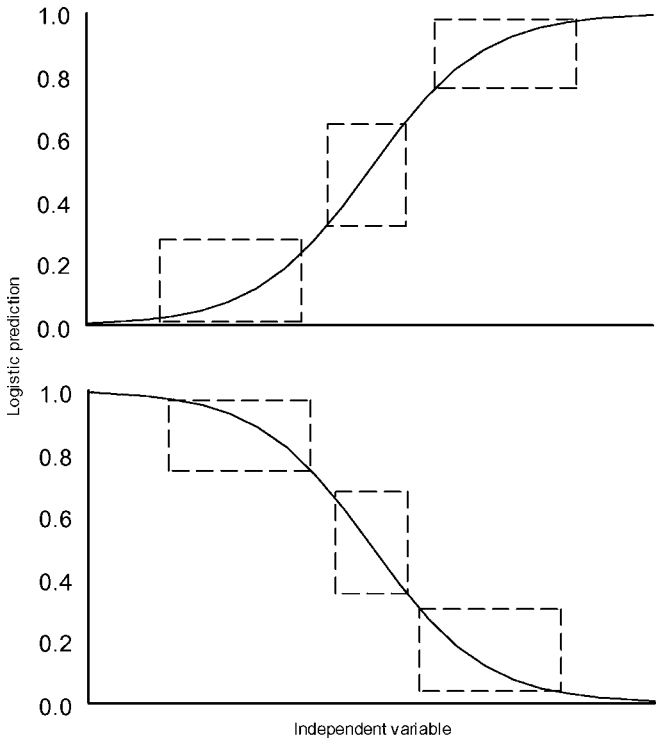

Figure 6. Forward (top) and reverse (bottom) logistic curves. The dashed boxes contain various manifestations of the logistic curve that might appear within the range of data values used in logistic regression modeling.

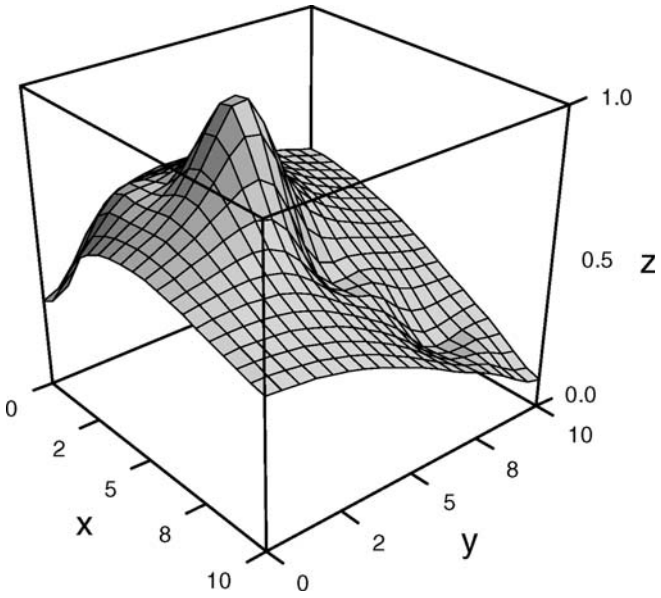

Figure 7. Three-dimensional graphic in 2 dimensions and an example of a response surface.

prehunt density declined with calling males per stop (x_3) on Area 1 (negative coeff.) and increased with this variable on Area 2 (positive coeff.). This result is self-contradictory. Why would prehunt density be negatively related with males per stop on one area and positively related on the other?

Upon interpretation, the models given above appear to be questionable even though they were statistically significant. The contradictory and counterintuitive results may have resulted from the independent variables being highly correlated resulting in the problem of multicollinearity, a condition where $\geq$1 independent variables are nearly a linear combination of the others. This condition may result in nonsensical coefficients in multiple regression analysis. With data pooled over both areas ($n = 16$), we find the correlation between x_1 and x_2 is 0.72, between x_1 and x_3 is 0.66, and between x_2 and x_3 is 0.93. The relatively high correlations among independent variables, especially x_2 and x_3, indicate that only one of them should be used for developing a prediction equation because in a modeling context these 2 variables are essentially the same.

