MONASH University

Information Technology

FIT5190 Introduction to IT Research Methods
Lecture 10
Quantitative Data Analysis
– Correlation and Regression

Slides prepared by
David Green, Frada Burstein, Jacques Steyn, Geoff Webb, Chung-Hsing Yeh

Learning objectives

- Understand
  - The purpose of correlation and regression analysis
  - Key issues in interpreting correlation and regression
- Be able to
  - Perform a simple regression analysis
  - Use regression to fit linear models to data
  - Use regression to fit nonlinear models to data

Overview

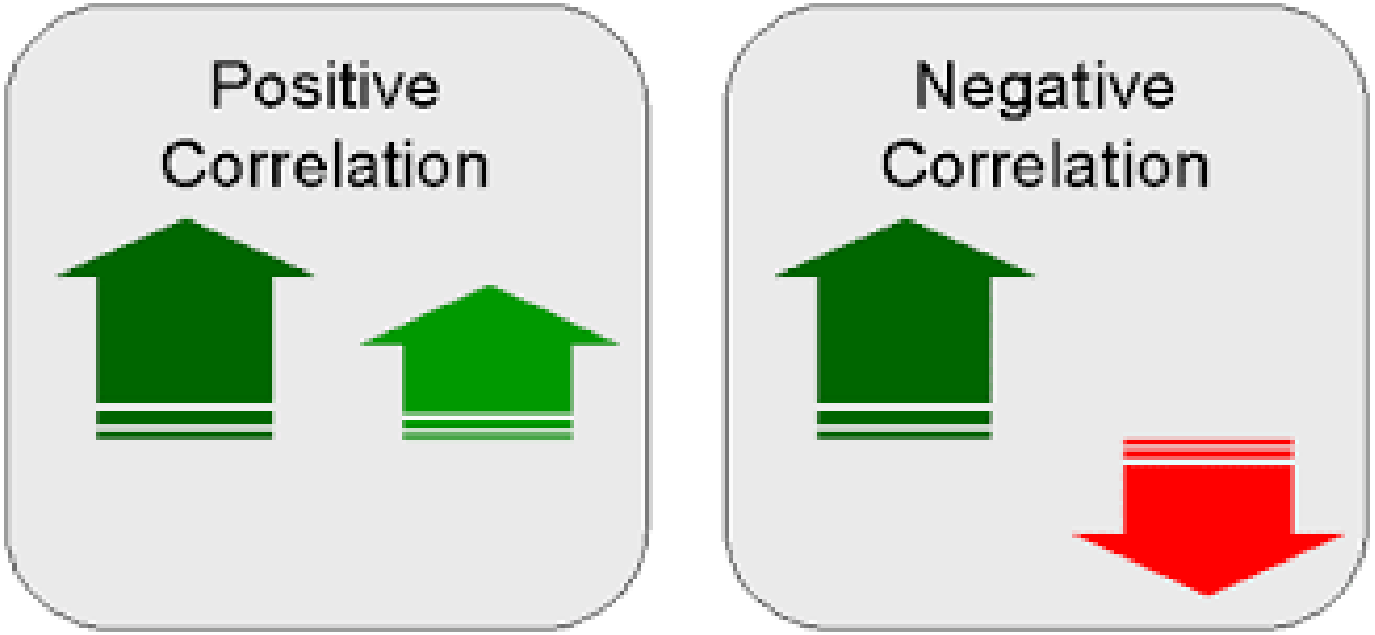
- This lecture concerns the use of quantitative data to detect and assess relationships between quantitative variables and to fit simple models to data.
- Tools used to do this include correlation and regression analysis.

Outline

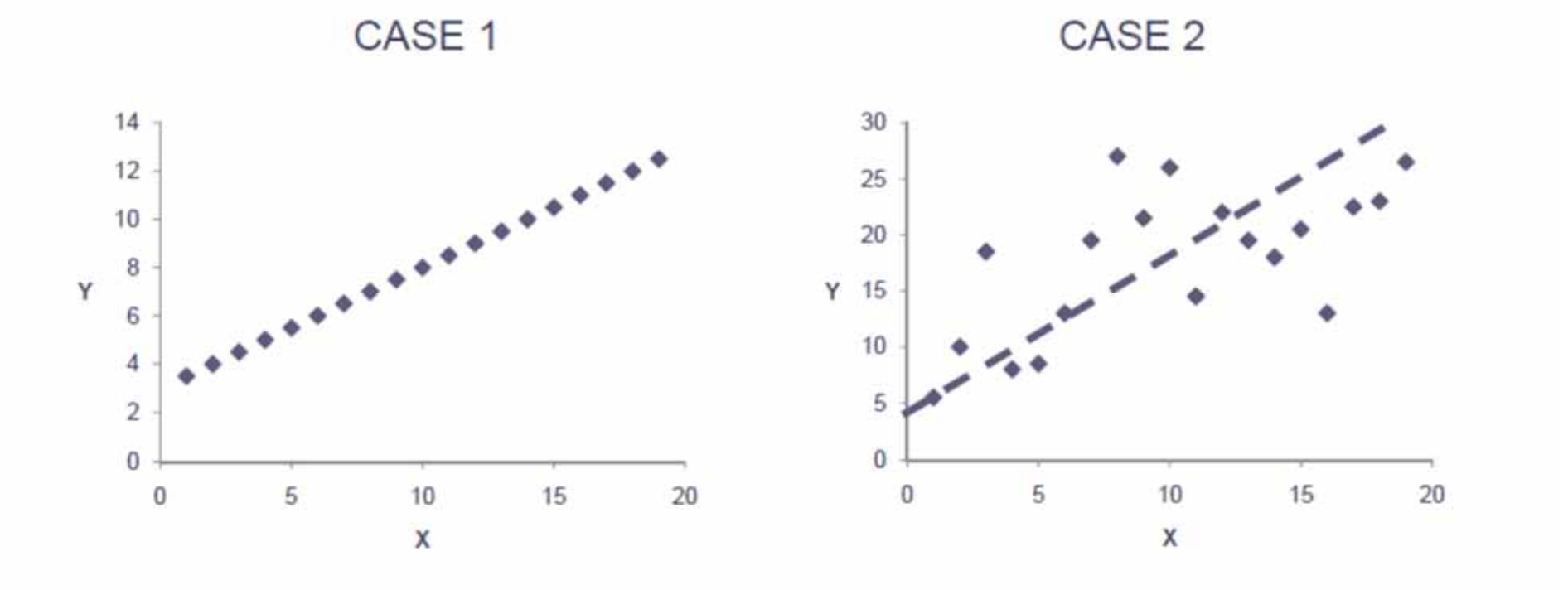
- **Correlation Analysis**
  - Parametric: Pearson
  - Non-parametric: Spearman

- **Regression Analysis**
  - Line and curve fitting
  - Multiple regression

Correlation



Are X and Y related?

