



EEC 686/785
Modeling & Performance Evaluation of
Computer Systems

Lecture 11

Wenbing Zhao

Department of Electrical and Computer Engineering
Cleveland State University

wenbing@ieee.org

(based on Dr. Raj Jain's lecture notes)



Outline

2

- Move 2nd midterm to Nov.14?
- Review of lecture 10
- Other regression models
- Experimental design

Midterm #1

P1	P2	P3	P4	P5	P6	P7	P8	Total
14	14	10	7	7.5	10	10	7.5	80
18	10	10	4.5	7.5	9	10	9.5	78.5
18	10	10	4.5	7.5	10	10	8	78
12	16	10	1.5	5	10	5	4.5	64
12	8	10	3	5	10	5	8.5	61.5
8	10	10	0	10	10	10	3.5	61.5

Workload Characterization Techniques

- **Workload characterization:** the process of studying the real-user environments, observe the key characteristics, and develop a workload model that can be used repeated
- The measured workload data consists of services requested or the resource demands of a number of users on the system
- The term “**user**” denotes the entity that makes the service requests at the SUT interface



Workload Characterization Techniques

- In workload characterization literature, the term **workload component** or **workload unit** is used instead of the user
 - The **workload component** should be at the SUT interface
- **Workload parameters** or **workload features**
 - Measured quantities, service requests, or resource demands
 - For example: transaction types, instructions, packet sizes, source-destinations of a packet, and page reference pattern
- Each component should represent as homogeneous a group as possible



Workload Characterization Techniques

- Averaging
- Single-parameter histograms
- Multiparameter histograms
- Principal component analysis
- Markov models
- Clustering



Principal Component Analysis

- Key idea: use a weighted sum of parameters to classify the components
- Let x_{ij} denote the i th parameter for j th component

$$y_j = \sum_{i=1}^n w_i x_{ij}$$

- Principal component analysis assigns weights w_i 's such that y_j 's provide the maximum discrimination among the components
- The quantity y_j is called the **principal factor**
- The factors are ordered. First factor explains the highest percentage of the variance



Finding Principal Factors

- Find the correlation matrix
- Find the eigenvalues of the matrix and sort them in the order of decreasing magnitude
- Find corresponding eigenvectors. These give the required loadings

Markov Models

- Markov => the next request depends only on the last request
- Described by a transition matrix

From/To	CPU	Disk	Terminal
CPU	0.6	0.3	0.1
Disk	0.9	0	0.1
Terminal	1	0	0

- Given the same relative frequency of requests of different types, it is possible to realize the frequency with several different transition matrices

Clustering

- Take a sample, that is, a subset of workload components
- Select workload parameters
- Select a distance measure
- Remove outliers
- Scale all observations
- Perform clustering
- Interpret results
- Change parameters, or number of clusters, and repeat 3-7
- Select representative components from each cluster

Multiple Linear Regression Models

- A multiple linear regression model allows one to predict a response variable y as a function of k predictor variables x_1, x_2, \dots, x_k using a linear model:
- Given a sample $\{(x_{11}, x_{21}, \dots, x_{k1}, y_1), \dots, (x_{1n}, x_{2n}, \dots, x_{kn}, y_n)\}$ of n observations, the model consists of the following n equations:

$$y_1 = b_0 + b_1 x_{11} + b_2 x_{21} + \dots + b_k x_{k1} + e_1$$

$$y_2 = b_0 + b_1 x_{12} + b_2 x_{22} + \dots + b_k x_{k2} + e_2$$

$$\vdots$$

$$y_n = b_0 + b_1 x_{1n} + b_2 x_{2n} + \dots + b_k x_{kn} + e_n$$

Multiple Linear Regression Models

- In vector notation, we have:

$$\begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \cdot \\ \cdot \\ b_n \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \cdot \\ \cdot \\ e_n \end{bmatrix}$$

or

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$

$$\mathbf{b} = (\mathbf{X}^T \mathbf{X})^{-1} (\mathbf{X}^T \mathbf{y})$$

Regression with Categorical Predictors

- Examples of categorical variables: CPU types
- To represent a categorical variable that can take only 2 values, we can define a binary variable, x_j , that takes 2 levels: +1 and -1, *i.e.*,

$$x_j = \begin{cases} -1 \Rightarrow \text{first value} \\ +1 \Rightarrow \text{second value} \end{cases}$$

- For a categorical variable that takes 3 values, cannot simply define a three-value variable because it implies an order

$$x_i = \begin{cases} 1 \Rightarrow \text{type A} \\ 2 \Rightarrow \text{type B} \\ 3 \Rightarrow \text{type C} \end{cases}$$

