LTA 15

NASA - AMATYC - NSF
Project Coalition

Kennedy Space Center

Population Size Matters:
Formulating a Mathematical Model for
Scrub-Jay Population at the Kennedy Space Center

Mathematics for Engineering Technology

Environmental Science

Capital Community College

A Florida Scrub Jay perches on a bush at The Merritt Island National Wildlife Refuge.
The 140,000 acre refuge within the boundaries of KSC is home to approximately
one-third of the Scrub Jays found in the Sunshine State.

LTA 15

Population Size Matters:
Formulating a Mathematical Model for Scrub-Jay Population at the Kennedy Space Center

Mathematics for
Environmental Science Engineering Technology

Brian Smith - AMATYC Writing Team Member
McGill University, Montreal, Quebec, **Canada**

Brian Smith taught at Dawson College for 27 years and is presently a faculty member at McGill University in Montreal, Canada. He was the chair of the Technology in Mathematics Education (TiME) Committee of AMATYC, and is the founder and list owner of MATHEDCC, an internet discussion group for college mathematics educators. Brian has published a book on teaching mathematics using graphing calculators.

Mario Triola - AMATYC Writing Team Member
Dutchess Community College, Poughkeepsie, **New York**

Mario Triola has been a mathematics faculty member at Dutchess Community College for 28 years. He is the author of several textbooks, including Elementary Statistics, Survey of Mathematics, and Mathematics and the Modern World. Mario is also a co-author of Elementary Business Statistics and Introduction to Technical Mathematics.

Janet Rebmann - NASA Scientist/Engineer
Kennedy Space Center, **Florida**

Janet Rebmann is a Remote Sensing Analyst and who collects and analyzes remote sensing data such as satellite imagery and spectroradiometric data. She is also an SLSTP Education Outreach Coordinator (ecological programs) and coordinates activities for the ecological programs division of NASA's Spaceflight and Life Sciences Training Program.

Project Grant Team

John S. Pazdar
Project Director
Capital Community College
Hartford, Connecticut

Patricia L. Hirschy
Principal Investigator
Asnuntuck Community College
Enfield, Connecticut

This project was supported, in part, by the
National Science Foundation
Opinions expressed are those of the authors
and not necessarily those of the Foundation

Peter A. Wursthorn
Principal Investigator
Capital Community College
Hartford, Connecticut

LTA 15

Population Size Matters: Formulating a Mathematical Model for Scrub-Jay Population at the Kennedy Space Center

NASA's stated policy is to "protect, preserve, and enhance Kennedy Space Center's unique natural environment". In keeping with that policy, scientists monitor various threatened and endangered species at Kennedy Space Center and Cape Canaveral Air Station, including the Florida Scrub-Jay, the Southeastern Beach Mouse, the Eastern Indigo snake, and the Gopher Tortoise. Most of these animals are classified as "scrub species" i.e., species whose natural habitats are located within the Florida scrub.

In this LTA we will focus our attention on the Florida Scrub-Jay. In addition to their intrinsic value as a threatened species, these birds are important as indicators of what is happening to many other species. In a paper on Scrub-Jay demography (Breininger, 1998), it is pointed out that the Florida Scrub-Jay "is an indicator species of suitable habitat conditions for many other scrub species". Breininger further notes that the Scrub-Jay population in the Kennedy Space Center/Merritt Island National Wildlife Refuge "has been declining for at least ten years and is expected to decline by 40% within the next 10 years". He asserts that the Scrub-Jay population in the region is threatened with extinction in 50 years unless a suitable habitat can be achieved.

Kennedy Space Center ecologists are attempting to create a favorable habitat for the Florida Scrub-Jay and other endangered scrub species by using a combination of cutting vegetation and setting prescribed fires, and generally creating the sort of "open landscape with short vegetation which is characteristic of marshes and scrub".

We will consider population data for the Florida Scrub-Jay as described in the table shown below. The population sizes after 1991 are projected values. Normally, we should not use projected values in developing a mathematical model, but we will treat those values as real for the purposes of this activity.

