Model Based Statistics in Biology.

Part III. The General Linear Model.

Chapter 9.7 Model Revision

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 Power Laws. Non linear regression
9.7 Model Revision

Data files & analysis
Arrh.out
Arrh.xls
Gleason.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 logarithms, to linearize the equation
Regression equations are inaccurate if relation not linear after taking logarithms
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 metabolic rate to body size (Kleiber's Law)
Goes back to late 19th century, with the work of Rubner.

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 the following relation of species to area

$$\left(\frac{Nsp}{Nsp_{ref}}\right)^{\beta} = \frac{A}{A_{ref}}$$

Nsp_{ref} is the number of species in a reference quadrat of area A_{ref}
Nsp is the number of species in larger areas A formed by combining quadrats.

Arrhenius, O. 1921. Species and area. Journal of Ecology 9: 95-99.

To obtain a power law in conventional notation, Arrhenius' relation is rewritten as

$$\frac{Nsp}{Nsp_{ref}} = \left(\frac{A}{A_{ref}}\right)^{1/\beta}$$

which becomes $$Nsp = \left[Nsp_{ref}\ A_{ref}{}^{-1/\beta}\right] A^{1/\beta}$$

This is rewritten as $$Nsp = c\,A^{z}$$

where $$c = \left[Nsp_{ref}\ A_{ref}{}^{-1/\beta}\right]$$

Arrhenius reported values of β in areas ranging in size from 0.02 m^2 to 1 m^2 in 14 different plant communities in Sweden.

GLM, regression. Application to power laws.
In the following year H.A. Gleason showed that species numbers from this power law could not be reasonably extrapolated to areas larger that 1 m^2. He then proposed an alternative relation of species to area.

$$\frac{Nsp - Nsp_{ref}}{\ln A - \ln A_{ref}} = \beta_G$$

which translates to

$$Nsp = \left[Nsp_{ref} - \beta_G \ln A_{ref} \right] + \beta_G \ln A$$
$$Nsp = \alpha + \beta_G \ln A$$

Gleason, H.A. 1922. On the relation between species and area. Ecology 3: 158-162.

Gleason reported values of β_G for areas ranging from 2 m^2 to 240 m^2 in Aspen woodlands.

Neither Arrhenius nor Gleason used regression methods to estimate their parameters. To illustrate power law regression, two data sets are analyzed. The first is Arrhenius' data for herb-*Pinus* wood in Sweden, which Gleason used to show that a power law cannot be extrapolated to large areas. The second is Gleason's data for aspen woodlands in Michigan.
Here is the Arrhenius data.

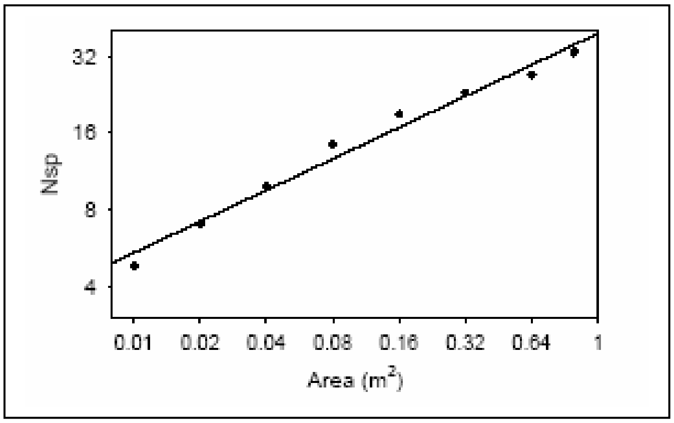
1.0000	4.8000
2.0000	7.0000
4.0000	9.8000
8.0000	14.3000
16.0000	18.9000
32.0000	23.0000
64.0000	27.0000
100.0000	33.0000
Area	Nsp

1. Construct model

Response variable is number of species *Nsp*
Explanatory variable is area *A*

Verbal model: Number of species increases with area according to a power law

Graphical model ...



