INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

Bell & Howell Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA

UMI[®]
800-521-0600

Estimating Resource Demands for Application Services

By

Xin Sun

**A thesis submitted to
the Faculty of Graduate Studies and Research
in partial fulfillment of
the requirements of the degree of
Master of Information and System Science**

**Ottawa-Carleton Institute for Electrical Engineering
Faculty of Engineering
Department of Systems and Computer Engineering
Carleton University
Ottawa, Ontario, Canada, K1S 5B6**

12 May 1999

© Copyright 1999, Xin Sun

National Library
of Canada

Acquisitions and
Bibliographic Services

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale
du Canada

Acquisitions et
services bibliographiques

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file *Votre référence*

Our file *Notre référence*

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-48495-5

Canada

The undersigned recommend to the Faculty of Graduate Studies and Research
acceptance of the thesis

Estimating Resource Demands for Application Services

submitted by Xin Sun, B.Sc., M.Sc.
in partial fulfillment of the requirements for
the degree of Master of Information and System Science

Dr. Jerome A. Rolia, Thesis Supervisor

Dr. Rafik A. Goubran, Chair, Department of Systems and Computer Engineering

Carleton University
27 May 1999

ABSTRACT

Recognizing and solving a distributed application's performance problems can require information about the resource demands of each of an application's services. However for most operating system platforms it is not possible to systematically measure the resource demands of application services directly. This information is necessary for the development of predicative performance models for tuning, load balancing, capacity planning, and system sizing.

In this thesis we present the results of experiments that illustrate the conditions under which three statistical techniques (ordinary least squares, the random coefficient model, and restricted least squares), instead of direct measurement, can be used to accurately correlate the use of application level services with resource demand information. Simulation experiments are used to evaluate the accuracy of the techniques and the relative importance of various experimental factors.

Acknowledgements

I would like to express my sincere appreciation to Professor Jerome Rolia for his wise guidance, patience and encouragement through my research. He spent a lot of time on either academic area or English for my thesis. This thesis wouldn't have been done so soon without his effort.

Also, I want to thank the other faculty members and students of the RADS group, especially Scratch and Peter, for their help and support. I am indebted to Vidar and Diwark in our group. Vidar provided the simulator, which is used in my research. Diwark helped me a lot set up and use software for my research.

I gratefully acknowledge the generous financial support from IBM Canada, Center for Advanced Studies and the partial funding from The Natural Science and Engineering Research Council.

CONTENTS

CHAPTER 1. INTRODUCTION	1
1.1.Challenges Measuring the Resource Demands of ApplicationServices	1
1.2. Goal of the Thesis	2
1.3. Thesis Organization	3
CHAPTER 2. LEAST SQUARES REGRESSION	5
2.1. Algorithm	5
2.1.1. The Multiple Linear Regression Model and Its Assumptions	5
2.1.2. Homoscedasticity and End Effects	7
2.1.3. Least Square Regression	8
2.1.4. Confidence Interval of Estimated Parameters	9
2.1.5. Analysis of Variance	11
2.1.6. Using Least Square Regression to Estimate Resource Demand	14
2.2. Results of Study	15
2.2.1. The Simulator	15
2.2.2. Case Study	18
2.2.3. Summary	22
CHAPTER 3. RANDOM COEFFICIENT MODEL REGRESSION	36
3.1. A Stochastic Parameter Regression Model	36
3.2. Kalman Filter	39
3.3. Random Coefficient Model	45
3.4. Results of Study	48
3.5. Summary	50
CHAPTER 4. MAXIMUM LIKELIHOOD ESTIMATION OF GENERALIZED REGRESSION MODEL	63
4.1. Generalized Regression Model	63
4.2. Maximum Likelihood Estimation	64
4.3. Results of Study	67
CHAPTER 5. RESTRICTED LEAST SQUARE REGRESSION	70
5.1. Algorithm	70
5.2. Results of Study	72

5.3. Summary	74
CHAPTER 6. SUMMARY AND FUTURE RESEARCH	89
6.1. Comparison of Three Statistic Technique	89
6.2. RCM and RLS Approaches for Real Systems	91
6.3. Summary for RCM	92
6.4. Future Research	93
APPENDIX A.	94
A.1. Package Diagram	94
A.2. Random Coefficient Model Regression Tool	96
A.3. Restricted Least Square Regression Tool	97
A.4. Maximum Likelihood Estimation Tool for Generalized Regression Model	98
REFERENCES	100

List of Tables

Table 2-1: Relative Regression Errors of OLS Model for Mean Service Times that Vary Linearly with a Range of 2, Duration 5 Seconds	24
Table 2-2: Relative Regression Errors of OLS Model for Mean Service Times that Vary Linearly with a Range of 10, Duration 5 Seconds	24
Table 2-3: Relative Regression Errors of OLS Model for Mean Service Times that Vary Linearly with a Range of 10, Duration 60 Seconds	25
Table 2-4: Relative Regression Errors of OLS Model for Mean Service Times that Vary Linearly with a Range of 10, Duration 300 Seconds	25
Table 2-5: Relative Regression Errors of OLS Model for Mean Service Times that Vary Geometrically with a Range of 10 and Equal Inter-arrival Times, Duration 5 Seconds	26
Table 2-6: Relative Regression Errors of OLS Model for Mean Service Times that Vary Geometrically with a Range of 10 and Equal Inter-arrival Times, Duration 60 Seconds	26
Table 2-7: Relative Regression Errors of OLS Model for Mean Service Times that Vary Geometrically with a Range of 10 and Equal Inter-arrival Times, Duration 300 Seconds	27
Table 2-8: Relative Regression Errors of OLS Model for Mean Service Times that Vary Geometrically with a Range of 10 and Increasing Inter-arrival Times, Duration 5 Seconds	27
Table 2-9: Relative Regression Errors of OLS Model for Mean Service Times that Vary Geometrically with a Range of 10 and Increasing Inter-arrival Times, Duration 60 Seconds	28
Table 2-10: Relative Regression Errors of OLS Model for Mean Service Times that Vary Geometrically with a Range of 10 and Increasing Inter-arrival Times, Duration 300 Seconds	28
Table 2-11: Relative Regression Errors of OLS Model for Different Numbers of Reporting Periods	29
Table 3-1: Relative Regression Errors of RCM Model for Mean Service Times that Vary Linearly with a Range of 2, Duration 5 Seconds	51
Table 3-2: Relative Regression Errors of RCM Model for Mean Service Times that Vary Linearly with a Range of 10, Duration 5 Seconds	51

Table 3-3: Relative Regression Errors of RCM Model for Mean Service Times that Vary Linearly with a Range of 10, Duration 60 Seconds	52
Table 3-4: Relative Regression Errors of RCM Model for Mean Service Times that Vary Linearly with a Range of 10, Duration 300 Seconds	52
Table 3-5: Relative Regression Errors of RCM Model for Mean Service Times that Vary Geometrically with a Range of 10 and Equal Inter-arrival Times, Duration 5 Seconds	53
Table 3-6: Relative Regression Errors of RCM Model for Mean Service Times that Vary Geometrically with a Range of 10 and Equal Inter-arrival Times, Duration 60 Seconds	53
Table 3-7: Relative Regression Errors of RCM Model for Mean Service Times that Vary Geometrically with a Range of 10 and Equal Inter-arrival Times, Duration 300 Seconds	54
Table 3-8: Relative Regression Errors of RCM Model for Mean Service Times that Vary Geometrically with a Range of 10 and Increasing Inter-arrival Times, Duration 5 Seconds	54
Table 3-9: Relative Regression Errors of RCM Model for Mean Service Times that Vary Geometrically with a Range of 10 and Increasing Inter-arrival Times, Duration 60 Seconds	55
Table 3-10: Relative Regression Errors of RCM Model for Mean Service Times that Vary Geometrically with a Range of 10 and Increasing Inter-arrival Times, Duration 300 Seconds	55
Table 3-11: Relative Regression Errors of RCM Model for Different Numbers of Reporting Periods	56
Table 4-1: Reported Errors for OLS and MLE Technique	68
Table 4-2: Reported Errors of MLE for Different Bucket Size	69
Table 5-1: Relative Regression Errors of RLS Model for Mean Service Times that Vary Linearly with a Range of 2, Duration 5 Seconds	75
Table 5-2: Relative Regression Errors of RLS Model for Mean Service Times that Vary Linearly with a Range of 10, Duration 5 Seconds	75
Table 5-3: Relative Regression Errors of RLS Model for Mean Service Times that Vary Linearly with a Range of 10, Duration 60 Seconds	76

Table 5-4: Relative Regression Errors of RLS Model for Mean Service Times that Vary Linearly with a Range of 10, Duration 300 Seconds	76
Table 5-5: Relative Regression Errors of RLS Model for Mean Service Times that Vary Geometrically with a Range of 10 and Equal Inter-arrival Times, Duration 5 Seconds	77
Table 5-6: Relative Regression Errors of RLS Model for Mean Service Times that Vary Geometrically with a Range of 10 and Equal Inter-arrival Times, Duration 60 Seconds	77
Table 5-7: Relative Regression Errors of RLS Model for Mean Service Times that Vary Geometrically with a Range of 10 and Equal Inter-arrival Times, Duration 300 Seconds	78
Table 5-8: Relative Regression Errors of RLS Model for Mean Service Times that Vary Geometrically with a Range of 10 and Increasing Inter-arrival Times, Duration 5 Seconds	78
Table 5-9: Relative Regression Errors of RLS Model for Mean Service Times that Vary Geometrically with a Range of 10 and Increasing Inter-arrival Times, Duration 60 Seconds	79
Table 5-10: Relative Regression Errors of RLS Model for Mean Service Times that Vary Geometrically with a Range of 10 and Increasing Inter-arrival Times, Duration 300 Seconds	79
Table 5-11: Relative Regression Errors of RLS Model for Different Numbers of Reporting Periods	80
Table 5-12: Relative Regression Errors of RLS Model with Ratio Varying 10% for Mean Service Times that Vary Geometrically, Duration 300 Seconds	81
Table 5-13: Relative Regression Errors of RLS Model with Ratio Varying 25% for Mean Service Times that Vary Geometrically, Duration 300 Seconds	81
Table 6-1: Performance of OLS, RLS and RCM Regression Tools, Time in ms.	90
Table 6-2: Relative Regression Errors of RCM Model for Three Services with Each Sharing the Same Mixed Distributions	92

List of Figures

Figure 2-1: The Structure of the Simulation Model	16
Figure 2-2: Relative Regression Errors of OLS Model for Mean Service Times that Vary Linearly with a Range of 2, Duration 5 Seconds	30
Figure 2-3: Relative Regression Errors of OLS Model for Mean Service Times that Vary Linearly with a Range of 10, Duration 5 Seconds	30
Figure 2-4: Relative Regression Errors of OLS Model for Mean Service Times that Vary Linearly with a Range of 10, Duration 60 Seconds	31
Figure 2-5: Relative Regression Errors of OLS Model for Mean Service Times that Vary Linearly with a Range of 10, Duration 300 Seconds	31
Figure 2-6: Relative Regression Errors of OLS Model for Mean Service Times that Vary Geometrically with a Range of 10 and Equal Inter-arrival Times, Duration 5 Seconds	32
Figure 2-7: Relative Regression Errors of OLS Model for Mean Service Times that Vary Geometrically with a Range of 10 and Equal Inter-arrival Times, Duration 60 Seconds	32
Figure 2-8: Relative Regression Errors of OLS Model for Mean Service Times that Vary Geometrically with a Range of 10 and Equal Inter-arrival Times, Duration 300 Seconds	33
Figure 2-9: Relative Regression Errors of OLS Model for Mean Service Times that Vary Geometrically with a Range of 10 and Increasing Inter-arrival Times, Duration 5 Seconds	33
Figure 2-10: Relative Regression Errors of OLS Model for Mean Service Times that Vary Geometrically with a Range of 10 and Increasing Inter-arrival Times, Duration 60 Seconds	34
Figure 2-11: Relative Regression Errors of OLS Model for Mean Service Times that Vary Geometrically with a Range of 10 and Increasing Inter-arrival Times, Duration 300 Seconds	34
Figure 2-12: Relative Regression Errors of OLS Model for Different Numbers of Reporting Periods	35
Figure 3-1: Relative Regression Errors of RCM Model for Mean Service Times that Vary Linearly with a Range of 2, Duration 5 Seconds	57

Figure 3-2: Relative Regression Errors of RCM Model for Mean Service Times that Vary Linearly with a Range of 10, Duration 5 Seconds	57
Figure 3-3: Relative Regression Errors of RCM Model for Mean Service Times that Vary Linearly with a Range of 10, Duration 60 Seconds	58
Figure 3-4: Relative Regression Errors of RCM Model for Mean Service Times that Vary Linearly with a Range of 10, Duration 300 Seconds	58
Figure 3-5: Relative Regression Errors of RCM Model for Mean Service Times that Vary Geometrically with a Range of 10 and Equal Inter-arrival Times, Duration 5 Seconds	59
Figure 3-6: Relative Regression Errors of RCM Model for Mean Service Times that Vary Geometrically with a Range of 10 and Equal Inter-arrival Times, Duration 60 Seconds	59
Figure 3-7: Relative Regression Errors of RCM Model for Mean Service Times that Vary Geometrically with a Range of 10 and Equal Inter-arrival Times, Duration 300 Seconds	60
Figure 3-8: Relative Regression Errors of RCM Model for Mean Service Times that Vary Geometrically with a Range of 10 and Increasing Inter-arrival Times, Duration 5 Seconds	60
Figure 3-9: Relative Regression Errors of RCM Model for Mean Service Times that Vary Geometrically with a Range of 10 and Increasing Inter-arrival Times, Duration 60 Seconds	61
Figure 3-10: Relative Regression Errors of RCM Model for Mean Service Times that Vary Geometrically with a Range of 10 and Increasing Inter-arrival Times, Duration 300 Seconds	61
Figure 3-11: Relative Regression Errors of RCM Model for Different Numbers of Reporting Periods	62
Figure 5-1: Relative Regression Errors of RLS Model for Mean Service Times that Vary Linearly with a Range of 2, Duration 5 Seconds	82
Figure 5-2: Relative Regression Errors of RLS Model for Mean Service Times that Vary Linearly with a Range of 10, Duration 5 Seconds	82
Figure 5-3: Relative Regression Errors of RLS Model for Mean Service Times that Vary Linearly with a Range of 10, Duration 60 Seconds	83
Table 5-4: Relative Regression Errors of RLS Model for Mean Service Times that Vary Linearly with a Range of 10, Duration 300 Seconds	83

Figure 5-5: Relative Regression Errors of RLS Model for Mean Service Times that Vary Geometrically with a Range of 10 and Equal Inter-arrival Times, Duration 5 Seconds	84
Figure 5-6: Relative Regression Errors of RLS Model for Mean Service Times that Vary Geometrically with a Range of 10 and Equal Inter-arrival Times, Duration 60 Seconds	84
Figure 5-7: Relative Regression Errors of RLS Model for Mean Service Times that Vary Geometrically with a Range of 10 and Equal Inter-arrival Times, Duration 300 Seconds	85
Figure 5-8: Relative Regression Errors of RLS Model for Mean Service Times that Vary Geometrically with a Range of 10 and Increasing Inter-arrival Times, Duration 5 Seconds	85
Figure 5-9: Relative Regression Errors of RLS Model for Mean Service Times that Vary Geometrically with a Range of 10 and Increasing Inter-arrival Times, Duration 60 Seconds	86
Figure 5-10: Relative Regression Errors of RLS Model for Mean Service Times that Vary Geometrically with a Range of 10 and Increasing Inter-arrival Times, Duration 300 Seconds	86
Figure 5-11: Relative Regression Errors of RLS Model for Different Numbers of Reporting Periods	87
Figure 5-12: Relative Regression Errors of RLS Model with Ratio Varying 10% for Mean Service Times that Vary Geometrically, Duration 300 Seconds	88
Figure 5-13: Relative Regression Errors of RLS Model with Ratio Varying 25% for Mean Service Times that Vary Geometrically, Duration 300 Seconds	88
Figure A-1: Package Diagram for Three Statistic Tools	95
Figure A-2: Class Diagram of RCM Tool	96
Figure A-3: Package Diagram of Maximum Likelihood Estimation Tool for Generalized Regression Model	99
Figure A-4: Class Diagram of MLE Tool	99

Chapter 1. Introduction

This chapter provides the motivation of the thesis which is to estimate resource demands for application services. It presents the goal of the thesis, describes the approach used and shows the organization of the thesis.

1.1. Challenges Measuring the Resource Demands of Application Services

Distributed applications consist of many software processes distributed over a network, cooperating to accomplish some overall application goals. When properly designed and managed they offer advantages in price-performance, availability, and resource sharing.

Information about the resource demands of each application service is necessary for the development of predicative performance models for tuning, load balancing, capacity planning, and system sizing. However for most platforms it is usually not possible to correlate application-level data with its corresponding resource demand data directly. The reason is that operating systems gather such data at the per-thread, per-process, or per-node levels of abstraction and have no knowledge of application-dependent services. The reasons given below explain why such measurements are so difficult for operating system vendors to provide [1]:

- CPU time is often collected using sampling techniques and is reported at the process or thread level. However, the process or thread may multiplex the execution of many service instances (due to context switching at the service instance level) during the execution of any service instance.
- Micro-kernel architectures make it hard to propagate enough information through parts of a system to enable accounting for resources managed outside the scope of the kernel.

- I/O is often bound to the virtual memory management system--physical I/O may not complete until the virtual memory pages associated with a written file are needed for some other purpose, this may be well after a process has terminated.
- I/O systems are complex--they may read ahead in predication of following read requests, and may delay writes to avoid unnecessary interactions with a physical disk. The impact of the I/O subsystem optimization techniques is load dependent. This can be a source of significant variation in physical disk demands.

Therefore, an operating system can only provide direct measurement of CPU demands at the levels of abstraction it has access to. For instance, Windows NT 4.0 provides physical I/O measures only at a per-disk level of abstraction, not on a per process or thread level. UNIX operating systems try to estimate the proportion of measured physical I/O to be assigned to each process. But, the operating systems sometimes provide poor answers.

Obviously, the direct measurement of resource demands for each application service is difficult and challenging. An alternative is to use statistical tools to estimate resource demands of each application service. We assume that an application level performance monitor periodically reports counts of the number of invocations for each application service and a total resource demand measure. These values are the inputs to the regression techniques.

1.2. Goal of the Thesis

Using statistical tools to estimate resource demand is not new. Bard [2] in 1978 used regression technique to estimate the CPU demand of system service routines that could not be measured individually.

In this thesis, statistical methods are used to estimate resource demands of application services. The following statistical techniques Least Square Regression (OLS), Random

Coefficient Model Regression (RCM), Restricted Least Square Regression (RLS) and Maximum Likelihood Estimation of Generalized Regression Model (MLE) are studied. We present the results of experiments that demonstrate the conditions under which three statistical techniques (OLS, RLS and RCM) can be used to accurately correlate application services with operating system resource demand measures.

Most mathematical references within this thesis are directly from [3, 6]. Assumptions that arise during the presentation of their methods are explained with the context of the problem addressed in this thesis. The main contributions of the thesis are finding the appropriate regression techniques to use, implementing them in Java, and evaluating their effectiveness.

1.3. Thesis Organization

There are in total six chapters and an appendix in this thesis. Chapter 2 provides the algorithm of least square regression, namely, OLS (Ordinary Least Squares). It also gives introduction of the simulator that was implemented by Vidar Vetland. Finally, experimental results that demonstrate the effects of different factors are given.

Chapter 3 introduces the stochastic parameter regression model. It focuses on the Kalman filter algorithm and random coefficient model. Experimental results are given.

Chapter 4 presents generalized regression model and the algorithm of maximum likelihood estimation for this model. Only a few experiments were performed. They are described.

Chapter 5 shows algorithm of restricted least square regression. Experiments are processed to explore factors that affect the accuracy of RLS tool.

Chapter 6 gives a summary of three statistical tools that are OLS, RLS and RCM. Their performance, advantages and disadvantages are presented. It also describes how RLS

and RCM can be used with real systems. Future research to improve the accuracy of these techniques is proposed.

Appendix A shows package diagram and class diagram of the Java software developed for this thesis. It implements the statistical tools, namely, OLS, RLS, RCM and MLE. These are a major contribution of the thesis.

Chapter 2. Least Squares Regression

This chapter introduces the multiple linear regression model, the least squares regression method to estimate parameters of this linear model and the possibility of using the least square regressions method to estimate resource demands of each application services. A simulator that can mimic a number of application services on a single-threaded server is described. Finally, using the least squares regression to estimate resource demands of each application services imitated by the simulator is explored.

2.1. Algorithm [3]

This section describes the multiple linear regression model, its assumptions and the least squares regression method to estimate coefficients of this model. We can treat resource demands of each application service as a coefficient of the multiple linear regression model and estimate the coefficients by the least square regression method. However in general, the resource demands of each application service are not deterministic. This violates one of the assumptions for the multiple linear regression model. So this model is not well adapted to our problem and may not give accurate result in some cases.

2.1.1. The Multiple Linear Regression Model and Its Assumptions

The multiple linear regression model is generally given as:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \varepsilon_i$$

In which β is a deterministic but unknown vector, x and y represent observation and ε is the error. Let the column vector x_{nk} be the n observations on variable x_k and we can put these data in an $n \times K$ matrix X . Finally let the column vector y be the n observations, $y_1, \dots, y_n$. We can rewrite the model as

$$y = x_1 \beta_1 + \dots x_K \beta_K + \varepsilon$$

or

$$y = X\beta + \varepsilon$$

In this equation, y is the regressand and X is the regressor. The x values give counts for invocations of application services, the y values give measured total resource demands. Each report from a performance monitor gives one row's worth of data. Assumptions about this linear model is given below:

1. Zero mean of the disturbance.

$$E[\varepsilon_i] = 0 \text{ or}$$

$$E[\varepsilon] = \begin{bmatrix} E[\varepsilon_1] \\ E[\varepsilon_2] \\ \vdots \\ E[\varepsilon_n] \end{bmatrix} = 0$$

which means that:

$$E[y] = X\beta$$

Disturbance describes the case that system behaves differently for different reporting periods. For example that some uninstrumented background processing is present and changing its load over reporting periods. The simulator does not include such background loads.

