Statistical Science.

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.
Explanatory variable measured with error.

Wrap-up

Regression is 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- Regression where the explanatory variable is measured with error.

Explanatory variables measured with error are common in observational studies, where there is often little opportunity to reduce error. An explanatory variable measured with error can result in biased estimates of the slope parameter β_X. We need to consider the magnitude of the bias when the explanatory variable is measured with error.

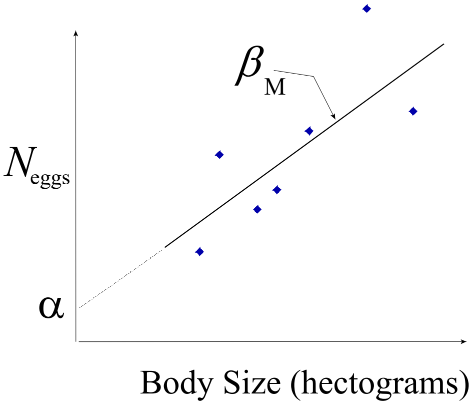
Fish egg example.

KiloEggs	Wt hectograms
61	14
7	17
65	24
69	25
54	27
3	33
87	34
89	37
100	40
90	41
97	42

The example is the number of eggs per female, in cabezon fish (*Scorpaenichthys marmoratus*) of several different sizes (Box 14.12, Sokal and Rohlf 1995).

We expect larger fish to produce more eggs than small fish. Once we frame the question in light of the biology we find that the analysis sits uneasily within the conventional logic of rejecting the null hypothesis. Rejecting the null hypothesis of no change in egg number with change in fish size is of little interest. Instead, we will focus on more plausible hypothesis, that egg number increases with body size in 1:1 relation.

1. Construct the model



Verbal model.

Egg number N_{eggs} increases with body mass.

Graphical model.

The simplest model is a 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

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 are already in model format: two columns N_{eggs} and M

The model statement is used to execute the analysis in a statistical package. The code we use will have a similar structure in any package.

Here are two examples, one in Minitab, one in R.

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

```
MTB> regress 'Neggs' 1 'Mass';
```

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

```
LM.Negg <- lm(Negg ~ Mass)
plot(LM.Negg)
anova(LM.Negg)
summary(LM.Negg)
```

2. Execute model. Estimate parameters.

Statistical packages report the parameter estimates as slope with intercept.

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

The package 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

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 obtain 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 non-existant or 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 used method is a formula that gives the estimate β_M by minimizing the sum of the squared deviations of the data points from the line. This is equivalent to maximum likelihood estimate when using a normal error.

2. Execute model. Compute fitted values and residuals.

Statistical packages typically produce parameter estimates and the ANOVA table. Diagnostic plots for the residuals need to be requested.

3. Evaluate model for downward bias.

Because this is an observational study where the explanatory was measured with error, there will be downward bias on the parameter β_M . The magnitude depends on the size of the error. We have no estimate of the total measurement error ε^* in this case. To illustrate the calculation we will use a component of measurement error that is always present, the absolute error, defined as half the resolution of the explanatory variable. In this case the absolute error is half a hectogram.

The model is:

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

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

Where ε^* is the absolute 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.

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

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 biased downward by measurement error.

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

$\sigma^2_{\mathrm{M}} + \sigma^2_{\mathrm{M*}} = 93.25 = \mathrm{var}(M)$

$\sigma^2_{\mathrm{M*}} = 0.5^2$

$\sigma^2_{\mathrm{M}} = 93$

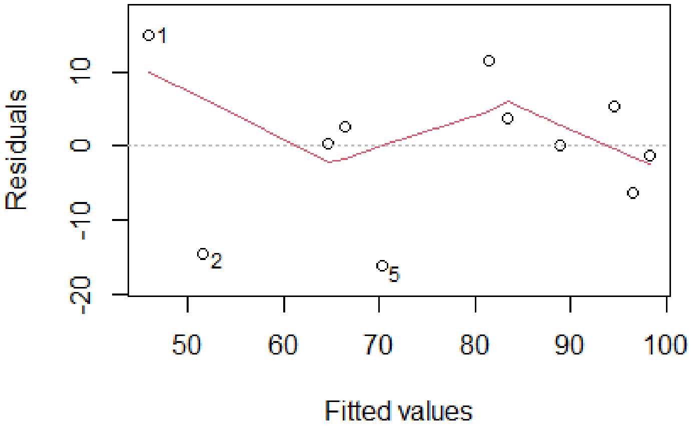
$k = 93/93.25 = 0.997$ on average.

