Outlines

- Correlation
  - Pearson's correlation
  - Kendall rank correlation
  - Spearman rank correlation
- Regression
  - Types of regression
  - Fitting a function – Criterion for best fit
  - Least squares
- Simple regression
- Multiple regression
- Model assessment and validation

CORRELATION

Preliminaries

- n observations for x and y variables (x_i, y_i)
- Sample means $\bar{x}$ and $\bar{y}$

$$\bar{x} = \frac{\sum x_i}{n} \qquad \bar{y} = \frac{\sum y_i}{n}$$

- Sample variances S_{xx} and S_{yy}

$$S_{xx} = \frac{1}{n}\sum(x_i - \bar{x})^2 \qquad S_{yy} = \frac{1}{n}\sum(y_i - \bar{y})^2$$

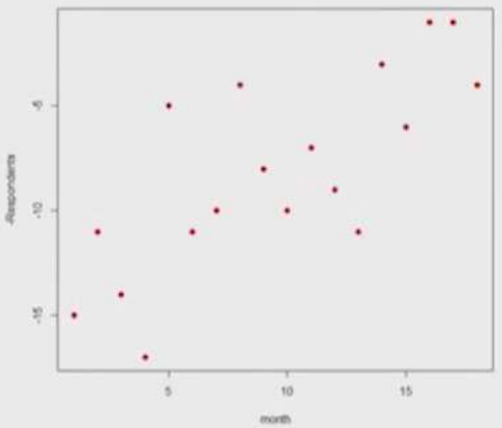
- Sample covariance S_{xy}

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

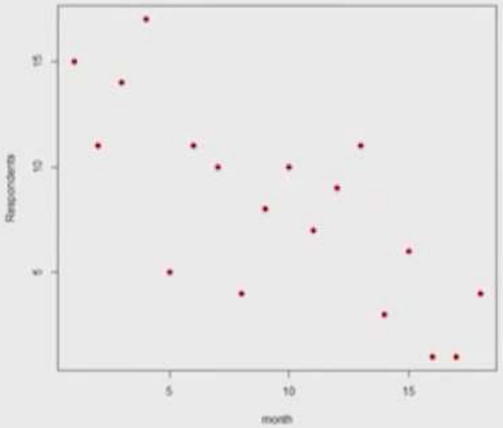
Correlation

- Correlation: the strength of association between two variables
- Correlation does not imply causation
- Visual representation of correlation: Scatter grams

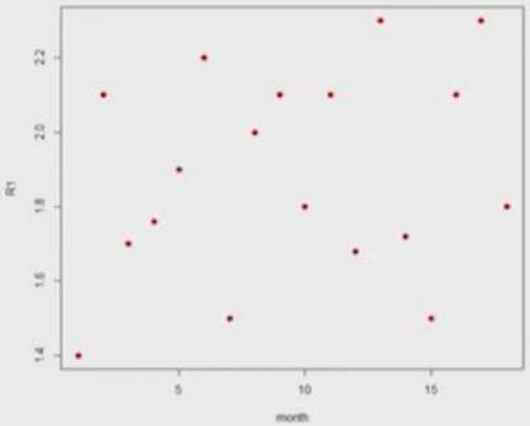
Positive trend



Negative trend



Little or no correlation



Quantitative Metric?

Pearson's Correlation

- n observations for x and y variables (x_i, y_i)
- Pearson's product-moment correlation coefficient (r_{xy})