2. Homoscedasticity and uncorrelatedness across observations.

Homoscedasticity means that error per observation (monitoring report) is independent of the value of the regressors. This assumption is not likely to be met for our problem. Future work includes exploiting transformation on the regressand to reduce the impact of this violation. In general reporting periods should be long enough so that subsequent reports are not correlated.

$$E[\varepsilon\varepsilon'] = \begin{bmatrix} E[\varepsilon_1\varepsilon_1] & E[\varepsilon_1\varepsilon_2] & E[\varepsilon_1\varepsilon_3] & \dots & E[\varepsilon_1\varepsilon_n] \\ E[\varepsilon_2\varepsilon_1] & E[\varepsilon_2\varepsilon_2] & E[\varepsilon_2\varepsilon_3] & \dots & E[\varepsilon_2\varepsilon_n] \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ E[\varepsilon_n\varepsilon_1] & E[\varepsilon_n\varepsilon_2] & E[\varepsilon_n\varepsilon_3] & \dots & E[\varepsilon_n\varepsilon_n] \end{bmatrix}$$

$$= \begin{bmatrix} \sigma^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma^2 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \sigma^2 \end{bmatrix}$$

$$E[\varepsilon\varepsilon'] = \sigma^2 \mathbf{I}$$

3. The disturbances are independent and identically normal distributed:

$$\varepsilon \sim N[0, \sigma^2 \mathbf{I}]$$

4. There are no linear relationships among the variables which imply that $\mathbf{X}$ has full column rank.

$$\text{Rank}(\mathbf{X}) = K$$

This means that the columns of $\mathbf{X}$ are linearly independent and we must have at least K observations. In our case, service invocations are linearly independent and there are enough observations for the regression. If two columns are linearly dependent then they must be merged. The regression reports an average service time for the combined column.

2.1.2. Homoscedasticity and End Effects

Homoscedasticity means that standard deviation of errors is constant. In another words, it means that the distribution of the errors does not depend on the predictor variables. In our case, the predictor variables are numbers of invocations of resource demands.

In last section, assumption 2, *homoscedasticity* is not likely to be met for our problem. It is because reporting periods are of finite duration so that successive reports are correlated. Furthermore each predictor variable may have a different impact on error.

To explain this, first we need to propose a new terminology called “*end-effect*.” As we will discuss in section 2.1.6, the resource demands are reported by the operating system during a specific period. Some application services may start in one period but end in another one. In this case, the number of invocation is counted in only one of the periods but its resource demands contribute to both periods, which leads to the error of results received from linear regression. Different services can cause different means and distributions of resource demands so they impact error differently.

For instance, there are three applications A, B and C in the graph above. Application A starts in reporting period 1 but end in 2. The invocation of application A is counted in reporting period 1 but its resource demands also contribute to reporting period 2. It is the same case for application C. If the number of invocation in reporting period 1 is zero, the estimation of resource demands for applications in reporting period 2 will be accurate because there is no end effect from applications in reporting period 1. Otherwise, there is error introduced by end effects of applications in reporting period 1.

From the discussion above, it is obvious that errors for the estimation of resource demands in a reporting period depend of the number and type of invocations within the reporting period. This violates the definition of *homoscedasticity*.

2.1.3. Least Squares Regression

The Least Squares Regression is the most common method to estimate parameters of linear models and the idea is to minimize the sum of squared errors. The sum of squared residuals can be expressed as:

$$\varepsilon'\varepsilon = (y - X\beta)'(y - X\beta)$$

where β is the arbitrary choice for the coefficient vector. All we need to do is to choose the β which can minimize $\varepsilon'\varepsilon$.

Let $S(\beta) = \varepsilon'\varepsilon = (y - X\beta)'(y - X\beta)$ and expanding this equation, we get:

$$\begin{aligned} S(\beta) &= y'y - \beta'X'y - y'X\beta + \beta'X'X\beta \\ &= y'y - 2\beta'X'y + \beta'X'X\beta \end{aligned}$$

The necessary condition for a minimum is:

$$\frac{\partial S(\beta)}{\partial \beta} = 2X'y + 2X'X\beta = 0.$$

Let b be the solution and b satisfies the least square normal equations

$$X'Xb = X'y$$

From the assumption 4 which is “ X has full column rank,” we can infer that the inverse of $(X'X)$ exists. Then the solution is:

$$b = (X'X)^{-1}X'y$$

The least square estimators are the best linear unbiased estimator of β under the classical assumptions. We infer this from **Gauss-Markov** theorem, which is given below:

“For any vector of constants, w , the minimum variance linear unbiased estimator of $w'\beta$ in the classical regression model is $w'b$, where b is the least squares estimator.”

2.1.4. Confidence Interval of Estimated Parameters

As we discussed before, $b = (X'X)^{-1}X'y$ and since $y = X\beta + \varepsilon$, we can get a new equation: $b = (X'X)^{-1}X'(X\beta + \varepsilon) = \beta + (X'X)^{-1}X'\varepsilon$. From this equation, we can see that b is a linear function of the normally distributed disturbance, ε and state as:

$$b \sim N[\beta, \sigma^2(X'X)^{-1}].$$

It is a multivariate normal distribution and each element of $\mathbf{b}$ is normally distributed:

$$b_i \sim N[\beta_i, \sigma^2(\mathbf{X}'\mathbf{X})_{ii}^{-1}].$$

From above, we can get a equation:

$$z_i = \frac{(b_i - \beta_i)}{\sigma^2 S^{ii}} \quad (2-1)$$

where S^{ii} is the i th diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$. Then statistic inference about β_i can be based on z_i . But σ^2 can not be known so this strategy doesn't work.

Another strategy is developed by using residuals. We know that the least square residuals are:

$$\begin{aligned} \mathbf{e} &= \mathbf{M}\mathbf{y} = \mathbf{M}[\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}] \\ &= \mathbf{M}\boldsymbol{\varepsilon} \end{aligned}$$

where $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and the above results are inferred by the condition of $\mathbf{M}\mathbf{X}=\mathbf{0}$. Using this result, the sum of squared residuals which is used to develop an estimator of σ^2 is :

$$\mathbf{e}'\mathbf{e} = \boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon}$$

The expected value of the sum of squared residuals is:

$$\begin{aligned} E[\mathbf{e}'\mathbf{e}] &= E[\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon}] \\ &= E[\text{tr}(\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon})] = E[\text{tr}(\mathbf{M}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')] \\ &= \text{tr}(\mathbf{M}E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}']) = \text{tr}(\mathbf{M}\sigma^2\mathbf{I}) \\ &= \sigma^2 \text{tr}(\mathbf{M}). \end{aligned}$$

The trace of $\mathbf{M}$ is:

$$\begin{aligned} \text{tr}[\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] &= \text{tr}(\mathbf{I}_n) - \text{tr}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}) \\ &= \text{tr}(\mathbf{I}_n) - \text{tr}(\mathbf{I}_k) = n - K. \end{aligned}$$

Finally,

$$E[\mathbf{e}'\mathbf{e}] = (n - K)\sigma^2$$

So an unbiased estimator of σ^2 is

$$s^2 = \frac{\mathbf{e}\mathbf{e}'}{n - K}$$

From s^2 , we can get the standard error of the regression s which, obviously is the square root of s^2 . Using s^2 instead of σ^2 , we can develop a new statistic to use rather than (2-1) which is:

$$\frac{(n-K)s^2}{\sigma^2} = \frac{\mathbf{e}'\mathbf{e}}{\sigma^2} = \left(\frac{\mathbf{e}}{\sigma}\right)' \mathbf{M} \left(\frac{\mathbf{e}}{\sigma}\right) \quad (2-2)$$

It has a chi-squared distribution with $n-K$ (rank of $\mathbf{M}$) degrees of freedom. The chi-squared variable in (2-2) is independent of the standard normal variable z_i in (2-1). According to the definition of t -distribution, we get:

$$t_i = \frac{\left[\frac{(b_i - \beta_i)}{(\sigma^2 S'')^{1/2}} \right]}{\left\{ \frac{\left[\frac{(n-K)s^2}{\sigma^2} \right]}{(n-k)} \right\}^{1/2}}$$

$$= \frac{(b_i - \beta_i)}{(s^2 S'')^{1/2}}$$

So, the ratio above has a t distribution with $(n-K)$ degrees of freedom. If we want to calculate $(1-\alpha)$ confidence interval of parameter b_i , we need to check $(1-\alpha/2)$ -quantile for a t -variate with $(n-K)$ degrees of freedom and follow the equation:

$$b_i = \beta_i + (s^2 S'')^{1/2} t_{[1-\alpha/2; n-K]}$$

2.1.5. Analysis of Variance

The error variance in the regression is equal to the variance of the response and the equation is given below:

$$\text{Error} = \varepsilon_i = \text{observed response} - \text{predicted response}$$

This quantity is called the total sum of squares (SST) which is expressed by the following equation:

$$SST = SSY - SS0$$

where SSY is the sum of squares of y and $SS0$ is the sum of squares of mean of y . In the other hand,

$$SST = SSR + SSE$$

$$SSE = \mathbf{y}^T \mathbf{y} - \mathbf{b}^T \mathbf{X}^T \mathbf{y}$$

So, the total variation SST consists two parts which are SSR and SSE . The SSR indicates the variation explained by the regression model and the SSE indicates the variation not explained by the regression. The fraction which is SSR to SST is called the coefficient of determination, R^2 , which determines the goodness of the regression. The formula of the coefficient of determination is given below:

$$R^2 = \frac{SSR}{SST} = \frac{(SST - SSE)}{SST}$$

The higher the R^2 , the better the regression is. Theoretically, when SSE is zero, which means that all observed value equals to the value expected by the regression model, the coefficient of determination is 1. On the other hand, when the regression model is too poor to reduce any variance, SSE is equal to SST and the coefficient of determination is 0.

As we discussed, the variation is divided into two parts: the explained part and the unexplained part. Then we need to know how much explained variation is good. A statistic test called Analysis of Variance is given. The analysis generally tests the hypothesis that the SSR is less than or equal to the SSE .

Each of the sum of squares has a degree freedom corresponding to the number of independent values with which the sum of squares is obtained. Various sum of squares and their associated degree freedom are as follow:

$$SST = SSY - SS0 = SSR + SSE$$

$$n - 1 = n - 1 = k + (n - k - 1)$$

SSY has n degrees of freedom since n observations can be independently chosen. The SSE has $n-k-1$ degrees of freedom because it is received by calculating $k+1$ regression parameters from the data. The SST has $n-1$ degrees of freedom since it is obtained after the mean of y is calculated from the data. The SSR has the remaining k degrees of freedom.

It is assumed that errors are independent, normally distributed and identically distributed. This infers that the y 's are also normally distributed because the x 's are nonstochastic. According to the definition of chi-square distribution, the sum of squares of normal variates has a chi-square distribution. Therefore, each sum of squares given above has a chi-square distribution with its degrees of freedom.

If there are two sum of squares SS_i and SS_j with v_i and v_j freedom, according to the definition of F -distribution, the ratio $(SS_i/v_i)/(SS_j/v_j)$ has an F distribution with v_i numerator degrees of freedom and v_j denominator degrees of freedom. The hypothesis that the SS_i is less than or equal to SS_j is rejected at the α significance level if the ratio is greater than the $1-\alpha$ quintile of the F -variate. Thus, we can compare the computed ratio with $F_{[1-\alpha; v_i, v_j]}$ found in the table of F quantiles and if the computed ratio is more than the one found in the table, the sum of squares is considered significantly different. We call the procedure F -test.

In our case, we need to check if the SSR is significantly higher than the SSE by computing the ratio $(SSR/v_R)/(SSE/v_e)$, where v_R and v_e are degrees of freedom for the SSR and SSE . We call (SSR/v_R) and (SSE/v_e) the mean square of the regression (MSR) and the mean square of the error (MSE), respectively.

$$MSR = \frac{SSR}{k}$$

$$MSE = \frac{SSE}{n - k - 1}$$

By the discussion before, the ratio MSR/MSE has an $F_{[k, n-k-1]}$ distribution, that is, an F distribution with k numerator degrees of freedom and $n-k-1$ denominator degrees of freedom. If the computed ratio is greater than the value found in the F -table, the predictor variables are assumed to explain a significant fraction of the response variation.

Also, the F -test is equivalent to testing the null hypothesis that y does not depend upon any x_j , that is

$$b_1 = b_2 = \dots = b_k = 0$$

against an alternate hypothesis that y depends upon at least one x_j , and thus, at least one $b_j \neq 0$.

2.1.6. Using Least Squares Regression to Estimate Resource Demand

As we mentioned in the introduction, using a statistical approach to estimate the resource demands in computer application service is not new. Bard [2] used regression to estimate the CPU time incurred by MVS operating system supervisory calls. [1] used regression to correlate Application Response Measurement (ARM) transactions with physical resource demands.

The model is given below:

In this model, there are m reporting periods worth of data from a performance monitor, which are from 1 to m . The regressand variables r_j are physical resource demands, such as CPU time or number of physical input or output operations, which were reported by the operating system during the j reporting period. The control variables $i_{j,k}$ for $k=1 \dots n$ are the number of times each of n application services is invoked during reporting period j . This regression model gives the solution $c_1 \dots c_n$ which are the estimates for the mean service times of application services from 1 to n .

2.2. Results of Study

This section introduces the structure of the simulator and parameters by which to describe simulations. In section 2.2.2, the Ordinary Least Squares regression is used to estimate resource demands for varied simulations and its ability is appraised by reported relative errors.

2.2.1. The Simulator

Vidar Vetland implemented a simulator using C++ on a Unix machine. The structure of the simulation model is given in Figure 2-1 [1]. From the figure given, we can see that there is one independent job generator for each application service. The jobs generated are put in the job queue and finally go to a single-threaded software server, which offers a number of application services. The reason why a single-threaded software server is chosen is that it can provides results with lower error than multi-threaded software server does. A high degree of multi-threaded server will increase the end-effects because many application services could be in progress when a transition between reporting periods happens.

Three kinds of parameters get involved in the simulation. Those are system description parameters, system load parameters and control parameters. *System description parameters* include the number of application services on the software server, the mean service time for each application services and the distribution for service times of application services. *System load parameters* include pattern of inter-arrival means,

distribution of inter-arrival times and the server utilization. *System control parameters* are the reporting period duration, the number of reporting periods and the number of replications.

Figure 2-1: The Structure of the Simulation Model

The model can produce the simulation with up to nine different factors. Because there are different number of services for each simulation, parameters are expressed for services systematically using patterns. A simulation needs a mean service time, pattern of service times, distribution of service time and a total server usage, from which the patterns can determine parameters for the remaining application services. To illustrate the question clearly, nine parameters [4] and their representations are given below:

- o Number of application services on the software server.
- l Length of the sample period

- p Number of samples to collect
- j Pattern of inter-arrival means
- k Distribution of inter-arrival means
- s Mean service time
- t Pattern of service time
- u Distribution of service times
- r Replication times.

For the distribution of service times, constant, normal and exponential are considered in our experiment. Constant means the service time is deterministic. Normal means that the service time abides the normal distribution, for example, normal (0.05) indicates that normal variates are generated with a standard deviation of $\pm 5\%$ of the mean service time. Negative values are discarded and redrawn. Exponential means that the service time follows the exponential distribution.

For the pattern of service time, linear and geometric patterns are used in our experiments. The series of service times from a linear progression for the linear pattern and a geometric progression for the geometric pattern.

From a pattern for arrival rates and a specified total server utilization level, we can get mean inter-arrival times for each application. Here is an example to produce a simulation:

```
regsim -o 5 -p 50 -l 1000 -j "geometric(0.6)" -k exp -s 3 -t "geometric(1.8)" -u
constant -U 0.75 -r 100
```

It specifies a five operations with fifty reporting periods and the length of the sample period is 1000 ms. The pattern of inter-arrival means is "geometric (0.6)" and the distribution of inter-arrival times is exponential. The initial of mean service time is three ms, the pattern of service times is "geometric (1.8)" and the distribution of service times

is constant. The server utilization is 0.75 and this simulation replicates the experiment 100 times.

2.2.2. Case Study

This section evaluates the ability of OLS (Ordinary Least Squares) technique to correlate process level resource demands with usage statistics for application services.

As mentioned at the beginning of Section 2.1, the resource demands of each application service is not deterministic. Therefore, in our research, we focus on varying service time distributions. Service time distribution can be normal, exponential, and constant. Normal and exponential violate the assumptions of the Ordinary Linear Regression. So the coefficient of determination R^2 can not be used as a measured of error. A new error measure formula is proposed and given as:

$$E_{rel} = \frac{\sum_e \frac{|S_{est,e} - S_{real,e}|}{S_{real,e}} \bullet i_e \bullet S_{real,e}}{\sum_e i_e \bullet S_{real,e}}$$

where $S_{est,e}$ is the estimate of mean service time, $S_{real,e}$ is the real mean service time and i_e is the measured number of invocations of the application service.

In this formula, the services responsible for most of the server utilization are weighted most heavily. The difference between estimated and real service time means is weighted by the simulated application service utilization, which is the product of number of invocations and the real mean service time.

We use the simulator to explore a broad space of factors and levels. The purpose is to discover how sensitive the results of the regression are to the patterns for service and invocation rates. A total eleven cases were considered. Eleven tables and eleven figures that correspond to the tables are given to illustrate the experimental results.

The uppermost result matrix in tables gives the basic parameter settings for the experiments. In each case we consider 3, 5, and 10 services. Each simulation run is for 50 reporting periods, each with duration of 5, 60 or 300 seconds. There are 30 replications of each run to help achieve statistical confidence for the results. The server utilization is chosen to be 75%.

The lower part of the tables are the average reported error measures and their 95% confidence intervals. In most case, when the number of replication is 30, we use normal distribution to get the confidence interval. However, in some cases, the number of replications is less than 30 because we omit all OLS regression results that predicted negative coefficients, for these cases we use t-distribution to get the confidence interval. Since the number of replication drops, the confidence interval in this case is worse than the former case. Generally speaking, the 95% confidence intervals of the average reported error measures are less than ± 20 percent of their reported mean values. However, in some cases (number of replication is less than 30) they are as wide as ± 33 percent. Fortunately these intervals are adequate for the comparison presented in this thesis.

We pose the following questions [5]:

1. How does the statistical distribution of service times affect the accuracy of the regression techniques?
2. How do the pattern of mean service times and the range of mean service times across different application services affect the accuracy of the techniques?
3. How do differing arrival rates for different services affect the accuracy of the techniques?
4. How does the choice of reporting period and the number of periods affect the accuracy of the technique?

We discuss these questions by the experimental results in the following subsections.

2.2.2.1 Effect of Service Time Distributions

We consider constant (deterministic), normal and exponential service time distributions from Table 2-1 to Table 2-11. The error measures of the tables show that service time distribution has a large impact of the OLS technique. The reported error measures increase in the order of Constant, Normal (0.05), Normal (0.15), Normal (0.25), Normal (0.35), and Exponential service time distributions. It is because the multiple linear regression techniques assume that the regressed variables, namely service times, are constant. Any deviations of mean service demand for an application service per reporting period from the expected mean is regarded as measurement error by the regression. As the service time variance increases, the extent of violation of the assumption for multiple linear model increases, which leads to worse accuracy of the regression. However, even in the constant service time case, there is error. This can be attributed to problems with end-effects. As the number of application services increases and service time variation increases, the accuracy of the OLS technique diminishes.

2.2.2.2 Effect of Mean Service Time Pattern And the Range of Mean Service Times Across Different Application Services

Linear Service Time Pattern is used for Tables 2-2 and 2-3, while the Geometric Service Time Pattern is used for Tables 2-5 to 2-7. All the other parameter settings including a range of 10 and equal inter-arrival rates are all the same. To make this more understandable, consider this demonstration. For Geometric Service Time Pattern, the mean service times for the 3 service case are 100, 316, and 1000 ms. While for Linear Service Time Pattern, the mean service times for the 3 services are 100, 550, and 1000 ms. From the results given in the tables, we can see that for the geometric case a subset of the service times is closer together than in the linear case. And we also can see that the mean service time pattern doesn't have any significant impact on the accuracy of the OLS technique. The confidence intervals for the errors reported in Linear Service Time Pattern cases and Geometric Service Time Pattern cases overlap. It can be explained that

using different mean service time pattern doesn't violate any assumptions of the multiple linear regression model and certainly doesn't bring any extra error for the regression.

The basic parameter settings are the same between Table 2-1 and Table 2-2 but the range of mean service times across different application services that is 2 in Table 2-1 and 10 in Table 2-2. For a range of 10 and 3 services the mean service times are 100, 550, and 1000 ms respectively. Compared with a range of 2 and 3-service case, the mean service times are 100, 150, and 200 ms respectively. It is clear from these tables that the errors reported increase dramatically as the range changed from 2 to 10. So the Range of Mean Service Times across different application services does appear to have a significant impact on the accuracy of the OLS technique. The more the range, the more the errors reported. However the correct reason is that the service times of 1000 ms are similar to the reporting period duration which was only 5000 ms. This causes significant end effects.

