Model Based Statistics in Biology.

Part III. The General Linear Model.

Chapter 9.3 Regression. Explanatory Variable Measured with Error.

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
SrBx1412.out
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 two examples, both with the explanatory variable X fixed, either by experiment or by definition of fixed classes.

Today:
Regression. Special case of the general linear model.
X variable measured with error.

Wrap-up

Regression a special case of the GLM.
When the explanatory variable is measured with error, parameters are estimated with bias, depending on the magnitude of the error.

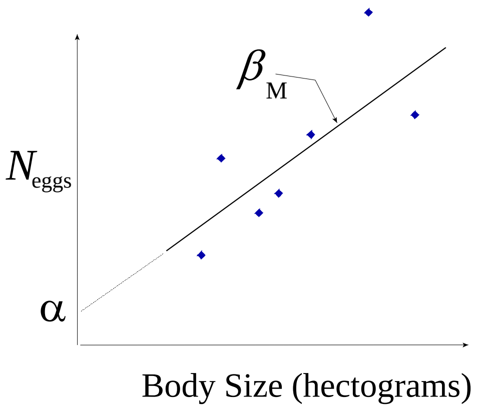
GLM, applied to regression

An explanatory variable measured with error results in biased estimates of regression parameters. Explanatory variables measured with error are common in observational studies, where there is often little opportunity to reduce error. An example (Box 14.12, Sokal and Rohlf 1995) is the number of eggs per female, in cabezon fish (*Scorpaenichthys marmoratus*) of several different sizes. For studies where the explanatory variable is measured with error, we need to consider the magnitude of the resulting bias.

Within a species we expect larger fish to produce more eggs than small fish. Of more interest is whether egg number increases in direct (1:1) proportion to body mass.

Once we frame the question in light of what we know, we find that the analysis sits uneasily within the conventional logic of rejecting the null hypothesis. The null hypothesis of no change in egg number with change in fish size is of little interest. Consequently, we will focus on the degree to which egg number changes with body size and on whether there is a 1:1 relation. We will focus on the analysis of parameters rather than on the machinery of hypothesis testing.

1. Construct the model



Verbal model.

Does egg number N_{eggs} depend on body mass M ?

Graphical model.

Draw picture of linear relation of N_{eggs} to M

Formal model.

Define variables.

Response variable is N_{eggs} the number of kiloeggs per fish (ratio scale)

Explanatory variable is M the body mass per fish, to nearest 100 grams (ratio scale)

Define symbols, units, type of measurement scale.

	Units	Dimensions	Type of measurement scale
N_{eggs}	kiloeggs	#	ratio
α	kiloeggs	#	ratio
M	hectograms	Mass	ratio
β_M	kiloeggs/hectogram	# M^{-1}	ratio

1. Construct the model

Write formal model.

For population $N_{eggs} = \alpha + \beta_M \cdot M + \varepsilon$

For sample $N_{eggs} = a + b_M \cdot M + error$

same as: $N_{eggs} = \hat{\alpha} + \hat{\beta}_M \cdot M + error$

2. Execute model. Place data in model format.

Data in two columns N_{eggs} and M

The model statement is used to execute the analysis in a statistical package.

$N_{eggs} = \alpha + \beta_M \cdot M + \varepsilon$

```
MTB> regress ‘Neggs’ 1 ‘Mass’
```

The typical results will be
parameter estimates and the anova table.

Estimate parameters and compute fitted values and residuals

Statistical packages report the parameter estimates as slope with intercept.

$N_{eggs} = 19.77 + 1.87\,M$

The package first estimates the parameters of the general linear model: $\hat{\beta}_o$ and $\hat{\beta}_M$

$N_{eggs} = \hat{\beta}_o + \hat{\beta}_M \cdot (M - \bar{M})$ + res general linear model for sample

The estimates are:

$\bar{M}$ = mean(M) = 30.36 hectograms,

$\hat{\beta}_o$ = mean(N_{eggs}) = 76.545 kiloeggs ,

$\hat{\beta}_M$ = 1.87 kiloeggs/hectogram = slope of line that minimizes vertical deviations

$\hat{\alpha} = \hat{\beta}_0 - \hat{\beta}_M$ (mean(M) = 76.545 – 1.87(30.36)

$\hat{\alpha}$ = 19.77 kiloeggs

To get the "best" estimate of the parameters of the regression line, we fit a line through the mean of all the data $\hat{\beta}_o$ = mean(N_{eggs}) and mean (M). We use this point because it is the best estimated point on the graph. We don’t make an estimate at the y-intercept = $\hat{\alpha}$, where the data are usually too sparse to obtain a good estimate.