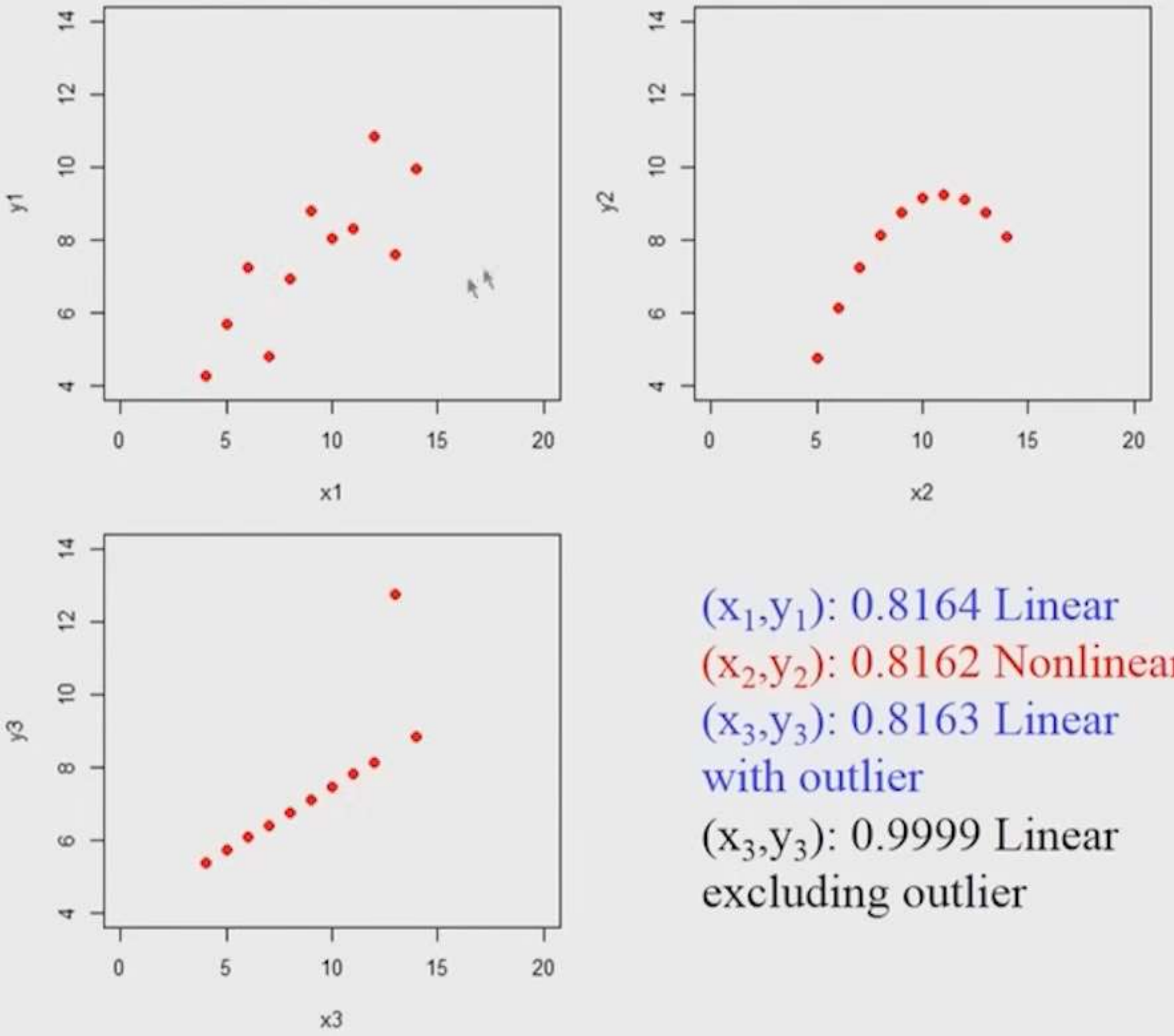
$$r_{xy} = \frac{\Sigma x_i y_i - n\bar{x}\bar{y}}{\sqrt{(\Sigma x_i^2 - n\bar{x}^2)}\sqrt{(\Sigma y_i^2 - n\bar{y}^2)}} = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$$

- r_{xy} takes a value between -1 (negative correlation) and 1 (positive correlation)
- $r_{xy} = 0$ means no correlation

Pearson's Correlation (Cont.)