2.2.2.3. Effect of Arrival Rate Pattern

The parameter settings are the same between the tables from 2-5 to 2-7 and the tables from 2-8 to 2-10 except for the arrival rate pattern. The tables from 2-5 to 2-7 use the equal inter-arrival times pattern, while the tables from 2-8 to 2-10 use the increasing inter-arrival times pattern. The equal inter-arrival times pattern means that inter-arrival times are the same for each application services. The increasing-arrival times pattern indicates that inter-arrival times are different for each application service and the services with the larger service times are visited least frequently.

There is no statistically significant change in errors with respect to the previous case with equal mean inter-arrival times for all application services. It can be explained by this way. The inter-arrival times pattern doesn't effect any assumptions of the regression techniques, so doesn't produce any extra error for the regression.

2.2.2.4. Effect of Reporting Period Duration and the Number of Reporting Periods

Each of the tables from 2-2 to 2-4 has the exact same parameter setting but the reporting duration. So are the tables from 2-5 to 2-7 and the ones from 2-8 to 2-10. The reporting duration is 5, 60 and 300 seconds respectively.

As the duration increases, the accuracy of the OLS technique gets improved and the extent of improvement decreases. But we have to notice that the number of reporting period is fixed in these cases. In another word, as the duration increases, the total measurement time increases. This is quite easily to be explained. It is because of end – effects that we discussed in Section 2.2.1. To increase the duration can decrease the end-effects. However, to some point, the end-effects diminish so to increase the duration in this situation has no help in the accuracy of the OLS technique.

Right now, let us consider the case, that is, the total measurement time is fixed. In another word, if we increase the duration, the number of reporting period will decrease. Table 2-11 gives the result of this experiment and Figure 2-12 is corresponding to this table. The total measurement time is fixed to 150 minutes (9000 seconds). The time is divided across 13...150 reporting periods. These give measurement durations ranging from over 10 minutes down to 1 minute. Application service costs range from 100 ms to 1000 ms. The 3 service case uses a linear service time pattern, while the other cases use the geometric service time pattern. Arrival rates are equal for all services.

From Table 2-11 and Figure 2-12, we can see that as the number of reporting period increases, the accuracy of the OLS technique gets improved. It can be explained as the tradeoffs between reporting period duration and the number of reporting periods. The minimum reporting duration which is one minute long is good enough to remove the end effects, while fewer but longer reporting periods do not give enough variation between periods to support regression.

2.3 Summary

This chapter described the algorithm of the Ordinary Least Regression (OLS), its assumptions and its approach to estimate resource demands.

A discrete-event simulator was introduced to model a single-threaded software server that offers a certain number of application services. The OLS technique was used to estimate resource demands for these simulated cases. Varied factors were related to different cases so that the ability of the OLS technique can be accessed. The evaluation is described in next paragraph.

The OLS technique can only provide the accuracy while the service time is constant. If this condition is violated, the accuracy will drop. The more the service time varies, the lower the accuracy. Also, as the number of application services increases, the accuracy drops. Mean service time and arrival rate pattern appear to have no effect on the accuracy of the OLS technique. Keeping the total number of reporting period as a constant, the longer duration gives the better accuracy. However, when the duration is long enough to remove the end-effects, to increase the duration at this time is no use to improve the accuracy. Keeping the total measurement time as a constant, to increase the number of reporting period can decrease or increase the accuracy depends upon the length of the duration. If the duration is long enough to remove the end-effects, decreasing the number of reporting period can decrease the accuracy. There exists an optimum point between these two factors.

It is obvious that the OLS technique doesn't give ideal solutions for non-deterministic case (service time is not constant). So other statistical techniques need to be explored. The next chapter describes the random coefficient model regression (RCM) which can deal with the situation where the coefficients are non-deterministic in the multiple linear model.

50 periods 5 seconds/period 75% server utilization 30 replications	3 Services		5 Services		10 Services		
	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time	
	Exp(2.200)	0.100	Exp(3.667)	0.100	Exp(7.333)	0.100	
	Exp(2.200)	0.150	Exp(3.667)	0.125	Exp(7.333)	0.111	
	Exp(2.200)	0.200	Exp(3.667)	0.150	Exp(7.333)	0.122	
			Exp(3.667)	0.175	Exp(7.333)	0.133	
			Exp(3.667)	0.200	Exp(7.333)	0.144	
					Exp(7.333)	0.156	
					Exp(7.333)	0.167	
					Exp(7.333)	0.178	
Exp(7.333)					0.189		
Exp(7.333)					0.200		
Linear Service Time Pattern		Average Error	Confidence Interval	Average Error	Confidence Interval	Average Error	Confidence Interval
Constant	0.020353544	0.003357281	0.023484061	0.002817707	0.041182062	0.003900389	
Normal(0.05)	0.018696116	0.003187525	0.028768929	0.003048241	0.044367375	0.004821715	
Normal(0.15)	0.026766487	0.005260018	0.044879804	0.008178841	0.071779129	0.007445674	
Normal(0.25)	0.043719114	0.009878868	0.067168473	0.008055409	0.096950527	0.008904365	
Normal(0.35)	0.068629007	0.00880804	0.080375287	0.014487581	0.153245321	0.014715255	
Exponential	0.149229607	0.027515576	0.22715518	0.031012049	0.343937678	0.031192235	

Table 2-1: Relative Regression Errors of OLS Model for Mean Service Times that Vary Linearly with a Range of 2, Duration 5 Seconds

50 periods 5 seconds/period 75% server utilization 30 replications	3 Services		5 Services		10 Services		
	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time	
	Exp(2.200)	0.100	Exp(3.667)	0.100	Exp(7.333)	0.100	
	Exp(2.200)	0.550	Exp(3.667)	0.325	Exp(7.333)	0.200	
	Exp(2.200)	1.000	Exp(3.667)	0.550	Exp(7.333)	0.300	
			Exp(3.667)	0.775	Exp(7.333)	0.400	
			Exp(3.667)	1.000	Exp(7.333)	0.500	
					Exp(7.333)	0.600	
Exp(7.333)					0.700		
Exp(7.333)					0.800		
Exp(7.333)					0.900		
Exp(7.333)					1.000		
Linear Service Time Pattern		Average Error	Confidence Interval	Average Error	Confidence Interval	Average Error	Confidence Interval
Constant	0.068760577	0.009749117	0.072586432	0.010713321	0.104626854	0.010193253	
Normal(0.05)	0.074330739	0.013100417	0.076262166	0.008322592	0.105005265	0.010431219	
Normal(0.15)	0.075276758	0.011915783	0.080775751	0.010333052	0.109674448	0.013479758	
Normal(0.25)	0.084826336	0.013185417	0.088102571	0.012535524	0.149254134	0.01749204	
Normal(0.35)	0.098886667	0.022157276	0.109025145	0.015597083	0.156775004	0.019290665	
Exponential	0.213193234	0.031112703	0.237619763	0.025338692	0.347751498	0.038536082	

Table 2-2: Relative Regression Errors of OLS Model for Mean Service Times that Vary Linearly with a Range of 10, Duration 5 Seconds

50 periods 60 seconds/period 75% server utilization 30 replications	3 Services		5 Services		10 Services	
	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time
	Exp(2.200)	0.100	Exp(3.667)	0.100	Exp(7.333)	0.100
	Exp(2.200)	0.550	Exp(3.667)	0.325	Exp(7.333)	0.200
	Exp(2.200)	1.000	Exp(3.667)	0.550	Exp(7.333)	0.300
			Exp(3.667)	0.775	Exp(7.333)	0.400
			Exp(3.667)	1.000	Exp(7.333)	0.500
					Exp(7.333)	0.600
					Exp(7.333)	0.700
					Exp(7.333)	0.800
					Exp(7.333)	0.900
					Exp(7.333)	1.000
Linear Service Time Pattern	Average Error	Confidence Interval	Average Error	Confidence Interval	Average Error	Confidence Interval
Constant	0.017042873	0.002981856	0.020683502	0.00256	0.028193704	0.002148766
Normal(0.05)	0.019719149	0.003529106	0.024600318	0.003221914	0.034673771	0.003414845
Normal(0.15)	0.032450374	0.006776169	0.048496134	0.006417504	0.065842096	0.006970305
Normal(0.25)	0.05051646	0.011127844	0.079192714	0.008884084	0.10616794	0.010033985
Normal(0.35)	0.073127135	0.013018113	0.09739968	0.017088081	0.156684818	0.013720237
Exponential	0.17491124	0.028323888	0.246072062	0.035168102	0.3431661	0.06425979

Table 2-3: Relative Regression Errors of OLS Model for Mean Service Times that Vary Linearly with a Range of 10, Duration 60 Seconds

50 periods 300 seconds/period 75% server utilization 30 replications	3 Services		5 Services		10 Services	
	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time
	Exp(2.200)	0.100	Exp(3.667)	0.100	Exp(7.333)	0.100
	Exp(2.200)	0.550	Exp(3.667)	0.325	Exp(7.333)	0.200
	Exp(2.200)	1.000	Exp(3.667)	0.550	Exp(7.333)	0.300
			Exp(3.667)	0.775	Exp(7.333)	0.400
			Exp(3.667)	1.000	Exp(7.333)	0.500
					Exp(7.333)	0.600
					Exp(7.333)	0.700
					Exp(7.333)	0.800
					Exp(7.333)	0.900
					Exp(7.333)	1.000
Linear Service Time Pattern	Average Error	Confidence Interval	Average Error	Confidence Interval	Average Error	Confidence Interval
Constant	0.006605668	0.001467652	0.008678808	0.001248825	0.013311641	0.001309566
Normal(0.05)	0.011973958	0.0020525	0.015676005	0.00201636	0.02479823	0.002156269
Normal(0.15)	0.028015902	0.006372728	0.042452139	0.006251371	0.068918625	0.005682227
Normal(0.25)	0.055003564	0.011504741	0.070883078	0.010334471	0.104550023	0.012858452
Normal(0.35)	0.076345515	0.013410798	0.098513581	0.012904878	0.144457845	0.013058782
Exponential	0.163585026	0.034798351	0.24182457	0.036330466	0.35278961	0.058038722

Table 2-4: Relative Regression Errors of OLS Model for Mean Service Times that Vary Linearly with a Range of 10, Duration 300 Seconds

50 periods 5 seconds/period 75% server utilization 30 replications	3 Services		5 Services		10 Services	
	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time
	Exp(1.888)	0.100	Exp(2.875)	0.100	Exp(5.448)	0.100
	Exp(1.888)	0.316	Exp(2.875)	0.178	Exp(5.448)	0.129
	Exp(1.888)	1.000	Exp(2.875)	0.316	Exp(5.448)	0.167
			Exp(2.875)	0.562	Exp(5.448)	0.215
			Exp(2.875)	1.000	Exp(5.448)	0.278
					Exp(5.448)	0.359
					Exp(5.448)	0.464
					Exp(5.448)	0.599
					Exp(5.448)	0.774
					Exp(5.448)	1.000
Geometric Service Time Pattern	Average Error	Confidence Interval	Average Error	Confidence Interval	Average Error	Confidence Interval
Constant	0.081073106	0.011446984	0.095664882	0.013920659	0.115998009	0.01324907
Normal(0.05)	0.074531119	0.01213839	0.099469299	0.011462662	0.117710995	0.011468557
Normal(0.15)	0.096168546	0.014190973	0.108181123	0.012838697	0.131405799	0.014838096
Normal(0.25)	0.091538504	0.020249263	0.124823555	0.012151821	0.157329989	0.01334858
Normal(0.35)	0.098474792	0.018665297	0.147115653	0.022158279	0.17116029	0.015774375
Exponential	0.217540207	0.034836084	0.236718401	0.023961277	0.352293896	0.036891843

Table 2-5: Relative Regression Errors of OLS Model for Mean Service Times that Vary Geometrically with a Range of 10 and Equal Inter-arrival Times, Duration 5 Seconds

50 periods 60 seconds/period 75% server utilization 30 replications	3 Services		5 Services		10 Services	
	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time
	Exp(1.888)	0.100	Exp(2.875)	0.100	Exp(5.448)	0.100
	Exp(1.888)	0.316	Exp(2.875)	0.178	Exp(5.448)	0.129
	Exp(1.888)	1.000	Exp(2.875)	0.316	Exp(5.448)	0.167
			Exp(2.875)	0.562	Exp(5.448)	0.215
			Exp(2.875)	1.000	Exp(5.448)	0.278
					Exp(5.448)	0.359
					Exp(5.448)	0.464
					Exp(5.448)	0.599
					Exp(5.448)	0.774
					Exp(5.448)	1.000
Geometric Service Time Pattern	Average Error	Confidence Interval	Average Error	Confidence Interval	Average Error	Confidence Interval
Constant	0.017062469	0.002912239	0.023553307	0.003544252	0.031915062	0.003923219
Normal(0.05)	0.019994201	0.005115098	0.02528316	0.002755032	0.03945357	0.004347424
Normal(0.15)	0.038167176	0.007316326	0.044224893	0.007710065	0.076830587	0.008213756
Normal(0.25)	0.059574217	0.011143908	0.076599518	0.01200872	0.120436827	0.010635219
Normal(0.35)	0.069622966	0.014814635	0.106559607	0.013430298	0.158462475	0.011202217
Exponential	0.169233979	0.034748583	0.296745787	0.051013149	0.353353733	0.092857486

Table 2-6: Relative Regression Errors of OLS Model for Mean Service Times that Vary Geometrically with a Range of 10 and Equal Inter-arrival Times, Duration 60 Seconds

50 periods 300 seconds/period 75% server utilization 30 replications	3 Services		5 Services		10 Services	
	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time
	Exp(1.888)	0.100	Exp(2.875)	0.100	Exp(5.448)	0.100
	Exp(1.888)	0.316	Exp(2.875)	0.178	Exp(5.448)	0.129
	Exp(1.888)	1.000	Exp(2.875)	0.316	Exp(5.448)	0.167
			Exp(2.875)	0.562	Exp(5.448)	0.215
Geometric Service Time Pattern	Average Error	Confidence Interval	Average Error	Confidence Interval	Average Error	Confidence Interval
	0.007173687	0.001475139	0.01034536	0.001508584	0.014562747	0.001474802
	0.014691004	0.002492508	0.019570055	0.002631581	0.025705824	0.002866844
	0.034483675	0.006758322	0.041961003	0.006868262	0.070838335	0.007297653
	0.053191615	0.009879093	0.08065	0.009901343	0.11303107	0.010195258
	0.081889919	0.012925454	0.094323324	0.012504709	0.141729544	0.013126959
Exponential	0.157102684	0.032862162	0.253060819	0.043271512	0.389217192	0.025818372

Table 2-7: Relative Regression Errors of OLS Model for Mean Service Times that Vary Geometrically with a Range of 10 and Equal Inter-arrival Times, Duration 300 Seconds

50 periods 5 seconds/period 75% server utilization 30 replications	3 Services		5 Services		10 Services	
	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time
	Exp(1.734)	0.100	Exp(2.442)	0.100	Exp(3.702)	0.100
	Exp(1.825)	0.316	Exp(2.571)	0.178	Exp(3.897)	0.129
	Exp(1.927)	1.000	Exp(2.713)	0.316	Exp(4.113)	0.167
			Exp(2.873)	0.562	Exp(4.355)	0.215
Varying Mean Interarrival Times	Average Error	Confidence Interval	Average Error	Confidence Interval	Average Error	Confidence Interval
	0.081588418	0.012156836	0.086555073	0.011698271	0.115040716	0.009590275
	0.09936515	0.017312058	0.088947881	0.010043898	0.118763567	0.013796454
	0.099265695	0.011672157	0.105641736	0.013482516	0.132457256	0.011152035
	0.07909227	0.012736935	0.120513425	0.013993813	0.14950156	0.014876731
	0.114127052	0.016935315	0.124511109	0.018796069	0.166380694	0.018316767
Exponential	0.247751671	0.0377873	0.279431335	0.03491236	0.376868324	0.034924379

Table 2-8: Relative Regression Errors of OLS Model for Mean Service Times that Vary Geometrically with a Range of 10 and Increasing Inter-arrival Times, Duration 5 Seconds

50 periods 60 seconds/period 75% server utilization 30 replications	3 Services		5 Services		10 Services	
	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time
	Exp(1.734)	0.100	Exp(2.442)	0.100	Exp(3.702)	0.100
	Exp(1.825)	0.316	Exp(2.571)	0.178	Exp(3.897)	0.129
	Exp(1.927)	1.000	Exp(2.713)	0.316	Exp(4.113)	0.167
			Exp(2.873)	0.562	Exp(4.355)	0.215
			Exp(3.053)	1.000	Exp(4.628)	0.278
					Exp(4.936)	0.359
					Exp(5.289)	0.464
					Exp(5.696)	0.599
					Exp(6.170)	0.774
					Exp(6.731)	1.000
Varying Mean Interarrival Times	Average Error	Confidence Interval	Average Error	Confidence Interval	Average Error	Confidence Interval
Constant	0.016434303	0.003091611	0.023115131	0.003173593	0.035051134	0.004255982
Normal(0.05)	0.01991762	0.004452925	0.027444863	0.00293833	0.0363936	0.003168069
Normal(0.15)	0.038322169	0.006075152	0.049629164	0.007380853	0.087247334	0.008244336
Normal(0.25)	0.062285517	0.013256963	0.090902788	0.009539424	0.129927221	0.009844851
Normal(0.35)	0.07255989	0.01776796	0.104458123	0.010550843	0.169570482	0.01285183
Exponential	0.219778631	0.038280808	0.259784871	0.035590593	0.400267525	0.046062861

Table 2-9: Relative Regression Errors of OLS Model for Mean Service Times that Vary Geometrically with a Range of 10 and Increasing Inter-arrival Times, Duration 60 Seconds

50 periods 300 seconds/period 75% server utilization 30 replications	3 Services		5 Services		10 Services	
	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time
	Exp(1.734)	0.100	Exp(2.442)	0.100	Exp(3.702)	0.100
	Exp(1.825)	0.316	Exp(2.571)	0.178	Exp(3.897)	0.129
	Exp(1.927)	1.000	Exp(2.713)	0.316	Exp(4.113)	0.167
			Exp(2.873)	0.562	Exp(4.355)	0.215
			Exp(3.053)	1.000	Exp(4.628)	0.278
					Exp(4.936)	0.359
					Exp(5.289)	0.464
					Exp(5.696)	0.599
					Exp(6.170)	0.774
					Exp(6.731)	1.000
Varying Mean Interarrival Times	Average Error	Confidence Interval	Average Error	Confidence Interval	Average Error	Confidence Interval
Constant	0.008321944	0.001792888	0.010067109	0.001651738	0.0135989	0.001413207
Normal(0.05)	0.012453052	0.00254253	0.019328915	0.002480593	0.028680866	0.003431841
Normal(0.15)	0.034675993	0.006407289	0.044533453	0.005454339	0.065188994	0.006449274
Normal(0.25)	0.049342106	0.008553081	0.076548161	0.012002946	0.117252914	0.009933059
Normal(0.35)	0.075237274	0.014004198	0.093900032	0.014312055	0.148126661	0.017375409
Exponential	0.197636172	0.041494303	0.275920025	0.039608216	0.41832986	0.174259714

Table 2-10: Relative Regression Errors of OLS Model for Mean Service Times that Vary Geometrically with a Range of 10 and Increasing Inter-arrival Times, Duration 300 Seconds

Total Time: 150 mins Server Utilization: 75% Replications: 30	3 Services Constant		10 Services Constant		10 Services Normal(0.05)	
	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time
	Exp(2.200)	0.100	Exp(5.448)	0.100	Exp(5.448)	0.100
	Exp(2.200)	0.550	Exp(5.448)	0.129	Exp(5.448)	0.129
	Exp(2.200)	1.000	Exp(5.448)	0.167	Exp(5.448)	0.167
			Exp(5.448)	0.215	Exp(5.448)	0.215
			Exp(5.448)	0.278	Exp(5.448)	0.278
			Exp(5.448)	0.359	Exp(5.448)	0.359
			Exp(5.448)	0.464	Exp(5.448)	0.464
			Exp(5.448)	0.599	Exp(5.448)	0.599
			Exp(5.448)	0.774	Exp(5.448)	0.774
			Exp(5.448)	1.000	Exp(5.448)	1.000
Number of Reporting Periods	Average Error	Confidence Interval	Average Error	Confidence Interval	Average Error	Confidence Interval
13	0.00977974	0.001942848	0.03351599	0.007270189	0.09030748	0.013550329
25	0.00803786	0.001660513	0.02265316	0.002062355	0.04525687	0.005741718
50	0.00836219	0.001445178	0.01788514	0.001386964	0.02979643	0.002681885
75	0.00837709	0.001766423	0.01644951	0.001682455	0.02405382	0.002199227
100	0.00840412	0.001395676	0.01812339	0.001397263	0.02325546	0.002080201
125	0.00768422	0.001881074	0.01648126	0.001727097	0.02416738	0.002587742
150	0.00884786	0.001456357	0.01764565	0.001531458	0.02081991	0.002137913

