

MA329 Statistical Linear Models

Prof. Jian Qing SHI, shijq@sustech.edu.cn

A little bit about me



Nanjing
Hong Kong
Warwick
Glasgow
Newcastle/London
Shen Zhen



CH1. Introduction—what is a regression model

Recall: Covariance

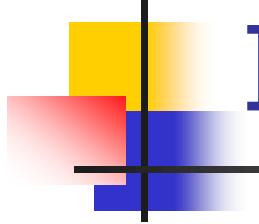
Sampling covariance

$$\text{cov}(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{X})(y_i - \bar{Y})}{n-1}$$

Remark 1.1

$n \rightarrow \infty$

$\text{cov}(\mathcal{D}, y)$



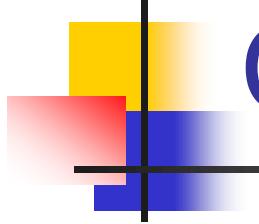
Interpreting Covariance

$\text{cov}(X,Y) > 0 \rightarrow X \text{ and } Y \text{ are positively correlated}$

$\text{cov}(X,Y) < 0 \rightarrow X \text{ and } Y \text{ are inversely correlated}$

$\text{cov}(X,Y) = 0 \rightarrow X \text{ and } Y \text{ are independent}$

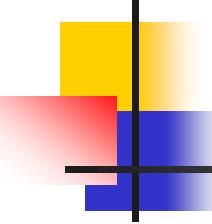
(un correlated)



Correlation coefficient

- Pearson's Correlation Coefficient is standardized covariance (unitless):

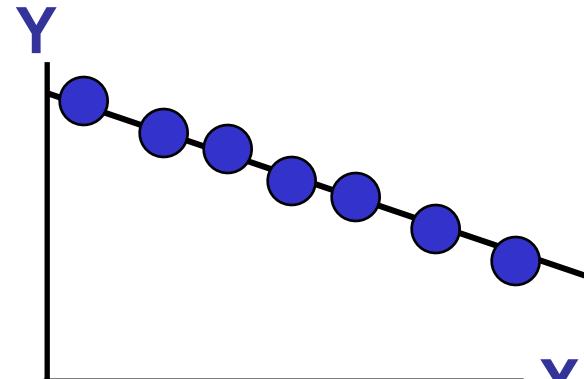
$$r = \frac{\text{covariance}(x, y)}{\sqrt{\text{var } x} \sqrt{\text{var } y}}$$



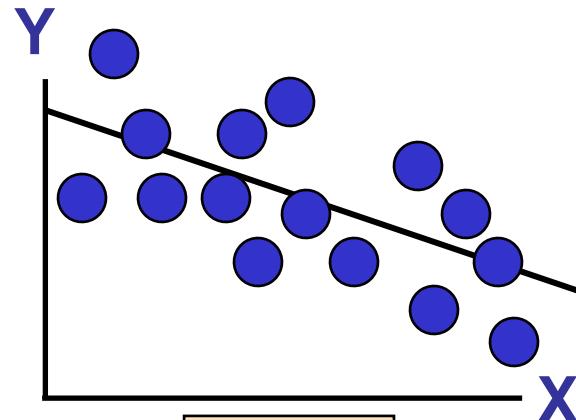
Correlation

- Measures the relative strength of the *linear* relationship between two variables
- Unit-less
- Ranges between -1 and 1
- The closer to -1 , the stronger the negative linear relationship
- The closer to 1 , the stronger the positive linear relationship
- The closer to 0 , the weaker any positive linear relationship

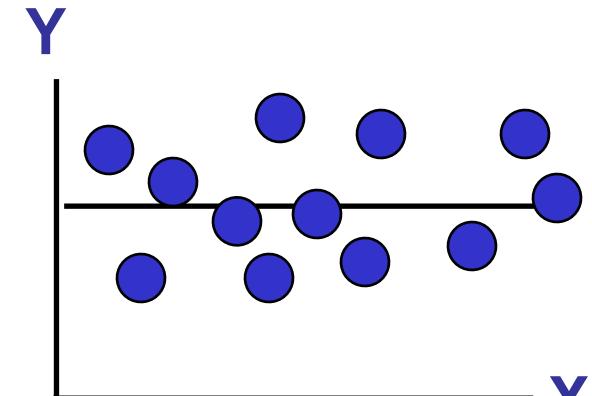
Scatter Plots of Data with Various Correlation Coefficients



