Chapter 21

Multiple Regression

Multiple regression is a statistical technique for examining the relationship between a dependent variable Y and multiple independent variables or regressors $X_1, X_2, \ldots, X_k$. Like with linear regression, the independent variables or regressors may be fixed values under experimental control, or random variables. One purpose of multiple regression is to determine whether changes in any of the independent variables cause changes in Y. This involves testing whether the slope β_j for a given independent variable X_j is significantly different from zero, for each of the independent variables. There is also an overall test that examines whether any of independent variables (alone or in combination) affect Y. Another purpose of multiple regression is prediction, using a set of values for the independent variable to predict the value of Y along with a confidence interval. A third use is model selection. The objective here is to find a model that approximates the data with the fewest variables, involving a trade-off between model fit and model complexity. We will examine a popular method of model selection that uses Akaike's Information Criterion or AIC (Akaike 1974; Anderson et al. 2000, Burnham & Anderson 2002).

We will first illustrate multiple regression using a relatively simple data set from a study of southern pine beetle, *Dendroctonus frontalis* (Reeve et al. 1998). We previously used this study to examine the relationship between the number of beetles added to caged trees and how this affected their attack density. We now examine how attack density and the density of a competitor, bluestain fungus, affects the survival rate of beetle offspring (from egg to emerging adult). High attack densities imply a high density of adult beetles within the tree, and this crowding could reduce survival of their offspring

(see also Coulson et al. 1976). High levels of bluestain fungus are also known to reduce survival, by interfering with the beetle's own symbiotic fungus (Hofstetter et al. 2006).

Table 21.1: Example 1 - Effects of attack density and bluestain fungus on the survival of *D. frontalis* brood from egg to emergence (Reeve et al. 1998). The dependent variable was the log-transformed survival rate of the beetle offspring, while attack density (attacks per 100 cm^2 of bark) and the proportion of bluestained phloem were the independent variables.

X_{1i} = Attack density	X_{2i} = Bluestain	Survival	Y_i = ln(Survival)	i
1.250	0.000	0.107	-2.235	1
2.656	0.481	0.715	-0.335	2
7.334	0.171	0.036	-3.324	3
1.603	0.352	0.188	-1.671	4
2.622	0.016	0.438	-0.826	5
1.000	0.000	0.585	-0.536	6
4.342	0.185	0.115	-2.163	7
5.233	0.018	0.257	-1.359	8
2.500	0.410	0.032	-3.442	9
3.250	0.015	0.350	-1.050	10
6.000	0.007	0.161	-1.826	11
4.750	0.000	0.073	-2.617	12
2.500	0.095	0.219	-1.519	13
8.750	0.033	0.028	-3.576	14
6.000	0.015	0.294	-1.224	15
5.000	0.105	0.207	-1.575	16
7.149	0.025	0.227	-1.483	17
6.750	0.015	0.040	-3.219	18
7.500	0.043	0.089	-2.419	19
2.500	0.073	0.176	-1.737	20
5.000	0.055	0.084	-2.477	21
2.250	0.023	0.203	-1.595	22
1.250	0.123	0.074	-2.604	23
4.750	0.035	0.126	-2.071	24
4.500	0.212	0.290	-1.238	25
9.557	0.166	0.010	-4.605	26
5.000	0.338	0.207	-1.575	27

We will use another data set to illustrate prediction in multiple regression. Soul et al. (2013) were interested in predicting endocranial volume (brain size) in extinct mammals, where only the skull length, height, and width are available. For this purpose, they developed a multiple regression model using existing species as the observations, with endocranial volume the dependent variable, and skull length, width and height the independent ones. A portion of these observations are listed below (see https://datadryad.org for the full data set). We will fit a multiple regression model to these observations, then use them to predict endocranial volume for two hypothetical fossils, a mouse and a bear.

Table 21.2: Example 2 - Skull length, width, height, endocranial volume, and species name (Soul et al. 2013). The dependent variable was endocranial volume, estimated using the mass of glass beads filling the skull.

Length (mm)	Width (mm)	Height (mm)	Volume (g)	i	Common name
15.04	11.29	6.61	0.38	1	Pygmy glider
52.40	30.94	25.68	12.36	2	Rufous kangaroo rat
75.87	52.79	39.45	56.70	3	Howler monkey
41.73	25.70	16.79	5.68	4	Scaley-tailed squirrel
39.71	26.87	17.13	5.92	5	Lord derby's flying squirrel
18.90	12.62	7.61	0.51	6	Yellow-footed antechinus
15.10	11.69	7.06	0.46	7	Brown antechinus
123.70	73.89	63.93	150.53	8	Pronghorn
46.75	28.70	18.45	6.51	9	Mountain beaver
154.32	103.77	71.95	284.03	10	Antarctic fur seal
133.39	59.75	72.60	128.49	11	Babiroussa
etc.					
32.90	19.83	14.73	3.19	185	Tree shrew
32.15	20.33	13.95	3.17	186	Painted tree shrew
200.23	98.99	84.53	358.82	187	Brown bear
179.70	95.48	75.51	302.72	188	Sloth bear
67.48	42.35	29.66	24.79	189	Ruffled lemur
70.78	30.98	28.08	17.91	190	Rasse
67.05	54.15	44.99	56.71	191	Wombat
70.36	45.09	37.72	38.43	192	Arctic fox
80.73	47.96	39.45	48.55	193	Fox
13.54	9.24	7.13	0.36	194	Meadow jumping mouse
13.15	9.05	7.00	-	195	Fossil mouse
190.17	97.32	80.31	-	196	Fossil bear

21.1 Multiple regression model

Suppose we want to model the observations for a data set like Example 1, where a dependent variable Y is observed along with two independent variables X_1 and X_2. Let Y_i, X_{1i}, and X_{2i} be the ith set of values. The multiple regression model takes the form

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i, \tag{21.1}$$

where β_0 is the intercept, β_1 and β_2 are the slopes or regression coefficients for X_1 and X_2, and $\epsilon_i \sim N(0, \sigma^2)$ (Draper & Smith 1981; Kutner et al. 2005, Sheather 2009). This equation defines a plane in three dimensions, which we will later visualize for the Example 1 data.

More generally, the model for k different independent variables $X_1, X_2, \ldots, X_k$ takes the form

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \epsilon_i. \tag{21.2}$$

Here, the parameters $\beta_1, \beta_2, \ldots, \beta_k$ are the slopes for each independent variable. While this model appears complicated, there is a simple interpretation of the regression coefficients. **The slope β_j can be thought of as the change in Y per unit change in X_j, while holding all the other variables constant.** There are also specific plots designed to visualize this model for any number of independent variables.

21.2 Multiple regression in matrix form

We will now show how the multiple regression model can be expressed in matrix form (Draper & Smith 1981). This will greatly simplify later developments, and in any event the matrix form of the model is commonly used in the statistical literature as well as software documentation. If you are unfamiliar with matrices, there are many online resources that provide an introduction to matrices and linear algebra. The textbook by Tabachnik and Fidell (2001) also provides a useful summary of essential concepts (see their Appendix A). In the following, we will briefly review various matrix operations and then apply them to multiple regression. Chapter 24 of this text lists a SAS program that carries out these operations using `proc iml` (SAS Institute Inc. 2018a).

A matrix is a rectangular collection of numbers (or other quantities) arranged in rows and columns, enclosed in a set of parentheses or brackets. A vector is a simple type of matrix consisting of a single column or row of numbers. Matrices and vectors can be added, multiplied, transposed, and even inverted in their own unique way, and these operations allow one to express the multiple regression model in a compact way as well as estimate the parameters of this model.

We will first make use of **matrix addition** and **multiplication** to write the multiple regression model. Suppose we have two vectors $\boldsymbol{A}$ and $\boldsymbol{B}$ of the following form:

$$\boldsymbol{A} = \begin{pmatrix} a \\ b \\ c \end{pmatrix} \text{ and } \boldsymbol{B} = \begin{pmatrix} d \\ e \\ f \end{pmatrix}. \tag{21.3}$$

To add these two vectors, we simply add the elements of each one to obtain

$$\boldsymbol{A} + \boldsymbol{B} = \begin{pmatrix} a+d \\ b+e \\ c+f \end{pmatrix}. \tag{21.4}$$

For example, suppose

$$\boldsymbol{A} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \text{ and } \boldsymbol{B} = \begin{pmatrix} 4 \\ 5 \\ 6 \end{pmatrix}. \tag{21.5}$$

Then

$$\boldsymbol{A} + \boldsymbol{B} = \begin{pmatrix} 1+4 \\ 2+5 \\ 3+6 \end{pmatrix} = \begin{pmatrix} 5 \\ 7 \\ 9 \end{pmatrix}. \tag{21.6}$$

Note that the two vectors (or matrices) must have the same dimensions or shape for addition to work.

For the multiple regression model, we will also need to multiply a matrix by a vector. Suppose that we have two matrices $\boldsymbol{C}$ and $\boldsymbol{D}$ of the following form:

$$\boldsymbol{C} = \begin{pmatrix} a & d \\ b & e \\ c & f \end{pmatrix} \text{ and } \boldsymbol{D} = \begin{pmatrix} g \\ h \end{pmatrix}. \tag{21.7}$$

Then

$$\boldsymbol{CD} = \begin{pmatrix} a & d \\ b & e \\ c & f \end{pmatrix} \times \begin{pmatrix} g \\ h \end{pmatrix} = \begin{pmatrix} ag + dh \\ bg + eh \\ cg + fh \end{pmatrix}. \tag{21.8}$$

Note the pattern in the multiplication process. You take the elements in each row of $\boldsymbol{C}$ and multiply them by the column elements of $\boldsymbol{D}$, then add the result to obtain $\boldsymbol{CD}$. In this case, the multiplication process takes a 3×2 matrix (3 rows and 2 columns) and a 2×1 matrix, and produces a 3×1 matrix. Thus, the numbers of rows and columns in the product depends on the number of rows in first matrix and columns in the second – this is true of matrix multiplication in general. The number of columns in the first matrix and rows in the second matrix must also match for matrix multiplication to be possible.

As an example of matrix multiplication, suppose that

$$\boldsymbol{C} = \begin{pmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{pmatrix} \text{ and } \boldsymbol{D} = \begin{pmatrix} 7 \\ 8 \end{pmatrix}. \tag{21.9}$$

Then

$$\boldsymbol{CD} = \begin{pmatrix} 1 \cdot 7 + 4 \cdot 8 \\ 2 \cdot 7 + 5 \cdot 8 \\ 3 \cdot 7 + 6 \cdot 8 \end{pmatrix} = \begin{pmatrix} 39 \\ 54 \\ 69 \end{pmatrix}. \tag{21.10}$$

Another matrix operation we will use later is the **transpose** of a matrix. This operation takes the columns of a matrix and turns them into the rows of a new matrix. For example, suppose we have a matrix

$$\boldsymbol{F} = \begin{pmatrix} a & e \\ b & f \\ c & g \\ d & h \end{pmatrix}. \tag{21.11}$$

The transpose of $\boldsymbol{F}$ (written as $\boldsymbol{F'}$) is defined to be

$$\boldsymbol{F'} = \begin{pmatrix} a & b & c & d \\ e & f & g & h \end{pmatrix}. \tag{21.12}$$

For example, suppose

$$\boldsymbol{F} = \begin{pmatrix} 1 & 5 \\ 2 & 6 \\ 3 & 7 \\ 4 & 8 \end{pmatrix}. \tag{21.13}$$

Then

$$\boldsymbol{F'} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{pmatrix}. \tag{21.14}$$

Now suppose we have a multiple regression problem with $k = 2$ independent variables and n observations, similar to the Example 1 data set. The standard model equation for this problem would be

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i, \tag{21.15}$$

for $i = 1$ to n. If we write out the full system of equations for each observation or value of i, we would obtain n different equations:

$$\begin{pmatrix} Y_1 = \beta_0 + \beta_1 X_{11} + \beta_2 X_{21} + \epsilon_1 \\ Y_2 = \beta_0 + \beta_1 X_{12} + \beta_2 X_{22} + \epsilon_2 \\ Y_3 = \beta_0 + \beta_1 X_{13} + \beta_2 X_{23} + \epsilon_3 \\ \vdots \\ Y_n = \beta_0 + \beta_1 X_{1n} + \beta_2 X_{2n} + \epsilon_n \end{pmatrix}. \tag{21.16}$$

Using the definition of matrix addition, these equations can be rewritten in matrix form as

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} \beta_0 + \beta_1 X_{11} + \beta_2 X_{21} \\ \beta_0 + \beta_1 X_{12} + \beta_2 X_{22} \\ \beta_0 + \beta_1 X_{13} + \beta_2 X_{23} \\ \vdots \\ \beta_0 + \beta_1 X_{1n} + \beta_2 X_{2n} \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{pmatrix}. \tag{21.17}$$

Using the definition of matrix multiplication, a further simplification is possible:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & X_{21} \\ 1 & X_{12} & X_{22} \\ 1 & X_{13} & X_{23} \\ \vdots & \vdots & \vdots \\ 1 & X_{1n} & X_{2n} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{pmatrix}. \tag{21.18}$$

As a final step, this equation can be written in the form

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon} \tag{21.19}$$

where

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}, \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{pmatrix}, \text{and} \tag{21.20}$$

$$\boldsymbol{X} = \begin{pmatrix} 1 & X_{11} & X_{21} \\ 1 & X_{12} & X_{22} \\ 1 & X_{13} & X_{23} \\ \vdots & \vdots & \vdots \\ 1 & X_{1n} & X_{2n} \end{pmatrix} \tag{21.21}$$

For an actual data set, these matrices and vectors would contain the values of Y, X_{1i}, and X_{2i}. The matrix $\boldsymbol{X}$ is often called the **design matrix**, because it basically describes the design of the study, including the values of the independent variables, their number, and the overall sample size.