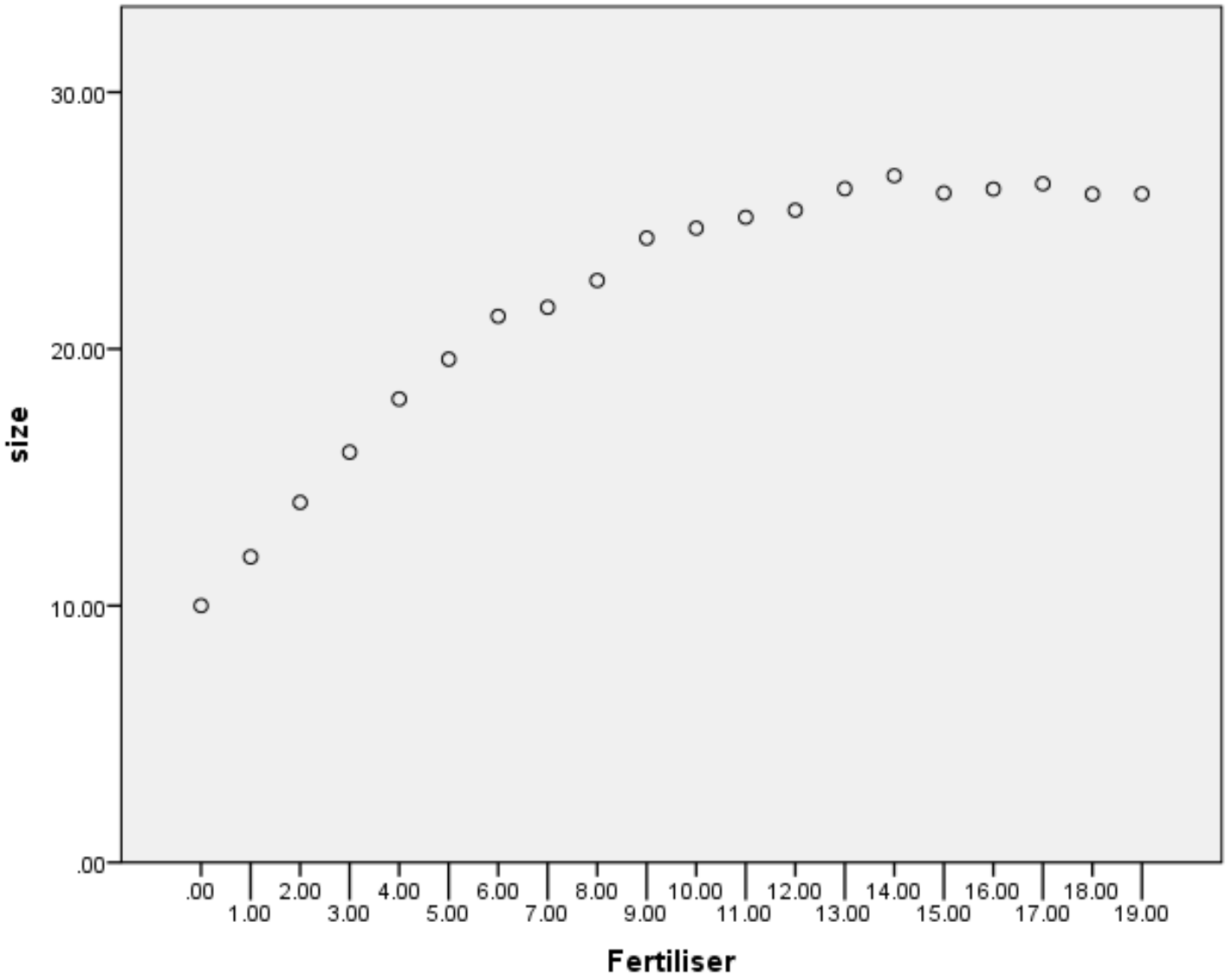


- In case 1, Y clearly increases as X does. But what about case 2?
- Correlation measures the strength of a relationship between variables.

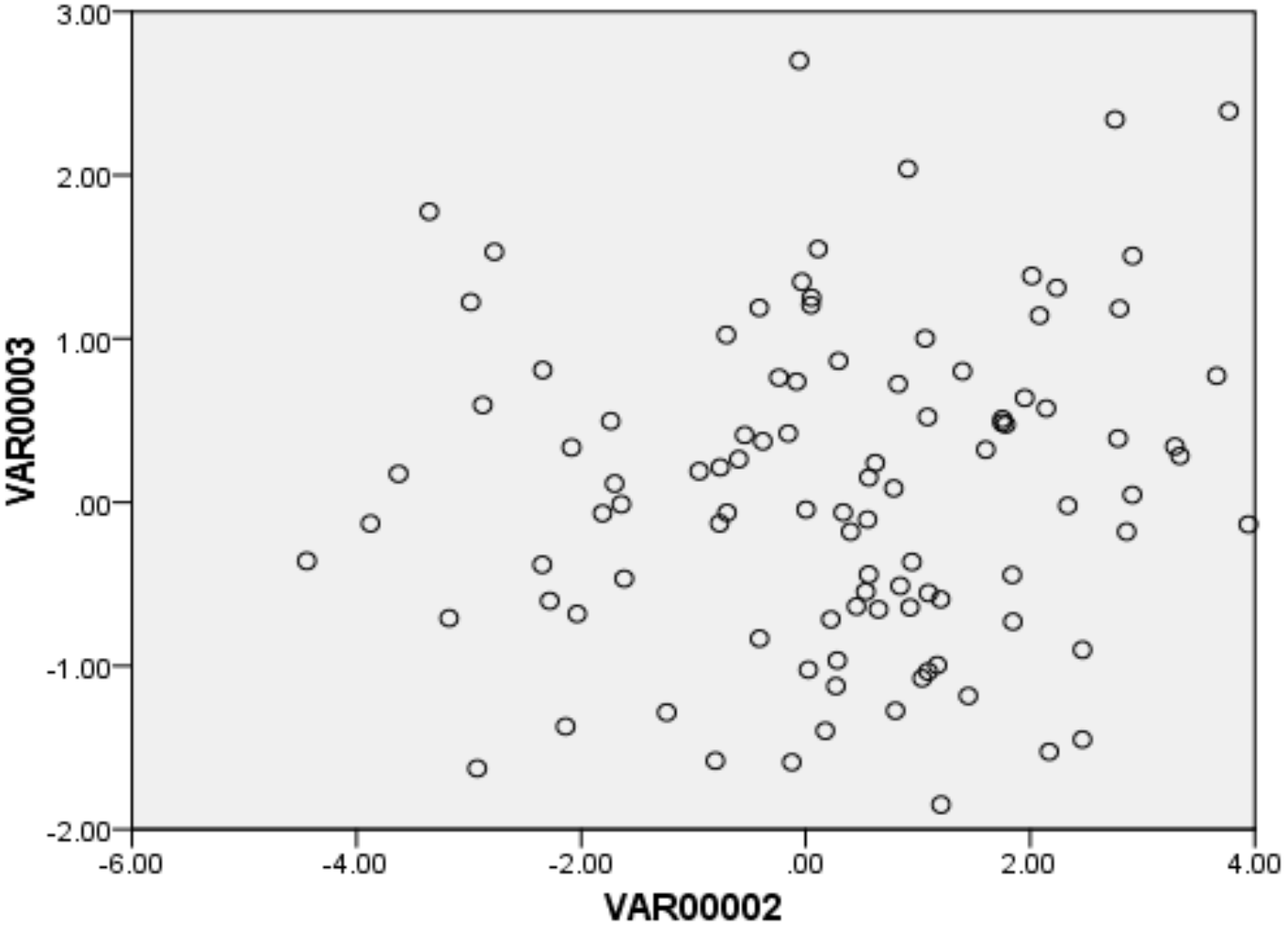
Case 1 $r = 1.0$
Case 2 $r = 0.55$

Correlation between numeric values

Fertiliser	Size
0	10
1	11.9
2	14.02
3	15.98
4	18.05
5	19.6
6	21.27
7	21.62
8	22.67
9	24.32
10	24.7
11	25.13
12	25.4
13	26.24
14	26.75
15	26.07
16	26.23
17	26.43
18	26.03
19	26.04



