THE APPLICATIONS OF REGRESSION ANALYSIS IN AUDITING AND COMPUTER SYSTEMS

THESIS

Presented to the Graduate Council of the North Texas State University in Partial Fulfillment of the Requirements

For the Degree of

MASTER OF SCIENCE

By

Larry D. Hubbard, B. S.

Denton, Texas

May, 1977

Hubbard, Larry D., The Applications of Regression Analysis in Auditing and Computer Systems. Master of Science (Computer Sciences), May, 1977, 88 pages, 5 tables, 2 illustrations, bibliography, 10 titles.

This thesis describes regression analysis and shows how it can be used in account auditing and in computer system performance analysis. The study first introduces regression analysis techniques and statistics. Then, the use of regression analysis in auditing to detect "out of line" accounts and to determine audit sample size is discussed. These applications led to the concept of using regression analysis to predict job completion times in a computer system. The feasibility of this application of regression analysis was tested by constructing a predictive model to estimate job completion times using a computer system simulator. The predictive model's performance for the various job streams simulated shows that job completion time prediction is a feasible application for regression analysis.

DAC

Burs

PREFACE

This investigation is concerned with the techniques of regression analysis and their applications in the fields of account auditing and computer systems. Chapter I presents a nonmathematical description of regression analysis techniques and of the statistics calculated by most computer programs that perform regression analysis. This chapter provides background information for the regression techniques discussed in the remainder of the thesis.

Chapter II describes two ways that regression analysis can be used in the field of auditing. These are the detection of "out of line" account balances and the determination of audit sample size. This chapter is included in the thesis to expound upon the techniques presented in Chapter I by describing some actual applications of regression analysis. The ideas presented here led to the concept of using regression analysis in computer systems as shown in the following chapter.

Chapter III describes how regression analysis might be used to predict the exit times of jobs running in a computer system. In this context, the "exit time" of a job refers to the time that the computer finishes processing the job. This use has practical application in that many computer users

request computer center personnel to make such a prediction. The chapter describes how the feasiblity of this regression analysis application was tested by the construction of a predictive model to estimate job exit time.

TABLE OF CONTENTS

Page

LIST OF TABLES iv

LIST OF ILLUSTRATIONS v

Chapter

I. INTRODUCTION TO REGRESSION ANALYSIS . . . 1

Types of Regression Analysis
Least Squares Method
Regression Statistics
Conclusion

II. REGRESSION ANALYSIS IN AUDITING 19

Ratio Analysis
Sample Size Determination
Conclusion

III. REGRESSION ANALYSIS IN COMPUTER SYSTEMS . . 32

Types of Mathematical Models
Computer System Model
Evolution of the Predictive Model
Conclusion

APPENDICES 60

BIBLIOGRAPHY 81

LIST OF TABLES

Table		Page
I.	DATA FOR EXAMPLE OF MULTIPLE LINEAR REGRESSION ANALYSIS	9
II.	F VALUES FROM A STEPWISE REGRESSION PROGRAM	15
III.	SAMPLE JOB STREAM	36
IV.	JOB TRACE FOR SAMPLE JOB STREAM	37
V.	SUMMARY OF SAMPLE JOB STREAM PROCESSING .	38

LIST OF ILLUSTRATIONS

Figure		Page
1.	Plot of Outside Diameter and Tensile Strength of Wire, and Regression Line	6
2.	Time Sequence Residuals Plot of Every Fourth Residual from Run Number Three	56

CHAPTER I

INTRODUCTION TO REGRESSION ANALYSIS

Regression analysis is a method of developing a mathematical equation to describe the relationships among a number of variables. This regression equation is formulated such that the value of one variable (called the dependent variable) can be estimated when the values of the other variables (called the independent, or predictor, variables) are known. Thus, the independent variables are used to estimate a value for the dependent variable. Regression analysis employs past period, or historical, data to build this regression equation. The equation is "built" by solving mathematical formulas involving the past period data. In using the equation, it is assumed that future data will act in much the same manner as the historical data.

According to Mason (2, p. 485), the word "regression" was introduced by Sir Francis Galton in 1877 during his study of heredity. He found that the heights of descendants of tall parents tended to regress (meaning to go back) toward the average height of the population. He developed a mathematical line, called the line of regression, to describe this tendency. The term "line of regression" is commonly used but, according to Mason, a "predictive equation," or

an "estimating equation," seems to be more appropriate (2, p. 485). The notion of regression analysis has not changed since Galton's time. It still means to develop a mathematical line that describes the tendency of one variable to regress toward another.

The variables mentioned may come from many different areas. In financial applications, selling expense may be estimated using the number of invoices processed, net sales dollars, number of salespeople, and their average hourly wage as predictor variables. In medical applications, the weight of a person's liver may be predicted based on his body weight, his height, and his age. Any number of other examples could be given. The point is that the variables in a regression analysis application are simply numbers, and it is the user's duty to assign meaning to them.

Types of Regression Analysis

There are four general types of regression analysis: (1) simple linear regression analysis; (2) multiple linear regression analysis; (3) simple nonlinear regression analysis; (4) multiple nonlinear regression analysis. The factors which distinguish the types from one another are the number of predictor variables in the equation and the power to which the predictor variables are raised. In the "simple" cases, a single predictor variable is used, and in the "multiple" cases, numerous predictor variables are used.

Furthermore, the predictor variables are all raised to the first power in the "linear" cases, and raised to a power greater than one in the "nonlinear" cases.

In all four types of regression analysis, the regressing equation is found by "regressing" upon the historical data to determine specific values to be used as the coefficients of the predictor variables in the equation. By calling X_1, X_2, X_3, . . ., X_k the predictor variables, Y_p the estimate of the dependent variable, and b_0, b_1, b_2, . . ., b_k the regression coefficients, general equations for the four types of regression equations can be given as follows:

a. Simple linear regression analysis:

$$Y_p = b_0 + b_1 \cdot X$$

b. Multiple linear regression analysis:

$$Y_p = b_0 + b_1 \cdot X_1 + b_2 \cdot X_2 + \ . \ . \ . \ + b_k \cdot X_k$$

c. Simple nonlinear regression analysis:

$$Y_p = b_0 + b_1 \cdot X_1^n$$

d. Multiple nonlinear regression analysis:

$$Y_p = b_0 + b_1 \cdot X^1 + b_2 \cdot X_2^m + \ . \ . \ . \ + b_k \cdot X_k^n$$

The two linear methods of regression analysis will be most important in the discussion that follows.

Prior to the introduction of the electronic computer, regression analysis was limited to about three independent variables because of the large number of calculations necessary to find the regression equation. Presently, most regression analysis programs available on large computers

will accept over twenty independent variables (2, p. 514). In the cases of nonlinear regression analysis, computational complexities are so immense that there are serious difficulties in solving them even with computers (3, p. 439).

Least Squares Method

The most popular way of finding the regression equation is called the least squares method. This method gives what is commonly referred to as the "best fitting" straight line based on the given historical data (2, p. 485). The best fit line for any set of data points depends upon how the user states his best fit criteria. In some applications the best fitting line may pass through each historical data point, while in other cases, such as with least squares, the best fitting line need not pass exactly through any of the points. In regression analysis the least squares method is said to produce the best fitting straight line because it minimizes the sum of the squares of the vertical deviations about the line. This least squares concept of best fit will be used in this paper.

In the simple linear regression case, this "best fit" line can be easily demonstrated, but the multiple linear regression case is more difficult to picture. The simple linear case will consist of one dependent and one independent variable, and the historical data will be given as two

sets of numbers. These sets of numbers can be thought of as x,y pairs in a cartesian plane. Then, the object of regression analysis is to find the equation of the straight line which comes closest to going through all the given points. In least squares, this means minimizing the sum of the squares of the vertical deviations about the line. As an illustration, assume the outside diameters and tensile strengths of three pieces of wire are measured, and the outside diameters are .3 inches, .4 inches, and .5 inches, and the tensile strengths are 8,000 pounds, 18,000 pounds, and 16,000 pounds. The results of plotting this information along with the least squares line are shown in Figure 1. In this example, the sum of the squares of the vertical deviation from the regression line can be calculated as $2.0^2+4.0^2+2.0^2=24.0$. Since this is a small example with few data points, it would not be difficult to verify that twenty-four is indeed the smallest possible sum of squared deviations. By drawing some other line to represent the three points, say by a freehand method, the sum of the squared deviations would be greater than twenty-four. The formulas used to find this "best fitting" line by the least squares method will be discussed later.

The meaning of the least squares method has been discussed with regard to the simple linear regression analysis

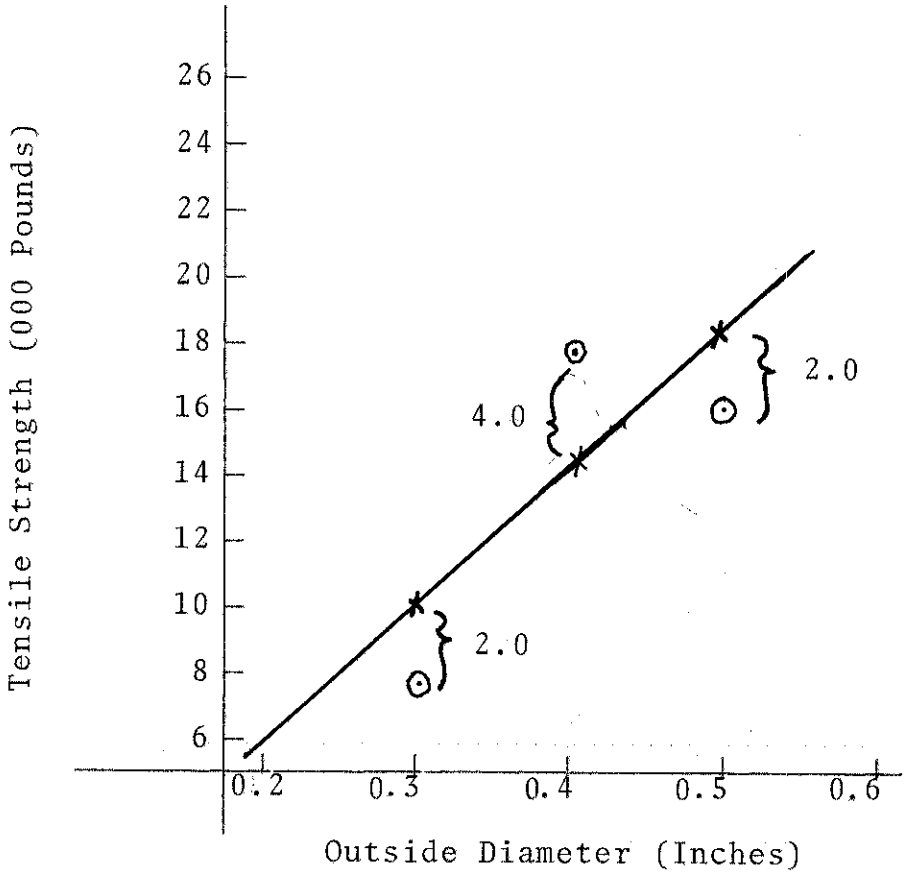

Figure 1--Plot of Outside Diameter and Tensile Strength of Wire, and Regression Line

case. For the multiple linear regression analysis case, the theory is the same, but graphical representation of the least squares line is very difficult. The object is still to minimize the sum of the squares of the vertical deviations about the regression line, but the multiple case involves one dependent variable and more than one independent variable.

Regression Analysis Formulas

The formulas for finding the least squares line in both the simple and multiple linear regression analysis cases will now be discussed. Because this is a nonmathematical discussion, proofs of the methods will not be given. These proofs are described in many statistics and econometrics texts, and involve using the calculus of partial derivatives to minimize certain mathematical equations. In the simple linear case, the values of the regression coefficients, b_0 and b_1, are found by using the following formulas:

$$b_1 = \frac{n(\Sigma XY) - (\Sigma X)(\Sigma Y)}{n(\Sigma X^2) - (\Sigma X)^2}$$

$$b_0 = \frac{\Sigma Y}{n} - b_1 \cdot \frac{\Sigma X}{n} \quad .$$

In these formulas, n is the number of historical data cases we have, and X and Y represent the actual historical data. (Recall that the general form of a simple linear regression equation is $Y_p = b_0 + b_1 \cdot X$, where Y_p is a prediction of the value of the dependent variable, Y.) Using the data for tensile strengths of wires as shown previously, the above calculations can be demonstrated.

$$b_1 = \frac{n(\Sigma XY) - (\Sigma X)(\Sigma Y)}{n(\Sigma X^2) - (\Sigma X)^2} \qquad b_0 = \frac{\Sigma Y}{n} - b_1 \cdot \frac{\Sigma X}{n}$$

$$= \frac{3(17.6) - (1.2)(42)}{(3)(.5) - (1.2)^2} \qquad = \frac{42}{3} - \frac{40 \cdot 1.2}{3}$$

$$= \frac{2.4}{.06} \qquad = 14 - 16$$

$$= 40 \qquad = -2$$

Thus, the regression equation is $Y_p = -2 + 40 \cdot X$. The graph of this line was shown in Figure 1, and can easily be verified by substituting the values three, four, and five for X in the equation. These formulas may be used to find the regression coefficients in any simple linear regression analysis problem that uses the least squares method.

