Statistics I – Supplements for Chapter 3 Measures of Linear Relationship

Ling-Chieh Kung

Department of Information Management
National Taiwan University

September 26, 2012

Introduction

- In Chapter 3, we have studied several descriptive measures of a **single** variable.
- How to describe a relationship between **two** variables?
- In particular, we will focus on identifying and describing a **linear relationship**.
  - A linear relationship is usually call a **correlation**.
  - For relationships among multiple variables: Multivariate statistical analysis.
  - For relationships among multiple variables with time index: Time series analysis.

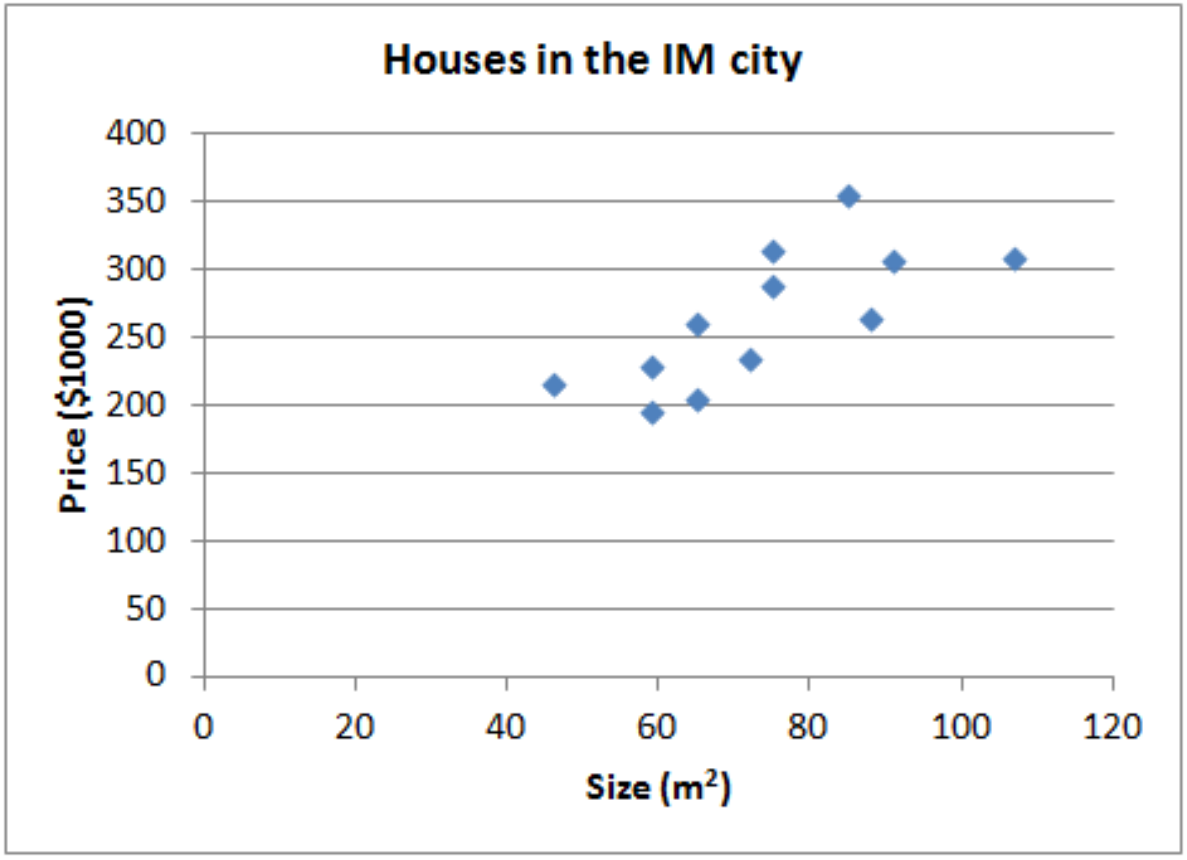
Introduction

- Suppose we have a set of **two-dimensional** quantitative data, $\{(x_i, y_i)\}_{i=1,\ldots,N}$.
  - How do we know whether there is a correlation between the two dimensions?
  - If yes, how does one dimension affect the other one?
  - If yes, how to quantify the strength?

Example: house sizes and prices

- Consider the size of a house and its price in the IM city:

Size (in m^2)	Price (in \$1000)
75	315
59	229
85	355
65	261
72	234
46	216
107	308
91	306
75	289
65	204
88	265
59	195