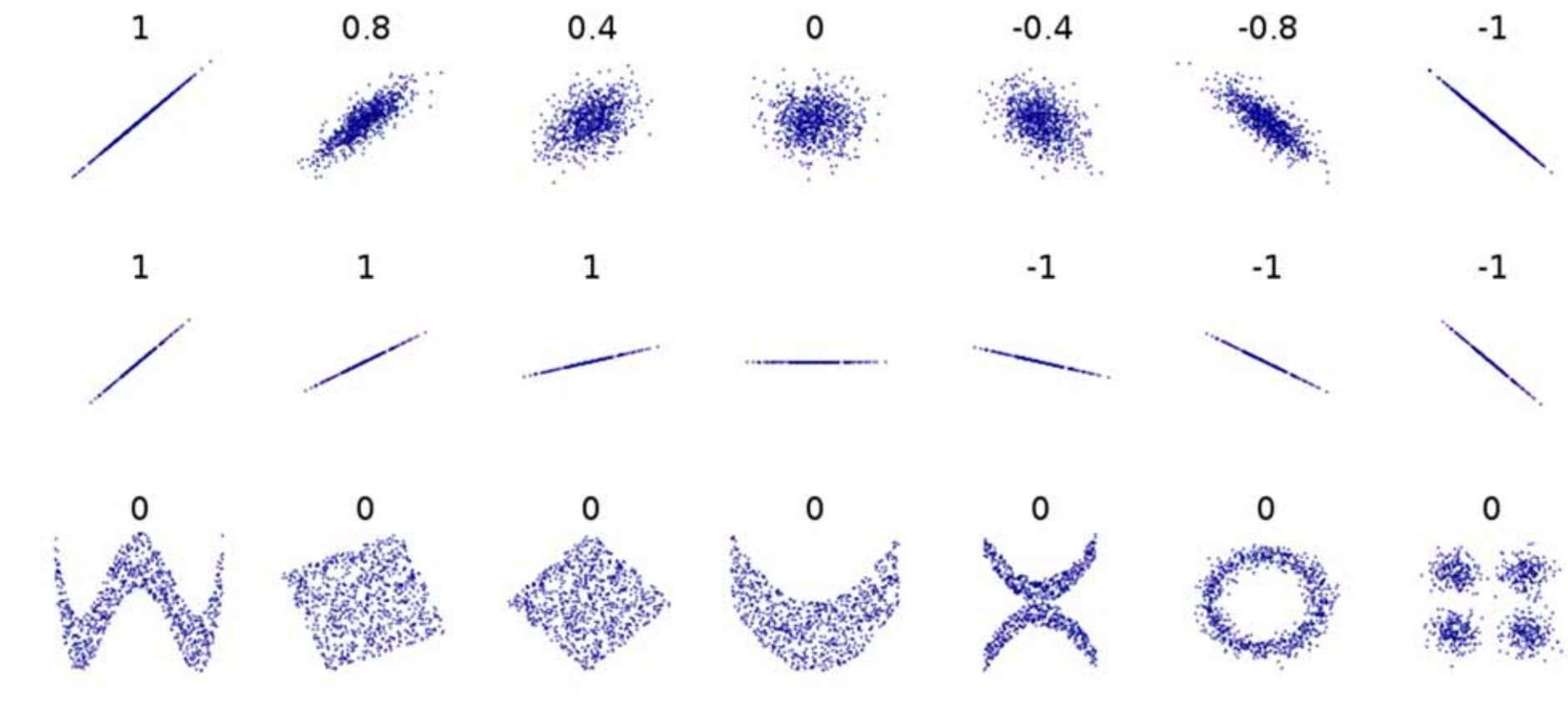
No correlation



Pearson's correlation coefficient

- Classical measure of correlation
- Parametric
  - Assumes variables are normally distributed.
- Susceptible to outliers
- Measures **linear** correlation
  - Extent to which one variable increases (or decreases) whenever the other increases (under assumption of normal distribution).
  - Positive values mean both tend to increase together.
    - 1.0 means perfect positive correlation.
  - Negative values mean one tends to decrease when the other increases.
    - -1 means perfect negative correlation.
  - 0 means no linear correlation.

Examples with Pearson correlation coefficients



Source: http://en.wikipedia.org/wiki/File:Correlation_examples2.svg

Pearson correlation coefficient (population)

- Assume:
  - X and Y are (normally distributed) random variables
  - The expected values are μ_X, μ_Y
  - The standard deviations are σ_X, σ_Y

- The Person's correlation coefficient is defined as:

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sqrt{E[(X-\mu_X)^2]}\sqrt{E[(Y-\mu_Y)^2]}}$$

Pearson correlation (sample)

- Assume:
  - Given a sample of n data points: $(x_1, y_1) \ldots (x_n, y_n)$
  - Let sample means: $\bar{x}, \bar{y}$
  - Let sample standard deviations: s_x, s_y

- The sample estimate for Pearson correlation coefficient:

$$r_{X,Y} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}}$$

Spearman's rank correlation coefficient

- Spearman's rho (ρ)
- Non-parametric test of correlation
  - Test extent to which one variable increases (or decreases) whenever the other increases.

- The Pearson correlation coefficient between the **ranks** is defined as

$$\rho_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$

  - x_i and y_i are ranks.

Example - Spearman's rank correlation

- Correlation between person's IQ and hours spent in front of TV per week

IQ	TV hours
106	7
86	0
100	27
101	50
99	28
103	29
97	20
113	12
112	6
110	17

convert to ranks

IQ	TV hours
1	1
2	6
3	8
4	7
5	10
6	9
7	3
8	5
9	2
10	4

compute Pearson corr

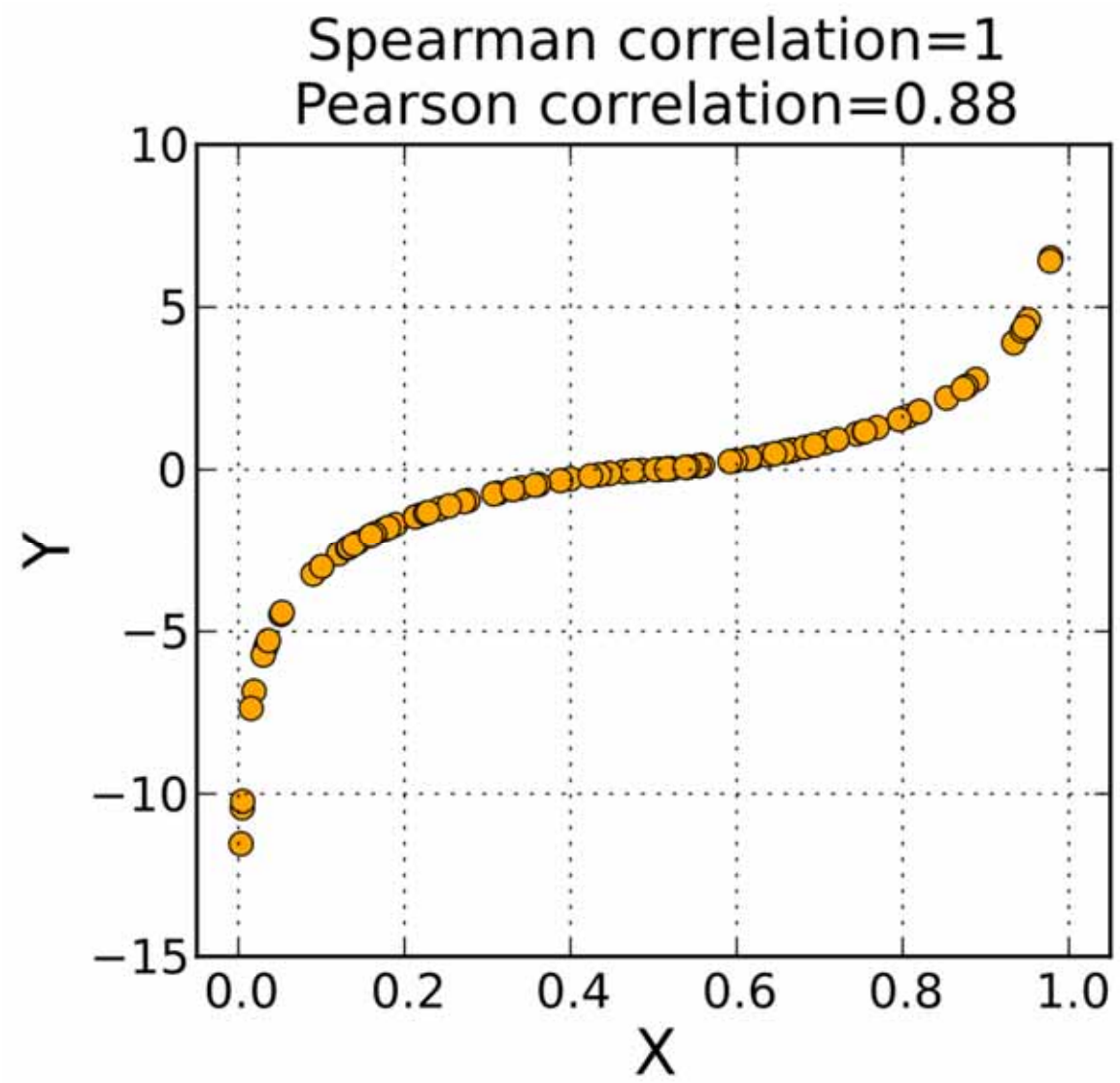
$\rho_{X,Y}$

$\rho = -0.1758$

- This low ρ value shows that the correlation between IQ and hours spent watching TV is very low, although the negative value suggests that the longer the time spent watching television the lower the IQ.

