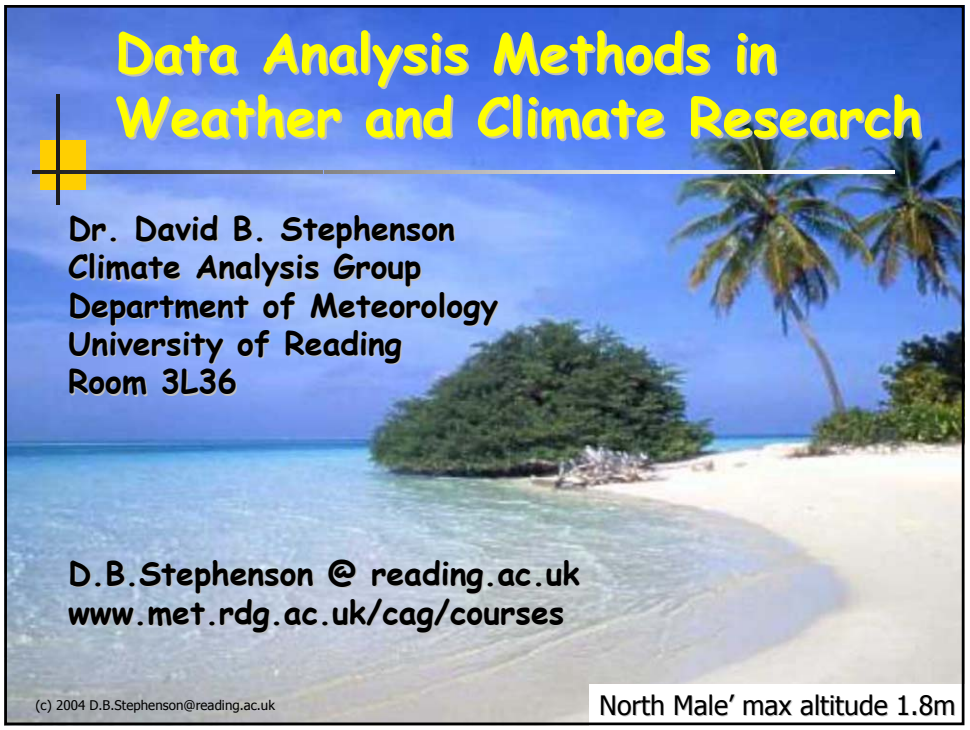
Data Analysis Methods in Weather and Climate Research

Dr. David B. Stephenson
Climate Analysis Group
Department of Meteorology
University of Reading
Room 3L36

D.B.Stephenson @ reading.ac.uk
www.met.rdg.ac.uk/cag/courses



North Male' max altitude 1.8m

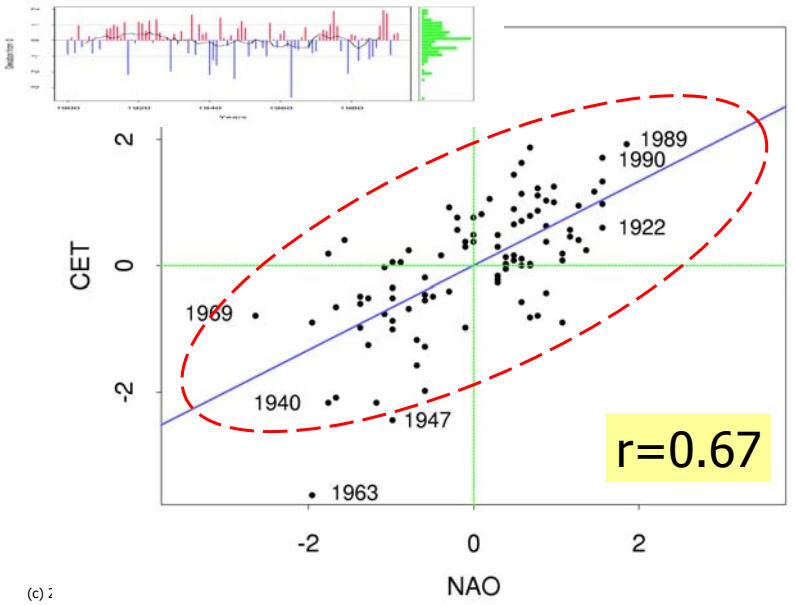
Course outline

1. Introduction
2. Descriptive sample statistics
3. Basic probability concepts
4. Probability distributions
5. Parameter estimation
6. Statistical hypothesis testing
7. Basic linear regression
8. More advanced regression
9. Introduction to time series

