Model Based Statistics in Biology.

Part III. The General Linear Model.

Chapter 9.4 Exponential Function, using Linear Regression

ReCap. Part I (Chapters 1,2,3,4)
ReCap Part II (Ch 5, 6, 7)
ReCap Part III
9.1 Explanatory Variable Fixed by Experiment
9.2 Explanatory Variable Fixed into Classes
9.3 Explanatory Variable Measured with Error
9.4 Exponential Functions
9.5 Power Laws. Linear Regression
9.6 Model Revision

Data files and analysis
Lungfish.xls
Ch9.xls

on chalk board

ReCap Part I (Chapters 1,2,3,4)
Quantitative reasoning: Example of scallops,
which combined models (what is the relation of scallop density to substrate?)
with statistics (how certain can we be?)

ReCap Part II (Chapters 5,6,7)
Data equations summarize pattern in data as a series of parameters (means, slopes).
Frequency distributions, a key concept in statistics, are used to quantify uncertainty.
Hypothesis testing uses the logic of the null hypothesis to make a decision about an unknown population parameter.
Estimation is concerned with the specific value of an unknown population parameter.

ReCap (Ch 9) The General Linear Model is more useful and flexible than a collection of special cases.
Regression is a special case of the GLM. We have seen examples with the explanatory variable X fixed and examples with the explanatory measured with error.

Today:
Linear Regression for Exponential Functions

Wrap-up

Exponential relations are common in biology.

Exponential growth of populations.
Exponential mortality of a cohort.
Exponential growth of organisms (over limited size ranges).

The relation of response to explanatory variable is non-linear.
To estimate exponential parameters we must either use non-linear regression or make the equation linear so we can apply linear regression.

GLM, regression. Application to exponential functions.
Exponential rates are common in biology.
An example: Intrinsic rate of population increase.

Data on population size N at time t
Draw graph (straight line, if logarithmic scale of N, but not for t).
Equation $N = N_o e^{rt}$
r is the intrinsic rate of increase. It has units of % time^{-1}
r is the slope of the regression of $\log_e N$ against time t.

Another example: specific growth rate from measurements of body mass M at two points in time $t = t_{final} - t_{initial}$.

M = initial weight (kg)
M_o = recapture weight (kg)
t = time in days from initial to recapture.
Equation $M = M_o e^{kt}$

$$\log_e (M / M_o) = k\, t$$

k = exponential growth rate = $\ln(M/M_o)/t$, with units of % / day

Data.
Growth of 6 lungfish in 2001 in Lake Baringo, Kenya.
Chrisestom Mlewa (2003) Biology of the African lungfish *Protopterus aethiopicus* Heckel 1851, and some aspects of its fishery in Lake Baringo, Kenya. Ph.D. Thesis, Department of Biology, Memorial University, St. John's, Canada.

kg	kg	Time
Initial	End	Days
1.32	1.46	50
1.30	1.48	64
1.60	1.84	65
0.76	0.90	56
0.60	0.65	20
2.74	2.86	48

1. Construct the model

Verbal model. Growth rate of lungfish is exponential, with fixed growth rate k.
Graphical model. Loglinear plot of relation of M/M_o to t.

Response variable is M/M_o the ratio of final to initial weight.
Explanatory variable is t = time in days from initial to recapture.

Formal model. $M = M_o e^{kt}$

This is a non-linear relation, hence to estimate k we must either use non-linear regression or make the equation linear so we can apply linear regression.

Here is the linearized model $\log_e (M / M_o) = k\, t$

For the linearized model we compute the intercept from the estimates of the slope and the grand mean of the response.

For population $\log_e (M / M_o) = \alpha + k\, t$
For sample $\log_e (M / M_o) = a + b_t \cdot t + error$
same as: $\log_e (M / M_o) = \hat{\alpha} + \hat{\beta} \cdot t + error$

1. Construct the model

Linearization is widely used but unfortunately, it introduces bias (Smith 1984, 1993, Packard 2009) that compromises the predictive power of the relationship of the response variable to the explanatory variable (Zar 1968, Smith 1980, 1984). Bias associated with log transformations include the magnification of the effects of outliers (Smith 1980, 1984), multiplicative error (Smith 1993), and inaccurate estimates of the dependent variable at large values of independent variable (Packard and Boardman 2008a). Advances in computer based graphics and statistical software allow estimates for non-linear functions (Packard 2009). We begin analysis with the classical approach, linearization. We then evaluate whether to undertake non-linear estimation of the growth parameter.

