

Looking at Relations between Continuous Variables

Primary tool: scatterplot (IOT test) of y versus x

- y dependent (response)
 - x independent (explanatory)
- e.g. Scatterplot Matrix

Geometric aspects

1. Linearity (Ellipticity?)
2. Uniform Scatter (Homoscedasticity)

Pearson's correlation coefficient:
$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

useful in summarizing linear (elliptical) patterns

e.g.

measure of strength and direction of linear association

r lies between -1 (perfect negative association) and +1 (perfect positive association)

r = 0 indicates no "trend" (i.e. no association)

Interpretation?

If (x,y) scatter is elliptical, r provides a quick summary of plot

$r \approx \pm .8$, strong association

$r \approx \pm .5$, moderate association

practical significance always unclear (association is not)

If scatterplot not elliptical, r may be misleading

Simple Linear Regression

Often wish to further describe aspects of linear association between y and x by "fitting a line" $y = a + b x$ that summarizes the pattern

a = intercept, value of y at $x = 0$

b = slope, $\frac{\Delta y}{\Delta x}$

Choosing most descriptive values for a and b

many possible methods, e.g. black string, minimum sum of absolute deviations

mathematically most "elegant", method of least squares

i.e. choose a and b so that sum of squared vertical deviations

= $\sum \{y_i - (a + b x_i)\}^2$ is as small as mathematically possible

Algebra shows that best values (relative to given data) are

$$b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\text{s.d. for y's}}{\text{s.d. for x's}} \times r$$

and

$a = \bar{y} - b \bar{x}$ - ensuring that line intersects (\bar{x}, \bar{y})

e.g.

Interpretation

For each observed value of x , x_i , yields a fitted or predicted value

$$\hat{y}_i = a + b x_i$$

e.g.

Describing variability (scatter) about the fitted line

consider magnitude of vertical deviations, $y_i - \hat{y}_i$ (residuals)

e.g.

similar in concept to deviations from mean, $x_i - \bar{x}$

by analogy define the residual standard deviation $s_{y \cdot x} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n - 2}}$