AN EVALUATION OF THE VALIDITY OF THE USE OF REGRESSION AND CORRELATION ANALYSIS AMONG SOME SELECTED M.B.A. PROJECTS, FACULTY OF COMMERCE, UNIVERSITY OF NAIROBI

BY

PETER NGEI KIILU

A MANAGEMENT RESEARCH PROJECT SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF BUSINESS AND ADMINISTRATION, FACULTY OF COMMERCE, UNIVERSITY OF NAIROBI.

JUNE 1990

DECLARATION

This Research Project is my original work and has not been presented for a degree in any other university.

PETER NGEI KIILU

This management project has been submitted for examination with my approval as university supervisor.

MR. GITURO WAINAINA

DEDICATION

To my mother , **Agatha Kavivi Kiilu** ; the lady without whose love I would not have gone this far.

TABLE OF CONTENTS

LIST OF TABLES

ACKNOWLEDGMENTS

I would like sincere gratitude to all the individuals who directly or indirectly contributed to the eventual completion of this project.

In particular I wish to thank most sincerely my supervisor **Mr. Gituro Wainaina** for his invaluable comments, guidance, keen interest and tireless effort that he devoted on this project from its conceptualisation up to the final stage of the production of this report.

I also feel indebted to sincerely thank **Professor Kimura** for his contribution during the proposal presentation stage of this project together with the other lecturers and my fellow **M.B.A II colleagues** whose contribution helped a great deal to lay a strong foundation for this project.

Special thanks go to my entire family for their encouragement, appreciation and patience that they have shown throughout my entire stay as a student.

ABSTRACT

This study sought to evaluate the validity of the use of regression and correlation analysis among some selected M.B.A. projects. The need for the study was aroused by the claim that many of the statistical analysis tools are being applied to the wrong situations which could better be done using other more appropriate analytical tools.

To achieve this objective , tests were carried out to determine whether the assumptions underlying the validity of ordinary least squares models held true. Significance tests were also carried out to determine whether the coefficients of the variables included in these models are significantly different from zero or they arise by chance.

Two M.B.A. projects were analysed ; one written by Onuong'a in 1988 and the other written by Kariuki in 1989.

The assumptions of ordinary least squares models all held true with the exception that there was autocorrelation in the case of Onuong'a's project. Significance tests on coefficients of the variables included by Onuong'a in his optimal model revealed that all the predictor variables were significantly related to the criterion variable. Thus it was concluded that the use of regression and correlation analysis was valid in this case.

In the case of Kariuki,s project the models he derived for metal work and carpentry sub-sectors of Jua Kali industry were valid. Tests on the assumptions made indicated that the assumptions held true.

However , for the motor vehicle sub-sector the tests revealed that the model specification was inadequate. Even though the three sub-sectors's models had a very weak explanatory power.

CHAPTER ONE

1. INTRODUCTION

1.1 Background

In the recent past, the Faculty of Commerce has experienced increased application of regression analysis as a tool of analysing data collected by Masters in Business and Administration (M.B.A) students doing their final management projects. Of interest to the faculty, the entire academia, and the external parties who would be intending to utilise the findings arrived at by these researchers is the validity and the reliability of the findings using the analytical tools that they choose to analyse their data with.

The term regression is often used to indicate "the return to a mean or average value" and regression analysis is an approach that may be used for the study of relations between variables-particularly for the purpose of understanding how one or more other variables relate [1]. Regression analysis models may be used for predictive as well as for descriptive purposes. They are descriptive if they are used to show the relationship existing between two or more variables and predictive if they can be used to estimate the values of the response variable given the value(s) of the predictor variable(s).

According to the Fontana Dictionary of Modern Thought, validity is the characteristic of an inference whose conclusion must be true if it premises are.

In the years to come, regression analysis may probably be the most common statistical tool used by M.B.A. students for data analysis in their projects. This has been evident in the number of projects that used regression analysis in the last two years 1988 and 1989 when eight M.B.A.students out of 31 used regression analysis models in analysing their data. In 1987, two students used this analytical tool to analyse their data in their projects. This increased popularity may have come about due to several factors; some technological and others due to the increased awareness of the power of regression analysis as a tool that may be used in analysing both qualitative as well as quantitative data.

The most remarkable boost to the popularity of regression analysis has been the advent of micro-computers in the portfolio of the faculty resources. These micro-computers have actually revolutionalised modes of data analysis in the faculty as a whole in the recent past. The micro-computers have loaded in them high quality software particularly the more commonly used one called STATGRAPHICS. This is a statistical as well as a graphics user-friendly package with wide variety of applications and capabilities. It is currently possible to carry out more sophisticated regression analysis than could have been possible before the faculty acquired these micros in 1987.

Another reason why there has been an increased application of regression analysis is due to the increased awareness of the M.B.A. students of the power and usefulness of regression analysis. It is not by any surprise that this is so because the

M.B.A. programme draws a substantial number of its students from the Faculty of Commerce and other faculties most of whom have at least an elementary knowledge of fundamentals of statistics and computer science.

However, many M.B.A. students may have tended to behave like lumberjacks who move around with axes on their shoulders looking for a tree to fell. This analogy is used in this study to describe a situation whereby a student carrying out a research study knows and appreciates the power and the usefulness of regression analysis and he/she opts to either; collect the data and manipulate it to suit regression analysis or collect the data and analyse it using regression analysis despite the fact that it may be an inappropriate tool for that purpose. It therefore follows that regression analysis could be one of the most abused tool of analysis due to the ease of its use.

This study strives to evaluate the validity of regression analysis results. It may be that this tool is very vulnerable to misuse because the users believe that it is simple to use and it produces useful and valid application or may be because of the advent and the availability of computers and useful software has influenced the users will to forget about the importance of identifying appropriate variables and the manner in which they are allowed to relate. This could lead to invalid research conclusions. This validity of the research findings arrived at by M.B.A. students in their projects using regression analysis has been the motivating factor for this study. The importance of valid research findings, needless to say, cannot be overlooked especially given that they are meant to be useful not only in the furtherance of the body knowledge but also for practical purposes.

For instance, the findings of the two studies that will be evaluated in this study are very important at this time both for the Faculty of Commerce and the relevant government agencies. The findings of Onuong'a (1988) can assist the faculty in making the proposal to the university senate on the extent to which the minimum entry requirements may be lowered if the university is to admit the targeted number of students.

For example , this may be necessary given the record mass failure by the form four students to attain the minimum entry requirements set earlier on before the Kenya Secondary Certificate of Education (KSCE) results were released. According to Onuong'a's study, the faculty can relax the minimum entry requirement in those subjects which were not identified as being the key success factors in the Bachelor of Commerce (BCOM) degree programme. This would ensure that the admitted students go through the programme smoothly while still ensuring that the numbers are maintained at the required level.

On the national front,the findings of Kariuki (1989) would also be of keen interest to organisations like Agricultural Finance Corporation, (AFC),Industrial and Commercial Development Corporation, (ICDC), Kenya Industrial Estates, (KIE), and other financing organisations in determining which small scale businessmen possess the key success socioeconomic qualities. This would enable these organisations to identify carefully people who are likely to succeed and save on the scarce funds earmarked for the small scale businesses in Kenya. Of particular need would be the allocation of the capital accumulated under the recently launched Small Enterprises Development Fund to those who deserve it.

Though the two studies would be very handy under the situations described above, this would only be so under the premises that the findings are valid and reliable hence the need for this study.

1.2. Statement of the problem.

The validity of the results of regression analysis is hinged on the significance of the estimates that have been computed and on how well the assumptions on which regression analysis is based hold true. Virtually every person who uses regression analysis for data analysis makes several assumptions. There are at least four assumptions that are commonly made and that must hold in order to make the results of regression analysis valid. If these assumptions do not hold true then the validity of the findings would be grossly jeopardised hence the need for this study. There are several reasons why these assumptions may not hold true. In the first place, the data being analysed may be inadequate with few observations only and second, the regression model in use may be misspecified in which case a wrong formulation of the model may be used instead of the right one. Another possible reason why the assumptions made may not hold is due to the omission of some vital variables from the model in use. These causes may not be of the researcher's own making but may be imposed on the researcher by inevitable circumstances. For whatever reasons that the assumptions are flouted, they render the research findings invalid and care has to be taken to ensure that there are minimum chances of flouting them.

The most important of the assumptions of regression analysis is the assumption that the expected value of the error term $(U_i)=0$, that is, $E(U_i)=0$ for all i's. This assumption is necessary for the estimates to be unbiased. The requirement that estimates be unbiased is relevant both for statistical inference and managerial decisions[2].

If the estimates are biased, the true but unknown values for which we are trying to estimate will be significantly different from the estimated values. Our estimates will therefore be meaningless and our managerial decision will be compromised. It follows therefore that any decision we make based on a biased estimate will be wrong.

The second assumption that is made in carrying out regression analysis is that $VAR(U_i)=\delta^2$ for all i's, that is, the variance of the error terms is constant for all observations. The violation of this assumption is commonly known as heteroscedasticity. If heteroscedasticity exists then the calculation of the error terms is wrong. Wrong error terms imply that our estimates are inefficient, that is, the variance of the estimated beta coefficients is not minimum. Hence the least squares estimates of the error terms would not be BLUE (Best Linear Unbiased Estimators).

The third assumption is that error terms are independent for any pair of observations. The violation of this assumption is commonly referred to as autocorrelation. Autocorrelation does not render the estimates biased but it does make them inefficient thus implying that the error term calculation is wrong. This means that it will not be of much use to apply the ordinary least squares regression models as they become inefficient.

The fourth assumption and the last to be highlighted in this section of the study is one usually made in multiple regression analysis. This is the assumption that there is no correlation between predictor variables in the models. This means that the predictor variables should be independent of one another. If this assumption does not hold true then a situation commonly referred to as multicollinearity is said to exist. Multicollinearity makes it difficult to obtain reliable estimates of the separate effects of the predictor variables. Holding other factors constant, as the correlation between predictor variables increases, the standard error of estimate also increases.

A test of significance for the effect of each predictor variable reveals that as the standard error increases, it becomes almost impossible to reject the null hypothesis of no effect for each predictor variable even though the actual effect may be insignificant.

Having pointed out briefly the major assumptions made in regression analysis and having highlighted some of the possible dangers that stand to be encountered if the assumptions do not hold true, it becomes very clear that the validity of regression analysis results stands at stake. Yet too often it is not uncommon to find a researcher who analyses his data using the tool of regression analysis making conclusions which are assumed to be valid without testing whether the assumptions of the model actually hold true.

1.3 Objectives of the study.

This study has two main objectives:

(1). To determine whether the coefficients of the variables used in the regression models are statistically significant (first-order regression conditions).

(2). To determine whether some of the assumptions of least square regression analysis made by the researchers are valid (second-order regression conditions). The second-order tests are done to determine whether the assumptions made in regression analysis actually hold true or whether they have been flouted. They help to determine the validity of the conclusions arrived at during the analysis.

To achieve these two objectives the following hypotheses have been developed all of which are stated in the null.

Hypothesis 1: All the regression model slope coefficients calculated by the researchers are significant to warrant reliance on them in making valid conclusions from the findings of their analysis.

Hypothesis 2: The least squares regression analysis assumptions made by the researchers in their data analysis hold true and can be relied upon to produce valid regression analysis results.

1.4 **Importance of the study.**

This study would provide a useful insight into regression analysis. Of particular importance, the study would :

(1). highlight the fact that substantive knowledge of regression analysis is required by the users before it is used for data analysis.

(2). bring into light the need to investigate the significance of estimates obtained from the regression analysis to see if the implications are reasonable.

(3). provide a framework that can be used to evaluate regression analysis results obtained by others as well as to produce such results on regression parameters that we obtain ourselves.

(4). assist in producing more valid regression analysis results by drawing the attention of possible users of the regression analysis tool of the importance of testing the assumptions on which the validity of their conclusions hinge.

(5). indicate the need for a more careful choice of an appropriate analytical tool that suits the situation at hand.

CHAPTER TWO

2. LITERATURE REVIEW

2.1. Findings of Onuong'a.

In his M.B.A. project submitted in 1988 for the partial fulfilment of the requirements for the award of the M.B.A. degree, Faculty of Commerce, University of Nairobi, John O. Onuong'a studied the relationship between certain Bachelor of Commerce student attributes and their performance in the final year of study at the University of Nairobi.

He carried out his study using a sample of 300 students divided into three sub-samples of 100 students each. The period covered in his study was from the 1973/4 academic year to the 1985/6 academic year.

The main statistical tools that he used to analyse his data were multiple and correlation analyses. He used a first-order ordinary least squares multiple regression analysis model of the form;

$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_n X_{ni} + U_i$, where,

Y_i=final year of study Gross Point Average (GPA),

β_i=model parameters,

U_i=error term,

X_{ij}'s=predictor variables.

This model was to be used to produce results which had some predictive value. However a model fitting derived from a sample could only give results from one set of data. Even if the fit was significant, the predictions could not be accurate for other sets of data. To overcome such a problem, cross-validation was done. Though there was cross-validation of the data with a sample of 100 students, the researcher did not test for the equality between the sets of coefficients in the linear regressions fitted from the study samples and those derived from the cross-validation sample. If this had been done, it would have enabled the researcher to conclude whether there existed any significant differences between the two study samples themselves, or between the study samples and the cross-validation sample.[3] The scores (GPA) obtained by students in their final year were used to measure performance and as the response variable. The predictor variables used included scores attained in subjects taken at ordinary level and two demographic variables; age and sex.

His findings were that scores in Mathematics, Commerce related subjects, Science subjects , O-level aggregate score, and age were strongly correlated with performance at the university. However scores in English language, Literature, General Sciences, other languages, and the liberal arts subjects showed a very weak association with the response variable.

Onuong'a, J. O. validated his model using a sample of 100 students. He also attempted a test on the assumptions underlying his conclusions. For instance he extensively used stepwise regression to eliminate those predictor variables that were highly correlated. However, he did not test as vigorously the other assumptions on which the validity of the least squares multiple regression analysis results depend on.

2.2. Findings of Kariuki, C.N.

This is the second project to be reviewed. It was presented by Charles N. Kariuki in 1989 in partial fulfilment of the requirements for the award of the M.B.A. degree, Faculty of Commerce, University of Nairobi.

Kariuki C.N. was studying the background socioeconomic characteristics that have a significant influence on performance of entrepreneurs in the Jua Kali (Informal) sector.

In his study Kariuki used a sample of 120 entrepreneurs, divided into three Jua Kali sub-sectors (metal work, motor vehicle and carpentry) of 40 each. All the sub-samples were selected from Eastlands of Nairobi. He collected primary data through a questionnaire.

He used multiple linear regression and correlation analyses for data analysis. He identified fifteen background socioeconomic characteristics for the analysis. He carried out two regression analyses using success as the response variable. Due to the inherent difficulty in measuring success, he used the number of workers in an organisation and the monthly profits as surrogates to measure success.

For each Jua Kali sub-sector he identified a number of factors that determine performance in the Jua Kali sector. For metal work sub-sector, primary school handicraft, formal technical training, number of years of technical training, previous employment in the formal sector, and initial capital had significant association with the performance of entrepreneurs.

For motor vehicle sub-sector, level of formal education, the entrepreneur's family size, preference for formal employment at entry, and distance to entrepreneur's ancestral home had significant association with the performance.

Formal technical training, initial capital, parental family size, preference for formal employment at entry, distance from ancestral home, primary school handicraft, and market consideration had significant influence on performance of entrepreneurs in the carpentry sub-sector.

Like Onuong'a, Kariuki also attempted to validate his model results by testing for autocorrelation. To do this he had residual plots for the three Jua Kali sub-sectors and his Durbin-Watson tests indicated that there was no autocorrelation. This could have come about due to the fact that Durbin-Watson statistic is inappropriate when the degrees of freedom are less than 40 as indicated in an earlier section of this study.

For Kariuki's case his observations were only 40 and the degrees of freedom were 38. This would prompt the use of a better test for autocorrelation.

Both Onuong'a and Kariuki made the following assumptions in using the least squares regression models:

i. The predictor variables are independent of one another.

ii. The error terms are independent of one another.

iii. The variance of the error terms is constant.

iv. The expected value of the error terms , E(Ui) = 0

For some of these assumptions they did not test whether they held true or not. If they did not hold true then their findings would not be valid hence the need to test for the validity of their findings in this study.

CHAPTER THREE

3. RESEARCH DESIGN

3.1 Data collection procedures.

The population to be studied is all M.B.A projects and the criteria used to define the sample of the M.B.A. projects that will be used in this study were all those projects which;

(i) were done between 1987 and 1989.
(ii) used regression analysis tool for data analysis.
(iii) retained at least some of the data used in the analysis.

In this study, two M.B.A projects will be evaluated for their validity because they are the only ones that meet the set criteria. The two projects are one written by Onuong'a, J. O. in 1988 and the second one was written by Kariuki, C. N. in 1989. Given that this study seeks to evaluate the validity of the results obtained when regression analysis is used as a tool for data analysis the data to be used will be the same data that was used by the two researchers in carrying out their analyses. It will be appreciated that the only best way to evaluate the validity of the findings of these two studies would be to use the findings of the researchers, but also carry out tests that would confirm whether the findings are valid or not. This calls for the utilisation of exactly the same data that the researcher used hence the definition of the sample.

3.2 Data analysis.

A total of six tests will be carried out in this section of the study. These tests will be done to test the hypotheses stated earlier on. The tests fall under two main categories as was stipulated under the objectives section of the study. These categories are second-order tests and significance tests. The methodology to be followed for each of these tests is given below.

Category one: Second-order tests.

(a) Test for multicollinearity

There are many ways of testing for multicollinearity, some of which are casual and others rigorous. Three tests will be done in this case and these are, the stepwise regression method, the variance inflation factors method and finally the ridge regression method. The stepwise regression method is the most widely used of the automatic search methods. It saves on the computational effort while arriving at a reasonable set of predictor variables[4].

The variance inflation factors $(VIF)_k$ will also be used as they are a measure of how much variances of estimated regression coefficients are inflated as compared to when predictor variables are not linearly related. $(VIF)_k$ for coefficient $b_k = (1 - R^2_k)^{-1}$, where k=1, 2, 3,..., P-1, and P is the number of coefficients. R^2_k is the coefficient of multiple determination when predictor variable X_k is regressed on the other P-2 predictor variables in the model. If a variable has a $(VIF)_k$ of more than 10 then we will exclude it from the model we are attempting to validate because it is an indication of serious multicollinearity problems[5]. If VIFs of least squares are large, it is appropriate to consider a biased estimation procedure such as ridge regression in order to minimise the effects of the predictor variable correlations and develop a set of stable coefficients.

Ridge regression will also be used in the selection of predictor variables. It helps to identify variables which might be dropped from the regression model. This method uses biased estimation in data analysis and when predictor variables are highly correlated, ridge regression produces coefficients which predict and extrapolate better than least squares and is a safe procedure for selecting variables[6]. It utilises the basic fact that it is meaningful to focus on the achievement of small mean square error (MSE) as the relevant criterion, if a major reduction in variance can be obtained as a result of allowing a little bias, where MSE = variance + $(bias)^2$. We shall consider for dropping variables whose ridge trace is unstable, with the coefficient tending towards the value of zero. We shall also consider for dropping variables whose ridge trace is stable but at a very small value. Also, variables with unstable ridge traces that do not tend towards zero will also be considered for dropping.

(b) <u>Test for autocorrelation</u>

The commonest test for autocorrelation is the use of Durbin-Watson statistic. This test employs the Von Neumann ratio of the least squares estimated disturbances (Q).

The Durbin-Watson statistic may be used for this purpose but due to the fact that it may be inconclusive in situations where, $Q_L \leq Q \leq Q_u$, the method will be abandoned in favour of a more effective method called Nagar and Theil method of testing the independence of regression disturbances.

Another reason of preferring the Nagar and Thiel method over the Durbin-Watson method is because it is not effective in cases where the degrees of freedom are less than 40. Given that data in one of the projects to be evaluated has only 40 observations, then the degrees of freedom will definitely be less than 40 and therefore the Durbin-Watson method cannot be applicable.

The Nagar and Thiel method is a " one-sided" test[7].It tests the hypotheses;

Ho:Residuals are independent.

Ha:Successive disturbances are positively correlated.

The Von Neumann ratio of the least squares estimated disturbances (Q), will be utilised. This is the same ratio that is used by Durbin-Watson statistic.

$$Q = \frac{\sum [U(t)-U(t-1)]^2}{\sum U(t)^2}$$

where U(t) = estimated disturbances at period t.

U(t-1) = estimated disturbances at period t-1

This calculated Q will then be compared with the theoretical Von Neumann ratio of the least squares estimated disturbances(Q*).

$$Q^* = 2\left(\frac{T-1}{T-A} - \frac{J}{T+2}\right)$$

where T = number of observations.

J = the theoretical t-value at a specified level of significance and degrees of freedom.

A = number of coefficients in the model.

The Ho will be rejected if Q<Q*, and consequently we will fail to reject Ho if Q>Q*. It will also be of paramount importance to plot the residuals so that we may get a rough idea of whether autocorrelation exists.

There is also need to test for negative autocorrelation. This will be done by testing the following hypotheses:

Ho: Residuals are independent.

Ha: Residuals are negatively correlated.

If $D_u \leq d \leq (4-D_u)$ we fail to reject the null hypothesis of no autocorrelation.

If $d > (4-D_L)$ we reject the null hypothesis of no autocorrelation and fail to reject that there is negative autocorrelation of the first order, where,

D_u = upper limits for the significance levels of Durbin-Watson statistic.

D_L = lower limits for the significance levels of Durbin-Watson statistic.

d = empirical Durbin-Watson statistic calculated from the residuals of the model being tested.

(c) Test for heteroscedasticity

In this study two methods of testing for heteroscedasticity will be used. A preliminary plot of residuals will be done to get a feel of the existence of heteroscedasticity before a formal test is done.

This first method that will be used to detect the presence of heteroscedasticity will be a rather casual one. The squares of the residuals will be plotted against the predictor variables. If there exists a clear pattern then we will conclude that there is heteroscedasticity, that is, $\delta^2_{ui} = f(X_i)$, where δ^2_{ui}=variance of the error term U_i, and X_i=predictor variable.

The second method which is more formal and rigorous in testing whether the variance of the error terms is constant was developed by Goldfeld and Quandt [8]. It is a method which is applicable in large samples. This method tests the hypotheses:

$Ho: U_i$'s are homoscedastic.

$Ha: U_i$'s are not homoscedastic.

The steps followed in this method are as follows;

(1). Order the observations according to the magnitude of the explanatory variable X_i.

(2). Select arbitrarily a certain number of central observations which will be omitted from the analysis. The remaining observations are divided into two sub-samples of equal size, one including the small values of X_i's and the other including the large values of X_i's.

(3). Fit separate regressions for each sample and obtain the sum squared residuals from each of them, that is, $\sum e^2_1$ and $\sum e^2_2$,where

$\sum e^2_1$ =sum squared residuals from the small sub-sample.

$\sum e^2_2$ =sum squared residuals from the large sub-sample.

(4). Calculate F-ratio $F= \dfrac{\sum e^2_2}{\sum e^2_1}$

(5). The theoretical F* is the value of F that defines the critical region of the test at $V_1=V_2=(n-c-2k)/2$ degrees of freedom, where

V_1 = numerator degrees of freedom.

V_2 = denominator degrees of freedom.

n = number of observations.

c = number of central observations omitted.

k = total number of parameters in the model.

If F > F* we fail to reject that there is heteroscedasticity and reject the null hypothesis

(d) <u>Test for the assumption that $E(U_i) = 0$.</u>

This is a rather difficult assumption to test. However, its validity is so vital that an attempt to test it will be made in this study. The test will only serve as an exposition only and it cannot be claimed to be an adequate one. In this study, we will make a further one assumption when we are testing this assumption. We will assume that if $E(U_i) \neq 0$ then this will be as a result of the misspecification of the functional relationship between the variables in model and not due to the omission of some important variable(s). However we know that in practice, more often than not, the problem is caused by the omission of an important variable. The original models will be expressed as:

(1). $Y_i = a_0 + b_1X_{1i} + b_2X_{2i} + \ldots\ldots + b_nX_{ni} + U_i$

(2). $Y_i = a_0 + b_1 \ln X_{1i} + b_2 \ln X_{2i} + \ldots + b_n \ln X_{ni} + U_i$

To carry out this test, we will specify a supermodel incorporating natural logarithmic variables in the form:

(3) $Y_i = a_0 + b_1X_{1i} + b_2X_{2i} + \ldots + b_nX_{ni} + b_{n+1}\ln X_{ni+1} + b_{n+2}\ln X_{n+2i} + \ldots + b_m\ln X_{mi} + U_i$

The supermodel will not be of interest in itself but would be used to determine the adequacy of the original models. The supermodel contains the union of all predictor variables. It will be used to determine the adequacy of each alternative model. One attractive feature of the supermodel is that it can enable us to reject model 1 and 2 if need arises. If this happens, we will conclude that neither model provides an adequate representation of the relationship.

We will then find the unadjusted and the adjusted R^2 for each of the three models. The model with the highest R^2 is actually the "best". We will not be interested in interpreting the slope coefficients of model with the highest R^2 which is more superior than the one with the lower R^2. After comparing model 1 and 2 in terms of their R^2 we will then be able to choose which of the two is better in terms of its explanatory power.

For example if equation 1 has a better explanatory power than equation 2, then we will compare it with the supermodel and the test statistic will be:

$$F = \frac{(R^2_3 - R^2_1)/(DF_1 - DF_3)}{(1-R^2_3)/DF_3}$$ where,

R^2_1 = Unadjusted R^2 for model 1.

R^2_3 = Unadjusted R^2 for model 3.

DF_1 = degrees of freedom left for model 1.

DF_3 = degrees of freedom left for model 3.

The supermodel 3 allows us to use the incremental R^2 test. We will then test the hypotheses that:

Ho: $\beta_1 = \beta_2 = \ldots\ldots = \beta_n = 0$

Ha: At least one of the $\beta_i \neq 0$, where

β_i = slope coefficients of the inferior model.

If the calculated F-value exceeds the critical value of F then we will reject the null hypothesis and conclude that the better of model 1 or 2 is an inadequate specification. Based on this test we will be able to conclude that both models are inadequate and we conclude that more initial thought is required to obtain an alternative specification of the model. All this is being done to point out the importance of having the correct model specification and the inclusion of all relevant variables in our models to ensure that the assumption E(Ui)= 0 is valid.

<u>Category two: Significance tests on the beta coefficients</u>

The significance tests that will be carried out in this study will involve not only the significance of the whole model, but also the significance of the individual betas of each particular predictor variable. The main aim of this test will be to dispose of those variables in our model which do not significantly affect the response variable. The two projects whose regression results will form the basis of our analysis have employed the general multiple regression model of the form,

$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_n X_{ni} + U_i$, where;

Y_i = response variable,

β_i = parameters of the model.

U_i = error term.

X_{ji} = predictor variables

(a) <u>Test for the overall significance of beta weights.</u>

In situations where there are very many predictor variables, the interpretation of the results of the regression analysis becomes difficult. This calls for the inclusion of only those variables which have a significant effect on the response variable. To be able to eliminate insignificant predictor variables from our models we shall need to carry out tests of significance. The procedure to be followed will be to test the hypotheses; for the significance of the whole model,

$H_o: \beta_{1i} = \beta_{2i} = \beta_{3i} = \ldots = \beta_{ni} = 0.$

Ha: At least one of the beta coefficients is not equal to zero.

The decision rule is that Ho will be rejected if the calculated F-value is greater than the theoretical F*-value at certain level of significance and degrees of freedom.

(b) <u>Test for the significance of individual beta weights.</u>

For the individual variables test of significance, the hypotheses to be tested are;

Ho: $_{ji}$ =0,

Ha: $_{ji}$ is not equal to zero.

Again as was the case with the test of significance for the whole model above, the decision rule is to reject Ho if the absolute calculated |t|-value is greater than the theoretical t*-value at certain level of confidence and degrees of freedom.

For both tests the conclusions will depend on the findings of the tests. If we find that some variables are not significantly different from zero then these could be discarded from our model. However, we cannot discard any variable straight away simply because the test indicates that its coefficient is not significantly different from zero.

This is because there are many reasons why the value for a given slope coefficient may be insignificant. First, the predictor variable has an effect that is different from the functional form assumed (incorrect functional form). Second the model excludes other relevant predictor variables (omitted variables).

Third, the predictor variable is highly correlated with one or more other predictor variables included in the model (multicollinearity). Fourth, the predictor variable has no relation with the criterion variable (irrelevance).

Insignificance due to either of the first two reasons requires us to investigate the existence of superior functional forms and additional predictor variables. The third reason requires the exploration of additional data sources, model reformulation, or alternative estimation procedures. Only the fourth reason is proper justification for eliminating a predictor variable from the model.

CHAPTER FOUR

4 . DATA ANALYSIS AND FINDINGS

4.1 Introduction

This chapter will be divided into two main sections. Section A will be on testing the validity of the of the findings of Onuong'a.

Section B will have three sub-sections. Sub-section (a) will comprise of tests on the validity of Kariuki's findings on metal work sub-sector. Sub-section (b) will comprise of tests on the validity of Kariuki's findings on motor vehicle sub-sector. Sub-section (c) will comprise of tests on the validity of Kariuki's findings on carpentry sub-sector.

For the two sections , A and B , the tests outlined in research design chapter of this study will be done for each of the sections. Each of the sections or sub-sections in the case of section B will be divided into two parts. Part 1 will be on testing the validity of the least squares assumptions made by the researchers. Part 2 of each section or sub-section will be on significance tests.

4.2 <u>Section A : Findings on Onuong'a's project</u>.

The following symbols will be used throughout section 1 of the data analysis chapter whenever encountered.

Y = Actual GPA

X1 = Age

X2 = Mathematics Score

X3 = Science Score

X4 = Commerce Score

X11 = Age observation in the small values sub-sample

X12 = Age observation in the large values sub-sample

X21 = Mathematics score in the small values sub-sample

X22 = Mathematics score in the large values sub-sample

X31 = Science Score in the small values sub-sample

X32 = Science score in the large values sub-sample

X41 = Commerce score in the small values sub-sample

X42 = Commerce score in the large values sub-sample

B_1 = Coefficient of Age

B_2 = Coefficient of Mathematics score.

B_3 = Coefficient of Science score.

B_4 = Coefficient of Commerce score.

Part 1: Tests for Least Squares Regression Assumptions

(a) Tests for multicollinearity

(i) Stepwise regression method

Stepwise regressions were run for the data at 90% and 95% significance levels. The results are shown on table 1. At these two levels of significance, only Science score, (X3) was found to be significantly related to the response variable, Actual GPA.

(ii) Variance inflation factors (VIFs) method.

$VIF_{x1} = (1 - R^2)^{-1}$

Regressions of each individual predictor variable on the rest of the other predictor variables were run. The results of these regressions are shown on tables 2 through 5. The following variance inflation factors for each predictor variable were calculated:

$VIF_{x1} = (1 - 0.113854)^{-1} = 1.1285$

$VIF_{x2} = (1 - 0.0796199)^{-1} = 1.0865$

$VIF_{x3} = (1 - 0.10163)^{-1} = 1.1131$

$VIF_{x4} = (1 - 0.103556)^{-1} = 1.11552$

Conclusion

From the above variance inflation factors calculations , it can be noted that all the VIFxi's are far below 10 and based on these findings one can safely conclude that there exists no serious multicollinearity problems. In actual fact for an ideal situation when there is absolutely no multicollinearity, the VIFxi's = 1. Our findings in this case are no far from this ideal situation.

(iii) Ridge regression method

Ridge traces were obtained by plotting the ridge regression coefficients against various levels of bias constant , k, where $0 < k < 1$. The smaller the value of k the better in evaluating the stability of the coefficients of the relevant variables.The value of k to be used is however determined subjectively as there does not exist a formal way of choosing the appropriate bias to be introduced. The ridge traces plot is shown on table 6. It is evident that Mathematics score is the most stable variable at k = 0.2 while the other variables were quite unstable at this bias level. When k =0 , the ridge regression coefficients are the same as the ordinary least squares regression coefficients.

At k = 0.4 , the coefficient of variable Age showed tendencies towards stabilisation. The other coefficients namely ; Science score and Commerce score coefficients continued to be unstable even at higher levels of bias.

