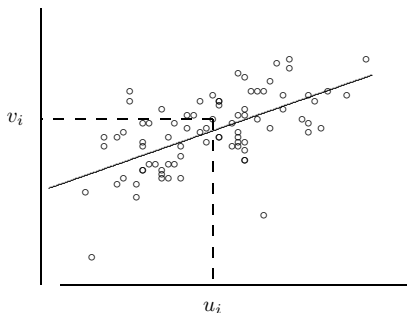


Discrete-Event Simulation: A First Course

Section 4.4: Correlation

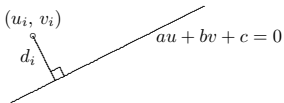
Section 4.4: Correlation

A display of the *paired sample data* (u_i, v_i) for $i = 1, 2, \dots, n$, as illustrated below, is called a *bivariate scatterplot*.



Bivariate Scatterplots and Paired Correlation

- In a bivariate scatterplot, sometimes the points (u_i, v_i) lie along or near a line
- *Linear* correlation shows us the line that “best fits” the data
- Consider the line defined by the equation $au + bv + c = 0$ and for each point (u_i, v_i)
- Let d_i be the *orthogonal* distance from this point to the line

$$d_i = \frac{|au_i + bv_i + c|}{\sqrt{a^2 + b^2}}$$


- Choose (a, b, c) line parameters that *minimize* the mean-square orthogonal distance

$$D = \frac{1}{n} \sum_{i=1}^n d_i^2 = \frac{1}{n(a^2 + b^2)} \sum_{i=1}^n (au_i + bv_i + c)^2.$$

Three-Parameter Minimization

- To minimize D with parameters (a, b, c)
- D can be written as

$$D = \frac{1}{n(a^2 + b^2)} \sum_{i=1}^n \left(a(u_i - \bar{u}) + b(v_i - \bar{v}) \right)^2 + \frac{1}{n(a^2 + b^2)} \sum_{i=1}^n (a\bar{u} + b\bar{v} + c)^2$$

where

$$\bar{u} = \frac{1}{n} \sum_{i=1}^n u_i \quad \text{and} \quad \bar{v} = \frac{1}{n} \sum_{i=1}^n v_i$$

are the sample means of the u and v data respectively

Three-Parameter Minimization

- Both terms in the new D equation are nonnegative
- The first term is independent of c , so c should be chosen to minimize the second term
- Choose c so that $a\bar{u} + b\bar{v} + c = 0$, minimizing the second term
- If $a\bar{u} + b\bar{v} + c = 0$, the line passes through the point $(u, v) = (\bar{u}, \bar{v})$
- The problem has been reduced to a two-parameter problem

Two-Parameter Minimization

- Because $c = -a\bar{u} - b\bar{v}$ the equation $au + bv + c = 0$ can be written equivalently as $a(u - \bar{u}) + b(v - \bar{v}) = 0$
- Simplify the equation for D by assuming that (a, b) are normalized so that $a^2 + b^2 = 1$

Theorem (4.4.1)

The line that best fits the data (u_i, v_i) for $i = 1, 2, \dots, n$ in a mean-squared orthogonal distance sense is given by the equation

$$a(u - \bar{u}) + b(v - \bar{v}) = 0$$

where the (a, b) line parameters are chosen to minimize

$$D = \frac{1}{n} \sum_{i=1}^n \left(a(u_i - \bar{u}) + b(v_i - \bar{v}) \right)^2$$

subject to the constraint $a^2 + b^2 = 1$.

Covariance and Correlation

Given the bivariate sample (u_i, v_i) for $i = 1, 2, \dots, n$

- the (linear) *sample covariance* is

$$c = \frac{1}{n} \sum_{i=1}^n (u_i - \bar{u})(v_i - \bar{v}) = \frac{1}{n} \left(\sum_{i=1}^n u_i v_i \right) - \bar{u} \bar{v}$$

(the second equation is the one-pass version)

- provided both s_u and s_v are not zero, the (linear) *sample correlation coefficient* is

$$r = \frac{c}{s_u s_v}$$

- \bar{u} , \bar{v} , $s_u^2 = \frac{1}{n} \sum_{i=1}^n (u_i - \bar{u})^2$, and $s_v^2 = \frac{1}{n} \sum_{i=1}^n (v_i - \bar{v})^2$ are the sample means and sample variances of the u and v data values respectively

