Making It Fit: Using Excel to Analyze Biological Systems

by

Robert Leaf and Brian R. Murphy
Fisheries and Wildlife Sciences Department
Virginia Polytechnic Institute and State University

Part I – The Age-at-Length Relation

> *… the islands are valued the world over for their spectacular coastlines and aquamarine waters, the industry of harvesting fish and other marine creatures for home aquariums is largely unregulated here, raising concerns over damage to the environment, the tourism industry and the aquarium fishery itself… a $50 permit allows collectors across most of Hawai'i to net as many of a species as they want, wherever they want and whenever they want. That sometimes means harvesting hundreds of thousands per year of a single species from a single bay.*
>
> —Tara Godvin, *The Honolulu Advertiser*, Sunday, October 9, 2005

Dr. Latimer, an ichthyologist working in Honolulu, Hawai'i, peered myopically at one of the tiny ear stones, the otolith, of the yellow tang *(Zebrasoma flavescens)* through his microscope. It was interesting for him to see the light and dark bands, called "annuli," that were formed as the fish grows. They reminded him of the rings of a tree and, like the rings on a tree, these could also be used to estimate the age of a fish. He had studied this species for years, and had often seen them racing by when he was diving on the local coral reefs and rocky areas around the island. Dr. Latimer was interested in their ecology, but today he was counting rings on otoliths from a collection of fish to provide information for fishery managers. The necessity of determining the age-at-length relationship was crucial because yellow tang and other reef fishes in Hawai'i are the target of a major fishery: hundreds of thousands of juvenile yellow tang and tens of thousands of Moorish idol *(Zanclus cornutus)*, spotted surgeonfish *(Ctenochaetus strigosus)*, and Achilles tang *(Acanthurus achilles)* are caught each year for the live aquarium trade, resulting in significant decreases in abundance (Tissot and Hallacher 2003). These fishes are shipped around the world to aquarium enthusiasts. In addition to determining the ecology of this species, Dr. Latimer's agency was charged with describing important life-history characteristics in the hope that an understanding of the species would help improve management practices.

Figure 1. Otolith of bluefish *Pomatomus saltatrix* showing the annuli.

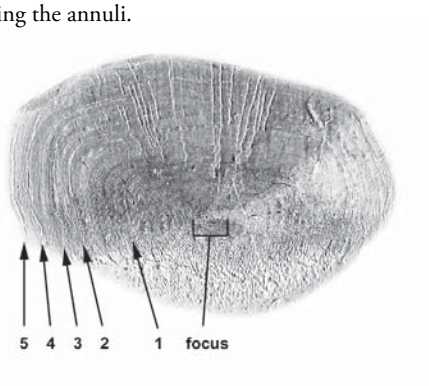


Figure 2. Photo of yellow tang *Zebrasoma flavescens.*

Image credits: Figure 1 used with permission from the Center for Quantitative Fisheries Ecology, Old Dominion University, Norfolk, VA. Figure 2 ©1997 Randall, J.E., Randall's underwater photos, licensed under a Creative Commons Attribution-Noncommercial 3.0 Unported License.

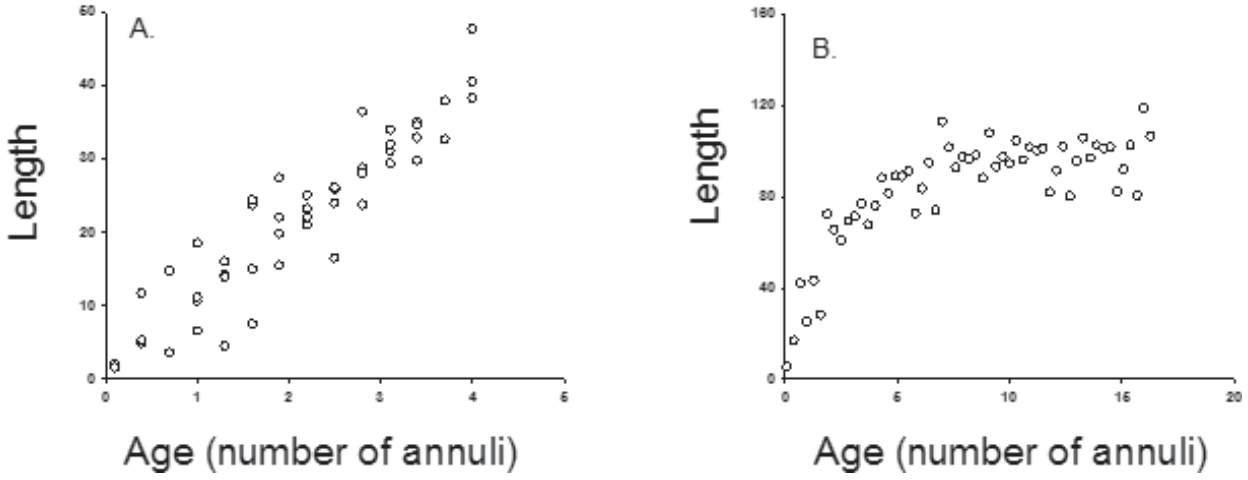
Dr. Latimer had had a long day. He was almost finished counting the rings of the otoliths for the collection of yellow tang he had collected. His next step was to figure out a way to present the data in a way that the managers could use. He knew that fishery managers

