STA 674

Regression Analysis And Design Of Experiments

Measuring Association between Two Variables – Lecture 3

STA 674, RADOE:
Measuring Association between Two Variables

- What is it?
    - Correlation

STA 674: Measuring Association between Two Variables
Correlation

- Correlation—definition:

$$R_{X,Y} = \frac{\sum_{i=1}^{n}[(x_i - \bar{x})(y_i - \bar{y})]}{(n-1)s_X s_Y}$$

- Where:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

And

$$s_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

STA 674: Measuring Association between Two Variables
Correlation

- Example: Average January temperature in US cities

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = -1001.7$$

$$R_{X,Y} = \frac{\sum_{i=1}^{n}[(x_i - \bar{x})(y_i - \bar{y})]}{(n-1)s_X s_Y}$$

$$= \frac{-1001.7}{(10-1)(7.0)(16.9)}$$

$$= -0.94$$

City		Latitude	Average Jan Temp
Louisville,	KY	39	27
Key West,	FL	25	65
New Orleans,	LA	30.8	45
Atlanta,	GA	33.9	37
Charlotte,	NC	35.9	34
Harrisburg,	PA	40.9	24
Omaha,	NE	41.9	13
Detroit,	MI	43.1	21
Burlington,	VT	45	7
Spokane,	WA	48.1	19
Means		38.4	29.2
Standard deviations		7	16.9

STA 674: Measuring Association between Two Variables
Correlation

- Interpretation—first, sign.
- Look at R: $R_{X,Y} = \frac{\sum_{i=1}^{n}[(x_i - \bar{x})(y_i - \bar{y})]}{(n-1)s_X s_Y}$
- Most of the factors are always positive: $(n-1)$, s_X, s_Y.
- So need to look at product: $(x_i - \bar{x})(y_i - \bar{y})$.
  - When is an individual term positive? $+\times+$ or $-\times-$ (meaning?)
  - When is it negative? $+\times-$ or $-\times+$ (ditto?)

STA 674: Measuring Association between Two Variables
Correlation

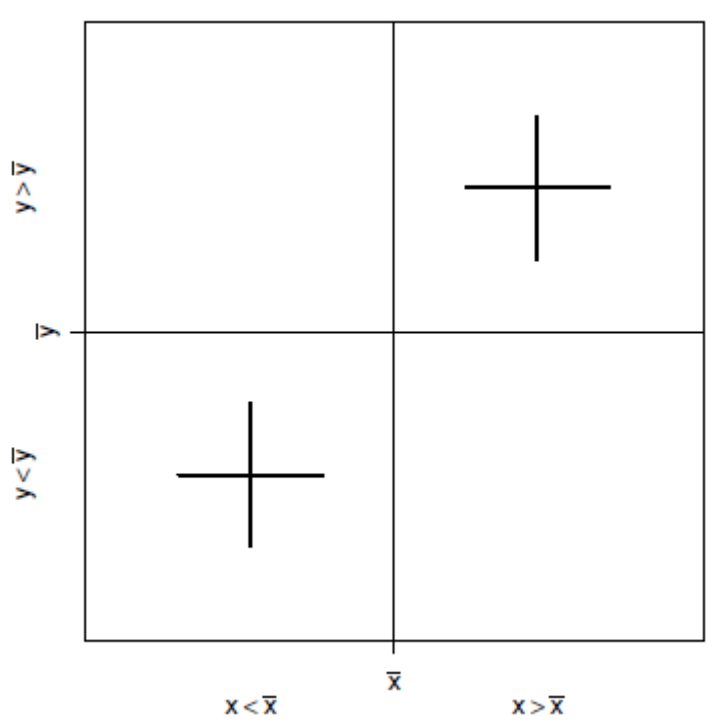
- Interpretation – sign.
- The product, $(x_i - \bar{x})(y_i - \bar{y})$, is

  positive if

  $x_i > \bar{x}$ and $y_i > \bar{y}$

  or if

  $x_i < \bar{x}$ and $y_i < \bar{y}$



STA 674: Measuring Association between Two Variables Correlation

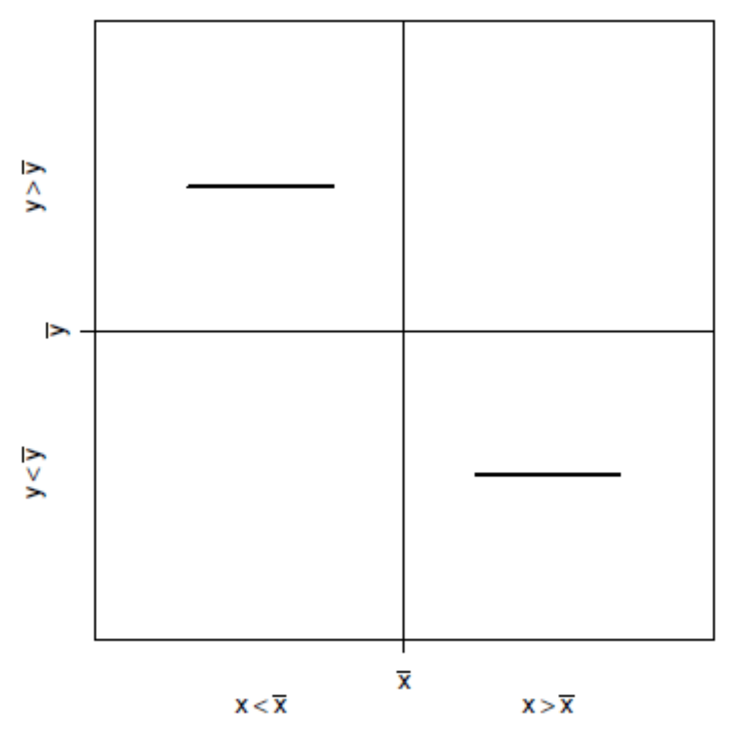
- Interpretation – sign.
- The product, $(x_i - \bar{x})(y_i - \bar{y})$, is

