Model Based Statistics in Biology.

Part III. The General Linear Model.

Chapter 9.2 Regression. Explanatory Variable Fixed into Classes

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
PrsnLee.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) Regression is a special case of GLM

Yesterday, we looked at regression of a response variable against an explanatory variable that was fixed by experimental manipulation.

Today:
Regression. Special case of the general linear model.
X variable from observational study, rather than experimental study.
Work through a generic recipe to illustrate the use of the general linear model.

Wrap-up

Regression a special case of the GLM.
This example was similar to previous, except that x variable fixed into classes.
Number of families per class differs, so means based on large number of families give n more weight than means based on less information, hence poorly estimated.

GLM, applied to regression X variable fixed into classes. Observational study.
Example. Galton's Law

The quantity of interest is the stature (height) of sons in relation to stature (height) of their fathers.

What is the relation of height of offspring to parents? How heritable is this trait ?

The data were collected by Francis Galton at end of the 19th century.
In 1903 K. Pearson and A. Lee reported the data, with analysis.

Pearson, K., A. Lee. 1903. On the laws of inheritance in man.
I. Inheritance of physical characters. *Biometrika* 2: 357-462.

This was the first application of regression, a method that Pearson invented to analyze Galton's data. Galton found that the height of sons 'regressed' toward the mean value of fathers in a height class.

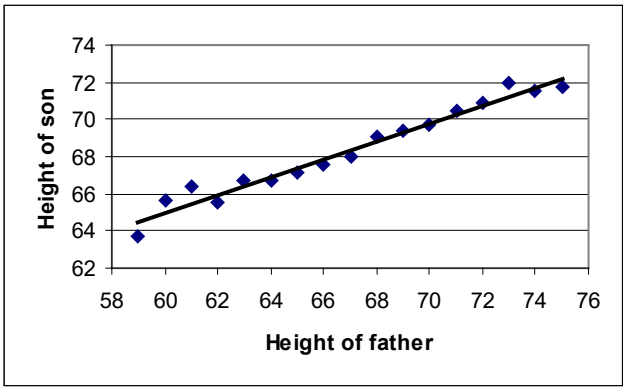
Hfather	Hson	Nfamily
59	63.667	3.0
60	65.643	3.5
61	66.344	8.0
62	65.559	17.0
63	66.679	33.5
64	66.740	61.5
65	67.186	95.5
66	67.606	142.0
67	67.951	137.5
68	69.078	154.0
69	69.385	141.5
70	69.744	116.0
71	70.497	78.0
72	70.872	49.0
73	72.000	28.5
74	71.500	4.0
75	71.727	5.5
		1078

1. Construct model

First, the verbal model:

There is a positive relation between heights of sons and fathers.

Next, the graphical model, a straight line.



Formal model. To construct this, we begin by distinguishing the response from the explanatory variable.

Hson Response variable is height of sons, in inches, from 1078 families
Hf Explanatory variable is height of father, in inches
Nfam Number of families at each stature interval

The data are taken from Table 22 in Pearson and Lee (1903).

1. Construct the model

This is an observational study in which the measurements on fathers and sons are both made with error. However the data were grouped into fixed size classes of the explanatory (independent) variable, height of fathers. This reduces the measurement error in the explanatory variable substantially because of the large number of fathers in each size class. To reduce this variability, we take the explanatory variable (heights of fathers) as the class mark (the midpoint of each category).

This example differs from the previous example in having an explanatory variable in fixed classes, rather than an explanatory variable at levels fixed by an experiment.

Symbol	Units	Dimensions	Notation
Hson	inches	Length [L]	Roman: observed values
	same as *Hson*	same as *Hson*	Greek: parameter
Hf	inches	Length [L]	Roman: observed values
β_{Hf}	none (in/in)	none [L/L]	Greek: parameter

Write formal model (write the GLM).

For population: $Hson = \alpha + \beta_{Hf} \cdot Hf + \varepsilon$

For sample: $Hson = \hat{\alpha} + \hat{\beta}_{Hf} \cdot Hf + \varepsilon$

same as: $Hson = a + b_{Hf} \cdot Hf + e$

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

Data column with response variable, *Hson*.
Data column with explanatory variable *Hf*
Data column with weights *Nfamily*

Use model statement in statistical package to code the GLM model

$Hson = \alpha + \beta_{Hf} \cdot Hf + \varepsilon$

```
MTB > GLM 'Hson' =      'Hf' ;
SUBC> weights 'Nfamily'.
```

Many packages have a graphics interface that assists in constructing the model (Minitab, SPSS). If you are using the graphics interface, you may want to look at the code produced by the interface, so that you understand how the model you wrote translates into a model statement in your package.