Regression with Categorical Predictors

- The recommended coding is to use 2 predictor variables:

$$x_1 = \begin{cases} 1 \Rightarrow \text{if type A} \\ 0 \Rightarrow \text{otherwise} \end{cases} \quad x_2 = \begin{cases} 1 \Rightarrow \text{if type B} \\ 0 \Rightarrow \text{otherwise} \end{cases}$$

- The 3 types can be represented by (x_1, x_2) pairs:

$$(x_1, x_2) = (1, 0) \Rightarrow \text{type A}$$

$$(x_1, x_2) = (0, 1) \Rightarrow \text{type B}$$

$$(x_1, x_2) = (0, 0) \Rightarrow \text{type C}$$

Regression with Categorical Predictors

- The regression model for a 3-level categorical variable:
 $y = b_0 + b_1x_1 + b_2x_2 + e$

- The average responses for the three types are:

$$\bar{y}_A = b_0 + b_1$$

$$\bar{y}_B = b_0 + b_2$$

$$\bar{y}_C = b_0$$

- Parameter b_1 represents the difference between average responses with types A and C
- Parameter b_2 represents the difference between average responses with types B and C
- Parameter b_0 represents average response with type C

Regression with Categorical Predictors

- To represent a categorical variable with k levels (or k categories), we need to define $k - 1$ binary variables as follows:

$$x_j = \begin{cases} 1 & \Rightarrow \text{if } j\text{th value} \\ 0 & \Rightarrow \text{otherwise} \end{cases}$$

- The k th value is defined by $x_1 = x_2 = \dots = x_{k-1} = 0$. The regression parameter b_0 represents the average response with the k th alternative
- The parameter b_j represents the difference between the average responses with alternatives j and k

Curvilinear Regression

- **Curvilinear regression:** if the nonlinear function can be converted into a linear form, then the regression can be carried out using the simple or multiple linear regression techniques

- Nonlinear

$$y = a + b/x$$

$$y = 1/(a+bx)$$

$$y = x/(a+bx)$$

$$y = abx$$

$$y = a + bx^n$$

$$y = bx^a$$

- Linear

$$y = a + b(1/x)$$

$$(1/y) = a + bx$$

$$(x/y) = a + bx$$

$$\ln y = \ln a + (\ln b)\ln x$$

$$y = a + b(x^n)$$

$$\ln y = \ln b + a \ln x$$

Transformations

- **Transformation:** when some function of the measured response variable y is used in place of y in a model

- E.g., square root transformation:

$$\sqrt{y} = b_0 + b_1x_1 + b_2x_2 + \dots + b_kx_k + e$$



When Transformations Are Needed

- If it is known from physical considerations of the system that a function of the response rather than the response itself is a better variable to use in the model
 - Measured the interarrival times y for requests and it is known that the number of requests per unit time ($1/y$) has a linear relationship to a certain predictor
- If the range of the data covers several orders of magnitude and the sample size is small \Rightarrow use transformation to reduce the range of variability
- If the homogeneous variance assumption of the residuals is violated



Experimental Design and Analysis

- Design a proper set of experiments for measurement or simulation
- Develop a model that best describes the data obtained
- Estimate the contribution of each alternative to the performance
- Isolate the measurement errors
- Estimate confidence intervals for model parameters
- Check if the alternatives are significantly different
- Check if the model is adequate



Terminology

- **Response Variables:** outcome
 - E.g., throughput, response time
- **Factors:** variables that affect the response variable
 - E.g., CPU type, memory size, number of disk drives, workload used, and user's educational level
 - Also called predictor variables or predictors
- **Levels:** the values that a factor can assume
 - E.g., the CPU type has three levels: 68000, 8080, Z80
 - Also called treatment



Terminology

- **Primary Factors:** the factors whose effects need to be quantified
 - E.g., CPU type, memory size only, and number of disk drives
- **Secondary Factors:** factors whose impact need not be quantified
- **Replication:** repetition of all or some experiments
- **Design:** the number of experiments, the factor level and number of replications for each experiment
 - E.g., full factorial design with 5 replications: $3 \times 3 \times 4 \times 3 \times 3$ or 324 experiments, each repeated five times

Terminology

- **Experimental Unit:** any entity that is used for experiments. Usually only those experimental units that are considered as one of the factors are of interest
 - E.g., users hired to use the workstation while measurements are being performed can be considered as the experimental unit.
 - Generally, no interest in comparing the units
 - Goal: minimize the impact of variation among the units

