Proc. 3rd North Am. Caribou Workshop.
Alaska Dep. Fish and Game. Juneau.
Wildl. Tech. Bull. No. 8: 89-118.

MULTIPLE REGRESSION METHODS FOR MODELING CARIBOU POPULATIONS

R. Terry Bowyer, Institute of Arctic Biology, and Department of Biology and Wildlife, University of Alaska Fairbanks, Fairbanks, AK 99775

Steven C. Amstrup, Department of Biology and Wildlife, University of Alaska Fairbanks, Fairbanks, AK 99775-1780, and Alaska Wildlife Research Center, U.S. Fish and Wildlife Service, 1011 E. Tudor Road, Anchorage, AK 99503

Julia G. Stahmann, Institute of Arctic Biology, and Department of Biology and Wildlife, University of Alaska Fairbanks, Fairbanks, AK 99775

Patricia E. Reynolds, Arctic National Wildlife Refuge, U.S. Fish and Wildlife Service, 101 12th Avenue, Fairbanks, AK 99701

Frank A. Burris, Institute of Arctic Biology, and Department of Biology and Wildlife, University of Alaska Fairbanks, Fairbanks, AK 99775

Abstract: Multiple linear regression (MLR) offers a useful method to model the dynamics of caribou (Rangifer tarandus) populations where changes in population parameters are brought about by several factors. Assumptions of using this statistical procedure are delineated and techniques necessary to meet these assumptions are discussed. Use of compounded independent variables is discouraged because they may violate assumptions of independence and artificially inflate values for coefficients of determination. Recommendations for determining an adequate sample size are provided. Multiple regression may not be useful when severe multicollinearity occurs among independent variables. All possible regressions, true stepping, or backward stepwise procedures are advocated for model building, depending upon the power of the available computing system. Examination of residuals is necessary to test the aptness of MLR models, and particular patterns may indicate the absence of important independent variables or that some variables do not meet the prescribed assumptions of this technique. Methods to correct such problems are suggested. Final presentation of the model should include the unstandardized equation with an adjusted multiple coefficient of determination (R^2_a) and a prediction error. Standardized regression coefficients should be presented as well as partial coefficients of determination. Interpretation of these parameters are dealt with relative to their biological significance. Finally, problems in testing the validity of the model are discussed.

Key Words: caribou, model, multicollinearity, multiple regressions, *Rangifer*, statistics

Recent advances in computer technology and statistical software have made complex biometric models readily available to wildlife biologists. Such models have been used to examine animal-habitat relationships, population dynamics, and distributional patterns (Capen 1981, Verner et al. 1986, Wehausen et al. 1987). Unfortunately, increasing ease of operating modern computing systems has not been matched by an equivalent understanding of assumptions inherent in application of these procedures. The purpose of this paper is to: (1) describe one kind of biometric model--multiple linear regression (MLR); (2) show how MLR can aid in understanding the dynamics of caribou (*Rangifer tarandus*) populations; and (3) summarize selected procedures, outcomes, and assumptions of MLR.

We thank X. P. Quang for his advice on statistical procedures. M. P. Gillingham and S. Fleischman provided many useful comments.

Why Multiple Linear Regression?

Many types of models other than MLR might be employed to investigate ungulate population dynamics (Starfield and Bleloch 1986). Further, complex statistical procedures should not be used when simple ones will suffice. For instance, McCullough (1979) explained most (85%) of the variation in reproductive rate of white-tailed deer (*Odocoileus virginianus*) with a single independent variable, population density. In that case, simple linear regression (SLR) was sufficient to explain the dynamics of the population.

Often, however, complex systems cannot be reduced to a single, important variable. MLR offers a reasonable approach for studying the response of populations to interactions of several environmental factors. Another advantage of MLR is the incorporation of laws of probability in model building rather than just relying on a subjective opinion about how well the model functions.

Regression analysis is a procedure by which an understanding of the statistical relationship between 2 or more variables can be gained. Once this relationship is understood, information on some of the variables can be used to predict or model the magnitude of another. Thus, MLR predicts change in the dependent variable (Y) as linear functions of independent variables (X's).

MLR Parameters

This paper outlines the regression process from experimental design to final model (Fig. 1). In this section we present the components of regression models and provide interpretations necessary to understand what follows.

For purposes of this discussion, 35 years of data for a hypothetical caribou population will be considered (Table 1). During those 35 years, the following 6 variables were recorded: calf:cow ratio; number of caribou; female harvest; number of wolves in the area; percentage cows in the population; and depth of snow on the ground in March. Long-lived animals like caribou might be influenced by weather patterns occurring over several years. Therefore, an additional compounded variable, 3-year running average of March snow depth, also was included. We would like to know what factors control recruitment of young into the population, thus the variable of greatest interest is the calf:cow ratio. Specifically, biologists want to know whether any of the other variables help explain variation observed in calf:cow ratio. The calf:cow ratio, then, is the "Y" or dependent variable, and the other 6 variables are "X's" or independent variables. Using the method of least squares (described in most statistics texts) and modeling procedures we discuss later, the following equation was constructed:

$$\text{Calves:Cow} = 0.772 - 0.00005\ (\text{no. of Caribou}) - 0.001\ (\text{Female Harvest}).$$