In general, the multiple regression model for k independent variables or regressors can be expressed in the same simple form

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{21.22}$$

where

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}, \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{pmatrix}, \text{and} \tag{21.23}$$

$$\boldsymbol{X} = \begin{pmatrix} 1 & X_{11} & X_{21} & \dots & X_{k1} \\ 1 & X_{12} & X_{22} & \dots & X_{k2} \\ 1 & X_{13} & X_{23} & \dots & X_{k3} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{1n} & X_{2n} & \dots & X_{kn} \end{pmatrix}. \tag{21.24}$$

21.3 Multiple regression and likelihood

We will use maximum likelihood to estimate the parameters in the multiple regression model, making use of the matrix form of the model. Suppose

we have $k = 2$ independent variables similar to the Example 1 data. The multiple regression model in this case would be

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i. \tag{21.25}$$

This model has four parameters to estimate, in particular β_0, β_1, β_2, and σ^2. Consider the first observation in the Example 1 data, for which $Y_1 = -2.235$, $X_{11} = 1.250$, and $X_{21} = 0.000$. For this observation, the model states that $Y_1 \sim N(\beta_0 + \beta_1 X_{11} + \beta_2 X_{21}, \sigma^2)$, and so the likelihood would be

$$L_1 = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(Y_1 - (\beta_0 + \beta_1 X_{11} + \beta_2 X_{21}))^2}{\sigma^2}} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(-2.235 - (\beta_0 + \beta_1 1.25 + \beta_2 0.000))^2}{\sigma^2}} \tag{21.26}$$

The overall likelihood is then defined as the product of the likelihoods for each observation, in particular

$$L(\beta_0, \beta_1, \beta_2, \sigma^2) = L_1 \times L_2 \times \ldots \times L_n. \tag{21.27}$$

Finding the maximum likelihood estimates involves maximizing this quantity with respect to the parameters β_0, β_1, β_2, and σ^2. Similar to linear regression, we can gain some insight into the estimation process by rearranging the likelihood function. It can be written in the form

$$L(\beta_0, \beta_1, \beta_2, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2}\frac{\sum_{i=1}^n (Y_i - (\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}))^2}{\sigma^2}}. \tag{21.28}$$

Focusing on the sum in this expression, we see that values of β_0, β_1, and β_2 that minimize the sum of the squared terms will maximize the overall likelihood. Similar to linear regression, these are also the **least squares** estimates because they minimize the sum of these squared terms (Draper & Smith 1981). We will later see that they minimize the sum of the squared residuals from the plane defined by β_0, β_1, and β_2.

Now consider the case where there are k independent variables, so that the model has $k + 2$ parameters $(\beta_0, \beta_1, \beta_2, \ldots, \beta_k, \sigma^2)$. The likelihood L would have the same structure as above, but with more parameters and independent variables. The maximum likelihood estimates can be found by taking the derivative of L (actually $\log L$) with respect to every parameter, setting these derivatives equal to zero, then solving for the parameter values that satisfy these equations. The result is a complex system of equations

involving the data set and parameters. Using matrix algebra, however, the equations for $\beta_0, \beta_1, \ldots, \beta_k$ can expressed in a very compact form:

$$\boldsymbol{X'X\beta = X'Y} \tag{21.29}$$

Here $\boldsymbol{X}, \boldsymbol{\beta}$, and $\boldsymbol{Y}$ are from the matrix version of the multiple regression model. The idea then is solve this equation for $\boldsymbol{\beta}$ using matrix operations. This set of equations are called the **normal equations** (Draper and Smith 1981; Kutner et al 2005; Sheather 2009). They look a bit like the simple equation $xb = y$, where x and y are known values. You would solve this equation for b by multiplying both sides by x^{-1}, to obtain $x^{-1}xb = x^{-1}y$, or $b = x^{-1}y = y/x$. What we need is the matrix equivalent of x^{-1}.

Note that the inverse of x has the property $x^{-1}x = 1$. The inverse of a matrix has the same property, but the equivalent of the number 1 is called the **identity matrix**, written as $\boldsymbol{I}$. It is defined as a square matrix with ones on the diagonal and zeroes everywhere else. For example, the 3×3 identity matrix is

$$\boldsymbol{I} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{21.30}$$

Similar to the number 1, if you multiply a matrix $\boldsymbol{A}$ by $\boldsymbol{I}$ the result is equal to $\boldsymbol{A}$. For example, suppose that $\boldsymbol{A}$ is defined by the matrix

$$\boldsymbol{A} = \begin{pmatrix} 1 & 6 & 4 \\ 3 & 7 & 6 \\ 4 & 1 & 9 \end{pmatrix}. \tag{21.31}$$

Then we have

$$\boldsymbol{AI} = \begin{pmatrix} 1 & 6 & 4 \\ 3 & 7 & 6 \\ 4 & 1 & 9 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{21.32}$$

$$= \begin{pmatrix} 1 \cdot 1 + 6 \cdot 0 + 4 \cdot 0 & 1 \cdot 0 + 6 \cdot 1 + 4 \cdot 0 & 1 \cdot 0 + 6 \cdot 0 + 4 \cdot 1 \\ 3 \cdot 1 + 7 \cdot 0 + 6 \cdot 0 & 3 \cdot 0 + 7 \cdot 1 + 6 \cdot 0 & 3 \cdot 0 + 7 \cdot 0 + 6 \cdot 1 \\ 4 \cdot 1 + 1 \cdot 0 + 9 \cdot 0 & 4 \cdot 0 + 1 \cdot 1 + 9 \cdot 0 & 4 \cdot 0 + 1 \cdot 0 + 9 \cdot 1 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 6 & 4 \\ 3 & 7 & 6 \\ 4 & 1 & 9 \end{pmatrix} = \boldsymbol{A}$$

Now we can define the inverse of a matrix for a square matrix like $\boldsymbol{A}$. The inverse of $\boldsymbol{A}$, written as $\boldsymbol{A^{-1}}$, is a matrix for which $\boldsymbol{A^{-1}A} = I$ and also $\boldsymbol{AA^{-1}} = I$. Note that the order of multiplication does not matter in this case, although it would for other types of matrices. There are a number of numerical techniques for finding the inverse of a matrix, but we will not be concerned with these details. The inverse of $\boldsymbol{A}$ is the matrix

$$\boldsymbol{A^{-1}} = \begin{pmatrix} -0.934 & 0.820 & -0.131 \\ 0.049 & 0.115 & -0.098 \\ 0.410 & -0.377 & 0.180 \end{pmatrix}. \tag{21.33}$$

Multiplying $\boldsymbol{A^{-1}}$ and $\boldsymbol{A}$, we obtain

$$\boldsymbol{A^{-1}A} = \begin{pmatrix} -0.934 \cdot 1 + 0.820 \cdot 3 - 0.131 \cdot 4 & \dots & \dots \\ 0.049 \cdot 1 + 0.115 \cdot 3 - 0.098 \cdot 4 & \dots & \dots \\ 0.041 \cdot 1 - 0.377 \cdot 3 + 0.180 \cdot 4 & \dots & \dots \end{pmatrix} \tag{21.34}$$

$$= \begin{pmatrix} 1.002 & 0.005 & 0.005 \\ 0.002 & 1.001 & 0.004 \\ -0.001 & 0.001 & 0.998 \end{pmatrix} \approx \boldsymbol{I}. \tag{21.35}$$

The result is not exact because of rounding in the values of $\boldsymbol{A^{-1}}$.

We are now ready to solve the normal equations for $\boldsymbol{\beta}$ using matrix operations. Recall that these equations are of the form

$$\boldsymbol{X'X\beta = X'Y} \tag{21.36}$$

Multiplying both sides of this equation by the inverse of $\boldsymbol{X'X}$, denoted by $\boldsymbol{(X'X)^{-1}}$, we obtain

$$\boldsymbol{(X'X)^{-1}X'X\beta = (X'X)^{-1}X'Y} \tag{21.37}$$

or

$$\boldsymbol{I\beta = (X'X)^{-1}X'Y} \tag{21.38}$$

from which it follows that

$$\boldsymbol{\hat{\beta} = (X'X)^{-1}X'Y} \tag{21.39}$$

Here $\boldsymbol{\hat{\beta}}$ is a vector containing the maximum likelihood (or least squares) estimates $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k$ of the model parameters, except for σ^2. This is

the method used by SAS and other statistical packages to estimate the model parameters. We will later see how the elements of $\boldsymbol{(X'X)^{-1}}$ are also used to calculate standard errors and confidence intervals. Similar methods are used to estimate the parameters for ANOVA models. In this case, the design matrix encodes the various treatment combinations and interactions.

The estimates of the model parameters can be used to generated a predicted value for each observation in the data set, of the form

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \ldots + \hat{\beta}_k X_{ki}. \tag{21.40}$$

The residual of each observation is the difference between the observed and predicted values, namely $Y_i - \hat{Y}_i$.

Maximum likelihood also provides an estimator of σ^2 similar to linear regression. Define an error sum of squares by the equation

$$SS_{error} = \sum_{i=1}^{n} \Big(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \ldots + \hat{\beta}_k X_{ki}) \Big)^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2. \tag{21.41}$$

The multiple regression form of MS_{error}, and an estimator of σ^2, is obtained by dividing SS_{error} by $n - k - 1$ degrees of freedom:

$$MS_{error} = \frac{SS_{error}}{n - k - 1} = \hat{\sigma}^2. \tag{21.42}$$

$SS_{regression}$ describes variation in the data explained by the regression model, similar to linear regression. It is defined as

$$SS_{regression} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 \tag{21.43}$$

and has k degrees of freedom. We therefore have

$$MS_{regression} = \frac{SS_{regression}}{k}. \tag{21.44}$$

$SS_{regression}$ and $MS_{regression}$ will be large if $\hat{Y}_i$ varies strongly with respect to one or more of the independent variables $(X_{1i}, X_{2i}, \ldots, X_{ki})$.

The total sum of squares for multiple regression is defined as

$$SS_{total} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \tag{21.45}$$

and has $n - 1$ degrees of freedom. Similar to linear regression, there is an additive relationship among the different sums of squares:

$$SS_{regression} + SS_{error} = SS_{total}. \tag{21.46}$$

We can use the two mean squares to construct an overall F test for the multiple regression, which tests $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$. This null hypothesis basically says none of the independent variables $(X_{1i}, \ldots, X_{ki})$ affect the dependent one (Y_i). The alternative hypothesis is that one or more slopes are different from zero ($H_1 : \beta_j \neq 0$ for some j). If this test is significant, it suggests one or more of the independent variables are affecting the dependent variable, but not which ones. The test statistic is

$$F_s = \frac{MS_{regression}}{MS_{error}}. \tag{21.47}$$

Under H_0, F_s has an F distribution with $df_1 = k$ and $df_2 = n - k - 1$ the degrees of freedom. Note that we encountered a similar test in the SAS output for ANOVA designs, but in ANOVA we were more concerned with tests of each treatment effect, not in testing the overall model. It is also a likelihood ratio test using the H_0 and H_1 models for the data (McCulloch & Searle 2001).

We can organize the different sum of squares and mean squares into an ANOVA table for multiple regression (Table 21.3). It lists the different sources of variation in the data (regression, error, and total), their degrees of freedom, as well as the overall F test.

Table 21.3: General ANOVA table for multiple regression, showing formulas for different mean squares and the overall F test.

Source	df	Sum of squares	Mean square	F_s
Regression	k	$SS_{regression}$	$MS_{regression} = SS_{regression}/k$	$MS_{regression}/MS_{error}$
Error	$n-k-1$	SS_{error}	$MS_{error} = SS_{error}/(n-k-1)$	
Total	$n-1$	SS_{total}		

21.4 Tests and confidence intervals for $\boldsymbol{\beta}$

We next develop tests and confidence intervals for the parameters of the multiple regression model, in particular the slope parameters $\beta_1, \beta_2, \ldots, \beta_k$ and also the intercept β_0. These will help us evaluate which (if any) of the independent variables affect the dependent one. These tests and confidence intervals are based on the maximum likelihood estimates of each β_j and its standard error s_{β_j}, given by the formula

$$s_{\hat{\beta}_j} = \sqrt{\hat{\sigma}^2 d_{j+1,j+1}}, \tag{21.48}$$

for $j = 0, 1, \ldots, k$. Here $\hat{\sigma}^2 = MS_{error}$ and $d_{j+1,j+1}$ is the entry in the $(j+1)th$ row and column of the matrix $\boldsymbol{(X'X)^{-1}}$, i.e., the diagonal entries of this matrix (Draper & Smith 1981). For example, for β_0 and $j = 0$ we would use $d_{0+1,0+1} = d_{11}$, the entry in the first row and column. It can then be shown that the quantity

$$\frac{\hat{\beta}_j - \beta_j}{s_{\hat{\beta}_j}} \tag{21.49}$$

has a t distribution with $n-k-1$ degrees of freedom, the same as for MS_{error}. This fact can be used to derive tests and confidence intervals for each β_j.

Suppose we want to test $H_0 : \beta_j = \beta_{j0}$ vs. $H_1 : \beta_j \neq \beta_{j0}$, where β_{j0} takes some value of interest. We would use the test statistic

$$T_s = \frac{\hat{\beta}_j - \beta_{j0}}{s_{\hat{\beta}_j}}. \tag{21.50}$$

Under H_0, T_s has a t distribution with $n - k - 1$ degrees of freedom, and we would reject H_0 for sufficiently large values of this statistic. The most commonly used null hypothesis tested is $H_0 : \beta_j = 0$ – if this test is significant it suggests the slope for X_j differs from zero, and so X_j is causing a change in Y. Note that this test examines the unique effect of X_j on Y with all the other independent variables in the model, in effect pitting X_j against all the other independent variables.

