A STATISTICAL ANALYSIS OF THE PERFORMANCE OF BATCH COMPUTING SYSTEMS

by

Justin Newland

Thesis submitted for the degree of
Doctor of Philosophy
in the University of London and
the Diploma of Membership of Imperial College

Imperial College of Science and Technology
Department of Mathematics

June, 1981

Abstract

The performance of batch computing systems is investigated from two perspectives: a descriptive analysis of job submission patterns, and the prediction of job turnaround time.

The aim of the first part of the study was to describe the submission patterns of different classes of jobs. A mathematical description of the large-scale, internal features of the job submission patterns was obtained by subjecting the data to two principal components analyses, first in the time domain, then in the frequency domain. The non-standard theory of principal components analysis in the frequency domain is presented for the general case. The method involved the construction and principal components analysis of consistent estimates of the spectral density matrix. The general level of demand at any time was found to be highly dependent on the previous short-term job submission level.

The aim of the second part of the study was to develop a parsimonious regression model to predict the turnaround time of a job, given the information available at the time of its submission. Because the least-squares approach was found to be inadequate, alternative estimation procedures were sought. Of these, a method of M-estimation proposed by Huber was found to work best. Least Absolute Deviations (LAD) estimation is also discussed. The algorithms available for deriving M- and LAD- estimates are compared. The distribution of a likelihood ratio test for M-estimators is

examined via a Monte Carlo sampling experiment; a second experiment examined the conjectured distribution of a proposed likelihood ratio test for LAD-estimators. The power of these tests are compared to that of the classical least-squares 'F-test'. A power study of the 't-tests' of these three estimators is also conducted. The main result of the applied study was that the turnaround time of a low (high) resource request job depended on the number of other low (high) resource request jobs in the system on its submission.

Preface

As a piece of applied work, this study is the outcome of a long period of research during which many approaches have been explored. Some of these were unilluminating, others proved to be extremely fruitful. As Box (1979) has pointed out, the scientific model building process is an iterative procedure, involving the postulation of a model structure, model criticism, then a re-assessment of the structure and approach in the light of the results. As the process rarely if ever results in the perfect model, the relative value of a model may be discussed in terms of its utility, and how effectively it clarifies the data under investigation and the mechanisms which generated them.

Hence, the motivations for the choice and treatment of the subject matter were entirely pragmatic. The following work is primarily a thesis in applied statistics, but it is also intended to contribute to the growing body of knowledge on computer performance evaluation. The methodology was deliberately chosen in such a way that it could, with little modification, be applied to any batch computing system, independent of the type of machine. The work was carried out with one eye at all times on the utility of the results.

As to the notion of clarification, the postulated models are by no means a perfect representation of the data. Their eventual inclusion in the thesis was dictated by how informative they were in elucidating the data under scrutiny. While the model building approach adopted here is only one of a possible constellation of approaches, the results are both revealing and useful.

Acknowledgements

I am endebted to Dr. B.D. Ripley for his perspicacious guidance and invaluable assistance. I would also like to thank Janie Coghill, who typed this manuscript.

The data for this work was kindly supplied by Mr. G. Adamson, of Rutherford High Energy Laboratory, Didcot, Oxfordshire, and Dr. T. Ratcliffe of Newcastle University, Tyneside.

The first year (1978-1979) of this work was supported by funds provided by the Science Research Council under grant B/RG/8722.8 to Prof. M.M. Lehman, of the Department of Computing and Control, Imperial College. The final two years of this study (1979-1981) were funded by a Research Studentship award, also from the Science Research Council.

The computer programs written by the author were all in FORTRAN. Some work was done on the IBM 360/65 machine (replaced in August, 1980) at University College, London. The bulk of the computing work was carried out on the CDC 6500/ Cyber 174 system at Imperial College Computer Centre.

CONTENTS

List of Tables 8

List of Figures 10

Chapter 1: PRELIMINARY CONSIDERATIONS 13

Chapter 2: ANALYSIS OF JOB SUBMISSION PATTERNS 19

2.1 Principal Components Analysis for Time-Series 21

2.1.1 Detrending the Data 22

2.1.2 Time Domain Analysis 23

2.1.3 Frequency Domain Analysis 27

2.2 Results of the Principal Components Analyses 33

2.2.1 Rutherford 33

2.2.2 Newcastle 46

Chapter 3: ROBUST REGRESSION TECHNIQUES 53

3.1 Least Absolute Deviation (LAD) Estimation 55

3.2 Huber's M-estimators 83

3.3 A Monte-Carlo Power Study 100

Chapter 4: PREDICTION OF JOB TURNAROUND 111

4.1 Preliminaries 111

4.2 Regression Results 116

4.2.1 Overview 117

4.2.2 NUMAC 122

4.2.3 Rutherford 139

Appendix 148

Chapter 5: CONCLUSIONS 152

5.1 Computer Performance Evaluation 152

5.2 Statistics 157

References 161

LIST OF TABLES

Table		Page
2.1	Classification Scheme (by resources consumed) for Rutherford jobs.	33
2.2	First five Eigenvalues from each of the three eigenanalyses for five weeks of Rutherford Data.	34
2.3	Classification Scheme (by resources consumed) for NUMAC jobs.	46
2.4	First five Eigenvalues from each of the three eigenanalyses for two weeks of NUMAC data.	47
3.1	Comparative Performance of three algorithms in fitting an LAD regression line to six data-sets.	76
4.1	Summary statistics of NUMAC job priorities; data-set N79W1.	123
4.2	Definition of NUMAC job classes.	123
4.3	Results of subset regression fits; data-set N79W1.	126
4.4	Mean absolute prediction errors (in minutes) for full and subset regressions; data-set N79W1.	136
4.5	Mean absolute prediction errors of turnaround times of jobs in N79W3 when these are predicted by the coefficients of the subset Huber models fitted to the N79W1 and N79W2 data-sets.	138
4.6	Summary statistics of Rutherford job priorities data-set R79W1	140

Table		Page
4.7	Definition of Rutherford job priorities.	141
4.8	Results of subset regression fits; data-set R78W1	143
4.9	Mean absolute prediction errors (in minutes) for full and subset regressions: data-set R78W1	147
4.1A	Results of subset regression fits: data-sets N79W2 and N79W3.	148
4.2A	Mean absolute prediction errors (in minutes) for full and subset regressions: data-set N79W2 and N79W3	151

LIST OF FIGURES

Figure		Page
2.1	Scatter plots of coefficients of first two eigenvectors: Rutherford data. (a) data treated as departures from class means: data-set R77W1. (b) data treated as departures from seasonal means: data-set R77W1. (c) data treated as residual processes: data-set R77W1 (d) data treated as departures from class means: data-set R77W2. (e) data treated as departures from seasonal means: data-set R77W2. (f) data-treated as departures from residual processes: data-set R77W2.	37
2.2	Log (base 10) auto-spectrum estimate of the detrended class A1 series (from R77W1) with two applications of the modified discrete Daniell filter with different half-widths. (a) half-widths ℓ = 8 and 6. (b) half-widths ℓ = 16 and 12. (c) half-widths ℓ = 32 and 16.	39
2.3	Log (base 10) of the first eigenvalue spectra: Rutherford data. (a) data-set R77W1 (b) data-set R77W2 (c) data-set R78W1 (d) data-set R78W2 (e) data-set R78W3	41
2.4	Log (base 10) auto-spectrum estimate of detrended class A1 series (from R77W1) before and after prewhitening.	43
2.5	Log (base 10) of the second eigenvalue spectra: Rutherford data. (a) data-set R77W1 (b) data-set R77W2 (c) data-set R78W1 (d) data-set R78W2 (d) data-set R78W3	45
2.6	Scatter plot of coefficients of first two eigenvectors:NUMAC data. (a) data treated as	49

Figure		Page
	departures from class means: data-set N79W1. (b) data treated as departures from seasonal means: data-set N79W1. (c) data treated as residual processes: data-set N79W1 (d) data treated as departures from class means: data-set N79W2. (e) data treated as departures from seasonal means: data-set N79W2. (f) data treated as residual processes: data-set N79W2.	
2.7	Log (base 10) of the first eigenvalue spectra: NUMAC data. (a) data-set N79W1. (b) data-set N79W2.	51
2.8	Log (base 10) of the second eigenvalue spectra: NUMAC data (a) data-set N79W1 (b) data-set N79W2	51
3.1	Cumulative frequency distribution of observed LRT statistic for LAD estimators. (a) errors sampled from a 'bootstrap'distribution. (b) errors sampled from a Normal (0,1) distribution (c) errors sampled from Cauchy (0,1) distribution. (d) errors sampled from a negative exponential (2) distribution.	81
3.2	Cumulative frequency distribution of observed LRT statistic for Huber's M-estimators. (a) errors sampled from a 'bootstrap' distribution. (b) errors sampled from a Normal (0,1) distribution: population λ. (c) errors sampled from a negative exponential (2)distribution:population λ. (e) errors sampled from a Cauchy (0,1) distribution: sample λ. (f) errors sampled from a negative exponential (2) distribution sample λ.	99
3.3	Empirical power curves of three likelihood	104

Figure		Page
	ratio tests when errors drawn from (a) a standard normal N(0,1) distribution. (b) a contaminated normal 8/9 N(0,1) + 1/9 N(0,8) distribution.	
3.4	Empirical power curves of three 't-tests' when errors drawn from (a) a standard normal N(0,1) distribution. (b) a contaminated normal 8/9 N(0,1) + 1/9 N(0,8) distribution.	108
4.1	Residuals vs. fitted values from linear model fits to class 1 (full sample) jobs: data-set N79W1. (a) Least-squares fit. (b) Huber fit.	129
4.2	Histogram of residuals from fits to class 1 (reduced sample) jobs: data-set N79W1. (a) Least-squares linear model fits. (b) Huber linear model fit. (c) LAD linear model fit. (d) Least-squares multiplicative model fit.	130
4.3	Histogram of residuals from fits to class 4 jobs: data-set N79W1. (a) Least-squares linear model fit (b) Huber linear model fit. (c) Least-squares multiplicative model fit.	132
4.4	Cumulative absolute error distributions: data-set N79W1. (a) class 1 (full sample) (b) class 1 (reduced sample) jobs. (c) class 2 jobs (d) class 3 jobs (e) class 4 jobs (f) class 5 jobs.	134
4.5	Cumulative absolute error distributions; data-set R78W1. (a) class 1 jobs. (b) class 2 jobs. (c) class 3 jobs (d) class 4 jobs	145

CHAPTER 1

PRELIMINARY CONSIDERATIONS

Data for this study were drawn from system housekeeping files at Rutherford High Energy Laboratory, and the Northumbrian' Universities Multiple Access Computer (NUMAC) system. Four weeks of data were collected from Rutherford during October 1977, and another four from the same month in 1978. At both dates, the Rutherford system had two I.B.M. 370/195 machines, each with two megabytes of core, and controlled by the I.B.M. Multiple Variable Tasks (MVT) Houston Automatic Spooling Program (HASP) operating system (OS). The default cpu limit for batch jobs is five seconds, and the default core limit is 80 kbytes. The system enhancements carried out between the two periods of observation resulted in a shift from high to low resource consumption jobs, as well as an improvement in turnaround time. In October 1977, the system handled between 10,000 and 15,000 batch jobs per week: by October 1978, this figure had increased to around 20,000. As the installation caters for a scientific community, predominantly using the FORTRAN programming language, many of the jobs invoke the FORTRAN 'H' compiler, which occupies approximately 210 kbytes of core.

The four weeks of NUMAC data were collected during May 1979. At that time, the NUMAC system had an I.B.M.360/67 machine with one megabyte of core, and an I.B.M. 370/168 with six megabytes of core. The 370/168 runs interactively

on the Michigan Terminal System (MTS) during the day, and under HASP OS MVT from 21.30 to 02.00 each night. The 360/67 also runs under OS MVT, which has only batch facilities, from 08.00 to 02.00. Default limits for HASP OS jobs are seven seconds cpu time (in equivalent 370/168 seconds) and 150 kbytes of core; for batch jobs submitted via MTS, the corresponding limits are five seconds cpu time and 80 kbytes of core. The NUMAC system handled between 4000 and 5000 jobs per week. Many of the jobs make use of the Link Editor which occupies approximately 80 kbytes of core, and the FORTRAN 'G' and FORTRAN Fort. Edit compilers, which occupy 118-124 and 256 kbytes of core respectively.

The observations made on each batch job were: the time and date of its submission; the cpu time requested and used; the maximum core size used and requested (in kbytes); and its turnaround time. The submission, cpu and turnaround times of each job were known to the nearest second. The measure of central memory usage (i.e. the maximum core size) of a job requires some elaboration. Our choice of a core measurement was primarily affected by the type of OS on which the jobs were processed. Both installations, we have seen, ran under the same type of I.B.M. OS. The MVT control program dynamically allocates core partitions throughout the sequence of job steps, where a job step is defined as a single program (IBM, 1972, pp 50-54). Thus, under MVT, a job may be allocated several core partitions, each at different locations in memory, during its execution. As a representative measure of the core consumption of a job,

we have employed the largest core partition size used over all job steps. This we refer to as the core consumption of a job. We define the core requested by a job in an analogous fashion. The resources consumed by a job (e.g. its core or cpu time) never exceed, and are sometimes much less than, the resources requested by it.

Chapter 2 deals with the analysis of job sumbission patterns. The underlying idea is that because low resource consumption jobs typically have shorter turnaround times than do high resource consumption jobs the former have different submission patterns to the latter. First of all, jobs are classified according to the amount of cpu time and core used. These two job characteristics represent a measure of the demand that a job has placed on the resources of a system. Job sumbissions are then aggregated over fifteen-minute intervals. Given m classes, and n intervals, we constructed m time-series $V_i(t)$, $i = 1, \ldots, m$, $t = 1, \ldots n$, where the element $V_i(t)$ is the number of class i jobs that arrived in the t^{th} quarter-hour interval of the week. Each of these time-series is then a numerical representation of the level of submissions of jobs that consume a given amount of the resources of the system. The mathematical tool best suited to describe the large-scale features of the submission data is principal components analysis. Section 2.1 deals with principal components analysis for time-series.

An essential preliminary to principal components analysis is to examine the data, and remove, or try to

remove, any known,systematic, sources of variation. These detrending techniques are described in section 2.1.1. Time-series data may be viewed from two different, complementary, perspectives in the time and frequency domain. A principal components analysis of the m time-series $V_i(\cdot)$ in the time-domain will reveal whether any meaningful submission patterns exist, and, if they do, will clarify their relationship to the resource consumption of a job (i.e. the job classes). A subsequent frequency domain analysis will then describe how these patterns fluctuate with frequency. The non-standard theory and practice of principal components analysis in the time and frequency domains is presented in Sections 2.1.2 and 2.1.3 respectively. The job classification schemes, and results of the analyses are described in section 2.2. Section 2.2.1 deals with the Rutherford data, section 2.2.2, the NUMAC data.

The nominal purpose of the second part of this opus was to predict the turnaround time of a job given the information available at the time of its arrival to the system. Thus jobs are classified according to the amount of resources requested, rather than the amount used. A multiple regression model, one for each job class, was constructed to predict job turnaround time. The models were fitted by a number of different estimation procedures, two of which are Least Absolute Deviations (LAD) and Huber's M-estimation procedure (Huber, 1964, 1973). The theory and method of these robust procedures are presented in Section 3.

Section 3.1 deals with LAD estimators. We develop the Linear Programming (LP) formulation of the LAD estimation problem, and consider certain properties of the LAD hyperplane. A partial review of LP algorithms to solve the LAD problem is presented; these are compared to the Iterated Weighted Least Squares (IWLS) approach, the method adopted in this study. Some new results on a likelihood ratio test for LAD-estimators are developed. The statistic, which is derived from maximum likelihood considerations, is conjectured to have a χ^2-distribution. The assumed distribution of the test is investigated in a Monte Carlo sampling experiment.

Huber's M-estimation procedure is presented in section 3.2. The computational algorithms available for deriving M-estimators are considered, and the (IWLS) method used in this study is described. The distribution of a likelihood ratio test for M-estimators, proposed by Schrader and Hettmansperger (1980), is examined via a Monte Carlo experiment.

The results of two additional Monte Carlo studies are presented in section 3.3. The first compares the power of the likelihood ratio tests for LAD- and Huber's M-estimators relative to that of the classical least-squares 'F-test'. The second is a comparative power study of the 't-tests' of these three estimators.

The explanatory variables in the turnaround model are of two kinds: first, the resources requested by each job, which are a measure of the amount of work each requires

the system to undertake; and second, the mix of other jobs awaiting execution when the job is submitted, which is a measure of the degree of system congestion. An overview of the results of the full regression fits to each of the turnaround models is given in section 4.1. Since the ultimate aim of the regression analysis was to produce models parsimonious in the parameters, we consider the problem of subset selection 4.2, which includes the results of the subset regression fits to the NUMAC (section 4.2.1) and Rutherford (section 4.2.2) data.

CHAPTER 2

ANALYSIS OF JOB SUBMISSION PATTERNS

The purpose of this analysis is to describe the way in which users submit jobs that consume different amounts of the resources of the computer sytem. Typically, the relationship between resource consumption and turnaround time is one of direct proportion. Moreover, it is reasonable to assume that low resource consumption jobs take less time to write, or to correct, than high resource consumption jobs. On the basis of these heuristic arguments, we would expect to find differences in the submission-resubmission strategies of users who submit jobs of different classes.

First, we require a numerical representation of job submissions according to the level of resource consumption. The natural measures of the resource consumption of a job are its cpu and core usage. In the following, we have employed a job classification scheme, based on these two job characteristics, of 20 categories for Rutherford, and nine for NUMAC. These are described in sections 2.2.1 and 2.2.2 respectively. In order to generate a time-series for each job class, we need to aggregate the submission times. The length of the aggregation interval is of crucial importance: if it is too short, say, less than five minutes, each interval will contain insufficient entries for a fruitful analysis of job submission patterns; if it is too large, say over an hour, the number of points in each time-series

would be small, with the result that short-term fluctuations in job submission patterns would be obscured. For example; assume that there are c20,000 jobs submitted to the Rutherford system in one week; with 20 classes, this gives c1,000 jobs per class. A short aggregation interval of five minutes implies that there are 1440 intervals over a five-day week, and hence less than one job per interval. An aggregation interval of two hours results in only 60 intervals, which is too small to conduct a satisfactory time-series analysis. On the basis of these arguments, and the results of a pilot study which are not reported here, it was decided to aggregate jobs from both installations over fifteen-minute periods; that is, a count was made of the number of job arrivals over successive quarter-hour intervals.

Assume there are m classes of jobs, and n intervals. We have a vector-valued function:

$$V(t) = (V_1(t),\ V_2(t),\ \ldots.\ V_m(t))^T \qquad t = 1,\ \ldots n \qquad (2.1)$$

where $V_i(t)$ i = 1, ..., m is real-valued, and the index t refers to the time of recording of the measurements. Element $V_i(t)$ is the number of jobs in class i submitted during the t^{th} unit time-interval. Since our period of observation is the five-day working week, n is an integer multiple of five; that is, n = 5r where r is the number of daily observations.

We intended to subject the multivariate system V(t)

to a principal components analysis. This is an internal data analytic tool, which describes the multivariate system by a filtered version of itself. If the filter is of reduced rank, the principal components analysis has the effect of summarising the large-scale features of the data in only a few dimensions (Craddock, 1965). Our interest centres on the principal components analysis of the m time-series, $V_i(t)$. Time-series data may be viewed from either the perspective of the time domain, or the frequency domain. A principal components analysis in the time domain will describe, in mathematical terms, the salient features of the submissions data, while a subsequent frequency domain analysis will describe how these submission patterns fluctuate with frequency. Section 2.1 below deals with principal components for time-series, while the results of the analyses are given in section 2.2.

2.1 Principal Components Analysis for Time-Series

In the following section (2.1.1) we report on the detrending techniques applied to the submissions data as a preparation for the principal components analysis. The adopted methodology and non-standard theory of principal components analysis in the time and frequency domains is discussed in sections 2.1.2 and 2.1.3 respectively.

2.1.1 Detrending the Data

An inspection of the raw data points of the class series $V_i(t)$, $t = 1, \ldots, n$, revealed that there was considerable variation between the within-day variances for each series. Clearly, some transformation was required to stabilise the various components of the overall variance. We required some transformation parameter $\lambda < 1$ such that:

$$W_i(t) = V_i(t)^{\lambda} \qquad i=1, \ldots, m; \quad t=1, \ldots, n$$

where $W_i(t)$ is the transformed series. A logarithmic transformation was tried, but was found to be too severe. In addition, there were no job submissions during some of the intervals. A square-root transformation (i.e. $\lambda = 0.5$) was found to reduce the variation between the within-day variances to an acceptable level. We denote the square-roots of the $V_i(t)$ series by:

$$W_i(t) = \sqrt{V_i(t)} \qquad i=1, \ldots, m; \quad t=1, \ldots, n \qquad (2.2)$$

Next, a linear trend was removed from each of the class series. This is a necessary preliminary, especially to the analysis in the frequency domain, as the presence of a linear trend (which is a kind of very slow cycle) can distort the low frequency spectrum estimates. Denote the resultant series by:

$$W_i^*(t) = W_i(t) - \hat{\alpha}_{oi} - t\hat{\alpha}_{1i} \qquad i=1, \ldots, m; \quad t=1, \ldots, n \qquad (2.3)$$

where the coefficients $\{\hat{\alpha}_{oi}, \hat{\alpha}_{1i}; i=1, \ldots, m\}$ are the least-squares parameter estimates.

The final detrending technique was intended to remove the daily, seasonal, variation from each class series. Each of these consists of five consecutive sets of r observations per day. A set of r estimated seasonal means was computed for each of the m series:

$$\hat{S}_i(j) = \sum_{k=1}^{5} W_i^*(j + r(k-1))/5 \qquad i=1, .., m; \; t=1,.., n$$

These were subtracted from their respective series to give the deseasonalised series:

$$X_i(t) = W_i^*(t) - \hat{S}_i(j) \qquad i=1, .., m; \; t=1,..,n; \qquad (2.4)$$
$$j=t \bmod r$$

This is equivalent to passing each of the m $W_i^*(t)$ series through the Buys-Ballot seasonal filter (Koopmans, 1974). The method has the useful advantage that no data points are sacrificed.

2.1.2 Time Domain Analysis

Principal components analysis in the time domain extracts those elements in the multivariate system that account for the most variation (i.e. carry the bulk of the information) and summarises them into the form of orthogonal mathematical functions (Craddock, 1965). This exploration

of the internal structure of the data identifies the main components, the relative importance of which may be assessed by examining the magnitudes of the leading eigenvalues. Physical meaning may then be attached to these first few dimensions by interpreting the pattern of coefficients in the leading eigenvectors, by which we mean the eigenvectors associated with the leading eigenvalues. This will result in a mathematical description of the more prominent submission patterns and their relationship to the job classes.

The standard theory of principal components analysis in the time domain may be found in any text on multivariate analysis, (e.g. Morrison, 1976, Chap.7). For the remainder of this section, we confine ourselves to a discussion of the adopted methodology.

The key issue in the estimation of the eigenvalues λ_i, i=1, ..., m and their corresponding eigenvectors $\underset{\sim}{v}_i = (v_{1i}, v_{2i}, \ldots, v_{mi})$ is whether to decompose the sample covariance or the sample correlation matrix. It is well-known that the two analyses are not equivalent, because the method of principal components analysis is not scale-invariant. An eigenanalysis of the sample covariance matrix would be scale-dependent, in so far as undue weight would be given to those classes with large variances. The effect of examining the sample correlation matrix would tend to equalise the weights given to each individual class series. Consequently, we decomposed the sample correlation matrix. In time-series nomenclature, the correlation matrix is the zero-lag autocorrelation function.

The next step is to determine which eigenvalues are significant. When they are arranged in order of magnitude, there is sometimes a clear cut-off point, beyond which the remaining eigenvalues are considerably smaller in magnitude than the leading group. We shall invoke a useful rule-of-thumb given by Jeffers (1964) for eigenvalues computed from correlation matrices, whereby only those eigenvalues that exceed unity are deemed as significant.

We carried out three principal components analyses on each data-set. In the first stage of the analysis, the data are treated as uncorrelated and non-seasonal; only the variation about the class means is removed. In the second stage, the data are treated as uncorrelated, in so far as the mean and seasonal variation is removed. In the third stage, the data are treated as residual or innovation processes; only the variation left after the mean variation, seasonality and local correlation have been removed remains in each components series. A straightforward application of principal components analysis to the submissions data ignores the dimension of time, by treating the observations as if they were independent. The purpose of this three-fold analysis is to introduce the element of time, and to take account of the sources of variation in the data. We intend to examine the changes in the magnitudes of the leading eigenvalues, and the changes in the pattern of coefficients in the eigenvectors with which they are associated, as the sources of variation are successively removed. We expect there to be changes in the magnitudes of the eigenvalues, although, since we are extracting the eigenvalues from a

correlation matrix, the total variation or sum of the eigenvalues will remain constant throughout the three stages. Nevertheless, there is no reason to suppose that the pattern of coefficients in the leading eigenvectors will remain constant. If they do, we shall be able to conclude that the components they describe are independent of the sources of variation, and therefore describe an essential characteristic of the data.

Principal components analyses were carried out on three different sample correlation matrices computed from the same data-set; these were when the data were treated as:

1) departures from class means.

The (m x m) correlation matrix was computed from the series:

$$Y_i^{(1)}(t) = W_i(t) - \bar{W}_i(\cdot) \qquad i=1,\ldots,m;\ t=1,\ldots,n \qquad (2.5)$$

where $W_i(t)$ is defined in equation (2.2) and

$$\bar{W}_i(\cdot) = n^{-1}\sum_{t=1}^{n} W_i(t), \quad i=1,\ldots,m.$$

2) departures from seasonal means.

The correlation matrix was constructed from the series:

$$Y_i^{(2)}(t) = X_i(t) \qquad i=1,\ldots,m;\ t=1,\ldots,n \qquad (2.6)$$

where $X_i(t)$ is the deseasonalised series defined in equation 2.4.

3) innovation or residual processes.

The correlation matrix was constructed from the series:

$$Y_i^{(3)}(t) = X_i(t) - \sum_{j=1}^{q} \hat{\beta}_{ji} X_i(t-j) \quad i=1, \ldots m; \quad t=q+1, q+2, \ldots, n \qquad (2.7)$$

where $\hat{\beta}_{ji}$, j=1, ..., q are the least-squares parameter estimates of the q^{th}-order autoregressive process fitted to the i^{th} class series.

The techniques for time-series model identification, estimation and residual diagnostics described in Box and Jenkins (1970) were employed to determine the order of the antoregressive process to be fitted to each series. We do not discuss the fitting procedure, nor do we report the results; these may be found in Newland (1979). The nature of the results suggested that a fifth-order model was sufficiently high to remove most of the local corrleation from within each series. Thus, a fifth-order autoregressive model was fitted to each of the m series (i.e. q = 5 in expression 2.7). This final data preparation technique prewhitens the data (Blackman and Tukey, 1959).

2.1.3 Frequency Domain Analysis

The estimation of the spectral density is the corner-stone of the method of principal components analysis in the frequency domain. The adopted approach is non-parametric. We begin with the data after prewhitening, the $Y_i^{(3)}(t)$

series given by expression 2.7. Scale considerations, as in the time domain, are important too in the frequency domain. Rather than work with the matrix of coherencies (the frequency domain analogue of the correlation matrix), we attempted to induce scale-invariance by normalising each component series in the usual way (i.e. by dividing each series by its standard deviation). Next, each series was tapered by 5% at either end, using the split-cosine bell window (Bloomfield, 1976, p.84). Denote the resultant series by $Z_i(t)$. The second-order cross-periodogram between series i and series j at frequency f_k is defined as:

$$I^{(T)}_{Z_i Z_j}(f_k) = (2\pi T)^{-1} \, d_{Z_i}{}^{(T)}(f_k) \, \overline{d_{Z_j}{}^{(T)}(f_k)}^{T} \qquad (2.8)$$

$$i, j = 1, \ldots, n$$

where $d_{Z_i}{}^{(T)}(f_k) = n^{-1} \sum_{t=1}^{n} Z_i(t) \exp(-itf_k)$ is the discrete Fourier transform of the $Z_i(t)$ series; the transpose of its complex conjugate is denoted by $\overline{d_{Z_i}{}^{(T)}(f_k)}^{T}$.

Frequency $f_k = 2\pi k/T$, for $k=0, 1, \ldots, T-1$ where $T=2^p$ for p the first integer such that $2^p > n$. The superscript (T) denotes that the function has been computed over T ordinates.

Expression 2.8 is an inconsistent estimate of the spectrum. To obtain a consistent estimate, we smooth the cross-periodogram over a number of adjacent frequencies. The cross-spectrum is defined as follows:

$$g_{Z_i Z_j}^{(T)}(f_k) = \sum_{j=-\ell}^{\ell} a_j \, I_{Z_i Z_j}^{(T)}(f_{k+j}) \qquad (2.9)$$

Smoothing was carried out using the modified discrete Daniell Filter, defined (for ℓ even) as:

$$a_j = \begin{cases} 1/2\ell & j = -\ell+1, \ldots, \ell-1 \\ 1/4\ell & j = -\ell, \ell \end{cases} \qquad (2.10)$$

The span of the window is then $(2\ell+1)$; its half-width is ℓ. The second-order cross-periodogram (eqn 2.8) and the cross-spectrum (eqn. 2.9) are both complex-valued functions. In all frequency plots that follow, we adjust the spectra to be within the frequency range (0.0, 0.5) cycles per unit time interval.

Define $C^{(T)}(f_k)$ as the estimate of the cross-spectral density matrix at frequency f_k derived from the multivariate system Z. The construction of such estimates was as follows:

Step 1: Construct estimates of the m real-valued auto-spectra (eqn. 2.9 with i=j) where each has been smoothed with the same number of filter applications and the same window spans.

Step 2: Construct estimates of the $\binom{m}{2} = m(m-1)/2$ cross-spectra by smoothing both the real and imaginary parts of the cross-periodograms as for the auto-spectra.

Step 3: Then the $(i, j)^{th}$ element of $C_Z^{(T)}(f_k)$ is $g_{Z_i Z_j}^{(T)}(f_k)$

From equation (2.9) $g_{Z_j Z_i}^{(T)}(f_k) = \overline{g_{Z_i Z_j}^{(T)}(f_k)}^{T}$

The complex matrix $C_Z^{(T)}(f_k)$ is Hermitian; that is, it satisfies:

$$C_Z^{(T)}(f_k) = \overline{C_Z^{(T)}(f_k)}^T$$

The above method yields (1 + T/2) (m x m) cross-spectral density matrices. If T is large, it is more practical and economic (in terms of computing time) to sample over a restricted set of frequencies. From the properties of Hermitian matrices (Fox, 1964; Wilkinson, 1965) it is known that the eigenvalues $\lambda_j(f_k)$, j = 1, ..., m of $C_Z^{(T)}(f_k)$ are real and non-negative. This property does not hold in general for the corresponding eigenvectors $\underset{\sim}{\nu}_j(f_k)$, the elements of which are usually complex-valued. This is one difference between principal components analysis in the time and frequency domains: in the former, the eigenvalues are of zero dimension, while in the latter, they are functions of frequency.

If, for each frequency f_k, the eigenvalues are arranged in descending order of magnitude $\lambda_{(1)}(f_k) > \lambda_{(2)}(f_k) > \ldots > \lambda_{(m)}(f_k)$, our interest centres on the leading few eigenvalue spectra, denoted as:

$$\{\lambda_{(j)}(f_k);\ f_k = 2\pi k/T,\ k=0, 1, \ldots T/2\} \text{ for } j \leq 2 \text{ or } 3 \quad (2.11)$$

The interpretation of such a series is that it is the spectrum of the j^{th} principal component series (Brillinger, 1975, p 351). On the assumption that $\lambda_j(f_k)$, j=1, ..., m are asymptotically jointly normally distributed, it can be shown that (Brillinger, 1975, theorem 9.4.3):

$$\text{var}\ (\lambda_j(f_k)) \approx b^2\ \lambda_j(f_k) \qquad (2.12)$$

where b^2 is a compound correction factor, defined as:

$$b^2 = (U_4/U_2^2)\ (T/n) \sum_{j=-\ell}^{\ell} a_j^2 \qquad (2.13)$$

The first term on the right-hand side is a correction factor for tapering (see Bloomfield, 1976, section 8.5) where U(t), t=1, ..., n are the taper weights. U_p is defined as $U_p = n^{-1} \sum_{t=1}^{n} U^p(t)$. The second term, (T/n), may be regarded as an adjustment for frequency spacing because the spectrum ordinates estimated by this method are more finely spaced than are the Fourier frequencies. The final term on the right-hand side of equation 2.13 is an adjustment for the amount of smoothing: the spectral weights a_j, $j = -\ell, \ldots \ell$ are defined in equation 2.10.

Expression 2.12 suggests that we take logarithms (to the base 10) of the eigenvalues. Thus:

$$\text{var}\ \{\log_{10}\ \lambda_j\ (f_k)\} \approx (b\ \log_{10}\ e)^2 \qquad (2.14)$$

Then, under the appropriate normality assumptions (Brillinger, 1975, section 9.4), approximate 95% confidence intervals for $\log_{10}\ \lambda_j(f_k)$ are:

$$\log_{10}\lambda_j(f_k) \pm 1.96\ b\ \log_{10}\ e \qquad (2.15)$$

The width of the interval is the same as for the log spectral density estimate. Expression 2.15 suggests that we should plot the j^{th} eigenvalue spectra given by (2.11)

on a logarithmic, rather than a linear, scale.

A crucial aspect in the estimation of the spectral density matrices $C_Z^{(T)}(f_k)$ is the degree of smoothing to be applied to the second-order periodograms. Consider the following argument: let $\lambda_j(f_k)$ and $\underset{\sim}{\gamma}_j(f_k)$, j=1, ...,m be the (distinct) eigenvalues and eigenvectors of the sample spectral density matrix $C_Z(f_k)$, an estimate of the population spectral density matrix $\Gamma_Z(f_k)$. Let $\mu_j(f_k)$ and $\underset{\sim}{\gamma}_j(f_k)$, j=1, ..., m be the eigenvalues and eigenvectors of $\Gamma_Z(f_k)$. On the condition that $C_Z(f_k)$ is a consistent estimate of $\Gamma_Z(f_k)$, it can be shown that (see Brillinger, 1975, theorem 9.4.2):

$$\lambda_j(f_k) = \mu_j(f_k) + \underset{\sim}{\nu}_j(f_k)^T D \underset{\sim}{\nu}_j(f_k) + o(D^2) \qquad (2.16)$$

where $D = C_Z(f_k) - \Gamma_Z(f_k)$

This first-order result follows from an application of perturbation theory (Wilkinson, 1965) to the eigenvalues of $C_Z(f_k)$ and $\Gamma_Z(f_k)$. Expression 2.16 shows that the extent of the asymptotic bias depends on the proximity of the sample estimate to $\Gamma_Z(f_k)$. Substantial bias may then be introduced if $C_Z(f_k)$ varies rapidly with frequency: hence the importance of smoothing.