7. Basic linear regression

- Modelling strategy
- Linear regression
- How to present the results
- Residual diagnostics
- Variations: weighted and robust regression

Central England Temperature versus NAO (winters 1900-1994)



Levels of explanation

- Purely descriptive: correlation r=0.67
- Data-analytic – least-squares minimisation

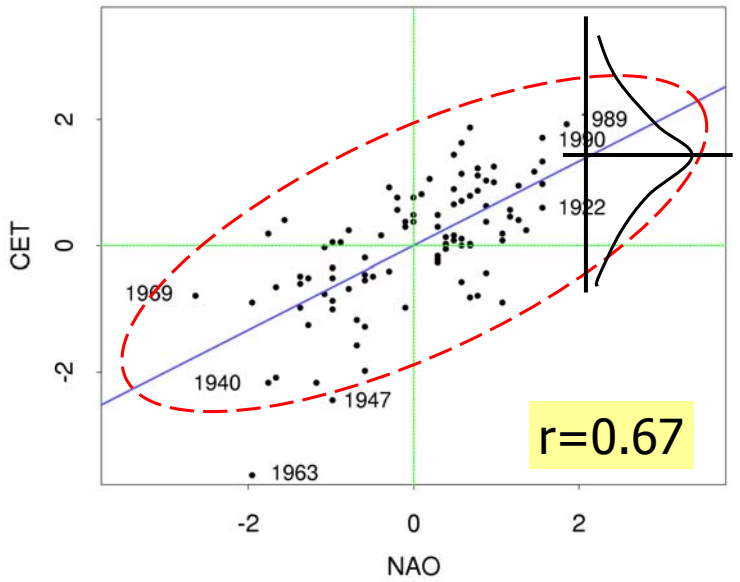
$$y_i = \alpha + \beta x_i + \varepsilon_i$$

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

- Probability model

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

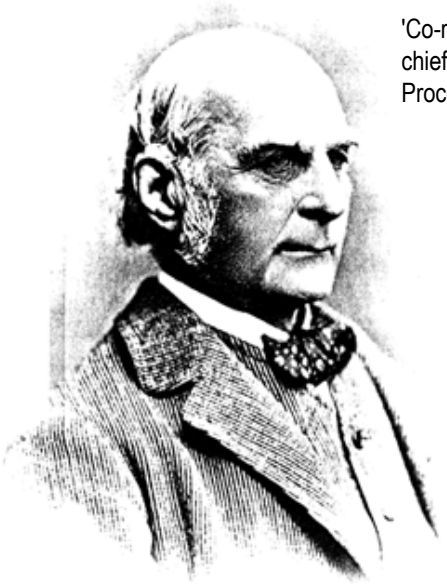
Central England Temperature versus NAO (winters 1900-1994)



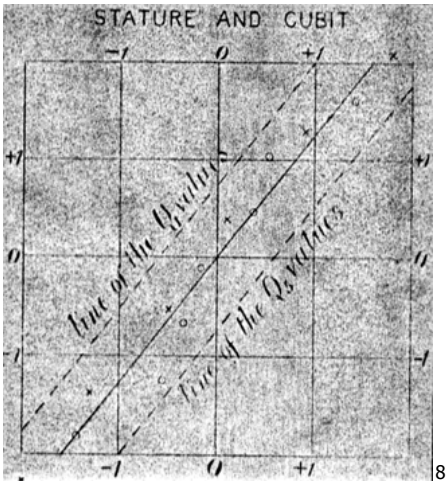
Correlation

- Original mathematical invention
- Introduction into climate studies
- Issues:
  - Sampling uncertainty
  - Multiple testing
  - Causality
- More model-based approaches

Sir Francis Galton FRS 1822-1911



'Co-relations and their measurement, chiefly from anthropometric data.'
Proceedings of the Royal Society 45, pp. 135-45. 1888