Logistic Regression

Natural resource scientists widely use logistic regression to model dependent variables that assume values between 0 and 1. Such variables might include survival or mortality rates, or the probability of a binary variable such as (absent, present), (survived, died), or (failed, succeeded). A logistic regression model is an approximate homologue of logistic population growth, except that the logistic regression model has a carrying capacity (K in population modeling) of 1. Simple (1 independent variable) logistic regression models have the typical S-shaped curve as the graph of the proportion or probability being modeled, the shape of which (forward, reverse) depends on the sign (+, −) of the coefficient of the independent variable (Fig. 6).

Before providing an example interpretation of a multivariable logistic regression model, we offer this prologue on the limitations of graphical analysis of multi-variable models. With respect to models, an equation with single dependent and independent variables has 2 dimensions. A model with a single dependent variable and 2 independent variables has 3 dimensions, assuming that neither independent variable is a function of the other. If the number of independent variables exceeds 2, the dimension of the model exceeds 3; a situation with $\geq$4 dimensions is sometimes called hyperspace.

To aid in interpretation of models, we can graphically visualize relationships in 2 or 3 dimensions on 2-dimensional paper in 3 ways. One is using commercially available software for plotting 3-dimensional graphs with 3-dimensional perspective in 2 dimensions (Fig. 7).

A second way is to view a curvilinear response surface (prediction surface for an equation; Fig. 7 is an example) as a homologue to landscape topography and then plot contours as is done on topographic maps. The technique first involves setting a contour value (arbitrary values of the dependent variable). Then one selects an arbitrary set of values for one independent variable and solves for the corresponding value of the other independent variable. The pairs of values for the 2 independent variables along a contour are then plotted. The contour technique shows peaks, valleys, and flatlands in the response surface just as a topographic map shows these landscape features. We have not seen this technique used in natural resource science but it is available.

The third alternative is simpler. It involves plotting the

dependent variable as a function of one independent variable at different fixed values of a second independent variable. We will see application of this technique in the logistic regression example that follows.

For ≥4 dimensions, we can graphically visualize only slices of hyperspace. By slices of hyperspace, we mean portions of the response surface that are infinitesimally thin because we held additional independent variables constant. For example, given a dependent variable and 3 independent variables, we could hold one of the independent variables constant, which yields a specific number. This number is added to the *y*-intercept and the resulting model (1 dependent, 2 independent variables) can be plotted as Figure 7. As to what values to use for constants, one would suppose modes or means would capture the general nature of the response to variables not held constant. However, our ability to reliably interpret multi-variable models wanes rapidly as the number of independent variables increases and, eventually, the models are beyond human comprehension. "It is impossible to imagine a four-dimensional space," observed Stephen W. Hawking (1988:24), the great theoretical physicist. "I personally find it hard enough to visualize three-dimensional space!"

The data for this example came from Curtis and Jensen (2004). They used logistic regression modeling to predict the presence (1) or absence (0) of beavers (*Castor canadensis*) at culverts and bridges in New York. The independent variables appearing in one logistic model were

x_1 = stream gradient (%),
x_2 = area without woody plant canopy (%),
x_3 = stream width (m), and
$x_4 = x_1x_2$ = the interaction of stream gradient and open area (%–%).

With these variables the model has 4 instead of 5 dimensions because of the interaction term ($x_4 = x_1x_2$ does not add a dimension).