Table 1: Scrub-Jay Population

Year (t)	Population Size (N)
1980	3697
1985	2512
1989	2176
1991	2100
1992	1922
1993	1857
1994	1860
1995	1689
1996	1603
1999	1127

Information such as that in Table 1 gives us an opportunity to learn about a situation and to acquire information that can be used to guide our decisions. Obtaining information often requires the development of a mathematical framework whose behavior reflects the real situation. Such a mathematical framework is called a model. In trying to find a mathematical model that fits the data best, there are many approaches that could be used. One such approach is to consider the sum of the squares of the residuals, which are differences between the actual
y values and the forecast y values; another is to conduct formal statistical tests. In this LTA, we will use the following steps to develop a mathematical model from data.

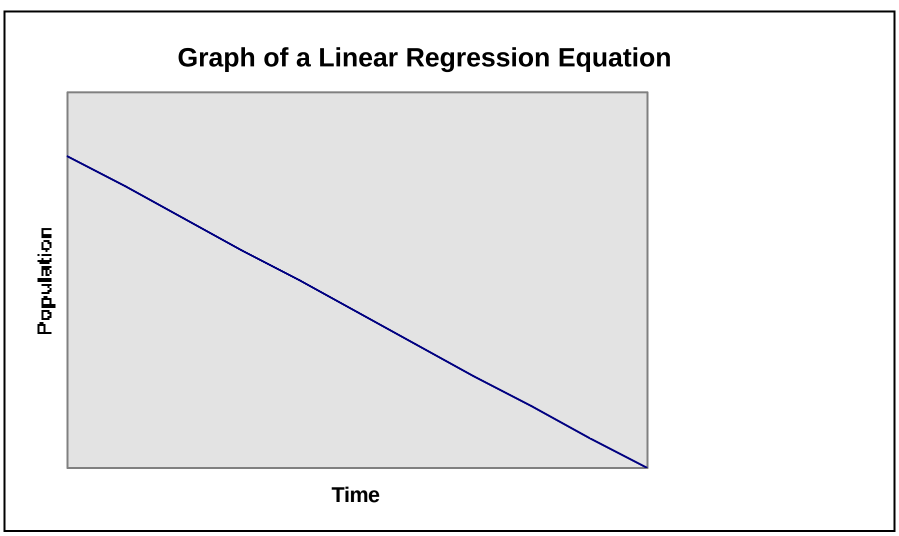
Step 1: Visually inspect a scatterplot of the data for patterns.
Step 2: Generate models to fit the data.
Step 3: Test model for feasibility.

We will illustrate these steps using the data from Table 1.

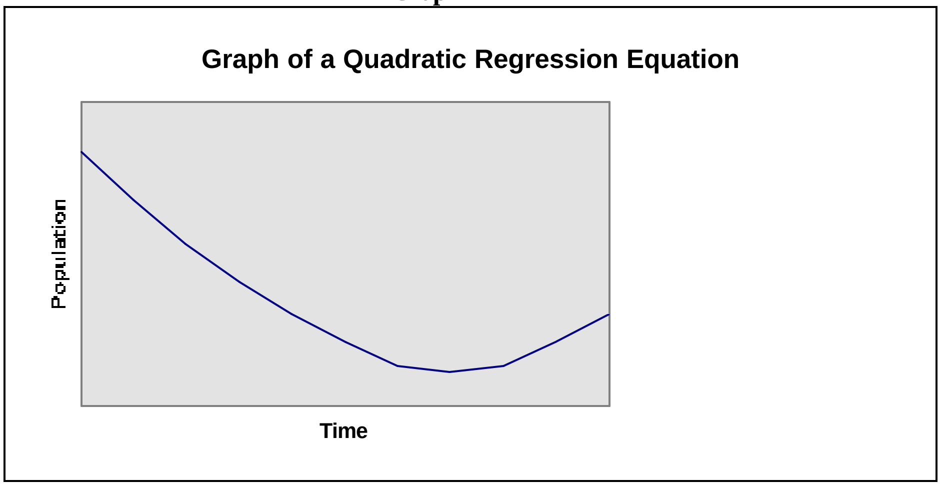
Step 1 Visually inspect a scatterplot of the data for patterns.