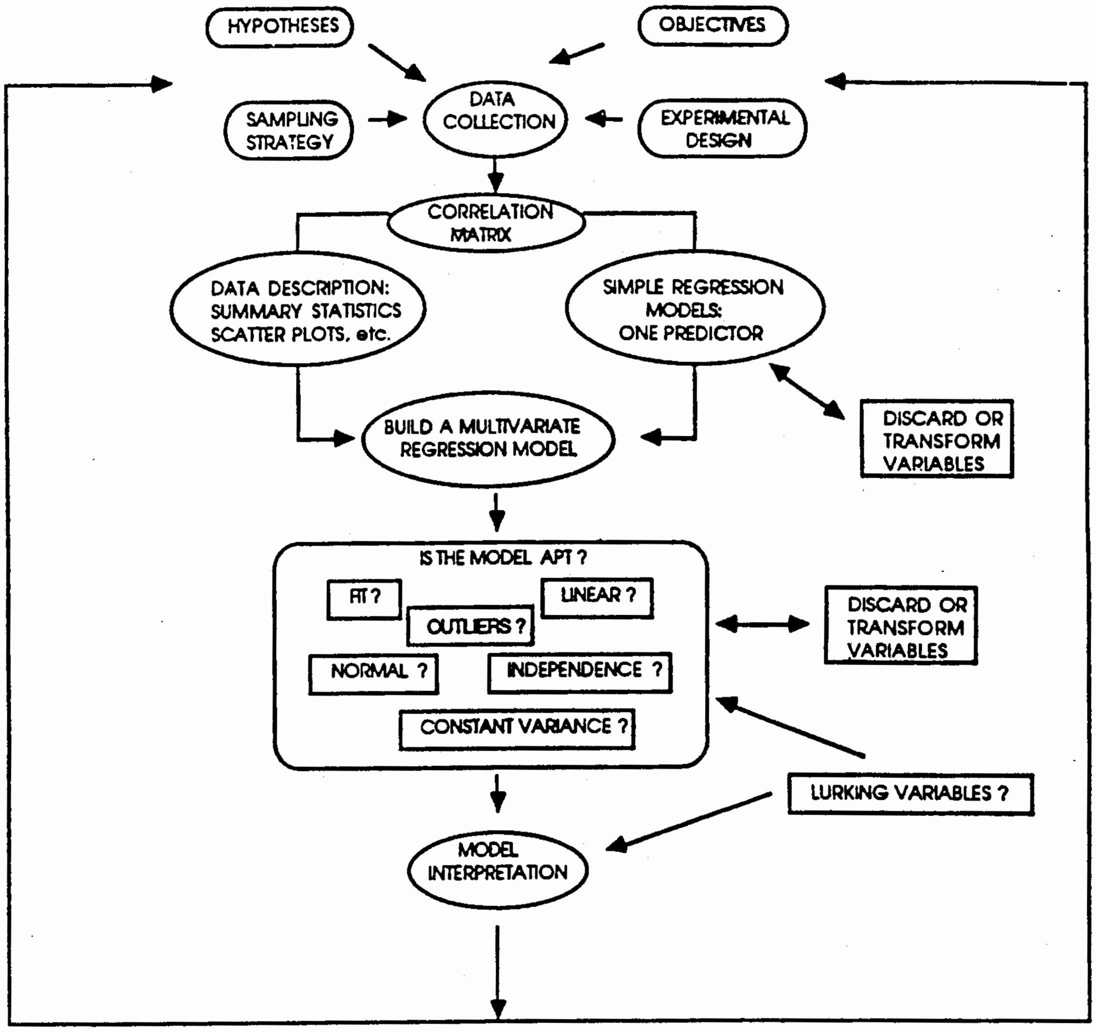


Fig. 1. A graphical summary of MLR model-building procedures described in this paper.

Table 1. Sample of raw data and resulting descriptive statistics for the hypothetical caribou population.

Year	No. of caribou	No. of females harvested	No. of wolves	No. of calves/ cow	% cows	March snow depth (in)	3-year average snow depth (in)
Raw Data							
1	5000	154	270	0.2829	52.0	5.00	
2	5000	168	270	0.3368	42.0	1.00	
3	4000	126	270	0.4477	34.0	18.0	8.00
4	3100	132	265	0.5069	27.0	13.0	10.667
5	2288	67	260	0.6034	30.0	7.00	12.667
-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-
35	5000	44	270	0.4429	36.0000	5.00	8.00
Descriptive Statistics							
N	35	35	35	35	35	35	33
Min.	2000	10	80	0.222	17	1	4
Max.	8083	168	270	0.669	66	22	13
Mean	4180.11	76.0	185.77	0.48	37.57	9.1	9.4
S. D.	1931.10	52.68	69.62	0.149	14.38	5.7	2.97
Coeff. var.(%)	46.2	69.3	37.5	31.0	38.3	62.6	31.6

This corresponds to the multiple regression equation: $Y = b_o + b_1X_1 + b_2X_2$. The b's in this equation are the parameters of the regression model, and commonly are called regression coefficients. The b_o term corresponds to the y-intercept, and b_1 indicates the degree of change that would be observed in Y if X_2 were held constant and X_1 varied. Likewise, b_2 indicates the change in Y expected if X_1 is held constant and X_2 varied. Such a model will estimate calf production rate for any year in which estimates of harvest and population size are available.

Associated with a regression model is the analysis of variance (ANOVA) table (Table 2). This table shows the partitioning of the deviations of the observed Y's around their mean value (SSTO for total sum of squares) into 2 components: (1) deviation of the fitted regression line about the mean value of Y (SSR for sum of squares explained by regression); and (2) deviation of the observed values of Y around the regression line (SSE for error sum of squares). The value of the regression model often is assessed by MSR/MSE, which is the ratio of the regression and error sums of squares divided by their respective degrees of freedom. If this ratio exceeds a tabled "$\underline{F}$"-value at the specified level of probability (e.g., $\alpha = 0.05$), the regression is significant. Put another way, number of caribou and harvest of females explain a significant portion of the variability in the calf:cow ratio.

Another measure of the utility of a regression model is the coefficient of determination ($\underline{r}^2$), which is the proportionate reduction in the total variation of Y that results from the information contained in the X variable (SSR/SSTO). The ratio SSR/SSTO ranges from $0 \leq \underline{r}^2 \leq 1$. The closer $\underline{r}^2$ is to 1, the greater the explanatory value of the regression model (e.g., more variation in Y is explained). $\underline{R}^2$ (the coefficient of multiple determination) represents the proportionate reduction in the variability of Y when 2 or more X-values are considered. This value is equivalent in concept to $\underline{r}^2$, and the 2 values are identical in simple linear regression (i.e., when there is a single independent variable).

Table 2. Analysis of variance table for multiple linear regression model predicting changes in Y (calves/cow) with values of X_1 (number of caribou) and X_2 (female harvest).[a]

Source of variation	Degrees of freedom (df)	Sum of squares (SS)	Mean square (SS/df)	$\underline{F}$-test	Probability of "F" as large or larger
Explained by	2	$\Sigma(\hat{Y}_i-Y)^2$	0.333	$\frac{MSR}{MSE}=121.6$	0.0001
Regression		0.665=SSR			
Error	32	$(Y_i-\hat{Y}_i)^2$	0.003		
Component		0.088=SSE			
TOTAL	34	$(Y_i-\hat{Y}_i)^2$ .753 - SSTO			
X_1	1	0.642	0.642	N/A	
X_2X_1	1	0.024	0.024	$\frac{MSR\ X_2X_1}{MSE}$ =8.672	0.01

---

[a] One degree of freedom is lost for each parameter entering the model (e.g. β_0, β_1, β_2). The contribution of X_2 given X_1 (X_2X_1) already in the model is assessed by the partial $\underline{F}$-ratio = (MSR X_2X_1)/MSE. A similar partial $\underline{F}$-test of (MSR X_1X_2)/MSE could be constructed in a second regression run in which X_2 is entered first. These partial $\underline{F}$-tests are equivalent to $\underline{F}$-to-enter tests performed by stepwise regression procedures. Thus, $\underline{F}$-to-enter of X_2 would be 8.672.

A significant F-test, or a high R^2, are often the only values considered by biologists using MLR, in part because most computer programs provide these statistics as basic output. Unfortunately, analyses often stop at that point without developing a further understanding of the assumptions and limitations of the model (Box 1966). The worth of a model cannot be assessed on the basis of these measures alone, and the utility of a model goes beyond those basic outcomes. In following sections, we discuss some of the broader ramifications of regression models.

Experimental Design

Selection of variables to be sampled and the manner in which data are collected will have a profound effect on the outcome and reliability of regression models. Thus, an adequate knowledge of the biology of the animal to be studied is the first (and most important) step in constructing a model. Although some required information may be available in the literature, it may be necessary to conduct a preliminary study to clarify some aspects of the animal's biology.

All too often data are collected without a particular method of analysis in mind. Such data may be difficult to analyze with MLR or, in some cases, with any statistical method. Remedial actions may be possible (Milliken and Johnson 1984), but it is better to collect data specifically for analysis by a particular method. Draper and Smith (1981:295) provide additional remarks about the use of such "unplanned" data. There is, however, no substitute for advanced planning.

Of course, the model selected will depend upon the hypotheses to be tested. A concise statement of the question(s) to be asked is usually necessary and always desirable for coherent and meaningful results (Green 1979).