Packard GC (2009) On the use of logarithmic transformations in allometric analyses. J Theor Biol 257(3): 515-518
Packard GC, Boardman TJ (2008a) Model selection and logarithmic transformation in allometric analysis. Physiol Biochem Zool 81(4): 496-507
Smith JR (1993) Logarithmic transformation bias in allometry. Am J Phys Anthropol 90: 215-228
Smith RJ (1980) Rethink allometry. J Theor Biol 87: 97-111
Smith RJ (1984) Allometric scaling in comparative biology: Problems of concept and method. Am Journal Phys Reg Int Comp Phys 246(2): 152-160
Zar JH (1968) Calculation and miscalculation of the allometric equation as a model in biological data. BioScience 18: 1118-1120

2. Execute model. For this example we will use a spreadsheet

Place data in model format.

Data in two columns, $\log_e$ *(M / M_o)* and *t*

Compute fitted values and residuals from parameter estimates.

Parameter estimates from functions in spreadsheet (cells D19 D20)
Fitted values from parameter estimates (column F).

	A	B	C	D	E	F	G	H
1	kg	kg	Time	ln(M/Mo)	fits	res	kg	
2	Initial	End	Days				predicted	
3	1.32	1.46	50	0.1008	0.1096	-0.0088	1.47	
4	1.30	1.48	64	0.1297	0.1313	-0.0016	1.48	
5	1.60	1.84	65	0.1398	0.1329	0.0069	1.83	
6	0.76	0.90	56	0.1691	0.1189	0.0502	0.86	
7	0.60	0.65	20	0.0800	0.0631	0.0170	0.64	
8	2.74	2.86	48	0.0429	0.1065	-0.0636	3.05	
9								
10		SS	1359.5	0.01025	0.00327	0.00698		
11		df		5	1	4		
12								

Move numbers to ANOVA table.

13	Source	df	SS	MS	F	p
14	Area	1	0.00327	0.00327	1.87	0.243
15	Residual	4	0.00698	0.00175		
16			0.01025			
17						
18		coeff	stdev	lower	upper	
19	slope	0.155%	0.113%	-0.160%	0.470%	
20	intercept	0.0321				
21						

Explanation of computations in spreadsheet.

Column D	=LN(B3/A3)	produces value of 0.1008 (paste from D4 to D8)
Column E	=INTERCEPT($D$3:$D$8,$C$3:$C$8)+SLOPE($D$3:$D$8,$C$3:$C$8)*C3	produces value of 0.1096 (paste from E4 to E8)
Column F	=D3-E3	produces value of -0.0088 (paste from F4 to F8)
Column G	=A3*EXP(E3)	produces value of 1.47 (paste from G4 to G8)
Cell C10	=DEVSQ(D3:D8)	produces value of 1359.5 (paste from D10 to F10)
Cell D11	=COUNT(D3:D8)-1	
Cell E11	=1	
Cell F11	=COUNT(F3:F8)-2	
Cells D14 and D15 from E11 and F11 respectively (df to ANOVA table)		
Cells E14 and E15 from E10 and F10 respectively (SS to ANOVA table)		
Cell F14	=E14/D14	MS
Cell F15	=E15/D15	MS
Cell G14	=F14/F15	F-ratio
Cell H14	=FDIST(G14,D14,D15)	p-value
Cell D19	=SLOPE($D$3:$D$8,$C$3:$C$8)	
Cell D20	=INTERCEPT($D$3:$D$8,$C$3:$C$8)	
Cell E19	=SQRT(F15/C10)	standard error of slope
Cell F19	=D19-E19*TINV(0.05,D15)	Lower Confidence Limit
Cell G19	=D19+E19*TINV(0.05,D15)	Upper Confidence Limit