2.2 Results of the Principal Components Analyses

2.2.1 Rutherford

We present the results of the analysis of five weeks batch job data. Two of these were from October, 1977 (the weeks beginning the third and 17th), and three consecutive weeks (beginning on the fourth) from October, 1978. We refer to these by the acronyms; R77W1, R77W2, R78W1, R78W2, R78W3.

Table 2.1 below shows the 20 class scheme used to classify Rutherford jobs.

TABLE 2.1 Classification Scheme (by resources consumed) for Rutherford Jobs

		CORE CATEGORY			
		A	B	C	D
		Max Core Size(Used) in Kilobytes			
		≤50	50,250	250,500	>500
	1 cpu ≤1	A1	B1	C1	D1
CPU	2 time 1, 5	A2	B2	C2	D2
	3 (used) 5, 20	A3	B3	C3	D3
CATEGORY	4 in 20, 60	A4	B4	C4	D4
	5 seconds >60	A5	B5	C5	D5

As the Rutherford system is quite heavily used throughout the day and night, we decided to include observations from all 24 hours. A week was defined to begin at midnight on Sunday and to end at midnight on Friday. This

gives a total of (n=) 480 fifteen-minute intervals for each of the (m=) 20 classes. In order to extract the daily cycle, each series was deseasonalised by a period of (r=) 96, as described in section 2.1.1.

In table 2.2 below, we present themain results of the time domain principal components analysis of the Rutherford data

Table 2.2 : First five eigenvalues from each of the three eigenanalyses for five weeks of Rutherford data

Total variance = 20.0

	Data treated as		
Data-set	Departures from class means	Departures from seasonal means	Residual processes
R77W1	8.62	6.39	3.82
	1.86	2.03	1.68
	1.06	1.12	1.23
	0.90	1.05	1.17
	0.84	0.95	1.10
R77W2	9.37	6.95	4.36
	1.24	1.55	1.36
	1.03	1.13	1.26
	0.83	1.02	1.11
	0.77	0.96	1.07
R78W1	8.31	2.56	2.12
	1.13	1.46	1.47
	1.07	1.41	1.40
	0.95	1.25	1.27
	0.91	1.18	1.22
R78W2	6.84	3.23	2.46
	1.43	1.51	1.51
	1.19	1.41	1.44
	1.13	1.31	1.31
	1.05	1.21	1.26
R78W3	6.50	1.96	1.85
	1.20	1.45	1.42
	1.17	1.34	1.33
	1.10	1.29	1.28
	1.09	1.24	1.25

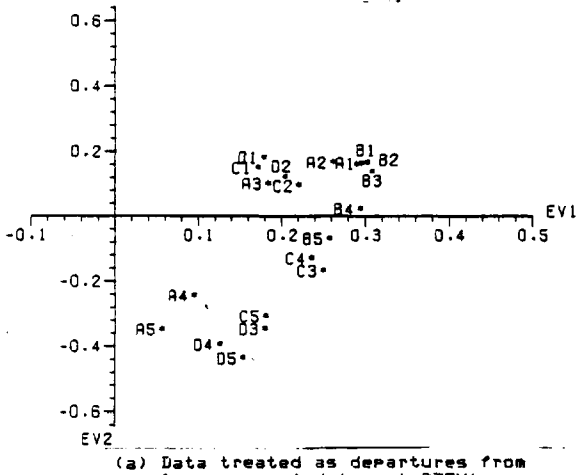
In each of the five analyses, the magnitude of the leading eigenvalue is reduced by the removal of the seasonal variation, and further reduced by removing some of the autocorrelation. The extraction of (some of) the daily cycle diminishes the percentage variation contributed by the largest eigenvalue by 11, 12, 28, 18 and 22% respectively for each data-set R77W1, R77W2, R78W1, R78W2 and R78W3. The corresponding percentages for the extraction of the local correlation are 13, 13, 2, 4 and 1% respectively. As the total variation remains constant throughout each stage of the analysis, the proportionate reduction in the variation explained by the first eigenvalue is counter-balanced by an increase in all the remaining eigenvalues (except the second largest) by approximately this same proportion. The behaviour of the second largest eigenvalue as the sources of variation are successively suppressed is anomalous. The reduction of some of the seasonality increases its percentage contribution, while extraction of the local correlation decreases its contribution.

Consequently, as all but the leading two eigenvalues behave in approximately the same way over the three stages of the analysis, we decided to centre our attention on only the first two dimensions. This step was further justified by an inspection of the coefficients of the eigenvectors associated with the third and fourth eigen-values. This revealed that the pattern of coefficients in these eigenvectors varied considerably as the seasonal variation and local correlation was reduced.

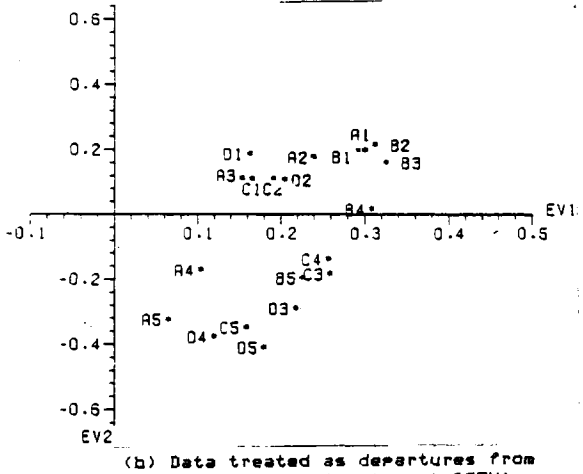
We present the coefficients of the eigenvectors associated with the first two eigenvalues $\lambda_{(1)}$ and $\lambda_{(2)}$ in the form of a scatter plot, one for each of the three stages of the analysis. The point with co-ordinate (x, y) refers to a class with coefficient x on the first eigen-vector, and y on the second. Where possible, the class associated with the point on the scatter plot has been marked to the left of the point. The advantages of such plots are that they condense the information in the first two eigenvectors into a single graph, and show how the different sets of points (classes) are dispersed .

The three scatter plots for the data-sets R77W1 and R77W2 are plotted in figures 2.1(a) to 2.1(c) and 2.1(d) to 2.1(f) respectively. The analogous plots for R78W1, R78W2 and R78W3 were very similar to these, and are omitted. In figures 2.1(a), (b) and (c) the points are grouped into four clusters; one for classes A1, B1, B2 and B3; another for C1, D1, A2, C2, D2 and A3, a third, consisting of C3, D3, C4 and C5; and the fourth, A4, D4, A5, C5, D5. The classes A1, B1, B2 and B3 have two things in common with each other; they have the largest average job submission rates (e.g. for R77W1, the largest average job submission rate is 6.33 jobs per 15 minutes for class B2), and refer to jobs with small cpu consumption. The second group contains the remaining small to medium cpu consumption classes. The third group refers to jobs with large cpu and core consumption,while the fourth is composed of jobs with large cpu consumption but only small to medium core usage. Approximately the same pattern of

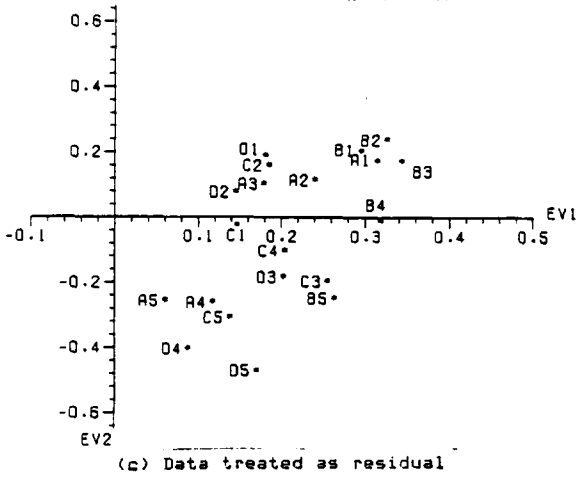
Figure 2.1 : Scatter plot of coefficients of first two eigenvectors ; Rutherford data.



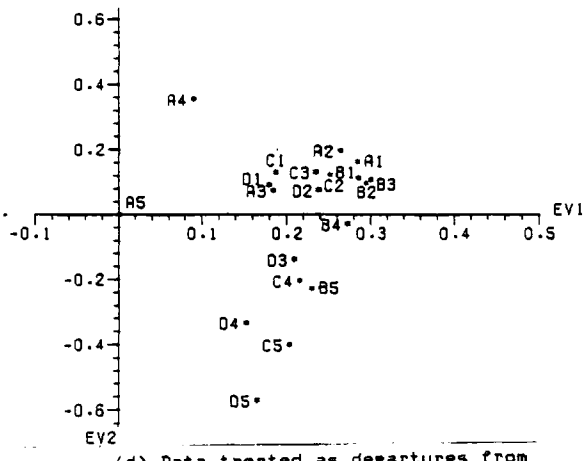
(a) Data treated as departures from class means ; data-set R77W1.



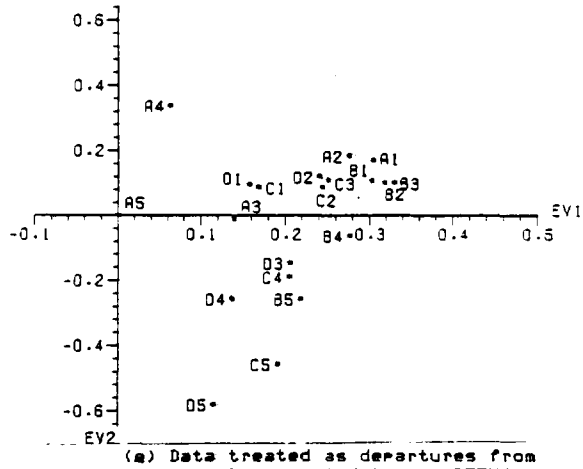
(b) Data treated as departures from seasonal means ; data-set R77W1.



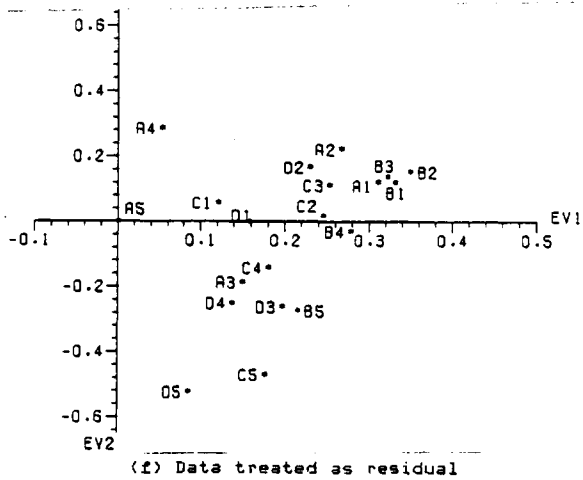
(c) Data treated as residual processes ; data-set R77W1.



(d) Data treated as departures from class means ; data-set R77W2.



(e) Data treated as departures from seasonal means ; data-set R77W2.



(f) Data treated as residual processes ; data-set R77W2.

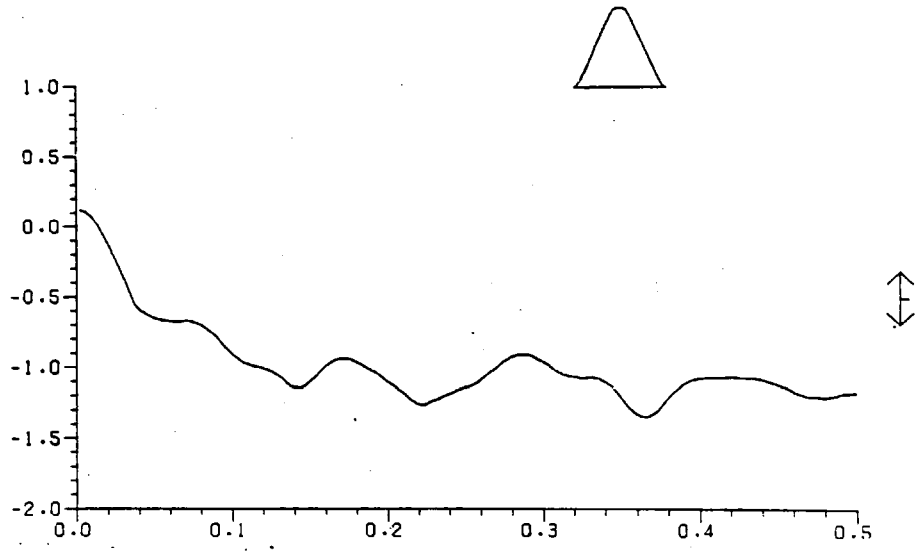
groupings of classes can be found in figures 2.1(d), (e) and (f).

From each of the figures, we see that job classes A1, B1, B2, B3 and B4 have the largest coefficients on the first eigenvector. Moreover, all the coefficients of the first eigenvector are positive in each of the plots. Consequently, the interpretation of this dimension is that is an index of the general level of demand, with especial emphasis on jobs from core category B.

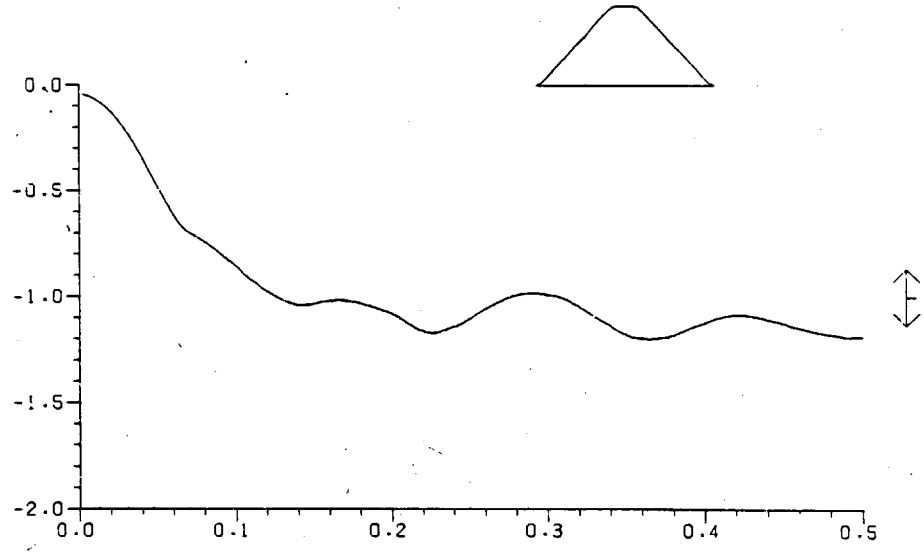
A rough assessment of the pattern of coefficients in the second eigenvector is that the coefficients associated with CPU categories 1 and 2 are positive, while those assoicated with CPU categories 4 and 5 are negative. The second dimension is interpreted as a contrast between small and large cpu jobs, that is, an index of the cpu activity of the job.

We turn to a study of the frequency domain behaviour of the first two principal components. With n = 480, T, the number of points over which the fast Fourier transform is computed, is 512. It follows that the second-order periodograms were computed over 1 + T/2 = 257 ordinates. The next step after the computation of the second-order cross-periodograms was to examine the effect of using different combinations of smoothing parameters. If the span of the spectral weights is too large, the fine structure in the spectrum, such as narrow peaks, is smoothed out and lost. Alternatively, if the span is too small, the spectrum ordinates will contain superfluous detail, with the result

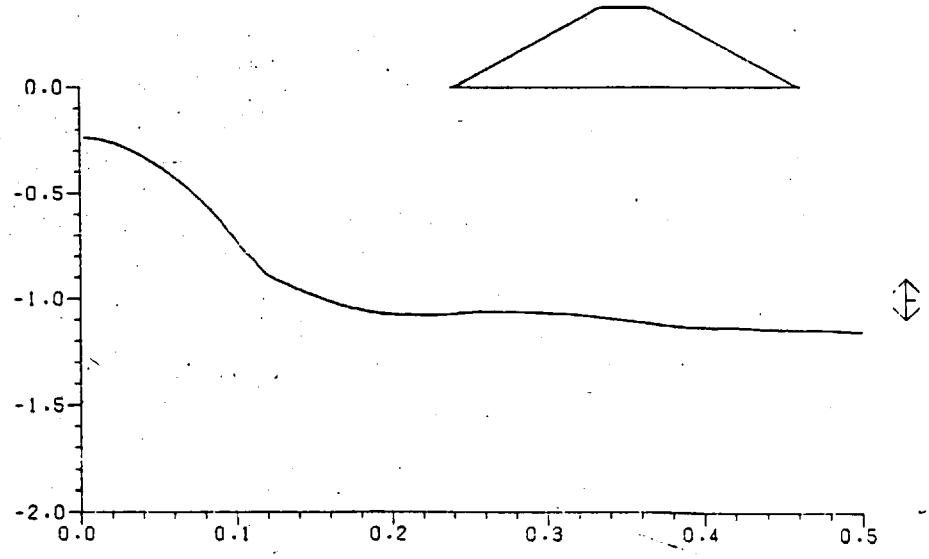
Figure 2.3 : Log(base 10) auto-spectrum estimate of the detrended class A1 series (from R77W1) with two applications of the modified discrete Daniel filter with different half-widths.



(a) half-widths l=8 and 6.



(b) half-withs l=16 and 12.



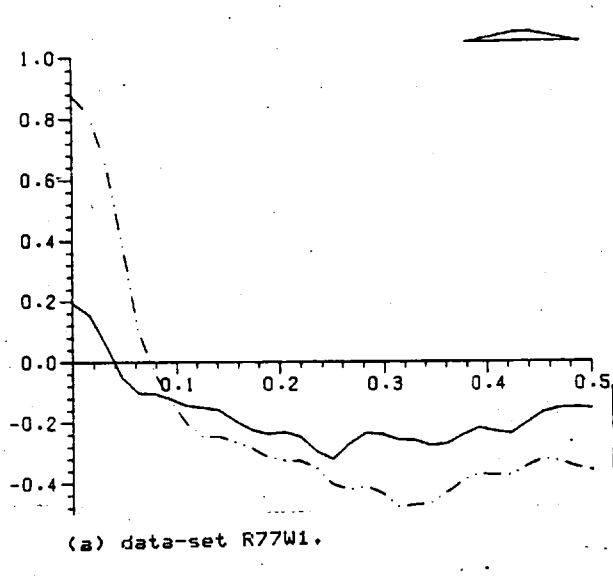
(c) half-widths l=32 and 24.

that bias is introduced into the estimates of the eigenvalue spectra. Clearly, a compromise between the two extremes is required.

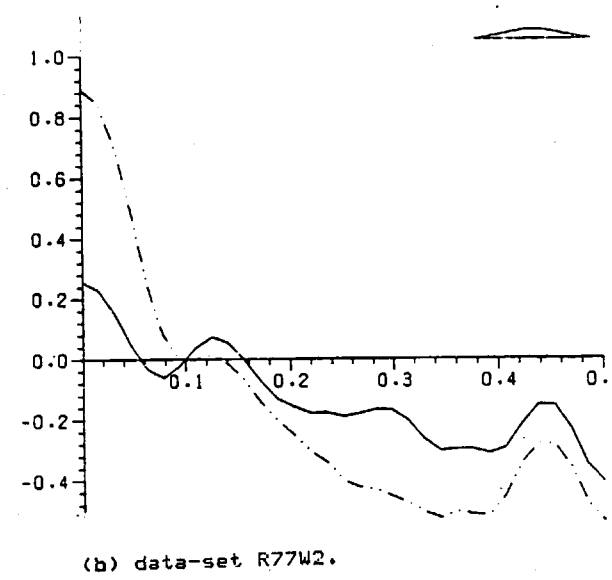
A thorough study with different numbers of passes of filters with different spans was carried out on the R77W1 data-set. As an example, we consider the log (to the base 10) of the auto-spectra of the detrended class A1 series (the $X_1(t)$ series given by eqn. 2.4), smoothed with two applications of the modified discrete Daniell filter with different half-widths, ℓ. Three examples, for ℓ = 8 and 6, ℓ = 16 and 12, and ℓ = 32 and 24 are shown in figures 2.2 (a), (b), and (c) respectively. The spectral window in each case is shown inset. Approximate 95% confidence intervals appear on the right-hand side of each plot. As the width of the spectral window increases, the auto-spectra becomes progressively smoother and flatter. On the basis of this, and analagous plots for other class series not presented here, we decided that two passes of the Daniell filter with half-widths equal to 16 and 12 gave the best estimates.

Rather than compute the eigenvalue spectra over the maximum possible number of frequencies (257), we chose a subset of frequencies. Estimates of 33 cross-spectral density matrices were constructed, that is $\{C_Z^{(T)}(f_k);\ f_k = 2\pi k/64,\ k=0,\ 1,\ \ldots,\ 32\}$ and each subjected to a principal components analysis. Figures 2.3(a) to 2.3(e) are the eigenvalue spectra of the first principal component series for each of the five data-sets; R77W1, R77W2, R78W1,

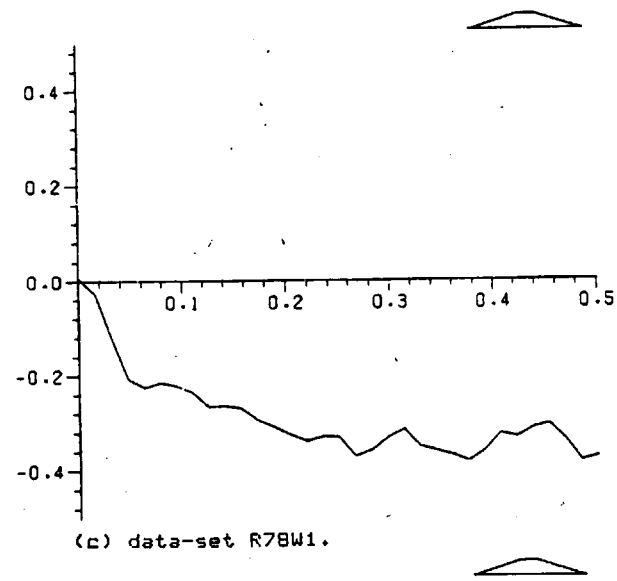
Figure 2.3 : Log(base 10) of first eigenvalue spectra ; Rutherford data.



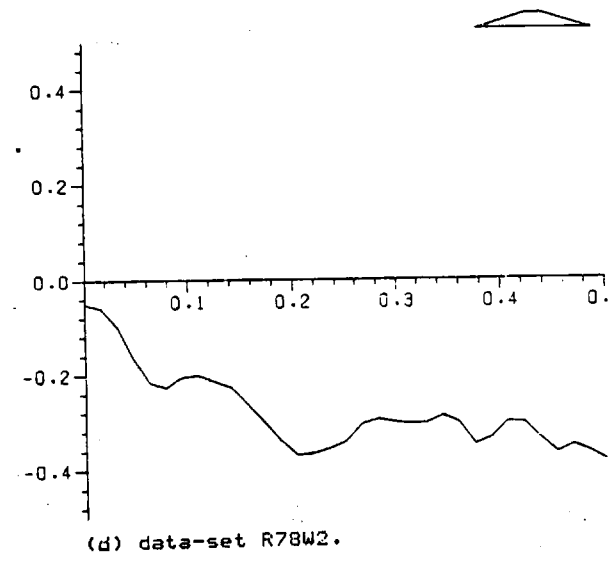
(a) data-set R77W1.



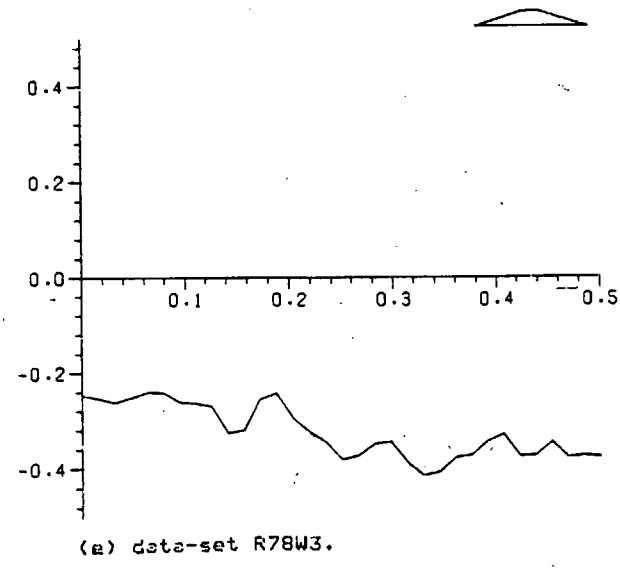
(b) data-set R77W2.



(c) data-set R78W1.



(d) data-set R78W2.



(e) data-set R78W3.

Legend

---·--- Detrended data before prewhitening

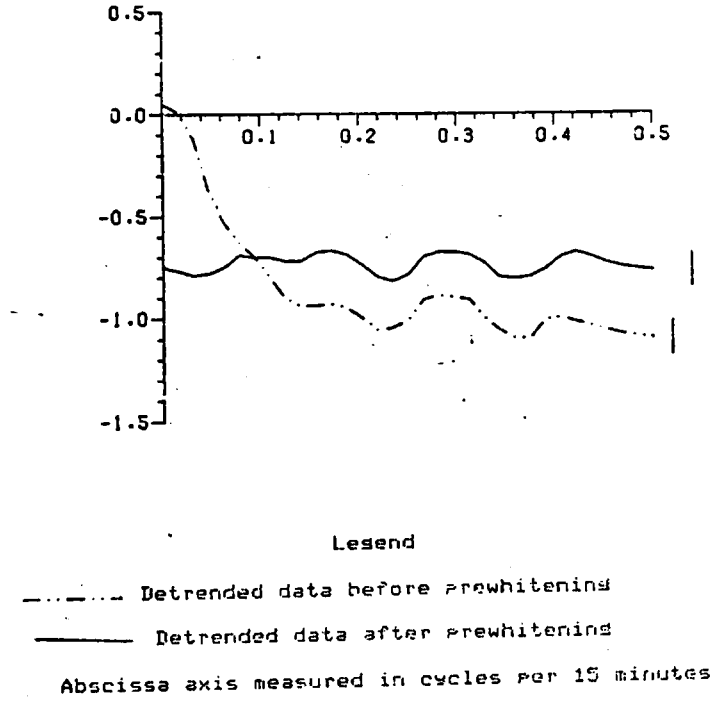
——— Detrended data after prewhitening

Abscissa axis measured in cycles per 15 minutes

R78W2 and R78W3 respectively. Note that the frequencies have been adjusted to be in the range (0.0, 0.5) cycles per unit time interval.

Figures 2.3(a) and 2.3(b)(for the data-sets R77W1 and R77W2 respectively) show the effect on the eigenvalue spectra of prefiltering the data,that is, removing some of the local autocorrelation from each series. Each of the spectra computed from the detrended data before prefiltering has a high concentration of power in the low frequency range, and a maximum at zero frequency. It is well-known that a spectrum estimated in the presence of low-order positive autocorrelation will exhibit these two features. The other three eigenvalue spectra (figures 2.3(c), (d) and (e)) show a gradual decline in power from left to right, which also suggests the presence of low-order positive autocorrelation. Indeed, the inverse relationship between power and frequency is the one common feature between the spectra (computed after prewhitening). There are two possible causes of this phenomenon; either the first principal component series genuinely contains low-order positive autocorrelation, or there still exists some unremoved autocorrelation in the component series. We decided to test the latter hypothesis by examining the effect of prewhitening the component series on the estimated auto-spectra of a number of class series. The auto-spectra were constructed from the detrended data before and after prewhitening. We show one example in figure 2.4; it is the logarithm (base 10) of the detrended class A1 series (R77W1 data-set) computed

Figure 2.4 : Log(base 10) auto-spectrum estimate of the detrended class A1 series (from R77W1) before and after prewhitening.



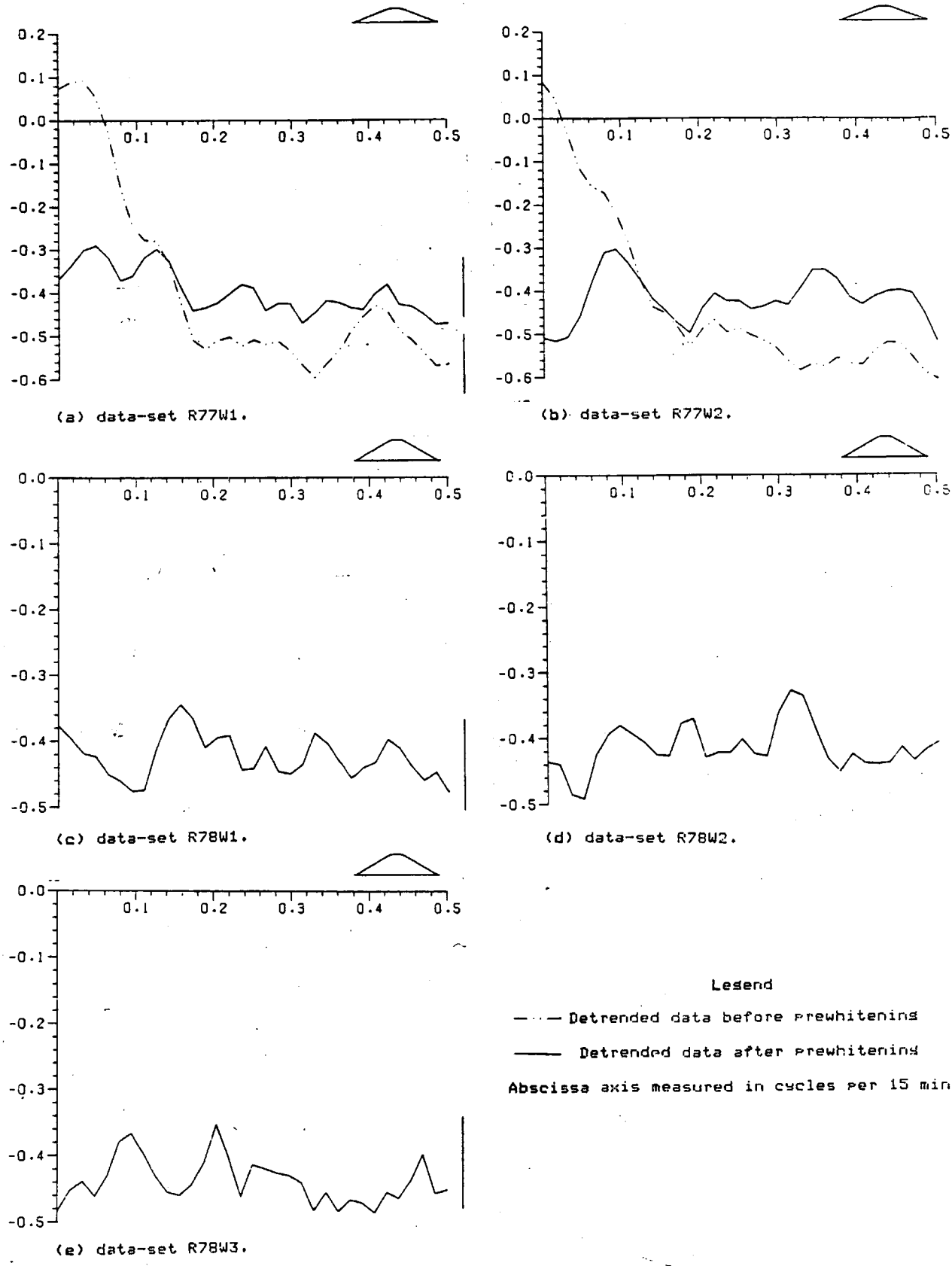
before and after prewhitening. The extraction of the low-order autocorrelation clearly reduces the high concentration of power in the low frequency range. This, and other evidence, strongly indicated that the significant degree of power in the first eigenvalue spectra computed after the component series had been prewhitened was not caused by unremoved local correlation in the component series.

In the time domain, the first dimension was interpreted as an index of the general level of demand. This interpretion is valid in the frequency domain only on the condition that the pattern of coefficients in the eigenvectors associated with the leading eigenvalue at each frequency is the same as the observed pattern of coefficients in the

leading eigenvector in the corresponding time domain analysis. This could be studied formally by constructing a plot of the estimated gain of the first principal component series $|\underset{\sim}{\nu}_{(1)}(f_k)|$ (see, e.g., Bartels, 1977, Chap. 7); in fact, only an informal check was made by a visual comparison of the pattern of coefficients in the relevant eigenvectors (we ignored their imaginary parts, which were generally near zero). Within tolerable limits, we were satisfied that the above condition had not been violated. Consequently, interpreting the first eigenvalue spectra as the power spectrum of the first principal component series, we conclude that the general level of demand at any time is heavily dependent on the previous short-term job submission level.

Figures 2.5(a) to 2.5(e) are the five log eigenvalue spectra of the second principal components series. As above, figures 2.5(a) and 2.5(b) (referring to data-sets R77W1 and R77W2 respectively) show the second eigenvalue spectra of the detrended data before and after prewhitening. When some of the autocorrelation is removed from the component series, there is a considerable reduction in power in the low frequency range, as we would expect. The dynamic range of the spectra computed after prewhitening is fairly narrow compared to the width of the approximate 95% confidence intervals. This leads us to conclude that none of the local fluctuations in any of the second log eigenvalue spectra are statistically significant. The second dimension was interpreted to be an index of the cpu activity of the job; combining the two results suggests

Figure 2.5 : Log(base 10) of second eigenvalue spectra ; Rutherford data.



that the overall cpu demand placed on the system is random.

2.2.2 Newcastle

Two weeks' data from Newcastle have been analysed: they are two consecutive weeks from the week beginning the fifth of May, 1979. We refer to these by the acronyms N79W1 and N79W2. The job classification scheme employed to discriminate between NUMAC jobs is given below in table 2.3

TABLE 2.3 Classification Scheme (by resources consumed) for NUMAC jobs

			CORE CATEGORY		
			A	B	C
			Max Core Size(Used)in Kilobytes		
			≤50	50-250	>250
CPU CATEGORY	1	cpu time (used) in seconds 5	A1	B1	C1
	2	5, 20	A2	B2	C2
	3	> 20	A3	B3	C3

As in the study of the Rutherford data, the week-end data was discarded. As the NUMAC system tended to be under-used during the night-time, we concentrated on job submissions during the prime-shift hours. Then, defining the beginning of the week to be at 08.30 on Monday, and the end at 19.30 on Friday, gives a total of (n=) 220 intervals for each of the (m=) 9 classes. The component series were

transformed, detrended and prewhitened in the way described in section 2.1.1. The seasonal period was r = 44.