Basic Assumptions of the Model

Five basic assumptions are necessary for the valid use of SLR or MLR: (1) for each value of X (independent variable) the associated values of Y (dependent variable) are distributed normally; (2) variances of Y-values are equal (homoscedastic) at all levels of X; (3) errors associated with Y-values sum to zero; (4) values of Y are independent of each other for all

X's; and (5) measurements of X-values are without error (Zar 1984:268). Fortunately, regression statistics are robust so long as violations of these basic assumptions are not too severe (Zar 1984:268, Neter et al. 1985:83).

Sample Size

Reliability of a model is reduced if sample sizes are too small for the number of variables included in the model (Noon 1986). As the number of independent variables (m) approaches the sample size (n), matrices used to calculate MLR statistics become singular (i.e., have a very large number of potential solutions). For this reason, it is recommended that sample sizes (n) be kept above a minimum of m + 10 (where none of the assumptions of MLR have been violated). For example, to predict a parameter for a caribou population such as "recruitment rate" with 4 independent variables (m), a minimum of 14 years (n) of observations is desirable. The confidence in predictions from MLR models increases with sample size. A sample size of at least n-m >50 is required for the valid use of MLR where the assumptions of this technique are violated substantially (Harris 1975:50). Few studies of caribou populations have lasted that many years! Nonetheless, we recommend about 10 samples (n) for each variable (m) in the final model (i.e., m X 10).

Preliminary Data Handling

Once data have been collected, the next step in building a regression model is a careful examination of the variables. Descriptive statistics ($\bar{x}$, SD, range) for each variable should be scrutinized and presented with the final model (Table 1). A complex biometric model without at least a summary description of data used to construct it may be difficult to interpret and certainly will be of limited value.

After examining summary statistics, SLR's should be performed to examine bivariate relationships between all combinations of variables. Such preliminary analyses can screen problems in data that might be difficult to tease from multivariate models. MLR assumes relationships between X's and Y are linear (although the resulting response surface for fitted values of Y in the final model need not be so). Variable pairs not exhibiting linearity require transformation. Nonlinear multiple regression techniques exist (Zar 1984:351) but are beyond the scope of this paper.

Independent variables (X's) also should be normally distributed and have equal variances to meet assumptions of MLR (Zar 1984:328). Unequal variances or "heteroscedasticity" may be detected by plots (Figs. 2 and 3) of residuals (differences between observed and fitted values). Heteroscedasticity also can be evaluated by comparisons of error mean squares in different ranges of the data set (Neter et al. 1985:123). Weighting of variables and or transformations may correct this problem. Histograms or plots of residuals against their expected values or normal scores may be used to detect deviation from linearity and normality (Fig. 4). A high coefficient of correlation ($\underline{r}$) between residuals and normal scores suggests normality (Neter et al. 1985:120). Several other formal procedures are available for assessing normality of variables (Conover 1980:359, Neter et al. 1985). The graphic method of Shapiro and Wilk (1965) is effective and easily understood; also see Sokal and Rohlf (1969:119-126). A condition of multivariate normality (a requirement for multiple correlation) may be inferred from a series of bivariate comparisons. Multivariate normality is difficult to determine, but if pairs of variables are not normally distributed, multivariate normality is unlikely to occur (Dunn 1981, Johnson 1981). This is especially important if it is uncertain which variable is dependent upon the other (e.g., do wolves control caribou or vice versa?). In this case, the procedure becomes multiple correlation by definition (i.e., the dependent variable is uncertain), and multivariate normality is required.

Several transformations are available to normalize data (Zar 1984:236-242; Neter et al. 1985:134-141). It is important to test for normality again after the transformation has been completed. Some transformations used to alleviate problems with heteroscedasticity and additivity also may normalize distributions (Zar 1984:236-242, Neter et al. 1985:120).

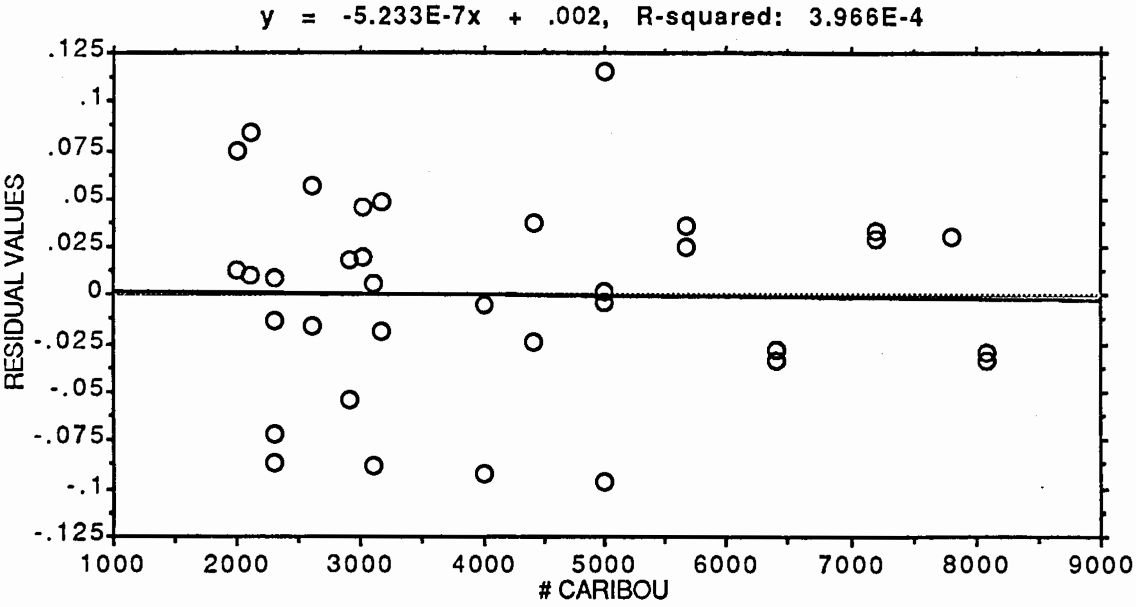


Fig. 2. Plot of residuals against X_1 showing that variances (error terms) tend to be greater when caribou numbers are smaller. Unequal variances (heteroscedasticity) must be corrected before proceeding with model building.