Terminology

- **Interaction:** effect of one factor depends upon the level of the other
 - Example: two factors A and B, each has two levels

Noninteracting Factors Interacting Factors

	A_1	A_2
B_1	3	5
B_2	6	8

	A_1	A_2
B_1	3	5
B_2	6	9

Types of Experimental Designs

- Given k factors, with i th factor having n_i levels
- **Simple Designs:** vary one factor at a time
 - Number of experiments $n = 1 + \sum_{i=1}^k (n_i - 1)$
 - Not statistically efficient
 - Wrong conclusions if the factors have interaction
 - Not recommended

Types of Experimental Designs

- **Full Factorial Design:** all combinations
 - Number of experiments $= \prod_{i=1}^k n_i$
 - Can find the effect of all factors
 - Too much time and money
 - Ways to reduce the number of experiments
 - Reduce the number of levels for each factor, e.g., 2 levels per factor
 - Reduce the number of factors
 - Use fractional factorial designs

Types of Experimental Designs

- **Fractional Factorial Designs:** save time and expense
 - Less information
 - May not get all interactions
 - Not a problem if negligible interactions

Types of Experimental Designs

- **A sample fractional factorial design**
 - 9 exprs (3^{4-2} design) instead of 81 (3^4 design)

Experiment Number	CPU	Memory Level	Workload Type	Educational Level
1	68000	512K	Managerial	High School
2	68000	2M	Scientific	Post-graduate
3	68000	8M	Secretarial	College
4	Z80	512K	Scientific	College
5	Z80	2M	Secretarial	High School
6	Z80	8M	Managerial	Post-graduate
7	8086	512K	Secretarial	Post-graduate
8	8086	2M	Managerial	College
9	8086	8M	Scientific	High School

2^k Factorial Designs

- k factors, each at two levels
- Easy to analyze
- Helps in sorting out impact of factors
- Good at the beginning of a study
- Valid only if the effect of a factor is unidirectional, i.e., the performance either continuously decreases or continuously increases as the factor is increased from min to max
 - E.g., memory size, the number of disk drives

2² Factorial Designs: Example

- Two factors, each at two levels

Performance in MIPS		
Cache Size	Memory Size	
	4M Bytes	16M Bytes
1K	15	45
2K	25	75

$$x_A = \begin{cases} -1 & \text{if 4M bytes memory} \\ 1 & \text{if 16M bytes memory} \end{cases}$$

$$x_B = \begin{cases} -1 & \text{if 1K bytes cache} \\ 1 & \text{if 2K bytes cache} \end{cases}$$

2² Factorial Designs: Model

- $y = q_0 + q_A x_A + q_B x_B + q_{AB} x_A x_B$

$$15 = q_0 - q_A - q_B + q_{AB}$$

$$45 = q_0 + q_A - q_B + q_{AB}$$

$$25 = q_0 - q_A + q_B - q_{AB}$$

$$75 = q_0 + q_A + q_B + q_{AB}$$
- Unique solution for q_A and q_B :

$$y = 40 + 20x_A + 10x_B + 5x_A x_B$$
- Interpretation:
 - Mean performance = 40 MIPS
 - Effect of memory = 20 MIPS
 - Effect of cache = 10 MIPS
 - Interaction between memory and cache = 5 MIPS

Computation of Effects

Experiment	A	B	y
1	-1	-1	y_1
2	1	-1	y_2
3	-1	1	y_3
4	1	1	y_4

- Model: $y = q_0 + q_A x_A + q_B x_B + q_{AB} x_A x_B$
- Substitution:

$$y_1 = q_0 - q_A - q_B + q_{AB}$$

$$y_2 = q_0 + q_A - q_B + q_{AB}$$

$$y_3 = q_0 - q_A + q_B - q_{AB}$$

$$y_4 = q_0 + q_A + q_B + q_{AB}$$

Computation of Effects

- Solution:

$$q_0 = \frac{1}{4} (y_1 + y_2 + y_3 + y_4)$$

$$q_A = \frac{1}{4} (-y_1 + y_2 - y_3 + y_4)$$

$$q_B = \frac{1}{4} (-y_1 + y_2 + y_3 + y_4)$$

$$q_{AB} = \frac{1}{4} (y_1 - y_2 - y_3 + y_4)$$

Experiment	A	B	y
1	-1	-1	y_1
2	1	-1	y_2
3	-1	1	y_3
4	1	1	y_4

- Notice that effects are linear combinations of responses. Sum of the coefficients is zero => contrasts