1. Construct model

Distinguish response from explanatory variables.
Response variable.

Nsp = number of species in single quadrat or conjoined quadrats in herb-*Pinus* wood in Sweden.

Explanatory variable.

A = area of combined quadrats, ranging from 0.02 to 1 m^2
each quadrat is (0.1 m)^2

Both variables are on a ratio type of scale, the explanatory variable (area) is fixed rather than measured.

The formal model is $Nsp = c\,A^z$ where *z* is the slope of the line

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

$$\ln(Nsp) = \ln c + z \ln A$$

For population $\ln(Nsp) = \alpha + z \cdot \ln A + \varepsilon$
For sample $\ln(Nsp) = \hat{\alpha} + \hat{z} \ln A + error$
Equivalently $\ln(Nsp) = \hat{\beta}_o + \hat{z} \ln A + error$

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 *c* will be calculated from the estimate of the y-intercept

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

2. Execute analysis. Place data in model format:

Column with response variable, ln(*Nsp*)
Column with explanatory variable ln(*A*)

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

$$\ln(\mathrm{Nsp}) = \alpha + z \cdot \ln A + \varepsilon$$

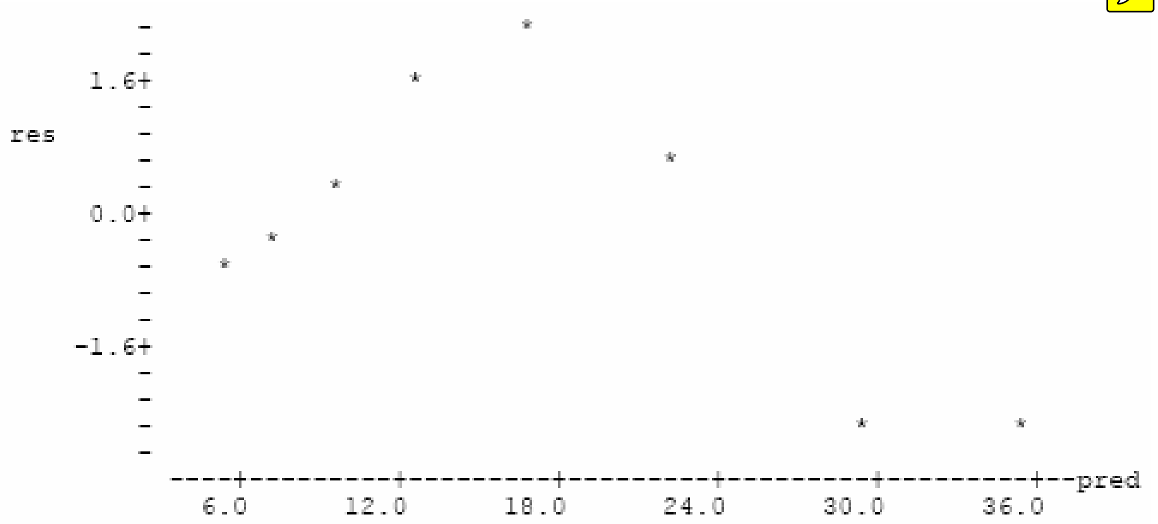
```
MTB > GLM 'lnNsp' = 'lnA' ;
SUBC> residuals c5;
SUBC> fits c6.
```

3. Evaluate model.

Structural model: Straight line?

Plot residuals against fitted values.

```
 MTB > plot c6 c5
```



```
         -                       *
         -
      1.6+                *
         -
res      -
         -                                *
         -           *
      0.0+
         -       *
         -    *
         -
         -
     -1.6+
         -
         -
         -                                            *         *
         -
           ----+---------+---------+---------+---------+---------+--pred
             6.0      12.0      18.0      24.0      30.0      36.0
```

Clearly, Arrhenius’ data do not fit a power law.
Because of the strong arch in the residuals, any extrapolation to larger areas will greatly overestimate species numbers.