In this example we use a weight command that takes into account the different number of cases at each value of of the explanatory variable *Hf*. Means based on a large number of families are given more weight than means based on less information. The data column *Nfamily* has the number of families at each value of *Hf*.

2. Execute the analysis. Compute fitted values and residuals.

Model based routines calculate residuals and fits as output.

Here are examples from Minitab.

```
MTB > GLM Hson = Hf;
SUBC> weights c3;
SUBC> fits c4;
SUBC> res c5.
```

```
MTB > regress c2 1 c1;
SUBC> weight c3;              #weighted by number of cases
SUBC> fits c4;
SUBC> res c5.

The regression equation is
Hson = 33.3 + 0.523 Hfather   #slope is 0.5
                              #stature regresses --> mean

Predictor       Coef       Stdev    t-ratio        p
Constant      33.284       1.643      20.26    0.000
Hfather      0.52254     0.02425      21.55    0.000
```

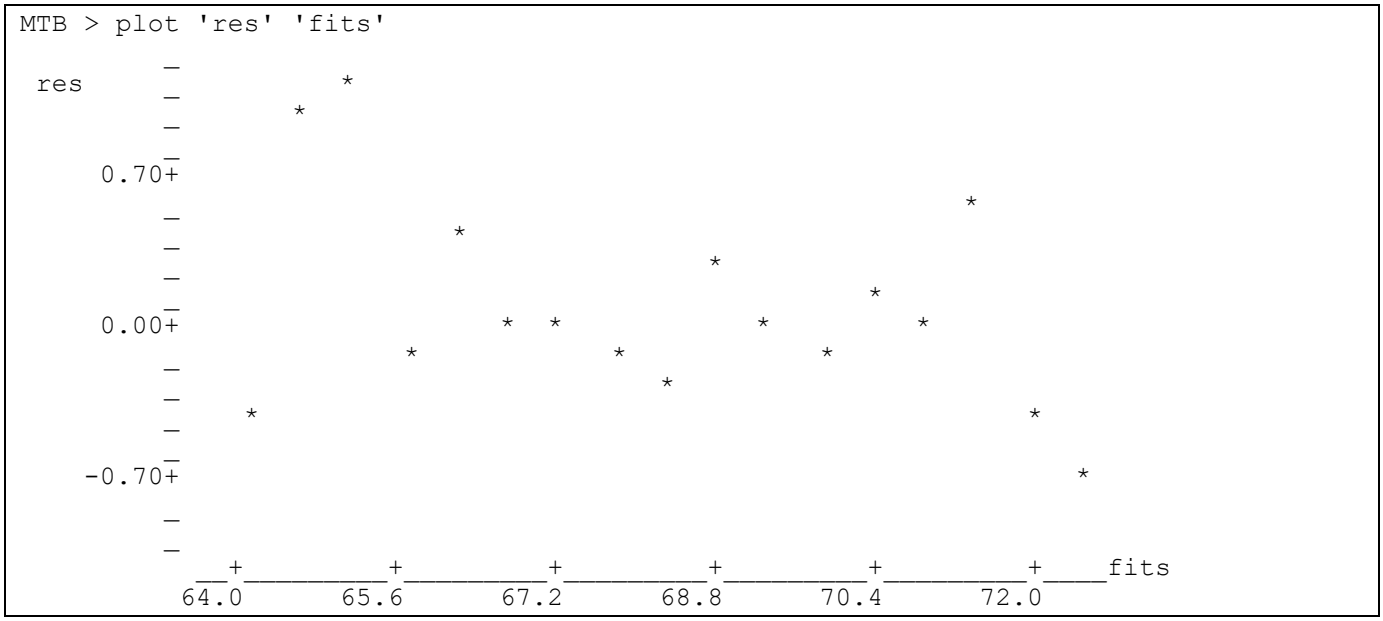
The residuals are calculated from fitted values, which were calculated from the parameter estimates.