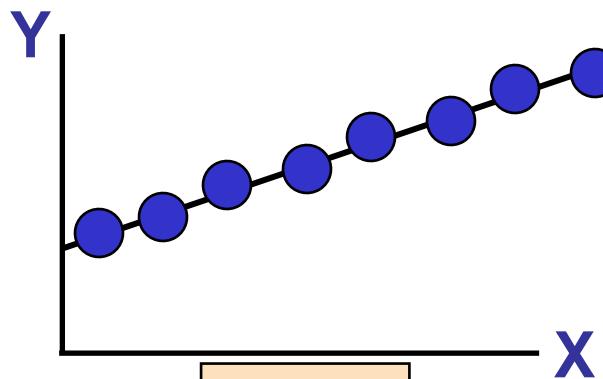
$$r = -1$$



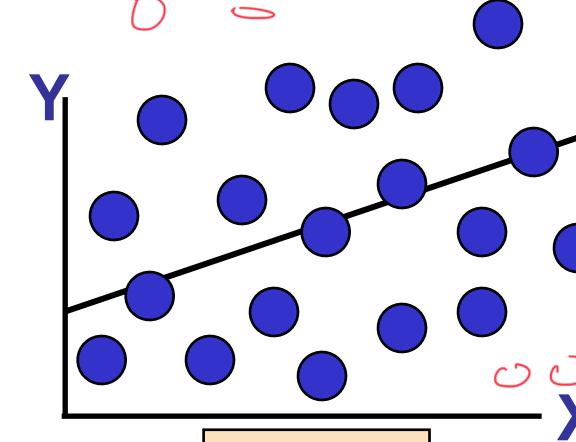
$$r = -.6$$



$$r = 0$$

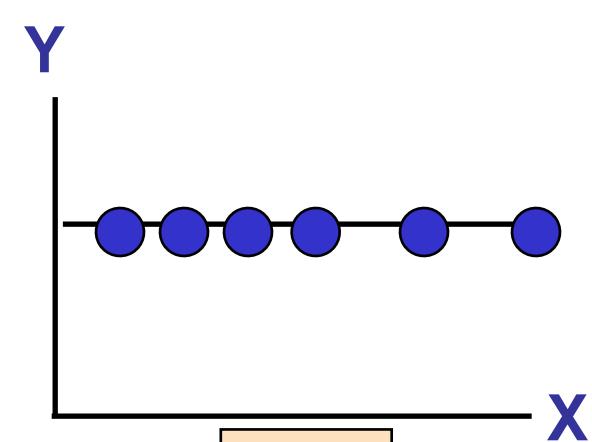


$$r = +1$$



$$r = +.3$$

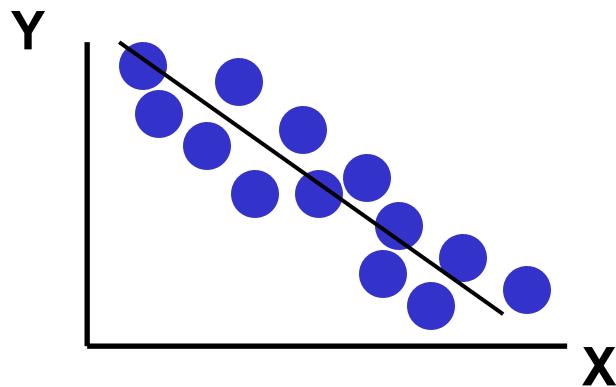
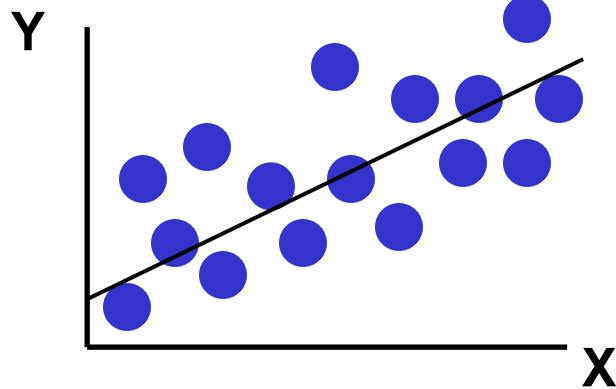
$H_0: \rho = 0$



