Statistical Science.

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 is a special case of the GLM.
Example today was similar to previous, except that explanatory variable today was fixed into classes.
The number of families per class differed.
So means are weighted by number of families
This gives more weight to means based on a large number of families.
It gives less weight to means based on few families, where there is less information.
Means based on few families are poorly estimated.

GLM, applied to regression
Example: Galton's law relating heights of sons to heights of their fathers.
The explanatory variable, heights of fathers, is fixed into classes.
This is an observational study.

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

The data were collected by Francis Galton at the 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.

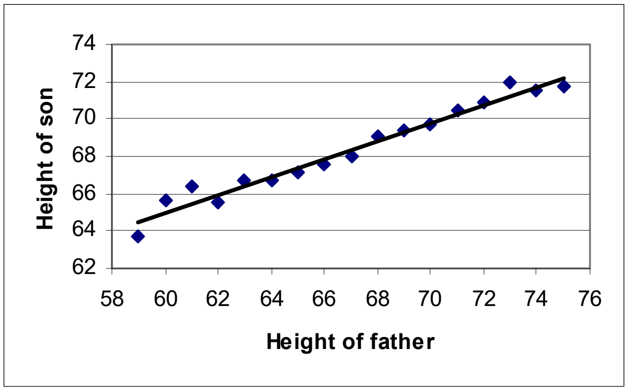
This was the first application of regression, a method that Pearson invented to analyze Galton's data. Galton found that the height of sons 'regressed' toward the mean value of fathers in a height class.

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.



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

Formal model. To construct this model, 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. We take the explanatory variable (heights of fathers) as the class mark (the midpoint of each category).

Taking the explanatory variable as fixed into classes differs from the previous example, where the explanatory variable was 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 (inch/inch)	none [L/L]	Greek: parameter

Write the formal model.

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 the model statement in statistical package to code the GLM model

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

```
HeightMod <- lm(Hson~Hf, weight= Nfam, data = GaltonDat)
```

If you are using a package with a graphics interface to construct the model, have a look at the code produced by the interface, so that you understand how the model you wrote in Step 1 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 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*. This weighted result will yield the same result as running each of the 1078 families separately.

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

Model based routines calculate residuals and fits as output.
Here is an example showing the regression equation..

```
MTB > regress Hson 1 Hf;
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, not 1:1
                              #stature regresses --> mean
```

The fitted values are calculated from the regression equation. The residuals are calculated from fitted values.

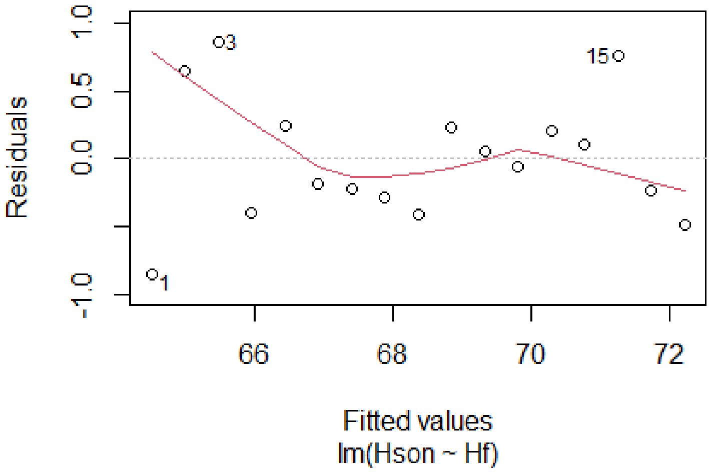
Fitted values: Fits = E[*Hson*] = $\hat{\alpha} + \hat{\beta}_{Hf} \cdot Hf$

Residuals: Res = *Hson* – Fits

3. Evaluate the structural model (the regression line). Plot the residuals against fitted values.

When fitting a line we evaluate whether there is some pattern of deviation from the line.