Table 2.4 ; First five eigenvalues from each of the three eigenanalyses for two weeks of NUMAC data

Total variance = 9.0

	Data treated as		
Data-set	Departures from class means	Departures from seasonal means	Residual processes
N79W1	2.85	2.36	2.19
	1.29	1.22	1.13
	1.04	1.09	1.04
	0.94	0.99	0.97
	0.81	0.84	0.88
N79W2	2.80	2.37	2.35
	1.30	1.28	1.25
	1.03	1.11	1.15
	0.93	1.02	0.99
	0.87	0.82	0.80

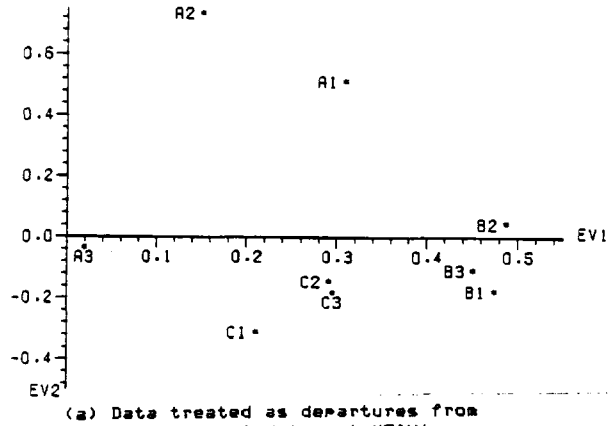
Deseasonalising each component series results in a reduction of 5.6% and 4.7% in the percentage variation accounted for by the first eigenvalue of weeks N79W1 and N79W2 respectively. The corresponding percentages for the second eigenvalue are 0.77 and 0.22. Fitting each deseasonalised series by a fifth-order autoregressive model reduces the percentage variation explained by the first eigenvalue by 1.88% and 0.22% respectively. The corresponding figures for the second eigenvalue are 0.22% and 0.33%. The removal of these two sources of variation has the effect of reducing the percent of variation explained the first two eigenvalues, but of increasing it for the remaining eigenvalues. On the basis of this evidence, and

a cursory inspection of the pattern of coefficients in the leading few eigenvectors, we concluded that the first two dimensions were of prominent importance.

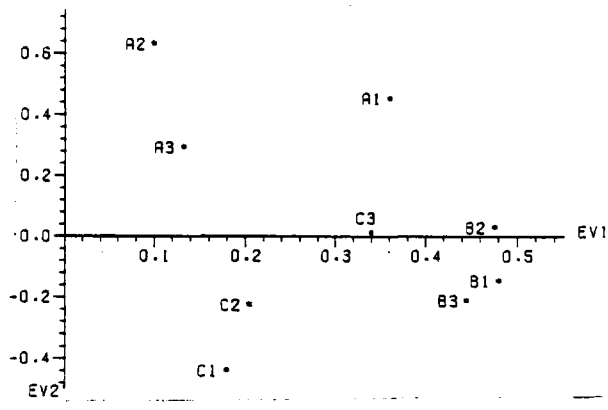
As for the Rutherford data, we present the coefficients of the eigenvectors associated with the first two eigenvalues in the form of a scatter plot. The resulting plots from the three stages of the analysis are shown in figures 2.6(a), (b) and (c) for the N79W1 data-set, and figures 2.6(d), (e) and (f) for the N79W2 data-set. The arrangement of the points in each of the graphs is roughly the same: the classes from core category B have the largest coefficients in the first eigenvector; the coefficients of the classes in core category A are positive, while those in core category C are negative. As the arrangement of these groupings is preserved through each of the three stages of the analysis, we infer that the groupings describe an underlying feature in the data.

As for the Rutherford data, the coefficients of the eigenvector associated with the leading eigenvalue are all positive in each of the three stages of the analysis. Again, we interpret this dimension as an index of the general level of demand, with particular emphasis on submissions to classes B1, B2 and B3. The second dimension is a contrast between the level of submissions of jobs in classes A1, A2 and A3 to those in classes C1, C2, and C3. The second dimension is therefore an index of the core activity of the job, or the general core demand placed on the system.

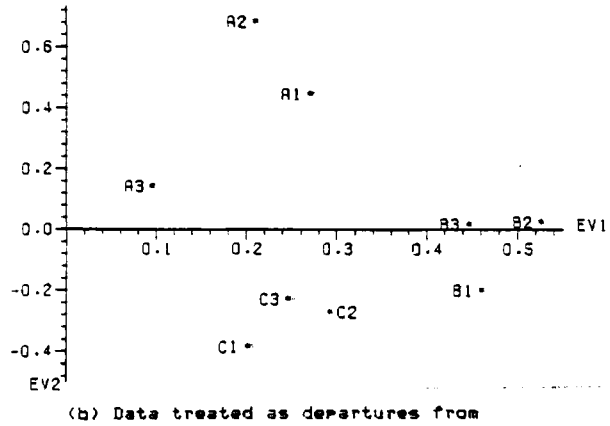
Figure 2.6 : Scatter plot of coefficients of first two eigenvectors ; NUMAC data.



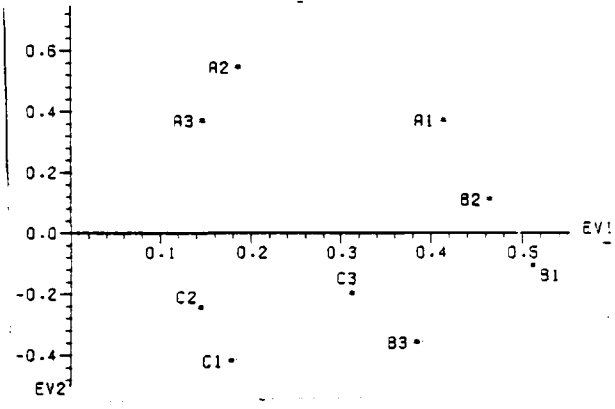
(a) Data treated as departures from class means ; data-set N79W1.



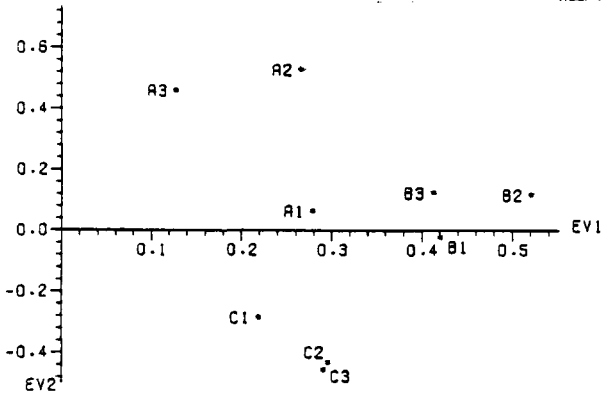
(d) Data treated as departures from class means ; data-set N79W2.



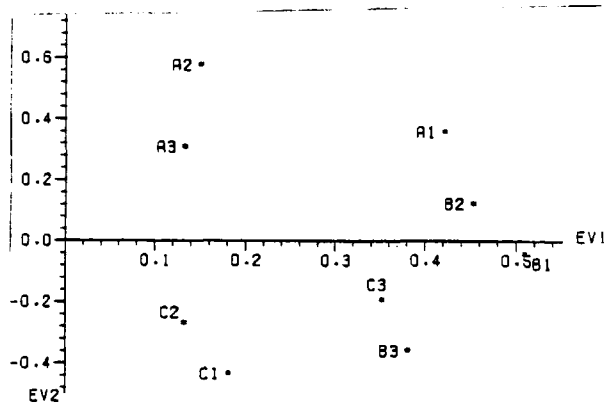
(b) Data treated as departures from seasonal means ; data-set N79W1.



(e) Data treated as departures from seasonal means ; data-set N79W2.



(c) Data treated as residual processes ; data-set N79W1.



(f) Data treated as residual processes ; data-set N79W2.

With n = 220, the number of fast Fourier transform ordinates, T, is 256. The second-order cross-periodograms and auto-and cross-spectrawere computed over 1+T/2 = 129 frequencies. As in the Rutherford study, a thorough investigation was carried out to determine the best set of smoothing perameters and number of passes. The most satisfactory results were obtained with two passes of the modified discrete Daniell window, with half-widths of six and four.

Thirty-three spectral density matrices were estimated at frequencies $f_k = 2\pi k/64$, k=0, 1, ... 32, and, as above, were adjusted to be in the frequency range (0.0, 0.5) cycles per unit time interval. Figures 2.7(a) and (b) show the spectra of the first principal components series for the N79W1 and N79W2 data-sets respectively. Figures 2.8(a) and (b) are the corresponding plots for the second principal components series. Each of the four figures shows the spectra evaluated in the presence and absence of the short-term autocorrelation. The spectral window is shown in the top right-hand corner; approximate 95% confidence intervals are inscribed on the far right-hand side of the plots.

We now discuss the features of the spectra evaluated after the component series have been prewhitened. The most important feature in figure 2.7(a) is the peak around frequency $f_k = 0.15625$ c.p. 15 minutes, which corresponds to a period of about 90 minutes. There is a considerable concentration of power in the low frequency range of the second eigenvalue spectra of week N79W1

Figure 2.7 : Log(base 10) of the first eigenvalue spectra ; NUMAC data.

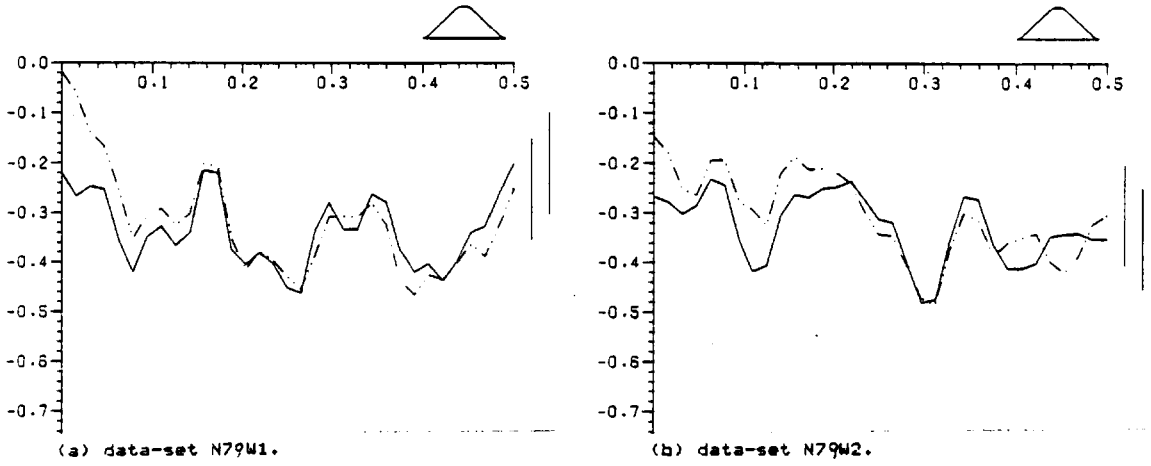
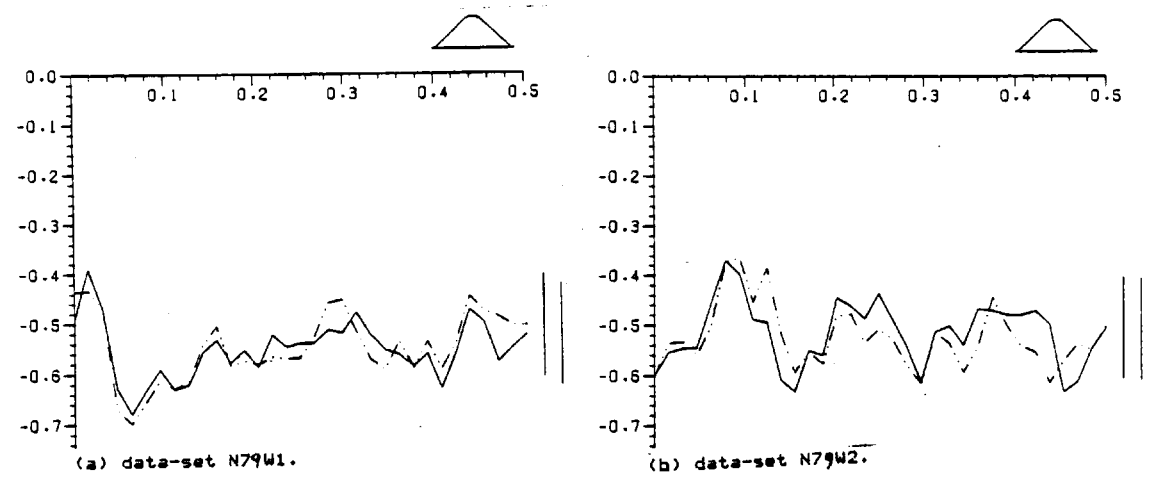


Figure 2.8 : Log(base 10) of the second eigenvalue spectra ; NUMAC data.



Legend

- - - Detrended data before prewhitening

——— Detrended data after prewhitening

Abscissa axis measured in cycles per 15 minutes

(figure 2.8(a)) that persists after the removal of the local correlation. The spectra of the first principal component series of data-set N79W2 (figure 2.7(b)) does not contain any significant peaks. In general, then, the dynamic range of the $\log_{10}$ spectrum ordinates for each of these four plots falls mostly within the approximate 95% confidence limits.

On the basis of these uniformative results, we conclude that the frequency domain analysis for the NUMAC data has added little to the results of the corresponding time domain analysis.

CHAPTER 3

ROBUST REGRESSION TECHNIQUES

It is well known that the least-squares estimation procedure is highly sample-dependent in the sense that it is very sensitive to large errors. Any outlying observations which are inconsistent with the remainder of the sample have the undesirable effect of pulling the fitted line or plane in their direction (see e.g., Tukey,1960). Minor departures from the variance homogeneity and normality assumptions of the least-squares estimation procedure impair the efficiency of the resulting estimates, while just a single outlier can spoil the least-squares estimate.

One way of overcoming these shortcomings in the least-squares estimation procedure is to use a robust alternative. In general, robust regression procedures are designed to accommodate a small proportion of non-Gaussian errors (which are said to 'contaminate' the observed error distribution) in such a way that they have considerably less influence on the final parameter estimates than in the least-squares case. A second related consideration is that robust estimation procedures also perform well when the error distribution is genuinely long-tailed. Two such robust procedures are Least Absolute Deviations (LAD) or ℓ_1-norm estimation and Huber's M-estimation. Both of these methods are invoked in Chapter 4 to predict job turnaround time.

LAD estimation is dealt with in section 3.1. First, we give the known distributional results for these estimators,

then give a partial review of some of the Monte Carlo work done on their finite and small sample properties. The Linear Programming (LP) formulation of the LAD estimation problem is then presented. In this context, the simplex method of solution is discussed, and necessary and sufficient conditions for optimality and uniqueness are given. This is followed by a brief survey of the LP algorithms that solve the LAD minimisation problem. The computational method adopted in this study was Iterated Weighted Least-Squares (IWLS). The procedure is described, and some methods to accelerate its rate of convergence are given. A new result on the distribution of a likelihood ratio test statistic for LAD estimators is developed. The conjectured distribution is examined via a Monte Carlo sampling experiment. The power of this test, relative to that of least-squares, is studied in section 3.3.

Huber's M-estimation is examined in section 3.2. The formulation of the minimisation problem involving Huber (1964)'s ρ function is set up, then some distributional results are given. A partial review of the computational algorithms available for deriving M-estimators is presented; one of these is an IWLS technique, the method adopted in this study. Other aspects of the IWLS method, such as the choice of starting-values and stopping-rules, are also considered. Next, we describe the derivation of a likelihood ratio test for M-estimators, following Schrader and Hettmansperger (1980); the distribution of the test statistic is investigated via Monte Carlo sampling.

Finally, section 3.3 reports on the results of a Monte

Carlo study of the power of the likelihood ratio tests for LAD estimators (developed in section 3.1) and M-estimators relative to that of the classical least-squares'F-test'. We also consider the relative power of the 't-tests' of least-squares, LAD and Huber's M-estimators.

3.1 Least Absolute Deviation (LAD) estimation

- Some Theoretical Results

Consider the usual multiple linear regression model:

$$\underset{\sim}{Y} = X \underset{\sim}{\beta} + \underset{\sim}{\varepsilon} \qquad (3.1)$$

where $\underset{\sim}{Y} = (y_1, \ldots, y_n)^T$ is an (n x 1) vector of responses; $X = (\underset{\sim}{x}_1^T, \ldots, \underset{\sim}{x}_n^T)$ is the (n x p) design matrix, where $\underset{\sim}{x}_i^T = (x_{i1}, \ldots, x_{ip})$, i=1, ..., n is the i^{th} row of X; $\underset{\sim}{\beta} = (\beta_1, \ldots, \beta_p)^T$ is the (p x 1) vector of unknown parameters and $\underset{\sim}{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ is the (n x 1) vector of unobservable disturbances.

When the errors in the linear model (3.1) are non-normally distributed, alternative estimators to least-squares may be desirable. One such alternative is the LAD or ℓ_1-norm estimator. An introduction to LAD estimation, and an extensive bibliography can be found in Gentle (1977). The LAD estimator, $\underset{\sim}{\beta}_L$, is given as the solution to the minimisation problem:

$$\min_{\underset{\sim}{\beta}} \sum_{i=1}^{n} |y_i - \underset{\sim}{x}_i^T \underset{\sim}{\beta}| \qquad (3.2)$$

The objective function $I = \sum_{i=1}^{n} |y_i - \underset{\sim}{x}_i^T \underset{\sim}{\beta}|$ is known as the ℓ_1-norm. Solutions $\hat{\underset{\sim}{\beta}}_L$ to (3.2) are also known as ℓ_1-estimators. Whereas the residuals arising from a least-squares fit have zero mean, those arising from an LAD fit have zero median. Similarly, just as the least-squares estimators $\underset{\sim}{\beta}_{LS}$ are maximum likelihood when the true error model is Gaussian, the LAD estimator $\underset{\sim}{\beta}_L$ may be derived by maximum likelihood when the errors follow a Double Exponential (or Laplace) distribution, whose density is given by $(2\sigma)^{-1} \exp\{-|\varepsilon_i|/\sigma\}$ where σ is a measure of the dispersion of the errors $\underset{\sim}{\varepsilon}$.

Since no tractable finite sample theory exists for LAD estimators, one approach to their distributional properties is to invoke asymptotic theory. Bassett and Koenker (1978) have derived the asymptotic distribution of LAD estimators under certain conditions. Let $\underset{\sim}{\beta}_L^*$ denote a sequence of unique solutions to the minimisation problem (3.2). The authors make two assumptions:

A1: The errors are independently and identically distributed with distribution function F, which is absolutely continuous with density f, positive at the median.

A2: There exists a positive definite matrix Q such that
$\lim_{n\to\infty} n^{-1}(X^TX) = Q$

Under these assumptions, the authors established that $\sqrt{n}(\underset{\sim}{\beta}_L^* - \underset{\sim}{\beta}) \xrightarrow{D} N(o, \omega^2 Q^{-1})$ where ω^2 is the asymptotic variance of

the sample median from a random sample from distribution function F. Rosenberg and Carlson (1977) used Monte Carlo sampling to test the assertion that, for symmetric error distributions, the LAD estimator is approximately normally distributed with variance given by $\lambda(X^TX)^{-1}$, where $n^{-1}\lambda$ is the variance of the median of a sample of size n. Simulation results indicated that the approximation was good for distributions of the explanatory variables or for distributions of the errors having low sample kurtosis, but less satisfactory in the case of high sample kurtosis.

These results suggest a means of obtaining (approximate) standard errors for LAD estimators. These will be required in order to assess the significance of the regression coefficients. If σ_F^2 is the asymptotic variance of the sample mean with distribution function F, and S_F^2 is the asymptotic variance of the sample median, (i.e. the limit as $n\to\infty$ of n x variance), we have:

$$\text{var}(\hat{\underset{\sim}{\beta}}_L)/S_F^2 = \text{var}(\hat{\underset{\sim}{\beta}}_{LS})/\sigma_F^2 \qquad (3.3)$$

If assumption A1 is satisfied, we have the well-known result (see, e.g. Cramer, 1946, p369):

$$S_F^2 = (2f(F^{-1}(\tfrac{1}{2})))^{-2} \qquad (3.4)$$

Expressions(3.3) and (3.4) were employed in this study to obtain estimates of the standard errors of the LAD estimators. When the error model is known a priori, expression (3.4) can be evaluated analytically. In our case, the error

structure was not known. Thus we require an estimate of the probability density function of the residuals evaluated at the median. We arbitrarily replaced the theoretical median $F^{-1}(\frac{1}{2})$, by the median of the sample residuals, which is zero. A kernel method using Epanechikov's parabolic window (Rosenblatt, 1971) was employed. The bandwidth of the window was of the form $b(n) = k\sigma n^{-\alpha}$ (Wertz, 1978). If k=1 and $\alpha = 0.2$, b(n) fulfils certain optimality criteria. As b(n) is proportional to σ, this ensures that the estimate is scale-invariant. Thus the bandwidth was set to $b(n)=\hat{\sigma}_F n^{-0.2}$

Rosenberg and Carlson (1977) suggest a different method to estimate S_F^2 involving numerical integration.

Other Monte Carlo studies have examined the bias and variability of LAD estimators for small sample sizes. Ashar and Wallace (1963) found no evidence of bias, even when the errors were serially correlated, but did conclude that there was a significant difference in the variability of ℓ_1- and ℓ_2- estimators. For Cauchy and Laplace errors, Kiountouzis (1973) found that the LAD estimators had significantly smaller variances than their least-squares counterparts, and also that the LAD estimators were not significantly biased. Pfaffenberger and Dinkel (1978) compared the performance of ℓ_1- and ℓ_2- estimators on the basis of the ratio of their sample mean square errors. Over a variety of sample sizes, the performance of the LAD estimator was superior to that of the ℓ_2-estimator for Cauchy errors, and vice-versa for normal errors. For Laplace errors, the LAD estimator was marginally superior in the majority of cases.

One fruitful method of investigating the properties of the LAD hyperplane is to consider the Linear Programming (LP) formulation of the minimisation problem (3.2). Since the ℓ_1-norm is convex, it can be shown that either a unique hyperplane exists, or there are an infinity of optimal hyperplanes (Barrowdale and Roberts, 1970). This is because any convex combination of optimal hyperplanes is also an optimal hyperplane. Thus, it is known that at least one optimal hyperplane exists. Furthermore, all extremal optimal hyperplanes will pass through at least (and usually exactly) p points (Gentle, Kennedy and Sposito, 1977a).

The simplex method of solution searches for a set of p points which satisfies the optimality condition (given below in Theorem 3.1). The solution is then determined by the plane that passes through these p points. In the degenerate case, which we consider in due course, an optimal solution passes through more than p points. In this case, it is possible to specify more than one optimal hyperplane each having the same estimated parameter vector $\hat{\underset{\sim}{\beta}}$. To avoid this possible ambiguity, it makes sense to specify a solution in ℓ_1 in terms of the set of p points (which is called the basis) through which the plane passes.

Let X_R and $\underset{\sim}{Y}_R$ respectively be the (p x p) submatrix of X and the (p x 1) vector of observations associated with the p points in the basis. Let X_S and $\underset{\sim}{Y}_S$ denote the remaining points, and partition $\underset{\sim}{\varepsilon}$ into $\underset{\sim}{\varepsilon}_S$ and $\underset{\sim}{\varepsilon}_R$ in the same way. The estimate of $\underset{\sim}{\beta}$ is thus the least-squares solution of the points in the R set, and is given by

$$\hat{\underset{\sim}{\beta}} = (X_R^T X_R)^{-1} X_R^T (\underset{\sim}{Y}_R - \underset{\sim}{\varepsilon}_R) \qquad (3.5)$$

where, initially, $\underset{\sim}{\varepsilon}_R = \underset{\sim}{0}$, a (px1) vector of zeroes.

Since, by definition,

$\underset{\sim}{Y}_S = X_S \hat{\underset{\sim}{\beta}} + \underset{\sim}{\varepsilon}_S$, then putting

$$A = X_S(X_R^T X_R)^{-1} X_R^T, \text{ an } (n-p) \times p \text{ matrix} \qquad (3.6)$$

we have that $\underset{\sim}{\varepsilon}_S = \underset{\sim}{Y}_S - X_S \hat{\underset{\sim}{\beta}}$

$$= \underset{\sim}{Y}_S - A \underset{\sim}{Y}_R - A\underset{\sim}{\varepsilon}_R \qquad (3.7)$$

A simplex iteration is defined by a change of points in the basis: depending on a certain criterion (which we give below) one of the p points in the basis is dropped, and replaced by one of the points in the S set by changing one co-ordinate of $\underset{\sim}{\varepsilon}_R$ until one co-ordinate of $\underset{\sim}{\varepsilon}_S$ becomes zero. This procedure continues until an optimal solution has been found. The test for optimality depends on ∇, the change in the norm.

Since we can write:

$\sum_{i=1}^{n} |\varepsilon_i| = \sum_{i \in S} |\varepsilon_i| + \sum_{i \in R} |\varepsilon_i|$, the change in the norm is

$$\nabla = \sum_{j \in R} (|\varepsilon_j| + \sum_{i \in S} \operatorname{sgn}(\varepsilon_i)(A\underset{\sim}{\varepsilon}_R)_i)$$

$$= \sum_{j \in R} (1 + \operatorname{sgn}(\varepsilon_j) \sum_{i \in S} \operatorname{sgn}(\varepsilon_i) A_{ij}) |\varepsilon_j| \qquad (3.8)$$

At each iteration, the simplex procedure changes one of the R set, finds a new parameter estimate using (3.5), then tests for optimality by employing the condition given below in Theorem 3.1.

Theorem 3.1: A necessary and sufficient condition for a solution to be optimal is that:

$$|D_j| \leq 1 \qquad \forall\ j \in R$$

$$\text{where } D_j = \sum_{i \in S} \text{sgn}(\varepsilon_i)\, A_{ij} \tag{3.9}$$

Proof: i) Necessity:: Using (3.8) and (3.9), the change in norm is:

$$\nabla = \sum_{j \in R} (1+\text{sgn}(\varepsilon_j)D_j)|\varepsilon_j| \tag{3.10}$$

Suppose the solution determined by R is optimal. Consider perturbing each ε_j for $j \in R$ in turn. Then, from (3.10):

$$1 + \text{sign}(\varepsilon_j)D_j \geq 0 \quad \forall\ j \in R$$

from which it follows that the solution is optimal only if $|D_j| \leq 1$.

ii) Sufficiency. If $|D_j| \leq 1\ \forall\ j \in R$, from (3.10), $\nabla \geq 0$ for all changes in ε_j, so the norm is minimal. Thus the solution is optimal if $|D_j| \leq 1$.

If it is found that for some $j \in R$, $|D_j| > 1$, the iterative sequence is continued. The observation with the

largest $|D_j|$ is dropped from the R set, to be replaced by an observation from the S set. When an optimal solution has been found by the simplex procedure (which guarantees that it passes through at least p points), it is tested for uniqueness according to the criterion given by Theorem 3.2 below.

Theorem 3.2: A necessary and sufficient condition for a a solution to be unique is that (provided the hyperplane passes through exactly p points):

$$|D_j| < 1 \qquad \forall\ j \in R \tag{3.11}$$

where D_j is given by (3.9).

Proof: i) Necessity: Suppose $|D_k| = 1$ for $k \in R$. Let $\varepsilon_j = 0 \quad \forall\ j \in R$ except $\varepsilon_k = -\delta \operatorname{sgn}(D_k)$ for some $\delta > 0$.

From (3.10), ∇, the change in the norm, is zero. Increase δ from 0 until some $\varepsilon_i = 0$ for $i \in S$, and $\nabla > 0$, at $\delta_o \leq \infty$. Then $\varepsilon_R = -\delta \operatorname{sgn}(D_K)\eta_k$ provides another solution for all $\delta \in (0, \delta_o)$, where η_k is a (p x 1) vector with a 1 in the k^{th} position, and zeroes everywhere else. Then the solution is unique if all $|D_j| < 1 \quad \forall\ j \in R$.

ii) Sufficiency: The solution is unique if all $|D_j| < 1$ since $\varepsilon_R \neq 0$ implies $\nabla > 0$ by 3.8

Remark: The solution found in this proof is genuinely

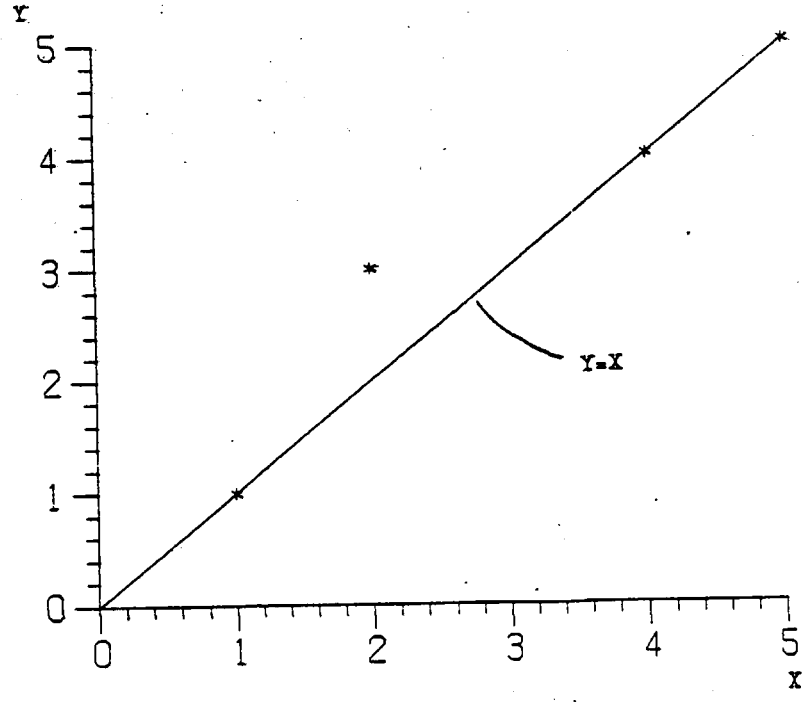
distinct from the original optimal solution provided that:

$$\tilde{\underset{\sim}{\beta}} = (X_R^T X_R)^{-1} X_R^T \ (\underset{\sim}{Y}_R - \underset{\sim}{\varepsilon}_R) \neq \hat{\underset{\sim}{\beta}} = (X_R^T X_R)^{-1} X_R^T \underset{\sim}{Y}_R$$

or equivalently, $\delta(X_R^T X_R)^{-1} X_R^T \underset{\sim}{\eta}_k \neq 0$ with $\delta \neq 0$, or $X_R^T \underset{\sim}{\eta}_k \neq 0$. Thus the conditions of Theorem 3.2 ensure a non-unique solution whenever $\delta \neq 0$ and X is of full rank. However, $\delta_o = 0$ can occur when more than p residuals are zero (i.e. the problem is degenerate). The conditions of theorem 3.2 give us a criterion for multiple optimal solutions, which occur when:

$$|D_j| = 1 \text{ for some } j \in R \qquad (3.12)$$

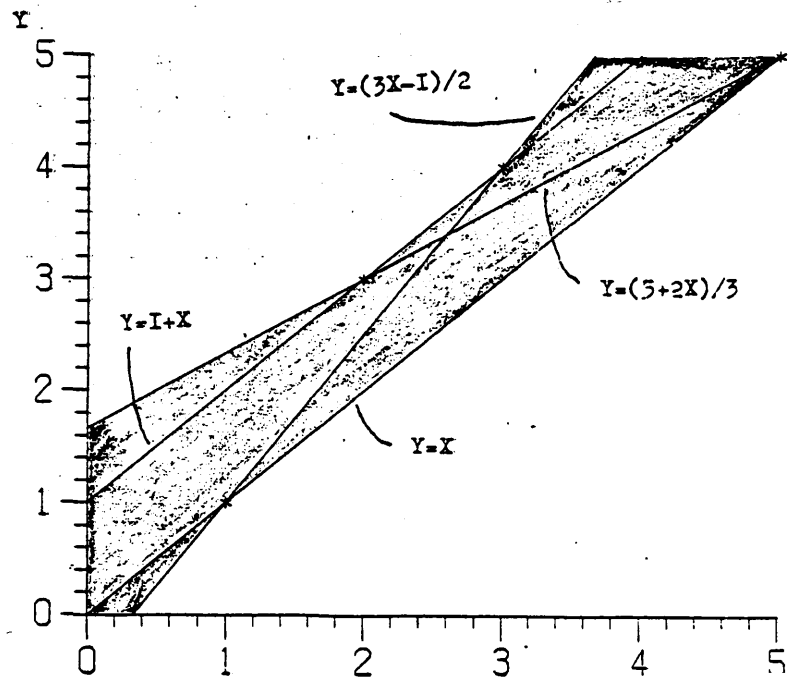
To illustrate some of these features, we consider some simple examples. First, consider the problem of fitting an LAD line through four points whose co-ordinates are (1,1) (2,3), (4,4) and (5,5). The LAD solution, shown in the figure below, is given by the equation Y = X.



This is a degenerate problem. If, for example, we assume that the points (1,1) and (5,5) constitute the final basis (the R set), the matrix A defined by (3.6) has elements $A_{11} = A_{22} = 0.75$ and $A_{12} = A_{21} = 0.25$. These values of A occur when the points are entered in the basis in the order given above. Then, $\underset{\sim}{\varepsilon}_S = (1,0)^T$

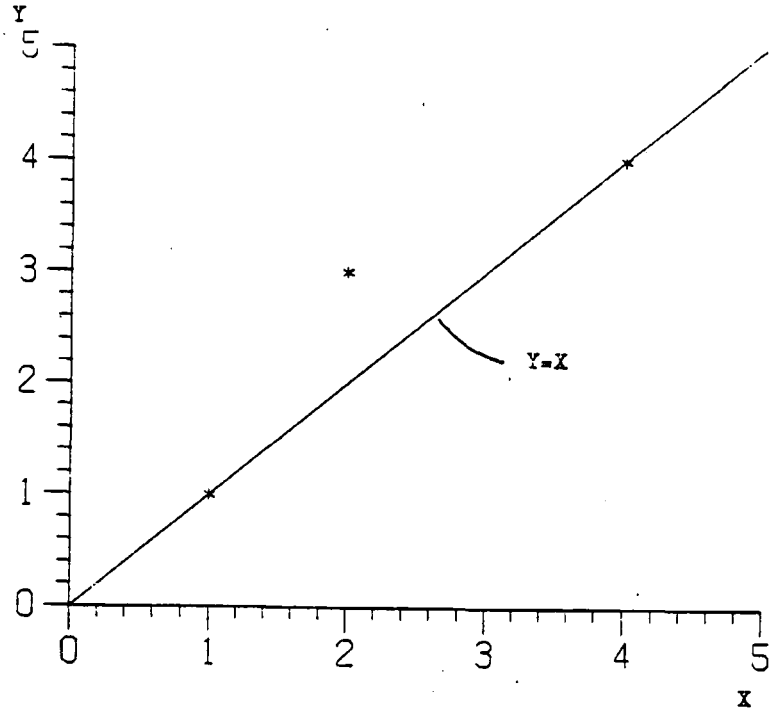
Then, $|D_j| = 1$ for $j = 1, 2$, where D_j is given by (3.9). Using condition (3.12), we have multiple optimal solutions. The same conclusion follows if the pairs (1,1) and (4,4) or (4,4) and (5,5) form the final basis. In each of the three cases, the estimated parameter vector $\hat{\underset{\sim}{\beta}}$ is the same, but the multiplicity of solutions has arisen because the points in the final basis are different. The robustness of the solution is evident in that the solution Y = X is the same, provided that the point which does not lie on the line lies within the range of the other three points.