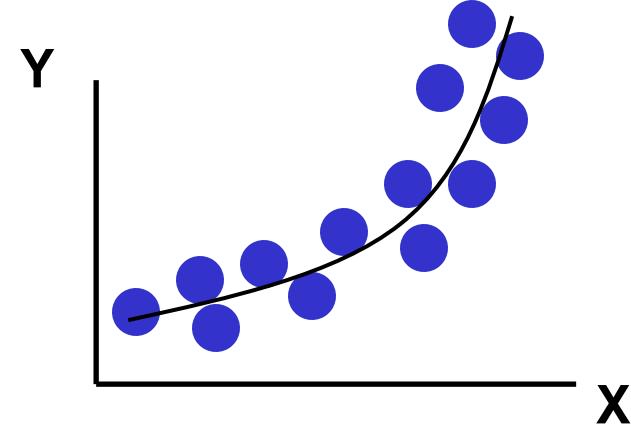
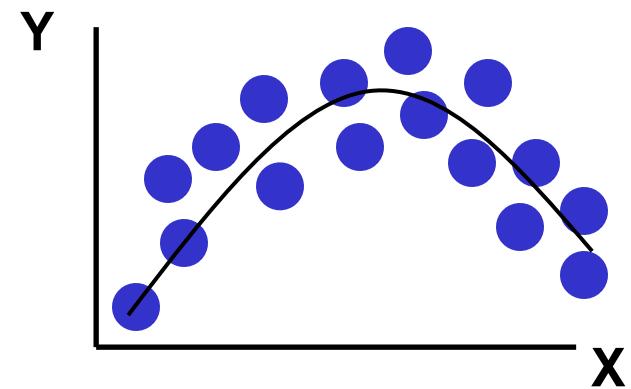
$$r = 0$$

Linear Correlation

Linear relationships

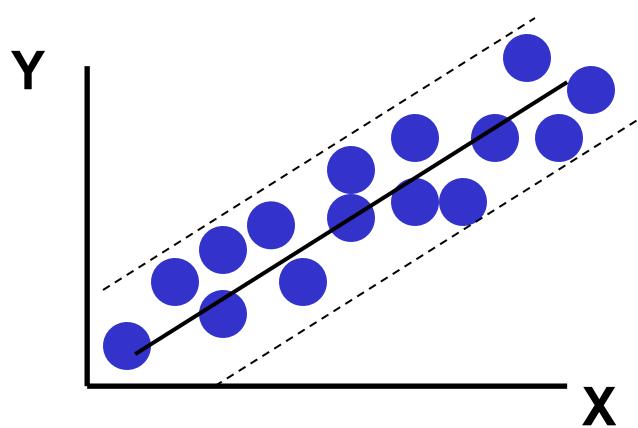


Curvilinear relationships

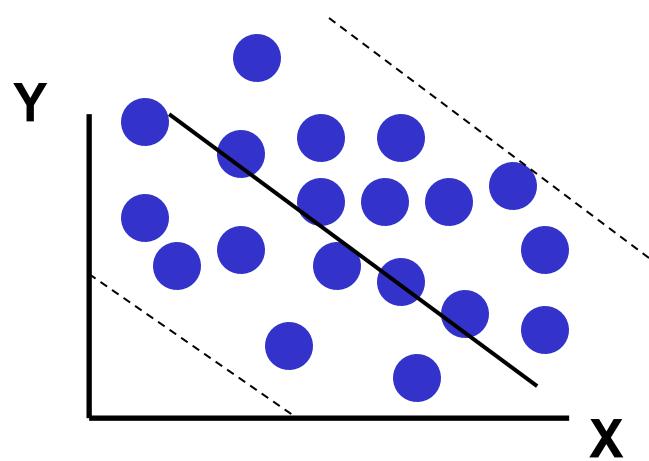
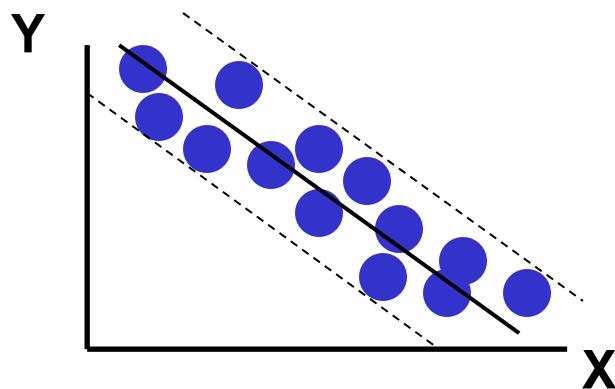
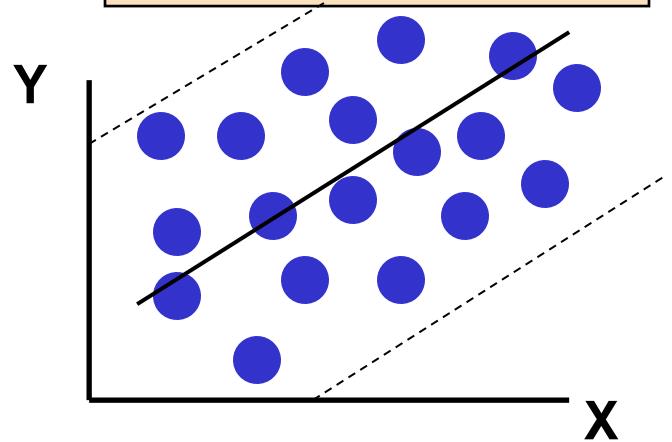


