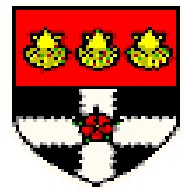
Data analysis methods in weather and climate research

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



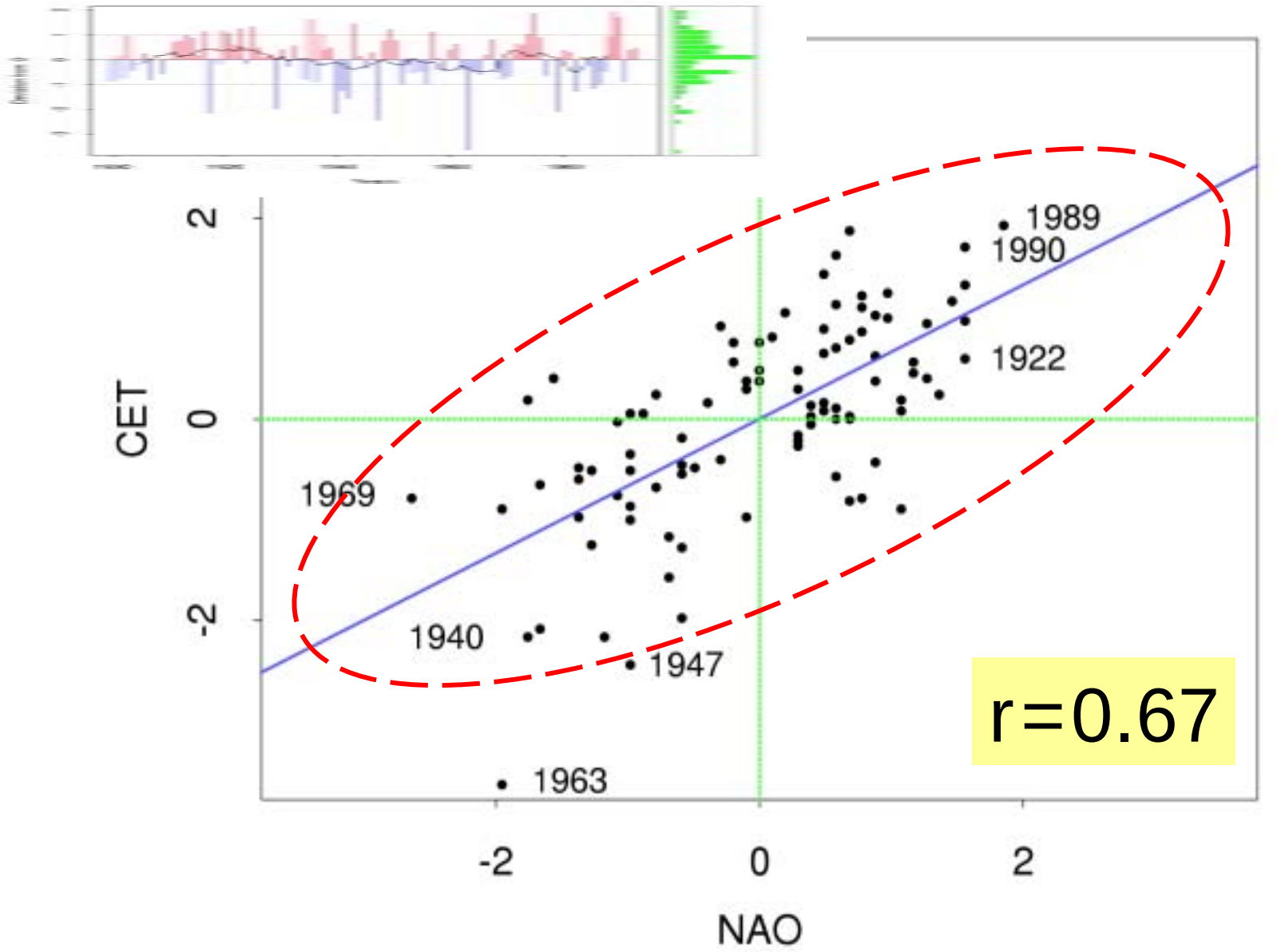
7. Basic linear regression

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

Modelling strategy

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

Central England Temperature versus NAO (winters 1900-1994)



Covariance of 2 variables

$$\mathrm{cov}(X,Y) = E((X - E(X))(Y - E(Y)))$$

$$= E(XY) - E(X)E(Y)$$

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

$$= \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right)$$

Correlation (product moment)