Fitted values: Fits = E[*Hson*] = $\hat{\alpha} + \hat{\beta}_{Hf} \cdot Hf$

Residuals: Res = *Hson* – Fits

3. Evaluate the structural model (the regression line).

Plot the residuals against fitted values. Whenever we fit a line we evaluate whether there is some pattern of deviation from the line.



The residuals show a downward tendency at large values,
However, they do not show simple bowls or arches.
The straight line model is acceptable.

3. Evaluate the error model.

Next, we evaluate the error model (homogeneous, normal, and independent errors).

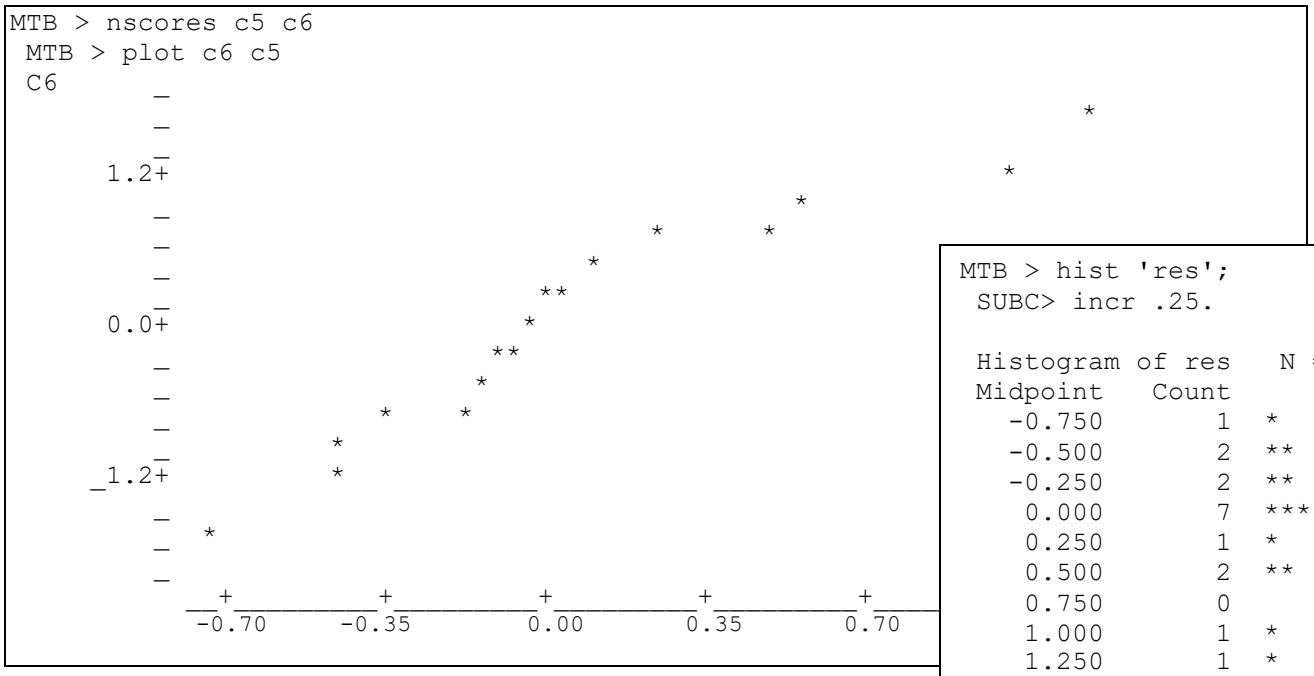
It is commonly believed that we check our assumptions before doing an analysis. This sounds like good practice but it turns out to be wrong. The assumptions apply to the errors. We cannot evaluate assumptions until we have the errors calculated.

Are the errors homogeneous?

The plot of residuals versus fitted values suggests that variance might be larger at low values than at intermediate values. But this is at best weak, compared to what we will see when assumption of homogeneity has been seriously violated.

Are the errors normal?

When the response variable consists of means, we expect the residuals to be normally distributed. As expected the residuals are close to normal. Both the histogram and the normal score plots show a slight tendency toward too many values clustered near zero, too few in the adjacent values.




```
MTB > hist 'res';
 SUBC> incr .25.

 Histogram of res   N = 17
 Midpoint   Count
   -0.750       1  *
   -0.500       2  **
   -0.250       2  **
    0.000       7  *******
    0.250       1  *
    0.500       2  **
    0.750       0
    1.000       1  *
    1.250       1  *
```

Are the errors independent?