- Notice:

$$q_A = \text{Column A} \times \text{Column y}$$

$$q_B = \text{Column B} \times \text{Column y}$$

$$q_{AB} = \text{Column A} \times \text{Column B} \times \text{Column y}$$

Sign Table Method

I	A	B	AB	y
1	-1	-1	1	15
1	1	-1	-1	45
1	-1	1	-1	25
1	1	1	1	75
160	80	40	20	Total
40	20	10	5	Total/4

- For a 2^2 design, the effects can be computed easily by preparing a 4×4 sign matrix as shown above
- Next, multiply the entries in column I by those in column y and put their sum under column I; perform similar calculation for other columns
- The sums under each column are divided by 4 to give the corresponding coefficients of the regression model

Allocation of Variation

- Importance of a factor = proportion of the *variation* explained

- Sample Variance of $y = s_y^2 = \frac{\sum_{i=1}^{2^2} (y_i - \bar{y})^2}{2^2 - 1}$

- Variation of $y \triangleq$ Numerator

$$= \sum_{i=1}^{2^2} (y_i - \bar{y})^2$$

= *sum of squares total (SST)*

Allocation of Variation

- For a 2^2 design

$$SST = 2^2 q_A^2 + 2^2 q_B^2 + 2^2 q_{AB}^2$$

- Variation due to A=SSA= $2^2 q_A^2$

- Variation due to B=SSB= $2^2 q_B^2$

- Variation due to interaction AB=SSAB= $2^2 q_{AB}^2$

$$SST = SSA + SSB + SSAB$$

$$\text{Fraction explained by A} = SSA / SST$$

- Variation \neq Variance

Derivation

- Model:

$$y_i = q_0 + q_A x_{Ai} + q_B x_{Bi} + q_{AB} x_{Ai} x_{Bi}$$

- Notice

- The sum of entries in each column is zero

$$\sum_{i=1}^4 x_{Ai} = 0; \sum_{i=1}^4 x_{Bi} = 0; \sum_{i=1}^4 x_{Ai} x_{Bi} = 0$$

- The sum of the squares of entries in each column is 4:

$$\sum_{i=1}^4 x_{Ai}^2 = 4; \sum_{i=1}^4 x_{Bi}^2 = 4; \sum_{i=1}^4 (x_{Ai} x_{Bi})^2 = 4$$

Derivation

- Notice (continued)

- The columns are orthogonal (inner product of any two columns is zero):

$$\sum_{i=1}^4 x_{Ai} x_{Bi} = 0; \sum_{i=1}^4 x_{Ai} (x_{Ai} x_{Bi}) = 0; \sum_{i=1}^4 x_{Bi} (x_{Ai} x_{Bi}) = 0$$

- Sample mean $\bar{y} = \frac{1}{4} \sum_{i=1}^4 y_i$

$$\begin{aligned} &= \frac{1}{4} \sum_{i=1}^4 (q_0 + q_A x_{Ai} + q_B x_{Bi} + q_{AB} x_{Ai} x_{Bi}) \\ &= \frac{1}{4} \sum_{i=1}^4 q_0 + \frac{1}{4} q_A \sum_{i=1}^4 x_{Ai} + \frac{1}{4} q_B \sum_{i=1}^4 x_{Bi} + \frac{1}{4} q_{AB} \sum_{i=1}^4 x_{Ai} x_{Bi} \\ &= q_0 \end{aligned}$$

Derivation

■ Variation of y

$$\begin{aligned}
 &= \sum_{i=1}^4 (y_i - \bar{y})^2 \\
 &= \sum_{i=1}^4 (q_A x_{Ai} + q_B x_{Bi} + q_{AB} x_{Ai} x_{Bi})^2 \\
 &= \sum_{i=1}^4 (q_A x_{Ai})^2 + \sum_{i=1}^4 (q_B x_{Bi})^2 + \sum_{i=1}^4 (q_{AB} x_{Ai} x_{Bi})^2 + \text{product terms} \\
 &= q_A^2 \sum_{i=1}^4 (x_{Ai})^2 + q_B^2 \sum_{i=1}^4 (x_{Bi})^2 + q_{AB}^2 \sum_{i=1}^4 (x_{Ai} x_{Bi})^2 + 0 \\
 &= 4q_A^2 + 4q_B^2 + 4q_{AB}^2
 \end{aligned}$$

Example

Cache Size	Performance in MIPS	
	Memory Size	
	4M Bytes	16M Bytes
1K	15	45
2K	25	75

