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

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 ?

Data collected by Francis Galton at end of the 19th century.
Reported by K. Pearson and A. Lee in 1903, with analysis.

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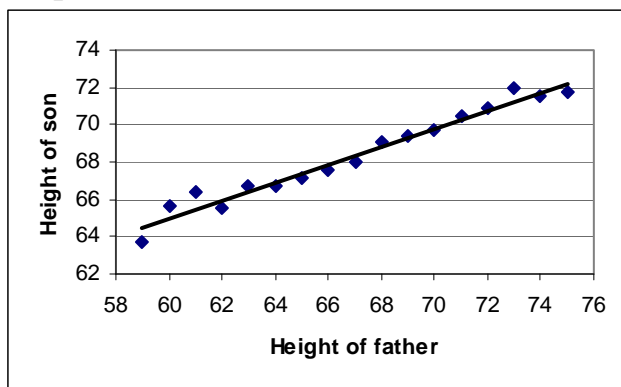I. Inheritance of physical characters. *Biometrika* 2: 357-462.

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. Hence the name regression for this statistical technique.

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
Begin with verbal model:
   There is a positive relation between heights of sons and fathers.
Graphical model.

Formal model.   Begin by distinguishing response from explanatory variables.

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

Data computed from Table 22 in Pearson and Lee (1903).

1. Construct 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. We take the explanatory variable (heights of fathers) as fixed in this observational study.

This example differs from the previous example in having an explanatory variable that is based on observations, rather than on 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 GLM)

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

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

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

2. Execute analysis. Place data in model format:
Column with response variable, $Hson$.
Column with explanatory variable Hf

Code model statement in statistical package according to the GLM

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

```
MTB > regress 'Hson' 1 'Hf';
SUBC> weights c3.
```
or
```
MTB > GLM 'Hson' = 'Hf' ;
SUBC> weights c3.
```

Many packages have a graphics interface that assists in constructing the model. If you are using the graphics interface, be sure 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 case 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 large number of families are given more weight than means based on less information. The data column (c3) has the number of families at each value of Hf.

2.   Execute 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
```

Residuals are calculated from fitted values, which were calculated from parameter
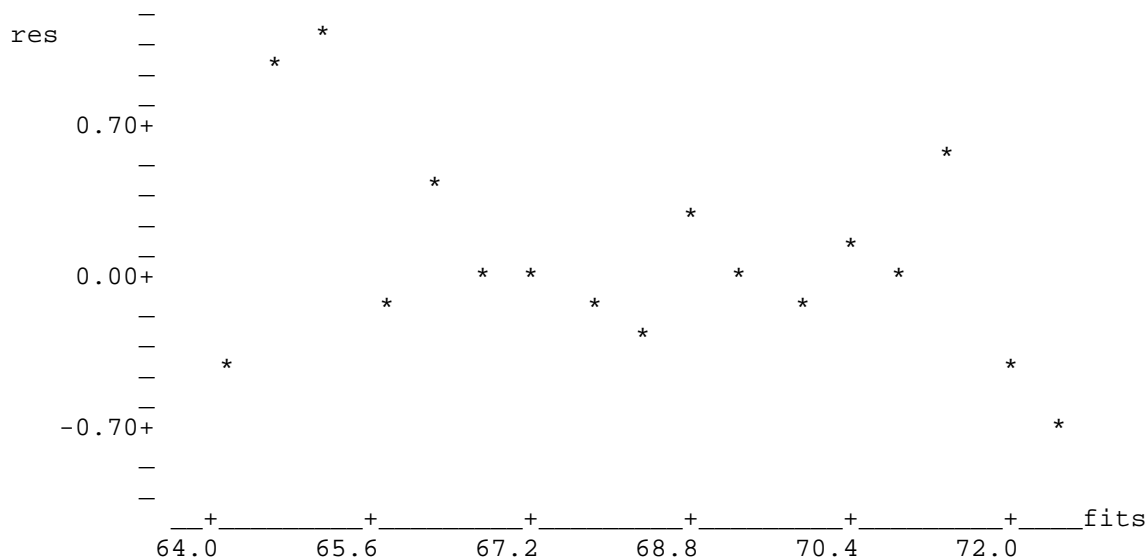estimates.

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

Residuals: Res = *Hson* − Fits

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