Product moment definition

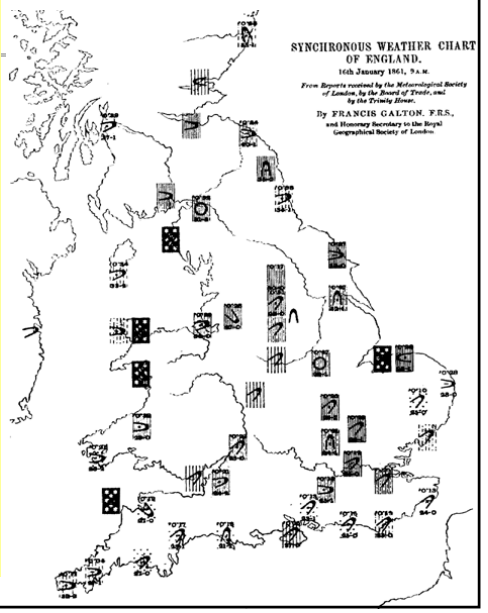
$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

where

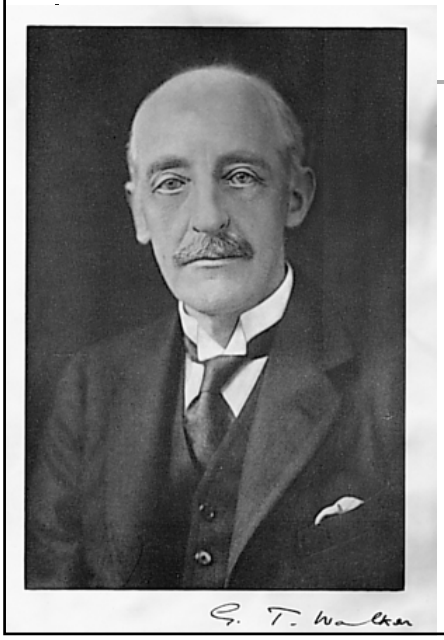
$$S_{xy} = \frac{1}{n} \sum_{i=1}^{n} (x - \bar{x})(y - \bar{y})$$

Galton's contributions to meteorology:

- "Meteorological charts" (1861) Philosophical Magazine, 22, pp.34-5 1861
- Anti-cyclone, Royal Soc. (1862)
- Meteorographica (1863)
- Meteorological Committee 1868-1904
- plus many others



Sir Gilbert Walker 1868-1958



- 1868 Born in Lancashire
- 1886-1890 Maths, Trinity College
- 1895-1903 Lecturer, Trinity College
- 1903-1924 India Met Department
- 1924-1934 Imperial College
- 1950 Last paper (Biometrika)
- 1958 Died in Surrey

"The relationships between weather over the Earth are so complex that it seems useless to try to derive them from theoretical considerations; and the only hope at present is that of ascertaining the facts and of arranging them in such a way that interpretation shall be possible."

- **Walker, G.T. (1910) Correlation in seasonal variation of climate, Mem. Ind. Met. Dept., 21 (2), 117-124.**
- **Walker, G.T. (1914) Correlation in seasonal variations of weather III. On the criterion for the reality of relationships or periodicities, 21 (Part 9) 13-15.**
- **Walker, G.T. (1923) Correlation in seasonal variation of weather VIII. A preliminary study of world weather, Mem. Ind. Met. Dept., 24, 75-131.**
- **Walker, G.T. (1924) Correlation in seasonal variation of weather IX, Mem. Ind. Met. Dept., 25, 275-332.**
- **Walker, G.T. and Bliss, E.W. (1932) World Weather V, Mem. Roy. Met. Soc., 4, 53-84.**

Katz, R.W., 2002: "Sir Gilbert Walker and a connection between El Nino and statistics." Statistical Science, 17, 97-112.

Stephenson et al. (2003)
The History of Scientific Research on the NAO, Chapter in AGU monograph on the North Atlantic Oscillation (Eds. Hurrell et al.)

