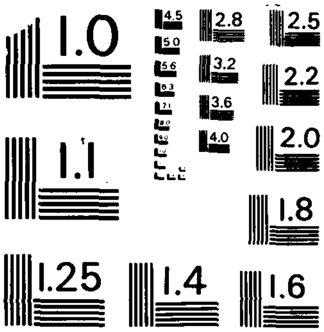


MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

2

NAVAL POSTGRADUATE SCHOOL

Monterey, California

THESIS

A MODEL FOR AND METHOD OF PREDICTING
HIGH QUALITY ARMY ENLISTMENT CONTRACTS

by

Jack E. Faires

September 1986

Thesis Advisor: Dan Boger

Approved for public release; distribution is unlimited.

ADA 175309

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a REPORT SECURITY CLASSIFICATION UNCLASSIFIED	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY 2b. DECLASSIFICATION / DOWNGRADING SCHEDULE	3 DISTRIBUTION / AVAILABILITY OF REPORT Approved for public release; distribution is unlimited.
4 PERFORMING ORGANIZATION REPORT NUMBER(S)	5 MONITORING ORGANIZATION REPORT NUMBER(S)
6a. NAME OF PERFORMING ORGANIZATION Naval Postgraduate School 6b OFFICE SYMBOL *(If applicable)* Code 55	7a. NAME OF MONITORING ORGANIZATION Naval Postgraduate School
6c. ADDRESS *(City, State, and ZIP Code)* Monterey, California 93943-5000	7b. ADDRESS *(City, State, and ZIP Code)* Monterey, California 93943-5000
8a. NAME OF FUNDING / SPONSORING ORGANIZATION 8b. OFFICE SYMBOL *(If applicable)*	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
8c. ADDRESS *(City, State, and ZIP Code)*	10 SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO PROJECT NO TASK NO WORK UNIT ACCESSION NO

11 TITLE *(Include Security Classification)*
A MODEL FOR AND METHOD OF PREDICTING HIGH QUALITY ARMY ENLISTMENT CONTRACTS

12 PERSONAL AUTHOR(S)
Faires, Jack E.

13a TYPE OF REPORT Master's Thesis	13b TIME COVERED FROM ________ TO ________	14 DATE OF REPORT *(Year, Month, Day)* 1986, September	15 PAGE COUNT 91

16 SUPPLEMENTARY NOTATION

17 COSATI CODES			18. SUBJECT TERMS *(Continue on reverse if necessary and identify by block number)*
FIELD	GROUP	SUB-GROUP	Regression Model for Predicting Army Enlistment Contracts.

19 ABSTRACT *(Continue on reverse if necessary and identify by block number)*

There are many variables that contribute to the explanation of why a person enlists in the Army. To efficiently manage personnel policy in regards to the recruitment process, the impact and significance of these variables needs to be fully understood. Ordinary least squares regression analysis is a powerful and useful tool in helping to explain the interaction of these variables. The understanding of the theories and methods behind this approach is essential. Army analysts apply regression derived results every day in a myriad of situations and operational contexts. Misuse or misunderstanding of these results can lead to inaccurate recommendations to the decision maker.

The thesis develops the framework for a parsimonious linear statistical model of quality enlistment contracts for the U.S. Army. There is a need for such a model that can be utilized by USAREC and DCSPER analysts to perform quick response analysis to 'what if' questions.

20 DISTRIBUTION / AVAILABILITY OF ABSTRACT ☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS	21 ABSTRACT SECURITY CLASSIFICATION UNCLASSIFIED
22a NAME OF RESPONSIBLE INDIVIDUAL Dan Boger	22b TELEPHONE *(Include Area Code)* (408) 646-3228 22c OFFICE SYMBOL Code 55Bo

83 APR edition may be used until exhausted
All other editions are obsolete

SECURITY CLASSIFICATION OF THIS PAGE

19. ABSTRACT

In order to facilitate further model enhancement and use, it is developed in a step-by-step fashion. The author uses a 'walk through' approach and thoroughly discusses the assumptions, procedures and analytical tools that were utilized in the model development. This approach was specifically requested by the Army analysts at USAREC.

Approved for public release; distribution is unlimited.

A Model for and Method of Predicting High Quality Army Enlistment Contracts

by

Jack E. Faires
Captain, United States Army
B.S., United States Military Academy, 1978

Submitted in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
September 1986

Author: Jack E. Faires
Jack E. Faires

Approved by: Dan C Boger
Dan Boger, Thesis Advisor

Harold Fredrickson
Harold Fredrickson, Second Reader

Peter Purdue
Peter Purdue, Chairman,
Department of Operations Research

Kneale T. Marshall
Kneale T. Marshall,
Dean of Information and Policy Sciences

ABSTRACT

There are many variables that contribute to the explanation of why a person enlists in the Army. To efficiently manage personnel policy in regards to the recruitment process, the impact and significance of these variables needs to be fully understood. Ordinary least squares regression analysis is a powerful and useful tool in helping to explain the interaction of these variables. The understanding of the theories and methods behind this approach is essential. .Army analysts apply regression derived results every day in a myriad of situations and operational contexts. Misuse or misunderstanding of these results can lead to inaccurate recommendations to the decision maker.

The thesis develops the framework for a parsimonious linear statistical model of quality enlistment contracts for the U.S. Army. There is a need for such a model that can be utilized by USAREC and DCSPER analysts to perform quick response analysis to 'what if' questions.

In order to facilitate further model enhancement and use, it is developed in a step-by-step fashion. The author uses a 'walk through' approach and thoroughly discusses the assumptions, procedures and analytical tools that were utilized in the model development. This approach was specifically requested by the Army analysts at USAREC.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

I. INTRODUCTION

The Commander, United States Army Recruiting Command (USAREC), is responsible for developing and issuing policies, procedures and standards for the recruitment of personnel into the United States Army. Each year, the Deputy Chief of Staff of Personnel (DCSPER) generates an accession mission based on the number of attritions and changes to the overall endstrength. This mission is then given to the Commander, USAREC. It is changed and updated throughout the year as policy decisions and fiscal and Congressional constraints dictate. This accession mission is broken down into several different categories relating to types (male, female, prior service, non-prior service) and quality (high school graduate, non-high school graduate, mental category I,II,IIIA,IIIB,IV,V). Historically, the largest problem in attaining these requirements has been in the enlistment of *male*, *high school graduate*, *non-prior service*, *mental category I-IIIA* (GSM I-IIIA) recruits. In this study, the problem of attempting to *predict* the number of these quality male recruits for future years is modelled. Ordinary least squares multiple linear regression analysis and stepwise regression analysis is utilized with an historical data base provided by USAREC.

A. PROBLEM STATEMENT

There are several objectives of this thesis. They vary in both scope and magnitude.

First and foremost is the near term need for the development of a predictive model to be used by the active duty Army 'green suit' analysts (hereafter refered to as Army analysts) stationed at USAREC headquarters and at the DCSPER, Department of the Army. At these agencies, major policy decisions are routinely contemplated. These decisions are usually concerned with aggregate responses to possible major personnel policy changes and/or budgetary realignments. There is a need for a quick response mechanism to answer various 'what if' questions concerning the quality of the force.

In this regard, it is desired to build a model that can be easily understood and quickly updated. Although a sufficient degree of complexity is an inherent desired feature of any proposed model, the true value of this particular model may be more in its ability to be maintained and updated, and its propensity for understanding by the (continuous) change of Army analysts that are stationed for a tour of duty at these agencies. The Army has initiated many studies in this field (usually through contracting) with various results. Where applicable, these studies will be referenced in the body of this thesis. There is an inherent problem, however, in the Army's ability to keep up with these efforts, either in the updating of the data base or in the level of understanding of the current, on-line Army analysts assigned to USAREC and DCSPER. It is thought by many that an in-house model, easily updated and universally understood, would be preferable to a more complex yet harder to comprehend effort. The need for simplicity for the analysts and understanding by the decision makers is a cornerstone on which this model will be derived.

It is not envisioned that this model will be a panacea to quality enlistment modeling. On the contrary, it will be promulgated as a 'first effort' on how to go about developing a model with the data base as given.

A concerted effort will be put forth on the whys and hows of going through the ordinary least squares and stepwise regression analysis used in developing this model. Most Army analysts have little knowledge or experience in the detailed theory of regression analysis. Their familiarity with the subject matter may be limited to graduate level studies (if at all) or to some contact with regression models in previous duty assignments. The community of experts in the manpower modeling field is small and few are in the active Army. The chapters of this thesis will cover the details of the model, some of the theory of its development and application, and possible sources of further study that needs to be accomplished. It is desired that an examination of this material, some of which will be heuristic in nature, will bridge this gap in knowledge. Hopefully, it will lead to a better understanding of the dynamics that affect the quality of the force and the accepted methods of modeling the interrelationships involved. Army analysts must be able to do more than just 'crunch the numbers' that they are given by other analysts. Forming a base for the understanding and refinement of this model is another major objective of this thesis.

B. BACKGROUND

In February, 1986, the Chief of Staff, USAREC, tasked the Programs, Analysis and Evaluation Directorate (PAE) to review the current list of enlistment supply models and to reevaluate and assess what factors (variables) were contributing significantly to explaining quality male enlistment contracts. This thesis is in partial fulfillment of that requirement. Although there have been many studies in this field, the need still exists for continuing development in order that the Programs Analysis and Evaluation Directorate may have an in-house model with current data and accessable to Army analysts. Other studies, such as the Enlistment Supply Model published by Daula and Smith, [Ref. 1] and the Recruiting Resources Allocation System by ABT Associates, Inc., [Ref. 2] are commendable. The problem is that they are neither readily accessable nor easily updated by USAREC or DCSPER personnel. Further, the level of understanding required is well beyond the expertise of the typical Army analyst. He must bear the burden of providing the day-to-day answers to various decision makers asking a plethora of questions on a litany of different issues. With his day-to-day plight in mind, the study objective of this thesis was conceived.

C. STUDY OBJECTIVE

The primary objective of this thesis is to develop a model using ordinary least squares multiple linear regression analysis and stepwise regression analysis to predict total Army male quality (GSM I-IIIA) contracts for future years. Special emphasis is placed on the explanation of the methods and techniques used to derive this model. All data elements must be readily obtainable and possess some potential for future prediction.

D. THE DATA

A longitudinal data base for this study was provided by PAE, USAREC (Table 1). The data is cross sectional in that it is broken down by recruiting battalions (1A,1B,...,6L) and time series in that it provides data for each of these battalions by year (1982,1983,1984,1985). Knowing the structure of the data has important implications as to the types of techniques that will be employed in the regression analysis. Of the 56 recruiting battalions of USAREC, data elements for 55 were made available (battalion 3L, San Juan, Puerto Rico was omitted). In all, the data base contained 19 variables. For a more detailed explanation of the data, to include variable descriptions, see Appendix B.

E. A REGRESSION REVIEW

If one accepts the premise that historical actualities can be used as a basis to predict future events, then regression analysis is a powerful tool that can provide much insight into the predicting phenomenon. The principle behind ordinary least squares is as follows.

TABLE 1
PARTIAL LIST OF DATA PROVIDED BY USAREC

BN	YEAR	CONT	RCTR	UNEMP	PROP	HSMMA	TOTPOP	WHIPOP
1A	1982	657	53.75	8.05	14.7	13931	2169022	2083422
1A	1983	805	52.25	7.93	15.1	13816	2180204	2093593
1A	1984	703	52.50	7.43	19.5	14153	2191385	2103765
1A	1985	724	50.00	6.20	15.7	13275	2202566	2113936
1B	1982	1585	155.00	8.60	13.4	37180	6109021	4448873
1B	1983	1977	165.00	7.53	13.4	36648	6157147	4481896
1B	1984	1733	161.00	6.15	15.3	29749	6205793	4515079
1B	1985	1611	150.50	5.55	17.4	28648	6254179	4548262
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
6L	1982	1217	103.25	11.08	8.5	26838	4103713	3743095
6L	1983	1278	103.00	11.18	8.4	26473	4153740	3785629
6L	1984	1060	106.00	9.95	8.8	25459	4203768	3828164
6L	1985	1396	97.00	8.63	8.5	24122	4253796	3870698

E1PAY	BLKPOP	HISPOP	INCOMPC	QMA	BNADV	PAYCO	ARMYMS	DOD-A
551.4	57346	25885	7610	555	720	6.99	0.33	1348
573.6	57840	26018	8715	555	633	10.49	0.37	1370
573.6	58334	26152	9056	753	555	3.91	0.39	1121
573.6	58828	26285	9396	753	738	3.76	0.41	1041
551.4	1483920	120148	10223	1338	2207	9.46	0.39	2509
573.6	1495431	121100	11888	1338	1837	12.26	0.42	2776
573.6	1506942	122051	12072	1922	1107	1.55	0.40	2647
573.6	1518453	123003	12256	1922	1357	1.52	0.41	2295
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
551.4	106922	119881	9416	960	1014	9.89	0.37	2066
573.6	108560	121342	10978	960	836	12.56	0.40	1879
573.6	110199	122804	11135	1281	1058	1.43	0.36	1866
573.6	111838	124265	11291	1281	1124	1.41	0.43	1869

Using some of the data for Contracts (CONT) and Propensity (PROP) from Table 1 above, draw a straight line through a cluster of the plotted data points on a scatter diagram (Figure 1.1). Then, for each point, find the vertical distance from the straight line, square this distance, and then add together all of the squared distances. Of all the straight lines that could be possibly drawn through the points on the graph, *the best-fitting line is the one with the smallest sum of the squared distances.* This line is called the regression line. The signed (positive or negative) distance from any point to the regression line is called the *residual.* It is the difference between the actual value of Contracts (1A ACTUAL) and the value of Contracts that the regression line predicts

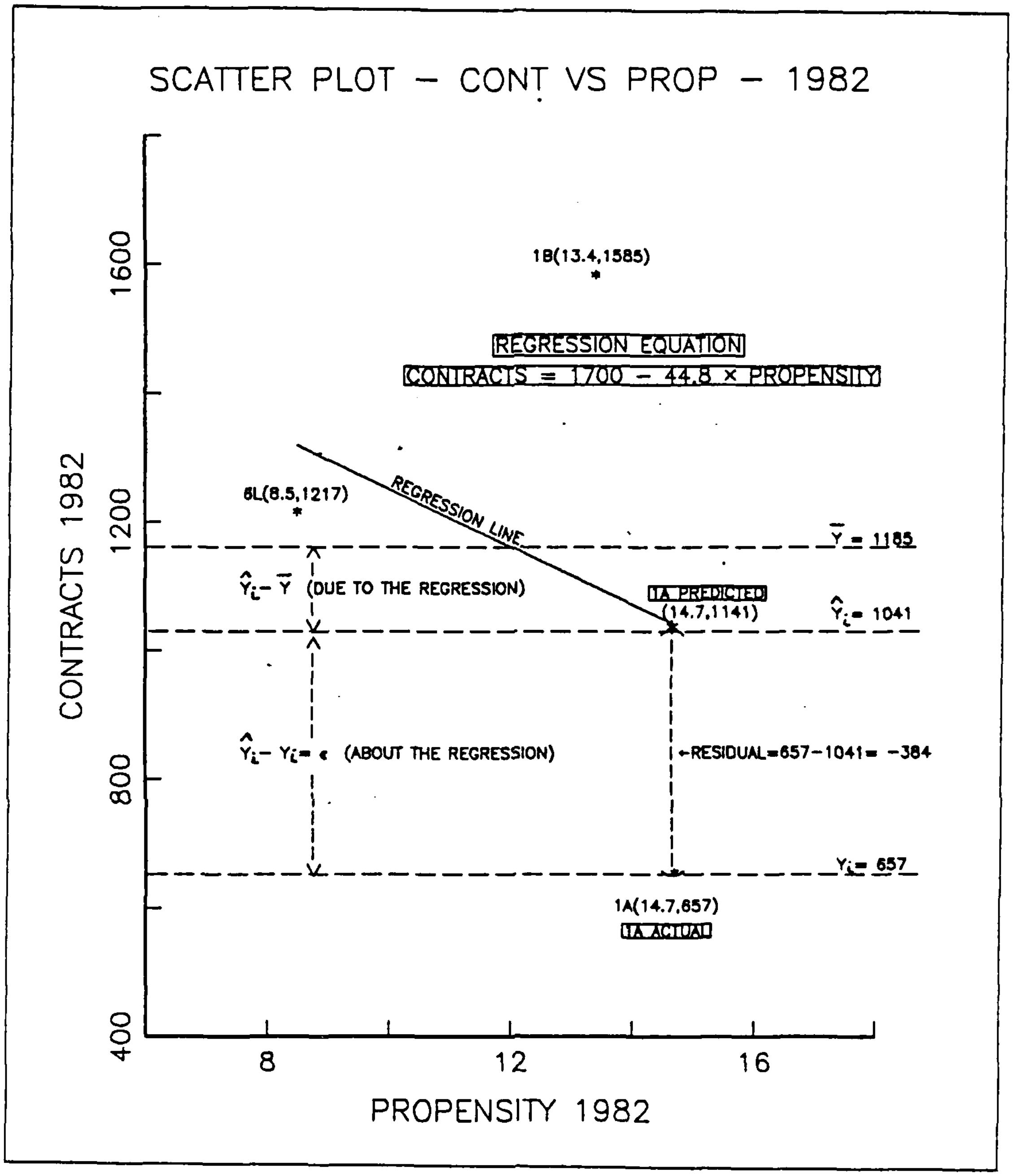


Figure 1.1 Graphical Representation of Ordinary Least Squares Regression

(1A PREDICTED). *The residuals represent the error in the model.* If there were no error in the model, and therefore, no residuals, then the regression line would pass through point 1A ACTUAL and the residual would equal zero. In Figure 1.1, the residual = -384 for BN 1A. The sum of all of the residuals squared is called the *sum of squares about the regression*, or $\sum (\hat{Y}_i - Y_i)^2$. [Ref. 3] Without the theory of regression, if asked to predict next year's contracts (or any year's contracts), one would choose the *mean or average* number of contracts as the best predictor. The mean is represented in Figure 1.1 as $\overline{Y} = 1185$. The square of the distance between the average value and the predicted value is called the *sum of squares due to regression*, or $\sum (\hat{Y}_i - \overline{Y})^2$. The mean is defined as the $\sum Y_i/n$, where n equals the number of data points. In this example, $\sum Y_i = 657 + 1585 + 1217$ and $n = 3$, so $\overline{Y} = 1185$. Another important term, called the *total sum of squares corrected for the mean* is equal to the addition of the sum of squares due to the regression plus the sum of squares about the regression. Algebraically, this is $\sum (Y_i - \overline{Y})^2 = \sum (\hat{Y}_i - \overline{Y})^2 + \sum (\hat{Y}_i - Y_i)^2$. It will be helpful to keep Figure 1.1 in mind as this thesis is read. Although the figure portrays a *simple* linear regression of two variables (CONT being the dependent variable on the vertical axis and PROP being the independent variable on the horizonal axis), it has direct translation to the theory of *multiple* linear regression. In multiple linear regression, the *objective is still to minimize the squares of the distance between the actual and the predicted values*, only now there are several (instead of two) dimensions. Graphical interpretations cannot be made above three dimensions. Above three dimensions, the regression line becomes a regression hyperplane in the hyperspace defined by the independent variables. The important thing to remember, however, is that all of the *mathematics* required to derive the regression line for simple regression are *still valid* for multiple regression. Therefore, the analysis of multiple regression will rely heavily on the interpretation of these mathematically derived values (or estimators). The mathematically derived estimators for the regression line in Figure 1.1 is called a *regression equation*. This regression equation is given in the form :

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

where the variables are:

Y = CONTRACTS = CONT = the dependent variable
X_1 = PROPENSITY = PROP = the independent variable

and the parameter estimators are:

β_0 = 1700 = the intercept with the dependent variable axis
β_1 = - 44.8 = the slope of the regression line

and the model error is represented by:

ε = residual with assumed distribution $N(0,\sigma^2)$
(residuals are also assumed to be independently distributed)

In looking at this particular equation, it seems counterintuitive that one would predict that, as the propensity for service goes up, the actual number of contracts goes down. This is because of the negative slope of the regression line which can be determined mathematically by the *negatively signed parameter estimator for the slope.* The signs of parameter estimates are important. The analyst must be cognizant of these anomalies and be prepared to think through the interpretation of his mathematical results. Hopefully, this thesis will explain this phenomenon. This study will outline many key estimators, how they are derived and their various uses. It is imperative, however, to understand Figure 1.1 before moving on into the body of this thesis.

F. INITIAL ASSUMPTIONS

There are several assumptions which should be explicitly stated. First of all, it is assumed that the data provided is accurate. This is imperative to the mechanics of model building and the analysis of the data.

More importantly, however, is the assumption that the personal and environmental statistical data upon which model is based have some effect on an individual's decision as to whether or not to enlist. Implicit in this assumption is that persons living in different areas of the country with different environments will behave differently. Also implicit is that different persons facing similar environments will behave in a similar manner. These assumptions, and the assumption that this behavior stays relatively stable across time, are fundamental to the cross sectional and time series regression analysis that will be required.

Finally, since a *linear* regression model is being built, it is necessary to assume that trends will continue exactly as they have in the past. Over the near term, this is a reasonable assumption. Over the long term, it is not. This implies that the predictions from the model will be more accurate for the next one or two time periods than for more future time periods. This is because real events rarely behave in a linear manner over long periods of time.

G. THESIS OUTLINE

This thesis develops and explains a model for the prediction of future GSM I-IIIA contracts. It is developed to predict the total contracts for a 'typical' Army recruiting battalion. In Chapter II, an outline is presented on how the regression model in this thesis will be built. Chapter III details some of the preliminary analysis and planning that led to the model formulation. Chapter IV continues through the development process and outlines many helpful statistical tools for data and regression analysis. Chapter V presents the model in detail and the results of the fitting of the model to the finalized data base. The last chapter, Chapter VI, lists the conclusions and recommendations of this study. Several Appendixes are included to enhance understanding and are referenced throughout the body of the thesis. A List of Appendixes is provided on page 6. Appendix A may be of particular interest. It is a select glossary of terms used in this study. If a certain term is unfamiliar, this is the first place one should look.

H. PROGRAMMING LANGUAGES AND STATISTICAL PACKAGES.

The programing languages used in the completion of this project are FORTRAN 77 (the 1977 update of the Formula Translation language) and APL (A Programing Language). The statistical packages used were GRAFSTAT (IBM Corporation) and the SAS-Statistical Analysis System Version V (SAS Institute Incorporated). With the realization that not all of these computational assets are readily available to most Army analysts, virtually all analysis and most of the required graphics that are presented can be accomplished using the SAS statistical package. This is in accordance with the current capabilities of both DCSPER and USAREC. Some GRAFSTAT graphics (such as Figure 1.1) will be presented only for the purpose of enhancing visual understanding.

II. BUILDING REGRESSION MODELS

Linear regression analysis is applicable to a vast array of subject matter. Linear regression models are built so that researchers can test the validity or falsity of hypothesized functional relationships. The purpose of the model that will be built in this thesis is to try to extract the main features of the relationships that are hidden or implied in the tabulated data in Table 1 on page 12.

Before one starts building a model, it is useful to have an outline of how to go about the process. This chapter provides the basic structure that will be followed in Chapters III, IV and V.

There are three distinct phases of building regression models. They are the Planning Phase, the Development Phase and the Verification and Maintenance Phase. [Ref. 3:p. 414]

Building a regression model is a time consuming task. It is made even more time consuming by the requirement to fully explain and document assumptions, methods, and results. Documentation is essential because one must be very careful in the use of multivariable regression analysis. Results from predictive models can be easily misinterpreted or misused. The analyst is wise to state his assumptions and desired goals of the model in order to minimize the potential for misunderstanding. The figures of this chapter provide flowcharts that can be followed when faced with building a regression model. Although these flowcharts are generic in nature, they detail the special problems encountered when dealing with cross sectional and time series data.

The regression review and Figure 1.1 in Chapter 1 discuss a simple regression approach. This thesis, however, will be describing some methods for building *multivariate* regression models. When analyizing multivariate models, the analyst must rely on many statistical indicators. Although these indicators will be mentioned in this chapter, *a more detailed explanation will be provided in Chapters III,IV and V.*

A. THE PLANNING PHASE

As can be seen in Figure 2.1, the first and foremost task in model building is to define the problem. Sometimes this is the most difficult step. What is the analyst *really* trying to accomplish? The problem statement must be specific, understandable and to the point.