At this point we return to step 1.

1. Construct the model

Do the data fit Gleason’s model ?

population $$Nsp = \alpha + \beta_G \ln A + \varepsilon$$

sample $$Nsp = \hat{\alpha} + \hat{\beta}_G \ln A + error$$

2. Execute analysis. Place Arrhenius data in model format:

Column with response variable *Nsp*.
Column with explanatory variable ln(*A*)

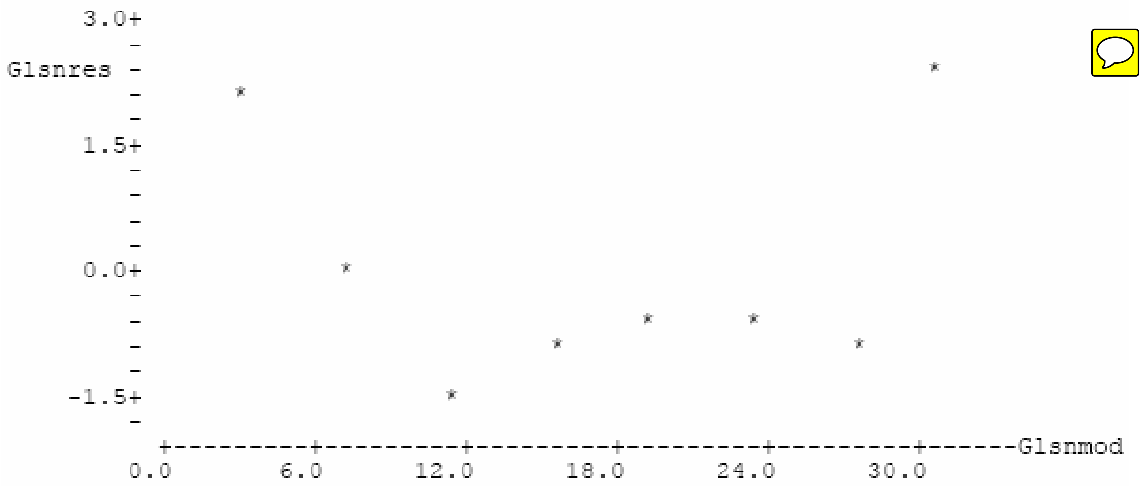
Code model statement in statistical package according to the GLM, compute residuals and fits.

$$Nsp = \alpha + z \cdot \ln A + \varepsilon$$

```
MTB > GLM ‘Nsp’ = ‘lnA’ ;
SUBC> residuals c5;
SUBC> fits c6.
```

3. Evaluate the model.

Plot residuals against fitted values.



There is a clear bowl in the plot of residuals. If we trim the high residuals at high or low areas, then we have an arch at intermediate sized areas.

Arrhenius’ data do not fit Gleason’s model. Rather than searching for a model appropriate to Arrhenius’ data, we examine Gleason’s data, beginning with Gleason’s model.

1. Gleason model

For population $Nsp = \alpha + z \cdot \ln A + \varepsilon$

For sample $Nsp = \hat{\alpha} + \hat{z} \ln A + error$

2. Execute analysis. Place Gleason data in model format:

Column with response variable *Nsp*

Column with explanatory variable ln(*A*)

