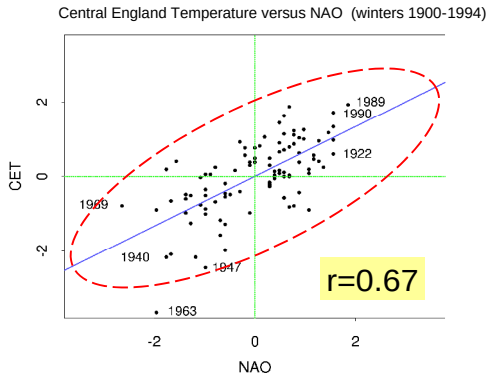
Data analysis methods in weather and climate research

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

7. Basic linear regression

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

7. Relationships between two variables (bivariate statistics)



7. Correlation (product moment)

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

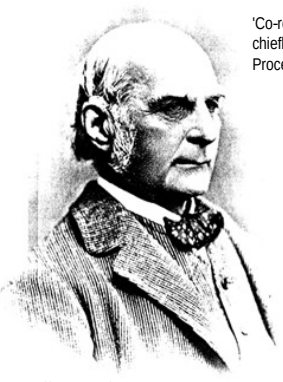
where

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

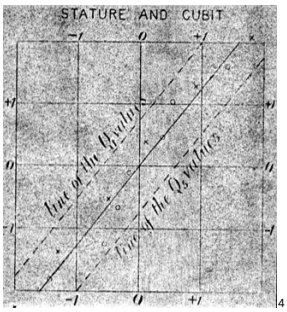
sample covariance

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

7. Sir Francis Galton FRS 1822-1911

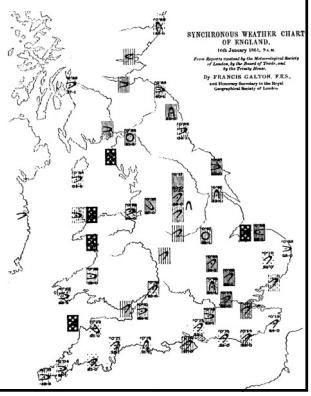


'Co-relations and their measurement, chiefly from anthropometric data.'
Proceedings of the Royal Society 45, pp. 135-45. 1888



7. 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



7. Linear regression of Y on X

Model one variable Y as the sum of a fixed response to X plus a random component:

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

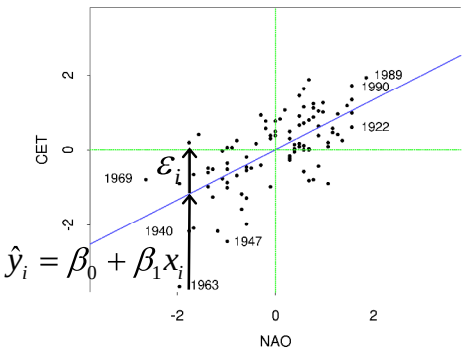
where

y_i = measured value of "response variable"

x_i = measured value of "explanatory variable"

ε_i = random normally distributed noise

7. Response = predicted part + residual



7

7. Ordinary Least Squares (OLS) Estimation

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

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

where

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

```
> # R commands
> fit<-lsfit(x,y)
```


8

7. OLS point estimates

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

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

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

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

9

7. 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

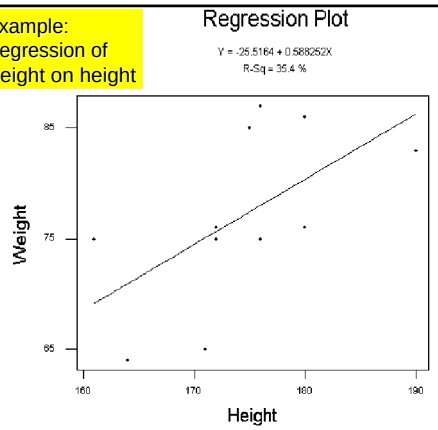
7. Standard errors of parameter estimates

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

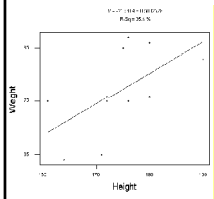
$$s_{\hat{\beta}_1} = \frac{s_\varepsilon}{s_x \sqrt{n}} = \text{std. error of slope}$$

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

Example: Regression of weight on height



7. Example of regression summary



The regression equation is

Weight = -25.5 + 0.588 height

Predictor	Param	Std.Err	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%

Linear regression of weight on height accounts for 35.4% of the variance in weight. The OLS estimate of the slope is 0.59±0.27kg/m and is not statistically different from zero at the 5% level of significance (p-value 0.053).

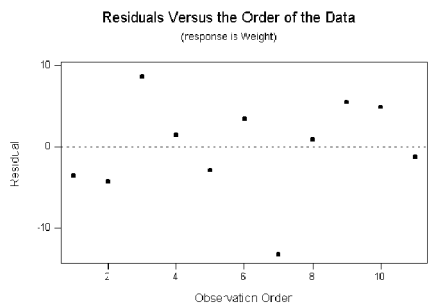
```
> # R commands
> ls.print(lsfit(x,y))
```


7. 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:

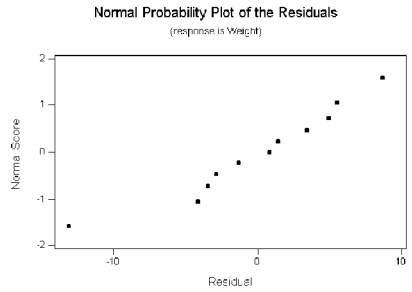
- Independent of each other
- Normally distributed ➔ Std. Resids ~ N(0,1)
- Independent of the fitted value

7. Residuals versus order



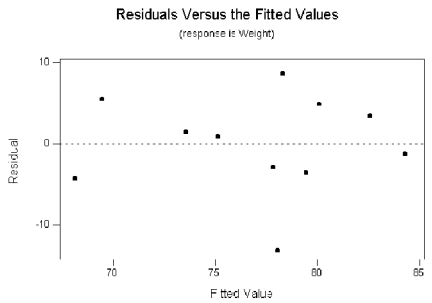
→ No obvious structure – so residuals are independent

7. Probability distribution of residuals



→ Close to a straight line so residuals normally distributed

7. Residuals versus fitted values



→ No obvious dependency so linear approximation ok

7. 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.

7. Summary

- Dependency of response Y on explanatory variable X can be modelled using linear regression
- OLS linear regression is a probability model for Y|X
- Model assumptions:
  - Y|X are independently normally distributed
  - The mean of Y|X is linearly related to X
  - The variance of Y|X is constant
- Assumptions SHOULD be checked by looking at residual diagnostics carefully