Confidence intervals can also be derived using the t distribution with $n - k - 1$ degrees of freedom. The interval

$$(\hat{\beta}_j - c_{\alpha,n-k-1} s_{\hat{\beta}_j}, \hat{\beta}_j + c_{\alpha,n-k-1} s_{\hat{\beta}_j}) \tag{21.51}$$

is a $100(1-\alpha)\%$ confidence interval for β_j, where $c_{\alpha,n-k-1}$ could be obtained from Table T (see Chapter 9 for details). We will let SAS handle the details for these confidence intervals as well as tests.

Chapter 24 of this text lists a SAS program that carries out a multiple regression analysis for the Example 1 data using `proc iml` and matrix operations. This includes constructing the design matrix $\boldsymbol{X}$ and vector $\boldsymbol{Y}$ from the observations, estimating $\boldsymbol{\beta}$, then calculating MS_{error}, $MS_{regression}$, and $s_{\hat{\beta}_j}$. It also conducts the overall F test of the model and t tests for the regression coefficients.

21.5 Standardized regression coefficients

The regression coefficient β_j is the change in Y per unit of X_j (the slope) given the other independent variables in the model. The magnitude of β_j is affected by the strength of this relationship as well as the units of measurement for the variables. This can make it difficult to compare the relative effects of the different independent variables on Y, because their units could be quite different. Standardized regression coefficients solve this problem by expressing the slope in units of the standard deviation of Y and X_j (Kutner et al. 2005). They are calculated using the formula

$$\hat{\beta}'_j = \hat{\beta}_j \frac{s_{X_j}}{s_Y}, \tag{21.52}$$

where s_{X_j} is the sample standard deviation of X_j and s_Y is the sample standard deviation of Y. As a result of this scaling the standardized coefficients are dimensionless, similar to a correlation coefficient (Chapter 18).

21.6 R^2 values

We can define an R^2 value for multiple regression similar to one for linear regression. It is the proportion of the total sum of squares explained by the regression model, or

$$R^2 = \frac{SS_{regression}}{SS_{total}} = \frac{SS_{regression}}{SS_{regression} + SS_{error}}. \tag{21.53}$$

Large R^2 values suggest the regression model explains most of the variation (sum of squares) in the data, and vice versa for small R^2 values.

21.7 Multiple regression for Example 1 - SAS demo

We next conduct a multiple regression analysis of the Example 1 data using `proc reg` (SAS Institute Inc. 2018b). See program below. It is similar in structure to previous linear regression and ANOVA programs, but we will use `proc reg` rather than `proc glm` because it has several useful features for multiple regression. We first input the observations using a `data` step, applying transformations if necessary. Theoretical models of competition suggest a linear relationship between the log of the survival rate and measures of density like attack density and bluestain levels, so we define `y = log(survival)` in the `data` step. The two independent variables are attack density (defined as `satkden`) and bluestain levels (`blueden`).

As a first step in the analysis, it is often useful to plot the values of the dependent variables vs. the independent ones, to see their individual effects. We will use `proc gplot` (SAS Institute Inc. 2016) for this purpose, using commands similar to the ones for linear regression (Chapter 17). The program fits a regression line through the points in each graph, but these are the lines for linear, not multiple, regression. Special techniques are needed visualize the fitted model for multiple regression, which we will later examine.

The next section of the program conducts the multiple regression using `proc reg`. The `plots=diagnostics` option generates graphs that are used to examine the assumptions of multiple regression, similar to ANOVA and linear regression. The `model` statement tells SAS the multiple regression model, including the dependent variable (`y`) and the two independent variables (`satkden` and `blueden`). Note the similarity of the `model` statement to the multiple regression model with two independent variables. The option `clb` requests confidence intervals for the model parameters while `stb` displays the standardized regression coefficients. We will examine the remaining options later.

Examining the two `proc gplot` graphs, we see that log survival rate appeared to decrease with attack density, while bluestain had no obvious effect (Fig. 21.2, 21.3). The `proc reg` output contains the overall F for the multiple regression as well as separate t tests for the independent variables (Fig. 21.4). We see that the overall test was significant ($F_{2,24} = 5.45, P = 0.0112$), suggesting one or more of the independent variables affected survival. The t test for attack density was highly significant ($t_{24} = -3.30, P = 0.0030$) while bluestain was nonsignificant ($t_{24} = -0.65, P = 0.5243$). The slope or

regression coefficient for attack density was negative ($\beta = -0.2391$), indicating survival decreases with attack density, as was the coefficient for bluestain ($\beta = -0.8096$). This suggests that bluestain actually had a greater effect than attack density, but this is because their units are quite different. If we examine the standardized regression coefficients, we see that attack density had a larger coefficient ($\beta' = -0.5682$) than bluestain ($\beta' = -0.1113$), and so had a larger effect on survival. The multiple regression model explained about 31% of the variation in the data ($R^2 = 0.3122$). The usual homogeneity of variances and normality assumptions also appear satisfied (Fig. 21.5).

SAS Program

```
* SPBsurvival.sas;
title "Multiple regression for SPB survival data";
data SPB;
    input satkden blueden survival;
    * Apply transformations here;
    y = log(survival);
    datalines;
1.250 0.000 0.107
2.656 0.481 0.715
7.334 0.171 0.036
1.603 0.352 0.188
2.622 0.016 0.438

etc.

5.000 0.338 0.207
;
run;
* Print data set;
proc print data=SPB;
run;
* Plot y vs. x variables;
proc gplot data=SPB;
    plot y*(satkden blueden) / vaxis=axis1 haxis=axis1;
    symbol1 i=rl v=star c=black height=2 width=2;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
run;
* Multiple regression analysis;
proc reg plots=diagnostics data=SPB;
    * Specify regression model and request residual-residual plots;
    model y = satkden blueden / clb stb tol vif partial;
run;
quit;
```

Multiple regression for SPB survival data

Obs	satkden	blueden	survival	y
1	1.250	0.000	0.107	-2.23493
2	2.656	0.481	0.715	-0.33547
3	7.334	0.171	0.036	-3.32424
4	1.603	0.352	0.188	-1.67131
5	2.622	0.016	0.438	-0.82554
6	1.000	0.000	0.585	-0.53614
7	4.342	0.185	0.115	-2.16282
8	5.233	0.018	0.257	-1.35868
9	2.500	0.410	0.032	-3.44202
10	3.250	0.015	0.350	-1.04982

etc.

Figure 21.1: `SPBsurvival.sas - proc print`

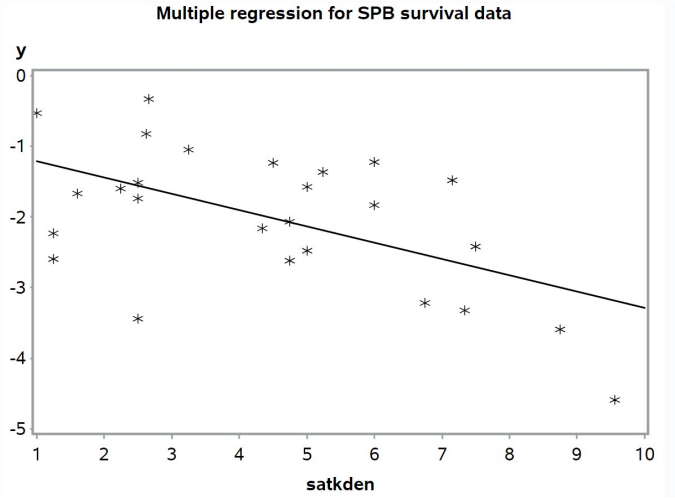


Figure 21.2: `SPBsurvival.sas - proc gplot`

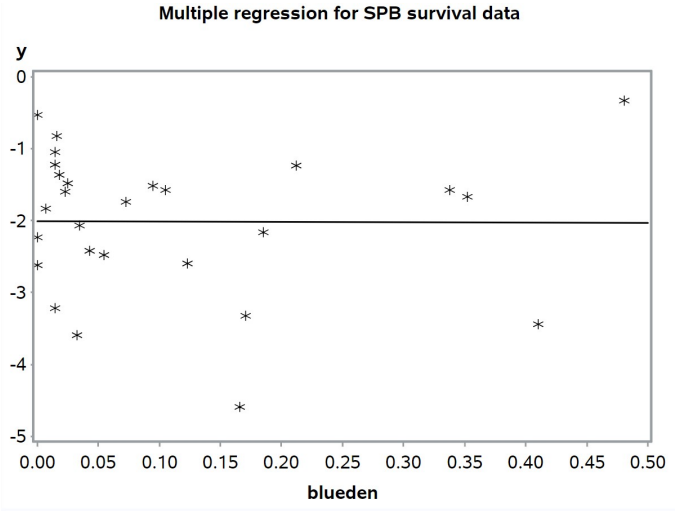
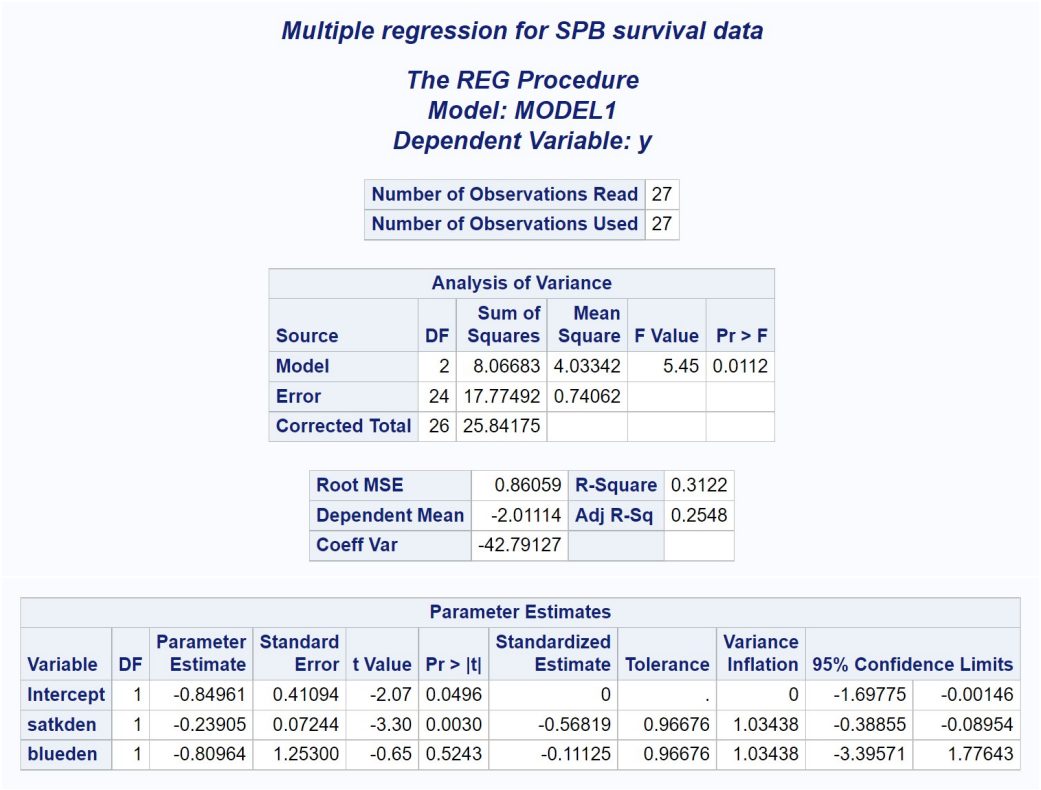


Figure 21.3: `SPBsurvival.sas - proc gplot`



Multiple regression for SPB survival data

The REG Procedure
Model: MODEL1
Dependent Variable: y

Number of Observations Read	27
Number of Observations Used	27

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	8.06683	4.03342	5.45	0.0112
Error	24	17.77492	0.74062		
Corrected Total	26	25.84175			

Root MSE	0.86059	R-Square	0.3122
Dependent Mean	-2.01114	Adj R-Sq	0.2548
Coeff Var	-42.79127		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Tolerance	Variance Inflation	95% Confidence Limits	
Intercept	1	-0.84961	0.41094	-2.07	0.0496	0	.	0	-1.69775	-0.00146
satkden	1	-0.23905	0.07244	-3.30	0.0030	-0.56819	0.96676	1.03438	-0.38855	-0.08954
blueden	1	-0.80964	1.25300	-0.65	0.5243	-0.11125	0.96676	1.03438	-3.39571	1.77643