Probable errors in correlations

Probable error in r :

$$e = 0.67449 \frac{1-r^2}{\sqrt{n}}$$

Pearson (1898) Phil. Trans., 191, p. 242

Probable error in $\max(r_1, r_2, \ldots, r_m)$:

$$e_m = k_m 0.67449 \frac{1-r^2}{\sqrt{n}}$$

m	k_m
1	1
2	1.56
10	2.72
100	4.01
1000	5.03

multiple testing $\Rightarrow k_m = \Phi^{-1}(0.5 + 0.5 \times 0.5^{\frac{1}{m}}) / \Phi^{-1}(0.75)$

Modelling strategy

- Exploratory Data Analysis (EDA)
- Model identification
- Parameter estimation
- Validation of model fit
- Out-of-sample prediction
- … and then iterate if necessary

The basic idea

Model the relationship between two random variables as:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

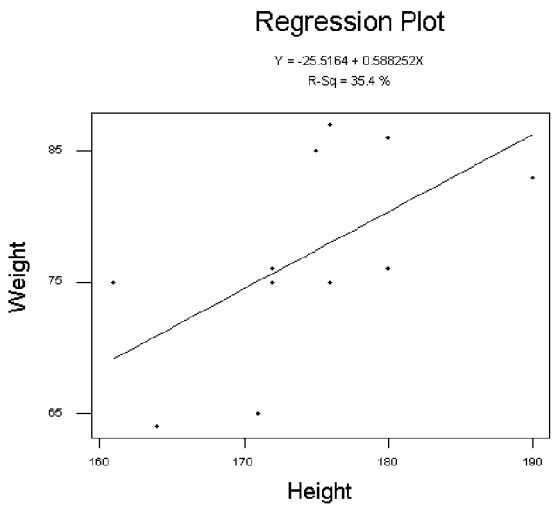
where

y_i = measured value of "response variable"

x_i = measured value of "explanatory variable"

ε_i = random normally distributed noise

Regression of weight on height



A deeper view …

Instead of just thinking of an additive signal+noise model, a deeper insight can be obtained by thinking of the regression like this:

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

or

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

Ordinary Least Squares Estimation

Find best parameters by minimising the sum of the squared residuals:

$$SS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

where

$$\hat{y}_i = \beta_0 + \beta_1 x_i = \text{predicted value}$$

OLS best estimates

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = r \frac{s_y}{s_x}$$

$$r = \frac{s_{xy}}{s_x s_y} = \text{sample correlation of x and y}$$

$$s_{xy} = \frac{1}{n} \sum_{i=1}^{n} (x - \bar{x})(y - \bar{y}) = \text{sample covariance(x, y)}$$

Confidence intervals …

$$s_{\hat{\beta}_0} = s_\varepsilon \sqrt{1 + \frac{\bar{x}^2}{s_x^2}} = \text{std. error of intercept}$$

$$s_{\hat{\beta}_1} = \frac{s_\varepsilon}{s_x \sqrt{n}} = \text{std. error of slope}$$

$$s_\varepsilon = s_y \sqrt{1 - r^2} = \text{std. deviation of noise}$$

Regression summary

The regression equation is

Weight = -25.5 + 0.588 height

Predictor	Coef	St.Dev	t	p-value
Constant	-25.52	46.19	-0.55	0.594
Height	0.5883	0.2648	2.22	0.053

S = 6.606 R-sq = 35.4% R-sq(adj) = 28.2%

Coefficient of Determination

Ratio of variance "explained" by the fit to the total variance of the response variable:

$$R^2 = \frac{s_{\hat{y}}^2}{s_y^2}$$

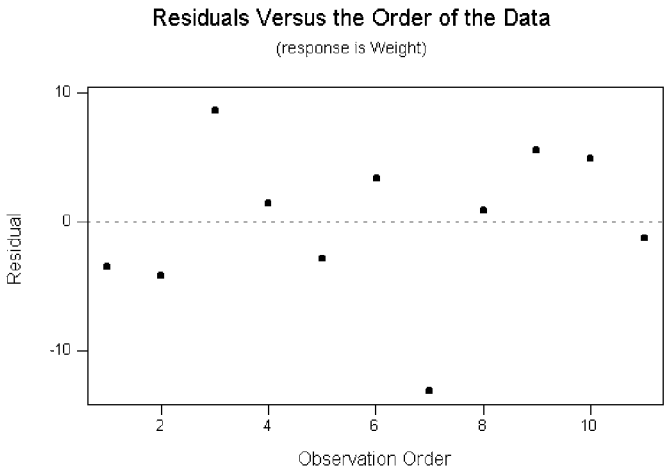
= square of the correlation coefficient

Model checking

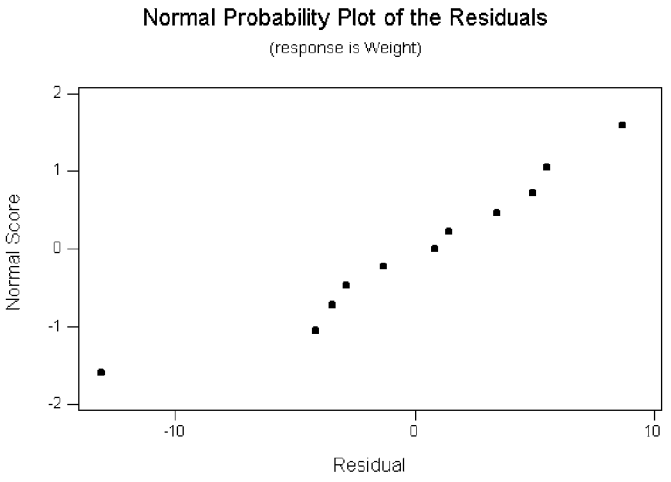
In addition to looking at R2 and p-value, it is also very important to check how well the model fits the data by looking at the residuals. The residuals should be:

- Independent of each other
- Normally distributed ➔ Std. Resids ~ N(0,1)
- Independent of the fitted value

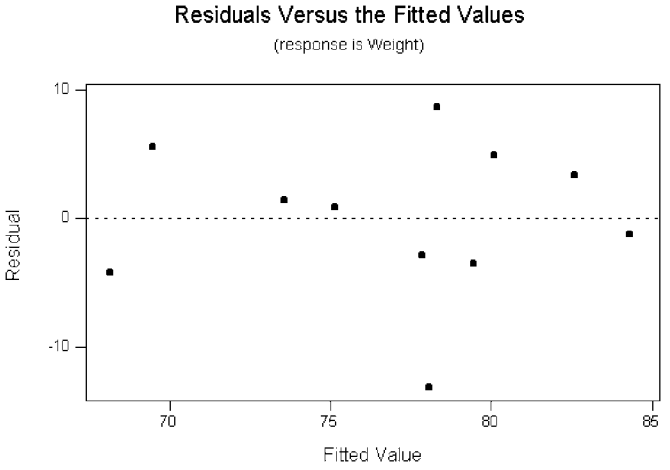
Residuals versus order



Residuals normally distributed?



Residuals versus fitted values



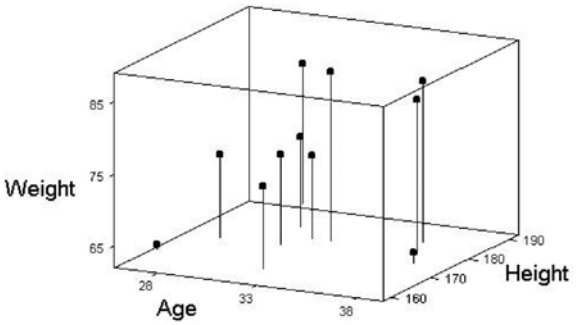
Influential observations

Some values far from the main cloud can have high leverage on the line of best fit and are known as influential observations. They are not necessarily outlier values in either x or y.