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

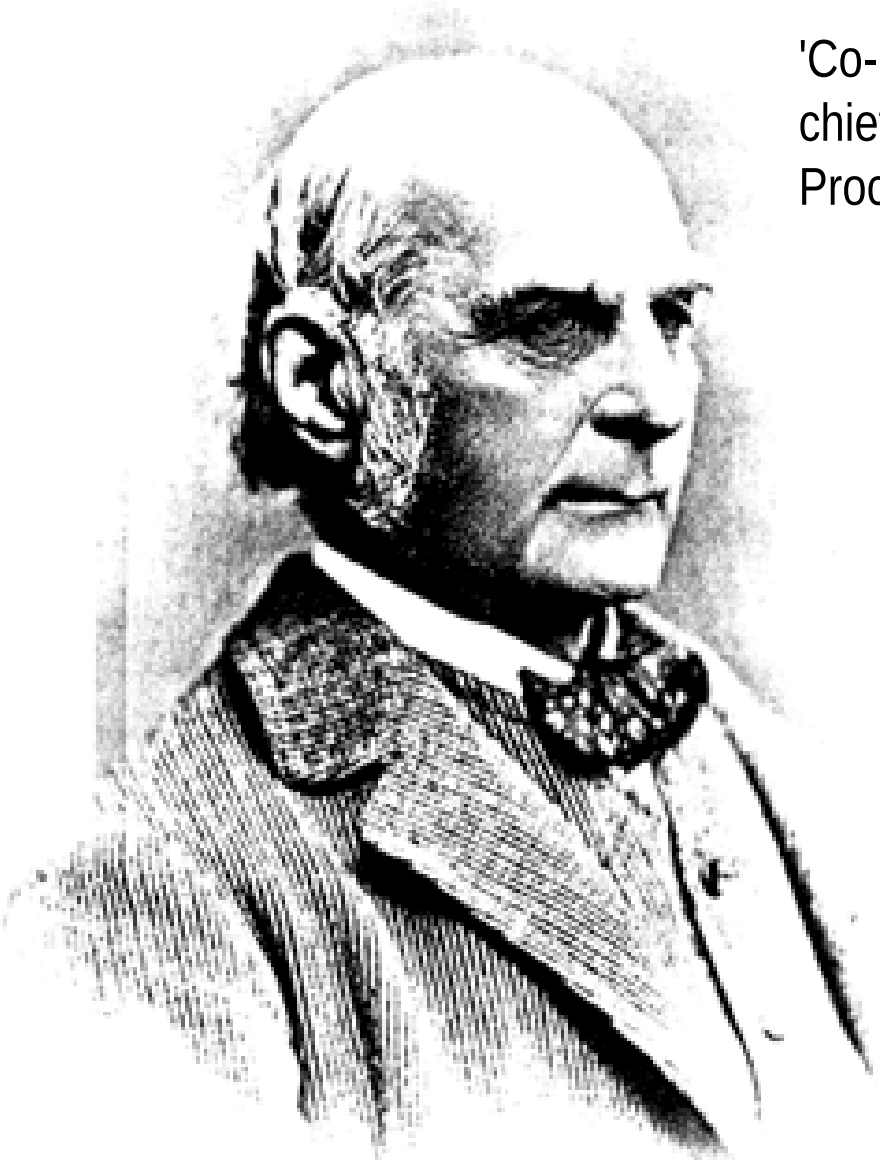
where

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

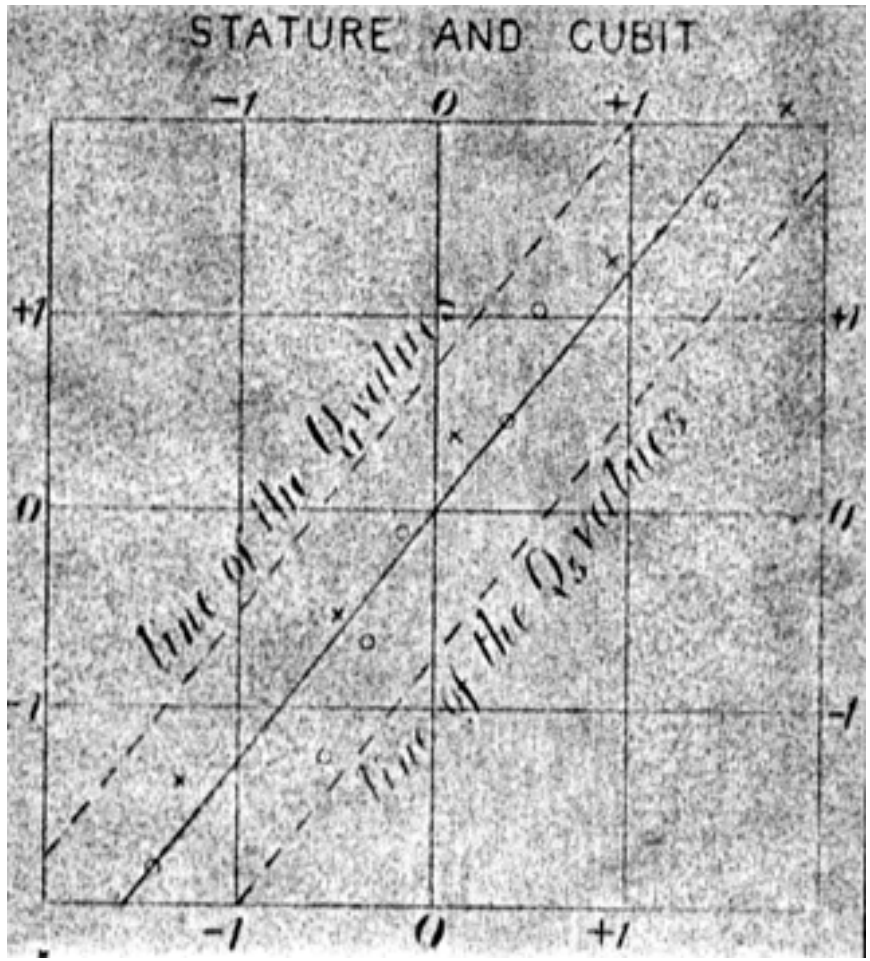
Measure of linear association between two variables