Total Time: 150 mins Server Utilization: 75% Replications: 30	10 Services Normal(0.15)		10 Services Normal(0.25)		10 Services Normal(0.35)		10 Services Exponential	
	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time
	Exp(5.448)	0.100	Exp(5.448)	0.100	Exp(5.448)	0.100	Exp(5.448)	0.100
	Exp(5.448)	0.129	Exp(5.448)	0.129	Exp(5.448)	0.129	Exp(5.448)	0.129
	Exp(5.448)	0.167	Exp(5.448)	0.167	Exp(5.448)	0.167	Exp(5.448)	0.167
	Exp(5.448)	0.215	Exp(5.448)	0.215	Exp(5.448)	0.215	Exp(5.448)	0.215
	Exp(5.448)	0.278	Exp(5.448)	0.278	Exp(5.448)	0.278	Exp(5.448)	0.278
	Exp(5.448)	0.359	Exp(5.448)	0.359	Exp(5.448)	0.359	Exp(5.448)	0.359
	Exp(5.448)	0.464	Exp(5.448)	0.464	Exp(5.448)	0.464	Exp(5.448)	0.464
	Exp(5.448)	0.599	Exp(5.448)	0.599	Exp(5.448)	0.599	Exp(5.448)	0.599
	Exp(5.448)	0.774	Exp(5.448)	0.774	Exp(5.448)	0.774	Exp(5.448)	0.774
	Exp(5.448)	1.000	Exp(5.448)	1.000	Exp(5.448)	1.000	Exp(5.448)	1.000
Number of Reporting Periods	Average Error	Confidence Interval	Average Error	Confidence Interval	Average Error	Confidence Interval	Average Error	Confidence Interval
13	0.21890223	0.034146544	0.24903003	0.048573258	0.34841277	0.070375633		
25	0.1242526	0.013766373	0.18219232	0.024240861	0.20600101	0.026931873	0.64088751	
50	0.07432098	0.007197995	0.12391675	0.012528642	0.15308034	0.016945036	0.35916365	0.069311155
75	0.05924906	0.004402396	0.09243006	0.008642827	0.12023885	0.013356942	0.32191225	0.060466771
100	0.05034979	0.004905106	0.07989153	0.009870045	0.10762776	0.009143613	0.27326943	0.027105509
125	0.04611729	0.004672835	0.07125425	0.006346368	0.09659653	0.009642331	0.2237621	0.025999435
150	0.04123762	0.003834604	0.06675542	0.006605973	0.08284927	0.007391884	0.22388113	0.022369528

Table 2-11: Relative Regression Errors of OLS Model for Different Numbers of Reporting Periods

Figure 2-2: Relative Regression Errors of OLS Model for Mean Service Times that Vary Linearly with a Range of 2, Duration 5 Seconds

Figure 2-3: Relative Regression Errors of OLS Model for Mean Service Times that Vary Linearly with a Range of 10, Duration 5 Seconds

Relative Regression Errors of OLS Model for Mean Service Times that Vary Linearly with a Range of 10, Duration 60 Seconds

Figure 2-4: Relative Regression Errors of OLS Model for Mean Service Times that Vary Linearly with a Range of 10, Duration 60 Seconds

Relative Regression Errors of OLS Model for Mean Service Times that Vary Linearly with a Range of 10, Duration 300 Seconds

Figure 2-5: Relative Regression Errors of OLS Model for Mean Service Times that Vary Linearly with a Range of 10, Duration 300 Seconds

Figure 2-6: Relative Regression Errors of OLS Model for Mean Service Times that Vary Geometrically with a Range of 10 and Equal Inter-arrival Times, Duration 5 Seconds

Figure 2-7: Relative Regression Errors of OLS Model for Mean Service Times that Vary Geometrically with a Range of 10 and Equal Inter-arrival Times, Duration 60 Seconds

Figure 2-8: Relative Regression Errors of OLS Model for Mean Service Times that Vary Geometrically with a Range of 10 and Equal Inter-arrival Times, Duration 300 Seconds

Figure 2-9: Relative Regression Errors of OLS Model for Mean Service Times that Vary Geometrically with a Range of 10 and Increasing Inter-arrival Times, Duration 5 Seconds

Figure 2-10: Relative Regression Errors of OLS Model for Mean Service Times that Vary Geometrically with a Range of 10 and Increasing Inter-arrival Times, Duration 60 Seconds

Figure 2-11: Relative Regression Errors of OLS Model for Mean Service Times that Vary Geometrically with a Range of 10 and Increasing Inter-arrival Times, Duration 300 Seconds

Relative Regression Errors of OLS Model for Different Numbers of Reporting Periods

Figure 2-12: Relative Regression Errors of OLS Model for Different Numbers of Reporting Periods

Chapter 3 Random Coefficient Model Regression

This chapter describes the stochastic parameters regression model. The algorithm to estimate parameters of this model is also introduced. It is the Kalman filter. As a special case of the stochastic parameters regression model, the random coefficient model is also introduced and the algorithm for this model is given. In Section 3.4, the RCM (random coefficient model) statistical tool is evaluated using the same test cases as in Section 2.2.

3.1 A Stochastic Parameters Regression Model [6]

Our goal in this thesis is to estimate resource demands of each application service. To fulfill the goal, we need to do modeling for this problem. In last chapter, we used multiple linear regression model to describe the problem. However it is not good since resource demands are not deterministic and break up the assumption of the model. So a new model needs to be found to suit this situation.

Let's first review linear regression model:

$$y_t = \beta_t X_t + e_t \quad (3-1)$$

The first assumption of this model is the coefficients β in the model are deterministic. What happens if this assumption is challenged, in the sense that the coefficients (in our case, resource demands) are not fixed and random variable instead. In equation (3-1), when the regression parameters are taken to be random variables, this model is called Stochastic Parameters Regression model. In this model, the parameters, β_t , are possibly autocorrelated random variables.

Let β_t be a random variable, which represents observations through time on a single entity. Some assumptions have to be made to study this process. A general assumption is stationarity. The series β_t is said to be stationary if the following three conditions hold:

- (1) The mean of β_t is the same for all time periods t .
- (2) The variance of β_t is the same finite value for all time periods t .

(3) The correlation between the random variable β_t and β_s depends only on their distance apart in time; that is, on $|t-s|$.

A common mean β through time is assumed since the process β_t is taken to be stationary.

$$E(\beta_t) = \beta \quad (3-2)$$

for all time periods t . Now consider uncorrelated β :

$$\beta_t - \beta = a_t \quad (3-3)$$

where the random variables a_t have zero mean, fixed variance through time, and, for $t \neq s$, a_t is uncorrelated with a_s . Such a process a_t is usually called white noise. If the time series β_t follows the simple model (3-3), then the β_t have common mean, common variance, and are not correlated with one another.

For time series that are autocorrelated, the result is different. We introduce a very simple model called first-order autoregressive process for this case. Let ϕ be the correlation between adjacent values in times of the series, so that

$$\text{Corr.}(\beta_t, \beta_{t-1}) = \phi$$

Then, suppose that the correlation between values two time periods apart is ϕ^2 , the correlation between values three times periods apart is ϕ^3 , and so on, so that

$$\text{Corr.}(\beta_t, \beta_{t-j}) = \phi^j \quad (j=1, 2, 3, \dots) \quad (3-4)$$

Such a process indicates that the strength of the correlation between values of a series is inversely related to their distance apart in time. Thus, we can generate the following equation:

$$(\beta_t - \beta) = \phi(\beta_{t-1} - \beta) + a_t \quad (3-5)$$

This model (3-5) is called a first-order autoregressive model. If the stationary time series β_t obeys this model, the mean of β_t is β and the variance of β_t is:

$$\sigma_{\beta}^2 = \frac{\sigma_a^2}{1 - \phi^2}$$

Where σ_a^2 is the variance of the white noise process a_t .

If a time series is autocorrelated, forecasts of the future can be predicated on the basis of current and past observations. Suppose we are at time T and want to predict l time period into the future, β_{T+l} from available observations $\beta_T, \beta_{T-1}, \beta_{T-2} \dots$. From equation (3-5), we can get:

$$\begin{aligned}(\beta_{T+l} - \beta) &= \phi(\beta_{T+l-1} - \beta) + a_{T+l} \\&= \phi^2(\beta_{T+l-2} - \beta) + a_{T+l} + \phi \cdot a_{T+l-1} \\&= \phi^3(\beta_{T+l-3} - \beta) + a_{T+l} + \phi \cdot a_{T+l-1} + \phi^2 \cdot a_{T+l-2}\end{aligned}$$

Keep doing this way, we can find

$$(\beta_{T+l} - \beta) = \phi^l(\beta_T - \beta) + (a_{T+l} + \phi \cdot a_{T+l-1} + \dots + \phi^{l-1} \cdot a_{T+1})$$

where future values a^{T+j} ($j \geq 1$) of the white noise process at will be unpredictable and so we can see that the optimal forecast of β_{T+l} is:

$$\beta + \phi^l(\beta_T - \beta)$$

It is obvious to see from the above equation that forecast of future values of the series depend only on the most recent observation β_T .

In the summary, the stochastic parameters regression model can be given as:

$$y_t = x_t' \beta_t + e_t \quad (3-6)$$

where the vectors of parameters, β_t , are random variables. There are a number of alternative time-series models available to describe this problem but only the first-order autoregressive model interests us. It is given in equation (3-5). When the matrix ϕ of this equation is a matrix of zeros, a very special equation is produced:

$$\beta_t - \beta = a_t \quad (3-7)$$

From equation (3-7), we can see that the parameters, β_t , are random variables which follow a normal distribution with mean β . This means that the future values of the parameters are unpredictable by past values. This model is known as the Random Coefficient Model. Hildreth and Houck (1968) did a lot of research in this topic. We will discuss this topic in Section 3.3.

3.2 Kalman Filter [7]

In last section, we introduced the stochastic parameters regression model. We will discuss the algorithms to estimate coefficients of the model in this section.

Recall the stochastic parameters regression model given in equation (3-6). Now supposed that there are T observations, $y_1, y_2, \dots, y_T$ on a dependent variable and observations $x_{i1}, x_{i2}, \dots, x_{iT}$ ($i=1, 2, \dots, K$) on each of K independent variables. A linear equation is given as below:

$$y_t = \beta_{1t}x_{1t} + \beta_{2t}x_{2t} + \dots + \beta_{Kt}x_{Kt} + e_t \quad (3-8)$$

where β_{it} are time-varying parameters and e_t is an error term which has zero mean, fixed variance and non-autocorrelated. Let $x_t' = (x_{1t}, x_{2t}, \dots, x_{Kt})$; $\beta_t' = (\beta_{1t}, \beta_{2t}, \dots, \beta_{Kt})$, then equation (3-8) is transformed into:

$$y_t = x_t' \beta_t + e_t$$

Which is the same as equation (3-6). The β_{it} are random variables and have fixed mean β_i through time. Therefor the mean vector for β_t' is:

$$\beta' = (\beta_1, \beta_2, \dots, \beta_K)$$

The parameters follow a first-order vector autoregressive process, so we can get:

$$\beta_t - \beta = \Phi(\beta_{t-1} - \beta) + a_t \quad (3-9)$$

Where Φ is a $K \times K$ matrix of fixed parameters and a_t is a $K \times 1$ vector of random errors with zero mean and fixed covariance matrix Ω through time. The a_t are supposed to be non-autocorrelated and to be uncorrelated with the error e_t . These two equations work together as the stochastic parameter regression model.

We can rewrite equation (3-6) as:

$$y_t = x_t' \beta + x_t' (\beta_t - \beta) + e_t \quad (3-10)$$

or

$$y_t = x_t' \beta + u_t \quad (3-11)$$

where

$$u_t = x_t'(\beta_t - \beta) + e_t \quad (3-12)$$

Equation (3-10) can be thought as a multiple regression model with fixed parameters β and a zero mean error term u_t . In the case the β_t are not fixed and equal to the mean β , that is β_t are stochastic, the error terms u_t will not have the same variance. The formula of the variance of u_t is given below:

$$Var(u_t) = x_t' \Omega_\beta x_t + \sigma^2 \quad (3-13)$$

where Ω_β is the covariance matrix of β_t . As long as the parameters are β_t autocorrelated, that is Φ in equation (3-9) is not a zero matrix, this stochastic parameter regression model can be considered as a regular multiple linear regression model with error terms showing both heteroscedasticity and autocorrelated errors. It indicates that if the parameters of the error covariance in equation (3-13) are known, we can use generalized least square to deal with this problem.

However this is not the only way. There is another technique known as Kalman filtering which provides a computationally efficient framework to solve this kind of problem. Before discuss the Kalman filter algorithm, we need to introduce some new terminology. For the distribution of β_t , given information up to time n , that is, the observations $y_1, y_2, \dots, y_n$, $\beta(t/n)$ will be used to represent the mean of this distribution and $P(t/n)$ to represent its $K \times K$ covariance matrix. Two definitions are given below:

$$\beta(t/n) = E[\beta_t | y_1, y_2, \dots, y_n]$$

$$P(t/n) = Var[\beta_t | y_1, y_2, \dots, y_n]$$

For the distribution of y_t , given information up to and including time $t-1$, its mean and variation will be given as:

$$y(t/t-1) = E[y_t | y_1, y_2, \dots, y_{t-1}]$$

$$h_t = Var[y_t | y_1, y_2, \dots, y_{t-1}]$$

Now, we go through the Kalman filter algorithm step by step. First, we need to find the joint distribution of β_t and y_t with information available at time $t-1$. From the

assumptions we made before, this distribution is multivariate normal. Therefore the distribution can be described by its mean and covariance matrix.

Let's deal with the mean of β_t first by taking expectations, conditional on the available information, in equation (3-9). The result is given below:

$$E(\beta_t | y_1, y_2, \dots, y_{t-1}) = \beta + \Phi E(\beta_{t-1} | y_1, y_2, \dots, y_{t-1}) - \Phi \beta + E(a_t | y_1, y_2, \dots, y_{t-1})$$

Because the errors are independent from observations to observations, the expectation of the error term is zero. So the above equation can be rewritten as:

$$\beta(t | t-1) = \Phi \beta(t-1 | t-1) + (I - \Phi) \beta \quad (3-14)$$

where I is the $K \times K$ identity matrix.

Also from the equation (3-9), the covariance matrix of β_t , given information available at time $t-1$, can be expressed as:

$$\text{Var}(\beta_t | y_1, y_2, \dots, y_{t-1}) = \Phi \cdot \text{Var}(\beta_{t-1} | y_1, y_2, \dots, y_{t-1}) \cdot \Phi' + \Omega$$

where Ω is the covariance matrix of a_t . We can rewrite this equation as:

$$P(t | t-1) = \Phi \cdot P(t-1 | t-1) \cdot \Phi' + \Omega \quad (3-15)$$

From the information up to time $t-1$, the mean of the distribution of y_t can be achieved by taking conditional expectation in equation (3-6). The result is given below:

$$E(y_t | y_1, y_2, \dots, y_{t-1}) = x_t' E(\beta_t | y_1, y_2, \dots, y_{t-1}) + E(e_t | y_1, y_2, \dots, y_{t-1})$$

As we know, the errors are independent from observation to observation, so the expectation of the error term is zero. Then we get the equation:

$$y(t | t-1) = x_t' \beta(t | t-1) \quad (3-16)$$

Also, we can use equation (3-6) to get the variance of y_t by information available at $t-1$. The result is given as:

$$Var(y_t | y_1, y_2, \dots, y_{t-1}) = x_t' Var(\beta_t | y_1, y_2, \dots, y_{t-1}) x_t + \sigma^2$$

where σ^2 is the variance of e_t . We can rewrite this equation as:

$$h_t = x_t' P(t | t-1) x_t + \sigma^2 \quad (3-17)$$

To get the covariance between β_t and y_t , we still use equation (3-6) to derive the equation as below:

$$\begin{aligned} Cov(\beta_t, y_t | y_1, y_2, \dots, y_{t-1}) &= Var(\beta_t | y_1, y_2, \dots, y_{t-1}) x_t \\ &= P(t | t-1) x_t \end{aligned} \quad (3-18)$$

Next, we need to figure out the distribution of β_t with information available at time t . We already know the joint distribution of β_t and y_t . To solve the problem, some known results of multivariate normal distribution need to be introduced. Let z_1 and z_2 be a pair of random variables, whose joint distribution is multivariate normal. The means and covariance matrices of these two random variables are μ_i and Σ_{ii} , which are expressed as:

$$E(z_i) = \mu_i$$

$$E[(z_i - \mu_i)(z_i - \mu_i)'] = \Sigma_{ii} \quad (i=1, 2)$$

The covariance between them is defined as Σ_{12} which is expressed in the following equation:

$$E[(z_1 - \mu_1)(z_2 - \mu_2)'] = \Sigma_{12}$$

This joint normal distribution can be expressed as:

$$\begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \sim N \left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} ; \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right] \quad (3-19)$$

It is also known that the distribution of z_1 given z_2 is multivariate normal whose mean and covariance matrices are:

$$E(z_1 | z_2) = \mu_1 + \Sigma_{12} \Sigma_{22}^{-1} (z_2 - \mu_2) \quad (3-20)$$

$$\text{Var}(z_1 | z_2) = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \quad (3-21)$$

In our case, the multivariate normal distribution is the joint distribution of β_t and y_t given information available at time $t-1$; that is given $y_1, y_2, \dots, y_{t-1}$. β_t and y_t are z_1 and z_2 respectively. Using the result from equation (3-14) to equation (3-15), we replace the corresponding terms in equation (3-19) and get the format as given below:

$$\begin{pmatrix} \beta_t \\ y_t \end{pmatrix}_{t-1} \sim N \left[\begin{pmatrix} \Phi\beta(t-1|t-1) + (\mathbf{I} - \Phi)\beta \\ x_t'\beta(t|t-1) \end{pmatrix} ; \begin{pmatrix} P(t|t-1) & P(t|t-1)x_t' \\ x_t'P(t|t-1) & h_t \end{pmatrix} \right] \quad (3-22)$$

So, we get the distribution of β_t given all information at time t ; that is given $y_1, y_2, \dots, y_{t-1}$ by conditioning the distribution of β_t at time $t-1$ or that of y_t at time $t-1$. As we discussed before, this distribution is also multivariate normal distribution. From equation (3-20), its mean is:

$$\beta(t|t) = \Phi\beta(t-1|t-1) + (\mathbf{I} - \Phi)\beta + P(t|t-1)x_t'h_t^{-1}[y_t - x_t'\beta(t|t-1)] \quad (3-23)$$

And also from equation (3-21), its covariance matrix is:

$$P(t|t) = P(t|t-1) - P(t|t-1)x_t'h_t^{-1}x_t'P(t|t-1) \quad (3-24)$$

Put the equation 3-14, 3-15, 3-17, 3-23 and 3-24 together, we have the following:

- (i) $\beta(t|t-1) = \Phi\beta(t-1|t-1) + (\mathbf{I} - \Phi)\beta$
- (ii) $P(t|t-1) = \Phi \cdot P(t-1|t-1) \cdot \Phi' + \Omega$
- (iii) $h_t = x_t'P(t|t-1)x_t + \sigma^2$ (3-25)
- (iv) $\beta(t|t) = \Phi\beta(t-1|t-1) + (\mathbf{I} - \Phi)\beta + P(t|t-1)x_t'h_t^{-1}[y_t - x_t'\beta(t|t-1)]$
- (v) $P(t|t) = P(t|t-1) - P(t|t-1)x_t'h_t^{-1}x_t'P(t|t-1)$

The set of these five equations derived by Kalman in 1960 is a special case of the Kalman filter algorithm. By using this set of equations, we can compute the conditional means and variance of the stochastic parameters, β_t . The computations begin with the $\beta(0|0)$

and $P(0|0)$ which are initial values and in most case are chosen arbitrarily. Given this initial values, we can find $\beta(1|0)$, $P(1|0)$, h_1 , $\beta(1|1)$ and $P(1|1)$ respectively through (i)-(v) of (3-35) by setting $t=1$. Then using these results and setting $t=2$, we can find $\beta(2|1)$, $P(2|1)$, h_2 , $\beta(2|2)$ and $P(2|2)$. Keep proceeding in this way, finally we can get the means and covariance of parameters, β_t .

Let us go back to the section 3.1, when the matrix Φ is zero, this model is changed into the Random Coefficient Model. It indicates that we can still use this algorithm discussed above to solve the Random Coefficient Model and the only difference is that there is one parameter Φ short.

Watching equation (3-25) sets carefully, however, we find that several fixed parameter need to be known to do the computation. These include the mean β of the stochastic parameters, the autoregressive matrix Φ of the process followed by these parameters, the error variance σ^2 of the measurement equation, and the covariance matrix Ω of the error terms in the transition equation. Given observations running over T time periods, we can use the method of maximum likelihood to obtain estimates for these parameters.

First, we need to derive the likelihood function which is the joint distribution of $y_1, y_2, \dots, y_T$, as a function of the fixed parameters of the model. The joint probability density function of $y_1, y_2, \dots, y_T$ can be expressed as the product of the conditional densities of y_t , given $y_1, y_2, \dots, y_{t-1}$, which is expressed as below:

$$p(y_1, y_2, \dots, y_T) = p(y_1)p(y_2 | y_1)p(y_3 | y_1, y_2) \dots p(y_T | y_1, y_2, \dots, y_{T-1})$$

Also, the distribution of y_t , given by $y_1, y_2, \dots, y_{t-1}$ is normal with mean as $x_t' \beta(t|t-1)$ and variance h_t given by equation 3-16 and 3-17 respectively. So the conditional density function is:

$$P(y_t | y_1, y_2, \dots, y_{t-1}) = (2\pi h_t)^{-\frac{1}{2}} \exp \left[-\frac{(y_t - x_t' \beta(t|t-1))^2}{2h_t} \right]$$

Then the likelihood function, or joint density function of the observations, is given by

$$L = (2\pi)^{-T/2} \prod_{t=1}^T h_t^{-1/2} \exp \left[- \sum_{t=1}^T (y_t - x_t' \beta(t|t-1))^2 / 2h_t \right].$$

Taking logarithms on both sides of the equation, we get the log likelihood function, which is expressed as below:

$$\log L = -\frac{T}{2} \log(2\pi) - \frac{1}{2} \sum_{t=1}^T \log h_t - \frac{1}{2} \sum_{t=1}^T (y_t - x_t' \beta(t|t-1))^2 h_t^{-1} \quad (3-26)$$

The function in equation(3-26) must be maximized to obtain maximum likelihood estimates of the fixed coefficients of the model. There are y_t , x_t , the conditional expectation $\beta(t|t-1)$ of the stochastic parameters, and the conditional variance h_t of the dependent variables in this equation. $\beta(t|t-1)$ and h_t can be evaluated as functions of the fixed parameters of the model through the Kalman algorithm (3-25).