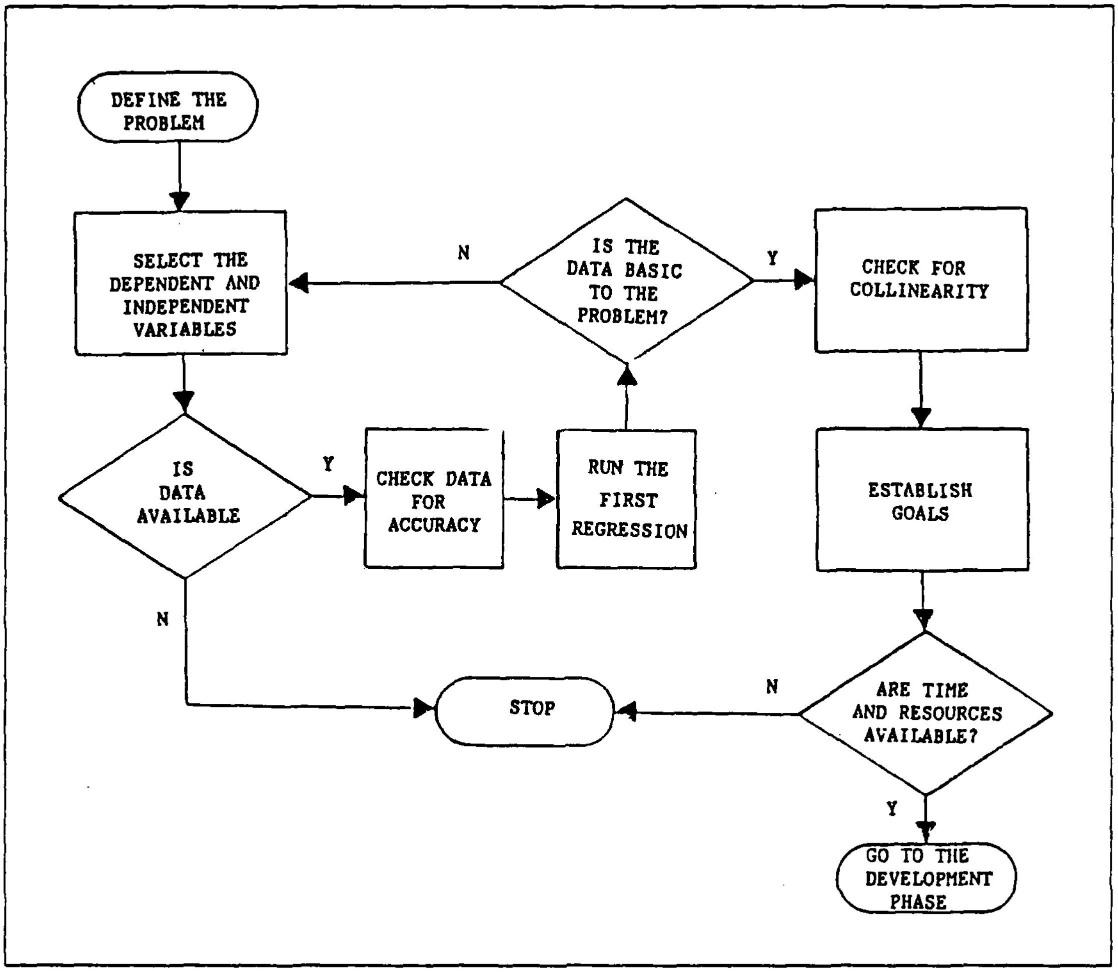

DEFINE THE PROBLEM
SELECT THE DEPENDENT AND INDEPENDENT VARIABLES
N
IS THE DATA BASIC TO THE PROBLEM?
Y
CHECK FOR COLLINEARITY
IS DATA AVAILABLE
Y
CHECK DATA FOR ACCURACY
RUN THE FIRST REGRESSION
ESTABLISH GOALS
N
STOP
N
ARE TIME AND RESOURCES AVAILABLE?
Y
GO TO THE DEVELOPMENT PHASE

Figure 2.1 The Planning Phase of Model Building

Next comes the data selection. Both the carrier (independent) and the response (dependent) variables must be clearly identifiable, readily available and as complete as possible. One should 'brainstorm' to try to think of any variable which might be relevant to the problem.

One of the first tasks is to check the data for validity. Histograms and scatter plots are excellent tools for this. Look at the data distribution. Pay close attention to the outliers. Ask if there are valid explanations as to why some of the data looks as if it does not belong. If necessary, consult the experts for advice. Also pay particular attention to the range of the data. Data that varies little will sometimes provide artificially high or artificially low values for the degree to which the model fits the data.

The regression hyperplane must fit through the hyperspace that is defined by the carrier variables. Small *relative* ranges tend to shrink this hyperspace and obtaining good predictions will become difficult.

Once the data has been verified, run the first regression. At first, it is only necessary to look at a few basic indicators. The analyst must be familiar with the information that the ANOVA table is providing. Stepwise regression is a powerful and widely accepted tool that can be extremely helpful when looking for significant variables that are basic to the problem. Stepwise regression is more fully explained in Appendix A. The analyst needs to become familiar with the ideas behind the correlation matrix and what it is indicating about multicollinearity. Multicollinearity arises whenever two or more independent variables used in the regression are not independent but are correlated. Among other things, the presence of multicollinearity will lead to larger standard errors in the model. Also it is helpful to understand the Variance Inflation Factor statistic and the Condition Index in the Variance Proportion Matrix. All of these indicators and procedures will be discussed in Chapter III. The first regression should provide a very rough indication of what kind of fits are going to be possible.

Finally, before leaving the Planning Stage, it needs to be determined whether there will be time and resources available to complete the task *correctly*. 'Half efforts' will lead to incorrect results and a lack of confidence in both the analyst and the regression procedures. The bottom line is that if time and resourses are not available, then stop. Again, Chapter III provides a 'walk through' of the procedures that are detailed in this section.

B. THE DEVELOPMENT PHASE

This section provides a brief outline of the development phase of model building. Chapter IV will discuss in detail the concepts and statistical indicators that are outlined in this section.

The first regression from the Planning Phase tells the analyst quite a bit about the behavior of the data in the model. Once the decision has been made to go ahead with the modelling effort, one moves to the Development Phase of model building. Many different approaches to the regression problem can occur during this phase.

The analyst may feel uneasy about some facet of the initial regression findings. The Development Phase is time consuming in that trial and error is the normal method

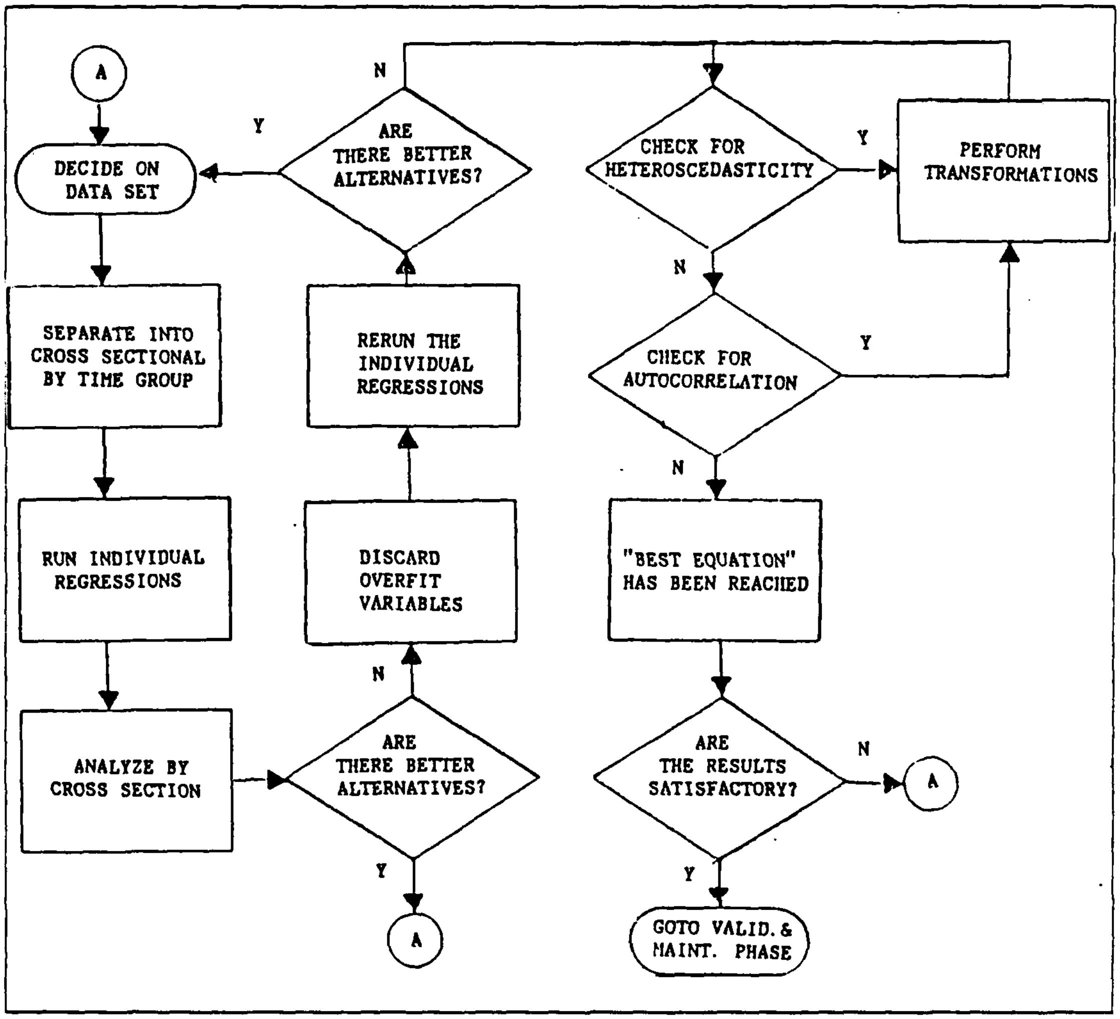

A
DECIDE ON DATA SET
SEPARATE INTO CROSS SECTIONAL BY TIME GROUP
RUN INDIVIDUAL REGRESSIONS
ANALYZE BY CROSS SECTION
ARE THERE BETTER ALTERNATIVES?
Y
A
N
DISCARD OVERFIT VARIABLES
RERUN THE INDIVIDUAL REGRESSIONS
ARE THERE BETTER ALTERNATIVES?
Y
N
CHECK FOR HETEROSCEDASTICITY
Y
PERFORM TRANSFORMATIONS
N
CHECK FOR AUTOCORRELATION
Y
N
"BEST EQUATION" HAS BEEN REACHED
ARE THE RESULTS SATISFACTORY?
N
A
Y
GOTO VALID. & MAINT. PHASE

Figure 2.2 The Development Phase of Model Building

of testing various ideas. Many times, ideas evolve from the results of previous experiments. This is the hallmark of the scientific process. Figure 2.2 outlines the Development Phase of regression model building.

Sometimes new variables are derived from raw data. This is usually because the analyst has some idea that it makes sense to do so, or because the original regressions are not behaving in an intuitive manner. This is similar to what happened in Figure 1.1 on page 13, where an increase in PROPENSITY resulted in a decrease in CONTRACTS. In the model that will be developed in this thesis, three out of the five variables that are finally utilized were derived from raw data.

Once the analyst is satisfied with the data, it must be separated into cross-sectional groupings (all battalions) by time period (year). For the data in Table 1 on page 12, this implies that it is separated into four distinct groups; all battalion data for 1982, all battalion data for 1983 and so on. *The purpose of this procedure is to check for heteroscedasticity without having the mathematical results biased by autocorrelation.* Heteroscedasticity is a condition where the error terms (ε) are not constant for all values of the independent variables. Autocorrelation is a condition where the error terms from different observations are correlated. Both of these conditions will affect the size of the standard error of the regression coefficient and therefore bias the results of the regression model.

Now each grouped (cross-sectional) data set is run through the regression procedures. The correlation matrix will indicate highly correlated carrier variables and the stepwise procedure will show which are the most significant in explaining the fit of the regression line. It is now time to drop those variables that are insignificant or are contributing the most to multicollinearity. Again, new variables may become apparent at any time. They should be included and scrutinized by the analyst until all practical possibilities have been exhausted.

Rerun the regression for all of the finalized groups of data. Look at the results and compare between time periods. Are the parameter estimates comparatively stable? Are they signed the same? Are the same variables significant in each time period? Are they comparable in magnitude? If the groups are different, are they *significantly* different? Most of the answers to these questions are judgment calls on the part of the analyst. Whatever the call, he should be able to justify his decision based upon the knowledge of the problem and the underlying data base. Next, plot the residuals versus the predicted values and look for any signs of heteroscedasticity. If heteroscedasticity is present, the results of the regression cannot be considered valid. *Unless the analyst has some valid reason to do otherwise, this should be the first time that he considers transforming the data.* Transformations inherently lead to a lack of understanding in the modeling process and should be avoided up until the point at which the benefit to the model derived by the transformation exceeds the detriment to the user in the understanding of the model. If heteroscedasticity is significant, then apply the appropriate variance stabilizing transformation to the groups of data. [Ref. 3:p. 238] If heteroscedasticity is not a problem, or if the transformation renders the problem insignificant, it is time to re-pool the data back to its original longitudinal structure.

The data set would look exactly like Table 1 again, except that now the analyst will be working only with those variables that were found to be significant in the cross-sectional analysis.

Run the regression on the entire pooled data set. Plot the residuals and check for *autocorrelation.* If autocorrelation is present, the results of the regression are biased and the standard error of the estimates is inaccurate. Accept or fail to accept the hypothesis on autocorrelation in the residuals using a runs test or the popular Durbin-Watson test. If autocorrelation seems to be a problem, then the true correlation coefficient of the data structure needs to be determined and another transformation on the data needs to be performed. Rerun the regression using the transformed data and then double check to ensure that the effects of autocorrelation are no longer present. The 'best regression equation' has now been determined.

Finally, check to see that the model is fulfilling the goals as set forth in the Planning Stage. If not, it may be time to start anew, possibly with new variables. Or, it may be time to re-access the goals of the model. Whatever the case, once the analyst has decided that the 'best equation' has been achieved, it is time to move to the model Validation and Maintenance Phase. Chapter IV details a step-by-step method for the development of the GSM I-IIIA model that is being built in this thesis.

C. VALIDATION AND MAINTENANCE PHASE

If the analyst feels comfortable about the achievement of the goals and the stability of the model after the Development Stage, then he has gone a long way towards the validation of the model. Figure 2.3 provides a step-by-step summary of this phase of model building. Chapter V details this phase as it applies to the regression model that is being built in this thesis. The concepts that are outlined in this section are more fully explained in Chapter V.

One last check needs to be performed to see if there is any systematic lack of fit in the model. Remember that the residuals contain all of the information on the lack of fit in the model and they should be checked for any possible pattern.

Next, validate the model. Validation merely implies checking to see if the model makes sense. Check the model by trying a few predictor variables and see if the response variable makes sense. For instance, try some data points near an extreme of the prediction space to see if the response is coherent with that extreme. There are many methods of validation and there is really no 'best method'. [Ref. 3:p. 420] As it is with variable selection, it is up to the judgment of the analyst.

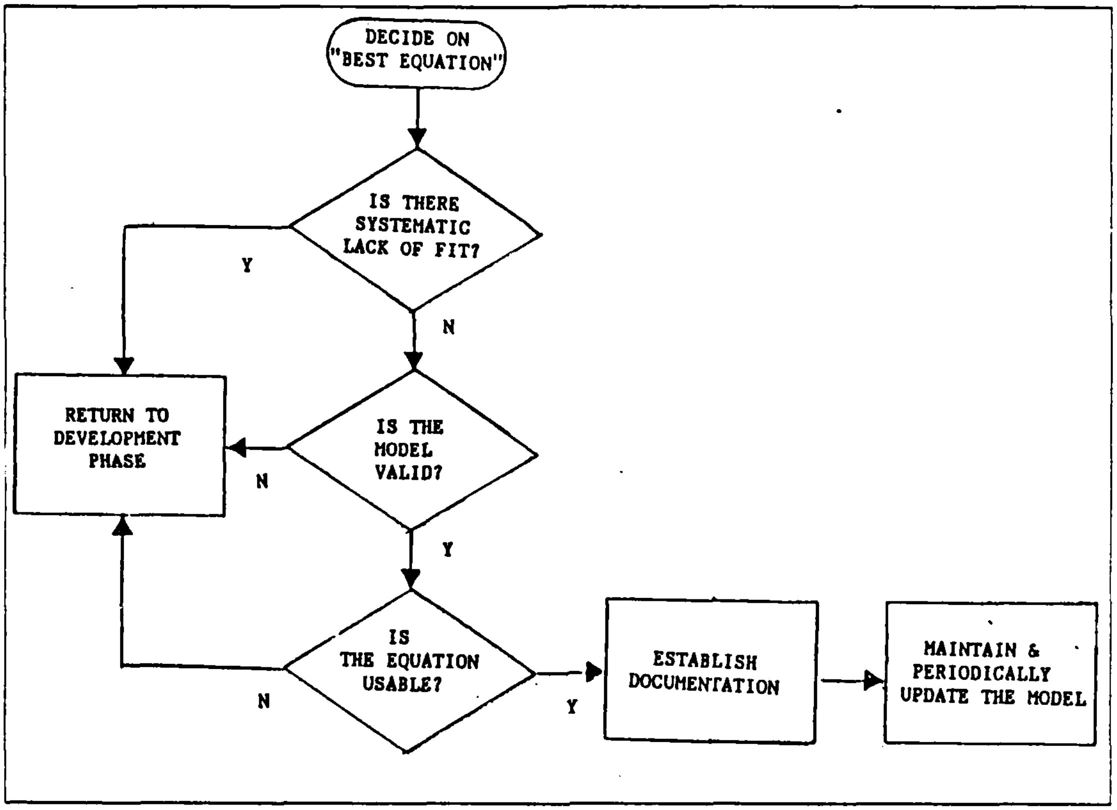

DECIDE ON "BEST EQUATION"
IS THERE SYSTEMATIC LACK OF FIT?
Y
N
RETURN TO DEVELOPMENT PHASE
IS THE MODEL VALID?
N
Y
IS THE EQUATION USABLE?
N
Y
ESTABLISH DOCUMENTATION
MAINTAIN & PERIODICALLY UPDATE THE MODEL

Figure 2.3 The Validation and Maintenance Phase of Model Building

Is this equation useful and are these parameters reasonable? This is the final validation test of the model. Does it pass the scrutiny of the experts? The final product should achieve the desired objectives as outlined in the initial problem statement. Obviously, the intermediate goals were either achieved or revised in order to get to this final stage. The only thing left to do is to establish the proper documentation for the model, this should include all assumptions and the ranges of the inputs for which the model is valid.

Finally, the model needs to be maintained, updated and periodically re-evaluated for accuracy and validity. This can be especially difficult for complex models that are to be maintained by Army analysts in a high turnover environment. One to the goals of this model has been to attempt to keep this maintenance procedure as simple as possible. It is now time to move on to Chapters III, IV, and V to see how well this goal was accomplished.

III. PLANNING THE GSM I-IIIA MODEL

This chapter explains the specifics of planning the GSM I-IIIA model. Much reference will be made to Figure 2.1 of Chapter 2 which provides an outline of the Planning Phase. It may be useful to review Figure 2.1 at this time.

A. DEFINING THE PROBLEM

The problem definition stems directly from the study objective. This thesis will detail a step-by-step procedure which can be used to build a predictive model for future year GSM I-IIIA contracts. The data for this model must be easily updated and readily available. The data should also have some potential for future prediction. This model will be developed to predict the results of a 'typical' Army recruiting battalion and is not designed for predicting any *specific* battalion results. Since one of the major objectives of the thesis is to provide a 'walk through' for the reader on the hows and whys of model building, the author has chosen the first person plural as the pronoun of choice. *We* will now attempt to build this model.

B. SELECTION OF THE INDEPENDENT AND DEPENDENT VARIABLES.

Data for this project was provided by the Programs Analysis and Evaluation (PAE) section of USAREC. It is as appears in Table 1 on page 12 and as described in Appendix B. Since this model is now in the Planning Phase we should be 'brainstorming' in order to try to think of any variable which might be relevant to the problem. We are trying to predict total contracts, and the variable CONT from Table 1 seems to be the logical and ideal choice for the dependent variable. Also, we figure that other variables, both endogenous and exogenous, may play some role in determining the number of contracts signed. Many variables, such as the Consumer Price Index (CPI), are contemplated. These variables, mostly of the exogenous variety, might be useful in capturing some of the social or demographic dynamics of the enlistment process. The problem is, however, that these statistics are not available at the cross-sectional (battalion) level and time specific (by year) period that would fit with the rest of the data structure.

C. CHECKING THE DATA

The final list of variables from the Planning Stage are as presented in Table 1. The only exception is with the battalion term, BN. Being alpha-numeric in nature, it can not be plotted in the multivariate hyperspace in order to determine a least squares fit. The analyst can substitute a numerical counterpart if he desires to use the battalion as a carrier variable. Therefore, the battalions are numbered from 1 to 55 instead of from 1A to 6L. This variable will be more thoroughly discussed as the model is developed. Table 1 is complete in that there are no missing data entries for any battalion during any year. Appendix B provides a detailed explanation of the data that will be used in this thesis. After checking the data using histograms and scatter plots and carefully verifying the outliers, the Planning Stage finalized matrix of longitudinal data appears below.

$$Y = \begin{array}{c} \text{CONT} \\ \left[\begin{array}{c} 657 \\ 805 \\ : \\ : \\ 1396 \end{array}\right] \end{array} \qquad X = \begin{array}{c} \begin{array}{ccccccc} & \text{BN} & \text{YEAR} & \text{RCTR} & \text{UNEMP} & & \text{DOD-A} \end{array} \\ \left[\begin{array}{ccccccc} 1 & 1 & 1982 & 53.75 & 8.05 & \ldots\ldots & 1348 \\ 1 & 1 & 1983 & 52.25 & 7.93 & \ldots\ldots & 1370 \\ : & : & : & : & : & : & : \\ : & : & : & : & : & : & : \\ 1 & 55 & 1985 & 97.00 & 8.63 & \ldots\ldots & 1869 \end{array}\right] \end{array} \qquad \beta = \left[\begin{array}{c} b_0 \\ b_1 \\ : \\ : \\ b_{17} \end{array}\right]$$

where Y = 220x 1 matrix (a column vector of the dependent variables)
X = 220x18 matrix (a column vector of 1's catonated with the 220x17 matrix of the independent variables)
β = 18x 1 matrix (a column vector of parameter estimates)

Notice that this is the initial matrix format required for the Normal Equations for Multiple Linear Regression (see definition in Appendix A). The column vector of 1's in the X matrix is required for the matrix multiplication of the b_i values in the β matrix.

D. THE FIRST REGRESSION

As stated in the introduction, SAS will be utilized as the statistical package for all of the analysis in this thesis.

Appendix C shows the basic format for the SAS input. Not every procedure was required for every step of the model development process. With few exceptions, Appendix C lists all of the steps that were used throughout Chapter III and some of Chapter IV. At each step in the Planning and Development Stage, this thesis will specify the procedure that is important to that particular step and provide a table of the output and diagnostics from SAS that are pertinent to that step.

Running the first regression with the data as in Table 1 (except 1 replaces 1A, 2 replaces 1B, etc), several outputted indicators are obtained.

E. DETERMINING IF THE DATA IS BASIC

Table 2 is the printout of the ANOVA table. The MODEL statement in SAS automatically provides this output. [Ref. 4] Reference is made to Figure 1.1 on page 13 for a graphical interpretation and to Appendix A for the algebraic interpretation of the values in the ANOVA table.

TABLE 2

ANALYSIS OF VARIANCE TABLE FROM SAS

```
DEP VARIABLE: CONT
                          ANALYSIS OF VARIANCE

                         SUM OF             MEAN
SOURCE     DF           SQUARES            SQUARE       F VALUE       PROB>F

MODEL      17          23653906           1391406       382.419       0.0001
ERROR     202            734963          3638.429
C TOTAL   219          24388868

ROOT MSE               60.31939          R-SQUARE        0.9699
DEP MEAN               1007.241          ADJ R-SQ        0.9673
C.V.                   5.988577
```

For illustrative purposes, the values in the ANOVA table in Table 2 are derived below. A few important facts to remember is that the MS ERROR is the best (unbiased) estimate of the variance of the residuals and, therefore, the ROOT MSE is the best (biased) estimate of the standard deviation of the residuals.

MODEL df = number of independent variables = 17

ERROR df = number of data lines - MODEL df - 1 = 220 - 17 - 1 = 202

CORRECTED TOTAL df = MODEL df + ERROR df = 17 + 202 = 219

SS MODEL = sum of squares due to regression = 23653906

SS ERROR = sum of squares about the regression = 734963

SS CORRECTED TOTAL = SS MODEL + SS ERROR = 24388868

MS MODEL = SS MODEL / MODEL df = 23653906 / 17 = 1391406

MS ERROR = SS ERROR / ERROR df = 734963 /202 = 3638.429 = σ^2

F VALUE = MS MODEL / MS ERROR = 1391406 / 3638.429 = 382.419

PROB > F = F distribution with 17 and 202 degrees of freedom = 0.0001

ROOT MSE = square root of MS ERROR = 60.31939 = σ

DEP MEAN = the average of the 220 values of CONT = 1007.241 = Y

COEFFICIENT OF VARIATION = (ROOT MSE / DEP MEAN) x 100 = 5.988577 = C.V.

RSQUARE = SS MODEL / SS CORRECTED TOTAL = 0.9699 = R^2

ADJ RSQ = 1 - (1-RSQUARE) x (n-1 / n - MODEL df + 1)

= 1 - (1-.9699) x (219 / 220 - 17 + 1)

= 1 - (.0301) x (1.0735) = 0.9673 = R_a^2

At this point in the planning stage, we are merely trying to determine if we have variables that are basic to the regression. To determine this, we look at the F VALUE and PROB > F statistics. If we did not have a regression, then we would not have a slope. As seen in equation 1.1 on page 13, the slope is equal to our β_i values (for i not = 0). By doing an F test (with 17 and 202 degrees of freedom), we postulate a null hypothesis that the β values all equal 0. A high F value tends to reject this null hypothesis, indicating that the β values do *not* equal 0. The PROB > F is the *actual level of significance*, α (actual), at which we reject this null hypothesis. What we are saying in this ANOVA table is that there is less than a .0001 probability of rejecting a true null hypothesis ($H_0 : \beta = 0$). In other words, there is statistically less than 1 chance in 10,000 that there is no slope and all of the β values equal 0.

We will use α (critical) = .1 as the critical level of significance when checking variable significance in this thesis. Since α (actual) = .0001 < α (critical) = .1, we continue with this data base knowing that there are some variables that are basic to the regression.