2. Execute model:

The least squares estimate of growth rate is 0.155% per day. The intercept is 3.21%. The general linear model is then:

$$\begin{array}{lllll} \text{GLM:} & M/M_o - 0.11037 = & & 0.00155\,(t - 50.5) & +\ \text{res} \\ \text{Regression Eq:} & M/M_o \quad = & 0.0321 & +\ 0.00155\,t & +\ \text{res} \end{array}$$

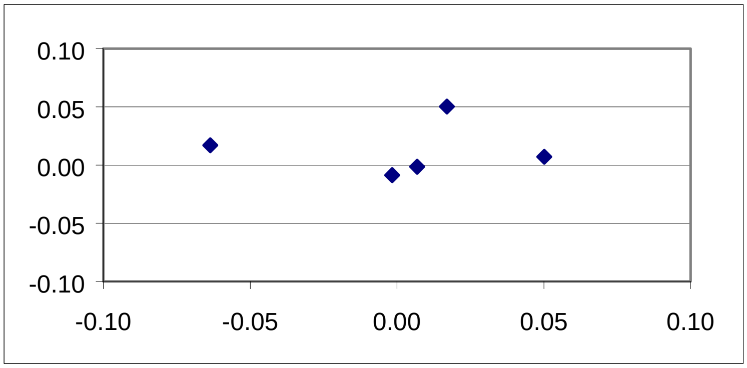
Compute residuals as observed – fitted.

3. Evaluate the model.

Check straight line assumption for regression.

No arches or bowls.

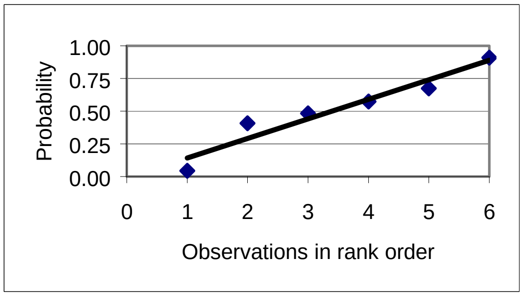
So linear model is acceptable.



Check error model (homogeneous, normal, independent errors).

Homogeneity is difficult to judge with only 6 residuals. There appears to be greater spread in the middle of the graph than on either side, but this is an artifact of only one residual to the left and only two residuals to the right. We will assume that residuals are homogeneous.

Normal errors ?
Too few residuals to construct a histogram.
Probability plot shows normal errors.



Independent errors ?
Fish recaptured on different dates so we will assume no influence of one measurement on another and hence independent errors.

4. State sample, population, and whether representative.

All lungfish ? Probably not.

All fish that could have been collected when the collection was made.

This is a more realistic statement of the population.

But it may not be defensible unless this collection was made at random, which is not likely.

All measurements that could have been made on 6 fish by this protocol.

This is an even more restrictive statement of the population.

This is a hypothetical rather than an enumerable biological population.

In this example, an enumerable population is not defensible.

So a hypothetical population, based on repeatable protocol, is used.

The results apply to other observational studies using the same measurement protocols.

The model to which we are inferring applies to egg number, given a knowledge of fish size.

5. Decide whether to use hypothesis testing.

The research objective is to estimate specific growth rate of fish.

We will examine the parameters and compute confidence limits (skip to step 10).

10. Examine parameters of biological interest.

Calculate confidence limits so as to include true value of β_M 95% of time.

$$s_b^2 = s_{y.x}^2/\Sigma x^2 = (0.00175/1359.5) = 0.00000128$$
$$s_b = \text{ square root of } s_b^2 = (0.00113)\ (0.113\ \%/\text{day})$$