Figure 21.4: `SPBsurvival.sas - proc reg`

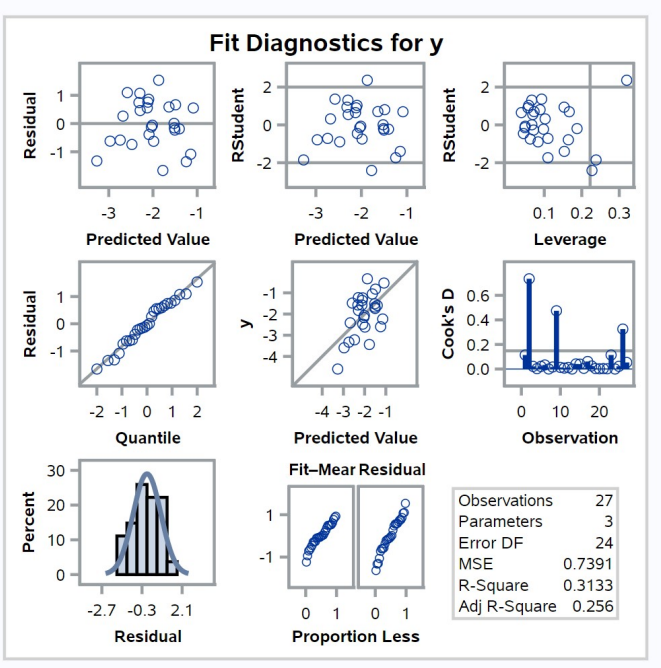


Figure 21.5: `SPBsurvival.sas - proc reg`

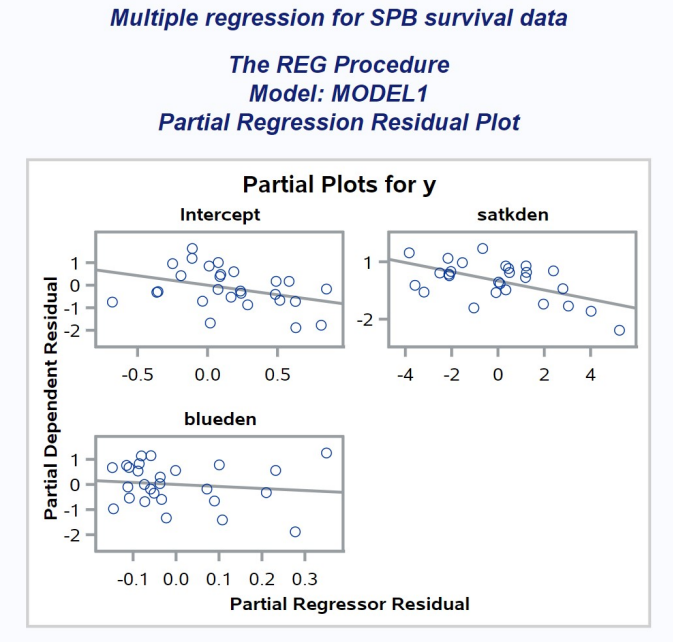


Figure 21.6: `SPBsurvival.sas - proc reg`

21.8 Visualizing the multiple regression model

We can visualize the model fitted to the Example 1 data using a three-dimensional scatter plot (Fig. 21.7). The maximum likelihood (and least squares) process minimizes the squared residuals between the observations and the plane defined by the multiple regression model. From this graph, we can see that survival decreased with increasing attack density while bluestain had a minimal effect. The slope of the plane with respect to attack density is the same as the estimated slope in Fig. 21.4, and similarly for bluestain. This kind of graph would not work for more than two independent variables, because it would have more than three dimensions.

Another type of graph that works for any number of independent variables are **residual-residual plots**, or added-variable plots (Kutner et al. 2005). As the name suggests, they are constructed using two sets of residuals. Suppose we are interested in visualizing the effect of X_1 on Y. The first set of residuals is obtained from a multiple regression of X_1 on $X_2, X_3, \ldots, X_k$, with X_1 the dependent variable. The second set of residuals is from a multiple regression of Y on $X_2, X_3, \ldots, X_k$, excluding X_1. This procedure essentially subtracts the effect of $X_2, X_3, \ldots, X_k$ on both Y and X_1. If we plot the two sets of residuals against each other, this would show the unique effect of X_1 on Y. If we were to fit a line through these residuals using linear regression, the slope of the line would be equal to $\hat{\beta}_1$ from the full multiple regression (Y vs. $X_1, X_2, \ldots, X_k$).

Residual-residual plots are requested in SAS using the `partial` option in the `model` statement for `proc reg`, generating the output in Fig. 21.6. Besides visualizing the relationships between the dependent and independent variables, these plots can be used to identify outliers and observations that strongly influence the regression lines, known as high **leverage** points (Sheather 2009). They can also be used to determine whether the relationship between Y and a given X variable is in fact linear, as assumed by the multiple regression model. Examining these plots for the Example 1 data, the relationship between survival rates and attack density (or bluestain) appeared linear and there were no large outliers.

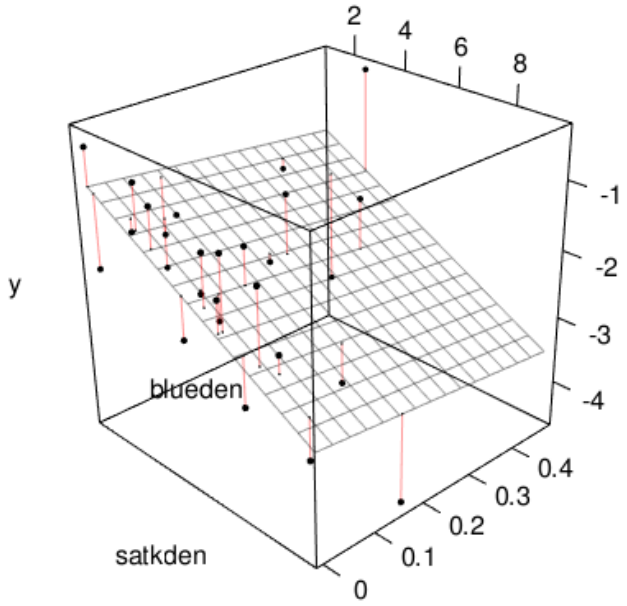


Figure 21.7: Multiple regression model fitted to the Example 1 data (see SAS program for variable definitions). The vertical red lines are the residuals for each observation $(Y_i - \hat{Y}_i)$. This plot used R code from Chang (2023).

21.9 Collinearity in multiple regression

In a multiple regression analysis, there may sometimes be strong linear relationships or correlations among two or more independent variables, a problem called **collinearity**. This can cause issues in estimating the regression coefficients, including large standard errors and confidence intervals, and potentially large values for the estimates themselves. Another symptom of collinearity are independent variables that are nonsignificant even though the overall F test is significant. See Kutner et al. (2005) and Sheather (2009) for further details.

One diagnostic tool for detecting collinearity are **tolerance values**. They are calculated as follows. Suppose we want the tolerance value for the independent variable X_1. We would run a multiple regression of X_1 on $X_2, \ldots, X_k$ and find the $R^2(X_1)$ value for this regression. The tolerance value for X_1 is defined as $1 - R^2(X_1)$. If X_1 is strongly collinear with one or more independent variables, it will have a small tolerance value because $R^2(X_1)$ will be large. Another common measure is the **variance inflation factor**, defined as $1/(1 - R^2(X_1))$. This is just the inverse of the tolerance value, and will be large if there is strong collinearity among the independent variables. A common rule of thumb is that collinearity is a problem when a variance inflation factor is sufficient large, say 5 or 10 (Kutner et al. 2005; Sheather 2009)

The tolerance and variance inflation factors are requested using the options `tol` and `vif` the `model` statement for `proc reg`. Examining these quantities for the Example 1 data set, we see the variance inflation factors were small for both independent variables (Fig. 21.4). The variance inflation factors were the same here because there were only two independent variables in the model.

21.10 Multiple regression for Example 2 - SAS demo

We now analyze the Example 2 data set using SAS and `proc reg` (see program below). Here the objective is to predict endocranial volume for fossil skulls using a multiple regression model fitted to existing species. We first log-transform all the variables in a `data` step. This makes intuitive sense, because we would expect endocranial volume to be the product of length, height, and width. After log-transform this would yield an additive model that can be

fitted using multiple regression. The dependent variable in the analysis is then `logV`, while `logL`, `logH`, and `logW` are the independent ones. Note the last two observations have missing values for endocranial volume – we will use multiple regression to predict it for these fossil skulls where endocranial volume is unavailable.

Plots generated using `proc gplot` show a strong linear relationship between `logV` and all three independent variables (Fig. 21.9-21.11). We then conduct the multiple regression using `proc reg` and the same syntax as in Example 1. Two new options for the `model` statement are `clm` and `cli`. The `clm` option generates a 95% confidence interval for the mean of Y_i for each observation, while `cli` generates a 95% prediction interval for a single Y_i (see Chapter 17). These intervals are calculated for all the observations, including the two fossil skulls. Examining the output (Fig. 21.12), we see that the overall F test was highly significant ($F_{3,190} = 8498.88, P < 0.0001$), as were the individual t tests for length ($t_{190} = 3.77, P = 0.0002$), height ($t_{190} = 9.55, P < 0.0001$), and width ($t_{190} = 14.11, P < 0.0001$). The standardized regression coefficients suggest that width had the greatest effect on endocranial volume ($\beta' = 0.5097$), followed by height ($\beta' = 0.3873$) and then width ($\beta' = 0.1052$). Combined, these three variables explained 99.3% of the variation in volume ($R^2 = 0.9926$), suggesting the model would be useful for prediction. The confidence and prediction intervals for the two fossil skulls are shown at the bottom of Fig. 21.13.

A possible concern with this analysis were large variance inflation factors for all three independent variables (Fig. 21.12). Despite these large values, the individual t tests for these variables were all highly significant, suggesting they each contribute something unique to the model. Kutner et al. (2005) also argue that collinearity is less important when prediction is primary goal of the analysis, as in the Example 2 regression.

SAS Program

```
* Endocranial4.sas;
title "Multiple regression for endocranial volume in mammals";
data ECVdat;
    input Length Width Height Volume Common_name :$30.;
    * Apply transformations here;
    logV = log(Volume);
    logL = log(Length);
    logH = log(Height);
    logW = log(Width);
    datalines;
15.04   11.29   6.61    0.38    Pygmy_glider
52.40   30.94   25.68   12.36   Rufous_kangaroo_rat
75.87   52.79   39.45   56.70   Howler_monkey
41.73   25.70   16.79   5.68    Scaley-tailed_squirrel
39.71   26.87   17.13   5.92    Lord_derby's_flying_squirrel

etc.

70.36   45.09   37.72   38.43   Arctic_fox
80.73   47.96   39.45   48.55   Fox
13.54   9.24    7.13    0.36    Meadow_jumping_mouse
13.15   9.05    7.00     .      Fossil_mouse
190.17  97.32   80.31    .      Fossil_bear
;
run;
* Print data set;
proc print data=ECVdat;
run;
* Plot y vs. x variables;
proc gplot data=ECVdat;
    plot logV*(logL logH logW) / vaxis=axis1 haxis=axis1;
    symbol1 i=rl v=star c=black height=2 width=2;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
run;
* Multiple regression;
proc reg plots=diagnostics data=ECVdat;
    * Specify variables in regression model;
    model logV = logL logH logW / clb stb tol vif partial clm cli;
run;
quit;
```

Multiple regression for endocranial volume in mammals

Obs	Length	Width	Height	Volume	Common_name	logV	logL	logH	logW
1	15.04	11.29	6.61	0.38	Pygmy_glider	-0.96758	2.71071	1.88858	2.42392
2	52.40	30.94	25.68	12.36	Rufous_kangaroo_rat	2.51447	3.95891	3.24571	3.43205
3	75.87	52.79	39.45	56.70	Howler_monkey	4.03777	4.32902	3.67503	3.96632
4	41.73	25.70	16.79	5.68	Scaley-tailed_squirrel	1.73695	3.73122	2.82078	3.24649
5	39.71	26.87	17.13	5.92	Lord_derby's_flying_squirrel	1.77834	3.68160	2.84083	3.29101
6	18.90	12.62	7.61	0.51	Yellow-footed_antechinus	-0.67334	2.93916	2.02946	2.53528
7	15.10	11.69	7.06	0.46	Brown_antechinus	-0.77653	2.71469	1.95445	2.45873
8	123.70	73.89	63.93	150.53	Pronghorn	5.01416	4.81786	4.15779	4.30258
9	46.75	28.70	18.45	6.51	Mountain_beaver	1.87334	3.84481	2.91506	3.35690
10	154.32	103.77	71.95	284.03	Antarctic_fur_seal	5.64908	5.03903	4.27597	4.64218

etc.

