MA329 Statistical Linear Models

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

CH1. Introduction—what is a regression model

Recall: Covariance

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

Interpreting Covariance

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

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

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

Correlation coefficient

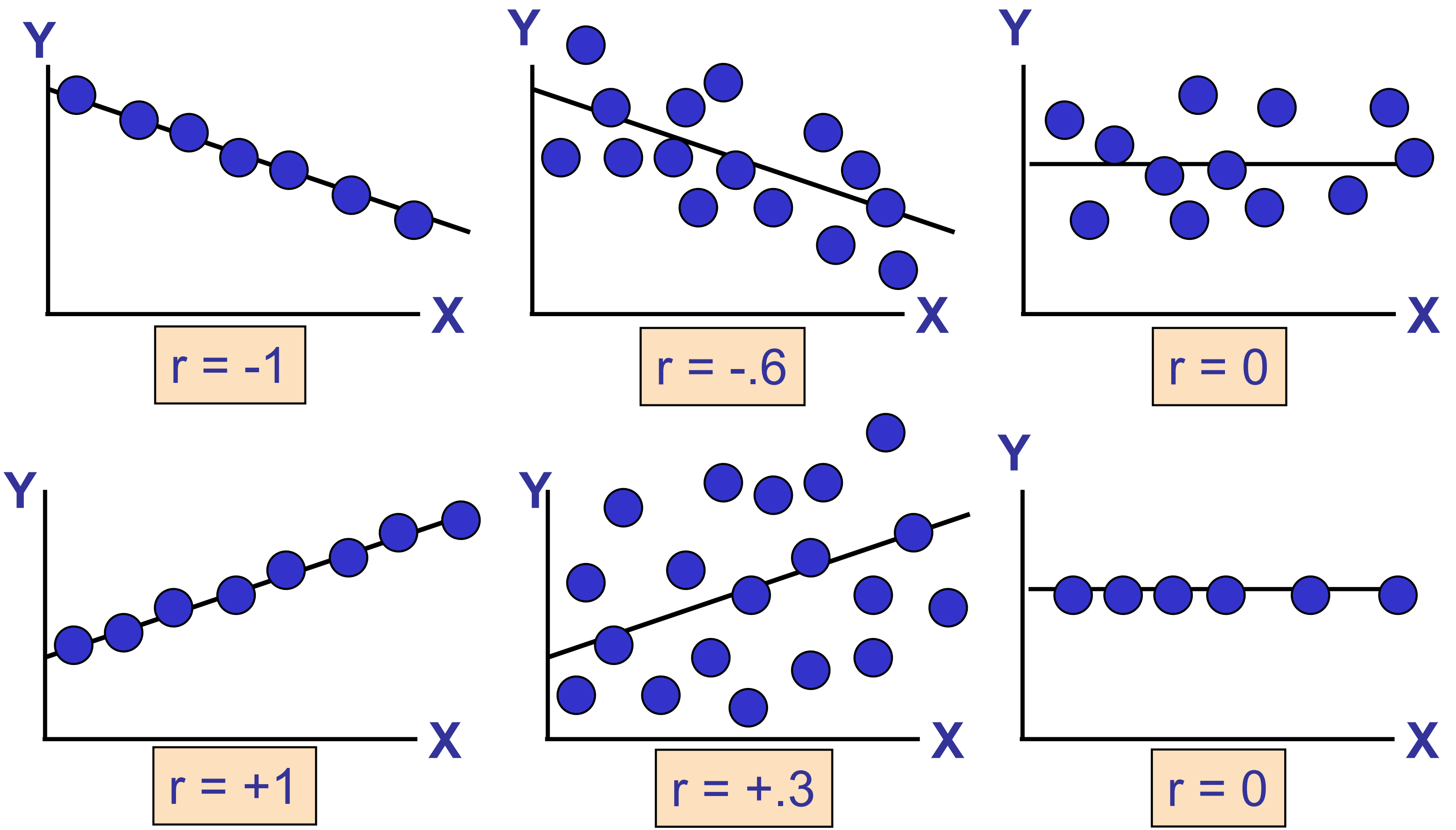
- Pearson's Correlation Coefficient is standardized covariance (unitless):

$$r = \frac{\mathrm{cov}ariance(x, y)}{\sqrt{\mathrm{var}\, x}\sqrt{\mathrm{var}\, y}}$$