- symmetric in x and y
- not affected by changes in mean
- not affected by changes in standard deviation

Sir Francis Galton FRS 1822-1911

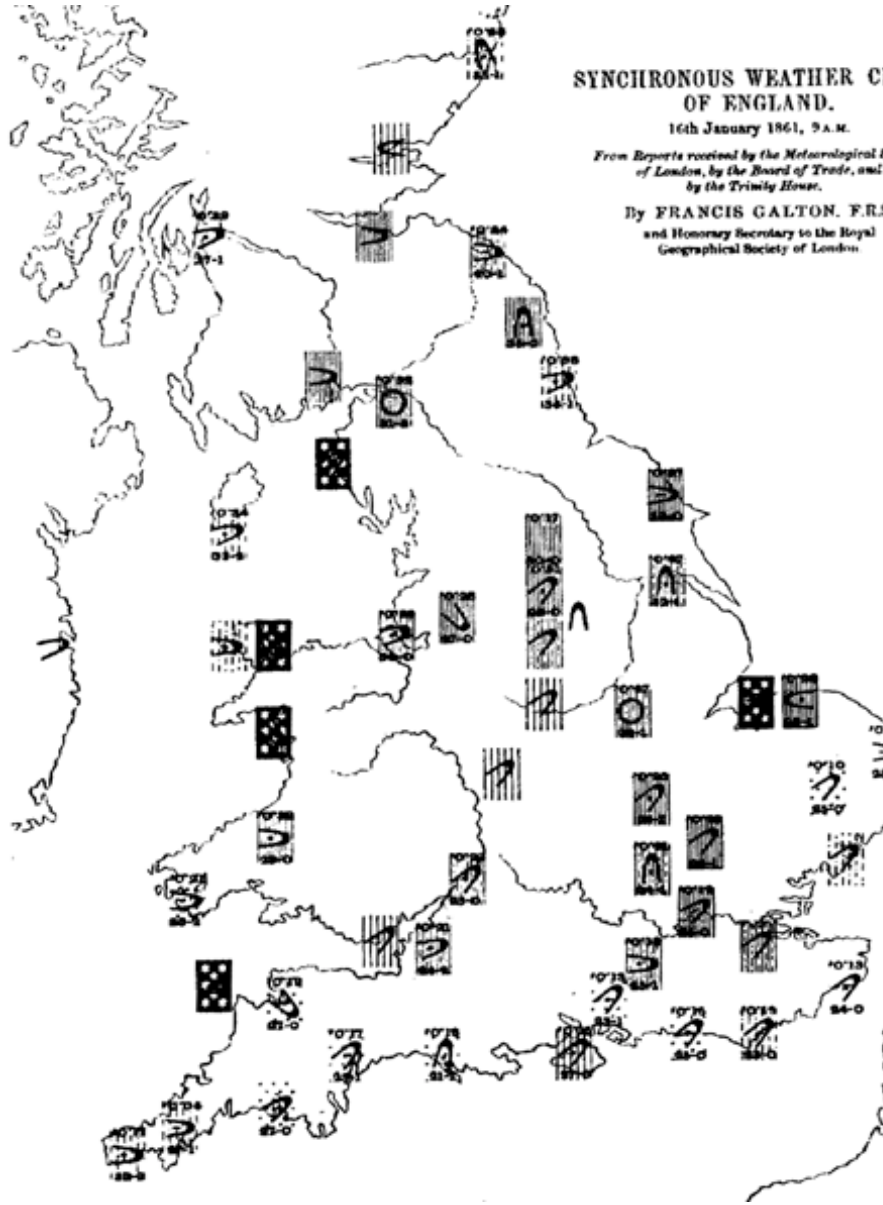


'Co-relations and their measurement,
chiefly from anthropometric data.'
Proceedings of the Royal Society 45, pp. 135-45. 188