Finding the regression coefficients in multiple linear regression analysis is a little more difficult. The least squares estimates of the coefficients in the multiple linear regression case are given by $a = (X' \cdot X)^{-1} \cdot X' \cdot Y$ (1, p. 52). Here a is a vector of the estimates of the regression coefficients, X is a matrix containing the historical observations and X' is its transpose, and Y is a vector containing the historical observations of the dependent variable Y. The symbol -1 above (X'X) indicates the inverse of this matrix.

An example of how the above formula can be used will now be given. Assume the set of values shown in Table I is given and the sums and sums of cross products for the data have been calculated. The object is to find a regression equation of the form $Y_p = b_0 \cdot X_0 + b_1 \cdot X_1 + b_2 \cdot X_2$ that describes the relationships between the dependent variable, Y, and the independent variables, X_1 and X_2.

TABLE I

DATA FOR EXAMPLE OF MULTIPLE LINEAR REGRESSION ANALYSIS

Variables		
Y	X_1	X_2
66.0	38.0	47.5
43.0	41.0	21.3
36.0	34.0	36.5
23.0	35.0	18.0
22.0	31.0	29.5
14.0	34.0	14.2
12.0	29.0	21.0
7.6	32.0	10.0
Sums and Cross Products		
$\Sigma Y = 223.6$	$\Sigma X_1 = 274.0$	$\Sigma X_2 = 198.0$
$\Sigma Y^2 = 8911.8$	$\Sigma X_1^2 = 9488.0$	$\Sigma X_2^2 = 5979.1$
$\Sigma X_1 \cdot Y = 8049.2$	$\Sigma X_2 \cdot Y = 6954.7$	$\Sigma X_1 \cdot X_2 = 6875.6$

(To facilitate matrix representation, X_0 is a dummy variable whose value is always unity.) The process begins by substituting the given data into the formula for finding the estimates of the coefficients in the regression equation.

$$a = (X' \cdot X)^{-1} \cdot X' \cdot Y$$

$$\begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \left\{ \begin{pmatrix} 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 \\ 38.0 & 41.0 & 34.0 & 35.0 & 31.0 & 34.0 & 29.0 & 32.0 \\ 47.5 & 21.3 & 36.5 & 18.0 & 29.5 & 14.2 & 21.0 & 10.0 \end{pmatrix} \cdot \begin{pmatrix} 1.0 & 38.0 & 47.5 \\ 1.0 & 41.0 & 21.3 \\ 1.0 & 34.0 & 36.5 \\ 1.0 & 35.0 & 18.0 \\ 1.0 & 31.0 & 29.5 \\ 1.0 & 34.0 & 14.2 \\ 1.0 & 29.0 & 21.0 \\ 1.0 & 32.0 & 10.0 \end{pmatrix} \right\}^{-1}$$

$$\cdot \begin{pmatrix} 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 \\ 38.0 & 41.0 & 34.0 & 35.0 & 31.0 & 34.0 & 29.0 & 32.0 \\ 47.5 & 21.3 & 36.5 & 18.0 & 29.5 & 14.2 & 21.0 & 10.0 \end{pmatrix} \cdot \begin{pmatrix} 66.0 \\ 43.0 \\ 36.0 \\ 23.0 \\ 22.0 \\ 14.0 \\ 12.0 \\ 7.6 \end{pmatrix}$$

The multiplication of these matrices yields:

$$\begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 8.0 & 274.0 & 198.0 \\ 274.0 & 9488.0 & 6875.6 \\ 198.0 & 6875.6 & 5979.1 \end{pmatrix}^{-1} \cdot \begin{pmatrix} 223.6 \\ 8049.2 \\ 6954.7 \end{pmatrix}$$

Continuing, any method available can be used to find the inverse of the above 3 by 3 matrix. When this is done, and the multiplication performed, the result is:

$$\begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} -94.6 \\ 2.8 \\ 1.1 \end{pmatrix}$$

Thus, the regression equation is given by:

$$Y_p = -94.6 + 2.8 \cdot X_1 + 1.1 \cdot X_2$$

The preceding method can be used to find the regression equation for any linear multiple regression problem. However, because of the many calculations involved, this method is impractical for problems using more than three independent variables.

When these matrix calculations are performed by a computer, they are not carried out using exactly the method shown above. One reason for this is the large rounding errors that may occur when this sequence is followed (1, p. 107). Rather than spend time discussing how electronic computers solve regression analysis problems, it is better to assume that packaged programs by computer manufacturers such as IBM are able to do so. With this assumption, the interpretation of certain statistics vital to regression analysis will be discussed.

Regression Statistics

Earlier it was stated that the least squares method yields the "best fit" regression line. The least squares method of linear regression analysis attempts to find a straight line that best describes the historical data that is given. Some sets of data values can be described very well by a straight line, but others cannot. There are

statistics that indicate how well this "best fit" straight line describes a particular set of data values. There are three such statistics that are of prime importance in this paper. These are: (1) the Pearson product moment correlation coefficient, (2) the standard error of the estimate, and (3) the F value of the equation. Each of these statistics will now be discussed.

The Pearson product moment correlation coefficient, symbolized by R, is a measure of the relationship between the dependent and independent variables. This measure is usually squared, R^2, to take on a value from zero to one proportional to the goodness with which the dependent variable, Y, can be predicted from a knowledge of the independent variables, X. For example, if $R^2=.92$, then 92 percent of the variation in Y is explained by X. If $R^2=.20$, then only 20 percent of the variation in Y is explained by X, and the remaining 80 percent is unexplained. This correlation coefficient, R^2, is printed out for most regression analysis problems solved by a computer. An examination of this statistic shows how well the regression equation fits the historical data.

The standard error of the estimate is a measure similar to the standard deviation normally encountered in statistics. The standard error is symbolized by SE and measures the dispersion of points about the regression line. As an example,

assume a regression equation that predicts amount of sales (in millions of dollars) based on several predictor variables has been developed, and that the computer run shows the standard error of the estimate to be 1.24. When the equation is used to predict sales, Y_p, the following statements can be made:

a. The probability is .68 that the sales are in the range $Y_p \pm \$1,240,000$.

b. The probability is .95 that the sales are in the range $Y_p \pm 1.96 \cdot (\$1,240,000)$.

c. The probability is .997 that the sales are in the range $Y_p \pm 3 \cdot (\$1,240,000)$.

These probabilities are based on the characteristics of the normal curve as described in many statistics books.

Another important use of the standard error of the estimate is apparent when computers use a "stepwise" method of finding a multiple linear regression equation. Most popular computer programs use this method, and it simply means that the predictor variables, X_1, X_2, . . . X_k, are entered into the regression equation one at a time. As each variable is entered, a regression "step" is completed. The computer program determines which variable to enter at each step by computing the correlation between each of the independent variables and the dependent variable. The

independent variables showing the highest correlation with the dependent variable are entered into the regression first. A matrix of these correlations is shown at the first of most stepwise multiple regression programs.

A regression equation and statistics about the regression are printed out at each step described above. The equation at each step includes all the predictor variables that have entered the regression thus far. This way, a user may decide at any step that the regression equation is "good enough" and obtain an equation that contains only the variables entered so far. The standard error can be used to tell when the regression equation is "good enough," because it indicates whether the estimation is getting better or worse. Since a small standard error is desirable, a predictor variable should be used in the regression equation only if the step in which it enters shows a smaller standard error than was shown at the previous step.

The F value of the equation is the final measure to be discussed. This is the ratio of the explained variance to the residual or unexplained variance. This ratio can be used in much the same way as the standard error of the estimate in a stepwise regression problem. If a variable entered into the regression at a step adds to the explained variance, that is, increases the F value, then it should be used in

the regression equation. For example, assume a stepwise regression analysis program has calculated the F values for each step in the regression as shown in Table II.

TABLE II

F VALUES FROM A STEPWISE REGRESSION PROGRAM

Step	Variables Used (X_k)	F Value Calculated
1	4	106.630
2	4,2	207.485
3	4,2,5	345.369
4	4,2,5,1	217.546
5	4,2,5,1,3	140.404

The predictor variable X_1 and X_3 would probably not be used in the final regression equation. This is because when they were entered, at step four and step five, they decreased rather than increased the F value, thus indicating that these two variables decreased the explained variance at their respective steps.

Another way the F value can be used is in testing the "statistical significance" of the regression (1, p. 64). Stating that a regression is "statistically significant," means that the portion of variance observed in the data,

and accounted for by the regression equation, is greater than would be expected by chance in $100 \cdot (1-\alpha)$ percent of the similar sets of data with the same number of observations and predictor variables. Here, α is a risk level specified by the user of the regression equation.

To further explain this statement, another statistic normally appearing on a computer printout of a regression run must be mentioned. This is the "degrees of freedom" (d.f.) in the regression and in the residual of the equation. These are two numbers determined during the computer run that relate to how many sets of historical data were used in a regression problem. For the purposes of this paper, it will be best to use these numbers without formally explaining their origin. (For further information, the reader may consult any introductory statistics text.) Assuming this, an example of testing the statistical significance of a regression equation can be given. Suppose a risk level, α , of .05 is chosen and that the degrees of freedom (d.f.) calculated in the regression is 10 and in the residual is 20. By consulting a statistical table giving F Distributions, it can be found that $F(10,20,.95) = 2.35$. This means that the F value calculated from the regression equation must exceed 2.35 in order for the regression to be considered statistically significant. In other words, the F value will exceed 2.35 if the regression equation does a better job of explaining the variances in

the data than could be done by mere chance. Studies have indicated that the F value should not merely exceed the selected percentage point of the F-distribution, but should be four times the selected percentage point (1, p. 64). Thus, the F value in this example would have to exceed 9.4 for the regression equation to be considered a better prediction tool than mere chance.

Conclusion

Regression analysis is a mathematical tool used to build an equation that describes the relationship among several variables. This equation is based upon the historical values of the dependent and independent variables in the equation. The least squares method is the most popular way to find the regression coefficients to be used in the equation. There are four types of regression analysis, and the two types most important to this paper are simple linear regression analysis and multiple linear regression analysis. These two types differ only in the number of independent, or predictor, variables used. Almost all regression analysis problems are solved by a computer, and most computer programs use a stepwise method to determine the regression equation. This enables the user to observe the values of certain statistics, such as the correlation coefficient, the standard error of the estimate, and the F value of the equation at each step and thus decide when the regression equation is satisfactory for his application.

CHAPTER BIBLIOGRAPHY

1. Draper, N. R. and H. Smith, Applied Regression Analysis, New York, John Wiley & Sons, Inc., 1967.

2. Mason, Robert D., Statistical Techniques in Business and Economics, Illinois, Richard D. Irwin, Inc., 1974.

3. Sterling, Theodor D. and Seymour V. Pollack, Introduction to Statistical Data Processing, New Jersey, Prentice-Hall, Inc., 1968.

CHAPTER II

REGRESSION ANALYSIS IN AUDITING

This chapter outlines how regression analysis has been applied in the field of auditing. More specifically, it shows how this tool has been used in auditing to predict sample size and to detect out of line accounts. These applications have special significance to this thesis in that they were the stimuli for the application of regression analysis in computer systems, as shown in Chapter III.

Business managers today retain more financial information about the activities of their companies than ever before. The need for keeping this large amount of information has always existed, but only recently has technology provided cost effective methods of doing this. As this retained information increases in amount, the independent auditor faces an increasingly difficult task when forming his opinion regarding the financial statements of a business. Auditors characteristically express this opinion after examining only a small portion of the underlying financial data. Often the auditor relies only on his informed judgement as to which specific data he should review. Two procedures that have been helpful in this selection process are statistical sampling and ratio analysis.

Statistical sampling allows the auditor to review a small number of transactions (called the sample size) and, based on the results of this examination, make a statement regarding all similar transactions (called the universe or population). This technique is used most notably in the examination of inventory and accounts receivable. Ratio analysis is performed by calculating percentages and ratios between account balances, and investigating those showing significant deviations from those of past periods. In this way, the auditor can direct more audit effort toward "out-of-line" accounts. Both of these procedures have disadvantages, however. In statistical sampling, the auditor usually has no quantifiable guidance as to whether he should increase or decrease the level of confidence (and thus the sample size) he specifies for his tests. Ratio analysis is not based on an understanding of the behavior of the individual accounts, and therefore can often be misleading.