Linear Correlation

Strong relationships

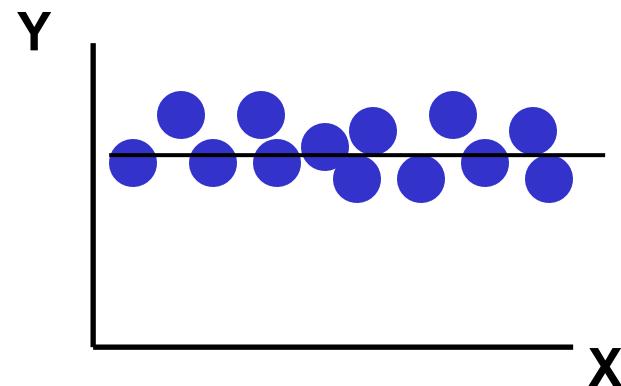
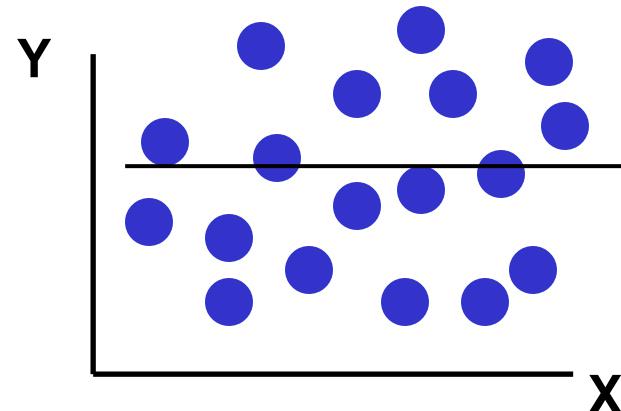


Weak relationships



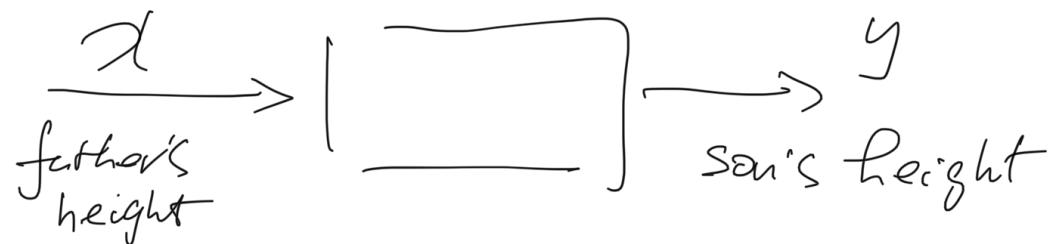
Linear Correlation

No relationship



Linear regression

In correlation, the two variables are treated as equals. In regression, one variable is considered independent (=predictor) variable (X) and the other the dependent (=outcome) variable Y .