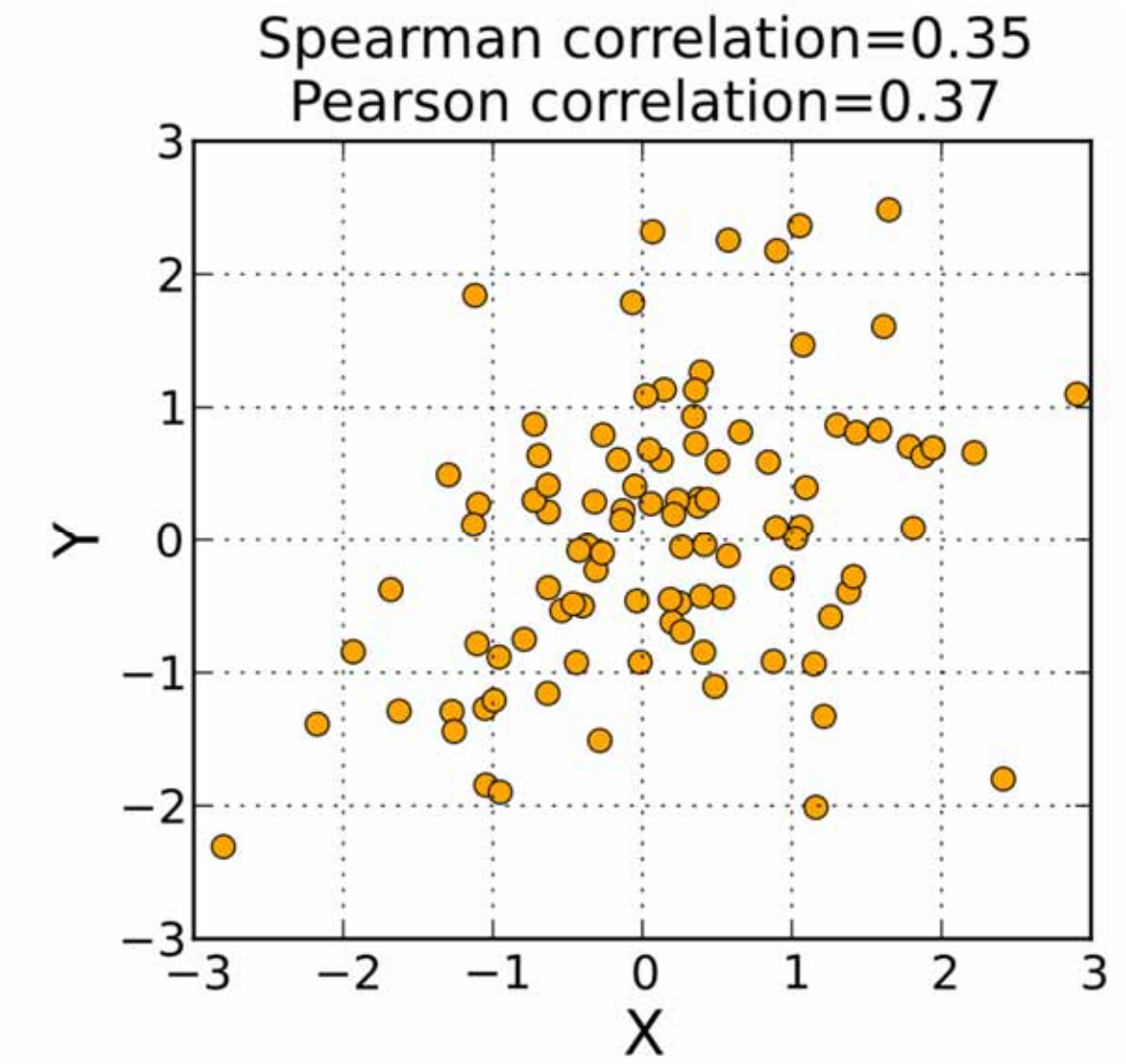
Source: http://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient

Spearman does not assume the relationship is linear



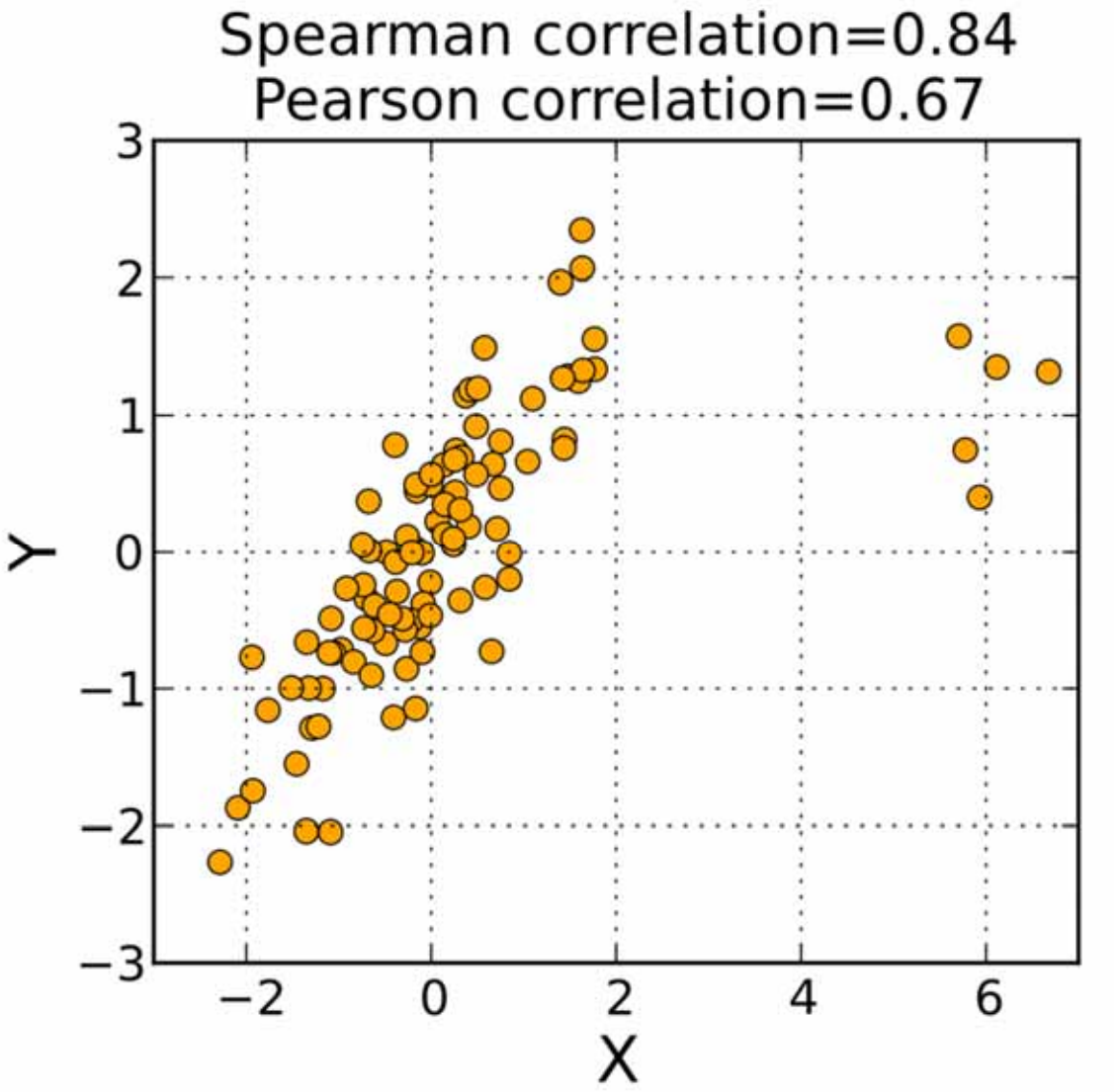
Source: http://en.wikipedia.org/wiki/File:Spearman_fig1.svg

