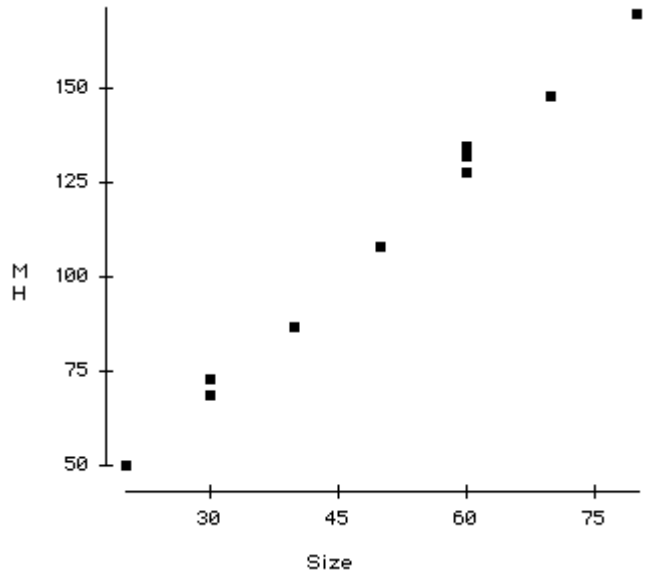
Lecture 13

Linear relationships. Regression. Correlation

The final topic we consider in this course is study of (linear) relationships between continuous variables.
We begin with the simple idea that X carries some information about Y.
There is a functional relationship but there also is error.

Convergence of the Gibbs' sampler. (numerical integration by simulation)

problem no	No of Parameters X	CPU min Y
1	30	73
2	20	50
3	60	128
4	80	170
5	40	87
6	50	108
7	60	135
8	30	69
9	70	148
10	60	132



Plagiarised from Neter and Wasserman.

Clearly there is a (strong) relationship between the size of problem (number of parameters) and the time to achieve the convergence criteria.

The relationship is stochastic – for example for 60 parameters we have we observe values 128, 135 and 132. i.e. not the same value every time.

Regression is a technique for extracting the systematic component of the relationship.

Although theoretically we would probably not believe the relationship to be linear.
As a first order approximation, within the range of values, we may propose a linear function. See graph.

We thus propose:

$$Y = a + b\,X + \text{error}$$

The statistical task is to estimate *a* and *b*.

The sensible thing to do is to minimise the errors in some way. Although other methods are possible and sometimes used. The nicest properties are got from a *least squares fit.*

$$Y_i = a + bX_i + R_i \quad \text{for i=1,10}$$

Choose *a* and *b* so that $\sum R_i^2$ is minimised.

The maths:
Solution to Least Squares $SSE = \sum R_i^2$
SSE - Sum of Squares Error (also called Residual sum of squares)
Task : choose a and b to minimise SSE

For reference
Algebraic theory tells us that this is achieved by solving:

$$\frac{\partial(SSE)}{\partial a} = 0$$

$$\frac{\partial(SSE)}{\partial b} = 0$$

$$SSE = \sum R_i^2 = \sum (y_i - a - bx_i)^2$$

$$\frac{\partial(SSE)}{\partial a} = \sum \frac{\partial}{\partial a}(y_i - a - bx_i)^2$$

(derivative of sum = sum of derivatives)

$$\frac{\partial}{\partial a}(y_i - a - bx_i)^2 = 2(y_i - a - bx_i)\frac{\partial}{\partial a}(y_i - a - bx_i)$$ (chain rule)

$$\frac{\partial}{\partial a}(y_i - a - bx_i) = -1$$

So

$$\frac{\partial(SSE)}{\partial a} = -2\sum (y_i - a - bx_i)$$

$$= -2(\sum y_i - na - b\sum x_i)$$

Equating this to 0 gives:

$$na = \sum y - b\sum x$$

or

$$a = \bar{y} - b\bar{x}$$ Equation 1

$\frac{\partial(SSE)}{\partial b} = -2\sum x_i(y_i - a - bx_i)$ and setting this to 0 gives:

$$\sum xy - a\sum x - b\sum x^2 = 0 \qquad \text{Equation 2}$$

Eqs 1 and 2 are called the normal equations - we shall use them again later:

Using 1 to get rid of a in 2 gives

$$\sum xy - (\bar{y} - b\bar{x})n\bar{x} - b\sum x^2 = 0$$

$$\sum xy - n\bar{x}\bar{y} = b(\sum x^2 - n\bar{x}^2)$$

or

$$b = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2}$$

Again:

$$b = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2}$$

$$a = \bar{y} - b\bar{x}$$

So to compute a and b we need:

$\sum xy,\ \sum x, \sum y, \sum x^2$ and n

In the example above:

$$\sum xy = 61800, \sum x = 500, \sum y = 1100, \sum x^2 = 28400, n = 10$$

$$b = \frac{61800 - 10 * 50 * 110}{28400 - 10 * 2500} = 2$$

$$a = 110 - 2 * 50 = 10$$

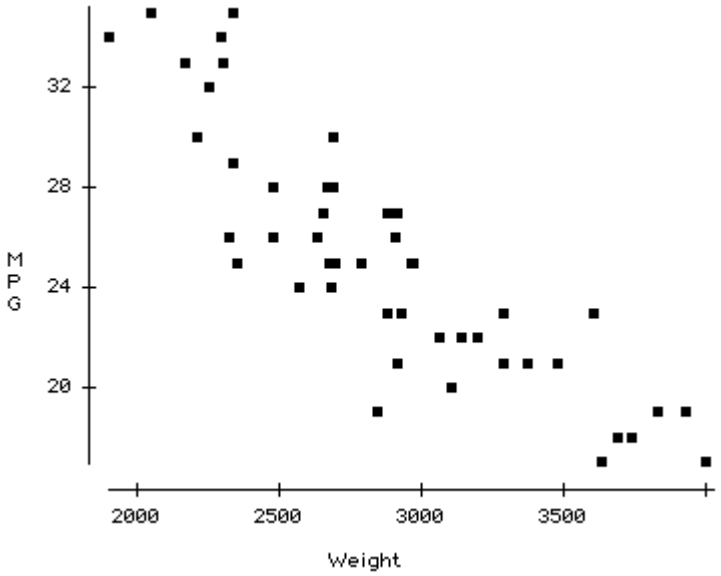
$$Y = 10 + 2 * X$$

10 – set up time.
2 CPU mins for every extra parameter.

It should be set at the outset that we are <u>not</u> proposing that this equation would hold for all X. But given the graph it is a good approximation within the X range 20 to 80.

In practice regression calculations are done in a package. You might get the following output:

Data Desk – MPG as a function of Car weight (kg)
MPG and weight. A sample of car models.



MPG on Weight

Variable	Coefficient	
Constant	48.7393	(a)
Weight	-0.00821362	(b)

(output edited)

Clearly this is just an "average" equation – but it can be used to identify "good and poor" performers within car classes (as defined by weight).

Equations such as these can be used to

1. Determine is there any (linear) relationship at all? (if not b=0)

2. Estimate the change in Y per unit change in X (b)

3. predict Y from a given X:

My problem has 53 parameters how long will it take?
My car weighs 1200Kg what would I expect the MPG to be?

These turn out to be quite simple problems – because of the very nice properties of the estimates a, b.

Rewriting the formula for b is useful

$$b = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2}$$

$$= \frac{1}{\sum x^2 - n\bar{x}^2}\left(\sum xy - \frac{n\bar{x}}{n}\sum y\right)$$

$$= \frac{1}{\sum x^2 - n\bar{x}^2}\sum (x_i - \bar{x})y_i$$

$$= \sum \frac{(x_i - \bar{x})}{\sum x^2 - n\bar{x}^2} y_i = \sum k_i y_i$$

$b = k_1 y_1 + k_2 y_2 \ldots$

b is a linear combination of y's (doesn't depend on y^2) this makes the properties of b simple (for testing, confidence intervals etc) (b is a weighted mean of Ys)

$a = \bar{y} - b\,\bar{x}$ so is also a lc. of Ys

$\hat{Y}_h = a + b\,x_h$ is also a lc of Ys

Statistical model.
In the algorithm above nothing depended on statistics. Its just a minimisation. To test, construct CI etc. we need distributional assumptions.

Model $E(Y) = a + bX$ or $E(error) = 0$.
That makes sense the error should not have a systematic component.

Error suggests Normal distribution .

$$R_i \sim N(0,\sigma^2) \quad \text{or } Y_i \sim N(\alpha + \beta X_i, \sigma^2)$$

where α and β are the true values which *a* and *b* estimate.

The Normal distribution likes lc. – they are also Normal so a, b and predictions will be Normally distributed.

All we have to do is compute means and variances. They are easy enough too.

What about the distribution for X? – NONE
The X are assumed to be known constants – not subject to error.

The number of parameters (X) is not a random quantity measured exactly. Y is a response to X.
It would be silly to estimate number of parameters from CPU time.

The MPG (Y) , Weight (X) is not so clear. It is reasonable to think that MPG is a function of Weight rather than the other way round. The errors in the measurement of weight will be small. We pretend they are not there.

If the data a paired random variables eg Height and Weight of people (H,W) Should regress H (Y) on W (X) or W(Y) on H(X) ? In fact (technically) you should do neither!

Regression is asymmetric:

$W = a + b\,H$ - minimise $(W - a - bH)^2$
and $H = c + d\,W$ - minimise $(H - c - dW)^2$
Will give inconsistent results.

A symmetric form would require:

minimise $(W-a-bH)^2 + (H+\frac{a}{b}-\frac{a}{b}W)^2$ (orthogonal distance of point from line).
This will have none of the nice properties of regression.

For reference

The key to computing means and variances lies in the ks defined earlier.

$$k_i = \frac{x_i - \bar{x}}{SSX} \text{ where } \mathrm{SSX} = \sum (\mathrm{x_i} - \bar{x})^2$$

$$\sum k_i = \frac{1}{SSX} \sum (x_i - \bar{x}) = 0$$

Sum of the k's is 0.

as $\sum (x_i - \bar{x}) = 0$

$$\sum k_i^2 = \sum \frac{(x_i - \bar{x})^2}{SSX^2} = \frac{SSX}{SSX^2} = \frac{1}{SSX}$$

$$\sum k_i x_i = \sum k_i (x_i - \bar{x}) \text{ adding } \bar{x} \sum k_i = 0$$

$$= \sum \frac{(x_i - \bar{x})(x_i - \bar{x})}{SSX} = \frac{1}{SSX} \sum (x_i - \bar{x})^2 = \frac{SSX}{SSX} = 1$$