A scatterplot of the data may exhibit a known pattern. A scatterplot can be constructed by considering the data as a set of ordered pairs with year as the independent variable and population size as the dependent variable. Each ordered pair is plotted as a point (x, y) on the x-y plane. The plot of the points is named a scatterplot, and it gives a visual display of the trends and patterns in the data set. The scatterplot will often display a recognizable pattern. Some common patterns are: straight line, quadratic curve, exponential curve, power curve, and logistic curve. These patterns are represented by the following graphs of population size vs time.

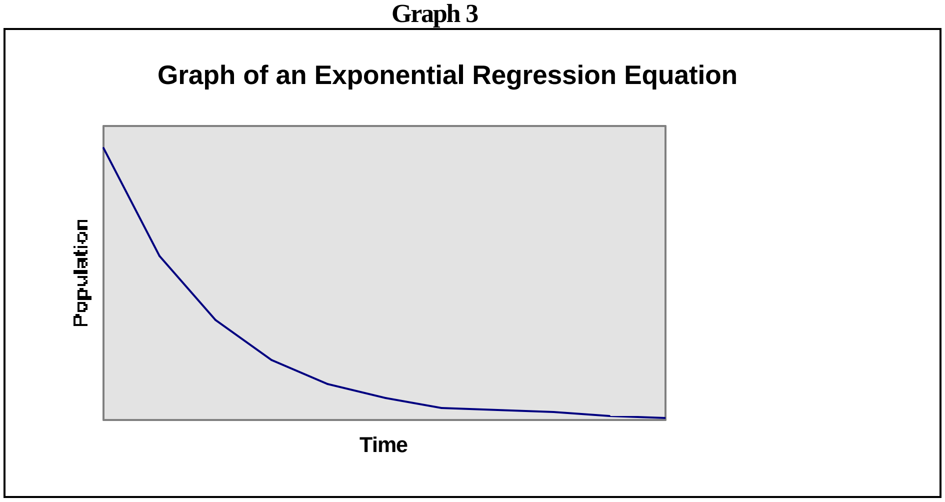
Graph 1



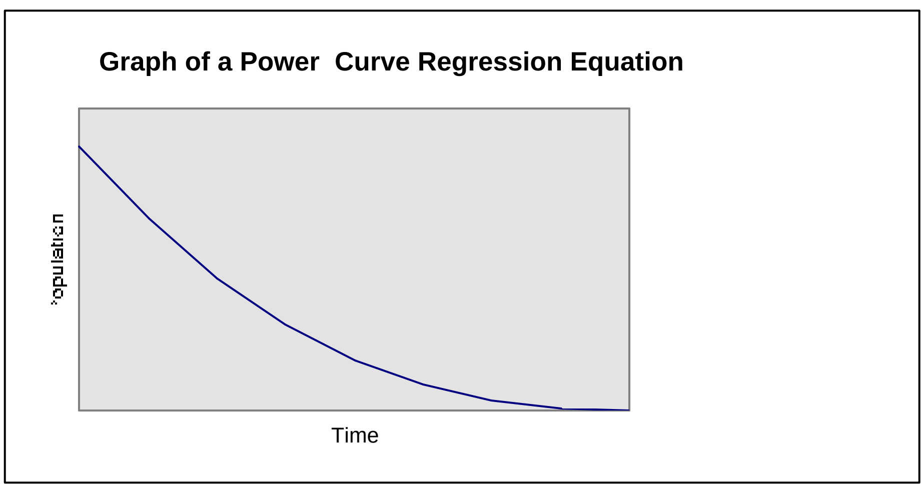
Graph 2



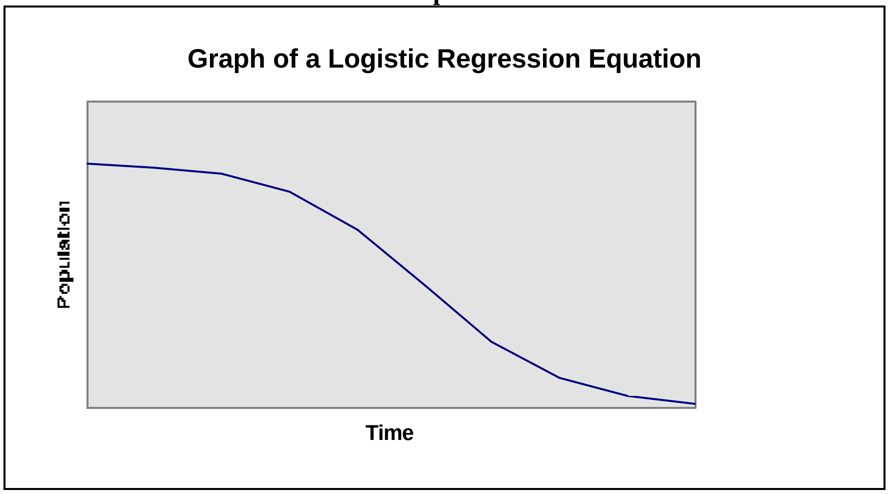
Graph 3



Graph 4



Graph 5



Your Turn

1) Sketch a scatterplot of the data from Table 1 using an x-y coordinate system with year as the independent variable and population size as the dependent variable. Does the pattern of the plotted points appear to fit any of above models?

Step 2 Generate models to fit the data.

There are many computer software packages and graphing calculators that are suitable for developing a mathematical model of sample data. We will illustrate Step 2 by using the TI-83™ graphing calculator, first to obtain an equation such as those listed below, and then to determine a number called the coefficient of determination that will indicate how well the equation fits the data in Table 1.