Computer-assisted logistic regression analysis yields the log-transformed (logit) version of the logistic model:

$$\begin{aligned}\text{logit } P(\boldsymbol{X}) &= \ln[P(\boldsymbol{X})/(1 - P(\boldsymbol{X}))] \\ &= a + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4\end{aligned}$$

where by exponentiating

$$P(\boldsymbol{X}) = 1/(1 + \exp\{-[\text{logit}P(\boldsymbol{X})]\}).$$

This looks intimidating but interpretation is fairly simple. The term $P(\boldsymbol{X})$ simply represents the logistic function with a carrying capacity of 1, and the boldface $\boldsymbol{X}$ indicates the list of parameters in the model (a, b_1, b_2, b_3, b_4). Notice the negative (–) sign preceding the argument (expression in parentheses) of the exponentiation operator in the expression for $P(\boldsymbol{X})$. The presence of this negative sign changes the sign of the intercept and all coefficients in the expression for logit $P(\boldsymbol{X})$.

Curtis and Jensen (2004) reported the logistic regression model

$$\begin{aligned}\text{logit } P(\boldsymbol{X}) = 7.423 &- 1.627x_1 - 0.094x_2 - 0.330x_3 \\ &+ 0.016x_4 .\end{aligned}$$

The logistic function derived from this model is

$$\begin{aligned}P(\boldsymbol{X}) = 1/[1 + \exp(-7.423 &+ 1.627x_1 + 0.094x_2 \\ &+ 0.330x_3 - 0.016x_4)] .\end{aligned}$$

Notice that we changed the signs of the intercept and coefficients in the logit model when we converted it to the logistic prediction model. If you plug in values for the x variables in the $P(\boldsymbol{X})$ function, you obtain a number between 0 and 1. You can view this number as an estimate of the probability of beaver presence, given the values of the independent variables used to calculate it.

As in multiple linear regression, it might be informative to estimate the prediction if all $x_i = 0$ (substitute 0 for all x_i in the equation). For the Curtis and Jensen (2004) model, we obtain

$$\begin{aligned}P(\boldsymbol{X}) &= 1/\{1 + \exp[-7.423 + 1.627(0) + 0.094(0) \\ &\qquad +0.330(0) - 0.016(0)]\} \\ &= 1/[1 + \exp(-7.423)] = 0.99.\end{aligned}$$

We have predicted presence of beavers on a placid stream (no gradient) with much food (0% open) that has no water (stream width = 0 m). This may be our familiar problem of extrapolating beyond the range of the data used to construct a model because a data point where all independent variables were zero probably did not occur. If you use model selection to identify best logistic regression models, and if the routine selects the intercept-only model (which happens), you are predicting a constant. This implies the prediction is independent of other variables tested as predictors.

Also as in multiple linear regression, the sign (–, +) of logit coefficients are directly interpretable if the model contains no independent variable that is a function of other independent variables, for example, interaction terms. If other variables are held constant, a negative logit coefficient indicates that $P(\boldsymbol{X})$ declines curvilinearly as x_i increases, and a positive logit coefficient indicates that $P(\boldsymbol{X})$ increases curvilinearly as x_i increases. You can use these facts to judge whether a logistic regression model makes sense. One has to be careful, however, in interpreting the sign of a logit coefficient because we are in a nonlinear realm that has flat parts, and the logistic equation itself provides a potpourri of curves (Fig. 6).

The Curtis and Jensen (2004) model is more difficult to interpret because of the interaction term ($x_4 = x_1x_2$). One way to interpret the implication of a variable contained in an interaction is to determine the sign of the effective coefficient at the mean value of the other variable in the interaction. Holding percent open area constant at its mean (approx. 33%), we obtain

$$b_1 + 33(b_4) = -1.627 + 33(0.016) = -1.099$$

as the coefficient for stream gradient. The negative value implies that the probability of beaver presence declines as