Covariance and Correlation

- The correlation coefficient r measures the “spread” (dispersion) of the u, v data about the line that best fits the data
- D can be written in terms of s_u^2 , s_v^2 , and r as

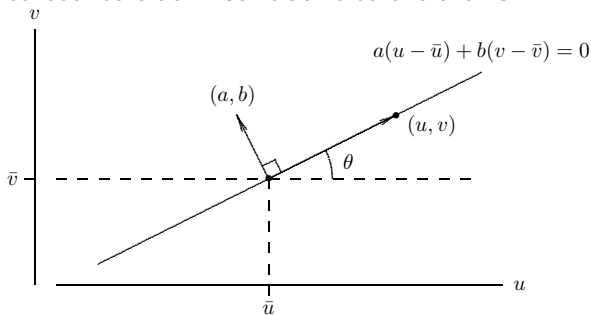
$$\begin{aligned}
 D &= \frac{1}{n} \sum_{i=1}^n \left(a(u_i - \bar{u}) + b(v_i - \bar{v}) \right)^2 \\
 &= \frac{1}{n} \sum_{i=1}^n \left(a^2(u_i - \bar{u})^2 + 2ab(u_i - \bar{u})(v_i - \bar{v}) + b^2(v_i - \bar{v})^2 \right) \\
 &= \frac{a^2}{n} \left(\sum_{i=1}^n (u_i - \bar{u})^2 \right) + \frac{2ab}{n} \left(\sum_{i=1}^n (u_i - \bar{u})(v_i - \bar{v}) \right) \\
 &\quad + \frac{b^2}{n} \left(\sum_{i=1}^n (v_i - \bar{v})^2 \right) \\
 &= a^2 s_u^2 + 2ab r s_u s_v + b^2 s_v^2.
 \end{aligned}$$

- Note:

$$|r| = 1 \quad \iff \quad D = 0 \quad \iff \quad \text{all the points } (u_i, v_i) \text{ lie on a line.}$$

One-Parameter Minimization

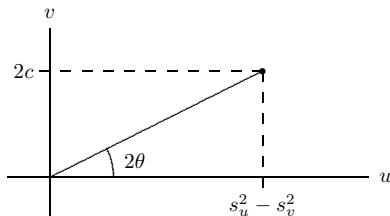
- Let θ be the angle of the line $a(u - \bar{u}) + b(v - \bar{v}) = 0$ measured counterclockwise relative to the u -axis



- The relation between (a, b) and θ using elementary trigonometry is $a = -\sin \theta$ and $b = \cos \theta$
- Must find θ that minimizes

$$D = \frac{1}{2} (s_u^2 + s_v^2) - c \sin 2\theta - \frac{1}{2} (s_u^2 - s_v^2) \cos 2\theta$$

Finding θ to Minimize D



The angle that minimizes D is

$$\theta = \frac{1}{2} \tan^{-1} (s_u^2 - s_v^2, 2c)$$

Theorem 4.4.2

Theorem (4.4.2)

The line that best fits the data (u_i, v_i) for $i = 1, 2, \dots, n$ in a mean-squared orthogonal distance sense passes through the point (\bar{u}, \bar{v}) at the angle

$$\theta = \frac{1}{2} \tan^{-1} (s_u^2 - s_v^2, 2c)$$

measured counterclockwise relative to the positive u -axis. The equation of the line is

$$v = (u - \bar{u}) \tan(\theta) + \bar{v}$$

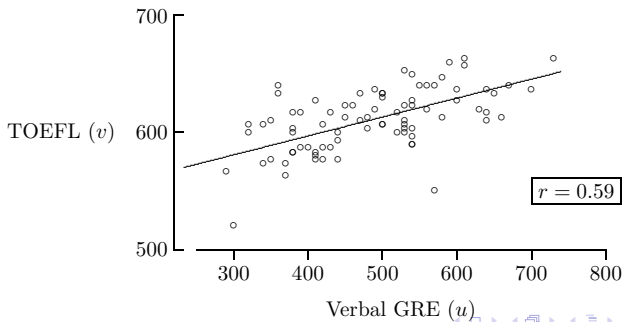
provided $\theta \neq \pi/2$. (By convention $-\pi < \tan^{-1}(u, v) \leq \pi$ so that $-\pi/2 < \theta \leq \pi/2$.)