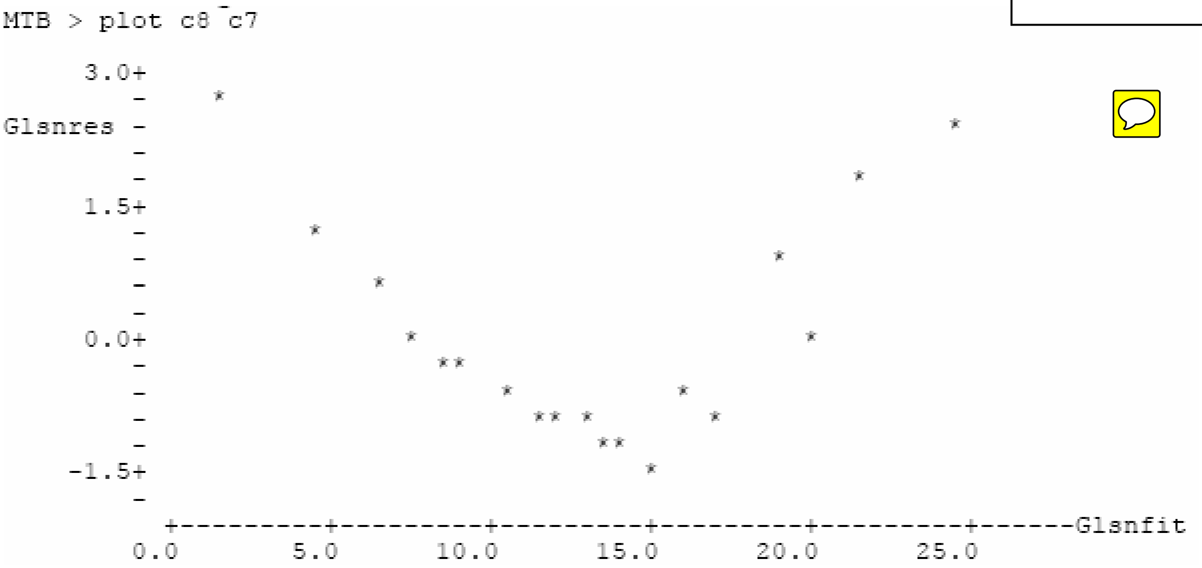
obsno	area(sq m)	Nsp
1	1	4.375
2	2	5.817
3	3	6.900
4	4	7.600
5	5	8.208
6	6	8.950
7	8	9.667
8	10	10.333
9	12	11.250
10	15	12.250
11	16	12.000
12	20	12.917
13	24	13.500
14	30	15.215
15	40	16.167
16	60	19.750
17	80	20.000
18	120	23.500
19	240	27.000

Code model statement, compute residuals and fits.

$$Nsp = \alpha + z \cdot \ln A + \varepsilon$$

```
MTB > GLM ‘Nsp’ = ‘lnA’ ;
SUBC> residuals c5;
SUBC> fits c6.
```

Plot residuals against fitted values.



```
MTB > plot c8 ¯c7

     3.0+
        -    *
Glsnres -                                                  *
        -
        -                                            *
     1.5+
        -          *
        -                                       *
        -              *
        -
     0.0+                *                        *
        -                  **
        -                      *          *
        -                        ** *       *
        -                            **
    -1.5+                               *
        -
          +---------+---------+---------+---------+---------+------Glsnfit
        0.0       5.0      10.0      15.0      20.0      25.0
```

3. Evaluate the model.

There is a clear bowl in the plot of residuals.
Gleason’s data do not fit Gleason’s model.
Try the Arrhenius power law.