These results tend to conform with the findings of Onuong'a that Mathematics score had the most significant relationship with the actual GPA.

<u>Conclusion</u>

The sample validation data does not exhibit acute multicollinearity problems. However from the ridge traces on table 6, the ordinary least squares coefficients cannot be relied upon since two of the variables have very unstable coefficients even at high levels of bias. There is therefore need to evaluate the possibility of dropping two of the variables if the other assumptions of least squares regression analysis are found to hold true , but of course being extra careful as there are many other reasons as to why the ridge traces were unstable.

(b) <u>Tests for autocorrelation</u>

(i) <u>Test for positive autocorrelation</u>

Ho : Residuals are independent.

Ha : Successive disturbances are positively correlated.

$\alpha = 0.05$

Critical $D^* = 2\left(\frac{T-1}{T-A} - \frac{J}{\sqrt{T+2}}\right) = 2\left(\frac{99}{96} - \frac{1.64485}{\sqrt{102}}\right) = 1.73677$

Decision rule : We will reject Ho if $Q < Q^*$

Calculated Q is as shown in table 7 and is = 1.68101

Statistical conclusion

Since $Q < Q^*$ at 95% level of significance, we reject the null hypothesis and conclude the alternative hypothesis.

Administrative conclusion

The successive disturbances or error terms are positively correlated. This means that the least squares regression estimates computed from this data are inefficient (not BLUE) and therefore the error term calculation is wrong.

(ii) Test for negative autocorrelation

Ho : Residuals are independent.

Ha : Residuals are negatively correlated.

$\alpha = 0.05$

Critical - $D_U = 1.76$

- $D_L = 1.59$

Decision rule : We will reject Ho if $d > (4 - D_L)$ and fail to reject Ho if $D_U < d < (4 - D_U)$

Calculated d is as shown in table 7 which is = 1.68

Statistical conclusion

Since $d < (4 - D_L)$ and $D_u > d < (4 - D_u)$ then these results are inconclusive at 95% confidence level.

Administrative conclusion

It is not possible to conclude the presence or absence of negative autocorrelation in this data since the test has yielded inconclusive results. This tends to confirm the results of the test for positive autocorrelation shown above.

(c) Test for heteroscedasticity

(i) Residual plots

Tables 8 through 11 show plots of residuals plotted against the various predictor variables. For all of them , no clear pattern emerges to suggest that there is heteroscedasticity in the error terms.

(ii) The Goldfeld and Quandt test for heteroscedasticity.

Eight simple regressions were run in this test. The data was divided into two sub-samples by omitting 20 of the central observations. Half of the regressions were for the response variable on the each of the predictor variables from the small values sub-sample and the other half was from the large values sub-sample.Details on of this test's methodology are given in the data analysis section of the research methodology chapter 3.

The simple regressions models run are of the form

$Yij = Xij$, where Y = response variable

X = predictor variable

i=1,2,3,4 and it identifies the particular predictor variable. For example X2j means that we are dealing with variable X2.

j = 1 for variables in the small sub-sample and j=2 for variables in the large sub-sample.

Table 12 shows regressions of the small (a) and the large (b) sub-samples of Y on X1.

$$\sum e^2_{x11} = 1190.02$$

$$\sum e^2_{x12} = 1182.81$$

Ho : Ui's are homoscedastic

Ha : Ui's are heteroscedastic

$\alpha = 0.05$

Critical $F^*_{(95,36,36)} = 1.765$

Decision rule : If $F > F^*$ we will reject the null hypothesis.

$$\text{Calculated } F = \left(\frac{1182.81}{1190.02}\right) = 0.994$$

Statistical conclusion.

Since F < F* we fail to reject the null hypothesis that the error terms are homoscedastic for the simple regression of Y on X1.

Table 13 shows regressions of the small (a) and the large (b) sub-samples of Y on X2.

$$\sum e^2_{x21} = 1155.14$$

$$\sum e^2_{x22} = 1521.70$$

Ho : Ui's are homoscedastic

Ha : Ui's are heteroscedastic

$\alpha = 0.05$

Critical $F^*_{(95,36,36)} = 1.765$

Decision rule : If F* < F we will fail to that there is heteroscedasticity and we shall reject the null hypothesis.

$$\text{Calculated } F = \left(\frac{1521.70}{1155.14}\right) = 1.317$$

Statistical conclusion.

Since F* > F we fail to reject the null hypothesis that the error terms are homoscedastic for the simple regression of Y on X2.

Table 14 shows regressions of the small (a) and the large (b) sub-samples of Y on X3.

$$\sum e^2_{x31} = 1466.64$$
$$\sum e^2_{x32} = 1164.68$$

Ho : Ui's are homoscedastic

Ha : Ui's are heteroscedastic

$\alpha = 0.05$

Critical $F^*_{(95,36,36)} = 1.765$

Decision rule : If F > F* we will fail to reject that there is heteroscedasticity and we shall reject the null hypothesis.

$$\text{Calculated } F = \left(\frac{1164.68}{1466.44}\right) = 0.794$$

Statistical conclusion.

Since F < F* we fail to reject the null hypothesis that the error terms are homoscedastic for the simple regression of Y on X3.

Table 15 shows regressions of the small (a) and the large (b) sub-samples of Y on X4.

$$\sum e^2_{x41} = 1705.12$$
$$\sum e^2_{x42} = 1083.44$$

Ho : Ui's are homoscedastic

Ha : Ui's are heteroscedastic

$\alpha = 0.05$

Critical $F^*_{(95,36,36)} = 1.765$

Decision rule : If $F > F^*$ we will fail to reject that there is heteroscedasticity and we shall reject the null hypothesis.

$$\text{Calculated } F = \left(\frac{1083.44}{1705.12}\right) = 0.635$$

Statistical conclusion

Since $F < F^*$ we fail to reject the null hypothesis that the error terms are homoscedastic for the simple regression of Y on X4.

(d) **Test for the assumption that E(Ui) = 0.**

The following three equations were formulated for the purposes of this test.

Equation 1 : $Y = \beta_1 X1 + \beta_2 X2 + \beta_3 X3 + \beta_4 X4 + Ui$

Equation 2 : $Y = \beta_1 \text{LOG } X1 + \beta_2 \text{LOG } X2 + \beta_3 X3^2 + \beta_4 X4^2 + Ui.$

Equation 3 : $Y = \beta_1 X1 + \beta_2 X2 + \beta_3 X3 + \beta_4 X4 + \beta_5 \text{LOG } X1 + \beta_6 \text{LOG } X2 + \beta_7 X3^2 + \beta_8 X4^2 + Ui.$

Tables 16 to 18 show the regression results for each of these three models. Judging from the value of the coefficient of determination , equation 1 is better than equation 2.

Ho : $\beta_5 = \beta_6 = \beta_7 = \beta_8 = 0$

Ha : At least one of the β_i's $\neq 0$

$\alpha = 0.05$

Critical $F^*_{(.95,4,95)} = 2.479$

Decision rule : If F < F* we will fail to reject Ho and we shall conclude that model 1 is an adequate representation of the relationship of interest.

$$\text{Calculated } F = \frac{(R^2_3 - R^2_1) / (DF_1 - DF_3)}{(1 - R^2_3)/DF_3}$$

$$= \frac{(0.1824 - 0.1076) / (95 - 91)}{(1 - 0.1824) / 91}$$

Calculated F = 2.081

<u>Statistical conclusion</u>

Since F < F* then we fail to reject the null hypothesis.

<u>Administrative conclusion</u>

Model 1 is an adequate representation of the relationship of interest. We would be tempted to conclude that E(Ui) = 0 . However there is need to explore an exhaustive model formulations with various functional forms in order to form a better opinion on this assumption.

Part 2 : Significance Tests

Overall significance tests

The final optimal equation derived by Onuong'a was ,

$Y = 77.49 - 0.55X1 - 1.6X2 - 0.24X3 - 0.31X4$

Table 19 shows the multiple regression results of this optimal equation which was derived by Onuong'a.

Ho : $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$

Ha : At least one of the β_i's $\neq 0$

$\alpha = 0.05$

Critical $F^*_{(0.95,4,195)} = 2.372$

Decision rule : We will reject Ho if $F > F^*$

Calculated F (from table 19) = 24.59

Statistical conclusion

Since $F > F^*$ we reject Ho and conclude that at 95% significance level the Overall model is significant.

Testing the significance of individual coefficients

Table 19 shows the results of multiple regression of the optimal equation as was derived by Onuong'a .

Ho : $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$

Ha : $\beta_1 = \beta_2 = \beta_3 = \beta_4 \neq 0$

$\alpha = 0.05$

Critical $t^* = 1.665$

Decision rule : We will reject the null hypothesis if

$|t| > t^*$.

Calculated $|t|$ for $\beta_1 = 2.432$

$\beta_2 = 7.496$

$\beta_3 = 1.70$

$\beta_4 = 2.453$

Statistical conclusion

Since $|t| > t^*$ for all the coefficients we will reject the null hypothesis and fail to reject the alternative hypothesis.

Administrative conclusion for both significance tests.

The optimal equation derived by Onuong'a J. O. was significant on the overall and each of the individual coefficients were significantly different from zero at 95% level of significance.

4.3 Section B: Findings on Kariuki's Project

The following symbols will be used throughout section B of the data analysis chapter whenever encountered.

X_4 = Number of workers(our response variable).

X_7 = Formal education

X_8 = Primary school handicraft

X_{11} = Formal technical training

X_{13} = number of years of technical training

X_{15} = formal employment before entry

X_{16} = Preference for formal job

X_{17} = Market consideration

X_{18} = Initial capital

X_{20} = Distance to ancestral home

X_{21} = Parental family size

This section will be divided into three sub-sections for purposes of data analysis as was the case in Kariuki's project. These three sub-sections are: metal work sub-sector, motor vehicle sub-sector and carpentry sub-sector.

The following were the optimal equations derived by Kariuki for each sub-sector:

Metal work Sub-sector:

$$X_4 = 1.21 + 1.993X_8 + 2.844X_{11} + 0.021XP_{13} + 2.642X_{15} - 0.71X_{18}$$

Motor vehicle sub-sector:

$$X_4 = 5.51 - 0.882X_7 - 1.347X_{16} + 0.595X_{20} + 0.568X_{21}$$

Carpentry sub-sector:

$$X_4 = 2.169 - 1.572X_8 + 0.98X_{11} - 1.43X_{16} - 1.834X_{17} + 1.032X_{18} - 0.65X_{20} + 0.536X_{21}$$

Metal Work sub-sector

Part 1 : Tests for Least Squares Regression Assumptions

(a) Tests for multicollinearity

(i) Stepwise regression method

A stepwise regression was run at 95% significance level as shown on table 20. At this level of significance, the following predictor variables were found to be significantly related to the response variable; primary school handicraft, formal technical training; formal employment before entry and initial capital. However, the number of years of technical training was found to be insignificantly related to the response variable.

(ii) **Variance inflation factors(VIFs) method**

$$VIF_{xi} = (1 - R^2)^{-1}$$

Regressions of each individual predictor variable on the rest of the other predictor variables were run. The results of these regressions are shown on table 21 through table 25. The following variance inflation factors for each predictor variable were calculated:

$VIF_{x8} = (1 - 0.0197)^{-1} = 1.02$

$VIF_{x11} = (1 - 0.1833)^{-1} = 1.22$

$VIF_{x13} = (1 - 0.1901)^{-1} = 1.23$

$VIF_{x15} = (1 - 0.1861)^{-1} = 1.23$

$VIF_{x18} = (1 - 0.161)^{-1} = 1.19$

Conclusion

From the above variance inflation factors calculations, it can be noted that all the VIF $_{xi}$'s were far below 10 and based on these findings one can safely conclude that there exists no serious multicollinearity problems in the data on metal work sub-sector.

(iii **Ridge regression method**

Ridge traces were obtained by plotting the ridge regression coefficients against various levels of bias constant, k, where 0 < k < 1. The ridge traces plot is shown on table 27. To assist in the labeling of the ridge traces, table 26 was run.

As can be noted from table 27, only variable X13 had a stable ridge trace for all values of k. The other variables' coefficients started to stabilise at k = 0.4.

Conclusion

The ridge traces obtained on table 27 indicate in general terms that the variables included in the final metal work sub-sector equation by Kariuki have stable coefficients, that is, they do not change drastically as small variations of bias are introduced. This is an indication that multicollinearity is not a serious problem in this equation and the coefficients derived are reliable.

(b) **Tests for autocorrelation**

(i) **Test for positive autocorrelation**

Ho : Residuals are independent

Ha : Successful disturbances are positively correlated

$\alpha = 0.05$

Critical $Q^* = 2\left(\frac{T-1}{T-A} - \frac{J}{\sqrt{T+2}}\right) = 2\left(\frac{39}{35} - \frac{1.64485}{\sqrt{42}}\right) = 1.72$

Decision rule: We will reject Ho if $Q < Q^*$

Calculated Q : (from table 28) = 2.03526

Statistical conclusion

Since $Q > Q^*$ we fail to reject the null hypothesis at 95% or above test, it seems evident that the error terms are independent. This in effects means that the calculation of the error terms is not wrong and our estimates are more likely to be efficient.

(ii) Test for negative autocorrelation

Ho : Residuals are independent

Ha : Residuals are negatively correlated

$\alpha = 0.05$

Critical - $D_U = 1.79$

- $D_L = 1.23$

Decision rule : we will reject Ho if $d > (4 - D_L)$

and fail to reject Ho if

$D_U \leq d \leq (4 - D_U)$

Calculated d as shown on table 28 = 2.035.

Statistical conclusion

Since $d < (4 - D_L$ and further $D_U < d < (4 - D_U)$ Then we fail to reject Ho.

Administrative conclusion

From the above test for negative autocorrelation, we note that the residuals are independent and are not negatively correlated. This indicates that our estimates from the metal work sub-sector may be efficient.

(c) Tests for heteroscedasticity

(i) Residual plots

Tables 29 through table 33 show plots of residuals plotted against each of the various predictor variables contained in the final equation for metal work sub-sector.

Tables 29 and 32 seem to indicate that the dispersion or variability of the residuals increases as X_{15} increases from 0 to 1.

On the other hand, table 30 shows that the residuals became less dispersed from zero as X_8 and X_{11} increased from 0 to 1. Only tables 31 and 33 do not seem to portray and pattern of the residual plots.

Conclusion

Tentatively, for those residual plots which seem to indicate a clear pattern, there is a likelihood of heteroscedasticity. For the five variables, three of them have residual plots which indicate that there is a chance of there being heteroscedasticity.

(ii) **The Goldfeld and Quandt test for heteroscedasticity**

Two simple regressions were run in this test. The data was divided into two sub-samples by omitting 10 of the central observations. Details of this test are given in the data analysis section of the research methodology chapter 3. After carrying out, that procedure, it was found out that for eight out of the ten possible simple regressions which were possible, the predictor variable was constant and no regression could be done. The two regressions which were done are shown on table 34 whereby the corresponding values of number of workers Y_{41} were regressed on the smaller (a) and larger (b) values of the number of years of technical training (X_{13}).

Using information in table 34 (a) and (b), we find that

$$\sum e^2 x_{13A} = 102$$
$$\sum e^2 x_{13B} = 273.3$$

where A is small value sample and

B is large value sample.

Ho : U_i's are homoscedastic

Ha : U_i's are heteroscedastic

$\alpha = 0.05$

Critical $F^*(.95,10,10) = 2.978$

Decision rule: We will reject Ho if $F > F^*$

$$\text{Calculated } F = \frac{\sum e^2 x_{13B}}{\sum e^2 x_{13A}} = \frac{273.3}{102} = 2.679$$

Statistical conclusion

Since $F < F^*$ we fail to reject Ho at 95% level of confidence.

Administrative conclusion

From the above test we can be able to conclude that the error terms are homoscedastic. However this is only as far one variable. Number of years of technical training (X_{13}) is concerned. For the other variables which whose regression we could not run because the values of the predictor variable were constant this conclusion may not apply.

(d) **Test for the assumption that $E(U_i) = 0$**

The following three equations were formulated for the purposes of this test. It can be noted that equation 2 and three contain some interactive terms.

Equation 1 : $Y = 1.585+2.3X_8+2.9X_{11}+0.4X_{13}+2.2X_{15}-0.5X_{18}$

Equation 2 : $Y = 1.63-0.36(X_8*X_{11})+0.77(X_{13}*X_{15})$
$+ 1.5(X_8 * X_{15})$

Equation 3 : $Y = 1.53 + 2.48X_8 + 3.9X_{11}+1.4X_{13} +2.68X_{15}$
$- 0.7X_{18}+1.1(X_8 * X_{11}) - 1.23(X_{13}* X_{15})$
$- 1.1(X_8 * X_{15})$

Table 35, 36 and 37 show the regression results from which equation 1,2 and 3 respectively are derived from. From these tables, it is clear that equation 1 has a higher coefficient of determination, (0.41) than equation 2 whose $R^2 = 0.118$. Using equation 3 as the supermodel, the following test was carried out to test whether the coefficients of the interactive terms (X_8*_{11}), $(X_{13} * X_{15})$ are significant. The coefficients of those three interactive variables will be denoted respectively as β_6, β_7 and β_8.

Ho : $\beta_6 = \beta_7 = \beta_8 = 0$

Ha : At least one of the βi's = 0

α = 0.05

Critical $F^*(.95,5,34) = 2.69$

Decision rule: We shall reject Ho if $F > F^*$

$$\text{Calculated } F = \frac{(R^2_3 - R^2_1)/(DF_1 - DF_3)}{(1 - R^2_3)/DF_3}$$

$$= \frac{(0.45 - 0.41)/(34 - 31)}{(1 - 0.45)/31}$$

$$= 0.75$$

Statistical conclusion

Since $F < F^*$ we fail to reject Ho.

Administrative conclusion

Equation 1 is an adequate presentation of the relationship of interest when compared to a model with interactive terms. This indicates that $E(ui) = 0$ as the correct model form seem to have been used. The calculated estimates are therefore most likely to unbiased.

Part 2: Significance tests

Table 38 shows regression results which were used in tests (i) and (ii)below.

(i) <u>Overall significance test</u>

Ho : $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5$

Ha : At least one of the β i's $\neq$ 0

α = 0.05

Critical F*(0.95,5,34) = 2.69

Decision rule: we will reject Ho if F > F*

Calculated F (from table 38) = 4.73

<u>Statistical conclusion</u>

Since F > F* then we reject Ho at α = 0.05

<u>Administrative conclusion</u>

The overall model derived by Kariuki for Metal work subsector is significantly different from zero.

(ii) <u>Individual coefficients significance test</u>

Ho : $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5$

Ha : None of the i's = 0

α = 0.05

Critical t*(.95,39) = 1.68

Decision rule : We shall reject Ho if $|t|$ > t*

Calculated $|t|$ for β_1 = 2.36

β_2 = 2.29

β_3 = 1.36

β_4 = 2.05

β_5 = 2.60

Statistical conclusion

For β_1, β_2, β_4 and β_5 we reject the null hypothesis since $|t| > t^*$ but for β_3 we fail to reject the null hypothesis since $|t| < t^*$

Administrative Conclusion

All variables contained in Kariuki's final optimal model for metal work sub-sector except number of years of technical training (X_{13}) are individually different from zero.

Motor Vehicle Sub-Sector

Part 1: Tests for Least Squares Regression Assumptions

(a) Tests for Multicollinearity

(i) Stepwise regression method

Stepwise regressions was run at 95% and 99% significance levels as shown on table 39. At 95% level of significance, none of the variables included by Kariuki in the final model for motor vehicle subsector were significantly related to the response variable.

At 99% significance level, only distance to ancestral home (X20) was found to be significantly related to the response variable. This state of affairs could be a signal of some multicollinearity problems.

(ii) **Variance inflation factors (VIFs) method.**

$VIF_{xi}=(1-R^2)^{-1}$

Regression for each individual predictor variable on the rest of the other predictor variables were run. The results of these regressions are shown on tables 40 through 43. The following variance inflation factors for each predictor variable were calculated:

$VIF_{x8}=(1-0.114)^{-1}=1.13$

$VIF_{x16}=(1-0.036)^{-1}=1.04$

$VIF_{x20}=(1-0.062)^{-1}=1.07$

$VIF_{x21}=(1-0.173)^{-1}=1.21$

Conclusion

From the above variance inflation factors, it can be noted that all the VIF_{xi}'s are far below 10. Based on these findings, one can safely conclude that there exists no serious multicollinearity problems in the data on motor vehicle subsector.

(iii) **Ridge regression method**

Ridge traces were obtained by plotting the ridge regression coefficients against various levels of bias constant, k, where $0 < k < 1$.

The ridge traces are shown on table 45. To assist in the labeling of the ridge traces, table 44 was run also. From the ridge traces on table 45, it can be seen that for the four predictor variables, though most of them start to stabilise at k=0.2 and by k = 0.4, virtually all the variable ridge traces stabilise.

Conclusion

Due to the general stability of the variables ridge regression coefficients at relatively low values of bias (k) evident on table 45, it is clear that there exists no serious multicollinearity problems in the data on motor vehicle subsector. The estimates that he carried at for this subsector are therefore likely to be efficient and reliable.

(b) **Tests for autocorrelation**

(i) **Test for positive autocorrelation.**

H_o: Residuals are independent

H_a: Successive disturbances are positively correlated

$\alpha = 0.05$

$$\text{Critical } Q^* = 2\left(\frac{T-1}{T-A} - \frac{J}{\sqrt{T+1}}\right) = 2\left(\frac{39}{36} - \frac{1.64485}{\sqrt{42}}\right) = 1.66$$

Decision rule: We will reject the H° if Q < Q* .

Calculated Q (from table 46)=1.67

Statistical conclusion

Since Q > Q* we fail to reject the H_o at α=0.05.

(ii) **Test for negative autocorrelation**

H_o: Residuals are independent

H_a: Residuals are negatively correlated.

$\alpha = 0.05$

Critical $-D_u = 1.72$

$-D_L = 1.29$

Decision rule: We will reject H_o if $d > (4-D_L)$ and

fail to reject H_o if $D_u <= d <= (4-D_u)$

Calculated d (as shown in table 46) = 1.67

Statistical conclusion

The test for negative autocorrelation is inconclusive because $d < (4-D_L)$ and $D_u > d < (4-D_u)$.

Administrative conclusion on autocorrelation

There is no positive autocorrelation in the data on motor vehicle subsector. There is no evidence of either presence or absence of negative autocorrelation. The error terms are independent and our estimates from the data are likely to be efficient.

(c) **Test for heteroscedasticity**

(i) **Residual plots**

Tables 47 through 50 show plots of residuals plotted against each of the various predictor variables contained in the final equation for motor vehicle subsector according to

Kariuki. For all the four residual plots none of them seems to have convincing clear pattern of the residuals.

Conclusion

None of the residual plots has a convincing and clear pattern that indicates that that the error term variance is not constant.

(iii) The Goldfeld and Quandt test for heteroscedasticity

Two simple regressions were run in this test. The data was divided into two sub-samples by omitting 10 of the central observations.Details of this test are given in the data analysis section of the research methodology chapter.After carrying out the laid down procedure on this test , only two out of the eight intended simple regressions could be done. The reason behind this is that it is not meaningful to regress the criterion variable on a constant predictor variable.

The two regressions are shown on table 51 whereby the corresponding values of number of workers (Y_{41}) were regressed on the smaller (a) and larger (b) values of the distance from ancestral home (X_{20}). Using information in table 51 (a) and (b), we find that;

$$\sum e^2 X_{20A} = 58.72$$

$$\sum e^2 X_{20B} = 137.69$$

Ho : Ui's are homoscedastic

Ha : Ui's are heteroscedastic

$\alpha = 0.05$

Critical $F^*_{(.95,10,10)} = 2.978$

Decision rule : We will reject Ho if $F > F^*$

Calculated $F = \frac{\sum e^2 X_{20B}}{\sum e^2 X_{20A}} = \frac{137.69}{58.72} = 2.345$

Statistical decision

Since $F < F^*$ we fail to reject Ho at 95% level of confidence.

Administrative conclusion

From the above test we can be able to conclude that the error terms are homoscedastic. This implies that the estimates we compute from the data on motor vehicle sub-sector are most likely to be efficient in the absence of heteroscedasticity.

(d) **Test for the assumption that E(Ui) = 0.**

For purposes of this test, the following equations were formulated;

Model 1 : $Y = 4.78 - 0.62X_7 - 1.05X_{16} + 0.48X_{20} + 0.52X_{21}$

Model 2 : $Y = 5.86 - 0.60(X_7.X_{16}) + 0.002(X_{20}.X_{21})$

Model 3 : $Y = 3.07 + 0.77X_7 + 2.68X_1 + 0.36X_{20} - 0.46X_{21}$
$-1.55(X_7.X_{16} + 0.002(X_{20}.X_{21})$.

Tables 52 , 53 and 54 show the results from which equation 1, 2 and 3 respectively were derived from. From these three tables it is clear that model 2 has a better explanatory power ($R^2_2 = 0.249$) than model 1 whose $R^2_1 = 0.202$. This does indicate that the original model 1 without interaction terms as used by Kariuki was not adequate. This can also be proved by using a more rigorous statistical test as shown below:

Let β_5 and β_6 represent coefficients of the interactive terms , $(X_7.X_{16})$ and $(X_{20}.X_{21})$ in model 3 respectively.

Ho : $\beta_5 = \beta_6 = 0$

Ha : At least one of the β_i's = 0

$\alpha = 0.05$

Critical $F^*_{(.95,4,35)} = 2.65$

Decision rule : We will reject Ho if $F > F^*$

$$\text{Calculated } F = \frac{(R^2_3 - R^2_1)/(DF_1 - DF_3)}{(1-R^2_3)/DF_3}$$

$$= \frac{(0.334-0.202)/(35-33)}{(1-0.334)/33}$$

$$= 3.27$$

Statistical conclusion

Since F > F* we reject the Ho at 95% level of confidence.

Administrative conclusion

The original model formulation without interactive terms as formulated by Kariuki, C. N. was inadequate. This means that E(Ui) = 0 and therefore any estimates calculated out of this data on motor vehicle sub-sector would be biased and therefore not BLUE. This would call for a better model formulation on this subsector or an inclusion of other variables in the model which were previously not included in the model.

Part 2: Significance Tests

Table 55 shows regression results which were used to carry out tests (i) and (ii) below :-

(i) Overall significance test

Ho : $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$

Ha : at least one of the β i's = 0

$\alpha = .05$

Critical $F^*_{(.95,4,35)} = 2.65$

Decision rule : we will reject Ho if F > F*.

Calculated F (from table 55) = 2.21

Statistical decision

Since $F < F^*$ we fail to reject the null hypothesis at 95% confidence level.

Administrative decision

The overall model is not significantly different from zero at 95% confidence level.

(ii) **Individual coefficients significance test**

Ho : $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$

Ha : None of the $\beta_{i's}$ = 0

$\alpha = 0.05$

Critical $t_{(.95,39)} = 1.68$

Decision rule : we shall reject Ho if $|t| > t^*$

Calculated t for $\beta_1 = 1.32$

$\beta_2 = 1.19$

$\beta_3 = 1.99$

$\beta_4 = 1.57$

Statistical conclusion

For β_1, β_2 and β_4, we fail to reject the null hypothesis and conclude the alternative hypothesis for β_3.

Administrative conclusion

Only distance to ancestral home (X_{20}) is significantly different from zero at 95% confidence level. The other variables are not significantly different from zero.

Carpentry sub-sector

Part 1 : Tests for Least Squares Regression Assumptions

(a) Tests for multicollinearity

(i) Stepwise regression method

Two stepwise regressions were run on the carpentry sub-sector data using those variables which were included in Kariuki's final model on this sub-sector. One of the regressions was run at 95% confidence level and the other at 99% confidence level . The results of these two regressions are found on table 56.

It was found that only initial capital (X_{18}) had significant relationship with the response variable, number of workers (X_4) . This could be an indication of some multicollinearity problems which could be confirmed in test a (ii) below .

(ii) Variance inflation factors method

$$VIF_{x_i} = (1-R^2)^{-1}$$

Regressions for each individual predictor variable on the rest of the other predictor variables were done. The results of these regressions are shown on tables 57 through 63 . The following variance inflation factors for each predictor variable were calculated :

$VIF_{x8} = (1-0.212)^{-1} = 1.27$

$VIF_{x11} = (1-0.155)^{-1} = 1.18$

$VIF_{x16} = (1-0.174)^{-1} = 1.21$

$VIF_{x17} = (1-0.083)^{-1} = 1.09$

$VIF_{x18} = (1-0.133)^{-1} = 1.15$

$VIF_{x20} = (1-0.141)^{-1} = 1.16$

$VIF_{x21} = (1-0.026)^{-1} = 1.03$

Conclusion

From the variance inflation factors,it can be noted that all the VIF_{xi}'s are far below 10. Based on these findings, one can safely conclude that there exists no serious multicollinearity problems in the carpentry data used to arrive at the final optimal model for this sub-sector by Kariuki.

(iii) **Ridge regression method**

Ridge traces were obtained by plotting the ridge regression coefficients against various levels of bias constant k , where $0 < k < 1$. The ridge traces are shown on table 65. To assist in the labeling of the ridge traces on table 65 , table 64 was run.

From the ridge traces on table 65, it can be seen that for the seven predictor variables included in Kariuki's carpentry sub-sector optimal model , six of them had stable coefficients at as low values of k as 0.2 . However the coefficient of initial capital (X18) was unstable for all levels of the bias constant k.

Conclusion

Due to the general stability of the coefficients of most of the predictor variables at low values of k as evident on table 65 , it can be concluded that there exists no serious multicollinearity problems in the data on carpentry sub-sector. Even though the initial capital (X18) has an unstable ridge coefficient , the values of this variable are greater than zero and could still be retained as being significant.

(b) **Tests for autocorrelation**

(i) **Test for positive autocorrelation**

Ho :Residuals are independent

Ha :Residuals are positively correlated

$\alpha = 0.05$

$$\text{Critical } Q^* = 2\left(\frac{T-1}{T-A} - \frac{J}{\sqrt{T+1}}\right) = 2\left(\frac{39}{33} - \frac{1.64485}{\sqrt{42}}\right)$$

$=1.86$

Decision rule:we will reject Ho if $Q < Q^*$

Calculated Q (from table 66) = 2.48

Statistical conclusion

Since $Q > Q^*$, we fail to reject Ho at 95% confidence level.

(ii) Test for negative autocorrelation.

Ho : Residuals are independent

Ha :Residuals are negatively correlated.

$\alpha = 0.05$

Critical Q* - $D_u = 1.72$

- $D_L = 1.29$

Decision rule: we will reject Ho if $d > (4-D_L)$ and fail to reject Ho if $D_u < d < (4-D_u)$

Calculated d (as shown on table 66) = 2.48

Statistical conclusion

The test for negative autocorrelation is inconclusive because $d < (4-D_u)$ and $D_u < d > (4-D_L)$.

Administrative conclusion on autocorrelation

There is no positive autocorrelation on the data on carpentry sub-sector.There is also no evidence of either the presence or absence of negative autocorrelation. Our error terms can be said to be independent and therefore our estimates from the data are likely to be efficient.

(c) Tests for heteroscedasticity

(i) Residual plots

Tables 67 through 73 show plots of residuals plotted against each of the various predictor variables included in the final model for carpentry sub-sector according to Kariuki.For all the seven residual plots none of them has a clear pattern to indicate that there is heteroscedasticity.

(ii) **The Goldfeld and Quandt test for heteroscedasticity**

Two simple regressions were run in this test. The data was divided into two sub-samples by omitting 10 of the central observations . Details of this test are given in the data analysis section of the research methodology chapter. After carrying out the laid down procedures on this test ,only two out of the 14 intended simple regressions could be done. The reason behind this is that on sorting out the predictor variable values in an ascending order and the omitting the 10 central observations, it was found out that 12 of the sub-samples had a constant value for its predictor variable.

The two regression results are shown on table 74 whereby the corresponding values of the number of workers (Y18i) were regressed on the larger (a) and the smaller (b) values of the Initial capital (X18) . Using information in table 74 (a) and (b) , we find that

$$\sum e^2_{x18A} = 72.86$$

$$\sum e^2_{x18B} = 104.4$$

Ho : Ui's are homoscedastic

Ha : Ui's are heteroscedastic

$\alpha = 0.05$

Critical $F^*_{(0.95,8,8)} = 3.44$

Decision rule :we will reject Ho if F > F*

Calculated F = $\frac{\sum e^2_{x18B}}{\sum e^2_{x18A}} = \frac{104.4}{72.86} = 1.43$

Statistical conclusion

Since F < F* we fail to reject Ho at 95% level of significance.