Linear regression equation: $y = a + bx$

Quadratic regression equation: $y = ax^2 + bx + c$

Exponential regression equation: $y = ab^x$

Power regression equation: $y = ax^b$

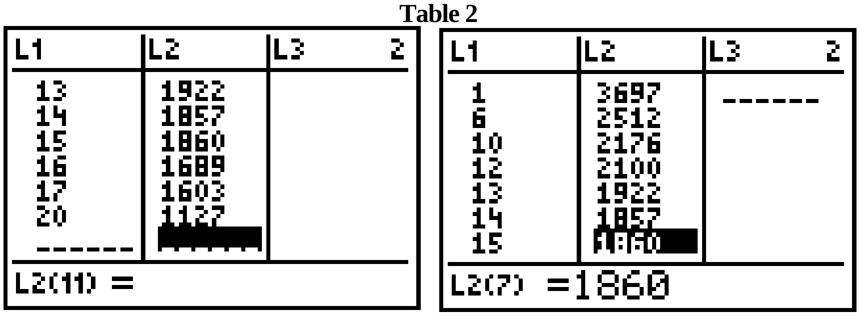
Logistic regression equation: $y = c/(1 + ae^{-bx})$

Using the TI-83™ to generate models.

Because the year values are used as exponents in some of the regression equations, it is helpful to encode the input data with numbers smaller than those in Table 1. This will avoid the creation of very large numbers and the corresponding loss of accuracy. Encode the input data by letting $t = 1$ represent 1980, $t = 6$ represent 1985, etc. The benefit of starting the variable t at 1 instead of 0 to represent 1980 is to avoid evaluating the power function at 0. At this value the power function is either equal to 0 or does not exist. To generate mathematical models based on the encoded data, follow these keystrokes for the TI-83™ calculator:

- Clear lists L1 and L2 by pressing STAT, then selecting item 4: ClrList. Press the ENTER key, then enter L1,L2 (with the comma) and press the ENTER key. A response of "DONE" indicates that lists L1 and L2 have been cleared.
- Press STAT, then select Edit, then press the ENTER key. Enter the coded x-values in list L1, and enter the y-values from Table 1 in list L2. Proceed to enter the data as illustrated in the following TI-83™ screens. Because the data do not fit on one screen, it is necessary to scroll down to obtain the complete set of values.