To determine which variables are basic to this particular regression, one would look at the matrix for parameter estimates in Table 3. It, like the ANOVA table, is printed automatically when the MODEL statement is requested in SAS. Looking down the column of PROB > |T|, we find nine variables that meet our criteria of α (actual) < α (critical). They are BN, RCTR, TOTPOP, WHIPOP, BLKPOP, HISPOP, QMA, ARMYMS and DODMA. This is an indication that these are the significant variables that are explaining this particular regression *when all of the variables are included at the same time.*

TABLE 3
PARAMETER ESTIMATES WITH VARIANCE INFLATION FACTORS

VARIABLE	DF	PARAMETER ESTIMATE	STANDARD ERROR	T FOR HO: PARAMTR=0	PROB >|T|	VARIANCE INFLATION
INTER	1	11842.283	21933.657	0.540	0.5899	0.00
YEAR	1	-7.058418	11.261912	-0.627	0.5315	9.58
BN	1	0.854557	0.341596	2.502	0.0132	1.77
RCTR	1	1.717983	0.516035	3.329	0.0010	11.17
UNEMP	1	-0.949183	2.454547	-0.387	0.6994	1.79
PROP	1	-0.559418	1.674842	-0.334	0.7387	3.44
HSMMA	1	0.0005085386	0.001407693	0.361	0.7183	8.52
PAYCO	1	-0.598691	2.458493	-0.244	0.8079	4.09
TOTPOP	1	-0.000156535	.00003336254	-4.692	0.0001	101.47
WHIPOP	1	0.0001638404	0.0000336013	4.876	0.0001	61.98
BLKPOP	1	0.0001727377	.00004012063	4.305	0.0001	16.81
HISPOP	1	.00007096273	.00002216656	3.201	0.0016	5.68
INCOMPC	1	0.002052645	0.005432364	0.378	0.7059	3.41
QMA	1	-0.053719	0.029942	-1.794	0.0743	8.67
BNADV	1	0.015437	0.015610	0.989	0.3239	1.92
E1PAY	1	1.346967	0.967061	1.393	0.1652	5.22
ARMYMS	1	3633.248	130.275	27.889	0.0001	1.86
DODMA	1	0.523631	0.015331	34.155	0.0001	4.57

Finally, we look at the result of the stepwise regression in Table 4. This comes from the PROC STEPWISE statement in Appendix C. SAS will print a complete ANOVA table as each variable is entered. Table 4 is the summary of relevant statistics from each of these ANOVA tables, which SAS also provides. The analyst has chosen to use the Stepwise Procedure, as opposed to the Forward Stepwise Procedure or the Backward Stepwise Elimination Procedure. A summary of these procedures can be found in Appendix A. The Stewise Procedure indicates that there are four variables that are significant at the α (critical) = .1 level *when only one variable is brought in at a time.* They are DODMA, ARMYMS, RCTR and QMA. All other variables fail to meet the .1 level of significance.

We conclude this section of the model planning with the knowledge that there exists data that is basic to the problem. The key indicators in Tables 2, 3 and 4 have provided the 'green light' to go ahead.

F. CHECKING FOR MULTICOLLINEARITY

The reason that we check for multicollinearity is because if there is a linear combination between the dependent variables in the X matrix (page 25), then our estimators will be unstable with high standard errors and we will probably calculate an artificially high R^2 value. The R^2 statistic is an indicator of how well the model fits

TABLE 4
SUMMARY OF STEPWISE OUTPUT FROM SAS

```
             STEPWISE REGRESSION PROCEDURE FOR DEPENDENT VARIABLE CONT

                  VARIABLE  NUMBER PARTIAL   MODEL
STEP  ENTERED      REMOVED    IN    R**2     R**2        C(P)     F VALUE    PROB>F

  1   DODMA                    1   0.7516   0.7516   1449.37      659.45     .0001
  2   ARMYMS                   2   0.2087   0.9602     52.79     1137.73     .0001
  3   RCTR                     3   0.0038   0.9640     29.24       22.84     .0001
  4   QMA                      4   0.0010   0.9650     24.45        6.23     .0133
  5   E1PAY                    5   0.0004   0.9654     23.80        2.45     .1188
  6   INCOMPC                  6   0.0004   0.9658     23.34        2.28     .1321
  7   BNADV                    7   0.0002   0.9660     24.18        1.07     .3010
  8   BN                       8   0.0001   0.9661     25.30        0.82     .3671
  9   HISPOP                   9   0.0001   0.9662     26.36        0.87     .3521
 10   WHIPOP                  10   0.0002   0.9664     26.92        1.34     .2480
 11   TOTPOP                  11   0.0001   0.9666     27.97        0.88     .3485
 12   BLKPOP                  12   0.0031   0.9697      9.04       21.34     .0001
 13   YEAR                    13   0.0000   0.9697     10.49        0.56     .4566
 14   HSMMA                   14   0.0001   0.9698     12.28        0.21     .6450
 15   UNEMP                   15   0.0000   0.9698     14.16        0.12     .7246
 16   PROP                    16   0.0000   0.9698     16.05        0.10     .7464
 17   PAYCO                   17   0.0000   0.9698     18.00        0.06     .8079
```

the data. An artifically high R^2 value is undesirable. A good example of multicollinearity (also known as collinearity) would be if the data base contained the measures of PERCENT MALES and PERCENT FEMALES per battalion. Clearly, these variables are not independent and if both were included in the regression model, the model would suffer from collinearity problems.

One indicator of multicollinearity is the Variance Inflation Factor (VIF) statistic, which is printed in the parameter estimates matrix. A SAS request of VIF in the MODEL statement provides this data in the Parameter Estimate Matrix (see Table 3). What is important to know about the VIF is that *big is bad.* Numbers of around 10 and over indicate multicollinearity. [Ref. 3:p. 416] Notice in Table 3 that there are several Variance Inflation Factors near or over 10.

Table 5 shows a partial output that is derived from SAS using the COLLIN procedure in the MODEL statement of SAS (Appendix C). Another key indicator is the Condition Index. Its derivation is somewhat involved. [Ref. 4:p. 55] As with the VIF, a big condition number is not a good sign. A condition index of 50 or more implies multicollinearity is a problem and the model suffers from multicollinearity. In this instance, there is an indication that at least five independent variables appear to be collinear.

TABLE 5

PARTIAL MATRIX OF COLLINEARITY DIAGNOSTICS FROM SAS

COLLINEARITY DIAGNOSTICS

NUMBER	EIGENVALUE	CONDITION INDEX	VAR PROP INTERCEP	VAR PROP YEAR	VAR PORP BN
1	15.811	1.000	0.0000	0.0000	0.0004
2	0.734740	4.639	0.0000	0.0000	0.0005
3	0.498840	5.630	0.0000	0.0000	0.0545
4	0.322976	6.997	0.0000	0.0000	0.0070
5	0.234600	8.209	0.0000	0.0000	0.0716
6	0.164497	9.804	0.0000	0.0000	0.4488
7	0.076709	14.357	0.0000	0.0000	0.0051
8	0.048103	18.130	0.0000	0.0000	0.0021
9	0.037163	20.626	0.0000	0.0000	0.1961
10	0.020718	27.625	0.0000	0.0000	0.0404
11	0.017691	29.895	0.0000	0.0000	0.0250
12	0.014183	33.387	0.0000	0.0000	0.0035
13	0.007865	44.835	0.0000	0.0000	0.0304
14	0.006012	51.280	0.0000	0.0000	0.0244
15	0.004832	57.201	0.0000	0.0000	0.0268
16	.000486733	180.230	0.0000	0.0000	0.0617
17	0.00007761	451.351	0.0001	0.0001	0.0001
18	1.687E-08	30611	0.9999	0.9999	0.0015

Table 6 is a printout of the correlation of estimates matrix. It is obtained from SAS by requesting CORRB in the MODEL statement. Its derivation is simply the $X'X^{-1}$ matrix scaled to unit diagonals. If you want to know which dependent variables are most highly correlated to each other, this is the place to look. Inspection shows that all of the population variables are highly correlated. This agrees with the VIF for TOTPOP, WHIPOP and BLKPOP, which also indicated a problem with these variables. The VIF also indicated a problem with RCTR and possibly YEAR, HSMMA and QMA. Checking Table 6 for these variables indicate that RCTR is most highly correlated with HSMMA (-0.4926); YEAR with PAYCO and E1PAY (0.5236 and -0.7072); HSMMA with RCTR (-.4926); and QMA with PAYCO (0.4771). An arbitrary level of $\rho > |0.4|$ was established by the analyst as an indicator of significant correlation. It is at this time that one needs to remember that the correlation coefficient shows only the extent to which two variables are linearly associated. It does not necessarily imply that there is any causal relationship between the two variables. Trying to figure out an explanation for the correlation between QMA and PAYCO could be difficult unless one was intimately familiar with the data gathering process and the demographics of these two variables. Even then, there may be no logical reason for the correlation. The only thing that is needed to know is that these two

TABLE 6
CORRELATION OF PARAMETER ESTIMATES FROM SAS

CORRB	INTER	YEAR	BN	RCTR	UNEMP	E1PAY
INTER	1.0000	-0.9999	0.0390	-0.2103	-0.1165	0.6963
YEAR	-0.9999	1.0000	-0.0390	0.2098	0.1156	-0.7072
BN	0.0390	-0.0390	1.0000	-0.0624	-0.3301	0.0422
RCTR	-0.2103	0.2098	-0.0624	1.0000	-0.0697	-0.0964
UNEMP	-0.1165	0.1156	-0.3301	-0.0697	1.0000	-0.0911
PROP	0.2255	-0.2227	0.0699	-0.0480	0.1364	-0.0292
HSMMA	0.0258	-0.0296	-0.0669	-0.4926	0.0798	0.1617
PAYCO	-0.5188	0.5236	0.0001	0.0443	-0.0225	-0.6104
TOTPOP	-0.1120	0.1104	-0.2427	0.1679	0.0794	0.0093
WHIPOP	0.1104	-0.1095	0.2592	-0.1646	-0.0508	0.0196
BLKPOP	0.0347	-0.0356	0.2948	-0.2737	-0.1356	0.0584
HISPOP	0.0894	-0.0887	0.0530	-0.1935	-0.0142	0.0107
INCOMPC	0.2461	-0.2387	0.0112	-0.2313	0.1816	-0.2366
QMA	0.2470	-0.2400	-0.0279	-0.3261	0.0490	-0.1627
BNADV	0.2829	-0.2857	-0.0678	-0.3092	0.1162	0.3273
E1PAY	0.6963	-0.7072	0.0422	-0.0964	-0.0911	1.0000
ARMYMS	0.1985	-0.1962	-0.2304	-0.1904	-0.1046	-0.0264
DODMA	-0.0461	0.0478	0.1412	-0.3699	-0.2819	-0.1164

CORRB	PROP	HSMMA	PAYCO	TOTPOP	WHIPOP	ARMYMS
INTER	0.2255	0.0258	-0.5188	-0.1120	0.1104	0.1985
YEAR	-0.2227	-0.0296	0.5236	0.1104	-0.1095	-0.1962
BN	0.0699	-0.0669	0.0001	-0.2427	0.2592	-0.2304
RCTR	-0.0480	-0.4926	0.0443	0.1679	-0.1646	-0.1904
UNEMP	0.1364	0.0798	-0.0225	0.0794	-0.0508	-0.1046
PROP	1.0000	0.2914	0.0465	0.1378	-0.1777	-0.2272
HSMMA	0.2914	1.0000	-0.1636	-0.1713	0.0828	-0.0085
PAYCO	0.0465	-0.1636	1.0000	0.0448	-0.0903	-0.0668
TOTPOP	0.1378	-0.1713	0.0448	1.0000	-0.9610	-0.1695
WHIPOP	-0.1777	0.0828	-0.0903	-0.9610	1.0000	0.1672
BLKPOP	-0.2622	0.3252	-0.1555	-0.9096	0.8809	0.1691
HISPOP	-0.1292	0.2237	-0.1067	-0.8436	0.7681	0.3010
INCOMPC	0.4867	0.0896	0.1104	-0.3007	0.2586	0.0108
QMA	0.1471	-0.1228	0.4771	-0.0434	-0.0516	0.1333
BNADV	-0.0387	0.1294	-0.2450	-0.0375	0.0393	0.0058
E1PAY	-0.0292	0.1617	-0.6104	0.0093	0.0196	-0.0264
ARMYMS	-0.2272	-0.0085	-0.0668	-0.1695	0.1672	1.0000
DODMA	-0.0815	-0.0814	-0.0807	0.0246	-0.0719	0.2523

CORRB	BLKPOP	HISPOP	INCOMPC	QMA	BNADV	DODMA
INTER	0.0347	0.0894	0.2461	0.2470	0.2829	-0.0461
YEAR	-0.0356	-0.0887	-0.2387	-0.2400	-0.2857	0.0478
BN	0.2948	0.0530	0.0112	-0.0279	-0.0678	0.1412
RCTR	-0.2737	-0.1935	-0.2313	-0.3261	-0.3092	-0.3699
UNEMP	-0.1356	-0.0142	0.1816	0.0490	0.1162	-0.2819
PROP	-0.2622	-0.1292	0.4867	0.1471	-0.0387	-0.0815
HSMMA	0.3252	0.2237	0.0896	-0.1228	0.1294	-0.0814
PAYCO	-0.1555	-0.1067	0.1104	0.4771	-0.2450	-0.0807
TOTPOP	-0.9096	-0.8436	-0.3007	-0.0434	-0.0375	0.0246
WHIPOP	0.8809	0.7681	0.2586	-0.0516	0.0393	-0.0719
BLKPOP	1.0000	0.7740	0.1558	-0.1520	0.0546	0.0639
HISPOP	0.7740	1.0000	0.1905	-0.0035	0.1148	0.1058
INCOMPC	0.1558	0.1905	1.0000	0.2612	-0.0564	0.0484
QMA	-0.1520	-0.0035	0.2612	1.0000	-0.0100	-0.0070
BNADV	0.0546	0.1148	-0.0564	-0.0100	1.0000	-0.1154
E1PAY	0.0584	0.0107	-0.2366	-0.1627	0.3273	-0.1164
ARMYMS	0.1691	0.3010	0.0108	0.1333	0.0058	0.2523
DODMA	0.0639	0.1058	0.0484	-0.0070	-0.1154	1.0000

variables are correlated and this relationship is possibly contributing towards an error in the parameter estimates. This same line of thought carries over to the model as a whole. When we postulate a $Y = \beta X + \varepsilon$ modei, we are merely implying that there is a linear association between the carrier and the response variables, not necessarily a causal relationship.

To summarize our first regression to this point, we know that there are basic variables to the model as proposed using the current dependent variable, CONT. Furthermore, the regression indicates some collinearity problems which will need to be scrutinized in the full development phase. With the rough indicators that have been derived thus far, we now need to access some preliminary goals for the model.

G. ESTABLISHING GOALS

When attempting to diagnose a problem using only statistical indicators, one must establish a standard by which results will be compared. This chapter has already discussed a few goals that are desired by our analysis. A complete statement of goals by the investigator is desirable at this point so that analytical results can be quickly and decisively interpreted.

1) NUMBER OF PREDICTOR VARIABLES = as few as possible.
2) SIGNIFICANCE OF FINAL VARIABLES < 0.1 (α critical).
3) ROOT MSE < 20% x DEP MEAN => C.V. < 20.
4) VIF < 8 for all variables.
5) CONDITION INDEX < 50 for all variables.
6) FINAL R^2 VALUE = as high as possible.
7) NO DISCERNABLE PATTERN IN THE PLOTTED RESIDUALS.

Figure 3.1 Goals of the GSM I-IIIA Model

With these preliminary goals as stated, the project now passes to the Development Phase.

IV. DEVELOPING THE GSM I-IIIA MODEL

In this chapter we will go into the specifics of developing the GSM I-IIIA model. Much reference will be made in this chapter to Figure 2.2 of Chapter 2. It may be useful to review Figure 2.2 at this time.

A. SEPARATING THE DATA

The first regression has provided information on some of the interactions among the variables. In dealing with longitudinal data, there needs to be checks for both heteroscedasticity and autocorrelation. Presently the data contains 19 carrier variables (the 18 as shown in Table 1 plus the 1 to 55 numerical representations for BN) on 55 battalions over a four year time period. It is desired to analyze this data and check for homogeneity without having the results biased by autocorrelation. The residuals contain *all* of the information concerning the fit of the model. Therefore, they can contain information on both heteroscedasticity and autocorrelation at the same time. By separating the data into time groups (by year) anJ running separate regressions on the individual sets of data, the effects of autocorrelation cannot be observed.

After separating the data base, we now have four separate response and four separate carrier matrices. For example, the matrices for 1982 are as shown below.

$$Y = \begin{bmatrix} 657 \\ 1585 \\ : \\ : \\ 1217 \end{bmatrix} \quad X = \begin{bmatrix} 1 & 1 & 53.75 & 8.05 & 14.7 & .... & 1348 \\ 1 & 2 & 155.00 & 8.60 & 13.4 & .... & 2509 \\ : & : & : & : & : & : & : \\ : & : & : & : & : & : & : \\ 1 & 55 & 103.25 & 11.08 & 8.50 & ...... & 2066 \end{bmatrix} \quad \beta = \begin{bmatrix} b_0 \\ b_1 \\ : \\ : \\ b_{15} \end{bmatrix}$$

where $Y =$ 55x 1 matrix (a column vector of the dependent variables)

$X =$ 55x16 matrix (a column vector of 1's catonated with the 55x15 matrix of the independent variables)

$\beta =$ 16x 1 matrix (a column vector of parameter estimates)

Notice that there are now only 15 carrier variables. First of all, only the numerical BN can be utilized in the least squares regression so the alpha-numerical representation had to be dropped. Also the variables for YEAR and E1PAY had to be dropped because there is no change in their values within any year across any battalion. Their inclusion would make the carrier matrix singular because it would not have full rank.

The restructuring of the data into year groups in order to obtain the carrier matrices can be accomplished by SAS. As shown in Appendix C, the use of the PROC SORT statement will sort the data. This model uses the year as the basic time unit, so our option is to sort the data BY YEAR .

B. ANALYSIS OF THE CROSS SECTIONAL DATA

After running the time grouped cross-sectional data, an analysis is performed in much the same way as was done for the first regression. First of all, it is desired to find basic variables. A summary of the stepwise regressions by year is presented in Table 7.

TABLE 7

BY YEAR STEPWISE SUMMARY OF FIRST REGRESSION DATA

STEPWISE REGRESSION PROCEDURE FOR DEPENDENT VARIABLE CONT

	1982		1983		1984		1985	
STEP	ENTERED	PROB>F	ENTERED	PROB>F	ENTERED	PROB>F	ENTERED	PROP>F
1	DODMA	.0001	DODMA	.0001	DODMA	.0001	DODMA	.0001
2	ARMYMS	.0001	ARMYMS	.0001	ARMYMS	.0001	ARMYMS	.0001
3	RCTR	.0179	RCTR	.0057	RCTR	.0244	RCTR	.0193
4	WHIPOP	.1241	TOTPOP	.1776	QMA	.0556	QMA	.1894
5	BN	.2295	PROP	.2285	BLKPOP	.1325	WHIPOP	.1100
6	PROP	.1657	WHIPOP	.2426	WHIPOP	.1088	BLKPOP	.1527
7	TOTPOP	.3880	PAYCO	.2816	TOTPOP	.0425	TOTPOP	.0328
8	BLKPOP	.1080	UNEMP	.2348	HISPOP	.0103	HISPOP	.1311
9	INCOMPC	.1429	BLKPOP	.5578	HSMMA	.0982	HSMMA	.0570
10	HISPOP	.1470	HISPOP	.2104	UNEMP	.2981	PAYCO	.2095
11	HSMMA	.1539	BN	.5019	PROP	.3227	BNADV	.3659
12	QMA	.2155	INCOMPC	.5001	BNADV	.3865	BN	.3981
13	PAYCO	.4769	BNADV	.6601	PAYCO	.9453	INCOMPC	.5063
14	BNADV	.5207	QMA	.8058	INCOMPC	.9813	PROP	.4273
15	UNEMP	.9208	HSMMA	.8815	BN	.9852	UNEMP	.6861

Table 8 contains the variables, their PROB > |T| statistics and their corresponding Variance Inflation Factors. This information came directly from the matrix of Parameter Estimates with Variance Inflation Factors similar to the one displayed in Table 3 on page 28.

It is time to stop and really think about what is happening in this model. For the proposed model using the dependent variable CONT, there are two dependent variables that are significant in *every* year in *both* the F-Test (Stepwise) and t-Test

TABLE 8

BY YEAR SIGNIFICANCE AND VIF FOR FIRST REGRESSION DATA

	1982		1983		1984		1985	
VARIABLE	PROB>|T|	VIF	PROB>|T|	VIF	PROB>|T|	VIF	PROB>|T|	VIF
INTERCEP	.0001	0.000	.0001	0.000	.0001	0.000	.0001	0.000
BN	.0410	2.165	.4659	2.090	.9852	2.390	.6549	2.139
RCTR	.1435	12.832	.0312	13.889	.0958	15.116	.0665	11.306
UNEMP	.9208	1.748	.5432	1.592	.3255	1.650	.6861	1.934
PROP	.2605	4.352	.4195	4.179	.3567	4.316	.3943	3.256
HSMMA	.0975	8.849	.8815	10.701	.1460	10.358	.1873	9.276
PAYCO	.5194	2.168	.6254	2.775	.9660	4.086	.3185	3.564
TOTPOP	.0087	151.139	.0331	106.092	.0046	120.374	.0062	123.036
WHIPOP	.0240	95.524	.0424	65.671	.0010	69.150	.0007	63.525
BLKPOP	.0089	22.228	.1324	18.541	.0002	18.271	.0013	16.490
HISPOP	.0756	6.619	.1873	6.316	.0039	6.390	.0820	5.906
INCOMPC	.2488	3.503	.5346	3.221	.9790	3.940	.4795	3.354
QMA	.2773	6.710	.7917	6.269	.0123	21.907	.0945	21.762
BNADV	.5362	2.841	.7366	4.064	.4405	2.400	.4212	3.241
ARMYMS	.0001	1.897	.0001	1.695	.0001	2.203	.0001	2.032
DODMA	.0001	6.011	.0001	5.685	.0001	7.014	.0001	6.522

(complete model) statistical analysis. They are DODMA and ARMYMS. There is now only one question that needs to be asked. Is this knowledge of any value to us? The answer is, probably not. First of all, DODMA and ARMYMS are derived ex post facto. Army recruiting battalion areas are unique to the Army. Recruiting areas are not uniform DOD wide. Therefore, it would be difficult and time consuming to attempt to gather data of the proper cross-sectional structure in order to try to predict these variables. This would violate one of the overall objectives of this particular model. Secondly, since the dependent variable, CONT, is utilized to *derive* these two variables, we would *expect* that would all help to explain each other. This is why, in Table 4, over 96% of the model has been explained (model R^2 = .9602) in the stepwise procedure after the introduction of these two variables. Similar results were obtained in the individual year stepwise regressions, with anywhere from R^2 = .953 for 1983 to R^2 = .981 in 1985 after the introduction of just these two variables.

The variable RCTR is significant in every stepwise procedure (Table 7) and every t-Test (Table 8) except for 1982. It seems to be a good predictor. It is easily obtainable and, to a certain extent, controllable. It has good potential for predictability. One only needs to look at present and proposed recruiter manning rosters. RCTR, however, does seem to have significant collinearity problems. It exceeds our goal of VIF < 8 for every year in Table 8. Checking the Correlation of Estimates Table (not shown here but similar to Table 6 of Chapter 3) RCTR is most highly correlated to HSMMA in 1982 (-.4325), HSMMA in 1983 (-.5209), DODMA and HSSMA in 1984 (-.5290 and -.4809 respectively) and INCOMPC, QMA and DODMA in 1985 (-.4187, -.4043 and -.4214 respectively).

Another noteworthy factor is that WHIPOP and TOTPOP in Table 7 seem to be more significant than any of the other population variables. Other studies have shown that areas of greater multiethnic population tend to attract significantly more recruits than other areas. [Ref. 6] This would lead us to believe that the higher range concentrations of WHIPOP would possibly have a detrimental effect on contracts. We cannot, however, surmise anything yet as to why these two variables might be significant. Our model has problems with collinearity with both WHIPOP and TOTPOP. Both have VIF substantially greater than 8 in Table 8. Other significant collinearity problems seem to be arising with HSMMA, QMA and BLKPOP.

Unemployment is not a significant indicator at all. In Table 7 for 1982 and 1985, it is the *least* significant of all of the predictor variables. Although this is counterintutive, it has also been shown in previous studies to be both significant and insignificant in explaining GSM I-IIIA accessions, depending upon the year and the dependent variable that is being studied. [Ref. 7] It may be that we are not using this statistic in the most appropriate manner and should be thinking about alternate possibilities of unemployment indicators for inclusion into the model.

Also, PROP is not a significant predictor. In Table 3 on page 28, the parameter estimate for the first regression (entire set of data) was equal to -0.559418. The negative sign of the parameter estimate is counterintutive (similar to the negative sign that we obtained with just 3 data points in Figure 1.1). This may be telling us something. Parameter estimates for PROP in each year group regression were positive for 1982 and 1983, but negative for 1984 and 1985. The α (actual) values for the t statistic (Table 8) ranged from .2605 for 1982 to .3943 for 1985. All of these values are outside of our model goals of α (critical) = .1. One reason that comes to mind when attempting to explain this may be that *propensity is high in smaller markets and low in larger markets.* Thus, although propensity may be high, it will not necessarily explain a high (in absolute terms) number of contracts.

There seems to be much work that needs to be done here. The results of the first regression, along with the results of the first set of time grouped regressions show many problems, especially with collinearity. Correlation is good if it is between the carrier and predictor variables. It is not good if it is just between the predictors.