Multiple linear regression analysis can be used along with statistical sampling and ratio analysis to make these procedures more effective. The use of regression analysis to detect "out-of-line" accounts will now be discussed. Following that, the use of regression analysis in conjunction with statistical sampling will be described.

Ratio Analysis

The first step in this application of regression analysis is the selection of independent variables (the X values) that can be used in predicting the value of the dependent variable (Y). For this application, the dependent variable represents the account balance the auditor wishes to test. The predictor variables chosen would normally be based, at least initially, upon the auditor's judgment. They could be derived from many different areas: internal or company data, industry statistics, or general economic indications. In order for the predictor variables to be used in cost related analysis, several conditions must be met. Some of these conditions will be mentioned here, but a more detailed list can be found in the works of Benston (1), Comiskey (2), and Jensen (6).

Because it is unlikely that the data the auditor is able to gather will meet all these requirements, the results of regression analysis can seldom be viewed as anything more than approximations. Violations of these formal requirements will usually have only minor impact on the results. Usually, the auditor's purpose will be adequately served if his data is in reasonable conformity with these requirements (4, p. 765).

As a general rule, the more historical observations the auditor can gather, the better his results will be.

Thus, it is helpful if the auditor can obtain data on a monthly or quarterly basis (3, p. 30). There should be uniformity in the method of recording this periodic information. This is an extension of the "matching" principle of accrual accounting: the revenues of a period should be matched with the expenses of that period. Because regression analysis uses historical, or past period, values to predict future amounts, it is important that no changes in accounting policy have occurred that may nullify this assumption. Finally, the auditor should attempt to find all major contributors to a particular account balance, and use all of these, at least initially, as predictors in his regression model.

The next step in the auditor's use of regression analysis is to determine the regression equation as based on the historical data. For this step it is necessary for the auditor to have access to a computer having a regression analysis program. Most computer manufacturers and service bureaus make these programs available to their users. As stated earlier, the stepwise regression method is used by most packaged programs and is the recommended variable selection procedure (5, p. 172). Using the stepwise method, the auditor can eliminate any predictor variables that he feels are unnecessary to the application, and thus reduce the complexity of the regression equation.

Once the auditor has determined the regression equation, it can be used to estimate the balance of the account he wishes to examine. This equation will have the form of the multiple linear regression equation shown earlier: $Y_p = b_0+b_1 \cdot X_1+b_2 \cdot X_2+ \ . \ . \ .+b_k \cdot X_k$. By evaluating this equation using the X values from the period being audited, the estimate, Y_p, of the current account balance can be found.

The final step the auditor must perform in his use of regression analysis is the interpretation of the estimated account balance, Y_p. As earlier stated, each regression equation has a standard error (SE) associated with it. Using this measure and the prediction, Y_p, several statements about the true account balance can be made. Based on the definition of standard error, sixty-eight percent of the time the true balance should be within a range of plus or minus one standard error of the predicted balance. Likewise, ninety percent of the time, it should be within a range of plus or minus 1.64 standard errors. For example, assume a regression model to predict shipping expense is built, and has a standard error of $1530. If the regression equation predicts shipping expense to be $11,640, then sixty-eight percent of the time the true balance should be within $11,640 plus or minus $1530. If the book value reported by the client is less than $10,110 or greater than $13,170, then the auditor would logically expect the account balance per the books to be in error.

Another way to interpret the results of using regression analysis in this manner involves calculating the difference between the book value (Y) and the predicted value (Y_p). This difference is called a "disturbance" (U_t), and over several time periods such disturbances should be normally distributed (3, p. 30). Thus, tests for the significance of a particular disturbance can be based on the characteristics of a normal curve.

The ratio between the disturbance and the standard error of the regression equation is known as a "z-value," thus: $z=U_t/SE$(3, p. 30). The significance of a computed z-value can be determined by referring to a table that gives the area under a normal curve. Because the disturbance is from a normal distribution, it can be shown that only five percent of the z-values are greater than 1.64 and only five percent are less than -1.64. Thus, a computed z-value of 1.64 implies the probability is five percent that the difference in predicted and book value can be attributed to random occurrences. In our previous example, Y_p was \$11,640 with a standard error of \$1530. If the book value is found to be \$11,990, the disturbance is Y minus Y_p, or \$350. The z-value for this disturbance is U_t/SE, or .23. Referring to a table giving the area under a normal curve, a z-value of .23 indicates the probability is eighty-two percent that the difference between predicted value and book value can be attributed to random occurrences.

In order to make use of this computed z-value, the auditor must establish a decision rule requiring him to investigate all accounts in which the disturbance has less than a certain percentage probability of resulting from random occurrences. This percentage probability is referred to as the auditor's "alpha level." If the auditor establishes his alpha level at ten percent, then he will investigate all accounts having a z-value of more than 1.64 or less than -1.64. With this cutoff value, there is only a ten percent probability of obtaining a z-value of greater than 1.64 or less than -1.64 if the book balances are correct. Likewise, for an alpha level of thirty-two percent, the auditor will investigate accounts whose z-value is more than 1 or less than -1.

If the auditor sets his alpha level at ten percent, then ten percent of the time he will investigate disturbances only to find that the account is stated correctly. On the other hand, at a five percent alpha level, nonproductive investigations drop to five percent, but the possibility of not investigating an incorrect balance is increased. Thus, to minimize his risk of failing to investigate accounts that are incorrect, the auditor must increase his risk of investigating accounts that turn out to be correct. This problem is similar to those encountered

in almost all statistical sampling decisions. Criteria for establishing decision rules such as this have been much discussed in the statistical sampling literature (3, p. 31).

Sample Size Determination

Another way that regression analysis can be used in auditing is in the determination of sample size. When regression analysis is used for auditing purposes, the detection of out of line accounts and sample size determination are closely related. Often, a regression equation will be constructed to find information about out of line accounts, and then this information will be used to determine sample size. In this way, a larger sample of items will be examined for accounts that are suspicious than from accounts that are in line.

The procedures necessary to determine sample size based on regression analysis can be rather complicated. Statistical hypothesis testing, statistical sampling techniques, and the use of Bayes' Formula to derive a set of posterior probabilities are all necessary for the complete application of this tool. The procedures will only be briefly outlined here. For further study, Deakin and Granof (4) have worked extensively in this area, as have Kinney and Bailey (7). These authors have cited excellent references and given examples of this application.

Once the auditor has formulated his regression equation, he can solve the equation and obtain a predicted value for an account. This value can be compared with the account balance as reported by the client. The auditor will establish two hypotheses: a null (H_0) and an alternative (H_α). The null hypothesis states that the true account balance is reflected by the client stated balance. The alternative hypothesis states that the client's balance differs significantly from the true account balance. Which of these two hypotheses the auditor accepts will depend upon his sampling program and the significance of the results of the regression analysis. In choosing a hypothesis, the auditor must weigh the costs of rejecting a client figure which is correct (α -risk) against that of accepting a client figure which is incorrect (β -risk). Deakin and Granof (4, p. 767) have formulated a decision rule for use in this decision. Thus far, the procedures described are like those used in the detection of out of line accounts.

The auditor now must test the significance of the variation in the predicted account balance, Y_p, and the client reported account balance, B_0. The procedure for this test is similar to the determination of a z-value as discussed earlier. The difference Y_p and B_0 can be evaluated from a table giving the area under the normal curve, using the standard error of the regression equation (SE). The value

that results from the calculation $z_0 = \frac{Y_p - B_0}{SE}$, can be converted to a probability using a table of areas under the normal curve. The amount thus determined is the probability that the observed z-value, z_0, came from the distribution stated by the null hypothesis. Or, the probability that the true account balance is reflected by the client's figure. The difference between Y_p and $B_0 \pm M$, the mean balance for the determination of accepting the alternative hypothesis, can also be evaluated. The value resulting from the calculation $z_\alpha = \frac{Y_p - (B_0 \pm M)}{SE}$, can be converted to a probability, as was z_0. This value is the probability of accepting a client figure which is incorrect.

The auditor can use these probabilities in either of two ways. First, he can use them to set heuristically his acceptable α- and β-risk levels for determination of sample size. Second, he can use the probabilities in a Bayesian sense to revise his prior probability estimates of a material misstatement of the account balance. The revised priors can then be used to find the conditional probabilities necessary for a given confidence level. If the auditor uses the first method above, he increases his investigation into accounts having high β-risks. Thus, he will increase sample size in high β-risk accounts and decrease sample size in low β-risk accounts.

If the auditor decides to use the Bayesian approach, he begins with the assumption that his prior probabilities of H_0 and H_α are equal. Then, the results of the regression equation evaluation will provide conditional probabilities for H_0 and H_α . These conditional probabilities, along with the prior probabilities, can be inserted in Bayes' Formula to derive a set of posterior probabilities. The auditor can use these posterior probabilities to select the appropriate α- and β -risk levels for his sampling plan. These values for α- and β -risks must permit the auditor to achieve a desired confidence level given his set of adjusted prior probabilities. Again Deakin and Granof (4, p. 768) have provided a decision table for use in this selection.

In their article, Deakin and Granof (4, pp. 768-770) present an example of using regression analysis to select sample size. Here, the auditor elects to evaluate cost of goods sold for a retail chain of four stores. By using certain predictor variables and historical data, a regression equation is constructed. Initially, the results of the regression analysis are not considered. By establishing equal prior probabilities of the accounts at each of the four stores being incorrect, the auditor determines his sample size by using a standard computational formula. This yields a total sample size of 543 items (151, 104, 241, and 47 for each of the four stores, respectively).

Next, the results of the regression analysis are used to revise the auditor's previously equal prior probabilities. By using the conditional probabilities and Bayes' Formula, the revised probabilities are found. This yields a total sample size of 364 items (173, 71, 91, and 29 for the four respective stores). Thus, by employing regression analysis, the auditor has reduced his total sample size by 33 percent. Also, the distribution of sampling effort was shifted. With equal prior probabilities, 27.8 percent of the items sampled were from Store One. After revising the sample selection, 47.8 percent of the sampling will be done at Store One. Likewise, revision of the prior probabilities resulted in reduction in sample size at Store Three from 44.4 percent to 25.0 percent. These changes in sample distribution indicate a higher probability of error at Store One than at Store Three.

Conclusion

This discussion has outlined how regression analysis can be used to identify out of line conditions and also how it can be used in the selection of audit sample size. These two procedures can be used either together or separately in an audit effort. Both applications are fairly new and have not been tested extensively in actual auditing practice. However, as the audit environment grows in complexity, tools such as regression analysis might be used to great advantage.

CHAPTER BIBLIOGRAPHY

1. Benston, George J., "Multiple Regression Analysis of Cost Behavior," The Accounting Review, XLI, No. 4 (October, 1966), 657-672.

2. Comiskey, Eugene E., "Cost Control by Regression Analysis, "The Accounting Review, XLI, No. 2 (April, 1966), 235-238.

3. Deakin, Edward R. and Michael H. Granof, "Directing Audit Effort Using Regression Analysis," The CPA Journal, XLVI, No. 2 (February, 1976), 29-33.

4. Deakin, Edward R. and Michael H. Granof, "Regression Analysis as a Means of Determining Audit Sample Size," The Accounting Review, XLIX, No. 4 (October, 1974), 764-771.

5. Draper, N. R. and H. Smith, Applied Regression Analysis, New York, John Wiley & Sons, Inc., 1967.

6. Jensen, Robert E., "A Multiple Regression Model for Cost Control - Assumptions and Limitations," The Accounting Review, XLII, No. 2 (April, 1967), 265-273.

7. Kinney, William R., Jr. and Andrew D. Bailey, Jr., "Regression Analysis as a Means of Determining Audit Sample Size: A Comment," The Accounting Review, LI, No. 2 (April, 1976), 396-401.

CHAPTER III

REGRESSION ANALYSIS IN COMPUTER SYSTEMS

The preceding chapter described how regression analysis can be used by the independent auditor in his work. This application suggested the idea of using regression analysis for another purpose, in the computer field. Programmers often ask computer center personnel to predict the time at which their job will have been completed, i.e., the time at which it will exit the machine. Traditionally, this prediction of exit time is based on intuition. Regression analysis might be used to produce quantifiable evidence for this prediction. Historical data of actual job run times could be used to produce a regression equation. By evaluating this equation when a job enters the computer system, a better exit time estimation might be possible. An investigation of the feasibility of such a regression model is described in this chapter. Draper and Smith (1, pp. 234-242) give a general outline for studies such as this, and also discuss several types of mathematical models that are important to this paper.