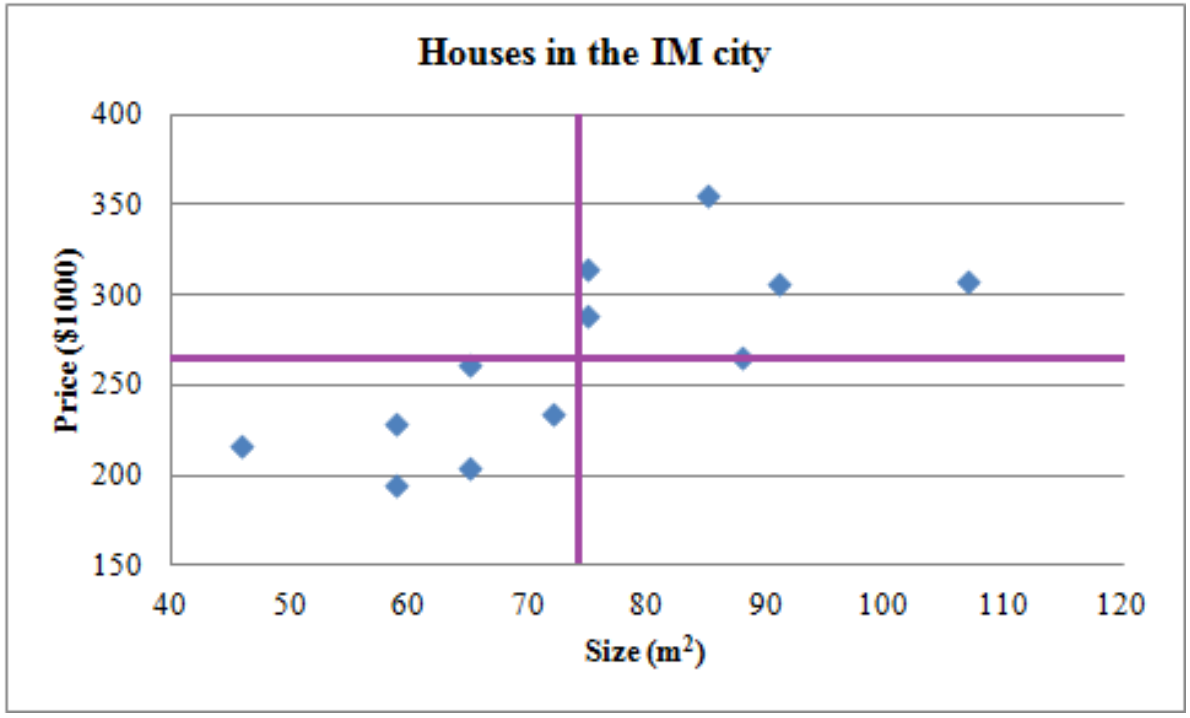


Intuition

- Suppose we want to see whether the following is true: "When one variable becomes larger, the other one tends to become larger as well."
- In this case, when x_i is larger than μ_x, the mean of the x_i's, intuitively what should happen between y_i and μ_y?
  - $y_i > \mu_y$?
  - $y_i = \mu_y$?
  - $y_i < \mu_y$?
- So let's separate the plane into four "quadrants" based on μ_x and μ_y.

Intuition

- The scatter plot with the two means:



- Most points fall in the **first** and **third quadrants**!

Covariances

- So we define the **covariance** of a set of two-dimensional **population** data as

$$\sigma_{xy} \equiv \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{N}.$$

  - If most points fall in the first and third quadrants, most $(x_i - \mu_x)(y - \mu_y)$ will be positive and σ_{xy} tends to be positive.
  - Otherwise, σ_{xy} tends to be negative.

Example: house sizes and prices

- For our example:

x_i	y_i	$x_i - \mu_x$	$y_i - \mu_y$	$(x_i - \mu_x)(y_i - \mu_y)$
75	315	1.08	50.25	54.44
59	229	−14.92	−35.75	533.27
85	355	11.08	90.25	1000.27
65	261	−8.92	−3.75	33.44
72	234	−1.92	−30.75	58.94
46	216	−27.92	−48.75	1360.94
107	308	33.08	43.25	1430.85
91	306	17.08	41.25	704.69
75	289	1.08	24.25	26.27
65	204	−8.92	−60.75	541.69
88	265	14.08	0.25	3.52
59	195	−14.92	−69.75	1040.44
$\mu_x = 73.92$	$\mu_y = 264.75$	–	–	$\sigma_{xy} = 565.73$

Covariances and variances

- Interestingly, the covariance of a single variable is its variance:

$$\begin{aligned}\sigma_{xx} &\equiv \frac{\sum_{i=1}^{N}(x_i - \mu_x)(x_i - \mu_x)}{N} \\ &= \frac{\sum_{i=1}^{N}(x_i - \mu_x)^2}{N} \equiv \sigma^2.\end{aligned}$$

Covariances and correlations

- For a pair of two variables, there are three possibilities:
  - They are **positively correlated** if the covariance is significantly **greater than zero**.
  - They are **negatively correlated** if the covariance is significantly **less than zero**.
  - They are **uncorrelated** if the covariance is **close to zero**.
- But how to define significance? How to compare the degrees of correlation among multiple pairs of variables?
- In particular, the **variability of each variable** itself affects the value of covariance. We need some kind of **normalization**.

Correlation coefficients

- The correlation coefficient of a set of two-dimensional data $\{(x_i, y_i)\}_{i=1,\ldots,N}$ is

$$\rho \equiv \frac{\sigma_{xy}}{\sigma_x \sigma_y},$$

  where σ_x and σ_y are the standard deviations of x_is and y_is, respectively.
- In essence, we normalize the **co-variability** by the **auto-variability** of the two variables.

Correlation coefficients

- Correlation coefficients have a very nice property:

Proposition 1

Let ρ be the correlation coefficient of a set of two-dimensional data, then

$$-1 \leq \rho \leq 1.$$

  - Intuition: When calculating σ_{xy}, positive and negative terms may cancel each others. But when calculating σ_x and σ_y, this never happens.

Correlation coefficients

- In practice, people often use the following rule to determine the degree of correlation based on $|\rho|$:
  - $0 \leq |\rho| < 0.25$: A weak correlation.
  - $0.25 \leq |\rho| < 0.5$: A moderately weak correlation.
  - $0.5 \leq |\rho| < 0.75$: A moderately strong correlation.
  - $0.75 \leq |\rho| \leq 1$: A strong correlation.
- Typically, we do not say "there is a correlation" if $|\rho| < 0.5$.

For sample data

- For a set of two-dimensional **sample** data $\{(x_i, y_i)\}_{i=1,\ldots,n}$:
  - The **sample covariance** is

$$s_{xy} \equiv \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}.$$

  - The **correlation coefficient** is

$$r \equiv \frac{s_{xy}}{s_x s_y}.$$