1. Arrhenius model

For population $\ln Nsp = \alpha + z \cdot \ln A + \varepsilon$

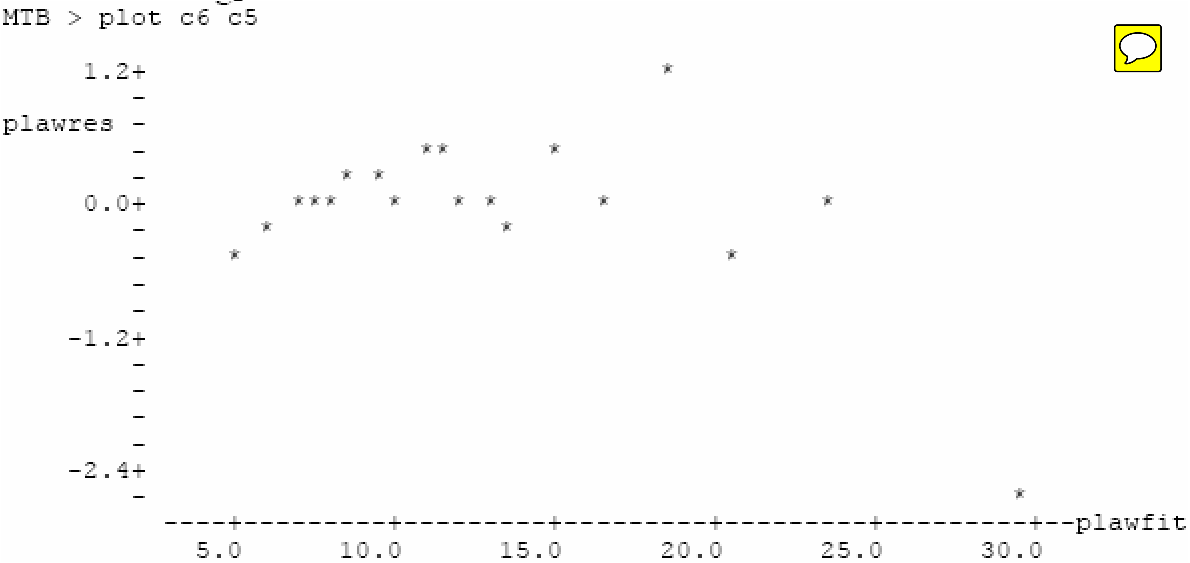
For sample $\ln Nsp = \hat{\alpha} + \hat{z} \ln A + error$

2. Execute analysis. Place data in model format:
Column with response variable, ln(*Nsp*).
Column with explanatory variable ln(*A*)

Code model statement, compute residuals and fits.
$\ln Nsp = \alpha + z \cdot \ln A + \varepsilon$

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

Plot residuals against fitted values.



```
MTB > plot c6 c5
```

3. Evaluate model. This looks promising at intermediate values.
Do Gleason's data follow (?) a power law at intermediate values (*i.e.*, without the largest and the two smallest areas).

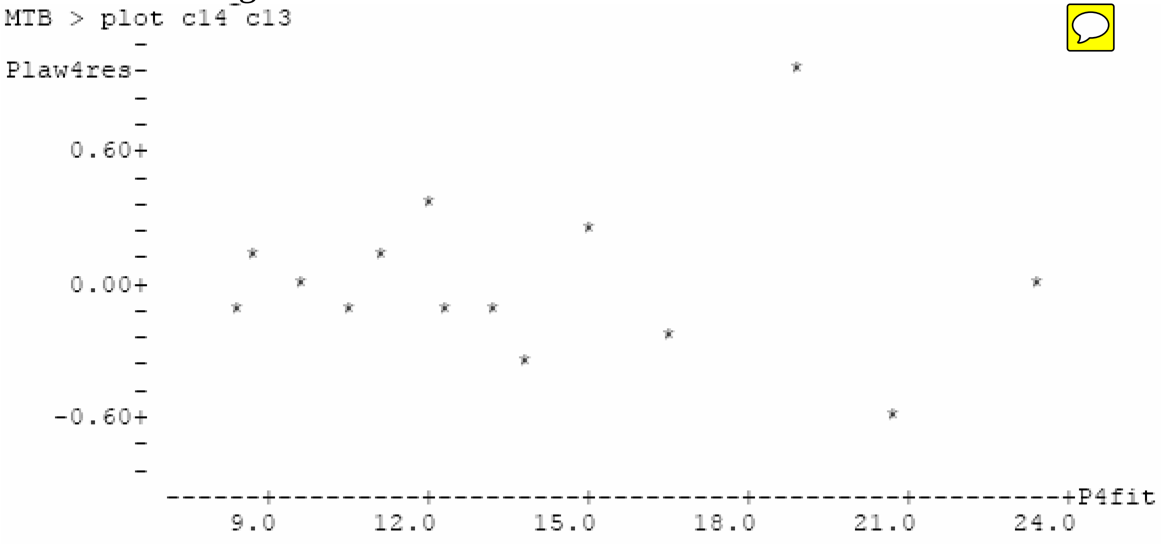
1. Arrhenius model

For population	$\ln Nsp = \alpha + z \cdot \ln A + \varepsilon$
For sample	$\ln Nsp = \hat{\alpha} + \hat{z} \ln A + error$