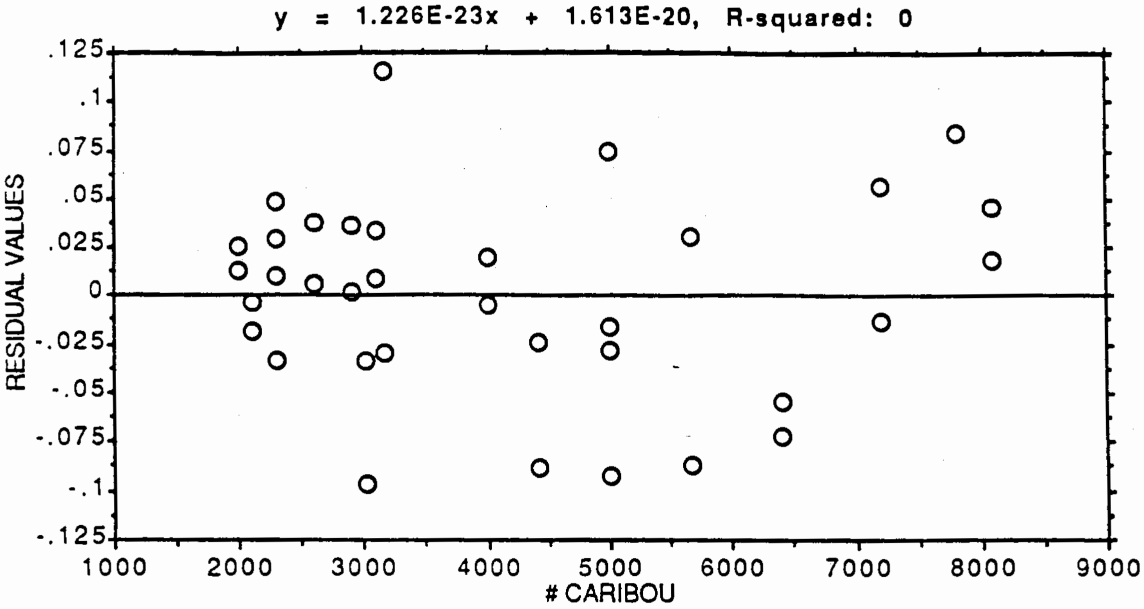


Fig. 3. Plot of residuals (Yi - $\bar{Y}$) against X_1 showing no trend in variances or error terms with population size.

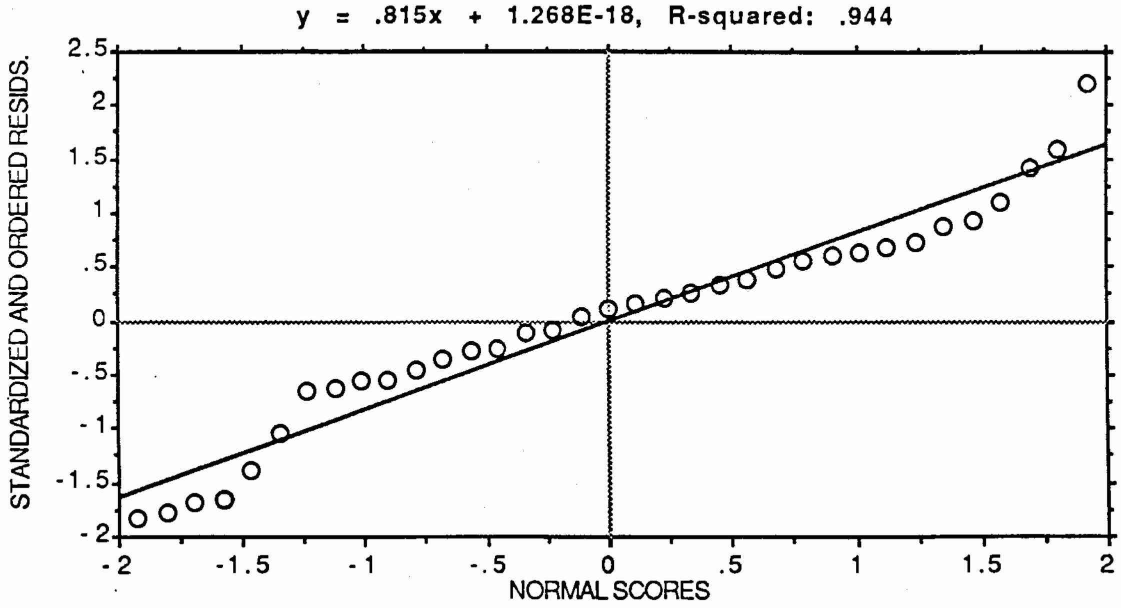


Fig. 4. Plot of standardized, ordered, residuals against normal scores. Points lying along a straight line suggest normality (this relationship may indicate a slight departure from normality typical of a logistic fit).

Preliminary examination and modification of data can be time consuming and frustrating, but is essential if a valid model is to be constructed.

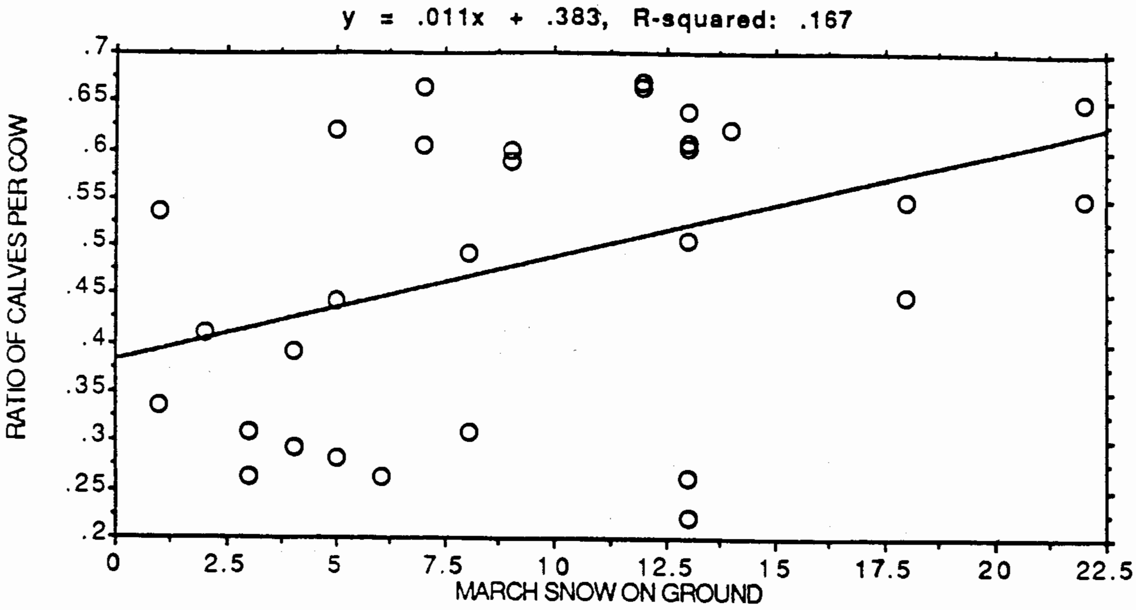
Compounded Independent Variables

Use of derived (or compounded) independent variables, such as running averages, should be avoided whenever possible (Green 1979:95). MLR requires that X-variables be known or measured without error. MLR also is applicable where X's are independent random variables (Neter et al. 1985:83). Because any level of X' (where X' denotes a compounded variable) is not independent of other values of X', this basic and important assumption of MLR is violated. Variables that are summed, averaged, or otherwise compounded across cases also may appear to provide more significant regression coefficients (e.g., coefficients differing significantly from zero) than the simple variables from which they were derived (Fig. 5) and should be avoided.

The importance of several years of snow might be assessed by assigning snowfall in the previous year (i.e., n-1) as a separate independent variable (e.g., X_1 = March snow in the same year as reproduction; X_2 = March snow in the previous year to reproduction; etc.). If some interaction in snowfall among years seems necessary to explain the biology, then, for example, $X_3 = X_1 X_2$ also might be examined as a separate variable (Neter et al. 1985:232-234).

Dealing with Multicollinearity

An important assumption of MLR is that independent variables are not substantially correlated. If multicollinearity among X variables exists, parameter estimates will have large variances (Neter et al. 1985:271-282, 382-390; Zar 1984:344). This problem is so severe that several authors (Green 1979:117, Pimentel 1979:45) have recommended against using a MLR approach. The removal of intercorrelated variables can be accomplished informally by examining a correlation matrix (Table 3) and eliminating 1 variable from each pair with $|r| \geq 0.70$. Selection of a variable from an intercorrelated pair to remove from consideration in model building can be based on the difficulty or cost of obtaining a particular measurement.