Next, consider the LAD fit through the points (1,1),(2,3) (3,4) and (5,5). The four solutions which each pass through two points are Y=X, Y=1 + X, Y=(3X-1)/2 and Y=(5+2X)/3. These are shown on the figure below.



The norm of each of these solutions is 2. Each of them is optimal by Theorem 3.1, because each has $|D_j| = 1$ for $j = 1, 2$. Any line which is a convex combination of these solutions is also optimal (e.g. the lines $Y = \alpha + X$ for $0 \leq \alpha \leq 1$). This illustrates the general principle that an optimal solution with p parameters does not necessarily have to pass through exactly p points. In the figure above, any solution contained within the shaded area is optimal. This is an example of the case where there are multiple optimal solutions to a non-degenerate problem.

The final case we consider is when the solution is unique. The unique LAD solution to the points (1,1), (2,3) and(4,4) is the line Y=X.



With $\underset{\sim}{\varepsilon}_S = (1)$, we have $|D_{(1,1)}| = 2/3$ and $|D_{(4,4)}| = 1/3$, which clearly satisfies the uniqueness condition (3.11) in Theorem 3.2.

- The primal and dual (LP) formulations of the LAD problem.

The idea of translating the minimisation problem (3.2) into a primal LP problem was due to Charnes, Cooper and Ferguson (1955). This is achieved by representing the i^{th} deviation ε_i as the difference between two non-negative variables $(P_i - Q_i)$, where $P_i = \max(\varepsilon_i, 0)$ and $Q_i = \max(-\varepsilon_i, 0)$. Thus we may write $|\varepsilon_i| = P_i + Q_i$. The LP formulation of (3.2) is:

$$\text{minimise } \sum_{i=1}^{n}(P_i + Q_i) \tag{3.13}$$

subject to the n constraints $y_i = \sum_{j=1}^{p} \beta_j x_{ij} + P_i - Q_i$, $i=1, \ldots, n$, with $(2n + p)$ variables, $P_i, Q_i \geq 0$ for $i = 1, \ldots, n$ and β_j, $j = 1, \ldots, p$ unrestricted.

A slight modification to the primal formulation (3.13) is required to convert it to a standard LP problem. This is done by writing $\beta_j = \gamma_j - \eta_j$ where $\gamma_j = \max(\beta_j, 0)$ and $\eta_j = \max(-\beta_j, 0)$. Then, the standard LP problem is:

$$\text{minimise } \sum_{i=1}^{n}(P_i + Q_i) \tag{3.14}$$

subject to the n constraints $y_i = \sum_{j=1}^{p}(\gamma_j - \eta_j)x_{ij} + P_i - Q_i$, $i=1, \ldots, n$, with $(2n + 2p)$ variables $P_i, Q_i \geq 0$, $i=1,\ldots,n$ and $\gamma_j, \eta_j \geq 0$, $j=1,\ldots,p$.

A straightforward application of the simplex algorithm to (3.14) results in a basis inverse of size (n x n) with (2n + p) variables. In large problems, the size of the matrices required for computing the solution is liable to become unmanageable. A more parsimonious approach, proposed

by Wagner (1959), is to solve the dual version of the problem. This is found by introducing Lagrange multipliers (dual variables), λ. Writing $\underset{\sim}{P} = \max(\underset{\sim}{\varepsilon}, \underset{\sim}{0})$ and $\underset{\sim}{Q} = \max(-\underset{\sim}{\varepsilon}, \underset{\sim}{0})$, we may re-write (3.13) as:

$$\text{maximise } -\underset{\sim}{1}^T(\underset{\sim}{P} + \underset{\sim}{Q})$$
$$\text{subject to } \underset{\sim}{Y} = X\underset{\sim}{\beta} + \underset{\sim}{P} - \underset{\sim}{Q}$$
$$\text{with } \underset{\sim}{P}, \underset{\sim}{Q} \geq 0 \text{ and } \underset{\sim}{\beta} \text{ unrestricted.}$$

If $\underset{\sim}{\lambda} = (\lambda_1, \ldots, \lambda_n)^T$ is the (n x 1) vector of Lagrange multipliers, we have:

$$\text{maximise } L(P, Q, X, \lambda) = -\underset{\sim}{1}^T(\underset{\sim}{P} + \underset{\sim}{Q}) + \underset{\sim}{\lambda}^T\underset{\sim}{Y} - \underset{\sim}{\lambda}^T X\underset{\sim}{\beta} - \underset{\sim}{\lambda}^T\underset{\sim}{P} + \underset{\sim}{\lambda}^T\underset{\sim}{Q}$$

$$= \underset{\sim}{\lambda}^T\underset{\sim}{Y} - \underset{\sim}{\lambda}^T X\underset{\sim}{\beta} - (\underset{\sim}{\lambda}+\underset{\sim}{1})^T\underset{\sim}{P} - (\underset{\sim}{\lambda}-\underset{\sim}{1})^T\underset{\sim}{Q}$$

The dual problem is then:

$$\text{minimise } \underset{\sim}{\lambda}^T\underset{\sim}{Y} \qquad (3.15)$$

$$\text{subject to } \sum_{i=1}^{n} \lambda_i = 0, \ \underset{\sim}{\lambda}^T X = \underset{\sim}{0} \text{ with } |\lambda_i| \leq 1, \ i=1,\ldots,n.$$

Then $P_i > 0$ only if $\lambda_i = -1$ and $Q_i > 0$ only if $\lambda_i = 1$. An optimal solution occurs if, and only if, $|\lambda_i| \leq 1 \ \forall i$ in the final basis; a unique solution occurs if, and only if, $|\lambda_i| < 1 \ \forall i$ in the final basis. Multiple optimal solutions arise when the dual variable λ_j associated with the j^{th} observation has an extreme point value (either + 1 or -1).

The bounded-variable LP problem (3.15) is more parsimonious in terms of storage because the size of the basis inverse depends not on n but on p.

Appa and Smith (1973) used LP theory to derive two fundamental properties of the LAD hyperplane when the model contains a constant term. The first is that there exists an LAD hyperplane that passes through p of the n points: this is a necessary, but not a sufficient, condition for uniqueness. The second property is that the solution cannot be optimal unless it satisfies the following condition:

$$|n_+ - n_-| \leq p \qquad (3.16)$$

where n_+ and n_- are respectively the number of points above and below the line. This second property assumes that there are no multiple optimal solutions (Sposito and Smith, 1973), which may, as we have seen, occur even in the full rank case. Gentle, Kennedy and Sposito (1977b) have pointed out that the two properties do not necessarily hold when there are constraints on the parameters.

- Computational Algorithms to solve the LAD minimisation problem.

We now review some of the LP algorithms to solve the linear regression problem that have appeared in the literature.

For the simple regression case, LP algorithms for (3.2) have been given by Sadovski (1974) (see also, Sposito, 1976), and Armstrong and Kung (1978).

For the multiple regression case, Barrowdale and Young

(1966) and Spyroupoulous et al (1973) have both given ALGOL codes to solve the primal version of the LP problem (given by 3.13). A new technique to solve the primal, which results in a substantial improvement in cpu solution time over the standard simplex method, was described by Barrowdale and Roberts (1973) (see also Kennedy and Gentle, 1980, section 11.2). A FORTRAN routine implementing this method is given by Barrowdale and Roberts (1974). The algorithm overcomes the large core requirements of the primal by exploiting the relationships amongst the variables. Armstrong Frome and Kung (1979) have also presented a FORTRAN code based on this method. Bartels and Conn (1977) have developed a different LP method for solving the problem, while Abdelmalek (1974, 1975) has given two algorithms that solve the dual version of the LP problem (given by 3.15).

Computational comparisons of some of these algorithms have been conducted by several authors. For example, Gilsinn et al (1977) found that, out of four algorithms, the one proposed by Armstrong and Frome (1977) was the quickest in terms of cpu solution time. Kennedy and Gentle (1977) examined the numerical stability of the Barrowdale-Roberts (1974) algorithm. The numerical accuracy of the results was found to be highly dependent on the conditioning of the design matrix, as least-squares results are known to be.

Improvements in the rate of convergence of some of these algorithms can be made in a number of ways. Results in McCormick and Sposito (1976) and Sposito, Hand and McCormick (1977) showed that the rate of convergence could be

accelerated if the least-squares solution is used as an initial adjustment. If $\hat{\underset{\sim}{\beta}}_{LS}$ is the vector of least-squares estimates, and the LAD solution is found for the secondary problem $(\underset{\sim}{Y} - X\hat{\underset{\sim}{\beta}}_{LS})$ giving parameter estimates $\hat{\underset{\sim}{\beta}}^*$, then the solution to the original problem $\hat{\underset{\sim}{\beta}}_L$ is given by the coefficients $\hat{\underset{\sim}{\beta}}_L = \hat{\underset{\sim}{\beta}}_{LS} + \hat{\underset{\sim}{\beta}}^*$.

The format of the Barrowdale and Roberts (1974) and Armstrong, Frome and Kung (1979) algorithms permits the use of this approach. The latter includes an additional option which allows the user to input an inital basis. A reasonable initial basis may be provided by the p least-squares residuals which are closest to zero. In the author's experience, this can lead to a reduction in total cpu time of approximately 10% for problems with large (n/p).

The LP algorithms discussed above give solutions to problems in which there are no side-constraints on the parameters. Algorithms to solve the restricted LAD case have been given by Barrowdale and Roberts (1978) and Armstrong and Hultz (1977). Barrowdale and Roberts (1977) have conducted a survey of the work in this area.

Roodman (1974) has developed a procedure, defined within the context of the LP problem, for obtaining the best subset of explanatory variables in LAD regression.

Another method to minimise (3.2) has been given by Abdelmalek (1971). It entails the determination of the ℓ_1-norm solution as the limit of ℓ_p-approximations (starting at ℓ_2) as $p \to 1$. Bartels, Conn and Sinclair (1978) have developed

a technique that minimises (3.2) directly, by minimising piecewise differentiable functions.

The main advantages of LP methods are that they allow for easy constraint imposition, and that the iterative sequence is known to terminate in a finite number of iterations.

A drawback is that sometimes the accumulation of round-off errors (or poor programming) can result in premature convergence to an incorrect solution. The validity of the solution may be checked in the following way: assume that the LP algorithm terminates at a solution $\hat{\underset{\sim}{\beta}}_L$. Then, recompute the solution for the problem $(\underset{\sim}{Y} - X\hat{\underset{\sim}{\beta}}_L)$. If the algorithm has not terminated prematurely, the resulting parameter estimates will all be zero. These convergence difficulties are prone to occur when the range of the X variables differs greatly. One method which may help in this case is to first solve a secondary problem in which the columns of the X matrix are rescaled in such a way that the largest element in each column is of the order of unity. The solution to the original problem is then retrieved by adjusting the resulting parameter estimates in the appropriate way (see NAG Mark 7, Vol. 2, E02GAF, P5).

The main competitor to LP methods is iterated weighted least-squares (IWLS). This method is popular amongst statisticians principally because it can be easily used in association with existing least-squares software packages.

The method is as follows: let

$$I^{(m)} = \sum_{i=1}^{n} |y_i - \underset{\sim}{x}_i^T \hat{\underset{\sim}{\beta}}^{(m)}| \qquad (3.17)$$

be the sum of absolute deviations after m iterations, and where $\hat{\underset{\sim}{\beta}}^{(m)}$ is the estimated parameter vector after m iterations. The LAD solution can be obtained by considering the weighted least-squares problem:

$$I^{(m+1)} = \sum_{i=1}^{n} w_i^{(m)} (y_i - \underset{\sim}{x}_i^T \hat{\underset{\sim}{\beta}}^{(m+1)})^2 \qquad (3.18)$$

where the weights on the $(m+1)^{st}$ iteration are defined by

$$w_i^{(m)} = 1/|y_i - \underset{\sim}{x}_i^T \hat{\underset{\sim}{\beta}}^{(m)}| \qquad i=1,\ldots, n \qquad (3.19)$$

Expressions (3.18) and (3.19) show that the weights employed on each iteration are computed by reciprocating the absolute deviations of the residuals arising from the previous iteration. It follows that small residuals are given large weights and vice-versa. From the LP formulation of the problem, we know that some of the residuals will tend to zero, in which case their corresponding weights will become very large. This can lead to numerical instabilities, which impede the rate of convergence. Schlossmacher (1973) suggested one method of overcoming this difficulty: since the contribution of any near zero residual to the total sum of absolute deviations is small, he advocated that any residual that has become small relative to the other residuals be set to zero weight. The weights would be computed according to the following scheme:

$$w_i^{(m)} = \begin{cases} 1/|y_i - \underset{\sim}{x}_i^T \hat{\underset{\sim}{\beta}}^{(m)}| & \text{if } |y_i - \underset{\sim}{x}_i^T \hat{\underset{\sim}{\beta}}^{(m)}| \geq \delta \\ 0 & \text{otherwise} \end{cases}$$

where δ is some small positive constant. However, the implementation of this scheme can itself lead to numerical instabilities. A heuristic argument is as follows: residuals which are tending toward zero, while their absolute values remain larger than δ, are allocated weights that increase after each iteration. When the absolute values of these residuals finally became less than δ, they are suddenly given a zero weight. The undesirable situation will then arise that gives an observation a maximum weight of the order of δ^{-1} on one iteration, and, on the next, gives it a zero weight.

An alternative method (and the one adopted here) is to Winsorize the residuals. Given a fixed maximum weight λ, which is of the order of 10^9, the scheme is:

$$w_i^{(m)} = \begin{cases} 1/\lambda & \text{if } |y_i - \underset{\sim}{x}_i^T \hat{\underset{\sim}{\beta}}^{(m)}| \geq \lambda^{-1} \\ 1/|y_i - \underset{\sim}{x}_i^T \hat{\underset{\sim}{\beta}}^{(m)}| & \text{otherwise} \end{cases} \quad (3.20)$$

A residual whose absolute value is less than λ^{-1} is thus accorded a weight of 10^9. This value should, of course, be adjusted according to the characteristics of the sample: in our case, the residuals were of order 10^2. This technique is more stable numerically than that proposed by Schlossmacher (1973).

Two important aspects in any iterative procedure are first, the starting-values, and second, the stopping-rule. Least-squares starting-values were used to begin the IWLS

iterative procedure to obtain the full LAD regression fits. The convergence criterion was based on the deviance statistic given by expression (3.39). A more detailed discussion of stopping-rules and starting-values is postponed until the next section (i.e. section 3.2) which deals with them in the context of Huber's M-estimators.

After fitting the full p-parameter LAD model, we required a technique to fit a q (<p) subset model. Rather than start with the least-squares root from the q-parameter fit, it was thought to be more reasonable to assume that the residuals from the full p-parameter LAD fit would be closer to the eventual solution. If this be the case, this technique of fitting a subset LAD model should reduce total cpu solution time. The adopted approach was as follows: if the p-parameter fit (convergent after k iterations) is

$$y_i = \underset{\sim}{x}_i^T \hat{\underset{\sim}{\beta}}^{(k)} + z_i^{(k)} \qquad i=1, \ldots, n$$

Then the initial weights to find the q-variable subset fit are:

$$w_i^{(o)} = \begin{cases} 1/|y_i - \underset{\sim}{x}_i^T \hat{\underset{\sim}{\beta}}^{(k)}| & \text{if } |y_i - \underset{\sim}{x}_i^T \hat{\underset{\sim}{\beta}}^{(k)}| \geq \lambda^{-1} \\ \lambda & \text{otherwise} \end{cases}$$

where $\lambda = 10^9$. The required solution was then found by following the IWLS technique described above.

One of the shortcomings of the IWLS procedure is that no absolute convergence theory has yet been proved for the LAD case; another is that the rate of convergence is often

slow. The rate of convergence may be accelerated by incorporating into the IWLS technique certain properties of the theoretical LAD hyperplane. From the LP formulation of the LAD problem, it is known that there is a solution to (3.2) in which the fitted line passes through at least p points. If a check is kept on the direction and magnitude of change of each residual after successive iterations, it may be possible to spot those residuals which are tending to zero. If these are set to zero prematurely, the rate of convergence may be improved. It is also known that a necessary (but not a sufficient) condition to minimise the sum of absolute deviations is given by expression (3.16), that is, $|n_+ - n_-| \leq p$. If, after each IWLS iteration, a count of n_+ and n_- (the numbers of points above and below the fitted line repectively) the rate of convergence may be accelerated by ensuring that condition (3.16) is not violated.

While the incorporation of these features into the IWLS algorithm may reduce total cpu solution time, it is certain that an increase in the storage requirements of the algorithm will ensure. This may adversely affect the main advantage of the IWLS algorithm over LP methods, which is that the former generally requires less storage than the latter. Armstrong and Frome (1976) compared the performance of a primal LP algorithm (similar to the one suggested by Barrowdale and Roberts, 1973, 1974) to the IWLS algorithm described by Schlossmacher (1973). The two algorithms were found to converge to the same optimal solution, within an acceptable tolerance level. Their results support the general concensus of opinion, which is that LP methods are relatively

quicker over all problem sizes. In table 3.1, we have compared the performance of three of these algorithms in fitting six data-sets by a linear regression model with a constant term and seven explanatory variables.

Table 3.1 : Comparative performance of three algorithms in fitting an LAD regression line to six series from data-set N79W1.

Number of parameters : 7.

	No.of	No.of iterations			CPU time (secs)			Mean absolute deviation		
Class	obs.	IWLS	AFK	BR	IWLS	AFK	BR	IWLS	AFK	BR
1-FS	875	42	41	43	47.4	16.2	21.6	16.77	16.77	16.78
1-RS	831	68	41	34	65.6	11.3	18.1	4.63	4.63	4.64
2	1138	67	39	39	80.9	12.8	27.9	58.84	58.84	58.84
3	880	63	41	38	67.9	9.7	20.9	64.92	64.92	65.04
4	815	36	29	35	36.3	7.9	18.6	166.40	166.40	177.72
5	482	48	40	34	30.2	4.8	9.7	136.42	136.42	136.49

Legend.

Algorithm

IWLS : Iterated Weighted Least-Squares

AFK ; Armstrong, Frome, and Kung (1979)

BR ; Barrowdale and Roberts (1974)

Method of solution

LS starting-values ;
Winsorised weights (see eqn.3.20) ;
convergence by deviance statistic (see eqn.3.39).

LS starting-values ; no initial basis.

LS starting-values.

FS - Full Sample.

RS - Reduced Sample.

The numerical accuracy of the results from the IWLS and AFK routines was found to be the same, within an acceptable tolerance level. The coefficients of the BR routine differed quite considerably from those arising from the other two routines. This is presumably due to the premature convergence of the BR routine; a comparison of the figures in the final three columns shows that this routine does not always converge to the minimum. The AFK routine is

clearly the quickest, requiring between 0.125 and 0.33 cp seconds per iteration: the IWLS routine required, on average, about one cp second per iteration.

- The development of a likelihood ratio test for LAD-estimators.

We now report on the results of a Monte Carlo study of the distribution of a likelihood ratio test statistic for LAD estimators. The statistic was used to determine the best subset LAD fit. First, we partition the linear model $\underset{\sim}{Y} = X\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$ given by equation 3.1 in the following way:

$$\underset{\sim}{Y} = X_1\underset{\sim}{\beta}_1 + X_2\underset{\sim}{\beta}_2 + \underset{\sim}{\varepsilon} \qquad (3.21)$$

where $X = (X_1, X_2)$ and $\underset{\sim}{\beta}^T = (\underset{\sim}{\beta}_1^T, \underset{\sim}{\beta}_2^T)$ where X_1 is (n x q), X_2 is n x (p-q), $\underset{\sim}{\beta}_1$ is (q x 1) and $\underset{\sim}{\beta}_2$ is (p-q) x 1. Then our interest is in the hypothesis:

$$H_o: \quad \underset{\sim}{\beta}_2 = \underset{\sim}{0} \qquad (3.22)$$

against H_1; $\underset{\sim}{\beta}_2 \neq \underset{\sim}{0}$

We make two assumptions, given by A1 and A2 above (see the beginning of section 3.1).

Define the LAD estimator $\underset{\sim}{\beta}_L$ as the solution to the minimisation problem:

$$\min_{\underset{\sim}{\beta}} \sum_{i=1}^{n} |y_i - \underset{\sim}{x}_i^T\underset{\sim}{\beta}| \qquad (3.23)$$

Since the estimator $\underset{\sim}{\beta}_L$ is maximum likelihood under the Double Exponential distribution, the likelihood of the full model (when H_1 is true and F is Laplace) is;

$$L_{H_1} = (2\sigma)^{-n} \exp\{-\sum_{i=1}^{n} |y_i - x_1\underset{\sim}{\beta}_1 - x_2\underset{\sim}{\beta}_2|/\sigma\}$$

where σ is the scale parameter. The likelihood of the reduced model (when H_o is true) is;

$$L_{H_o} = (2\sigma)^{-n} \exp\{-\sum_{i=1}^{n} |y_i - x_1\underset{\sim}{\beta}_1|/\sigma\}$$

Let V(F) be the minimum of (3.23) when H_1 is true, and V(R) be the minimum of (3.23) subject to H_o; V(F) and V(R) are the ℓ_1-norms of the full and reduced models respectively. Consider the likelihood ratio:

$$\Lambda = L_{H_o}/L_{H_1}$$

$$= \exp\{-V(R)/\sigma\}/\exp\{-V(F)/\sigma\}$$

If the full model has fitted parameters $\underset{\sim}{\beta}_L^*$, and the reduced model, $\hat{\underset{\sim}{\beta}}_L$, the usual likelihood ratio test of (3.21) would lead us to consider the statistic:

$$-2 \log_e \Lambda = 2\sigma^{-1}\{V(\hat{\underset{\sim}{\beta}}_L) - V(\underset{\sim}{\beta}_L^*)\}/(p-q)$$

$$= 2\sigma^{-1}\{D(\hat{\underset{\sim}{\beta}})\}/(p-q)$$

where $D(\hat{\underset{\sim}{\beta}})$ is the reduction is the ℓ_1-norm obtained by passing from the reduced to the full model. Asymptotic distribution theory for LAD estimators suggests that the constant of

proportionality is ω^{-1}, where ω^{-2} is the asymptotic variance of the sample median under distribution function F. Thus

$$\omega^2 = (2f(F^{-1}(\tfrac{1}{2})))^{-2} \qquad (3.24)$$

We conjecture that a test of (3.22) for LAD estimators may be based on the statistic;

$$\Lambda = 2\omega^{-1}\{V(R) - V(F)\}/(p-q) \qquad (3.25)$$

Using assumption A1 and expression(3.24), we would consider estimating (3.25) by:

$$\hat{\Lambda} = 4\hat{f}(o)\{V(R) - V(F)\}/(p-q) \qquad (3.26)$$

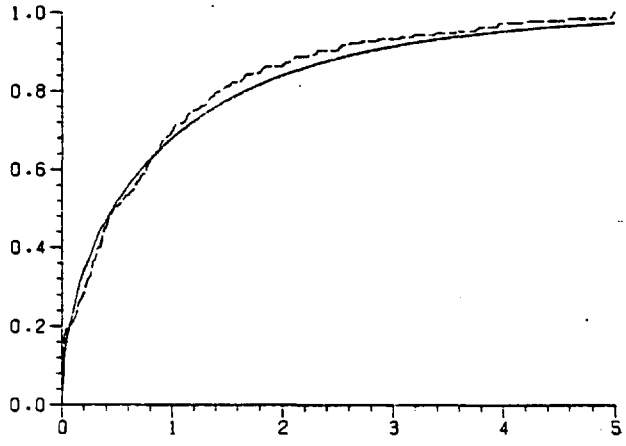
where $\hat{f}(o)$ is an estimate of the error density evaluated at the origin. Here, as before, we have arbitrarily replaced the theoretical median $F^{-1}(\frac{1}{2})$ by the sample residual median, which is zero. From the maximum likelihood derivation of the statistic Λ, we conjecture that its asymptotic distribution is approximated by that of a χ^2 variable on (p-q) degrees of freedom. The assumed distribution of this test was studied via a Monte Carlo sampling experiment.

The fixed design matrix was one replicate of a 3^3 factorial layout with seven unknown parameters, $\beta_j = 1.0$, j=1, ..., 6, and a constant term $\mu = 1.0$. The null hypothesis was H_o: $\beta_6 = 0$, so (p-q) = 1. Each observation y_i, i=1, ..., 27, was generated from the linear regression model $y_i = \mu + \sum_{j=i}^{5} \beta_j x_{ij} + \epsilon_i$ i=1, ..., 27

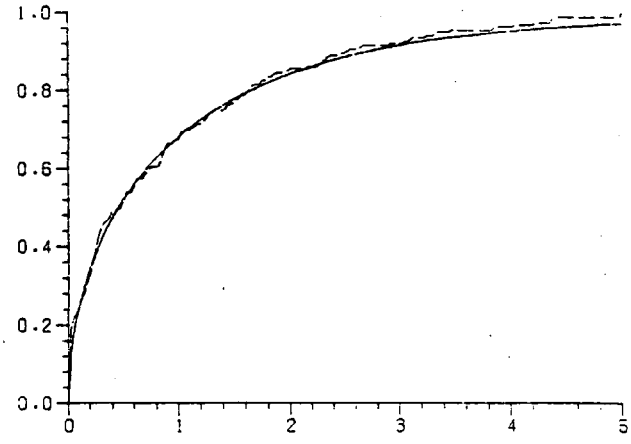
Each of the four experiments conducted consisted of 200 simulations. The LAD estimators were fitted to the full and reduced models using an algorithm given by Armstrong, Frome and Kung (1979). After each simulation, the quantities V(R) and V(F) were computed, and the statistic L given by (3.26) was calculated. For the four experiments, the cumulative frequency distribution of the observed Likelihood Ratio Test (LRT) statistic was plotted against its conjectured theoretical distribution, in this case a χ^2 distribution on one degree of freedom. The incomplete gamma integral was computed by a method proposed by Chi-Leung (1980): Stirling's well-known formula was used to evaluate the Gamma (factorial) function. Each theoretical cumulative frequency distribution is plotted up to the 97.5% point.

The first experiment was conducted using a 'bootstrap' procedure (see, e.g. Efrom, 1979), whereby the errors were sampled from the residuals arising from one of the LAD fits to some NUMAC turnaround data. The histogram of the sample (which has 831 points) is shown in figure 4.2(c). The sampling was performed without replacement within each experimental sample, but with replacement between each experimental sample. This avoided the possibility of selecting the same point more than once within each Monte Carlo sample. In this case, the estimate of the density of experimental residuals (those arising from each LAD fit in the Monte Carlo experiment) at the median was taken to be the density of the entire sample evaluated at the median. This gave $\hat{f}(o) = 0.09814$. Figure 3.1(a) shows that the distribution of the observed LRT statistic is approximately the same as the

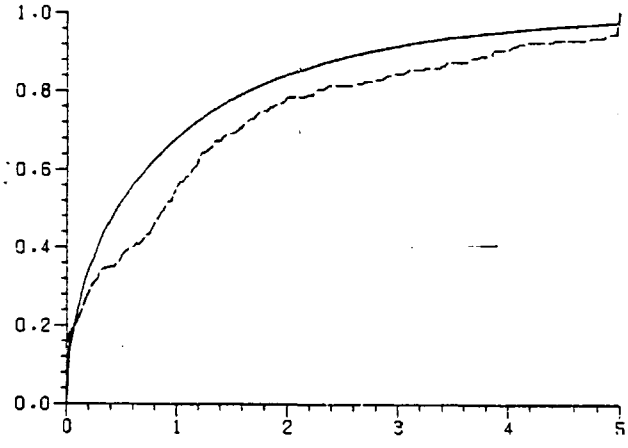
Figure 3.1 : Cumulative frequency distribution of observed likelihood ratio test statistic for LAN estimators.



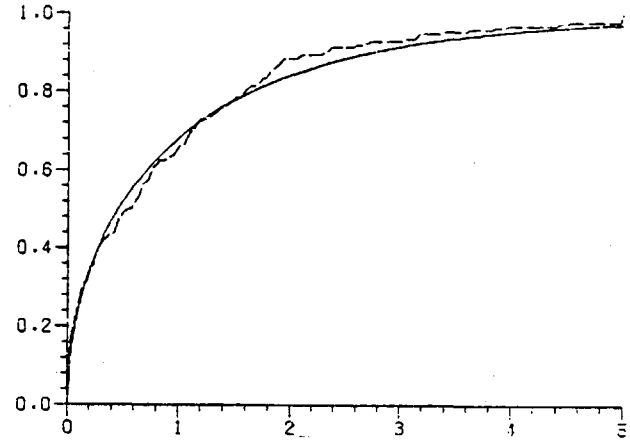
(a) errors sampled from a 'bootstrap' distribution.



(b) errors sampled from a Normal(0,1) distribution.



(c) errors sampled from a Cauchy(0,1) distribution.



(d) errors sampled from a negative exponential(2) distribution.

Legend

—— Theoretical Chi-squared distribution on 1 degree of freedom.

---- Observed frequency distribution.

conjectured χ_1^2 distribution.

In the last three experiments, the disturbances ε_i were sampled from the following distributions:

1. Normal, mean zero, variance $\sigma^2 = 1$.
2. Cauchy, location parameter zero, scale parameter $\delta = 1$.
3. Negative exponential, mean parameter $\eta = 2$.

The errors were generated by using the appropriate NAG library subroutines.

Rather than estimate the quantity $\hat{f}(o)$ from each set of experimental residuals, we used its known theoretical value: for the three distributions specified above, these were $(2\pi\sigma)^{-0.5}$, $(\pi\delta)^{-1}$, and $(2\eta)^{-1}$ respectively.

For normal errors (figure 3.1(b)), the fit between the empirical and theoretical distributions is good. When the errors are sampled from the heavy-tailed Cauchy distribution, figure 3.1(c) shows that the empirical frequency distribution tends to underestimate the theoretical distribution. Figure 3.1(d) suggests that the conjectured χ_1^2 distribution is an adequate representation even when the errors are sampled from an asymmetric distribution.

These preliminary results suggest that the χ^2 approximate for the distribution of the LRT statistic Λ given by (3.25), for LAD estimators is good for normal and asymmetric error models, but not for very heavy-tailed error distributions as typified by the Cauchy distribution.

The results of power study of this test are reported in section 3.3

3.2 Huber's M-estimators

- Some theoretical results.

A general definition of a robust estimate T is as follows: suppose that H is a conjectured distribution of the observations and $\{G_i\}$ is a family of true observation distributions where each $\{G_i\}$ is fairly close to H according to some appropriate metric. Then T is robust at H if its performance is good over the family of distributions $\{H\} \cup \{G_i\}$(Kleiner et al, 1979). For example the G_i may be symmetric and long-tailed, and H Gaussian. Alternatively the G_i may also be Gaussian with different variance to that of H. In both these cases, the result is the so-called contaminated normal distribution.

Because exact finite sample theory in this situation is usually impossible, Huber (1964) suggested that the robustness of T should be gauged according to its asymptotic properties. Consider two examples, the sample mean and the sample median: the former has asymptotic variance $\int(t-\mu)^2F(dt)$, and the latter $(2F'(0))^{-2}$, where F is the underlying distribution which is assumed to be absolutely continuous and to have a positive derivative at the origin. These expressions show that the asymptotic variance of the sample mean is extremely sensitive to large outliers appearing in the tails of the distribution, and that the asymptotic variance of the sample median ignores the behaviour of F everywhere except about its median. Huber (1964) proposed a compromise between the two; it both provides protection against large outlying points and takes more account of

the behaviour of F at points in the interval around its median. The sample mean, despite its inadequacies, does have optimal properties (it is. B.L.U.E.) when the (finite) variance is constant and the observations are independent. Similarly, the sample median, which is the location analogue of the LAD-estimator in regression, does offer resistance to large disturbances. Huber (1964)'s ρ function is a synthesis of the two:

$$\rho(t) = \begin{cases} t^2/2 & |t| \leq c \\ c|t| - c^2/2 & |t| > c \end{cases} \qquad (3.27)$$

As $c \to 0$, this gives the LAD-estimator, and as $c \to \infty$, the least-squares estimator

If $y_1, \ldots, y_n$ are a sample of n independent random variables with common distribution function, F, and an estimate T is defined by the property that is minimises the function $\sum_{i=1}^{n} \rho(y_i - T)$, where $\rho(t)$ is given in (3.27), then $\sqrt{n}T$ attains the maximal asymptotic variance when G_i is symmetric and H is Gaussian (Huber, 1964, p82). For any non-constant function $\rho(t)$, the general class of estimators that minimise $\sum_{i=1}^{n} \rho(y_i - T)$ defines all maximum likelihood type (or M-) estimators.

In the regression context, an M-estimator $\hat{\underset{\sim}{\beta}}_M$ is given as the solution to the minimisation problem:

$$\underset{\underset{\sim}{\beta}}{\text{minimise}} \sum_{i=1}^{n} \rho\{(y_i - \underset{\sim}{x}_i^T \underset{\sim}{\beta})/S\} \qquad (3.28)$$

where S is a (robust) estimate of residual scale. Thus, for a trim constant c, any residual whose absolute value exceeds

cS is automatically Winsorised. We discuss the estimation of S, in due course (see, e.g. equation 3.37).

If (3.28) is differentiated, we have, defining $\psi(t) = \rho'(t) = \max(-c, \min(c, t))$,

$$\sum_{i=1}^{n} x_{ij}\, \psi\{(y_i - \underset{\sim}{x}_i^T \underset{\sim}{\beta})/S\} = 0 \qquad j=1, \ldots, p \qquad (3.29)$$

Apart from a multiplicative factor depending upon the underlying distribution function F, the function $\psi(t)$ is Hampel (1974)'s influence curve. If $\rho(t)$ is convex, as it is for Huber's ρ function, (3.29) is an equivalent formulation of (3.28). There are two classes of $\rho(\cdot)$ functions: convex and non-convex. Convex $\rho(\cdot)$ restricts $\psi(\cdot)$ to be monotone. Non-monotone $\psi(\cdot)$ functions with $\psi(t) \to 0$ as $|t| \to \infty$, are sometimes called redescending ψ functions.

The Huber ρ function given by (3.27) was used in this study. It corresponds to maximum likelihood for an error distribution which is normal in the centre with double-exponential tails. The following distributional results are taken from Huber (1973): M-estimators defined by (3.28) are consistent and asymptotically normally distributed when three regularity conditions (Huber, 1973, p 804) hold. An expression for the asymptotic variance of an M-estimator β_j is:

$$V_{\psi,F}(\beta_j) = \frac{E_F(\psi^2\{(y_i - \underset{\sim}{x}_i^T \underset{\sim}{\beta})/S\})((X^TX)^{-1})_{jj}}{E_F(\psi'\{(y_i - \underset{\sim}{x}_i^T \underset{\sim}{\beta})/S\})^2} \qquad (3.30)$$

when the errors have distribution function F, and ψ' is the derivative of ψ.