When data normally distributed without outliers Spearman & Pearson are similar



Source: http://en.wikipedia.org/wiki/File:Spearman_fig2.svg

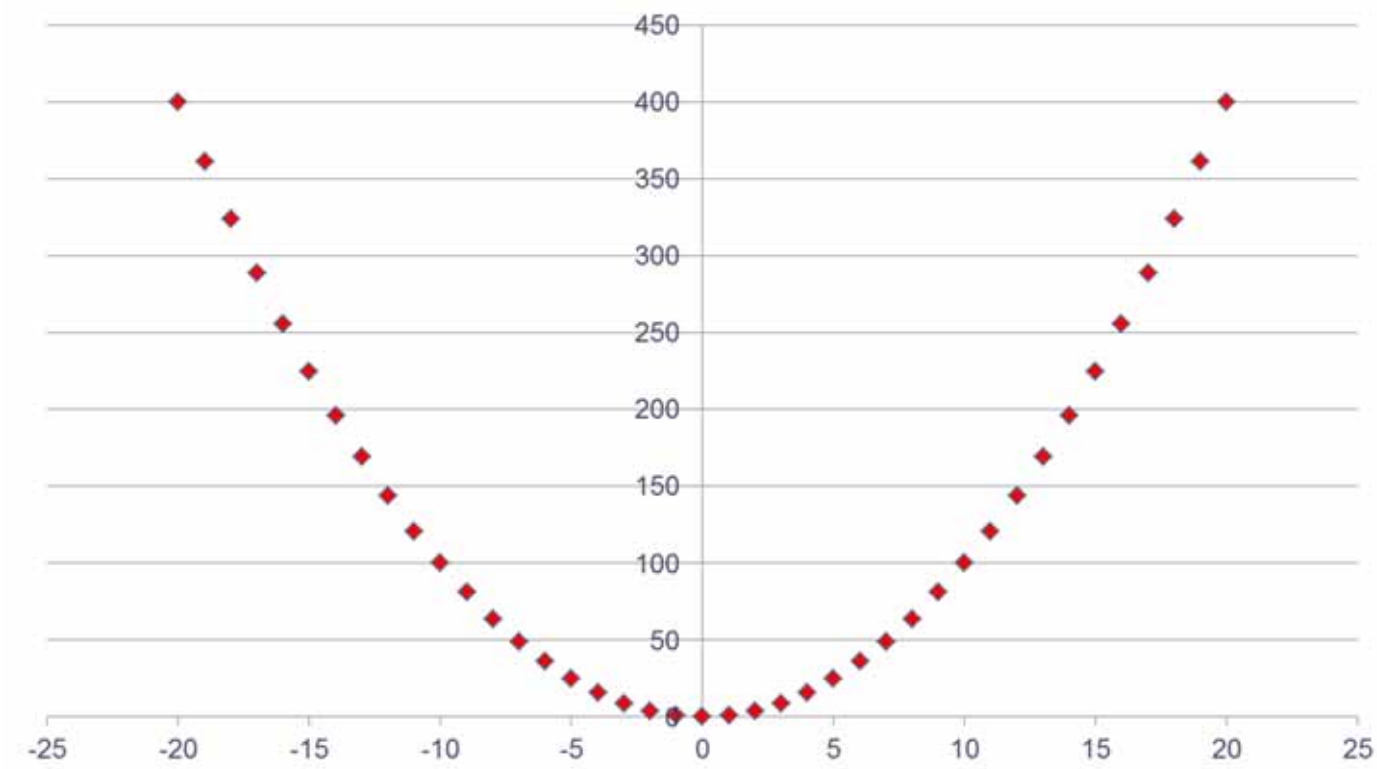
Spearman less sensitive to outliers



Source: http://en.wikipedia.org/wiki/File:Spearman_fig3.svg

Always do a scatter plot first

- It is pointless doing either Pearson or Spearman if the relationship is not monotonic (consistently either increasing or decreasing).
- Example: both Pearson & Spearman are 0



Correlation summary

- Pearson and Spearman correlation coefficients provide measures of correlation between numeric variables.
  - Measure degree to which one variable consistently increases or consistently decreases as the other increases.
  - Spearman makes fewer assumptions than Pearson.
  - Pearson is more popular.
  - They only measure one type of correlation, so check your data with a scatter plot before doing them.

Regression analysis

Linear regression

- Regression is the practice of describing the relationship between 2 or more quantitative variables.
- Thus, if we know the value of one variable, we can estimate the value of the related variable of interest.
- Origin:
  - The 19th century scientist Francis Galton collected data on the heights of fathers and their sons.
  - He found that tall fathers had slightly shorter sons and that short fathers had slightly taller sons.
  - Thus in each case there was a *regression to the mean*.
- The term *regression* has, over time, come to mean the process of fitting the model, rather than the original observations.

The purpose of regression

- To find the underlying relationship between variables.
- We may use this to develop a model (or equation) of our data.
- We may use the model for prediction or extrapolation.
- Regression based methods are particularly suited to long-term forecasting.
- For example:
  - The time it takes to serve a customer given the number of items they have purchased.
  - The industry standard for the number of staff employed given the annual turnover of the company.
  - The relationship between advertising and sales for a company.

Regression

- If we know how variables are related to one another then we should be able to use one to predict the other.
  - Prediction may not be perfect because there may be other relevant factors about which we do not have information.
  - e.g., a life insurance company can predict life expectancy from current age, medical history, occupation, etc., but not perfectly.

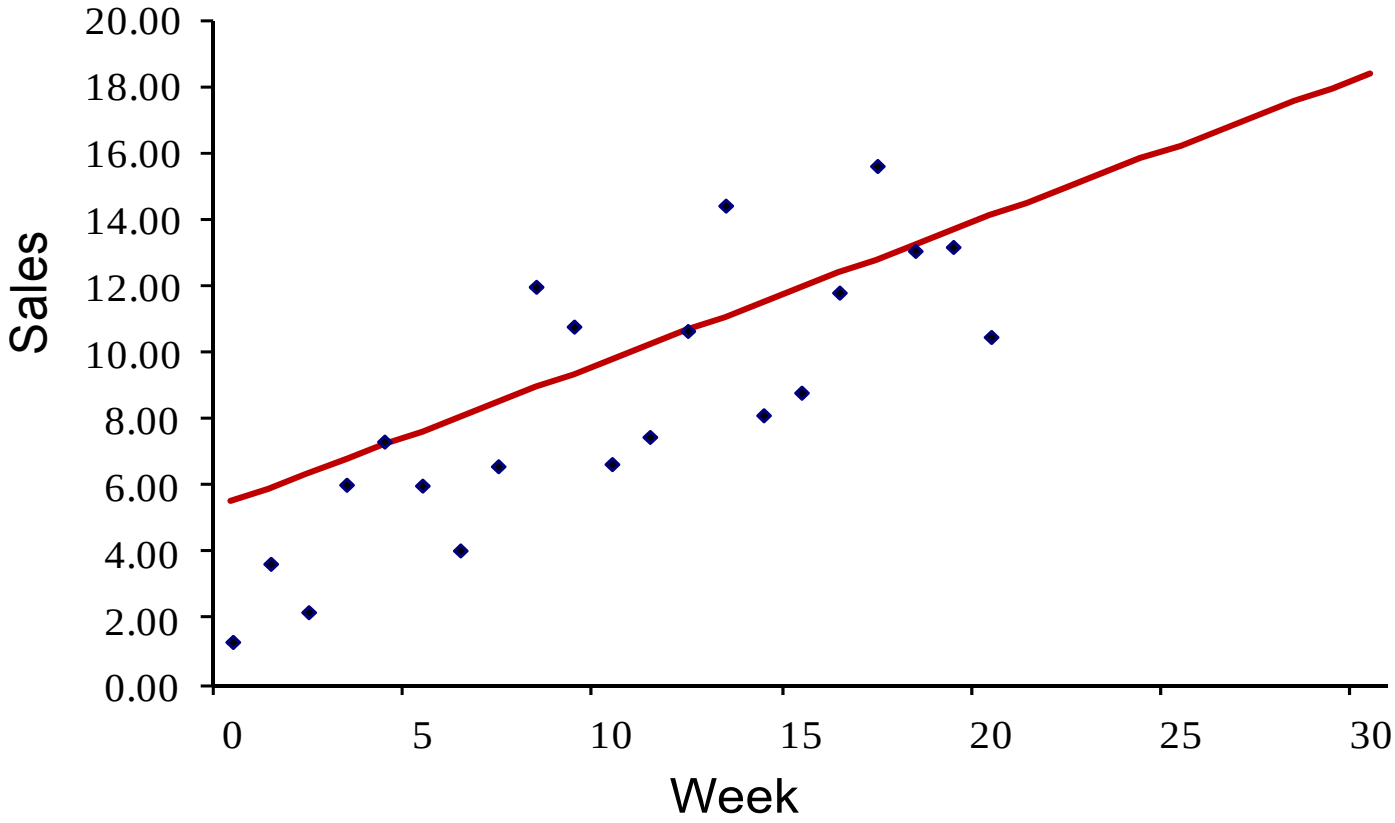
Regression

- Regression predicts a numeric value (the dependent variable) from a collection of other values (the independent variables).

- The most common form is linear regression.
  - Assume that the relationship can be described by a line with added noise.
  - Simple linear regression assumes a straight line.