2. Execute analysis. Place data in model format:
Column with response variable, ln(*Nsp*)
Column with explanatory variable ln(*A*), where $3\ \text{m}^2 \leq A \leq 120\ \text{m}^2$

3. Evaluate the model.

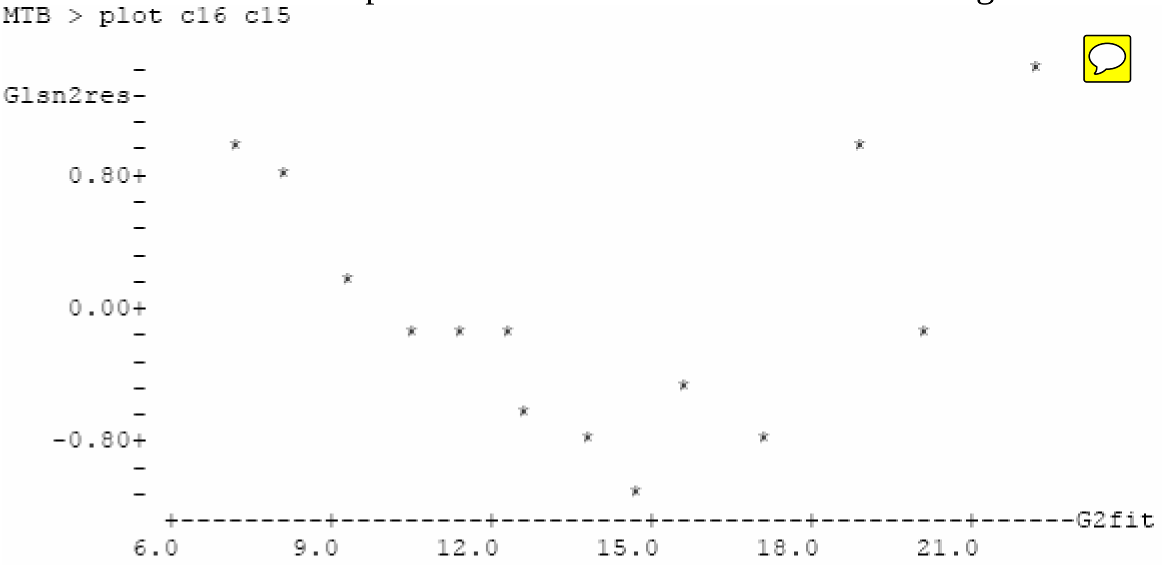
Plot residuals against fitted values.



```
MTB > plot c14 c13
```

Straight line acceptable? Yes, because no bowls or arches.
A power law is an acceptable model of Gleason’s data in areas ranging from 3 m^2 to 120 m^2.

Is Gleason’s model acceptable for Gleason’s data in the same range ?



```
MTB > plot c16 c15
```

Gleason’s model not appropriate for Gleason’s data.
Returning to the power law model for Gleason’s data we complete the analysis.