Linear Regression Line and More

- The line that best fits the data (u_i, v_i) for $i = 1, 2, \dots, n$ is known as the (mean-square orthogonal distance) *linear regression line*
- The correlation coefficient satisfies the inequality $-1 \leq r \leq 1$
- The closer $|r|$ is to 1, the smaller the dispersion of the (u, v) data about the regression line, and the better the (linear) fit
- All the (u_i, v_i) points lie on the regression line if and only if $D_{\min} = 0$ or equivalently if and only if $|r| = 1$

Example 4.4.1

- The scatterplot corresponds to 82 student scores on two standardized tests of English verbal skills: the Test of English as a Foreign Language (TOEFL) and Graduate Record Examination (GRE)
- In this case $r = 0.59$
- The consistency between the two tests is certainly less than desirable

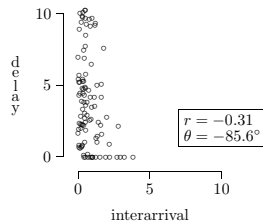
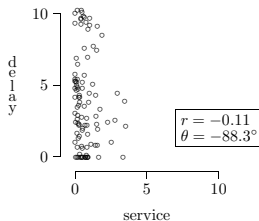
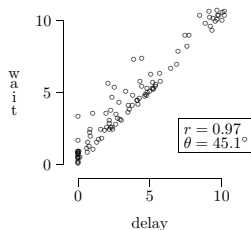
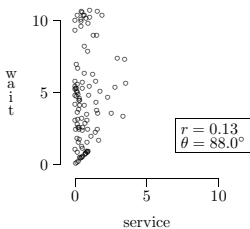


Significance

- The magnitude of $|r|$ measures the linear relation between u and v
- If $r \neq 0$ then the slope of the regression line is positive ($\theta > 0$) if and only if $r > 0$ and the slope of the regression line is negative ($\theta < 0$) if and only if $r < 0$.
- If r is close to $+1$ the data is said to be *positively correlated*.
- If r is close to -1 the data is said to be *negatively correlated*.
- If r is close to 0 then the data is said to be *uncorrelated*.

Example 4.4.2

Bivariate scatterplots for interarrival, service, delay, and wait times for a steady-state sample of 100 jobs passing through an $M/M/1$ service node with arrival rate 1.0 and service rate 1.25 (both service and interarrival are *Exponential*)



Computational Considerations

Theorem (4.4.2)

Let \bar{u}_i and \bar{v}_i denote the sample means of u_1, u_2, \dots, u_i and v_1, v_2, \dots, v_i respectively and define

$$w_i = (u_1 - \bar{u}_i)(v_1 - \bar{v}_i) + (u_2 - \bar{u}_i)(v_2 - \bar{v}_i) + \dots + (u_i - \bar{u}_i)(v_i - \bar{v}_i)$$

for $i = 1, 2, \dots, n$ where w_i/i is the covariance of the first i data pairs. Then, with the initial condition $w_0 = 0$,

$$w_i = w_{i-1} + \left(\frac{i-1}{i} \right) (u_i - \bar{u}_{i-1})(v_i - \bar{v}_{i-1}) \quad i = 1, 2, \dots, n$$

which provides a one-pass recursive algorithm to compute $c_{uv} = w_n/n$.

- Program bvs is based upon the extended version of Welford's algorithm in Theorem 4.4.3
- This program illustrates the calculation of the *bivariate* sample statistics \bar{u} , s_u , \bar{v} , s_v , r , and the linear regression line angle θ

Serial Correlation

- Frequently we are interested in how a set of data is *auto-correlated* (e.g., self-correlated)
- For example, in a steady-state analysis of the waits experienced by consecutive jobs entering a service node
- If the utilization of the service node is high, there will be a high positive correlation between the wait w_i experienced by the i^{th} job and the wait w_{i+1} experienced by the next job
- There will be a statistically significant positive correlation between w_i and w_{i+j} for some range of small, positive j values

Autocorrelation Lag

- Let x_1, x_2, \dots, x_n be data which is presumed to represent n consecutive observations of some stochastic process
- Pick a (small) fixed positive integer $j \ll n$ and then associate u_i with x_i and v_i with x_{i+j}

$$\begin{array}{rcccccccccccccccc}
 u & : & & & & x_1 & x_2 & x_3 & \cdots & x_i & \cdots & x_{n-j} & x_{n-j+1} & \cdots & x_n \\
 v & : & x_1 & \cdots & x_j & x_{1+j} & x_{2+j} & x_{3+j} & \cdots & x_{i+j} & \cdots & x_n & & &
 \end{array}$$