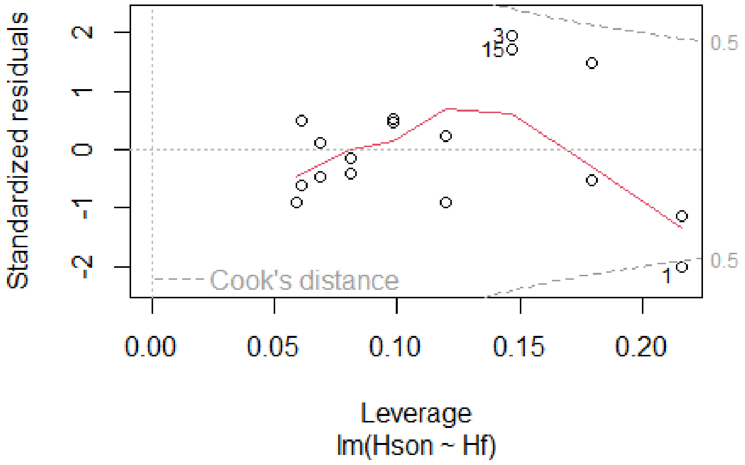
The residuals here show a slight downward tendency at small and large values. However, the plot does not show a simple pattern of a bowl or arch.



3. Evaluate the error model.

We use the residual vs fit plot to evaluate the homogeneity assumption. We find the vertical spread is greater at low fitted values than at high fitted values. The residuals are not homogeneous.

To evaluate the influence the larg residuals have on the estimate of the slope, we plot standardized residuals versus their leverage. Residuals at either end of the regression line (far from the mean) have greater leverage. Values of Cook's D greater than 0.5 have some influence on the slope estimate. Note the residual sitting below the 0,5 boundary. Values greater than 1 have substantial influence. We judge that the largest residual does not have substeantial influence on the slope estimate.

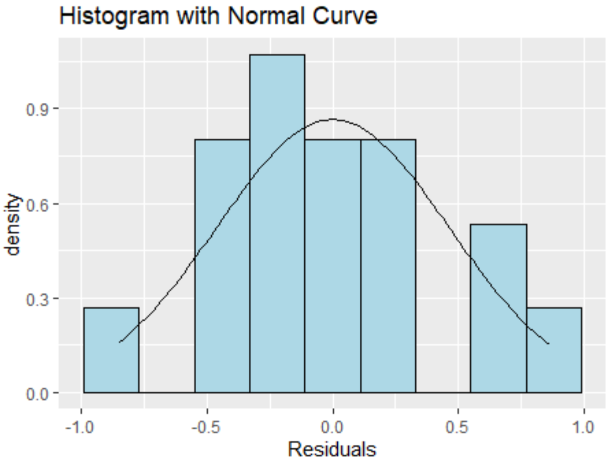


3. Evaluate the error model.

Next, we evaluate whether the errors are normally distributed.

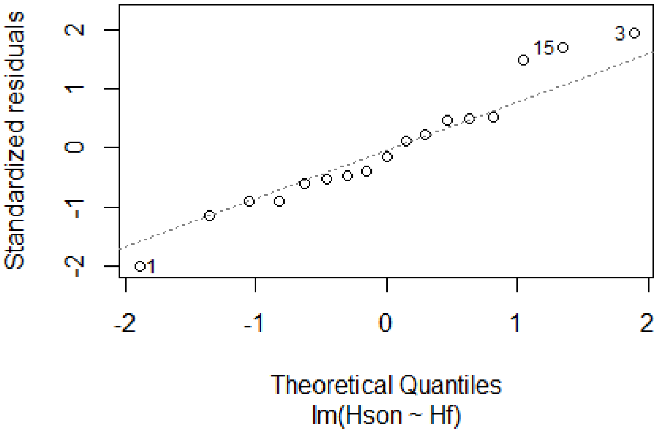
It is commonly believed that we check this assumption before doing an analysis. This sounds like good practice but it turns out to be wrong. The assumptions apply to the errors. We cannot evaluate assumptions until we have the errors calculated.