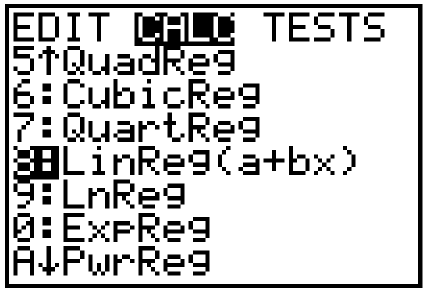
Table 2



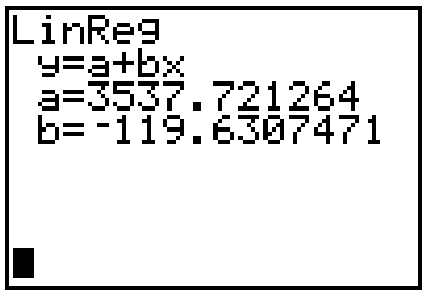
- To select a particular mathematical model, press STAT, then select CALC. Scroll down to select functions such as these:

LinReg(a + bx)	Linear regression model
QuadReg	Quadratic regression model
ExpReg	Exponential regression model
PwrReg	Power regression model
Logistic	Logistic regression model

The selection of the linear regression model is shown below.



If you press the ENTER key twice, you will get the following display.



From the above screen display, we see that the linear regression equation for the sample data is $y = 3537.7 - 119.63x$.

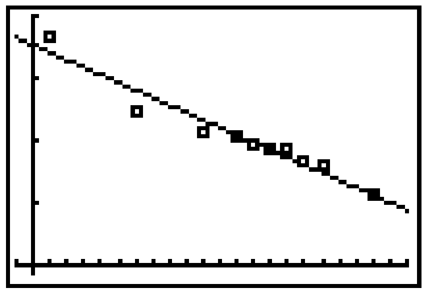
Note that if you attempt to find the logistic regression equation, the TI-83™ will indicate a domain error. The logistic regression equation is discussed in the Appendix.

Your Turn

2) Find the following regression equations for the coded data in Table 2.
   a) Quadratic regression model
   b) Exponential regression model
   c) Power regression model

Scatterplots and Graphs of Equations

Now that we have generated several possible equations for the information in Table 1, how do we decide which equation gives the best fit to the data? One method is to superimpose a graph of the equation on the scatterplot, and see how closely the graph fits the scatterplot. Consider for example the following graph of the linear equation and the scatterplot.



Notice that both the linear graph and the scatterplot share a downward trend, and that the points of the scatterplot tend to cluster on or around the line. These observations indicate that the linear equation provides a good match to the Scrub-Jay population data.

Your Turn

3) Sketch graphs of the quadratic regression, exponential regression, and power regression equations superimposed on the scatterplot of the Scrub-Jay data. Visually inspect the graphs and describe how well each regression equation fits the data.

Coefficient of Determination (r^2 or R^2)

Models can also be compared according to how well they fit a set of data by using a quantity named the coefficient of determination. The notation r^2 is usually used for linear, exponential, and power function models, while R^2 is used for other types of models. The formula for R^2 (or r^2) is shown below, where the values of y are from the original set of sample data, $\hat{y}$ denotes values predicted by the model, and $\bar{y}$ is the mean of the original y values.

$$R^2 = 1 - \frac{\Sigma(y - \hat{y})^2}{\Sigma(y - \bar{y})^2}$$

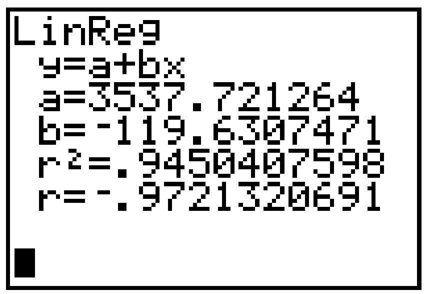
Although this formula is not complicated, it is cumbersome, which makes the manual calculation of R^2 time consuming. Fortunately, the same calculator or software program that gives us a regression equation often provides the coefficient of determination. Thus, we can focus on interpreting values of R^2 instead of performing the calculations.