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

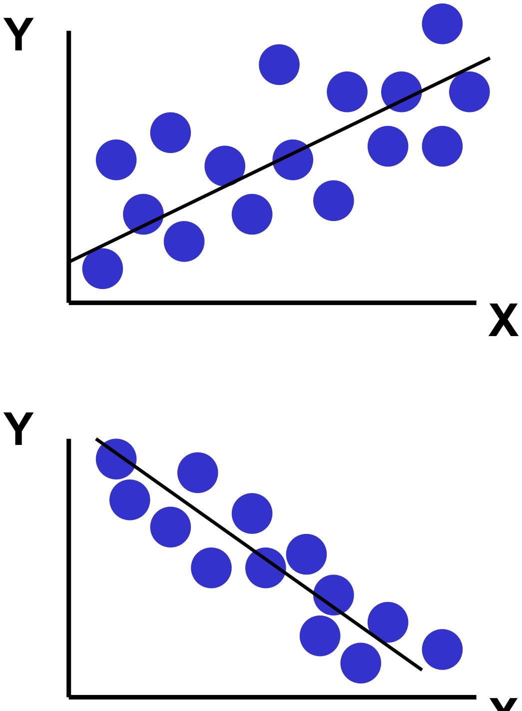
Scatter Plots of Data with Various Correlation Coefficients



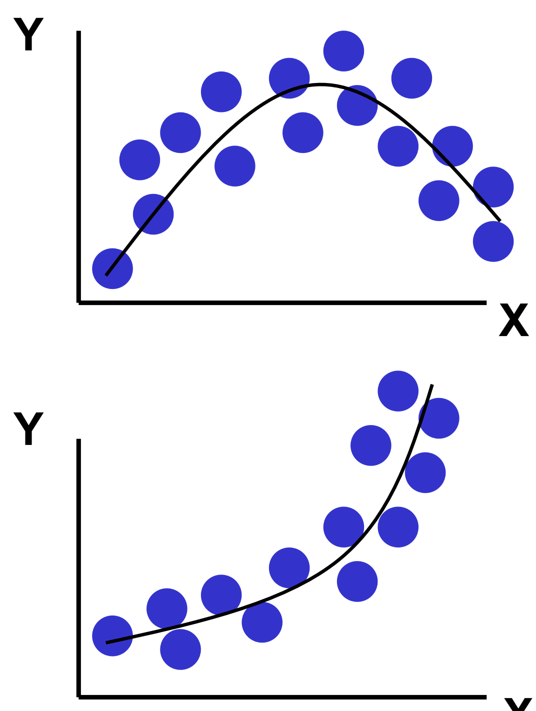
▪Slide from: Statistics for Managers Using Microsoft® Excel 4th Edition, 2004 Prentice-Hall