- A measure for the degree of linear dependence between x and y
- Cannot be applied to ordinal variables
- Sample size: Moderate (20-30) for good estimate
- Robustness: Outliers can lead to misleading values

Pearson's Correlation: Anscombe's data



(x_1,y_1): 0.8164 Linear

(x_2,y_2): 0.8162 Nonlinear

(x_3,y_3): 0.8163 Linear with outlier

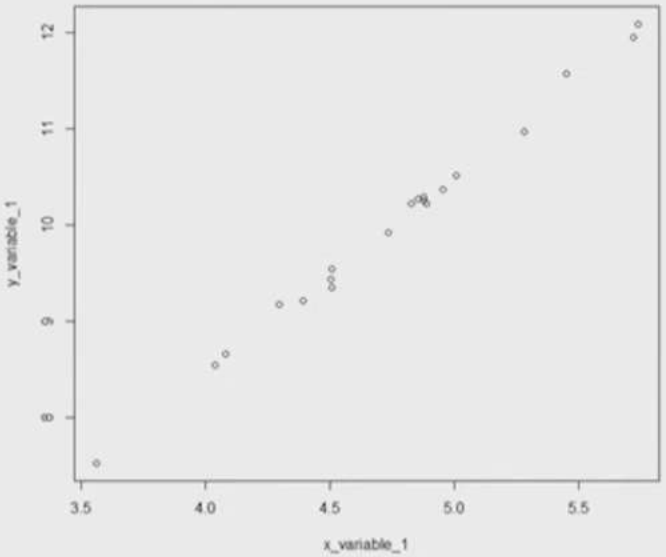
(x_3,y_3): 0.9999 Linear excluding outlier

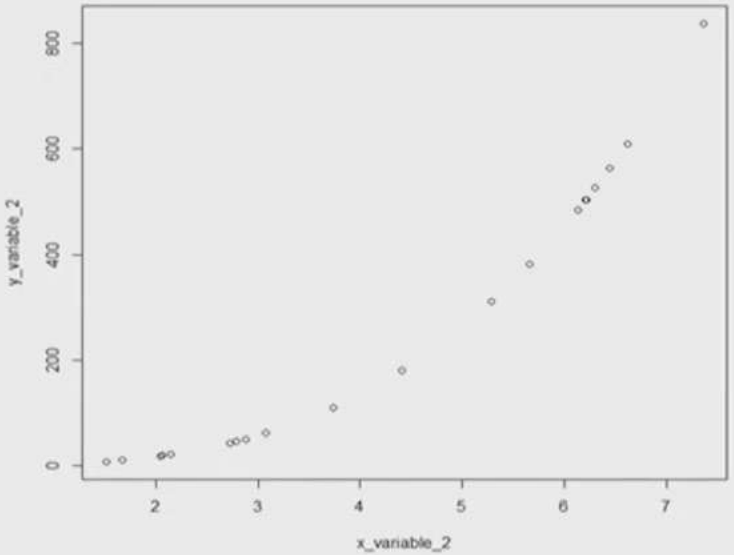
Pearson's Correlation (Cont.)

- Example: Nonlinear
  - x = 125 equally spaced values between $[0, 2\pi]$
  - $y = \cos(x)$
  - $r_{xy} = -0.0536$
- Example: Nonlinear
  - $x = 0{:}0.5{:}20$; $y = x^2$; $r_{xy} = 0.967$
  - $x = -10{:}0.5{:}10$; $y = x^2$; $r_{xy} = 0.0$

Spearman Rank Correlation

- Degree of association between two variables
- Linear or nonlinear association
- x increases, y increases or decreases monotonically





Spearman Rank Correlation

- Spearman rank correlation computation for n observations:

$$r_s = 1 - \frac{6\Sigma d_i^2}{n(n^2-1)}$$

d_i is the difference in the ranks given to the two variables values for each item of the data

- Example:

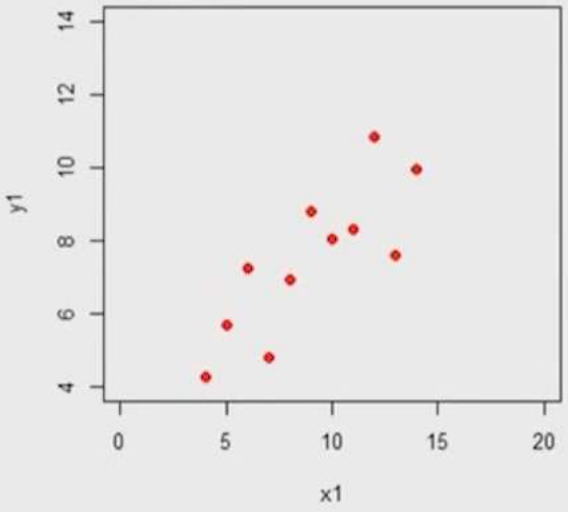
Number	1	2	3	4	5	6	7	8	9	10
X_1	7	6	4	5	8	7	10	3	9	2
Y_1	5	4	5	6	10	7	9	2	8	1
Rank X1	6.5	5	3	4	8	6.5	10	2	9	1
Rank Y1	4.5	3	4.5	6	10	7	9	2	8	1
d^2	4	4	2.25	4	4	0.25	1	0	1	0