generally fitted curved lines to these types of data, and he was puzzled about how to proceed. He was familiar with plotting straight lines (linear model) to data, but wondered if he could use the same principles of linear regression to plot curved lines.

Questions

1. Other than using the otolith, what alternatives are there for determining the age of a fish?
2. If we want to predict the length of a fish based on its age, how would we plot such a relationship on a graph with two axes (*x* and *y*)? How do we choose which value to appear on the vertical axis, and which to appear on the horizontal axis?
3. Consider the two panels (A and B) in Figure 3 below that include length and age (estimated from the number of annuli) data collected from a fish population. Discuss the criteria you would use to determine the "best-fit" line for each of the figures.
4. Why might you get such data? For example, when might you expect a straight line, and when might you expect a curved line? Justify your answers.
5. Draw a straight line through the points in Figure A and a curve through the points in Figure B. Both of these should start at or near the origin (0,0). Describe why you drew the line you did to determine the "best-fit" curve.

Figure 3.



References

Tissot, B. N., and L. E. Hallacher. 2003. Effects of aquarium collectors on coral reef fishes in Kona, Hawaii. *Conservation Biology* 17:1759–1768.

Part II – The von Bertalanffy Growth Function

Dr. Latimer looked through the textbooks that he had and found that:

> *"Theoretical nonlinear curves are often used to describe the relation of age to length in fishes. One such curve that assumes indeterminate, asymptotic growth and is often used to describe length-at-age is the von Bertalanffy growth function."*

The von Bertalanffy growth function (VBGF) can be expressed in the following form:

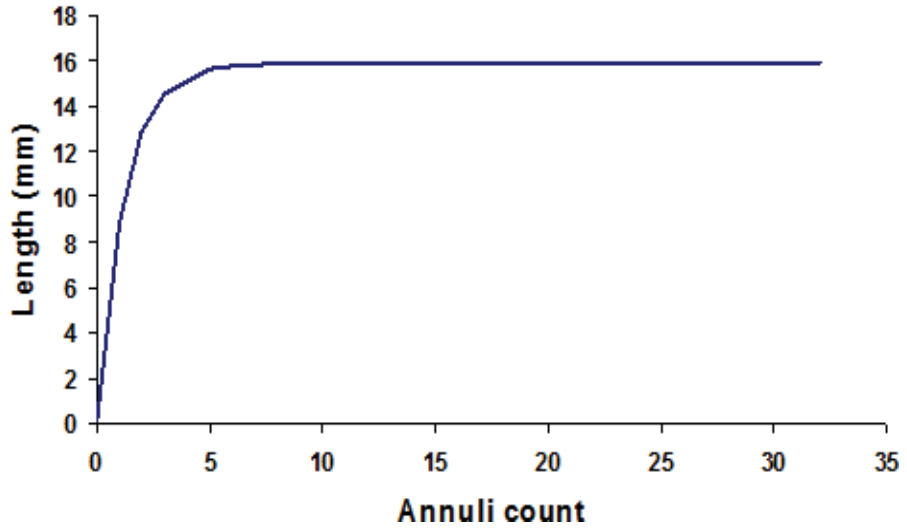
Figure 4.

$$L_i = L_\infty(1 - e^{-kt_i}) + \varepsilon_i$$

Karl Ludwig von Bertalanffy was an Austrian born biologist and intellectual who published the above model in 1938. This model (presented above in its two-parameter form) is perhaps one of the most widely used models to describe age and growth relationships of fishes. The above equation describes how an individual's (the i^{th} individual's) length, L, is predicted as a function of its age (t) as a result of the values of the parameters k and L_∞. k is the instantaneous growth rate (units of time, most often an annual period, year^{-1}) and L_∞ is the maximum length (units are length, such as inches, mm, or cm) attained by an average individual in the population. Below we present the age as the annuli count, which are often assumed to be annual.