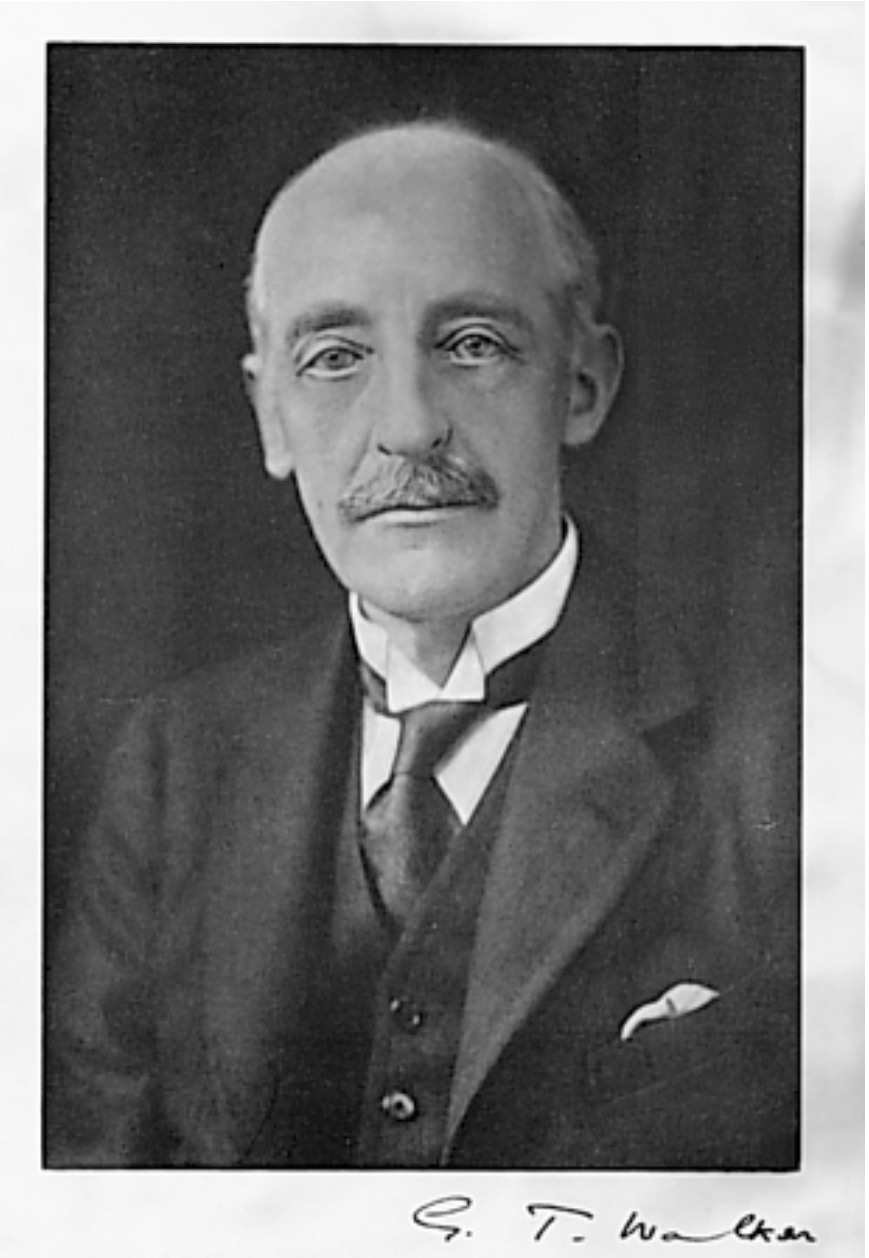


Galton's contributions to meteorology:

- "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 $\max(r_1, r_2, ..., r_m)$:

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

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

multiple testing $\Rightarrow k_m = \Phi^{-1}(0.5 + 0.5 \times 0.5^{\frac{1}{m}}) / \Phi^{-1}(0.7$

The basic idea

Model the relationship between two random variables as:

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

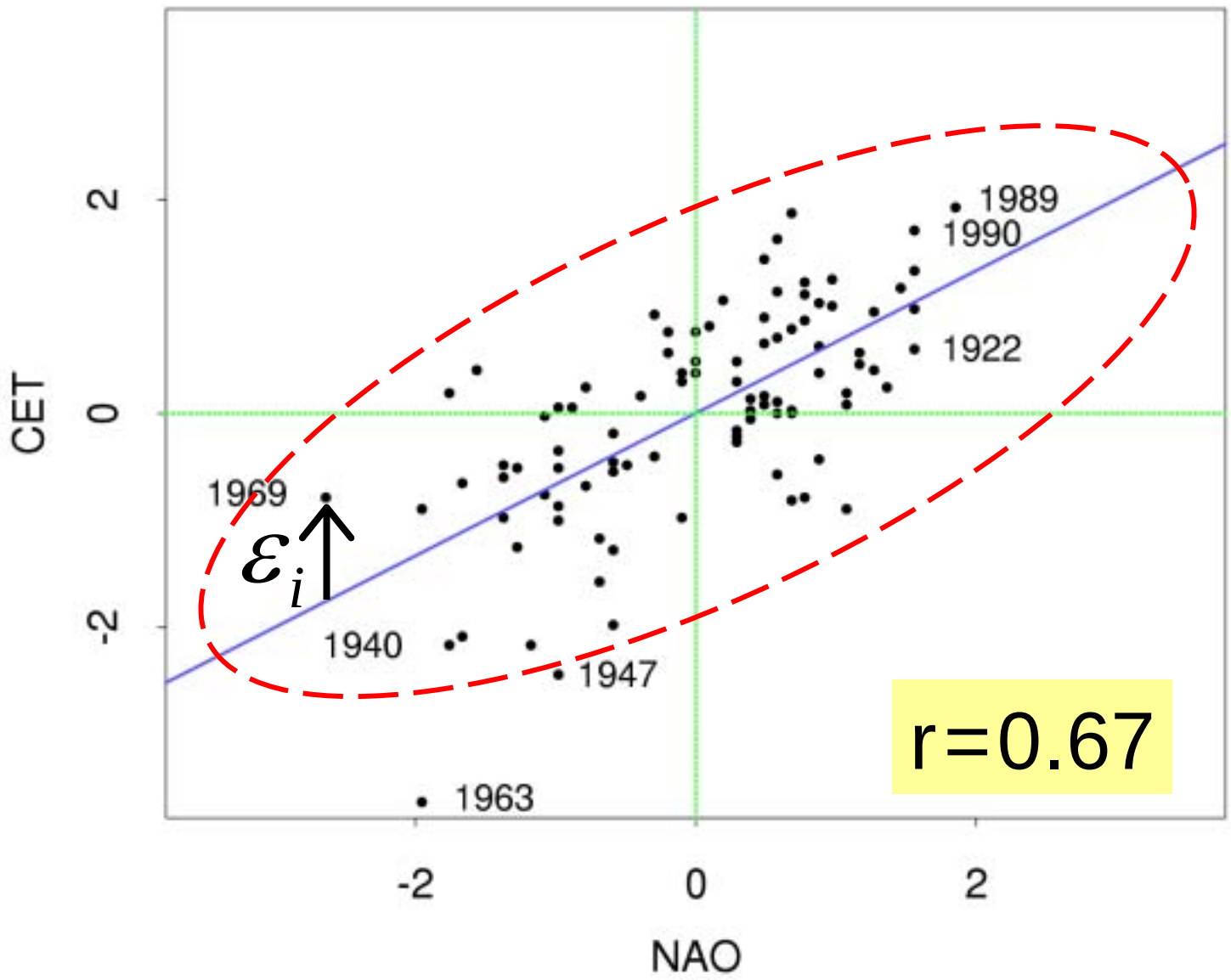
where

y_i = measured value of "response variable"

x_i = measured value of "explanatory variable"

ε_i = random normally distributed noise

Central England Temperature versus NAO (winters 1900-1994)



Ordinary Least Squares (OLS) 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