The response variable consists of means, so we expect the residuals to be normally distributed. As expected, the residuals are close to normal.



When we compare the histogram to the QQ plot we see some indication of slight deviation from normality in the upper (positive) tail of the distribution.

There is a slight tendency toward large residuals for very tall and very short sons. The deviations are more evident in the QQ plot than in the histogram.

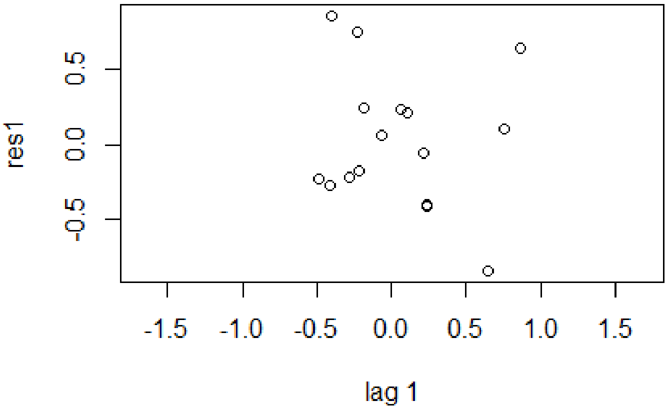


3. Evaluate the error model.

Are the residuals independent?

We lack information on temporal order or spatial layout of samples, to evaluate this assumption. As a check we can plot the residuals ordered from short fathers to tall fathers. To do the check we offset (lag) the ordered residuals so that each residual is plotted against its adjacent residual in the ordered list. Here are the lagged residuals.

Hfather	Res	Lag 1
59	-0.447	
60	1.006	0.447
61	1.185	1.006
62	-0.123	1.185
63	0.475	0.123



When the residuals are plotted against the adjacent residual (lag1) we see no pattern of dependence. The residuals ae independent when ordered by the explanatory variable, height of the fathers.

4. Partition df and SS according to model.

$$\begin{array}{rcl} Hson & = & \alpha + \beta_{Hf} \cdot Hf + \varepsilon \\ 17\text{-}1 & = & 1 \quad + 15 \\ 2248 & = & 2177.9 \quad + 70.34 \end{array}$$

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 the null model, 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. Th protocol is applicable to any group of people.

6. State test statistic, and tolerance for Type I error, α

From everyday experience with fathers and sons we know there is a positive relation. And from animal husbandry in Galton's time, we know in general there is a positive relation of offspring to parents. So testing the null hypothesis (no relation) is superfluous. A plausible model is a 1:1 relation of heights of sons to fathers. Instead of using a test statistic to test a null hypothesis, we will use confidence interval to evaluate 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.

Skip ANOVA table (step 7) and randomized p-value (step 8).

9. Report measure of evidence and measure of certainty

The likelihood ratio (step 4 above) in favor of a positive relation of heights of sons to fathers is enormous for a positive relation of heights of sons to heights of fathers, as measured by the slope parameter. Our measure of uncertainty will be the standard error of the slope parameter. We will use this to calculate a confidence limits, which allow comparison to a 1:1 relation.

9. Report measure of evidence and measure of certainty

GLM routines reports the parameter with a standard error:

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

The measure of certainty for this estimate is the standard error 0.02425 inches.

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

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

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

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

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

The confidence limits exclude several hypotheses about change in height of sons with change in height of fathers (ΔHson / ΔHfather $= \beta_{Hf}$)
They exclude the hypothesis of no relation: $\beta_{Hf} = 0$.
They exclude a 1:1 relation, which is what we might have expected.
They do not exclude a simple rule of height inheritance: $\beta_{Hf} = 0.5$

Report 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 are narrow, include a theoretical value of 0.5.

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

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

10. Report science conclusion. Interpret parameters of biological interest

For each unit of increase in height of fathers there was not an 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: Why might we expect a value of 0.5 ?