An estimate of (3.30), unbiased in terms up to order $(p/n)^2$, is

$$\hat{V}_{\psi,F}(\hat{\beta}_j) = (n-p)^{-1}(k/\hat{\mu})^2 \sum_{i=1}^{n} \psi^2\{(y_i - \underset{\sim}{x}_i^T \underset{\sim}{\beta})/\hat{S}\} S^2((X^TX)^{-1})_{jj} \qquad (3.31)$$

where k is a correction factor (to be used when (p/n) is not too small). For Huber's ψ function,

$$k = 1 + p(1-\hat{\mu})/n\hat{\mu} \qquad \text{where} \qquad (3.32)$$

$$\hat{\mu} = n^{-1} \sum_{i=1}^{n} \psi'\{(y_i - \underset{\sim}{x}_i^T \underset{\sim}{\beta})/S\} \qquad (3.33)$$

Expression (3.33) gives the proportion of non-Winsorised residuals in the sample; it is an estimate of $E_F(\psi'\{(y_i - \underset{\sim}{x}_i^T \underset{\sim}{\beta})/S\})$. Expression (3.31) was employed in this study to obtain estimates of the standard errors of the Huber parameter estimates. Huber (1973 , p 814) has given two other plausible estimates of (3.30), each of which reduces to the standard formula in the least-squares case.

Other aspects of the distributional properties of M-estimators have been studied by Monte Carlo sampling; see, for example, Hill and Holland (1977) and Denby and Larson (1977). Caroll (1979) investigated the effects of asymmetry on estimates of variance of robust estimators. If the error distribution is highly skewed, the usual estimates of variance of M-estimators of location (given in Huber, 1964) were found to be too small. The same conclusion applied to estimates of the variance of the intercept term in linear

regression. The common variance estimate of the variance of the slope parameter in a simple regression problem is consistent even under asymmetry. Monte Carlo results testified that robust estimates of the slope parameter are negligibly influenced by asymmetry when n>>p. We mention these results on asymmetric bias because some of the error distributions encountered in the prediction of job turnaround time (in Chap. 4) are heavily skewed.

- Computational algorithms to solve the Huber M-estimation problem.

We now consider methods of solution of the minimisation problem (3.28).

Three principal algorithms have been proposed. They are Newton's method (see, e.g. Chambers, 1977): an algorithm known as the Huber (or H-) method, which is discussed by Huber (1977), Bickel (1975) and Kennedy and Gentle (1980); and an IWLS method generally attributed to Beaton and Tukey (1974). Each of these algorithms is discussed in Holland and Welsch (1977) and Huber (1977). Agee and Turner (1979) have described Newton's method and the IWLS technique, while Dutter (1977) discusses each in some detail.

Newton's method provides the most accurate (numerically exact) solutions for Huber's ρ function, for which it gives the largest adjustment on the first iteration. The salient difficulties with this method are that it requires the existence of $\psi'(t)$ and is not easily used in association with existing least-squares packages. The H-method has the computational advantage that the generalised inverse $(X^TX)^{-1}X^T$

need be computed only once. The IWLS and H-method are relatively simple to program.

Dutter (1977) has conducted a computational comparison for these three algorithms, and their variants, with Huber's ρ function. His results are based on the solution time and number of iterations required by each of the algorithms to solve a variety of test problems. A version of the H-method was found to be superior to the others: for example, it was nearly twice as fast (in terms of solution time) as a method based on the Fletcher-Powell minimisation algorithm.

The adopted approach is IWLS. The estimating equation (3.29) may be re-written as a weighted least-squares equation:

$$\sum_{i=1}^{n} w_i x_{ij}(y_i - \underset{\sim}{x}_i^T \underset{\sim}{\beta})/S = 0 \qquad j=1,\ldots, p.$$

where the weights w_i are sample-dependent. If the weight function is defined as $w(t) = \psi(t)/t$, the weights input to the $(m+1)^{st}$ iteration are:

$$w_i^{(m)} = \frac{\psi((y_i - \underset{\sim}{x}_i^T \hat{\underset{\sim}{\beta}}^{(m)})/\hat{S}^{(m)})}{((y_i - \underset{\sim}{x}_i^T \hat{\underset{\sim}{\beta}}^{(m)})/\hat{S}^{(m)})} \qquad i=1, \ldots, n \qquad (3.34)$$

where $\hat{\underset{\sim}{\beta}}^{(m)}$ is the robust parameter estimate after the m^{th} iteration and $\hat{S}^{(m)}$ is the robust estimate of residual scale after the m^{th} iteration (see expression 3.37 for the updating formula for $S^{(m)}$).

As the method of solution is iterative, a set of starting ing-values for the (p+1) unknown parameters $\hat{S}^{(o)}$, $\hat{\underset{\sim}{\beta}}^{(o)}$ is needed. The usual estimate of the variance about the fitted

least-squares line is an unsatisfactory estimate of $\hat{S}^{(o)}$, because it is sensitive to outliers. Thus a more robust starting estimate of residual scale is required. For our purposes, we took

$$\hat{S}^{(o)} = \underset{i}{\text{median}} \left(|y_i - \underset{\sim}{x}_i^T \hat{\underset{\sim}{\beta}}^{(o)}| \right) / \ 0.6745$$

where $\hat{\underset{\sim}{\beta}}^{(o)} = \hat{\underset{\sim}{\beta}}_{LS}$, the least-squares estimate.

Division by the factor 0.6745 (the median of the absolute values of a standard normal sample) makes the estimate of $\hat{S}^{(o)}$ approximately unbiased when the true error model is Gaussian. Other suggestions for robust estimates of scale can be found in Andrews et al (1972), Bickel (1976), Hinich and Talwar (1975), Hill and Holland (1977), Holland and Welsch (1977), Huber (1977) and Hogg (1977).

As we have indicated, we employed the least-squares values of the coefficient $\underset{\sim}{\beta}$ as preliminary estimator to obtain both Huber and LAD estimates. For convex ρ functions, this usually provides an adequate starting-point if the proportion of outliers in the sample is not excessive.

The choice of preliminary estimators is particularly important for non-convex ρ functions. If a poor starting-value is employed; the eventual solution may converge to a local minimum relatively remote from the global minimum. The LAD solution is to be preferred for non-convex ρ functions because it is less influenced by outliers than is the least-squares solution.

The LAD solution is also recommended by Holland and

Welsch (1977) for non-convex ρ functions. If scale is to be iterated, or the least-squares root used as a starting-point, these authors advocate the following procedure: iterate to convergence with the Huber ρ function, then iterate once, or at most twice, with the non-convex ρ function. Hinch and Talwar (1975), Ramsey (1977) and Birch (1980) have suggested different preliminary estimators. Despite the complications in the computational procedure using non-convex ρ functions, Andrews et al (1972) recommend their use under certain conditions because they are known to yield higher efficiencies for heavy-tailed distributions than do convex ρ functions.

For Huber's ρ function and other non-homogenious ρ functions, the solution to the estimating equation (3.29) is not scale-invariant. Scale-invariance can be induced by solving (3.29) simultaneously with

$$\sum_{i=1}^{n} \psi^2\{(y_i - \underset{\sim}{x}_i^T \underset{\sim}{\beta})/S\} = (n-p)A \qquad (3.35)$$

where the constant $A = E\{\psi^2(Z)\}$ (3.36)

and where Z is a standard normal deviate. It is inadvisable to iterate on $\hat{S}$, except for Huber's ρ function, for which a convergence theory exists (Huber, 1977).

The estimate of $\hat{S}$ on the $(m + 1)^{st}$ iteration is given by Huber as:

$$(\hat{S}^{(m+1)})^2 = ((n-p)A)^{-1} \sum_{i=1}^{n} \psi^2\{(y_i - \underset{\sim}{x}_i^T \hat{\underset{\sim}{\beta}}^{(m)})/S^{(m)}\}(S^{(m)})^2 \qquad (3.37)$$

The constant A in expression (3.35) is estimated in such a way that the estimate $\hat{S}^2$ is consistent for normal errors. For example, when $\psi(t) = t$, A = 1.0 and (3.35) gives the usual estimate of the error variance $\hat{S}^2 = \hat{\sigma}^2$. The value of A defined in (3.36) may be computed by considering the equation:

$$A = \int_{-c}^{c} z^2 \phi(z)\, dz + c^2 P(|Z| > c) \qquad (3.38)$$

where $\phi(z)$ is the standard normal density function. Expression (3.38) may be evaluated either by numerical integration, or (as used here) Monte Carlo sampling. The constant A depends on c, the trim constant, through the function ψ(e.g. for Huber's ρ function with c = 1.0, Huber (1973) gives A = 0.51606). The value of the trim constant c was set to 1.0 for all the regressions in this study. This low value was chosen primarily because the observed degree of contamination in some of the samples was seen to be quite high. Huber (1981 , p87) indicates that this value of c should be used when there is approximately 15% contamination in the normal error model (this result applies to the location model). Holland and Welsch (1977) have listed the values of c for eight ρ functions such that the asymptotic efficiency in the location model is 95% when the true error is Gaussian. For Huber's ρ function, they give c = 1.345.

The value of c effectively controls the amount of trimming in the sample. In most applications, it is usually chosen fairly arbitrarily, and is fixed throughout the iterative sequence. More efficient estimators may result if the value

c is adjusted after each iteration. Hogg (1974, 1979) has suggested that an estimate of the trim constant could be derived from the sample kurtosis of the residuals. Denby and Mallows (1977) have introduced two graphical techniques that exhibit the way in which robust regression coefficients and residuals vary with the size of the trim constant. This approach is called adaptive estimation. The underlying idea is that not only the trim constant, but the actual form of the $\rho(\cdot)$ could be selected after observing the sample. For example, Harter (1975) has proposed the following scheme to determine the estimation procedure: compute the kurtosis k from a preliminary regression (he suggests either ℓ_1 or ℓ_2). If $k > 3.8$, use LAD; if $k < 2.2$, use L_∞ (minimise the maximum deviation); otherwise use least-squares. Such an approach is only feasible in the case of large sample sizes, because a considerable amount of data is required to derive a satisfactory estimate of kurtosis. Wegman and Caroll (1977) have reported favourably on the use of adaptive estimation in the location case.

Once the iterative sequence has begun, the next question concerns the point at which it is to be terminated. One simple rule is to stop after a fixed, prespecified number of iterations. This has been used in Monte Carlo studies by Holland and Welsch (1977) and Wegman and Caroll (1977). The stopping-rule employed in this study, for both LAD and Huber estimates, depends on the deviance statistic (Baker and Nelder, 1978). First, define the norm after m iterations: for Huber's ρ, given by (3.27), this is $I^{(m)} = \sum_{i=1}^{n} \rho\{(y_i - \underset{\sim}{x}_i^T \hat{\underset{\sim}{\beta}}^{(m)})/S^{(o)}\}$;

in the LAD case, it is $I^{(m)}=\sum_{i=1}^{n}|y_i-\underset{\sim}{x}_i^T\hat{\underset{\sim}{\beta}}^{(m)}|$. The rule is, for $I^{(m)}\leq I^{(m-1)}$,

$$\text{stop when } (I^{(m-1)} - I^{(m)}) < \gamma I^{(m)} \qquad (3.39)$$

In this study, the tolerance level was set to $\gamma = 10^{-4}$. This criterion is closely related to the following rule:

$$\text{stop when } \max_j (|\hat{\beta}_j^{(m-1)}-\hat{\beta}_j^{(m)}|)<\tau s.\hat{e}.(\hat{\beta}_j^{(m)}) \quad j=1,..,\ p \qquad (3.40)$$

If the tolerance level $\gamma = \tau^2$, criterion (3.40) will give approximately the same degree of numerical accuracy in the final solution as criterion (3.39). This approximation is suggested because, in the least-squares case, (3.39) is a quadratic equvalent of (3.40). Criterion (3.39) is clearly more economical to use, since it involves the calculation of only one quantity after each iteration, whereas criterion (3.40) requires estimates of the standard errors of the coefficients after each iteration. Armstrong and Frome (1976), Ramsey (1977) and Birch (1980) have used a relative change stopping-rule, this is, for m>1,

$$\text{stop when } \max_j (|\hat{\beta}_j^{(m-1)}-\beta_j^{(m)}|)<\eta(|\hat{\beta}_j^{(m)}|) \qquad j=1,\ \ldots,\ p$$

where $\eta = 10^{-7}$ (Armstrong and Frome), $\eta = 10^{-3}$ (Ramsey) and $\eta=5 \times 10^{-4}$(Birch).

An alternative criterion, based on a scale-dependent measure of the gradient, has been proposed by Coleman et al (1980). While each of these criteria may be used in

association with the IWLS method, Huber (1977) has given a stopping-rule to be used with the H-method. Birch (1980) has studied the relationship between starting-values and stopping-rules in the IWLS context. The results of that author's Monte Carlo sampling experiment emphasised the importance of this relationship, not only for determining actual estimators, but also in cost terms when any large-scale Monte Carlo study involving iterative schemes is conducted.

As in the LAD case, we require a method by which a subset q(<p) variable Huber regression model may be fitted. The technique is analogous to that employed in the LAD case. The starting-values for the subset fit were supplied by the solution to the full p parameter Huber fit. Assume this has converged after k > 1 iterations, and denote it by

$$y_i = \underset{\sim}{x}_i^T \hat{\underset{\sim}{\beta}}^{(k)} + z_i^{(k)} \quad i = 1, \ldots, n.$$

Then the input weights to the first iteration of the subset fit are (from eqn. 3.34):

$$w_i^{(o)} = \psi(z_i^{(k)} / \hat{S}^{(o)}) / (z_i^{(k)} / \hat{S}^{(o)}) \quad i=1, \ldots, n$$

If $(\hat{\sigma}^{(k)})^2$ is the final estimate of the scale of the residuals $z_i^{(k)}$ (obtained from expression 3.37), then the initial estimate of the residual scale in the first iteration is simply $\hat{S}^{(o)} = \hat{\sigma}^{(k)}$. The fully-iterated estimates of the subset model were found by following the IWLS procedure described above.

Thus we required a method of comparing the significance

of any two Huber models, where one is nested inside the other. Recently, Huber (1981, sect. 7.10) has suggested one analysis-of-variance technique for M-estimators. However, we have employed a likelihood ratio test statistic for Huber's M-estimators developed by Schrader and Hettmansperger (1980).

- A likelihood ratio test for Huber's M-estimators.

The basis of their approach, which is analogous to that used in the least-squares case, is the extent of the reduction in the norm $\sum_{i=1}^{n}\rho\{(y_i - \underset{\sim}{x}_i^T\hat{\underset{\sim}{\beta}})/S\}$ when passing from the reduced to the full model. The linear model (3.1) is partitioned into $\underset{\sim}{Y} = X_1\underset{\sim}{\beta}_1 + X_2\underset{\sim}{\beta}_2 + \underset{\sim}{\varepsilon}$, as in expression (3.21); the interest is in the hypothesis $H_o: \underset{\sim}{\beta}_2 = 0$. Define Huber's ρ by (3.27) and consider the norm:

$$\sum_{i=1}^{n}\rho\{(y_i - \underset{\sim}{x}_i^T\hat{\underset{\sim}{\beta}})/S\} \qquad (3.41)$$

Define D(F) as the deviance under the full model (i.e. as the minimum of 3.41) and D(R) as the deviance under the reduced model (i.e. as the minimum of 3.41 when H_o is true). Schrader and Hettmansperger (1980) proposed a test based on the likelihood ratio:

$$\Lambda = \exp\{-D(R)\}/\exp\{-D(F)\}$$

This is the true likelihood ratio test (LRT) when the error model has a density of the form $\exp\{-\rho(t/\sigma)\}$. The main results states that, in distribution,

$-2\lambda^{-1} \log \Lambda = 2\lambda^{-1}\{D(R) - D(F)\} \rightarrow \chi^2_{p-q}$, where λ, the constant of proportionality, is given as

$$\lambda = \frac{E_F(\psi^2\{(y_i - \underset{\sim}{x}_i^T\underset{\sim}{\beta})/S\})}{E_F(\psi'\{(y_i - \underset{\sim}{x}_i^T\underset{\sim}{\beta})/S\})} \qquad (3.42)$$

Then a generalisation of the classical F-test for H_o may be based on the statistic:

$$F_M = \hat{2\lambda}^{-1} \{D(R) - D(F)\}/(p-q) \qquad (3.43)$$

where the estimate $\hat{\lambda}$ is given as

$$\hat{\lambda} = (\hat{\mu}(n-p))^{-1} \sum_{i=1}^{n} \psi^2\{(y_i - \underset{\sim}{x}_i^T\hat{\underset{\sim}{\beta}})/\hat{S}\} \qquad (3.44)$$

When H_o is true, F_M is shown to be distributed according to an F-distribution on (p-q) and (N-p) degrees of freedom. This result holds only if the three regularity conditions given by Huber (1973, p804) hold.

The performance of this LRT statistic was examined by a Monte Carlo sampling experiment. The specifications of the experiment were identical to those employed in the investigation of the conjectured distribution of the LRT statistic for LAD estimators. The full model had seven unknown parameters (including the constant term) and the reduced model had six (i.e. p-q = 1). There were 27 observations in each sample. In this case, each experiment consisted of 150 simulations.

Each model was fitted by the IWLS procedure described above with the trim constant c = 1.0. The starting-values

for each fit were supplied by the residuals from the corresponding LAD solution, which was found by implementing the Armstrong, Frome and Kung (1979) algorithm. The convergence criterion was the deviance statistic given by expression 3.39, with tolerance level $\gamma = 10^{-5}$. The quantities D(F) and D(R) were computed after each run. The errors were sampled from the same four distributions as in the LAD case. These were: normal, mean zero, variance one; Cauchy, location parameter zero, scale parameter one; negative exponential, mean parameter two; and a 'bootstrap' distribution, whose histogram is shown in figure 4.2(b). Population estimates of the constant of proportionality (given by 3.44) for the three theoretical distributions were obtained by using the constants set out in Huber (1973, Table 3). These are given below:

Error Model	$E_F\{\psi^2(\cdot)\}$	$E_F\{\psi'(\cdot)\}$	λ
Normal (0,1)	0.51606	0.68269	0.75592
Cauchy (0,1)	0.63262	0.5	1.26524
Negative Exponential (2)	0.39077	0.84141	0.46442

For the negative exponential distribution, the errors were first adjusted so that $E\{\psi(\cdot)\} = 0$. The estimate of the constant of proportionality, λ, for the bootstrap distribution, was taken over the entire sample; this gave $\hat{\lambda} = 0.73631$.

For each run, the statistic F_M given by expression (3.43) was calculated. To compare the observed and theoretical distribution of the LRT statistic, we constructed the cumulative frequency distribution of the observed test

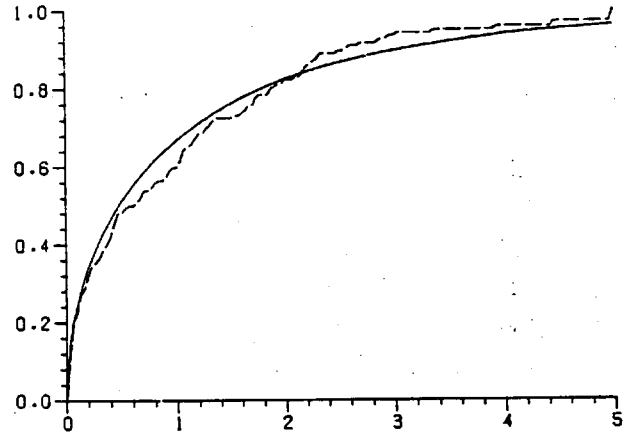
statistic and plotted it against its theoretical distribution, an F-distribution on 1 and 20 degrees of freedom. The distribution function of the F-distribution was computed by a FORTRAN program given by Kennedy and Gentle (1980, pp113-114).

The resulting plots are shown in figures 3.2(a) to 3.2(d). For the 'bootstrap' and normal error models (figures 3.2(a) and (b)) the fits are quite good. The discrepancy between the empirical and theoretical distributions for the Cauchy and negative exponential error models(figures 3.2(c) and (d)) is somewhat larger.

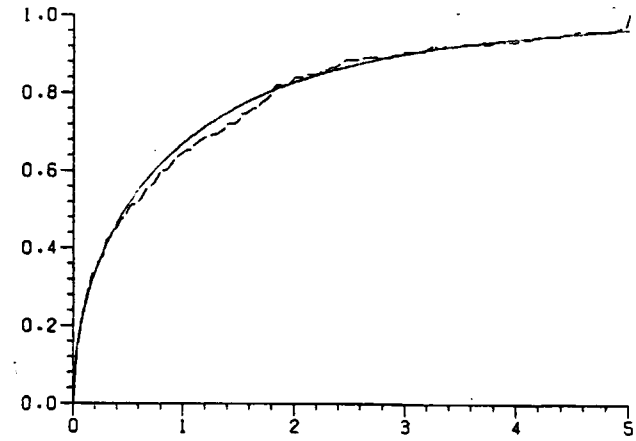
We noted that the overall shape of the cumulative frequency distribution of the observed LRT statistic for the Cauchy and negative exponential error models is correct. This suggested that the population constant of proportionality over-estimated the sample values in the Cauchy case, and under-estimated it in the negative exponential case. This hypothesis was tested by re-computing the LRT statistic F_M in the same way as above, except that the sample rather than the population estimate of λ was used. (This variation would be that used in practice). The resulting plots, which show a definite improvement, especially in the negative exponential case, are shown in figures 3.2(e) (for Cauchy errors) and 3.2(f) (for negative exponential errors).

The results of a power study of this test are reported in the next sub-section.

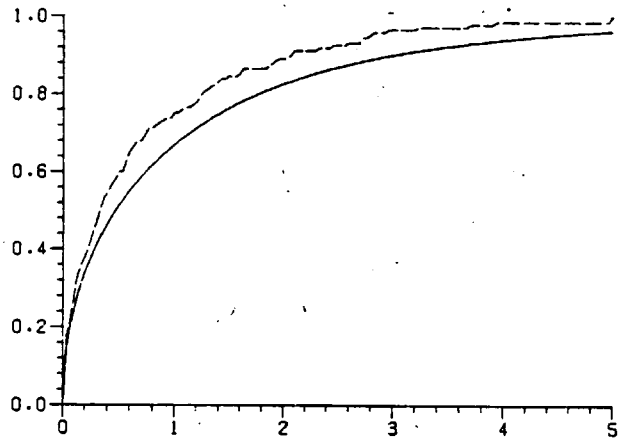
Figure 3.2 ; Cumulative frequency distribution of observed likelihood ratio test statistic for Huber s M-estimators.



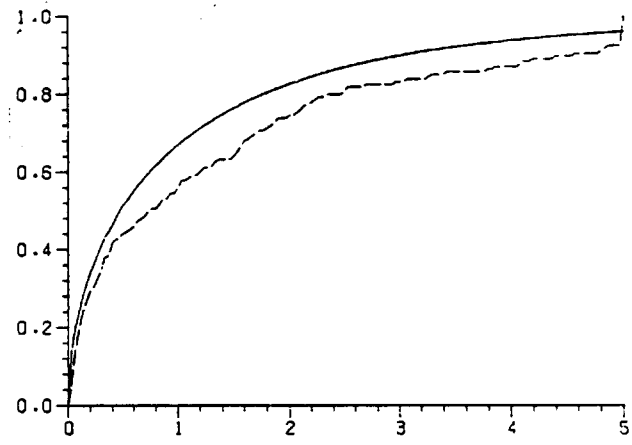
(a) errors sampled from a 'bootstrap' distribution.



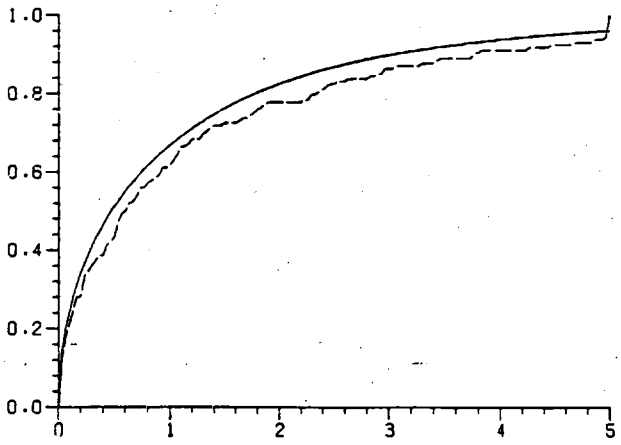
(b) errors sampled from a Normal(0,1) distribution ; population λ.



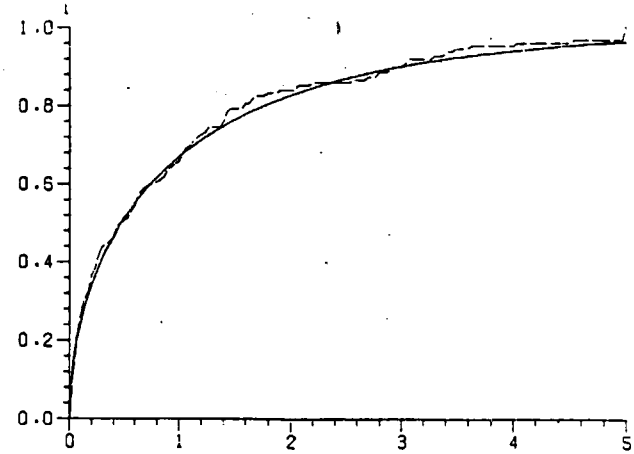
(c) errors sampled from a Cauchy(0,1) distribution ; population λ.



(d) errors sampled from a negative exponential(2) distribution ; population λ.



(e) errors sampled from a Cauchy(0,1) distribution ; sample λ.



(f) errors sampled from a negative exponential(2) distribution ; sample λ.

Legend

—— Theoretical F distribution on 1 and 20 degrees of freedom.

– – – Observed frequency distribution.

λ Constant of proportionality.

3.3 Monte Carlo Power Study

This section reports on the results of a Monte Carlo study of the power of two tests associated with the two robust estimators considered in sections 3.1 and 3.2 (i.e. LAD and Huber's M-estimators) relative to the power of the corresponding least-squares tests. The two tests are the likelihood ratio test ('F-test') and the 't-test'. The specifications of each experiment, that is, the design matrix, the linear regression model, the null and alternative hypotheses, and the estimation procedures, were identical to those employed in the two Monte Carlo experiments described in sections 3.1 and 3.2. The Armstrong, Frome and Kung (1979) LP algorithm was used to estimate the LAD parameters, while Huber's M-estimators were found by IWLS.

We know that, for normal errors, the least-squares 'F-test' (which we denote by the 'F_{LS}-test') has most power. But of particular interest is the behaviour of the power of the LAD and Huber's M-estimator likelihood ratio test relative to that of the 'F_{LS}-test' when the normal errors are contaminated by disturbances drawn from a second normal distribution with larger variance than the first.

Thus, we shall consider α-contaminated normal errors, in which successive points are drawn from a standard normal (i.e. N(0,1)) distribution with probability α, and from an $N(0, \sigma_o^2)$ distribution with probability $(1 - \alpha)$, where $0 \leq \alpha \leq 1$ and $\sigma_o^2 \geq 1$ is the variance of the contaminating normal distribution. The error variance is then $\alpha + (1-\alpha)\sigma_o^2$.

For the power studies of both tests we conducted two

experiments each consisting of 200 simulations. In the first, we sampled exclusively from an N(0, 1) distribution, thus the sampling probability α=1; in the second, we set α = 8/9 and σ_o^2 = 8, which gives an overall error variance of 16/9. Thus the standard normal errors are only mildly contaminated. To ensure that the results from the two experiments on each test were compatible, we standardised the error variance in the second experiment by dividing each sampled error by the population standard deviation (i.e. 4/3). We would expect to sample three out of 27 points from the N(0,8) error distribution in the second experiment, but since the α is a sampling probability and not a fixed sampling rate, not all the experimental samples are likely to draw exactly three points from it. In addition, note that the sequence of random variables used in the two experiments conducted on the two tests were not the same.

We report first on the results of the power study of the 'F-tests'. For each simulation, we computed three test statistics:

i) $F_{LS} = (\{SS(R) - SS(F)\}/(p-q))/\{SS(F)/(N-p)\}$
where SS(R) and SS(F) are, respectively, the residual sum of squares from the reduced and full models. As before, p-q = 1, and N = 27.

ii) The LAD likelihood ratio statistic L, defined as:
$L = 4\, f(o)\{V(R) - V(F)\}/(p-q)$
where f(o) is the density of the errors evaluated at the origin and V(R) - V(F) is the reduction in the ℓ_1-norm caused by passing from the reduced to

the full model.

iii) The likelihood ratio test for Huber's M-estimators defined as:
$F_M = 2\lambda^{-1}\{D(R) - D(F)\}/(p-q)$
where D(R) - D(F) is the reduction in the Huber-norm (defined by 3.28) when passing from the reduced to the full model and λ, the constant of proportionality, is given by:
$\lambda = E\{\psi^2(\cdot)\}/E\{\psi'(\cdot)\}$. Note that the Huber-norms D(R) and D(F) were computed using sample estimates of the robust scale parameter $\hat{s}$.

As in the previous Monte Carlo experiments, we tested $H_o: \beta_6 = 0$ against the alternative $H_1: \beta_6 \neq 0$, but allowed the true value of β_6 to vary over the interval 0.0(0.25)1.5. With the error variance of each parameter of the order of unity, there would be a high probability of rejecting H_o when the true value of β_6 was high.

With the significance level fixed at 5%, the empirical power curves for each of the tests were constructed by counting the number of times the null hypothesis was rejected for each of the true values of β_6, then by converting this figure to a probability in the obvious way. The observed F_{LS} and F_M statistics were compared to the relevant 5% points of an 'F'-distribution. An initial scan of the results for the power of the statistic L, which we had conjectured to have a χ^2 distribution, showed that the size of the test (when the errors were sampled from a standard normal distribution) was suspiciously large. To adjust the results, we estimated the

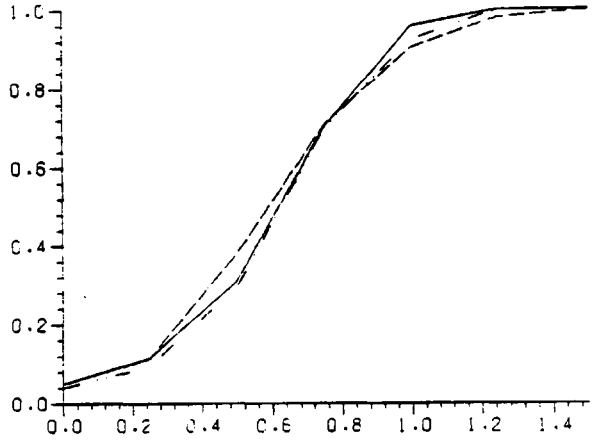
critical point by the 5% quantile of the distribution of 200 simulations with the true value of $\beta_6 = 0$. This value (4.33) was then used to replace the theoretical 5% point (3.84) for the LAD 'L-test'.

Where possible, we used population estimates in computing the test statistics. For the L statistic, we used the population value of f(o), and for the statistic F_M, we evaluated the shape parameters $E\{\psi^2(\cdot)\}$ and $E\{\psi'(\cdot)\}$ by Monte Carlo sampling. When the standard normal errors were contaminated by errors drawn from the N(0,8) distribution, these three constants were adjusted to account for the standardisation of the error variance. The two sets are values are as follows:

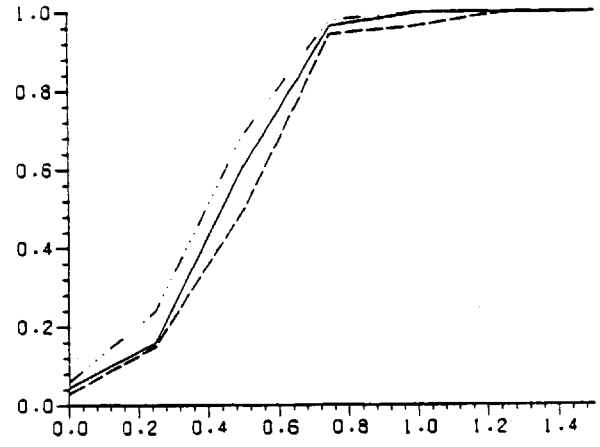
Error distribution	Specification	f(o)	$E\{\psi^2(\cdot)\}$	$E\{\psi'(\cdot)\}$
Standard normal	N(0,1)	0.39894	0.51606	0.68269
Contaminated normal	0.75N(0,1)w.p.8/9 0.75N(0,8)w.p.1/9	0.49372	0.43656	0.76740

The three empirical power curves under standard normal errors are shown in figure 3.3(a). The power curve of the 'F_M-test' closely follows that of the 'F_{LS}-test', while the power curve of the 'L-test' falls off after the true value of β_6 increases beyond 0.75. The 'F_{LS}-test' has most power at all points, except when the true value of $\beta_6 = 0.5$, where the 'L-test' has most power. In order to check whether the difference in power at this point was within the bounds of sampling variability, we constructed the 2 x 2 contingency table shown below:

Figure 3.3 : Empirical power curves of three likelihood ratio tests.



(a) Errors sampled from a standard normal N(0,1) distribution.



(b) Errors sampled from a contaminated normal
8/9N(0,1) + 1/9N(0,8) distribution.

Legend.

——— Least-squares 'F-test.'

— — — Likelihood ratio test for LAD-estimators.

— · — · — Likelihood ratio test for Huber's M-estimators.

x axis : true value of parameter .

y axis : (empirical) power.

True value of $\beta_6 = 0.5$

		'L-test'		
		H_o accepted	H_o rejected	row totals
'F_{LS}-test'	H_o accepted	108	30	138
	H_o rejected	16	46	62
column totals		124	76	200

This table shows that, for example, when the true value of β_6 was 0.5, there were 30 occasions (out of 200) on which the null hypothesis was rejected by the 'L-test' when H_o was accepted by the 'F_{LS}-test'at the 5% level. Because we have used the same random numbers within each experimental sample, we invoke a paired test of marginal equality (Bishop, Fienberg and Holland, 1975, p.258) in order to test whether the probability of obtaining a rejection under the 'L-test' when the 'F_{LS}-test' has accepted H_o is the same as the probability of the 'F_{LS}-test' rejecting when the 'L-test' has accepted H_o. The test was carried out using a χ^2 approximation with correction (McNamar's test). The observed test statistic was 3.67, which is not significant at the 5% level when compared to the critical point of a χ_1^2 distribution.

Turning to the power of the tests when the normals are contaminated, the results of which are shown in figure 3.3(b) a clear picture of the relative power of the three tests emerges. The 'F_M-test' has most power, with the 'F_{LS}-test' next, then the 'L-test'.