- Memory-cache study: $\bar{y} = \frac{1}{4}(15 + 55 + 25 + 75) = 40$

- Total variation $= \sum_{i=1}^4 (y_i - \bar{y})^2$

$$\begin{aligned}
 &= (25^2 + 15^2 + 15^2 + 35^2) \\
 &= 2100 \\
 &= 4 \times 20^2 + 4 \times 10^2 + 4 \times 5^2
 \end{aligned}$$

- Total variation = 2100
 Variation due to memory = 1600 (76%)
 Variation due to cache = 400 (19%)
 Variation due to interaction = 100 (5%)



Case Study: Interconnection Nets

- Memory interconnection networks:
 - Omega and Crossbar
- Memory reference patterns
 - Random (with uniform probability)
 - Matrix multiplication problem in which each process is doing a part of the multiplication



Case Study: Interconnection Nets

- Fixed factors:
 - Number of processors was fixed at 16
 - Queued requests were not buffered but blocked
 - Circuit switching instead of packet switching
 - Random arbitration instead of round robin
 - Infinite interleaving of memory => no memory bank contention

2² Design for Interconnection Networks

Factors Used in the Interconnection Network Study

Symbol	Factor	Level	
		-1	1
A	Type of the network	Crossbar	Omega
B	Address Pattern Used	Random	Matrix

		Response		
A	B	Throughput T	90% Transit N	Response R
-1	-1	0.0641	3	1.655
1	-1	0.4220	5	2.378
-1	1	0.7922	2	1.262
1	1	0.4717	4	2.190

17 October 2005

EEC686/785

Wenbing Zhao

Interpretation of Results

Parameter	Mean Estimate			Variation Explained		
	T	N	R	T	N	R
q_0	0.5725	3.5	1.871			
q_A	0.0595	-0.5	-0.145	17.2%	20%	10.9%
q_B	-0.1257	1.0	0.413	77.0%	80%	87.8%
q_{AB}	-0.0346	0.0	0.051	5.8%	0%	1.3%

- Average throughput = 0.5725
- More effective factor = B = reference pattern
=> The address patterns chosen are very different
- Reference pattern explains ± 0.1257 (77%) of variation
- Effect of network type = 0.0595
Omega networks = average + 0.0595
Crossbar networks = average - 0.0595
Difference between the two = 0.119
- Slight interaction (0.0346) between reference pattern and network type

17 October 2005

EEC686/785

Wenbing Zhao

General 2^k Factorial Designs

- k factors at two levels each
- 2^k experiments
- 2^k effects:
 - k main effects
 - $\binom{k}{2}$ two factor interactions
 - $\binom{k}{3}$ three factor interactions

2^k Design Example

- Three factors in designing a machine
 - Cache size
 - Memory size
 - Number of processors

Factor	Level -1	Level 1
A Memory Size	4MB	16MB
B Cache Size	1kB	2kB
C Number of Processors	1	2

2^k Design Example

Cache Size	4M Bytes		16M Bytes	
	1 Proc	2 Proc	1 Proc	2 Proc
1K Byte	14	46	22	58
2K Byte	10	50	34	86

<i>I</i>	<i>A</i>	<i>B</i>	<i>C</i>	<i>AB</i>	<i>AC</i>	<i>BC</i>	<i>ABC</i>	<i>y</i>
1	-1	-1	-1	1	1	1	-1	14
1	1	-1	-1	-1	-1	1	1	22
1	-1	1	-1	-1	1	-1	1	10
1	1	1	-1	1	-1	-1	-1	34
1	-1	-1	1	1	-1	-1	1	46
1	1	-1	1	-1	1	-1	-1	58
1	-1	1	1	-1	-1	1	-1	50
1	1	1	1	1	1	1	1	86
320	80	40	160	40	16	24	9	Total
40	10	5	20	5	2	3	1	Total/8

2^k Design Example - Analysis

$$\begin{aligned}
 SST &= 2^3(q_A^2 + q_B^2 + q_C^2 + q_{AB}^2 + q_{BC}^2 + q_{AC}^2 + q_{ABC}^2) \\
 &= 8(10^2 + 5^2 + 120^2 + 5^2 + 2^2 + 3^2 + 1^2) \\
 &= 800 + 200 + 3200 + 200 + 32 + 72 + 8 \\
 &= 18\% + 4\% + 71\% + 4\% + 1\% + 2\% + 0\% \\
 &= 100\%
 \end{aligned}$$

- Number of processors (C) is the most important factor