Model Based Statistics in Biology.

Part III. The General Linear Model.

Chapter 9.5 Power Law 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 & analysis
Kleiber.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 an example with the explanatory variable X fixed, an example with the explanatory measured with error, and an example for a non-linear (exponential) relation of response to explanatory variable.

Today:
Linear Regression for Power Laws, another non-linear relation.

Wrap-up
Power laws are common in biology.
- Number of species in relation to area
- Metabolic rate in relation to body size
- Perimeter of a convoluted object (shoreline, leaf edge, etc).

Power laws are usually analyzed taking logs, to linearize the equation
Regression equations are inaccurate if linear assumption not correct.
Residual analysis is especially important in analysis of power laws.
If the first model tried is not appropriate, based on residual analysis, an iterative approach is taken to arrive at an appropriate model.

GLM, regression. Application to power laws.
Power laws are common in biology. An example is the allometric relation of part of the body to the entire body (Gould ref). Goes back to Huxley 1932.

Another example is the relation of species to area.
As a rule of thumb species numbers will double for each tenfold increase in area.
Species - area relations have a long history in biology.
The first quantitative treatment was by Olof Arrhenius, who proposed a power law relation of species number to area
Arrhenius, O. 1921. Species and area. Journal of Ecology 9: 95-99.

Another example is the relation of metabolic rate to body size (Rubner's Law)

Kleiber (1932) reviewed the relation of the metabolic rate to body mass

$$\frac{\dot{E}}{\dot{E}_{ref}} = \left(\frac{M}{M_{ref}} \right)^{\beta}$$

$\dot{E}_{ref}$ is the metabolic rate (kcal/day) of the reference unit (organism) of mass M_{ref} (kg)
$\dot{E}$ is the metabolic rate of other units (organisms) each with mass M
$\dot{E} / M$ is the mass-specific metabolic rate of each organism.
To obtain a power law in conventional notation:

$$\dot{E} = \left[\dot{E}_{ref} \; M_{ref}{}^{-\beta} \right] M^{\beta}$$

where

$$\dot{E} = \alpha \; M^{\beta}$$

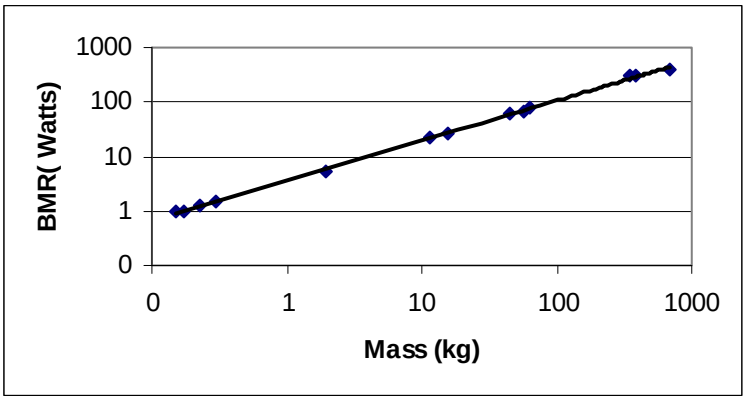
$$\alpha = \left[\dot{E}_{ref} \; M_{ref}{}^{-\beta} \right]$$

Table 1 in Kleiber (1932) reports average values of metabolic rate (kcal/day) and body weight (kg) for birds and mammals ranging in size from a ring dove to a steer. The averages were based on numbers of organisms ranging from 2 to 136.

	Mass	BMR Kcal/day	Norganism	BMR Watts
Ring Dove	0.15	19.5	9	0.94
Female Rat	0.173	20.2	18	0.98
Male Rat	0.226	25.5	23	1.24
Pigeon	0.3	30.8	3	1.49
Hen	1.96	106.0	14	5.14
Female Dog	11.6	443.0	11	21.47
Male Dog	15.5	525.0	10	25.44
Sheep	45.6	1219.9	7	59.11
Woman	56.5	1349.0	103	65.37
Man	64.1	1632.0	136	79.08
Cow	388	6421.0	4	311.15
Steer	342	6255.0	4	303.11
Steer	679	8274.0	2	400.94

1. Construct model.

Response variable is metabolic rate $\dot{E}$ (Watts)
Explanatory variable is Weight M (kg)
Verbal model: Metabolic rate is a power law function of body mass.



The graphical model is a line ---->

The formal model: $\dot{E} = k\, M^{\beta}$

Where β is the slope of the line on a log-log plot.