In order to infer maximum likelihood estimates of these parameters, we require the value of the log likelihood function in equation(3-26) is a maximum, or equivalently minimize the function

$$\sum_{t=1}^T \left[\log h_t + (y_t - x_t' \beta(t|t-1))^2 h_t^{-1} \right]$$

Since $\beta(t|t-1)$ and h_t are quite complicated functions of the fixed parameters of the model, it is impossible to solve this function minimization problem analytically. Numerical function optimization algorithms have to be used. So the numerical maximization of the log likelihood function can produce point estimates of the fixed parameters of the model.

3.3 Random Coefficient Model [8]

Considering that current resource demands have no effects on future ones, we use the random coefficient model to our problem.

As we discussed in Section 3.1, the random coefficient model is a special case of the first-order autoregressive model and can be solved by using Kalman filter algorithm.

However, the numerical function optimization algorithm in this process is cumbersome. So a simple alternative needs to be found.

Let's talk about the Random Coefficient Model which is given as below:

$$y_i = \mathbf{X}_i \beta_i + \varepsilon_i$$

where

$$\begin{aligned} \beta_i &= \beta + v_i, \\ E[v_i] &= 0, \\ E[v_i v_i'] &= \Gamma. \end{aligned} \tag{3-27}$$

The β_i , which belongs to a bucket, is the outcome of a random process with mean vector β and covariance matrix Γ . Substitute β_i expression into the equation (3-27), we can get the model:

$$\begin{aligned} y_i &= \mathbf{X}_i \beta + (\varepsilon_i + \mathbf{X}_i v_i) \\ &= \mathbf{X}_i \beta + w_i \end{aligned}$$

where

$$E[w] = 0$$

and

$$\begin{aligned} E[w_i w_i'] &= \sigma_i^2 \mathbf{I} + \mathbf{X}_i \Gamma \mathbf{X}_i' \\ &= \Pi_i. \end{aligned}$$

In our case, mean service times are not constant. We assume mean service times are randomly changed from observation to observation.

This model fits directly into the framework of groupwise heteroscedasticity. For the full sample of observations,

$$\mathbf{V} = \begin{bmatrix} \Pi_1 & 0 & 0 & \dots & 0 \\ 0 & \Pi_2 & 0 & \dots & 0 \\ & & \cdot & & \\ & & \cdot & & \\ 0 & 0 & 0 & & \Pi_n \end{bmatrix}$$

To deal with this model, let $\mathbf{b}_i$ be the i th ordinary least squares coefficient vector and the equation as below be the covariance matrix of $\mathbf{b}_i$

$$\mathbf{V}_i + \Gamma = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \Pi_i \mathbf{X} (\mathbf{X}'\mathbf{X})^{-1} \quad (3-28)$$

where

$$\mathbf{V}_i = \sigma_i^2 [\mathbf{X}_i' \mathbf{X}_i]^{-1} \quad (3-29)$$

The Generalized Least Squares (GLS) estimator in this groupwise heteroscedastic model can be written as:

$$\hat{\beta} = \sum_i \mathbf{w}_i \mathbf{b}_i \quad (3-30)$$

where

$$\mathbf{w}_i = \left\{ \sum_{i=1}^n [\Gamma + \mathbf{V}_i]^{-1} \right\}^{-1} [\Gamma + \mathbf{V}_i]^{-1}.$$

Equation (3-30) shows that the GLS estimator is a matrix weighted average of the OLS estimators. The asymptotic covariance matrix for $\hat{\beta}$ is the inverse matrix in braces in equation(3-31).

The main hindrance for this algorithm is the unknown parameters in Γ and $\mathbf{V}_i$. Swamy suggested a two –step approach in 1971. It begins with ordinary least squares. Let $\mathbf{b}_i$ be the specific OLS coefficient vector and $\hat{\mathbf{V}}_i$ be the sample covariance matrix which can be calculated from the following equation:

$$\hat{\mathbf{V}}_i = s_i^2 (\mathbf{X}_i' \mathbf{X}_i)^{-1} \quad (3-31)$$

where

$$s_i^2 = \frac{\mathbf{e}_i' \mathbf{e}_i}{T - K} \quad (3-32)$$

Equation (3-32) is actually from the ordinary linear regression.

Let

$$\bar{\mathbf{b}} = \frac{1}{n} \sum_i \mathbf{b}_i \quad (3-33)$$

Then,

$$\hat{\Gamma} = \mathbf{G} = \frac{1}{n-1} \left(\sum_i \mathbf{b}_i \mathbf{b}_i' - n \bar{\mathbf{b}} \bar{\mathbf{b}}' \right) - \frac{1}{n} \sum_i \hat{\mathbf{v}}_i. \quad (3-34)$$

This algorithm is quite good but still a potential problem exists. It is that $\mathbf{G}$ may not be positive definite. The two matrices in $\mathbf{G}$ are of order 1 and $1/(nT)$. So in terms of the large sample behavior of $\mathbf{G}$, the second matrix will be negligible. Swamy suggests various fixups for a non-positive definite $\mathbf{G}$. However, a simple and asymptotically valid way is to drop the second part. So equation (3-34) is changed to

$$\hat{\Gamma} = \mathbf{G} = \frac{1}{n-1} \left(\sum_i \mathbf{b}_i \mathbf{b}_i' - n \bar{\mathbf{b}} \bar{\mathbf{b}}' \right). \quad (3-35).$$

3.4 Results of Study

We use the algorithm in Section 3.3 to implement the random coefficient model regression (RCM) tool. To run RCM, we need to divide the whole data into a few buckets and then follow the method in Section 3.3. Note that different bucket sizes can give different results. The RCM tool can automatically choose different bucket size and find the optimum bucket size by trial and error that gives the minimum reported error.

All the parameter settings are exactly the same as ones in Section 2.2.2. Eleven tables and eleven figures, which correspond to the previous tables respectively, are given. Some confidence intervals of reported errors are not perfect since the RCM tool caused exceptions, especially in short duration cases where it could not finish 30 replications.

We pose four questions in Section 2.2.2 for the RCM tool. The RCM behaves the same way as the OLS.

Service time distribution has significant effects on the RCM technique. Linear Service Time Pattern and Geometric Service Time Pattern are used in tables from 3-2 to 3-4 and tables from 3-5 to 3-7 respectively. It is obvious that the mean service time pattern doesn't have any impact on the accuracy of the RCM technique. The range of mean service times across different application services is 2 in Table 3-1 and 10 in Table 3-2. These tables show that the errors reported increase as the range changed from 2 to 10. The reason for this is end-effects. The tables from 3-5 to 3-7 has the equal mean inter-arrival times pattern, while the tables from 3-8 to 3-10 has the increasing inter-arrival times pattern. From data in the tables, we can see that the arrival rate pattern has no impact on the RCM.

The tables from 3-2 to 3-4 have the same parameter setting and the reporting duration 5, 60 and 300 seconds respectively. So are the tables from 3-5 to 3-7 and the ones from 3-8 to 3-10. We can see from tables that as the duration increases, the accuracy of the RCM technique gets improved. From Table 3-11, we can see that as the number of reporting period increases (the total measurement time is fixed), the accuracy of the RCM technique gets improved.

The reason of these phenomena is that the RCM uses the OLS in each bucket so that the effects of factors on the OLS transfer to the RCM.

Now we focus on the improvement of the RCM over the OLS. Table 2-4 and Table 3-4 have Linear Service Time Pattern and duration of 300 seconds. Table 2-7 and Table 3-7 have Geometric Service Time Pattern, equal inter-arrival times and duration of 300 seconds. Table 2-10 and Table 3-10 have Geometric Service Time Pattern, increasing inter-arrival times and duration of 300 seconds. We use these cases for comparison between the RCM and the OLS. From these tables, we can see that the RCM gives better accuracy for low service time variation than the OLS. Also, as the number of service

increases, the extent of improvement decreases. In 3 services case, even for the exponential case, the RCM provides better estimates. But as the number of services increases to 10, the RCM doesn't give better accuracy than the OLS for the exponential case.

3.5 Summary

This chapter introduced the stochastic parameters regression model and the algorithm, Kalman filter, to estimate parameters for this model. As a special form of the stochastic parameters regression model, the random coefficient model and its algorithm were also described

The RCM (random coefficient model) was used to the simulations in Section 2.2.2. In general the RCM technique can provide better results than the OLS. The extent of improvement drops as the service time varies more and the number of service increases. Finally, in 10 services exponential case, the RCM hardly gives any improvement in accuracy over OLS. Besides these, the RCM technique shows the same characteristics as the OLS.

Although the RCM gives some improvement over the OLS, it does not meet our requirement. We need to try other possible statistical techniques. The next chapter provides maximum likelihood estimation of generalized regression model and Chapter 5 provides restricted least square regression.

50 periods 5 seconds/period 75% server utilization 30 replications	3 Services		5 Services		10 Services	
	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time
	Exp(2.200)	0.100	Exp(3.667)	0.100	Exp(7.333)	0.100
	Exp(2.200)	0.150	Exp(3.667)	0.125	Exp(7.333)	0.111
	Exp(2.200)	0.200	Exp(3.667)	0.150	Exp(7.333)	0.122
			Exp(3.667)	0.175	Exp(7.333)	0.133
			Exp(3.667)	0.200	Exp(7.333)	0.144
					Exp(7.333)	0.156
					Exp(7.333)	0.167
					Exp(7.333)	0.178
					Exp(7.333)	0.189
					Exp(7.333)	0.200
Linear Service Time Pattern	Average Error	Confidence Interval	Average Error	Confidence Interval	Average Error	Confidence Interval
Constant	0.012883908	0.002980664	0.019179615	0.003488463	0.040388759	0.004921862
Normal(0.05)	0.012988828	0.002820483	0.023153637	0.003432314	0.043876334	0.005086987
Normal(0.15)	0.01514715	0.005044463	0.03398298	0.006524929	0.072264217	0.006474451
Normal(0.25)	0.025492055	0.009147485	0.051880265	0.00827009	0.093774742	0.00977523
Normal(0.35)	0.043767284	0.00819866	0.057441459	0.009123197	0.145055753	0.014966859
Exponential	0.082542462	0.024262091	0.181220407	0.031359683	0.38045752	0.046409054

Table 3-1: Relative Regression Errors of RCM Model for Mean Service Times that Vary Linearly with a Range of 2, Duration 5 Seconds

50 periods 5 seconds/period 75% server utilization 30 replications	3 Services		5 Services		10 Services	
	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time
	Exp(2.200)	0.100	Exp(3.667)	0.100	Exp(7.333)	0.100
	Exp(2.200)	0.550	Exp(3.667)	0.325	Exp(7.333)	0.200
	Exp(2.200)	1.000	Exp(3.667)	0.550	Exp(7.333)	0.300
			Exp(3.667)	0.775	Exp(7.333)	0.400
			Exp(3.667)	1.000	Exp(7.333)	0.500
					Exp(7.333)	0.600
					Exp(7.333)	0.700
					Exp(7.333)	0.800
					Exp(7.333)	0.900
					Exp(7.333)	1.000
Linear Service Time Pattern	Average Error	Confidence Interval	Average Error	Confidence Interval	Average Error	Confidence Interval
Constant	0.048219071	0.009136466	0.066257925	0.021726564	0.100950277	0.01877215
Normal(0.05)	0.064956458	0.106939196	0.070256355	0.014935193	0.105636315	0.00891635
Normal(0.15)	0.050878343	0.011187226	0.061663462	0.009647338	0.112393919	0.012540371
Normal(0.25)	0.061832948	0.019494718	0.073883439	0.011203689	0.16132291	0.081255613
Normal(0.35)	0.057102047	0.033443337	0.07597516	0.012991239	0.161987705	0.019982025
Exponential	0.130601538	0.024313188	0.2138575	0.029033611	0.343332689	0.035500497

Table 3-2: Relative Regression Errors of RCM Model for Mean Service Times that Vary Linearly with a Range of 10, Duration 5 Seconds

50 periods 60 seconds/period 75% server utilization 30 replications	3 Services		5 Services		10 Services	
	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time
	Exp(2.200)	0.100	Exp(3.667)	0.100	Exp(7.333)	0.100
	Exp(2.200)	0.550	Exp(3.667)	0.325	Exp(7.333)	0.200
	Exp(2.200)	1.000	Exp(3.667)	0.550	Exp(7.333)	0.300
			Exp(3.667)	0.775	Exp(7.333)	0.400
			Exp(3.667)	1.000	Exp(7.333)	0.500
					Exp(7.333)	0.600
					Exp(7.333)	0.700
					Exp(7.333)	0.800
					Exp(7.333)	0.900
					Exp(7.333)	1.000
Linear Service Time Pattern	Average Error	Confidence Interval	Average Error	Confidence Interval	Average Error	Confidence Interval
Constant	0.009783639	0.003102796	0.015197275	0.00285846	0.029922781	0.0028245
Normal(0.05)	0.012036724	0.00282982	0.018932337	0.002892252	0.034384945	0.002772586
Normal(0.15)	0.019770083	0.00549092	0.036143979	0.007628973	0.067899154	0.007114224
Normal(0.25)	0.02703543	0.010419015	0.054574252	0.007739309	0.107441752	0.011091542
Normal(0.35)	0.045961426	0.012479699	0.070868691	0.013897352	0.144855789	0.013557266
Exponential	0.119038649	0.028680695	0.195919841	0.023255653	0.448803139	0.041681967

Table 3-3: Relative Regression Errors of RCM Model for Mean Service Times that Vary Linearly with a Range of 10, Duration 60 Seconds

50 periods 300 seconds/period 75% server utilization 30 replications	3 Services		5 Services		10 Services	
	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time
	Exp(2.200)	0.100	Exp(3.667)	0.100	Exp(7.333)	0.100
	Exp(2.200)	0.550	Exp(3.667)	0.325	Exp(7.333)	0.200
	Exp(2.200)	1.000	Exp(3.667)	0.550	Exp(7.333)	0.300
			Exp(3.667)	0.775	Exp(7.333)	0.400
			Exp(3.667)	1.000	Exp(7.333)	0.500
					Exp(7.333)	0.600
					Exp(7.333)	0.700
					Exp(7.333)	0.800
					Exp(7.333)	0.900
					Exp(7.333)	1.000
Linear Service Time Pattern	Average Error	Confidence Interval	Average Error	Confidence Interval	Average Error	Confidence Interval
Constant	0.004422069	0.001229022	0.006660793	0.001018876	0.01265849	0.001264526
Normal(0.05)	0.007468858	0.002148574	0.010877399	0.00169645	0.023233748	0.002109643
Normal(0.15)	0.017499202	0.005822339	0.03255758	0.004541463	0.067685025	0.005989231
Normal(0.25)	0.035117397	0.010328853	0.05324178	0.008810263	0.095801481	0.011119234
Normal(0.35)	0.044472368	0.010125874	0.075795709	0.011698967	0.150603508	0.014793346
Exponential	0.104550336	0.031100964	0.189448996	0.029951836	0.400792764	0.034307627

Table 3-4: Relative Regression Errors of RCM Model for Mean Service Times that Vary Linearly with a Range of 10, Duration 300 Seconds

50 periods 5 seconds/period 75% server utilization 30 replications	3 Services		5 Services		10 Services	
	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time
	Exp(1.888)	0.100	Exp(2.875)	0.100	Exp(5.448)	0.100
	Exp(1.888)	0.316	Exp(2.875)	0.178	Exp(5.448)	0.129
	Exp(1.888)	1.000	Exp(2.875)	0.316	Exp(5.448)	0.167
			Exp(2.875)	0.562	Exp(5.448)	0.215
			Exp(2.875)	1.000	Exp(5.448)	0.278
					Exp(5.448)	0.359
					Exp(5.448)	0.464
					Exp(5.448)	0.599
					Exp(5.448)	0.774
					Exp(5.448)	1.000
Geometric Service Time Pattern	Average Error	Confidence Interval	Average Error	Confidence Interval	Average Error	Confidence Interval
Constant	0.050970893	0.009739465	0.076111329	0.012493874	0.116680982	0.013098868
Normal(0.05)	0.050947205	0.012518973	0.075735438	0.012902475	0.123406221	0.012829713
Normal(0.15)	0.065713183	0.015517366	0.081734381	0.011451591	0.129517823	0.014585974
Normal(0.25)	0.063395338	0.019930378	0.094366013	0.013279367	0.158805866	0.017104893
Normal(0.35)	0.07018432	0.017144069	0.116653096	0.018943183	0.172747496	0.017339805
Exponential	0.128484461	0.02352236	0.182133375	0.023962682	0.400883622	0.038772034

Table 3-5: Relative Errors of RCM Model for Mean Service Times that Vary Geometrically with a Range of 10 and Equal Inter-arrival Times, Duration 5 Seconds

50 periods 60 seconds/period 75% server utilization 30 replications	3 Services		5 Services		10 Services	
	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time
	Exp(1.888)	0.100	Exp(2.875)	0.100	Exp(5.448)	0.100
	Exp(1.888)	0.316	Exp(2.875)	0.178	Exp(5.448)	0.129
	Exp(1.888)	1.000	Exp(2.875)	0.316	Exp(5.448)	0.167
			Exp(2.875)	0.562	Exp(5.448)	0.215
			Exp(2.875)	1.000	Exp(5.448)	0.278
					Exp(5.448)	0.359
					Exp(5.448)	0.464
					Exp(5.448)	0.599
					Exp(5.448)	0.774
					Exp(5.448)	1.000
Geometric Service Time Pattern	Average Error	Confidence Interval	Average Error	Confidence Interval	Average Error	Confidence Interval
Constant	0.011123931	0.002964654	0.017663977	0.003060744	0.030298686	0.003733103
Normal(0.05)	0.013458371	0.004713812	0.019712289	0.002394724	0.041148189	0.003831533
Normal(0.15)	0.022648661	0.005804812	0.033184589	0.006185534	0.077637429	0.009021227
Normal(0.25)	0.034641311	0.009067949	0.061163056	0.009387403	0.119685217	0.011950861
Normal(0.35)	0.0428723	0.013880387	0.085562922	0.014704259	0.1581391	0.012496098
Exponential	0.085903808	0.022788631	0.260279451	0.040628873	0.462968466	0.052561773

Table 3-6: Relative Errors of RCM Model for Mean Service Times that Vary Geometrically with a Range of 10 and Equal Inter-arrival Times, Duration 60 Seconds

50 periods 300 seconds/period 75% server utilization 30 replications	3 Services		5 Services		10 Services	
	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time
	Exp(1.888)	0.100	Exp(2.875)	0.100	Exp(5.448)	0.100
	Exp(1.888)	0.316	Exp(2.875)	0.178	Exp(5.448)	0.129
	Exp(1.888)	1.000	Exp(2.875)	0.316	Exp(5.448)	0.167
			Exp(2.875)	0.562	Exp(5.448)	0.215
Geometric Service Time Pattern	Average Error	Confidence Interval	Average Error	Confidence Interval	Average Error	Confidence Interval
	0.004126337	0.001226436	0.008200346	0.001195014	0.014869565	0.001733788
	0.008421729	0.002341636	0.015019646	0.002511325	0.026402851	0.002855097
	0.021463414	0.006351707	0.031118136	0.006666209	0.069108682	0.008438385
	0.032364777	0.007982463	0.060960956	0.009676663	0.108692278	0.011772235
	0.042782691	0.010173516	0.076815297	0.012778269	0.13922716	0.01396321
Exponential	0.117333962	0.029491195	0.205883347	0.029494296	0.415024061	0.048381851

Table 3-7: Relative Errors of RCM Model for Mean Service Times that Vary Geometrically with a Range of 10 and Equal Inter-arrival Times, Duration 300 Seconds

50 periods 5 seconds/period 75% server utilization 30 replications	3 Services		5 Services		10 Services	
	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time
	Exp(1.734)	0.100	Exp(2.442)	0.100	Exp(3.702)	0.100
	Exp(1.825)	0.316	Exp(2.571)	0.178	Exp(3.897)	0.129
	Exp(1.927)	1.000	Exp(2.713)	0.316	Exp(4.113)	0.167
			Exp(2.873)	0.562	Exp(4.355)	0.215
Varying Mean Interarrival Times	Average Error	Confidence Interval	Average Error	Confidence Interval	Average Error	Confidence Interval
	0.057726661	0.012421885	0.073492573	0.009659167	0.10479525	0.011517065
	0.076297467	0.016412847	0.068540699	0.009382171	0.11888668	0.011727325
	0.057107935	0.011054203	0.079162484	0.0093364	0.134359151	0.014139688
	0.03911269	0.010800706	0.088024555	0.013463621	0.154211941	0.017958888
	0.064034304	0.015228124	0.095299805	0.016260718	0.184585864	0.022502806
Exponential	0.183277052	0.032444597	0.233840895	0.032624192	0.40660093	0.042755114

Table 3-8: Relative Errors of RCM Model for Mean Service Times that Vary Geometrically with a Range of 10 and Increasing Inter-arrival Times, Duration 5 Seconds

50 periods 60 seconds/period 75% server utilization 30 replications	3 Services		5 Services		10 Services	
	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time
	Exp(1.734)	0.100	Exp(2.442)	0.100	Exp(3.702)	0.100
	Exp(1.825)	0.316	Exp(2.571)	0.178	Exp(3.897)	0.129
	Exp(1.927)	1.000	Exp(2.713)	0.316	Exp(4.113)	0.167
			Exp(2.873)	0.562	Exp(4.355)	0.215
			Exp(3.053)	1.000	Exp(4.628)	0.278
					Exp(4.936)	0.359
					Exp(5.289)	0.464
					Exp(5.696)	0.599
					Exp(6.170)	0.774
					Exp(6.731)	1.000
Varying Mean Interarrival Times	Average Error	Confidence Interval	Average Error	Confidence Interval	Average Error	Confidence Interval
Constant	0.009990768	0.002872015	0.018560299	0.002741843	0.033044611	0.003643295
Normal(0.05)	0.012794562	0.003580988	0.021365123	0.002788121	0.03689879	0.004031785
Normal(0.15)	0.023088092	0.005954955	0.037841786	0.007891715	0.085534943	0.008882659
Normal(0.25)	0.031224372	0.010186105	0.070883543	0.012093144	0.117113664	0.011053324
Normal(0.35)	0.044375267	0.014506053	0.085090743	0.010771496	0.15308951	0.011141908
Exponential	0.151538437	0.041614156	0.230913844	0.029912379	0.470646795	0.035240609