Administrative conclusion

From the above two tests (i) and (ii) , we can be able to conclude that the error terms are homoscedastic . This implies that the estimates that were computed from the data on the carpentry sub-sector were most likely to be efficient in the absence of heteroscedasticity.

(d) **Test for the assumption that E(Ui) = 0**

For purposes of this test , the following equations were formulated ;

Model 1 : Y = 5.2 -1.5X8 +1.3X11 - 1.3X16 - 1.4X17 + 0.97X18 - 0.57X20 + 0.21X21

Model 2 : Y = 5 - 0.55(X8.X11) - 2.6(X16.X17)

Model 3 : Y = 3.5 - 1.5X8 +1.1X11 -2.2X16 - 0.7X17 + 1.1X18 - 0.5X20 + 0.2X21 - 0.7(X8.X11)

Tables 75 ,76 ,and 77 show the results from which model 1 ,2, and 3 respectively were derived from .From these three tables , it is clear that model 1 has a better explanatory power (R^2_1=0.476) than model 2 whose R^2_2 = 0.05.

The following test was carried out to test whether the coefficients of (X8*X11) and (X16*X17) in the supermodel are significant. Let β_8 and β_9 be the coefficients of (X8.X1) and (X16.X17) respectively.

Ho : $\beta_8 = \beta_9 = 0$

Ha : At least one of the B_i's = 0.

α = 0.05

Critical $F^*_{(0.95,7,32)}$ = 2.33

Decision rule : we will reject Ho if F > F*

$$\text{Calculated } F = \frac{(R^2_3 - R^2_1)/(DF_1 - DF_3)}{(1 - R^2_3)/DF_3}$$

$$= \frac{(0.5 - 0.46)/(32-29)}{(1-0.5)/29}$$

$$= 0.77$$

Statistical conclusion

Since F < F* we fail to reject Ho at 95% confidence level.

Administrative conclusion

The original model formulation without interactive terms as formulated by Kariuki on the carpentry sub-sector was adequate.This means that E(Ui) = 0 and therefore any estimates computed from this data would not be biased.

Part 2 : Significance Tests

Table 78 shows regression results which were used to carry out tests (i) and (ii) below

(i) Overall significance test

Ho : $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$

Ha : At least one of the $\beta_{i's} \neq 0$

$\alpha = 0.05$

Critical F*(.95,7,32) =2.33

Decision rule: We will reject Ho if F > F*

Calculated F (from table 78) = 4.144

Statistical decision

Since F > F* we reject Ho at 95% level of confidence.

Administrative conclusion

The overall carpentry sub-sector model is significant.

(ii) Individual coefficients significance tests

Ho : $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$

Ha : None of the $\beta_{i's} \neq 0$

$\alpha = 0.05$

Critical t* (.95,39) = 1.68

Decision rule: We shall reject Ho if $|t| > t^*$.

Calculated $|t|$ for $\beta_1 = 1.71^*$

$\beta_2 = 1.52$

$\beta_3 = 1.38$

$\beta_4 = 1.67$

$\beta_5 = 4.24^*$

$\beta_6 = 2.24^*$

$\beta_7 = 1.41$

* Significant at 95% confidence level

Administrative Conclusion

Out of the seven predictor variables that were included in the final optimal model on carpentry sub-sector by Kariuki, only primary school handicraft(X8), initial capital (X18) and distance to ancestral home (X20) were found to be significantly different from zero. The other four variables were found to be insignificant in the model.

CHAPTER FIVE

5. CONCLUSIONS AND RECOMMENDATIONS

5.1 Summary

Most econometric problems can be characterised as situations in which one or more of the four assumptions tested in this study are violated.Once the researcher finds that one or more of the assumptions are violated,he has to determine whether the ordinary least squares estimators retain their desirable properties such as unbiasedness and minimum mean square error to a larger extent or not in order to evaluate their validity .

It is against this background that it is thought wise to evaluate the relative importance of each of these assumptions on the general validity of the results obtained by Onuong'a and Kariuki in their studies in case one of them has been violated.The violation of at least one of these assumptions is almost obvious in a real life situation. Therefore it is important to mention at this juncture which of the assumptions form the corner-stones of the validity of our ordinary least squares (OLS) models.

Briefly,the most important of all the four assumptions is that of zero expected disturbance.The violation of this assumption leads to the departure from the key desirable properties of the OLS models.The violation of this assumption implies that either the model formulation has been misspecified or some important variables have been omitted from the model or both of them. For instance,the violation of this assumption leads to the production of biased estimators and also, the expected value of the disturbance or error term is not constant. The disturbance term varies with the omitted predictor variable. This causes all the OLS estimates to be inefficient. Further, the specification error of the omitted variable causes the estimates to be biased.

The next in importance is the assumption that the predictor variables are independent of one another. This implies that there should be no exact linear relationships between the predictor variables and that there are at least as many observations as the predictor variables. If either half of this assumption is violated, it is mechanically impossible to compute the OLS estimates, that is, the estimating procedure simply breaks down for mechanical reasons, just as if someone attempted to divide by zero.

Cases of exact linear relationships between the predictor variables are rare. However, approximate linear relationships among predictor variables are very common among economic or social oriented variables. Although the OLS estimator in the presence of multicollinearity remains still unbiased and in fact Best Linear Unbiased Estimators (BLUE) and the R^2 statistic is unaffected, the major undesirable consequence of multicollinearity is that the variances of the OLS estimates of the parameters of the collinear variables are quite large. The effect of each of the predictor variables on the response variable cannot be discerned clearly because the parameters computed give a mixed effect of the predictor variables amongst themselves and on the response variable. The OLS estimating procedure is not given enough independent variation in a variable to calculate with confidence the effect it has on the response variable.

The other two assumptions ; independent error terms and constant variance of the error terms do not have as much impact on the validity of the results of OLS regression analysis as the first two discussed above. It is unlikely that there will be serious heteroscedastic problems if the assumption that there is zero expected disturbance holds true. On the other hand, in economic situations, autocorrelation is rarely absent. This is because observations in real life are hardly unrelated in one way or the other. Both of these two assumptions once violated only affect the efficiency of our estimators but not their unbiasedness.

5.2 Conclusion On Onuong'a's Findings

The various tests done on the findings of Onuong'a indicated that all the four assumptions except that on autocorrelation had not been violated. The tests on the significance of the optimal model and each of the individual predictor variables indicated that they were all statistically significant. As indicated earlier-on , autocorrelative situations in real life are not uncommon. This case cannot be taken to be an unusual one as it is common in economics or other social sciences for a variable to be affected by its own value in the previous periods.

It can therefore be concluded safely that the coefficients of the variables used by Onuong'a in his optimal regression model were statistically significant. It can also be concluded that the assumptions of OLS regression analysis made by the researcher held true safe for autocorrelation. This then implies that the conclusions reached by Onuong'a on the pre-entry performance predictors for Bachelor of Commerce Students at the University of Nairobi were to a larger extent valid .

<u>**Policy Recommendations**</u>

The findings of Onuong'a would definitely be of interest to the University of Nairobi academia and administration in their endeavour to figure out the best admission criteria for the students wishing to take a course leading to the award of Bachelor of Commerce degree. The findings of this study have not revealed any notion to suggest that Onuong,a,s findings were not valid either because his predictor variables were not significantly related to the criterion variable or because the premises of his least squares model did not hold true . It is therefore recommended that his findings could lay a good basis to the solution of the Faculty of Commerce student admission criteria. It should be noted that this recommendation is subject to the limitations noted by Onuong'a in his study.

5.3 Conclusion on Kariuki's Findings

5.3.1 Metal Work Sub-Sector

Conclusions

The tests carried out on the validity of the conclusion arrived at by Kariuki on the apriori determinants of performance of entrepreneurs in metal work sub-sector revealed that the assumptions of OLS were not violated. The significance tests carried out indicated that the overall optimal model and each of the individual parameters were statistically significant.

It follows therefore that based on the results of the tests carried out, the findings of Kariuki and the conclusions thereon as concerns this sub-sector were valid as the tests performed did not indicate any serious issue to purport the contrary.

Policy Recommendations

For the users of Kariuki's findings there exists some comfort in using these findings for the entrepreneurs in this sub-sector. The tests which were carried out for this sub-sector indicated that the parameters of the model were significant and the assumptions of OLS stood unflouted. This means that the findings of Kariuki on the apriori determinants of performance of entrepreneurs in this sub-sector are valid. The funding of the potential entrepreneurs in this sub-sector by the various interested parties such as the government and the other non-governmental organisations could therefore be based on the factors identified by Kariuki.

5.3.2 Motor Vehicle Sub-Sector

Conclusions

The data analysis on the findings of Kariuki on this sub-sector indicated that the absence of the following problems in the data used ; multicollinearity, autocorrelation, and heteroscedasticity. However the analysis indicated that a very important assumption had been flouted. This was the assumption of zero expected disturbances. The analysis revealed that a wrong or rather an inadequate model formulation had been used for this sub-sector.

It therefore follows that any estimators calculated for this sub-sector are biased to a greater extent. This problem also proliferates into other undesirable properties of ordinary least squares models like inefficient estimators, that is, the variance of the error terms is not minimum.

On the significance of the coefficients of the variables included in the optimal model on this sub-sector, the overall model was found to be statistically significant while only one out of the four variables included in the optimal model on this sub-sector by Kariuki was found not to be statistically significant.

Policy Recommendations

The findings on Kariuki's conclusions on this sub-sector suggest that the validity of his findings is at stake. Given the serious attention that the informal sector commonly referred to as "Jua Kali" in Kenya is receiving currently there is a lot of temptation for those parties who may be interested in the promotion of this sector to use any research findings that they can lay their hands on. A word of caution is given here that ,although the findings of Kariuki on this sub-sector could be of some use , they should be taken with pinch of some salt . We should not discard them outrightly but at the same time they should not be taken as gospel truth.It is recommended that they be used in conjunction with other known facts on this sub-sector.

5.3.3 Carpentry Sub-Sector

Conclusions

The tests carried out on the validity of the conclusion arrived at by Kariuki on the apriori determinants of performance of entrepreneurs in carpentry sub-sector revealed that the assumptions of OLS had not been violated. The significance tests carried out indicated that the overall optimal model and each of the individual parameters were statistically significant.

It follows therefore that based on the results of the tests carried out, the findings of Kariuki and the conclusions thereon as concerns this sub-sector were valid as the tests performed did not indicate any serious issue to purport the contrary.

Policy recommendations

For the users of Kariuki's findings there exists some comfort in using these findings on the entrepreneurs in this sub-sector. The tests which were carried out for this sub-sector indicated that the parameters of the model were significant and the assumptions of OLS stood unflouted.This means that the findings of Kariuki on the apriori determinants of performance of entrepreneurs in this sub-sector are valid. The funding of the potential entrepreneurs in this sub-sector by the various interested parties such as the government and the other non-governmental organisations could therefore be based on the factors identified by Kariuki.

5.4 <u>Limitations of This Study</u>

1. Data analysis on Onuong'a's findings was done based on the sample model validation data which the researcher used. The actual data which was used by Onuong'a was not available as it could not be obtained. This data was used on the strength of the researcher's assertion that the validation of model revealed that the sample data was derived from the same population as the actual sample used to compute the model parameters. This presents a limitation on the reliability of the results obtained from the analysis.

2. There are many more tests which would have been apt in this study. Due to the pressure of time, the study concentrated on only those tests which in the opinion of the researcher were more important than the rest. However no matter how auxiliary a test may seem relative to the others, it is still very important in bringing out the exact picture that the researcher had sought to achieve.

3. The tests carried out in this study revealed some violations of the basic OLS assumptions. Remedies exist to these problems but would require extensive amount of time to research on their appropriate solutions. Thus,although it would have been more desirable to find the remedies to the assumptions violated, this was not possible given the inherent want of time.

5.5 Suggestions For Further Research

An analysis on the findings of Onuong'a have indicated that there exists autocorrelation in his data. This has the effect of rendering the OLS estimators inefficient which is an undesirable property. This is an indication that some variables included in Onuong'a's final optimal model in his study have their values affected by their previous values. It stands clear that a students performance in a particular course may be a function of many factors, some present and others having a past bearing.

Thus further research is required to unearth the relationship between student's performance and other time series related factors. Some variable transformations will be definitely inevitable together with lagging of some variables to see the timing effect on the model results.

The so-called Jua Kali sector has in the resent past received a lot of emphasis both by the government of Kenya and a large number of non-governmental organisations (NGOs). This sector is seen to be a major source of employment in most of the developing countries , Kenya included. Kariuki's study must have come at an opportune moment and his findings are therefore of major importance.

The analysis carried out in this study to determine the validity of his conclusions on each of the three Jua Kali sub-sectors he studied indicated that his conclusions on the metal work and the carpentry sub-sectors were valid as the assumptions of OLS were found to hold true while the parameters of his models were statistically significant.

On the other hand the violation of the zero expected disturbances assumption in the case of motor vehicle sub-sector rendered his conclusions on this sub-sector subject to doubt and therefore the conclusions arrived at cannot be said to be valid. Further research is required on the correct model formulation on this sub-sector. The violation of this assumption could also be attributed to the omission of some very vital variables.

Looking at the explanatory power of the three models none of the models derived by Kariuki and the one derived by Onuong'a had an R^2 greater than 0.6. This indicates that some variables explaining over 50% of the determinants of performance of Jua Kali sector entrepreneurs were omitted. There is need therefore to research on all the possible predictor variables which must have been omitted. In fact although tests were carried out on whether the assumption of zero expected disturbances held true this was only as far as model specification was concerned. Other possible causes of violation of this model like the omission of vital variables were not investigated and they remain fertile areas for future research.

FOOTNOTES

1. Wittink,D.R.,The Application of Regression Analysis.Allyn and Bacon,Inc.,Boston,1988,pp.1.

2. Ibid pp.176.

3. Gujarati,D. Use of Dummy Variables in Testing for Equality Between Sets of Coefficients in Linear Regressions:A Generalisation.The American Statistician,Vol.24, No.5,December 1970,pp.18-22.

4. Wasserman,W.etal,Applied Linear Statistical Models. Irwin,Homewood,Illinois,1985,pp.392.

5. Ibid pp. x.

6. Marquardt,D.W. and Snee,R.D."Ridge Regression in Practice". The American Statistician.February 1975,Vol.29,No.1.

7. Thiel,H.,and Nagar,A.L,"Testing the Independence of Regression Disturbances".Journal of the American Statistical Association. 56 (1961),pp.763-806.

8. Goldfeld,S.M.,and Quandt,R.E.,"Some Tests for Homoscedasticity".Journal of American Statistical Association,Vol.60,1965,pp.539-47.

Table 1

(a) Stepwise Regression at 95% confidence Level

```
                              STEPWISE REGRESSION
SELECTION:     FORWARD                                     CONTROL:      AUTOMATIC
F-TO-ENTER =  2.713          MAX STEPS =  50               F-TO-REMOVE = 2.713
                                      STEP 1
R-SQUARED = 0.0905206
R-SQUARED (ADJ.) = 0.0812402                     MSE = 33.8527 WITH 98 D.F.
 VARIABLES CURRENTLY IN MODEL               VARIABLES CURRENTLY NOT IN MODEL
VARIABLE                  COEFF.  F-REMOVE VARIABLE         PARTIAL CORR.   F-ENTER
  3. X3                 -1.07591    9.7540  1. X1                  -.0758     .5612
                                            2. X2                  -.0880     .7562
                                            4. X4                   .0552     .2968
```

(b) Stepwise Regression at 99% Confidence Level

```
                              STEPWISE REGRESSION
SELECTION:     FORWARD                                     CONTROL:      AUTOMATIC
F-TO-ENTER =  4.87           MAX STEPS =  50               F-TO-REMOVE = 4.87
                                      STEP 1
R-SQUARED = 0.0905206
R-SQUARED (ADJ.) = 0.0812402                     MSE = 33.8527 WITH 98 D.F.
 VARIABLES CURRENTLY IN MODEL               VARIABLES CURRENTLY NOT IN MODEL
VARIABLE                  COEFF.  F-REMOVE VARIABLE         PARTIAL CORR.   F-ENTER
  3. X3                 -1.07591    9.7540  1. X1                  -.0758     .5612
                                            2. X2                  -.0880     .7562
                                            4. X4                   .0552     .2968
```

Table 2

Regression of Age on the Other Predictor Variables

MODEL FITTING RESULTS

VARIABLE	COEFFICIENT	STND. ERROR	T-VALUE	PROB(>|T|)
CONSTANT	20.442445	0.766112	26.6834	.0000
X2	0.151654	0.142095	1.0673	.2884
X3	-0.229971	0.141349	-1.6270	.1069
X4	0.418035	0.132771	3.1485	.0022

0 CASES WITH MISSING VALUES WERE EXCLUDED.

RESIDUALS PLACED IN VARIABLE: RESIDUALS

ANALYSIS OF VARIANCE FOR THE FULL REGRESSION

SOURCE	SUM OF SQUARES	DF	MEAN SQUARE	F-RATIO	PROB(>F)
MODEL	64.892200	3	21.630733	4.111429	.008630
ERROR	505.06780	96	5.26112		
TOTAL (CORR.)	569.96000	99			

R-SQUARED = 0.113854
R-SQUARED (ADJ. FOR D.F.) = 0.0861619
STND. ERROR OF EST. = 2.29371

Press ENTER to continue.

1HELP 2LABEL 3SAVSC 4RECORD 5 6 7 8 9REVIEW 10QUIT
INPUT SAT JUN 23 1990 09:20:00 PM VERSION 1.1 REC:OFF

Table 3

Regression of Mathematics Score on Other Predictor Variables

MODEL FITTING RESULTS

VARIABLE	COEFFICIENT	STND. ERROR	T-VALUE	PROB(>|T|)
CONSTANT	0.851307	1.584659	.5372	.5923
X1	0.077322	0.072448	1.0673	.2884
X3	0.272048	0.098472	2.7627	.0068
X4	-0.016024	0.099566	-.1609	.8725

0 CASES WITH MISSING VALUES WERE EXCLUDED.

RESIDUALS PLACED IN VARIABLE: RESIDUALS

ANALYSIS OF VARIANCE FOR THE FULL REGRESSION

SOURCE	SUM OF SQUARES	DF	MEAN SQUARE	F-RATIO	PROB(>F)
MODEL	22.276863	3	7.425621	2.768246	.045888
ERROR	257.51314	96	2.68243		
TOTAL (CORR.)	279.79000	99			

R-SQUARED = 0.0796199
R-SQUARED (ADJ. FOR D.F.) = 0.0508581
STND. ERROR OF EST. = 1.63781

Press ENTER to continue.

1HELP 2LABEL 3SAVSC 4RECORD 5 6 7 8 9REVIEW 10QUIT
INPUT SAT JUN 23 1990 09:22:00 PM VERSION 1.1 REC:OFF

Table 4

Regression of Science Score on Other Predictor Variables

MODEL FITTING RESULTS

VARIABLE	COEFFICIENT	STND. ERROR	T-VALUE	PROB(>|T|)
CONSTANT	5.453506	1.482106	3.6796	.0004
X1	-0.116682	0.071717	-1.6270	.1069
X2	0.270724	0.097992	2.7627	.0068
X4	0.135834	0.098364	1.3809	.1704

0 CASES WITH MISSING VALUES WERE EXCLUDED.

RESIDUALS PLACED IN VARIABLE: RESIDUALS

ANALYSIS OF VARIANCE FOR THE FULL REGRESSION

SOURCE	SUM OF SQUARES	DF	MEAN SQUARE	F-RATIO	PROB(>F)
MODEL	28.989864	3	9.66328	3.620054	.015876
ERROR	256.26014	96	2.66938		
TOTAL (CORR.)	285.25000	99			

R-SQUARED = 0.10163
R-SQUARED (ADJ. FOR D.F.) = 0.0735556
STND. ERROR OF EST. = 1.63382

Press ENTER to continue.

1HELP 2LABEL 3SAVSC 4RECORD 5 6 7 8 9REVIEW 10QUIT
INPUT SAT JUN 23 1990 09:25:00 PM VERSION 1.1 REC:OFF

Table 5

Regression of Commerce Score on Other Predictor Variables

MODEL FITTING RESULTS

VARIABLE	COEFFICIENT	STND. ERROR	T-VALUE	PROB(>|T|)
CONSTANT	-2.820631	1.600935	-1.7619	.0812
X1	0.223901	0.071113	3.1485	.0022
X2	-0.016833	0.104593	-.1609	.8725
X3	0.14339	0.103836	1.3809	.1704

0 CASES WITH MISSING VALUES WERE EXCLUDED.

RESIDUALS PLACED IN VARIABLE: RESIDUALS

ANALYSIS OF VARIANCE FOR THE FULL REGRESSION

SOURCE	SUM OF SQUARES	DF	MEAN SQUARE	F-RATIO	PROB(>F)
MODEL	31.249559	3	10.416520	3.696593	.014435
ERROR	270.51554	96	2.81787		
TOTAL (CORR.)	301.76510	99			

R-SQUARED = 0.103556
R-SQUARED (ADJ. FOR D.F.) = 0.075542
STND. ERROR OF EST. = 1.67865

Press ENTER to continue.

1HELP 2LABEL 3SAVSC 4RECORD 5 6 7 8 9REVIEW 10QUIT
INPUT SAT JUN 23 1990 09:27:00 PM VERSION 1.1 REC:OFF

Table 6

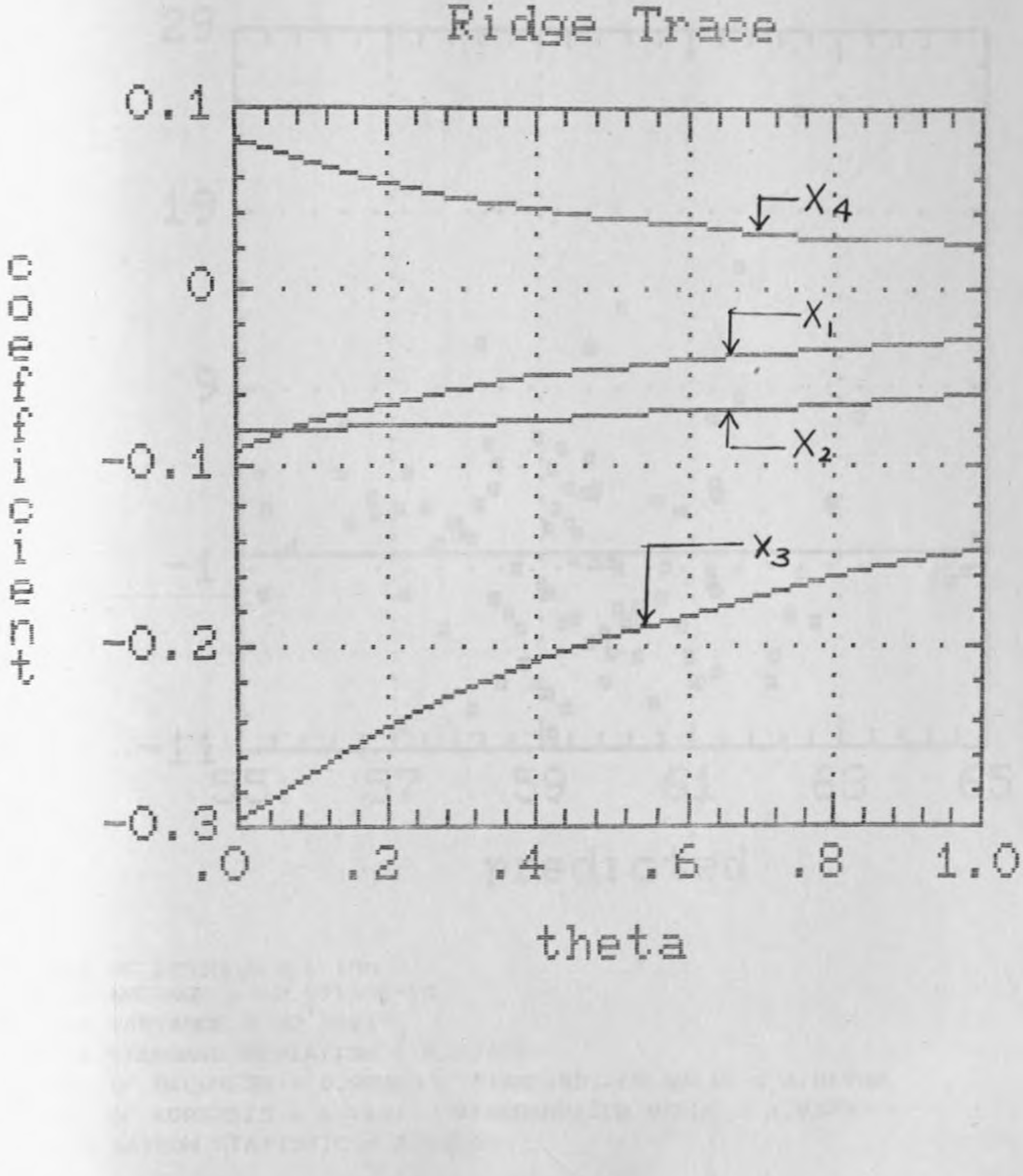
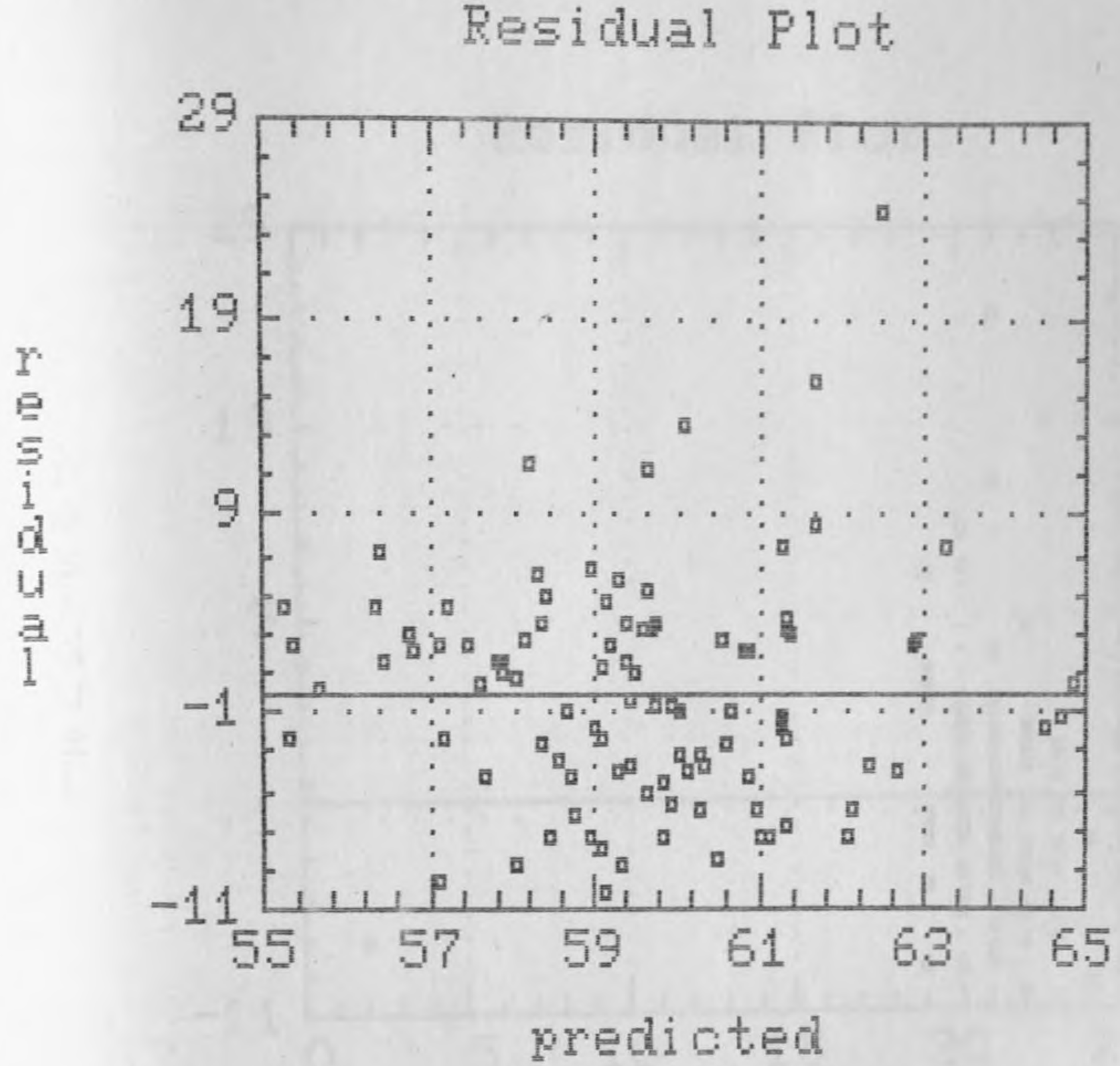

Ridge Trace
0.1
0
-0.1
-0.2
-0.3
.0
.2
.4
.6
.8
1.0
theta
Coefficient
X4
X1
X2
X3

Table 7



NUMBER OF RESIDUALS = 100
SAMPLE AVERAGE = -2.99138E-14
SAMPLE VARIANCE = 32.8821
SAMPLE STANDARD DEVIATION = 5.73429
COEFF. OF SKEWNESS = 0.98465 STANDARDIZED VALUE = 4.01982
COEFF. OF KURTOSIS = 5.4191 STANDARDIZED VALUE = 4.93797
DURBIN-WATSON STATISTIC = 1.68101