There are several ways of estimating the slope β_M. Texts on mathematical statistics describe the methods. A common and widely accepted method is to estimate β_M by minimizing the sum of the squared deviations of the data points from the line. Statistical packages use standard formulae to make these estimates.

GLM routines report fits and residuals if requested.
These are computed from parameter estimates.

3. Evaluate structural model.

Downward bias on the parameter estimate. Because this is an observational study where the explanatory was measured with error, we will evaluate the resulting downward bias on the parameter β_M. The model is:

$$N_{eggs} = \alpha + \beta_M \cdot M + \varepsilon$$

$$M^* = M + \varepsilon^*$$

Where ε^* is the measurement error.

If ε, ε^*, and M^* are normally and independently distributed the regression coefficient $\beta_M{}^*$ will be smaller than β_M by a factor k.

$$\beta_M{}^* = k \cdot \beta_M \qquad k = \sigma^2_M / (\sigma^2_M + \sigma^2_{M*})$$

The factor k is called the reliability ratio, or sometimes just reliability. It is always less then unity. It describes the degree to which the true relation β_M is based downward by measurement error.

In this example we have no independent estimate of the measurement error ε^* but we can make a rough estimate by considering resolution used in measuring the variable M. The error is presumably no worse than 1 hectogram and hence the standard deviation will be less than 1 as well.

We take var(M) as an estimate of the true variance ($\sigma^2_M + \sigma^2_{M*}$)

$\sigma^2_M + \sigma^2_{M*} = \text{var}(M) = 78$

$\sigma^2_{M*} < 1$

$\sigma^2_M > 77$

$k > 77/78 = 0.987$

Downward bias due to measurement error is not a large concern with this data.

Next we evaluate the straight line assumption, using the residual vs fit plot.

```
MTB > plot c4 c3
        -
        -
res     -   *
      12+                                      *
        -
        -
        -                                        *           *
        -                        *
       0+                      *                       *
        -                                                       *
        -
        -                                                     *
        -
     -12+
        -         *
        -                           *
        -
          ------+---------+---------+---------+---------+---------+fits
               50        60        70        80        90       100
```

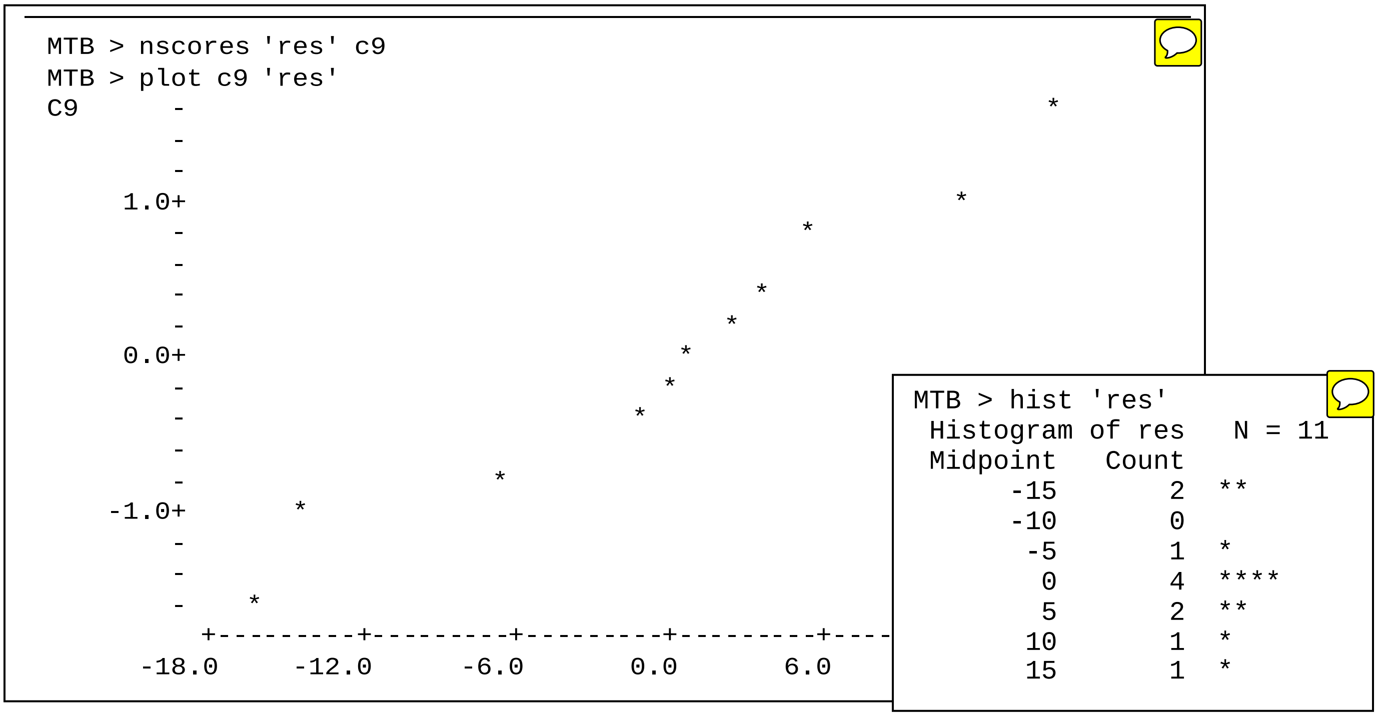
No arches or bowls. So linear model is acceptable.