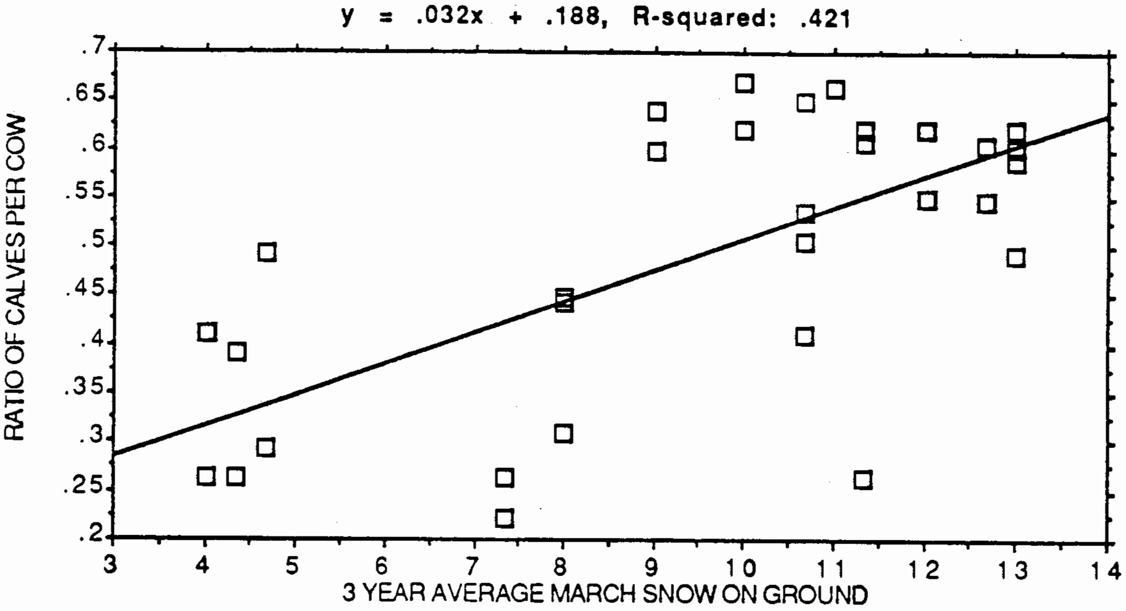


Fig. 5. Graphic representation of the effect of compounding variables. Compounded variables (below) often will have regression coefficients greater than simple variables (above) from which they were derived.

Table 3. Correlation matrix showing correlation coefficients (r) between all variable pairs considered in the caribou population.

Year	No. of caribou	No. of females harvested	No. of wolves	No. of calves/ cows	% cows	March snow depth (in)	3-year average snow depth (in)
	1	.769	.032	-.938	.776	-.319	-.557
		1	.024	-.788	.607	-.195	-.615
			1	.046	-.213	.116	.157
				1	-.758	.366	.649
					1	-.156	-.394
						1	.450
							1

Note: 2 cases deleted with missing values.

Construction of a correlation matrix is an essential step in building a valid regression model and should be presented with the final model.

After intercorrelated variables have been removed, and a model constructed with the remaining independent variables, further elimination of independent variables may be necessary. Another justification for removing one of a pair of similar variables is a change in sign of the regression coefficient associated with 1 independent variable when another independent variable enters the model. Reasons for this change in algebraic sign are discussed elsewhere (Mullet 1976). Large fluctuations in estimated variances of regression coefficients as other variables enter or leave the model also are indicative of multicollinearity among independent variables (Neter et al. 1985:390).

A protocol sometimes used to remedy multicollinearity is principal components analysis (PCA). PCA combines interrelated variables into a single independent variable (Draper and Smith 1981:258). We do not recommend this procedure because: (1) it is largely a descriptive technique in which statistical inference is not attempted; (2) the hodgepodge of variables loaded on a PCA axis often defies biological interpretation; and (3) this technique gives similar results from analyses of real and random data (Karr and Martin 1981). Mitchell-Olds and Shaw (1987) provide additional cautions about the use of PCA for regression analysis.

Several other methods have been suggested to remedy problems with multicollinearity after the model has been constructed (Belsley et al. 1980, Neter et al. 1985:390-400, Willan and Watts 1978). Whichever method is used, a final check of variance inflation factors (VIF's) is recommended (Table 4). A VIF >10 for any regression coefficient is thought to indicate problems with multicollinearity (Neter et al. 1985:391-393). This check is necessary because not all intercorrelated variables are readily apparent in the correlation matrix (i.e., X_1 may be strongly related to X_2 and X_3 together, but not singly) and may not have been removed from the model. If a computer package does not provide VIF, they may be calculated as the reciprocal of the tolerance value for each β-coefficient. Of course, VIF cannot be determined until a regression model is built; this topic is discussed in the following section.

Table 4. Multiple regression model predicting Y (calves/cow) with X-variables (number of caribou = X_1, female harvest = X_2).[a]

Parameter	Coefficient value	S.D.	Beta values	Partials	R^2	R^2_a	VIF
Intercept (b_o)	0.771	0.0215	-	-	-	-	
# Caribou (b_1)	-0.000056	0.000007	-0.727	0.671	-	-	2.2
Fem. Harvest (b_2)	-0.00074	0.00025	-0.265	0.213	0.884	0.876	2.2

[a] In addition to the basic model, adjusted (R_a^2) and unadjusted (R^2) coefficients of multiple determination, standardized regression coefficients (beta values), coefficients of partial determination (partials), variance inflation factors (VIF), and standard deviations of the regression coefficients are shown. Partial coefficients and β- values suggest the same relative importance of X_1 and X_2 (note that partial coefficients only may be positive values).

Selecting a MLR Procedure

Most statistical computing packages offer several multiple regression options. Opinions vary as to the "best" method for choosing a set of independent variables to be included in the model (Draper and Smith 1981:294-380, Neter et al. 1985:417-443). We recommend 1 of 3 general methods for building a MLR model: (1) backward elimination; (2) "true" stepping; or (3) all possible regressions (Draper and Smith 1981:294-380, Neter et al. 1985:417-443).

Backward elimination of independent variables starts with all variables in the model and eliminates them one at a time based on their partial *F*-values (often *F*-to-remove = 3.996). This procedure has advantages of requiring less computing time and of initially including all variables so that none are "missed." Once a variable exits the model, however, it cannot re-enter the analysis. Backward elimination (as well as the other search algorithms we recommend) does not require the arbitrary selection of the first independent variable to enter the model. Order of variable elimination and variables not re-entering the model may affect the final outcome. If variables are not free to enter or leave at all stages, the first variable removed or entered may predetermine whether other independent variables become part the final model (Draper and Smith 1981:307-308).

"True" stepping possesses the advantage of allowing independent variables to enter and exit the model based on their partial *F*-values (typically *F*-to-enter = 4.00, *F*-to-remove = 3.996). Other numeric values of partial *F* might be used based on degrees of freedom associated with the mean square error (MSE), but the *F*-to-enter value *must be larger* than that of the *F*-to-remove (Neter et al. 1985:435). If values of α are used, however, α-to-enter should always be *smaller* than α-to-remove (i.e., α decreases as *F* increases). True stepping requires less computing time than an "all possible regressions" approach, but does not show the researcher all the models that might be built.