$$1^{\circ} \quad y = \beta_0 + \beta_1 x + \xi$$

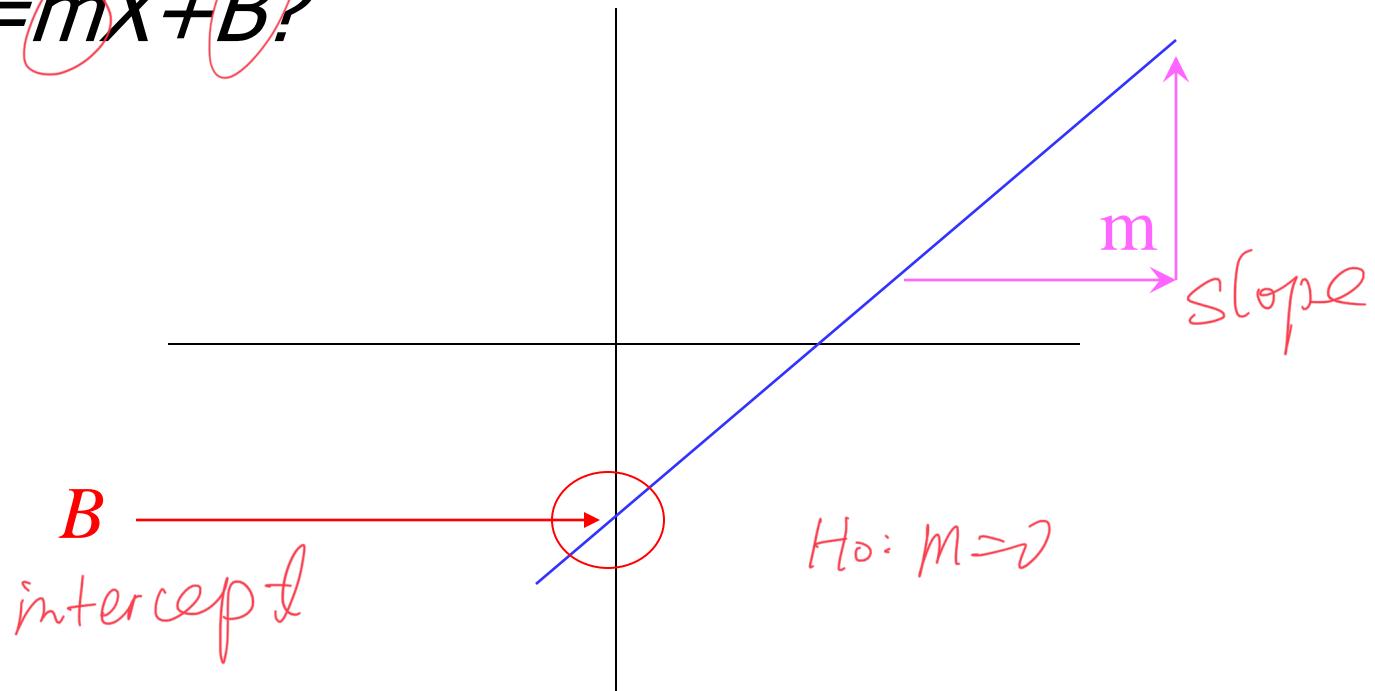
$$2^{\circ} \quad y = f(x, \underline{\theta}) + \xi$$

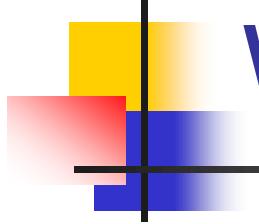
$f(\cdot)$ — given
N.T. estimate $\underline{\theta}$

$f(\cdot)$ — unknown
N.T. estimate $\hat{f}(\cdot)$

What is “Linear”?

- Remember this:
- $Y = mX + B?$





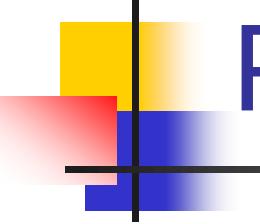
What's Slope?

A slope of 2 means that every 1-unit change in X yields a 2-unit change in Y.

Prediction

If you know something about X, this knowledge helps you predict something about Y. (Sound familiar?...sound like conditional probabilities?)





Regression equation...

Expected value of y at a given level of x =

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

- 1^o How to estimate β_0 - β_1 ?
- 2^o How good is the estimation?

Predicted value for an individual...