3. Evaluate error model.

Next we evaluate the error model (homogeneous, normal, independent errors).
We used the normal error model to estimate parameters. We will use it again when computing confidence limits or Type I error in hypothesis testing.

First assumption: homogeneous errors? The residual vs fit plot shows that dispersion of the residuals was slightly less at large fitted than at small fitted values. However, this is minor. There is no convincing evidence of heterogeneity.

Second assumption: Residuals normal?
The residuals are not symmetrically distributed around zero in the histogram.
The nscores versus residual plot departs somewhat from straight line.



```
MTB > nscores 'res' c9
MTB > plot c9 'res'
C9      -                                                        *
        -
        -
     1.0+                                                  *
        -                                        *
        -
        -                                     *
        -                                   *
     0.0+                                *
        -                               *
        -                             *
        -
        -                    *
    -1.0+       *
        -
        -
        -    *
          +---------+---------+---------+---------+----
      -18.0     -12.0      -6.0       0.0       6.0
```

```
MTB > hist 'res'
 Histogram of res    N = 11
 Midpoint   Count
      -15       2  **
      -10       0
       -5       1  *
        0       4  ****
        5       2  **
       10       1  *
       15       1  *
```

Residuals show some deviation from normal.

Third assumption: Independent errors ?
We have no information on temporal sequence or on spatial arrangement of samples to evaluate this assumption. If residuals are small at small fitted values, and increase at larger fitted values, then residuals are potentially non-independent. We evaluate this by ordering the observations from small to large, copying residuals from one column into an adjacent column, then plotting the residuals against the adjacent column.

3. Evaluate error model.

Third assumption: Independent errors ?

```
MTB > let c8 = lag('res')
MTB > name c8 'lag1_res'
MTB > plot c8 'res'
lag1_res-
        -
        -         *
      12+                                       *
        -
        -
        -                      *          *
        -      *
       0+                                     *    *
        -
        -
        -                               *
        -
     -12+
        -                                  *
        -                                                    *
        -
          --+---------+---------+---------+---------+---------+----res
        -18.0     -12.0      -6.0       0.0       6.0      12.0
```

The plot shows some evidence of a trend downward. There is weak evidence of non-independence.

Fourth assumption: Errors sum to zero? No need to check this because statistical packages produce parameter estimates where the residuals sum to zero.

4. State sample, population, and whether representative.

All cabezon fish ? 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 11 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 question is whether relation of fish egg number deviates from a 1:1 relation with fish body size. Rather then accepting/rejecting the null hypothesis, we will examine the parameters and compute confidence limits (skip to step 10).

10. Examine parameters of biological interest.

The purpose of this analysis was to estimate parameters in a situation where a relation was expected to exist based on the biology of fish. Our research hypothesis was that egg number increases with body size, perhaps in 1:1 proportion with body size. Hence confidence limits are more appropriate than hypothesis testing.

Compute confidence limits around estimate β_M so as to include true value β_M 95% of time.

$s_b^2 = s_{y.x}^2\ \Sigma/x^2 = (103.0962/932.55) = 0.1106$

s_b = square root of s_b^2 = sqrt(0.1106) = 0.3325 kiloeggs/hectogram

L = Lower limit = $\hat{\beta}_M - t_{\alpha/2[\nu]}s_b$

U = Upper limit = $\hat{\beta}_M + t_{\alpha/2[\nu]}s_b$

for 95% limits use $t_{0.05/2[9]}$ because df = 9 = ν

```
MTB > invcdf .025;
SUBC> t 9.
      .025   −2.2622
```

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

```
MTB > invcdf .975;
SUBC> t 9.
      .975    2.2622
```

Some tables give both tails of the t-distribution e.g. Rohlf and Sokal give $t_{0.05[9]} = 2.622$

L = 1.87 – (2.2622)(0.3325) = 1.12 kiloeggs/hectogram

U = 1.87 + (2.2622)(0.3325) = 2.62 kiloeggs/hectogram

We assumed that the minor violations of the assumptions would have little effect on the confidence limits computed from a t-distribution. To check this judgment the confidence limits were computed by randomization. To do this, the errors are randomly assigned to the fitted values, producing new ‘observed’ values. These values were then regressed against the explanatory variable to obtain a randomized estimate of $\hat{\beta}_M$. This was repeated, to accumulate thousands of randomized estimates. The confidence limits were then identified as the values of $\hat{\beta}_M$ that encompass 95% of the estimates from randomization.