3. Evaluate the error model.

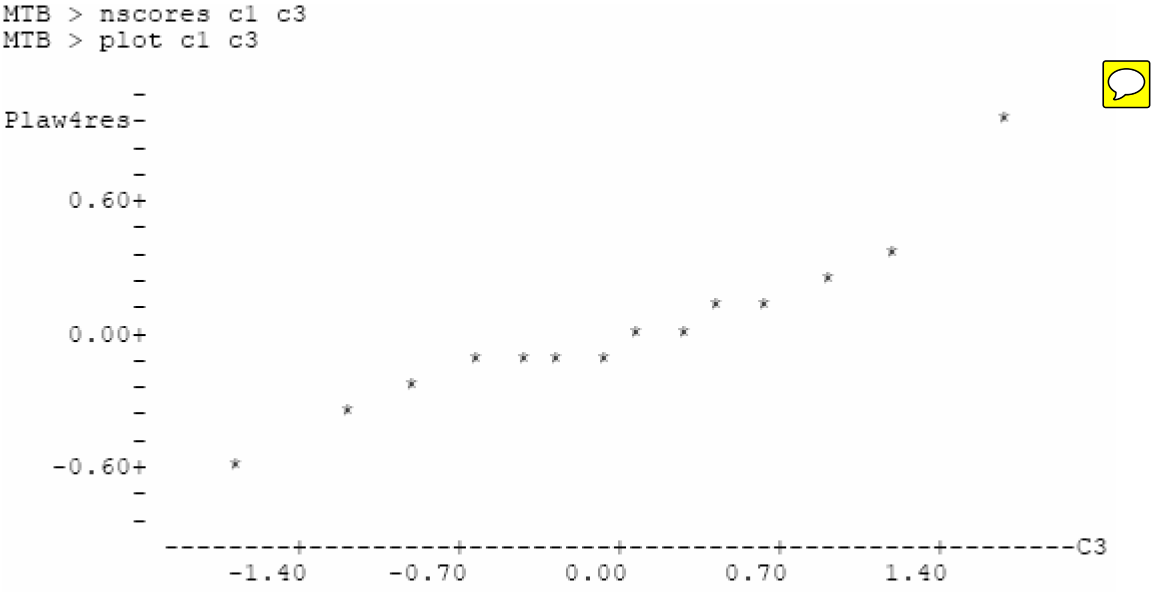
Homogeneity?

Plot of residuals versus fitted values shows little evidence of heterogeneity of variance.

Normal errors ?

```
 Histogram of Plaw4res   N = 14

 Midpoint   Count
     -0.6       1  *
     -0.4       1  *
     -0.2       4  ****
     -0.0       3  ***
      0.2       3  ***
      0.4       1  *
      0.6       0
      0.8       0
      1.0       1  *
```



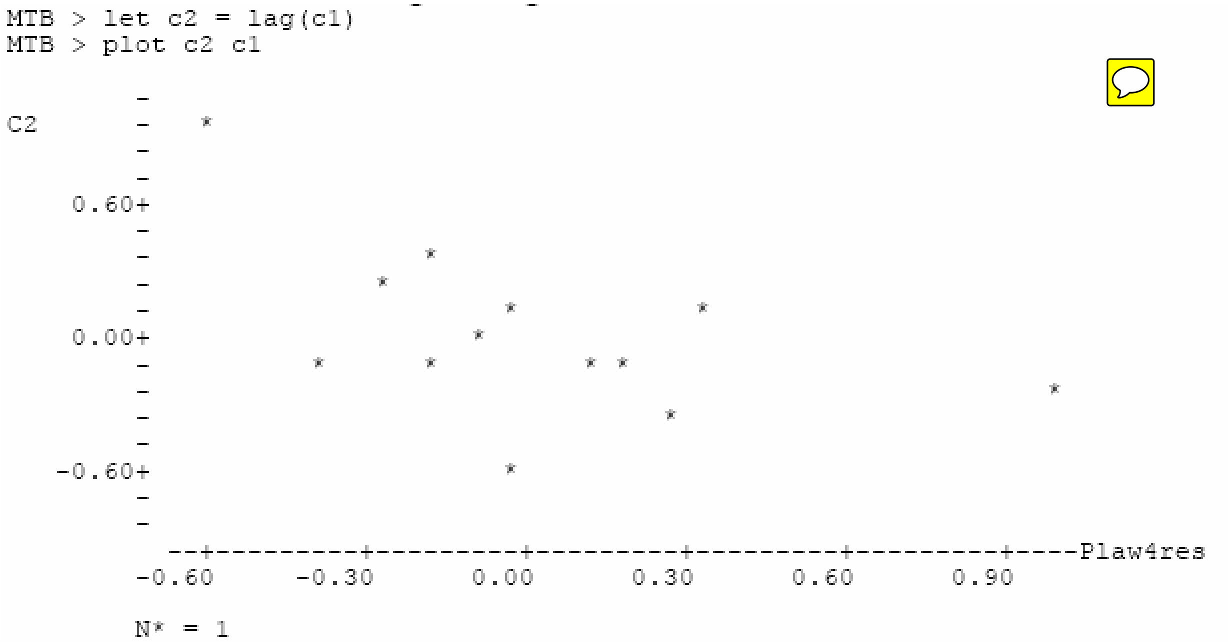
```
MTB > nscores c1 c3
MTB > plot c1 c3
```