C. THE SECOND REGRESSIONS

At this time we decide to drop both DODMA and ARMYMS and rerun the regressions. This series of regressions will be referred to as the second regression. In order to circumvent the obvious problem of multicollinearity between WHIPOP and TOTPOP, yet still retain them in the predictor matrix, a new variable is adopted. This new term, PERCWI (for percent white) is merely the WHIPOP divided by TOTPOP. In SAS, this is easily produced by the algebraic equation immediately following the INPUT line (Appendix D). Also dropped is the QMA variable. QMA was displaying some problems with collinearity. In looking at Appendix B, it is noticed that QMA is usually derived as a straight percentage of TOTPOP and only updated once every other year, whereas HSMMA is a number based on actual counts that are performed by recruiters and verified at certain non-specific time intervals by the Area Recruiting Zone (ARZ) verification teams. All else being equal, HSMMA is a prefered statistic because of its perceived accuracy. Since QMA and HSMMA are closely related, and since there is also a problem with collirearity in the HSMMA variable, it is anticipated that dropping QMA might help to alleviate this collinearity problem with HSMMA as well.

The results of the second regression are only slightly encouraging. Tables 9 and 10 present the summary of the second regression results for the overall and year grouped data bases. The regressions modeled 13 dependent variables versus CONT. The far left column of Table 10 lists the independent variables used in these regressions. These tables present the results as compared to the preliminary established goals of the model as outlined in Figure 3.1 of Chapter 3.

The R^2 values all fell substantially, but this was to be expected after dropping the two derived variables, DODMA and ARMYMS. The t statistic indicates that RCTR is significant in every year, as does the stepwise regression procedure. The new variable, PERCWI, is significant in every year with the stepwise procedure. Furthermore, none of the population parameters are showing any signs of collinearity problems. UNEMP and PROP, two variables that have been historically good indicators, are significant in some years, but not in others. The VIF and Condition Index (C.I.) indicate multicollinearity, especially with RCTR and HSMMA. Until this problem can be solved, many of the key indicators are suspect in their accuracy.

There are several issues that arise from the second regression. The first is the question of why BN would be a significant variable. BN is merely an ordinal number

TABLE 9

SECOND REGRESSION RESULTS VERSUS ESTABLISHED GOALS

	1982	1983	1984	1985
R^2=	.82	.75	.60	.60
VARIABLES WHOSE PROB>|T| WAS < 0.1	RCTR PROP HISPOP UNEMP PERCWI BN BNADV	RCTR PROP HISPOP	RCTR PROP HISPOP	RCTR PAYCO BNADV
VARIABLES w/C.I. > 50 (TOTAL #)	2	2	1	1
VARIABLES w/VIF > 8	RCTR	RCTR HSSMA	RCTR HSMMA	-
C.V. <20	YES	YES	YES	YES

TABLE 10

SECOND REGRESSION STEPWISE RESULTS FOR VARIABLES WITH PROB>F < 0.1

PARAMETER ESTIMATES OF SIGNIFICANT VARIABLES FROM STEPWISE REGRESSION

	1982	1983	1984	1985
BN	3.116	-	-	2.262
RCTR	7.601	13.37	10.63	5.291
UNEMP	26.56	25.05	-	-
PROP	17.56	27.36	16.84	-
HSMMA	-	-	-	-
PAYCO	-	-	-	-
PERCWI	1292	208.4	189.6	283.3
BLKPOP	-	-	-	-
HISPOP	-17E-5	-27E-5	-17E-5	-
INCOMPC	-	-	-	-
BNADV	.3034	-	-	.1468

given to the alpha-numeric battalion names. One must be *very* careful when using substitute ordinal level data in a regression equation. In this instance, however, it is signifying an interesting phenomenon. Why does the mere battalion name signify contracts? Part of the answer has to do with the concept of lurking or latent variables. As stated previously, there is no possible way in which one can collect numerical data on all possible aspects of the recruiting process. There are many undefinable or uncaptureable nuances that lead to the decision to enlist in the Army. Intangables such as leadership within the recruiting battalion, a wealth of overachieving recruiters, favorable local school officials or the mere history of being the 'best', 'worst' or an 'also ran' battalion may have significant impact. The fact that BN is showing up as a significant variable implies that battalions are doing the way they are just because they are named that battalion. In an attempt to capture this phenomenon and to discard the substitute numbering system for the battalions, the analyst checked several indicators of battalion output history over the four years covered by this study. Instead of merely using the (constant interval) BN number, another variable was contemplated that would more readily capture the 'spread' between the battalions. After several trials, the variable BNPER (meaning battalion percent) was adopted. It is the number of contracts signed by a battalion in a particular year, divided by the total number of contracts signed in that year. For example, BN 1A signed 657 contracts in 1982. There were a total of 51,431 contracts signed in 1982. Therefore, BN 1A is given a new variable of 657/51431 = 0.0127744. In looking at all of the battalions over all of the years, the standard deviation of this indicator is less than one third of its mean and it is fairly normally distributed with no significant skewing. Some battalions are always near the top percent of total recruits, and some are always near the bottom.. This variable allows the analyst to control his inputs at the battalion level based on his knowledge of a particular unit. For instance, although a particular battalion usually recruits about 2.5 % of the total mission, a leadership change or a high recruiter turnover rate or a particularly disastrous local situation may force the analyst to decrease that number and re-distribute it to another more favorable location. Or, some demographic phenomenon may lead to an entire region (or Brigade) having their inputted numbers shifted. If this much detail is not desired, we can merely plug in the percent of total mission that has been assigned to that unit as a result of the latest Enlisted Personnel Model (EPM) run.

There are some valid concerns with using proportions as predictor variables. First of all, their average value will never change (it will always be 1.00 total number of battalions in this case). Secondly, this particular variable could not be used with the dependent variable, CONT, because they are linear functions of one another. It would be just like artificially plugging in equalities on both sides of the hypothesized linear regression equation. We are still, however, in the trial and error mode, so maybe we will be able to utilize this new variable in a future regression run.

The second issue is that PROP is now becoming a significant variable. As stated earlier, it is speculated that propensity may be more of a proportion indicator than an absolute value indicator. This might be due to higher propensities in smaller market areas and vice versa. In Table 10, it is now seen that PROP has all positive parameter values. The reason that PROP would now have all positive parameter values when in the first regression, it had both positive *and* negative values has to do with the concept of the *costock*. [Ref. 5] In speaking of the costock of a independent variable, we are refering to all of the other independent variables in a particular regression. For example, if we were modeling CONT versus RCTR, UNEMP and PROP, the costock of PROP is RCTR and UNEMP. The thing to remember is that *the value of a parameter estimate of a particular independent variable may have more to do with the data values of its costock than it does with its own data values.* In other words, as given in the example above, the derived parameter estimates for PROP may be more a function of the data values of RCTR and UNEMP than the data values of PROP itself.

With this in mind, we look at another aspect of the second regression. In Table 9 and 10 we notice that there are different significant variables in different years. As a matter of fact, there are no two years in which the significant variables are the same. We know that the costock has a lot to do with the values of a particular regression equation. All else being equal, we would certainly prefer that the regression equations for each year contain the same variables at the same level of significance. If this were to happen, we could *compare* parameter estimates with some degree of validity. One of the largest abuses of regression analysis is when an attempt is made to try to compare parameter estimates that have been derived from two different regressions using two different costocks. *These types of comparisons are not valid.*

Finally, the second regression is somewhat unstable across time periods in the R^2 values that are achieved (see Table 9). These R^2 values are not necessarily bad, but since we are building a predictive model, a higher R^2 value is prefered. We are not sure just how high of an R^2 value can be obtained from this particular data base. If there are any ties in the data values of a particular independent variable in the carrier matrix, the R^2 value can never attain unity. This is because the regression hyperplane would be trying to fit itself through the two different points in the same plane, which cannot be done. This phenomenon is known as pure error. If pure error is present in a data base, the R^2 value can never be 1.0. We do not know how much pure error is present in this regression, but higher R^2 values will be prefered.

D. THE THIRD AND SUBSEQUENT REGRESSIONS

A third regression is now planned. In order to check the PROP variable against our suspicions that it is a proportion indicator, we contemplate changing the dependent variable. Again, we must remember that the overall goal of the model is to predict total GSM I-IIIA contracts. Perhaps a dependent variable of CONT/TOTPOP or CONT/QMA would give us some indication of the proportion of a specific population that a recruiting battalion is actually enlisting. One term that is utilized by the recruiting community is that of Penetration. Penetration is the proportion of contracts that are signed per the market of GSM I-IIIA available. We adopt the term PENT, which equals CONT/HSMMA. This looks to be an ideal response variable because we have seen that there is definitely collinearity between HSMMA and the other predictor variables (see Table 9). By putting HSMMA on the response side and dropping it from the predictor side, we expect to decrease the problem with multicollinearity. Also, we can now utilize the variable BNPER since there is no longer a strict linear function between it and PENT. Since this is an entirely new approach with a new dependent variable, we will keep all of the other carrier variables for this regression.

The results of this regression are much more encouraging. Tables 11 and 12 present the summary of the third regression results for the year grouped data bases. The far left column of Table 12 lists the independent variables used in these regressions versus the dependent variable PENT.

TABLE 11

THIRD REGRESSION RESULTS VERSUS ESTABLISHED GOALS

	1982	1983	1984	1985
R^2=	.81	.85	.84	.77
VARIABLES WHOSE PROB>|T| WAS < 0.1	PROP BNPER RCTR BLKPOP	PROP BNPER RCTR BLKPOP INCOMPC	PROP BNPER RCTR PERCWI	PROP BNPER RCTR PERCWI INCOMPC
VARIABLES w/C.I. > 50 (TOTAL #)	2	2	1	1
VARIABLES w/VIF > 8	-	RCTR	-	-
C.V. <20	YES	YES	YES	YES

TABLE 12

THIRD REGRESSION STEPWISE RESULTS FOR VARIABLES WITH PROB>F < 0.1

PARAMETER ESTIMATES OF SIGNIFICANT VARIABLES
FROM STEPWISE REGRESSION

	1982	1983	1984	1985
PROP	11E-4	10E-4	90E-5	13E-4
BNPER	2.658	3.386	2.602	2.289
RCTR	-62E-5	-71E-5	-54E-5	-45E-5
PERCWI	-13E-3	-27E-3	-53E-3	-28E-3
UNEMP	-	-	-	-
PAYCO	-	-	-	-
BLKPOP	12E-9	-15E-9	-	-
HISPOP	-	-	-	-
INCOMPC	-	-25E-7	-16E-7	-24E-7
BNADV	-	-	-	-

As compared with Table 9, the R^2 values have increased for most years and are more stable. There is more stability in the variables across the years in that PROP, BNPER and RCTR appear in every year using both the t-Test and the stepwise F Test. PERCWI also shows up every year in the stepwise procedure. There is only one VIF greater than 8, and that is for RCTR in 1983. There is still collinearity problems in every year according to the Condition Index numbers.

Checking for collinearity in the Correlation of Parameter Estimates Matrix for this regression (see Table 13) it is noted that there are several variables that indicate a $\rho > |0.4|$. Our collinearity problems are very probably arising with one of these relationships. Since UNEMP, PAYCO, HISPOP, BNADV and E1PAY are not significant in any year in Tables 11 and 12, these are the first candidate variables to be dropped in the next regression attempt. Checking these variables against Table 13, it is seen that UNEMP, PAYCO and BNADV are not highly correlated with any other variable, HISPOP is negatively correlated with RCTR (-0.4331), and E1PAY is correlated with PROP and INCOMPC (-0.4481 and -0.6207 respectively).

Dropping these five insignificant variables and running a fourth regression still indicated a condition index greater than 50 for one variable. Since BLKPOP and PERCWI are highly correlated ($\rho > .7$), these are the two suspect variables as to the probable cause of this indicator of multicollinearity. In trying to determine which of these variables to drop, it is decided that BLKPOP should go because it has been shown to be the least significant in more years than PERCWI.

TABLE 13

CORRELATION OF PARAMETER ESTIMATES FROM SAS FOR THE THIRD REGRESSION

CORRB	INTER	PROP	BNPER	RCTR	HISPOP	INCOMPC
INTERCEP	1.0000	0.3493	-0.0150	0.0992	-0.1431	0.4520
PROP	0.3493	1.0000	-0.1865	0.4271	-0.0858	0.4853
BNPER	-0.0150	-0.1865	1.0000	-0.6623	0.2929	0.0071
RCTR	0.0992	0.4271	-0.6623	1.0000	-0.4331	-0.1051
PERCWI	-0.1230	-0.1614	-0.0952	-0.3369	0.3454	0.1932
UNEMP	-0.0517	0.1821	-0.2426	0.0179	0.0919	0.2350
PAYCO	0.0200	0.2116	0.1169	-0.0075	-0.0291	0.2443
BLKPOP	-0.2060	-0.5186	0.1636	-0.5566	0.2739	-0.0208
HISPOP	-0.1431	-0.0858	0.2929	-0.4331	1.0000	-0.0580
INCOMPC	0.4520	0.4853	0.0071.	-0.1051	-0.0580	1.0000
BNADV	-0.1457	-0.1561	-0.1208	-0.3229	0.1673	-0.1193

CORRB	PERCWI	UNEMP	PAYCO	BLKPOP	BNADV
INTER	-0.1230	-0.0517	0.0200	-0.2060	-0.1457
PROP	-0.1614	0.1821	0.2116	-0.5186	-0.1561
BNPER	-0.0952	-0.2426	0.1169	0.1636	-0.1208
RCTR	-0.3369	0.0179	-0.0075	-0.5566	-0.3229
PERCWI	1.0000	0.0633	-0.0043	0.7239	0.1211
UNEMP	0.0633	1.0000	-0.2970	-0.0648	0.1453
PAYCO	-0.0043	-0.2970	1.0000	-0.0528	-0.0941
BLKPOP	0.7239	-0.0648	-0.0528	1.0000	0.1129
HISPOP	0.3454	0.0919	-0.0291	0.2739	0.1673
INCOMPC	0.1932	0.2350	0.2443	-0.0208	-0.1193
BNADV	0.1211	0.1438	-0.0941	0.1129	1.0000

Now a fifth regression was run. The independent variables were PROP, BNPER, RCTR, PERCWI and INCOMPC. The dependent variable was PENT. For every year except 1985, INCOMPC was the last variable to enter the stepwise regression. It was also an insignificant variable in 1982 according to the t-Test. Every other variable for every other year was significant for both tests. There was, however, still a collinearity problem. A single condition index of greater than 50 was noted for every separate year regression.

Several combinations using four of the five independent variables listed above were then tried. This is because one of our goals in this model is to use as few predictor variables as possible. It must be remembered that for every variable that is included in the model, the analyst must take the time and effort to predict that variable. It is hoped that a combination of four could be found that was 'as good as' the above combination of five. Any combination chosen had to meet all of the goal criteria as set forth in Figure 3.1. Finally, one 'best equation' was chosen. It was decided that INCOMPC could be dropped with no substantial loss to the model. This was determined when checking the partial R^2 values as given in the stepwise summary (similar to Table 4). The partial R^2 values for INCOMPC ranged from 0.0001 in 1982 to 0.03 in 1985. These added values to the overall R^2 were considered insignificant. The dropping of this variable also solved the condition index collinearity problem, with the highest index value being 32.16 for 1982 which is well below our goal of 50.

Before moving on a few issues need to be addressed. Although we have named the regressions first, second, etc., this is really a misnomer. There have actually been scores of regressions run to this point, each checking a different aspect of the problem or verifying the intuitions of the analyst. One can do this to the point where the data tends to dictate the 'next move' of the analyst. If this happens, we will end up with a model that will only fit the data that is contained in the data base. A predetermined set of goals (such as Figure 3.1) tends to counter this problem. Also, the validation phase contains provisions to check the model with *different* data to assure the model's validity.

The most notable work with the other regressions was with the unemployment variable, UNEMP. For the time span of this study, UNEMP was not a significant variable except for a few regressions, mostly in 1982. This is counterintutive to most USAREC analysts. An attempt was made to transform this variable in two distinct ways.

First of all, a variable called CHUNEMP was attempted. This variable was actually the change in unemployment within a battalion between years. This was derived by using the following formula.

$$\text{CHUNEMP}_t = (\text{UNEMP}_t - \text{UNEMP}_{t-1}) / \text{UNEMP}_{t-1}$$

$$\text{where } t = 1983, 1984, 1985$$

This variable did not prove to be any more significant than the UNEMP variable.

Also, a dummy variable was defined as a battalion either being above or below the average national unemployment as calculated by the Bureau of Labor Statistics. It was hypothesized that although perspective accessions might not be familiar with their particular unemployment rate, they could be cognizant of whether they were in an area that was higher or lower than the national average as reported in the local media. This dummy variable also did not prove to be significant.

The only logical explanation for this is that the costock of UNEMP is carrying the signal from UNEMP. It is thought that PROP is the predominant carrier of the signal since PROP is the most significant variable in all of the regressions and is a variable that is designed to capture several signals that may or may not be otherwise measured.

The bottom line at this point is that although this study has discussed five regressions to end up with four variables, the trials and thought processes that have actually taken place significantly exceeds that which is discussed in the text.

E. CHECKING FOR LEVERAGE

In regression model building, one should check *every* regression equation for possible lack of fit due to outliers. Outliers may cause an effect called *leverage* which can cause a significant decrease in R^2 values.

One method of finding outliers is to look at the "studentized" residuals. These residuals are produced when a P or R is requested in the option section of the MODEL statement in SAS (see Appendix D). Studentized residuals are merely the actual residuals that have been set to a normal distribution with a variance of one. Therefore, we would expect their values to range from about -3.0 to +3.0. With the sample size of 55 battalions per year that we have, we would expect that approximately two residual values per year would exceed |1.96|. Looking down the list of studentized residuals in Table 14, we notice that there are two residuals that are outliers in 1982 (6E and 6J), eleven in 1983 (3B, 3D, 3E, 3F, 3G, 3H, 3J, 3K, 5A, 5B, 6G) none in 1984 and one in 1985 (6G). These battalions should be rechecked to insure that their underlying data base is accurate. If it seems to be proper, the analyst should attempt to explain the deviation that these samples are displaying.

Another more powerful indicator of lack of fit due to leverage is the Cook's D statistic. [Ref. 4] It is also located in Table 14. It measures two things at once. Cook's D will get large when (1) the residual gets large and (2) when there is an outlier data point that is lying outside of the data cloud in the carrier hyperspace and is exerting some leverage on the regression plane. In Table 14, we notice that the Cook's D statistic is significantly larger in 1982 for 6E; in 1983 for 3B, 3D, 3K, 5A, 6E and 6G; in 1984 for 1N, 6E and 6G; and in 1985 for 6G and 6H.

Discarding data from the data base is a judgement call on the part of the analyst. One should never discard data from the data base without significant reason. The

TABLE 14

YEARLY VALUES OF "STUDENTIZED" RESIDUALS AND COOK'S D STATISTICS

	1982		1983		1984		1985	
	STDNT RESID	COOKS D	STDNT RESID	COOKS D	STDNT RESID	COOKS D	STDNT RESID	COOKS D
1A	-0.152	0.000	0.973	0.004	-1.068	0.006	0.222	0.000
1B	-1.376	0.018	0.412	0.002	-0.318	0.001	-0.238	0.001
1C	-0.935	0.002	0.728	0.002	0.643	0.002	0.044	0.000
1D	-0.146	0.000	0.265	0.000	-0.734	0.004	-0.266	0.000
1E	-1.327	0.005	-0.536	0.001	-0.749	0.002	-0.270	0.000
1F	-0.148	0.000	0.287	0.000	-0.348	0.001	-0.711	0.002
1G	0.804	0.006	1.660	0.025	0.545	0.002	-0.123	0.000
1H	-0.894	0.005	0.421	0.001	0.205	0.000	0.094	0.000
1I	-0.491	0.001	-0.132	0.000	-0.825	0.003	-0.505	0.001
1K	-0.056	0.000	0.490	0.001	0.741	0.002	0.597	0.001
1L	-1.026	0.002	0.406	0.000	-0.901	0.003	-0.997	0.006
1N	-1.220	0.006	0.210	0.000	-1.737	0.012	-1.244	0.004
3A	-1.567	0.015	0.341	0.001	-1.273	0.006	-0.961	0.007
3B	-0.267	0.001	2.777	0.050	0.280	0.000	0.629	0.003
3C	-1.108	0.004	0.063	0.000	-0.570	0.001	-1.149	0.010
3D	0.292	0.001	2.679	0.049	-0.050	0.000	0.142	0.000
3E	-0.853	0.004	2.143	0.029	-0.031	0.000	0.968	0.007
3F	0.299	0.000	2.278	0.012	0.524	0.001	0.725	0.002
3G	0.632	0.001	2.046	0.008	-0.258	0.000	0.997	0.003
3H	0.135	0.000	2.376	0.019	1.122	0.004	1.536	0.009
3I	-1.126	0.004	1.174	0.002	0.027	0.000	0.387	0.000
3J	-1.779	0.019	2.197	0.028	-0.280	0.001	-0.315	0.001
3K	-0.519	0.002	2.874	0.043	0.478	0.001	-0.148	0.000
4A	-0.829	0.002	1.336	0.005	0.552	0.001	1.066	0.004
4C	-1.217	0.007	-0.613	0.001	-1.001	0.002	-1.345	0.003
4D	-0.467	0.001	1.012	0.003	0.492	0.001	1.006	0.005
4E	-1.460	0.007	-0.230	0.000	-1.299	0.007	-1.113	0.006
4F	-0.584	0.001	1.830	0.015	0.250	0.000	0.554	0.002
4G	-0.475	0.000	0.846	0.002	0.122	0.000	0.411	0.000
4H	-1.310	0.004	0.342	0.000	-0.421	0.001	-0.108	0.000
4I	-1.641	0.015	0.343	0.000	-0.475	0.001	0.268	0.000
4J	-0.968	0.004	0.183	0.000	-0.295	0.000	0.256	0.000
4K	-1.088	0.005	0.415	0.001	0.297	0.000	0.829	0.001
5A	1.146	0.009	2.169	0.032	0.426	0.001	0.288	0.001
5B	1.105	0.003	2.169	0.008	1.593	0.004	0.948	0.002
5C	-0.505	0.001	1.509	0.013	-0.147	0.000	0.336	0.001
5D	-0.822	0.001	0.445	0.000	-0.614	0.001	-0.198	0.000
5E	-0.151	0.000	0.897	0.003	0.166	0.000	0.147	0.000
5F	-1.020	0.009	-0.078	0.000	-0.108	0.000	0.811	0.002
5H	-0.475	0.000	1.085	0.002	0.207	0.000	0.040	0.000
5I	-1.711	0.012	0.050	0.000	-0.412	0.001	0.366	0.000
5J	-1.100	0.003	0.350	0.000	-0.395	0.000	0.284	0.000
5K	-1.735	0.015	0.947	0.004	0.138	0.000	0.383	0.000
5L	-1.113	0.003	0.186	0.000	-0.609	0.001	0.206	0.000
5M	-1.541	0.005	0.253	0.000	0.125	0.000	0.431	0.001
5N	-0.435	0.000	1.578	0.011	0.242	0.000	-0.338	0.000
6A	-0.534	0.002	0.053	0.000	-0.835	0.002	-0.852	0.002
6E	-2.251	0.122	-1.212	0.034	-0.799	0.015	-0.303	0.002
6F	-1.021	0.009	-0.036	0.000	-0.392	0.001	-1.057	0.008
6G	0.883	0.004	2.927	0.040	1.899	0.020	2.207	0.027
6H	-0.444	0.002	0.834	0.005	-1.381	0.007	-1.737	0.021
6I	-1.017	0.002	0.424	0.000	0.023	0.000	-1.319	0.004
6J	-2.044	0.027	-0.626	0.002	-1.411	0.008	-1.754	0.009
6K	-1.388	0.003	-0.141	0.000	-0.293	0.000	-0.913	0.002
6L	0.111	0.000	1.512	0.006	0.811	0.002	0.769	0.003

biggest perpetrators of lack of fit for this model seems to be battalions 6E, 6G and 6F. It is the judgement of the analyst to discard 6E and to keep the rest. The reasoning for this is that battalion 6E represents Honolulu, which is an extreme point in almost every statistical variable that is included in the model. Also, its actual contributions to contracts (approximately one-half of one percent) is negligible. In consulting with experienced USAREC analysts, Honolulu (along with San Juan, P.R.) are seldom used in other regression models due to their peculiar demographics and unique characteristics.

On the other hand, 6G and 6F represent the Phoenix and the Portland battalions. Phoenix is undoubtedly and outlier due to its low PERCWI value and Portland due to its low PROP value. In any event, their exclusion is not deemed appropriate due to the fact that they contribute significantly more total contracts than does Honolulu. In fact, their inclusion (with associated range of carrier variables) may tend to add to the robustness of the model.

F. THE FINAL REGRESSIONS

After discarding the values for battalion 6E and rerunning the regression, an across the board increase in R^2 values is obtained. Partial R^2 increases ranged from .0034 in 1985 to .0232 in 1983.

Table 15 shows the pertinent regression statistics for the final regression of 1985. Other years were nearly identical. In every year the stepwise procedure brought in the variables in the same order (PROP,BNPER,RCTR then PERCWI). Tables 16 and 17 display the results of the final regressions which determined our 'best separate equations'. A detailed discussion of these result will be provided later in the text.

Notice that every variable is significant in each test in each year (each was significant at the 0.0001 level). All parameters are equivalent in magnitude and signed the same. The regressions are stable across time periods and indicate fairly good R^2 values for cross-sectional data. Since they each contain the same costocks, their parameter estimates are comparable. We are satisfied that these regressions have achieved our preliminary goals as specified in Figure 3.1. It is now time to check the underlying assumptions of multivariable regression analysis to insure that these equations are valid.