Linear Correlation

Linear relationships

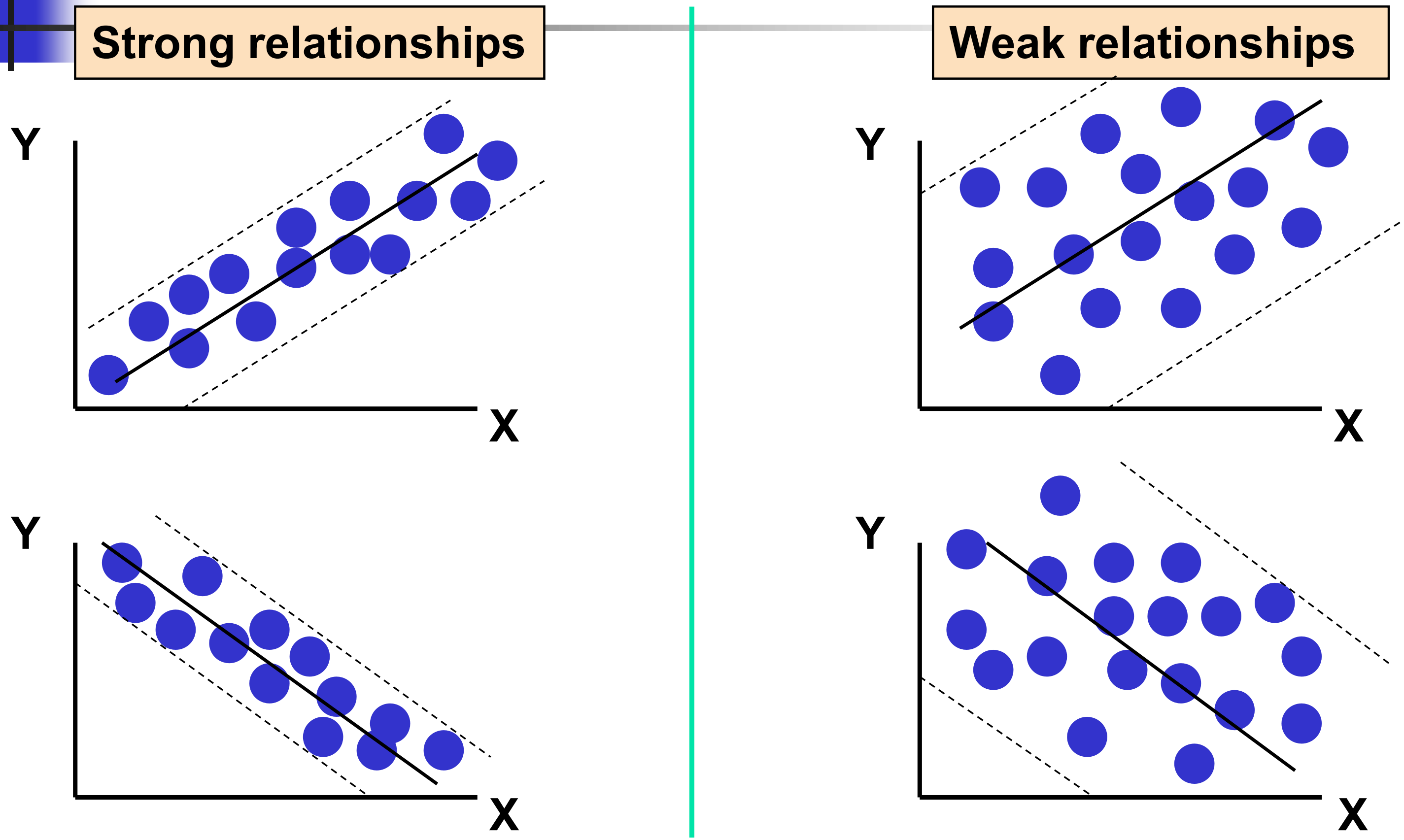


Curvilinear relationships



- Slide from: Statistics for Managers Using Microsoft® Excel 4th Edition, 2004 Prentice-Hall

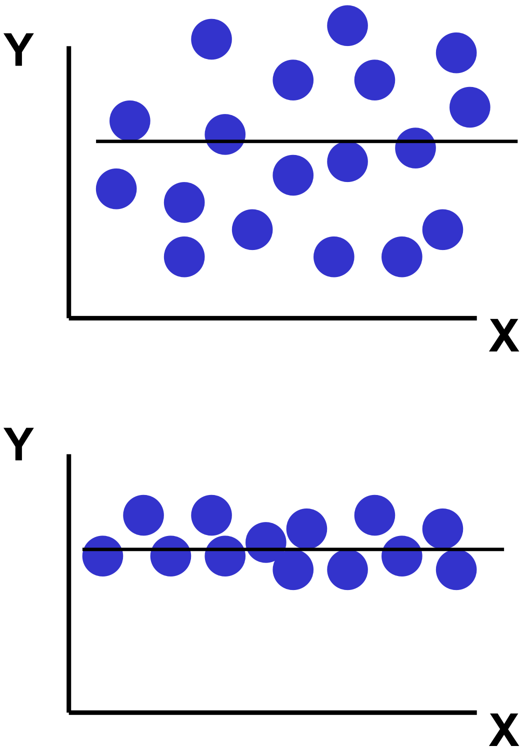
Linear Correlation



- Slide from: Statistics for Managers Using Microsoft® Excel 4th Edition, 2004 Prentice-Hall