To estimate the parameters by linear regression, the equation is rewritten in linear form by taking logarithms

$$\ln(\dot{E}) = \ln k + \beta\ \ln M$$

The model for the population $\ln(\dot{E}) = \alpha + \beta\ \ln M + \varepsilon$

The model for the sample $\ln(\dot{E}) = \hat{\alpha} + \hat{\beta}\ \ln M + \varepsilon$

This is equivalent to $\ln(\dot{E}) = \hat{\beta}_o + \hat{\beta}\ \ln M + \varepsilon$

The y-intercept, α, will be calculated from the estimate of the slope and the estimate of the grand mean, $\hat{\beta}_o$. The estimate of *k* will be calculated from the estimate of the y-intercept

$$\hat{k} = e^{\hat{\alpha}}$$

2. Execute analysis. Place data in model format:

Column with response variable $\ln(\dot{E})$
Column with explanatory variable $\ln(M)$

Code model statement in statistical package according to the GLM, compute residuals and fits.

$\ln(\dot{E}) = \beta_o + \beta\ \ln M + \varepsilon$

```
MTB > regress 'lnNsp' 1 'lnA';
SUBC> residuals c5;
SUBC> fits c6.
```

or

```
MTB > GLM 'lnNsp' = 'lnA' ;
SUBC> residuals c5;
SUBC> fits c6.
```

2. Execute analysis.

To obtain the fitted values, we convert from logarithms back to the original scale.

Formula Fits = exp (ln[fits])
Example 0.880 = exp (–0.13)

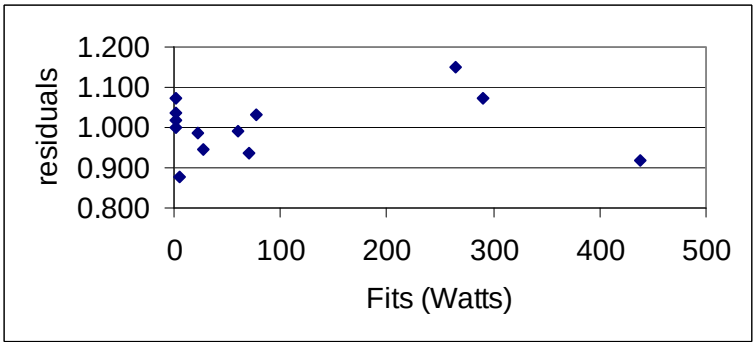
Note that after log transformation, the data equations become:

Data = Fits * Residuals

ln(Data)	=ln(Fits)	+ ln(Res)	Data	= Fits	x Res
-0.06	-0.13	0.07	0.945	0.880	1.073
-0.02	-0.02	0.00	0.979	0.978	1.001
0.21	0.17	0.04	1.236	1.191	1.037
0.40	0.38	0.02	1.493	1.468	1.017
1.64	1.77	-0.13	5.137	5.860	0.877
3.07	3.08	-0.01	21.467	21.748	0.987
3.24	3.29	-0.06	25.441	26.932	0.945
4.08	4.09	-0.01	59.114	59.691	0.990
4.18	4.25	-0.07	65.370	69.914	0.935
4.37	4.34	0.03	79.084	76.734	1.031
5.74	5.67	0.07	311.151	289.555	1.075
5.71	5.58	0.14	303.107	263.821	1.149
5.99	6.08	-0.09	400.944	437.505	0.916

3. Evaluate structural model.

Straight line?
We plot the back calculated residuals against the back calculated fitted values.

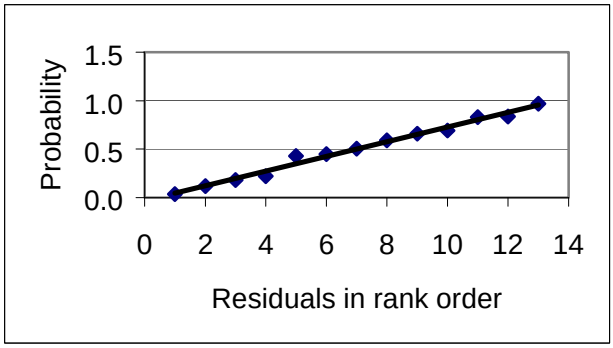


There is no convincing evidence of bowl or arch in the plot. Straight line assumption accepted.

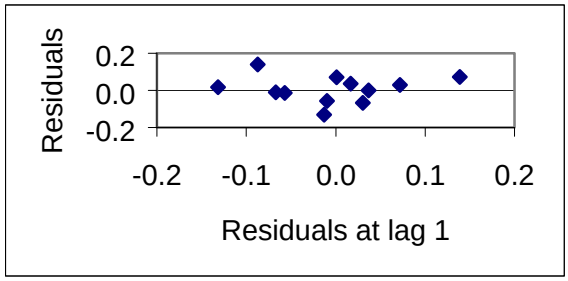
3. Evaluate error model.

Homogeneous errors ? Yes.
Dispersion around zero similar from left to right in residual vs fit plot.

Normal errors?
Yes, residuals close to line for normal distribution.



Independent errors?
We have no information on temporal sequence or on spatial arrangement. We order the observations from small to large, to check this possible source of non-independence. The plot of errors versus neighboring error value shows no upward or downward trend. We assume errors are independent.