This curve looks like:

Figure 5.



In this model, the slope of the line defined by the VBGF changes (decreases) as age increases. At some age (around age 8), the individual stops growing.

Although it had been a while since Dr. Latimer had done this kind of work, he thought that he could make some headway if he compared it to a similar problem—how to fit a straight line to data. A straight line that describes the relationship of data is called a linear model and is described mathematically with two parameters, m and b. m describes the slope of the line and b is the location that the line crosses the y axis.

$y_i = mx_i + b + \varepsilon_i$

The ε_i term describes the amount of error, or residual error, that is not accounted for by the prediction (the difference between the observed length, *yi*, and the expected length, *mxi* + *b*, predicted by the model). He knew that in order to fit a line, the value of the difference of the observed data (that collected in the field) and expected data (the line) needed to be as small as possible. This is called "minimizing the sum of squared residuals." Thus the value of ε for each individual is the residual.

Linear model: $y_i = mx_i + b + \varepsilon_i$

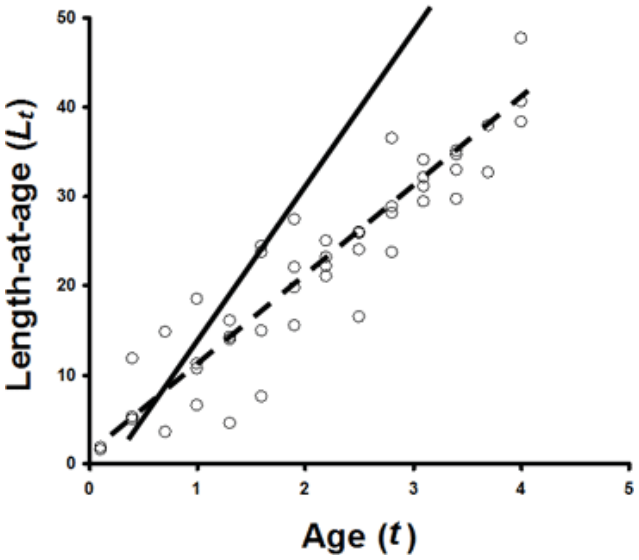
von Bertalanffy growth function: $L_i = L_\infty (1 - e^{-kt_i}) + \varepsilon_i$

The observed data are on the left side of the equal sign, and the equation for the line is specified on the right side of the equation. The shapes of the non-linear or linear models are determined by the parameters.

Questions

1. Discuss the difference between "data" and "parameters." What are parameters and why are they important for our understanding of what a line looks like?

Figure 6.

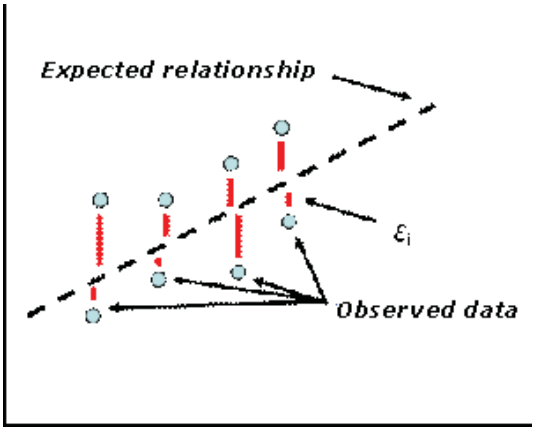


2. Why would we prefer the dashed line if we were trying to describe the trends in the data? What exactly are we describing with the dashed line?
3. Which line (solid or dashed) has a greater slope *(m)* and which has the greater *y*-intercept *(b)*?
4. Using estimates of the parameters, *m* and *b*, for the linear model, what is the equation for the dashed line (L_t is the length at age = *t* and age is *t*)?