Linear Correlation

No relationship



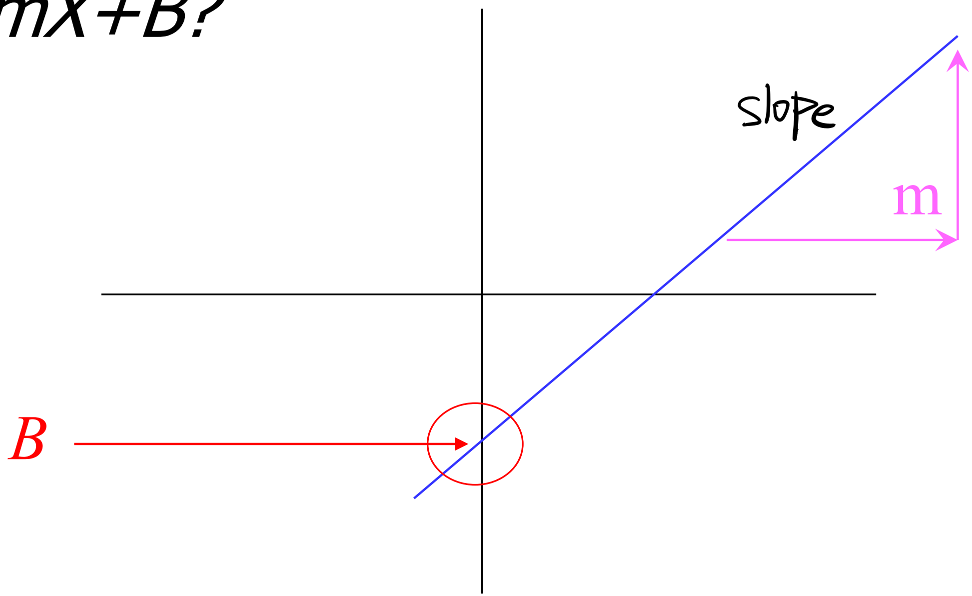
▪Slide from: Statistics for Managers Using Microsoft® Excel 4th Edition, 2004 Prentice-Hall

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.

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?)

$$\text{your height} \longrightarrow \underset{\text{Predict}}{\boxed{\quad\quad}} \longrightarrow \text{your son's height}$$

Regression equation…

Expected value of y at a given level of *x*=

$$\boxed{\underline{Y}_i = \beta_0 + \beta_1 \underline{X}_i + \varepsilon_i}$$

mine

father's

Questions: 1°. how to estimate β_0, β_1?

2°. how good is the estimation?

Predicted value for an individual…

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

Follows a normal distribution

error ↓ ⇒ estimation越好

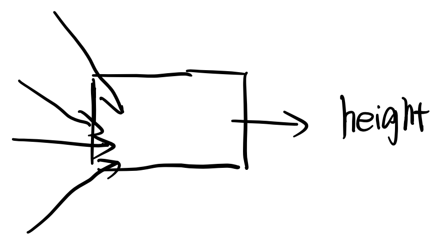
Assumptions (or the fine print)

- Linear regression assumes that…
  - first and strong 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

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon$$

$$= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

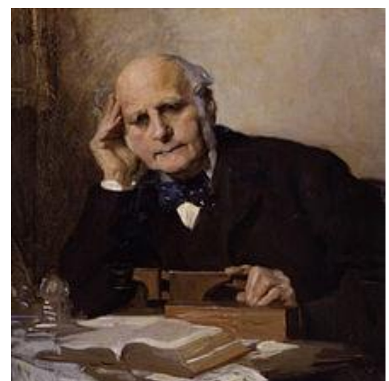
$$(X_2 = x^2, \ X_3 = x^3)$$



Linear Regression – a brief history

12/13

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