Table 8

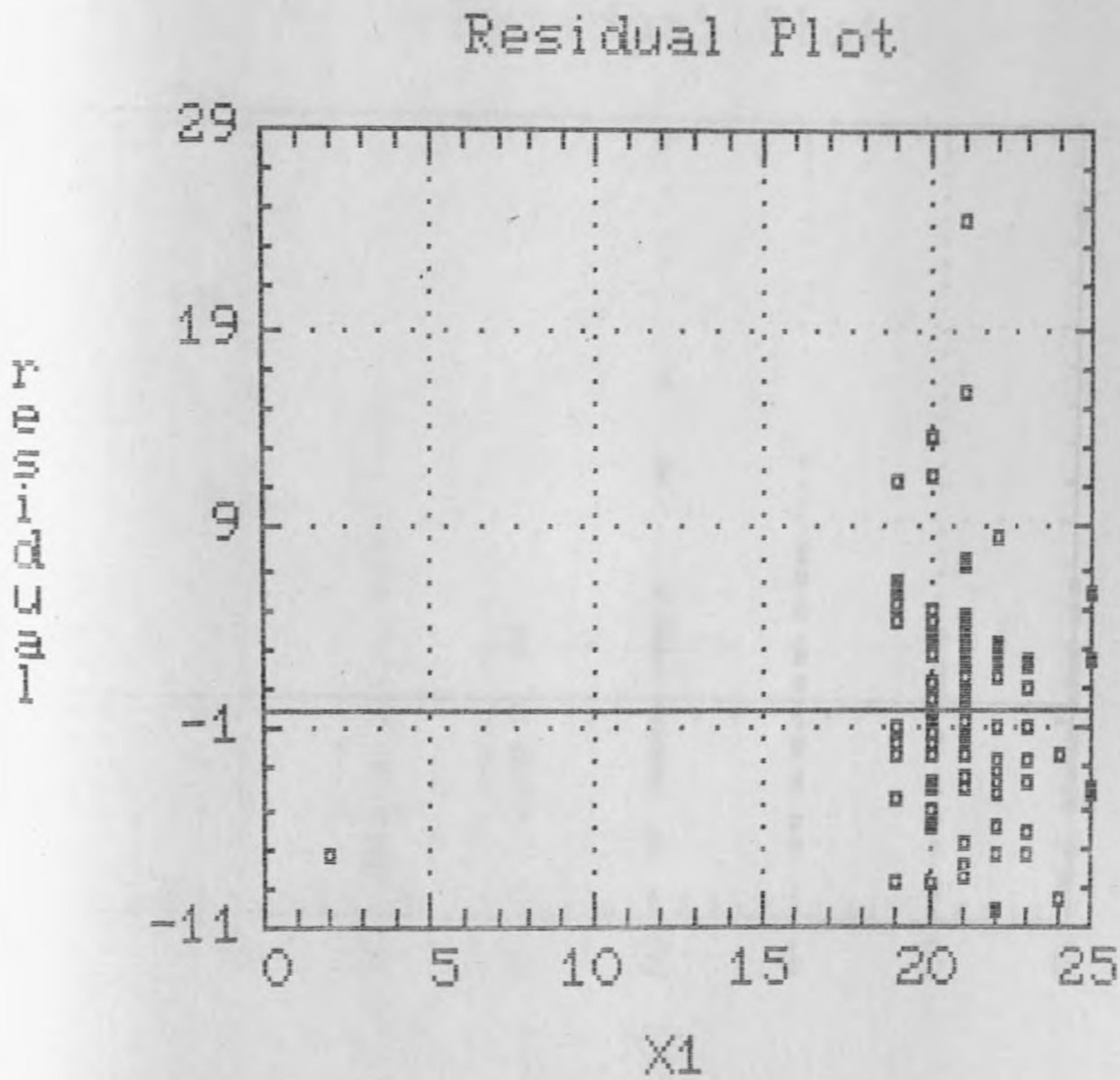


Table 9

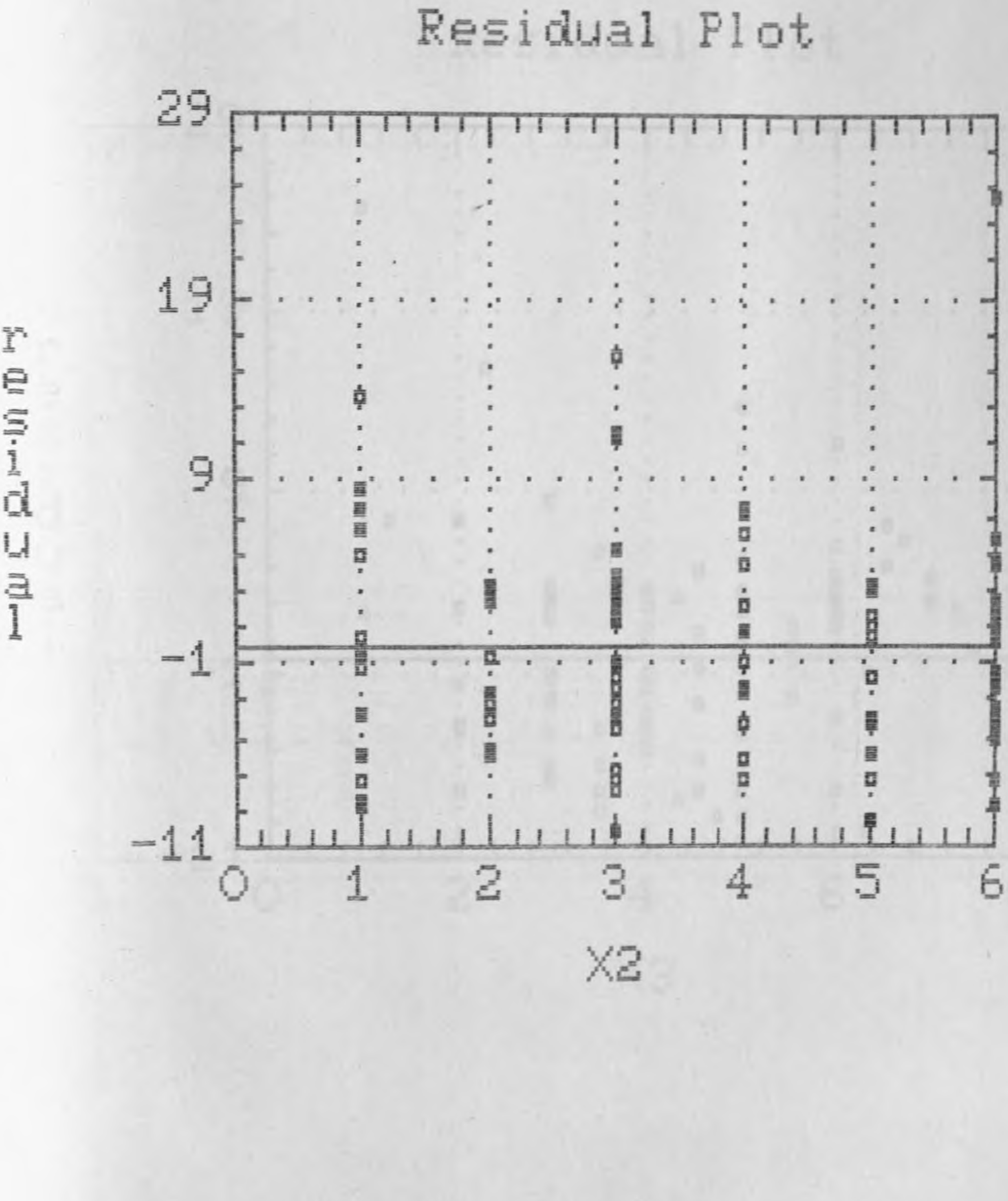


Table 10

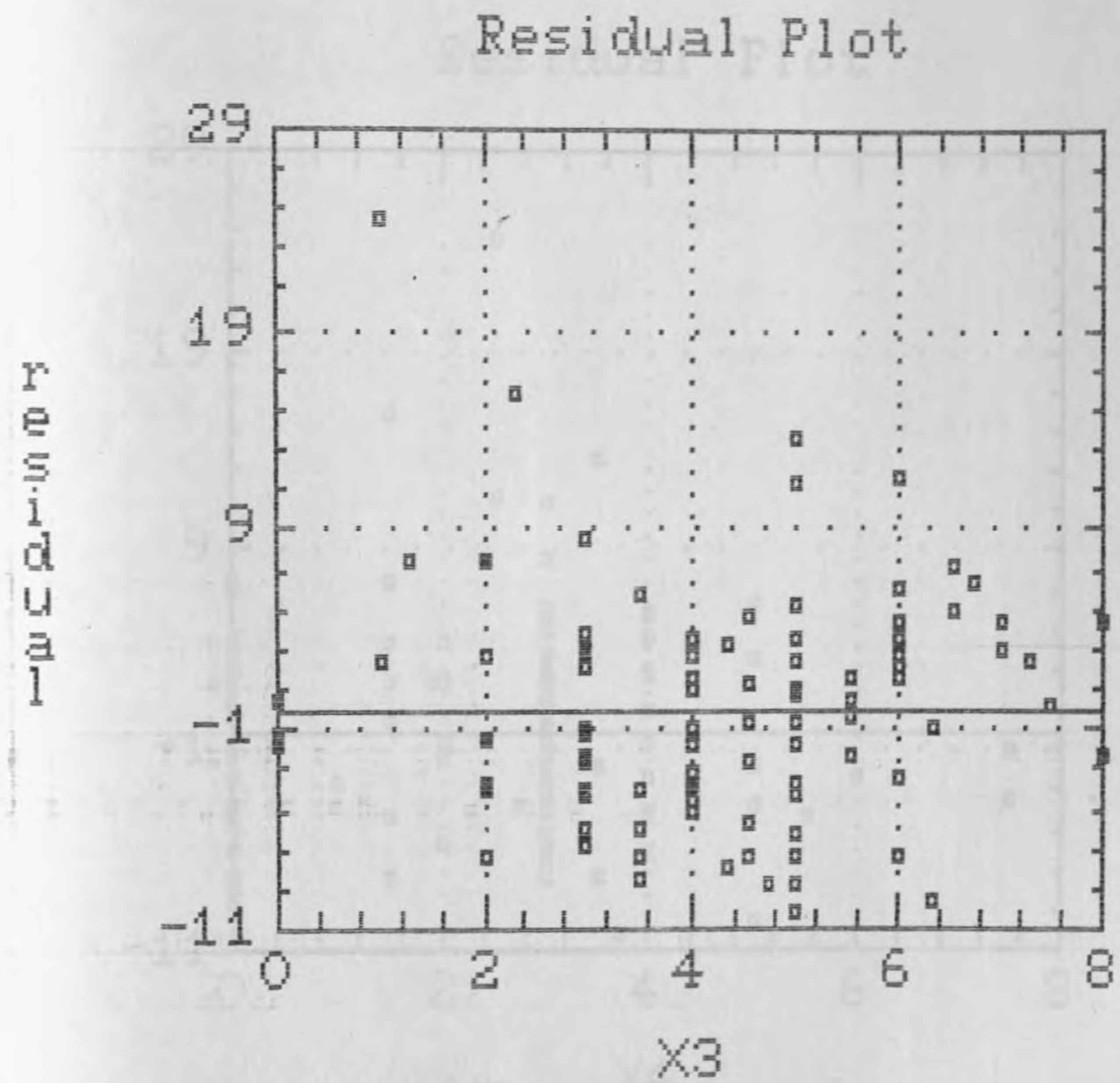


Table 11

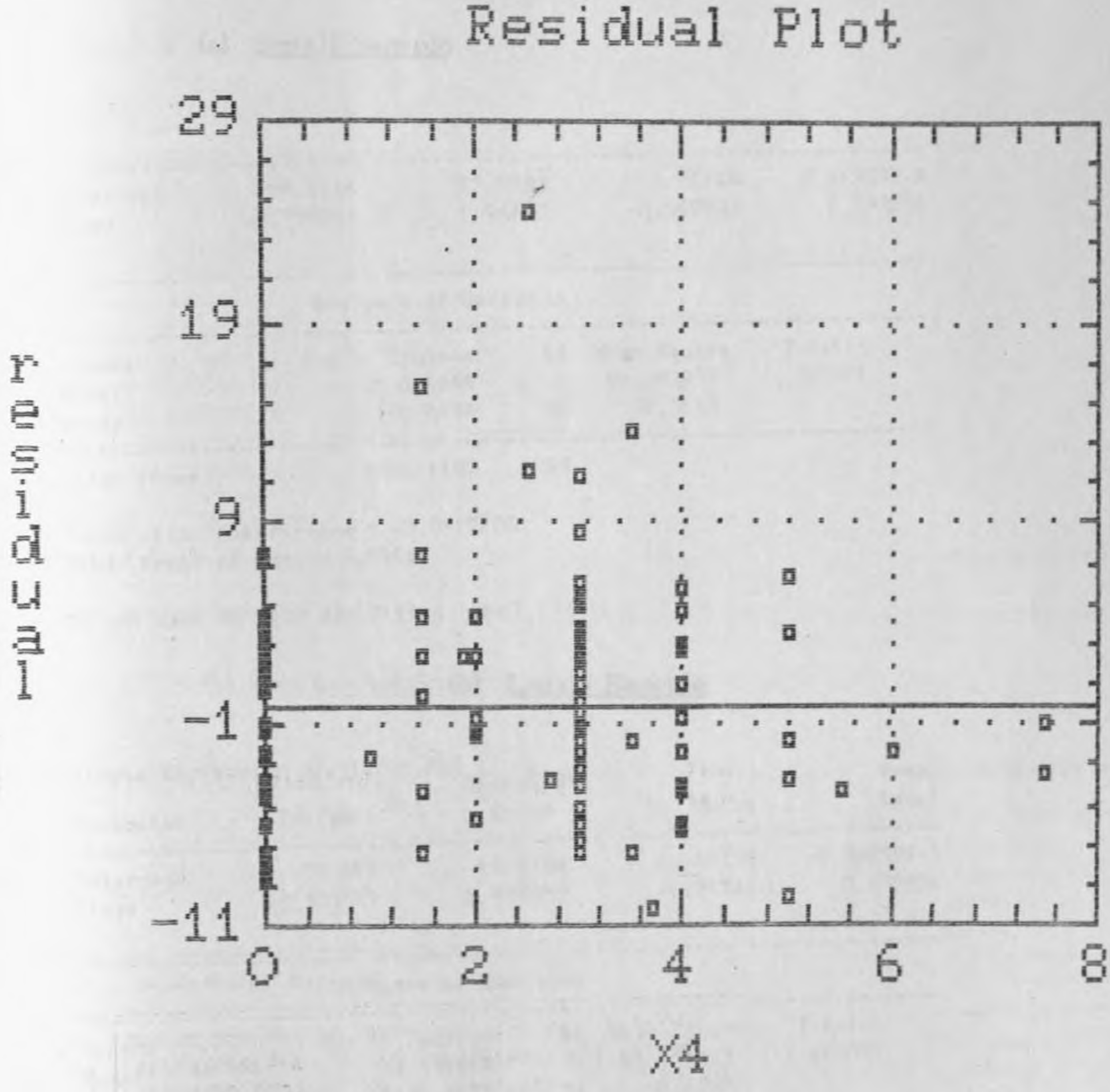


Table 12

(a) Small Sample

~

Intercept	75.2185	27.2384	2.76149	8.81329E-3
Slope	-0.775708	1.36633	-0.567733	0.573556

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio
Model	10.093899	1	10.093899	.322321
Error	1190.0199	38	31.3163	
Total (Corr.)	1200.1137	39		

Correlation Coefficient = -0.0917103
Stnd. Error of Est. = 5.5961

Do you want to plot the fitted line? (Y/N):

(b) Large Sample

Simple Regression of Y12 on X12

Parameter	Estimate	Standard Error	T Value	Prob. Level
Intercept	79.857	18.1354	4.40337	8.38758E-5
Slope	-0.972527	0.798852	-1.21741	0.230956

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio
Model	46.131814	1	46.131814	1.482075
Error	1182.8072	38	31.1265	
Total (Corr.)	1228.9390	39		

Correlation Coefficient = -0.193747
Stnd. Error of Est. = 5.57911

Do you want to plot the fitted line? (Y/N):

1HELP 2LABEL 3

Table 13

(a) Small Sub-sample

Simple Regression of Y21 on X21

Parameter	Estimate	Standard Error	T Value	Prob. Level
Intercept	60.4847	17.9572	3.36827	1.74438E-3
Slope	0.0231676	0.791	0.029289	0.976787

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio
Model	.0261794	1	.0261794	.0008578
Error	1159.6698	38	30.5176	
Total (Corr.)	1159.6960	39		

Correlation Coefficient = 4.75125E-3
Stnd. Error of Est. = 5.52428

(b) Large Sub-sample

Simple Regression of Y22 on X22

Parameter	Estimate	Standard Error	T Value	Prob. Level
Intercept	57.911	7.11474	8.13959	7.4785E-10
Slope	0.075974	1.31664	0.0577029	0.954288

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio
Model	.1333344	1	.1333344	.0033296
Error	1521.7044	38	40.0449	
Total (Corr.)	1521.8377	39		

Correlation Coefficient = 9.36024E-3
Stnd. Error of Est. = 6.3281

Do you want to plot the fitted line? (Y/N):

1HELP 2LABEL 3SAVSC 4RECORD 5 6 7 8 9REVIEW 10QUIT
INPUT SAT JUN 23 1990 09:45:00 PM VERSION 1.1 REC:OFF

Table 14

(a) Small Sub-Sample

Simple Regression of Y31 on X31

Parameter	Estimate	Standard Error	T Value	Prob. Level
Intercept	69.2524	2.62178	26.4143	0
Slope	-2.90428	0.912123	-3.18409	2.89632E-3

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio
Model	391.30104	1	391.30104	10.13842
Error	1466.6427	38	38.5959	
Total (Corr.)	1857.9437	39		

Correlation Coefficient = -0.458922
Stnd. Error of Est. = 6.21256

Do you [illegible]

(b) Large Sub- Sample

Simple Regression of Y32 on X32

Parameter	Estimate	Standard Error	T Value	Prob. Level
Intercept	60.5602	6.10442	9.92071	4.25748E-12
Slope	-0.261628	1.02787	-0.254532	0.800457

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio
Model	1.9856877	1	1.9856877	.0647868
Error	1164.6841	38	30.6496	
Total (Corr.)	1166.6697	39		

Correlation Coefficient = -0.0412555
Stnd. Error of Est. = 5.53621

Do you want to plot the fitted line? (Y/N):

1HELP 2LABEL 3SAVSC 4RECORD 5 6 7 8 9REVIEW 10QUIT
INPUT SAT JUN 23 1990 09:47:00 PM VERSION 1.1 REC:OFF

Table 15

(a) Small Sub-Sample

Simple Regression of Y41 on X41

Parameter	Estimate	Standard Error	T Value	Prob. Level
Intercept	57.4999	1.34963	42.6042	0
Slope	3.05492	1.22115	2.50168	0.0167853

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio
Model	280.82475	1	280.82475	6.25840
Error	1705.1230	38	44.8717	
Total (Corr.)	1985.9477	39		

Correlation Coefficient = 0.37604
Stnd. Error of Est. = 6.69863

Do you want to plot the fitted line? (Y/N):

(b) Large Sub-Sample

Simple Regression of Y42 on X42

Parameter	Estimate	Standard Error	T Value	Prob. Level
Intercept	65.6009	3.02668	21.6742	0
Slope	-1.44729	0.737234	-1.96313	0.0569807

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio
Model	109.88032	1	109.88032	3.85389
Error	1083.4387	38	28.5115	
Total (Corr.)	1193.3190	39		

Correlation Coefficient = -0.303446
Stnd. Error of Est. = 5.33962

Do you want to plot the fitted line? (Y/N):

1HELP 2LABEL 3SAVSC 4RECORD 5 6 7 8 9REVIEW 10QUIT
INPUT SAT JUN 23 1990 09:50:00 PM VERSION 1.1 REC:OFF

Table 16

Model 1

MODEL FITTING RESULTS

VARIABLE	COEFFICIENT	STND. ERROR	T-VALUE	PROB(>|T|)
CONSTANT	69.251169	5.672299	12.2087	.0000
X1	-0.226317	0.260472	-.8689	.3870
X2	-0.283061	0.364784	-.7760	.4396
X3	-1.065864	0.365675	-2.9148	.0044
X4	0.283966	0.355909	.7979	.4269

0 CASES WITH MISSING VALUES WERE EXCLUDED.

RESIDUALS PLACED IN VARIABLE: RESIDUALS

FURTHER ANOVA FOR VARIABLES IN THE ORDER FITTED

SOURCE	SUM OF SQUARES	DF	MEAN SQ.	F-RATIO	PROB(>F)
X1	6.07519	1	6.0752	.1773	.6792
X2	89.88711	1	89.8871	2.6232	.1086
X3	274.65945	1	274.6594	8.0154	.0057
X4	21.81345	1	21.8134	.6366	.4354
MODEL	392.43520	4			

b~:F8DD11!!~8@~8@

ANALYSIS OF VARIANCE FOR THE FULL REGRESSION

SOURCE	SUM OF SQUARES	DF	MEAN SQUARE	F-RATIO	PROB(>F)
MODEL	392.43520	4	98.10880	2.86310	.02742
ERROR	3255.3264	95	34.2666		
TOTAL (CORR.)	3647.7616	99			

R-SQUARED = 0.107582
R-SQUARED (ADJ. FOR D.F.) = 0.070007
STND. ERROR OF EST. = 5.85377

Press ENTER to continue.

Table 17

Model 2

MODEL FITTING RESULTS

VARIABLE	COEFFICIENT	STND. ERROR	T-VALUE	PROB(>|T|)
CONSTANT	58.344816	7.697206	7.5800	.0000
LOG X1	1.056463	2.497248	.4231	.6732
LOG X2	-1.404214	1.077013	-1.3038	.1953
EXP X3	-0.001013	0.001253	-.8089	.4205
EXP X4	-0.001732	0.002404	-.7207	.4728

0 CASES WITH MISSING VALUES WERE EXCLUDED.

RESIDUALS PLACED IN VARIABLE: RESIDUALS

ANALYSIS OF VARIANCE FOR THE FULL REGRESSION

SOURCE	SUM OF SQUARES	DF	MEAN SQUARE	F-RATIO	PROB(>F)
MODEL	135.71179	4	33.92795	.91774	.45700
ERROR	3512.0498	95	36.9689		
TOTAL (CORR.)	3647.7616	99			

R-SQUARED = 0.0372041
R-SQUARED (ADJ. FOR D.F.) = 0
STND. ERROR OF EST. = 6.08021

Press ENTER to continue.

1HELP 2LABEL 3SAVSC 4RECORD 5 6 7 8 9REVIEW 10QUIT
INPUT SAT JUN 23 1990 10:02:00 PM VERSION 1.1 REC:OFF

Table 18

Supermodel Equation 3

MODEL FITTING RESULTS

VARIABLE	COEFFICIENT	STND. ERROR	T-VALUE	PROB(>|T|)
CONSTANT	54.443055	8.985519	6.0590	.0000
LOG X1	16.498433	6.880164	2.3980	.0184
LOG X2	0.49465	3.660477	.1351	.8928
EXP X3	0.002339	0.001568	1.4916	.1390
EXP X4	-0.002396	0.002698	-.8880	.3767
X1	-1.905811	0.725469	-2.6270	.0100
X2	-0.175752	1.253926	-.1402	.8888
X3	-1.555526	0.477474	-3.2578	.0015
X4	0.64226	0.415937	1.5441	.1257

0 CASES WITH MISSING VALUES WERE EXCLUDED.

RESIDUALS PLACED IN VARIABLE: RESIDUALS

ANALYSIS OF VARIANCE FOR THE FULL REGRESSION

SOURCE	SUM OF SQUARES	DF	MEAN SQUARE	F-RATIO	PROB(>F)
MODEL	665.38298	8	83.17287	2.53782	.01532
ERROR	2982.3786	91	32.7734		
TOTAL (CORR.)	3647.7616	99			

R-SQUARED = 0.182409
R-SQUARED (ADJ. FOR D.F.) = 0.110532
STND. ERROR OF EST. = 5.7248

Press ENTER to continue.

1HELP 2LABEL 3SAVSC 4RECORD 5 6 7 8 9REVIEW 10QUIT
INPUT SAT JUN 23 1990 10:06:00 PM VERSION 1.1 REC:OFF

Table 19

Multiple Regression Results, Optimal Model

MODEL FITTING RESULTS

VARIABLE	COEFFICIENT	STND. ERROR	T-VALUE	PROB(>|T|)
CONSTANT	77.48639	4.900672	15.8114	.0000
age	-0.546303	0.224659	-2.4317	.0159
sci	-0.241409	0.142008	-1.7000	.0907
com	-0.305014	0.124339	-2.4531	.0150
mat	-1.601736	0.213674	-7.4961	.0000

0 CASES WITH MISSING VALUES WERE EXCLUDED.

RESIDUALS PLACED IN VARIABLE: RESIDUALS

ANALYSIS OF VARIANCE FOR THE FULL REGRESSION

SOURCE	SUM OF SQUARES	DF	MEAN SQUARE	F-RATIO	PROB(>F)
MODEL	1745.1543	4	436.2886	24.5931	.0000
ERROR	3459.3615	195	17.7403		
TOTAL (CORR.)	5204.5158	199			

R-SQUARED = 0.335315
R-SQUARED (ADJ. FOR D.F.) = 0.321681
STND. ERROR OF EST. = 4.21193
NUMBER OF RESIDUALS = 200
SAMPLE AVERAGE = -1.15108E-14
SAMPLE VARIANCE = 17.3837
SAMPLE STANDARD DEVIATION = 4.16938
COEFF. OF SKEWNESS = 0.249053   STANDARDIZED VALUE = 1.43791
COEFF. OF KURTOSIS = 3.38031   STANDARDIZED VALUE = 1.09787
DURBIN-WATSON STATISTIC = 1.85719

Table 20

Stepwise Regression at 95% Significance Level

```
                                  STEPWISE REGRESSION
SELECTION:    FORWARD                                            CONTROL:       AUTOMATIC
F-TO-ENTER =  2.648          MAX STEPS =  50                     F-TO-REMOVE = 2.648
                                       STEP 4
R-SQUARED = 0.378283
R-SQUARED (ADJ.) = 0.30723                          MSE = 9.48518 WITH 35 D.F.
 VARIABLES CURRENTLY IN MODEL               VARIABLES CURRENTLY NOT IN MODEL
VARIABLE                    COEFF.   F-REMOVE VARIABLE             PARTIAL CORR.   F-ENTER
  1. X8                    2.21507    5.0431  3. X13                       .2271    1.8486
  2. X11                   3.56423    8.8300
  4. X15                   2.31687    4.6325
  5. X18                   -.50061    5.9986




SPECIFY FINISH, ENTER K,
OR REMOVE K:

 1HELP   2LABEL  3SAVSC  4RECORD 5        6       7       8       9REVIEW 10QUIT
INPUT    SAT JUN 23  1990 10:12:00 PM  VERSION 1.1                            REC:OFF
```

Table 21

Regression of Primary School Handcraft (X8) on Other Predictor Variables.

MODEL FITTING RESULTS

VARIABLE	COEFFICIENT	STND. ERROR	T-VALUE	PROB(>|T|)
CONSTANT	0.597393	0.152447	3.9187	.0003
X11	-0.09911	0.219989	-.4505	.6548
X13	-0.020897	0.052488	-.3981	.6927
X15	-0.032658	0.184726	-.1768	.8606
X18	0.00992	0.035086	.2827	.7789

0 CASES WITH MISSING VALUES WERE EXCLUDED.

RESIDUALS PLACED IN VARIABLE: RESIDUALS

ANALYSIS OF VARIANCE FOR THE FULL REGRESSION

SOURCE	SUM OF SQUARES	DF	MEAN SQUARE	F-RATIO	PROB(>F)
MODEL	.1948381	4	.0487095	.1756625	.9494341
ERROR	9.705162	35	.277290		
TOTAL (CORR.)	9.900000	39			

R-SQUARED = 0.0196806
R-SQUARED (ADJ. FOR D.F.) = 0
STND. ERROR OF EST. = 0.526584

Press ENTER to continue.

Table 22

Regression of Formal Technical Training (X11) on Other Predictor Variables.

MODEL FITTING RESULTS

VARIABLE	COEFFICIENT	STND. ERROR	T-VALUE	PROB(>|T|)
CONSTANT	0.041459	0.139919	.2963	.7686
X8	-0.058175	0.129128	-.4505	.6548
X13	0.087914	0.037465	2.3465	.0241
X15	0.108497	0.140397	.7728	.4443
X18	-0.001175	0.02691	-.0437	.9654

0 CASES WITH MISSING VALUES WERE EXCLUDED.

RESIDUALS PLACED IN VARIABLE: RESIDUALS

ANALYSIS OF VARIANCE FOR THE FULL REGRESSION

SOURCE	SUM OF SQUARES	DF	MEAN SQUARE	F-RATIO	PROB(>F)
MODEL	1.2783052	4	.3195763	1.9634491	.1217231
ERROR	5.696695	35	.162763		
TOTAL (CORR.)	6.975000	39			

R-SQUARED = 0.18327
R-SQUARED (ADJ. FOR D.F.) = 0.0899289
STND. ERROR OF EST. = 0.403439

Press ENTER to continue.

Table 23

Regression of Number of Years of Technical Training on Other Predictor Variables.

MODEL FITTING RESULTS

VARIABLE	COEFFICIENT	STND. ERROR	T-VALUE	PROB(>|T|)
CONSTANT	1.321366	0.543423	2.4316	.0197
X8	-0.215742	0.541883	-.3981	.6927
X11	1.546244	0.658946	2.3465	.0241
X15	0.317991	0.591367	.5377	.5938
X18	0.067079	0.11229	.5974	.5537

0 CASES WITH MISSING VALUES WERE EXCLUDED.

RESIDUALS PLACED IN VARIABLE: RESIDUALS

ANALYSIS OF VARIANCE FOR THE FULL REGRESSION

SOURCE	SUM OF SQUARES	DF	MEAN SQUARE	F-RATIO	PROB(>F)
MODEL	23.516278	4	5.879069	2.053677	.108089
ERROR	100.19466	35	2.86270		
TOTAL (CORR.)	123.71094	39			

R-SQUARED = 0.190091
R-SQUARED (ADJ. FOR D.F.) = 0.0975294
STND. ERROR OF EST. = 1.69195

Press ENTER to continue.

1HELP 2LABEL 3SAVSC 4RECORD 5 6 7 8 9REVIEW 10QUIT
INPUT SAT JUN 23 1990 10:19:00 PM VERSION 1.1 REC:OFF

Table 24

Regression of Formal Employment Before Entry on Other Predictor Variables

MODEL FITTING RESULTS

VARIABLE	COEFFICIENT	STND. ERROR	T-VALUE	PROB(>|T|)
CONSTANT	0.19473	0.163974	1.1876	.2422
X8	-0.02732	0.154531	-.1768	.8606
X11	0.154627	0.20009	.7728	.4443
X13	0.025767	0.047918	.5377	.5938
X18	0.069307	0.029915	2.3168	.0259

0 CASES WITH MISSING VALUES WERE EXCLUDED.

RESIDUALS PLACED IN VARIABLE: RESIDUALS

Use F10 to Quit, Cursor keys or Page Number:

ANALYSIS OF VARIANCE FOR THE FULL REGRESSION

SOURCE	SUM OF SQUARES	DF	MEAN SQUARE	F-RATIO	PROB(>F)
MODEL	1.8562514	4	.4640629	2.0005792	.1159162
ERROR	8.118749	35	.231964		
TOTAL (CORR.)	9.975000	39			

R-SQUARED = 0.18609
R-SQUARED (ADJ. FOR D.F.) = 0.0930721
STND. ERROR OF EST. = 0.481627

Press ENTER to continue.

Table 25

Regression of Initial Capital on Other Predictor Variables.

MODEL FITTING RESULTS

VARIABLE	COEFFICIENT	STND. ERROR	T-VALUE	PROB(>|T|)
CONSTANT	1.735945	0.829576	2.0926	.0429
X8	0.229712	0.812481	.2827	.7789
X11	-0.046362	1.061664	-.0437	.9654
X13	0.150464	0.251875	.5974	.5537
X15	1.918574	0.828099	2.3168	.0259

0 CASES WITH MISSING VALUES WERE EXCLUDED.

RESIDUALS PLACED IN VARIABLE: RESIDUALS

ANALYSIS OF VARIANCE FOR THE FULL REGRESSION

SOURCE	SUM OF SQUARES	DF	MEAN SQUARE	F-RATIO	PROB(>F)
MODEL	43.155805	4	10.788951	1.680192	.176597
ERROR	224.74419	35	6.42126		
TOTAL (CORR.)	267.90000	39			

R-SQUARED = 0.161089
R-SQUARED (ADJ. FOR D.F.) = 0.0652137
STND. ERROR OF EST. = 2.53402

Press ENTER to continue.