True stepping should not be confused with "forward selection" of variables, in which the biologist is asked to specify the order in which variables enter the model and sometimes the number of steps to be taken. Forward selection will not allow removal of a variable that already has been added to the model, as will a true stepping procedure (Neter et al. 1985:435).

An "all possible regressions" procedure builds a series of models that includes all potential combinations and arrangements of independent variables. Clearly, the "best" model will be among those generated by this method, but if there are many independent variables (m), it may be overlooked by a biologist scanning pages of computer output. There will be 2^m regressions, so that 10 independent variables (m) will result in 1,024 regression equations. This is time-consuming and usually wasteful of computer time because many potential combinations of variables would have otherwise been rejected without subjecting them to MLR analysis (Draper and Smith 1981:302). Many statistical packages incorporate a procedure to search for the "best" subset of equations (Garside 1971, Hocking 1972, Furnival and Wilson 1974). Thus, several equations representing the "best" one-variable solution, two-variable solution, three-variable solution, and so forth, are presented. The investigator then chooses the "best" model (i.e., the one explaining the most variation with the fewest variables).

Several methods are available to aid in selecting the best model. For situations in which a large number of independent variables ($m > 10$) are incorporated, a plot of the residual mean square error against the number or parameters in the model may be helpful. The best model occurs at the point where variation in the residual mean square error stabilizes (Draper and Smith 1981:298-299).

Another method involves Mallows' Cp statistic. Here, Cp-values are plotted against the number of parameters (p) in each model (Kennard 1971, Draper and Smith 1981:302, Neter et al. 1985:427). The model with the fewest variables lying closest to the line $Cp = p$ is considered best.

A final method is to select the model with the highest adjusted multiple coefficient of determination ($\underline{R}^2_a$) (Table 4). $\underline{R}^2$ only can increase as new variables are added to the model, regardless of the contribution made by these new variables. Indeed, the value for $\underline{R}^2$ when none of the independent variables are related to the dependent variable is m/n-1 (Harris 1975:46). $\underline{R}^2_a$ adjusts the gain in explanatory power by the number of parameters in the model and will only increase if the new information is more important than the loss of degrees of freedom (Draper and Smith 1981:303, Neter et al. 1985:423-425).

These search procedures and their statistical outputs are related mathematically and should result in selection of a similar best model. In practice, it is worthwhile to combine procedures in selecting a final model. For instance, several equations with high $\underline{R}^2_a$ and few independent variables might be compared with Mallows' $\underline{Cp}$. Similarly, the use of different procedures for determining which X's to include (e.g., backwards, true, or all models) should result in selection of the same independent variables; great discrepancies in variables selected may indicate problems.

Testing for Aptness

Once a regression model with appropriate independent variables has been selected, it is necessary to judge whether the model is apt. This procedure involves further examination of residuals. Residuals are plotted against fitted values of the dependent variable ($\hat{Y}$) and inspected for a pattern (Fig. 6). Residuals should not be plotted against observed values of Y with which they often are correlated (Draper and Smith 1981:147, Framstad et al. 1985). A cloud of points around the line Y = zero on the residual axis would be indicative of an apt model, whereas a positive or negative slope, parabola, or trapazoidal pattern of points may indicate violations of assumptions (Draper and Smith 1981:145-146, Zar 1984:288, Neter et al. 1985:113). Again, departures from the model that may be detected by an analysis of residuals include: (1) nonlinearity; (2) lack of constant variance in error terms; (3) error terms not being normally distributed; (4) outliers; (5) lack of independence in error terms; and (6) omission of

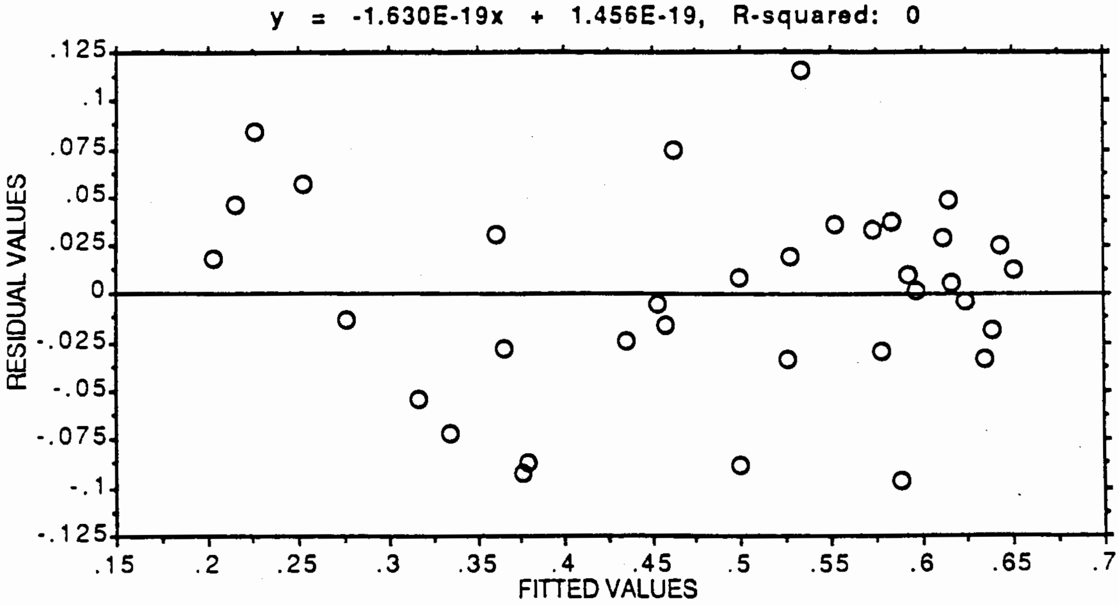


Fig. 6. Testing whether the final model is apt requires a plot of residual (Yi - $\bar{Y}$i) against the values fitted ($\bar{Y}$) by the multiple regression model.

important independent variables (Neter et al. 1985:111). Because of previous testing during the selection of independent variables, problems in 1-3 would not be expected at this time but are worth considering.

Outcomes from regression analyses may be strongly affected by outliers (observations that make extreme departures from expected values), especially when sample sizes are small (Draper and Smith 1981:152-153, Neter et al. 1985:114-115). In addition to the inspection of residual plots to locate outliers, a number of more formal procedures exist (Neter et al. 1985:400-411). Outliers should not be deleted unless measurement or recording errors are suspected. Outliers may add significant information to the model, such as when they result from interactions with variables omitted from the analysis (Neter et al. 1985:115).

Because studies of population dynamics typically require data collected over a number of years, it is worthwhile to plot residuals against time to ensure that no correlation exists among error terms (Fig. 7). Lack of independence sometimes may be corrected by incorporating time (years) as an independent variable in the model (Neter et al. 1985:115-118).