Types of Mathematical Models

Three main types of mathematical models are often used by scientists: (1) the functional model, (2) the control

model, and (3) the predictive model (1, pp. 234-235). The functional model is used when a true functional relationship that exists between a dependent and the independent variables is known. In practice, there are very few models that fit into this category. The second type of model, the control model, contains variables that are under the control of the experimenter. Usually this type of model requires a designed experiment using the controlled variables. Often in practice a controlled experiment is not feasible. Regression analysis techniques have made their greatest contribution in the construction of the third type of model, the predictive model. This type of model, though in some senses unrealistic, reproduces the main features of the behavior of the variable under study. The model is not ordinarily functional, and need not be useful for control purposes. The primary purposes of the predictive model are to provide guidelines for further experimentation, to pinpoint important variables, and to act as a variable screening device.

The mathematical model built in this thesis to estimate job exit time is a predictive model, and has the same purpose as the predictive model described above. It is not meant to be a functional model, but is designed to provide insight for further research and identify variables that are most important to the regression model. This predictive

model is designed to show the general behavior of a real situation. The historical data for this regression model was obtained by simulating a computer system. In this way, different job streams and levels of activity could be easily arranged to test the regression model. This computer system model will now be described.

Computer System Model

There is a tremendous number of different computer systems in existence, and to choose one typical system to model would be difficult. For the purposes of this study, a choice of this nature is not necessary. Since this is to be a predictive type model to estimate job exit time, a computer system model that shows the general behavior of a real system is sufficient. A major contributor to the behavior of a computer system is the memory management scheme it uses. This is especially true in a small, simple system such as the one simulated in this paper.

Memory Management in the Computer System Model

There exist many different methods of allocating computer memory to the jobs that run in a system. Madnick and Donovan (2, pp. 105-198) describe seven important memory management schemes in their book about operating systems. One of these schemes, relocatable partitioned allocation,

is used in the simulated computer system that was built for this thesis. This method was chosen because it is relatively easy to understand and to simulate.

This method allocates a partition (area) of computer memory to a job, the size of the partition being equal to the size of the job, and then relocates this partition as necessary to avoid memory fragmentation. (Fragmentation can be defined as the development of a large number of separate, unused areas of computer memory (2, p. 121). Although the total amount of free memory is large, this memory is not contiguous and therefore cannot be used by the system.) Instead of a detailed explanation of relocatable partitioned memory management, an example of how a job stream is handled by this scheme will be given.

In order to facilitate the explanation, several assumptions will be made. These are: (1) the jobs do no input or output, (2) the time necessary to relocate a partition is ignored, (3) a first-in first-out (FIFO) method of starting jobs is used, and (4) multiprogramming exists, so that if two or more jobs are in the system they all get equal CPU time. With these assumptions, only the arrival time, CPU time required, and core required for the jobs will affect the total time to process the jobs. The job stream for this example consists of the jobs shown in Table III. The Arrival Time given corresponds to the time a job becomes

available to enter the system. (This may not be equal to the time the Central Processing Unit (CPU) actually begins to process the job.) The CPU Time Required is the total amount of processing time a job requires, and the Memory Required is the total contiguous memory (core) needed by a job. It is assumed the computer system has 100,000 bytes (100K) of usable memory.

TABLE III

SAMPLE JOB STREAM

Job	Arrival Time	CPU Time Required (minutes)	Memory Required (000 bytes)
1	0.0	1.0	50
2	0.1	1.0	20
3	0.5	2.0	30
4	1.0	1.0	20

The job trace for this job stream is shown in Table IV. The symbols such as 2(1.0) indicate the job number (two) and the required CPU time remaining (one minute). Job One arrives at time zero and immediately gets 0.1 minute of CPU time. It does not get more time than this since Job Two arrives at time 0.1 and is put into memory. After Job Two enters the memory, the total occupied memory is 70K. These two jobs will split CPU time until one of two things happens:

TABLE IV

JOB TRACE FOR SAMPLE JOB STREAM

Elapsed Time (minutes)	CPU Time Given to Each Job (minutes)	Jobs Being Processed			
0.0	0.1	1(1.0)	2(1.0)		
0.1	0.2	1(0.9)	2(0.8)	3(2.0)	
0.5	0.7	1(0.7)	2(0.1)	3(1.3)	4(1.0)
2.6	0.1	1(0.0)	2(0.0)	3(1.2)	4(0.9)
2.9	0.9			3(0.3)	4(0.0)
4.7	0.3			3(0.0)	
5.0					

(1) another job arrives that will fit into core, or (2) either Job One or Job Two finishes processing. In this example, Job Three arrives at time 0.5, and will fit into core, so it is entered into the processing. Total occupied memory is now 100K, and therefore no more jobs can enter until one of these three jobs finishes processing. Job One has the least remaining CPU time required (0.7 minutes) and finishes at time 2.6. Since Job One occupied 50K of core, this amount of memory is freed when Job One finishes. Therefore, at this time Job Four can be put into the memory. No more jobs are available in the job stream, so Job Two, Job Three, and Job Four split CPU time until each job finishes.

The processing of this job stream is summarized in Table V. The Arrival Time, CPU Time Required, and Memory

Required for each job are given, along with the time processing was actually begun and finished for each job. This type of job trace can be used to simulate the processing of any job stream, given the assumptions previously mentioned.

TABLE V

SUMMARY OF SAMPLE JOB STREAM PROCESSING

Job	Arrival Time	CPU Time Required (minutes)	Memory Required (000 bytes)	Time Processing Began	Time Processing Finished
1	0.0	1.0	50	0.0	2.6
2	0.1	1.0	20	0.1	2.9
3	0.5	2.0	30	0.5	5.0
4	1.0	1.0	20	2.6	4.7

Programs in the Computer System Model

The computer system model used for this study consists of two FORTRAN IV computer programs. FORTRAN IV was used because it is a widely accepted language for simulations such as this, and because the International Business Machine (IBM) regression analysis program to be used later is written in FORTRAN IV. The two programs in the model are a job stream generator and a memory simulator. These will now be described.

Job stream generator.--This computer program is essentially a random number generator. It generates a stream of jobs to be processed by the memory simulator program. The only input to the generator is a number representing how many jobs are to be included in the job stream. The output from the generator consists of an arrival time, required CPU time, and required memory size for each job in the stream. This is the same type of information as shown in Table II, and as described in the section on memory management.

The job stream produced by the generator can be changed by adjusting parameters within the computer program. In this manner, different type job streams can be simulated to represent the varied activity levels and job mixes that often exist in a real computer system. This capability will be described later in this chapter. A sample listing of the job stream generator is shown in Appendix I.

Memory simulator.--This computer program processes the job stream generated by the previous program. It processes this job stream in the manner described in the section of this chapter discussing memory management; thus it is simply a program that constructs the job traces as shown in Table III. This way, it simulates the processing of any given job stream according to the assumptions discussed previously.

The amount of memory used in the model is 250,000 bytes. The output from the simulator is like that shown in Table IV, except that more statistics about the processing of the job stream are gathered. These statistics will be discussed later in this chapter. A sample listing of the memory simulator is shown in Appendix II.

Computer Processing of the Computer System Model

The computer system model is processed by the IBM 360/50 computer at North Texas State University (NTSU). The output from the model is used as historical data to be analyzed by a regression analysis program. The regression analysis program used is program number ST041 in the Statistical Library of the IBM 360 computer at NTSU. The computer system model and the regression analysis program are processed as a single job on the computer, thus making it possible to generate a job stream, simulate its processing, and analyze the results with regression analysis in a single computer run.

Evolution of the Predictive Model

The procedure for constructing the model to predict job exit time began with finding a regression equation to describe a simple, random job stream. After this, more complicated job streams, such as those with changes in activity level, were considered. This evolution process allowed many different data items to be tested and either accepted or

rejected as good predictor variables. By doing this, one of the important functions of a predictive model, the screening of variables, was accomplished. The result of this process was the identification of several of the variables most important in job exit time prediction. The final step in the development was to analyze the model and to draw conclusions as to the feasibility of an operational model of this type and as to the knowledge gained by the study. These conclusions should provide insight for further research on this subject.

Many runs of the computer system model and regression program were necessary in order to draw the conclusions mentioned above. The objective in each of these runs was to find the best variables for predicting where a job would exit. This was accomplished by carefully selecting possible predictor variables, and testing them with a specific stream of jobs. When a set of good predictor variables was found for a specific job stream, these same variables were used with a different job stream to see if they were still good predictors. The process of testing new predictor variables with different job streams continued until the important predictors of job exit time were found. Thus the predictive model evolved from the testing of many different variables and many different job streams.

The job streams used in the test ranged from a simple, random stream to a more complicated stream with varying levels of activity. In the early job streams, the job arrival times were random occurrences, while in the final streams, jobs arrived rapidly for awhile, and then slowly. This simulated the cyclic levels of activity that normally occur in a computer center. That is, during "busy" hours, jobs may arrive only seconds apart, while during "slow" hours, jobs may arrive many minutes apart. In real job streams, when a job finishes processing depends heavily upon how many other jobs are in the system with it. Thus, an important test of a model to predict job exit time is how well it works with these complicated job streams. For this reason, the final test of the predictive model as it evolved was its ability to perform well in cyclic job streams. As stated earlier, many runs of the computer system model and regression analysis program were necessary to develop the predictive model to estimate job exit times. Only three of these computer runs will be discussed in this chapter. These runs show how the model evolved as the job streams became more complex and different predictor variables were tested.

The explanation of each run will include a discussion of the job stream used, the predictor variables used, the results of the run, and the conclusions drawn from the run. To facilitate these discussions, variable names used in

the computer system model and in the regression program will be used in the text of the discussion. The following variable names are used:

a. ARRT is the time that a job arrives at the computer system to be run.

b. CPUT is the amount of Central Processing Unit (CPU) time that a job requires.

c. CORE is the amount of computer memory that a job requires.

d. PREVJB is the number of jobs arriving during a predetermined time interval before the arrival of a given job. (This is explained more fully in the discussion of Run Number Two.)

e. CLASS is a number from one to four used instead of CPUT to indicate the amount of CPU time a job requires.

f. SE is the standard error of the regression equation, as described in Chapter I.

g. R^2 is the correlation coefficient of the regression equation, as described in Chapter I.

h. Y_p is the job exit time as predicted by the regression equation, as described in Chapter I.

i. RN is a random number used to determine characteristics of the job stream.

Run Number One

The first run that will be described occurred early in the evolution of the model and tested the most basic predictor variables for importance.

Job stream used.--A job stream with random arrival times is used in this run of the system. This is the simplest type of stream used in any of the runs, and probably is not typical in any real computer center. However, this job stream serves an important step in the evolution of the predictive model. The job stream consists of twenty jobs arriving randomly and having random CPU and CORE requirements. The arrival times, ARRT, begin at zero for the first job in the stream, and increase by a random number (RN) from the interval [0.1,0.9] for each following job. The arrival times range from zero for the first job to 8.3 for the last. The job exit times as calculated by the computer system model range from 23.03 for the first job to 103.83 for the last. The CPU time required, CPUT, for each job is a random number from the interval [1,9], and the memory requirement, CORE, is a random number from the interval [1,99]. These characteristics of the job stream can be described by a shorter notation as follows:

a. ARRT=ARRT+RN, where RN $\in$ [0.1,0.9].

b. CPUT=RN, where RN $\in$ [1,9].

c. CORE=RN, where RN $\in$ [1,99].

This notation will be used in describing the characteristics of future job streams.

Predictor variables tested.--The variables being tested for their importance in predicting job exit time are the arrival time, CPU requirement, and CORE requirements of the jobs. These are the most basic variables in the system and are used by the computer system model to generate the historical data used in the regression program. Thus, it may seem odd to test these variables for importance. But recall that the object of the regression analysis is to produce an equation to predict job exit times, and the object of the computer system model is to calculate the actual exit times for the jobs in the stream. The object of testing the variables ARRT, CPUT, and CORE here is to determine their usefulness in the predictive model, not in the computer system model.

Results of the run.--The output from the regression analysis program for this run is shown in Appendix III. Since three independent or predictor variables are being tested, the stepwise regression program has three steps. During each step, one variable is entered into the regression. In this program, the variables numbered one, two, and three refer to ARRT, CPUT, and CORE, respectively. As each variable enters the regression, many statistics about the regression equation are calculated. The three most

important statistics were discussed in Chapter I and are the correlation coefficient, R^2, the standard error, SE, and the F value for the equation. By observing the changes in these values from step to step, the best regression equation for the given historical data and predictor variables can be determined. A variable that improves the regression equation causes R^2 to increase, SE to decrease, and the F value to increase. This run shows that the variables ARRT and CPUT improve the equation when entered, but that CORE does not improve the equation. Thus, the best regression equation for this run includes only the variables ARRT and CPUT, and is $Y_p = -2.7+11.3 \cdot ARRT+4.03 \cdot CPUT$. This is shown in step two of the regression run.