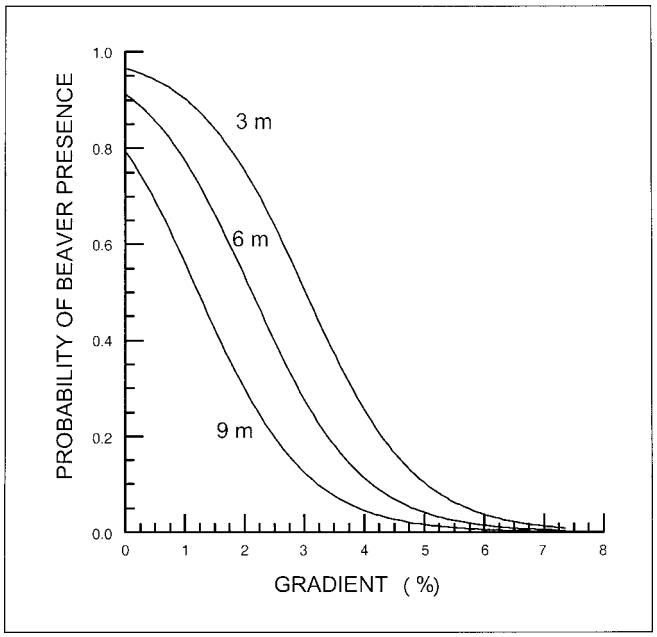

Figure 8. Logistic predictions of the probability of beaver presence as a function of stream gradient and width at bridges and culverts in New York, USA (Curtis and Jensen 2004). A third predictive variable, percent open area, was held constant at its mean.

gradient increases if percent open area is constant at its mean. To interpret the implication of percent open area if gradient is constant at its mean (approx. 1.3%), we have

$$b_2 + 1.3(b_4) = -0.094 + 1.3(0.016) = -0.073$$

as the coefficient for percent open area. If gradient is held constant at its mean, the probability of beaver presence declines as percent open area increases. Under average conditions, we would expect beavers to disappear as gradient increases and food becomes less plentiful. The negative value for stream width is difficult to interpret as is the positive value for the interaction coefficient. As a point of interest, the model predicts a low probability (0.003) of presence of beavers in waterfalls with no food (100% gradient, 100% open).

A final diagnostic in interpreting logistic regression models is the magnitude of a coefficient. Coefficients near zero (e.g., −0.003, 0.005) might imply the x variable associated with the coefficient has little or no predictive value. This is because such coefficients might, for all intents and purposes, predict a number that is nearly constant [exp(−0.003) = 0.997, exp(0) = 1, exp(0.005) = 1.005]. Indeed, logistic regression coefficients near zero might indicate an essentially null response, despite whether they are significant or a member of the Akaike best family of parameters. On the other hand, coefficients near zero can be quite meaningful, depending on the units of measure for both the independent and dependent variable.

Let us graphically examine the predictions associated with stream gradient at different stream widths (3 m, 6 m, 9 m) with percent open area held constant at its mean (approx. 33%; Fig. 8). The relationship would change for different stream widths or if open area was held constant at a different value (the problem of hyperspace). For the interaction term, we used $33x_1$ ($\bar{x}$ % open area × gradient). So in interpreting this figure, we should qualify: "under average conditions for percent open area . . .." Under this proviso, Fig. 8 indicates that the probability of beaver presence declined with gradient at all values of stream width examined. This result is the same as obtained by interpreting the logit coefficient, but the graphic provides information on the shape of the relationship. Likewise, at any fixed gradient, except in the right tails of the curves, the probability of beaver presence declined as stream width increased. This can be interpreted to indicate that beavers were less sensitive to gradient in narrower streams (or more sensitive to gradient in broader streams). To reiterate, Fig. 8 shows a method of examining a 3-dimensional problem in 2 dimensions.

CONCLUDING REMARKS

Before reviewing the tools of model interpretation and concluding, we wish to observe that all published models receive some interpretation. At a minimum, models are interpreted with respect to statistical significance or Akaike best-ness; often, but not always, scientists judge the merit of a model relative to the data used to construct it on the basis of an r^2 statistic or percentage of correct predictions. Many authors already practice the kind of full interpretation we advocate, which involves assessing the meaning and merit of parameters in models and graphical relationships implied by models. Absent such full interpretation, however, implausible models appear in print and the information contained in plausible models is not fully retrieved.