```
MTB > plot 'res' 'fits'
        —
 res    —            *
        —         *
        —
        —
   0.70+
        —                                              *
        —            *
        —                        *
        —                              *
        —
   0.00+          *    *         *        *
        —       *         *         *
        —                    *
        —
        —    *                            *
        —
  -0.70+                                      *
        —
        —
        __+_____+_____+_____+_____+_____+_____fits
          64.0      65.6      67.2      68.8      70.4      72.0
```

Residuals show acceptable band, no bowls or arches.
The straight line model is acceptable.
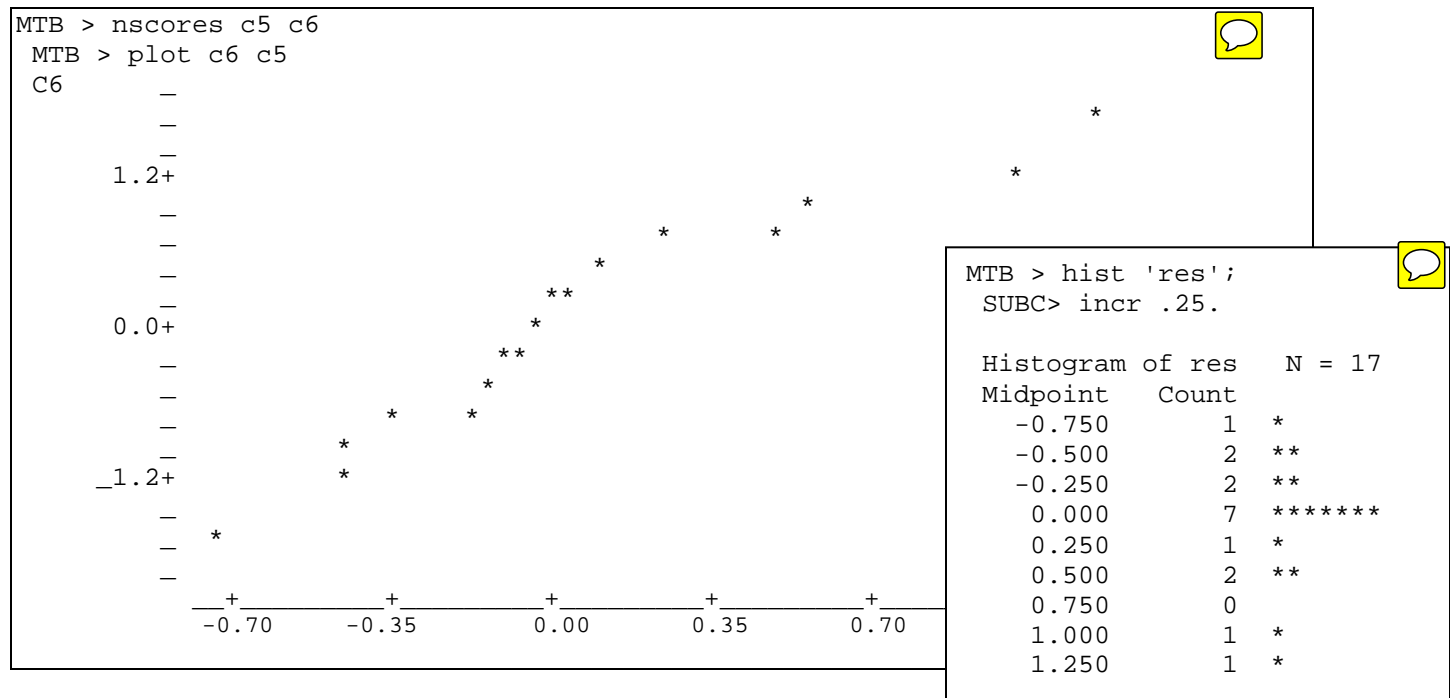
4

3. Evaluate model.

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

Errors homogeneous?

    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.

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.

```
MTB > nscores c5 c6
 MTB > plot c6 c5
  C6     _
         _
         _
     1.2+                                              *
         _                                          *
         _                               *
         _                         *        *
         _                   *
         _                 **
     0.0+               *
         _             **
         _           *
         _      *     *
         _   *
    _1.2+      *
         _
         _   *
         _
         __+_____+_____+_____+_____+____
          -0.70    -0.35     0.00      0.35      0.70
```

```
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    *
```

Errors independent?

    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.

We conclude that the straight line model and the error model are both acceptable.

4. State population and whether sample is representative.

    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 that the authors were prepared to infer to all people in the world.

5.   Decide on mode of inference. Is hypothesis testing appropriate?
We expect stature of sons to be positively related to stature of fathers. As a guess, we might a 1 to 1 relation.
Undertake hypothesis testing ?
No, use confidence limits to evaluate H_o and 1:1 hypothesis.

10.   Report and interpret parameters of biological interest.
Compute confidence limits from standard error of the slope parameter.

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

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

Lower	$=$	$\hat{\beta}_{Hf}$	$-$	$t_{0.025[df]}$ * st.err.	
Lower	$=$	0.52254	$-$	2.1315 * 0.02425	$= 0.471$
Upper	$=$	$\hat{\beta}_{Hf}$	$+$	$t_{0.025[df]}$ * st.err.	
Upper	$=$	0.52254	$+$	2.1315 * 0.02425	$= 0.574$

Which hypotheses are excluded by CI ?
These confidence limits exclude several hypotheses about change in height of sons with change in height of fathers (ΔHson / ΔHfather $= \beta_{Hf}$)

They exclude the hypothesis of no relation, that $\beta_{Hf} = 0$.
They exclude a 1:1 relation, which is what we might have expected
They are consistent with a simple rule of height inheritance, that $\beta_{Hf} = 0.5$
The interpretation of this regression by 0.5 toward the mean value is that fathers contribute half of the genetic component of height to their sons, the other half comes from mothers.

Report conclusion with statistical evidence: *Hson* $= 33.284 + 0.52254$ *Hf*
There was as a half unit increase in height of sons for each unit of increase in height of fathers.   That is, sons 'regress toward the mean.' They tend to be closer to the mean (shorter than father if father tall, taller than father if father short).   The confidence limits, which are narrow, include a value of 0.5, consistent with half of heritable component of height from fathers.

$0.471 \le \beta_{Hf} \le 0.574$ for $= 5\%$