The statistics generated by the program at this step tell more about the equation than which variables are most important. The correlation coefficient, R^2, for the equation equals .9843. (Recall from Chapter I that $R^2 \in [0,1]$ and that a value near one indicates a good fit. This means that about 98 percent of the variance in the exit times of the job stream is explained by the regression equation. Thus the regression equation fits the historical data very well. The standard error, SE, of the regression equation is 4.14 minutes. As shown in Chapter I, the following statements are possible using this measure:

a. The probability is .68 that a given job will exit during the time interval $[Y_p-4.14, Y_p+4.14]$. (Recall Y_p is the job's exit time as predicted by the regression equation.)

b. The probability is .95 that a given job will exit during the time interval $[Y_p-1.96 \cdot 4.14, Y_p+1.96 \cdot 4.14]$.

In other words, for a specific job that arrives at time 0.1 and has a CPU requirement of 6.0, there is a 68 percent chance that the job will exit in the time interval [18.52, 26.80], and a 95 percent chance it will exit in the time interval [14.51,30.81]. These intervals were found by substituting 0.1 and 6.0 into the regression equation for ARRT and CPUT, respectively, and then calculating the error term as shown above.

Conclusions drawn from the run.--The results of the run show that the two variables ARRT and CPUT are important in the prediction process while the third variable, CORE, is not. The regression equation has a high correlation coefficient and this shows the equation fits the historical data well. The standard error is small, only 4.14 minutes, so the error involved in using the equation to predict job exit times is probably insignificant. Based on these facts, it can be concluded that the regression equation performs well when used to predict job exit times in a random job stream.

Run Number Two

Satisfactory results from the early runs of the system as shown by Run Number One made it possible to progress to more complicated job streams. Run Number Two shows one of the early attempts to develop a regression equation for use in a cyclic job stream.

Job stream used.--A job stream with varying levels of activity is used in this run. The stream consists of twenty-five jobs arriving as follows:

a. For the first five jobs, ARRT=ARRT+RN, where RN $\in$ [0.1,0.9].

b. For jobs six through ten, ARRT=ARRT+.1.

c. For jobs eleven through fifteen, ARRT=ARRT+RN, where RN $\in$ [0.1,0.9].

d. For jobs sixteen through twenty, ARRT=ARRT+2.0.

e. For the last five jobs, ARRT=ARRT+RN, where RN $\in$ [0.1,0.9].

This arrival time distribution gives a pattern of random, fast, random, slow, and then random arrival times in the job stream. The arrival times range from zero for the first job to 16.70 for the last job. The job exit times as calculated by the computer system model range from 23.03 for the first job to 119.80 for the last. The CPU and CORE requirements for each job are determined just as they were in Job Number One, that is:

a. CPUT=RN, where RN $\in$ [1,9].

b. CORE=RN, where RN $\in$ [1,99].

Predictor variables tested.--Two of the variables tested in this run, ARRT and CPUT, were used in the regression equation developed in Run Number One. In addition, two other variables are tested. For each job in the stream, the jobs arriving in the previous ten minutes are counted and shown as the variable PREVJB. This variable indicates whether the jobs in the stream are arriving rapidly or slowly. The other variable, CLASS, is used to represent the CPU requirements of each job and is related to the variable CPUT. Most programmers can not accurately predict how much CPU time their job will require. By establishing classes to represent ranges of CPU requirements, the importance of an accurate estimate of this type can be reduced. In this run, jobs in the stream are assigned to a class from one to four depending on the variable CPUT. The classification is made as follows:

a. Class one indicates that CPUT is less than three minutes.

b. Class two indicates that CPUT is between three and five minutes.

c. Class three indicates that CPUT is between five and seven minutes.

d. Class four indicates that CPUT is greater than seven minutes.

Results of the run.--The output from the regression analysis program for this run is shown in Appendix IV. The four variables tested, ARRT, CPUT, PREVJB, and CLASS, are represented by the variables numbered one, two, three and four, respectively. The regression program was run twice, once to test the variables ARRT, CPUT, and PREVJB, and again to test the variables ARRT, CLASS, and PREVJB. These two runs are shown as selection one and selection two on the output. The first selection, using CPUT instead of CLASS, shows that R^2=.963 and SE=7.65 minutes. The second selection, using CLASS, shows that R^2=.967 and SE=7.19 minutes. The higher correlation coefficient and smaller standard error in the second selection indicate a better estimate is possible by using the variables ARRT,CLASS, and PREVJB than by using the variables ARRT, CPUT, and PREVJB. The regression equation for this selection can be stated as Y_p = -1.50+ 5.19·ARRT+2.62·PREVJB+8.02·CLASS. This is shown in step three of selection two.

The statistics R^2 and SE for the equation can be interpreted as they were for Run Number One. The correlation coefficient indicates that about 96 percent of the variance in the exit times is explained by the regression equation. The standard error of 7.65 minutes indicates the probability is .68 that a given job will exit during the time interval $[Y_p-7.65, Y_p+7.65]$, and .95 that it will exit during the time interval $[Y_p-14.99, Y_p+14.99]$.

Conclusions drawn from the run.--The results of Run Number Two indicate that a good regression equation to predict job exit time can be constructed for a cyclic job stream. The analysis of the standard error shows the error range to be about fifteen minutes at the 68 percent level and about thirty minutes at the 95 percent level. These ranges are small enough that the predictive model should be useful in estimating a job's exit time.

Run Number Three

As stated earlier, many runs of the computer system model and regression program were necessary in order to find the important predictor variables for use in a cyclic job stream. Many of the variables that were tested and rejected as important predictors will not be mentioned in the three runs described in this chapter. The three most important predictor variables found in the prior runs were the arrival time, ARRT, the number of jobs arriving before a particular job, PREVJB, and a classification of the amount of CPU time required, CLASS. The last run shown here is the first test for these three variables in a cyclic job stream.

Job stream used.--The job stream used in this run consists of seventy jobs arriving in a cyclic pattern. The arrival times for the jobs are determined as follows:

a. For the first ten jobs, ARRT=ARRT+RN, where RN $\in$ [0.1,0.9].

b. For jobs ten through thirty, ARRT=ARRT+0.1.

c. For jobs thirty-one through fifty, ARRT=ARRT+RN, where RN $\in$ [0.1,0.9].

d. For the last twenty jobs, ARRT=ARRT+2.0.

This arrival time distribution forms a pattern similar to the job stream in Run Number Two, that is, random, fast, random, and then slow. The arrival times range from zero for the first job to 52.40 for the last job. The exit times as calculated by the computer system model range from 23.03 for the first job to 371.00 for the last job. This stream contains more jobs than the previous job streams, and is therefore more realistic.

Predictor variables tested.--The variables tested for their importance in predicting job exit times are the arrival times, ARRT, the classification of the CPU requirements, CLASS, and the count of the jobs arriving in the twenty minutes before each job, PREVJB. These have the same meaning as the variables tested in the previous run, except for PREVJB, which here is a count of the jobs arriving in the last twenty minutes rather than the last ten minutes as in the last run. The runs of the system made before Run Number Three, but not shown in this chapter, indicated that PREVJB should be changed in this manner.

Results of the run.--The output from the regression analysis program for this run is shown in Appendix V. The three variables tested, ARRT, PREVJB, and CLASS, are represented by the variables numbered one, two, and three, respectively. The correlation coefficient, R^2, is equal to .955, thus 95 percent of the variance in the exit times of the jobs is explained by the regression equation. The standard error of 23.56 minutes means the probability is .95 that a job will exit during the interval [Y_p-46.18, Y_p+46.18], and .68 that it will exit during the interval [Y_p-23.56,Y_p+23.56].

Conclusions drawn from the run.--The high correlation coefficient for this run shows that the regression equation fits the historical data well. Thus, even for a complicated job stream such as this, the development of a good regression equation is possible. The standard error of 23.56 minutes is rather high, but not so high that it would make an exit time prediction useless. The range of this error term at a 68 percent confidence level is 48 minutes, and this knowledge would provide at least some quantifiable evidence to support the job exit time estimates made by computer center personnel.

A measure not yet discussed in any of the runs is the F value of the regression equation. This was described in

Chapter I as the ratio of the explained variance to the unexplained variance in the historical data. This ratio can be used in stepwise regression to indicate whether or not a variable should be used in the regression equation. The F value was used, along with R^2 and SE, to determine the variables most important in predicting job exit time.

Another way the F value can be used is in testing the statistical significance of the regression. The mechanics of this test were described in Chapter I. Using these procedures, the statistical significance of Run Number Three can be tested. As shown in step three of the regression run in Appendix V, the degrees of freedom (d.f.) in the regression equals 3 and in the residual equals 66. Consulting a table of F Distributions, and using a 95 percent confidence level, it can be found that F(3,66,.95)=2.76. The calculated F value of the equation must exceed four times this number, or 11.04, for the equation to be considered statistically significant. The calculated F value for Run Number Three is 470.57, so it is concluded that the equation is statistically significant. This means that the equation does a better job of predicting job exit times than could be done by mere chance.

Residuals Calculation

The regression equation constructed in Run Number Three is the final form of the predictive model for job exit times.

To help interpret the usefulness of this model, the residuals from the job stream in Run Number Three were calculated and are shown in Appendix VI. This printout shows the difference between the actual and predicted exit times for each job in the stream. The negative residuals indicate an overestimate of the actual exit time. That is, the job exited before the time predicted by the model. Conversely, the positive residuals show that the job exited later than the time predicted by the model. Thirty-one of the jobs in the stream have negative residuals, while the other thirty-nine have positive residuals. The largest residual is 59.10 minutes, meaning that this job exited about one hour later than the predictive model estimated. Likewise, the largest negative residual is -53.83 minutes, so the job finished about one hour before the time predicted by the model. Both of these jobs required about five hours to process; therefore, for these "worst" cases, the exit time prediction is about 20 percent off.

Draper and Smith (1, pp. 89-90) suggest that a good way to analyze residuals is to plot them in time sequence. This was done for every fourth residual from Run Number Three, and is shown in Figure 2. The general trend of this plot is for the residuals to increase with time. According to Draper and Smith, this means that a weighted least

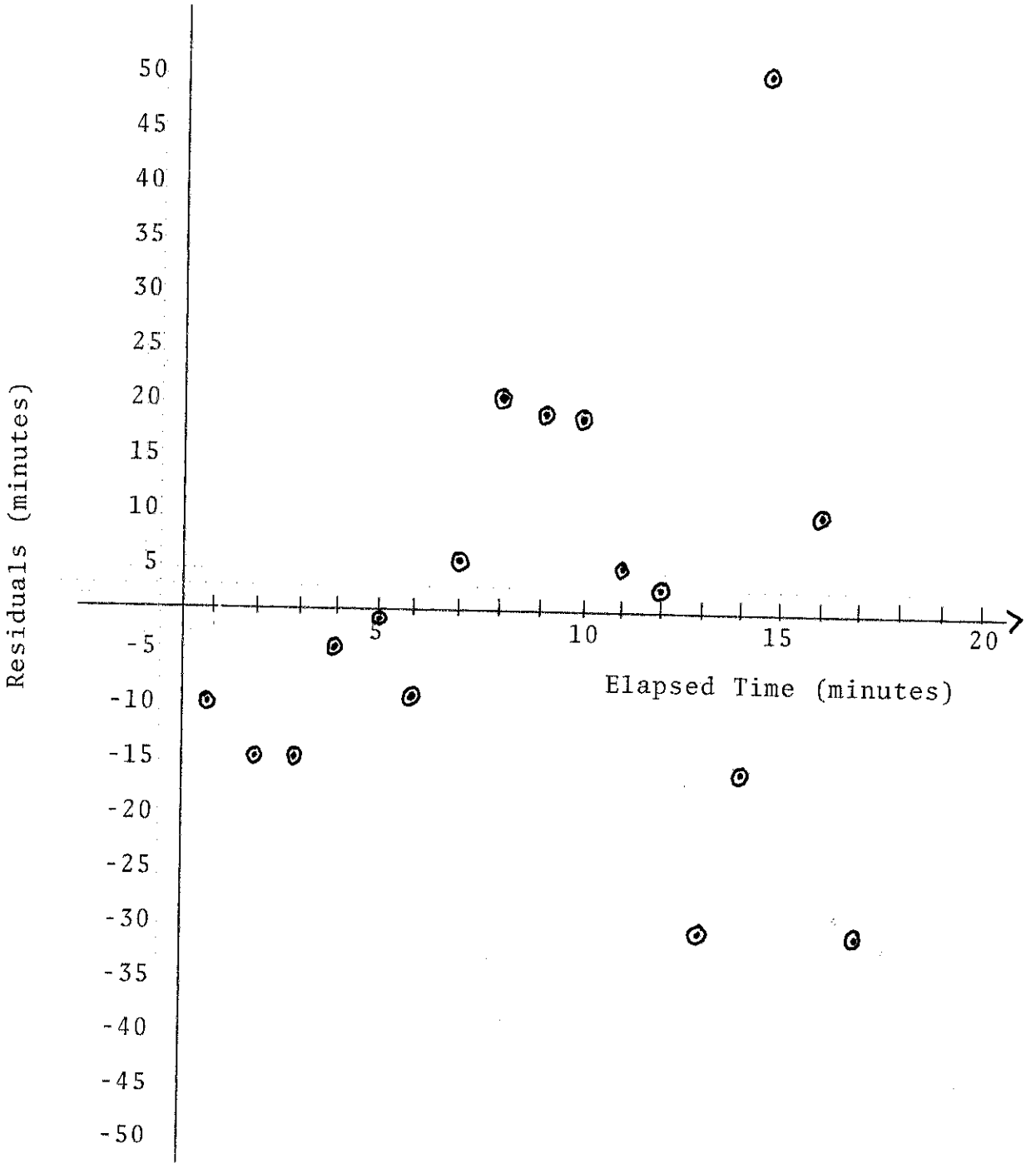

Figure 2--Time Sequence Residuals Plot for Every Fourth Residual from Run Number Three

squares method should have been used in the regression analysis. This method involves a transformation of the historical data before and after the regression analysis.