8000 randomizations. 200 (2.5%) were less than 1.28 kiloeggs/hectogram
200 (2.5%) were greater than 2.48 kiloeggs/hectogram

The confidence limits via randomization, which are free of assumptions, were somewhat narrower than the confidence limits from the t-distribution. Our judgment was correct, that violations of assumptions were minor, with little effect on the calculation of the confidence limits from the t-distribution.

10. Examine parameters of biological interest.
Report conclusions about parameters, with evidence.

The confidence limits do not include zero and so we can reject the null hypothesis of no relation. The null hypothesis in this case is of little interest because we expect a large fish to produce more eggs that a small fish of the same species.

Of more interest is that the confidence limits exclude a 1:1 ratio of egg number to body mass. We conclude that in this species, large fish invest disproportionately more in eggs (per unit of body mass) than do small fish.

We report the regression equation.

$$N_{eggs} = 19.77 + 1.87\, M$$

We also report the confidence limits, which show that large fish produce more eggs (per unit of body mass) than small fish ($\hat{\beta}_M > 1$)

$\hat{\beta}_M = 1.87$ with 95% confidence limits of 1.28 to 2.48 kiloeggs/hectogram

We report the confidence limits from randomization, having made the effort to calculate limits that are free of assumptions.

Extra material

Texts (e.g Sokal and Rohlf 1995) contain several methods for regression when the explanatory variable is measured with error.

One of the most common is

reduced major axis regression.	kiloEggs = 12.19366 + 2.11937*'Wt'
Others are major axis regression	kiloEggs = 6.65668 + 2.30173*'Wt'
Bartlett's 3 group regression,	kiloEggs = 21.89091 + 1.80000*'Wt'
Kendall's robust regression.	kiloEggs = 26.68421 + 1.68421*'Wt'

Another example where explanatory variable measured with error:
Does methyl mercury meHg in the blood (ng/g) depend on methyl Hg intake from fish (µg/day) ? Daniel (1995 p 408)

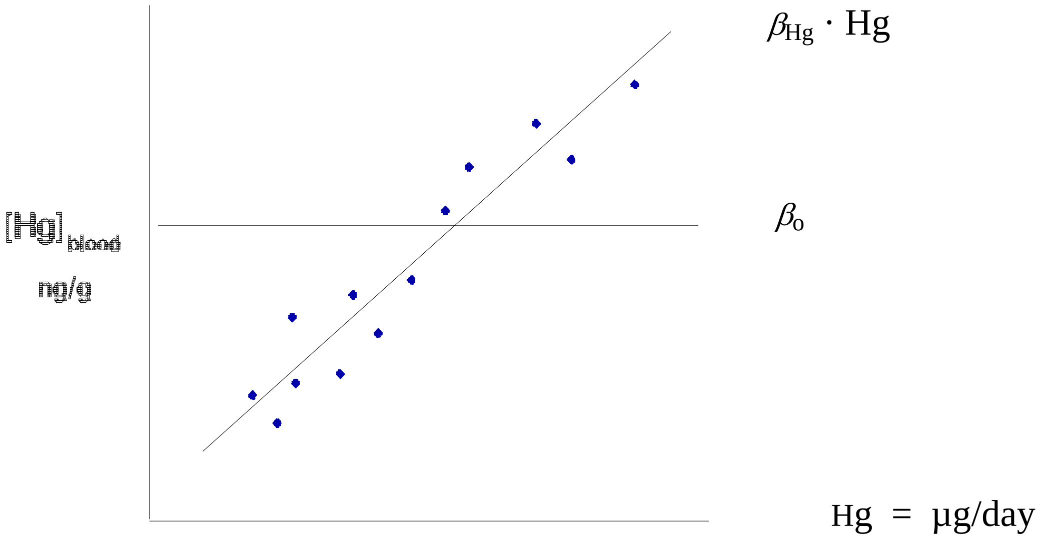


Fig L13c

This is evaluated with a model: $[Hg]_{blood} = \beta_o + \beta_{Hg} * Hg + \varepsilon$

H_A: $\beta_{Hg} > 0$ (ie term retained in the model)

H_o: $\beta_{Hg} = 0$ (ie term not in model and $[Hg]_{blood} = \beta_o$

How does the variance explained by the model compare to the unexplained or residual variance $Var(\varepsilon)$?
$F = Var(\beta_{Hg} \cdot Hg) / Var(\varepsilon)$