Table 3-9: Relative Errors of RCM Model for Mean Service Times that Vary Geometrically with a Range of 10 and Increasing Inter-arrival Times, Duration 60 Seconds

50 periods 300 seconds/period 75% server utilization 30 replications	3 Services		5 Services		10 Services	
	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time
	Exp(1.734)	0.100	Exp(2.442)	0.100	Exp(3.702)	0.100
	Exp(1.825)	0.316	Exp(2.571)	0.178	Exp(3.897)	0.129
	Exp(1.927)	1.000	Exp(2.713)	0.316	Exp(4.113)	0.167
			Exp(2.873)	0.562	Exp(4.355)	0.215
			Exp(3.053)	1.000	Exp(4.628)	0.278
					Exp(4.936)	0.359
					Exp(5.289)	0.464
					Exp(5.696)	0.599
					Exp(6.170)	0.774
					Exp(6.731)	1.000
Varying Mean Interarrival Times	Average Error	Confidence Interval	Average Error	Confidence Interval	Average Error	Confidence Interval
Constant	0.005328321	0.00156445	0.007991193	0.001666211	0.013464519	0.001449157
Normal(0.05)	0.008924478	0.002341695	0.014767757	0.002502621	0.024351635	0.002137679
Normal(0.15)	0.019735594	0.005605441	0.034602583	0.005001177	0.059029529	0.004758408
Normal(0.25)	0.028237602	0.007267473	0.062052006	0.010639629	0.108620804	0.009336736
Normal(0.35)	0.045866332	0.014391745	0.074663201	0.012181213	0.148583262	0.016567079
Exponential	0.110860942	0.036982471	0.241060029	0.037421829	0.455396608	0.038996053

Table 3-10: Relative Errors of RCM Model for Mean Service Times that Vary Geometrically with a Range of 10 and Increasing Inter-arrival Times, Duration 300 Seconds

Total Time: 150 mins Server Utilization: 75% Replications: 30	3 Services	Constant	10 Services	Constant	10 Services	Normal(0.05)	
	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time	
	Exp(2.200)	0.100	Exp(5.448)	0.100	Exp(5.448)	0.100	
	Exp(2.200)	0.550	Exp(5.448)	0.129	Exp(5.448)	0.129	
	Exp(2.200)	1.000	Exp(5.448)	0.167	Exp(5.448)	0.167	
			Exp(5.448)	0.215	Exp(5.448)	0.215	
			Exp(5.448)	0.278	Exp(5.448)	0.278	
			Exp(5.448)	0.359	Exp(5.448)	0.359	
			Exp(5.448)	0.464	Exp(5.448)	0.464	
			Exp(5.448)	0.599	Exp(5.448)	0.599	
			Exp(5.448)	0.774	Exp(5.448)	0.774	
			Exp(5.448)	1.000	Exp(5.448)	1.000	
Number of Reporting Periods		Average Error	Confidence Interval	Average Error	Confidence Interval	Average Error	Confidence Interval
13	0.01105655	0.00339824					
25	0.00647895	0.001619309	0.04130844	0.007587671	0.09912256	0.018430536	
50	0.00549393	0.001294138	0.01746596	0.001789083	0.03201989	0.002359843	
75	0.00480888	0.001376774	0.01497996	0.001612089	0.02204168	0.002482763	
100	0.00479082	0.001263913	0.01562962	0.001176381	0.02009974	0.001793407	
125	0.00479698	0.001665745	0.01383929	0.001401188	0.02003262	0.002375466	
150	0.00510118	0.001483442	0.01464597	0.001577108	0.01835531	0.00181251	

Total Time: 150 mins Server Utilization: 75% Replications: 30	10 Services	Normal(0.15)	10 Services	Normal(0.25)	10 Services	Normal(0.35)	10 Services	Exponential
	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time
	Exp(5.448)	0.100	Exp(5.448)	0.100	Exp(5.448)	0.100	Exp(5.448)	0.100
	Exp(5.448)	0.129	Exp(5.448)	0.129	Exp(5.448)	0.129	Exp(5.448)	0.129
	Exp(5.448)	0.167	Exp(5.448)	0.167	Exp(5.448)	0.167	Exp(5.448)	0.167
	Exp(5.448)	0.215	Exp(5.448)	0.215	Exp(5.448)	0.215	Exp(5.448)	0.215
	Exp(5.448)	0.278	Exp(5.448)	0.278	Exp(5.448)	0.278	Exp(5.448)	0.278
	Exp(5.448)	0.359	Exp(5.448)	0.359	Exp(5.448)	0.359	Exp(5.448)	0.359
	Exp(5.448)	0.464	Exp(5.448)	0.464	Exp(5.448)	0.464	Exp(5.448)	0.464
	Exp(5.448)	0.599	Exp(5.448)	0.599	Exp(5.448)	0.599	Exp(5.448)	0.599
	Exp(5.448)	0.774	Exp(5.448)	0.774	Exp(5.448)	0.774	Exp(5.448)	0.774
	Exp(5.448)	1.000	Exp(5.448)	1.000	Exp(5.448)	1.000	Exp(5.448)	1.000
Number of Reporting Periods	Average Error	Confidence Interval	Average Error	Confidence Interval	Average Error	Confidence Interval	Average Error	Confidence Interval
13								
25	0.38885594	0.335648253	0.33938849	0.0425304	0.45134045	0.086621851	1.59075045	0.261158119
50	0.07182047	0.006387789	0.12031437	0.014286265	0.15763432	0.01209955	0.40974349	0.042689132
75	0.05316459	0.00483626	0.08254747	0.008107684	0.10772057	0.010479319	0.31847356	0.027856874
100	0.04265063	0.00446942	0.07187047	0.009590009	0.09227385	0.010898738	0.26200713	0.029308349
125	0.03918899	0.003961569	0.06164483	0.005164386	0.08185165	0.008594661	0.22770855	0.023413126
150	0.03100094	0.002928601	0.05381961	0.006280978	0.07301368	0.007853909	0.19667969	0.014710045

Table 3-11: Relative Regression Errors of RCM Model for Different Number of Reporting Periods

Relative Regression Errors of RCM Model for Mean Service Times that Vary Linearly with a Range of 2, Duration 5 Seconds

Figure 3-1: Relative Regression Errors of RCM Model for Mean Service Times that Vary Linearly with a Range of 2, Duration 5 Seconds

Relative Regression Errors of RCM Model for Mean Service Times that Vary Linearly with a Range of 10, Duration 5 Seconds

Figure 3-2: Relative Regression Errors of RCM Model for Mean Service Times that Vary Linearly with a Range of 10, Duration 5 Seconds

Relative Regression Errors of RCM Model for Mean Service Times that Vary Linearly with a Range of 10,
Duration 60 Seconds

Figure 3-3: Relative Regression Errors of RCM Model for Mean Service Times that Vary Linearly with a Range of 10, Duration 60 Seconds

Relative Regression Errors of RCM Model for Mean Service Times that Vary Linearly with a Range of 10,
Duration 300 Seconds

Figure 3-4: Relative Regression Errors of RCM Model for Mean Service Times that Vary Linearly with a Range of 10, Duration 300 Seconds

Figure 3-5: Relative Errors of RCM Model for Mean Service Times that Vary Geometrically with a Range of 10 and Equal Inter-arrival Times, Duration 5 Seconds

Figure 3-6: Relative Errors of RCM Model for Mean Service Times that Vary Geometrically with a Range of 10 and Equal Inter-arrival Times, Duration 60 Seconds

Figure 3-7: Relative Errors of RCM Model for Mean Service Times that Vary Geometrically with a Range of 10 and Equal Inter-arrival Times, Duration 300 Seconds

Figure 3-8: Relative Regression Errors of RCM Model for Mean Service Times that Vary Geometrically with a Range of 10 and Increasing Inter-arrival Times, Duration 5 Seconds

Figure 3-9: Relative Regression Errors of RCM Model for Mean Service Times that Vary Geometrically with a Range of 10 and Increasing Inter-arrival Times, Duration 60 Seconds

Figure3-10: Relative Regression Errors of RCM Model for Mean Service Times that Vary Geometrically with a Range of 10 and Increasing Inter-arrival Times, Duration 300 Seconds

Relative Regression Errors of RCM Model for Different Numbers of Reporting Periods

Figure 3-11: Relative Regression Errors of RCM Model for Different Number of Reporting Periods

Chapter 4 Maximum Likelihood Estimation of Generalized Regression Model

This chapter describes the generalized regression model and the algorithm, the maximum likelihood estimation (MLE), with which to estimate coefficients for this model. One simulated case is used to test the MLE statistical technique to see whether it is worth doing further research or not.

4.1. Generalized Regression Model [9]

In chapter 2, we discussed the multiple linear regression model. The generalized regression model is basically the same model but its disturbance that violates the classical assumptions. It is given below:

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \\ E[\boldsymbol{\varepsilon}] &= \mathbf{0}, \\ E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'] &= \sigma^2 \boldsymbol{\Omega}, \end{aligned} \tag{4-1}$$

where $\boldsymbol{\Omega}$ is a positive definite matrix. As we know, resource demands in our problem are not constant which violates the assumption of the multiple linear regression model. The generalized regression model could fit our problem better than the multiple linear regression model.

We only consider heteroscedasticity though autocorrelation is important in some cases. Disturbance is heteroscedastic when it has different variance and usually happens in cross-section data where the scale of the dependent variable and the explanatory power of the model tend to vary across observations. The disturbance is still thought to be uncorrelated across observations, therefore $\boldsymbol{\Omega}$ is:

$$\sigma^2 \Omega = \begin{bmatrix} \sigma_1^2 & 0 & . & . & . & 0 \\ 0 & \sigma_2^2 & . & . & . & 0 \\ & & . & & & \\ & & . & & & \\ & & . & & & \\ 0 & 0 & . & . & . & \sigma_n^2 \end{bmatrix} \quad (4-2)$$

4.2 Maximum Likelihood Estimation

To obtain the estimation of β in the general regression model needs to know Ω . To keep the question simple, we assume Ω is a known, symmetric, positive definite matrix. However, in most case, Ω is unknown and must also be estimated [10].

If the disturbances are multivariate normally distributed, the log-likelihood function of the sample is given below:

$$\ln L = -\frac{n}{2} \ln(2\pi) - \frac{1}{2} \ln |\sigma^2 \Omega| - \frac{1}{2} \varepsilon' (\sigma^2 \Omega)^{-1} \varepsilon \quad (4-3)$$

Using $\varepsilon = \mathbf{y} - \mathbf{X}\beta$ to substitute ε , we get

$$\ln L = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\beta)' \Omega^{-1} (\mathbf{y} - \mathbf{X}\beta) - \frac{1}{2} \ln |\Omega| \quad (4-4)$$

Because Ω is a matrix of known constants, the maximum likelihood estimator of β is the vector that minimizes the generalized sum of squares. This can be expressed below:

$$S_*(\beta) = (\mathbf{y} - \mathbf{X}\beta)' \Omega^{-1} (\mathbf{y} - \mathbf{X}\beta) \quad (4-5)$$

The necessary conditions for maximizing L are:

$$\frac{\partial \ln L}{\partial \beta} = \frac{1}{\sigma^2} \mathbf{X}' \Omega^{-1} (\mathbf{y} - \mathbf{X}\beta) = \frac{1}{\sigma^2} \mathbf{X}' (\mathbf{y} - \mathbf{X}\beta) \quad (4-6)$$

and

$$\begin{aligned} \frac{\partial \ln L}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} (\mathbf{y} - \mathbf{X}\beta)' \Omega^{-1} (\mathbf{y} - \mathbf{X}\beta) \\ &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} (\mathbf{y} - \mathbf{X}\beta)' (\mathbf{y} - \mathbf{X}\beta). \end{aligned} \quad (4-7)$$

where

$$\mathbf{y}_. = \mathbf{P}\mathbf{y}$$

$$\mathbf{X}_. = \mathbf{P}\mathbf{X}$$

$$\mathbf{P} = \mathbf{\Omega}^{-\frac{1}{2}}$$

The solutions are:

$$\hat{\boldsymbol{\beta}}_{ML} = (\mathbf{X}_.'\mathbf{X}_.)^{-1}\mathbf{X}_.'\mathbf{y}_. = (\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{y} \quad (4-8)$$

$$\begin{aligned} \hat{\sigma}_{ML}^2 &= \frac{(\mathbf{y}_. - \mathbf{X}_.\hat{\boldsymbol{\beta}})'(\mathbf{y}_. - \mathbf{X}_.\hat{\boldsymbol{\beta}})}{n} \\ &= \frac{(\mathbf{y}_. - \mathbf{X}_.\hat{\boldsymbol{\beta}})'\mathbf{\Omega}^{-1}(\mathbf{y}_. - \mathbf{X}_.\hat{\boldsymbol{\beta}})}{n} \end{aligned} \quad (4-9)$$

All the solutions above are based on the condition that $\mathbf{\Omega}$ is known. What happens if $\mathbf{\Omega}$ is not known? First, to make the computation simple, we introduce a new matrix $\mathbf{\Gamma}$, which is defined as below [11]:

$$\mathbf{\Gamma} = \mathbf{\Omega}^{-1}$$

Also, we can use the results

$$\log|\mathbf{\Gamma}| = -\log|\mathbf{\Omega}|,$$

$$\frac{\partial \log|\mathbf{\Gamma}|}{\partial \mathbf{\Gamma}} = \mathbf{\Gamma}^{-1},$$

$$\frac{\partial \boldsymbol{\varepsilon}'\mathbf{\Gamma}\boldsymbol{\varepsilon}}{\partial \mathbf{\Gamma}} = \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'.$$

Replacing $\mathbf{\Omega}$ with $\mathbf{\Gamma}^{-1}$ in equation (4-3), we can get

$$\ln L = -\frac{n}{2}(\ln 2\pi + \ln \sigma^2) - \frac{1}{2\sigma^2}\boldsymbol{\varepsilon}'\mathbf{\Gamma}\boldsymbol{\varepsilon} + \frac{1}{2}\ln|\mathbf{\Gamma}| \quad (4-10)$$

We do the derivation to $\boldsymbol{\beta}$, σ^2 , L and get the equations, which are the necessary conditions to maximum likelihood. Those equations are given below:

$$\frac{\partial \ln L}{\partial \boldsymbol{\beta}} = \frac{1}{\sigma^2}\mathbf{X}'\mathbf{\Gamma}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \quad (4-11)$$

$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' \Gamma (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \quad (4-12)$$

$$\begin{aligned} \frac{\partial \ln L}{\partial \Gamma} &= \frac{1}{2} \left[\Gamma^{-1} - \left(\frac{1}{\sigma^2} \right) \varepsilon \varepsilon' \right] \\ &= \frac{1}{2\sigma^2} \left[\sigma^2 \Omega - \varepsilon \varepsilon' \right] \end{aligned} \quad (4-13)$$

We have to restrict Ω in some way to make the estimation problem practical. Let assume, that $\Omega = \Omega(\theta)$, where θ is a vector of a small number of parameters and is not a function of any elements of β .

Using equation (4-13), for a given set of β and σ^2 , it is straightforward to get estimates of θ . Oberhofer and Kmenta proposed the iterative two-step method in 1974. Given an estimate $\hat{\theta}$, we use step1 (4-11) and step2 (4-12) to estimate β and σ^2 . With the new estimates β and σ^2 , we can reestimate θ by equation (4-13). Then we repeat the iteration until satisfactory convergence has been gained. If the initial $\hat{\theta}$ is a consistent estimator of θ , it will be asymptotically efficient at every step.

Now, Let us discuss groupwise heteroscedasticity. A groupwise heteroscedasticity regression model is given below:

$$\begin{aligned} y_i &= \boldsymbol{\beta}' \mathbf{x}_i + \varepsilon_i, \quad i = 1, \dots, n \\ E[\varepsilon_i] &= 0, \quad i = 1, \dots, n \end{aligned} \quad (4-15)$$

The n observations are divided into G groups, each with n_g observations. The variance within group g is:

$$\text{Var}[\varepsilon_{ig}] = \sigma_g^2, \quad i = 1, \dots, n_g.$$

The log-likelihood function is

$$\ln L = -\frac{n}{2} \ln(2\pi) - \frac{1}{2} \sum_g n_g \ln \sigma_g^2 - \frac{1}{2} \sum_g \sum_{i=1}^{n_g} \frac{\varepsilon_{ig}^2}{\sigma_g^2} \quad (4-16)$$

To get the estimates of parameters β , we need to know Ω . The weighting matrix is:

$$\Omega = \begin{bmatrix} \sigma_1^2 \mathbf{I} & \mathbf{0} & . & . & . & \mathbf{0} \\ \mathbf{0} & \sigma_2^2 \mathbf{I} & . & . & . & \mathbf{0} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ \mathbf{0} & \mathbf{0} & . & . & . & \sigma_G^2 \mathbf{I} \end{bmatrix}, \quad (4-17)$$

Therefore the GLS estimator is:

$$\hat{\beta} = \left[\sum_g \frac{1}{\hat{\sigma}_g^2} \mathbf{X}_g' \mathbf{X}_g \right]^{-1} \left[\sum_g \frac{1}{\hat{\sigma}_g^2} \mathbf{X}_g' \mathbf{y}_g \right] \quad (4-18)$$

For each variance parameter, the likelihood equation is:

$$\frac{\partial \ln L}{\partial \sigma_g^2} = -\frac{n_g}{2\sigma_g^2} + \frac{1}{2\sigma_g^4} \sum_i \varepsilon_{ig}^2$$

So that given the current value of $\hat{\beta}$,

$$\hat{\sigma}_g^2 = \frac{\mathbf{e}_g' \mathbf{e}_g}{n_g} \quad (4-19)$$

To get this maximum likelihood estimator, an iterative procedure that satisfies the Oberhofer-Kmenta requirements is given below:

1. Pool the data and estimate β by Ordinary Least Square regression (OLS).
2. Estimate the disturbance variances respectively with $\mathbf{e}_g' \mathbf{e}_g / n_g$, where $\mathbf{e}_g = \mathbf{y}_g - \mathbf{X}_g \mathbf{b}$.
3. Compute $\hat{\beta}$
4. If $\hat{\beta}$ has not converged, go to step 2; otherwise, exit.

4.3 Results of Study

To run the MLE technique, we need to do something similar to the RCM. Divide whole data into a number of buckets of data. Different bucket size gives different result.

Also, we need a set of initial values. The initial values don't effect the final result but only effect the time needed to approach the final result. We can use OLS results as the initial values.

Since the time is limited, we only study one case. Parameters of this case are 3 services, Linear Service Time Pattern, range of 10, equal inter-arrival times and duration of 60 seconds. We only run it in one replications and run OLS at the same time. The purpose is to see whether there is a trend that the MLE give much better result than the OLS, if it is, we will do the research further. The results are given in Table 4-1 and the bucket size is 10.

Service Time Distribution	E_{OLS}	E_{MLE}
Constant	0.005613	0.004208
Normal(0.05)	0.02179	0.02386
Normal(0.15)	0.05611	0.06394
Normal(0.25)	0.003474	0.02728
Normal(0.35)	0.03783	0.06068
Exponential	0.1736	0.2187

Table 4-1: Reported Errors for OLS and MLE Technique.

From Table 4-1, we can see that in general MLE does not give a better result than OLS. But this is the only result with bucket size 10. What if bucket size changes?

We choose the exponential case in the Table 4-1 and vary the bucket size. The result is given in Table 4-1. From Table 4-2, we can see that all bucket sizes but bucket size 15 gives the worse accuracy than OLS.

Bucket Size	OLS	MLE
4	0.1736	0.2760
5	0.1736	0.5832
6	0.1736	0.3780
7	0.1736	0.2278
8	0.1736	0.2516
9	0.1736	0.2462
10	0.1736	0.2187
15	0.1736	0.1714
25	0.1736	0.2098

Table 4-2: Reported Errors of MLE for Different Bucket Size.

In most case, this method is even worse than OLS. So no more research is needed.

Chapter 5. Restricted Least Square Regression

This chapter describes the restricted least square regression and its algorithm. The restricted least square regression (RLS) technique is assessed using the same experiments as in Section 2.2.2. Besides that, the impact of the perturbation of ratios for resource demands is explored.

5.1 Algorithm [12]

In the former chapters, we introduced three statistical techniques and the accuracy of these tools is not good enough. If we know some linear relationship exists among resource demands, is there any statistical model that can deal with this special case? The answer is positive and the model is the restricted least square regression.

Let's recall the multiple regression model. If some restrictions are imposed to this model, we would call it restricted multiple regression model. The model is described as below:

$$\mathbf{y} = \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (5-1)$$

A set of linear restrictions given below are imposed to the model (equation 5-1)

$$\begin{aligned} r_{11}\beta_1 + r_{12}\beta_2 + \dots + r_{1K}\beta_K &= q_1 \\ r_{21}\beta_1 + r_{22}\beta_2 + \dots + r_{2K}\beta_K &= q_2 \\ &\vdots \\ r_{J1}\beta_1 + r_{J2}\beta_2 + \dots + r_{JK}\beta_K &= q_J. \end{aligned}$$

These equations can be combined into a single equation:

$$\mathbf{R}\boldsymbol{\beta} = \mathbf{q}. \quad (5-2)$$

where the matrix $\mathbf{R}$ must have K columns to be conformable with $\boldsymbol{\beta}$, J rows for a total of J restrictions. Each row is the coefficients in one of the restrictions. Also $\mathbf{R}$ must have full row rank, which means that J must be less than or equal to K . The rows of $\mathbf{R}$ must be linear independent.