Figure 21.8: `Endocranial4.sas - proc print`

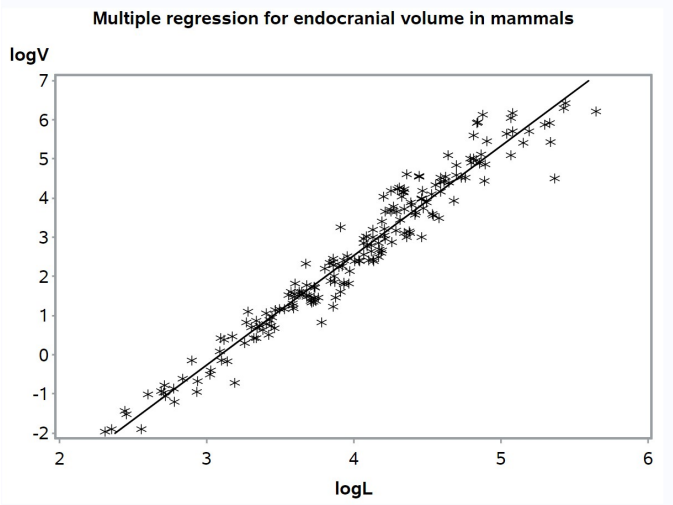


Figure 21.9: `Endocranial4.sas - proc gplot`

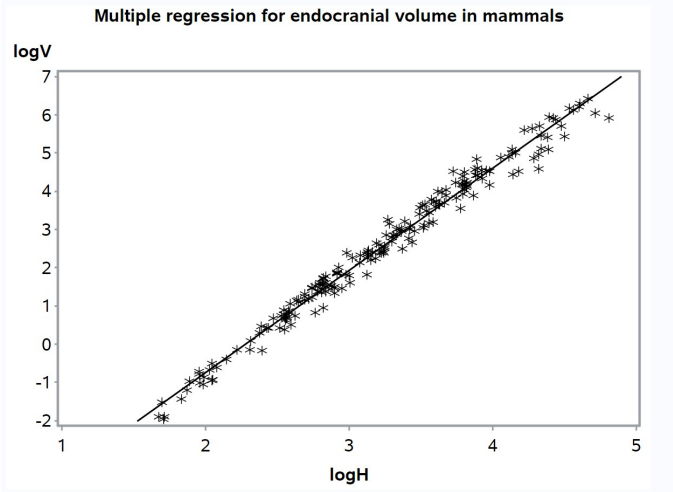


Figure 21.10: `Endocranial4.sas - proc gplot`

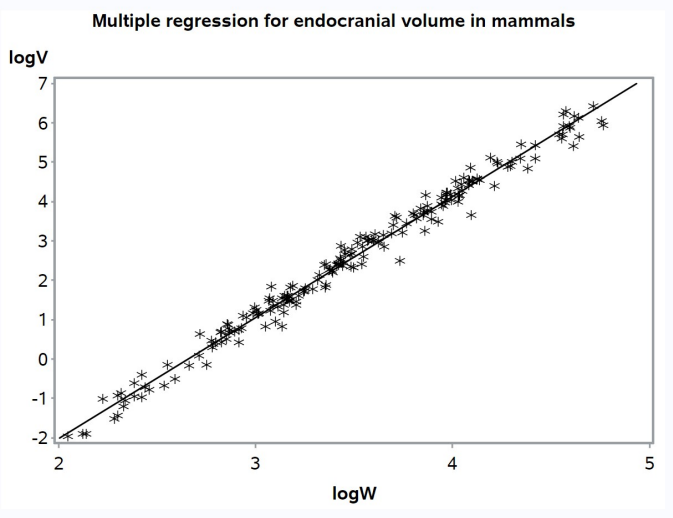
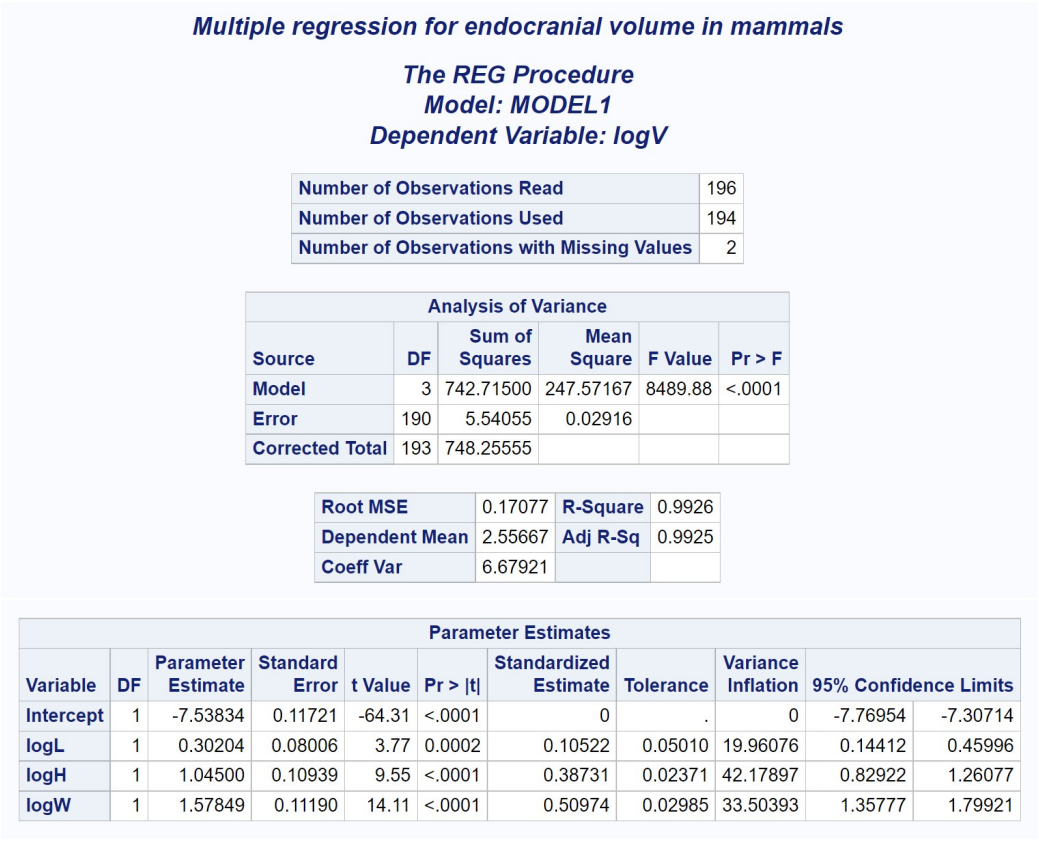


Figure 21.11: `Endocranial4.sas - proc gplot`



Multiple regression for endocranial volume in mammals

The REG Procedure
Model: MODEL1
Dependent Variable: logV

Number of Observations Read	196
Number of Observations Used	194
Number of Observations with Missing Values	2

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	742.71500	247.57167	8489.88	<.0001
Error	190	5.54055	0.02916		
Corrected Total	193	748.25555			

Root MSE	0.17077	R-Square	0.9926
Dependent Mean	2.55667	Adj R-Sq	0.9925
Coeff Var	6.67921		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Tolerance	Variance Inflation	95% Confidence Limits	
Intercept	1	-7.53834	0.11721	-64.31	<.0001	0	.	0	-7.76954	-7.30714
logL	1	0.30204	0.08006	3.77	0.0002	0.10522	0.05010	19.96076	0.14412	0.45996
logH	1	1.04500	0.10939	9.55	<.0001	0.38731	0.02371	42.17897	0.82922	1.26077
logW	1	1.57849	0.11190	14.11	<.0001	0.50974	0.02985	33.50393	1.35777	1.79921

Figure 21.12: `Endocranial4.sas - proc reg`

Multiple regression for endocranial volume in mammals

The REG Procedure
Model: MODEL1
Dependent Variable: logV

Output Statistics								
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	-0.9676	-0.9199	0.0286	-0.9763	-0.8635	-1.2614	-0.5784	-0.0477
2	2.5145	2.4666	0.0144	2.4381	2.4951	2.1286	2.8047	0.0478
3	4.0378	3.8704	0.0202	3.8306	3.9102	3.5312	4.2096	0.1674
4	1.7370	1.6609	0.0198	1.6219	1.6999	1.3218	2.0000	0.0760
5	1.7783	1.7371	0.0209	1.6959	1.7784	1.3978	2.0765	0.0412
6	-0.6733	-0.5279	0.0254	-0.5780	-0.4778	-0.8684	-0.1873	-0.1455
7	-0.7765	-0.7949	0.0283	-0.8507	-0.7391	-1.1363	-0.4535	0.0184
8	5.0142	5.0533	0.0202	5.0134	5.0932	4.7141	5.3925	-0.0391
9	1.8733	1.9680	0.0220	1.9247	2.0113	1.6284	2.3076	-0.0947
10	5.6491	5.7797	0.0359	5.7089	5.8504	5.4355	6.1238	-0.1306
				etc.				
193	3.8826	3.7377	0.0144	3.7092	3.7661	3.3997	4.0757	0.1449
194	-1.0217	-1.1888	0.0346	-1.2570	-1.1206	-1.5325	-0.8451	0.1671
195	.	-1.2496	0.0355	-1.3196	-1.1797	-1.5937	-0.9056	.
196	.	5.8563	0.0286	5.7999	5.9127	5.5148	6.1979	.

Figure 21.13: `Endocranial4.sas - proc reg`

21.11 Power analysis for multiple regression

The appropriate sample size for a multiple regression study can be determined through a power analysis. Similar to power analysis in ANOVA, we must specify the Type I error rate α, the desired power level, and the size of the effect we wish to detect. The effect size in power analyses for multiple regression is often expressed in terms of an R^2 value, which combines the effects of the independent variables (through $SS_{regression}$) and the variability of the observations (SS_{error}).

The SAS procedure `power` can do a power analysis for multiple regression using the `multreg` option. We first specify the Type I error rate and desired power using the `alpha` and `power` options (see SAS program below). We will be interested in the sample sizes needed for the overall F test of $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$, which is equivalent to testing $H_0 : R^2 = 0$. This value of R^2 is specified using the `rquaredreduced` option. The values of R^2 under the alternative hypothesis ($H_1 : \beta_j \neq 0$ for some j) are then specified using the `rsquarefull` option. Some plausible values for ecological or behavioral data are 0.1, 0.3, and 0.6, but any value can be used. We must also specify the number of independent variables (k) under H_0 and H_1, using the `nreducedpredictors` and `nfullpredictors` options. We set the `ntotal` option to a missing value, which tells `power` to solve for the sample size n that gives the desired power.

SAS Program

```
* multreg_power.sas;
title 'Power Analysis for Multiple Regression';
proc power;
    multreg
    model = fixed
    alpha = 0.05
    power = 0.8
    rsquarereduced = 0
    rsquarefull = 0.1 0.3 0.6
    nreducedpredictors = 0
    nfullpredictors = 1 2 3 4 5 6 7 8 9 10 20 30 40 50
    ntotal = . ;
run;
quit;
```

Table 21.4 summarizes the result of this analysis, with the entries the sample size n to obtain the desired power. Note that the effect size (R^2 under H_1) strongly influences sample size, and that more observations are necessary to maintain power as the number of independent variables (k) is increased. For one predictor, the sample size specified is for a simple linear regression.

The `power` procedure can be used to find the sample size for other scenarios, including tests of the individual regression coefficients ($H_0 : \beta_j = 0$). The basic idea is to specify an R^2 value with and without X_j in the model, with the number of predictors in the full and reduced model differing by 1.

Table 21.4: Power for Multiple Regression - Effect of R^2 and the number of independent variables (k) on the sample size n for the overall F test of the model ($H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$). See text for further details.

	R^2		
k	0.1	0.3	0.6
1	73	21	8
2	90	26	11
3	103	30	12
4	113	33	14
5	122	36	16
6	130	39	17
7	137	42	18
8	144	44	20
9	150	46	21
10	156	48	22
20	205	67	34
30	244	82	45
40	278	97	55
50	308	110	66

21.12 Polynomial regression

In a linear regression, we sometimes saw a nonlinear relationship between Y and X for some data sets (Chapter 17). This problem could often be fixed by applying a transformation to Y or X , but this approach sometimes fails. An alternative solution is to fit a flexible polynomial in X to the data. The observations would be modeled using the equation

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + \ldots + \epsilon_i. \tag{21.54}$$

This is a **polynomial regression** model. It is similar in structure to multiple regression, except the independent variables $X_1, X_2, \ldots, X_k$ are replaced with increasing powers of X.

As we add more powers of X, the polynomial regression model becomes increasingly flexible. A model using only X and X^2 would fit a quadratic polynomial (a parabola) to the data, while one with X, X^2, and X^3 would fit a cubic one, which is S-shaped. While higher powers of X would allow even more flexibility, they are seldom needed to obtain an adequate fit. Another issue is extrapolation beyond the range of X values, where higher order polynomials can generate unrealistic estimates (Kutner et al. 2005). For these reasons, it is desirable to find the lowest order polynomial that adequately describes the data.

One issue with using the powers of X in a regression is that they are collinear with one another. For example, we would expect X, X^2, and X^3 to be strongly correlated. A common strategy is to use **centered polynomials** to reduce this collinearity. This is accomplished by centering the independent variable around its mean before finding the power. In particular, we first define $x = X - \bar{X}$ and then raise x to the desired power, using these centered variables in the polynomial regression.

21.13 Population growth experiment - SAS demo

As an example of polynomial regression, we will analyze data from a hypothetical experiment on a stored grain insect, where varying numbers of adult insects (N) are added to a container with grain, and then the number of offspring per adult estimated (R). We would expect that R would decrease

as N was increased because of intraspecific competition among the insects. The Ricker model is often used as a simple description of intraspecific competition and could be suitable for these data (Ricker 1954). The model has two parameters, the intrinsic growth rate r of the organism and its carrying capacity K. For this model, we would expect the following relationship between $\log R$ and N:

$$\log R = r(1 - (N/K)) = r - (r/K)N = \alpha - \beta N, \tag{21.55}$$

where $\alpha = r$ and $\beta = r/K$. This is essentially a linear regression model for $\log R$ vs. N. What we would like to determine is whether this model is adequate, or whether a more complex nonlinear one is needed. We can answer this question using a polynomial regression model with different powers of N. If the tests for these terms are significant, it suggests a more complex model is needed for these observations.

The SAS program below lists the observations from this hypothetical experiment in a `data` step. Also listed is the mean of value of N (`nbar = 50.455`). This is used in the centering process, which first calculates a centered density `x` and then the powers of `x` (`x2`, `x3`). The data are then plotted along with a smooth line using `proc gplot` and the `symbol1 i=sm70` option. The smooth line is constructed using cubic splines, which are themselves a kind of polynomial. This graph helps visualize the relationship between `logR` and `n`.

We then use `proc glm` to conduct the polynomial regression (SAS Institute Inc. 2018b). We use this procedure rather than `proc reg` because it can generate Type I sums of squares and tests. These are produced by sequentially fitting the different terms in the `model` statement, and can be used to determine the lowest order polynomial needed to describe the data. For example, the Type I test for `x3` tests whether this power is needed with `x` and `x2` already in the model.

