Lecture Notes in Quantitative Biology

Exploratory Data Analysis -- Model revision

Chapter 19.2 Revised 25 November 1997

ReCap:Exploratory Data Analysis
What is it ?

on chalk board

```
Today:  Exploratory data analysis.  Introduction.
```

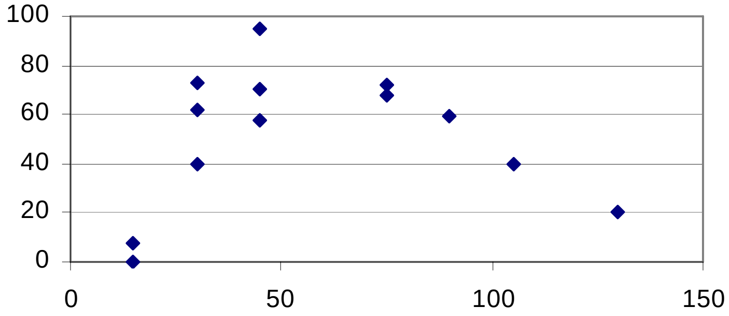
Wrap-up Purpose of exploratory analysis is to discover pattern.
4 tactics: drop, add, combine variables, discover variables via residuals
Today, we looked at distinction between explatory and confirmatory.
We also learned about box and arrow diagrams,
to simplify and bring out the relation of variables.

EDA. Model revision

Pattern can be graphical, as in the search for better model of relation of aphid stem mother length to thorax width. A series of rescalings of ltot and lthor were plotted. See handout SRBX15_7.out.

Here is another example of one form of exploratory analysis, successive approximations.
Downing, D.L. and S.K. Allen. 1987. Aquaculture 63: 1-21.
percent triploidy as function of time after fertilization in Pacific Oysters treated with Cytocholasin B at 20°C.

$$P = 0.02T^{2.90}e^{-0.62T}$$



Draw axes: vertical = % triploidy L26b
horizontal = time
Draw 2 straight lines thru points.

The initial description of pattern here is 2 straight lines. This is only a caricature. Very few points fall on these lines.
But they do capture one of the main features of this batch of data, which is that there is increase, then decrease in triploidy with time.

Erase, draw 3 lines. L26c
one rising, one across, one dropping

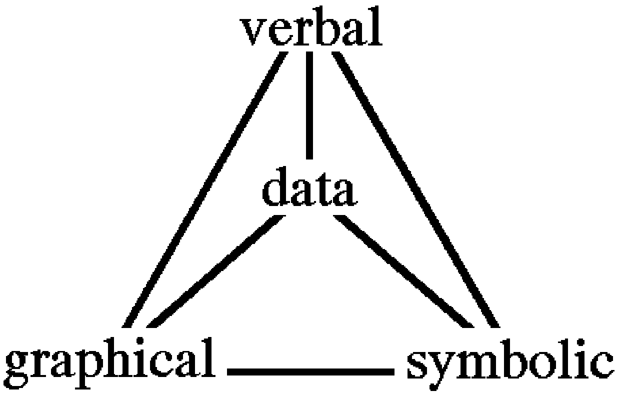
A somewhat more accurate graphical model of pattern: 3 lines.

Erase the three lines, L26d
draw in single sigmoid curve.

An equally accurate and possibly more biologically realistic pattern is a curve, representing the idea that with time, triploidy increases rapidly, then tapers off.

This illustrates the idea of exploratory analysis as successive approximations.

EDA--Another example of model revision



In the above example (oyster triploidy) we used both verbal and graphical models to describe the pattern in the data.

A formal model can also be used. Here are three successive models (caricatures written in the form of equations).

1. Seeds/tree = k * Altitude (straight line)

2. Seeds/tree = k* Alt^2 if Alt < 1000 m

   Seeds/tree = k* $\text{Alt}^{1/2}$ if Alt > 1000 m

3. Seeds/tree = $k\,(1 + e^{-r(\text{Alt}-1000)})^{-1}$ (logistic equation)

This last equation looks formidable. But it is not that hard to apply the model with a calculator that takes exponents. It is the same equation you might have met for exponential growth in population biology or ecology. It is the same equation used in dosage-response curves in medicine.

The three equations (line, pair of curves, logistic) are succesive descriptions of pattern in the seed production data. They are formal, rather than graphical representations of the pattern in the data.