4. State sample, population, and whether sample is representative.
Population is all possible measurements, given the measurement protocol.
Sample assumed to be representative.

5. Decide on mode of inference. Is hypothesis testing appropriate?
No. At this point we are more interested in the magnitude of the exponent than we are in whether there is a relation. We already know that metabolism increases with body size, a statistical test of the null hypothesis is of no interest. Kleiber was interested in whether the exponent departs from 2/3, as expected if metabolism in a volume depends on the surface area of that volume. Skip to step 10, analysis of parameters.

10. Analyze parameters of biological interest.
There was no evidence of violation of assumptions so confidence limits will be calculated from t-distribution.

Compute confidence limits from standard error of the slope parameter.

GLM routine reports $\hat{\beta} = 0.73755 \pm 0.007136$

$P\{\text{Lower} \leq \beta \leq \text{Upper}\} = 1 - \alpha = 95\%$

$\text{Lower} = \hat{\beta} - t_{0.025[df]} * \text{st.err.}$

$\text{Lower} = 0.73755 - 2.201 * 0.007136 = 0.722$

$\text{Upper} = \hat{\beta} + t_{0.025[df]} * \text{st.err.}$

$\text{Upper} = 0.73755 + 2.201 * 0.007136 = 0.753$

Which hypotheses are excluded by confidence limits ?

The confidence limits exclude an exponent of 2/3
They also exclude an exponent of zero (the null hypothesis).
They also exclude an exponent of 1 (1:1 relation of metabolism to mass).
The confidence limits do not exclude an exponent of 3/4.

Kleiber's 1932 paper was a landmark in establishing that the exponent relating metabolic rate to body mass was higher than the 2/3 value expected from the Euclidean geometry of flux into or out of a volume across its surface. Subsequent theoretical work focused on explanation of the 3/4 power law (e.g. West *et al*. 1997).

Exercises.

1. The Kleiber data consists of means computed from sample sizes ranging from 2 to 136. Estimates based on few organisms are less reliable than those based on larger numbers. Most statistical packages allow the placing of greater weight on means from large samples than from small samples. Use weighted regression to estimate the exponent that relates metabolic rate to body mass. Does weighted regression change the results of analysis?

2. Rubner examined the relation of the mass specific metabolic rate to body mass in 7 dogs, ranging in mass from 3 to 30 kg.
   Rubner M. (1883) Über die Einflus der Körpergrösse auf Stoff und Kraftwechsel. *Z. Biol*. 19: 535-562

$$\frac{\dot{E} / M}{\dot{E}_{ref} / M_{ref}} = \left(\frac{M}{M_{ref}} \right)^{\beta}$$

$\dot{E}_{ref}$ is the metabolic rate (kcal/day) of the reference unit (organism) of mass M_{ref} (kg)
$\dot{E}$ is the metabolic rate of other units (organisms) each with mass M
$\dot{E} / M$ is the mass-specific metabolic rate of each organism.
To obtain a power law in conventional notation:

Rubner's relation is rewritten as $\frac{\dot{E}}{\dot{E}_{ref}} = \left(\frac{M}{M_{ref}} \right)^{\beta-1}$

which becomes $\dot{E} = \left[\dot{E}_{ref} \; M_{ref}{}^{\beta-1} \right] M^{\beta-1}$

This is rewritten as $\dot{E} = \alpha \, M^{\beta-1}$

where $\alpha = \left[\dot{E}_{ref} \; M_{ref}{}^{\beta-1} \right]$

If the exponent that relates metabolic rate to body mass is 2/3 (as Rubner expected), then $\beta - 1 = 2/3$ and $\beta = -1/3$.

Dog	M (kg)	E/M	E (kcal/day)
1	31.20	35.68	1113.22
2	24.00	40.91	981.84
3	19.80	45.87	908.23
4	18.20	46.20	840.84
5	9.61	65.16	626.19
6	6.50	66.07	429.46
7	3.19	88.07	280.94

Rubner computed and reported the mass-specific metabolic rate ($\dot{E} / M$). As a result, the response variable ($\dot{E} / M$) has a built-in dependence on the explanatory variable *M*. The strength of this computationally induced correlation depends on variation in *M*, relative to variation in $\dot{E}$ .Computing the metabolic rate $\dot{E}$ from the mass-specific rate $\dot{E} / M$ (as in the table above) aggravates the problem by introducing computationally induced correlation of the response variable $\dot{E}$ with the explanatory variable *M*. Does computationally induced correlation affect the parameter estimates? Compare the regression of $\dot{E}$ =f(*M*) to the regression of $\dot{E} / M$ =f(*M*) with respect to (a) the parameter estimates; (b) explained variance r^2; (c) F ratio; (d) p-value; (e) confidence limits; (f) whether a straight line model is appropriate, as judged from the residual versus fit plot.

Summarize by stating which components (a-f) are affected by computationally induced correlation and which are not.