Conclusion

This chapter describes how regression analysis might be used in a computer system to predict job exit times. The feasibility of this application is tested by using regression analysis to construct a predictive model to estimate job exit times. Some other purposes of the predictive model are to provide guidelines for further investigation, to pinpoint important variables, and to act as a variables screening device.

A computer system simulator was programmed to provide the historical data used in constructing the model. The output from this simulator is input to a stepwise multiple regression analysis which builds a regression equation to predict job exit times. The results of three runs of the computer system simulator and regression program are shown in this chapter.

Based upon the performance of the predictive model with the various job streams that were tested, it is concluded that job exit time prediction is a feasible application for regression analysis. The results of the tests show that persons doing further research should be aware that a

weighted rather than normal least squares method of regression analysis may be necessary. The most important variables in predicting a job's exit time is its arrival time, a classification of its CPU time requirement, and the number of jobs arriving immediately prior to the job.

The next step for researchers in this area is to obtain historical job stream data from a small computer center and use the data to construct an operational model to predict job exit times. The actual application of this model may reveal other variables that are important to the prediction process, and the "fine tuning" of the model will be a time-consuming process. Nevertheless, based upon the results of the study described in this chapter, such a model is feasible and will be a valuable addition to the computer center.

CHAPTER BIBLIOGRAPHY

1. Draper, N. R. and H. Smith, Applied Regression Analysis, New York, John Wiley & Sons, Inc., 1967.

2. Madnick, Stuart E. and John J. Donovan, Operating Systems, New York, McGraw-Hill Book Company, 1974.

APPENDICES

I. COMPUTER LISTING OF JOB STREAM GENERATOR

II. COMPUTER LISTING OF MEMORY SIMULATOR

III. REGRESSION ANALYSIS FOR RUN NUMBER ONE

IV. REGRESSION ANALYSIS FOR RUN NUMBER TWO

V. REGRESSION ANALYSIS FOR RUN NUMBER THREE

VI. RESIDUALS CALCULATIONS FOR RUN NUMBER THREE

APPENDIX I

COMPUTER LISTING OF JOB STREAM GENERATOR

```
                C JOB STREAM GENERATOR
0001                REAL ARRT/0./,CPUT/0./,CORE/0./
0002                INTEGER NJOBS,KX/123456789/
0003             20 FORMAT(I3)
0004             50 FORMAT('1','JOB',2X,'ARRIVAL TIME',2X,'CPU TIME',2X,'CORE SIZE')
0005             40 FORMAT(' ',I3,3X,F7.2,6X,F7.2,3X,F7.2)
0006             60 FORMAT(' XNMB=',F12.9)
0007             70 FORMAT(3F8.2)
0008                READ(5,20) NJOBS
0009                WRITE(2,20) NJOBS
0010                WRITE(6,50)
0011                XNMB = XRANC(KX)
0012                KX = 0
0013                DO 30 I = 1,NJOBS
              C FIND ARRIVAL TIME
0014                XNMB = ABS(XRAND(KX))
0015                WRITE(8,60) XNMB
0016                KNMB = IFIX(XNMB * 10.)
0017                XNMB = FLOAT(KNMB) / 10.
0018                ARRT = ARRT + XNMB
              C FIND CPU TIME
0019                XNMB = ABS(XRAND(KX))
0020                WRITE(8,60) XNMB
0021                KNMB = IFIX(XNMB * 10.)
0022                IF (KNMB.EQ.0) KNMB = IFIX(XNMB * 100.)
0023                CPUT = KNMB
              C FIND CORE SIZE
0024                XNMB = ABS(XRAND(KX))
0025                WRITE(8,60) XNMB
0026                KNMB = XNMB * 100
0027                IF (KNMB.EQ.0) KNMB = XNMB * 1000
0028                CORE = KNMB
0029                WRITE(6,40) I,ARRT,CPUT,CORE
0030                WRITE(2,70) ARRT,CPUT,CORE
0031             30 CONTINUE
0032                END FILE 2
0033                STOP
0034                END

                C GENERATES RANDOM NUMBERS
0001                FUNCTION XRAND(KX)
0002                IF (KX.GT.0) IX = KX
0003                IY = 65539 * IX
0004                IF (IY.LT.0) IY = IY + 214748367 + 1
0005                XRAND = .4656613E-9 * FLOAT(IY)
0006                IX = IY
0007                RETURN
0008                END
```

JOB	ARRIVAL TIME	CPU TIME	CORE SIZE
1	0.0	5.00	40.00
2	0.10	6.00	42.00
3	0.60	2.00	56.00
4	0.80	4.00	40.00
5	1.20	4.00	34.00
6	1.50	9.00	68.00
7	1.90	5.00	95.00
8	2.60	2.00	70.00
9	3.40	6.00	4.00
10	3.50	3.00	91.00
11	4.00	7.00	4.00
12	4.50	6.00	57.00
13	5.00	8.00	72.00
14	5.20	7.00	21.00
15	5.40	3.00	95.00
16	6.20	4.00	60.00
17	6.70	7.00	32.00
18	7.10	6.00	1.00
19	7.80	2.00	74.00
20	8.30	4.00	66.00

APPENDIX II

COMPUTER LISTING OF MEMORY SIMULATOR

```
                C MEMORY SIMULATOR
0001                REAL ARRT(100),CPUT(100),CORE(100),EXEC(100),FINT(100)
0002                REAL RUNT(100)
0003                REAL SRESID/0./
0004             10 FORMAT(I3)
0005             12 FORMAT(3F8.2)
0006            114 FORMAT('1','JOB     ARRT     CPUT     CORE     FINT')
0007            255 FORMAT(5F8.2)
0008            115 FORMAT(' ',I3,4F8.2)
                C INITILIZATION
0009                READ(2,10) NJOBS
0010                READ(2,12) (ARRT(I),CPUT(I),CORE(I),I=1,NJOBS)
0011                CORSIZ = 250.
0012                ELAPT = 0.
0013                DO 20 I=1,100
0014             20 EXEC(I) = 0
0015                NRUN = 0
0016                JOBCNT = 0
0017                ELAPT = ARRT(1)
0018                CORSIZ = CORSIZ - CORE(1)
0019                NRUN = 1
0020                EXEC(1) = CPUT(1)
0021                JOBCNT = 1
0022                FINT(1)=ELAPT
                C FIND NEXT TAB POINT AND DETERMINE IF CAN START JOB
0023            100 EXEMIN = EXESUB(EXEC,NRUN)
0024                IF(JOBCNT.EQ.NJOBS) GO TO 30
0025                ARRMIN = ARRSUB(ELAPT,ARRT,JOBCNT)
0026                IF(CORSIZ.GE.CORE(JOBCNT + 1)) GO TO 35
0027                TAB = EXEMIN
0028                GO TO 40
0029             35 TAB = AMIN1(EXEMIN,ARRMIN)
0030                GO TO 40
0031             30 IF(EXEMIN.EQ.0.) GO TO 200
0032                TAB = EXEMIN
                C ADJUST EXECUTION TIMES FOR TAB POINT
0033             40 ELAPT = ELAPT + TAB
0034                NRUNT = NRUN
0035                LIMIT = JOBCNT
0036                DO 50 I = 1,LIMIT
0037                IF(EXEC(I).EQ.0) GO TO 50
0038                EXEC(I) = EXEC(I) - TAB/NRUN
0039                IF(EXEC(I).GT.0) GO TO 43
                C DELETE A JOB
0040                FINT(I)=ELAPT
0041                EXEC(I) = 0.
0042                NRUNT = NRUNT - 1
0043                CORSIZ = CORSIZ + CORE(I)
0044             43 IF(JOBCNT.GE.NJOBS) GO TO 50
0045                IF(CORE(JOBCNT+1).LE.CORSIZ.AND.ARRT(JOBCNT+1).LE.ELAPT) GO TO 45
0046                GO TO 50
                C START A JOB
0047             45 CORSIZ = CORSIZ - CORE(JOBCNT + 1)
0048                JOBCNT = JOBCNT + 1
0049                NRUNT = NRUNT + 1
0050                EXEC(JOBCNT) = CPUT(JOBCNT)
0051             50 CONTINUE
0052                NRUN = NRUNT
```

```
0053                GO TO 100
0054            200 WRITE(6,114)
0055                KJOBS = NJOBS - 1
0056                DO 201 K = 1,KJOBS
0057                PREVJB = 0.
0058                IF (K.EQ.1) GO TO 305
0059            305 RUNT(K) = FINT(K) - ARRT(K)
                C PRINT RESULTS OF SIMULATION
0060                WRITE(9,255)   ARRT(K),CPUT(K),CORE(K),FINT(K)
0061                WRITE(6,115) K,ARRT(K),CPUT(K),CORE(K),FINT(K)
0062            201 CONTINUE
0063                ENDIND = 99.
0064                WRITE(9,255) (ENDIND,I=1,4)
0065                END FILE 9
0066                STOP
0067                END
```

```
                C FIND MINIMUM AMOUNT OF EXECUTION TIME LEFT
0001                FUNCTION EXESUB(EXEC,NRUN)
0002                REAL EXEC(100)
0003                EXESUB = 9999.
0004                DO 20 J=1,100
0005                IF(EXEC(J).EQ.0.) GO TO 20
0006                IF(EXEC(J).LT.EXESUB) EXESUB = EXEC(J)
0007             20 CONTINUE
0008                IF(EXESUB.EQ.9999.) EXESUB = 0.
0009                EXESUB = EXESUB * NRUN
0010                RETURN
0011                END
```

```
                C FIND ARRIVAL TIME OF NEXT JOB
0001                FUNCTION ARRSUB(ELAPT,ARRT,JOBCNT)
0002                REAL ARRT(100)
0003                ARRSUB = ARRT(JOBCNT + 1) - ELAPT
0004                IF(ARRSUB.LT.0.) ARRSUB = 0.
0005                RETURN
0006                END
```

JOB	ARRT	CPUT	CORE	FINT
1	0.0	5.00	40.00	23.03
2	0.10	6.00	42.00	26.33
3	0.60	2.00	56.00	10.37
4	0.80	4.00	40.00	20.70
5	1.20	4.00	34.00	21.10
6	1.50	9.00	68.00	47.75
7	1.90	5.00	95.00	38.58
8	2.60	2.00	70.00	34.33
9	3.40	6.00	4.00	52.25
10	3.50	3.00	91.00	54.33
11	4.00	7.00	4.00	75.33
12	4.50	6.00	57.00	69.33
13	5.00	8.00	72.00	89.50
14	5.20	7.00	21.00	84.50
15	5.40	3.00	95.00	69.33
16	6.20	4.00	60.00	90.33
17	6.70	7.00	32.00	102.17
18	7.10	6.00	1.00	99.17
19	7.80	2.00	74.00	94.50
20	8.30	4.00	66.00	103.83

APPENDIX III

REGRESSION ANALYSIS FOR RUN NUMBER ONE

STEPWISE REGRESSION - RUN NUMBER ONE

NUMBER OF VARIABLES.....	4
NUMBER OF SELECTIONS.....	1
END OF DATA INDICATOR.....	99
DATA INPUT DEVICE.....	9

NO MINIMUM VARIANCE REQUIRED.

