Finding a Fit
Regression and Data Analysis

Background

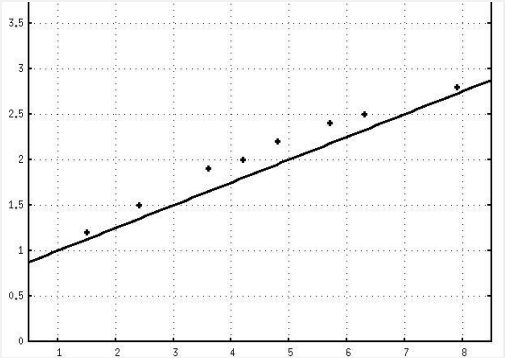
- Suppose we have carried out a *set of observations*, eg the average time taken by a program on a range of randomly chosen data sets of various sizes.
- How do we *present* these data *effectively* and *clearly*?
- By a *table*?
- Might be very *lengthy* and *unclear*.
- By a *plot/curve* on a *coordinate axis*?
- How do we find the "*best-fitting*" line and/or *function*?
- This is the topic we now consider.

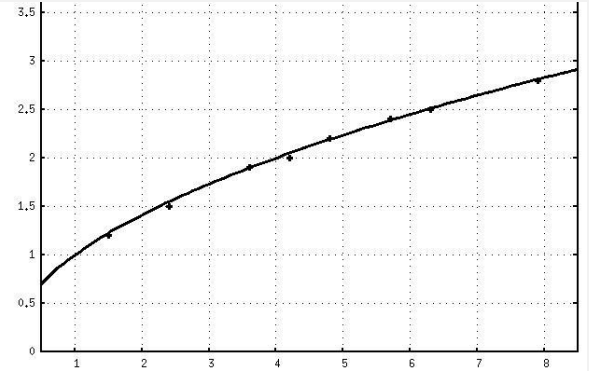
An Example

- A table of observations (Y) seen from a set of values (X)

X	1.5	2.4	3.6	4.2	4.8	5.7	6.3	7.9
Y	1.2	1.5	1.9	2.0	2.2	2.4	2.5	2.8

- Which best describes this data set?





A sense of “best”

- The standard interpretation of “*best fit*” is “*least squares*”.
- Meaning?
- We have *data-observation* pairs: $\{\langle x_k, y_k \rangle : 1 \leq k \leq n\}$.
- We want a function $F : R \to R$ that “*best*” describes these, ie that *minimizes*

$$\sum_{k=1}^{n} \left(y_k - F(x_k)\right)^2$$

Linear Functions & Regression

- While not perfect, line functions often provide useful support.
- The "*best-fit line*" function can be found "*easily*".
- Other classes of function, after some *data transformations* may be captured in terms of "*best-fit line function*".
- What is a line?
- A *function of its input* characterized by *two parameters*:

$$F(x) = ax + b$$

- Finding the "*best-line*" is finding the best pair (a, b).

Combining everything

- We have *data-observation* pairs: $\{\langle x_k, y_k \rangle : 1 \leq k \leq n\}$.
- We want to find (a, b) to *minimize*

$$F(a,b) = \sum_{k=1}^{n} (y_k - ax_k - b)^2$$

- Notice that the values $\{\langle x_k, y_k \rangle : 1 \leq k \leq n\}$ are constants.
- We *know* how to find the optimising (a, b).
- Use *partial differentiation*: find $\left(\frac{\partial F}{\partial a}, \frac{\partial F}{\partial b}\right)$; and find the *critical points* for these.

What happens?

- We find that the best fit line function is

$$\left(\frac{nW_{xy} - W_xW_y}{nW_{xx} - (W_x)^2}\right)x + \left(\frac{W_{xx}W_y - W_{xy}W_x}{nW_{xx} - (W_x)^2}\right)$$

- Where

$$W_x = \sum_{k=1}^{n} x_k \; ; \; W_y = \sum_{k=1}^{n} y_k$$

$$W_{xx} = \sum_{k=1}^{n} x_k^2 \; ; \; W_{xy} = \sum_{k=1}^{n} x_k y_k$$

- Test your understanding of partial derivatives: *do this*.

Other Functions?

- On occasion a line will not be the best fitting curve.
- We may need to look at functions such:
- Powers: ax^b
- Exponentials: $a \exp bx$
- Logarithms: $a \ln x + b$
- Linear Rational functions: $\frac{1}{ax+b}$
- We may have *more than one* parameter to optimize:
- Quadratic functions: $ax^2 + bx + c$.

Dealing with these cases

- The two variable cases, although in principle amenable to exactly the same approach are most easily dealt with by some rewriting. These are discussed on pages 344–350 of the text.
- Cases involving 3 or more variable become increasingly onerous. Quadratic regression is feasible (pp. 353–4). Larger numbers of parameters must deal with solving systems of equations as arise in multivariable partial derivatives, page 338 gives an overview.

Other Data Analysis Issues

- The methods discussed do not require the function used *exactly* to match pairs of data and observation.
- When this is of interest techniques such as *Lagrange Interpolation* (pages 350–353) deal with *polynomials*.
- The related topic of *trigonometric interpolation* is of interest in *electronics* and *signal processing*.
- Many of these ideas will be found in later specialist CS modules such as COMP229.