The following are some key points about the coefficient of determination:

- Values of R^2 are always between 0 and 1.
- If $R^2 = 0$, the model fits the sample data as poorly as is possible.
- If $R^2 = 1$, the model fits the sample data perfectly.
- In general, better models result in values of R^2 that are closer to 1.

If you are using a TI-83™ calculator, you can display values of R^2 along with the regression equations by setting the "DiagnosticOn" option. This option may be set by pressing the CATALOG key (above the number 0 on the TI-83™ keypad), scrolling down until you reach DiagnosticOn, and pressing Enter.

Consider the display for the linear model.



Since the value of r^2 is 0.945 which is close to 1, the linear model provides a close fit to the Scrub-Jay data. This conclusion is consistent with our observations about the close fit of the linear graph superimposed on the scatterplot.

Your Turn

4) Using your calculator or computer software, retrieve the values of the coefficient of determination for each of the models considered in Step 2. Use the coefficients of determination to decide which model best fits the data.

Step 3 Test model for Feasibility

In Step 2, we compared possible models of the Scrub-Jay data by visually inspecting the graphs superimposed on the scatterplot and by comparing the coefficients of determination. It's also important for the model to provide reasonable predictive values for future, past, or missing years. To test your models for this feature, consider the following approach.

- Use the regression equation to calculate future values of the population size, and decide whether these predicted results can reasonably occur.
- Use the regression equation to calculate past values of the population size, and decide whether these results seem reasonable.

- Use the regression equation to calculate the values of the population size for missing years, and determine whether the results appear to be reasonable.

Let's use the linear regression equation from Step 2 to illustrate this approach by predicting the population for the year 2100. First calculate the coded t value, t = 2100 – 1979 = 121, then substitute the t value into the linear equation:

$$N = 3537.7 - 119.63(121) = -10{,}938$$

It's not at all feasible to expect that in the year 2100, there will be -10,938 Florida Scrub-Jays, so the linear model is not good for predicting future values.

Your Turn

5) Predict the population size for the year 2100 using each of the regression equations you determined in Step 2. Determine which equations best fit the data on the basis of the reasonableness of the answers.

After considering the graphs, the values of the coefficients of determination r^2, and the predicted values, it appears that the exponential regression equation provides the best fit to the data.

Exercises

1) The original table of data does not include a population size for the year 1990. Estimate the 1990 population using the indicated regression model. Which estimate is likely to be the most reasonable? Why?
   a) Linear
   b) Quadratic
   c) Exponential
   d) Power

2) The original table of data does not include a population size for the year 2000. Forecast the year 2000 population using the indicated regression model. Which estimate is likely to be the most reasonable? Why?
   a) Linear
   b) Quadratic
   c) Exponential
   d) Power

3) The accompanying table lists year and population size data for a beach mouse in a region of southern Florida. Use computer software or a graphing calculator to find the indicated regression equations and corresponding coefficients of determination. Determine which model gives the best fit to the data, and find the best predicted population size in year 7.

Year (t)	1	2	3	4	5	6
Population	63	90	110	141	165	187

   a) Linear
   b) Quadratic
   c) Exponential
   d) Power

4) The accompanying table lists year and population size data for a beach mouse in a region of southern Florida. Use computer software or a graphing calculator to find the indicated regression equations and corresponding coefficients of determination.
   Determine which model gives the best fit to the data, and find the best predicted population size in year 7.

Year (t)	1	2	3	4	5	6
Population	24	65	165	377	944	2509

   a) Linear
   b) Quadratic
   c) Exponential

d) Power

5) The shad fish population has been declining in the Hudson River for several years. Some biologists have joined in a recommendation that ocean shad fishing be prohibited so that these fish can return to the Hudson River for spawning. The accompanying table lists the pounds of harvested shad in the Hudson River for several consecutive years. Those weights are based on the annual reports from commercial Hudson River fishers to the New York State Department of Environmental Conservation.