For 95% limits use $t_{0.05/2[4]}$ because df = 4 = ν

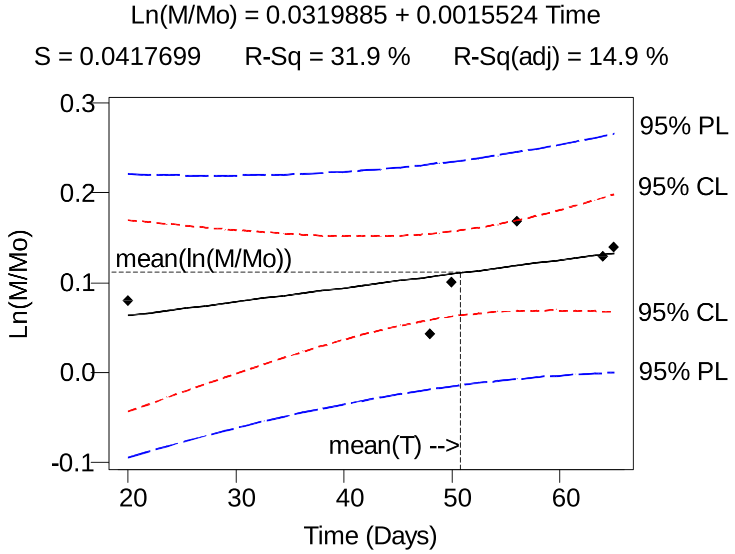
$$L = \text{Lower limit} = \hat{\beta}_M - t_{\alpha/2[\nu]}s_b = 0.00115 - 0.00113*2.776 = -0.160\ \%/\text{day}$$
$$U = \text{Upper limit} = \hat{\beta}_M + t_{\alpha/2[\nu]}s_b = 0.00115 + 0.00113*2.776 = 0.470\ \%/\text{day}$$

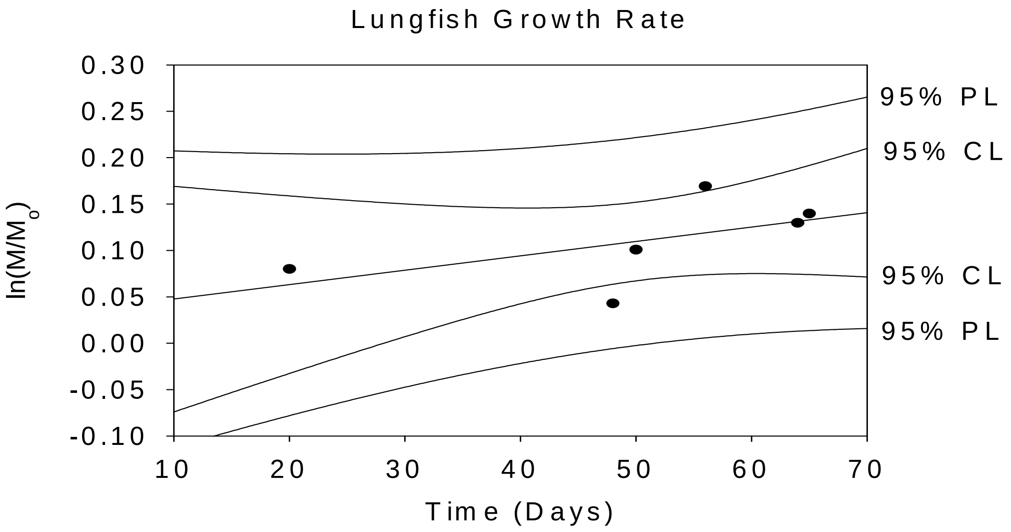
Draw cdf, arrows going from p-value (vertical axis) over to curve and down to t statistic (horizontal axis).

The confidence limits include zero, leading to the conclusion that there was no growth. However, all 6 fish were larger upon recapture than initially. This is an improbable result ($0.5^6 = 0.0156$) under the null hypothesis of no growth (binomial test). Consequently we can exclude the hypothesis of no increase in mass (a biologically possible outcome in fish). However, we cannot exclude the hypothesis of no exponential increase in biomass, based on the confidence limits for the growth parameter *k*.

The estimate of growth rate is approximately 0.1%/day, or about 3% per month. The estimate is however, not very reliable because we have so few fish.

This unreliability can be seen when we plot the confidence limits on the growth equation. The confidence limits for the equation are not the same as for the parameters. The confidence limits become wider the further the distance from the mean value of X. This shape accommodates a range of slopes, all running through the same point, the mean value of the Y and X variable.





Given the uncertainty and the absence of outliers to distort the estimates, a better estimate of the growth parameter *k* via non-linear estimation is not necessary.