Downward bias due to absolute error is rarely of any concern.

3. Evaluate structural model.

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

No arches or bowls.
The straight line model is acceptable.



3. Evaluate error model.

We then evaluate the error model assumptions: homogeneous, normal, and independent errors.

<u>First assumption: </u>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.

3. Evaluate error model.

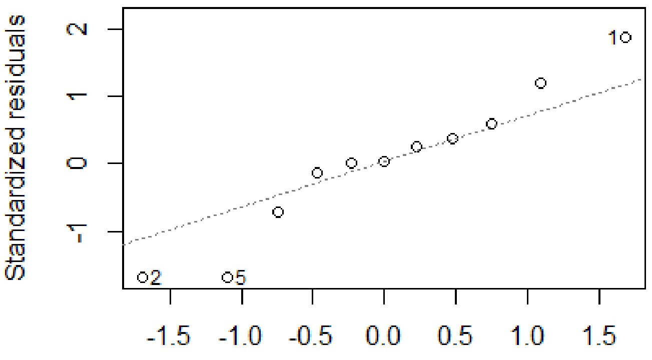
Second assumption:

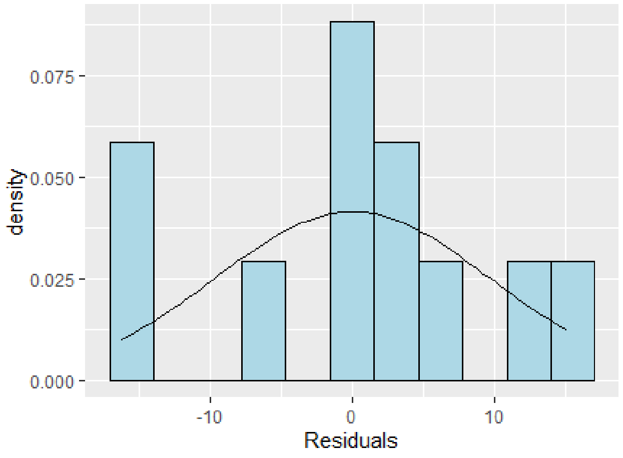
Residuals normal?

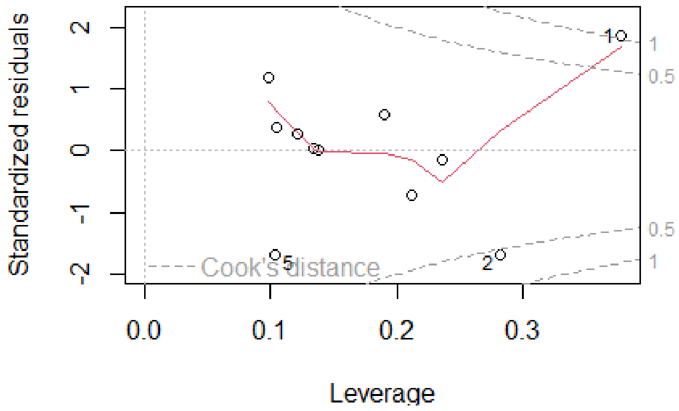
The QQ plot shows outliers.

The normal curve overlay on the histogram identifies the same outliers. These outliers are far to the left, away from the mean value of zero. They have considerable leverage on the parameter estimates. The outlier to the far left of the histogram has a leverage of 1, as measured by Cook's D statistic.