Table 26

Some Selected Ridge Coefficients at Various Levels of Bias (Theta)

```
ENTER NAME OF DEPENDENT VARIABLE:  X4
ENTER NAME OF FIRST INDEPENDENT VARIABLE:  X8
ENTER NAME OF NEXT INDEPENDENT VARIABLE:  X11
ENTER NAME OF NEXT INDEPENDENT VARIABLE:  X13
ENTER NAME OF NEXT INDEPENDENT VARIABLE:  X15
ENTER NAME OF NEXT INDEPENDENT VARIABLE:  X18


0.137749 0.17874 0.119362 0.12362 -0.136619←
ENTER VALUE FOR THETA (QUIT):  0
PARAMETER ESTIMATES FOR THETA = 0:
0.313751 0.33431 0.198961 0.298698 -0.374231
ENTER VALUE FOR THETA (QUIT):  .1
PARAMETER ESTIMATES FOR THETA = 0.1:
0.278748 0.307193 0.186706 0.258766 -0.32107
ENTER VALUE FOR THETA (QUIT):  .2
PARAMETER ESTIMATES FOR THETA = 0.2:
0.250655 0.28418 0.175799 0.22894 -0.28069
ENTER VALUE FOR THETA (QUIT):  .3
PARAMETER ESTIMATES FOR THETA = 0.3:
0.227614 0.264438 0.16605 0.205764 -0.24899
ENTER VALUE FOR THETA (QUIT):  .4
PARAMETER ESTIMATES FOR THETA = 0.4:
0.208381 0.247323 0.157296 0.187197 -0.223463
ENTER VALUE FOR THETA (QUIT):  .5
PARAMETER ESTIMATES FOR THETA = 0.5:
0.192088 0.232344 0.149401 0.171956 -0.202483
ENTER VALUE FOR THETA (QUIT):  1
PARAMETER ESTIMATES FOR THETA = 1:
0.137749 0.17874 0.119362 0.12362 -0.136619
ENTER VALUE FOR THETA (QUIT):
 1HELP   2LABEL  3SAVSC  4RECORD 5      6      7       8      9REVIEW 10QUIT
INPUT    SAT JUN 23  1990 10:30:00 PM  VERSION 1.1                    REC:OFF
```

Table 27

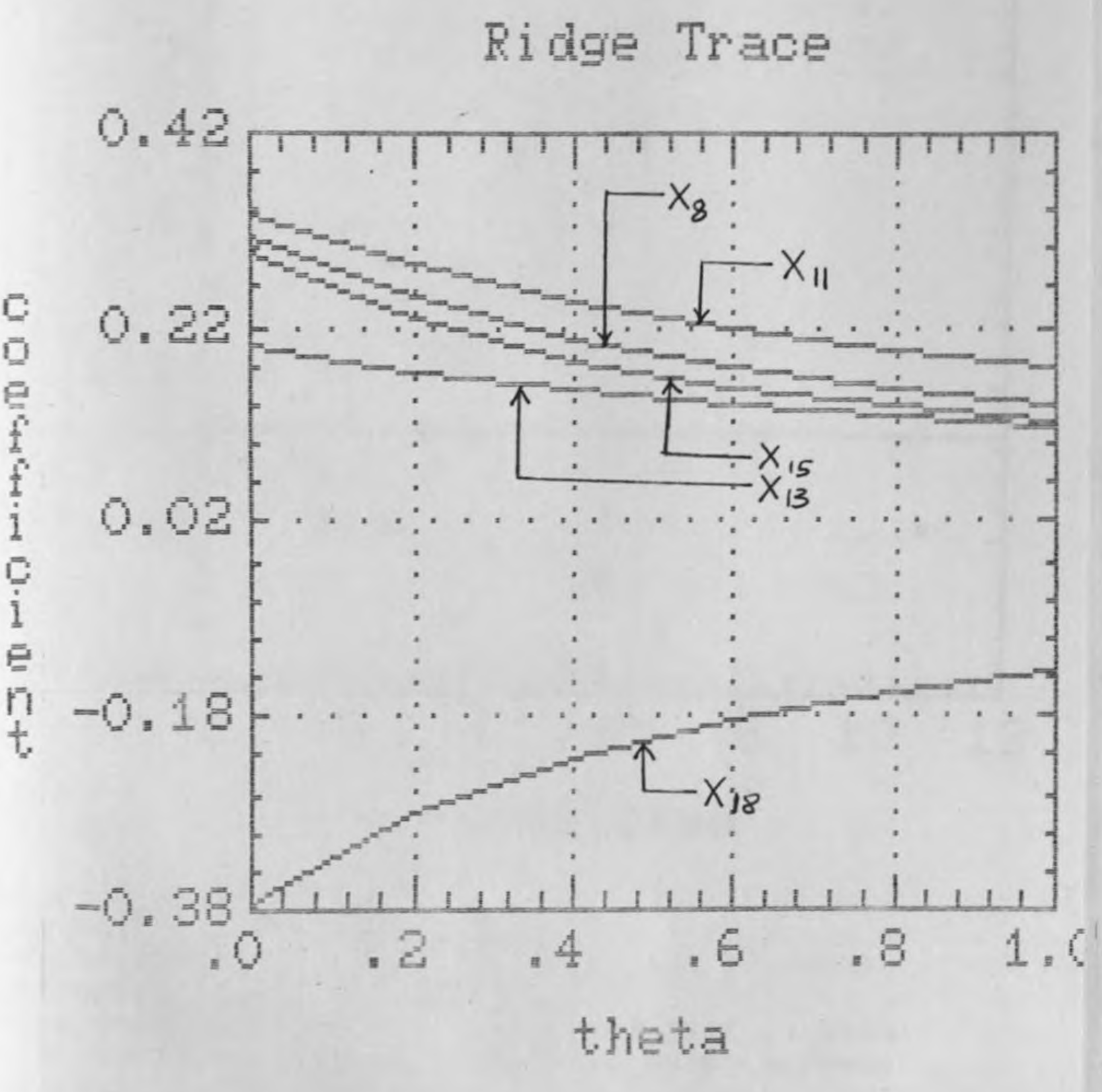
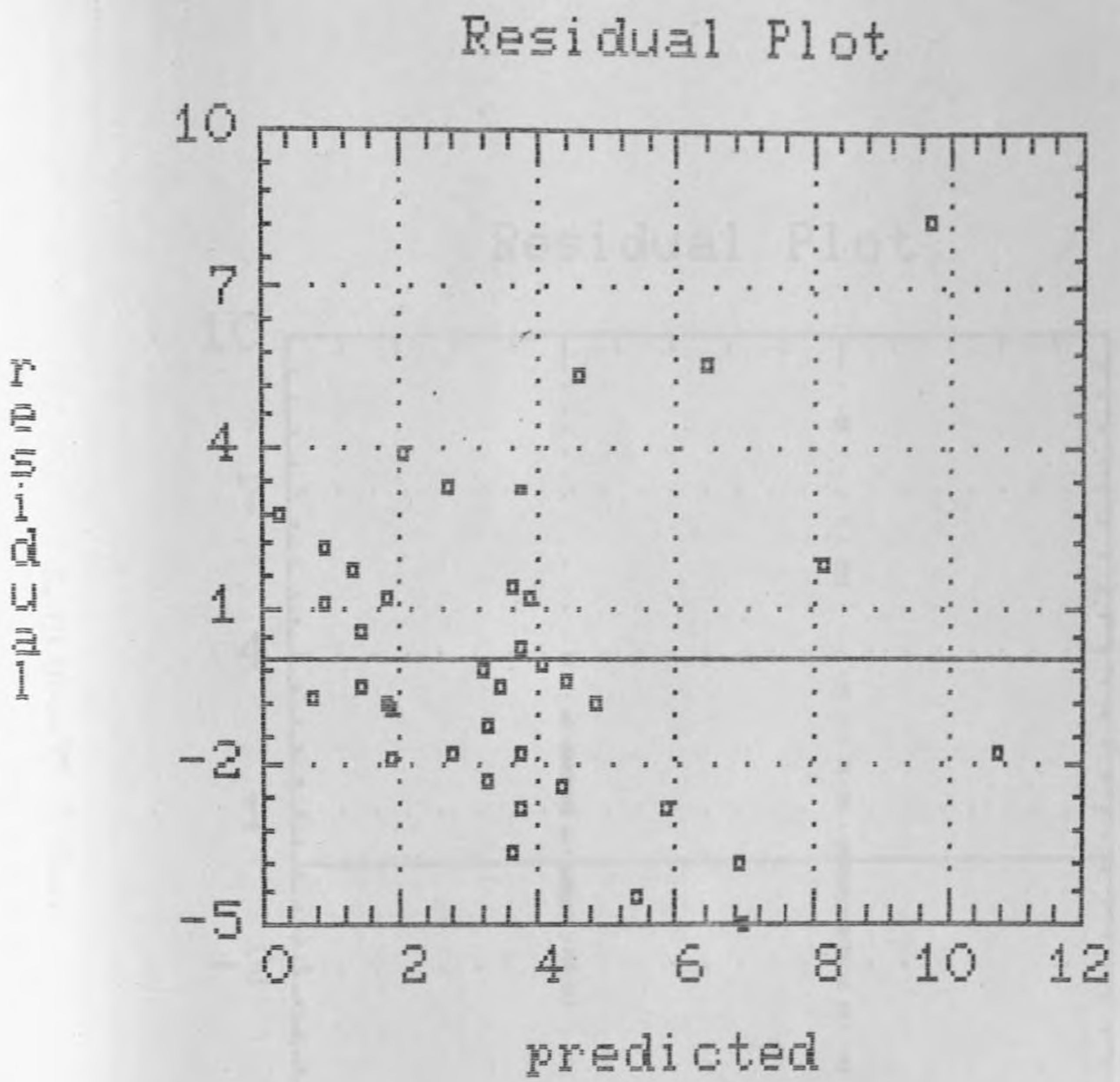

Ridge Trace
0.42
0.22
0.02
-0.18
-0.38
.0
.2
.4
.6
.8
1.0
coefficient
theta
X8
X11
X15
X13
X18

Table 28



NUMBER OF RESIDUALS = 40
SAMPLE AVERAGE = 1.62093E-15
SAMPLE VARIANCE = 8.07337
SAMPLE STANDARD DEVIATION = 2.84137
COEFF. OF SKEWNESS = 0.736155   STANDARDIZED VALUE = 1.90074
COEFF. OF KURTOSIS = 3.55761   STANDARDIZED VALUE = 0.719871
DURBIN-WATSON STATISTIC = 2.03526

Table 29

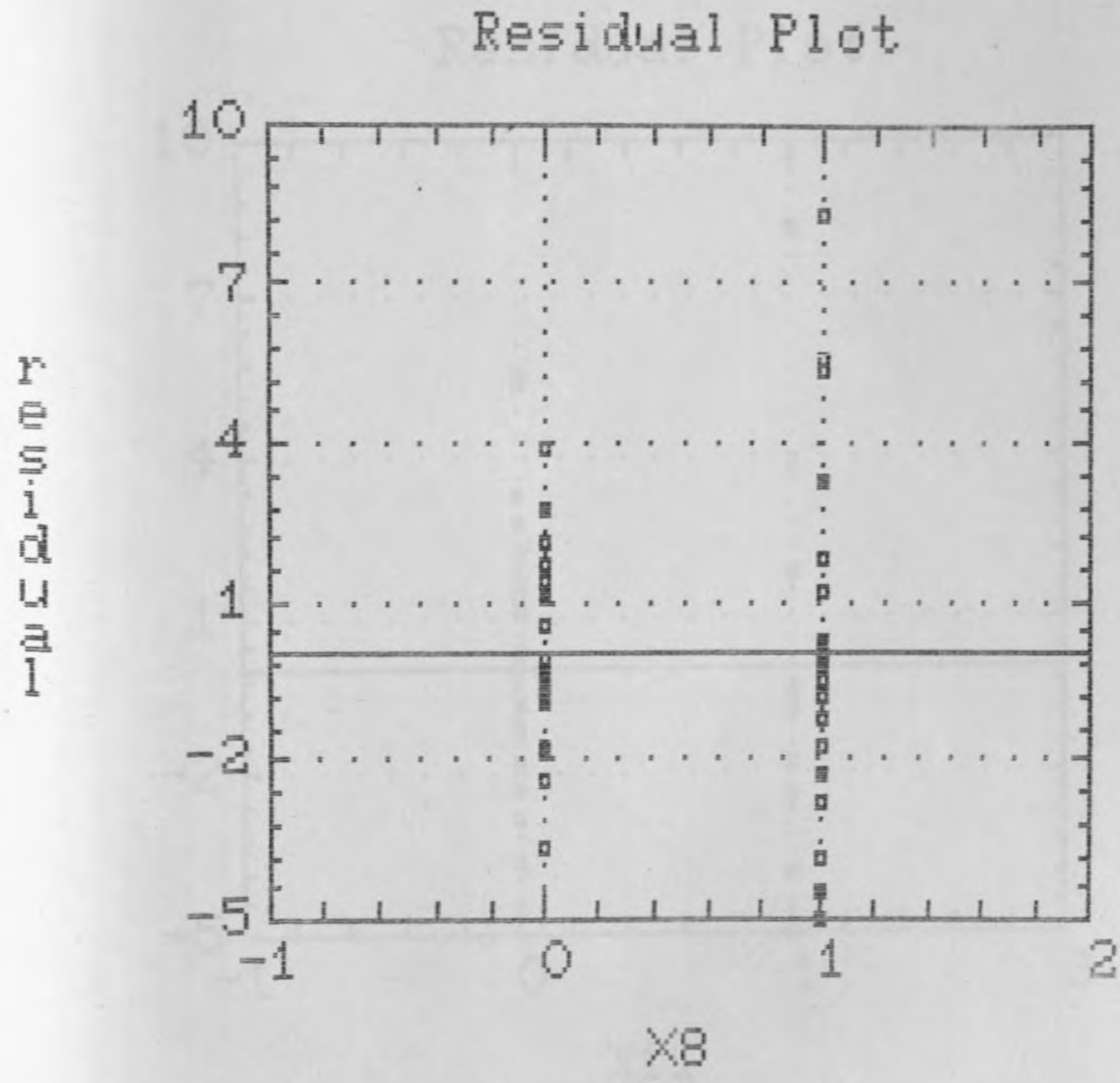


Table 30

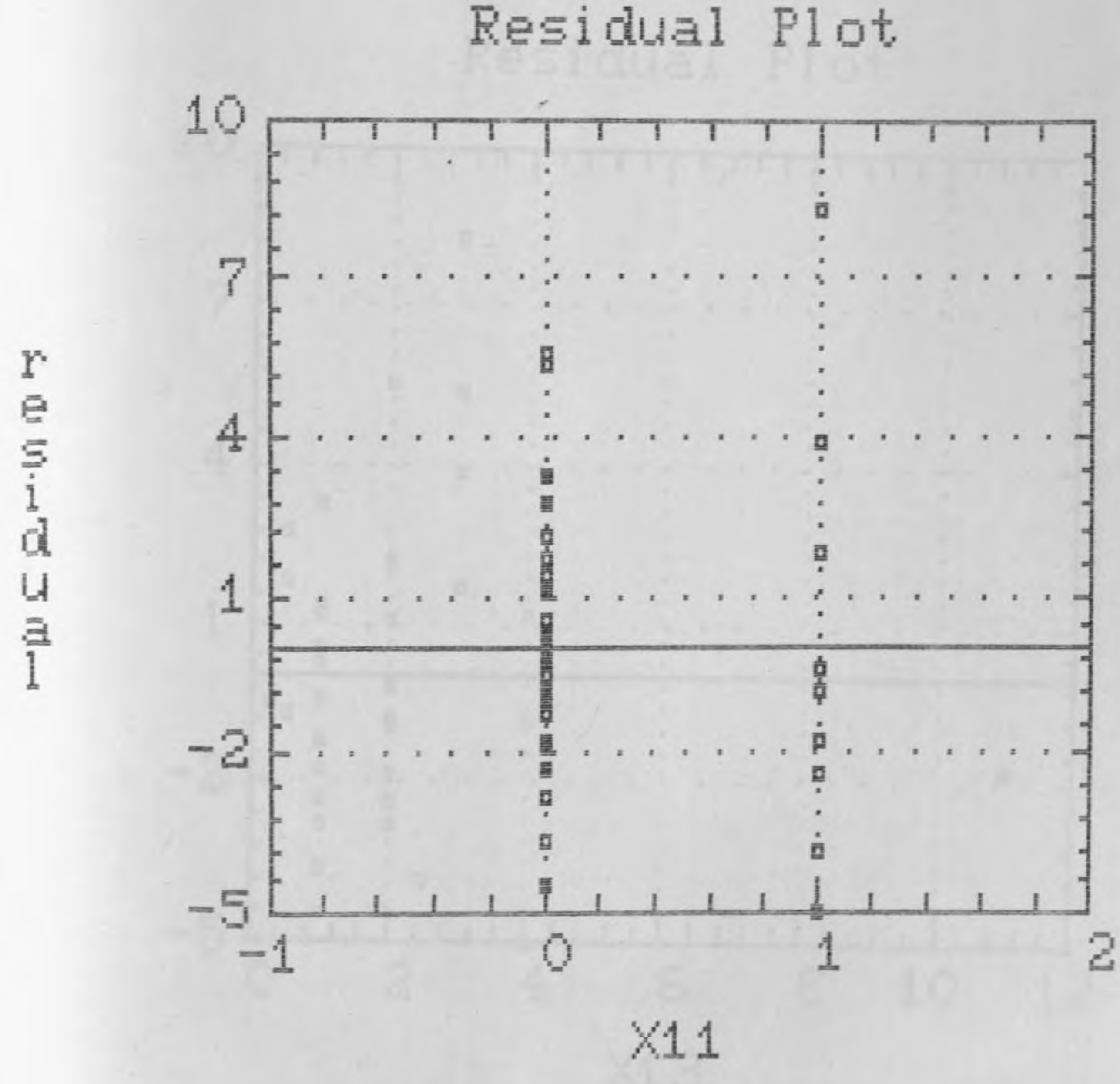


Table 31

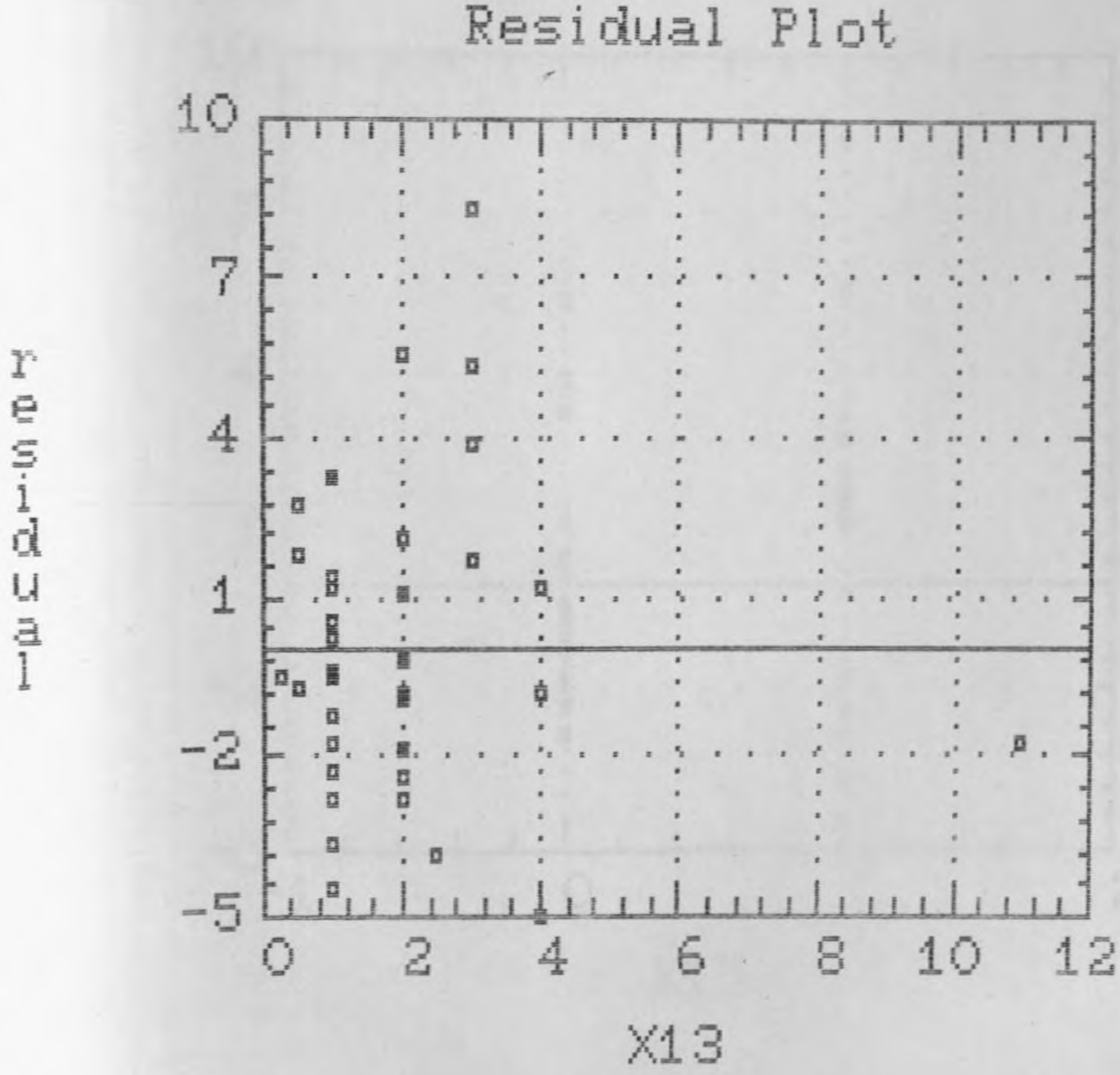


Table 32

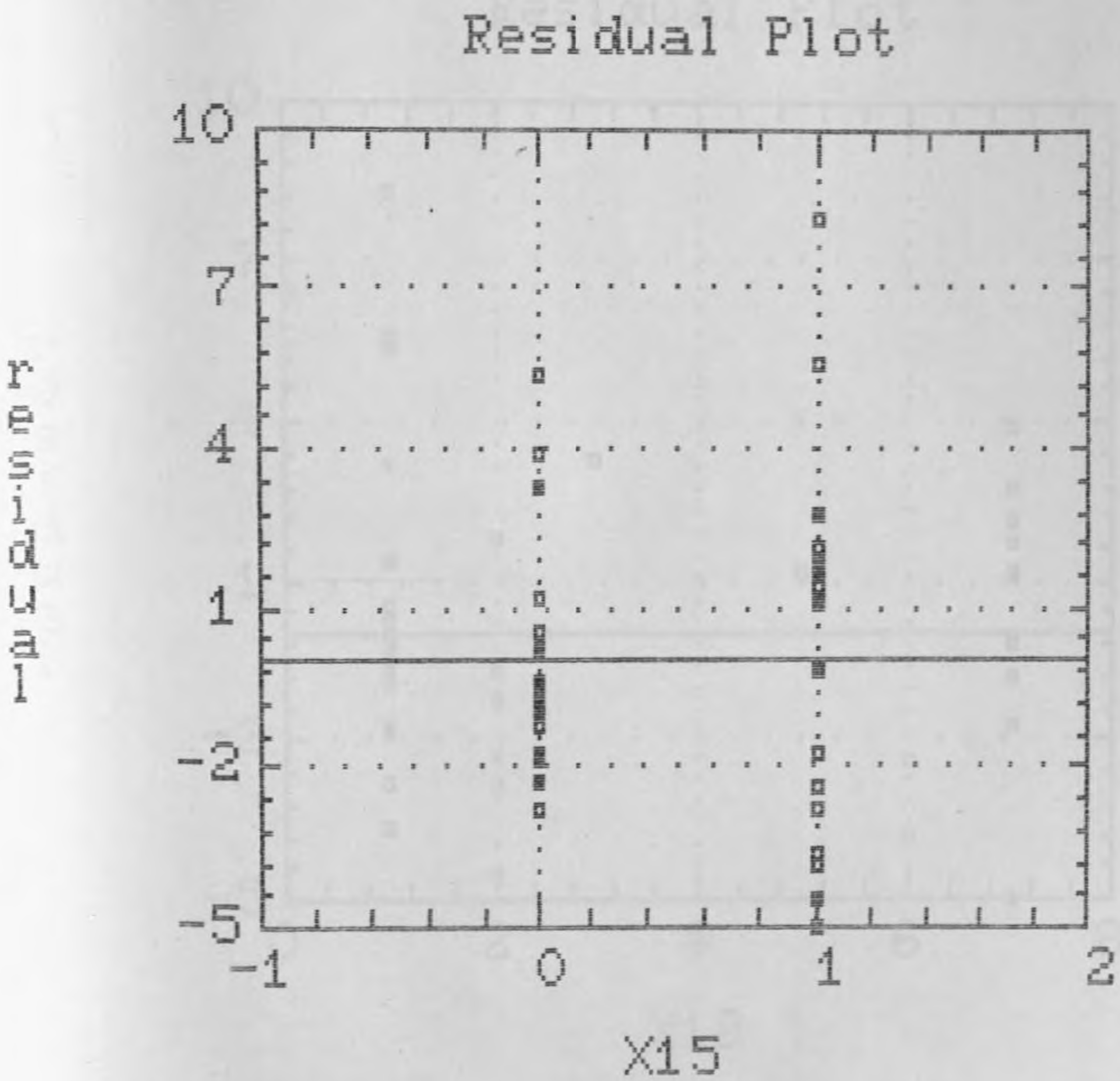


Table 33

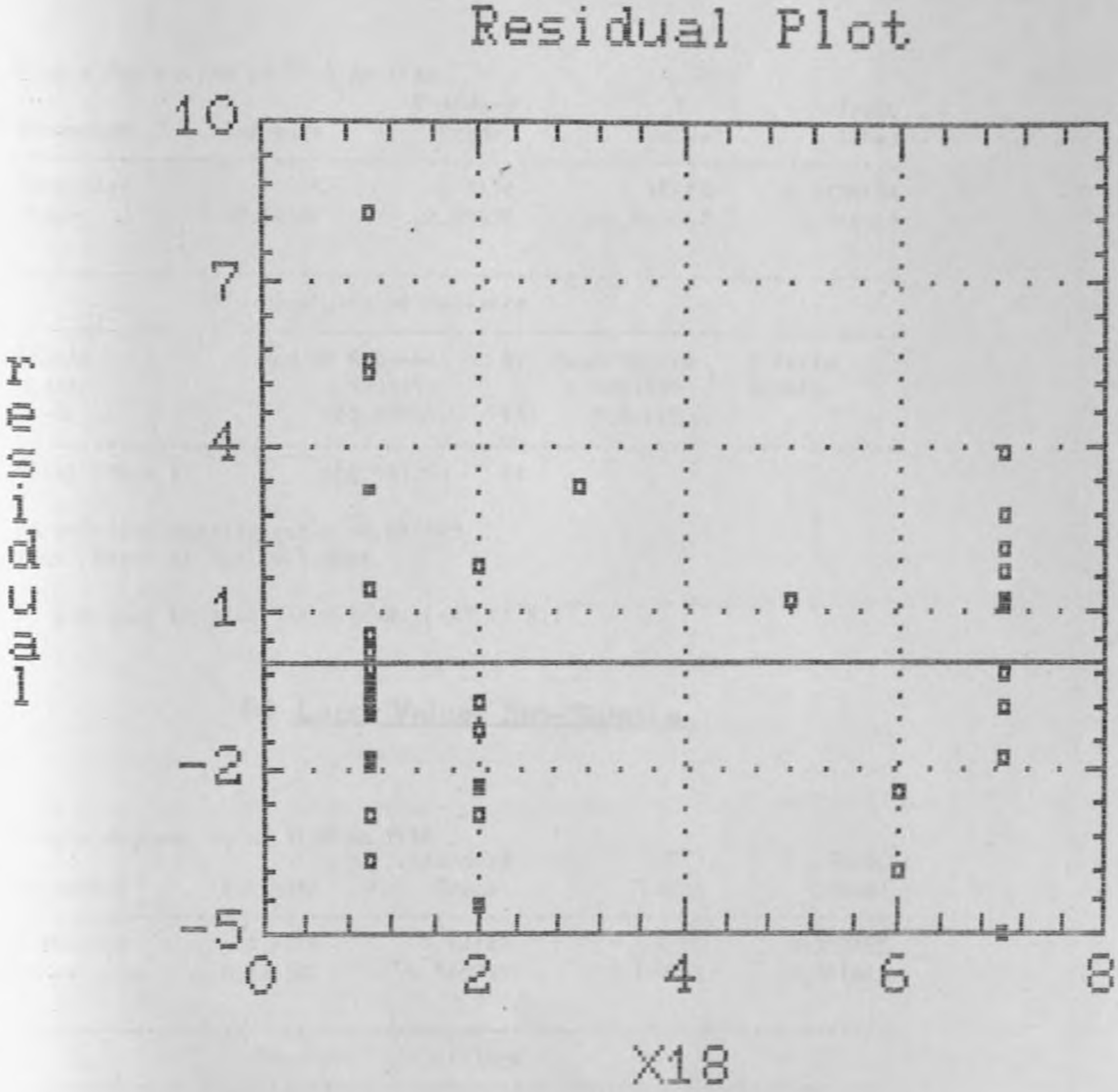


Table 34

(a) Small Values Sub-Sample

Simple Regression of Y13A on X13A

Parameter	Estimate	Standard Error	T Value	Prob. Level
Intercept	5	2.5174	1.98618	0.0685134
Slope	-2.66667	2.83678	-0.940032	0.364341

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio
Model	6.9333333	1	6.9333333	.8836601
Error	102.00000	13	7.84615	
Total (Corr.)	108.93333	14		

Correlation Coefficient = -0.252285
Stnd. Error of Est. = 2.8011

Do you want to plot the fitted line? (Y/N):

(b) Large Values Sub-Sample

Simple Regression of Y13B on X13B

Parameter	Estimate	Standard Error	T Value	Prob. Level
Intercept	2.6815	2.14142	1.25221	0.232555
Slope	0.581363	0.540733	1.07514	0.301863

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio
Model	24.300974	1	24.300974	1.155923
Error	273.29903	13	21.02300	
Total (Corr.)	297.60000	14		

Correlation Coefficient = 0.285756
Stnd. Error of Est. = 4.58508

Do you want to plot the fitted line? (Y/N):

Table 35

Model 1 Regression Results

MODEL FITTING RESULTS

VARIABLE	COEFFICIENT	STND. ERROR	T-VALUE	PROB(>|T|)
CONSTANT	1.584922	1.05673	1.4998	.1417
X8	2.304245	0.97683	2.3589	.0234
X11	2.92508	1.274996	2.2942	.0272
X13	0.413356	0.304017	1.3596	.1818
X15	2.185425	1.068011	2.0463	.0475
X18	-0.52834	0.202991	-2.6028	.0130

0 CASES WITH MISSING VALUES WERE EXCLUDED.

RESIDUALS PLACED IN VARIABLE: RESIDUALS

ANALYSIS OF VARIANCE FOR THE FULL REGRESSION

SOURCE	SUM OF SQUARES	DF	MEAN SQUARE	F-RATIO	PROB(>F)
MODEL	219.11346	5	43.82269	4.73215	.00218
ERROR	314.86154	34	9.26063		
TOTAL (CORR.)	533.97500	39			

R-SQUARED = 0.410344
R-SQUARED (ADJ. FOR D.F.) = 0.32363
STND. ERROR OF EST. = 3.04313

Press ENTER to continue.

1HELP 2LABEL 3SAVSC 4RECORD 5 6 7 8 9REVIEW 10QUIT
INPUT SAT JUN 23 1990 10:44:00 PM VERSION 1.1 REC:OFF

Table 36

Model 2 Regression Results

MODEL FITTING RESULTS

VARIABLE	COEFFICIENT	STND. ERROR	T-VALUE	PROB(>|T|)
CONSTANT	1.634467	2.376534	.6878	.4957
X8*X11	-0.363919	1.8542	-.1963	.8454
X13*X15	0.768867	0.38793	1.9820	.0546
X8*X15	1.5086	1.467746	1.0278	.3104

0 CASES WITH MISSING VALUES WERE EXCLUDED.

RESIDUALS PLACED IN VARIABLE: RESIDUALS

ANALYSIS OF VARIANCE FOR THE FULL REGRESSION

SOURCE	SUM OF SQUARES	DF	MEAN SQUARE	F-RATIO	PROB(>F)
MODEL	62.823497	3	20.941166	1.600084	.206364
ERROR	471.15150	36	13.08754		
TOTAL (CORR.)	533.97500	39			

R-SQUARED = 0.117653
R-SQUARED (ADJ. FOR D.F.) = 0.0441235
STND. ERROR OF EST. = 3.61767

Press ENTER to continue.

1HELP 2LABEL 3SAVSC 4RECORD 5 6 7 8 9REVIEW 10QUIT
INPUT SAT JUN 23 1990 10:50:00 PM VERSION 1.1 REC:OFF

Table 37

Supermodel 3 Rebression Results

MODEL FITTING RESULTS

VARIABLE	COEFFICIENT	STND. ERROR	T-VALUE	PROB(>|T|)
CONSTANT	1.532185	2.763429	.5545	.5824
X8	2.483742	1.496105	1.6601	.1049
X11	3.914057	2.012808	1.9446	.0591
X13	1.395045	0.731491	1.9071	.0639
X15	2.676389	1.708744	1.5663	.1254
X18	-0.700912	0.235204	-2.9800	.0049
X8*X11	1.111733	2.612345	.4256	.6728
X13*X15	-1.225769	0.820123	-1.4946	.1431
X8*X15	-1.087647	2.195826	-.4953	.6232

0 CASES WITH MISSING VALUES WERE EXCLUDED.

RESIDUALS PLACED IN VARIABLE: RESIDUALS

ANALYSIS OF VARIANCE FOR THE FULL REGRESSION

SOURCE	SUM OF SQUARES	DF	MEAN SQUARE	F-RATIO	PROB(>F)
MODEL	242.40726	8	30.30091	3.22165	.00880
ERROR	291.56774	31	9.40541		
TOTAL (CORR.)	533.97500	39			

R-SQUARED = 0.453967
R-SQUARED (ADJ. FOR D.F.) = 0.313056
STND. ERROR OF EST. = 3.06682

Press ENTER to continue.