Searching for patterns in the residuals that result from a failure to include an important independent variable is essential to building a statistically sound and biologically meaningful model. Otherwise, controlling for multicollinearity through scrutiny of a correlation matrix or VIF's could result in elimination of the "wrong" variable in an intercorrelated pair. The Durbin-Watson test (Neter et al. 1985:450-460) offers a formal method by which temporal patterns in residuals may be detected. A significant outcome from this test will require re-examination of the correlation matrix, building a new model, and checking VIF's. If the problem is not thereby resolved, a search for another, and perhaps previously unconsidered, independent variable is in order. Failure to meet assumptions concerning residuals is evidence that the model is not apt, and reason to abandon a MLR approach. Conversely, a lack of pattern in the residuals and nonsignificant

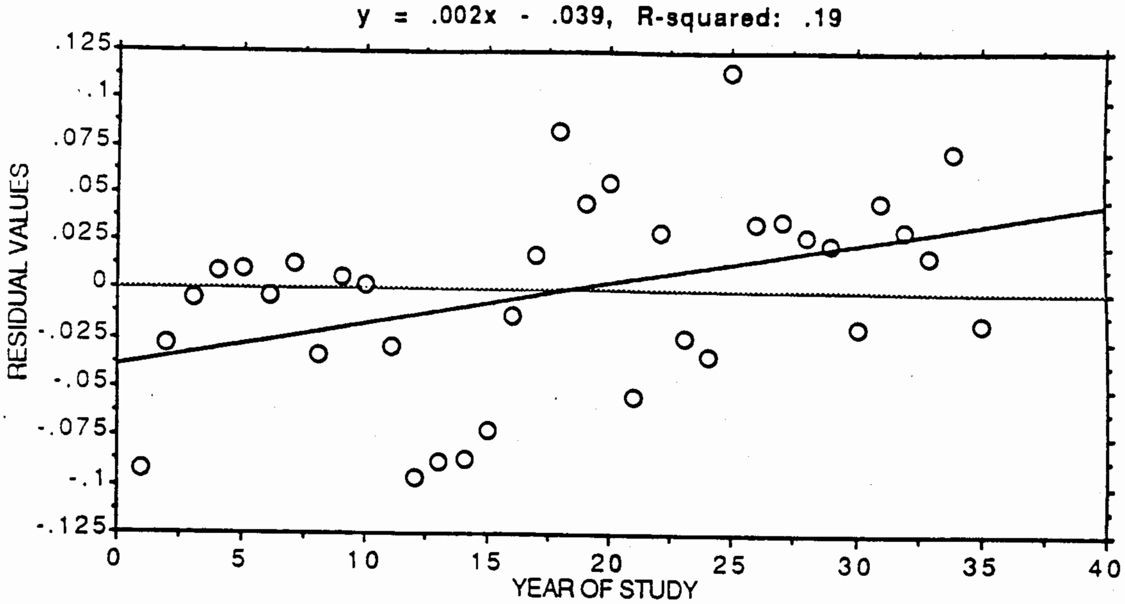


Fig. 7. Plot of residuals (Yi - $\bar{Y}$i) against time suggesting a possible relationship between year of study and size of variances (error terms).

outcome from a Durbin-Watson test suggest that selection of independent variables was reasonable, and modeling may proceed.

An important test for aptness of a regression model is the F-test for lack-of-fit (Marzluff 1986). This test is conducted after the MSE given by the regression analysis is partitioned into pure error and lack of fit components. Fit should be evaluated whenever possible. Partitioning of MSE, however, requires true, or at least closely approximate, replications (i.e., more than one observed Y for each value of the X's) (Draper and Smith 1981:42). For MLR applications to wildlife data, however, replicates are not likely because all X-values must be repeated. Thus, there may be little opportunity to test for lack-of-fit.

Model Interpretation and Validation

It is desirable to distinguish between models that are useful for prediction and those that are significant but lack predictive value. In general, a model is useful for prediction if the F-ratio is at least 4-5 times greater than that required for significance at the chosen level of α (e.g., $\alpha = 0.05$). Draper and Smith (1981:129-133) discuss and provide tables for determination of the F-ratio desired for prediction.

Great care should be taken in interpreting regression coefficients because their relative magnitudes are influenced by the units of measurement. For instance, if b_1 was millimeters of snowfall, and b_2 was number of wolves, they would not be directly comparable because of their different units. This problem may be resolved by calculating standardized regression coefficients or beta values (Zar 1984:338, Neter et al. 1985:261-263), which are unitless measures that may be compared directly so long as problems with multicollinearity have been rectified. Both unstandardized and standardized (which lacks b_o) regression equations should be presented for the final model.

The coefficient of multiple determination (R^2), mentioned earlier, also can be partitioned to evaluate the relative contribution of each independent variable to the regression. Coefficients of partial determination measure the contribution of each independent variable when all others have been included in the model (Neter et al. 1985:286). Coefficients of partial determination reveal information similar to that provided by standardized regression coefficients. These partial coefficients are less susceptible (but not immune) to problems with multicollinearity because they measure the contribution of each variable after adjustment for the degree of linear relationship with other variables (Neter et al. 1985:288). Nonetheless, coefficients of partial determination rely on output from the same matrices used for calculating other MLR statistics; thus rounding errors resulting from multicollinearity in these matrices also would affect coefficients of partial determination.

Both standardized beta-values and coefficients of partial determination suggest which variables are most influential in the final model. Keep in mind, however, that coefficients of partial determination will always have a positive algebraic sign, whereas standardized regression coefficients may be either positive or negative. Failure of both procedures to suggest the same order of influence (ignoring sign) for independent variables may be an indication of multicollinearity.

Just as the coefficient of multiple determination ($\underline{R}^2$) and $\underline{F}$-test reveal the predictive value of the total regression model, the significance of each independent variable also can be evaluated. Whether individual regression coefficients differ from zero (i.e., do they contribute significantly to the model?), from each other, or from some arbitrary value can be determined with partial $\underline{F}$-tests. Many computer programs present extra sums of squares that allow a variety of tests regarding the regression coefficients without additional computer runs or manual computations. Tests of significance for individual regression coefficients are performed in stepwise regression procedures as each variable is considered for entry into the model. These $\underline{F}$-to-enter tests are equivalent to partial $\underline{F}$-tests performed with derived extra sums of squares. Interval estimates for these regression coefficients can be derived from the standard deviations provided by most computer programs (Table 4).

Biologists often are interested in estimating the mean response (Y) at particular values of X-variables. Likewise, we sometimes wish to predict the response at some new level(s) of X's. Some computer programs will provide these values along with the other outcomes so long as the desired prediction levels of X's are provided. If these values are not generated automatically, confidence bands on predicted or estimated values of the dependent variables can be constructed from intermediate computer outputs. If the model is used for predictive purposes, it should be kept in mind that confidence intervals for predicted values of new observations (e.g., in future years) of $\hat{Y}$ are wider than those for expected (mean) values (Zar 1984:272-276, 344-346; Neter et al. 1985: Chs 3, 5, 7).

One question that should concern biologists using MLR is the applicability of the final model over the range of interest. This can be evaluated by examining stability of partial regression coefficients (β's). For instance, the entire data set might be subdivided into 5 periods, and β's recalculated for each time-interval (provided there are not too many independent variables in the model). If these β's varied greatly among periods, as in a step function, or exhibited a trend, it would be unwise to use all the data points to build 1 model for predictive purposes (Draper and Smith 1981:419); in this case it may be necessary to construct more than 1 model from these data. "Jacknife" and other resampling procedures also are available to examine the stability of β's (Wu 1986); another option is to use the "PRESS" statistic. Likewise, it is hazardous to attempt to use a model to predict the dependent variable when values of important independent variables lie outside the range used to construct the original model (i.e., be careful not to extrapolate too far beyond the range of observed values).