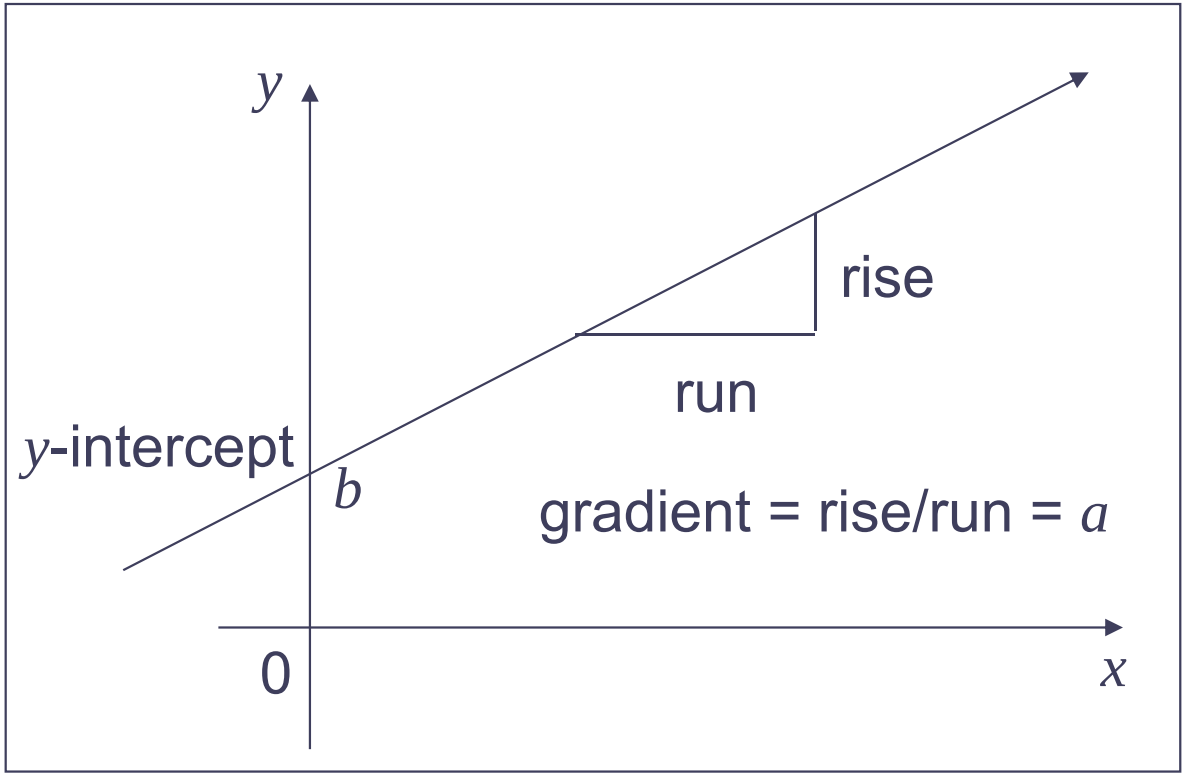
Regression based prediction

- The data should be linear.
- A straight line with the general equation $y = ax + b$ is fitted to the data.
- The equation of this line then forms the basis for prediction.



The equation of a straight line

- We can use the basic equation of a straight line as the model for our regression equation.
- A line with gradient a and y-intercept b has the equation $y = ax + b$.



Least squares regression

- The basic idea is to find the straight line which minimises the squared differences between actual points and those predicted by the model.
- To express the regression of y on x as $y = ax + b$, we calculate:

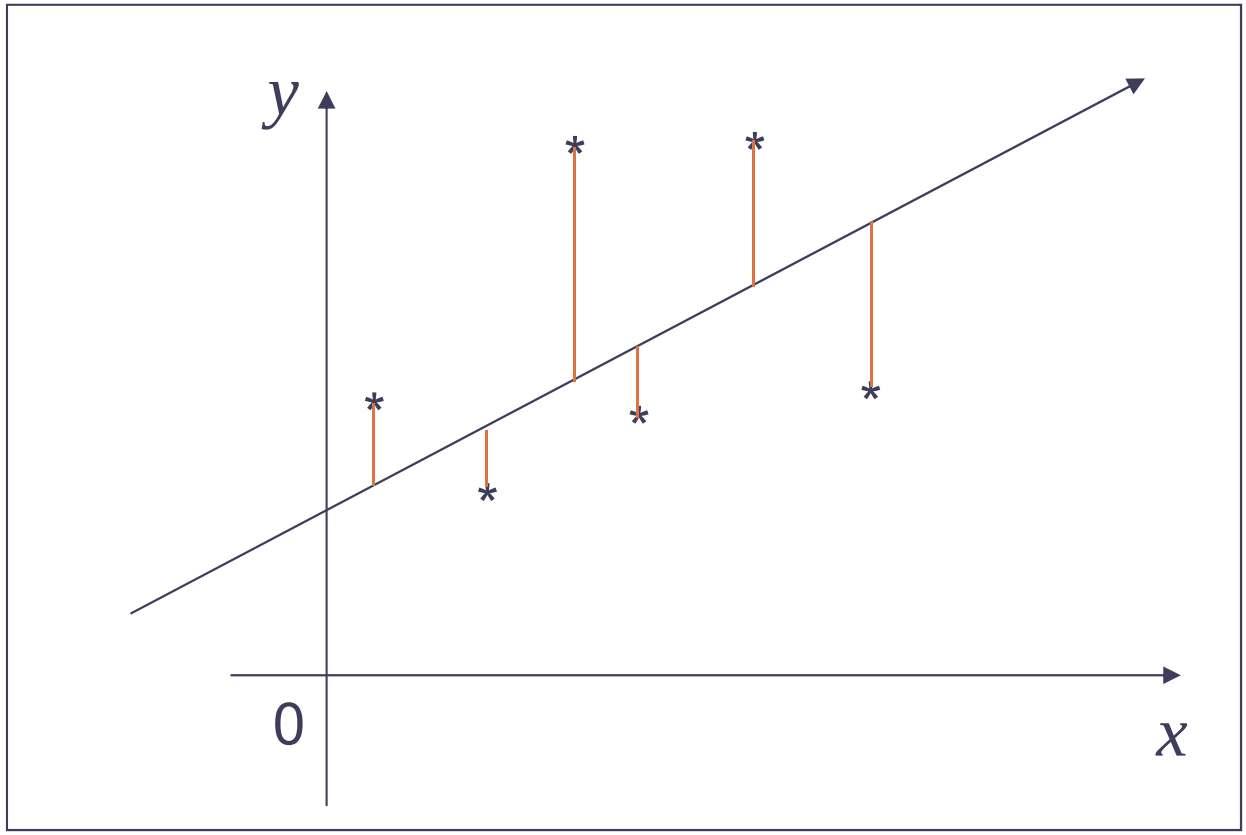
$$a = \frac{s_{xy}}{s_x^2} = \frac{\Sigma xy - \frac{\Sigma x \Sigma y}{n}}{\Sigma x^2 - \frac{(\Sigma x)^2}{n}} \text{ and } b = \bar{y} - a\bar{x}$$

or

$$a = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

The basic idea of least squares regression

- We want to minimise the sum of the squared errors, or differences between the fitted model (line) and the data.



Regression in EXCEL

- Regression is one of the analysis functions in EXCEL.
- However, you can also calculate the formulas manually with the following built-in formulas: $(y = ax + b)$

  a = SLOPE(y values, x values)

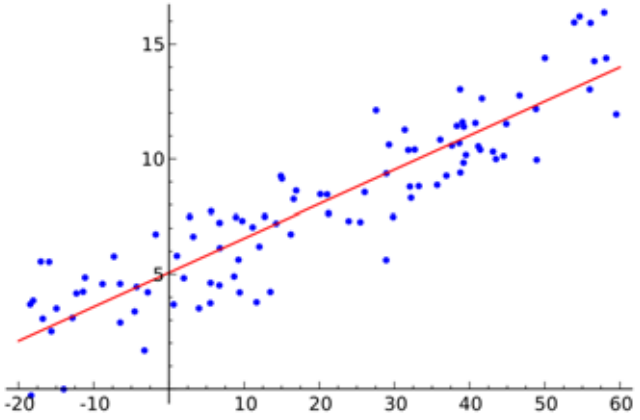
  b = INTERCEPT(y values, x values)

- Also, the correlation coefficient $r^2 = C = \frac{S_{xy}}{S_x S_y}$

  where S_x and S_y are the standard deviation of x and y, S_{xy} is the covariance of x and y.

  r^2 = CORREL(y values, x values)^2

Multiple linear regression



- Assume:
  - The input variable is a vector $(x_1, \ldots, x_d)$
  - the output variable is a real number y
- The relationship between the input and output:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_d x_d + \varepsilon$$

  - β_i's are the coefficients.
  - Most commonly found by minimizing the mean squared error.
  - ε captures the missing information (noise).