TABLE 15

FINAL RESULTS OF 1985 YEAR GROUP REGRESSION

```
                                          SAS
                                       YEAR=1985
 DEP VARIABLE:  PENT
                          SUM OF            MEAN
 SOURCE      DF          SQUARES          SQUARE        F VALUE         PROB>F
 MODEL        4      0.008142583     0.002035646         35.863         0.0001
 ERROR       49      0.002781353   .00005676231
 C TOTAL     53         0.010924
     ROOT MSE        0.007534076        R-SQUARE         0.7454
     DEP MEAN           0.055076        ADJ R-SQ         0.7246
     C.V.               13.67944


                   PARAMETER       STANDARD     T FOR HO:                  VARIANCE
VARIABLE  DF        ESTIMATE          ERROR    PARAMTR=0   PROB>|T|       INFLATION

INTERCEP   1        0.070790       0.013362        5.298     0.0001        0.000000
PROP       1     0.001561433   0.0002565583        6.086     0.0001        1.312525
BNPER      1        2.568842       0.350706        7.325     0.0001        3.068071
RCTR       1    -0.000515104  .00007407063       -6.954     0.0001        3.136923
PERCWI     1       -0.051608       0.012295       -4.197     0.0001        1.410820


                                CORRELATION OF ESTIMATES
 CORRB          INTERCEP           PROP          BNPER            RCTR          PERCWI

 INTERCEP         1.0000        -0.6737         0.3081         -0.5463         -0.8637
 PROP            -0.6737         1.0000        -0.1487          0.3528          0.3389
 BNPER            0.3081        -0.1487         1.0000         -0.7903         -0.4657
 RCTR            -0.5463         0.3528        -0.7903          1.0000          0.4311
 PERCWI          -0.8637         0.3389        -0.4657          0.4311          1.0000

COLLINEARITY DIAGNOSTICS          VARIANCE PROPORTIONS

                    CONDITION    PORTION    PORTION    PORTION    PORTION    PORTION
NUMB  EIGENVALUE        INDEX   INTERCEP       PROP      BNPER       RCTR     PERCWI

1          4.815        1.000     0.0003     0.0022     0.0010     0.0010     0.0004
2       0.128379        6.124     0.0010     0.2313     0.0365     0.C447     0.0004
3       0.034521       11.811     0.0191     0.3913     0.0210     0.0974     0.1328
4       0.018265       16.237     0.0294     0.0277     0.7019     0.4610     0.0029
5       0.003529       36.941     0.9503     0.3474     0.2397     0.3960     0.8635
```

G. CHECKING FOR HOMOGENEITY IN THE RESIDUALS

Our 'best separate equations' to this point are of the form:

$$PENT_t = \beta_{0,t} + \beta_{1,t}PROP + \beta_{2,t}BNPER + \beta_{3,t}RCTR + \beta_{4,t}PERCWI + \varepsilon_t$$

where t = 1982, 1983, 1984, 1985

These equations were derived under the assumption that the residual errors are independent, that they have a mean of zero, that they have a constant variance (known

TABLE 16

FINAL REGRESSION RESULTS VERSUS ESTABLISHED GOALS

	1982	1983	1984	1985
R^2=	.78	.83	.82	.74
VARIABLES WHOSE PROB>|T| WAS < 0.1	PROP BNPER RCTR PERCWI	PROP BNPER RCTR PERCWI	PROP BNPER RCTR PERCWI	PROP BNPER RCTR PERCWI
VARIABLES w/C.I.> 50 (TOTAL #)	0	0	0	0
VARIABLES w/VIF > 8	-	-	-	-
C.V. <20	YES	YES	YES	YES

TABLE 17

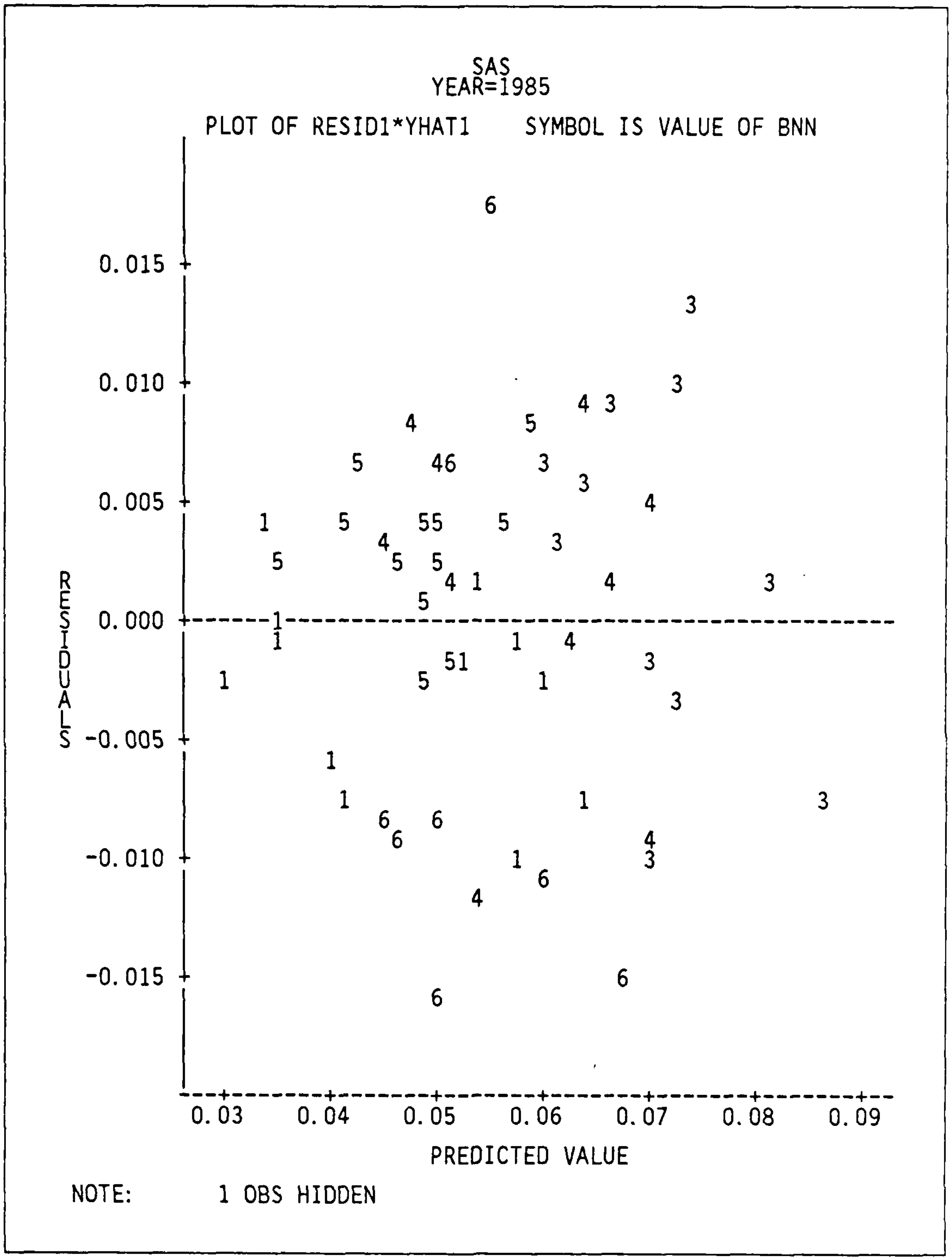
FINAL REGRESSION STEPWISE RESULTS FOR VARIABLES WITH PROB> F < 0.1

PARAMETER ESTIMATES OF SIGNIFICANT VARIABLES
FROM STEPWISE REGRESSION

	1982	1983	1984	1985
PROP	11E-4	16E-4	12E-5	15E-4
BNPER	2.659	3.376	2.771	2.568
RCTR	-56E-5	-69E-5	-55E-5	-51E-5
PERCWI	-58E-3	-74E-3	-66E-3	-51E-3

as homogeneity) and that they conform to a normal distribution. Heteroscedasticity is where the model fails to meet the assumption of constant variance. The easiest method of checking these regressions for heteroscedasticity is by plotting the residuals.

The most common residual plot is the plot of the residuals versus the predicted values. The reason for this is because the covariance between the residuals and the predicted values is equal to zero. The procedure PROC PLOT in Appendix D indicates how to get these residual plots from SAS. Each individual year has to be generated and checked. Figure 4.1 is the graph of the residuals versus the predicted values for the year 1985. This is actually a three dimensional graph in that the plotted



```
NOTE:       1 OBS HIDDEN
```

Figure 4.1 1985 Plot of Residuals vs Predicted Values data points indicates which battalion is being plotted. The resolution of SAS is only down to the first number of the battalion (a plot of 1 can indicate from battalion 1A to 1N), but it can give a quick indication of which general region is contributing the most to the error in the model. In this particular graph, most of the 1's are lying below the zero reference line and most of the 5's are lying above. This quickly gives us an indication that the First Brigade is below the regression plane for Penetration and Fourth Brigade is lying above. Similar results were obtained for the other three years. There is no discernable pattern in this year (nor were there in any other years) and we can tentatively conclude that there is no heteroscedasticity within the year groupings.

Plotting the residuals against the predicted values is not the only plot that can or should be used. Plotting the residuals against the independent variables can give some indication as to whether a transformation of the variables is needed. If the residuals plot out in a megaphone type shape (close to each other on one side of the graph and spread apart on the other) then there is a problem with constant variance. If this pattern is apparent, then a transformation on the response variables may be needed or a weighted least squares regression method is required. [Ref. 3:p. 148] An archlike pattern may indicate the need for extra terms (such as a quadratic). Figure 4.2 shows such a plot for each independent variable for a different year. Appendix D specifies how to produce these plots from SAS. Again, each plot in each year must be checked. These plots indicated no discernable pattern and heteroscedasticy is not indicated.

H. CHECKING FOR NORMALITY IN THE RESIDUALS

One of the most important indicators that the model is correct is in the checking of the residuals for normality. This is an initial assumption for the derivation of the regression equations and is crucial for the validity of using F-Tests as key statistical indicators. Furthermore, if there is no discernable pattern in the residuals and if the residuals can be shown to follow a normal distribution, then there is no graphical or statistical indication that heteroscedasticity is present in the proposed models.

One of the quickest methods of checking for normality is to plot the residuals and visually determine if the pattern follows a normal bell-shaped distribution. SAS can accomplish this using the PROC CHART statement as presented in Appendix D. The output for this procedure for 1985 is as shown in Figure 4.3 . This figure tends to support the assumption of a normal distribution, as did the charts of the other years.

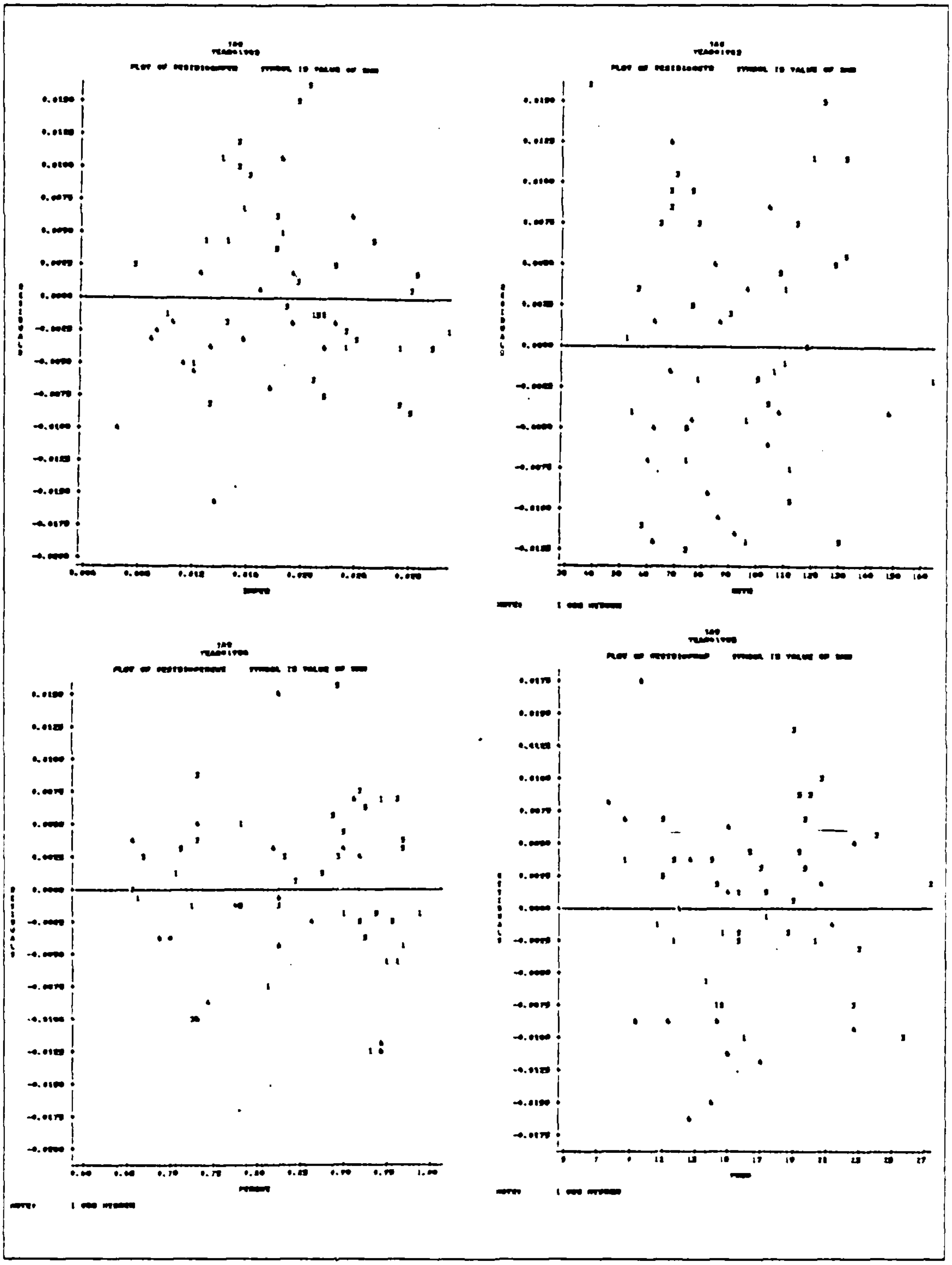
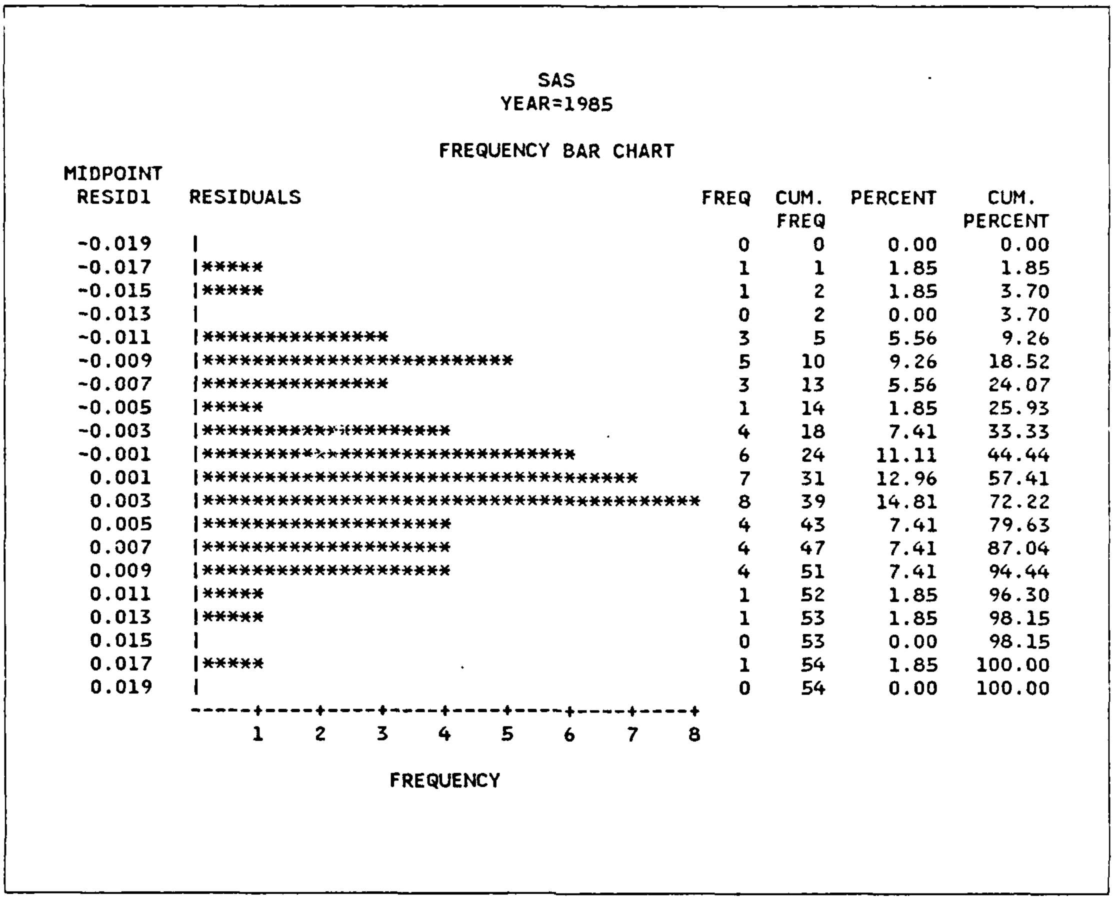


Figure 4.2 Selected Plots of Residuals vs Independent Variables



```
                              SAS
                           YEAR=1985

                       FREQUENCY BAR CHART
MIDPOINT
 RESID1   RESIDUALS                                  FREQ  CUM.  PERCENT    CUM.
                                                           FREQ           PERCENT
 -0.019   |                                            0     0     0.00     0.00
 -0.017   |*****                                       1     1     1.85     1.85
 -0.015   |*****                                       1     2     1.85     3.70
 -0.013   |                                            0     2     0.00     3.70
 -0.011   |***************                             3     5     5.56     9.26
 -0.009   |*************************                   5    10     9.26    18.52
 -0.007   |***************                             3    13     5.56    24.07
 -0.005   |*****                                       1    14     1.85    25.93
 -0.003   |********************                        4    18     7.41    33.33
 -0.001   |******************************              6    24    11.11    44.44
  0.001   |***********************************         7    31    12.96    57.41
  0.003   |****************************************    8    39    14.81    72.22
  0.005   |********************                        4    43     7.41    79.63
  0.007   |********************                        4    47     7.41    87.04
  0.009   |********************                        4    51     7.41    94.44
  0.011   |*****                                       1    52     1.85    96.30
  0.013   |*****                                       1    53     1.85    98.15
  0.015   |                                            0    53     0.00    98.15
  0.017   |*****                                       1    54     1.85   100.00
  0.019   |                                            0    54     0.00   100.00
          ----+----+----+----+----+----+----+----+
              1    2    3    4    5    6    7    8

                       FREQUENCY
```

Figure 4.3 Graphical Inspection for Residual Normality - 1985

We can use a Chi-Squared goodness-of-fit test to further support the hypothesis of a normal distribution. The null hypothesis is H_0 : The residuals are distributed Normal (0, σ^2). The results of the Chi-Squared test for each year are as follows:

1982 - α (actual) = .262

1983 - α (actual) = .580

1984 - α (actual) = .527

1985 - α (actual) = .319

Since these values of α (actual) are greater than α (critical) = 0.1, we fail to reject the null hypothesis that the residuals for each year group are normally distributed.

To summarize the progress on the planning and developing of the GSM I-IIIA model to this point, the following steps have been accomplished.

1) First regression run. Basic variables present.

2) Data separated into time groups to nullify effects of possible autocorrelation.
3) Subsequent regressions to reduce the effects of multicollinearity.
4) Subsequent regressions to determine significant variables per time group.
5) Subsequent regressions to determine final 'best separate equation' per time group.
6) Check for leverage from insignificant outliers per time group.
7) Plots of residuals. Visual check in each time group for heteroscedasticy.
8) Check for normality in each time group using charts and statistical tests.

It is now time to repool the data back into its original longitudinal structure. The data set has the same basic structure as in Table 1 on page 12, except that now we will be working with only the four independent variables that were found to be significant in the cross-sectional analysis.

Another regression is performed using these four variables. An overall R^2 value of 0.7171 is obtained. As expected, each of the variables in the individual year groups is significant in the overall regression using both the t-Test and the stepwise F-Test. Again, multicollinearity is not a problem as the Condition Index and Variance Inflation Factors are well below the model goals. It is now time to check the residuals of this overall regression for any signs of autocorrelation.

I. CHECKING FOR AUTOCORRELATION

Autocorrelation is a problem that sometimes arises with time series data. Positive autocorrelation tends to underestimate the standard error of the estimated coefficients and could lead to an indication of significance (i.e., slope not = 0) when actually the coefficients are not significant.

Once the data is restructured and the regression is accomplished, one of the first indicators for autocorrelation is for the residuals (in the overall regression) to become non-normal. In our particular model, we will now be checking a total of 216 residuals (54 battalions x 4 years) for normality. This is quite a large sample size to be trying to determine a goodness-of-fit for any known distribution. If the statistical indicators come out to confirm a normal distribution, it would be a very good sign. If not , it could be due to the sample size or it could be the fact that the residuals are carrying certain biasing information concerning autocorrelation. There are several methods to check for autocorrelation which will be covered in this section.

The results of a Chi-Square goodness-of-fit test for the re-pooled residuals indicate an α (actual) equal to .055. This is less than α (critical) so we fail to accept the hypothesis that the residuals are distributed normally. This is the first bad sign.

One very quick way of checking for autocorrelation is to look at the residual plots in Table 18. These plots are given by SAS when the request of R is indicated in the option section of the MODEL statement. These are actually plots of the studentized residuals (similar to those presented in Table 14) of the overall regression. A residual that is within 0.5 standard deviations of the mean is left blank; between 0.5 and 1.0 standard deviations gets a single *; between 1.0 and 1.5 gets **; and so forth. When checking for autocorrelation, we look for *patterns* in these residuals. A graphical example of this is given in Table 18. The GOOD is a hypothetical example that is presented for illustrative purposes. The BAD are selected segments of actual results from our newly (repooled) postulated model. Notice that there is a distinctive pattern of a definitive series of positive or negative residuals in the actual (BAD) results. What we are looking for is something similar to the GOOD results where there is a seemingly random shift between the positively and negatively plotted residuals.

TABLE 18

PLOTS OF STUDENTIZED RESIDUALS FROM SAS

```
|--- THE GOOD ----|      |---------------- THE BAD ----------------|
BN    -2-1-0 1 2         BN    -2-1-0 1 2        BN    -2-1-0 1 2
UU  |     *|       |     3C  |    **|       |    6A  |     *|       |
UU  |      |*      |     3C  |      |       |    6A  |      |       |
UU  |      |       |     3C  |     *|       |    6A  |      |       |
UU  |      |*      |     3C  |    **|       |    6A  |    **|       |
VV  |      |*      |     3D  |      |       |    6F  |    **|       |
VV  |      |*      |     3D  |      |*****  |    6F  |      |       |
VV  |     *|       |     3D  |      |       |    6F  |      |       |
VV  |      |       |     3D  |      |       |    6F  |    **|       |
WW  |      |       |     3E  |     *|       |    6G  |      |*      |
WW  |     *|       |     3E  |      |****   |    6G  |      |****   |
WW  |      |*      |     3E  |      |       |    6G  |      |***    |
WW  |      |       |     3E  |      |**     |    6G  |      |***    |
XX  |     *|       |     3F  |      |       |    6H  |     *|       |
XX  |     *|       |     3F  |      |****   |    6H  |      |*      |
XX  |      |*      |     3F  |      |*      |    6H  |    **|       |
XX  |     *|       |     3F  |      |*      |    6H  |   ***|       |
YY  |      |***    |     3G  |      |       |    6I  |    **|       |
YY  |     *|       |     3G  |      |***    |    6I  |      |       |
YY  |      |*      |     3G  |      |       |    6I  |      |       |
YY  |      |       |     3G  |      |**     |    6I  |    **|       |
ZZ  |     *|       |     3H  |      |       |    6J  |  ****|       |
ZZ  |      |**     |     3H  |      |****   |    6J  |     *|       |
ZZ  |      |*      |     3H  |      |*      |    6J  |    **|       |
ZZ  |     *|       |     3H  |      |**     |    6J  |   ***|       |
```

Another graphical method is provided by SAS and is shown in Figure 4.4. The PROC PLOT procedure is again used. This time we will plot the residuals from one year versus the residuals of the previous year. The idea is that if autocorrelation is *not* present, then the only discernable pattern should be a cloud of residual plots centered around the (0,0) coordinate. Otherwise, we can assume that the two plotted residuals are pairwise correlated and therefore not independent. Figure 4.4 does not look very promising. The fact that many negative residuals are being plotted against other negative residuals, and many positive residuals are being plotted against other positive residuals indicates that positive correlation is very probable (negative correlation would have been centered on the complimentary northwest to southeast axis).

One should seldom rely on graphical methods alone, however. Another test that is easy to perform is the runs test. It is a simple non-parametric test based on probability theory. Reference is made to Figure 4.5.

Our data is structured over a four year time period. If we place the residuals for each battalion in a row over this four year period and *if these residuals are independent and randomly distributed* we would expect them to fall in a distribution that is similar to the distribution that is depicted at the bottom of Figure 4.5. In Figure 4.5, if we have four columns of residuals (where each column equates to a year) and each residual can be either positive (+) or negative (-), then probability theory indicates that there are 16 different ways (2^4 combinations) that these four columns of positive and negative residuals can be arranged. By looking at the actual arrangement versus the theoretical arrangement, we compare to see if there is independence or non-independence. Independence is indicated if the distributions are statistically identical. Too *few* runs (a run being defined as a string of positive or negative residuals) indicates a positive autocorrelation between the year groups. This means that the variables in one time period will be high if the variables in the previous time period were high and low if the previous time period were low. Too *many* runs indicate that there is a negative correlation and that one year's highs will cause the next year's to be low, and vice versa.