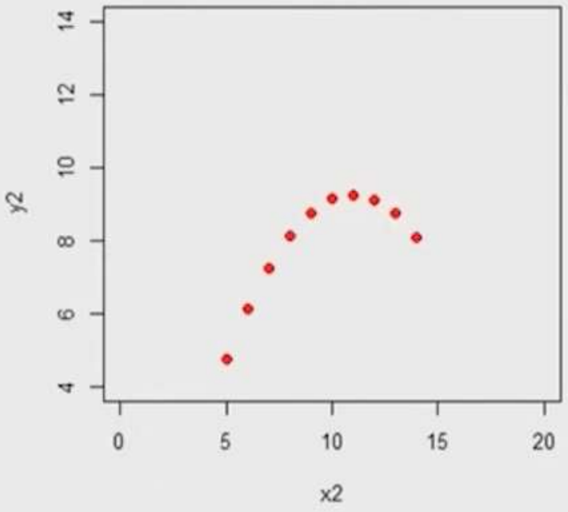
r_s=0.88

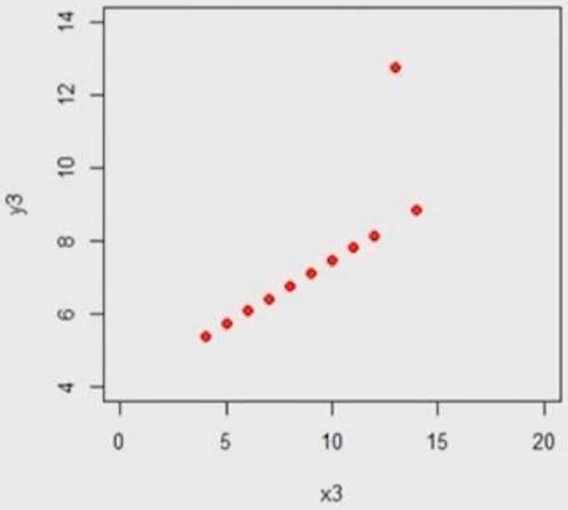
Spearman Rank Correlation

- r_s takes a value between -1 (negative association) and 1 (positive association)
- $r_s = 0$ means no association
- Monotonically increasing $r_s = 1$
- Monotonically decreasing $r_s = -1$
- Can be used when association is nonlinear
- Can be applied for ordinal variables

Spearman Rank Correlation: Anscombe's data







(x_1,y_1): 0.8182 Linear

(x_2,y_2): 0.6909 Nonlinear

(x_2,y_2): 1 Nonlinear excluding last three points

(x_3,y_3): 0.9909 Linear

Kendall rank correlation coefficient

- Correlation coefficient to measure association between two ordinal variables
- Concordant Pair: A pair of observations (x_1, y_1) and (x_2, y_2) that follows the property $x_1 > x_2$ and $y_1 > y_2$ or $x_1 < x_2$ and $y_1 < y_2$
- Discordant Pair: A pair of observations (x_1, y_1) and (x_2, y_2) that follows the property $x_1 > x_2$ and $y_1 < y_2$ or $x_1 < x_2$ and $y_1 > y_2$