In our case, we can use ratios between mean service times to set up linear restrictions and can make these restrictions linear independent. Also we can choose more observations than the number of linear restrictions. Generally speaking, we can meet those assumptions of RLS.

How to solve this model? Recall the multiple linear regression model we discussed in Chapter 2, we used the least squares regression, that is, choose the β which can minimize the sum of squared residuals

$$\text{Minimize}_{\tilde{\beta}} S(\tilde{\beta}) = (y - X\tilde{\beta})' (y - X\tilde{\beta}) \quad (5-3)$$

To solve this restricted multiple model, we just choose the β which can minimize the sum of squared residuals (equation 5-3) and also subject to the set of linear restrictions (equation 5-2).

A Lagrangean solution for this problem can be described as below:

$$\mathbf{b}_* = \text{Min}_{\tilde{\beta}} S^*(\tilde{\beta}) = (y - X\tilde{\beta})' (y - X\tilde{\beta}) + 2\lambda' (R\tilde{\beta} - q). \quad (5-4)$$

where the solution $\mathbf{b}_*$ and λ will satisfy the necessary conditions:

$$\frac{\partial S^*}{\partial \mathbf{b}_*} = 2X'(y - X\mathbf{b}_*) + 2R'\lambda = 0, \quad (5-5)$$

$$\frac{\partial S^*}{\partial \lambda} = 2(R\mathbf{b}_* - q) = 0. \quad (5-6)$$

To get $\mathbf{b}_*$, we use $X'y = X'Xb$ to make the substitution in equation (5-5), where $\mathbf{b}$ is the unrestricted least squares coefficient vector. Then a result for $\mathbf{b}_*$ in terms of $\mathbf{b}$ is given below:

$$\mathbf{b}_* = \mathbf{b} + (X'X)^{-1} R'\lambda. \quad (5-7)$$

Multiplying both sides of the above equation by $\mathbf{R}$ and using the restrictions $R\mathbf{b}_* = q$, we can get a solution for λ :

$$\lambda = -[R(X'X)^{-1} R']^{-1} (R\mathbf{b} - q) \quad (5-8)$$

Finally, we can use the result of equation (5-8) to replace λ in equation (5-7) and get the solution as below:

$$\mathbf{b}_* = \mathbf{b} - (\mathbf{X}'\mathbf{X})^{-1} \mathbf{R} [\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}]^{-1} (\mathbf{R}\mathbf{b} - \mathbf{q}) \quad (5-9)$$

5.2 Results of Study

To run the RLS, we need to know estimates for the ratio of per-service cost values. These ratios can be expressed as linear constraints. For example, Table 5-1, three service with linear service time pattern in a range 2. Supposed that three service time is c_1 , c_2 and c_3 where $c_2 = 1.5 c_1$ and $c_3 = 2 c_1$. The linearly independent constraints can be given as below:

$$c_2 - 1.5c_1 = 0$$

$$c_3 - 2c_1 = 0$$

We use the simulations in Section 2.2.2. Tables 5-1 to 5-11 and Figures 5-1 to 5-11, which correspond to the tables respectively, are given to describe the results of these simulations. Besides these, two extra cases are used. Tables 5-12 and 5-13 where the ratios of means are perturbed ± 10 percent and ± 25 percent express these two cases respectively. Figure 5-12 and 5-13 corresponds to Table 5-12 and 5-13 separately. Their parameter settings are Geometric Mean Service Time Pattern, equal arrival rates, range of 10 and duration of 300 seconds. In addition those four questions proposed in section 2.2.2, we bring another question as below:

5. What is the sensitivity of the RLS to estimates for ratios?

The RLS shows the same behaviors as the OLS in some aspects. It is because that the RLS uses the parameters estimated by the OLS as input. Some factors that have impacts on the OLS could effect the RLS in the same manner by the OLS. So the explanation for the OLS is also suited for the RLS.

Let's talk about these in detail. Service time distribution is related to the accuracy of the RLS technique. However, RLS does remarkably well even for the 10 service

exponentially distributed service time cases. Its E_{rel} (duration 300 seconds) remains below 2% whereas the other techniques have over 35% error. The reason is that the known ratios effectively reduce the problem to that of solving for a single variable.

Parameter settings in the tables from 5-2 to 5-4 are Linear Service Time Pattern and Geometric Service Time Pattern for the tables from 5-5 to 5-7. From results given, we can see that the mean service time pattern has no effect on the RLS technique. Table 5-1 and Table 5-2 have the same parameter settings but range. The range for tables, 5-1 and 5-2, is 2 and 10 respectively. We can see from the experimental results that the accuracy of the RLS goes down as the range moving from 2 to 10. This case is due to end effects for these set of factor loads.

Arrival rate pattern of the tables from 5-5 to 5-7 is equal and increasing for the tables from 5-8 to 5-10. The other parameters are the same. It is clear from those tables that the arrival rate pattern doesn't cause any significant change of the accuracy of the RLS.

Tables 5-2 to 5-4 have duration of 5, 60 and 300 seconds respectively. Other parameters are identical. It is similar for the Tables 5-5 to 5-7 and the ones from 5-8 to 5-10. It can be seen that as the duration increases, the reported errors of the RLS diminish. Table 5-11 has the fixed total measurement time and varied number of reporting periods. We can see that as the number of reporting period increases, there is no significant change in the accuracy. It can be explained as the variation between periods has no impact on the RLS regression and also the minimum reporting duration which is one minute long is good enough to remove the end effects.

To investigate the sensitivity of RLS to the ratios of mean service times we randomly perturb each of the 30 replications' mean service time ratios as input to RLS by ± 10 percent and ± 25 percent. For instance, if the ratio between c_1 and c_3 is 2 and the perturbation is 10% then the ratio is chosen uniformly between 1.8 and 2.2 using a pseudo-random number generator. Each run had a separate set of ratios. From Table 5-12 and Table 5-13, we can see that the perturbations have a greater impact than service time

distribution. However the reported errors remain low 2% to 4% for ± 10 percent perturbation and 5% to 11% for ± 25 percent perturbation.

5.3 Summary

This chapter introduced the restricted least square regression and its algorithm. Total 13 simulations were used to test the ability of the RLS technique.

In general the RLS technique can provide much better results than OLS and RCM while the service time is not constant. However, to do RLS, we need to know some ratios of parameters. The ratios don't have to be too accurate since RLS is robust with respect to errors in ratios. Besides these, RLS technique shows some characteristics, which are the same as the OLS and RCM. Those characteristics include effect of service time distributions, effect of mean service time pattern and the range of mean service times across different application services and effect of arrival rate pattern. Notice that keeping the total measurement time as a constant, to increase the number of reporting period has little impact on the accuracy of the RLS technique. This point is different from OLS and RCM.

The RLS gives the satisfying accuracy of estimating the resource demands of each application service though it needs information of the ratio of resource demands. It is possible to estimate these ratios in real systems. Section 6.2 describes this issue. In conclusion, the RLS technique enables the development of software models [13][14] that characterize the resource demands of many services within application processes.

50 periods 5 seconds/period 75% server utilization 30 replications	3 Services		5 Services		10 Services	
	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time
	Exp(2.200)	0.100	Exp(3.667)	0.100	Exp(7.333)	0.100
	Exp(2.200)	0.150	Exp(3.667)	0.125	Exp(7.333)	0.111
	Exp(2.200)	0.200	Exp(3.667)	0.150	Exp(7.333)	0.122
			Exp(3.667)	0.175	Exp(7.333)	0.133
			Exp(3.667)	0.200	Exp(7.333)	0.144
					Exp(7.333)	0.156
					Exp(7.333)	0.167
					Exp(7.333)	0.178
					Exp(7.333)	0.189
					Exp(7.333)	0.200
Linear Service Time Pattern	Average Error	Confidence Interval	Average Error	Confidence Interval	Average Error	Confidence Interval
Constant	7.46E-04	1.49E-04	5.66E-04	1.56E-04	9.21E-04	1.74E-04
Normal(0.05)	0.00166866	3.38E-04	0.002359464	2.85E-04	0.00359109	3.42E-04
Normal(0.15)	0.005250454	7.85E-04	0.006122042	8.04E-04	0.009997228	9.06E-04
Normal(0.25)	0.008251921	0.001572422	0.010840716	0.001036055	0.016973186	0.001677648
Normal(0.35)	0.010154198	0.001759261	0.014417697	0.002084769	0.023844374	0.001693985
Exponential	0.028558164	0.004519605	0.04801092	0.005529974	0.071161233	0.007121588

Table 5-1: Relative Regression Errors of RLS Model for Mean Service Times that Vary Linearly with a Range of 2, Duration 5 Seconds

50 periods 5 seconds/period 75% server utilization 30 replications	3 Services		5 Services		10 Services	
	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time
	Exp(2.200)	0.100	Exp(3.667)	0.100	Exp(7.333)	0.100
	Exp(2.200)	0.550	Exp(3.667)	0.325	Exp(7.333)	0.200
	Exp(2.200)	1.000	Exp(3.667)	0.550	Exp(7.333)	0.300
			Exp(3.667)	0.775	Exp(7.333)	0.400
			Exp(3.667)	1.000	Exp(7.333)	0.500
					Exp(7.333)	0.600
					Exp(7.333)	0.700
					Exp(7.333)	0.800
					Exp(7.333)	0.900
					Exp(7.333)	1.000
Linear Service Time Pattern	Average Error	Confidence Interval	Average Error	Confidence Interval	Average Error	Confidence Interval
Constant	0.008192206	0.001284105	0.007591019	0.00123879	0.008012691	0.001303781
Normal(0.05)	0.009518955	0.001591839	0.007827737	0.001049694	0.010394442	0.001007374
Normal(0.15)	0.011987494	0.001715585	0.01368366	0.001674836	0.019371384	0.001734243
Normal(0.25)	0.01788291	0.002395299	0.02381772	0.002917912	0.033885863	0.002904695
Normal(0.35)	0.025074267	0.00285866	0.029763412	0.002719375	0.049214092	0.005000187
Exponential	0.112379195	0.012383841	0.119756555	0.01352916	0.143125032	0.012151121

Table 5-2: Relative Regression Errors of RLS Model for Mean Service Times that Vary Linearly with a Range of 10, Duration 5 Seconds

50 periods 60 seconds/period 75% server utilization 30 replications	3 Services		5 Services		10 Services	
	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time
	Exp(2.200)	0.100	Exp(3.667)	0.100	Exp(7.333)	0.100
	Exp(2.200)	0.550	Exp(3.667)	0.325	Exp(7.333)	0.200
	Exp(2.200)	1.000	Exp(3.667)	0.550	Exp(7.333)	0.300
			Exp(3.667)	0.775	Exp(7.333)	0.400
			Exp(3.667)	1.000	Exp(7.333)	0.500
					Exp(7.333)	0.600
					Exp(7.333)	0.700
					Exp(7.333)	0.800
					Exp(7.333)	0.900
					Exp(7.333)	1.000
Linear Service Time Pattern	Average Error	Confidence Interval	Average Error	Confidence Interval	Average Error	Confidence Interval
Constant	1.89E-04	3.36E-05	2.26E-04	5.06E-05	2.61E-04	8.51E-05
Normal(0.05)	9.01E-04	2.18E-04	0.001183399	2.44E-04	0.001671024	1.82E-04
Normal(0.15)	0.002722195	6.59E-04	0.003279234	4.97E-04	0.005480565	5.77E-04
Normal(0.25)	0.003639893	9.24E-04	0.006242813	8.13E-04	0.009022521	9.99E-04
Normal(0.35)	0.005569941	0.001083682	0.008653481	0.001296691	0.013896868	0.001411463
Exponential	0.015406693	0.003236045	0.026671111	0.003633187	0.035105191	0.003818766

Table 5-3: Relative Regression Errors of RLS Model for Mean Service Times that Vary Linearly with a Range of 10, Duration 60 Seconds

50 periods 300 seconds/period 75% server utilization 30 replications	3 Services		5 Services		10 Services	
	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time
	Exp(2.200)	0.100	Exp(3.667)	0.100	Exp(7.333)	0.100
	Exp(2.200)	0.550	Exp(3.667)	0.325	Exp(7.333)	0.200
	Exp(2.200)	1.000	Exp(3.667)	0.550	Exp(7.333)	0.300
			Exp(3.667)	0.775	Exp(7.333)	0.400
			Exp(3.667)	1.000	Exp(7.333)	0.500
					Exp(7.333)	0.600
					Exp(7.333)	0.700
					Exp(7.333)	0.800
					Exp(7.333)	0.900
					Exp(7.333)	1.000
Linear Service Time Pattern	Average Error	Confidence Interval	Average Error	Confidence Interval	Average Error	Confidence Interval
Constant	2.80E-05	5.28E-06	3.91E-05	1.11E-05	3.87E-05	1.31E-05
Normal(0.05)	2.98E-04	8.34E-05	5.92E-04	7.95E-05	8.49E-04	8.94E-05
Normal(0.15)	8.87E-04	2.74E-04	0.001483846	2.85E-04	0.00240651	2.16E-04
Normal(0.25)	0.001725367	4.18E-04	0.002958212	4.83E-04	0.004389689	4.39E-04
Normal(0.35)	0.002049282	5.19E-04	0.003990585	6.52E-04	0.006043414	6.34E-04
Exponential	0.006812243	0.00181385	0.010104696	0.001724814	0.015403404	0.001849051

Table 5-4: Relative Regression Errors of RLS Model for Mean Service Times that Vary Linearly with a Range of 10, Duration 300 Seconds

50 periods 5 seconds/period 75% server utilization 30 replications	3 Services		5 Services		10 Services	
	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time
	Exp(1.888)	0.100	Exp(2.875)	0.100	Exp(5.448)	0.100
	Exp(1.888)	0.316	Exp(2.875)	0.178	Exp(5.448)	0.129
	Exp(1.888)	1.000	Exp(2.875)	0.316	Exp(5.448)	0.167
			Exp(2.875)	0.562	Exp(5.448)	0.215
			Exp(2.875)	1.000	Exp(5.448)	0.278
					Exp(5.448)	0.359
					Exp(5.448)	0.464
					Exp(5.448)	0.599
					Exp(5.448)	0.774
					Exp(5.448)	1.000
Geometric Service Time Pattern	Average Error	Confidence Interval	Average Error	Confidence Interval	Average Error	Confidence Interval
Constant	0.010040276	0.001593485	0.007491821	0.001142561	0.006619481	0.001164389
Normal(0.05)	0.007746092	0.0010759	0.009069315	0.001361957	0.008399408	0.001017803
Normal(0.15)	0.014096533	0.001489047	0.014760831	0.002002425	0.016091703	0.001688781
Normal(0.25)	0.01794177	0.002104755	0.022512679	0.003241641	0.027057221	0.003524555
Normal(0.35)	0.02164261	0.002795342	0.029023978	0.003466681	0.039289079	0.003712261
Exponential	0.107892989	0.00933539	0.095028151	0.007873728	0.12776299	0.007627472

Table 5-5: Relative Errors of RLS Model for Mean Service Times that Vary Geometrically with a Range of 10 and Equal Inter-arrival Times, Duration 5 Seconds

50 periods 60 seconds/period 75% server utilization 30 replications	3 Services		5 Services		10 Services	
	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time
	Exp(1.888)	0.100	Exp(2.875)	0.100	Exp(5.448)	0.100
	Exp(1.888)	0.316	Exp(2.875)	0.178	Exp(5.448)	0.129
	Exp(1.888)	1.000	Exp(2.875)	0.316	Exp(5.448)	0.167
			Exp(2.875)	0.562	Exp(5.448)	0.215
			Exp(2.875)	1.000	Exp(5.448)	0.278
					Exp(5.448)	0.359
					Exp(5.448)	0.464
					Exp(5.448)	0.599
					Exp(5.448)	0.774
					Exp(5.448)	1.000
Geometric Service Time Pattern	Average Error	Confidence Interval	Average Error	Confidence Interval	Average Error	Confidence Interval
Constant	1.57E-04	2.86E-05	1.93E-04	4.11E-05	2.04E-04	6.19E-05
Normal(0.05)	5.80E-04	1.46E-04	0.001135416	1.70E-04	0.001458123	1.21E-04
Normal(0.15)	0.002115266	4.89E-04	0.0030777	5.58E-04	0.004629767	6.27E-04
Normal(0.25)	0.002761348	5.73E-04	0.004746547	9.79E-04	0.007949462	9.74E-04
Normal(0.35)	0.004379797	0.001034239	0.006424702	0.001166054	0.011769706	0.001309883
Exponential	0.010403328	0.001709228	0.018658787	0.003113431	0.031157626	0.004120819

Table 5-6: Relative Errors of RLS Model for Mean Service Times that Vary Geometrically with a Range of 10 and Equal Inter-arrival Times, Duration 60 Seconds

50 periods 300 seconds/period 75% server utilization 30 replications	3 Services		5 Services		10 Services	
	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time
	Exp(1.888)	0.100	Exp(2.875)	0.100	Exp(5.448)	0.100
	Exp(1.888)	0.316	Exp(2.875)	0.178	Exp(5.448)	0.129
	Exp(1.888)	1.000	Exp(2.875)	0.316	Exp(5.448)	0.167
			Exp(2.875)	0.562	Exp(5.448)	0.215
			Exp(2.875)	1.000	Exp(5.448)	0.278
					Exp(5.448)	0.359
					Exp(5.448)	0.464
					Exp(5.448)	0.599
					Exp(5.448)	0.774
					Exp(5.448)	1.000
Geometric Service Time Pattern	Average Error	Confidence Interval	Average Error	Confidence Interval	Average Error	Confidence Interval
Constant	2.17E-05	5.24E-06	2.97E-05	7.26E-06	4.19E-05	1.26E-05
Normal(0.05)	2.99E-04	7.33E-05	4.40E-04	6.25E-05	6.57E-04	5.19E-05
Normal(0.15)	7.33E-04	1.87E-04	0.001430341	2.42E-04	0.002338676	1.73E-04
Normal(0.25)	0.001385308	3.54E-04	0.001878212	2.82E-04	0.003791782	3.65E-04
Normal(0.35)	0.001760496	3.44E-04	0.002943226	6.30E-04	0.004788803	5.61E-04
Exponential	0.005270072	0.001265523	0.009535065	0.001707654	0.01305245	0.001502033

Table 5-7: Relative Errors of RLS Model for Mean Service Times that Vary Geometrically with a Range of 10 and Equal Inter-arrival Times, Duration 300 Seconds

50 periods 5 seconds/period 75% server utilization 30 replications	3 Services		5 Services		10 Services	
	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time
	Exp(1.734)	0.100	Exp(2.442)	0.100	Exp(3.702)	0.100
	Exp(1.825)	0.316	Exp(2.571)	0.178	Exp(3.897)	0.129
	Exp(1.927)	1.000	Exp(2.713)	0.316	Exp(4.113)	0.167
			Exp(2.873)	0.562	Exp(4.355)	0.215
			Exp(3.053)	1.000	Exp(4.628)	0.278
					Exp(4.936)	0.359
					Exp(5.289)	0.464
					Exp(5.696)	0.599
					Exp(6.170)	0.774
					Exp(6.731)	1.000
Varying Mean Interarrival Times	Average Error	Confidence Interval	Average Error	Confidence Interval	Average Error	Confidence Interval
Constant	0.009792105	0.001727109	0.006925893	0.001110749	0.006476236	0.001155435
Normal(0.05)	0.009603376	0.001567968	0.008405744	0.001275424	0.008794102	8.05E-04
Normal(0.15)	0.011723604	0.001346853	0.012652315	0.001814395	0.018925158	0.001656275
Normal(0.25)	0.014411288	0.001929519	0.021813275	0.002840093	0.025721384	0.002346574
Normal(0.35)	0.024465802	0.002894711	0.028010649	0.003532467	0.038854382	0.00415778
Exponential	0.101700855	0.010583468	0.103301341	0.009316359	0.124043143	0.007283432

Table 5-8: Relative Errors of RLS Model for Mean Service Times that Vary Geometrically with a Range of 10 and Increasing Inter-arrival Times, Duration 5 Seconds

Total Time: 150 mins Server Utilization: 75% Replications: 30	3 Services Constant		10 Services Constant		10 Services Normal(0.05)	
	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time
	Exp(2.200)	0.100	Exp(5.448)	0.100	Exp(5.448)	0.100
	Exp(2.200)	0.550	Exp(5.448)	0.129	Exp(5.448)	0.129
	Exp(2.200)	1.000	Exp(5.448)	0.167	Exp(5.448)	0.167
			Exp(5.448)	0.215	Exp(5.448)	0.215
			Exp(5.448)	0.278	Exp(5.448)	0.278
			Exp(5.448)	0.359	Exp(5.448)	0.359
			Exp(5.448)	0.464	Exp(5.448)	0.464
			Exp(5.448)	0.599	Exp(5.448)	0.599
			Exp(5.448)	0.774	Exp(5.448)	0.774
			Exp(5.448)	1.000	Exp(5.448)	1.000
Number of Reporting Periods	Average Error	Confidence Interval	Average Error	Confidence Interval	Average Error	Confidence Interval
13	3.27E-05	1.35E-05	5.77E-05	2.02E-05	9.57E-04	8.67E-05
25	5.11E-05	1.28E-05	7.05E-05	2.32E-05	0.00100773	1.12E-04
50	5.39E-05	1.16E-05	7.67E-05	2.19E-05	0.0010077	1.12E-04
75	5.97E-05	1.18E-05	7.81E-05	2.22E-05	0.00100731	1.12E-04
100	7.36E-05	1.38E-05	8.79E-05	2.29E-05	0.00100724	1.12E-04
125	8.31E-05	1.31E-05	9.22E-05	2.63E-05	0.00100927	1.12E-04
150	8.60E-05	1.42E-05	9.64E-05	2.34E-05	0.00100976	1.12E-04