Table 38

Multiple Regression Results, Optimal Model

MODEL FITTING RESULTS

VARIABLE	COEFFICIENT	STND. ERROR	T-VALUE	PROB(>|T|)
CONSTANT	1:584922	1.05673	1.4998	.1417
X8	2.304245	0.97683	2.3589	.0234
X11	2.92508	1.274996	2.2942	.0272
X13	0.413356	0.304017	1.3596	.1818
X15	2.185425	1.068011	2.0463	.0475
X18	-0.52834	0.202991	-2.6028	.0130

0 CASES WITH MISSING VALUES WERE EXCLUDED.

RESIDUALS PLACED IN VARIABLE: RESIDUALS

ANALYSIS OF VARIANCE FOR THE FULL REGRESSION

SOURCE	SUM OF SQUARES	DF	MEAN SQUARE	F-RATIO	PROB(>F)
MODEL	219.11346	5	43.82269	4.73215	.00218
ERROR	314.86154	34	9.26063		
TOTAL (CORR.)	533.97500	39			

R-SQUARED = 0.410344
R-SQUARED (ADJ. FOR D.F.) = 0.32363
STND. ERROR OF EST. = 3.04313

Press ENTER to continue.

Table 39

a) Stepwise Regression at 95% Significance Level

```
                              STEPWISE REGRESSION
SELECTION:    FORWARD                                       CONTROL:      AUTOMATIC
F-TO-ENTER =  3.92        MAX STEPS =  50                   F-TO-REMOVE = 3.92
                                   STEP 0
R-SQUARED = 0
R-SQUARED (ADJ.) = 0                              MSE = 8.25577 WITH 39 D.F.
 VARIABLES CURRENTLY IN MODEL              VARIABLES CURRENTLY NOT IN MODEL
VARIABLE                 COEFF.  F-REMOVE VARIABLE           PARTIAL CORR.   F-ENTER
                                           1. X7                    -.1463     .8314
                                           2. X16                   -.2482    2.4949
                                           3. X20                    .2826    3.2975
                                           4. X21                    .1468     .8373
```

b) Stepwise Regression at 99% Significance Level

```
                              STEPWISE REGRESSION
SELECTION:    FORWARD                                       CONTROL:      AUTOMATIC
F-TO-ENTER =  2.65        MAX STEPS =  50                   F-TO-REMOVE = 2.65
                                   STEP 1
R-SQUARED = 0.0798471
R-SQUARED (ADJ.) = 0.0556325                      MSE = 7.79648 WITH 38 D.F.
 VARIABLES CURRENTLY IN MODEL              VARIABLES CURRENTLY NOT IN MODEL
VARIABLE                 COEFF.  F-REMOVE VARIABLE           PARTIAL CORR.   F-ENTER
  3. X20                 .44003    3.2975  1. X7                    -.1349     .6854
                                           2. X16                   -.2329    2.1229
                                           4. X21                    .2223    1.9232
```

SPECIFY F

Table 40

Regression of Formal Education on Other Predictor Variables.

MODEL FITTING RESULTS

VARIABLE	COEFFICIENT	STND. ERROR	T-VALUE	PROB(>|T|)
CONSTANT	3.422518	0.501478	6.8249	.0000
X16	0.122408	0.312694	.3915	.6976
X20	0.010408	0.08617	.1208	.9045
X21	0.236288	0.111577	2.1177	.0406

0 CASES WITH MISSING VALUES WERE EXCLUDED.

RESIDUALS PLACED IN VARIABLE: RESIDUALS

ANALYSIS OF VARIANCE FOR THE FULL REGRESSION

SOURCE	SUM OF SQUARES	DF	MEAN SQUARE	F-RATIO	PROB(>F)
MODEL	4.2924088	3	1.4308029	1.5464614	.2192457
ERROR	33.307591	36	.925211		
TOTAL (CORR.)	37.600000	39			

R-SQUARED = 0.11416
R-SQUARED (ADJ. FOR D.F.) = 0.0403398
STND. ERROR OF EST. = 0.961879

Press ENTER to continue.

Table 41

Regression of Preference for Formal Job on Other Predictor Variables

MODEL FITTING RESULTS

VARIABLE	COEFFICIENT	STND. ERROR	T-VALUE	PROB(>|T|)
CONSTANT	0.671921	0.38813	1.7312	.0913
X7	0.034628	0.088457	.3915	.6976
X20	-0.034171	0.045486	-.7512	.4570
X21	-0.063128	0.062047	-1.0174	.3152

0 CASES WITH MISSING VALUES WERE EXCLUDED.

RESIDUALS PLACED IN VARIABLE: RESIDUALS

ANALYSIS OF VARIANCE FOR THE FULL REGRESSION

SOURCE	SUM OF SQUARES	DF	MEAN SQUARE	F-RATIO	PROB(>F)
MODEL	.3527150	3	.1175717	.4492095	.7193892
ERROR	9.422285	36	.261730		
TOTAL (CORR.)	9.775000	39			

R-SQUARED = 0.0360834
R-SQUARED (ADJ. FOR D.F.) = 0
STND. ERROR OF EST. = 0.511596

Press ENTER to continue.

Table 42

Regression of Distance from Ancestral Home on Other Predictor Variables

MODEL FITTING RESULTS

VARIABLE	COEFFICIENT	STND. ERROR	T-VALUE	PROB(>|T|)
CONSTANT	4.512236	1.261573	3.5767	.0009
X7	0.03892	0.322229	.1208	.9045
X16	-0.451698	0.601268	-.7512	.4570
X21	-0.309225	0.22293	-1.3871	.1733

0 CASES WITH MISSING VALUES WERE EXCLUDED.

RESIDUALS PLACED IN VARIABLE: RESIDUALS

ANALYSIS OF VARIANCE FOR THE FULL REGRESSION

SOURCE	SUM OF SQUARES	DF	MEAN SQUARE	F-RATIO	PROB(>F)
MODEL	8.2231228	3	2.7410409	.7922600	.5062158
ERROR	124.55188	36	3.45977		
TOTAL (CORR.)	132.77500	39			

R-SQUARED = 0.0619328
R-SQUARED (ADJ. FOR D.F.) = 0
STND. ERROR OF EST. = 1.86005

Press ENTER to continue.

Table 43

Regression of Parental Family Size (X21) on Other Predictor Variables

MODEL FITTING RESULTS

VARIABLE	COEFFICIENT	STND. ERROR	T-VALUE	PROB(>|T|)
CONSTANT	0.620186	1.064821	.5824	.5636
X7	0.468812	0.221377	2.1177	.0406
X16	-0.442759	0.435176	-1.0174	.3152
X20	-0.164068	0.118282	-1.3871	.1733

0 CASES WITH MISSING VALUES WERE EXCLUDED.

RESIDUALS PLACED IN VARIABLE: RESIDUALS

ANALYSIS OF VARIANCE FOR THE FULL REGRESSION

SOURCE	SUM OF SQUARES	DF	MEAN SQUARE	F-RATIO	PROB(>F)
MODEL	13.815372	3	4.605124	2.508669	.074276
ERROR	66.084628	36	1.835684		
TOTAL (CORR.)	79.900000	39			

R-SQUARED = 0.172908
R-SQUARED (ADJ. FOR D.F.) = 0.103984
STND. ERROR OF EST. = 1.35487

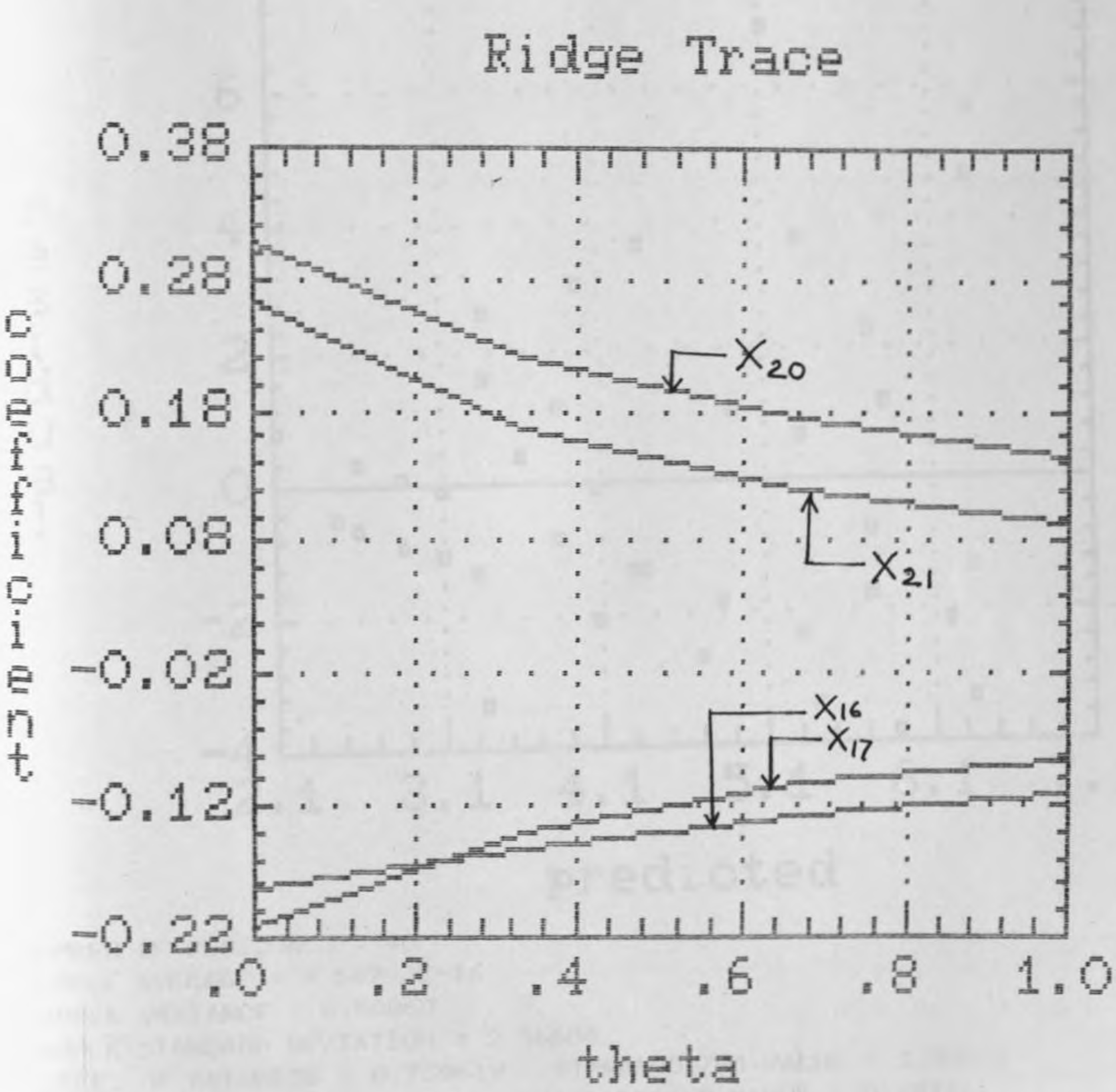
Press ENTER to continue.

Table 44

Some Selected Ridge Coefficients

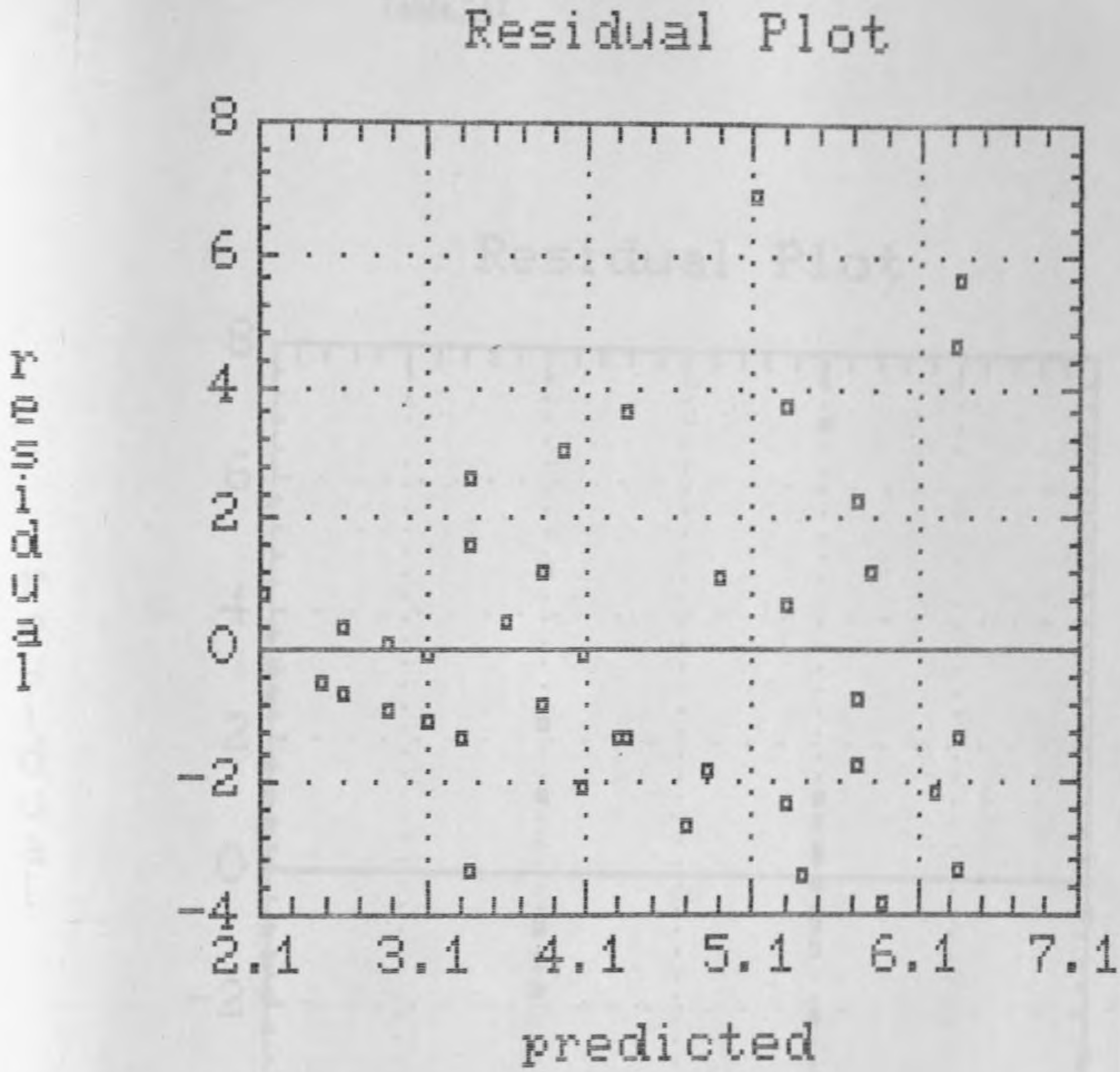
```
ENTER NAME OF DEPENDENT VARIABLE:  X4
ENTER NAME OF FIRST INDEPENDENT VARIABLE:  X7
ENTER NAME OF NEXT INDEPENDENT VARIABLE:  X16
ENTER NAME OF NEXT INDEPENDENT VARIABLE:  X20
ENTER NAME OF NEXT INDEPENDENT VARIABLE:  X21



ENTER NAME OF NEXT INDEPENDENT VARIABLE:
ENTER VALUE FOR THETA (QUIT):  0
PARAMETER ESTIMATES FOR THETA = 0:
-0.211025 -0.182785 0.309575 0.260201
ENTER VALUE FOR THETA (QUIT):  .1
PARAMETER ESTIMATES FOR THETA = 0.1:
-0.182401 -0.173813 0.276302 0.222407
ENTER VALUE FOR THETA (QUIT):  .2
PARAMETER ESTIMATES FOR THETA = 0.2:
-0.160833 -0.164683 0.249949 0.194134
ENTER VALUE FOR THETA (QUIT):  .3
PARAMETER ESTIMATES FOR THETA = 0.3:
-0.143966 -0.155926 0.228461 0.172185
ENTER VALUE FOR THETA (QUIT):  .5
PARAMETER ESTIMATES FOR THETA = 0.5:
-0.119227 -0.140174 0.19535 0.140333
ENTER VALUE FOR THETA (QUIT):  .6
PARAMETER ESTIMATES FOR THETA = 0.6:
-0.109868 -0.133221 0.182275 0.128414
ENTER VALUE FOR THETA (QUIT):  1
PARAMETER ESTIMATES FOR THETA = 1:
-0.0838071 -0.11065 0.144094 0.0957356
ENTER VALUE FOR THETA (QUIT):
 1HELP   2LABEL  3SAUSC  4RECORD 5        6        7        8        9REVIEW 10QUIT
INPUT    SAT JUN 23  1990 11:13:00 PM  VERSION 1.1                            REC:OFF
```

Table 45



Ridge Trace
0.38
0.28
0.18
0.08
-0.02
-0.12
-0.22
coefficient
.0
.2
.4
.6
.8
1.0
theta
X20
X21
X16
X17

Table 46



NUMBER OF RESIDUALS = 40
SAMPLE AVERAGE = 9.54792E-16
SAMPLE VARIANCE = 6.58868
SAMPLE STANDARD DEVIATION = 2.56684
COEFF. OF SKEWNESS = 0.758619   STANDARDIZED VALUE = 1.95875
COEFF. OF KURTOSIS = 3.1403   STANDARDIZED VALUE = 0.181121
DURBIN-WATSON STATISTIC = 1.67165

Press ENTER to continue.

Table 47

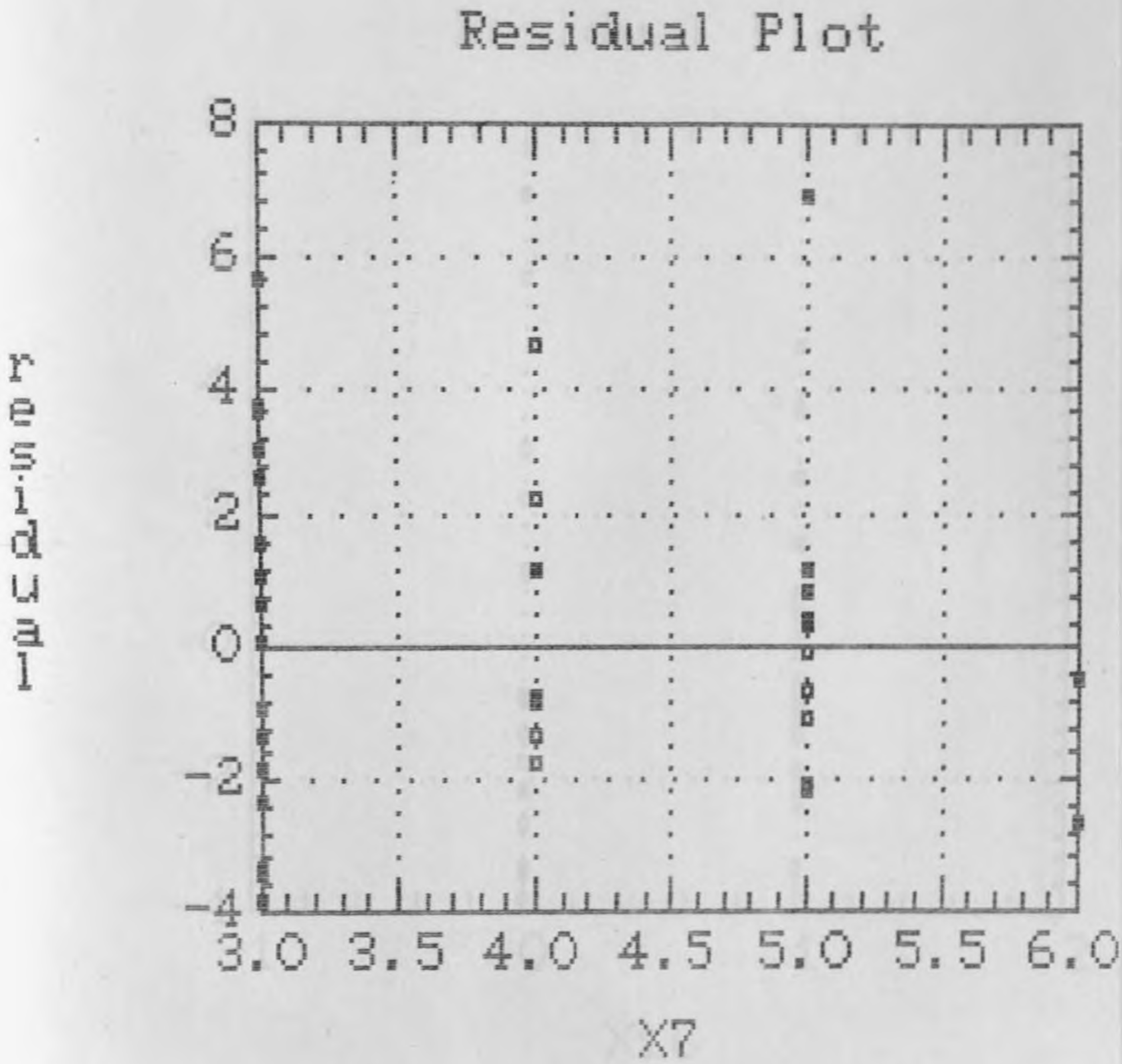


Table 48

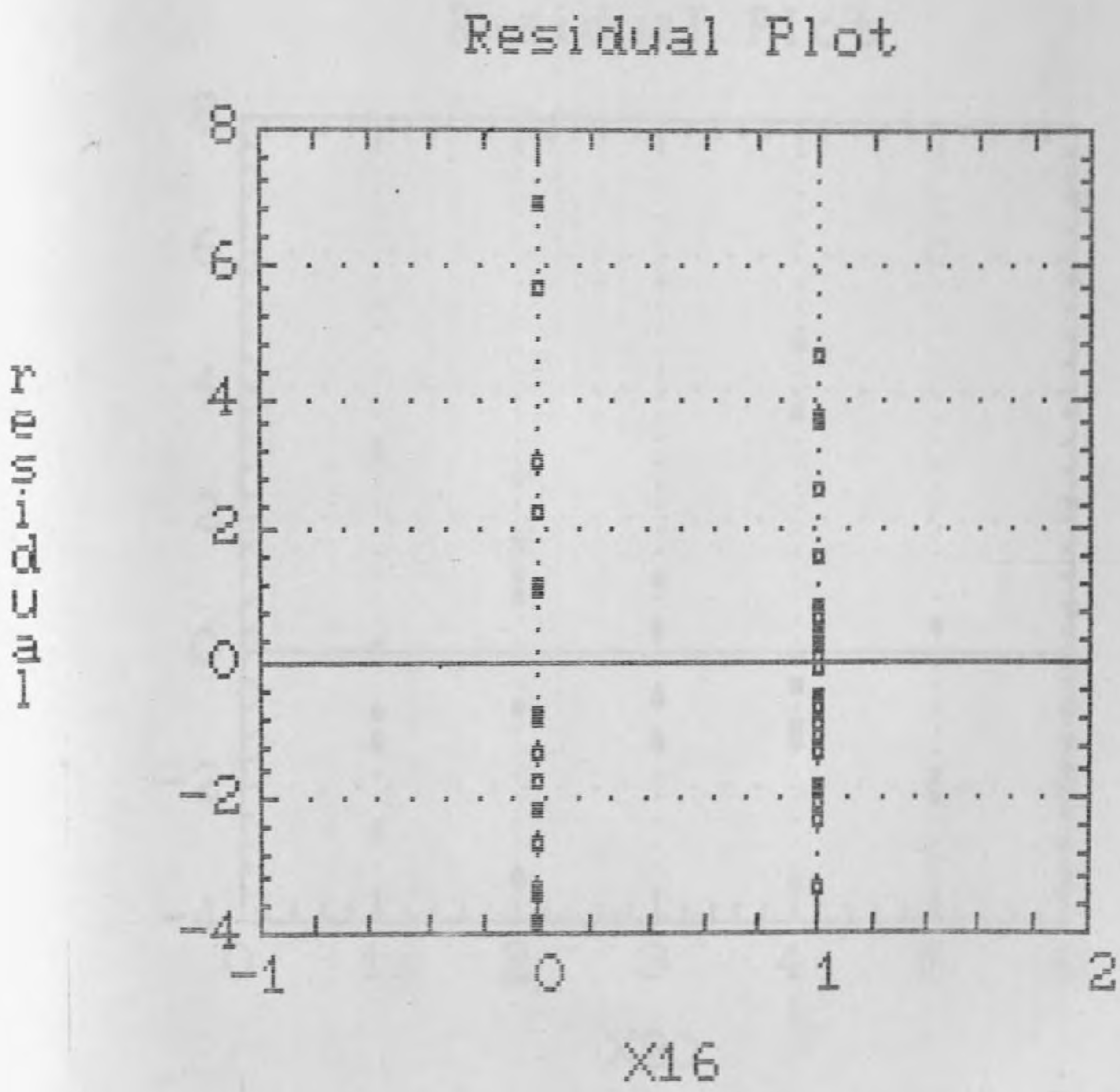


Table 49

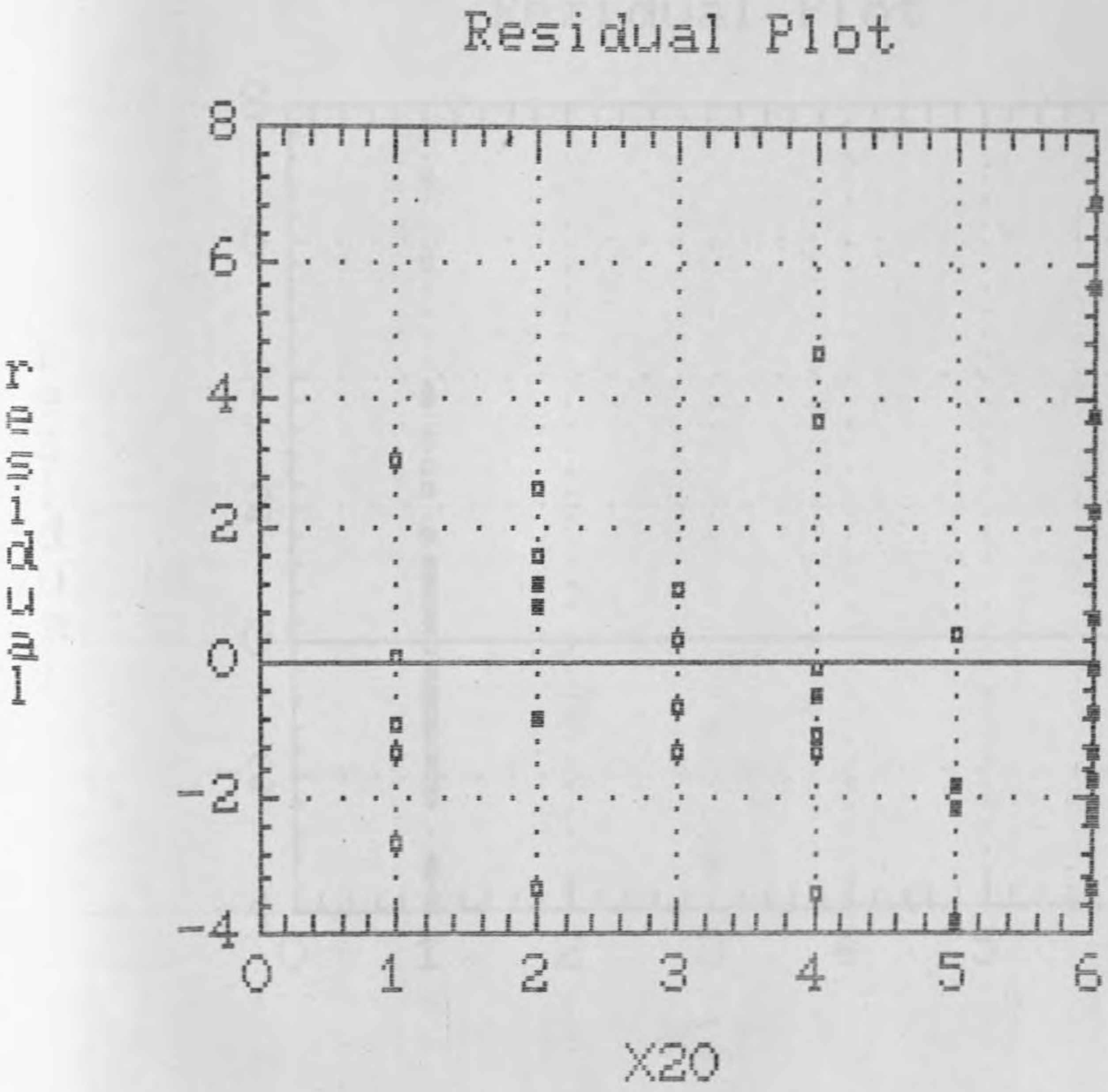


Table 50

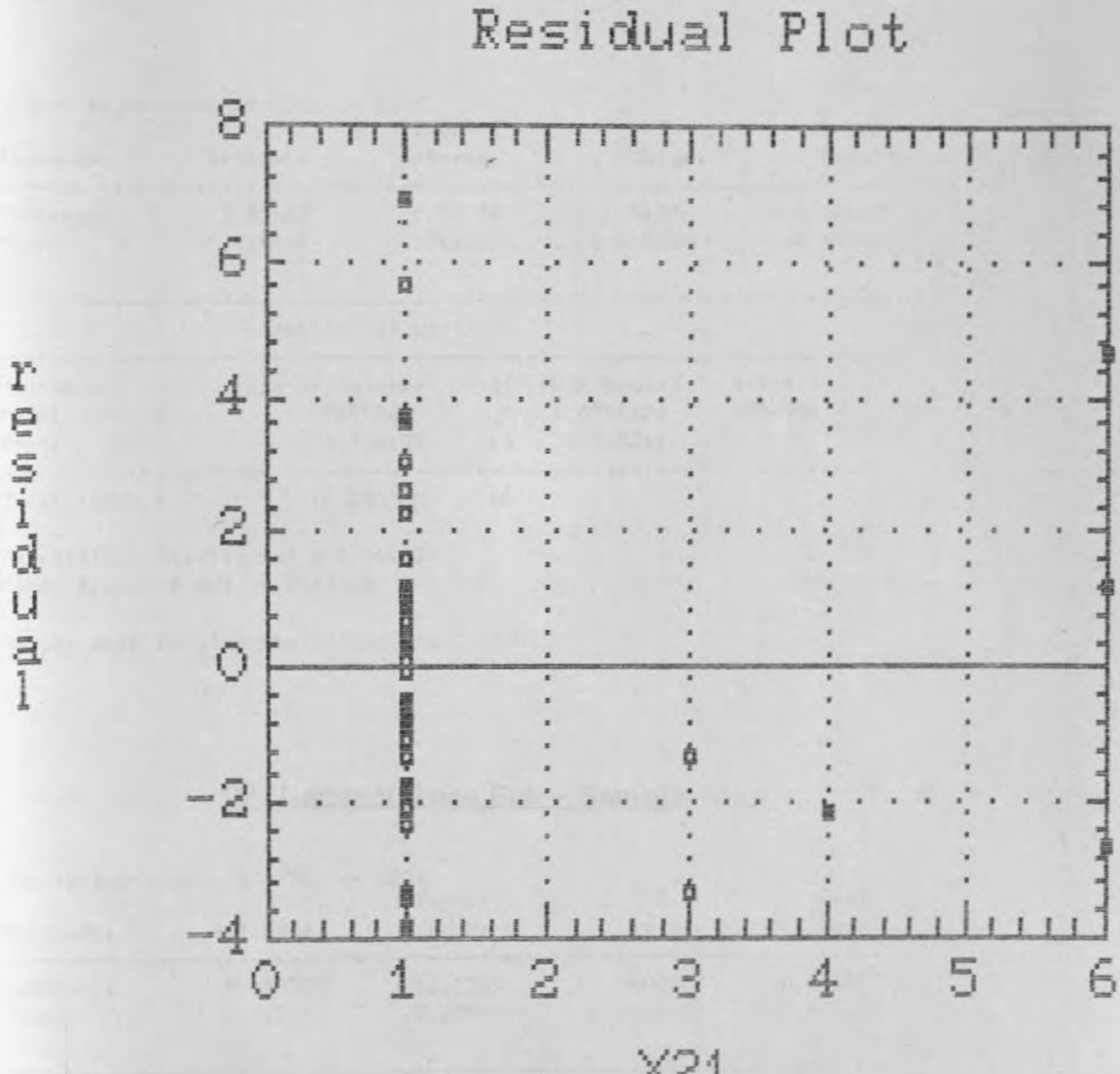


Table 51

(a) Small Values Sub - Sample

Simple Regression of Y20A on X20A

Parameter	Estimate	Standard Error	T Value	Prob. Level
Intercept	2.93103	1.52856	1.91751	0.0774102
Slope	0.465517	0.764281	0.609091	0.55296

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio
Model	1.6758621	1	1.6758621	.3709924
Error	58.724138	13	4.517241	
Total (Corr.)	60.400000	14		

Correlation Coefficient = 0.166571
Stnd. Error of Est. = 2.12538

Do you want to plot the fitted line? (Y/N):

(b) Large Values Sub - Sample

Simple Regression of Y20B on X20B

Parameter	Estimate	Standard Error	T Value	Prob. Level
Intercept	-5.23077	14.5265	-0.360084	0.724567
Slope	1.84615	2.47196	0.746837	0.468455

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio
Model	5.9076923	1	5.9076923	.5577654
Error	137.69231	13	10.59172	
Total (Corr.)	143.60000	14		

Correlation Coefficient = 0.20283
Stnd. Error of Est. = 3.25449

Do you want to plot the fitted line? (Y/N):

Table 52

Regression Results Model 1

MODEL FITTING RESULTS

VARIABLE	COEFFICIENT	STND. ERROR	T-VALUE	PROB(>|T|)
CONSTANT	4.784754	2.139497	2.2364	.0311
X7	-0.617521	0.469489	-1.3153	.1961
X16	-1.049045	0.882713	-1.1884	.2418
X20	0.48208	0.242785	1.9856	.0541
X21	0.522332	0.333309	1.5671	.1252

0 CASES WITH MISSING VALUES WERE EXCLUDED.

RESIDUALS PLACED IN VARIABLE: RESIDUALS

ANALYSIS OF VARIANCE FOR THE FULL REGRESSION

SOURCE	SUM OF SQUARES	DF	MEAN SQUARE	F-RATIO	PROB(>F)
MODEL	65.016493	4	16.254123	2.213954	.087534
ERROR	256.95851	35	7.34167		
TOTAL (CORR.)	321.97500	39			

R-SQUARED = 0.20193
R-SQUARED (ADJ. FOR D.F.) = 0.110722
STND. ERROR OF EST. = 2.70955

Press ENTER to continue.

