Data analysis methods in weather and climate research

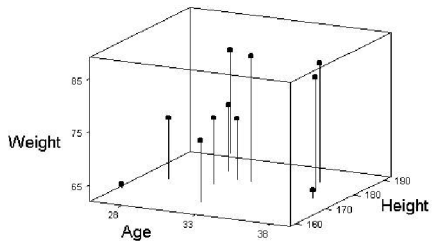
Dr. David B. Stephenson
Department of Meteorology
University of Reading
www.met.rdg.ac.uk/cag

8. Multiple linear regression and non-linear regression

- Multiple regression
- Multivariate regression – the General Linear Model
- Non-linear responses – Generalized Linear Models
- Non-parametric regression

8. Multiple regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

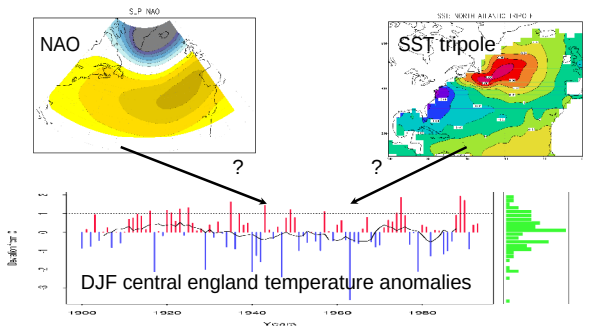



```
> fit<-lsfit(cbind(x1,x2),y)
> ls.print(fit)
```

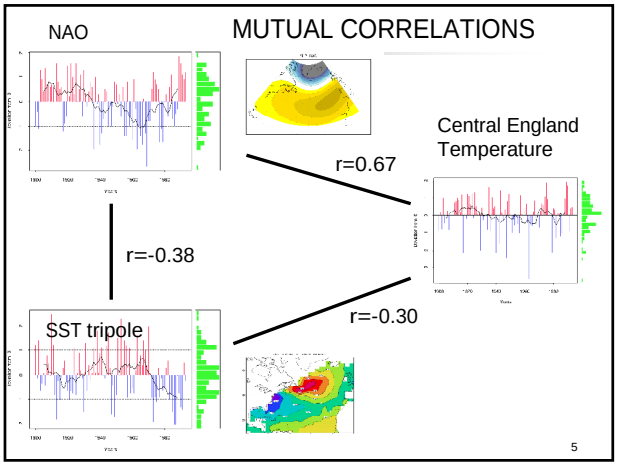
8. Regression summary

```
The regression equation is
Weight = -40.4 + 0.517 Age + 0.577 Height

Predictor   Coef        St.Dev      t           p-value
Constant    -40.36      49.20       -0.82       0.436
Age         0.5167      0.5552      0.93        0.379
Height      0.5769      0.2671      2.16        0.063

S = 6.655   R-sq = 41.7%      R-sq(adj) = 27.1%

Analysis of variance
Source           DF    SS          MS    F     p-value
Regression  2    253.66           126.8 2.86  0.115
Residual    8    354.34           44.29
Total       10   608.00
```

8. Climate example: role of the N. Atl ocean



Stephenson, D.B. and M. Junge, 2003: Int. J. Climatology, 23, 245-261
www.met.rdg.ac.uk.cag/publications



8. The linear modelling approach

To unravel the indirect from the direct effects we need to go beyond decriptive methods (correlation analysis) and introduce a model:

$$CET = \beta_1 NAO + \beta_2 SST + \varepsilon$$

Using data from 1900-1994, we obtain estimates of:

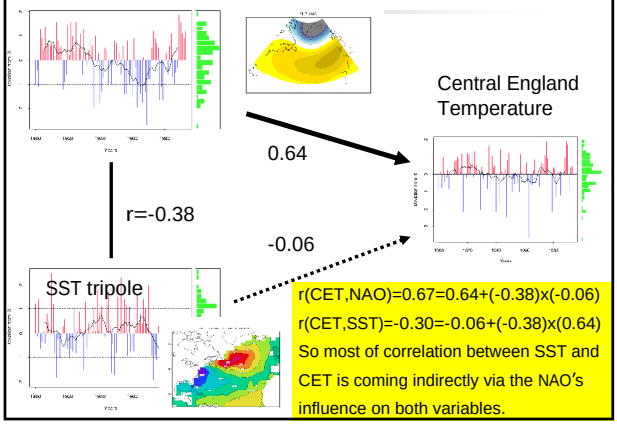
$$\hat{\beta}_1 = +0.64 \pm 0.08$$

$$\hat{\beta}_2 = -0.06 \pm 0.08$$

The fit explains 45% of the total CET variance and is statistically significant at p<0.001

6

8. Direct and indirect effects



r(CET,NAO)=0.67=0.64+(-0.38)x(-0.06)
r(CET,SST)=-0.30=-0.06+(-0.38)x(0.64)
So most of correlation between SST and CET is coming indirectly via the NAO's influence on both variables.

8. Multivariate regression

The General Linear Model:

$$Y = X\beta + \varepsilon$$

$Y = (n \times p)$ matrix of p response variables

$X = (n \times q)$ matrix of q explanatory variables

$\beta = (q \times p)$ matrix of model parameters

$\varepsilon = (n \times p)$ matrix of normally distributed errors

Multiple regression extended to more than 1 response

```
> # R command
> lm(y~x)
```


8. Regression as a probability model

- Purely descriptive correlation r=0.67
- Least-squares regression

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

$$\text{Minimise} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

- Probability model

$$Y \mid X \sim N(\alpha + \beta X, \sigma_\varepsilon^2)$$

8. A deeper probabilistic view …

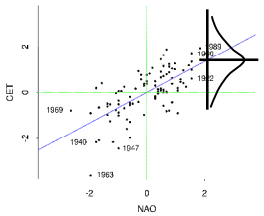
Instead of just thinking of an additive signal plus noise model, a deeper insight can be obtained by thinking of the regression as a model for the conditional probability of Y|X:

$$Y \mid X \sim N(\beta_0 + \beta_1 X, \sigma_\varepsilon^2)$$

or

$$E(Y \mid X) = \beta_0 + \beta_1 X$$

$$Var(Y \mid X) = \sigma_\varepsilon^2$$



8. Non-linear/non-normal responses

The Generalised Linear Model (GLM):

$$Y \mid X \sim F(\theta)$$

$$g(\theta) = X\beta$$

```
> # R command
> glm(y~x)
```

Some examples:

Linear regression: $F = N(\mu, \sigma^2)$ $g(\mu) = \mu$

Gamma regression: $F = G(\alpha, \beta)$ $g(\mu = \alpha / \beta) = 1/\mu$

Logistic regression: $F = Be(\pi)$ $g(\pi) = \log(\pi / 1 - \pi)$

Poisson regression: $F = Po(\mu)$ $g(\mu) = \log \mu$

8. Non-parametric regression

Know that Y depends on X but have no idea what the functional relationship is except that it is smooth.

$$Y = f(X) + \varepsilon$$

Need to use *smoothing methods* to estimate the unknown function f(X). Main approaches are:

- local robust polynomial fits `>lowess(x,y)`
- smoothing splines `>smooth.spline(x,y)`
- kernel smoothing `>kernel(.)`

8. Kernel estimation

Smooth local composites:

$$\hat{f}(x) = \frac{\sum_{i=1}^{n} w(x - x_i) y_i}{\sum_{i=1}^{n} w(x - x_i)}$$

$$w(.) = \text{kernel weights}$$

8. Summary

- Regression models the conditional probability p(Y|X)
- Many such models can be used depending upon the type of response(s), type of explanatory variable(s), and knowledge of distributional form.
- Linear regression can be extended to include multiple explanatory variables (multiple regression), multiple responses (multivariate regression), and non-linear/normal responses (Generalised Linear Models).
- Non-linear unknown relationships can be estimated using non-parametric regression approaches.