As might be expected, the overall power of each of the tests increases when the standard normal errors are contaminated

by errors drawn from a normal distribution with larger variance. The 'L-test' has the smallest increase in power, the 'F_M-test' the largest, with the 'F_{LS}-test' between the two. We would have anticipated that the 'F_M-test' would be more powerful than the 'F_{LS}-test' when the standard normal errors are contaminated, as is the case. The results for the uncontaminated case suggest that the 'F_M-test' is almost as powerful as the 'F_{LS}-test', which also conforms to our intuitive expectations.

However, the behaviour of the power of the 'L-test' requires some clarification. The need for an initial adjustment of the 5% point, and the relatively small increase in power of the test when passing from the uncontaminated to the contaminated case suggest that the χ^2 approximation is too gross, at least for small sample sizes.

The second experiment compared the power of the 't-tests' of the least-squares, LAD-and Huber's M-estimators. For each simulation, we computed the t-statistics $t(\beta_o)$, given by:

$$t(\beta_o) = |\hat{\beta}_6 - \beta_o|/\hat{s.e.}(\hat{\beta}_6)$$

where $\hat{s.e.}(\hat{\beta}_6)$ is the estimated standard error of $\hat{\beta}_6$. We then allowed β_o to vary over the range of values 0.0(0.25) 1.5 and computed the empirical power curves in the same manner as above.

Again we used population estimates where possible, but only when computing the final test statistics, and not during the iterative robust estimation procedures. Both the

asymptotic variances of the LAD-estimator (β_L) and Huber's M-estimator(β_M) can be related to the variance of the corresponding least-squares estimator (β_{LS}).

Using expressions (3.3) and (3.4) we have that (asymptotically)

$$\text{var}(\beta_L) = (2f(o)\sigma_F)^{-2} \quad \text{var}(\beta_{LS})$$

where σ_F^2 is the asymptotic variance of the sample mean with distribution function F. In practice, this was replaced by a sample estimate. We did use the population estimate of f(o) given in the text-table above. Because scale is assumed known in this case, the LAD 't-test' is distributed according to a standard normal and not a 't' distribution. Expression 3.30 shows that the (asymptotic) variance of an M-estimator can be written as:

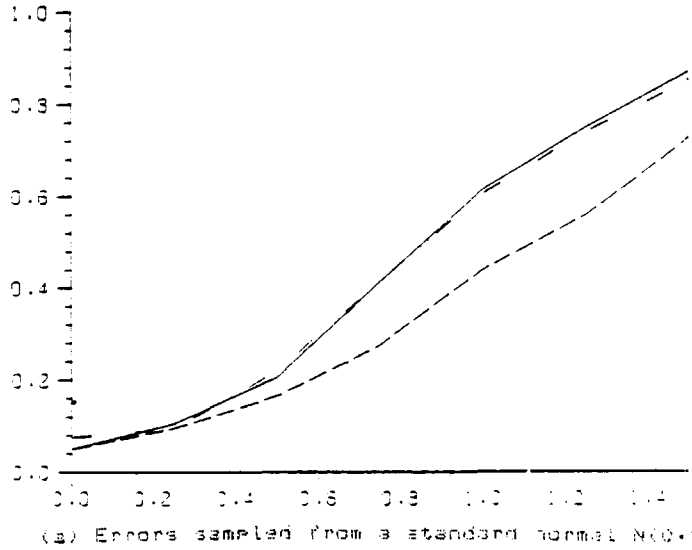
$$\text{var}(\beta_M) = \text{var}(\beta_{LS})\ E\{\psi^2(\cdot)\}/(\sigma_F\ E\{\psi'(\cdot)\})^2$$

In order to make the Huber 't-test' comparable to the 'F_M-test,', we did not employ population estimates of the constants $E\{\psi^2(\cdot)\}$ and $E\{\psi'(\cdot)\}$ in this case. Instead, they were computed using the sample estimate of robust scale parameter $\hat{s}$. The sample estimate of σ_F^2 was also employed. Thus, because the scale is assumed unknown for the Huber and least-squares 't-tests', both their observed t-statistics were compared to the 5% point of a 't-distribution' on (N-p =) 20 degrees of freedom.

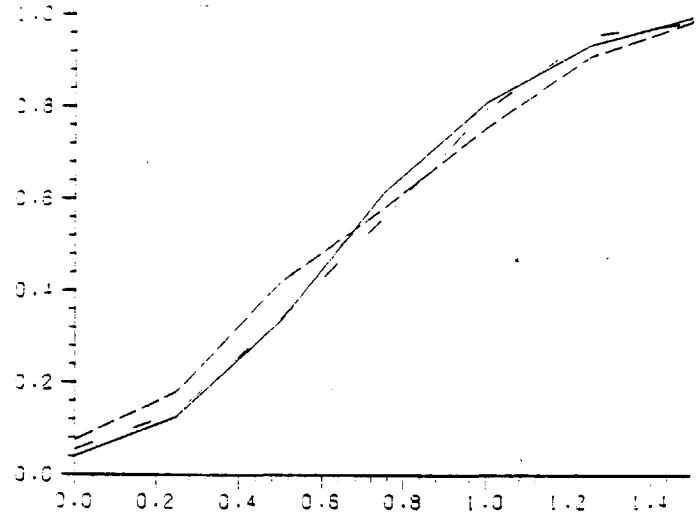
However, it was found that the size of the LAD 't-test' when the errors were sampled from a standard normal

distribution was too large. As for the 'L-test,' the results were adjusted by estimating the critical point of the 5% quantile of the distribution with $\beta_6 = 0$. This value was found to be 2.14, and was used to replace the 5% point of its theoretical distribution, a standard normal.

Figure 3.4 : Empirical power curves of three 't-tests.'



(a) Errors sampled from a standard normal N(0,1) distribution.



(b) Errors sampled from a contaminated normal 8/9N(0,1) + 1/9N(0,8) distribution.

Legend.

——— Least-squares 't-test.'

— — — 't-test' for LAD estimators (adjusted normal test).

— — · 't-test' for Huber's M-estimators.

x axis : true value of parameter .

y axis : (empirical) power.

The resulting plots of the power curves for the standard normal and contaminated normalerrors are shown in figures 3.4(a) and 3.4(b) respectively. Figure 3.4(a) shows that the power of the Huber 't-test' closely follows that of the least-squares 't-test', which is the most powerful of the three. Relative to the power of the other two tests, the LAD 't-test' is lacking in power. For the contaminated normal case, figure 3.4(b) shows that the LAD 't-test' as the most powerful until the true value of β_6 increases beyond the value 0.5. The power curve of the Huber 't-test' again closely follows that of the least-squares 't-test' and, apart from two cases when $\beta_6 = 0.75$ and 1.0, is marginally more powerful.

In conclusion, this power study has shown that, for standard normal errors, the 't' and 'F-tests' of M-estimators are almost as powerful as the corresponding least-squares tests. Moreover, the results for the contaminated normals have confirmed that the 't' and 'F-tests' of Huber's M-estimators are more robust than their least-squares equivalents.

It remains to consider the power of the likelihood ratio and 't-tests' of LAD-estimators. The results of the power study of the 'L-test' suggest that the χ^2 approximation for the distribution of the statistic requires some adjustment, at least for small sample sizes. The results of the 't-test' indicated that, under both contaminated and uncontaminated errors, the power of the LAD 't-test' is reasonably high, but only when the true value of the unknown parameter is not

large relative to its variance.

Finally, we consider the relative merits of the 't' and 'F-tests' of the two robust estimators. Comparison of figures 3.3(a) and 3.4(a), and 3.3(b) and 3.4(b) indicates that the empirical power curves of all three likelihood ratio tests have more power than their associated 't-tests' under both contaminated and uncontaminated normal errors. For standard normal errors, the power of the LAD 't-test' is considerably less than the power of the 'L-test', so much so that it would seem inadvisable to use the former in this case. The cause of this phenomenon is probably the inadequacy of the asymptotic formula linking the variance of an LAD-estimator to that of a least-squares estimator when the sample size is small. We are led to a similar conclusion for the 't-test' of Huber's M-estimators: the 'F_M-test' has greater power for standard normal errors than the corresponding 't-test'. Moreover, when the normal errors are contaminated, the likelihood ratio tests of the robust estimators are more powerful than the corresponding 't-tests'.

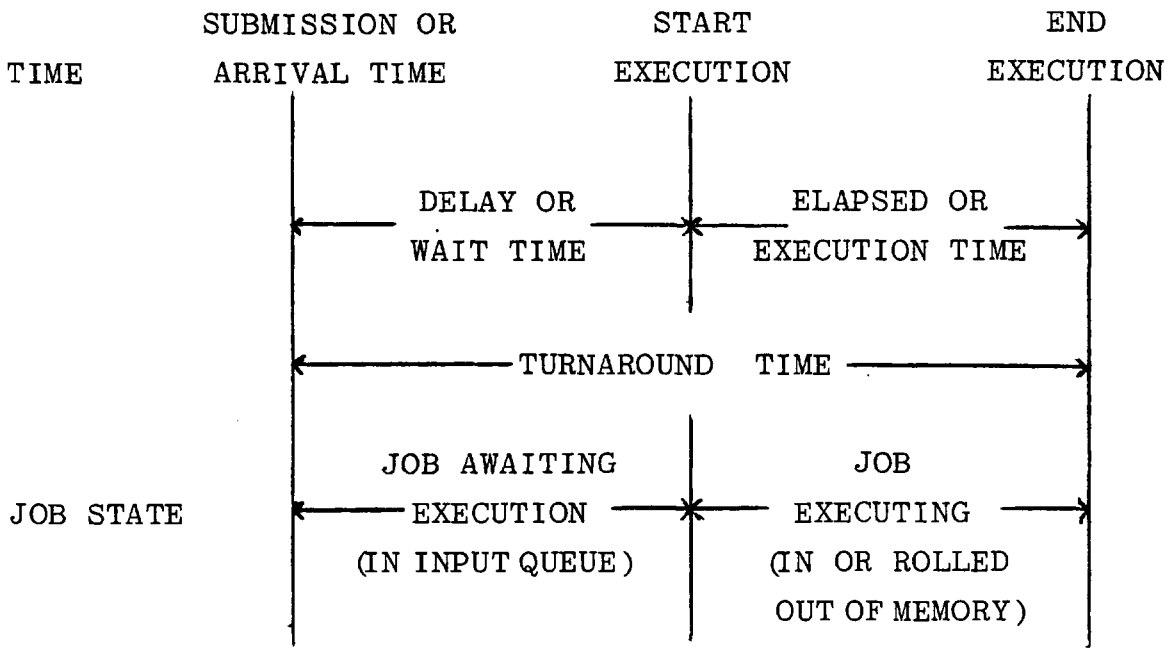
CHAPTER 4

PREDICTION OF JOB TURNAROUND TIME

4.1 Preliminaries

The main objective of this part of the study is to find a multiple regression model that will predict the turnaround time of a job from the information available at the time of its submission. Our criterion of success will be that the model accurately predicts the bulk of the sample. Note that the definition of this criterion contains within it the restriction that the chosen measure of predictive accuracy should not severely penalise any wayward predictions. One measure of goodness-of-fit that meets these two demands is the absolute prediction error, which is the one adopted in this study. The model should be realistic, parsimonious in the parameters, and contain explanatory variables that are easy to interpret. We describe below the two types of explanatory variables in the turnaround models. Once the general form of the model had been decided upon, we examined one data-set (from NUMAC) in order to formulate a general approach to other analyses. This is described in section 4.2.1. The results of the regression analyses are presented in section 4.2.2 (for NUMAC) and section 4.2.3 (for Rutherford).

First, we define batch job turnaround time: it is composed of two non-overlapping periods, job delay and job elapsed time.



The diagram above shows that, during its delay time, a job awaits execution in an input queue, and that during its elapsed time it is either in memory (i.e. executing), or temporarily rolled out of memory (e.g. while awaiting access to a peripheral device). Since the turnaround of a job is split into these two phases, it makes sense to examine the variables that affect the duration of each. To this end, we trace the path of a job through the system. The job control language of a newly-arrived job is checked for errors: if any are found, the job is aborted. Abort jobs place negligible demands on the resources of the system, and so were omitted from the analysis. The valid jobs are then classified by the resources they request. The NUMAC scheduler assigns jobs to one of 11 'priorities', the Rutherford scheduler, to one of 12'priorities'. We present the details of these classification schemes in due course. Competition

for the resources of the system is the main contributor to job delay time. The start of the execution may be delayed either because it has been assigned to the back of a non-empty queue, or because of interference from jobs of other priorities. Alternatively, if a job requests the use of a magnetic tape, it is assigned to a "set-up" queue. The job then experiences further delays while the operator finds and mounts the required tape(s). A magnetic tape job may be subjected to additional delays if there are no magnetic tape drives available. When a job begins execution, its processing time is mainly influenced by the amount of work it requires the system to undertake, that is, its resource consumption.

In summary, contributions to job turnaround time are made by the resources it requests and consumes, the length of the input queue to which it is assigned, and the length of all other input queues. The mix of jobs of different priorities resident in the system at any one time we refer to as the degree of system congestion. As our stated aim is to predict the turnaround time of a job from the information available at the time of its submission, we are restricted to using the degree the system congestion at job arrival time. The resources consumed by the job are not available to us. We may employ only the resources requested by the job as explanatory variables: these are its cpu time, maximum core partition size, and its magnetic tape requirements, if any. If no resources are requested on the job card, we have used the default cpu and core limits assumed

by the job scheduler.

We define the first two explanatory variables in the turnaround model by the mnemonics: REQCPU, to denote the cpu time (in seconds) requested by the job, and REQCORE, to denote the maximum core partition size (in kbytes) requested by the job.

On submission, a job is assigned to a "set-up" queue if it requires the use of magnetic tapes; otherwise, it joins one of the "priority" input queues. Because magnetic tape jobs are subject to extraordinary delays caused by operator intervention, we intend to treat them as a separate "priority". This gives a total of 12 possible "priorities" for NUMAC jobs, and 13 for Rutherford jobs.

At this stage, we could formulate a turnaround model with 14 explanatory variables for NUMAC, and 15 for Rutherford, where the explanatory variables are the REQCPU and REQCORE of the job, and one variable for each of the input queue lengths of jobs awaiting execution at the time of the submission of the job. If a separate model were formulated for each of the job priorities, we would have 12 regression equations for NUMAC jobs, and 13 for Rutherford.

Two considerations demand that we reduce the number of priority classes, and hence the number of regression equations and explanatory variables: first, the principle of parsimony, and second, the fact that some priority classes, (typically those that contain large resource request jobs) have only a few entries over the one week period of

observation. This reduction was achieved by merging some of the job priority classes. We present the details of the merging procedure for the NUMAC and Rutherford systems in the relevant sections below. This expedient allowed us to reduce the number of distinct classes of jobs from 12 to 5 for NUMAC, and from 13 to 4 for Rutherford. Thus, the final model formulation for NUMAC contained seven explanatory variables, and six for Rutherford, (excluding the constant term).

For both installations, we have included in the analysis only jobs submitted during the 11 prime-shift hours 08.30 to 19.30 over the five day working week. Two advantages accrue from this restriction: first, the day-time workload mix is not affected by the processing of system housekeeping jobs, which tend to be run in batches at the same time every night, or the same night every week; second, the night-time and week-end turnaround data have statistical characteristics that differ widely from those of our chosen sample.

We have made a further simplification concerning the measurement of the input queue lengths: rather than compute them for the exact time of submission for each job (our measurements are accurate to the nearest second), we have approximated the job arrival time by the minute interval in which it fell. This may be envisaged as taking "snap-shots" of the number of jobs in each input queue at one minute intervals. For each class, we computed the input queue lengths for each of the 5 x 11 x 60 = 3300 one-minute intervals over the observation period. This simplification

implicitly assumes that the change in the input queue lengths of jobs in each designated class is negligible over one-minute intervals. At a later stage, we offer empirical justification for this expediency, which was principally aimed at reducing the amount of computing time required to generate the queue lengths.

4.2 Regression Results

In this section, we present the results of the four regression models fitted to each job class turnaround model for both installations. We begin (in section 4.2.1) with an overview of the analyses of one week's data-set. The purpose of this analysis was to formulate a general approach for analyses of additional data-sets. In addition, we examine this data-set for the presence of multicollinearity, and assess the extent of its effects on the regression results. The results of the analyses of three NUMAC data-sets and one Rutherford data-set are given in sections 4.2.2 and 4.2.3 respectively. These results pertain to the subset, not the full, regression fits. The variable-selection techniques employed to arrive at the "best" subset models for the NUMAC and Rutherford data-sets are described in these two sub-sections. We also discuss the technique employed to amalgamate certain job "priorities".

4.2.1 Overview

We refer to the three weeks of NUMAC data which have been analysed by the acronyms N79W1, N79W2 and N79W3. This sub-section reports on the results of the full analysis of the N79W1 data-set.

First, a linear model structure was postulated. A priori, this was thought to be the best representation of the relationship between job turnaround time and the explanatory variables. Each of the five regression models was then fitted by least-squares, using all seven explanatory variables plus a constant term. A cursory examination of the model residuals indicated that there were severe departures from the normality assumption. The observed error distribution from each of the fits (except for class 4) were highly skewed. As a result, the overall accuracy of the predictions (in terms of the mean absolute prediction error in relation to the mean turnaround time) was unsatisfactory. For example, table 4.4 below shows that the mean absolute prediction of the full least-squares fit to the complete class 1 sample (i.e. 27 mins) is over 50% more than the mean turnaround time for class 1 jobs (i.e. 18 mins). Clearly, we would want to be able to improve on forecasts at this level of accuracy.

As an attempt to overcome the failure of the least-squares model approach, the turnaround times were transformed by taking logarithms (to base 10) and the multiplicative model re-fitted by least-squares. The logarithmic trans-

formation was invoked to produce nearly normal residuals as well as to stablise the error variance. A subsequent inspection of the fit indicated that these aims had been partially fulfilled. In addition, there was a marked improvement in the accuracy of the predictions of jobs in classes 1, 2 and 3. The fits to jobs in classes 4 and 5 generated forecasts that were, on average, less accurate than the corresponding linear model predictions. For example, table 4.4 shows that the multiplicative fit to the class 4 model had an average prediction error of 228 minutes, while the corresponding figure for the linear model least-squares fit was 169 minutes. It was found that the multiplicative fits to classes 4 and 5 produced some extreme overestimates. This feature highlighted the conceptual difficulty involved in specifying a multiplicative model structure. It presupposes that the turnaround time of a job is exponentially related to its resource requirements and to the vector of input queue lengths. As this assumption was thought to be intuitively unrealistic, and coupled with the empirical evidence of the breakdown of the multiplicative model in certain circumstances, we attempted to overcome the inadequacy of the linear least-squares model by other means.

Retaining the linear model structure, we decided to employ the two robust estimation techniques discussed above in Chapter 3, that is, Least Absolute Deviations (LAD) and Huber's M-estimation.

Thus, each of the five classes of jobs was fitted in four different ways in order to compare the eventual results.

To recap, these four methods were as follows: the linear model was fitted by least-squares, the Huber and the LAD estimation procedures, and the multiplicative model by least-squares.

The residuals from the least-squares linear model fit to the turnaround times of class 1 jobs were found to contain some unusually large values associated with abnormally large turnaround times (see, e.g., fig 4.1(a) , a plot of the least-squares residuals against their respective fitted values). A further analysis was carried out on a reduced sample (RS) of class 1 jobs from which all jobs whose turn-around times exceeded one hour were omitted. The purpose of this analysis was to see whether the robust procedure could still improve on the least-squares fit. The results in table 4.4 confirm that it does make an improvement.

All the data-sets were also examined for multicollinearity. On the basis of one of the main results of the principal components analysis in section 2, which indicated that the general submission level of jobs to all classes increased or decreased in unison, we expected to find some evidence of collinearity between the queue length variables. The method we have employed to detect multicollinearity is as follows; construct the sample correlation matrix of the explanatory variables, then subject it to a principal components analysis (see, e.g. Seber, 1976, p81). The coefficients of the eigenvectors associated with any near-zero eigenvalues are then examined. Any variables associated with large positive or negative coefficients

are nearly collinear.

For each of the five N7QW1 data-sets, there was a clear cut-off point beyond which the eigenvalues were considerably smaller than the leading group. For example, in the class 4 data-set, the percentage contributions of the seven eigenvalues were 2.1, 2.4, 10.1, 11.9, 16.1, 16.6 and 40.8. A similar pattern was revealed for the other data-sets. For the same data-set as above, the coefficients in the eigenvector corresponding to the smallest eigenvalue were (0.02, -0.02, -0.06, -0.73, 0.67, 0.13, -0.08) and the second smallest, (-0.04, 0.01, 0.11, 0.01, -0.20, 0.74, -0.63), where the first two coefficients in these eigenvectors refer to the REQCPU and REQCORE variables, while the remaining five refer to the queue length variables QLC1 - QLC5. Thus the pattern of the coefficients in these eigenvectors, and in those of the other data-sets, indicated that there were near collinear relationships between the QLC2 and QLC3 variables, and between the QLC4 and QLC5 variables.

Consequently, we would also expect to find further evidence of collinearity between these two pairs of variables manifested in the behaviour of the regression coefficient of one of the pair when the other is omitted from the list of regressors. This is the well-known indicator of collinearity, where we look for a large change in the magnitude of one variable relative to its standard error. For example, in the (least-squares) fit to the Class 4 linear model, the coefficient of the QLC4 variable (where the other variables were REQCPU, REQCORE and QLC1) was -0.7483 with standard

error 0.20, but with the inclusion of the QLC5 variable into the list of regressors, it changed to 1.137, with standard error 0.33. Here the magnitude of the change is rather large, indicating the presence of collinearity between the two variables. However, because, in this case, the QLC5 coefficient is highly significant, its addition to the list of explanatory variables reduced the average absolute prediction error from 177 minutes to 169 minutes.

The net effect of these collinear relationships on the variables included in the best subset fits is mixed. Referring to table 4.3 above, we see that the QLC2 and QLC3 variables are included in the final list of regressors only three times out of a possible twenty; these are the fits to the class 3 data-set, where the QLC3 variable is automatically included. Thus these are the only three cases when the inclusion of both of the collinear variables resulted in an improvement in the fit. On the other hand, the QLC4 and QLC5 variables appear together in the final list of regressors in all the four fits to Class 4 jobs, and three of the four fits to class 5 jobs. The explanation of this phenomenon is as before: both variables are included because the small amount of information added by the inclusion of the second of the pair is offset by the improvement in the overall fit when both are included.

4.2.2 NUMAC

The NUMAC scheduler allocates "priorities" to batch jobs solely according to the amount of cpu time requested by the job. This assignment of job priorities is as follows:

CPU TIME requested (in 370/168 seconds)	Priority
0-2	14
2-4	11
4-8	10
8-16	7
16-32	6
32-64	5
64-128	4
128-156	3
256-512	2
512-1024	1
>1024	0

Priority 14 (or P14) jobs are called batch monitor jobs and receive "instant" turnaround, that is, their delay time is generally only dependent on the length of the input queue of other P14 jobs. Once the priority has been allocated, the job is assigned to the back of the queue of jobs with the same priority.

The purpose of the preliminary study of the data was to reduce the number of priority-classes. We first assumed that the priority scheme discrminates between jobs making markedly different demands on the resources of the system. We anticipated that the turnaround times of jobs with different priorities would exhibit markedly varying character-

istics. Hence, we sought to group together those priorities of jobs the turnaround times of which had similar statistical features. This ad hoc approach consisted of an examination of the summary statistics of the N79W1 data shown in Table 4.1 below:

Table 4.1 : Summary statistics of NUMAC Job priorities ; data-set N79W1.

Priority	No. of Jobs	Turnaround time (in minutes)	
		mean	std. dev.
14	750	19.8	77.9
11	125	16.0	53.4
10	1138	66.6	117.6
7	350	77.2	118.9
6	530	83.5	121.8
5	247	160.6	170.3
4	160	216.3	252.7
3	204	391.3	186.0
2	125	559.0	235.2
1	79	613.0	244.9
0	0		
Magnetic tape Jobs	482	262.7	259.5

On the basis of our criterion for merging priorities together, it was decided to group jobs assigned to priorities 11 and 14, as well as those assigned to priorities 6 and 7. Thus, we propose to divide NUMAC jobs in the following way:

Table 4.2 : Definition of NUMAC Job classes.

Class	Amalgamated Priorities	Definitive Characteristic
1	11,14	Jobs requesting 0-4 CPU seconds.
2	10	Jobs requesting 4-8 CPU seconds.
3	6,7	Jobs requesting 8-32 CPU seconds.
4	0-5	Jobs requesting more than 32 CPU seconds.
5		Jobs requesting the use of one or more magnetic tapes.

Henceforth, we refer to the amalgamated "priorities" as job classes. We have already proposed two explanatory variables for each turnaround model, (the REQCORE and REQCPU variables); to these, we add five system congestion variables. Unlike the resource request variables (which are unique to each job), these will be implicit functions of time. For example, we refer to the queue length of class 1 jobs awaiting execution at the submission time of a job by the acronym QLC1, with analogous acronyms for the other four queue length variables (i.e. QLC2, QLC3, QLC4, QLC5). Thus these queue length variables are not a property of the job, but are determined solely by the state of the system encountered by the job on its arrival. We have explained above the method of taking "snap-shots" of the system, by which the five queue lengths are numerically determined.

For the N79W1 data-set, table 4.1 shows that the second class has the largest number of entries with 1138 jobs. Spread over 3300 intervals, this results in an average submission rate of about one job every three minutes, which suggests that we are empirically justified in assuming that the system is stable over one-minute intervals.

We next consider the problem of determining those variables to be included in the subset regression fits. The full regression fits were examined with a view to compiling, for each class, a set of variables with a physical basis that were all (or nearly all) significant in each of the four different models. For example, the REQCPU, REQCORE and QLC1 variables were found to exert a strong influence on the turnaround times of jobs from all the classes,

irrespective of the estimation procedure employed. A close relationship was also found to exist between the turnaround time of a job and the length of the queue to which it was assigned on arrival. These two results suggested that a first approximation to the best subset of variables would be given by the following group of variables.

Job Class	Physical basis of variables
1	REQCPU, REQCORE, QLC1
2	" , " , QLC1, QLC2
3	" , " , " , QLC3
4	" , " , " , QLC4
5	" , " , " , QLC5

Each turnaround series was fitted by the explanatory variables indicated by the physical basis. For each of the four models, a sequence of likelihood ratio tests was carried out, testing one variable at a time. If any of the tests were significant at the 5% level the variable with the highest significance level was added to the physical basis. The process was repeated until the best subset regression fit was found. The likelihood ratio test, in the least-squares context, is the usual F-test of variance. For the Huber and LAD procedures, the usual likelihood ratio test is inappropriate. The form of these test statistics has been discussed above in sections 3.1 and 3.2.

As a consequence of the selection strategy, the lists of explanatory variables in the four models fitted to each class series are not always identical, nor are they always significant (the results of the subset fits to the N7QW1 data-

Table 4.3 ; Results of subset regression fits ; data-set N79W1.

Data-set Class	Model/Estimation Procedure	EXPLANATORY VARIABLE Constant	REQCPU	REQCORE	QLC1	QLC2	QLC3	QLC4	QLC5
N79W1 Class 1-FS	Linear/Least-Squares	74.55*	4.088	-0.3879*	1.862*				-0.7439*
		(11)	(3.3)	(0.05)	(0.54)				(0.18)
	Linear/Huber	-1.485*	0.6452*	0.0019	0.7068*	0.0351*			
		(0.70)	(0.21)	(0.003)	(0.03)	(0.02)			
	Linear/LAD	-2.863*	0.7022*	0.0066	0.6236*	0.0448*			
		(0.88)	(0.26)	(0.004)	(0.04)	(0.02)			
	Multiplicative/ Least-Squares	0.4128*	0.0925*	-0.0011*	0.0469*				-0.0028*
		(0.08)	(0.02)	(0.0003)	(0.004)				(0.001)
N79W1 Class 1-RS	Linear/Least-Squares	0.2060	0.5579	0.0066	0.7526*				
		(1.3)	(0.39)	(0.01)	(0.07)				
	Linear/Huber	-3.050*	0.5323*	0.0138*	0.5909*	0.0402*			
		(0.60)	(0.17)	(0.003)	(0.03)	(0.02)			
	Linear/LAD	-3.565*	0.6718*	0.0123*	0.5525*	0.0457*			
		(0.60)	(0.17)	(0.003)	(0.03)	(0.02)			
	Multiplicative/ Least-Squares	-0.0668	0.0695*	0.0013*	0.0413*	0.0029*			
		(0.07)	(0.02)	(0.0003)	(0.003)	(0.001)			
N79W1 Class 2	Linear/Least-Squares	-21.73	4.089	0.2264*	3.888*	1.504*			-1.284*
		(42)	(5.7)	(0.07)	(0.97)	(0.25)			(0.29)
	Linear/Huber	-7.667	1.484	0.0595*	2.088*	0.6743*			-0.3264*
		(13)	(1.8)	(0.02)	(0.31)	(0.13)			(0.14)
	Linear/LAD	-24.52*	2.150*	0.0498*	2.220*	0.5149*			
		(10)	(0.2)	(0.02)	(0.25)	(0.10)			
	Multiplicative/ Least-Squares	0.3240	0.0354	0.0020*	0.0396*	0.0148*			-0.0075*
		(0.26)	(0.04)	(0.004)	(0.006)	(0.002)			(0.002)
N79W1 Class 3	Linear/Least-Squares	-93.94*	1.319*	0.6478*	3.639*		2.260*		-1.704*
		(19)	(0.53)	(0.07)	(0.96)		(0.28)		(0.35)
	Linear/Huber	-46.44*	0.8892*	0.2354*	2.6459*	0.7996*	0.1369		
		(9.5)	(0.25)	(0.03)	(0.47)	(0.22)	(0.24)		
	Linear/LAD	-43.02*	0.9151*	0.1484*	2.589*	0.5490*	0.5989*		
		(9.4)	(0.21)	(0.03)	(0.38)	(0.18)	(0.20)		
	Multiplicative/ Least-Squares	0.2488*	0.0105*	0.0031*	0.0286*	0.0068*	0.0062*		
		(0.11)	(0.003)	(0.0004)	(0.006)	(0.003)	(0.003)		

Table 4.3 (continued) ; Results of subset regression fits ; data-set N79W1.

Data-set Class	Model/Estimation Procedure	EXPLANATORY VARIABLE							
		Constant	REQCPU	REQCORE	QLC1	QLC2	QLC3	QLC4	QLC5
N79W1 Class 4	Linear/Least-Squares	152.9*	0.7377*	0.6069*	4.831*			1.137*	-7.194*
		(30)	(0.04)	(0.11)	(2.2)			(0.33)	(1.0)
	Linear/Huber	102.2*	0.8192*	0.6654*	4.006			1.302*	-7.073*
		(33)	(0.04)	(0.11)	(2.3)			(0.38)	(1.0)
	Linear/LAD	166.1*	0.7883*	0.6794*	-0.1619	-1.006*		1.447*	-8.422*
		(22)	(0.03)	(0.08)	(1.7)	(0.65)		(0.28)	(0.76)
	Multiplicative/ Least-Squares	1.604*	0.0015*	0.0014*	0.0105*		0.0054*	0.0021*	-0.0090*
		(0.08)	(0.0001)	(0.0003)	(0.0054)		(0.0014)	(0.0009)	(0.0025)
N79W1 Class 5	Linear/Least-Squares	-26.70	0.8773*	1.130*	2.839		1.884*		-1.872
		(38)	(0.09)	(0.18)	(2.8)		(0.78)		(0.88)
	Linear/Huber	-57.27*	0.8845*	0.8159*	5.340*			0.5714*	-0.3968
		(24)	(0.06)	(0.11)	(1.8)			(0.29)	(0.72)
	Linear/LAD	-42.29*	0.9622*	0.5359*	5.810*			0.5548*	0.0576
		(20)	(0.05)	(0.10)	(1.6)			(0.25)	(0.61)
	Multiplicative/ Least-Squares	1.355*	0.0014*	0.0025*	0.0127*				0.0054*
		(0.077)	(0.0002)	(0.0004)	(0.006)				(0.002)

Legend.

* denotes significance at the 5% level.

The parameter estimate is the upper figure in each cell ; below it, in parentheses, is its estimated standard error.

Class 1-FS ; Full Sample Class 1 Jobs.

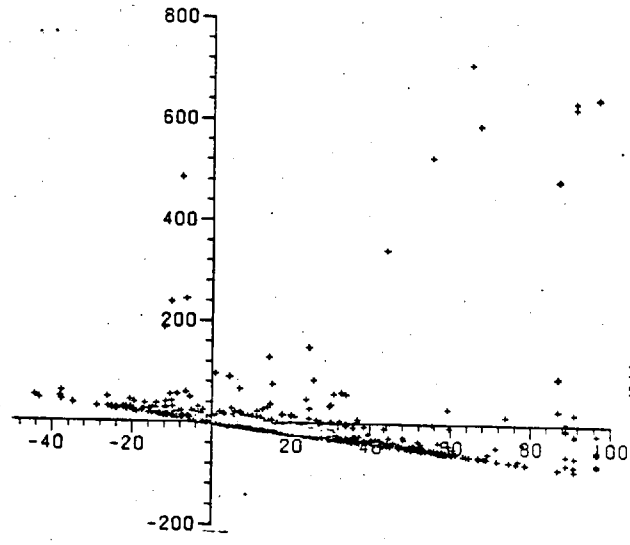
Class 1-RS ; Reduced Sample Class 1 Jobs.

set are given in table 4.3, those of the subset fits to the N79W2 and N79W3 data-sets appear in table 4.1A, which is appended to this chapter). Most of the coefficients of the explanatory variables are positive, indicating that the turn-around time of a job increases as its resource demands increase, and as the system becomes more congested. Since this accords with what we would have expected intuitively, the result is a corroboration of the general form of the model. There are several exceptions, notably the coefficients of the QLC4 and QLC5 variables, which tend to have negative coefficients when they are included in a model. The overall pattern of variables included in each model suggests that the turnaround times of jobs in classes 1, 2, and 3 are directly proportional to the queue lengths of jobs in the same classes, while the turnaround times of jobs in classes 4 and 5 depend only on the number of other class 4 and 5 jobs awaiting execution at the time of their submission. Another feature revealed in this table is the considerable importance of the REQCORE and REQCPU variables in affecting the turn-around time of jobs from all classes.