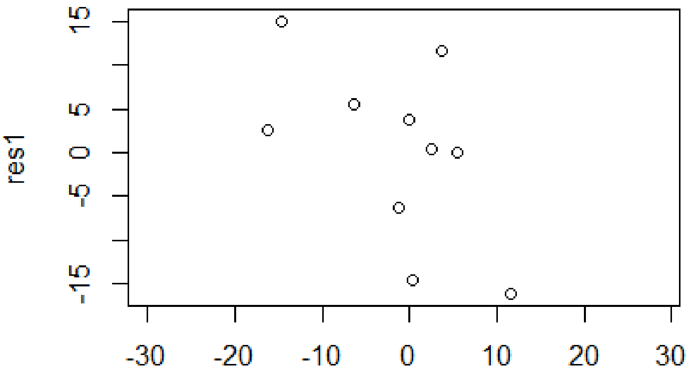
Values of 1 or above are of concern.







Third assumption: Independent errors ?



We have no information about temporal sequence, spatial arrangement of samples, or other common sources of non-independence. If we had this information we would order the observations (and hence the residuals) by temporal sequence or by proximity. The data are listed from small fish to large. To evaluate independence with respect to fish size we copy the residuals to an adjacent column, offset downward by one, so that neighbors are matched by row. The plot shows that residuals are independent. The plot shows no strong upward or downward trends.

Fourth assumption: Errors sum to zero? We do not need to check this because statistical packages produce parameter estimates where the residuals sum to zero.

Conclusion: Errors are homogeneous with substantial deviation from normal distribution due to an outlier.

4. Partition df and SS according to model.

Calculate df according to rules

Obtain sums of squares from ANOVA table.

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

$N_{eggs} = 19.77 + 1.87 \cdot M + \varepsilon$

$11\text{-}1 = 1 + 9$

$4188.73 = 3260.86 + 927.87$

Source	df	SS
Wt	1	3260.86
Residual	9	927.87
		4188.73

4. State the full (null) and reduced (alternative) model pair.

Full model H_o: $\beta_M = 1$

Reduced (alternative) model H_A: $\beta_M > 1$

4. Calculate likelihood ratio for omnibus model.

LR = $(927.87/4188.73)^{-11/2} = 3984$

The slope of 1.87 kiloeggs/hectogram is 3984 times more likely than no relation.

A test against the 1:1 ratio is of more interest than a test against zero slope.

5. State sample, population, and whether representative.

All cabezon fish ? Probably not.

All fish from a finite (enumerable) population? Almost certainly not.

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

This is a more realistic statement of the population.

Is the sample representative?

Random sample in this case has high cost, the sample was likely haphazard.

We'll assume a haphazard sample is unbiased with respect to egg number in relation to fish biomass

This is a more restrictive statement of the population than "all fish."

This is a <u>hypothetical</u> or <u>notional</u> population.

So a hypothetical population, based on repeatable protocol, will be used.

All measurements that could have been made on 11 fish by this protocol?

This is more restrictive and more defensible.

It leaves aside the question of whether these fish are representative of other fish.

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. The null hypothesis (no relation of egg number to fish body mass) is hardly plausible for fish. Rather than rejecting an implausible hypothesis, we will report parameter estimates with 95% confidence limits to evaluate $\beta_M = 1$.

10. Evaluate 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, plausibly in 1:1 proportion with body size. Confidence limits are more informative than hypothesis tests against a single value.

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

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

sy.x2 = $MS_{residual} = SS_{residual}$ / df = 927.89 / 9 = 103.096 (see step 4)