8. Multiple and nonlinear regression

- Multiple regression
- Multivariate regression
- Non-linear responses
- Parametric vs. non-parametric regression

Multiple regression



Regression summary

The regression equation is

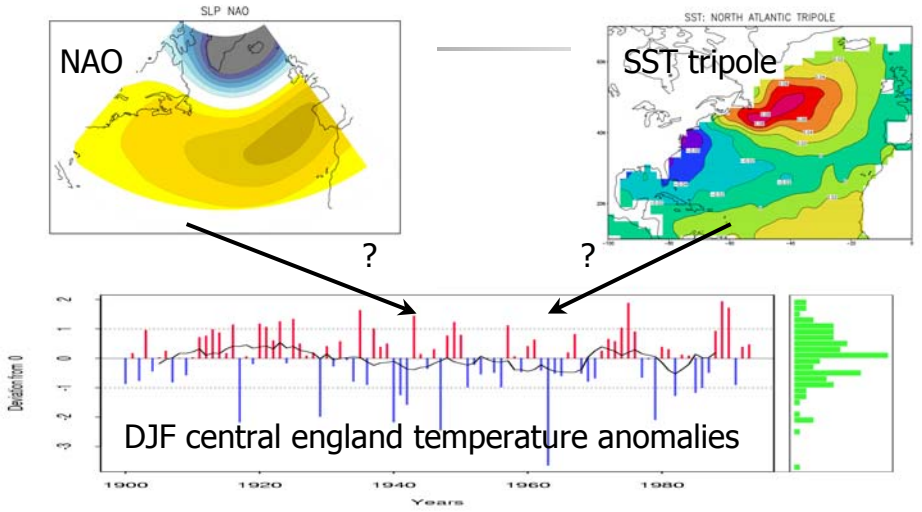
Weight = -40.4 + 0.517 Age + 0.577 Height

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

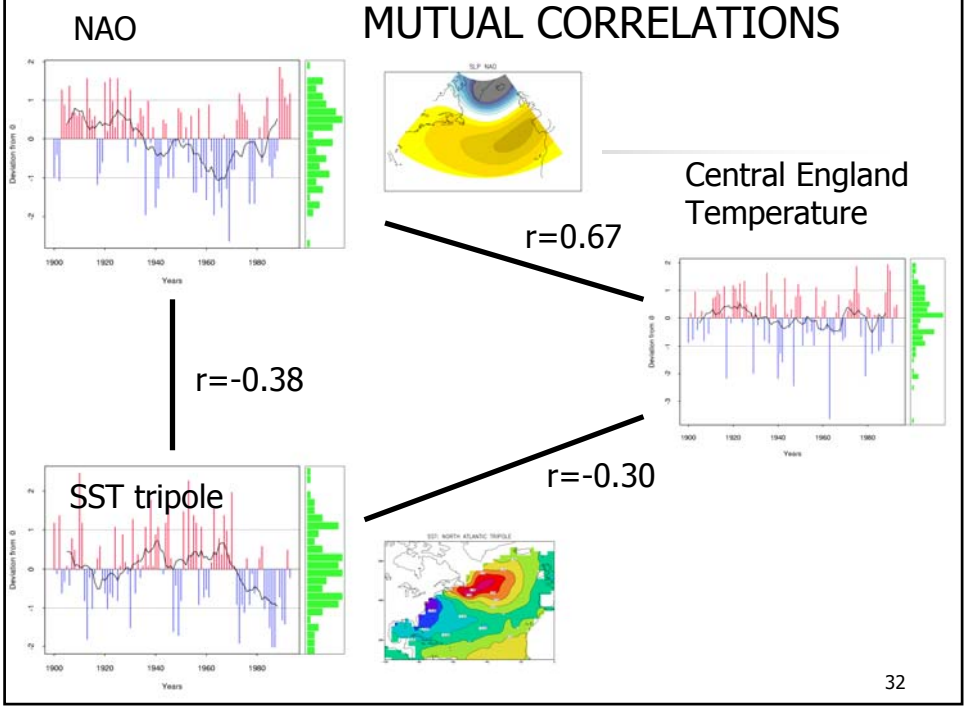
Source	DF	SS	MS	F	p-value
Regression	2	253.66	126.8	2.86	0.115
Residual	8	354.34	44.29		
Total	10	608.00			