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

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

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

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

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

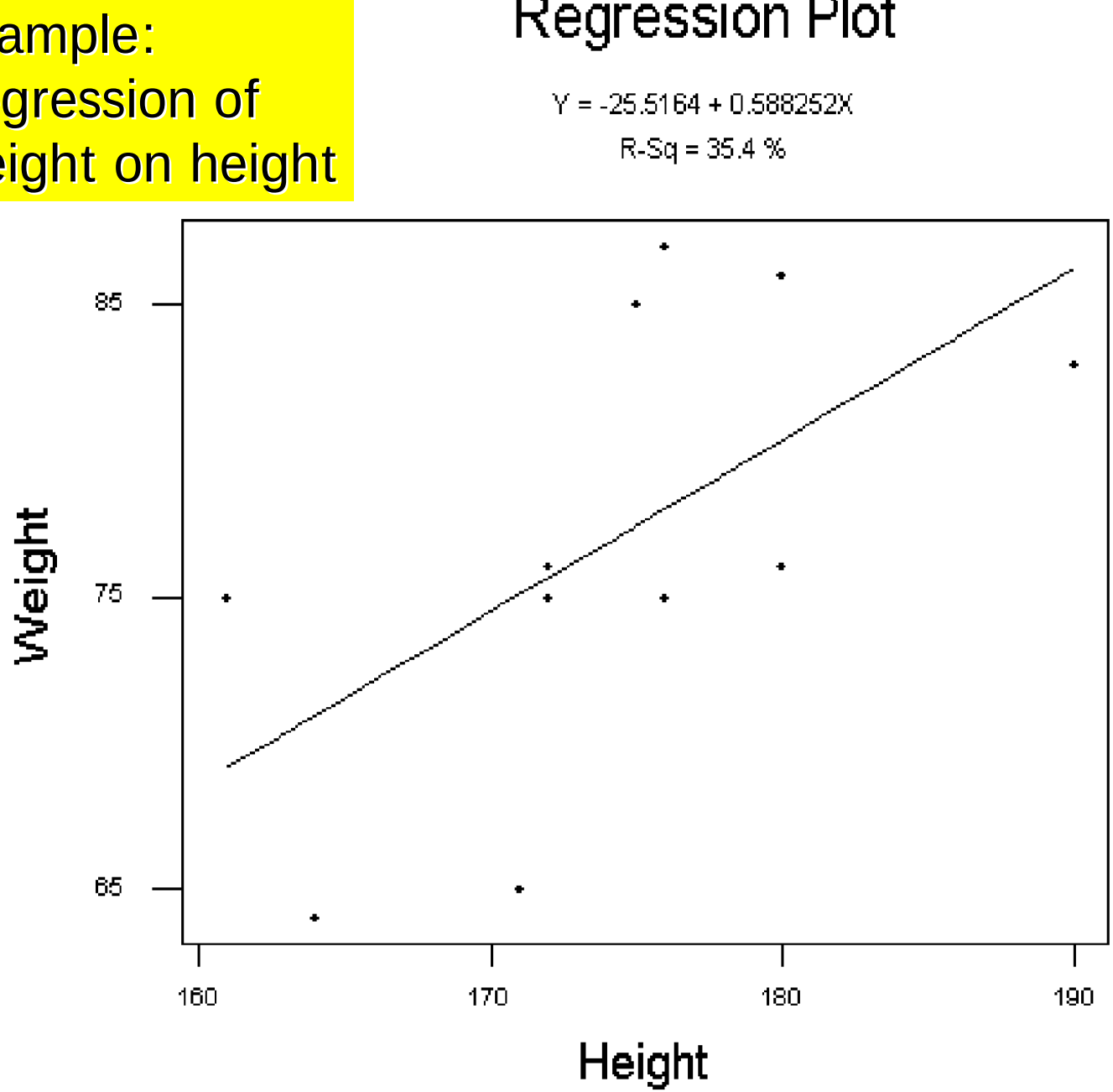
Standard errors on estimates

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

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

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

Example: Regression of weight on height



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%

A deeper view …

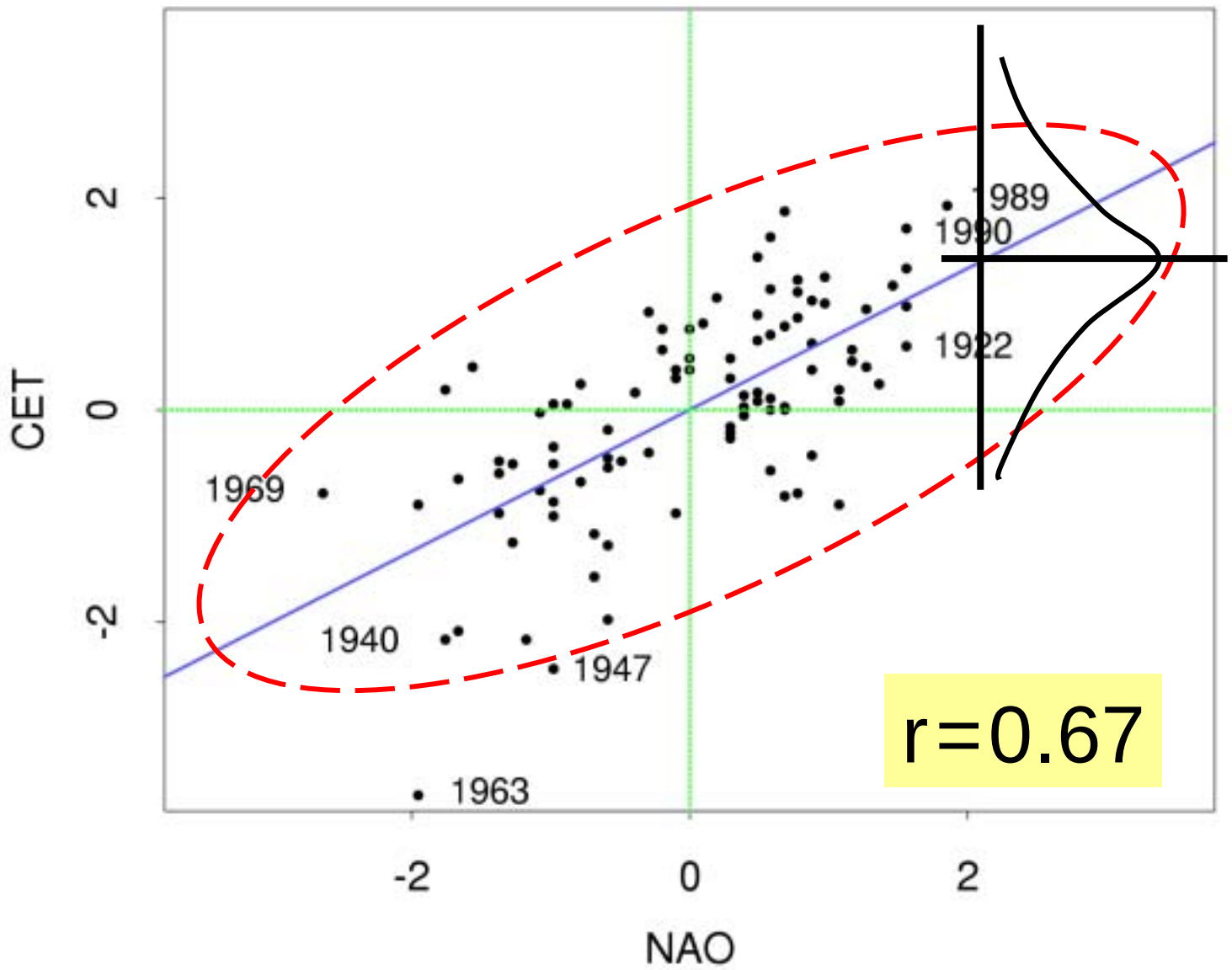
Instead of just thinking of an additive signal+noise model, a deeper insight can be obtained by thinking of the regression like this:

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

or

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

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

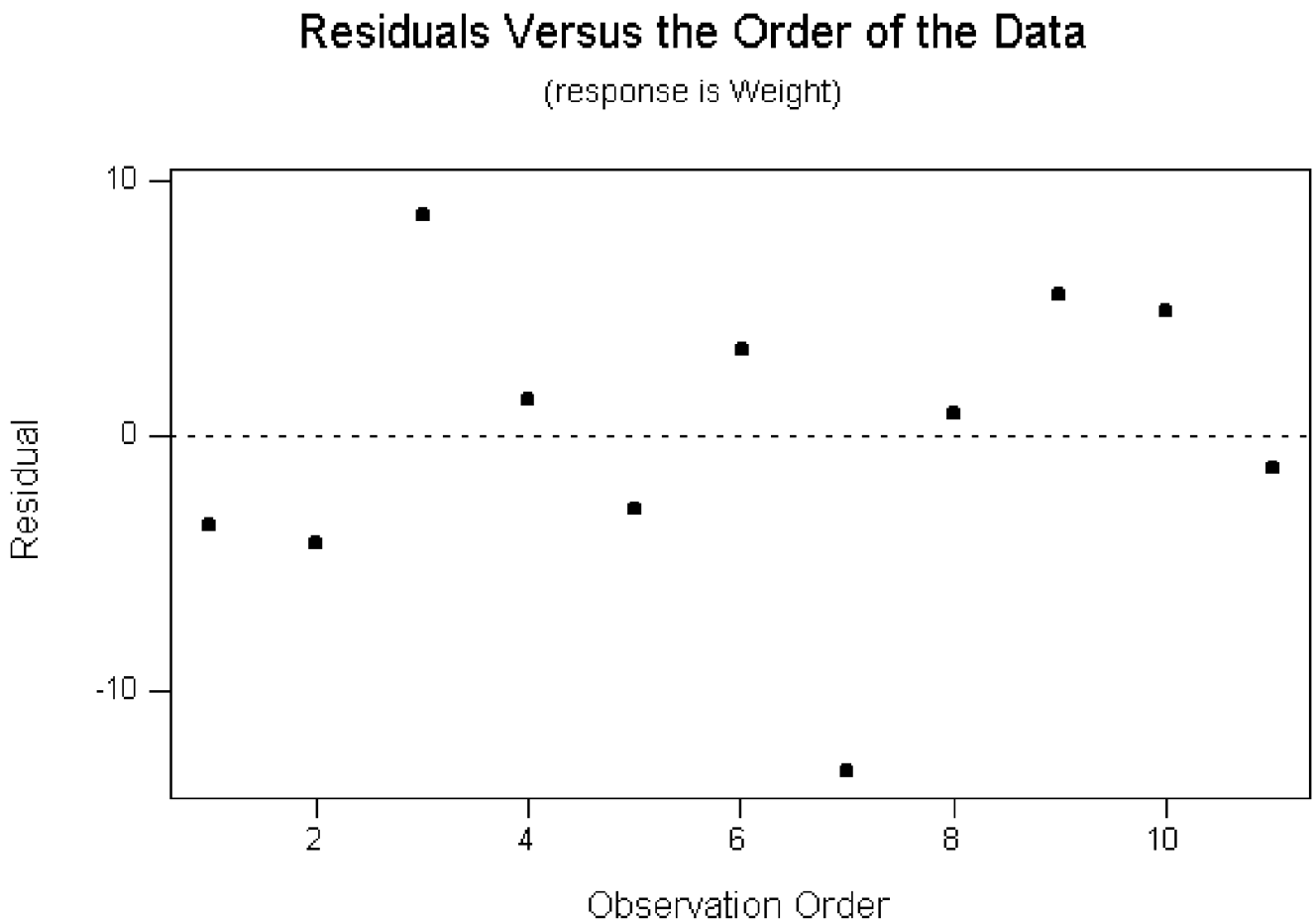
$$Y \sim N(\alpha + \beta X, \sigma_\varepsilon^2)$$

Model checking

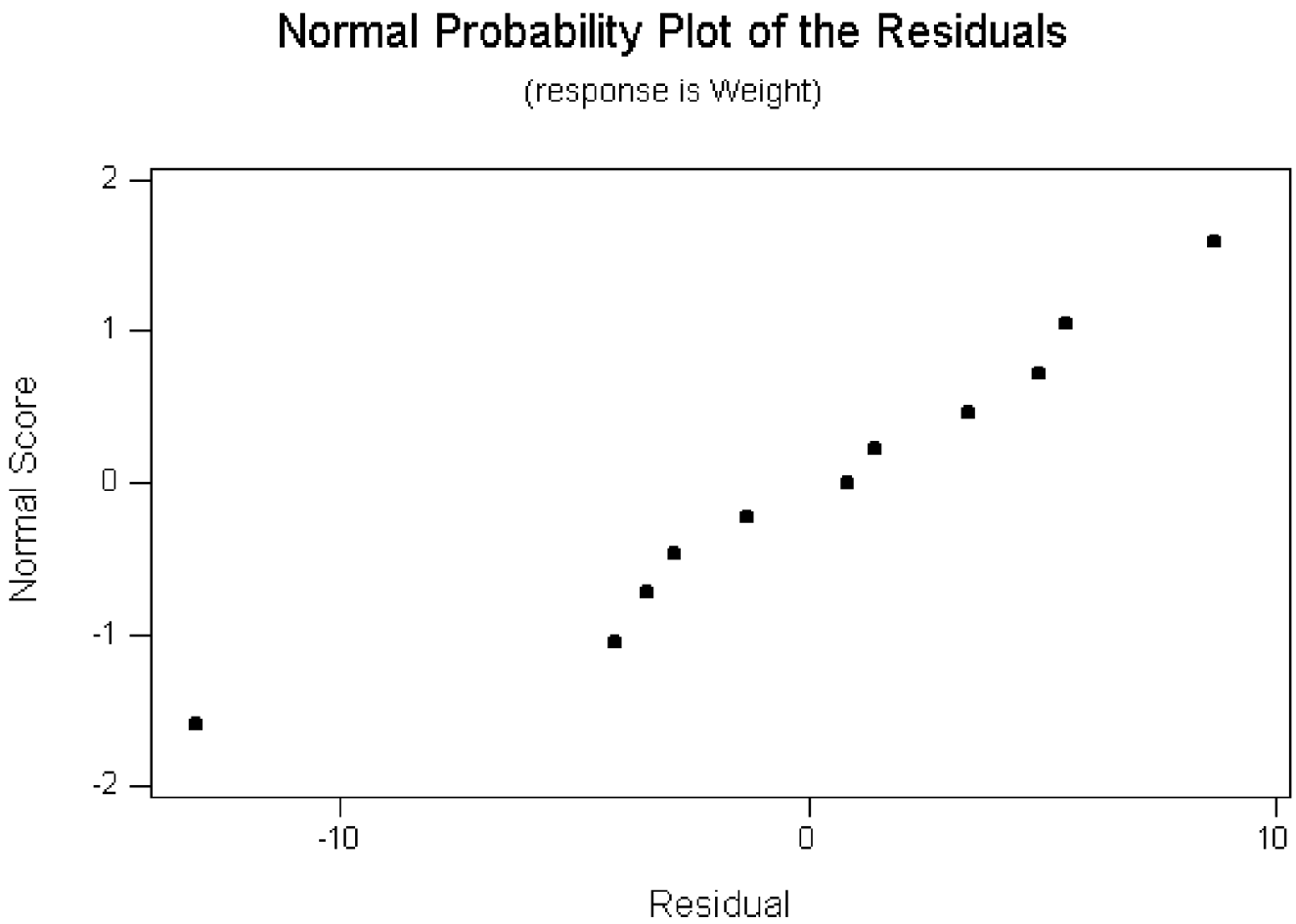
In addition to looking at R2 and p-value, it is also very important to check how well the model fits the data by looking at the residuals. The residuals should be:

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

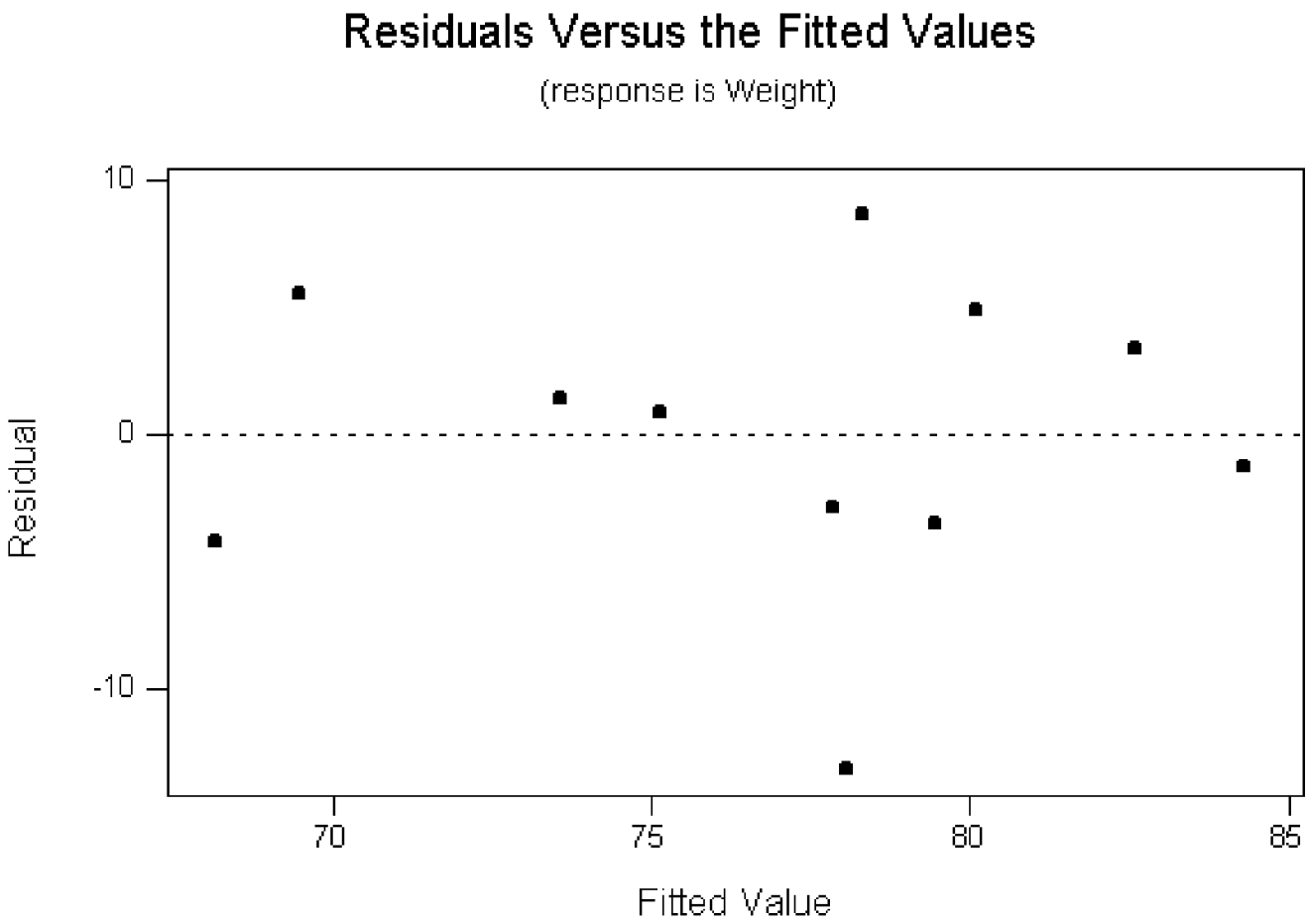
Residuals versus order



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.

Summary