We have no information on temporal order or spatial layout of samples, to evaluate this assumption. Observations can have non-independent errors for many reasons. In contrast, we do not expect means from groups to have non-independent errors. So we judge that the error model are both acceptable.

4. Partition df and SS according to model.

$$Hson = \alpha + \beta_{Hf} \cdot Hf + \varepsilon$$

$$17\text{-}1 = 1 + 15$$

$$2248 = 2177.9 + 70.34$$

4. Calculate likelihood ratio for omnibus model

Full model / reduced model = 2248/70.34 = 31.96

$LR = 31.96^{17/2} = 6 \times 10^{12}$

The reduced model is 6×10^{12} more likely than no relation.

The likelihood ratio puts a number to the obvious, when looking at the graph of mean heights of sons in relation to fathers. Heights of sons increase in linear fashion as the heights of fathers increase.

5. State population and whether sample is representative.

The population is all possible measurements, given the measurement protocol, if we repeated the study thousands of times. We will infer to a population consisting of thousands of runs of the same experiment, using the same protocol.

What if we ran the study elsewhere in the world, rather than just England? From the title of the publication the authors were prepared to infer to all people in the world.

5. Decide on mode of inference. Is hypothesis testing appropriate?
If yes step 6 (frequentist inference). If no, step 10 (evidentialist inference).

Hypothesis testing is appropriate, the measurement protocol is readily repeated. The population is many repeats of the study in the relatively well-off members of Galton's social circle. At the time of Galton's study the state of science knowledge was no evidence or theory, and so no relation. However, from everyday experience with fathers and sons we expect a positive relation. And from animal husbandry in Galton's time, we expect a positive relation of offspring to parents.

However, testing the null hypothesis in this case is not appropriate. The relation is obvious from the graph; the null hypothesis is of no interest. A more interesting (an plausible) model is a 1:1 relation of heights of sons to fathers. We will use confidence the 1:1 hypothesis. Confidence limits, like Neyman-Pearson hypothesis testing , start with fixed tolerances of Type I error. For a 95% confidence limit we are tolerating a 5% error rate in falsely rejecting hypotheses outside the confidence limits.

10. State evidential support. Report effects sizes. Interpret parameters of biological interest.

We conclude (step 4 above) that a positive relation of heights of sons to fathers is far more likely than no relation.

The effect size is the change in heights of sons relative to change in heights of fathers.

GLM routines reports the parameter with a standard error:

$\hat{\beta}_{Hf} = 0.52254 \pm 0.02425$

To evaluate multiple hypotheses we use the standard error 0.02425 to compute confidence limits for the slope parameter,

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

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

$\text{Lower} = 0.52254 - 2.1315 * 0.02425 = 0.471$

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

$\text{Upper} = 0.52254 + 2.1315 * 0.02425 = 0.574$

The confidence limits exclude several hypotheses about change in height of sons with change in height of fathers ($\Delta\text{Hson} / \Delta\text{Hfather} = \beta_{Hf}$)

They exclude the hypothesis of no relation: $\beta_{Hf} = 0$.

They exclude a 1:1 relation, which is what we might have expected.

The confidence limits are consistent with a simple rule of height inheritance: $\beta_{Hf} = 0.5$

Summary: Report evidence, effect size, confidence limits, and sample size.

Likelihood ratio. $LR = 6 \times 10^{12}$

Effect size. $Hson = 33.284 + 0.52254\ Hf$

The 95% confidence limits, which were very narrow, include a value of 0.5.

$0.471 \leq \beta_{Hf} \leq 0.574$

N = Two measurements from 1078 families, grouped into 17 height classes.

For each unit of increase in height of fathers there was not a equal increase in heights of sons Instead there was almost exactly a half unit increase in height of sons per unit increase in heights of fathers. Galton noticed that sons tend to be closer to the mean (shorter than father if father tall, taller than father if father short). He called this 'regression to the mean.' Galton's concept of regression to the mean became attached to Pearson's estimation method. Estimating the rate of change in one variable with change in another is now called regression.

Why does the relation of heights of sons to fathers follow 0.5:1 relation instead of a 1:1 relation? Hint: how many genes does a son inherit from his father?