Mean squared error

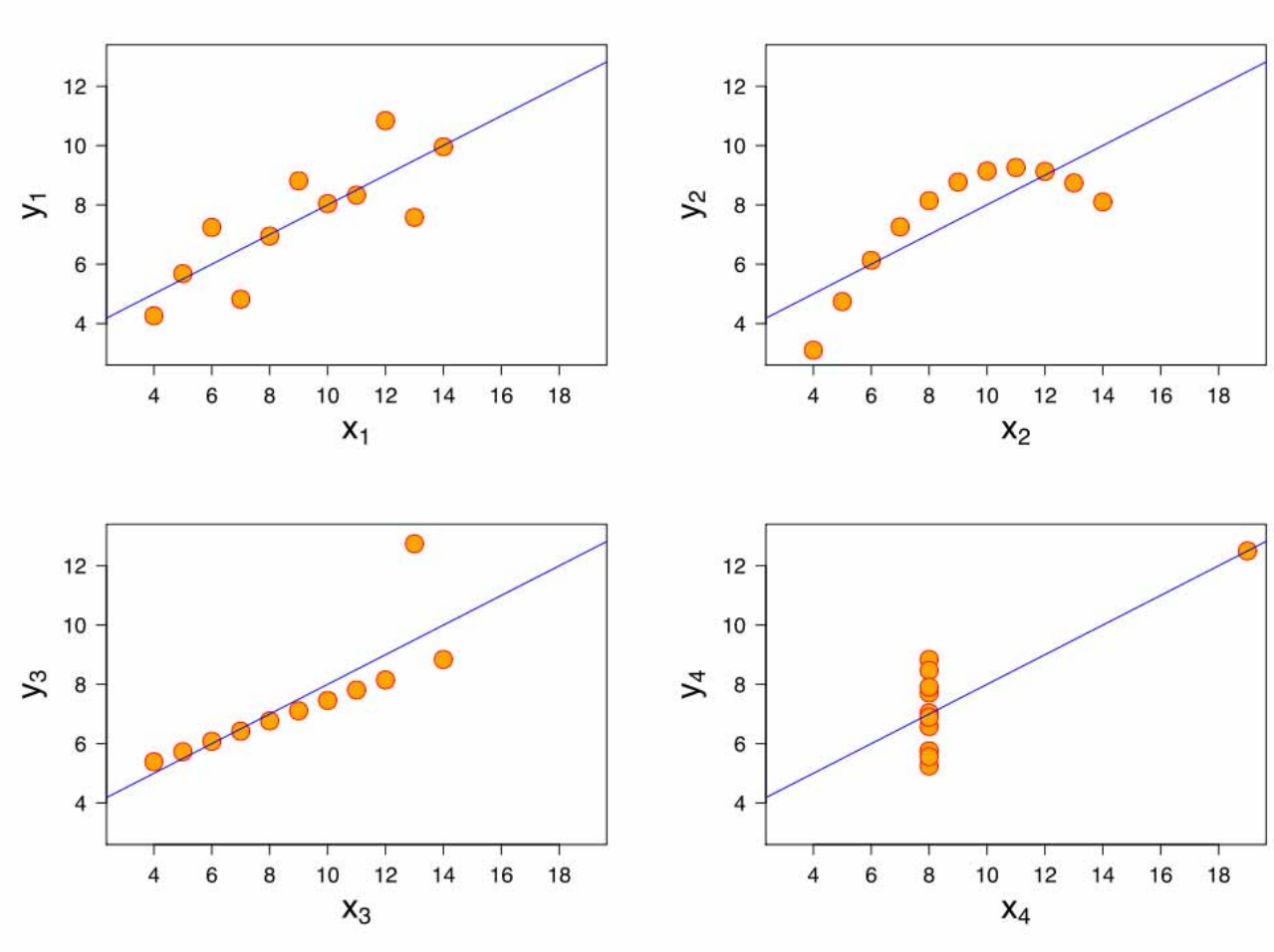
- Often used as a measure of how well a simple line describes the relationship between the input the dependent and independent variables.
  - For each data point, the error is the difference between the predicted value y' and the actual value y :
  - Mean squared error (MSE)
    - Take the average of the squared errors:

$$\text{MSE} = \sum_{i=1}^{n} (y' - y)^2$$

- Avoid large error more than small errors.
  - Hence outliers have large effect.

Simple linear regression assumes that the relationship is a straight line

- 4 distributions with the same regression line
- Note fit to a curve (top-right)
- Effect of outliers (bottom row)



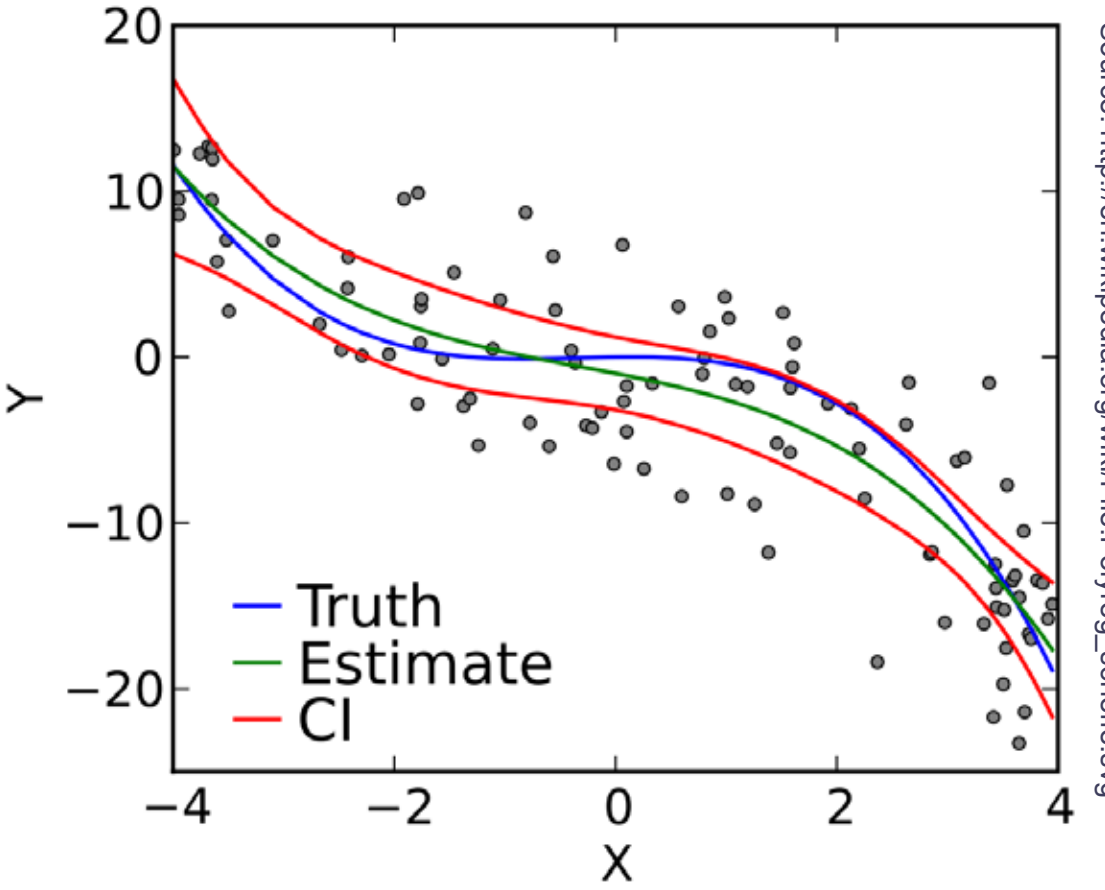
Source: http://en.wikipedia.org/wiki/File:Anscombe%27s_quartet_3.svg

Polynomial regression

- Assume:
  - Input a real value: x (no vector)
  - Output a real value: y
  - The relationship between input and output is a polynomial of degree 2

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

- Can be implemented by creating new data set:
  - Input as vector: (x, x^2)
  - Output: y



