R80-29

OSP 87493
32591

EMPIRICAL TEMPERATURE FORECASTING: EXTENSIONS OF THE MODEL OUTPUT STATISTICS METHOD

by
David E. Langseth
and
Rafael L. Bras

RALPH M. PARSONS LABORATORY
FOR
WATER RESOURCES AND HYDRODYNAMICS

Report No. 257

Prepared with the support of the
U.S. Department of Energy Through
The M.I.T. Energy Laboratory
and
The Gilbert Winslow Career Development Chair

June 1980

MIT

DEPARTMENT
OF
CIVIL
ENGINEERING

SCHOOL OF ENGINEERING
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Cambridge, Massachusetts 02139

R 80-29

OSP 87448
32591

EMPIRICAL TEMPERATURE FORECASTING:

EXTENSIONS OF THE MODEL OUTPUT STATISTICS METHOD

by

David E. Langseth

and

Rafael L. Bras

RALPH M. PARSONS LABORATORY
FOR
WATER RESOURCES AND HYDRODYNAMICS

Department of Civil Engineering

Report No. 257

Prepared with the support of the

U.S. Department of Energy Through

The M.I.T. Energy Laboratory

and also

The Gilbert Winslow Career Development Chair

June 1980

ABSTRACT

EMPIRICAL TEMPERATURE FORECASTING:
EXTENSIONS OF THE MODEL OUTPUT STATISTICS METHOD

by

DAVID E. LANGSETH

and

RAFAEL L. BRAS

Deterministic models of complex natural phenomena such as streamflow or weather events are usually either unknown or unwieldy and thus are often augmented or replaced by stochastic or empirical models. For example, the National Weather Service (NWS) uses a combination of deterministic and empirical models to predict several weather parameters. An approximate deterministic model of the atmosphere provides predictions of some meteorological parameters at the grid points used in the numerical solution of the model. Some of these deterministic predictions, along with recent measured data, are then used as input variables to an empirical prediction equation. The National Weather Service uses a stepwise least-squares regression algorithm to develop the empirical equations.

The prediction of maximum surface air temperature is investigated in this work. The NWS currently uses 10 variable linear models to predict maximum temperatures. The 10 variable restriction is based on research and the linear restriction is based primarily on the prohibitive amount of time and effort required to develop non linear models. The potential model improvements from relaxing these two restrictions are examined in this work. Data from Huntsville, Alabama, supplied by the NWS, is used. Non linear models are created by applying a non linear model identification algorithm called the Group Method of Data Handling to the data. Two linear model identification algorithms are also used. The usefulness of the removal of harmonic components and the identification of principal components were investigated along with each of the model identification algorithms.

It is shown that, for the site investigated, the linear restriction does not hurt model quality and that while 10 is a reasonable number of variables, models with fewer variables can also perform well. Also, modeling the mean trends separately from the more transient effects improves model quality.

ACKNOWLEDGEMENTS

This work was supported in part by the Department of Energy under contract number EX-76-A-01-2295. The authors wish to thank Dr. Harry Glahn and Gary Carter of the National Weather Service Techniques Development Laboratory for providing the data used in this work and Pedro Restrepo and Kevin Curry for assistance in computer programming.

The patient and skillful typing of Zigrida Garnis, Carolyn Jundzilo Comer, and Anne Clee are gratefully acknowledged.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

Weather influences many activities and weather forecasts influence many decisions. Some of these decisions, such as whether to plan a picnic for a particular day, are made informally. Others, such as raisin growers' decisions to set grapes out to dry, have major consequences and are often carefully analyzed with regard to weather forecasts. Examples of decision making using weather forecasts are given in Howe and Cochrane (1976), Kernan (1975), Carter (1972), Helbush (1968), Glahn (1964), Lave (1963), and Kolb and Rapp (1962).

Weather related decision making involves interaction between the forecasters, or forecasting system, and the forecast consumer. The forecasters must decide what, when, how, and how well to forecast. The forecast consumers must decide how best to use the forecast. For example, the basic output of a physical model of the atmosphere may be humidity, but raisin growers and construction contractors both want forecasts of rainfall. The forecaster must then decide whether to attempt a forecast of rainfall or let the consumers use the humidity forecast directly. The contractor may need a 3 day advance notice of a single dry day while the raisin grower may need a one day advance notice of 3 consecutive dry days. Either may desire a forecast of the probability of rain or may prefer an unqualified statement whether or not it will rain. Forecasters can tailor their forecasts to a particular consumer or try

to produce a forecast of more general usefulness (Murphy, 1977, and Nelson and Winter, 1964). This work is restricted to an examination of some forecasting methods.

The numerous methods of forecasting weather can generally be classified as either objective or subjective. We will adopt the definition given by Allen and Vernon (1951) that objective forecasts are those which are uniquely determined by a set of data. Subjective forecasts are those in which human judgment is used. This distinction is sometimes fuzzy, as, for example, when some of the data used in an otherwise objective forecast procedure have been subjectively derived or adjusted. Objective forecasts are frequently used as guidance for subjective forecasts. Subjective forecasts are not discussed in this work.