Year	Pounds
1980	1,313,100
1981	620,200
1982	378,900
1983	459,400
1984	701,400
1985	756,064
1986	798,768
1987	684,182
1988	782,932
1989	485,700
1990	463,529
1991	329,368
1992	265,598
1993	138,210
1994	157,672
1995	190,607
1996	135,629
1997	93,688

Use computer software or a graphing calculator to find the indicated regression equations and corresponding coefficients of determination. Determine which model gives the best fit to the data, and find the best predicted population size in year 1998.

a) Linear
b) Quadratic
c) Exponential
d) Power

6) Using the Internet, find data describing the world population. Use computer software or a graphing calculator to find the indicated regression equations and corresponding coefficients of determination. Determine which model gives the best fit to the data, and find the best predicted population size in years 2100 and 3000.

a) Linear
b) Quadratic
c) Exponential
d) Power

Appendix

The Logistic Equation

A logistic regression equation is often quite good as a mathematical model describing the size of a population. From the graph of the logistic regression equation given earlier, we can see that the population approaches a limiting value, which is realistic in the sense that the population cannot decrease below 0. The TI-83™ calculator offers the Logistic Regression as an option in its statistics menu, but that option is very limited in the types of data that can be used. This appears to happen because the built-in algorithm on the TI-83™ fails to converge in the number of iterations supported by the calculator. (See Martinez-Cruz and Ratliff, 1998.) Although the
TI-83™ is not capable of generating the specific logistic regression equation based on the data in Table 1, we will use the technique suggested by Martinez-Cruz and Ratliff to obtain a logistic model for the Scrub-Jay population data. Note that other types of software may be used to perform a logistic regression analysis, and the statistical software package SPSS™ is particularly effective in dealing with logistic regression equations.

Consider the logistic model in the form $N = \dfrac{c}{1 - e^{a+bt}}$
On the TI-83™ we will write it as

$$y = \frac{c}{1 - e^{a+bx}}$$

We can rewrite the equation as

$$1 - \frac{c}{y} = e^{a+bx}$$

and, taking logarithms of both sides, we obtain

$$\ln\left(1 - \frac{c}{y}\right) = a + bx$$

Because the right side of this equation is linear, we can use linear regression to fit $\ln\left(1 - \dfrac{c}{y}\right)$

to a + bx and hence compute values of a and b. The catch here is that this will only work if we know the value of c.

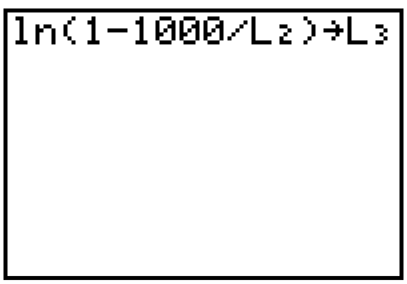
Notice that $\lim_{x \to \infty} y = c$, provided b < 0. Thus c represents the asymptotic limit of y.

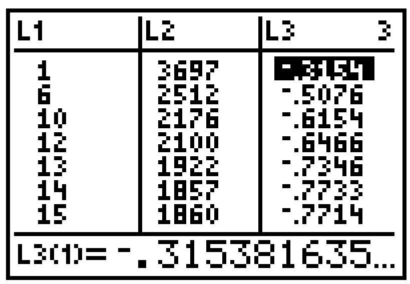
Although c is unknown (in fact it is the value that we are attempting to estimate) we will find the logistic model which best fits our data by computing the regression equation for selected values of c and choosing the model with the highest value of the coefficient of determination, r^2.

We illustrate the procedure for the case when c = 1,000.

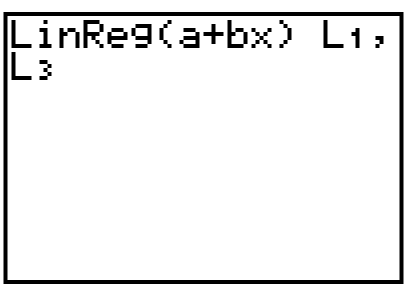
We must first transform the values of y for the Scrub-Jay population, stored in list L_2, into the values $\ln\left(1-\frac{c}{y}\right)$, which we will store in list L_3.