Part III – Minimized Sum of Squared Residuals

Dr. Latimer knew that some lines fit the data better than others. During his research, he noticed that the differences in the line and the points should be minimized, but he was puzzled about how this could be done. He also knew that in order to determine the best fit curve, he would need an independent criterion to determine the fit of line to data. Could he use the process of "minimizing the sum of squared residual" as a criterion?

Figure 7.



He knew from his previous work (in Part II) that the equation for the straight line, if it were fitted to the length-at-age data, would be:

$$L_i = mt_i + b + \varepsilon_i$$

and is similar to:

$$L_i = L_\infty(1 - e^{-kt_i}) + \varepsilon_i$$

The parameters in both curves are on one side of the equation and define the expected portion of the line. The observed data is on the left side of the equation.

Questions

1. We could present the equation to express the residual error term as the difference of the observed and expected terms. Rewrite the equation below to solve for ε.

   $L_i = mt_i + b + \varepsilon_i$

   $\varepsilon_i =$

2. How could we express the von Bertalanffy growth function if we wanted to have the residual error value be equal to the difference in the observed and expected values?

   $L_i = L_\infty(1 - e^{-kt_i}) + \varepsilon_i$

   $\varepsilon_i =$

3. Taking this a step further; what is the equation of ε^2 for each of the equations?

   von Bertalanffy $\varepsilon_i^2 =$

   Linear $\varepsilon_i^2 =$

4. How would these calculations work for a sample data set? Fill in the expected lengths for each of the given equations:

Age *(t)*	Observed Length *(Lt)*	$L_\infty = 26$, $k = 0.6$
1	7.2	
2	21	
3	27.5	

5. Using the "expected" values that you calculated above for the von Bertalanffy growth function, fill in the table below:

Age *(t)*	Observed Length *(Lt)*	Observed – Expected Length	(Observed – Expected Length)^2
1	7.2		
2	21		
3	27.5		

6. What is the *sum* of all of the elements in the (observed – expected)^2 column (this will be the "minimized sum of squared residuals")? Use some alternative values for the *k* parameter and compare the values of the sum of the squared residuals. Which model do you think fits the data better? Why?
7. How can the poor fitting model be made to fit the data better?

Part IV –Help from Excel

Dr. Latimer knew that in order to determine the proper parameters it was not adequate to use arbitrarily selected values. Computer programs, such as Excel, are used to determine these. He entered the data into a spreadsheet called Age_Length.xls.

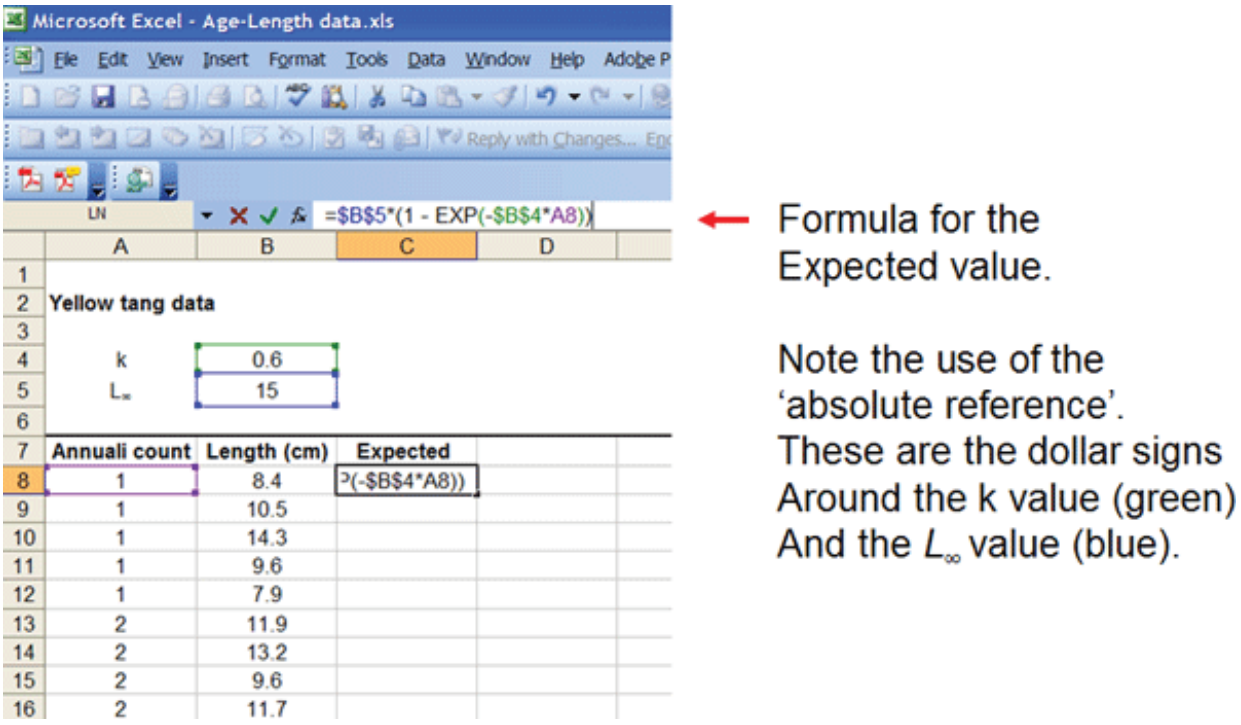
Following the same methods you used in Part III, there are a number of columns that need to be added to the spreadsheet in order to determine the sum of the (observed – expected)^2. What are the contents in these columns? *Hint:* If you follow the methods in Part III, you will need to insert formulas into three columns and add the numbers of the last column to get a single value (the sum of the squared residuals).