Table 53

Regression Results Model 2

MODEL FITTING RESULTS

VARIABLE	COEFFICIENT	STND. ERROR	T-VALUE	PROB(>|T|)
CONSTANT	5.864111	0.775442	7.5623	.0000
X7*X16	-0.603239	0.248688	-2.4257	.0200
X20*X21	0.001747	0.000632	2.7666	.0086

0 CASES WITH MISSING VALUES WERE EXCLUDED.

RESIDUALS PLACED IN VARIABLE: RESIDUALS

ANALYSIS OF VARIANCE FOR THE FULL REGRESSION

SOURCE	SUM OF SQUARES	DF	MEAN SQUARE	F-RATIO	PROB(>F)
MODEL	80.248306	2	40.124153	6.141621	.004975
ERROR	241.72669	37	6.53315		
TOTAL (CORR.)	321.97500	39			

R-SQUARED = 0.249238
R-SQUARED (ADJ. FOR D.F.) = 0.208656
STND. ERROR OF EST. = 2.556

Press ENTER to continue.

Table 54

Regression Results Model 3

MODEL FITTING RESULTS

VARIABLE	COEFFICIENT	STND. ERROR	T-VALUE	PROB(>|T|)
CONSTANT	3.066455	2.597714	1.1804	.2450
X7	0.773792	0.958172	.8076	.4242
X16	2.684967	3.15336	.8515	.3997
X20	0.360243	0.234759	1.5345	.1330
X21	-0.458921	0.509707	-.9004	.3735
X7*X16	-1.549453	1.10844	-1.3979	.1701
X20*X21	0.002249	0.00089	2.5278	.0156

0 CASES WITH MISSING VALUES WERE EXCLUDED.

RESIDUALS PLACED IN VARIABLE: RESIDUALS

ANALYSIS OF VARIANCE FOR THE FULL REGRESSION

SOURCE	SUM OF SQUARES	DF	MEAN SQUARE	F-RATIO	PROB(>F)
MODEL	107.61593	6	17.93599	2.76120	.02752
ERROR	214.35907	33	6.49573		
TOTAL (CORR.)	321.97500	39			

R-SQUARED = 0.334237
R-SQUARED (ADJ. FOR D.F.) = 0.213189
STND. ERROR OF EST. = 2.54867

Press ENTER to continue.

Table 55

Regression Results Optimal Model

MODEL FITTING RESULTS

VARIABLE	COEFFICIENT	STND. ERROR	T-VALUE	PROB(>|T|)
CONSTANT	4.784754	2.139497	2.2364	.0311
X7	-0.617521	0.469489	-1.3153	.1961
X16	-1.049045	0.882713	-1.1884	.2418
X20	0.48208	0.242785	1.9856	.0541
X21	0.522332	0.333309	1.5671	.1252

0 CASES WITH MISSING VALUES WERE EXCLUDED.

RESIDUALS PLACED IN VARIABLE: RESIDUALS

ANALYSIS OF VARIANCE FOR THE FULL REGRESSION

SOURCE	SUM OF SQUARES	DF	MEAN SQUARE	F-RATIO	PROB(>F)
MODEL	65.016493	4	16.254123	2.213954	.087534
ERROR	256.95851	35	7.34167		
TOTAL (CORR.)	321.97500	39			

R-SQUARED = 0.20193
R-SQUARED (ADJ. FOR D.F.) = 0.110722
STND. ERROR OF EST. = 2.70955

Press ENTER to continue.

Table 56

(a) Stepwise Regression at 95% Confidence Level

```
                                  STEPWISE REGRESSION
SELECTION:     FORWARD                                           CONTROL:       AUTOMATIC
F-TO-ENTER =   2.317          MAX STEPS =  50                    F-TO-REMOVE = 2.317
                                     STEP 1
R-SQUARED = 0.270599
R-SQUARED (ADJ.) = 0.251404                          MSE = 7.08239 WITH 38 D.F.
 VARIABLES CURRENTLY IN MODEL                 VARIABLES CURRENTLY NOT IN MODEL
VARIABLE                    COEFF.  F-REMOVE VARIABLE            PARTIAL CORR.   F-ENTER
  5. X18                   .86652   14.0976  1. X8                      -.0701      .1827
                                             2. X11                      .1932     1.4347
                                             3. X16                     -.2202     1.8863
                                             4. X17                     -.2069     1.6547
                                             6. X20                     -.2040     1.6063
                                             7. X21                      .1726     1.1356
```

(b) Stepwise Regression at 99% Confidence Level

```
                                  STEPWISE REGRESSION
SELECTION:     FORWARD                                           CONTROL:       AUTOMATIC
F-TO-ENTER =   3.27           MAX STEPS =  50                    F-TO-REMOVE = 3.27
                                     STEP 1
R-SQUARED = 0.270599
R-SQUARED (ADJ.) = 0.251404                          MSE = 7.08239 WITH 38 D.F.
 VARIABLES CURRENTLY IN MODEL                 VARIABLES CURRENTLY NOT IN MODEL
VARIABLE                    COEFF.  F-REMOVE VARIABLE            PARTIAL CORR.   F-ENTER
  5. X18                   .86652   14.0976  1. X8                      -.0701      .1827
                                             2. X11                      .1932     1.4347
                                             3. X16                     -.2202     1.8863
                                             4. X17                     -.2069     1.6547
                                             6. X20                     -.2040     1.6063
                                             7. X21                      .1726     1.1356
```

Table 57

Regression of Primary School Handcraft on Other Predictor Variables

MODEL FITTING RESULTS

VARIABLE	COEFFICIENT	STND. ERROR	T-VALUE	PROB(>|T|)
CONSTANT	0.807032	0.353468	2.2832	.0280
X11	0.15181	0.169375	.8963	.3756
X16	-0.159551	0.179203	-.8903	.3787
X17	-0.086556	0.162304	-.5333	.5969
X18	0.06477	0.044041	1.4707	.1494
X20	-0.097226	0.047562	-2.0442	.0477
X21	0.019933	0.032866	.6065	.5477

0 CASES WITH MISSING VALUES WERE EXCLUDED.

RESIDUALS PLACED IN VARIABLE: RESIDUALS

ANALYSIS OF VARIANCE FOR THE FULL REGRESSION

SOURCE	SUM OF SQUARES	DF	MEAN SQUARE	F-RATIO	PROB(>F)
MODEL	2.1126493	6	.3521082	1.4778749	.2161678
ERROR	7.862351	33	.238253		
TOTAL (CORR.)	9.975000	39			

R-SQUARED = 0.211794
R-SQUARED (ADJ. FOR D.F.) = 0.0684843
STND. ERROR OF EST. = 0.488112

Press ENTER to continue.

Table 58

Regression of Formal Technical Education on Other Predictor Variables

MODEL FITTING RESULTS

VARIABLE	COEFFICIENT	STND. ERROR	T-VALUE	PROB(>|T|)
CONSTANT	0.493101	0.376591	1.3094	.1981
X8	0.156546	0.174659	.8963	.3756
X16	-0.308127	0.176165	-1.7491	.0881
X17	0.169242	0.162882	1.0390	.3052
X18	-0.041794	0.045589	-.9168	.3649
X20	0.026857	0.051051	.5261	.6018
X21	-0.018703	0.033402	-.5599	.5787

0 CASES WITH MISSING VALUES WERE EXCLUDED.

RESIDUALS PLACED IN VARIABLE: RESIDUALS

ANALYSIS OF VARIANCE FOR THE FULL REGRESSION

SOURCE	SUM OF SQUARES	DF	MEAN SQUARE	F-RATIO	PROB(>F)
MODEL	1.4923827	6	.2487304	1.0123942	.4342152
ERROR	8.107617	33	.245685		
TOTAL (CORR.)	9.600000	39			

R-SQUARED = 0.155457
R-SQUARED (ADJ. FOR D.F.) = 1.90317E-3
STND. ERROR OF EST. = 0.495667

Press ENTER to continue.

Table 59

Regression of Preference for Formal Job on Other Predictor Variables

MODEL FITTING RESULTS

VARIABLE	COEFFICIENT	STND. ERROR	T-VALUE	PROB(>|T|)
CONSTANT	1.122421	0.308446	3.6390	.0008
X8	-0.147023	0.165132	-.8903	.3787
X11	-0.275344	0.157421	-1.7491	.0881
X17	0.098344	0.155533	.6323	.5309
X18	-0.042943	0.042995	-.9988	.3241
X20	-0.042438	0.047894	-.8861	.3810
X21	0.007424	0.031698	.2342	.8160

0 CASES WITH MISSING VALUES WERE EXCLUDED.

RESIDUALS PLACED IN VARIABLE: RESIDUALS

ANALYSIS OF VARIANCE FOR THE FULL REGRESSION

SOURCE	SUM OF SQUARES	DF	MEAN SQUARE	F-RATIO	PROB(>F)
MODEL	1.5299998	6	.2550000	1.1614905	.3502316
ERROR	7.245000	33	.219545		
TOTAL (CORR.)	8.775000	39			

R-SQUARED = 0.174359
R-SQUARED (ADJ. FOR D.F.) = 0.0242424
STND. ERROR OF EST. = 0.468557

Press ENTER to continue.

Table 60

Regression of Market Consideration on Other Predictor Variables

MODEL FITTING RESULTS

VARIABLE	COEFFICIENT	STND. ERROR	T-VALUE	PROB(>|T|)
CONSTANT	0.379624	0.400795	.9472	.3494
X8	-0.098718	0.18511	-.5333	.5969
X11	0.187183	0.180149	1.0390	.3052
X16	0.12172	0.192502	.6323	.5309
X18	0.056606	0.04754	1.1907	.2410
X20	-0.039779	0.053467	-.7440	.4613
X21	-0.010172	0.03525	-.2886	.7744

0 CASES WITH MISSING VALUES WERE EXCLUDED.

RESIDUALS PLACED IN VARIABLE: RESIDUALS

ANALYSIS OF VARIANCE FOR THE FULL REGRESSION

SOURCE	SUM OF SQUARES	DF	MEAN SQUARE	F-RATIO	PROB(>F)
MODEL	.8079098	6	.1346516	.4955346	.8069857
ERROR	8.967090	33	.271730		
TOTAL (CORR.)	9.775000	39			

R-SQUARED = 0.0826506
R-SQUARED (ADJ. FOR D.F.) = 0
STND. ERROR OF EST. = 0.521277

Press ENTER to continue.

Table 61

Regression of - Initial Capital on Other Predictor Variables

MODEL FITTING RESULTS

VARIABLE	COEFFICIENT	STND. ERROR	T-VALUE	PROB(>|T|)
CONSTANT	2.492848	1.390297	1.7930	.0807
X8	0.949658	0.645736	1.4707	.1494
X11	-0.594245	0.648199	-.9168	.3649
X16	-0.683278	0.684117	-.9988	.3241
X17	0.727711	0.611162	1.1907	.2410
X20	0.052676	0.193088	.2728	.7864
X21	-0.01519	0.126518	-.1201	.9051

0 CASES WITH MISSING VALUES WERE EXCLUDED.

RESIDUALS PLACED IN VARIABLE: RESIDUALS

ANALYSIS OF VARIANCE FOR THE FULL REGRESSION

SOURCE	SUM OF SQUARES	DF	MEAN SQUARE	F-RATIO	PROB(>F)
MODEL	17.697045	6	2.949507	.844340	.545138
ERROR	115.27796	33	3.49327		
TOTAL (CORR.)	132.97500	39			

R-SQUARED = 0.133086
R-SQUARED (ADJ. FOR D.F.) = 0
STND. ERROR OF EST. = 1.86903

Press ENTER to continue.

Table 62

Regression of Distance to Ancestral Home on Other Predictor Variables

MODEL FITTING RESULTS

VARIABLE	COEFFICIENT	STND. ERROR	T-VALUE	PROB(>|T|)
CONSTANT	5.708421	0.856021	6.6686	.0000
X8	-1.156042	0.565519	-2.0442	.0477
X11	0.30967	0.588646	.5261	.6018
X16	-0.5476	0.618001	-.8861	.3810
X17	-0.414709	0.557412	-.7440	.4613
X18	0.042718	0.156586	.2728	.7864
X21	0.027125	0.113861	.2382	.8130

0 CASES WITH MISSING VALUES WERE EXCLUDED.

RESIDUALS PLACED IN VARIABLE: RESIDUALS

ANALYSIS OF VARIANCE FOR THE FULL REGRESSION

SOURCE	SUM OF SQUARES	DF	MEAN SQUARE	F-RATIO	PROB(>F)
MODEL	15.289800	6	2.548300	.899542	.506926
ERROR	93.485200	33	2.832885		
TOTAL (CORR.)	108.77500	39			

R-SQUARED = 0.140564
R-SQUARED (ADJ. FOR D.F.) = 0
STND. ERROR OF EST. = 1.68312

Press ENTER to continue.

Table 63

Regression of Parental Family Size on Other Predictor Variables

MODEL FITTING RESULTS

VARIABLE	COEFFICIENT	STND. ERROR	T-VALUE	PROB(>|T|)
CONSTANT	2.210926	1.966187	1.1245	.2677
X8	0.553036	0.911859	.6065	.5477
X11	-0.503221	0.898692	-.5599	.5787
X16	0.223529	0.954402	.2342	.8160
X17	-0.247448	0.857508	-.2886	.7744
X18	-0.028744	0.23941	-.1201	.9051
X20	0.063294	0.265685	.2382	.8130

0 CASES WITH MISSING VALUES WERE EXCLUDED.

RESIDUALS PLACED IN VARIABLE: RESIDUALS

ANALYSIS OF VARIANCE FOR THE FULL REGRESSION

SOURCE	SUM OF SQUARES	DF	MEAN SQUARE	F-RATIO	PROB(>F)
MODEL	5.7595977	6	.9599330	.1452174	.9887617
ERROR	218.14040	33	6.61032		
TOTAL (CORR.)	223.90000	39			

R-SQUARED = 0.025724
R-SQUARED (ADJ. FOR D.F.) = 0
STND. ERROR OF EST. = 2.57105

Press ENTER to continue.

Table 64

Some Selected Ridge Coefficients

```
ENTER NAME OF DEPENDENT VARIABLE:  X4
ENTER NAME OF FIRST INDEPENDENT VARIABLE:  X8
ENTER NAME OF NEXT INDEPENDENT VARIABLE:  X11
ENTER NAME OF NEXT INDEPENDENT VARIABLE:  X16
ENTER NAME OF NEXT INDEPENDENT VARIABLE:  X17
ENTER NAME OF NEXT INDEPENDENT VARIABLE:  X18
ENTER NAME OF NEXT INDEPENDENT VARIABLE:  X20
ENTER NAME OF NEXT INDEPENDENT VARIABLE:  X21
ENTER NAME OF NEXT INDEPENDENT VARIABLE:
```

```
ENTER NAME OF NEXT INDEPENDENT VARIABLE:
ENTER VALUE FOR THETA (QUIT):  0
PARAMETER ESTIMATES FOR THETA = 0:
-0.247289 0.211209 -0.194121 -0.223794 0.583025 -0.309198 0.18288
ENTER VALUE FOR THETA (QUIT):  .1
PARAMETER ESTIMATES FOR THETA = 0.1:
-0.186898 0.175546 -0.181943 -0.18568 0.512708 -0.25994 0.160143
ENTER VALUE FOR THETA (QUIT):  .2
PARAMETER ESTIMATES FOR THETA = 0.2:
-0.145458 0.150742 -0.170875 -0.158402 0.459358 -0.224717 0.142851
ENTER VALUE FOR THETA (QUIT):  .3
PARAMETER ESTIMATES FOR THETA = 0.3:
-0.115672 0.132501 -0.160937 -0.137921 0.417146 -0.198251 0.129164
ENTER VALUE FOR THETA (QUIT):  .4
PARAMETER ESTIMATES FOR THETA = 0.4:
-0.0935055 0.11851 -0.15203 -0.121985 0.382715 -0.177611 0.118009
ENTER VALUE FOR THETA (QUIT):  .5
PARAMETER ESTIMATES FOR THETA = 0.5:
-0.0765543 0.107423 -0.144028 -0.10924 0.353977 -0.161044 0.108715
ENTER VALUE FOR THETA (QUIT):  1
PARAMETER ESTIMATES FOR THETA = 1:
-0.0315057 0.0744399 -0.11396 -0.0711521 0.259516 -0.110776 0.078379
ENTER VALUE FOR THETA (QUIT):
 1HELP   2LABEL  3SAVSC  4RECORD 5       6       7       8       9REVIEW 10QUIT
INPUT    SUN JUN 24  1990 12:06:00 AM  VERSION 1.1                     REC:OFF
```

Table 65

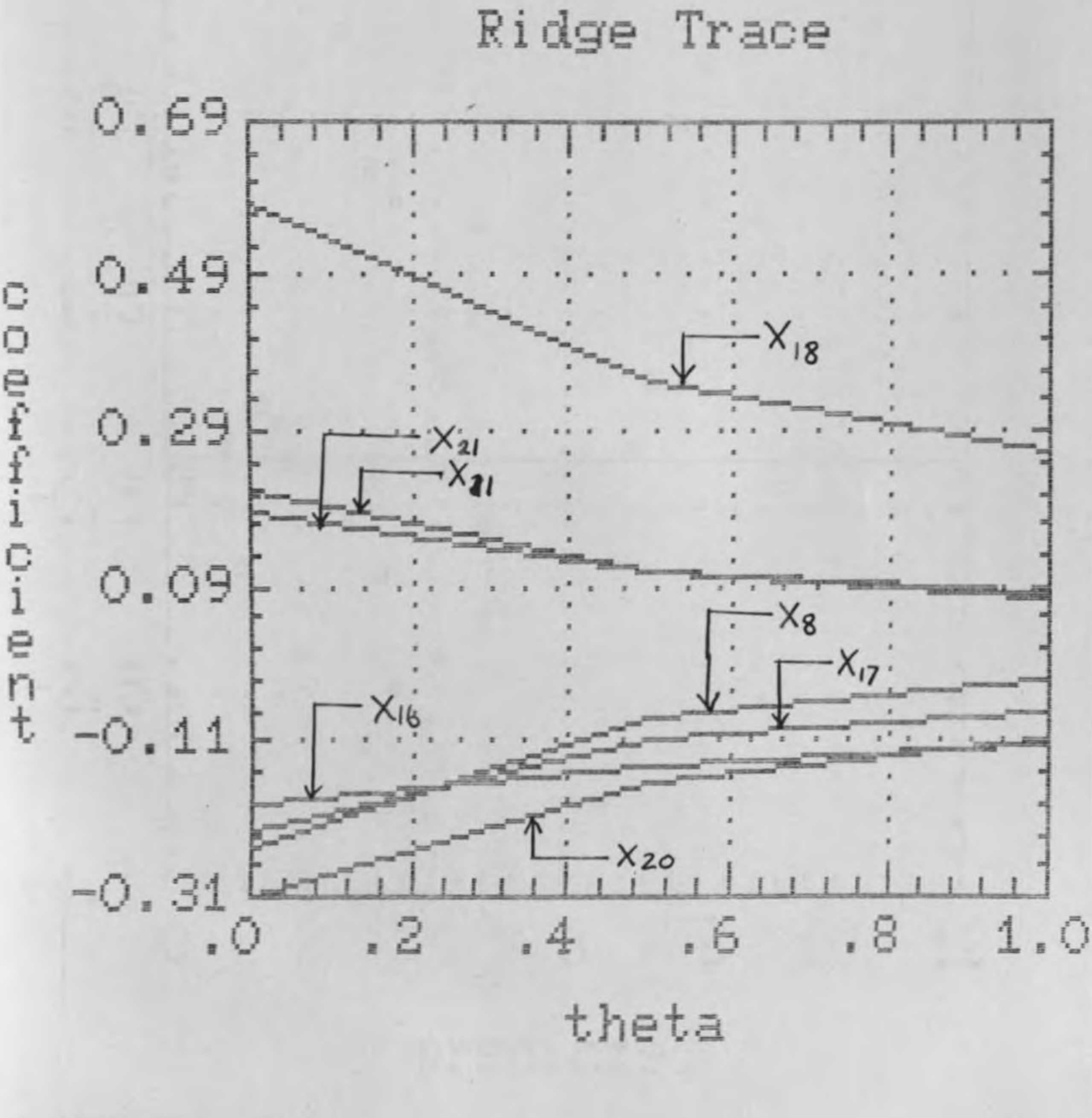
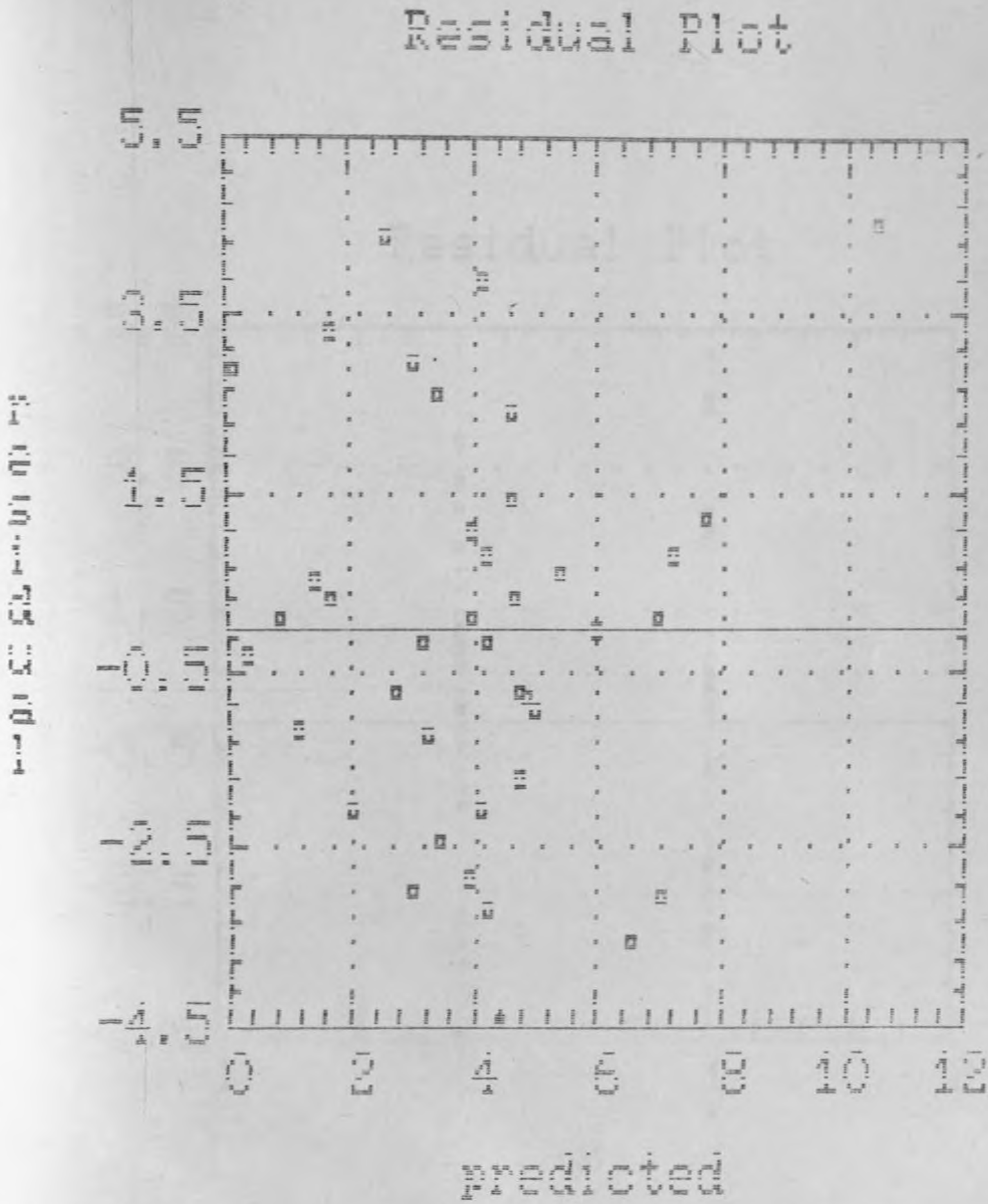

Ridge Trace
0.69
0.49
0.29
0.09
-0.11
-0.31
coefficient
.0
.2
.4
.6
.8
1.0
theta
X18
X21
X21
X8
X17
X16
X20

Table 66



NUMBER OF RESIDUALS = 40
SAMPLE AVERAGE = -1.72085E-15
SAMPLE VARIANCE = 4.96223
SAMPLE STANDARD DEVIATION = 2.22761
COEFF. OF SKEWNESS = 0.188396   STANDARDIZED VALUE = 0.486436
COEFF. OF KURTOSIS = 2.48315   STANDARDIZED VALUE = -0.667244
DURBIN-WATSON STATISTIC = 2.48068

Press ENTER to continue.