In looking at Table 19, our overall analysis of the regression residuals indicate too few runs. This signifies positive correlation. By inspection, the actual cumulative probability distribution in Table 19 is not identical to the theoretical cumulative distribution in Figure 4.5, therefore the residuals are not independent and autocorrelation is possible. This supports our observations from the BAD.

One final check could be the Durbin-Watson Test. It is the most popular of the autocorrelation tests. The Durbin-Watson test is a test which postulates a hypothesis that there is no correlation in the residuals (H_0: $\rho = 0$ between adjoining periods).

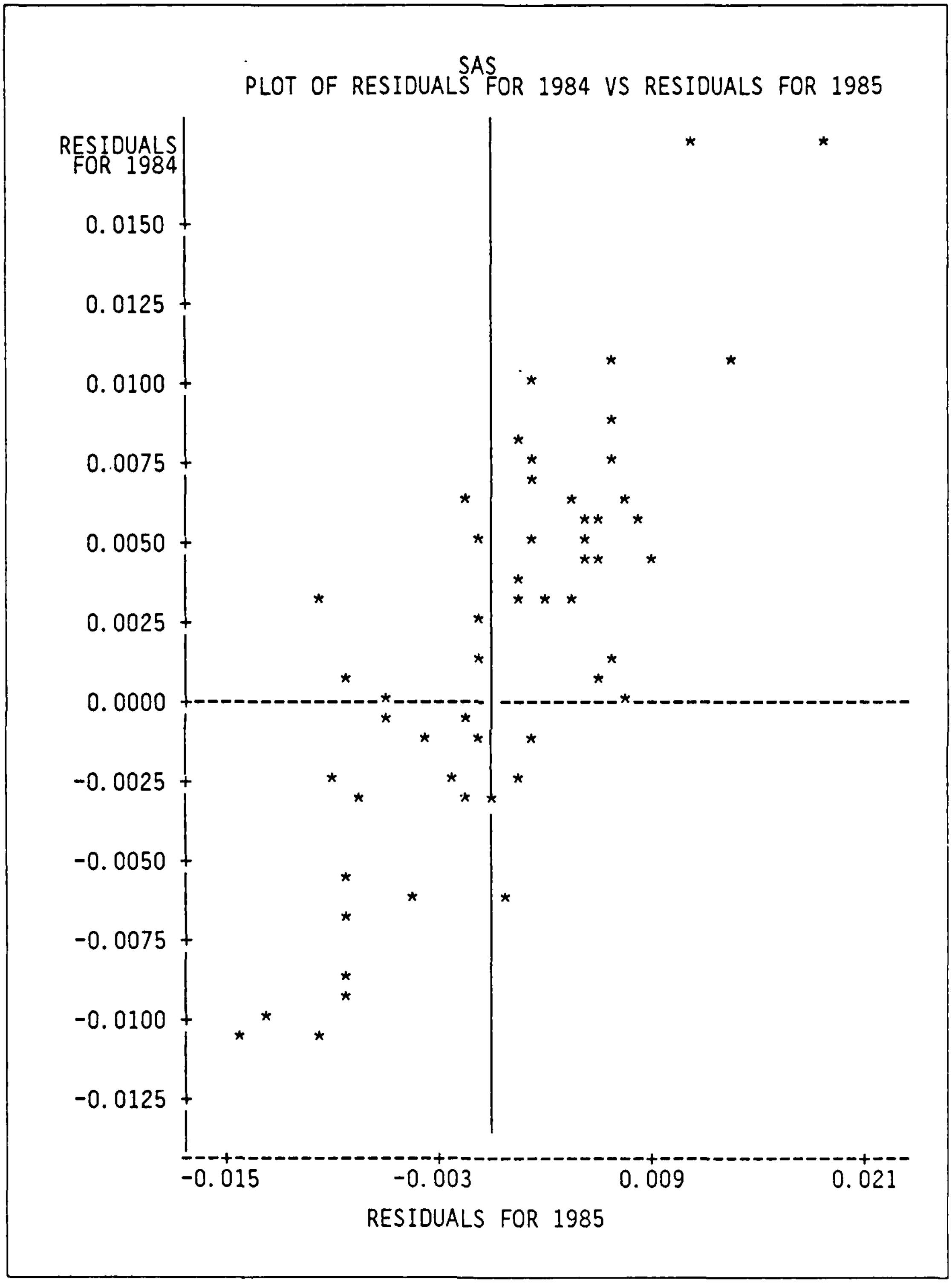


Figure 4.4 Plot of Lag-One Residuals for 1984 vs 1985 from SAS

POSSIBLE COMBINATIONS = 2^4 = 16

	(**** SIGNS ***)				RUNS
1)	+	+	+	+	1
2)	+	+	+	-	2
3)	+	+	-	-	2
4)	+	+	-	+	3
5)	+	-	+	+	3
6)	+	-	-	+	3
7)	+	-	-	-	2
8)	+	-	+	-	4
9)	-	-	-	-	1
10)	-	-	-	+	2
11)	-	-	+	+	2
12)	-	-	+	-	3
13)	-	+	-	-	3
14)	-	+	+	-	3
15)	-	+	+	+	2
16)	-	+	-	+	4

RUNS	1	2	3	4
FREQUENCY	2	6	6	2
PROBABILITY	.125	.375	.375	.125
CUMULATIVE PROBABILITY	.125	.50	.875	1.00

Figure 4.5 Theoretical Distribution for Runs Test

SAS has an option (DW) which will calculate a Durbin-Watson statistic. If the underlying data base was purely time-series in structure, then this option would be ideal. The underlying data base for this regression, however, is longitudinal. Furthermore, the time span of the serial portion of the data is only four years. This is not enough units of sample size in order to do a Durbin-Watson Test with any degree of accuracy.

J. TRANSFORMATION OF THE VARIABLES

All of the graphical and statistical techniques that we have employed indicate autocorrelation. This implies that a transformation of the data is is required. The idea behind the transformation that we will use is to *subtract out* the effects of the previous year's correlation from the present year's data, and use this resultant *transformed* data for building the finalized regression model. First, a determination of the actual correlation is required. The calculation of the true (actual) correlation coefficient, ρ_a,

TABLE 19

RUNS TEST RESULTS FOR OVERALL REGRESSION

BN	R82	R83	R84	R85	(*** SIGNS ***)				RUNS
1A	-0.001241	0.008685	-0.008464	0.002043	-	+	-	+	4
1B	-0.011716	0.004273	-0.002061	-0.001031	-	+	-	-	3
1C	-0.007402	0.007605	0.006852	0.001432	-	+	+	+	2
1D	-0.000309	0.003754	-0.004902	-0.000976	-	+	-	-	3
1E	-0.010784	-0.003933	-0.005845	-0.001013	-	-	-	-	1
1F	-0.001267	0.002864	-0.002646	-0.005989	-	+	-	-	3
1G	0.005885	0.013278	0.003000	-0.001671	+	+	+	-	2
1H	-0.010284	0.001426	-0.000047	-0.001306	-	+	-	-	3
1I	-0.005516	-0.002165	-0.008134	-0.005148	-	-	-	-	1
1K	-0.000963	0.003649	0.005655	0.003974	-	+	+	+	2
1L	-0.008081	0.004893	-0.006942	-0.007875	-	+	-	-	3
1N	-0.010382	0.002084	-0.014450	-0.009868	-	+	-	-	3
3A	-0.015223	0.001774	-0.012066	-0.009367	-	+	-	-	3
3B	-0.002737	0.024288	0.003528	0.007530	-	+	+	+	2
3C	-0.010129	0.000804	-0.004841	-0.009029	-	+	-	-	3
3D	0.001749	0.022454	-0.000884	0.000942	+	+	-	+	3
3E	-0.007807	0.019081	0.000352	0.009787	-	+	+	+	2
3F	0.002805	0.020199	0.005236	0.007384	+	+	+	+	1
3G	0.004176	0.017198	-0.001996	0.008889	+	+	-	+	3
3H	-0.000892	0.018997	0.008046	0.012272	-	+	+	+	2
3I	-0.010070	0.011057	0.000395	0.003383	-	+	+	+	2
3J	-0.017311	0.017266	-0.003279	-0.004264	-	+	-	-	3
3K	-0.006674	0.023623	0.002226	-0.002977	-	+	+	-	3
4A	-0.008884	0.010292	0.003525	0.008390	-	+	+	+	2
4C	-0.010280	-0.005494	-0.008715	-0.011747	-	-	-	-	1
4D	-0.004799	0.008260	0.003489	0.007306	-	+	+	+	2
4E	-0.013210	-0.002234	-0.011172	-0.009604	-	-	-	-	1
4F	-0.006490	0.015066	0.001456	0.003886	-	+	+	+	2
4G	-0.004083	0.007727	0.001667	0.004095	-	+	+	+	2
4H	-0.012256	0.002925	-0.003732	-0.000890	-	+	-	-	3
4I	-0.016221	0.001357	-0.005137	0.001106	-	+	-	+	4
4J	-0.008587	0.001020	-0.003359	0.001753	-	+	-	+	4
4K	-0.009324	0.004253	0.002135	0.006641	-	+	+	+	2
5A	0.009027	0.018026	0.003143	0.003796	+	+	+	+	1
5B	0.009407	0.019036	0.014146	0.009152	+	+	+	+	1
5C	-0.004602	0.013281	-0.000192	0.004489	-	+	-	+	4
5D	-0.007355	0.003645	-0.005200	-0.001307	-	+	-	-	3
5E	-0.001404	0.008321	0.002072	0.002669	-	+	+	+	2
5F	-0.010517	-0.001710	-0.001200	0.006828	-	-	-	+	2
5H	-0.003820	0.010553	0.003917	0.002602	-	+	+	+	2
5I	-0.015058	0.000153	-0.003294	0.004511	-	+	-	+	4
5J	-0.009090	0.004122	-0.002474	0.003986	-	+	-	+	4
5K	-0.014821	0.00959	0.002669	0.004858	-	+	+	+	2
5L	-0.008416	0.003143	-0.003995	0.002807	-	+	-	+	4
5M	-0.013881	0.003025	0.002460	0.005870	-	+	+	+	2
5N	-0.003045	0.015589	0.004272	-0.001644	-	+	+	-	3
6A	-0.007684	-0.002056	-0.009474	-0.009681	-	-	-	-	1
6F	-0.009216	-0.000314	-0.003385	-0.009416	-	-	-	-	1
6G	0.005392	0.023753	0.015176	0.016055	+	+	+	+	1
6H	-0.005602	0.005791	-0.012722	-0.015993	-	+	-	-	3
6I	-0.009914	0.003006	-0.000451	-0.011823	-	+	-	-	3
6J	-0.019204	-0.006437	-0.012900	-0.015844	-	-	-	-	1
6K	-0.012839	-0.001868	-0.003404	-0.009066	-	-	-	-	1
6L	0.000567	0.013283	0.007282	0.005887	+	+	+	+	1

RUNS	1	2	3	4	
FREQUENCY	13	17	17	7	(TOTAL = 54)
PROBABILITY	.24	.31	.31	.13	
CUMULATIVE PROBABILITY	.24	.55	.86	1.00	

for the data base for our overall regression is according to the following formula. [Ref. 8:p. 510]

$$\rho_a = \frac{\varepsilon_{1A,85} x \varepsilon_{1A,84} + \varepsilon_{1A,84} x \varepsilon_{1A,83} + \varepsilon_{1A,83} x \varepsilon_{1A,82} + \varepsilon_{1B,85} x \varepsilon_{1B,84} + \ldots\ldots}{\varepsilon_{1A,84}^{2} + \varepsilon_{1A,83}^{2} + \varepsilon_{1A,82}^{2} + \varepsilon_{1B,84}^{2} + \ldots\ldots}$$

Substituting the residuals from the regression (Table 19), this implies that the true correlation coefficient for this overall regression is

$$\rho_a = \frac{(.002043)(-.008464) + (-.008464)(.008645) + (.008645)(-.001214) + (-.001031)(-.002061) + \ldots\ldots}{(-.008464)^2 + (.008685)^2 + (-.001241)^2 + (-.001031)^2 + \ldots\ldots}$$

$$= .175482$$

A positive value for ρ_a is consistent with all of the other indications of correlation.

For the first data line (BN 1A, 1982) the transformation of the independent and dependent variables are according to the following formulas. [Ref. 8:p. 510]

$$x^*_{i,1} = (1-\rho_a{}^2)^{1/2}\ x_{i,1}$$

$$y^*_1 = (1-\rho_a{}^2)^{1/2}\ y_1 \tag{4.1}$$

where i = PROP,BNPER,RCTR,PERCWI

For the last 215 data lines, the following equations are utilized.

$$x^*_{i,j} = x_{i,j} - \rho_a\, x_{i,j-1}$$

$$y^*_j = y_j - \rho_a\, y_{j-1} \tag{4.2}$$

where i = PROP,BNPER,RCTR,PERCWI
j = 2,3,...,216

Again, these transformations are to nullify the effect of previous year correlation on the next year's data.

Although we assume (and there is in fact) independence (and therefore no correlation) between one battalion in 1985 and another battalion in 1982, the data structure dictates that a transformation between these two variables is warranted. For instance, there is no correlation between battalion 1A in 1985 and battalion 1B in 1982. However, equations 4.2 dictate that

$$x^{*}_{i,1B,1982} = x_{i,1B,1982} - \rho_a x_{i,1A,1985}$$

and

$$y^{*}_{1B,1982} = y_{1B,1982} - \rho_a y_{1A,1985}$$

where i = PROP,BNPER,RCTR,PERCWI

After transforming all of the variables in the data base of the final model, we arrive with the Development Phase finalized matrix of longitudinal data. It appears as below.

$$Y = \begin{array}{c} \text{PENT} \\ \left[\begin{array}{c} 0.406 \\ 0.049 \\ : \\ : \\ 0.050 \end{array}\right] \end{array} \qquad X = \begin{array}{c} \begin{array}{ccccc} & \text{PROP} & \text{BNPER} & \text{RCTR} & \text{PERCWI} \end{array} \\ \left[\begin{array}{ccccc} 1 & 14.471 & 0.012 & 52.915 & 0.944 \\ 1 & 12.520 & 0.010 & 42.817 & 0.791 \\ : & : & : & : & : \\ : & : & : & : & : \\ 1 & 6.955 & 0.021 & 78.398 & 0.750 \end{array}\right] \end{array} \qquad \beta = \left[\begin{array}{c} b_0 \\ b_1 \\ : \\ : \\ b_4 \end{array}\right]$$

where Y = 216x 1 matrix (a column vector of the dependent variables)

X = 216x 5 matrix (a column vector of 1's catonated with the 216x 4 matrix of the independent variables)

β = 5 x 1 matrix (a column vector of parameter estimates)

K. INSPECTING THE RESULTS

A regression on these matrices is now performed with the results as displayed in Table 20. The 'T' on the end of the variable names now indicate a transformed variable.

TABLE 20
RESULTS OF REGRESSION ON TRANSFORMED DATA

```
                                              SAS
DEP VARIABLE: PENTTRANS
                         SUM OF              MEAN
SOURCE      DF          SQUARES            SQUARE       F VALUE       PROB>F
MODEL        4         0.030239       0.007559647        99.795       0.0001
ERROR      211         0.015984      0.0000757518
C TOTAL    215         0.046222
  ROOT MSE          0.008703551          R-SQUARE        0.6542
  DEP MEAN             0.042223          ADJ R-SQ        0.6476
  C.V.                 20.61326

                  PARAMETER        STANDARD    T FOR H0:                  VARIANCE
VARIABLE   DF      ESTIMATE           ERROR  PARAMTR=0      PROB>|T|    INFLATION

INTERCEP    1      0.062179     0.007412616      8.388        0.0001     0.000000
PROPT       1   0.001531895    0.0001770035      8.655        0.0001     1.408339
BNPERT      1      2.685632        0.203836     13.175        0.0001     2.790335
RCTRT       1  -0.000537034    .00004744603    -11.319        0.0001     3.190078
PERCWIT     1     -0.056823     0.007936443     -7.160        0.0001     1.363656

                                 CORRELATION OF ESTIMATES
CORRB         INTERCEP            PROPT        BNPERT          RCTRT        PERCWIT

INTERCEP        1.0000          -0.6785        0.3278        -0.6033        -0.8795
PROPT          -0.6785           1.0000       -0.2127         0.4426         0.3689
BNPERT          0.3278          -0.2127        1.0000        -0.7775        -0.4271
RCTRT          -0.6033           0.4426       -0.7775         1.0000         0.4452
PERCWIT        -0.8795           0.3689       -0.4271         0.4452         1.0000

COLLINEARITY DIAGNOSTICS            VARIANCE PROPORTIONS

                        CONDITION    PORTION    PORTION    PORTION    PORTION    PORTION
NUMB   EIGENVALUE           INDEX   INTERCEP      PROPT     BNPERT      RCTRT    PERCWIT

 1          4.763           1.000     0.0003     0.0028     0.0014     0.0011     0.0005
 2       0.165250           5.369     0.0006     0.2364     0.0399     0.0371     0.0003
 3       0.043872          10.419     0.0199     0.3818     0.0849     0.0387     0.1173
 4       0.024188          14.032     0.0125     0.0001     0.6515     0.4867     0.0197
 5       0.003914          34.884     0.9667     0.3789     0.2223     0.4364     0.8621
```

This indicates that our 'best regression' equation is

$$PENTT = 0.062179 + 0.001531\ PROPT + 2.68563\ BNPERT$$
$$- 0.000537\ RCTRT - 0.056823\ PERCWIT + \varepsilon$$

After checking for heteroscedasticy, leverage and then repooling the data with the final four independent variables, we obtained a pre-transformed R^2 value of .7171. The R^2 value of the transformed data is now .6549. This drop is to be expected after reducing the variables via the transformations due to the positive autocorrelation. The final model of the transformed data fulfills all of the preliminary goals as outlined in

Figure 3.1. The positive parameter estimates for PROPT and BNPERT are reassuring. We would expect that the Penetration would increase as the Propensity and Battalion percent of mission increases. The negative signs for RCTRT and PERCWIT , however, are worthy of discussion.

If RCTRT has a negative value, then USAREC is probably experiencing negative returns to scale in the employment of recruiters. These results have been empirically substantiated by previous studies. [Ref. 6] This finding was not apparent in the initial regressions when CONT was the dependent variable. Obviously, more recruiters bring in more contracts. With PENETRATION as the dependent variable, however, the slope of the regression plane through the RCTR dimension in the carrier hyperspace is negative, indicating negative returns to scale in the market penetration.

The negative slope for PERCWIT is a little more difficult to explain. It must be remembered that this variable was always the least significant of the four significant variables in the stepwise regressions (it was always brought in last). Again, it is very possible that its parameter estimate is being heavily influenced by the costock of variables. Furthermore, its absolute magnitude is relatively high. In checking with Appendix B, the maximum value of PERCWI is .99. A maximum PERCWIT input value of $x^* = .816272$ would decrease PENTT by a total of .046830 (β x PERCWIT = -.056823 x .816272 = 0.046830). The maximum PERCWIT input value would be derived by a battalion with a 99% white population that is transformed. This is calculated as

$$x^* = .99 - (.175482 \times .99) = .816272$$
$$\text{where} \quad \rho = .175482$$

A total decrease in Penetration of .046830 is significant when one considers that the average value of Penetration is .051157. This further supports the theory that the parameter estimate for PERCWIT is highly influenced by its costock.

After satisfying ourselves that the 'best equation' has been obtained to this point, it is now time to move into the Validation and Maintenance Phase of the GSM I-IIIA model.

V. VALIDATION AND MAINTENANCE OF THE GMA-I-IIIA MODEL

In this section we will discuss a few techniques for verifying and updating the GSM I-IIIA model. It may be useful to review Figure 2.3 of Chapter 2 at this time.

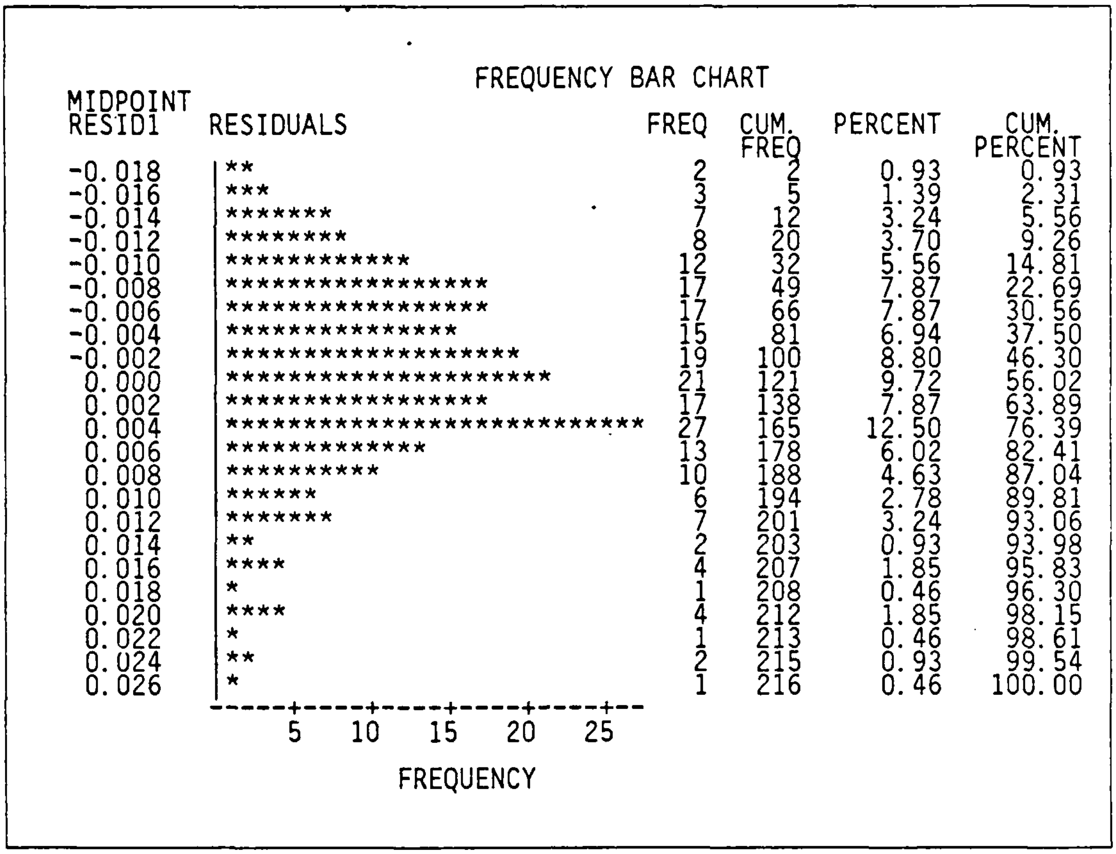
A. CHECK FOR SYSTEMATIC LACK OF FIT

Much work has been accomplished towards the development of this model. Many checks and balances have been performed along the way for compliance with the application of the theory of multivariable regression analysis. As was indicated in Table 20, we have achieved a final R^2 value of .6542 for the transformed data model. A few final checks need to be performed to ensure that there is no lingering systematic lack of fit.

First of all, a plot of the residuals to check for normality is shown in Figure 5.1. A normal, symmetric distribution seems to be indicated. A Chi-Squared goodness of fit test is performed on these residuals. The hypothesis is H_0: the residuals are normally distributed. The level of significance of this test is α (actual) = .4003. Since α (actual) > α (critical), and since the graphical representation indicates no apparent problems, we fail to reject the null hypothesis that the residuals are normally distributed.

Secondly, we need to ensure that the transformation that was applied using equations 4.1 and 4.2 on page 60 is effective in nullifying the effects of autocorrelation.

Longitudinal data presents special problems due to its structure. Autocorrelation is a almost always a *time series problem*, and we have a mixture of cross sectional and time series data. The runs test is especially applicable to this type of data structure. A runs test was performed on the residuals from the transformed data and the results are as appears in Table 21. Comparing Table 21 with Table 19 indicates that there is much less of a problem now with too few runs. In fact, the middle distributions of two and three runs has shifted dramatically toward the three runs side. A distribution like this indicates possible negative correlation. This would really be considered a weak indication, however, because the skewness of the distribution in Table 21 is weighted more in the center than in the tails. A better indicator might be a check of the final calculation of ρ_a.



```
                              FREQUENCY BAR CHART
MIDPOINT
RESID1     RESIDUALS                                FREQ   CUM.    PERCENT     CUM.
                                                           FREQ              PERCENT
-0.018    |**                                          2      2       0.93       0.93
-0.016    |***                                         3      5       1.39       2.31
-0.014    |*******                                     7     12       3.24       5.56
-0.012    |********                                    8     20       3.70       9.26
-0.010    |************                               12     32       5.56      14.81
-0.008    |*****************                          17     49       7.87      22.69
-0.006    |*****************                          17     66       7.87      30.56
-0.004    |***************                            15     81       6.94      37.50
-0.002    |*******************                        19    100       8.80      46.30
 0.000    |*********************                      21    121       9.72      56.02
 0.002    |*****************                          17    138       7.87      63.89
 0.004    |***************************                27    165      12.50      76.39
 0.006    |*************                              13    178       6.02      82.41
 0.008    |**********                                 10    188       4.63      87.04
 0.010    |******                                      6    194       2.78      89.81
 0.012    |*******                                     7    201       3.24      93.06
 0.014    |**                                          2    203       0.93      93.98
 0.016    |****                                        4    207       1.85      95.83
 0.018    |*                                           1    208       0.46      96.30
 0.020    |****                                        4    212       1.85      98.15
 0.022    |*                                           1    213       0.46      98.61
 0.024    |**                                          2    215       0.93      99.54
 0.026    |*                                           1    216       0.46     100.00
           -----+----+----+----+----+--
                5   10   15   20   25

                     FREQUENCY
```

Figure 5.1 Graphical Inspection for Residual Normality - Transformed Data

Calculating ρ_a in the exact same manner as before, we derive a value of ρ_a = -0.0335. The negative sign confirms our suspicions of possible negative correlation, but, by inspection, the *magnitude* of ρ_a indicates that autocorrelation has been removed from the model.