Another validation method is to randomly subdivide the data set and test the model derived from 1 subset against the remainder (e.g., percentage of observations predicted correctly) (Snee 1977, Draper and Smith 1981:419). Studies of caribou populations seldom will include a sufficient number of cases (years) to permit this procedure, or an opportunity to collect additional years of data.

It may be unwise to apply the results from modeling 1 caribou population to another, especially if variables fall outside previously measured ranges. Further, new and unmeasured conditions affecting dependent variables may be operating in the second population. If a second population model is built with the same variables contained in the first, however, it is possible to test it against the first model (Zar 1984:347-349).

A cautious approach to interpreting the final model is advisable. Even a valid and highly predictive equation may lack biologically important variables. For instance, "number of adult males" might not enter a model designed to predict recruitment rate (perhaps because sexual segregation reduces competition of adult males with young), but a population without these males certainly would fail to increase at the expected rate!

Certain variables that are highly intercorrelated with other variables included in the model may be overlooked even though they are directly responsible for the observed changes in the dependent variable. Such "lurking" variables are the reason that a MLR approach cannot be used to infer cause and effect.

We believe it unlikely that the dynamics of caribou populations will be understood fully by relying on an inductive approach that searches for past examples that "fit" existing hypotheses. Likewise, MLR analysis of unplanned data collected for other purposes has limited usefulness. The need for experimental (manipulative) studies of wildlife is well-documented (McCullough 1979, James and McCulloch 1985, Noon 1986). Certainly, a hypothetico-deductive approach to test factors regulating caribou populations will require such manipulations. Nonetheless, MLR offers a reasonable approach to identify important variables for such experiments, especially where complex interactions occur among many potential factors.

LITERATURE CITED

Belsley, D. A., E. Kuh, and R. E. Welsch. 1980. Regression diagnostics: Identifying influential data and sources of collinearity. John Wiley and Sons, Inc., New York.

Capen, D. E., ed. 1981. The use of multivariate statistics in studies of wildlife habitat. USDA, For. Serv., Gen. Tech. Rep. RM-87, Rocky Mountain Forest and Range Exp. Sta., Fort Collins, Colorado. 249pp.

Conover, W. J. 1980. Practical nonparametric statistics. John Wiley, New York. 483pp.

Box, G. E. P. 1966. Use and abuse of regression. Technometrics 8:625-629.

Draper, N. R., and H. Smith. 1981. Applied regression analysis, 2nd ed., John Wiley and Sons, Inc., New York. 709pp.

Dunn, J. E. 1981. Data-based transformations in multivariate analysis. Pages 93-103 in D. E. Capen, ed. The use of multivariate statistics in studies of wildlife habitat. USDA, For. Serv., Gen. Tech. Rep. RM-87, Rocky Mountain Forest and Range Exp. Sta., Fort Collins, Colorado. 249pp.

Framstad, E., S. Engen, and N. C. Stenseth. 1985. Regression analysis, residual analysis and missing variables in regression models. Oikos 44:319-323.

Furnival, G. M., and R. W. Wilson. 1974. Regression by leaps and bounds. Technometrics 16:499-551.

Garside, M. J. 1971. Best subset search. Applied Statistics 20:112-115.

Green, R. H. 1979. Sampling design and statistical methods for environmental biologists. John Wiley and Sons, Inc., New York. 257pp.

Harris, R. J. 1975. A primer of multivariate statistics. Academic Press, New York. 332 pp.

Hocking, R. R. 1972. Criteria for selection of a subset regression: Which one should be used? Technometrics 14:967-970.

James, F. C., and C. E. McCulloch. 1985. Data analysis and the design of experiments in ornithology. Pages 1-102 in R. F. Johnston, ed. Current ornithology, Vol. 2, Plenum, New York.

Johnson, D. H. 1981. The use and misuse of statistics in wildlife habitat studies. Pages 11-19 in D. E. Capen, ed. The use of multivariate statistics in studies of wildlife habitat. USDA For.

Serv., Gen. Tech. Rep. Rm-87, Rocky Mountain Forest and Range Exp. Sta., Fort Collins, Colorado. 249pp.

Karr, J. R., and T. E. Martin. 1981. Random numbers and principle components: further searches for the unicorn? Pages 20-25 in D. E. Capen, ed. The use of multivariate statistics in studies of wildlife habitat. USDA For. Serv., Gen. Tech. Rep. RM-87. Rocky Mountain Forest and Range Exp. Sta., Fort Collins, Colorado. 249pp.

Kennard, R. W. 1971. A note on the Cp statistic. Technometrics 12:69-92.

Marzluff, J. M. 1986. Assumptions and design of regression experiments: The importance of lack-of-fit testing. Pages 165-170 in J. Verner, M. L. Morrison, and C. J. Ralph, eds. Wildlife 2000 modeling habitat relationships of terrestrial vertebrates. Univ. Wisconsin Press, Madison. 470 pp.

McCullough, D. R. 1979. The George Reserve deer herd: population ecology of a k-selected species. Univ. Michigan Press, Ann Arbor. 271pp.

Milliken, G. A., and D. F. Johnson. 1984. Analysis of messy data, Vol I: designed experiments. Van Nostrand Reinhold Co., New York. 473pp.

Mitchell-Olds, T., and R. G. Shaw. 1987. Regression analysis of natural selection: Statistical inference and biological interpretation. Evolution 41:1149-1161.

Mullet, G. M. 1976. Why regression coefficients have the wrong sign. J. Qual. Tech. 8:121-126.

Neter, J., W. Wasserman, and M. H. Kutner. 1985. Applied linear statistical models: regression, analysis of variance and experimental designs. Irwin, Homewood, Illinois. 1127pp.

Noon, B. R. 1986. Summary: biometric approaches to modeling - the researcher's viewpoint. Pages 197-202 in J. Verner, M. L. Morrison, and C. J. Ralph, eds. Wildlife 2000, modeling habitat relationships of terrestrial vertebrates. Univ. Wisconsin Press, Madison. 470pp.

Pimentel, R. A. 1979. Morphometrics: the multivariate analysis of biological data. Kendall Hunt Inc., Dubuque, Iowa. 276pp.

Shapiro, S. S., and M. B. Wilk. 1965. An analysis test for normality (complete samples). Biometrika 52:591-611.

Snee, R. D. 1977. Validation of regression models: methods and examples. Technometrics 19:415-428.

Sokal, R. R., and F. J. Rohlf. 1969. Biometry: the principles and practice of statistics in biological research. W. H. Freeman Co., San Francisco, Calif. 776pp.

Starfield, A. M., and A. L. Bleloch. 1986. Building models for conservation and wildlife management. MacMillan, New York. 253pp.

Verner, J., M. L. Morrison, and C. J. Ralph, eds. 1986. Wildlife 2000, Modeling habitat relationships of terrestrial vertebrates. Univ. Wisconsin. Press, Madison. 470pp.

Wehausen, J. D., V. C. Bleich, B. Blong, and T. L. Russi. 1987. Recruitment dynamics in a southern California sheep population. J. Wildl. Manage. 51:86-98.

Willan, A. W., and D. G. Watts. 1978. Meaningful multicollinearity measures. Technometrics 20:407-417.

Wu, C. F. 1986. Jacknife, bootstrap, and other resampling methods in regression analysis. Ann. Sta. 14:1261-1295.

Zar, J. H. 1984. Biostatistical analysis. 2nd ed., Prentice-Hall Inc., Englewood Cliffs, New Jersey. 718pp.