A climate example: the role of the ocean



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

MUTUAL CORRELATIONS



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:

$$CET = a \times NAO + b \times SST + \varepsilon$$

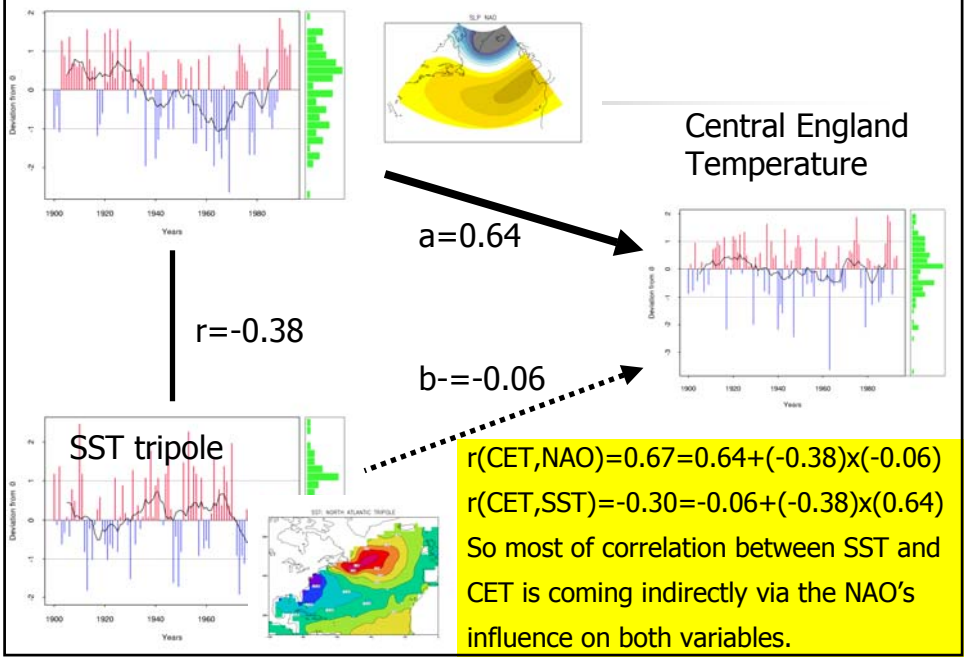
Using data from 1900-1994, we obtain estimates of:

$$\hat{a} = +0.64 \pm 0.08$$

$$\hat{b} = -0.06 \pm 0.08$$

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

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. Introduction to time series

- Basic concepts
- Trend, periodic, and irregular components
- Filtering and smoothing
- Serial correlation
- Autoregressive time series models

Basic concepts

- Discrete time series = sequence of values (x1,x2,x3,…,xn) recorded at discrete times (t1,t2,t3,…,tn) (usually regular).
- Explore, forecast, and control systems
- Time domain vs. frequency domain ?

Further reading: C. Chatfield little red book

Time series components

X = Trend + Periodic cpts. + Irregular

- Trend=smooth long-term changes in mean
- Periodic cpt=cyclic terms such as annual cycle, diurnal cycle, etc.
- Irregular cpt=noisey random part

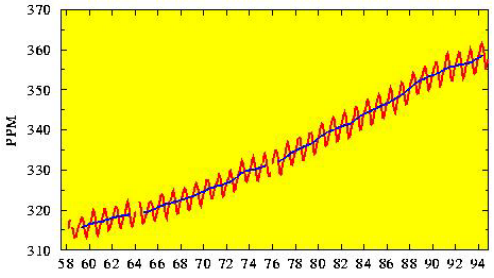
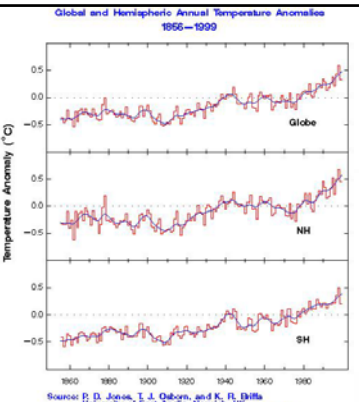
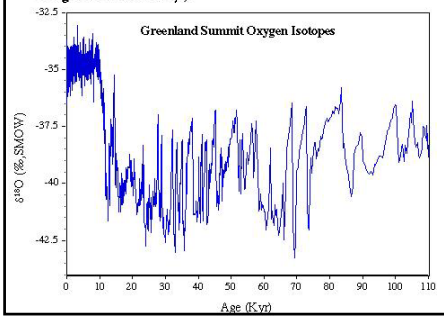
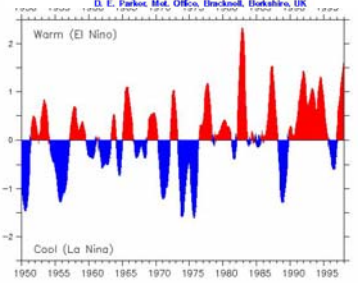


Figure 1.18 Monthly mean (red line) carbon dioxide concentrations (ppm) measured at Mauna Loa, Hawaii, 1958-1994. Blue line is the twelve-month running mean. The data through 1973 are from C. D. Keeling at Scripps Institute of Oceanography. (Data provided by the Climate Monitoring Diagnostics Laboratory.)