Histogram close to symmetrical,
Probability plot shows some evidence of non-normal residuals but not severe.

Independent errors? Yes.

Plot of errors versus neighboring value.

3. Evaluate the error model.



```
MTB > let c2 = lag(c1)
MTB > plot c2 c1

        -
C2      -   *
        -
        -
    0.60+
        -
        -                 *
        -              *
        -                      *           *
    0.00+                    *
        -          *      *         * *
        -                                                        *
        -                                *
        -
   -0.60+                      *
        -
        -
          --+---------+---------+---------+---------+---------+----Plaw4res
        -0.60     -0.30      0.00      0.30      0.60      0.90

        N* = 1
```

There is a tendency toward negative correlation of errors, but it is slight.

4. State sample, population, and whether sample is representative.

Population is all possible measurements, given the measurement protocol.
The power law was meant to be general.

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. When log species number is plotted against log area there is no question that species number increases with area. Skip to step 10.

10. Analyze parameters of biological interest.

There was some evidence of non-normal residuals and perhaps some heterogeneity and non-normality of the residuals. The violations are slight, and so confidence limits will not be recomputed by randomization methods.

Compute confidence limits from standard deviation of the slope parameter.

```
MTB > regress ‘lnNsp' = ‘lnA' ;
 Predictor        Coef       Stdev    t-ratio        p
 Constant      1.59065     0.02166      73.45    0.000
 lnA          0.327485    0.006854      47.78    0.000

 s = 0.02384     R-sq = 99.5%     R-sq(adj) = 99.4%
```

10. Analyze parameters of biological interest.

GLM routine reports $\hat{z}$ = 0.327 ±0.006854

P{Lower $\leq z \leq$ Upper} = 1 – α = 95%

Lower = $\hat{z}$ – $t_{0.025[df]}$ * st.err.

Lower = 0.327 – 2.1788* 0.006854/sqrt(14) = 0.331

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

Upper = 0.327 + 2.1788* 0.006854/sqrt(14) = 0.323

Which hypotheses are excluded by CI ?
The CI excludes 1:1 relation of Nsp with Area.
Does the CI for Gleason's data include the conventional value ?
The conventional value is based on doubling of species number with 10 fold increase in area.
Arrhenius' Law

$$\frac{Nsp}{Nsp_{ref}} = \left(\frac{A}{A_{ref}}\right)^{z}$$

$$\frac{2}{1} = \left(\frac{10}{1}\right)^{z}$$

z = log(2)/log(10) = log(2)/1 = 0.3

The CI includes an exponent of 0.3, the conventional value of the exponent of the species area curve at this spatial scale.

The species area curve for Gleason's (1922) data is: $Nsp = e^{1.59065} A^{0.327}$

It is a curious irony that Gleason's data fit Arrhenius' power law better that Arrhenius' data. It is a further irony that Arrhenius data does not fit a power law (for herb-*Pinus* woodland). These conclusions are based on a powerful technique not known to either investigator, that of residual analysis.