DATA FORMAT = (4F8.2)

NUMBER OF OBSERVATIONS 20

VARIABLE	MEAN	STANDARD DEVIATION
1	3.79000	2.64095
2	5.00000	2.09196
3	51.10000	29.24470
4	60.33800	31.29021

SIMPLE CORRELATIONS

	1	2	3	4
1	1.0000	0.0058	0.0048	0.9562
2	0.0058	1.0000	-0.4087	0.2703
3	0.0048	-0.4087	1.0000	-0.0778
4	0.9562	0.2703	-0.0778	1.0000

SELECTION 1

DEPENDENT VARIABLE.....	4
NUMBER OF FORCED VARIABLES..	0
NUMBER OF FREE VARIABLES....	3
MAXIMUM NUMBER OF STEPS.....	3

STEP 1

VARIABLE ENTERED 1

MULTIPLE R	0.9562	MULTIPLE R ADJUSTED FOR DEGREES OF FREEDOM	0.9537
MULTIPLE R-SQUARE	0.9143	R-SQUARE ADJUSTED FOR DEGREES OF FREEDOM	0.9095
INCREASE IN R-SQUARE	0.9143	INCREASE IN ADJUSTED R-SQUARE	0.9095
STANDARD ERROR OF ESTIMATE	9.4128	ADJUSTED STANDARD ERROR OF ESTIMATE	9.4128

*** ANALYSIS OF VARIANCE ***

SOURCE	DEGREES OF FREEDOM	SUM OF SQUARES	MEAN SQUARES	F	P
REGRESSION	1	17007.647	17007.647	191.9568	0.0000
RESIDUAL	18	1594.826	88.601		
TOTAL	19	18602.473			

*** REGRESSION EQUATION ***

VARIABLE	RAW COEFFICIENT	STANDARD COEFFICIENT	STANDARD ERROR	F	P
1 (FREE)	11.32882	0.95617	0.81768	191.9568	0.0000
CONSTANT	17.40177	0.57059			

STEP 2

VARIABLE ENTERED 2

MULTIPLE R	0.9921	MULTIPLE R ADJUSTED FOR DEGREES OF FREEDOM	0.9912
MULTIPLE R-SQUARE	0.9843	R-SQUARE ADJUSTED FOR DEGREES OF FREEDOM	0.9825
INCREASE IN R-SQUARE	0.0701	INCREASE IN ADJUSTED R-SQUARE	0.0730
STANDARD ERROR OF ESTIMATE	4.1406	ADJUSTED STANDARD ERROR OF ESTIMATE	4.2540

*** ANALYSIS OF VARIANCE ***

SOURCE	DEGREES OF FREEDOM	SUM OF SQUARES	MEAN SQUARES	F	P
REGRESSION	2	18311.019	9155.509	534.0236	0.0000
RESIDUAL	17	291.455	17.144		
TOTAL	19	18602.473			

*** REGRESSION EQUATION ***

VARIABLE	RAW COEFFICIENT	STANDARD COEFFICIENT	STANDARD ERROR	F	P
1 (FREE)	11.31054	0.95463	0.35969	988.7923	0.0000
2 (FREE)	4.03642	0.26470	0.46294	76.0232	0.0000
CONSTANT	-2.71107	-0.08889			

STEP 3

VARIABLE ENTERED 3

MULTIPLE R	0.9925	MULTIPLE R ADJUSTED FOR DEGREES OF FREEDOM	0.9911
MULTIPLE R-SQUARE	0.9851	R-SQUARE ADJUSTED FOR DEGREES OF FREEDOM	0.9823
INCREASE IN R-SQUARE	0.0008	INCREASE IN ADJUSTED R-SQUARE	-0.0001
STANDARD ERROR OF ESTIMATE	4.1572	ADJUSTED STANDARD ERROR OF ESTIMATE	4.3949

*** ANALYSIS OF VARIANCE ***

SOURCE	DEGREES OF FREEDOM	SUM OF SQUARES	MEAN SQUARES	F	P
REGRESSION	3	18325.962	6108.654	353.4694	0.0000
RESIDUAL	16	276.512	17.282		
TOTAL	19	18602.473			

*** REGRESSION EQUATION ***

VARIABLE	RAW COEFFICIENT	STANDARD COEFFICIENT	STANDARD ERROR	F	P
1 (FREE)	11.30791	0.95441	0.36114	980.4021	0.0000
2 (FREE)	4.22999	0.27740	0.50928	68.9866	0.0000
3 (FREE)	0.03323	0.03106	0.03573	0.8647	0.3663
CONSTANT	-5.36684	-0.17597			

*** SUMMARY TABLE ***

STEP	VARIABLE ENTERED	MULTIPLE R-SQUARE	ADJUSTED R-SQUARE	R-SQUARE INCREASE	ADJUSTED INCREASE	F	P
1	1 (FREE)	0.9143	0.9095	0.9143	0.9095	191.9568	0.0000
2	2 (FREE)	0.9843	0.9825	0.0701	0.0730	534.0236	0.0000
3	3 (FREE)	0.9851	0.9823	0.0008	-0.0001	353.4694	0.0000

APPENDIX IV

REGRESSION ANALYSIS FOR RUN NUMBER TWO

STEPWISE REGRESSION – RUN NUMBER TWO

NUMBER OF VARIABLES..... 5
NUMBER OF SELECTIONS..... 2
END OF DATA INDICATOR..... 99
DATA INPUT DEVICE..... 9

NO MINIMUM VARIANCE REQUIRED.

DATA FORMAT = (5F8.2)

NUMBER OF OBSERVATIONS 25

VARIABLE	MEAN	STANDARD DEVIATION
1	6.08400	6.30195
2	5.04000	2.24499
3	7.92000	4.67190
4	2.68000	1.06927
5	72.33640	37.18948

SIMPLE CORRELATIONS

	1	2	3	4	5
1	1.0000	-0.0269	0.1150	-0.0253	0.8788
2	-0.0269	1.0000	0.1672	0.9602	0.2527
3	0.1150	0.1672	1.0000	0.2282	0.4791
4	-0.0253	0.9602	0.2282	1.0000	0.2842
5	0.8788	0.2527	0.4791	0.2842	1.0000

SELECTION 1

DEPENDENT VARIABLE..... 5
NUMBER OF FORCED VARIABLES.. 0
NUMBER OF FREE VARIABLES.... 3
MAXIMUM NUMBER OF STEPS..... 3

STEP 1

VARIABLE ENTERED 1

MULTIPLE R	0.8788	MULTIPLE R ADJUSTED FOR DEGREES OF FREEDOM	0.8731
MULTIPLE R-SQUARE	0.7722	R-SQUARE ADJUSTED FOR DEGREES OF FREEDOM	0.7623
INCREASE IN R-SQUARE	0.7722	INCREASE IN ADJUSTED R-SQUARE	0.7623
STANDARD ERROR OF ESTIMATE	18.1299	ADJUSTED STANDARD ERROR OF ESTIMATE	18.1299

*** ANALYSIS OF VARIANCE ***

SOURCE	DEGREES OF FREEDOM	SUM OF SQUARES	MEAN SQUARES	F	P
REGRESSION	1	25633.445	25633.445	77.9860	0.0000
RESIDUAL	23	7559.940	328.693		
TOTAL	24	33193.386			

*** REGRESSION EQUATION ***

VARIABLE	RAW COEFFICIENT	STANDARD COEFFICIENT	STANDARD ERROR	F	P
1 (FREE)	5.39120	0.87878	0.61049	77.9860	0.0000
CONSTANT	39.53633	1.08503			

STEP 2

VARIABLE ENTERED 3

MULTIPLE R	0.9576	MULTIPLE R ADJUSTED FOR DEGREES OF FREEDOM	0.9537
MULTIPLE R-SQUARE	0.9171	R-SQUARE ADJUSTED FOR DEGREES OF FREEDOM	0.9095
INCREASE IN R-SQUARE	0.1448	INCREASE IN ADJUSTED R-SQUARE	0.1472
STANDARD ERROR OF ESTIMATE	11.1854	ADJUSTED STANDARD ERROR OF ESTIMATE	11.4260

*** ANALYSIS OF VARIANCE ***

SOURCE	DEGREES OF FREEDOM	SUM OF SQUARES	MEAN SQUARES	F	P
REGRESSION	2	30440.873	15220.437	121.6523	0.0000
RESIDUAL	22	2752.513	125.114		
TOTAL	24	33193.386			

*** REGRESSION EQUATION ***

VARIABLE	RAW COEFFICIENT	STANDARD COEFFICIENT	STANDARD ERROR	F	P
1 (FREE)	5.12090	0.63472	0.37916	182.4065	0.0000
3 (FREE)	3.04964	0.38311	0.49198	38.4243	0.0000
CONSTANT	17.02769	0.46730			

STEP 3

VARIABLE ENTERED 2

MULTIPLE R	0.9813	MULTIPLE R ADJUSTED FOR DEGREES OF FREEDOM	0.9786
MULTIPLE R-SQUARE	0.9630	R-SQUARE ADJUSTED FOR DEGREES OF FREEDOM	0.9577
INCREASE IN R-SQUARE	0.0459	INCREASE IN ADJUSTED R-SQUARE	0.0482
STANDARD ERROR OF ESTIMATE	7.6453	ADJUSTED STANDARD ERROR OF ESTIMATE	7.9852

*** ANALYSIS OF VARIANCE ***

SOURCE	DEGREES OF FREEDOM	SUM OF SQUARES	MEAN SQUARES	F	P
REGRESSION	3	31965.928	10655.309	182.2967	0.0000
RESIDUAL	21	1227.458	58.450		
TOTAL	24	33193.386			

*** REGRESSION EQUATION ***

VARIABLE	RAW COEFFICIENT	STANDARD COEFFICIENT	STANDARD ERROR	F	P
1 (FREE)	5.16331	0.64469	0.25945	399.1321	0.0000
3 (FREE)	2.75070	0.34555	0.34132	64.9466	0.0000
2 (FREE)	3.60544	0.21765	0.70585	26.0915	0.0000
CONSTANT	0.84417	0.02317			

*** SUMMARY TABLE ***

STEP	VARIABLE ENTERED	MULTIPLE R-SQUARE	ADJUSTED R-SQUARE	R-SQUARE INCREASE	ADJUSTED INCREASE	F	P
1	1 (FREE)	0.7722	0.7623	0.7722	0.7623	77.9860	0.0000
2	3 (FREE)	0.9171	0.9095	0.1446	0.1472	121.6523	0.0000
3	2 (FREE)	0.9630	0.9577	0.0459	0.0482	182.2967	0.0000

1 VARIABLES DELETED: 4

SELECTION 2

DEPENDENT VARIABLE..... 5
NUMBER OF FORCED VARIABLES.. 0
NUMBER OF FREE VARIABLES.... 3
MAXIMUM NUMBER OF STEPS..... 3

STEP 1

VARIABLE ENTERED 1

MULTIPLE R	0.8788	MULTIPLE R ADJUSTED FOR DEGREES OF FREEDOM	0.8731
MULTIPLE R-SQUARE	0.7722	R-SQUARE ADJUSTED FOR DEGREES OF FREEDOM	0.7623
INCREASE IN R-SQUARE	0.7722	INCREASE IN ADJUSTED R-SQUARE	0.7623
STANDARD ERROR OF ESTIMATE	18.1299	ADJUSTED STANDARD ERROR OF ESTIMATE	18.1299

*** ANALYSIS OF VARIANCE ***

SOURCE	DEGREES OF FREEDOM	SUM OF SQUARES	MEAN SQUARES	F	P
REGRESSION	1	25633.445	25633.445	77.9860	0.0000
RESIDUAL	23	7559.940	328.693		
TOTAL	24	33193.386			

*** REGRESSION EQUATION ***

VARIABLE	RAW COEFFICIENT	STANDARD COEFFICIENT	STANDARD ERROR	F	P
1 (FREE)	5.39120	0.87878	0.61049	77.9860	0.0000
CONSTANT	39.53633	1.08503			

STEP 2

VARIABLE ENTERED 3

MULTIPLE R	0.9576	MULTIPLE R ADJUSTED FOR DEGREES OF FREEDOM	0.9537
MULTIPLE R-SQUARE	0.9171	R-SQUARE ADJUSTED FOR DEGREES OF FREEDOM	0.9095
INCREASE IN R-SQUARE	0.1448	INCREASE IN ADJUSTED R-SQUARE	0.1472
STANDARD ERROR OF ESTIMATE	11.1854	ADJUSTED STANDARD ERROR OF ESTIMATE	11.4260