Total Time: 150 mins Server Utilization: 75% Replications: 30	10 Services Normal(0.15)		10 Services Normal(0.25)		10 Services Normal(0.35)		10 Services Exponential	
	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time
	Exp(5.448)	0.100	Exp(5.448)	0.100	Exp(5.448)	0.100	Exp(5.448)	0.100
	Exp(5.448)	0.129	Exp(5.448)	0.129	Exp(5.448)	0.129	Exp(5.448)	0.129
	Exp(5.448)	0.167	Exp(5.448)	0.167	Exp(5.448)	0.167	Exp(5.448)	0.167
	Exp(5.448)	0.215	Exp(5.448)	0.215	Exp(5.448)	0.215	Exp(5.448)	0.215
	Exp(5.448)	0.278	Exp(5.448)	0.278	Exp(5.448)	0.278	Exp(5.448)	0.278
	Exp(5.448)	0.359	Exp(5.448)	0.359	Exp(5.448)	0.359	Exp(5.448)	0.359
	Exp(5.448)	0.464	Exp(5.448)	0.464	Exp(5.448)	0.464	Exp(5.448)	0.464
	Exp(5.448)	0.599	Exp(5.448)	0.599	Exp(5.448)	0.599	Exp(5.448)	0.599
	Exp(5.448)	0.774	Exp(5.448)	0.774	Exp(5.448)	0.774	Exp(5.448)	0.774
	Exp(5.448)	1.000	Exp(5.448)	1.000	Exp(5.448)	1.000	Exp(5.448)	1.000
Number of Reporting Periods	Average Error	Confidence Interval	Average Error	Confidence Interval	Average Error	Confidence Interval	Average Error	Confidence Interval
13	0.00248521	2.26E-04	0.00469891	5.30E-04	0.00652693	6.35E-04	0.01813674	0.001703543
25	0.00270826	2.02E-04	0.00482497	5.10E-04	0.00651147	5.69E-04	0.01706965	0.002336281
50	0.00270773	2.02E-04	0.0048294	5.11E-04	0.00651206	5.67E-04	0.0170562	0.002337401
75	0.00270774	2.03E-04	0.00482279	5.09E-04	0.00652671	5.67E-04	0.01706555	0.002335433
100	0.00270407	2.03E-04	0.00481913	5.15E-04	0.00652818	5.68E-04	0.01703545	0.002348051
125	0.00271391	2.00E-04	0.00482008	5.12E-04	0.00653686	5.73E-04	0.01702613	0.002343312
150	0.00270669	2.02E-04	0.0048254	5.11E-04	0.00653671	5.69E-04	0.01707718	0.002353845

Table 5-11: Relative Regression Errors of RLS Model for Different Number of Reporting Periods

50 periods 300 seconds/period 75% server utilization 30 replications	3 Services		5 Services		10 Services	
	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time
	Exp(1.888)	0.100	Exp(2.875)	0.100	Exp(5.448)	0.100
	Exp(1.888)	0.316	Exp(2.875)	0.178	Exp(5.448)	0.129
	Exp(1.888)	1.000	Exp(2.875)	0.316	Exp(5.448)	0.167
			Exp(2.875)	0.562	Exp(5.448)	0.215
			Exp(2.875)	1.000	Exp(5.448)	0.278
					Exp(5.448)	0.359
					Exp(5.448)	0.464
					Exp(5.448)	0.599
					Exp(5.448)	0.774
					Exp(5.448)	1.000
Perturb Ratios of Means From -10% to +10%	Average Error	Confidence Interval	Average Error	Confidence Interval	Average Error	Confidence Interval
Constant	0.023011719	0.004592909	0.040352109	0.005283478	0.04305147	0.003976231
Normal(0.05)	0.025005011	0.005664773	0.036687136	0.004982387	0.045155999	0.004402304
Normal(0.15)	0.022590873	0.005396481	0.03671457	0.004469548	0.043317135	0.003816775
Normal(0.25)	0.019695067	0.005236223	0.035502605	0.005102134	0.042104254	0.003595104
Normal(0.35)	0.020770461	0.005604783	0.039357783	0.004547905	0.041418464	0.00350378
Exponential	0.030761808	0.007000595	0.039130185	0.006140262	0.042583799	0.003979219

Table 5-12: Relative Regression Errors of RLS Model with Ratio Varying 10% for Mean Service Times that Vary Geometrically, Duration 300 Seconds

50 periods 300 seconds/period 75% server utilization 30 replications	3 Services		5 Services		10 Services	
	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time	Inter-arrival time distr.	Mean service time
	Exp(1.888)	0.100	Exp(2.875)	0.100	Exp(5.448)	0.100
	Exp(1.888)	0.316	Exp(2.875)	0.178	Exp(5.448)	0.129
	Exp(1.888)	1.000	Exp(2.875)	0.316	Exp(5.448)	0.167
			Exp(2.875)	0.562	Exp(5.448)	0.215
			Exp(2.875)	1.000	Exp(5.448)	0.278
					Exp(5.448)	0.359
					Exp(5.448)	0.464
					Exp(5.448)	0.599
					Exp(5.448)	0.774
					Exp(5.448)	1.000
Perturb Ratios of Means From -25% to +25%	Average Error	Confidence Interval	Average Error	Confidence Interval	Average Error	Confidence Interval
Constant	0.05221809	0.009877085	0.094569285	0.016347848	0.111191582	0.008984018
Normal(0.05)	0.067963045	0.013712362	0.095426979	0.014084361	0.117942057	0.01305981
Normal(0.15)	0.073258897	0.017594265	0.10248698	0.018238424	0.108750818	0.011054979
Normal(0.25)	0.064736676	0.012558361	0.087478097	0.011844015	0.118108592	0.014012852
Normal(0.35)	0.049018674	0.01269852	0.092489897	0.015138774	0.119713352	0.009616844
Exponential	0.080054948	0.014693823	0.100998736	0.013932221	0.10866196	0.011335129

Table 5-13: Relative Regression Errors of RLS Model with Ratio Varying 25% for Mean Service Times that Vary Geometrically, Duration 300 Seconds

Figure 5-1: Relative Regression Errors of RLS Model for Mean Service Times that Vary Linearly with a Range of 2, Duration 5 Seconds

Figure 5-2: Relative Regression Errors of RLS Model for Mean Service Times that Vary Linearly with a Range of 10, Duration 5 Seconds

Figure 5-3: Relative Regression Errors of RLS Model for Mean Service Times that Vary Linearly with a Range of 10, Duration 60 Seconds

Figure 5-4: Relative Regression Errors of RLS Model for Mean Service Times that Vary Linearly with a Range of 10, Duration 300 Seconds

Figure 5-5: Relative Errors of RLS Model for Mean Service Times that Vary Geometrically with a Range of 10 and Equal Inter-arrival Times, Duration 5 Seconds

Figure 5-6: Relative Errors of RLS Model for Mean Service Times that Vary Geometrically with a Range of 10 and Equal Inter-arrival Times, Duration 60 Seconds

Relative Regression Errors of RLS Model for Mean Service Times that Vary Geometrically with a Range of 10, Duration 300 Seconds

Figure 5-7: Relative Errors of RLS Model for Mean Service Times that Vary Geometrically with a Range of 10 and Equal Inter-arrival Times, Duration 300 Seconds

Relative Regression Errors of RLS Model for Mean Service Times that Vary Geometrically with a Range of 10 and Increasing Inter-arrival Times, Duration 5 Seconds

Figure 5-8: Relative Errors of RLS Model for Mean Service Times that Vary Geometrically with a Range of 10 and Increasing Inter-arrival Times, Duration 5 Seconds

Figure 5-9: Relative Errors of RLS Model for Mean Service Times that Vary Geometrically with a Range of 10 and Equal Inter-arrival Times, Duration 60 Seconds

Figure 5-10: Relative Errors of RLS Model for Mean Service Times that Vary Geometrically with a Range of 10 and Equal Inter-arrival Times, Duration 300 Seconds

Relative Regression Errors of RLS Model for Different Number of Reporting Periods

Figure 5-11: Relative Regression Errors of RLS Model for Different Number of Reporting Periods

Relative Regression Errors of RLS Model with Ratio Varying 10% for Mean Service Times that Vary Geometrically, Duration 300 Seconds.

Figure 5-12: Relative Regression Errors of RLS Model with Ratio Varying 10% for Mean Service Times that Vary Geometrically, Duration 300 Seconds

Relative Regression Errors of RLS Model with Ratio Varying 25% for Mean Service Times that Vary Geometrically with a Range of 10, Duration 300 Seconds.

Figure 5-13: Relative Regression Errors of RLS Model with Ratio Varying 25% for Mean Service Times that Vary Geometrically, Duration 300 Seconds

Chapter 6. Summary and Future Research

In this chapter, we give the summary of whole thesis. It includes comparison of three statistical tools, which are OLS, RLS and RCM. Section 6.2 describes how to run RCM and RLS in real systems. Section 6.3 provides the simulated cases which best fit to the RCM model. Future work and possible improvement for these three statistical tools are explored.

6.1 Comparison of Three Statistic Technique

In this thesis, we use OLS, RLS and RCM to estimate mean service times for application services. The main hindrance to statistical approaches is that service times are not deterministic, which violate the assumption of the ordinary least squares (OLS).

The RCM provides better estimates for mean service times than the OLS in most case. However, in exponential case, the RCM shows no improvement than the OLS. It is because that mean service times' varying exponentially violates assumptions of the RCM model. One of assumptions is that coefficients vary normally.

The RLS outperforms the other regression techniques over all of the cases that we studied. But we need to know the ratios for the mean service times of the application services. We perturb the known ratio for RLS and get the satisfying results. These suggest that RLS is robust with respect to errors in ratio. More over the ratios are not constant in some case. Consider the following application of the technique to estimate per-application-service physical input/output operations for some system. The state of the system's cache and virtual memory could cause more or less physical output for each service. If the impact is small with respect mean number of input-output operations then the ratios will remain consistent. If they are large the ratios will be misleading. RLS constraints can be expressed to model these kinds of state dependent system behavior as

well. For example if the impact of the behavior on all services is expected to be of magnitude p for service 1 and 2, an RLS constraint could be expressed as:

$$(c_1 + p) - (2c_2 + p) = 0$$

p must be measured and/or estimated by the monitoring system.

Let's talk about cost for each of these techniques. We use the same case as Table 2-10, that is Geometric Service Time Pattern, increasing inter-arrival times, range of 10 and duration of 300 seconds, to test the performance for each tool. This experiment is done in an NT system with 300MHZ CPU. Table 6-1 describes cost of OLS, RLS and RCM. All confidence intervals are less than ± 5 percent of their reported mean value.

Distribution	3 Services			5 Services			10 Services		
	OLS Cost	RLS Cost	RCM Cost	OLS Cost	RLS Cost	RCM Cost	OLS Cost	RLS Cost	RCM Cost
Constant	148.57	149.57	10790.20	193.97	197.30	13868.33	416.90	435.23	22149.87
Normal(0.05)	148.90	150.57	10816.97	191.60	193.60	13916.97	403.90	427.67	22313.70
Normal(0.15)	147.83	149.83	10838.57	192.00	195.33	13949.37	406.41	429.90	22929.70
Normal(0.25)	149.57	150.57	10838.17	192.00	195.00	14036.47	413.37	435.03	23427.63
Normal(0.35)	147.23	149.57	10842.93	191.60	196.60	13986.33	440.97	464.63	21936.87
Exponential	147.23	149.57	10877.97	191.00	193.67	14062.53	411.97	430.97	22089.77

Table 6-1: Performance of OLS, RLS and RCM Regression Tools, Time in ms.

From the table, we can see that as the number of services increases, the cost of each tool gets increased. The cost of RLS and OLS is almost the same, while the cost of RCM is as 50-75 times as the one of RLS and OLS.

Now it is obviously that the RLS appears to be robust with respect to service time variation, the number of operations and monitoring control parameters. And also the RLS does cost a little compared with RCM. The drawback is that we need to know the ratios in advance.

The RCM does give some improvement in some cases and doesn't need so much more information in advance but with a higher cost. The OLS does provide the worst answer but it needs nothing and costs a little. This is similar to real life. You can't gain without loss. Though RCM and RLS need some more information in advance, it doesn't mean

we can't do it in real systems. The next section gives some approaches to do RCM and RLS in real case.

6.2 RCM and RLS Approaches for Real Systems

In Chapter 3, we found the best answer of the RCM by trial and error because we already knew the true parameters. How can we do that in real systems? There is no way we can know true parameters in advance.

To solve this problem, we need to recall the relative error. There is another relation between resources demands and the relative error. The formula is given below:

$$E_{rel} \geq \frac{\sum_{k=1}^n U_{est,k} - \sum_{k=1}^n U_{true,k}}{\sum_{k=1}^n U_{true,k}} = \frac{U_{est} - U_{true}}{U_{true}} \quad (6-1)$$

We can know the total resources demands because operating systems can provide this information. For the total estimated resources demands, we can use estimated service times and their invocations to get it. Following the algorithm in Chapter 3, we try different bucket size and get total estimated resources demands. Using the equation (6-1), we can get different ratios. The set of estimated parameters, which leads to the minimum ratio, can be chosen the best-estimated parameters of the RCM model.

Now let's talk about the RLS. This tool needs ratios of parameters in advance. It is feasible in real systems. The ratios could be captured by keeping a historical record of successful applications of the OLS or the RCM, from data gathered during low system loads or offline, or possibly by the infrequent use of operating system kernel tracing tools. As we discussed in chapters before, OLS and RCM techniques caused less error in less number of applications cases. The ratios derived from parameters of these cases are closed to true values.

6.3 Summary for RCM

As we discussed before, the RCM doesn't give as good result as we expect. One of reason is that the RCM uses covariance of each group to decide weights of parameters for each group in a weighted mean. The simulator produces simulations that generally have similar situations in each reporting period. This results in that each group of data has similar covariance and parameters, which leads to that the RCM doesn't get so much improvement than the OLS. The key point is that we need some groups are "stable." In another word, we need some groups, which can give good estimation of parameters with small covariance. Probably, this is the case in real systems. For example, low system loads can give good estimates. The whole data we collect may include different system loads for each reporting period. We can cluster the data into different groups by system loads and then do RCM.

Distribution	Range of 2		Range of 10	
	E_{RCM}	E_{OLS}	E_{RCM}	E_{OLS}
Normal(0.05)	0.041776	0.048158	0.030647	0.037772
Normal(0.15)				
Normal(0.35)				
Normal(0.05)	0.068793	0.082939	0.076404	0.100292
Normal(0.30)				
Normal(0.90)				
Normal(0.05)	0.152377	0.206086	0.182942	0.247286
Normal(0.5)				
Normal(3)				

Table 6-2: Relative Regression Errors of RCM Model for Three Services with Each Sharing the Same Mixed Distributions

We do the following experiments to show this idea. First we mix three block data with different distributions and then do RCM and OLS to them. For the RCM, we divided data into different groups by different distribution. All results are given in Table 6-2. As

the variation of service time increases, the RCM shows more advantage than the OLS. So the RCM is still a good tool as only as we find the right situation to use it.

6.4 Future Research

For the RCM, we need to develop some functions to do calculation in equation (6-1) for real systems. Also, we can use the simulator to produce blocks of data with different distributions and mix them together. We can apply RCM to study this mixed data further.

As we knew in Chapter 3, we used the OLS to each subgroup of data in the RCM model regression. We can use maximum likelihood estimation instead of the OLS to each subgroup of data in the RCM model. From a theoretical view, using maximum likelihood estimation to each subgroup in the RCM is more reasonable than using the OLS [15].

For maximum likelihood estimation of generalized regression model (MLE), we need to test it with all the cases we used to the RCM and compare results with the other techniques.

In future, we can develop a Kalman Filter tool to our problem. In real systems, because of end effect, Kalman Filter may be a better solution. The algorithm of Kalman Filter and the situation fitted to this model are given in Chapter 3.

For the RLS, we need to implement some functions to fulfill the purpose of estimating ratios of parameters. I would say the process of estimating the ratios is the “learning process” which is done in situations where the OLS or the RCM can be done successfully, such as low system loads or offline.

Appendix A

This appendix describes the implementation of the statistical packages. Package diagrams are given for OLS, RLS, RCM and MLE. Also, class diagrams are illustrated for RCM and MLE. Class specifications are provided for RLS, RCM and MLE.

A.1 Package Diagram

Three statistic tools, namely, RCM, RLS and OLS were written in Java by the author. However, the simulator is written by C++ by Vidar Vetland. An integrating code was done to read the output of the simulator and edit data to appropriate form which can be accepted by three statistic tools. The basic implementation is given in Figure A-1. All these packages are driven by a script in Unix system and a batch file in Windows NT system. We can see from Figure A-1, the Simulator produces the output of experiment, ReadVidar edits and passes that to statistic tools. If we use the OLS, just call OLS package. When we use RCM and RLS, we call them and also call OLS through them.

There is some subtle difference in each IntegarteTool for three statistic tools. For example, IntegrateTool for the RCM includes features of finding the optimum solution by trial and error. IntegrateTool for the RLS contains features of setting up independent linear equations for constraints.

Written in Java, all three statistic tools are platform independent and portable. It is convenient for future use in distributed systems, network or electronic commerce systems.

Figure A-1: Package Diagram for Three Statistic Tools

Figure A-2: Class Diagram of RCM Tool

A.2 Random Coefficient Model Regression Tool

The class diagram is given in Figure A-2. Since I do not write Matrix class, I am not going to describe it. Also, as you may know, RCM needs to use OLS. I do not implement OLS and just do some modification so I won't describe it either.

Class specification is expressed as below:

Name: RCM

Responsibilities:

Get covariance matrix for each bucket.

Operations:

RCM()

doVi()

Attributes:

SiSquare

MatVi

Name: RCMP

Responsibilities:

Figure out results of RCM regression model and get relative errors.

Operations:

RCMP()

getGamma()

getAverage()

getSum()

getW()

getAllBi()

calculateB()

getError()

Attributes:

allV

allViG

allW

allBi

averageParameterError

A.3 Restricted Least Square Regression Tool

It's quite hard to describe RLS without OLS. Basically, what I did is to modify OLS by adding some new methods. So only new methods are expressed as below:

Name: RLS

Responsibilities:

Calculate parameters by RLS regression model and give relative errors.

Operations:

doRestrictedRegression ()

getConsError()

Attributes:

reConsError

matR

matQ

matConsResult

A.4 Maximum Likelihood Estimation Tool for Generalized Regression Model

Maximum Likelihood Estimation Tool for Generalized Regression Model (MLE) is different from RLS and RCM. MLE doesn't need to do OLS. The package diagram is given in Figure 7-3. From the figure, we can see an IntegrateTool drives this tool. ReadVidar inside IntegrateTool package reads output of Simulator and passes that to MLE, which finally does calculation.

The class diagram is shown in Figure A-4. The specification of MLE is expressed as below:

Name: MLE

Responsibilities:

Calculate parameters by MLE tool of generalized regression model and give relative errors.

Operations:

MLE()

startIteration()

doVariance()

computeB()

getError()

Attributes:

reError

matY

matX

matB

Figure A-3: Package Diagram of Maximum Likelihood Estimation Tool for Generalized Regression Model

Figure A-4: Class Diagram of MLE Tool

References

- [1] J. Rolia and V. Vetland, "Correlating Resource Demand Information with ARM Data for Application Services," Proceedings of the *ACM Workshop on Software and Performance (WOSP '98)*, October 22-28, 1998, Santa Fe, NM, USA, pages 219-230.
- [2] Y. Bard and M Schatzoff, "Statistical Methods in Computer Performance Analysis," in Chandy and Yeh (eds.) *Current Trends in Programming Methodology*, Vol. III, Prentice-Hall, Englewood Cliffs, N.J., 1978.
- [3] William H. Greene, "Econometrics Analysis," Chapter 6, Macmillan Publishing Company, 1990.
- [4] README file of Vidar Vetland's Simulation Tool.
- [5] Xin Sun, Jerome Rolia and Vidar Vetland "Correlating Usage Statistics for Application Services with Measurable Resource Demand Information," Submitted to *IEEE Transactions on Software Engineering*, March, 1999.
- [6] Paul Newbold and Theodore Bos, "Stochastic Parameter Regression Models," Chapter1, Sage Publications.
- [7] Paul Newbold and Theodore Bos, "Stochastic Parameter Regression Models," Chapter2, Sage Publications.
- [8] William H. Greene, "Econometrics Analysis," Page 476-477, Macmillan Publishing Company, 1990.
- [9] William H. Greene, "Econometrics Analysis," Page 381-382, Macmillan Publishing Company, 1990.
- [10] William H. Greene, "Econometrics Analysis," Page 387-388, Macmillan Publishing Company, 1990.
- [11] William H. Greene, "Econometrics Analysis," Page 390-393, Macmillan Publishing Company, 1990.
- [12] William H. Greene, "Econometrics Analysis," Page 212-213, Macmillan Publishing Company, 1990.
- [13] J.A. Rolia and K.C. Sevick, "The Method of Layers," *IEEE Transaction on Software Engineering* Vol. 21, Num. 8, August 1995, pages 689-700.

[14] Woodside C.M., Neilson J.E., Petriu D.C., and Majumdar S., "The Stochastic Rendezvous Network Model for the Performance of Synchronous Client-Server-like Distributed Software," *IEEE Transactions on Computer*, 44(1), pp.20-34, January 1995.

[15] Longford, Nicholas T., "Random coefficient models," 1955- Oxford, Clarendon Press. 1993.