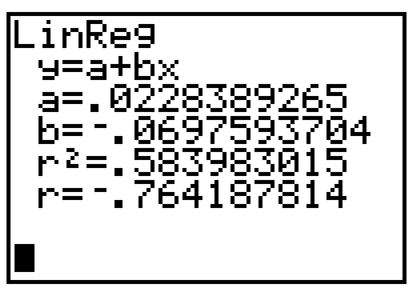
From the home screen on the TI-83™, perform the command $\ln\left(1-\frac{1000}{L_2}\right) \to L_3$. The command and the resulting lists are shown below:





Next, perform a linear regression on L_1, L_3 as shown in the next two screens:

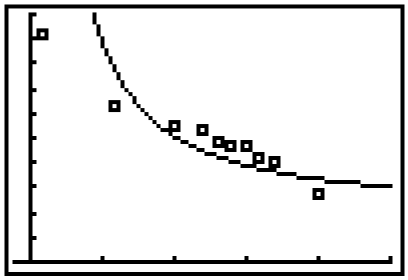




Rounding to four decimal places, we see that a = 0.0228 and b = -0.0698. Hence the logistic model is

$$y = \frac{1000}{1 - e^{.0228 - .0698x}}$$

Since the value of r^2 = .58398, this logistic equation does not appear to be a convincing model for the population data. The next screen shows the graphical representation of this model.



Based on the above analysis it appears that a logistic model which predicts a final population value of c = 1,000 is not appropriate. The following table shows the results when we try different values for c.

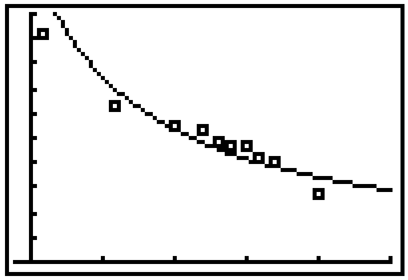
c	a	b	r^2
1000	0.0228	-0.0698	0.58398
800	-0.0876	-0.0397	0.69997
600	-0.0946	-0.0241	0.76007
400	-0.0726	-0.0136	0.79753
200	-0.0388	-0.0060	0.82314
100	-0.0198	-0.0028	0.83310
50	-0.00999	-0.00136	0.83755
10	-0.0020	-2.6694E-4	0.84088
1	-2.011529E-4	-2.6564195E-5	0.84161

Notice that as c decreases, r^2 increases. Since the case when c = 0 is degenerate, the smallest integer value that c can assume is 1. Clearly, if there is only one member of a population left alive, the population will soon be extinct.

We see that the best logistic model is obtained when c = 1 and

$$y = \frac{1}{1 - e^{-.0002011529 - .000026564195x}}$$

The graphical representation of this model is shown below:



While this appears to be a reasonable fit to the population data, the r^2 value of 0.84161 is noticeably smaller than the values previously obtained for the linear (r^2 = 0.9450),
quadratic (R^2 = 0.9601), and exponential (r^2 = 0.9457) regression models.

It is worth noting that the predicted value for the year 2100 (t = 121) is N = 293. Thus, the logistic model, like the exponential model, is approaching zero as a limiting value, but the convergence is slower for the logistic model.

However, because of the larger value of r^2 we prefer the exponential model. Therefore, we can reasonably conclude that this logistic model is not the most appropriate one for the Scrub-Jay population data.

REFERENCES

Breininger, D.R.: Florida Scrub-Jay Demography in an Urban Metapopulation. Unpublished manuscript, Kennedy Space Center, 1998.

Martinez-Cruz, A.M. & Ratliff, M. I.: Beyond Modeling World Records with a Graphing Calculator: Assessing the Appropriateness of Models. Mathematics and Computer Education, 32(2), 143-153, 1998.