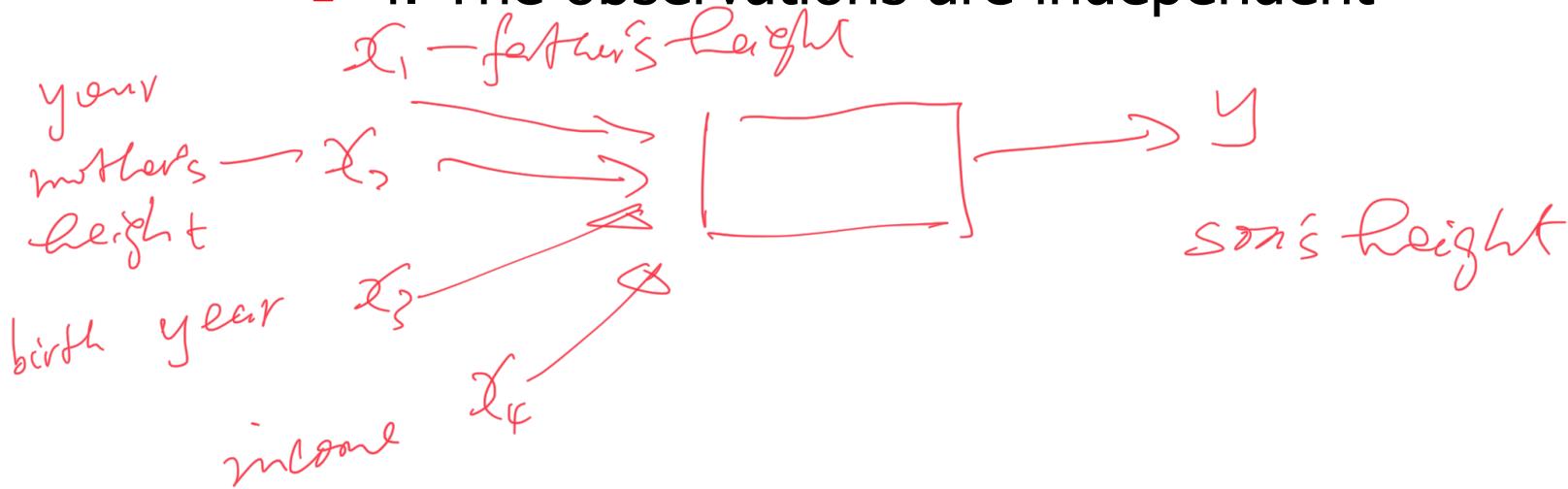
$$\hat{y}_i = \underbrace{\beta_0 + \beta_1 x_i}_{\text{Fixed – exactly on the line}} + \text{random error}_i$$

Fixed –
exactly
on the
line

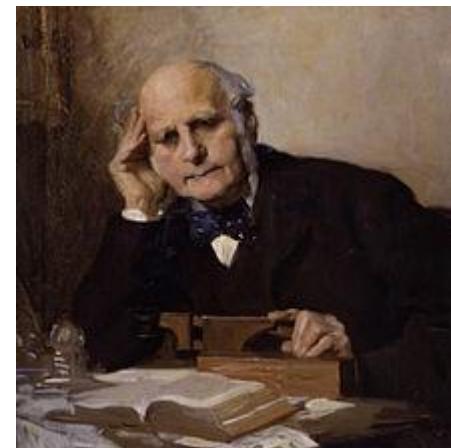
Follows a normal
distribution

Assumptions (or the fine print)

- Linear regression assumes that...
 - 1. The relationship between X and Y is linear
 - 2. Y is distributed normally at each value of X
 - 3. The variance of Y at every value of X is the same (homogeneity of variances)
 - 4. The observations are independent

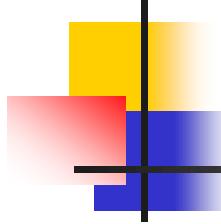


Linear Regression – a brief history



The earliest form of regression was the [method of least squares](#), which was published by [Legendre](#) in 1805, and by [Gauss](#) in 1809. Legendre and Gauss both applied the method to the problem of determining, from astronomical observations, the orbits of bodies about the Sun (mostly comets, but also later the then newly discovered minor planets). Gauss published a further development of the theory of least squares in 1821, including a version of the [Gauss–Markov theorem](#).

The term "regression" was coined by [Francis Galton](#) in the 19th century to describe a biological phenomenon. The phenomenon was that the heights of descendants of tall ancestors tend to regress down towards a normal average (a phenomenon also known as [regression toward the mean](#))



Contents

- Ch1. Introduction
- CH2. Simple linear regression: Model, estimation and testing
- Ch3. Matrix algebra, generalized inverse
- Ch4. Random Vector and Matrices
- Ch5. Multivariate normal distribution
- Ch6. Quadratic Forms
- **Mid-term test**
- Ch7. Multiple regression (I): Model and estimation
- Ch8. Multiple regression (II): Hypothesis testing
- Ch9. Multiple regression (III): Diagnostics and model-building
- Ch10. Analysis of Variance Models