The results from `proc glm` output suggested a quadratic polynomial provides an adequate description of these data (see discussion below). The remainder of the program plots the observations with a quadratic polynomial line plus a confidence interval (`proc gplot` with the `symbol` option `i=rqclm`). It then uses `proc reg` to finish the analysis, using syntax similar to previous multiple regression analyses.

SAS Program

```
* Ricker_polynomial.sas;
title "Polynomial regression for Ricker data";
data ricker;
    input n logR;
    * For centered polynomials, you'll need the mean X value;
    nbar = 50.455;
    x = n-nbar;
    x2 = x**2;
    x3 = x**3;
    datalines;
 5  0.42
10  0.33
20  0.48
30  0.03
40 -0.18
50 -0.16
60  0.08
70 -1.20
80 -1.45
90 -1.72
100 -2.67
;
run;
* Print data set;
proc print data=ricker;
run;
* Plot data and fit smooth line;
proc gplot data=ricker;
    plot logR*n / vaxis=axis1 haxis=axis1;
    symbol1 i=sm70 v=star c=black height=2 width=2;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
run;
* Polynomial regression;
proc glm data=ricker;
    * Look at Type I tests to determine order of polynomial;
    model logR = x x2 x3;
run;
* Preceding analysis suggests second-order polynomial adequate;
* Plot the data and second-order polynomial;
proc gplot data=ricker;
    plot logR*n / vaxis=axis1 haxis=axis1;
    symbol1 i=rqclm v=star c=black height=2 width=2;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
```

```
run;
* Polynomial regression with second-order polynomial;
proc reg data=ricker;
    model logR = x x2 / clb stb tol vif partial;
run;
quit;
```

Examining the first `proc gplot` graph, we observe `logR` decreased with density `n`, suggesting that reproduction was affected by intraspecific competition. The relationship appears curved, however, and so the Ricker model may not be adequate (Fig. 21.15). The Type I tests from `proc glm` yielded highly significant results for `x` ($F_{1,7} = 109.81, P < 0.0001$) and `x2` ($F_{1,7} = 12.45, P = 0.0096$), but a nonsignificant one for `x3` ($F_{1,7} = 0.43, P = 0.5328$) (Fig. 21.16). This pattern suggests a quadratic polynomial would be sufficient to describe these data. In addition, the highly significant test for `x2` means we can definitively reject the linear Ricker model.

The second `proc gplot` graph shows that a quadratic provides a reasonable approximation to the observations (Fig. 21.17). Examining the `proc reg` output (Fig. 21.18), we see that overall F test was highly significant ($F_{3,8} = 40.89, P < 0.0001$), as were the individual tests for `x` ($t_8 = -10.59, P < 0.0001$) and `x2` ($t_8 = -3.66, P = 0.0064$). The polynomial regression model explained about 94% of the variation in the data ($R^2 = 0.943$). Due to centering, the variance inflation factors show no collinearity issues with `x` and `x2`.

Polynomial regression for Ricker data

Obs	n	logR	nbar	x	x2	x3
1	5	0.42	50.455	-45.455	2066.16	-93917.17
2	10	0.33	50.455	-40.455	1636.61	-66208.94
3	20	0.48	50.455	-30.455	927.51	-28247.23
4	30	0.03	50.455	-20.455	418.41	-8558.52
5	40	-0.18	50.455	-10.455	109.31	-1142.80
6	50	-0.16	50.455	-0.455	0.21	-0.09
7	60	0.08	50.455	9.545	91.11	869.62
8	70	-1.20	50.455	19.545	382.01	7466.33
9	80	-1.45	50.455	29.545	872.91	25790.04
10	90	-1.72	50.455	39.545	1563.81	61840.75
11	100	-2.67	50.455	49.545	2454.71	121618.46

Figure 21.14: `Ricker_polynomial.sas - proc print`

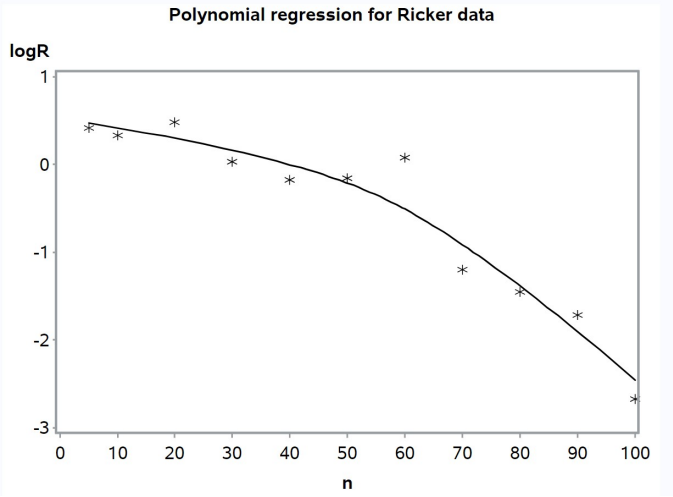


Figure 21.15: `Ricker_polynomial.sas - proc gplot (1)`

Polynomial regression for Ricker data

The GLM Procedure

Dependent Variable: logR

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	10.30624576	3.43541525	40.89	<.0001
Error	7	0.58804515	0.08400645		
Corrected Total	10	10.89429091			

R-Square	Coeff Var	Root MSE	logR Mean
0.946023	-52.78519	0.289839	-0.549091

Source	DF	Type I SS	Mean Square	F Value	Pr > F
x	1	9.22443700	9.22443700	109.81	<.0001
x2	1	1.04566094	1.04566094	12.45	0.0096
x3	1	0.03614783	0.03614783	0.43	0.5328

Figure 21.16: `Ricker_polynomial.sas - proc glm`

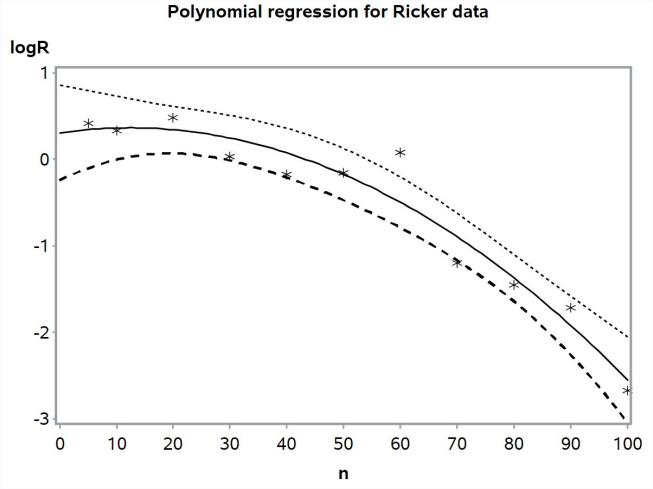


Figure 21.17: `Ricker_polynomial.sas - proc gplot (2)`

Polynomial regression for Ricker data

The REG Procedure
Model: MODEL1
Dependent Variable: logR

Number of Observations Read	11
Number of Observations Used	11

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	10.27010	5.13505	65.81	<.0001
Error	8	0.62419	0.07802		
Corrected Total	10	10.89429			

Root MSE	0.27933	R-Square	0.9427
Dependent Mean	-0.54909	Adj R-Sq	0.9284
Coeff Var	-50.87099		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Tolerance	Variance Inflation	95% Confidence Limits	
Intercept	1	-0.18655	0.13000	-1.43	0.1892	0	.	0	-0.48634	0.11324
x	1	-0.02890	0.00273	-10.59	<.0001	-0.89833	0.99505	1.00497	-0.03520	-0.02261
x2	1	-0.00037900	0.00010353	-3.66	0.0064	-0.31058	0.99505	1.00497	-0.00061774	-0.00014026

Figure 21.18: `Ricker_polynomial.sas - proc reg`

21.14 Model selection using information criteria

There is a substantial literature on problems with hypothesis testing and P values in scientific research, as well as defenses of this approach. Recent papers that summarize these issues include Aho et al. (2014), Burnham and Anderson (2014), Murtaugh (2014), and de Valpine (2014), in an ecological context. The most common alternative to hypothesis testing is model selection using Akaike's Information Criterion, or AIC (Akaike 1974; Anderson et al. 2000; Burnham and Anderson 2002; Burnham and Anderson 2014). The basic idea is to formulate a collection of models to describe the data, and then choose the best one based on AIC values, defined as the model with the smallest AIC. For example, in a multiple regression setting we might be interested in determining the best model among different subsets of the independent variables. There is no explicit hypothesis testing in this approach, but confidence intervals can be calculated to describe the magnitude of an effect.

While the hypothesis testing and AIC approaches seem different, they often use similar statistical models with the same sets of assumptions. There is also a common scenario, nested models, where the two approaches would produce similar results. Models are nested when a simpler model is a special case of a more complex one, with fewer parameters or variables. Procedures like ANOVA and multiple regression utilize nested models, with the tests constructed using a simpler H_0 model nested within a more complex H_1 model. Murtaugh (2014) showed there is a direct relationship between P values and changes in AIC under these conditions. Suppose that a test comparing H_0 and H_1 was highly significant, favoring the H_1 model. The AIC value for the H_1 model would also be substantially smaller than H_0, and so this approach would also select the H_1 model. However, it is important to note there are scenarios where the models are not nested, which precludes hypothesis testing and P values but where AIC is useful. For example, Burnham & Anderson (2002) used AIC to compare nine different nonlinear models of the relationship between the number of bird species and sample size, with the models of such different forms they could not be nested.

So what is AIC? The AIC uses the concept of **Kullback-Leibler information**. We suppose that the data have some probability distribution f, and we would like to approximate it with another distribution g. These

two distributions can be thought of as different models for the data, with f the true one. Kullback-Leibler information is a measure of the distance between f and g, denoted as $I(f, g)$. In mathematical terms, it is defined as the expected value of $\ln(f/g)$, where the expected value is calculated using the f distribution:

$$I(f, g) = E_f \left[\ln \left(\frac{f(x)}{g(x|\boldsymbol{\theta})} \right) \right]. \tag{21.56}$$

(Akaike 1974; Anderson et al. 2000; Burnham and Anderson 2002). The notation $g(x|\boldsymbol{\theta})$ is used to emphasize that g has a number of parameters (say θ_1, θ_2, etc.) that could affect $I(f, g)$. $I(f, g)$ is always positive unless $f = g$, for which $I(f, g) = 0$. Because the true distribution f and the parameters of g are typically unknown, $I(f, g)$ is not useful in this form because it cannot be calculated.

To see how $I(f, g)$ behaves, suppose that f and g are simple continuous distributions like the normal. Equation 21.56 can then be expressed as an integral of the form

$$I(f, g) = \int f(x) \ln \left(\frac{f(x)}{g(x|\boldsymbol{\theta})} \right) dx. \tag{21.57}$$

If f and g are quite distinct from each other $I(f, g)$ will be large, because positive values of $\ln(f/g)$ will mostly coincide with f, and so receive more weight in the integral (Fig. 21.19). This effect is diminished when f and g are closely overlapping. One can think of $I(f, g)$ as measuring the mismatch between the two distributions, or more formally as the loss of information when approximating the true distribution f using $g(x|\boldsymbol{\theta})$.

We can break the expected value in Equation 21.56 into two pieces, using the fact that $\ln(a/b) = \ln a - \ln b$ and formulas for the expected value of a sum (see Chapter 7). We have

$$I(f, g) = E_f[\ln f(x)] - E_f[\ln g(x|\boldsymbol{\theta})]. \tag{21.58}$$

The first term in this equation does not involve g, and in any event would be a constant because f is fixed. This suggests that to minimize $I(f, g)$, we should compare the relative values of the second term. It can be shown that smaller values of $-E_f[\ln g(x|\boldsymbol{\theta})]$ would make $I(f, g)$ smaller, minimizing the loss of information (Burnham and Anderson 2002).

The contribution of Akaike (1974) was to find an estimator of $-E_f[\ln g(x|\boldsymbol{\theta})]$ using maximum likelihood, which also provides estimates of the parameters

of g. Suppose we have a data set that could be used to estimate $\boldsymbol{\theta}$ using maximum likelihood (see Chapter 8). He showed that $-E_f[\ln g(x|\boldsymbol{\theta})]$ could be estimated using

$$AIC = -2 \ln L(\hat{\boldsymbol{\theta}}) + 2K, \tag{21.59}$$

where $L(\hat{\boldsymbol{\theta}})$ is the likelihood function for g at the maximum likelihood estimate of $\boldsymbol{\theta}$, by definition the largest value of L (Akaike 1974; Anderson et al. 2000; Burnham and Anderson 2002). Here K is the number of parameters in $\boldsymbol{\theta}$. An interesting feature of the AIC is that K is actually a bias correction for this estimate.

Now suppose we have a number of different g distributions that are models for our data, with different numbers of parameters. Models with the smallest value of AIC would also have the smallest $I(f, g)$, and so the smallest loss of information in approximating f by g. We can gain further insight into this process by examining the two terms in the AIC formula. Models with more parameters could potentially fit the data better, generating a larger L and so smaller $-2 \ln L$, but they would also have larger values of $2K$. Thus, the AIC imposes a tradeoff between the fit of the model and its complexity.

In multiple regression, ANOVA, and other general linear models, $-2 \ln L$ and so AIC are a function of SS_{error} and the number of parameters in the model. In particular, for models of this type we have

$$AIC = n \ln(SS_{error}/n) + 2K \tag{21.60}$$

where n is sample size. We can see from this expression that better models will tend to have smaller values of SS_{error} and also fewer parameters, for a given sample size. Note that different software packages may count K and calculate AIC in different ways, so that the values of reported are different. These differences, while confusing, have no effect on the relative ranking of models by AIC.