Questions

1. Plot the observed data: length and age. Using Excel: Insert → Chart → XY (Scatter) → Next → Series Tab → select X (Annuli count) and Y (Length).
2. As we previously did, we will use the equation of the von Bertalanffy growth function to determine expected values. This time, you will program Excel to do the work for you. Use the parameters k and L_∞ from cell B3 and B4 to calculate the values in column C.

   You will use the absolute reference feature in Excel to do this:

Figure 8.



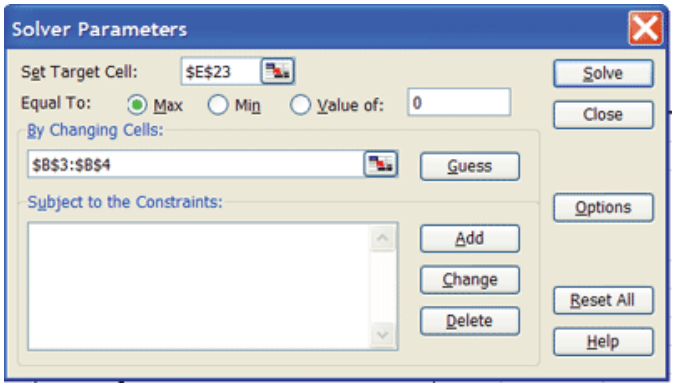
3. Plot the expected data (expected length). How do you think that the fit of the curve can be improved?
4. Change the values of k and L_∞ in the cells B3 and B4. Describe how your estimate of the expected lengths changes as the parameters are altered. For example, change the values of k to smaller values. Observe the fit of the corresponding curve. How does this compare to the first graph?

Part V –Help from Solver

Dr. Latimer knew that there was a tool in Excel for minimizing the sum of squared residuals called "Solver." He opened the built in Solver function using Tools→Solver.

In order to minimize the sum of squared residuals, he chose that cell as the target cell (this will likely be the sum of the final column you added to the spreadsheet). The cells that correspond to k and L_∞ are the target cells; they will be changed to find their value, which results in the minimum value of the sum of squared residuals.

Figure 9.



Questions

1. Experiment with Solver using your data set and describe how you think it works.
2. Compare the values that Solver estimated with those of your classmates that analyzed the age and growth data fish species—did you get the same values?
3. As discussed in Part I, you might estimate age and growth by using tagging and recapturing methods in the field. If you did this, what would you include in Excel to analyze these data?
4. You have now seen that Solver can be used to determine the parameters of a non-linear curve. Fisheries biology (and biology in general) is full of examples of non-linear curves. Discuss how you would set up the Excel spreadsheet (what are the parameters, how would you calculate the difference in the observed data, etc.) to determine the parameters for length-weight curve below (these are data collected by the author of Japanese medaka, *Oryzias latipes*, a small freshwater fish):

   Equation is: $W = aL^b$

Figure 10.