Since there is no suggestion of systematic lack of fit in the model, we can assume that the statistical tests that were utilized to derive the parameter estimates were valid. Now it is time to check these parameter estimates.

B. MODEL RANGES AND VALIDATION

There are several methods which can be employed to validate our model equation. As stated in Chapter 4, the equation is of the following form.

$$\text{PENTT} = 0.062179 + 0.001531\ \text{PROPT} + 2.68563\ \text{BNPERT} - 0.000537\ \text{RCTRT} - 0.056823\ \text{PERCWIT} + \varepsilon$$

TABLE 21

RUNS TEST RESULTS FOR TRANSFORMED DATA REGRESSION

1A	0.010457	0.008586	-0.010425	0.003279	+	+	-	+	3
1B	-0.011753	0.006372	-0.002550	-0.000714	-	+	-	-	3
1C	-0.007723	0.008380	0.004769	-0.000287	-	+	+	-	3
1D	-0.000885	0.003093	-0.006209	-0.000612	-	+	-	-	3
1E	-0.010802	-0.002197	-0.005321	-0.000323	-	-	-	-	1
1F	-0.001459	0.002551	-0.003726	-0.005976	-	+	-	-	3
1G	0.006778	0.012082	0.000716	-0.002583	+	+	+	-	2
1H	-0.009648	0.003526	-0.000262	-0.001198	-	+	-	-	3
1I	-0.005554	-0.001443	-0.008009	-0.003968	-	-	-	-	1
1K	-0.000175	0.003744	0.004849	0.003003	-	+	+	+	2
1L	-0.009204	0.005889	-0.007881	-0.006555	-	+	-	-	3
1N	-0.008958	0.004106	-0.014852	-0.007569	-	+	-	-	3
3A	-0.012683	0.005024	-0.012097	-0.006631	-	+	-	-	3
3B	-0.001632	0.024406	-0.001163	0.006198	-	+	-	+	4
3C	-0.011272	0.002459	-0.004985	-0.008590	-	+	-	-	3
3D	0.003759	0.022579	-0.004615	0.001265	+	+	-	+	3
3E	-0.007486	0.020729	-0.002826	0.009715	-	+	-	+	4
3F	0.000833	0.019648	0.001369	0.006114	+	+	+	+	1
3G	0.003206	0.016546	-0.005273	0.009253	+	+	-	+	3
3H	-0.001774	0.019725	0.005222	0.011323	-	+	+	+	2
3I	-0.012090	0.012593	-0.001590	0.003293	-	+	-	+	4
3J	-0.017347	0.020900	-0.006018	-0.003382	-	+	-	-	3
3K	-0.005122	0.025383	-0.001271	-0.002976	-	+	-	-	3
4A	-0.008369	0.011944	0.001900	0.007899	-	+	+	+	2
4C	-0.012380	-0.003886	-0.007995	-0.010329	-	-	-	-	1
4D	-0.002797	0.009086	0.002063	0.007047	-	+	+	+	2
4E	-0.014781	-0.000016	-0.010916	-0.007581	-	-	-	-	1
4F	-0.004567	0.016403	-0.001033	0.003902	-	+	-	+	4
4G	-0.005086	0.008043	0.000023	0.003563	-	+	+	+	2
4H	-0.013032	0.004767	-0.004349	-0.000339	-	+	-	-	3
4I	-0.016027	0.004291	-0.005444	0.002247	-	+	-	+	4
4J	-0.009373	0.002289	-0.003637	0.002267	-	+	-	+	4
4K	-0.010106	0.005358	0.001300	0.006201	-	+	+	+	2
5A	0.007874	0.016325	-0.000111	0.002726	+	+	-	+	3
5B	0.008970	0.017374	0.010726	0.006338	+	+	+	+	1
5C	-0.005943	0.014285	-0.002616	0.004309	-	+	-	+	4
5D	-0.008178	0.004904	-0.006028	-0.000690	-	+	-	-	3
5E	-0.001323	0.008089	0.000236	0.001816	-	+	+	+	2
5F	-0.010071	0.000757	-0.000846	0.007033	-	+	-	+	4
5H	-0.005214	0.010879	0.001381	0.001187	-	+	+	+	2
5I	-0.015174	0.003128	-0.003256	0.004723	-	+	-	+	4
5J	-0.010100	0.005254	-0.003589	0.003879	-	+	-	+	4
5K	-0.015408	0.011757	0.000451	0.003992	-	+	+	+	2
5L	-0.009754	0.004019	-0.005002	0.003235	-	+	-	+	4
5M	-0.014262	0.005128	0.001499	0.004761	-	+	+	+	2
5N	-0.004140	0.015539	0.000897	-0.002710	-	+	+	-	3
6A	-0.006741	-0.000386	-0.008856	-0.007702	-	-	-	-	1
6F	-0.007629	0.001181	-0.003418	-0.008621	-	+	-	-	3
6G	0.007668	0.023303	0.011498	0.013911	+	+	+	+	1
6H	-0.008195	0.006858	-0.013745	-0.013408	-	+	-	-	3
6I	-0.006979	0.004854	-0.001002	-0.011557	-	+	-	-	3
6J	-0.017226	-0.003314	-0.012041	-0.013655	-	-	-	-	1
6K	-0.010154	0.000221	-0.003127	-0.008226	-	+	-	-	3
6L	0.002192	0.012931	0.004692	0.004992	+	+	+	+	1

RUNS	1	2	3	4	
FREQUENCY	10	11	22	11	(TOTAL = 54)
PROBABILITY	.19	.20	.41	.20	
CUMULATIVE PROBABILITY	.19	.39	.80	1.00	

One of the quickest and easiest methods is to check the equation at the midpoint and at the extremes of the data ranges. By inserting the mean values of the independent variables on the right hand side of the above equation, we would expect the resultant equality to be equal to the mean value of Penetration. This is because, by definition, $\overline{Y} = \beta \overline{X}$. Another check is to look at the minimum and maximum values of the dependent variable. First we choose the battalion with the lowest value of Penetration. Then we insert into the equation the data that corresponds to this minimum value. We would expect that the resultant value of PENTT from this equation would be moving away from the mean and towards the minimum value of Penetration. The same logic also applies for the maximum value of Penetration.

Appendix B provides all of the relevant data that is required to initiate these tests. Appendix B also contains the data ranges for which this model is valid. Regression theory dictates that the regression equation is relatively reliable near the means of the inputted data ranges. At the extremes it is much less accurate. *For any inputted data values outside of the data range, the model can be considered to have no predictive value.* From Appendix B, the means of the data ranges are as follows:

DATA AT THE MEAN = >

	PENT	PROP	BNPER	RCTR	PERCWI
(MEAN)	0.05115	14.48	0.0183	88.69	0.8429

Taking the minimum and maximum values of PENT from Appendix B, we search the data base to find the corresponding input variables for these values. The minimum Penetration over the four year time span was obtained by battalion 6J in 1982. The maximum Penetration was by battalion 3D in 1983. The variables for these two extreme values of Penetration are as follows:

DATA AT THE MIN = >

	PENT	PROP	BNPER	RCTR	PERCWI
(6J;1982)	0.01967	8.0	0.0136	59.0	0.9431

DATA AT THE MAX = >

	PENT	PROP	BNPER	RCTR	PERCWI
(3D;1983)	0.10396	23.9	0.0163	68.25	0.6676

Applying the transform where $x^* = x - (\rho_a)\, x$ for all of the above variables (where $\rho_a = .175482$), the following transformed variables are derived.

	PENT	PROP	BNPER	RCTR	PERCWI
(MEAN)	0.04218	11.93	0.0150	73.12	0.6950
(6J;1982)	0.01622	6.59	0.0112	48.63	0.7776
(3D;1983)	0.08572	19.71	0.0134	56.27	0.5504

Inserting the values of the independent variables in the regression equation supplies the following results.

TEST AT THE MEAN =>

$$0.062179 + 0.001531\,(11.93) + 2.68563\,(.0150) - 0.000537\,(73.12) - 0.056823\,(.6950) = 0.04221$$

TEST AT THE MIN =>

$$0.062179 + 0.001531\,(6.59) + 2.68563\,(0.0112) - 0.000537\,(48.63) - 0.05682\,(.7776) = 0.03204$$

TEST AT THE MAX =>

$$0.062179 + 0.001531\,(19.71) + 2.68563\,(.0134) - 0.000537\,(56.27) - 0.056823\,(.5404) = 0.06741$$

As can be readily seen, the test at the mean provides an estimate of the dependent variable (0.04221) that is extremely close to the mean value of the transformed dependent variable (0.04218). The discrepancy is due purely to roundoff error. At the extremes, the magnitude is not nearly so close. This lack of accuracy is not, however, unexpected. At the extremes, we are satisfied that the equations provide predictions that are in the correct *direction*.

C. USING THE REGRESSION EQUATIONS

Once that we are satisfied that the regression equations are behaving correctly, we can begin to utilize the model as a tool for predicting GSM I-IIIA contracts.

As was previously stated in this thesis, one of the primary objectives is to minimize the number of input variables in the model. For every independent variable that is included in the model, the analyst must devise some scheme to *predict* that input variable. It does not matter how close of a fit one can achieve with a predicting regression model. The results can only be as accurate as the inputted data.

Ways of predicting the independent variables for this particular model could be the subject for several more theses. The desired complexity is left totally to the discretion of the analyst.

Some variables, such as RCTR and PERCWI are relatively stable and fairly predictable. Predicting experienced recruiters for a future year may merely entail looking at unit manning rosters. The use of the prior year estimate for PERCWI might be the most logical choice for the next year's prediction.

The variable BNPER is relatively stable for some battalions, but suffers a wide variance in others. Again, unless the analyst has some reason to feel otherwise, possibly using the previous year's data for next year's prediction might be the most reasonable choice.

Propensity is the most significant variable in the regression equation. We would like to be as accurate as possible in the prediction of this variable. The variance for this variable has been dissipated due to the fact that we are using a four year moving average. Propensity may be particularly attractive to more complex regression techniques since it is a 'catch all' type variable and may be partially explained by several other controllable variables.

There are numerous methods that an analyst can utilize to predict future year carrier variables. For illustrative purposes, this study will make a few simple assumptions for a 1986 data base and apply the proper methods of applying the regression equation. If the analyst wishes to predict the propensity for any one particular battalion, he should follow the same methodology that was utilized in testing the minimum and maximum values. That is, merely estimate the values of the independent variables for the battalion under consideration, transform and insert these values into the regression equation. If the analyst wishes to predict contracts for the entire Army, he must estimate values for the entire data base. A simple example of this procedure is provided.

The following assumptions will be utilized to determine the 1986 data base for the GSM I-IIIA model. These assumptions are merely hypothetical and are not based on any factual data or observations.

1) PROP - Assume a 2% across the board drop in propensity from 1985 levels for every battalion.

2) BNPER - Due to changing economic conditions, allocate an increase of 0.02 % to each battalion in the 5th Brigade (except 4A and 4C)

and a 0.02 % decrease in each battalion in the 6th Brigade from 1985 levels.

3) RCTR - Assume a net gain of two recruiters per battalion over 1985 recruiter endstrengths.

4) PERCWI- Assume the same white percentage population as in 1985.

The data base for 1986, under the above assumptions, would be structured as shown below. A comparison with Table 1 on page 12 displays the differences between the 1985 data base and this assumed 1986 data base.

BN	PROP86	BNPER86	RCTR86	PERCWI86
1A	13.7	0.0129422	52.00	0.959761
1B	15.4	0.0287982	152.50	0.727236
:	:	:	:	:
:	:	:	:	:
6L	6.5	0.0247549	99.00	0.909940

After applying the necessary transformations as specified in equations 4.1 and 4.2 on page 60, the finalized matrices for the assumed 1986 data base are as shown below.

$$Y^*_{86} = \begin{bmatrix} y^*_{1A} \\ y^*_{1B} \\ \vdots \\ \vdots \\ y^*_{6L} \end{bmatrix} \quad X^*_{86} = \begin{bmatrix} 1 & 13.48 & 0.0127 & 51.193 & 0.944 \\ 1 & 11.29 & 0.0107 & 42.874 & 0.791 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 6.101 & 0.018 & 88.841 & 0.677 \end{bmatrix} \quad \beta = \begin{bmatrix} 0.062179 \\ 0.001531 \\ 2.68563 \\ -0.000537 \\ -0.056823 \end{bmatrix}$$

where Y^*_{86} = 54 x 1 matrix (a column vector of the dependent variables)

X^*_{86} = 54 x 5 matrix (a column vector of 1's catonated with the 54 x 4 matrix of the independent variables)

β = 5 x 1 matrix (a column vector of parameter estimates)

Multiplying the X matrix times the β matrix will result in a 54 x 1 matrix of the *transformed* y values (PENTT). This matrix represents the model's predictions for transformed penetration in each battalion in 1986. In order to solve for total contracts, we need to 'untransform' the y values and multiply the resultant matrix times the estimated number of HSSMA for each battalion. Since we transformed the data by $y^*_t = (y_t) - \rho\,(y_{t-1})$, we 'untransform' using the following equation.

$$y_t = \rho\,(y_{t-1}) + X^*_t\,(\beta) \quad =>$$

$$\begin{aligned} y_{86} &= \rho\,(y_{85}) + X^*_{86}\,(\beta) \\ &= \rho\,(y_{85}) + y^*_{86} \end{aligned}$$

This implies that

$$Y^{86} = .175482 \begin{bmatrix} .0543 \\ .0562 \\ : \\ : \\ .0378 \end{bmatrix} + \begin{bmatrix} .03586 \\ .04010 \\ : \\ : \\ .01967 \end{bmatrix} = \begin{bmatrix} .04543 \\ .05001 \\ : \\ : \\ .04395 \end{bmatrix}$$

Using the USAREC estimates (as of 20 June, 1986) for the number of 1986 High School Male Market Available ($HSSMA_{86}$), the following matrix equations will provide the number of contracts per battalion for each of the 54 battalions represented in the model.

$$Y_{86} \text{ x } HSSMA_{86} = \begin{bmatrix} .04543 \\ .05001 \\ : \\ : \\ .04395 \end{bmatrix} \text{ x } \begin{bmatrix} 12396 & 27547 & . . . & 22784 \end{bmatrix}$$

$$= \begin{bmatrix} 563 & 1377 & .... & 1001 \end{bmatrix}$$

Taking the sum of all of the individual battalion contracts will result in the aggregate number of Army contracts predicted in 1986.

$$\begin{aligned} \text{Total Army Contracts} &= 563 + 1377 + ... + 1001 \\ &= 50{,}132 \end{aligned}$$

Therefore, under the assumptions that we specified for the 1986 data base, total Army GSM I-IIIA contracts for the 54 included battalions in 1986 should equal 50,132. This compares with 50,794 in 1982; 62,781 in 1983; 51,359 in 1984; and 55,098 in 1985.

VI. CONCLUSIONS AND RECOMMENDATIONS

In this thesis, the problem of building a predictive model in order to determine high quality Army enlistment contracts was formulated and solved using stepwise and ordinary least squares linear regression analysis.

The model was developed using a readily available data base and easily obtained variables. It is simple in structure and requires the analyst to predict only a limited number of input variables. All of these aspects contribute towards the desired goal of developing an easy-to-understand and easy-to-update regression model.

This model could be used as a framework for the continued development and refinement of a predictive model to be used by USAREC and DCSPER analysts. There is a need for a 'quick look' predictive tool for getting fast answers to a variety of proposed policy changes. Army analysts at USAREC and DCSPER are trying to upgrade and refine their capabilities in this area.

In concluding this study, a few recommendations are in order. First of all, there needs to be a concerted effort to continually maintain and update the relevant data bases under USAREC control. The mathematical formulations and theories that are used in the technical analysis are useless without an accurate data base. Furthermore, the data maintained by USAREC is highly susceptible to the effects of autocorrelation. In order to efficiently counteract this undesirable side effect, *all* of the data *must* be assimilated in *time specific* intervals. Monthly, quarterly or yearly data bases need to be established. Some conscientious and straightforward method needs to be developed in order to measure or estimate the variables. After this methodology is developed, it needs to be well-documented. A universal understanding of the data by both the on-line analysts and potential external/contractor analytical assistants is essential.

Also, much work could be done towards predicting input variables for this model. Propensity is the most significant variable in this model and there are probably several variables in the data base which affect the propensity of individuals to join the Army. Discovering how income per capita or unemployment rates are reflected in the propensity for service could lead to some insight into the enlistment process.

A more accurate assessment of the behavior of individual battalions could be a worthwhile project. This study models the 'typical' battalion and is useful in interpreting and comparing against the average. A more detailed study of each individual battalion could prove to be fruitful in leading to an understanding of the variances in the cross-sectional behavior over time.

Finally, there needs to be a continued emphasis on the efficient allocation of recruiters. It is the one variable that is most easily controlled by the Army personnel establishment. The negative returns to scale that were discovered in the development phase of this model is somewhat unsettling. In a large and dispersed organization such as USAREC, some negative returns may be unavoidable. This is especially true when mission takes priority over costs. Its existence needs to be recognized, however, and positive control measures need to be implemented, continually assessed, and updated.

APPENDIX A
SELECTED GLOSSARY OF REGRESSION TERMS

Definitions of selected regression terms are presented as follows:

Adjusted R^2 ($R_a{}^2$) - A statistic where an adjustment has been made for the corresponding degrees of freedom of the two quantities, the Residual Sum of Squares (RSS) and the Corrected Total Sum of Squares (CTSS). The idea behind the $R_a{}^2$ is that this statistic can be used to compare equations fit not only to a specific set of data but also to two or more entirely different sets of data. This statistic is usually used only as an initial gross indicator. [Ref. 3:p. 92]

Analysis of Variance (ANOVA) Table - Format for the presentation of key statistics of a regression model. Typically, it is given as follows: [Ref. 3:p. 20]

Source of Variation	Degrees of Freedom (df)	Sum of Squares (SS)	Mean Square (MS)	F Value	Prob > F
Due to the Regression (MODEL)	x	$\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2$	MS_{Reg} (= SS/df)	MS_{Reg}/s^2	
About the Regression (ERROR)	n-x+1	$\sum_{i=1}^{n}(\hat{Y}_i-Y_i)^2$	$s^2 = SS/(n-2)$		
Total, Corrected for the Mean	n-1	$\sum_{i=1}^{n}(Y_i-\bar{Y})^2$			

where Y_i = Y (actual) $\hat{Y}_i$ = Y (predicted) $\bar{Y}$ = Y (average)
n = number of observations x = number of predictor variables

Alpha (α) - α is the level of significance. It is the maximum probability of rejecting a true null hypothesis (H_0). [Ref. 9:p.78]

Autocorrelation - Autocorrelation is a situation, usually found in time series data, in which the impact of a independent variable on the dependent variable is not always completely instantaneous. This implies that there is a a correlation, usually over time. Also known as Serial Correlation. [Ref. 10:p. 289]

Backward Stepwise Elimination Procedure - A procedure that tries to examine only the 'best' regressions containing a certain number of variables. The basic procedure is as follows:

1. A regression equation containing all of the variables is computed.
2. The partial F-test value is calculated for every predictor variable treated as though it were the last variable to enter the regression equation.
3. The lowest partial F-test value, say Fl, is compared with a preselected significance level, say Fo.
   a. If Fl < FO, remove the variable which rose Fl from consideration and recompute the regression equation in the remaining variables. Then reenter stage (2).
   b. If Fl > FO, adopt the regression equation as calculated. [Ref. 3:p. 305]

Carrier Variables - See Independent Variables.

Coefficient of Determination - See R^2 [Ref. 10:p. 146]

Confidence Coefficient - Confidence Coefficients are used when speaking of confidence intervals. The confidence coefficient is the number (1-α) x 100 percent. Therefore, at an α equal to .05, the confidence coefficient is equal to 95 percent. [Ref. 10:p. 55]

Corrected Sum of Squares - The Corrected Sum of Squares (CSS) is the value obtained when the Correction for the Mean is subtracted from the Uncorrected Sum of Squares. Notationally, this is CSS = $\sum X_i^2$ - $(\sum X_i^2)/n$ and is called the Corrected Sum of Squares for the X's. [Ref. 3:p. 14]

Corrected Sum of Products - The Corrected Sum of Products (CSP) is the value obtained when the Correction for the Mean is subtracted from the Uncorrected Sum of

Products. Notationally, this is CSP = $\sum X_iY_i$ - $(\sum X_i)(\sum Y_i)/n$ and is called the Corrected Sum of Products for X and Y. [Ref. 3:p. 14]

Correlation Coefficient - The correlation coefficient, ρ_{uw}, provides an empirical measure of the linear association between U and W. Its values can be between -1 and 1. When ρ_{uw} is nonzero, this means that there exist's a linear association between the specifics values of x_i and y_i in the data. The value of a correlation ρ_{xy} shows only the extent to which x and y are linearly associated. It does not by itself imply that any sort of causal relationship exists between x and y. [Ref. 3:p. 43]

C(P) Statistic - The C(P) statistic is used to assess the fit of a regression equation. It is closely related to the R^2 and adjusted R^2 statistic. A close fitting model will have a low C(P) value close to P, where P is the number of parameters in the model including β_0. If several models are being contemplated, one method to determine the "best" model is to plot C(P) vs P for all of the models and then choose the model where C(P) falls closest to the P line. One word of caution, however, is that smaller models have smaller values of C(P), but larger models have C(P) values closer to P. If a low C(P) value close to P is not clear cut, then the analyst must make a decision. See reference for more complete details. [Ref. 3:p. 299]

Degrees of Freedom - Degrees of freedom (in regression) is a number that is associated with any sum of squares. This number indicates how many independent pieces of information involving the n independent numbers Y1, Y2, Y3, ... are needed to compile the sum of squares. [Ref. 3:p. 19]

Dependent Variable - The receptor of changes that are deliberately made or that simply happen to the independent variables. Also called the Response Variable, it is the value that a regression model is trying to predict or control. [Ref. 3:p. 3]

Dummy Variable - A variable used as an independent variable that is arbitrarily picked by the analyst. It is introduced to factor two or more distinct levels of data that may have separate deterministic effects on the dependent variables. They are usually (but not always) unrelated to the any physical levels that might exist in the factors themselves. [Ref. 3:p. 241]

Endogenous Variables - Variables that are jointly determined or that have outcome values determined through the joint interaction of other variables within the system. [Ref. 10:p. 339]

Exogenous Variables - Exogenous variables affect the outcome of the endogenous variables, but are determined outside of the system. [Ref. 10:p. 339]

F Test for the ANOVA Table - F equals the ratio of the Mean Square due to the Regression divided by the Mean Square about to Regression. Algebraically, it is F = MS_{Reg} / s^2. (see Analysis of Variance definition). This value is then compared to the 100(1-α) % point of an F distribution with (N_r - N_e) and N_e degrees of freedom. If the ratio is significant (ie -prob> F in ANOVA Table is greater than the selected 100(1-α)%) than the model is probably inadequate and attempts should be made to discover when and how the inadequacy occurs. If the F value is insignificant (ie - prob> F in ANOVA Table is less than the selected 100(1-α)%), then it is reasonable to assume that the model is accurate and that the pure error (or residual error - S^2) and the lack of fit (MS) mean squares can be used as estimates of σ^2. [Ref. 3:p. 37]

Forward Stepwise Regression Procedure - A technique which begins with no variables in a model. For each independent variable, a F statistic is calculated to reflect that particular variables contribution to the model if it is included. Variables are then included in the order of most significant to least significant. [Ref. 4:p. 102]

General Linear Hypothesis - The General Linear Hypothesis is of the form -- $Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$, where y is the dependent variable, X_1 and X_2 are the independent variables, β_o is the intercept value, β_1 and β_2 are the 'coefficients' or parameter estimates and ε is the error term. [Ref. 3:p. 102]

Heteroscedasticity - Heteroscedasticity is a situation in which the random errors (ε_i's) from the statistical regression model have different (non-constant) variances. [Ref. 10:p. 289]

Homoscedastic - A situation where there is an identical variance in the random errors. Homoscedastic is the converse of heteroscedastic. [Ref. 10:p. 119]

Indempotent Matrix - An indempotent matrix is a special form of a matrix that is symmetric and that holds the following two properties. [Ref. 10:p. 31]

1) $M = M'$ and

2) $MxM = M^2 = M$

Idependent Variable - Variables that can either be set to a desired value or else take on values that can be observed but not controlled. Also known as Carrier or Predictor Variables. [Ref. 3:p. 3]

Lack of Fit - A situation in which a postulated model is not correct. Lack of fit is present when the residuals contain both random AND systematic errors. [Ref. 3:p. 34]

Level of Significance - See α

Least Squares - A concept having to do with minimizing the square of the distance between an actual and predicted value. See Chapter 1, Section E for a detailed explanation.

Latent Variables - Variables that are not incorporated in a regression equation (or, perhaps, are not even measured) that contribute to the error in the model. Also called Lurking variables. [Ref. 3:p. 295]

Lurking Variables - See Latent Variables.