*** ANALYSIS OF VARIANCE ***

SOURCE	DEGREES OF FREEDOM	SUM OF SQUARES	MEAN SQUARES	F	P
REGRESSION	2	30440.873	15220.437	121.6525	0.0000
RESIDUAL	22	2752.513	125.114		
TOTAL	24	33193.386			

*** REGRESSION EQUATION ***

VARIABLE	RAW COEFFICIENT	STANDARD COEFFICIENT	STANDARD ERROR	F	P
1 (FREE)	5.12090	0.83472	0.37916	182.4065	0.0000
3 (FREE)	3.04964	0.38311	0.49198	38.4243	0.0000
CONSTANT	17.02769	0.46730			

STEP 3

VARIABLE ENTERED 4

MULTIPLE R	0.9835	MULTIPLE R ADJUSTED FOR DEGREES OF FREEDOM	0.9811
MULTIPLE R-SQUARE	0.9673	R-SQUARE ADJUSTED FOR DEGREES OF FREEDOM	0.9626
INCREASE IN R-SQUARE	0.0502	INCREASE IN ADJUSTED R-SQUARE	0.0531
STANDARD ERROR OF ESTIMATE	7.1906	ADJUSTED STANDARD ERROR OF ESTIMATE	7.5103

*** ANALYSIS OF VARIANCE ***

SOURCE	DEGREES OF FREEDOM	SUM OF SQUARES	MEAN SQUARES	F	P
REGRESSION	3	32107.584	10702.528	206.9928	0.0000
RESIDUAL	21	1085.802	51.705		
TOTAL	24	33193.386			

*** REGRESSION EQUATION ***

VARIABLE	RAW COEFFICIENT	STANDARD COEFFICIENT	STANDARD ERROR	F	P
1 (FREE)	5.19469	0.84674	0.24409	452.9071	0.0000
3 (FREE)	2.61994	0.32913	0.32520	64.9062	0.0000
4 (FREE)	8.01619	0.23048	1.41190	32.2351	0.0000
CONSTANT	-1.50145	-0.04121			

*** SUMMARY TABLE ***

STEP	VARIABLE ENTERED	MULTIPLE R-SQUARE	ADJUSTED R-SQUARE	R-SQUARE INCREASE	ADJUSTED INCREASE	F	P
1	1 (FREE)	0.7722	0.7623	0.7722	0.7623	77.9860	0.0000
2	3 (FREE)	0.9171	0.9095	0.1448	0.1472	121.6523	0.0000
3	4 (FREE)	0.9673	0.9626	0.0502	0.0531	206.9928	0.0000

1 VARIABLES DELETED: 2

APPENDIX V

REGRESSION ANALYSIS FOR RUN NUMBER THREE

STEPWISE REGRESSION - RUN NUMBER THREE

NUMBER OF VARIABLES..... 4
NUMBER OF SELECTIONS..... 1
END OF DATA INDICATOR..... 99
DATA INPUT DEVICE..... 9

NO MINIMUM VARIANCE REQUIRED.

DATA FORMAT = (4F8.2)

NUMBER OF OBSERVATIONS 70

VARIABLE	MEAN	STANDARD DEVIATION
1	13.57143	14.32214
2	23.82857	15.52742
3	2.70571	1.07562
4	194.97357	109.01965

SIMPLE CORRELATIONS

	1	2	3	4
1	1.0000	-0.1473	-0.0151	0.8511
2	-0.1473	1.0000	0.0802	0.3443
3	-0.0151	0.0802	1.0000	0.0996
4	0.8511	0.3443	0.0996	1.0000

SELECTION 1

DEPENDENT VARIABLE..... 4
NUMBER OF FORCED VARIABLES.. 0
NUMBER OF FREE VARIABLES.... 3
MAXIMUM NUMBER OF STEPS..... 3

STEP 1

VARIABLE ENTERED 1

MULTIPLE R	0.8511	MULTIPLE R ADJUSTED FOR DEGREES OF FREEDOM	0.8488
MULTIPLE R-SQUARE	0.7245	R-SQUARE ADJUSTED FOR DEGREES OF FREEDOM	0.7204
INCREASE IN R-SQUARE	0.7245	INCREASE IN ADJUSTED R-SQUARE	0.7204
STANDARD ERROR OF ESTIMATE	57.6462	ADJUSTED STANDARD ERROR OF ESTIMATE	57.6462

*** ANALYSIS OF VARIANCE ***

SOURCE	DEGREES OF FREEDOM	SUM OF SQUARES	MEAN SQUARES	F	P
REGRESSION	1	594114.906	594114.906	178.7843	0.0000
RESIDUAL	68	225969.634	3323.083		
TOTAL	69	820084.540			

*** REGRESSION EQUATION ***

VARIABLE	RAW COEFFICIENT	STANDARD COEFFICIENT	STANDARD ERROR	F	P
1 (FREE)	6.47892	0.85115	0.48455	178.7843	0.0000
CONSTANT	107.04737	0.98900			

STEP 2

VARIABLE ENTERED 2

MULTIPLE R	0.9745	MULTIPLE R ADJUSTED FOR DEGREES OF FREEDOM	0.9737
MULTIPLE R-SQUARE	0.9497	R-SQUARE ADJUSTED FOR DEGREES OF FREEDOM	0.9482
INCREASE IN R-SQUARE	0.2252	INCREASE IN ADJUSTED R-SQUARE	0.2278
STANDARD ERROR OF ESTIMATE	24.8181	ADJUSTED STANDARD ERROR OF ESTIMATE	24.9999

*** ANALYSIS OF VARIANCE ***

SOURCE	DEGREES OF FREEDOM	SUM OF SQUARES	MEAN SQUARES	F	P
REGRESSION	2	778816.627	389408.313	632.2190	0.0000
RESIDUAL	67	41267.914	615.939		
TOTAL	69	820084.540			

*** REGRESSION EQUATION ***

VARIABLE	RAW COEFFICIENT	STANDARD COEFFICIENT	STANDARD ERROR	F	P
1 (FREE)	7.01583	0.92168	0.21090	1106.6143	0.0000
2 (FREE)	3.36865	0.47979	0.19453	299.8701	0.0000
CONSTANT	19.49056	0.16007			

STEP 3

VARIABLE ENTERED 3

MULTIPLE R	0.9774	MULTIPLE R ADJUSTED FOR DEGREES OF FREEDOM	0.9764
MULTIPLE R-SQUARE	0.9553	R-SQUARE ADJUSTED FOR DEGREES OF FREEDOM	0.9533
INCREASE IN R-SQUARE	0.0057	INCREASE IN ADJUSTED R-SQUARE	0.0051
STANDARD ERROR OF ESTIMATE	23.5579	ADJUSTED STANDARD ERROR OF ESTIMATE	23.9069

*** ANALYSIS OF VARIANCE ***

SOURCE	DEGREES OF FREEDOM	SUM OF SQUARES	MEAN SQUARES	F	P
REGRESSION	3	783456.271	261152.090	470.5665	0.0000
RESIDUAL	66	36628.269	554.974		
TOTAL	69	820084.540			

*** REGRESSION EQUATION ***

VARIABLE	RAW COEFFICIENT	STANDARD COEFFICIENT	STANDARD ERROR	F	P
1 (FREE)	7.01776	0.92194	0.20019	1228.8411	0.0000
2 (FREE)	3.32641	0.47377	0.18523	322.4980	0.0000
3 (FREE)	7.65040	0.07546	2.64593	8.3601	0.0052
CONSTANT	-0.84088	-0.00777			

*** SUMMARY TABLE ***

STEP	VARIABLE ENTERED	MULTIPLE R-SQUARE	ADJUSTED R-SQUARE	R-SQUARE INCREASE	ADJUSTED INCREASE	F	P
1	1 (FREE)	0.7245	0.7204	0.7245	0.7204	178.7843	0.0000
2	2 (FREE)	0.9497	0.9482	0.2252	0.2278	632.2190	0.0000
3	3 (FREE)	0.9553	0.9533	0.0057	0.0051	470.5665	0.0000

APPENDIX VI

RESIDUALS CALCULATIONS FOR RUN NUMBER THREE

RESIDUALS FROM RUN NUMBER THREE			
ACTUAL	ESTIMATED	RESIDUAL	SUM RESIDUAL
23.03	22.11	0.92	0.92
26.33	26.14	0.19	1.11
10.37	17.68	-7.32	-6.20
20.70	30.07	-9.37	-15.57
21.10	36.20	-15.10	-30.67
47.75	56.94	-9.19	-39.86
38.58	55.43	-16.84	-56.71
34.33	48.37	-14.04	-70.74
52.25	72.62	-20.37	-91.11
54.33	69.00	-14.67	-105.78
75.33	88.33	-13.00	-118.78
69.33	84.71	-15.38	-134.16
89.50	96.40	-6.90	-141.05
84.50	100.43	-15.93	-156.98
69.33	89.16	-19.83	-176.81
90.33	93.19	-2.86	-179.67
105.33	112.52	-7.19	-186.86
101.33	108.91	-7.57	-194.43
94.50	97.64	-3.14	-197.57
108.67	109.32	-0.65	-198.22
117.33	121.00	-3.67	-201.89
122.83	132.68	-9.85	-211.74
100.50	113.77	-13.27	-225.01
131.50	140.75	-9.25	-234.26
119.83	129.48	-9.65	-243.90
130.83	141.16	-10.33	-254.23
164.67	152.84	11.82	-242.41
154.83	149.23	5.61	-236.80
164.67	160.91	3.76	-233.04
155.50	157.29	-1.79	-234.83
201.67	169.67	31.99	-202.84
179.83	157.70	22.13	-180.71
195.83	174.30	21.53	-159.18
201.67	178.33	23.33	-135.84
218.50	192.12	26.38	-109.46
218.50	198.96	19.54	-89.92
201.67	195.34	6.32	-83.60
207.50	195.23	12.27	-71.33
207.50	199.27	8.23	-63.10
239.50	218.60	20.90	-42.20
255.33	221.93	33.41	-8.79
255.33	230.87	24.46	15.67
240.50	227.96	12.54	28.21
224.50	219.50	5.00	33.21
224.50	222.83	1.67	34.88
230.50	227.56	2.94	37.82
235.50	236.51	-1.01	36.81
261.67	260.05	1.61	38.43
254.67	259.24	-4.58	33.85
258.67	264.68	-6.01	27.84
283.17	297.35	-14.18	13.66
283.17	314.72	-31.55	-17.89
291.67	332.09	-40.42	-58.31
273.67	327.50	-53.83	-112.14
307.17	343.52	-36.35	-148.50
301.17	316.61	-15.44	-163.94
309.67	287.36	22.31	-141.63
348.17	289.07	59.10	-82.53
310.33	260.18	50.15	-32.38
328.17	275.21	52.96	20.58
359.17	304.55	54.62	75.20
322.33	292.31	30.02	105.23
336.17	314.00	22.17	127.39
348.17	335.69	12.48	139.87
360.00	349.73	10.27	150.14
329.17	348.47	-19.30	130.84
348.17	370.16	-21.99	108.85
368.00	399.50	-31.50	77.35
377.00	413.54	-36.54	40.81
371.00	419.93	-48.93	-8.12

BIBLIOGRAPHY

Books

Draper, N. R. and H. Smith, Applied Regression Analysis, New York, John Wiley & Sons, Inc., 1967.

Madnick, Stuart E. and John J. Donovan, Operating Systems, New York, McGraw-Hill Book Company, 1974.

Mason, Robert D., Statistical Techniques in Business and Economics, Illinois, Richard D. Irwin, Inc., 1974.

Sterling, Theodor D. and Seymour V. Pollack, Introduction to Statistical Data Processing, New Jersey, Prentice-Hall, Inc., 1968.

Articles

Benston, George J., "Multiple Regression Analysis of Cost Behavior," The Accounting Review, Vol. XLI, No. 4 (October, 1966), 657-672.

Comiskey, Eugene E., "Cost Control by Regression Analysis," The Accounting Review, Vol. XLI, No. 2 (April, 1966), 235-238.

Deakin, Edward R. and Michael H. Granof, "Directing Audit Effort Using Regression Analysis," The CPA Journal, Vol. XLVI, No. 2 (February, 1976), 29-33.

Deakin, Edward R. and Michael H. Granof, "Regression Analysis as a Means of Determining Audit Sample Size," The Accounting Review, Vol. XLIX, No. 4 (October, 1974), 764-771.

Jensen, Robert E., "A Multiple Regression Model for Cost Control - Assumptions and Limitations," The Accounting Review, Vol. XLII, No. 2 (April, 1967), 265-273.

Kinney, William R., Jr. and Andrew D. Bailey, Jr., "Regression Analysis as a Means of Determining Audit Sample Size: A Comment, "The Accounting Review, Vol. LI, No. 2 (April, 1976), 396-401.