Kendall rank correlation coefficient

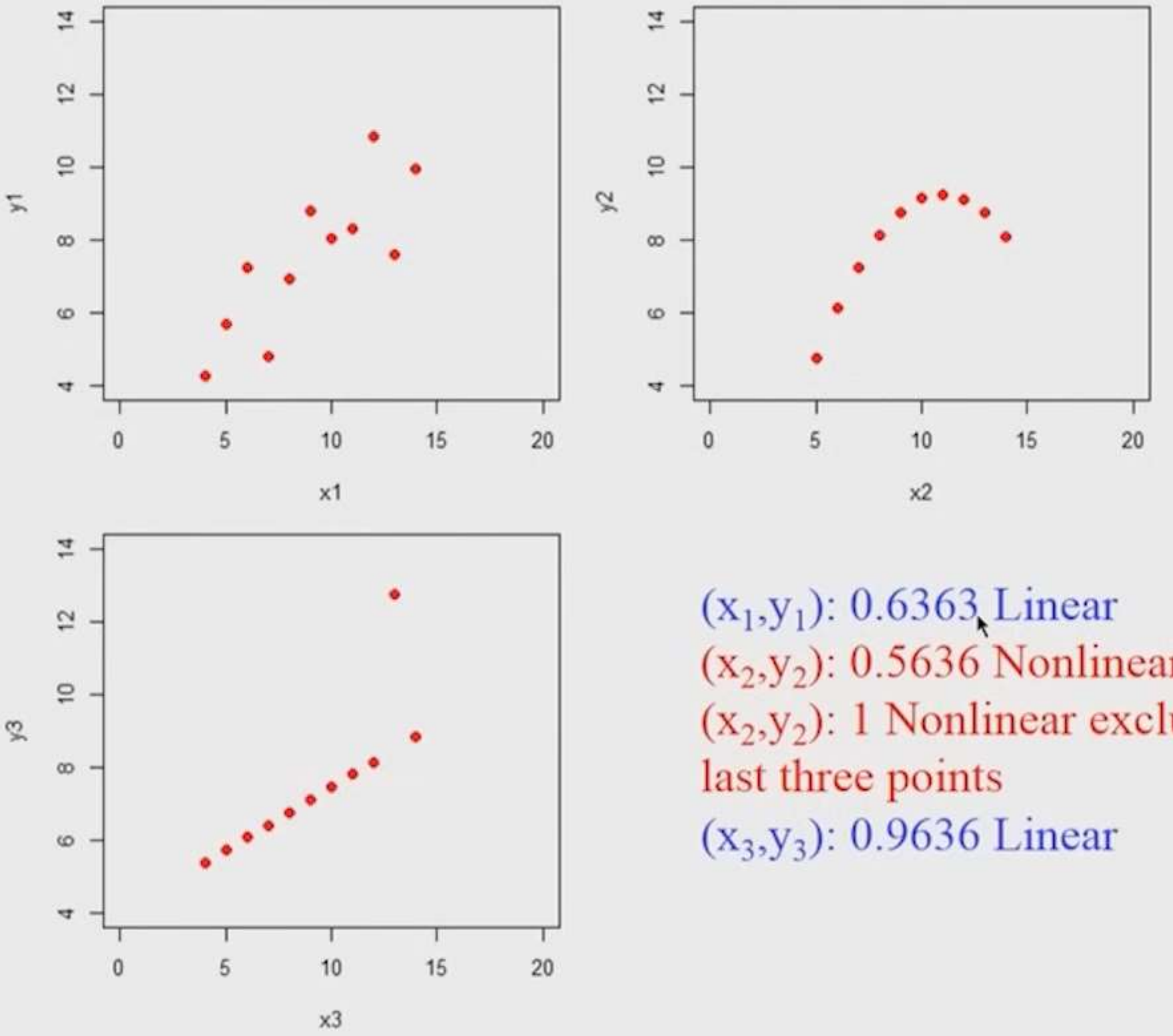
Example: Two experts ranking on food items

Items	Expert 1	Expert 2
1	1	1
2	2	3
3	3	6
4	4	2
5	5	7
6	6	4
7	7	5

1							
2	C						
3	C	C					
4	C	D	D				
5	C	C	C	C			
6	C	C	D	C	D		
7	C	C	D	C	D	C	
	1	2	3	4	5	6	7

$$\tau = \frac{15 - 6}{21} = 0.42857$$

Kendall rank Correlation: Anscombe's data



(x_1,y_1): 0.6363 Linear

(x_2,y_2): 0.5636 Nonlinear

(x_2,y_2): 1 Nonlinear excluding last three points

(x_3,y_3): 0.9636 Linear