  negative if

  $x_i > \bar{x}$ and $y_i < \bar{y}$

  or if

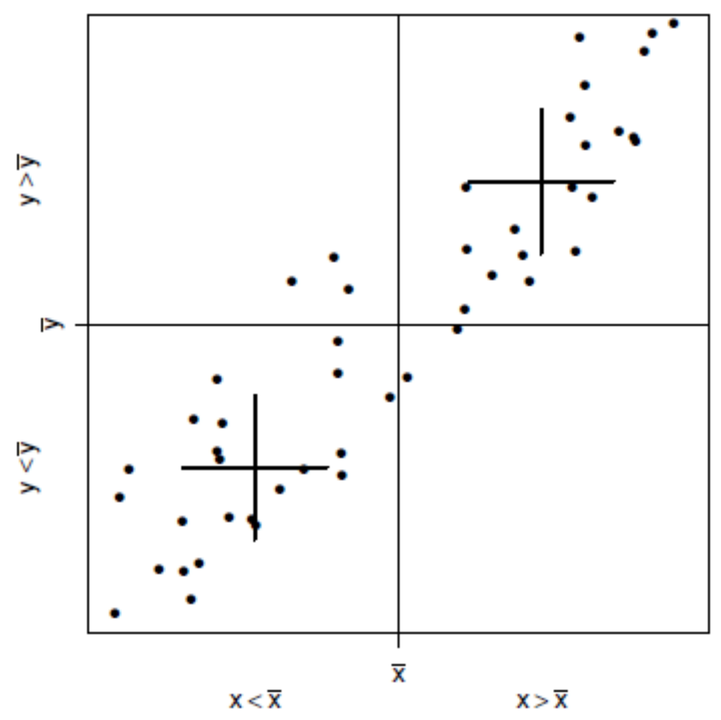
  $x_i < \bar{x}$ and $y_i > \bar{y}$



STA 674: Measuring Association between Two Variables
Correlation

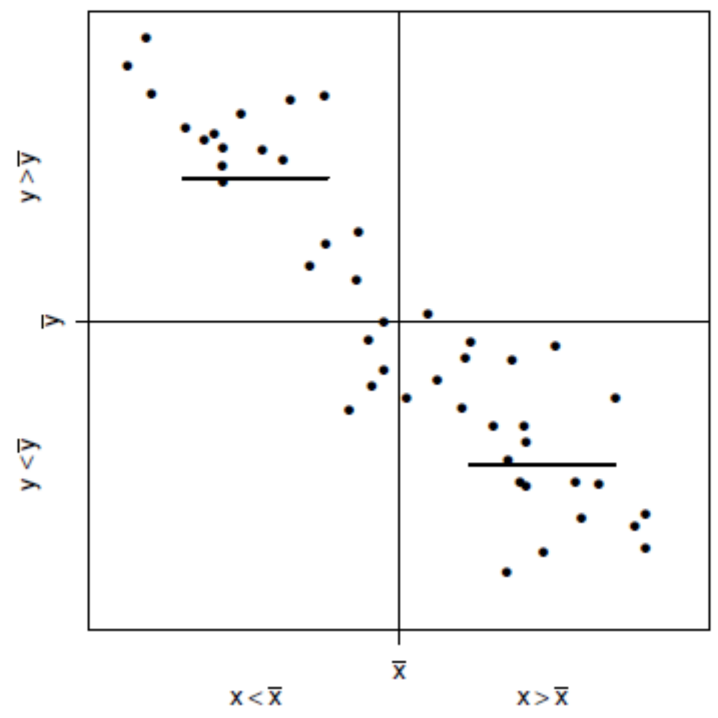
- Interpretation, finally:
- Sign of $R_{X,Y} = \frac{\sum_{i=1}^{n}[(x_i - \bar{x})(y_i - \bar{y})]}{(n-1)s_X s_Y}$

  will be **positive** if most of the data falls in the **upper-right** or **lower-left** quadrants relative to the mean.



STA 674: Measuring Association between Two Variables
Correlation

- Interpretation:
- Sign of $R_{X,Y}$ will be **negative** if most of the data falls in the **lower-right** or **upper-left** quadrants relative to the mean.



STA 674: Measuring Association between Two Variables Correlation

- Interpretation:
- The correlation, $R_{X,Y}$, will be **close to zero** if most of the data is **evenly distributed** to all four quadrants.