$\Sigma x^2 = \Sigma(M- \bar{M})^2 = 932.55$

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

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

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

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

```
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 $t_{0.05[9]} = 2.622$

Lower limit = 1.87 − (2.2622)(0.3325) = 1.12 kiloeggs/hectogram
Upper limit = 1.87 + (2.2622)(0.3325) = 2.62 kiloeggs/hectogram

From the residuals we judged that violation of the assumption of normality was potentially serious due to the leverage of an outlier. To evaluate the influence of the outlier we generate an empirical distribution of estimates of the slope parameter. First, 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 except that of representative sample, were somewhat narrower than the confidence limits from the t-distribution. The outlier had little effect on our measure of uncertainty, the confidence interval.

10. Evaluate parameters of biological interest.

The confidence limits exclude the (implausible) null hypothesis of no relation. Of more interest is that the confidence limits exclude a 1:1 ratio of egg number to body mass. The evidence supports an estimate greater than 1:1. In other words disproportionately more eggs (per unit of body mass) in large than in small fish.

We report the regression equation with confidence limits on β_M

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

95% confidence limits of 1.28 to 2.48 kiloeggs/hectogram

We report the evidence (LR = 4 x 10^3) along with a measure of uncertainty, the confidence limits from randomization.

* * *

Extra material

Texts (*e.g* Sokal and Rohlf 2012) contain several methods for regression when the explanatory variable is measured with error.

One of the most common is

Reduced major axis regression	kiloEggs = 12.19 + 2.12*Mass

Others are

Major axis regression	kiloEggs = 6.66 + 2.30*Mass
Bartlett's 3 group regression,	kiloEggs = 21.89 + 1.80*Mass
Kendall's robust regression.	kiloEggs = 26.68 + 1.68*Mass

Some of these methods persist in widely used statistical packages and still appear in the published literature. These methods address, in different ways, the problem of an explanatory regression variable measured with error. These methods produce different (in some cases disturbingly different) estimates of the parameter of interest kiloeggs/hectogram of body mass. These methods were developed during a period of limited computationl power, in the context of null hypothesis testing rather than parameter estimates with the best statistical support, as from a likelihood ratio. They were never intended for use in applied contexts, such as the calculations that underpin fisheries management.

GLM applied to regression where the explanatory variable is measured with error.

Fish mercury example.
Here is another example where the explanatory variable is measured with error.

Methyl mercury is an environmental neurotoxin that results in loss of physical coordination, difficulty in speech, narrowing of the visual field, hearing impairment, blindness, and in extreme cases, death. Chronic exposure to mercury vapor was an occupational hazard of hatters in 17th century France and England (Mad hatter disease). Methyl mercury poisoning of fishing families in Minamata, Japan in the mid-20th century (Minamata disease) brought worldwide attention to the problem to bio-amplification of fat soluble food chain contaminants such as methyl mercurye (meHg) and DDT. Consumption of fish from hydro reservoirs is a continuing health risk in the 21st century, because impoundment of lakes results in fish with high levels of meHg. Young mothers are strongly advised against consuming fish from reservoirs, as their milk contains high levels of meHg if they consume large number of fish from reservoirs.

One link in the causal change from fish consumption to health risk is passage of meHg from food to the blood. Daniel (1995 p 408) reports methyl mercury meHg in the blood (ng/g) relative to methyl Hg intake (µg/day). Here are the data.

Hg intake µg/day	Hg in whole blood ng/g
100	90
200	120
230	125
410	290
600	310
550	290
275	170
580	375
105	70
250	105
460	205
650	480

Graph the relation of Hg in the blood to Hg intake.

Construct a model to quantify the relation of Hg in the blood to Hg intake

Report parameter estimates, with their units.

Is downward bias due to measurement error a concern with this data?

Evaluate the linearity assumption as a model of blood Hg to Hg consumption.

Evaluate the homogeneity and normality assumptions for the error model.

Obtain confidence limits for the estimate of the regression coefficient.

Can you exclude the null hypothesis of no relation?

Can you exclude the hypothesis of a 1:1 relation of Hg in the blood to Hg in food?

Following impoundment of Cat Arm lake to a reservoir in Newfoundland in 1982, the meHg level in fish (trout and charr) rose from less than 0.2 µg/g fish to 0.5 µg/g fish (Brook trout) and 0.8 µg/g in Arctic charr (Environmental Pollution 101: 33-42). Calculate the intake of meHg at a consumption rate of 200 g/day of fish protein. The CFIA (Canadian Food Inspection Agency) standard for mercury in fish is 1 ppm (1µg/g) of fish.