Filtering and smoothing

$$y_t = \sum_{lag=-k^{y_t}}^{lag=+k} w_{lag} x_{t+lag}$$

- Low-pass (smoothing)
- Band-pass
- High pass

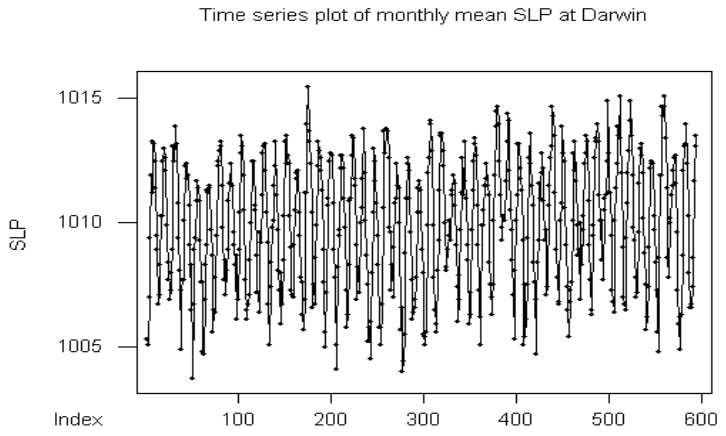
Serial correlation

Correlation between successive values in the time series caused by persistence.

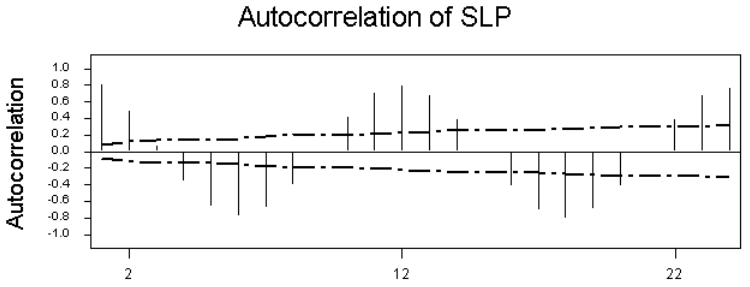
Can be measured using autocorrelations:

$$r_{lag} = cor(x_t, x_{t+lag})$$

Sea-level pressure at Darwin



Autocorrelation function



Lag	Corr	T	LBQ	Lag	Corr	T	LBQ	Lag	Corr	T	LBQ	Lag	Corr	T	LBQ
1	0.80	19.61	386.62	8	-0.38	-3.83	1555.20	15	-0.01	-0.11	2719.90	22	0.39	2.58	3971.56
2	0.49	7.86	529.32	9	0.01	0.13	1555.31	16	-0.40	-3.11	2819.61	23	0.67	4.38	4253.18
3	0.07	0.98	532.01	10	0.41	4.06	1658.96	17	-0.69	-5.22	3110.16	24	0.77	4.83	4618.84
4	-0.34	-5.04	603.38	11	0.70	6.69	1956.57	18	-0.79	-5.72	3491.75				
5	-0.64	-9.04	852.56	12	0.80	7.11	2343.91	19	-0.68	-4.72	3780.65				
6	-0.76	-9.45	1200.63	13	0.68	5.64	2629.47	20	-0.39	-2.62	3876.55				
7	-0.66	-7.25	1467.48	14	0.38	3.01	2719.78	21	-0.00	-0.00	3876.55				

Autoregressive models

Simplest model is Autoregressive order 1 (AR1) red noise model:

$$x_{t+1} = \varphi x_t + \varepsilon_t$$

Plus there are many many other time series models …