Table 67

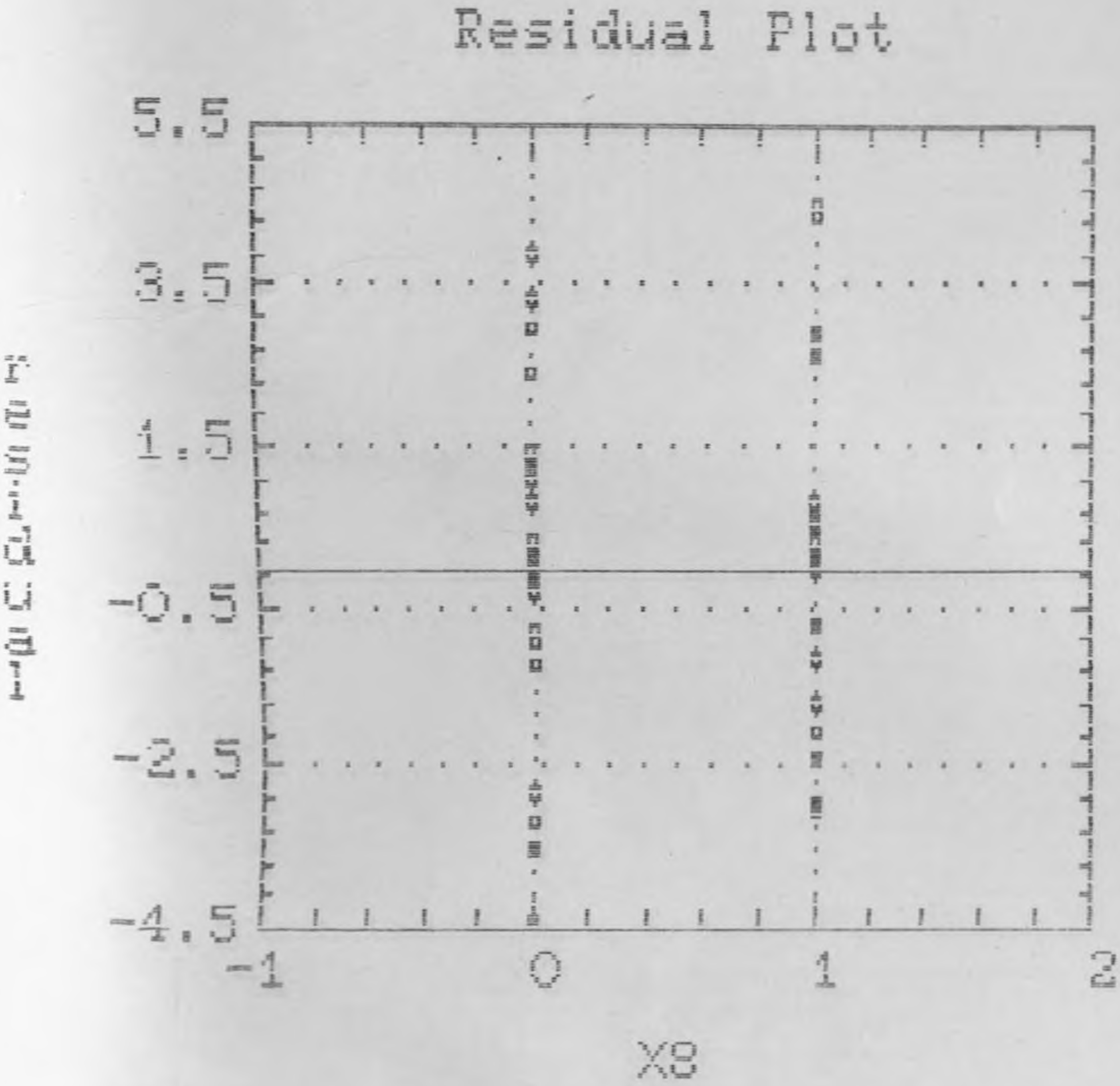


Table 68

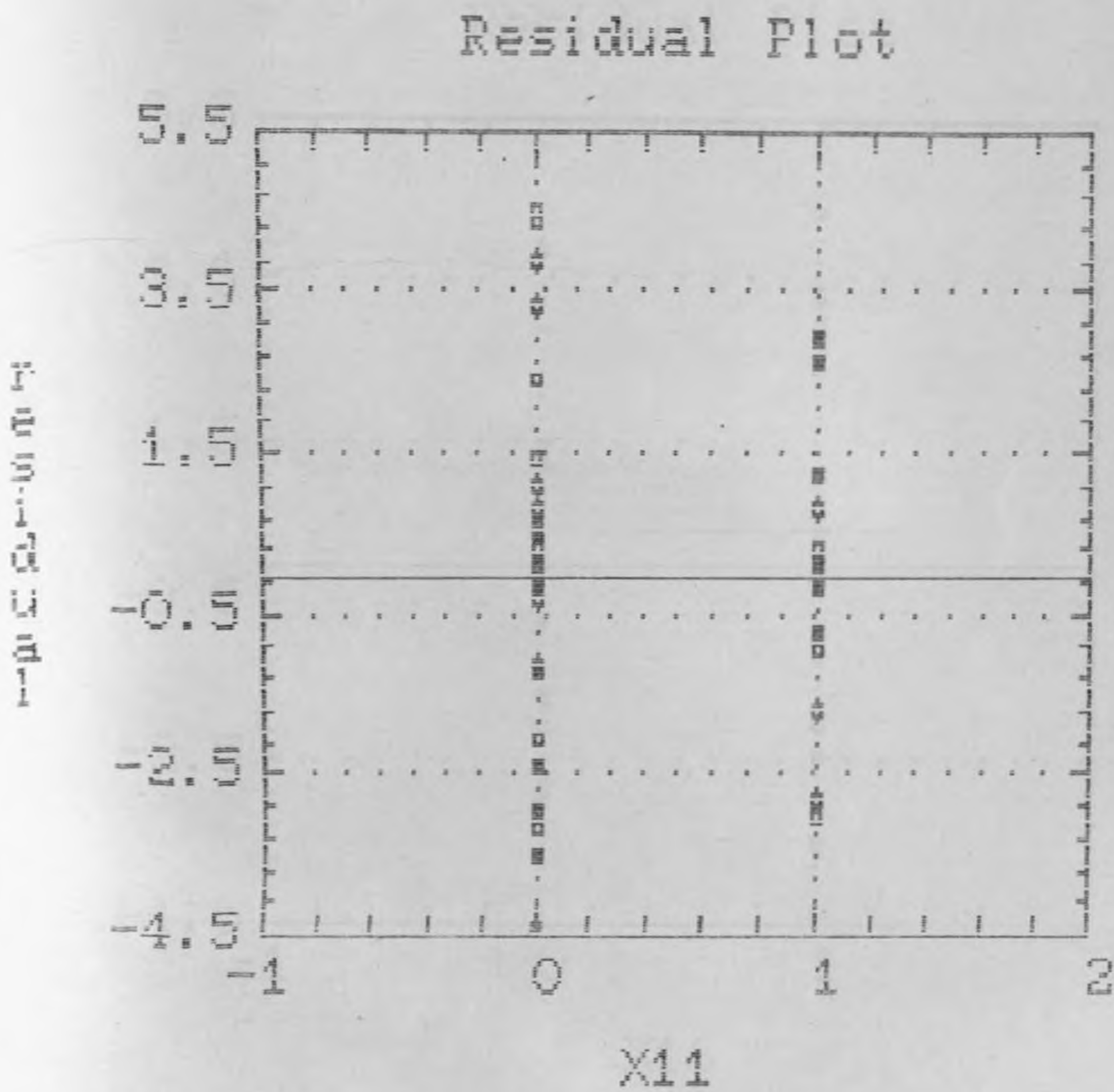


Table 69

Residual Plot

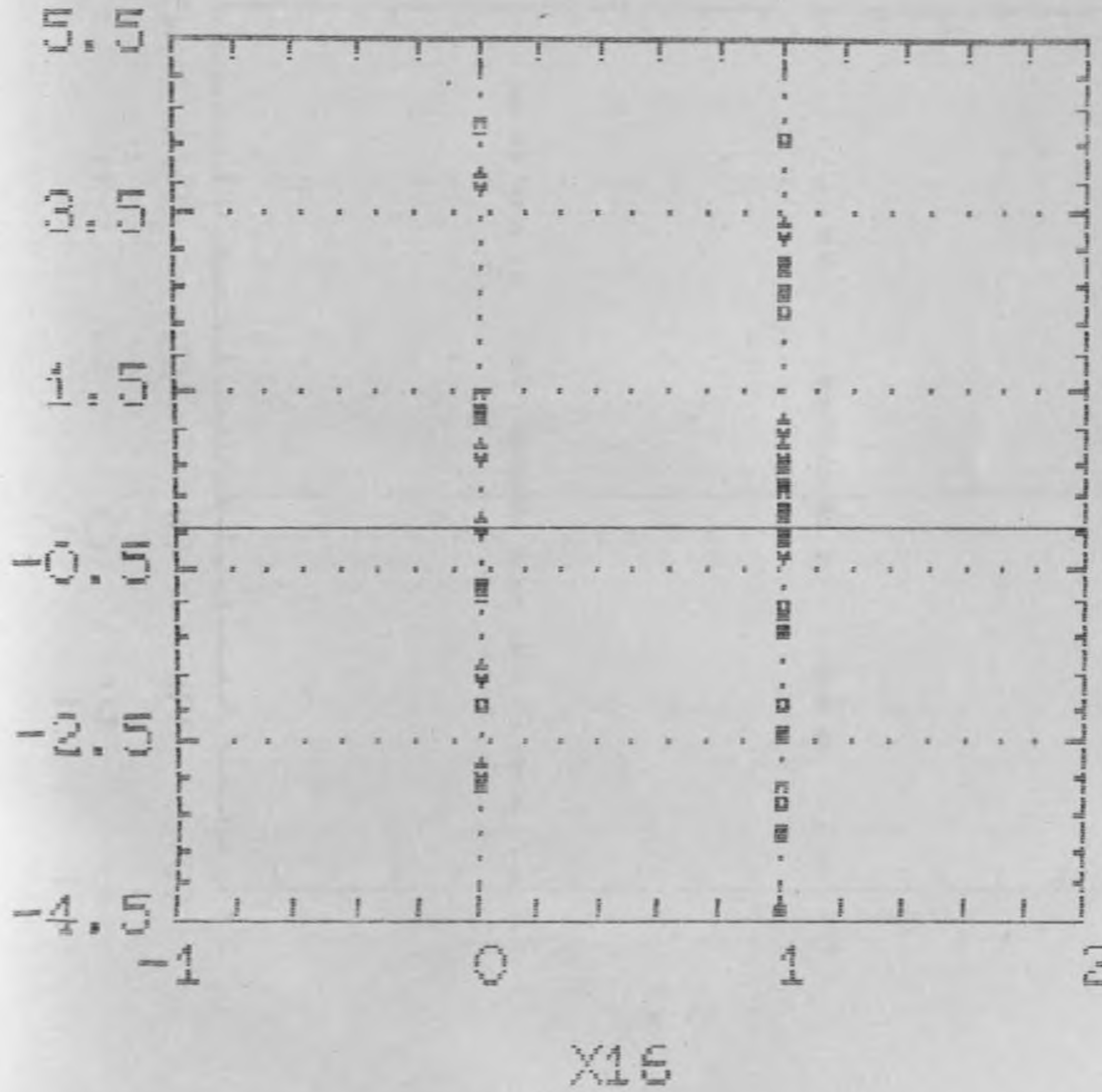


Table 70

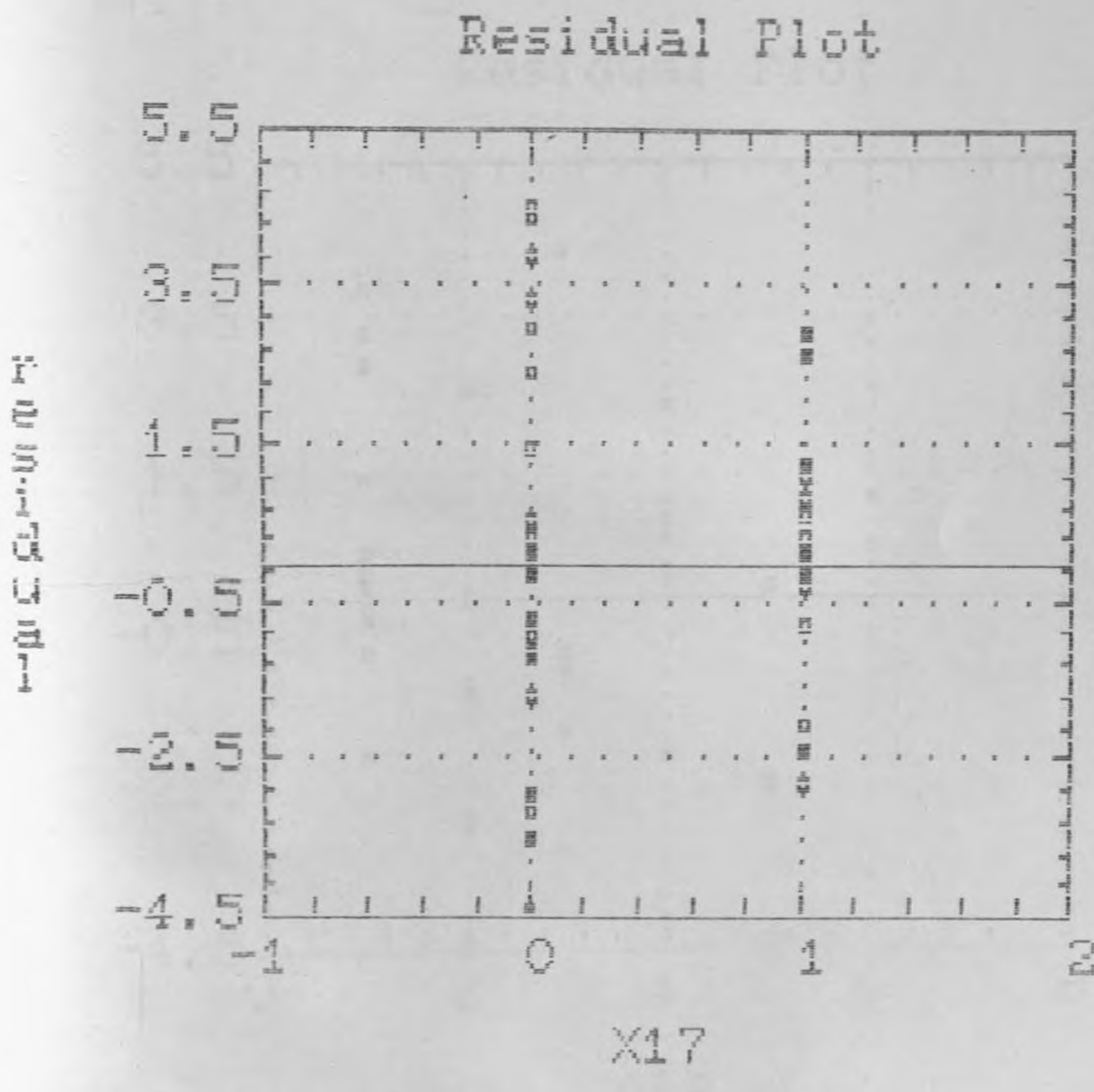


Table 71

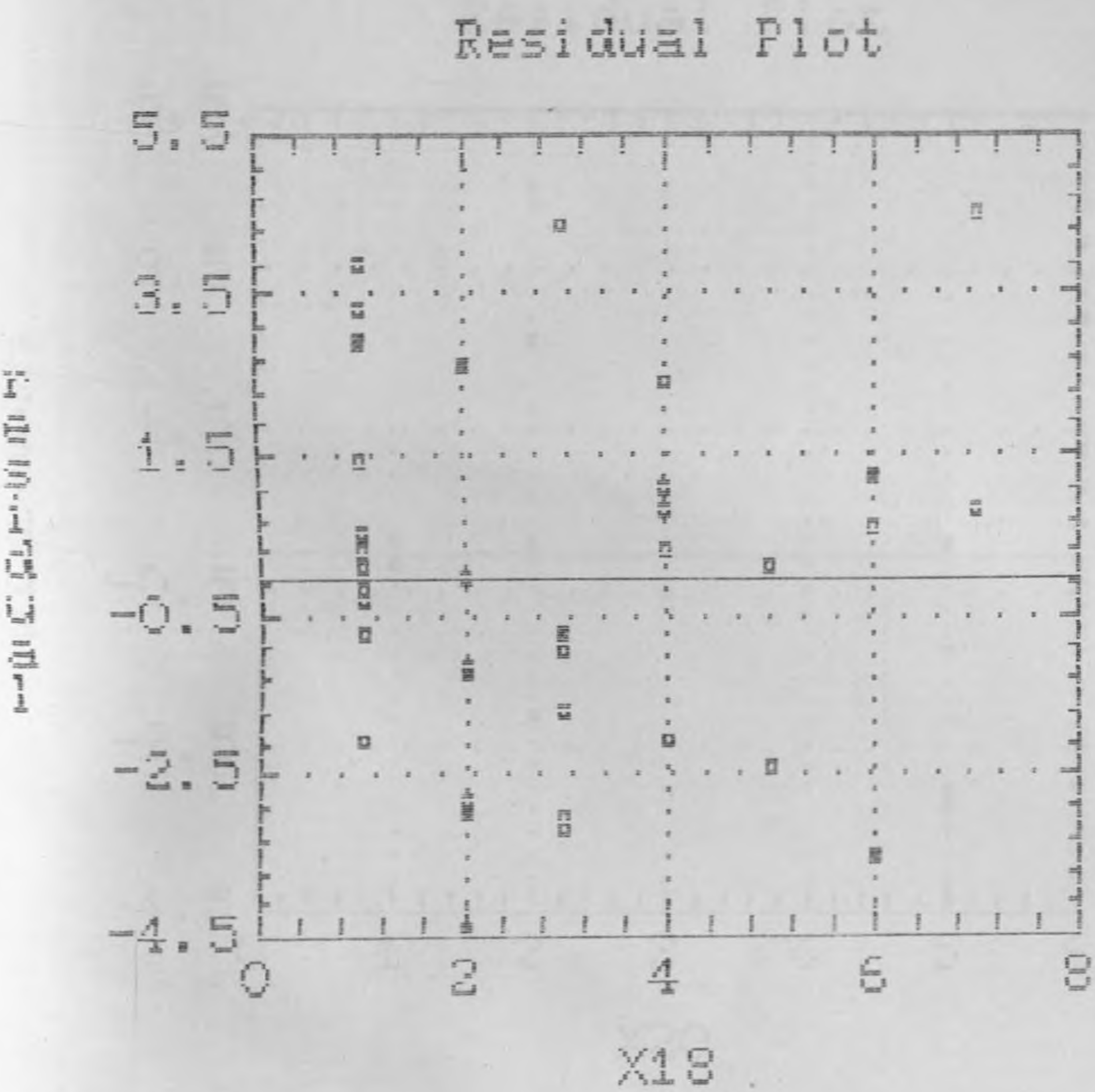


Table 72

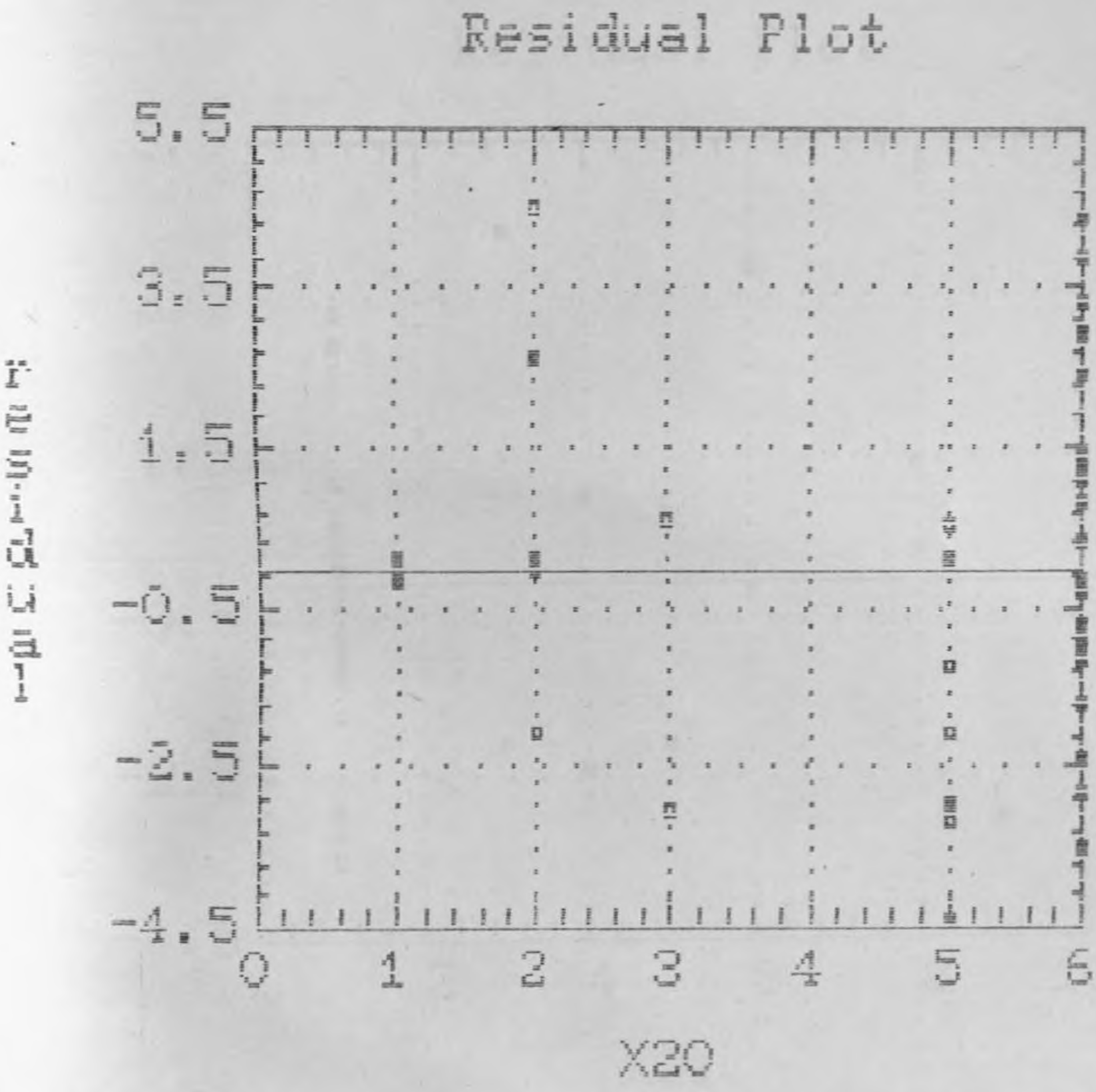


Table 73

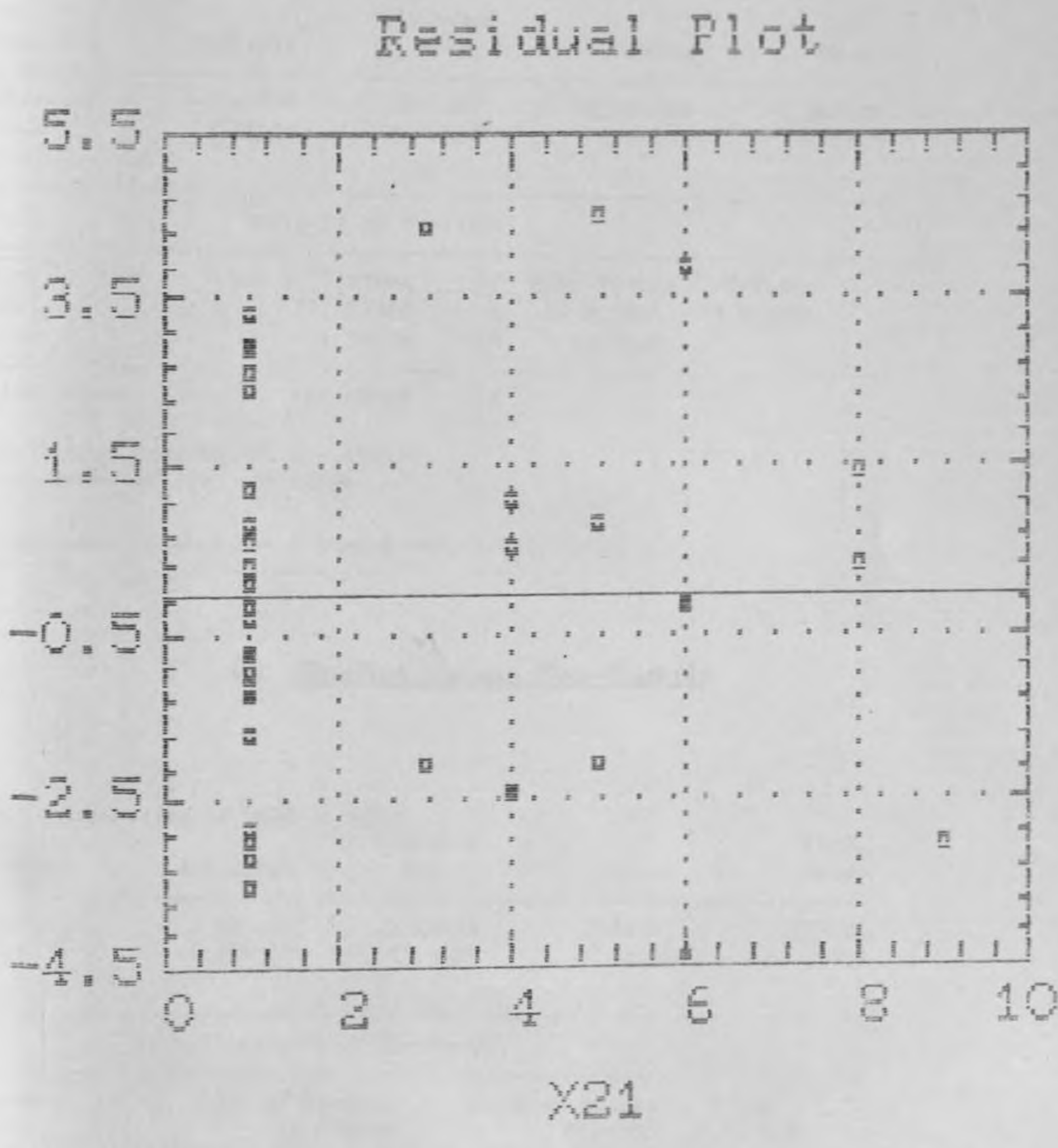


Table 74

(a) Larger Values Sub-Sample

Simple Regression of Y18B on X18B

Parameter	Estimate	Standard Error	T Value	Prob. Level
Intercept	-2.61386	2.68617	-0.973082	0.348272
Slope	1.69307	0.546041	3.10063	8.43674E-3

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio
Model	77.203960	1	77.203960	9.613885
Error	104.39604	13	8.03046	
Total (Corr.)	181.60000	14		

Correlation Coefficient = 0.652021
Stnd. Error of Est. = 2.83381

Do you want to plot the fitted line? (Y/N):

(b) Smaller Values Sub-Sample

Simple Regression of Y18A on X18A

Parameter	Estimate	Standard Error	T Value	Prob. Level
Intercept	6.57143	2.68433	2.44807	0.0293194
Slope	-3.28571	2.45045	-1.34086	0.202925

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio
Model	10.076190	1	10.076190	1.797908
Error	72.857143	13	5.604396	
Total (Corr.)	82.933333	14		

Correlation Coefficient = -0.348565
Stnd. Error of Est. = 2.36736

Do you want to plot the fitted line? (Y/N):

1HELP 2LABEL 3:

Table 75

Model 1 Multiple Regression

MODEL FITTING RESULTS

VARIABLE	COEFFICIENT	STND. ERROR	T-VALUE	PROB(>|T|)
CONSTANT	5.160616	1.916348	2.6929	.0104
X8	-1.503997	0.877041	-1.7149	.0943
X11	1.309411	0.863673	1.5161	.1376
X16	-1.258773	0.913643	-1.3778	.1761
X17	-1.374954	0.82124	-1.6742	.1021
X18	0.971181	0.229046	4.2401	.0001
X20	-0.56947	0.254346	-2.2390	.0309
X21	0.234768	0.166505	1.4100	.1665

0 CASES WITH MISSING VALUES WERE EXCLUDED.

RESIDUALS PLACED IN VARIABLE: RESIDUALS

ANALYSIS OF VARIANCE FOR THE FULL REGRESSION

SOURCE	SUM OF SQUARES	DF	MEAN SQUARE	F-RATIO	PROB(>F)
MODEL	175.44789	7	25.06398	4.14437	.00240
ERROR	193.52711	32	6.04772		
TOTAL (CORR.)	368.97500	39			

R-SQUARED = 0.475501
R-SQUARED (ADJ. FOR D.F.) = 0.360767
STND. ERROR OF EST. = 2.45921

Press ENTER to continue.

Table 76

Model 2 Multiple Regression

MODEL FITTING RESULTS

VARIABLE	COEFFICIENT	STND. ERROR	T-VALUE	PROB(>|T|)
CONSTANT	5.020165	1.507627	3.3298	.0019
X8*X11	-0.552746	1.435881	-.3850	.7024
X16*X17	-0.785531	1.646304	-.4771	.6359
X18*X21	0.000057	0.000046	1.2419	.2217

0 CASES WITH MISSING VALUES WERE EXCLUDED.

RESIDUALS PLACED IN VARIABLE: RESIDUALS

ANALYSIS OF VARIANCE FOR THE FULL REGRESSION

SOURCE	SUM OF SQUARES	DF	MEAN SQUARE	F-RATIO	PROB(>F)
MODEL	19.212323	3	6.404108	.659155	.582556
ERROR	349.76268	36	9.71563		
TOTAL (CORR.)	368.97500	39			

R-SQUARED = 0.0520694
R-SQUARED (ADJ. FOR D.F.) = 0
STND. ERROR OF EST. = 3.11699

Press ENTER to continue.

Table 77

Supermodel 3 Regression Results

MODEL FITTING RESULTS

VARIABLE	COEFFICIENT	STND. ERROR	T-VALUE	PROB(>|T|)
CONSTANT	3.466841	3.036997	1.1415	.2606
X8	-1.535259	1.151456	-1.3333	.1902
X11	1.090136	1.321122	.8252	.4143
X16	-2.214708	1.226444	-1.8058	.0787
X17	-0.682609	1.039097	-.6569	.5151
X18	1.059662	0.259862	4.0778	.0002
X20	-0.494141	0.270858	-1.8244	.0758
X21	0.181759	0.20921	.8688	.3903
X8*X11	-0.733369	1.847405	-.3970	.6936
X16*X17	2.568063	2.174369	1.1811	.2447
X18*X21	9.20057E-6	0.00005	.1825	.8561

0 CASES WITH MISSING VALUES WERE EXCLUDED.

RESIDUALS PLACED IN VARIABLE: RESIDUALS

ANALYSIS OF VARIANCE FOR THE FULL REGRESSION

SOURCE	SUM OF SQUARES	DF	MEAN SQUARE	F-RATIO	PROB(>F)
MODEL	184.84078	10	18.48408	2.91113	.0[illegible]
ERROR	184.13428	29	6.34946		
TOTAL (CORR.)	368.97500	39			

R-SQUARED = 0.500957

R-SQUARED (ADJ. FOR D.F.) = 0.328874

STND. ERROR OF EST. = 2.51981

Press ENTER to continue.

Table 78

Multiple Regression Optimal Model

MODEL FITTING RESULTS

VARIABLE	COEFFICIENT	STND. ERROR	T-VALUE	PROB(>|T|)
CONSTANT	5.160616	1.916348	2.6929	.0106
X8	-1.503997	0.877041	-1.7149	.0943
X11	1.309411	0.863673	1.5161	.1376
X16	-1.258773	0.913643	-1.3778	.1761
X17	-1.374954	0.82124	-1.6742	.1021
X18	0.971181	0.229046	4.2401	.0001
X20	-0.56847	0.254346	-2.2390	.0309
X21	0.234768	0.166505	1.4100	.1665

0 CASES WITH MISSING VALUES WERE EXCLUDED.

RESIDUALS PLACED IN VARIABLE: RESIDUALS

ANALYSIS OF VARIANCE FOR THE FULL REGRESSION

SOURCE	SUM OF SQUARES	DF	MEAN SQUARE	F-RATIO	PROB(>F)
MODEL	175.44789	7	25.06398	4.14437	.00240
ERROR	193.52711	32	6.04772		
TOTAL (CORR.)	368.97500	39			

R-SQUARED = 0.475501
R-SQUARED (ADJ. FOR D.F.) = 0.360767
STND. ERROR OF EST. = 2.45921

Press ENTER to continue.

BIBLIOGRAPHY

Anderson , D.R.,et al ,Statistics for Business and Ecomomics.West publishing company ,N.Y , 1981.

Byrkit , D.R, Elements of Statistics.D.Van Nostrand Company , N.Y.,1975.

Clark ,C.T. ,Statistical Analysis for Administrative Decisions. South-West Publishing Company,Cincinnati,1974.

Dixon ,W.J and Massey, F.J.,Introduction to Statistical Analysis. McGraw-Hill Book company Inc. N.Y, 1983.

Ekeblad , F.A.,The Statistical Method in Business :Application of Probability and Inference to Business and other Problems.John Wiley and sons , Inc. N.Y.,1962.

Fallik ,F. and Brown ,B. Statistics for Behavioral Sciences.The Dorsey Press , Homewood,Illinois,1983.

Freedman,D. et al,Statistics. W.W. Norton and Company, N.Y.,1978.

Gitlow ,H.S.,Stat City : Understanding Statistics through Realistic Applications.Richard, D. Irwin,Inc., Homewood,Illinois,1982.

Goldfeld,S.M.,and Quandt,R.E.,"Some Tests for Homoscedasticity". Journal of American Statistical Association,Vol.60, 1965,pp.539-47.

Gujarati.D. Use of Dummy Variables in Testing for Equality Between Sets of Coefficients in Linear Regressions:A Generalisation.The American Statistician,Vol.24,No.5, December 1970,pp.18-22.

Kariuki,C.N."Apriori determinants of performance in Jua Kali Sector:The case of Nairobi Eastlands".Unpublished M.B.A. Project,University of Nairobi,1989.

Kennedy ,P.,A Guide to Econometrics. MIT Press ,Cambridge , Massachusetts ,1980

Koutsoyiannis. A. Theory of Econometrics. Macmillan Publishers Ltd,Hong Kong ,1977.

Lapin, L. Statistics: Meaning and Method.Harcourt Brace Jovanovich, Inc.,N.Y.,1975.

Leabo,D.A.,Basic Statistics.Richard D. Irwin,Inc. ,Homewood , Illinois, 1972.

Lindgren,B.W., Statistical Theory.The Macmillan Co.,N.Y.,1965.

Lippman,S.A.,Elements of Probability and Statistics.Holt,Rinehart and Winston ,Inc. N.Y. 1971.

Marquardt,D.W. and Snee,R.D."Ridge Regression in Practice". The American Statistician.February 1975,Vol.29,No.1.

Mills ,F.C.,Statistical Methods. Henry Holt and Co. , N.Y.,1955.

Newton,B.L.,Statistics for Business.Science Research Associates ,Inc.,Chicago,1973.

Onuong'a,J.O. "An evaluation of pre-entry predictors for Bachelor of Commerce students at the University of Nairobi".Unpublished M.B.A Project,University of Nairobi,1988.

Sandblom,C.,Elementary Statistics for Business and Economics. Walter de Gruyter,N.Y.,1983.

Sanders,D.H,et al,Statistics:A Fresh approach .McGraw-Hill Book Co.,N.Y. 1976.

Schlaifer ,R. Probability and Statistics for Business Decisions: An Introduction to Managerial economics Under Uncertainty.McGraw-Hill Book Co.,Inc., N.Y.,1959.

Spiegel,M.R.,Theory and Problems of Statistics.Schaum's Outline Series , McGraw-Hill Book Co.,Inc., N.Y.,1961.

Spirer,H.F.,Business Statistics :A Problem -Solving Approach. Richard,D. Irwin,Inc.,Homewood,Illinois,1975.

Suits ,D.B.,Statistics : An Introduction to Quantitative Economic Research. Rand,McNally and Co.,Chicago,1963.

Summers,G.W.,Student Suppliment for Basic Statistics in Business and Economics.Wadsworth Publishing Co.,Inc., Belmomt,California,1977.

Thiel,H.,and Nagar,A.L,"Testing the Independence of Regression Disturbances".Journal of the American Statistical Association. 56 (1961),pp.763-806.

Wasserman,W. et al,Applied Linear Statistical Models.Irwin, Homewood,Illinois,1985.

Wittink,D.R.,The Application of Regression Analysis.Allyn and Bacon,Inc.,Boston,1988.

Yamane,T.,Statistics : An Introductory Analysis.Harper and Row Publishers,N.Y.,1964.