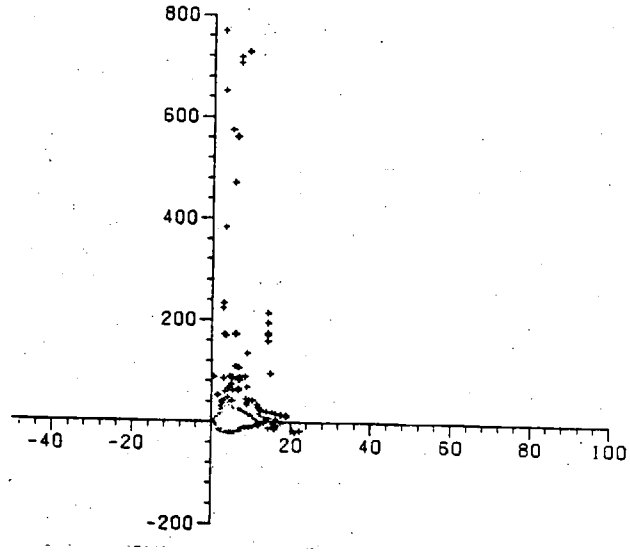
We turn now to a graphical examination of the residuals arising from the different fits. We show a number of figures relating to the different class series from the N79W1 data-set, the data-set which was discussed in section 4.1 above. In this case, all the plots have been constructed from the subset regression fits listed in Table 4.3 above.

Figure 4.1(a) is a plot of the least-squares residuals against their respective fitted values from the full sample

Figure 4.1 : Residuals vs. fitted values from linear model fits to class 1 (full sample) jobs ; data-set N79W1.



(a) Least-squares fit.

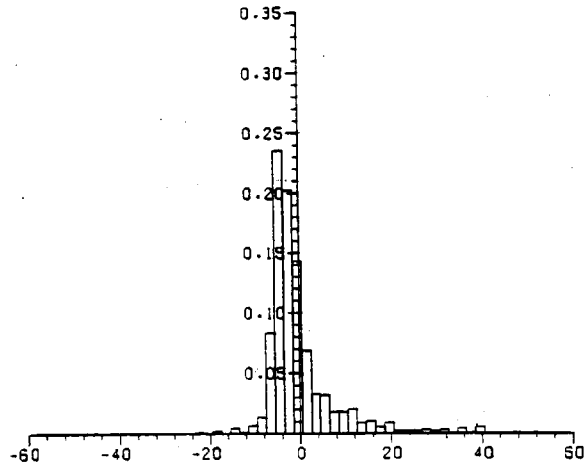


(b) Huber fit.

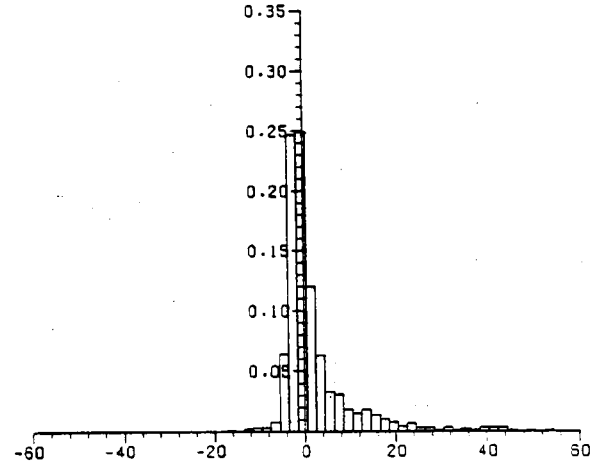
Legend.

x and y axes measured in minutes.

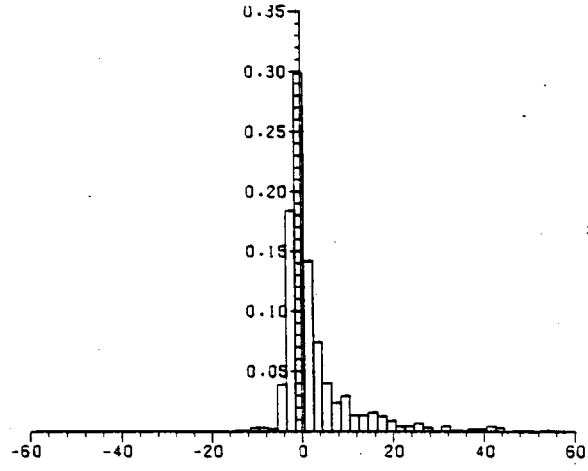
Figure 4.2 : Histograms of residuals from fits to class 1 (reduced sample) jobs ; data-set N79W1.



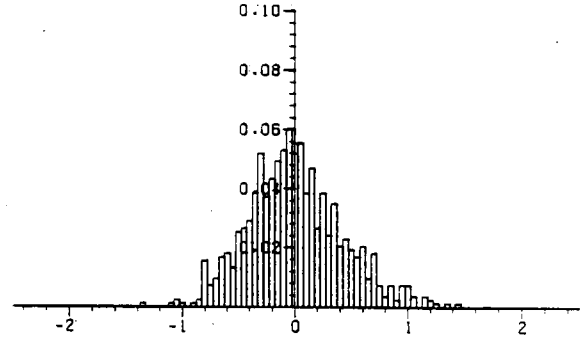
(a) Least-squares linear model fit.



(b) Huber linear model fit.



(c) LAD linear model fit.



(d) Least-squares multiplicative model fit.

Legend.

x axis measured in minutes for linear model fits.

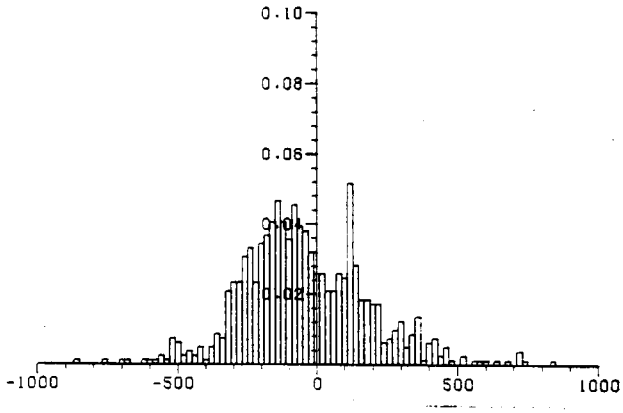
x axis measured in log(base 10) minutes for multiplicative model fit.

(FS) class 1 linear model. There is a small proportion of extremely large outliers, all severely underpredicted by the model. The least-squares estimation procedure will have been decisively influenced by these outlying points. The apparent linear relationship between the residuals and their fitted values is caused by the fact that the Y-observations are constrained to be positive. The model also produces some negative predictions. Figure 4.1(b) is the analogous plot for the Huber fit to the same data-set. The robust estimation procedure, less sensitive to large disturbances than least-squares, produces a more meaningful set of predictions. The dispersion of the fitted values has been reduced.

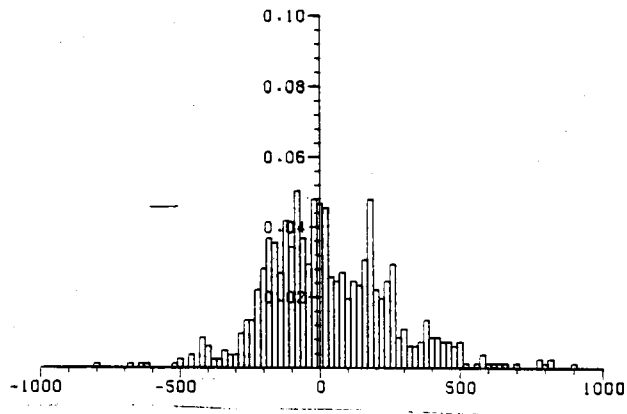
Figures 4.2(a) to 4.2(d) show the histogram of the residuals from the fits to the class 1 reduced sample (RS), that is, the class 1 sample with all jobs whose turnaround times exceed one hour omitted. Figure 4.2(a) shows that the least-squares residuals from the linear model fit to this sample have a distribution that is skewed to the left of the origin. The corresponding histogram for the Huber residuals (fig 4.2(b)) still has a long right-hand tail, but the peak has been shifted toward the origin. The histogram of the LAD residuals (fig 4.2(c)) has its peak centered at the origin. This is a natural outcome of moving from a least-squares, to a Huber, to an LAD fit (i.e. from a mean- to a median-based regression) if the distribution of the least-squares residuals is skew .

Figure 4.2(d) is a histogram of the least-squares

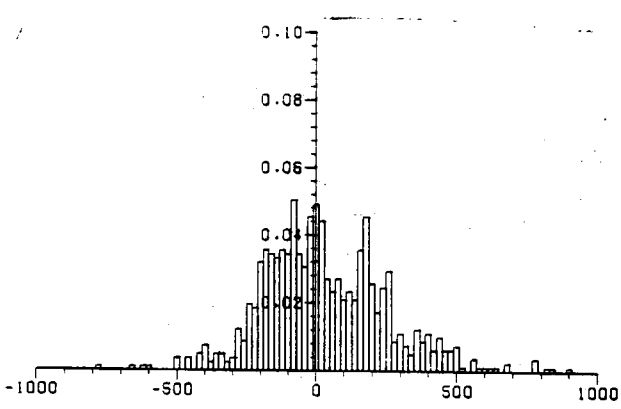
Figure 4.3 : Histograms of residuals from fits to class 4 jobs ; data-set N79W1.



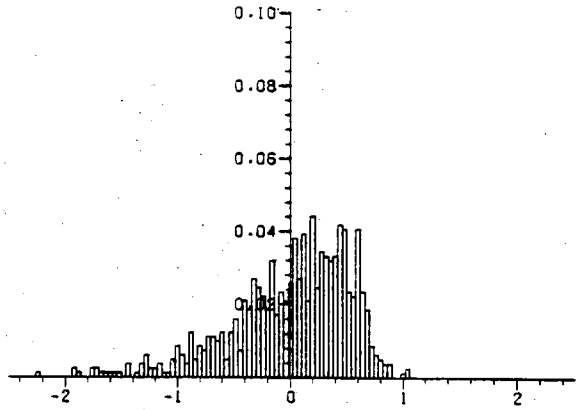
(a) Least-squares linear model fit.



(b) Huber linear model fit.



(c) LAD linear model fit.



(d) Least-squares multiplicative model fit.

Legend.

x axis measured in minutes for linear model fits.

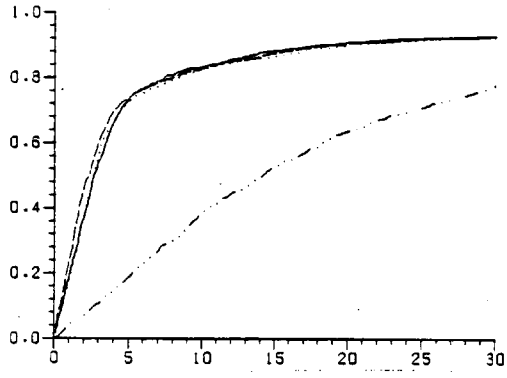
x axis measured in log(base 10) minutes for multiplicative model fit.

residuals arising from the multiplicative model fit to the class 1 reduced sample. The logarithmic transformation did produce near-normal residuals for this sample, and succeeded in doing so for the other class series, with the exception of the class 4 series, to which we now turn. The histograms of the least-squares, Huber and LAD residuals arising from the linear model fits to the class 4 sample are shown respectively in figures 4.3(a), (b) and (c). The differences between the three plots are hardly discernible. Clearly, this is because the three sets of parameter estimates are very close to each other, as the results of Table 4.3 confirm. The robust solutions have not signifcantly improved on the least-squares one, because the distribution of the residuals from the least-squares fit shows no serious departures from normality. As a result, when the logarithmic transformation is carried out, the distribution of the residuals arising from the least-squares fit is clearly non-Gaussian (see figure 4.3(d)).

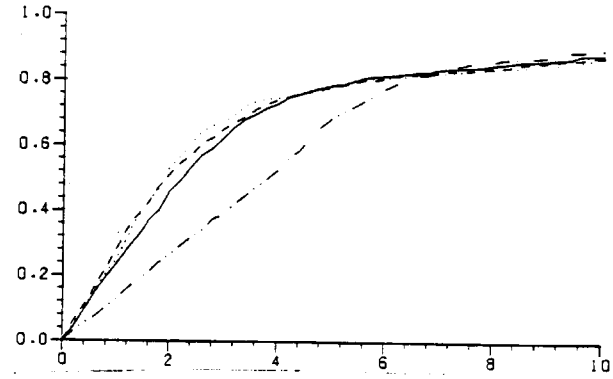
Finally, we examine the predictive performance of the models. The choice of a measure of predictive accuracy (i.e. the absolute prediction error) has been discussed above. In order to compare the accuracy of the four fits to each of the N79W1 series, the cumulative absolute error distributions for each fit were computed and plotted. Each of the figures (4.4(a) to 4.4(f)) compares these distributions for the residuals arising from the Huber, LAD , linear/least-squares and multiplicative/least-squares fits.

The relevant plot for the full sample (FS) class 1 series is shown in figure 4.4(a). The inferiority of the

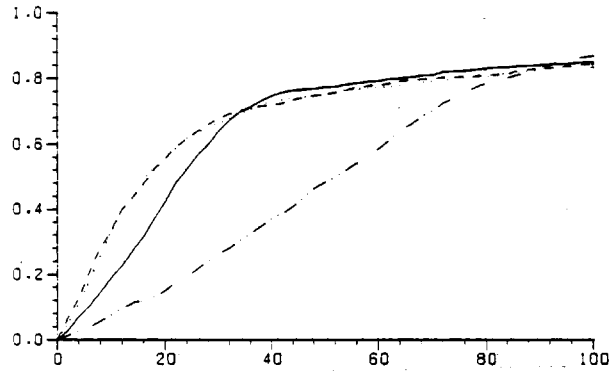
Figure 4.4 : Cumulative absolute error distributions ; data-set N79W1.



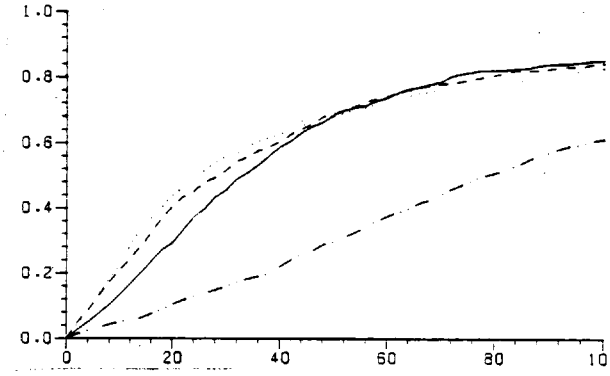
(a) Class 1 (full sample) Jobs.



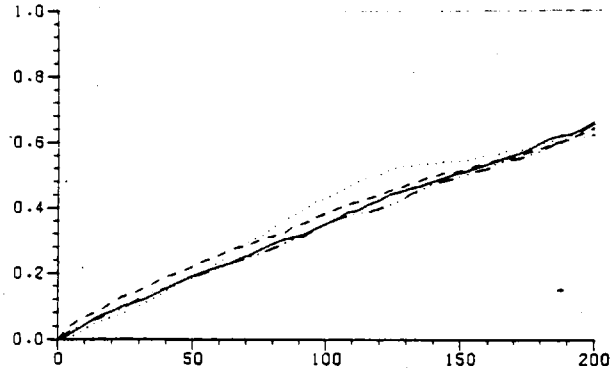
(b) Class 1 (reduced sample) Jobs.



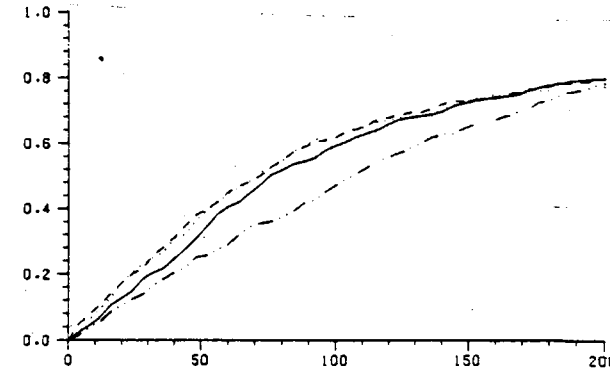
(c) Class 2 Jobs.



(d) Class 3 Jobs.



(e) Class 4 Jobs.



(f) Class 5 Jobs.

Legend.

Least-squares linear model fit.

Huber linear model fit.

LAD linear model fit.

Least-squares multiplicative model fit.

x axis measured in minutes.

least-squares linear model fit, relative to the other three, is self-evident. The other three fits, between which there are negligible differences, are also superior to the least-squares linear model fit for the class 1 reduced sample (fig. 4.4(b)). The results exhibit a similar pattern for the fits to the class 2 and 3 samples (figs. 4.4(c) and (d) respectively). Figures 4.4(e) and 4.4(f) pertain to the fits to the class 4 and 5 samples respectively. Over the range of values shown, there are negligible differences between the Huber, LAD and multiplicative/least-squares models. However, predictions of jobs with very large turnaround times (over four hours), all of which are class 4 or 5 jobs, tended to be seriously overestimated by the multiplicative model. For example, the maximum turnaround predicted by this model for a class 4 job was $10^{3.7}$ minutes and $10^{3.5}$ minutes for a class 5 job. While no job in either of the samples has an observed turnaround of greater than 24 hours, both these forecasts, when translated back to the original measurement scale, exceed 100 hours. This feature of the multiplicative model is clearly a grave shortcoming.

A further comparison of the goodness-of-fit of the models was made by examining the mean absolute prediction error for all the fits. For data-set N79W1, these are given in table 4.4, while the mean absolute predictions of the fits to data-sets N79W2 and N79W3 appear in table 4.2A, which is appended to the main body of this chapter.

Table 4.4 : Mean absolute prediction errors (in minutes) for full and subset regression fits ; data-set N79W1.

Class	Mean Turnaround time	FULL REGRESSION (7 VARIABLES)				SUBSET REGRESSION			
		LIN/LS	LIN/HUB	LIN/LAD	MULT/LS	LIN/LS	LIN/HUBER	LIN/LAD	MULT/LS
1-FS	18.31	27.69	16.82	16.77	16.90	27.47	16.83	16.77	16.94
1-RS	7.02	5.30	4.67	4.63	4.66	5.30	4.67	4.64	4.66
2	66.63	71.82	59.76	57.07	59.32	72.13	59.79	58.84	59.54
3	80.82	74.77	65.66	64.92	65.46	74.54	65.81	64.97	65.64
4	334.21	169.12	167.56	166.40	228.74	169.31	168.05	166.53	232.11
5	262.74	149.91	137.94	136.42	163.16	147.28	137.86	136.46	162.95

Legend.

LIN/LS Linear model/Least-squares estimation procedure.

LIN/HUBER Linear model/Huber estimation procedure.

LIN/LAD Linear model/Least absolute deviations estimation procedure.

MULT/LS Multiplicative model/Least-squares estimation procedure.

With an absolute, rather than a squared, loss function, it is possible for the subset fits to produce more accurate sets of predictions than their corresponding full regression fits (with the exception of the LAD fits). For most of the classes, the Huber and LAD fits substantially improve on the accuracy of the least-squares linear model fits.

The overall effectiveness of each fit may be gauged both in relation to the average absolute prediction error (a.p.e.) of the other fits, and by comparison with the mean turnaround time. In the context of the former criterion, the results of Table 4.4 show that the average a.p.e of the Huber fits for the first three classes is marginally smaller than that of the multiplicative least-squares model fits, but for job classes 4 and 5, the Huber fits are considerably superior. This is undoubtedly due to the breakdown of the multiplicative least-squares model in predicting jobs with very large turnaround times. In the context of the latter

criterion, the mean a.p.e. for the Huber fits to job classes 1, 2 and 3 are approximately of the same order of magnitude as the mean turnaround times for the corresponding classes. For job classes 4 and 5, the average a.p.e. is about half the mean turnaround time for these classes when fitted by the Huber procedure.

We briefly discuss the advantages and shortcomings of each of the four models. The linear model structure has the advantage that it provides the better representation of the real world, because the coefficients in the linear models have an unambiguous physical meaning. The least-squares fit to the linear model failed because not only were the residuals non-normal, but also the goodness-of-fit was poor. The Huber technique produced a sound set of predictions for all the class series. Even in the case of the trimmed class 1 sample of the N79W1 data-set, it provided a more accurate fit than the corresponding linear least-squares model, although the extent of the improvement was not large in absolute terms. Recent work has improved the understanding of the computational solution for the LAD estimators. But the theoretical framework of the LAD estimators, and their estimated covariances, is not as sound as it is for the Huber approach. In all the class series, the difference between the overall accuracy of the predictions generated by the two robust procedures was minimal. The Huber approach is flexible and versatile.

The multiplicative model works adequately for the first three class series, but breaks down in the case of

large turnaround jobs in classes 4 and 5. However, apart from this inadequacy, the fundamental objection to the model is that it is an unrealistic representation of the relationship between the turnaround time of a job and the explanatory variables.

We conclude that the most appropriate paradigm for modelling turnaround times (from the NUMAC system) is a linear model the parameters of which are estimated by the Huber procedure.

Finally, in order to gauge the variation between the Huber fits to each of the three data-sets, the coefficients of the subset Huber models for the N79W1 and N79W2 data-sets were employed to predict the turnaround times of jobs in the N79W3 data-set. The results are given below in table 4.5.

Table 4.5 : Mean absolute prediction errors of turnaround times of jobs in data-set N79W3 when these are predicted by the coefficients of the subset Huber fits to data-sets N79W1 and N79W2.

		N79W3 turnaround times predicted by coefficients from subset Huber fits to data-set		
Job	Class	N79W1	N79W2	N79W3
	1	26.23	26.54	26.05
	2	48.98	73.77	48.71
	3	58.65	123.13	55.24
	4	219.04	251.91	204.30
	5	147.05	192.36	142.78

Legend.

Mean absolute prediction errors measured in minutes.

The final column on the right-hand side lists the actual mean a.p.e. of the subset Huber fits to the N79W3 data-set. Clearly, the N79W1 and N79W3 data-sets are very similar, while the N79W2 data-set is somewhat anomolous compared to the other two. The coefficients from N79W2 data-set produce predictions for job classes 2 and 3 which are particularly poor. Nevertheless, the results indicate that it is not practicable to formulate a static turnaround model, whose coefficients are estimated once and for all, to predict different weeks of NUMAC data. Some updating procedure, whereby the coefficients of the model are amended in the light of additional information, is to be preferred. The prediction of job turnaround times from week to week should therefore be dynamic.

4.2.3 Rutherford

The Rutherford scheduler allocates priorities to batch jobs according to the amount of CPU time and CORE requested. The assignment of job priorities is made in accordance with the following scheme:

		Maximum Core Size Requested (Kbytes)			
		0-210	210-350	350-560	>560
CPU TIME	0-10	P1	P2	P3	P4
REQUESTED	10-90	P5	P6	P7	P8
(in 370/195 seconds)	>90	P9	P10	P11	P12

Turnaround guidelines issued by user support at Rutherford indicate that, for example, the system is tuned

to turnaround a P1 job in approximately 15 minutes. Only one week (from October 1978) of Rutherford workload data has been analysed - we refer to it by the acronym R78W1. As for the NUMAC data, a preliminary study, whose objective it was to reduce the number of priorities, was carried out. In table 4.6 below we present the summary statistics of the turnaround times of jobs in each of the priorities, and a separate class for jobs using magnetic tapes.

Table 4.6 : Summary statistics of Rutherford Job Priorities ; data-set R78W1.

Priority	No. of Jobs	Turnaround time (in minutes)	
		mean	std. dev.
1	4258	14.3	56.4
2	89	126.5	186.3
3	42	117.8	168.8
4	1	695.0	
5	3121	70.6	148.5
6	434	143.2	214.9
7	317	245.6	262.6
8	24	544.0	252.7
9	135	515.5	602.9
10	83	415.9	401.9
11	81	744.3	292.7
12	60	817.8	136.1
Magnetic tape jobs	1130	225.6	306.7

We decided to use two boundary points, 10 seconds requested cpu time and 210 kbytes of core. We also discarded certain priority classes which had few entries. We retained P1 jobs as a class of their own, merged priorities P5 and P9, discarded jobs with priorities P2, P3 and P4, and

amalgamated jobs with the remaining priorities into a separate class. Rutherford jobs were divided in the following way:

Table 4.7 : Definition of Rutherford Job classes.

Class	Amalgamated Priorities	Definitive Characteristic
1	1	Jobs requesting 0-10 cpu seconds and 0-210 kbytes of core.
2	.5,9	Jobs requesting more than 10 cpu seconds and 0-210 kbytes of core.
3	6-8,10-12	Jobs requesting more than 10 cpu seconds and more than 210 kbytes of core
4		Jobs requesting the use of one or more magnetic tapes.

Then, the six explanatory variables (excluding the constant term) to appear in each of the four turnaround models are: REQCPU, REQCORE, QLC1, QLC2, QLC3, and QLC4, where, for example, the QLC1 variable represents the queue length of class 1 jobs. From Table 4.6, it can be seen that the number of jobs appearing in each of the four classes are, respectively, 4258, 3256, 999 and 1130. To reduce the computing time, and the amount of storage required to analyse the jobs in classes 1 and 2, we decided to examine a subset of these samples which consisted of those jobs submitted on the first day of the week. This expediency reduced the size of the class 1 and 2 samples to more manageable proportions.

Each of the four classes was then fitted by the four models. The subset selection procedure was slightly different from that employed for the NUMAC models. The full regression fits were again examined with a view to compiling a physical basis of variables. In this case, the REQCPU

and REQCORE were found to be important influences on job turnaround. Of the system congestion variables, the same clear relationship was found to exist between the turnaround time of a job and the queue to which it was assigned on arrival. However, the QLC1 variable, representing the queue length of "instant turnaround" jobs, did not exert a strong influence on the turnaround times of jobs from all classes, which is contrary to the conclusions from the NUMAC data. In particular, the relationship between the QLC1 variable and the turnaround times of jobs in classes 3 and 4 was found to be tenuous. One reason for this phenomenon is that, for the Rutherford system, many of the class 3 and 4 jobs experience much longer turnaround times than do jobs of comparable size at NUMAC. Thus we would not expect the number of jobs whose average turnaround time is of the order of 15 minutes (i.e. class 1 jobs) to heavily influence jobs whose average turnaround is of the order of five hours or more (i.e. class 3 and 4 jobs). These considerations led us to the following approximation to the best subset of variables:

JOB CLASS	PHYSICAL BASIS
1	REQCPU, REQCORE, QLC1
2	REQCPU, REQCORE, QLC2
3	REQCPU, REQCORE, QLC3
4	REQCPU, REQCORE, QLC4

Each job class series was then re-fitted by the explanatory variables listed in the physical basis. Using a sequence of likelihood ratio tests, the best subset fits

Table 4.8 : Results of subset regression fits ; data-set R78W1.

Data-set Class	Model/Estimation Procedure	EXPLANATORY VARIABLE Constant	REQCPU	REQCORE	QLC1	QLC2	QLC3	QLC4
	Linear/Least-Squares	34.02*	-2.485*	-0.0197	-0.0680			0.0789*
		(7)	(0.6)	(0.02)	(0.18)			(0.04)
	Linear/Huber	3.152*	-0.1524*	0.0156*	0.1591*	-0.0373*		
R78W1		(0.4)	(0.04)	(0.001)	(0.002)	(0.01)		
Class 1	Linear/LAD	2.476	-0.1317	0.0178*	0.1747*	-0.0446*		
		(2.8)	(0.23)	(0.008)	(0.08)	(0.02)		
	Multiplicative/	0.7415*	-0.0355*	0.00120*	0.0135*	-0.00200*		
	Least-Squares	(0.07)	(0.01)	(0.0002)	(0.002)	(0.001)		
	Linear/Least-Squares	-0.9236	0.0962*	0.3316*		0.1765		
		(18.3)	(0.01)	(0.08)		(0.24)		
	Linear/Huber	-6.020	0.0890*	0.1050*	0.8663*	0.1232	-0.1613	
R78W1		(3.7)	(0.003)	(0.02)	(0.19)	(0.07)	(0.06)	
Class 2	Linear/LAD	-8.073	0.0865*	0.0898*	1.236*	0.0865	-0.2513*	
		(9.3)	(0.01)	(0.04)	(0.5)	(0.28)	(0.12)	
	Multiplicative/	0.4437*	0.000343*	0.00387*		0.00996*	-0.00703*	
	Least-Squares	(0.09)	(0.00006)	(0.0004)		(0.002)	(0.002)	
	Linear/Least-Squares	-104.9*	0.2561*	0.9908*			-0.0787*	
		(28)	(0.03)	(0.07)			(0.04)	
	Linear/Huber	-244.3*	0.3758	1.100*			2.840*	-1.747*
R78W1		(24)	(0.02)	(0.1)			(0.2)	(0.7)
Class 3	Linear/LAD	-330.8*	0.3712	1.269*			-0.000059	
		(24)	(0.02)	(0.1)			(0.036)	
	Multiplicative/	0.9831*	0.000494*	0.00224*	0.0191*	-0.00274*	0.0000446	
	Least-Squares	(0.08)	(0.00007)	(0.0002)	(0.005)	(0.0009)	(0.0001)	
	Linear/Least-Squares	8.326	0.3046*	0.8866*				-0.0258
		(18.4)	(0.03)	(0.07)				(0.13)
	Linear/Huber	-95.85*	0.4685*	0.7757*	2.261*			0.1572
R78W1		(13)	(0.02)	(0.04)	(0.6)			(0.08)
Class 4	Linear/LAD	-88.50*	0.5110*	0.5335*	2.584*			0.1814*
		(15)	(0.02)	(0.05)	(0.7)			(0.09)
	Multiplicative/	1.018*	0.000371*	0.00236*	0.0126*			0.00117*
	Least-Squares	(0.06)	(0.00007)	(0.0002)	(0.002)			(0.0003)

Legend.

* denotes significance at the 5% level.

The parameter estimate is the upper figure in each cell ;

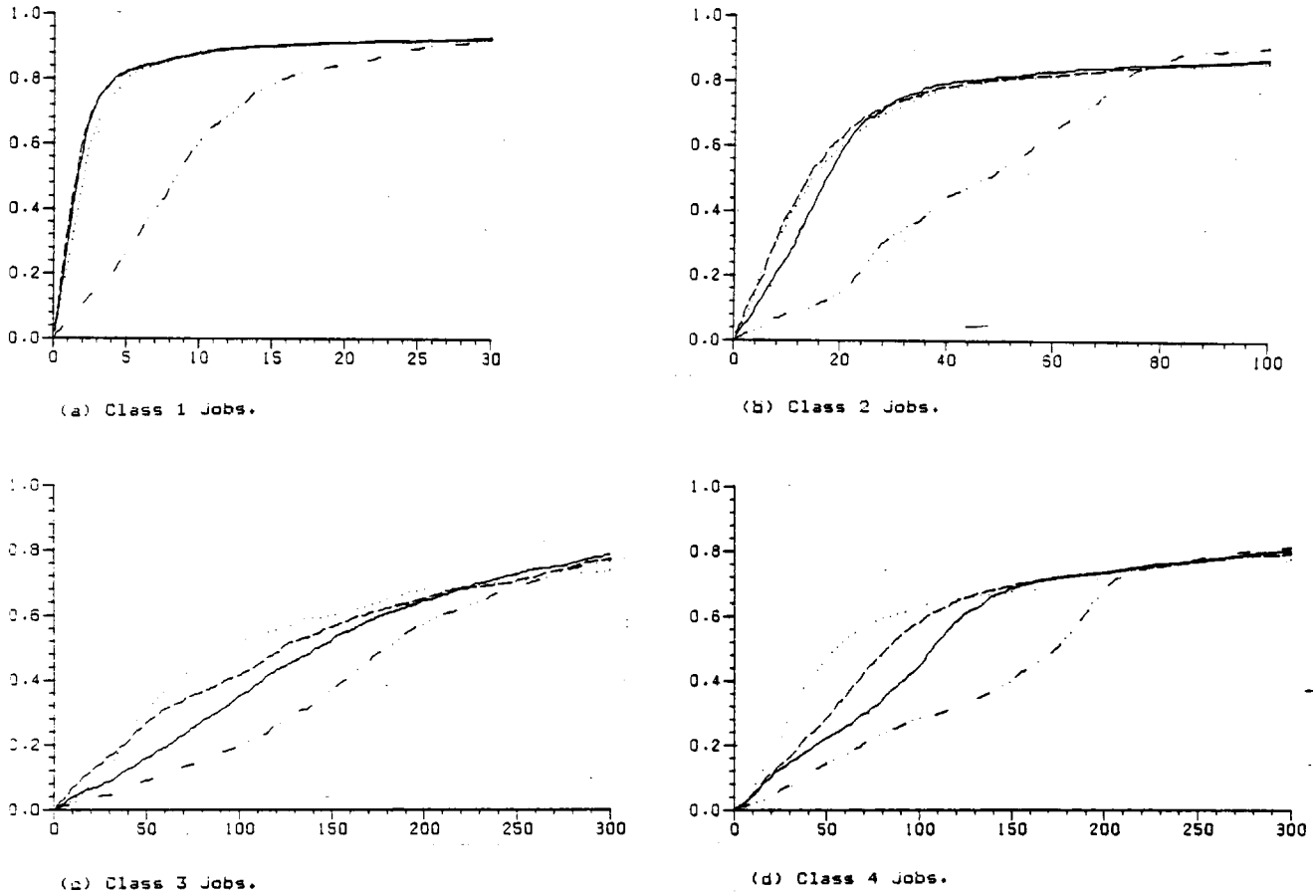
below it, in parentheses, is its estimated standard error.

were found as for the NUMAC data.

The results of the subset regression fits are displayed in Table 4.8. The turnaround time of a class 1 job, according to these results, is inversely related (in all the fits) to the amount of core requested by the job. With this exception, the turnaround times of all other jobs are directly related to the amount of resources requested by the job. The results indicate that the longer the queue to which a class 1 job is assigned, the longer is its turnaround time. The turnaround time of class 2 jobs (those requesting more than 10 seconds cpu time but less than 210 kbytes of core) is inversely related to the length of the queue of class 3 jobs. However, the results of the fits to the class 3 turnaround series indicate that class 3 jobs are unaffected by the number of class 2 jobs in the system at the time of their submission. The Huber fit to the class 3 series has a significant coefficient for the QLC3 variable. Apart from the cpu time and core requested by them, turnaround times of magnetic tape (class 4) jobs are heavily influenced by the number of class 1 jobs in the system at their submission time.

We do not present a graphical examination of the residuals amoung from any of these fits, as they were found to behave in a similar way to the residual plots that were constructed for the NUMAC data. Finally, we investigated the predictive performance of the models. As before, we employed the mean absolute prediction error (a.p.e.) as our criterion of goodness-of-fit. The cumulative absolute error distributions for each fit were computed and plotted, one

Figure 4.5 : Cumulative absolute error distributions ; data-set R78W1.



for each of the job classes. These are shown in figures 4.5(a) to 4.5(d). In each of these, the inadequacy of the least-squares fit is readily apparent. Over the range of values shown, the plots suggest that the best results are obtained by the LAD fit, then the multiplicative least-squares fit, then the Huber fit, although for classes 1 and 2 there is little to choose between these plots. However, the plots for classes 3 and 4 (figures 4.5(c) and (d)) conceal the fact that the accuracy of the predictions of the multiplicate least-squares model is seriously undermined by the occurrence of a small proportion of extremely large errors. This deficiency in the multiplicative model is best shown in the context of the mean absolute prediction errors for each of the fits, listed in Table 4.9. Here the average absolute prediction error of the multiplicative model fit to classes 3 and 4 is considerably larger than the corresponding figures for the linear model Huber and LAD fits.