A quantity related to AIC is the Bayesian Information Criterion or BIC (Schwarz 1978). The BIC was derived using the Bayesian interpretation of probability as a belief, but is valid outside this framework. BIC is calculated using the formula

$$BIC = -2 \ln L(\hat{\theta}) + \ln(n)K \tag{21.61}$$

where as before n is sample size and K the number of parameters. The only difference between the formulas for AIC and BIC is the multiplier for K – it is a constant (2) for AIC but $\ln(n)$ for BIC. In terms of regression and

ANOVA models, BIC can be calculated using the formula

$$BIC = n \ln(SS_{error}/n) + \ln(n)K \tag{21.62}$$

The BIC is used in the same fashion as AIC, with smaller values indicating a better model. It is clear from this formula that BIC penalizes complex models more heavily as sample size increases, because of the $\ln(n)$ multiplier.

Which criterion, AIC vs. BIC, performs best in model selection? Brewer et al. (2016) compared the two methods using simulated data intended to mimic the hidden heterogeneity likely present in real data sets, where the data could be mixture of observations with different parameter values. Performance was measured by how well the selected models predicted the observations of similar data sets, separate from the ones used in model selection. This tests how well the predictions of the model generalize to new observations. When heterogeneity was low AIC generally performed best, but BIC was better when heterogenity was large, so there was no clear winner.

We will use a more complex data set to illustrate model selection using AIC. Kaul and Wilsey (2020) wanted to determine which factors affect the success of tallgrass prairie restorations located in Iowa, USA. These prairies were restored using seed mixes, and as one measure of success they compared the species diversity of the seed mix with the diversity at the restored site, using the Bray-Curtis dissimilarity index as the dependent variable. This index ranges from 0 (all species shared) to 1 (none in common), so larger values suggest the restoration has failed. The independent variables were the age of the site and its linearity (shape), soil pH and organic matter, temperature and precipitation at establishment as well as annual averages, and exotic species abundance. A subset of these observations is shown in Table 21.5 (see https://datadryad.org for the full data set).

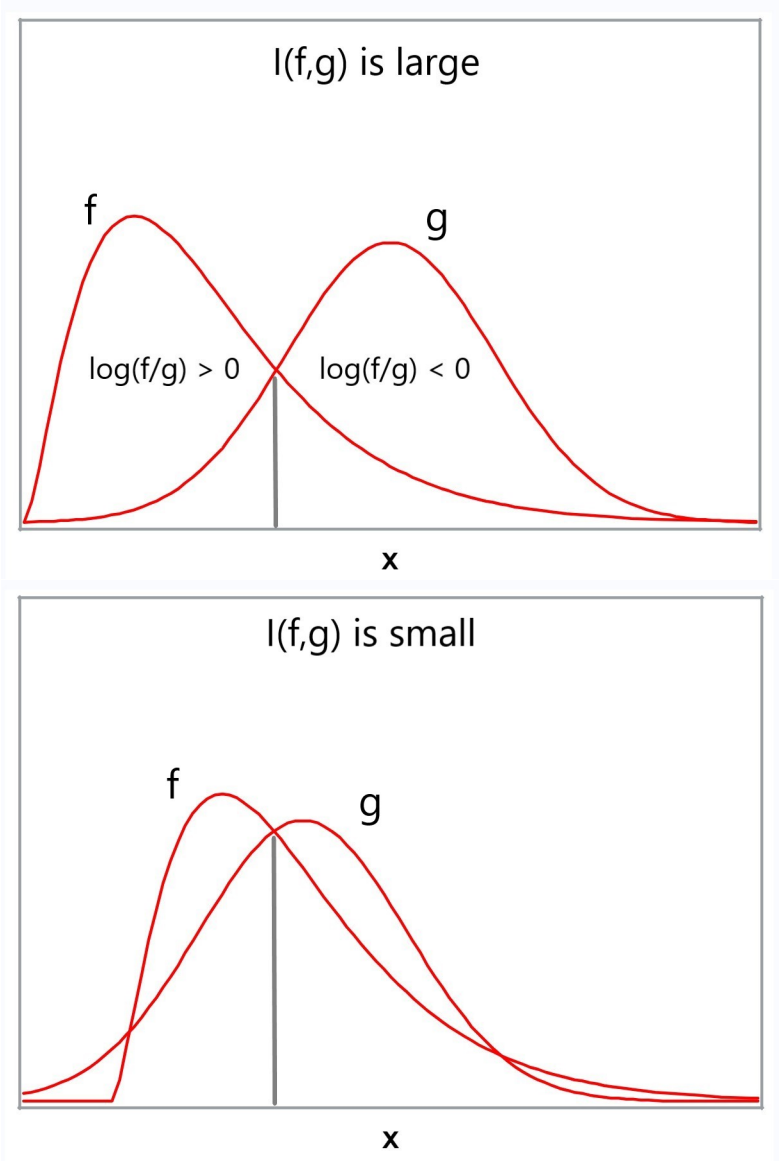


Figure 21.19: Graphical illustration of $I(f, g)$ under two scenarios.

Table 21.5: Example 3 - Site variables for restored prairies and Bray-Curtis dissimilarity (Kaul and Wilsey 2020). Here TE = temperature at establishment, PE = precipitation at establishment, TA = average annual temperature, and PA = average annual precipitation. See text for further details.

Site	Age	Linearity	pH	Organic	TE	PE	TA	PA	Exotic	Bray-Curtis	i
3	5	1.29	7.33	12.28	12.25	31.77	9.94	35.61	24.05	0.982	1
4	6	1.39	7.93	8.54	8.69	45.50	8.89	36.80	19.87	0.898	2
5	14	1.21	7.90	7.11	9.31	31.83	9.11	35.52	3.68	0.791	3
7	5	1.24	8.03	6.28	9.42	28.26	8.00	36.48	10.25	1.000	4
8	3	1.28	7.67	5.69	6.61	43.13	8.00	36.48	18.00	0.998	5
etc.											
100	17	1.30	7.87	10.42	9.17	33.58	10.72	37.59	25.03	0.970	40
101	3	1.26	7.93	8.64	8.56	40.66	10.72	37.59	15.37	0.772	41
102	11	1.25	8.10	3.61	11.69	40.36	10.00	36.28	9.57	0.972	42
105	13	1.05	7.93	14.54	8.53	35.57	8.00	36.48	13.77	0.718	43
106	10	1.02	6.97	8.19	7.64	47.79	8.00	36.48	14.25	0.624	44

21.15 Model selection for Example 3 - SAS demo

We will use AIC to select the best model for the Example 3 data, with multiple regression the underlying model (see program below). We first input the observations using a `data` step, selecting the Bray-Curtis index (`bc`) as the dependent variable `y`. We then plot `y` vs. all the independent variables (`age`, `linear`, ..., `exotic`) using `proc gplot`.

The next section of the program conducts a standard multiple regression using `proc reg`. We will later compare the results of this analysis with that generated by `proc glmselect`, a SAS procedure that implements various types of model selection (SAS Institute Inc. 2018b). The `model` statement for `proc glmselect` is similar to `proc reg`, but with a `class` statement it can also accomodate ANOVA-like factors. Model selection using AIC is implemented using the `selection=stepwise(select=AICC)` option. Stepwise refers to the search method, with the procedure adding or dropping individual variables until it finds the best model. The option `AICC` requests a version of AIC corrected for small sample sizes. Model selection using BIC could be requested using the `select=SBC` option (Schwarz's Bayesian Criterion or BIC).

Examining a subset of `proc gplot` graphs, we see that the Bray-Curtis dissimilarity index (`y`) increased with the linearity of the site (`linear`) and exotic species abundance (`exotic`), and decreased with precipitation during establishment (`PE`) (Fig. 21.21-21.23). From the `proc reg` output (Fig. 21.24), we see that the overall model was highly significant ($F_{9,34} = 11.11, P < 0.0001$) as were the individual tests for linearity ($t_{34} = 3.43, P = 0.0016$), exotic abundance ($t_{34} = 4.19, P = 0.0002$), and precipitation during establishment ($t_{34} = -3.07, P = 0.0042$). These variables also had the largest standardized regression coefficients. No other variables approached significance.

Model selection using AIC and `proc glmselect` chose linearity, exotic abundance, and precipitation during establishment for the best model (Fig. 21.25). These were the same variables that were significant in the multiple regression. Kaul and Wilsey (2020) found these same three variables in their model search using stepwise regression, a method of model selection where variables are added or removed based on repeated tests at some α level ($\alpha = 0.15$ in this case). The different model selection methods all yielded the same result, suggesting it is a robust one. These authors conclude that

high exotic species abundance interfered with the restoration process, so that the restored site shared fewer species with the seed mix used to restore it. Linearity also affected restoration, likely because highly linear sites had more edges for exotic species to invade. Presumably precipitation during establishment aided the initial success of the seed mix, and so had the opposite effect.

Note that `proc glmselect` does not provide P values for the t tests of the independent variables, because they would not be valid in this context. The Type I error rates for these tests assume a single multiple regression analysis, not a selection process where many different models were considered.

SAS Program

```
* Restored6.sas;
title "Model selection for restored prairie data";
data RPdat;
    input site_id $ age linear ph organic TE PE TA PA exotic bc;
    * Bray-Curtis (bc) measures dissimilarity of the site vs.
    restoration seed mix;
    * 0 = all species in common, 1 = none in common;
    * Kaul and Wilsey (2020) say similarity in paper;
    * Apply transformations here;
    y = bc;
    datalines;
3       5       1.29   7.33 12.28 12.25   31.77  9.94    35.61   24.05 0.982
4       6       1.39   7.93 8.54  8.69    45.50  8.89    36.80   19.87 0.898
5       14      1.21   7.90 7.11  9.31    31.83  9.11    35.52   3.68  0.791
7       5       1.24   8.03 6.28  9.42    28.26  8.00    36.48   10.25 1.000
8       3       1.28   7.67 5.69  6.61    43.13  8.00    36.48   18.00 0.998

etc.

100     17      1.30   7.87 10.42 9.17    33.58  10.72   37.59   25.03 0.970
101     3       1.26   7.93 8.64  8.56    40.66  10.72   37.59   15.37 0.772
102     11      1.25   8.10 3.61  11.69   40.36  10.00   36.28   9.57  0.972
105     13      1.05   7.93 14.54 8.53    35.57  8.00    36.48   13.77 0.718
106     10      1.02   6.97 8.19  7.64    47.79  8.00    36.48   14.25 0.624
;
run;
* Print data set;
proc print data=RPdat;
run;
* Plot y vs. x variables;
proc gplot data=RPdat;
```

```
    plot y*(age linear ph organic TE PE TA PA exotic) / vaxis=axis1
haxis=axis1;
    symbol1 i=rl v=star c=black height=2 width=2;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
    symbol1 i=rl v=star c=black;
run;
* Multiple regression;
proc reg data=RPdat;
    * Specify variables in regression model;
    model y = age linear ph organic TE PE TA PA exotic / clb stb tol vif partial;
run;
* Model selection using AICc (stepwise);
proc glmselect data=RPdat;
    * Specify variables in regression model and method of selection;
    model y = age linear ph organic TE PE TA PA exotic /
    selection=stepwise(select=AICC);
run;
quit;
```

Model selection for restored prairie data

Obs	site_id	age	linear	ph	organic	TE	PE	TA	PA	exotic	bc	y
1	3	5	1.29	7.33	12.28	12.25	31.77	9.94	35.61	24.05	0.982	0.982
2	4	6	1.39	7.93	8.54	8.69	45.50	8.89	36.80	19.87	0.898	0.898
3	5	14	1.21	7.90	7.11	9.31	31.83	9.11	35.52	3.68	0.791	0.791
4	7	5	1.24	8.03	6.28	9.42	28.26	8.00	36.48	10.25	1.000	1.000
5	8	3	1.28	7.67	5.69	6.61	43.13	8.00	36.48	18.00	0.998	0.998
6	9	10	1.12	7.97	7.92	8.97	41.19	8.00	36.48	4.50	0.712	0.712
7	10	8	1.04	8.10	11.04	7.19	28.99	7.22	29.60	0.90	0.623	0.623
8	13	2	1.22	7.53	9.43	7.06	33.20	8.94	30.46	17.70	0.892	0.892
9	14	6	1.43	7.80	7.81	7.42	37.36	8.94	30.46	19.70	0.841	0.841
10	15	8	1.10	6.70	11.44	7.72	38.73	7.94	33.94	12.97	0.646	0.646

etc.