Source: http://en.wikipedia.org/wiki/File:Polyreg_scheffe.svg

Can extend to any order of polynomial

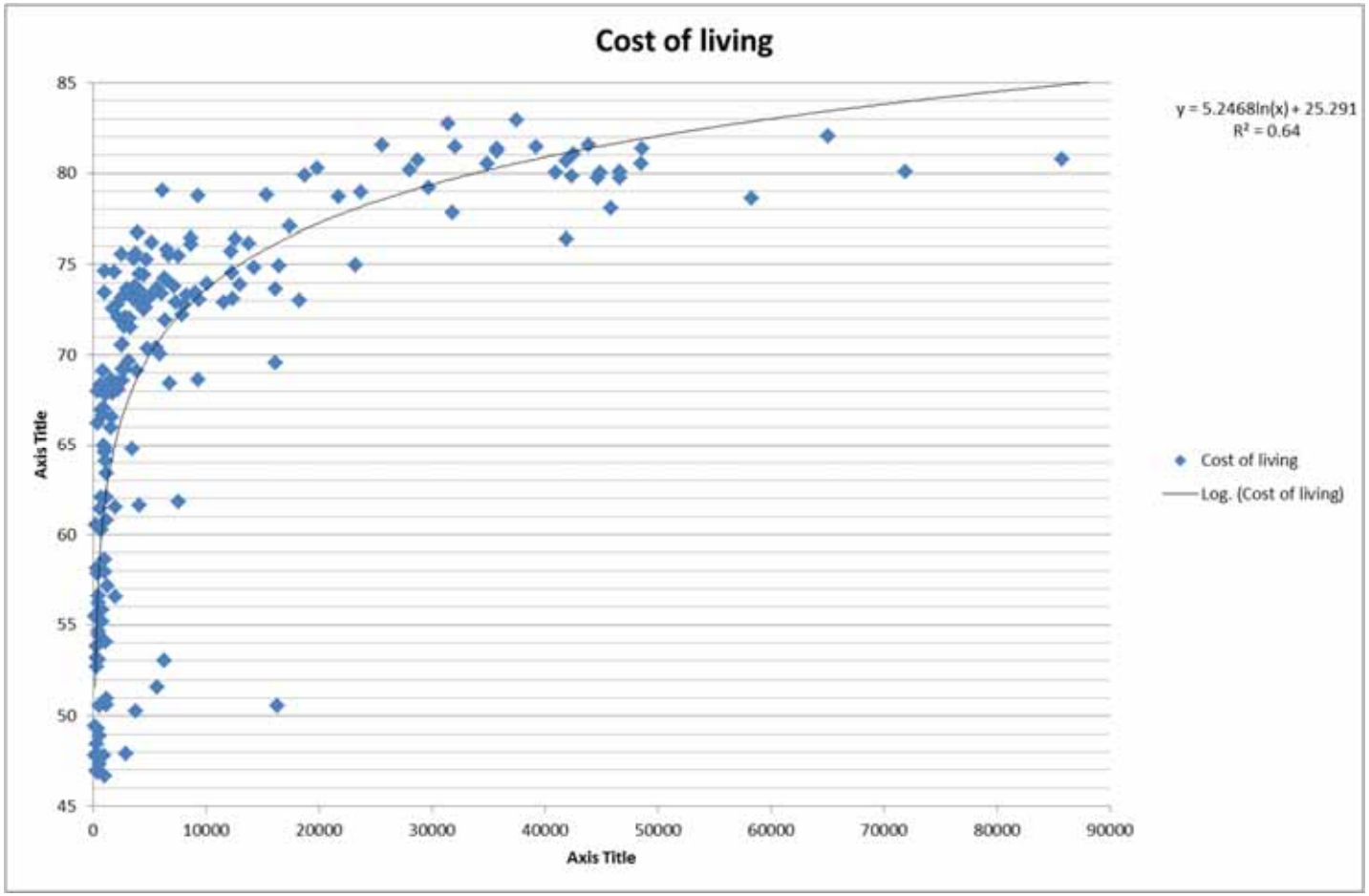
- Can be easily extended to higher order polynomials

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_d x^d + \varepsilon$$

  - The more variables there are, the more data is required to get a good line.
  - Higher order polynomials with many variables require very large number of examples to get a good fit.

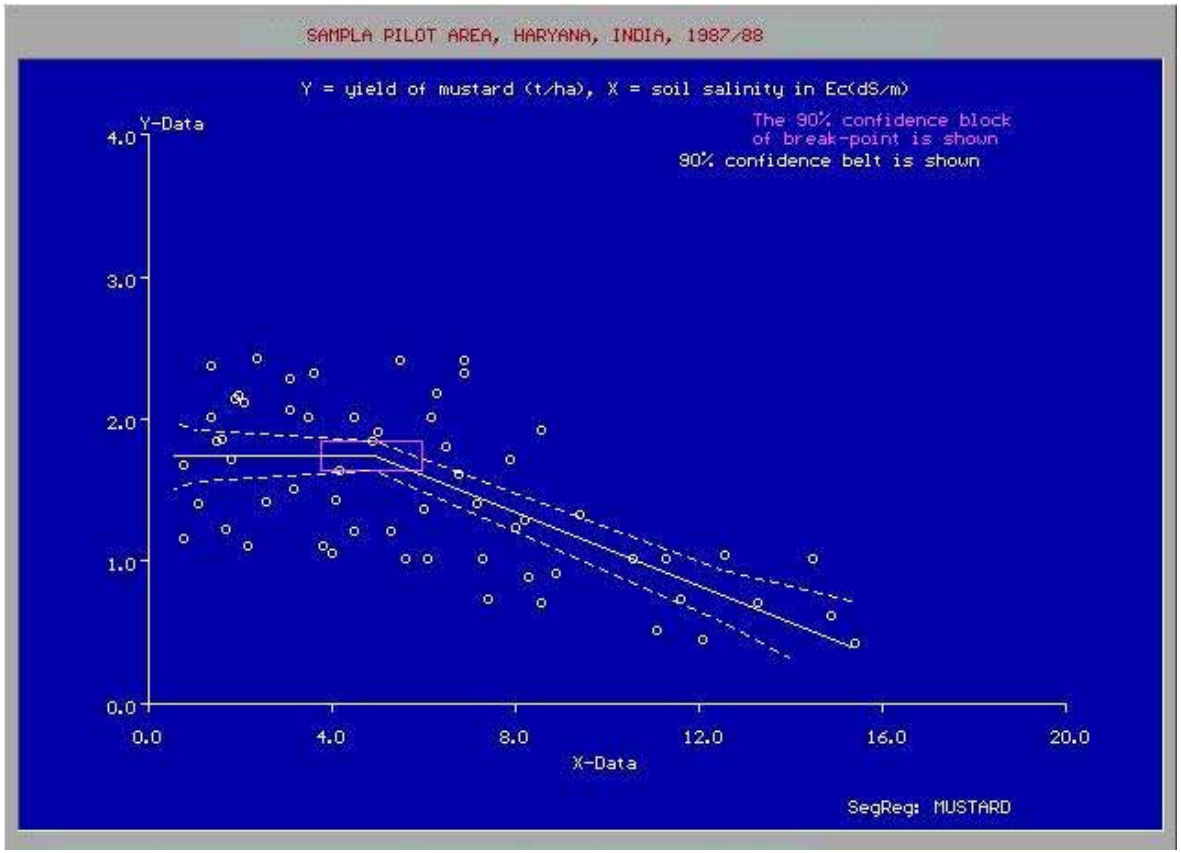
Can fit any other form of equation

$$y = \beta_0 + \beta_1 \ln x_1 + \beta_2 \ln x_2 + \ldots + \beta_d \ln x_d + \varepsilon$$



Other approaches to regression

- There are many other approaches to regression
  - Regression trees
  - Piecewise (segmented) regression



Source: http://en.wikipedia.org/wiki/File:Segmented_linear_regression_graph_showing_yield_of_mustard_plants_vs_soil_salinity_in_Haryana,_India,_1987%E2%80%931988.jpg

Regression summary

- Regression predicts one numeric variable from other variables.
  - Data must include the dependent variable in order to discover the regression function, but can then be predicted for future data.
- Simple linear regression finds the straight line that minimises error.
  - Polynomial regression fits higher order lines
  - Alternative approaches fit other kinds of functions
- A regression model can help understand the nature of a correlation, whereas correlation analysis only assesses its strength.

Classification learning

- A large body of techniques exist for regression-like prediction of categorical values
  - Machine learning
  - Classification learning
  - Data mining

- If you need to do this, Monash has many experts!
  - Have a look on the Machine Learning Flagship page:

    http://www.infotech.monash.edu.au/research/about/flagships/machine-learning/