- The integer $j > 0$ is called the *autocorrelation lag* (or *shift*)
- The value $j = 1$ is generally of primary interest
- It is also conventional to calculate the serial correlation for a range of lag values $j = 1, 2, \dots, k$ where $k \ll n$

Sample Autocorrelation

- To resolve the “non-overlap” in the data at the beginning and end — ignore the extreme data values
- Define the sample autocovariance for lag j , based only on the $n - j$ overlapping values, as

$$c_j = \frac{1}{n-j} \sum_{i=1}^{n-j} (x_i - \bar{x})(x_{i+j} - \bar{x}) \quad j = 1, 2, \dots, k,$$

where the sample mean, based on all n values, is

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

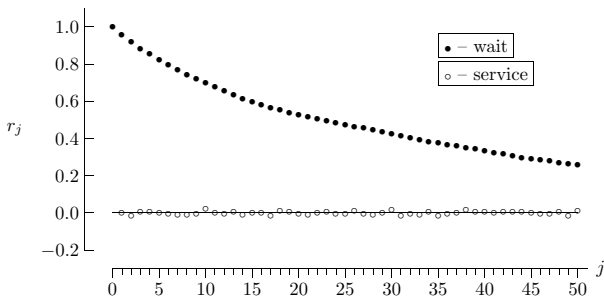
- Define the *sample autocorrelation* for lag j is $r_j = \frac{c_j}{c_0}$ for $j = 1, 2, \dots, k$ where the sample variance is $c_0 = s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$

Computational Considerations

- The one-pass version of the sample autocovariance for lag j is

$$c_j = \left(\frac{1}{n-j} \sum_{i=1}^{n-j} x_i x_{i+j} \right) - \bar{x}^2 \quad j = 1, 2, \dots, k.$$

- Example 4.4.3: 10 000 consecutive jobs processed through an $M/M/1$ service node, in steady-state, with arrival rate 1.0, service rate 1.25, and utilization $1/1.25 = 0.8$



Program acs

Algorithm 4.4.1: One Pass Algorithm for Fixed Lag j

A circular queue is initially filled with $x_1, x_2, \dots, x_k, x_{k+1}$, as illustrated by the boxed elements below. The lagged products $x_1 x_{1+j}$ are computed for all $j = 0, 1, \dots, k$ thereby initializing the $k + 1$ cosums. Then the next data value is read into the (old) head of the queue location, p is incremented by 1 to define a new head of the queue location, the lagged products $x_2 x_{2+j}$ are computed for all $j = 0, 1, \dots, k$, and the cosums are updated. This process is continued until all the data has been read and processed. (The case $n \bmod (k + 1) = 2$ is illustrated.)

$(i = k + 1)$	x_1	x_2	x_3	\dots	x_{k-1}	x_k	x_{k+1}	$(p = 0)$
$(i = k + 2)$	x_{k+2}	x_2	x_3	\dots	x_{k-1}	x_k	x_{k+1}	$(p = 1)$
$(i = k + 3)$	x_{k+2}	x_{k+3}	x_3	\dots	x_{k-1}	x_k	x_{k+1}	$(p = 2)$
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
$(i = 2k)$	x_{k+2}	x_{k+3}	x_{k+4}	\dots	x_{2k}	x_k	x_{k+1}	$(p = k)$
$(i = 2k + 1)$	x_{k+2}	x_{k+3}	x_{k+4}	\dots	x_{2k}	x_{2k+1}	x_{k+1}	$(p = k + 1)$
$(i = 2k + 2)$	x_{k+2}	x_{k+3}	x_{k+4}	\dots	x_{2k}	x_{2k+1}	x_{2k+2}	$(p = 0)$
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
$(i = n)$	x_{n-1}	x_n	x_{n-k}	\dots	x_{n-4}	x_{n-3}	x_{n-2}	$(p = 2)$

After the last data value, x_n , has been read, the associated lagged products computed, and the cosums updated, all that remains is to “empty” the queue. This can be accomplished by effectively reading k additional 0-valued data values. For more details, see program acs.