Figure 21.20: `Restored6.sas - proc print`

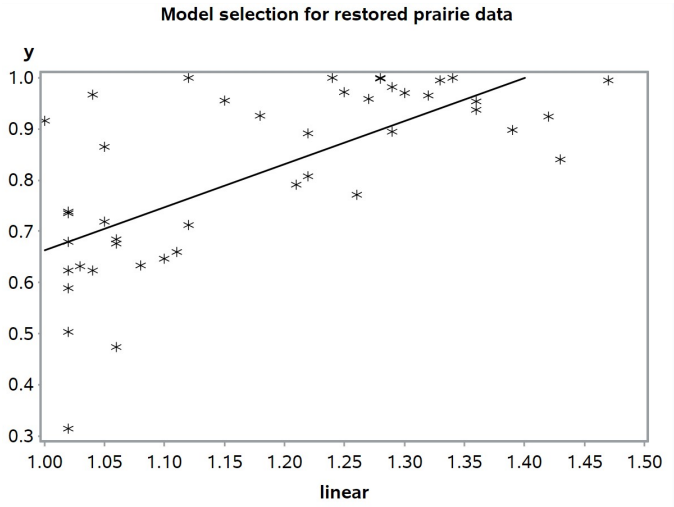


Figure 21.21: `Restored6.sas - proc gplot`

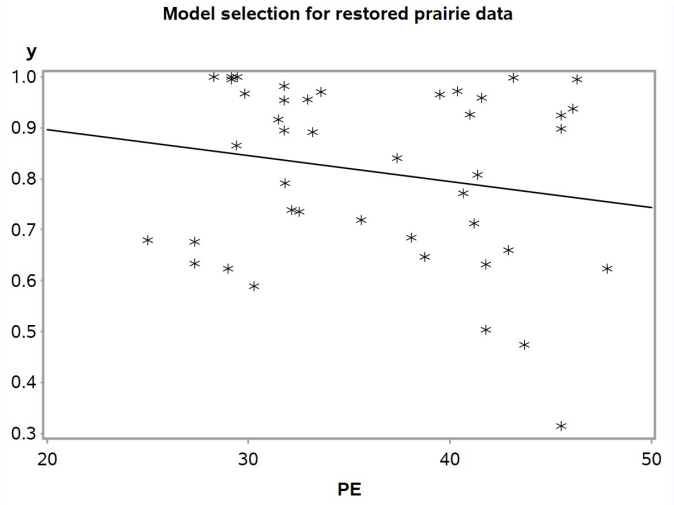


Figure 21.22: `Restored6.sas - proc gplot`

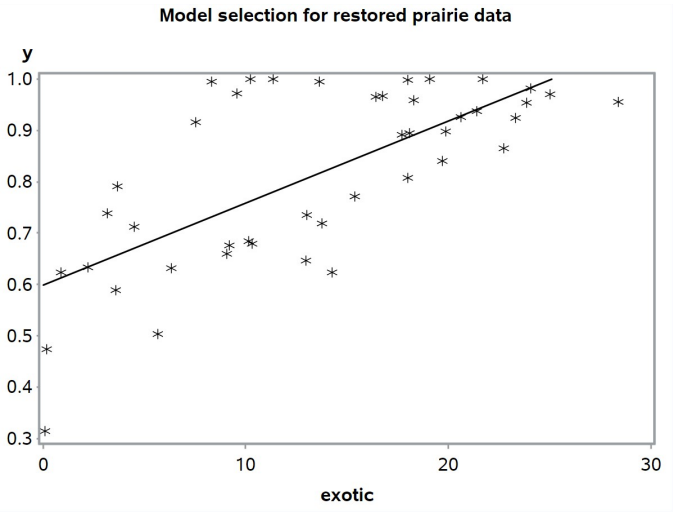


Figure 21.23: `Restored6.sas - proc gplot`

Model selection for restored prairie data

The REG Procedure
Model: MODEL1
Dependent Variable: y

Number of Observations Read	44
Number of Observations Used	44

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	0.95793	0.10644	11.11	<.0001
Error	34	0.32582	0.00958		
Corrected Total	43	1.28375			

Root MSE	0.09789	R-Square	0.7462
Dependent Mean	0.81402	Adj R-Sq	0.6790
Coeff Var	12.02586		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Tolerance	Variance Inflation	95% Confidence Limits	
Intercept	1	-0.06835	0.33572	-0.20	0.8399	0	.	0	-0.75061	0.61392
age	1	0.00358	0.00364	0.98	0.3327	0.09162	0.85869	1.16456	-0.00382	0.01098
linear	1	0.58627	0.17117	3.43	0.0016	0.47598	0.38652	2.58720	0.23841	0.93413
ph	1	0.01789	0.03498	0.51	0.6123	0.05646	0.61263	1.63230	-0.05320	0.08899
organic	1	-0.00646	0.00543	-1.19	0.2425	-0.10727	0.91794	1.08940	-0.01750	0.00458
TE	1	-0.00771	0.01633	-0.47	0.6400	-0.07283	0.31353	3.18951	-0.04090	0.02548
PE	1	-0.01024	0.00333	-3.07	0.0042	-0.38766	0.46845	2.13470	-0.01702	-0.00346
TA	1	0.00021430	0.02467	0.01	0.9931	0.00115	0.42670	2.34355	-0.04992	0.05034
PA	1	0.01084	0.00806	1.35	0.1874	0.13906	0.69869	1.43124	-0.00554	0.02722
exotic	1	0.01060	0.00253	4.19	0.0002	0.46410	0.60956	1.64054	0.00546	0.01573

Figure 21.24: `Restored6.sas - proc reg`

Model selection for restored prairie data

The GLMSELECT Procedure
Selected Model

The selected model is the model at the last step (Step 3).

Effects:	Intercept linear PE exotic

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	3	0.90599	0.30200	31.98
Error	40	0.37776	0.00944	
Corrected Total	43	1.28375		

Root MSE	0.09718
Dependent Mean	0.81402
R-Square	0.7057
Adj R-Sq	0.6837
AIC	-155.33791
AICC	-153.75897
SBC	-194.20115

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	0.245894	0.147395	1.67
linear	1	0.605915	0.127482	4.75
PE	1	-0.007851	0.002301	-3.41
exotic	1	0.010194	0.002334	4.37

Figure 21.25: `Restored6.sas - proc glmselect`

21.16 References

Aho, K., Derryberry, D., & Peterson, T. (2014) Model selection for ecologists: the worldviews of AIC and BIC. *Ecology* 95: 631-636.

Akaike, H. (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control* AC-19: 716-723.

Anderson, D. R., Burnham, K. P., & Thompson, W. L. (2000) Null hypothesis testing: problems, prevalence, and an alternative approach. *The Journal of Wildlife Management* 64: 912-923.

Brewer, M. J., Butler, A., & Cooksley, S. L. (2016) The relative performance of AIC, AICc and BIC in the presence of unobserved heterogeneity. *Methods in Ecology and Evolution* 7: 679-692.

Burnham, K. P. & Anderson, D. R. (2002) *Model Selection and Inference: A Practical Information-Theoretic Approach, Second Edition.* Springer-Verlag, New York, NY.

Burnham, K. P. & Anderson, D. R. (2014) *P* values are only an index to evidence: 20th- vs. 21st-century statistical science. *Ecology* 95: 627-630.

Chang, W. (2023) *R Graphics Cookbook, Second Edition.* https://r-graphics.org/

Coulson, R. N., Mayyasi, A. M., Foltz, J. L., Hain, F. P. & Martin, W. C. (1976) Resource utilization by the southern pine beetle, *Dendroctonus frontalis* (Coleoptera: Scolytidae). *Canadian Entomologist* 108: 353-362.

de Valpine, P. (2014) The common sense of *P* values. *Ecology* 95: 617-621.

Draper, N. R. & Smith, H. (1981) *Applied Regression Analysis, Second Edition.* John Wiley & Sons, New York, NY.

Gotelli, N. J. (2008) *A Primer of Ecology, Fourth Edition.* Sinauer Associates, Inc., Sunderland, MA.

Hofstetter, R. W., Klepzig, K. D., Moser, J. C. & Ayres, M. P. (2006) Seasonal dynamics of mites and fungi and their interaction with southern pine beetle. *Enviromental Entomology* 35: 22-30.

Kaul, A. D., & Wilsey, B. J. (2020) Exotic species drive patterns of plant species diversity in 93 restored tallgrass prairies. *Ecological Applications* (2020): e2252. doi: 10.1002/eap.2252.

Kutner, M. H., Nachtsheim, C. J., Neter, J. & Li, W. (2005) *Applied Linear Statistical Models, Fifth Edition.*

McCulloch, C. E. & Searle, S. R. (2001) *Generalized, Linear, and Mixed Models.* John Wiley & Sons, Inc., New York, NY.

Murtaugh, P. A. (2014) In defense of *P* values. *Ecology* 95: 611-617.

Reeve, J. D., Rhodes, D. J. & Turchin, P. (1998) Scramble competition in southern pine beetle (Coleoptera: Scolytidae). *Ecological Entomology* 23: 433-443.

Ricker, W. E. (1954) Stock and recruitment. *Journal of the Fisheries Research Board of Canada* 11: 559623.

SAS Institute Inc. (2016) *SAS/GRAPH 9.4: Reference, Fifth Edition.* SAS Institute Inc., Cary, NC.

SAS Institute Inc. (2018a) *SAS/IML 15.1 User's Guide.* SAS Institute Inc., Cary, NC.

SAS Institute Inc. (2018b) *SAS/STAT 15.1 Users Guide.* SAS Institute Inc., Cary, NC.

Schwarz, G. (1978) Estimating the dimension of a model. *Annals of Statistics* 6: 461-464.

Sheather, S. J. (2009) *A Modern Approach to Regression with R.* Springer Science+Business Media LLC, New York, NY.

Soul, L. D., Benson, R. B. J., & Weisbecker, V. (2013) Multiple regression modeling for estimating the endocranial volume in extinct Mammalia. *Paleobiology* 39: 149-162.

Tabachnick, B. G., & Fidell, L. S. (2001) *Using Multivariate Statistics, Fourth Edition.* Allyn and Bacon, Boston, MA.

21.17 Problems

1. This problem involves the matrix calculations for linear regression, a special case of multiple regression. Suppose you have a data set with four observations:

Y_i	X_i
4	1
6	2
9	3
10	4

   (a) What is the design matrix $\boldsymbol{X}$ and the vector $\boldsymbol{Y}$ for this data set?

   (b) What is the transpose of $\boldsymbol{X}$, or $\boldsymbol{X'}$? The answer should be a 2×4 matrix.

   (c) Calculate $\boldsymbol{X'X}$ using matrix multiplication. The answer should be a 2×2 matrix.

   (d) Show that the matrix below is the inverse of $\boldsymbol{X'X}$, by multiplying them together to obtain $\boldsymbol{I}$ (the identity matrix).

$$\boldsymbol{(X'X)^{-1}} = \begin{pmatrix} 1.5 & -0.5 \\ -0.5 & 0.2 \end{pmatrix} \tag{21.63}$$

   (e) Calculate $\boldsymbol{(X'X)^{-1}X'}$ using matrix multiplication. The answer should be a 2×4 matrix.

   (f) Finally, calculate $\boldsymbol{\beta = (X'X)^{-1}X'Y}$ using matrix multiplication. The answer should be a 2×1 matrix, with elements equal to the regression intercept and slope. You can check your answer by running a linear regression (see Chapter 17).

2. Ecologists who study predator-prey interactions are often interested in the mortality inflicted by the predator as a function of prey abundance. Data were collected on the proportion of prey eaten by a single predator as the number of prey were increased, in a laboratory experiment (see table below). The proportion eaten was an average over multiple replicates.

   (a) Fit a flexible model to these observations using polynomial regression and SAS. What order polynomial was needed to describe these observations? Attach your program and output.

   (b) Use the polynomial model to predict the proportion eaten for 35 and 45 prey, including confidence intervals for the predictions.

   (c) The proportion eaten vs. prey curve can take different shapes depending on the **functional response** of the predator. For example, the curve would be flat for a Type I response, strictly decreasing for a Type II response, and hump-shaped for a Type III response (Gotelli 2008). How would you classify the response in this experiment?

Number of Prey	Proportion Eaten
1	0.00
2	0.05
3	0.10
4	0.13
5	0.14
7	0.21
10	0.24
15	0.31
20	0.39
25	0.39
30	0.42
40	0.40
50	0.30

3. Data were collected on the abundance of an insect species (Y) as a function of five environmental variables (X_1, X_2, X_3, X_4, and X_5). See table below.

   (a) Conduct a multiple regression analysis of these data using SAS, with Y the dependent variable and X_1, X_2, X_3, X_4, and X_5 the regressors. Discuss the significance of the overall test of the model and the tests for each independent variable. Attach your program and output.

   (b) Use standardized regression coefficients to compare the size and direction of the different effects, especially the significant ones. Which independent variables have the most effect on insect abundance? Which ones increase or decrease it?

   (c) Select the best model for these data using AIC, and write the answer below. Attach your SAS program and output.

X_1	X_2	X_3	X_4	X_5	Y
13.4	10.4	12.7	126	15.4	18.6
11.4	11.9	8.7	115	12.8	21.2
9.2	13.3	10.3	143	11.2	24.1
15.5	11.0	12.6	88	14.9	18.9
13.1	9.9	10.3	156	11.9	26.4
16.3	13.2	9.2	146	14.0	19.8
10.1	15.5	8.9	135	11.3	21.8
7.6	10.0	16.7	128	7.0	26.7
12.9	11.7	16.4	89	13.2	18.6
11.0	11.3	14.8	171	12.3	20.8
11.1	13.3	6.4	128	14.1	21.6
14.3	10.1	9.3	92	15.1	18.4
10.1	13.7	11.3	129	13.0	18.5
12.2	8.9	11.9	143	12.9	24.8
10.9	10.3	12.8	154	12.3	27.9
12.6	13.4	13.7	177	16.1	22.6
12.6	12.4	11.1	165	16.4	26.5
12.4	10.2	7.1	118	13.9	23.9
15.2	13.4	8.1	111	14.2	19.5
13.7	14.0	11.3	123	12.3	16.3
14.1	16.5	7.0	58	7.8	7.1
20.5	8.5	8.1	143	11.9	27.3
5.9	12.0	8.8	149	8.2	27.2
12.7	13.8	19.3	122	12.7	19.7
15.2	9.5	11.9	126	9.2	24.0
17.5	14.0	16.0	130	15.0	19.1
7.7	10.0	10.0	81	9.9	16.0
16.7	13.8	9.3	132	11.2	18.4
14.9	16.9	11.1	124	11.4	17.2
10.6	15.3	13.1	145	15.4	19.4