It is evident that the multiplicative model breaks down when either the resources requested are very large (this is more likely the case), or when the system becomes over-congested. At any rate, this shortcoming in the model belies the fact that it is not an appropriate representation of the real world situation.

As for the NUMAC system, the results listed in Table 4.9 indicate that the Huber linear model fit is superior to both the linear and multiplicative least-squares fits, and is on a par with the LAD linear model fit. For the same reasons as before, we conclude that the Huber linear model

fit is the most appropriate paradigm to model the turnaround of Rutherford jobs.

Table 4.9 : Mean absolute prediction errors (in minutes) for full and subset regression fits ; data-set R7CW1.

Class	Mean Turnaround time	FULL REGRESSION (6 VARIABLES)				SUBSET REGRESSION			
		LIN/LS	LIN/HUB	LIN/LAD	MULT/LS	LIN/LS	LIN/HUBER	LIN/LAD	MULT/LS
1	13.86	15.90	10.81	10.79	11.04	15.87	10.81	10.79	11.02
2	65.71	69.61	55.22	54.67	59.00	69.66	55.22	54.67	59.22
3	297.27	227.70	211.74	207.61	234.12	237.55	211.89	207.98	228.47
4	225.61	200.80	181.77	179.93	233.31	201.69	182.38	180.45	233.40

Legend.

LIN/LS Linear model/Least-squares estimation procedure.

LIN/HUBER Linear model/Huber estimation procedure.

LIN/LAD Linear model/Least absolute deviations estimation procedure.

MULT/LS Multiplicative model/Least-squares estimation procedure.

Appendix

Table 4.1A : Results of subset regression fits ; data-sets N79W2 and N79W3.

Data-set Class	Model/Estimation Procedure	EXPLANATORY VARIABLE							
		Constant	REQCPU	REQCORE	QLC1	QLC2	QLC3	QLC4	QLC5
	Linear/Least-Squares	70.50*	10.38*	-0.3168*	0.8701			-0.5064*	
		(16)	(5.1)	(0.06)	(0.73)			(0.12)	
	Linear/Huber	-2.920*	2.720*	-0.00346	0.8991*		-0.1040*		0.1277*
N79W2		(1.3)	(0.4)	(0.005)	(0.06)		(0.03)		(0.04)
Class 1	Linear/LAD	-3.773	2.623	-0.00722	0.740*				
		(8.0)	(2.4)	(0.03)	(0.36)				
	Multiplicative/ Least-Squares	0.4648*	0.1816*	-0.000896*	0.0396*			-0.00466*	
		(0.10)	(0.03)	(0.0004)	(0.005)			(0.0007)	
	Linear/Least-Squares	-66.76	10.61	0.6363*	4.411*	1.476*	-1.537*		
		(53)	(7)	(0.10)	(1.0)	(0.4)	(0.4)		
	Linear/Huber	-60.43	10.61*	0.3539*	3.532*	0.2131			
N79W2		(36)	(5)	(0.07)	(0.7)	(0.15)			
Class 2	Linear/LAD	-105.5*	12.00*	0.3751*	3.870*	1.281*			
		(43)	(6)	(0.08)	(0.8)	(0.3)			
	Multiplicative/ Least-Squares	0.5904*	0.0692*	0.00244*	0.0307*	0.00755*	-0.00652*		
		(0.22)	(0.03)	(0.0004)	(0.004)	(0.002)	(0.002)		
	Linear/Least-Squares	-0.9284	-0.5137	1.102*	4.857*	1.235*	-2.326*		
		(34)	(0.92)	(0.1)	(1.5)	(0.6)	(0.6)		
	Linear/Huber	18.06	-0.9099	0.7434*	5.724*		-0.4795		
N79W2		(31)	(0.84)	(0.12)	(1.3)		(0.30)		
Class 3	Linear/LAD	-29.05	0.00168	0.8094*	5.207*		-0.3265		
		(42)	(1.1)	(0.16)	(1.8)		(0.40)		
	Multiplicative/ Least-Squares	1.419*	-0.00255	0.00205*	0.0207*	0.00429*	-0.00363		
		(0.1)	(0.003)	(0.0005)	(0.005)	(0.002)	(0.002)		
	Linear/Least-Squares	351.9*	0.6837*	-0.2114	5.599*			-1.159*	
		(43)	(0.05)	(0.17)	(2.2)			(0.30)	
	Linear/Huber	249.9*	0.7620	-0.2458	8.360*			-2.090*	4.491*
N79W2		(55)	(0.06)	(0.19)	(2.7)			(0.7)	(1.8)
Class 4	Linear/LAD	98.22	0.8183*	0.0278	6.809*	2.359*		-3.214*	6.207*
		(61)	(0.06)	(0.21)	(3.0)	(1.0)		(0.7)	(2.0)
	Multiplicative/ Least-Squares	2.210*	0.00112*	-0.000372	0.0205*			0.000414	
		(0.1)	(0.0001)	(0.0004)	(0.005)			(0.0006)	

Table 4.1A (continued) ; Results of subset regression fits ; data-sets N79W2 and N79W3.

Data-set Class	Model/Estimation Procedure	EXPLANATORY VARIABLE Constant	REQCPU	REQCORE	QLC1	QLC2	QLC3	QLC4	QLC5
	Linear/Least-Squares	-49.65	1.431*	1.451*	-2.908			-1.988*	3.249*
		(50)	(0.1)	(0.2)	(2.4)			(0.5)	(1.6)
	Linear/Huber	-24.14	1.723*	0.9475*	-1.482*			-1.719*	2.814*
N79W2		(32)	(0.1)	(0.16)	(0.4)			(0.4)	(1.2)
Class 5	Linear/LAD	18.67	1.736*	0.6870*	-1.614			-1.399*	1.664
		(41)	(0.1)	(0.20)	(2.1)			(0.5)	(1.4)
	Multiplicative/ Least-Squares	1.469*	0.00166*	0.00325*	0.00777			-0.00279*	0.00763*
		(0.1)	(0.0002)	(0.0004)	(0.004)			(0.0001)	(0.003)
	Linear/Least-Squares	107.7*	-2.959	-0.4283*	2.850*				-1.417*
		(22)	(6.7)	(0.07)	(1.4)				(0.33)
	Linear/Huber	-1.708	1.609*	-0.00812*	1.504*	-0.0892*			
N79W3		(1.5)	(0.5)	(0.004)	(0.1)	(0.04)			
Class 1	Linear/LAD	-5.033	2.071	0.000861	1.136				
		(9.8)	(3.2)	(0.032)	(0.7)				
	Multiplicative/ Least-Squares	0.3039*	0.1432*	-0.00185*	0.0982*	-0.00840*			
		(0.12)	(0.04)	(0.0004)	(0.009)	(0.003)			
	Linear/Least-Squares	-147.6*	14.80*	0.5748*	2.862*	-0.2198			
		(38)	(5.3)	(0.06)	(1.0)	(0.29)			
	Linear/Huber	-57.65*	6.553*	0.1788*	1.458*	0.5134*			
N78W3		(15)	(2.0)	(0.02)	(0.4)	(0.11)			
Class 2	Linear/LAD	-48.66*	6.072*	0.0844*	1.247*	0.7053*			
		(19)	(2.7)	(0.03)	(0.6)	(0.17)			
	Multiplicative/ Least-Squares	-0.3674*	0.1244*	0.00268*	0.0244*	0.0130*			
		(0.24)	(0.03)	(0.0004)	(0.007)	(0.002)			
	Linear/Least-Squares	-18.68	-0.6306	0.4338*	3.990*	-1.168*	1.001*		
		(20)	(0.52)	(0.08)	(1.3)	(0.5)	(0.4)		
	Linear/Huber	4.595	0.7899*	-0.0291	2.676	-0.5108*	1.157*		
N78W3		(7.4)	(0.18)	(0.03)	(0.5)	(0.20)	(0.1)		
Class 3	Linear/LAD	-2.363	0.8515*	-0.0492	2.065*		0.9703*		
		(11.1)	(0.29)	(0.04)	(0.6)		(0.18)		
	Multiplicative/ Least-Squares	0.8111*	0.0102*	0.000672	0.0255*		0.00917*		
		(0.10)	(0.003)	(0.0004)	(0.006)		(0.0016)		

Table 4.1A (continued) : Results of subset regression fits ; data-sets N7$W2 and N7$W3.

Data-set Class	Model/Estimation Procedure	EXPLANATORY VARIABLE Constant	REQCPU	REQCORE	QLC1	QLC2	QLC3	QLC4	QLC5
	Linear/Least-Squares	53.42 (45)	0.9845* (0.06)	1.102* (0.2)	-3.051 (3.5)			0.7986 (0.77)	-6.703* (1.4)
	Linear/Huber	-17.80 (39)	1.158* (0.05)	0.6780* (0.15)	1.621 (3.8)			0.9087 (0.68)	-4.220* (1.4)
N79W3 Class 4	Linear/LAD	-87.58* (40)	1.290* (0.05)	0.5190* (0.15)	1.518 (3.1)			1.496* (0.7)	-3.774* (1.2)
	Multiplicative/ Least-Squares	1.338* (0.1)	0.00191* (0.0001)	0.00166* (0.0004)	0.0138 (0.008)			0.00179 (0.001)	
	Linear/Least-Squares	66.05 (50)	0.9122* (0.07)	0.7416* (0.21)	2.413 (4.3)			-2.355* (0.90)	0.2311 (1.6)
	Linear/Huber	-39.32 (34)	0.9338* (0.05)	0.6184* (0.14)	8.924* (3.6)			-1.152* (0.6)	1.275 (1.2)
N79W3 Class 5	Linear/LAD	14.62 (42)	1.007* (0.06)	0.2705 (0.18)	6.398 (3.6)			-1.294* (0.6)	1.831 (1.2)
	Multiplicative/ Least-Squares	1.362* (0.1)	0.00127* (0.0002)	0.00204* (0.0005)	0.0308 (0.009)				0.00432* (0.002)

Legend.

* denotes significance at the 5% level.

The parameter estimate is the upper figure in each cell ; below it, in parentheses, is its estimated standard error.

Table 4.2A : Mean absolute prediction errors (in minutes) for full and subset regression fits ; data-sets N79W2 and N79W3.

Class	Mean Turnaround time	FULL REGRESSION (7 VARIABLES)				SUBSET REGRESSION			
		LIN/LS	LIN/HUB	LIN/LAD	MULT/LS	LIN/LS	LIN/HUBER	LIN/LAD	MULT/LS
W2-C1	29.89	39.89	26.43	28.38	26.65	39.87	26.43	26.53	26.69
W2-C2	142.96	124.03	115.49	113.57	115.39	124.21	115.91	114.21	115.52
W2-C3	196.69	153.68	148.57	149.65	153.07	153.64	149.60	149.00	152.89
W2-C4	422.12	205.15	203.55	202.62	254.78	209.07	206.78	206.78	255.48
W2-C5	308.20	158.95	148.83	146.85	175.87	159.35	149.28	147.46	177.39
W3-C1	28.75	44.67	26.07	25.95	25.83	44.55	26.05	26.00	26.04
W3-C2	57.93	56.91	48.61	48.22	49.13	56.99	48.71	48.27	49.08
W3-C3	69.84	67.79	55.20	54.54	55.28	68.00	55.24	54.74	55.40
W3-C4	356.43	213.66	203.98	199.38	320.38	213.52	204.30	199.54	317.05
W3-C5	255.63	147.08	138.68	137.31	161.50	148.57	142.78	137.86	160.21

Legend.

LIN/LS Linear model/Least-squares estimation procedure.

LIN/HUBER Linear model/Huber estimation procedure.

LIN/LAD Linear model/Least absolute deviations estimation procedure.

MULT/LS Multiplicative model/Least-squares estimation procedure.

W2-C1 ; Numac data-set N79W2, Class 1.

CHAPTER 5

CONCLUSIONS

Since this study involves both statistical theory and method, and an application of statistics to computer performance evaluation, we have divided the conclusions into two sections: the first (section 5.1) deals with the results of the analysis in the context of computer performance evaluation, while the second (section 5.2) considers conclusions of a statistical nature.

5.1 Computer Performance Evaluation

As far as the author is aware, few computer performance analyses (at least of IBM systems) involving job turnaround time have been carried out, for the following reason: on an IBM system, there are two principal records compiled on each job. The first is a System Management Facility (SMF) record (type 132) that is written just after job submission time: the second is the so-called Condensed Record, written just after job termination. The former contains information such as the cpu and core requested and job submission time, but not job end time. The latter contains all the information pertaining to the demands placed on the resources of the system by the job (e.g. job start and end time, cpu and core consumed), but not the submission time of the job. The difficulty is that the order in which the two sets of records are written is not the same. Thus, even before the required

data on turnaround time (which requires knowledge of both job submission and job termination time) could be extracted, the two sets of records had to be matched.

The general method that has been applied to the evaluation of computer performance in this study has been to arrange the structure of the model according to that suggested by the characteristics of the data. Thus, for example, we have defined job classes in an analogous way to the actual definitions used at the computer installations. This approach has two advantages: first, since there is a direct correspondence between the installation and the model, the latter is endowed with a certain realism or physical meaning: and second, because the approach is fairly general, it may be applied, with suitable alterations, to analyse the performance of any batch computing system.

Nevertheless, there are also a number of shortcomings in the approach. In particular, there is the problem of jobs that make no specific request for resources on their job card, and, as a result, are allocated the default core or cpu limit by the system. This is bound to introduce inaccuracies. Since, in general, the requested resources of a job over-estimate its actual needs, one plausible method would be to replace these default values by some estimate (e.g. the mean) of the resources requested, or to treat the situation as a missing-value problem. Another problem is amalgamating the classes; we did not employ any formal mathematical method (such as cluster analysis) to determine which priorities should be amalgamated because

these methods typically give different results for different data-sets. The adopted ad hoc approach was preferred because of its flexibility.

Another source of error that was built into the turn-around model took the form of the simplification whereby the queue lengths of jobs awaiting execution were assumed stable over one-minute intervals. Intuition suggests that this simplification would be a large source of error only in the case of the prediction of low resource consumption jobs; in the case of large jobs, the size of the error (relative to the turnaround time of the job) introduced by small changes in the queue lengths would be smaller than in the case of low resource consumption jobs.

The major conclusion of the principal components analysis in the time domain was that, in both installations, the underlying statistical feature of the submissions data could be expressed in terms of an index of the general level of demand for the resources of the system. This knowledge is especially useful to the installation manager: for example, he may use it to assess the effects of changes in the size of the user population. Since the latter necessarily changes the submission rates of jobs to different classes, the result may be used to give an indication of the overall change in the general level of submissions that will follow from a change in the size of the user population. Secondly, knowledge of the composition of the general level of demand is useful for adjusting and 'tuning' the scheduler, and in formulating the relative costs of different jobs. Finally,

the result may be used to describe the changes in the composition of the general level of demand over time; for example, a shift from Input/Output bound jobs to cpu-bound jobs will be reflected by a change in the magnitudes of the coefficients associated with certain job classes in the first eigenvector.

One puzzling result was that the second dimension arising from the (time-domain) principal components analysis of the two installations was different: for the Rutherford system, it was an index of the general cpu demand, and for NUMAC, an index of the general core demand. There does not seem to be any obvious reason why this should be the case, although it is probably related to differences in the computing requirements of the two user communities.

The principal components analysis in the frequency domain did not uncover as much knowledge about the job submission rates as the analysis in the time domain. The one significant conclusion concerned the dependence of the general level of demand for the resources of the system on the previous short-term job submission level. The length of the period, about 90 minutes, could consist of two components: turnaround plus 'think' time. The 90-minute period is thus an average measure of the length of submission-resubmission time taken by a large number of users who submit their jobs, await the output, then, either submit a different job using the new results, or modify the same program, then resubmit it.

The major strengths in the design of the turnaround model are that it allows the parameters of the model to be

updated from week to week, thus reflecting current job turn-around behaviour, and that it is sensitive to large or unusual changes in job submissions (e.g. the effect on the turnaround time of a job submitted just after the simultaneous submission of a large number of jobs of the same class).

One of the main sources of error that upset the estimation of the parameters in the turnaround model was the excessive variation in the observed turnaround times of jobs submitted in the same one-minute interval and having similar resource requirements. The most probable cause of this phenomena is the advent of a temporary sytem breakdown after one of the jobs has executed, but while the second is still awaiting execution. The use of the robust estimation procedures partially overcame this problem.

As we have already assessed the accuracy of the predictions produced by the turnaround models (in sec. 4), it remains to consider the utility of the turnaround model. For example, we may ask what benefits accrue to the user if he has a good estimate of the turnaround time of a job. Clearly, the user will waste less time retrieving output, but this saving is marginal. The major utility of the model to the user is that it gives an immediate and quantitative indication of the degree of system congestion: for example, it may cause the user to delay the submission of a job until the system is less heavily loaded. Just as the turnaround time of a job may be seen as a measure of the performance of the system on which it is run, so the utility of a turn-around model may also be analysed in terms of the performance

of the system. In this sense, it would be possible, given the parameters of the models for different classes, to use the turnaround model as a theoretical tool: for example, to assess the effects of different mixes of jobs on the turnaround times of jobs in particular classes.

Finally, the results of the turnaround analysis have been helpful in clarifying the relationships between the explanatory variables, in particular, between the queue length variables. For instance, a study of the variables appearing in the final subset model for each class suggests that the turnaround time of a low resource request job is mainly influenced by the number of other low resource request jobs in the system at the time of its submission, with the analogous result for high resource request and magnetic tape jobs.

5.2 Statistical Conclusions

The main statistical contribution of a non-standard nature in the first part of this study was the presentation of a coherent theory of principal components analysis in the frequency domain. In addition, a method to obtain a consistent estimate of the cross-spectral density matrix was developed. A central feature of the time and frequency domain principal component analysis was the removal of the variation about the mean, the seasonality, and the local correlation in each series. Each of these sources of variation had to

be removed before the underlying statistical features of the submissions data could be uncovered.

Another important step in the principal components analysis was the prior transformation of each series. No formal mathematical technique (such as the well-known Box-Cox (1964) transformation analysis) was employed here: the eventual choice of transformation (i.e. square root) was determined by a simple graphical analysis, and an examination of the within-day variances. In our case, we did not use the Box-Cox method because it is doubtful whether it would have given the same transformation parameter for each data-set. Since a unified approach was required (i.e. one that resulted in the use of the same transformation parameter for each data-set) we decided to sacrifice the increased reduction in variation that would have resulted from the use of a Box-Cox analysis, for computational convenience and overall simplicity. Additional justification for the square-root transformation was that the mean submission level for each class appeared to be related to the variance.

The analysis of job submission patterns described in chapter 2 was essentially a case study in the application of principal components in the time and frequency domains. One important aspect that arose out of the frequency domain study was the development and formulation of a sound, flexible methodology: the first step being to detrend and prewhiten the data; a further study to determine the smoothing parameters of the cross-spectra, then finally the construction and decomposition of the cross-spectral density matrices. Finally, the study showed that a principal component analysis of time-

series data in both the time and frequency domains can be useful and revealing.

The main theoretical results appeared in Chapter 3, which reviewed in detail the current state of the art of robust regression. Part of section 3.1 dealt with the theory underlying the Linear Programming (LP) formulation of the LAD minimisation problem: and, in the context of the simplex method of solution, conditions for uniqueness and optimality were given. One reason why LAD regression has not been popular amongst statisticians is because the absolute value function is difficult to work with mathematically. Another is that there is no tractable finite sample theory for LAD-estimators. Indeed, before these estimators come to be used as regularly as least-squares, the distribution theory must be given a firm base. This study has made some inroads into this problem; a likelihood ratio test was developed, and its distribution tested via Monte Carlo sampling. While these results suggested that the χ^2-distribution was a good approximation for light-tailed and asymmetric error distributions, the results of the subsequent power study of the test tended to undermine the very validity of the approximation. Clearly, more work on the size and power of this test is required. For example, this study has considered only the one-dimensional case (i.e. when the difference between the number of parameters in the full and reduced models is one); it is equally important to consider the behaviour of the power of the test in problems of high dimensions.

Other results in section 3.3 suggested that the 't-test' of LAD-estimator lacks power against departures from the null

hypothesis. This conclusion is also an indication that the asymptotic formula linking the variance of an LAD-estimator to that of a least-squares estimator does not hold good for small samples.

The second main robust technique considered in Chapter 3 was Huber's M-estimation procedure. The M-estimator, has attracted considerable interest over the past decade, and so much more is known about it. Yet, in section 3.2, some new results on the adequacy of the distribution of a likelihood ratio test for M-estimators were obtained. A further study of the power of the test in section 3.3 revealed that its behaviour relative to the power of the least-squares 'F-test' conformed to theoretical expectations.

Much of the content of Chapter 3 also dealt with computational considerations. Robust estimation procedures are necessarily iterative. Thus, all aspects of computational solutions were considered: starting values and convergence criterion, the speed and core requirements of the algorithm, its availability, and its performance relative to other algorithms. For most LAD problems, the algorithm proposed by Armstrong, Frome and Kung (1979) was found to be the most satisfactory. However, occassionally, the algorithm converged prematurely, because of problems of numerical accuracy and tolerance limits. For Huber's M-estimators, the IWLS procedure is probably the easiest to program, and since the least-squares solution will usually be required to obtain estimates of the standard errors of the parameters, this procedure is recommended.

REFERENCES

Abdelmalek, N.N. (1971) Linear ℓ_1 - approximations for a discrete set and ℓ_1 solutions of overdetermined linear equations. Jour. Assoc. Comp. Mach., 18, 41-47.

Abdelmalek, N.N. (1974). On the discrete ℓ_1 approximation and ℓ_1 solutions of overdetermined linear equations. Jour. of Approx. Theory, 11, 38-53

Abdelmalek, N.N. (1975). An efficient method for the discrete ℓ_1 - approximation problem. Math. Comp., 29, 844-850.

Agee, W.S. and Turner, R.H. (1979). Application of robust regression to trajectory data reduction. In Launer, R.L. and Wilkinson, G.N., Eds., "Robustness in Statistics", New York: Academic Press. pp107-126.

Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber P.J., Rogers, W.H. and Tukey, J.W. (1972). Robust Estimates of Location; Survey and Advances. Princeton Univ. Press: Princeton, N.J.

Appa, G. and Smith, C. (1973). On ℓ_1 and Chebychev estimation. Jour. Math. Programming, 5, 73-87

Armstrong, R.D., and Frome, E.L. (1976). A comparison of two algorithms for absolute deviations curve-fitting. Jour. Amer. Statist. Assoc., 71, 328-330.

Armstrong, R.D., Frome, E.L. (1977). A special purpose linear programming algorithm for obtaining Least Absolute Deviation estimates in a linear model with dummy variables. Commun. Statist., B,6, 383-398

Armstrong, R.D., Frome, E.L., and Kung, D.S. (1979). A revised simplex algorithm for the absolute deviation

curve-fitting problem. Commun.Statist, B,8, 175-190.

Armstrong, R.D. and Hultz, J.W. (1977). An algorithm for a restricted discrete approximation in the ℓ_1-norm. S.I.A.M. Jour. Numer. Anal., 14, 555-565.

Armstrong, R.D., and Kung, M.T. (1978). Least absolute value estimates for a simple linear regression problem: Algorithm AS 132. Appl. Statist., 27, 363-366.

Ashar, V.G., and Wallace, T.D. (1963). A sampling study of minimum absolute deviations estimators. Operat.Res, 11, 747-748.

Baker, R.J., and Nelder, J.A. (1978). The GLIM System, Release 3. Oxford:Numerical Algorithms Group.

Barrowdale, I., and Robers, F.D.K. (1970). Applications of mathematical programming to ℓ_p-approximation. In Rosen, J.B., Mangasarian, D.L., and Ritter, K., Eds, "Nonlinear programming." New York; Academic Press.

Barrowdale, I., and Roberts, F.D.K. (1973). An improved algorithm for discrete ℓ_1-approximation. S.I.A.M. Jour. Numer. Anal., 10, 843-848

Barrowdale, I., and Roberts, F.D.K. (1974). Algorithm 478: Solution of an overdetermined system of equations in the ℓ_1-norm. Commun. A.C.M., 17, 319-320.

Barrowdale, I., and RobertsF.D.K. (1977). Algorithms for restricted LAV estimation. Commun. Statist., B,6, 353-363.

Barrowdale, I., and Roberts, F.D.K. (1978). An efficient algorithm for discrete ℓ_1 linear approximation with linear constraints. S.I.A.M. Jour. Numer. Anal., 15 603-611.

Barrowdale, I., and Young, A. (1966). Algorithms for best ℓ_1 and ℓ_∞ linear approximation on a discrete set. Numer. Math., 8, 295-306.

Bartels, C.P. (1977). Economic Aspects of Regional Welfare, Income Distribution, and Unemployment. Lieden; Nijhoff.

Bartels, R.H., and Conn, A.R. (1977). LAV regression: a special case of piecewise linear minimisation. Commun. Statist., B, 6, 329-339.

Bartels, R.H., Conn, A.R., and Sinclair, J.N. (1978). Minimisation techniques for piecewise differentiable functions: the ℓ_1 solution to an overdetermined linear system. S.I.A.M. Jour.Numer.Anal., 15. 224-241.

Bassett, G.W.Jnr., and Koenker, R. (1978). Asymptotic theory of least absolute error regression. Jour. Amer. Statist. Assoc., 73, 618-622.

Beaton, A.E., and Tukey, J.W. (1974). The fitting of power series, meaning polynomials, illustrated on band spectro-scopic data. Technometrics, 16, 147-192.

Bickel, P.J. (1975). One-step Huber estimates in the linear model. Jour. Amer. Statist. Assoc., 70, 428-434.

Bickel, P.J. (1976). Another look at robustness: a review of reviews and some new developments. Scand. Jour. Statist, 3, 145-168.

Birch, J.B. (1980). Effects of the starting-values and stopping-rule on robust estimates obtained by iterated weighted least-squares. Commun. Statist., B,9, 141-154.

Bishop, Y.M.M., Fienberg, W.D., and Holland, P.W.(1975). Discrete Multivariate Analysis: Theory and Practice. Cambridge, Mass.; MIT Press.

Blackman, R.B., and Tukey, J.W. (1959). The Measurement of of Power Spectra. New York: Dover

Bloomfield, P. (1976). Fourier Analysis of Time-Series: an Introduction. New York: John Wiley.

Box, G.E.P., (1979). Robustness in the strategy of scientific model building. In Laurner, R.L. and Wilkinson, G.N., Eds., "Robustness in Statistics", New York: Academic Press

Box, G.E.P., and Cox, D.R. (1964). The analysis of transformations. Jour. Roy. Statist. Soc., B, 26, 211-251.

Box, G.E.P., and Jenkins, G.M. (1970). Time Series Analysis; Forecasting and Control. San Francisco: Holden-Day.

Brillinger, D.R. (1975). Time Series; Data Analysis and Theory. New York: Holt, Rhinehart and Wilson.

Carroll, R.J. (1979). Estimating variances of robust estimators when the errors are asymmetric. Jour. Amer. Statist. Assoc., 74, 674-679.

Chambers, J.M. (1977). Computational Methods for Data Analysis. New York; John Wiley.

Charnes, A., Cooper, W.W., and Ferguson, R.D. (1955). Optimal estimation of executive compensation by linear programming. Manag. Sci., 1, 138-151.

Chi-Leung, L. (1980). A simple series for the incomplete gamma integral. Appl. Statist., 29, 113-114.

Coleman, D., Holland, P. Kaden, N., Klema, V., and Peters,S.C. (1980). A system of subroutines for IRLS computations. A.C.M. Trans. on Math. Software, 6. 327-336.

Craddock, J.M. (1965). The analysis of meteorological time-series for use in forecasting. Statistician, 15, 167-190.

Cramér, H. (1946). Mathematical Methods of Statistics. Princeton,New Jersey; Princeton Univ. Press.

Denby, L., and Larsen, W. (1977). Robust regression estimators compared via Monte Carlo. Commun. Statist., A,6, 335-362.

Denby, L., and Mallows, C.L. (1977). Two diagnostic displays for robust regression analysis. Technometrics, 19, 1-13.

Dutter, R. (1977). Numerical solution of robust regression problems; computational aspects, a comparison. Jour. Statist. Comput. Simul, 5, 207-238.

Efron, B. (1979). Bootstrap methods; another look at the jack-knife. Annals of Statistics, 7, 1-26.

Fox, L. (1964). An Introduction to Numerical Linear Algebra. London; Oxford Univ. Press.

Gentle, J.E. (1977). Least absolute values estimation: an introduction. Commun. Statist., B,6, 313-328.

Gentle, J.E., Kennedy, W.J., and Sposito, V.A. (1977a). On least absolute values estimation. Commun. Statist., A,6, 839-845.

Gentle, J.E., Kennedy, W.J., and Sposito, V.A. (1977b). On properties of ℓ_1-estimators. Math. Programming, 12, 139-140.

Gilsinn, J., Hoffman, K., Jackson, R.H.F., Leyendecker, E., Saunders, P., and Sheer, D. (1977). Methodology and analysis for comparing discrete linear ℓ_1-approximation codes. Commun. Statist., B,6, 399-413.

Hampel, F. (1974). The influence curve and its role in robust estimation. Jour. Amer. Statist. Assoc., 69 383-393.

Harter, H.L. (1975). The method of least-squares and some alternatives: Part V. Int. Statist. Rev., 43, 269-278.

Harvey, A.C. (1977). A comparison of preliminary estimators for robust regression. Jour. Amer. Statist. Assoc., 72, 910-913.

Hill, R.W. and Holland, P.W. (1977). Two robust alternatives to least-squares. Jour. Amer. Statist. Assoc., 72. 828-833.

Hinich, M.J., and Talwar, P.P. (1975). A simple method for robust regression. Jour. Amer. Statist. Assoc., 70 113-119.

Hogg, R.V. (1974). Adaptive robust procedures: a partial review and some suggestions for future applications and theory. Jour. Amer. Statist. Assoc., 69, 909-923.

Hogg, R.V. (1977). An introduction to robust procedures. Commun. Statist., A,6, 789-794.

Hogg, R.V. (1979). An introduction to robust estimation. In Launer, R.L., and Wilkinson, W.N., Eds., "Robustness in Statistics," New York; Academic Press.

Holland, P.W. and Welsch, R.E. (1977). Robust regression using iteratively weighted least-squares. Commun. Statist., A,6 813-827.

Huber, P.J. (1964). Robust estimation of a location parameter. Ann. Math. Statist., 35, 73-101.

Huber, P.J. (1972). Robust statistics; a review. Ann. Math. Statist., 43, 1041-1067.

Huber, P.J. (1973). Robust regression; asymptotic conjectures and Monte Carlo. Ann. Statist., 1, 799-821

Huber, P.J. (1977). Robust Statistical Procedures. Philadelphia Society for Industrial and Applied Mathematics.

Huber, P.J. (1981). Robust Statistics. New York; John Wiley.

IBM System/360 Operating System. Introduction (IBM Systems Reference Library) Pub. no. GC28-6534-4. San José, California; IBM Corp.

Kennedy, W.J., and Gentle, J.E. (1977). Examining rounding error in LAV regression computations. Commun. Statist., B,6, 415-420.

Kennedy, W.J., and Gentle, J.E. (1980). Statistical Computing. New York; Marcel Dekker Inc.

Kiountouzis, E.A. (1973). Linear programming techniques in regression analysis. Appl. Statist., 20, 69-73.

Kleiner, B., Martin, R.D., and Thomson, D.J. (1979). Robust estimation of power spectra. Jour. Roy. Statist. Soc., B,41, 313-351 (with discussion).

Koopmans, L.H. (1974). The Spectral Analysis of Time Series New York; Academic Press.

Mc Cormick, G.F., and Sposito, V.A. (1976). Using the ℓ_2-estimator in ℓ_1-estimation. S.I.A.M. Jour. Numer. Anal., 13, 337-343.

Morrison, D.F. (1976). Multivariate Statistical Methods. 2nd ed., New York: McGraw Hill Book Co..

Newland, J. (1979). Job class submission patterns: a principal components analysis I. Research report 79/27, Dept. of Computing and Control, Imperial College, London.

Pfaffenberger, R.C., and Dinkel, J.J. (1978). Absolute deviations curve-fitting: an alternative to least-squares. In David, H.A., Ed., "Contributions to Survey Sampling and Applied Statistics", New York; Academic Press

Ramsey, J.O. (1977). A comparative study of several robust estimates of slope, intercept and scale in linear regression. Jour. Amer. Statist. Assoc., 72,608-615.

Roodman, G.M. (1974). A procedure for optimal step-wise MSAE regression analysis. Operat. Res., 22, 393-399.

Rosenberg, B., and Carlson, D. (1977). A simple approximation to the sampling distribution of least absolute residuals regression estimates. Commun. Statist., B,6, 421-437.

Rosenblatt, M. (1971). Curve estimates. Annals. Math. Statist., 42, 1815-1842.

Sadovski, A.N. (1974). L_1-norm fit to a straight line: algorithm AS 74. Appl. Statist., 23, 244-248.

Schlossmacher, E.J. (1973). An iterative technique for absolute deviations curve-fitting. Jour. Amer. Statist. Assoc., 68, 857-859.

Schrader, R.M., and Hettmansperger, T.D. (1980). Robust analysis of variance based on a likelihood ratio criterion. Biometrika,67, 93-101.

Seber, G.A.F. (1977). Linear Regression Analysis. New York. John Wiley.

Sposito, V.A. (1976). Remark on algorithm AS 74: ℓ_1-norm fit to a straight line. Appl. Statist., 25, 96-97.

Sposito, V.A., Hand, M.L., and McCormick, G.F. (1977). Using an approximate ℓ_1-estimator. Appl. Statist., 25, 154-157.

Sposito, V.A., and Smith, W.C. (1976). On a sufficient and necessary condition for ℓ_1-estimation. Appl. Statist., 25, 154-157.

Spyroupoulos, K., Kiountouzis, E., and Young, A. (1973). Discrete approximations in the ℓ_1-norm. Computer Jour., 16, 180-186.

Tukey, J.W. (1960). A survey of sampling from contaminated distributions. In Olkin, I., Ed., "Contributions to Probability and Statistics,". Stanford, California; Stanford Univ. Press.

Wagner, H.M. (1959). Linear programming techniques for regression analysis. Jour. Amer. Statist. Assoc., 54, 206-212.

Wegman, E.J., and Carroll, R.J., (1977). A Monte Carlo study of robust estimators of location. Commun. Statist., A, 6. 795-812.

Wektz, W. (1978). Statistical Density Estimation; A Survey. Gottinger: Vandenhoeck and Ruprecht.

Wilkinson, J.M. (1965). The Algebraic Eigenvalue Problem. London: Oxford Univ. Press