We gave examples of assessing model plausibility and retrieving information from plausible models. We first interpreted the y-intercept to see whether it made sense. In some cases it did not (e.g., heavy hunter effort in the absence of quail). We attributed nonsensical y-intercepts to extrapolation beyond the range of the data. The x-intercept might also be of interpretive interest, as it was in the case of coyote predation on white-tailed deer. We then interpreted the sign (+, −) of regression coefficients and asked whether the sign made sense. Is it reasonable to suppose that the dependent variable increases with an independent variable (+ sign)? Is it reasonable to suppose that it decreases (− sign)? We gave an example where some of the signs of coefficients in a multiple regression model made no sense. Finally, in the case of a multi-variable logistic regression model, we interpreted logit coefficients to determine whether the model was reasonable and where it was unreasonable. We showed how to extract additional information from a polynomial and a logistic regression model with graphical analysis.

We conclude with this observation. Statistical significance or Akaike best-ness implies only that a regression model might be meritorious. These properties do not prevent the appearance of implausible models and many such models sully our literature. We recommend that members of the peer review process be sensitive to full interpretation of

regression models to forestall bad models and maximize information retrieval from good models.

MANAGEMENT IMPLICATIONS

Regression models are products of field research, which is one basis for wildlife management decisions. Model interpretation not only exposes implausible models but also reveals additional information that models contain. Thus, model interpretation enhances the scientific basis of wildlife management.

ACKNOWLEDGMENTS

We thank C. R. McKinley, R. Houchin, G. E. Mecozzi, M. S. Mecozzi, K. S. Reyna, R. D. Elmore, and M. L. Morrison for commenting on a draft of this manuscript. The Oklahoma Agricultural Experiment supported Guthery's contribution and approved this manuscript for publication. The Caesar Kleberg Wildlife Research Institute supported Bingham's contribution.

LITERATURE CITED

Burnham, K. P., and D. R. Anderson. 2002. Model selection and multimodel inference. Springer, New York, New York, USA.

Curtis, P. D., and P. G. Jensen. 2004. Habitat features affecting beaver occupancy along roadsides in New York state. Journal of Wildlife Management 68:278–287.

Ellis, J. A., K. P. Thomas, and P. Moore. 1972. Bobwhite whistling activity and population density on two public hunting areas in Illinois. Proceedings of the National Bobwhite Quail Symposium 1:282–288.

Guthery, F. S., M. J. Peterson, J. L. Lusk, M. J. Rabe, S. J. DeMaso, M. Sams, R. D. Applegate, and T. V. Dailey. 2004. Multi-state analysis of fixed, liberal regulations in quail harvest management. Journal of Wildlife Management 68:1104–1113.

Hawking, S. W. 1988. A brief history of time. Bantam, New York, New York, USA.

Hayne, D. W. 1949. Two methods for estimating populations from trapping records. Journal of Mammalogy 30:399–411.

Merrill, E. H., J. J. Hartigan, and M. W. Meyer. 2005. Does prey biomass or mercury exposure affect loon chick survival in Wisconsin? Journal of Wildlife Management 69:57–67.

Patterson, B. R., and F. Messier. 2000. Factors influencing killing rates of white-tailed deer by coyotes in eastern Canada. Journal of Wildlife Management 64:721–732.

Rolstad, J., E. Rolstad, and Ø. Sæteven. Black woodpecker nest sites: characteristics, selection, and reproductive success. Journal of Wildlife Management 64:1053–1066.

Shafer, C. L. 1990. Nature preserves. Smithsonian Press, Washington, D.C., USA.

Silver, B. L. 1998. The ascent of science. Solomon Press, New York, New York, USA.

Stewart, K. M., T. E. Fulbright, and D. L. Drawe. 2000. White-tailed deer use of clearings relative to forage availability. Journal of Wildlife Management 64:733–741.

Associate Editor: Morrison.