Multicollinearity - Also known as ill conditioning, multicollinearity is a situation in which there is an interrelationship amongst the predictor (or carrier) variables. These interrelationships will adversly affect statistical results which may cause estimated values to be far from the true values. [Ref. 10:p. 610]

Multiple Regression - Regression using more than one explanatory (or carrier) variable.

Nonsingular Matrix - A square matrix whose determinate is nonzero. Nonsingular matrices have full row rank (all rows and columns are linearly independent). [Ref. 11]

Normal Equation for Multiple Linear Regression - The general linear equation for multiple linear regression in matrix form is as follows. [Ref. 3:p. 74]

$$X'X\beta = X'Y$$

Overfitting - The fitting of regression equations that involve more predictor variables than are necessary to obtain a satisfactory fit to the data. [Ref. 3:p. 298]

Outliers - An outlier is a point that is far from the mean in absolute value and is, perhaps, several standard deviations away from the mean. In regression analysis, a residual that is an outlier comes under close scrutiny in order to determine if its peculiarity can be established. [Ref. 3:p. 152]

Parameter Equations for Simple Linear Regression - The general equations for estimating simple linear regression parameters are as follows. [Ref. 3:p. 14]

$$\beta_1 = S_{xy}/S_{xx} \qquad \text{and} \qquad \beta_0 = Y - \beta_1$$

where:

$$S_{xy} = \sum X_i Y_i - nXY$$
$$S_{xx} = \sum X_i^2 - nX^2$$

Residuals - Residuals (often denoted ε_i) is the difference between the actual value of y and the predicted value of y. Algebraically, this is denoted as $Y_i - Y_i$. The residuals contain all the information on the way in which the regression model fails to explain the observed variation in the dependent variable. [Ref. 3:p. 34]

Residual Plots - Plots of residuals versus other parameters in the regression. For analytical purposes, the plot of ε_i versus Y is common. The reason that the residuals are plotted against the predicted values is because the covariance between these two values ($\mathrm{Cov}(\varepsilon,Y)$) is equal to 0, whereas the covariance between the residuals and the actual values is not. (actually, $\mathrm{cov}(\varepsilon_i,Y_i) = \sigma^2 (I - X(Y'Y)^{-1}X')$).

Ridge Regression - A regression procedure that is intended to overcome certain lack of fit situations where correlations between the various carrier variables in the model cause the X'X matrix to become close to singular, giving rise to unstable parameter estimates. (The estimates may, for example, have the wrong sign or be much larger than physical or practical considerations would deem appropriate). [Ref. 3:p. 313]

R^2 - R^2 measures the proportion of total variation about the mean Y explained by the regression. Algebraically, R^2 = (SS due to the Regression)/(Total SS, corrected for the mean Y) = $\sum(\hat{Y}_i - \bar{Y})^2 / \sum(Y_i - \bar{Y})^2$ As more variables are added to the regression, R^2 (unlike adjusted R^2) will never decrease. [Ref. 3:p. 19]

Stepwise Regression Procedure - A technique which begins with no variables in a model. For each independent variable, a F statistic is calculated to reflect that particular variables contribution to the model if it is included. Variables are then included one by one in the order of most significant to least significant. Unlike the Forward Stepwise Regression Procedure, however, once a variable is entered, a regression is performed on all of the variables that are currently in the model, and any variables that may now have an F statistic which is less significant than the newly entered variable will be removed from the model. [Ref. 4:p. 102]

Weighted Least Squares - A regression technique used when some of the carrier observations are 'less reliable' than others. This is usually indicated when the variances of the observations are unequal or, sometimes, if the various observations are correlated. The basic idea is to use a transform of the observations to other variables that do fit the basic assumptions of the ordinary least squares model and then apply the usual (unweighted) analysis to these new variables. [Ref. 3:p. 108]

X'X Matrix - Matrix notation format for determining the $\sum X_i$, $\sum X_i^2$ and n. It is of particular use in multiple regression for ease of computation. The X'X matrix is determined as follows and is used in the Normal Equation for Multiple Linear Regression (see definition). [Ref. 3:p. 74]

$$X'Y = \begin{bmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{bmatrix}$$

X'Y Matrix - Matrix notation format for determining the $\sum Y_i$, $\sum X_i\ Y_i$. It is of particular use in multiple regression for ease of computation. The X'Y matrix is determined as follows and is used in the Normal Equation for Multiple Linear Regression (see definition). [Ref. 3:p. 74]

$$X'Y = \begin{bmatrix} \sum Y_i \\ \sum X_i Y_i \end{bmatrix}$$

X'X inverse Matrix $(X'X)^{-1}$ - The X'X inverse matrix is an extremely important concept in multiple regression calculations. The calculation of this matrix allows for the solving of the multiple regression equations. This matrix must be nonsingular. When both sides of the Normal Equation for Multiple Linear Regression are multiplied by $X'X^{-1}$ the resultant matrix is the matrix of the estimators of the coefficients, β. The X'X inverse matrix is calculated as follows. [Ref. 3:p. 78]

$$(X'X)^{-1} = (1/n\sum(X_i - X)^2) \begin{bmatrix} \sum X_i^2 & -\sum X_i \\ -\sum X_i & n \end{bmatrix}$$

APPENDIX B

THE VARIABLES

A detailed listing of variable information, definitions and statistical data follows. Statistical information does not include BN 6E (Honolulu). Variables that appear in the final model are analyzed first, complete with histograms. The other variables appear later with a less rigorous summary. There was no attempt to weight any data elements. All estimates are derived from performing statistical analysis on the raw data as given.

The following variables appear in the finalized model.

****************************** PENETRATION ****************************

VARIABLE NAME: PENT

DESCRIPTION: Contracts divided by HSMMA by battalion by year. Penetration actually shows what percent of the market that actually contracted with the Army.

UPDATED: As Contracts and HSMMA are updated.

MEAN	MEDIAN	ST. DEVIATION	MIN	MAX
.051157	.048616	.016318	.019678	.10397

***************************** PROPENSITY *****************************

VARIABLE NAME: PROP

DESCRIPTION: Army Positive Propensity measure. Four year moving average of the percent of positive respondents to questions about military and Army service on the Youth Attitude Tracking Survey (YATS). The data is presented as percent times 100.

UPDATED: Fall quarter (actual), other quarters (estimated)

MEAN	MEDIAN	ST. DEVIATION	MIN	MAX
14.48	14.1	4.5095	6.4	27.3

************************** BATTALION PERCENT *************************

VARIABLE NAME: BNPER

DESCRIPTION: Contracts divided by the total number of contracts signed
in any given year. BNPER tells what percent of the total number of incoming GSM I-IIIA recruits were accessed by that particular battalion.

UPDATED: Daily as contracts are updated.

MEAN	MEDIAN	ST. DEVIATION	MIN	MAX
.0183	.0179	.0056139	.0062	.0334

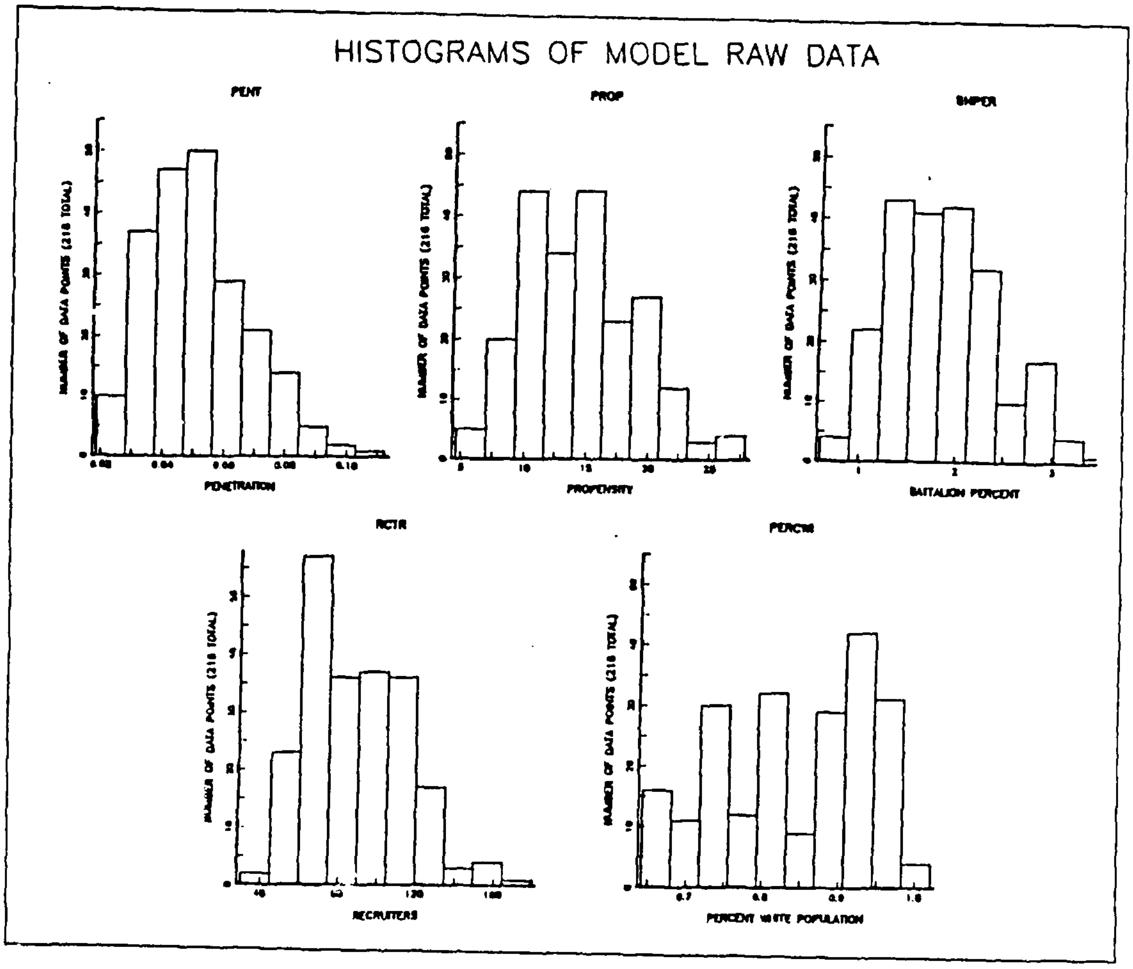


Figure B.1 Histogrm Distribution of Final Model Variables

************************** RECRUITERS ********************************

VARIABLE NAME: RCTR
DESCRIPTION: Average number of on-production recruiters assigned. On-production means all recruiters actively recruiting and assigned contract quotas (missions).

UPDATED: Yearly, or as desired by checking unit manning rosters.

MEAN	MEDIAN	ST. DEVIATION	MIN	MAX
88.69	83	25.59	30.5	165

*********************** PERCENT WHITE ********************************

VARIABLE NAME: PERCWI
DESCRIPTION: WHIPOP divided by TOTPOP. PERCWI tells the percentage of total population within a battalion are white.

UPDATED: Every census (actual), each year (estimated) with 5-year projections available every year.

MEAN	MEDIAN	ST. DEVIATION	MIN	MAX
.84296	.86	.098515	.65	.99

The following variables were considered for the model. Some were used in the derivation of other variables. These data points are maintained at USAREC headquarters at Fort Sheridan, Illinois.

*************************** BATTALION ******************************

VARIABLE NAME: BN

DESCRIPTION: USAREC recruiting battalion reference codes (BN 3L not provided)

UPDATED: As organizational realignments dictate

****************************** YEAR *******************************

VARIABLE NAME: YR

DESCRIPTION: fiscal year (1982 to 1985)

UPDATED: 1 October of each year

*************************** CONTRACTS ******************************

VARIABLE NAME: CONT

DESCRIPTION: Number of GSM I-IIIA contracts actually written per year

UPDATED: daily throughout the year

MEAN	MEDIAN	ST. DEVIATION	MIN	MAX
1018.5	975	325.27	319	2021

*************************** UNEMPLOYMENT *****************************

VARIABLE NAME: UNEM

DESCRIPTION: Average total unemployment in a given battalion for a for a given year. The data is presented as percent times 100.

UPDATED: Yearly by the Bureau of Labor Statistics with subsequent (by zipcode) updates by USAREC to fit into battalion structure.

MEAN	MEDIAN	ST. DEVIATION	MIN	MAX
8.68	8.53	2.23	3.33	15.43

************ HIGH SCHOOL MALE MARKET AVAILABLE (CAT I-IIIA) **********

VARIABLE NAME: HSMMA

DESCRIPTION: Measured or predicted size of available pool of high school seniors or high school graduates within the last two years that are in mental category I-IIIA. Also known as the market. All variables were as given by USAREC except for HSMMA for 1985. HSMMA for 1985 was the average value for HSMMA84 and HSMMA86 (as of June 25, 1986)

UPDATED: Random times throughout the year by USAREC.

MEAN	MEDIAN	ST. DEVIATION	MIN	MAX
21783	21525	8393	8172	46120

********************** PAY COMPATIBILITY *****************************

VARIABLE NAME: PAYCO

DESCRIPTION: Civilian to military pay compatibility. This is the difference in the year-to-year percent changes between income per capita and the Basic pay for an E-1 under four months of active duty service. Data is given in percent times 100.

UPDATED: As INCOMPC and E-1 PAY is updated.

MEAN	MEDIAN	ST. DEVIATION	MIN	MAX
6.45	5.61	3.34	.21	12.7

********************** TOTAL POPULATION ******************************

VARIABLE NAME: TOTPOP

DESCRIPTION: Total population within a battalion area.

UPDATED: Every census (actual), each year (estimated) with 5-year projections available every year.

MEAN	MEDIAN	ST. DEVIATION	MIN	MAX
4.15E6	4.02E6	1.18E6	2.06E6	8.92E6

********************** WHITE POPULATION ******************************

VARIABLE NAME: WHIPOP

DESCRIPTION: Total white population within a battalion area.

UPDATED: Every census (actual), each year (estimated) with 5-year projections available every year.

MEAN	MEDIAN	ST. DEVIATION	MIN	MAX
3.45E6	3.44E6	8.83E6	1.81E6	6.10E6

********************** BLACK POPULATION ******************************

VARIABLE NAME: BLKPOP

DESCRIPTION: Total black population within a battalion area.

UPDATED: Every census (actual), each year (estimated) with 5-year projections available every year.

MEAN	MEDIAN	ST. DEVIATION	MIN	MAX
4.90E5	2.70E5	4.15E5	6782	1.52E6

******************** HISPANIC POPULATION ****************************

VARIABLE NAME: HISPOP

DESCRIPTION: Total hispanic population within a battalion area.

UPDATED: Every census (actual), each year (estimated) with 5-year projections available every year.

MEAN	MEDIAN	ST. DEVIATION	MIN	MAX
2.77E5	76325	4.40E5	10496	2.36E6

******************** INCOME PER CAPITA ******************************

VARIABLE NAME: INCOMPC

DESCRIPTION: Average income per capita (in dollars) within a battalion area.

UPDATED: Yearly by the Bureau of Labor Statistics with subsequent (by zipcode) updates by USAREC to fit into battalion structure.

MEAN	MEDIAN	ST. DEVIATION	MIN	MAX
9429	9394	1374	6255	13105

******************* QUALIFIED MILITARY AVAILABLE ********************

VARIABLE NAME: QMA

DESCRIPTION: Predicted number (times 100) of physically, mentally and morally qualified for service males within a battalion area. Normally predicted as a straight percentage of the total male population.

UPDATED: Every two years.

MEAN	MEDIAN	ST. DEVIATION	MIN	MAX
1183	1165	387	336	2658

**************** BATTALION ADVERTISEMENT EXPENDITURES ***************

VARIABLE NAME: BNADV

DESCRIPTION: Battalion level expenditures (in hundreds of dollars) that were spent on advertising within the battalion. Does not include any national advertising expenditures.

UPDATED: Yearly

MEAN	MEDIAN	ST. DEVIATION	MIN	MAX
969	903	352.9	273	2211

************************* E-1 PAY ************************************

VARIABLE NAME: E1PAY

DESCRIPTION: Basic pay of an enlisted rank 1 (E-1) with under four months of active federal service.

UPDATED: Yearly as congressions pay changes mandate.

MEAN	MEDIAN	ST. DEVIATION	MIN	MAX
568.5	573.6	9.613	551.4	573.6

********************** ARMY MARKET SHARE ***************************

VARIABLE NAME: ARMYMS

DESCRIPTION: The total number of contracts by the Army divided by the total number of Department of Defense contracts within a battalion.

UPDATED: Yearly when DOD-A is updated.

MEAN	MEDIAN	ST. DEVIATION	MIN	MAX
.3811	.38	.04146	.26	.47

**************** DEPARTMENT OF DEFENSE MINUS ARMY *******************

VARIABLE NAME: DOD-A

DESCRIPTION: The total number of military contracts minus the total number of Army contracts within the battalion.

UPDATED: Yearly by the Department of Defense with subsequent (by zipcode) updates by USAREC to fit into battalion structure.

MEAN	MEDIAN	ST. DEVIATION	MIN	MAX
1658.4	1522	549.89	533	3597

APPENDIX C

SAS INPUT PROGRAM FOR INITIAL REGRESSIONS

```
//JACK JOB (0438,9999),'THESISOUT',CLASS=A
//*MAIN SYSTEM=SY2
//    EXEC SASV5
//SYSIN DD *
OPTIONS LINESIZE = 80;
DATA DATA1;
  INPUT BNN $ YEAR BN CONT RCTR UNEMP PROP HSMMA PAYCO TOTPOP WHIPOP ;
  cards;
 1A  1982 1    657    53.75    8.05  14.7  13931   6.99  2169022  2083422
 1A  1983 1    805    52.25    7.93  15.1  13816  10.49  2180204  2093593
 :    :   :     :       :        :     :     :       :       :        :
 :    :   :     :       :        :     :     :       :       :        :
 6L  1984 55  1060   106.00    9.95   8.8  25459   1.43  4203768  3828164
 6L  1985 55  1396    97.00    8.63   8.5  24122   1.41  4253796  3870698
;
DATA DATA2;
  INPUT BLKPOP HISPOP INCOMPC QMA BNADV E1PAY ARMYMS DODMA ;
  cards;
 57346   25885     7610     555      720    551.4    0.33     1348 1A
 57840   26018     8715     555      633    573.6    0.37     1370 1A
   :       :         :       :        :       :        :         :   :
   :       :         :       :        :       :        :         :   :
110199  122804    11135    1281     1058    573.6    0.36     1866 6L
111838  124265    11291    1281     1124    573.6    0.43     1869 6L
;
DATA ALLYEARS;
  MERGE DATA1 DATA2;

PROC REG DATA=ALLYEARS;
  MODEL CONT=YEAR BN RCTR UNEMP PROP HSMMA PAYCO TOTPOP WHIPOP
             BLKPOP HISPOP INCOMPC QMA BNADV E1PAY ARMYMS DODMA /
             R CORRB COLLIN VIF ;
  ID BNN;

PROC STEPWISE DATA=ALLYEARS;
  MODEL CONT=YEAR BN RCTR UNEMP PROP HSMMA PAYCO TOTPOP WHIPOP
             BLKPOP HISPOP INCOMPC QMA BNADV E1PAY ARMYMS DODMA /
             SLE=1 SLS=1;

PROC SORT DATA = ALLYEARS;
  BY YEAR;

PROC REG DATA=ALLYEARS;
  MODEL CONT=BN RCTR UNEMP PROP HSMMA PAYCO TOTPOP WHIPOP
             BLKPOP HISPOP INCOMPC QMA BNADV ARMYMS DODMA /
  BY YEAR;

PROC STEPWISE DATA=ALLYEARS;
  MODEL CONT=BN RCTR UNEMP PROP HSMMA PAYCO TOTPOP WHIPOP
             BLKPOP HISPOP INCOMPC QMA BNADV ARMYMS DODMA /
             SLE=1 SLS=1;
  BY YEAR;
/*
//
```

APPENDIX D

SAS INPUT PROGRAM FOR INTERMEDIATE REGRESSIONS

```
//JACK JOB (0438,9999),'THESISOUT',CLASS=A
//*MAIN SYSTEM=SY2
//     EXEC SAS
//SYSIN DD *
OPTIONS LINESIZE = 80;
DATA DATA1;
   INPUT BNN $ YEAR BN CONT RCTR UNEMP PROP HSMMA PAYCO TOTPOP WHIPOP ;
  PERCWI=WHIPOP/TOTPOP;
  PENT=CONT/HSMMA;
  IF YEAR EQ 1982 THEN BNPER = CONT/51431;
  IF YEAR EQ 1983 THEN BNPER = CONT/63498;
  IF YEAR EQ 1984 THEN BNPER = CONT/52299;
  IF YEAR EQ 1985 THEN BNPER = CONT/55941;
  cards;
;
  (data)
DATA DATA2;
   INPUT BLKPOP HISPOP INCOMPC QMA BNADV E1PAY ARMYMS DODMA ;
   cards;
    (data)
;
DATA ALLYEARS;
  MERGE DATA1 DATA2;
PROC REG DATA=ALLYEARS;
  MODEL PENT=PROP BNPER RCTR PERCWI /R CORRB COLLIN VIF ;
      OUTPUT OUT=OUT1 P=YHAT1 R=RESID;
      ID BNN;

PROC STEPWISE DATA=ALLYEARS;
  MODEL PENT=PROP BNPER RCTR PERCWI
               / SLE=1 SLS=1;

DATA OUT82;
     SET OUT1;
     IF YEAR NE 1982 THEN DELETE;
      R82=RESID1;

DATA OUT83;
     SET OUT1;
     IF YEAR NE 1983 THEN DELETE;
      R83=RESID1;

DATA OUT84;
     SET OUT1;
     IF YEAR NE 1984 THEN DELETE;
      R84=RESID1;

DATA OUT85;
     SET OUT1;
     IF YEAR NE 1985 THEN DELETE;
      R85=RESID1;

DATA LAG83;
     MERGE OUT82 OUT83;

DATA LAG84;
     MERGE OUT83 OUT84;

DATA LAG85;
     MERGE OUT84 OUT85;
```

```
PROC PLOT
     DATA= LAG83;
     PLOT R82*R83='*' / VREF=0 HREF=0;

PROC PLOT
     DATA= LAG84;
     PLOT R83*R84='*' / VREF=0 HREF=0;

PROC PLOT
     DATA= LAG85;
     PLOT R84*R85='*' / VREF=0 HREF=0;
              / SLE=1 SLS=1;

PROC SORT DATA=OUT1;
BY YEAR;

PROC PLOT
     DATA=OUT1;
     PLOT RESID1*YHAT1=BNN/VREF=0;
  BY YEAR;

PROC PLOT
     DATA=OUT1;
     PLOT RESID1*PROP=BNN/VREF=0;
  BY YEAR;

PROC PLOT
     DATA=OUT1;
     PLOT RESID1*BNPER=BNN/VREF=0;
  BY YEAR;

PROC PLOT
     DATA=OUT1;
     PLOT RESID1*RCTR=BNN/VREF=0;
  BY YEAR;

PROC PLOT
     DATA=OUT1;
     PLOT RESID1*PERCWI=BNN/VREF=0;
  BY YEAR;

PROC CHART
     DATA=OUT1;
     HBAR RESID1/MIDPOINTS=-.021 TO .021 BY .0020;
  BY YEAR;
/*
//
```

LIST OF REFERENCES

1. Daula, Thomas V. and Smith, Alton D., *Recruiting Goals, Enlistment Supply and Enlistments in the U.S. Army*, USAREC Publication, October, 1984.

2. ABT Associates Inc., *The Recruiting Resource Allocation System: Final Report*, December 31, 1984.

3. Draper, N. R. and Smith, H., *Applied Regression Analysis*, 2d ed., John Wiley and Sons, 1981.

4. SAS Institute Inc., *SAS User's Guide: Statistics*, Version IV ed., 1982.

5. Mosteller, F. and Tukey, J. W., *Data Analysis and Regression*, Addison-Wesley Publishing Company, 1977.

6. ABT Associates Inc., *A Review of Military Enlistment Supply Models: In Search of Further Improvements* July 1983.

7. Argonne National Laboratory, *Socioeconomic Analysis of U.S. Army Recruitment*, December 1985.

8. Kmenta, Jan, *Elements of Econometrics*, The Macmillan Company, 1971.

9. Conover, W.J., *Practical Nonparametric Statistics*, John Wiley & Sons, 1980.

10. Judge, George G., et al. *Introduction to the Theory and Practice of Econometrics*, John Wiley and Sons, 1982.

11. Anton, Howard, *Elementary Linear Algebra*, John Wiley & Sons, 1984.

12. Barr, Donald R. and Zehna, Peter W., *Probability: Modeling Uncertainty*, Addison-Wesley, 1983.

13. Koutsoyiannis, A., *Theory of Econometrics, 2nd ed.*, MacMillian Ltd, 1983.

INITIAL DISTRIBUTION LIST

		No. Copies
1.	Defense Technical Information Center Cameron Station Alexandria, Virginina 22304-6145	2
2.	Library,Code 0142 Naval Postgraduate School Monterey, California 93943-5002	2
3.	Deputy Undersecretary of the Army for Operations Research Room 2E261, Pentagon Washington, D.C. 20310	2
4.	Commander, United States Army Recruiting Command Attn: USAR-PAE-PA (Captain Patchell) Fort Sheridan, Illionis 60037	10
5.	Headquarters, United States Army Deputy Chief of Staff for Personnel Room 2D724, Pentagon Attn: DAPE-PAE (Captain McKenna) Washington, D.C. 20310	5
6.	PROF Dan Boger, Code 55BO Department of Operations Research Naval Postgraduate School Monterey, California 93943-5000	2
7.	PROF Harold Fredrickson,Code 53FS Department of Mathematics Naval Postgraduate School Monterey, California 93943-5000	2
8.	CPT Jack E Faires 54 Beechrock Drive South Zanesville, Ohio 43701	2

END

2-81

DTIC