Physical and empirical models are the basic objective forecast methods. Physical models use fundamental equations of motion, thermodynamics, and continuity, along with prescribed boundary and initial conditions, to forecast future conditions. Empirical models use data to calibrate relations between variables known at the forecast issue time and the forecast variable. The forms of empirical models are usually chosen for convenience and only implicitly describe the physical relations. Empirical models thus rely primarily on data, rather than physics, to connect the past to the future. Data may also be used in physical models, either for calibration or to assign initial and boundary conditions. Subjective judgment is normally required to construct either a physical or empirical model, but the models are still called objective because once they are constructed, only one forecast can be produced

from a given set of input data.

Physical forecast models were proposed by Bjerknes in 1904 and first attempted during World War I by Richardson. The development of high speed computers made their routine use possible. Petterssen (1957) reviews the history of physical forecasting methods and Rieck, et al. (1976) gives brief descriptions of several physical models used by the National Weather Service (NWS). The physical models related to this work are described in Section 2.1.

Many empirical objective methods have been used to forecast weather. Some uses of scatter diagrams, discriminant analysis, adaptive logic, multiple linear regression, and orthogonal functions are described in Glahn (1965). Scatter diagrams are an approximate, but statistically robust, method of deriving relations between variables. The use of scatter diagrams in temperature forecasting is described in Dickey (1960). Discriminant analysis is generally useful when the predictand is one of a set of categories, rather than a continuous variable. A good description and example use of discriminant analysis in weather forecasting is given in Miller (1962). A form of adaptive logic is part of the motivation for the structure of the GMDH, a model identification method used in this work and described in Chapter 4. Multiple linear regression and orthogonal functions are used in this work and are described in Chapter 3.

The primary advantage of physical models over empirical models is their relatively general applicability. Empirical models frequently have severely limited prediction capabilities outside the range of the

data used to calibrate the model. The primary advantage of empirical models over physical models is their ability to capture complex relations without precise specifications of the underlying processes. Empirical models are thus frequently much simpler in form and easier to use than physical models.

Some of the most successful short range (less than 3 days) forecast models are empirically derived linear combinations of physical model forecasts and other meteorological variables. Such models are the objective analogs of subjective forecast methods and have the advantages over subjective forecast methods of a nearly perfect and neutral memory and being transferable between forecasters, but have the disadvantage of not being able to capture the full range of relations implicit in a forecasters experience. The Perfect Prog (PP) and Model Output Statistics (MOS) models discussed in Section 2.2 have this form.

The models developed in this work are extensions of the MOS modeling method. MOS models for temperature forecasting are linear combinations of 10 variables chosen from 70 to 120 (depending on the particular variable being forecast) potential predictor variables. The 10 term restriction is based on research by Annet et al. (1972) and Bocchieri and Glahn (1972). The linear restriction is dictated by the large number of models which must be produced by the NWS. A variety of models with different numbers and transformations of the potential predictor variables are developed and examined in this work.

A secondary emphasis in this work is model validation.

The NWS produces too many models to consider applying validation procedures to every model. Validation procedures are applied to a few of the models developed in this work both to illustrate the procedures and to suggest how other models might perform when subjected to the same procedures.

The temperature forecasting methods used by the NWS are described in Chapter 2. Most of the empirical modeling techniques used in this work are described in Chapter 3. A model identification method called the Group Method of Data Handling is described in Chapter 4. The details of the procedures used to generate alternative models are described in Chapter 5 and those models are analyzed in Chapter 6. Chapter 7 contains the summary and conclusions.

Chapter 2

OBJECTIVE TEMPERATURE FORECASTING

Model Output Statistics (MOS) is the most successful short range objective temperature forecasting method in current use. Perfect Prog (PP) models were the direct predecessors of MOS models and are still used for some forecasts. MOS and PP models are empirical models which use both observed conditions and physical model forecasts as predictor variables. Models developed in this work are based on the MOS modeling method. The physical models whose forecasts are used in the MOS and PP models are described in Section 2.1, the PP and MOS modeling methods are described in Section 2.2, the current schedule for MOS temperature forecasts is described in Section 2.3, and the data used to develop MOS temperature forecasting models are described in Section 2.4. Some methods of evaluating the quality of temperature forecasting models are described in Section 2.5.

2.1 Physical Models

Forecasts from the Seven Layer Primitive Equation (7LPE), Limited Area Fine Mesh (LFM-II), and Trajectory Models are used in MOS and PP models. These 3 physical models are described briefly in this section.

The 7LPE model is the basic physical model used by the National Meteorological Center (NMC) for routine forecasting. The 7LPE model takes its name from the number of vertical layers in the grid over which the model is solved and the nature of the equations used

in the model. Until January 1978 the NMC used a similar model which had 6 vertical layers and was called the 6LPE model. The 6LPE model is described in Stackpole (1975) and Schuman and Hovermale (1968). The 7LPE model is described in Brown (1977 and 1977a).

The dynamics of the atmosphere are described in the 7LPE model by equations of motion in 3 dimensions, thermodynamics of potential temperature, and continuity of dry air and water vapor. Complementary equations describe forces to which the air is subjected, the heat budget, and sources and sinks of air and water. The equations are written in horizontal coordinates related to lines of latitude and longitude and a vertical coordinate perpendicular to the surface of the earth.

The model is solved numerically over a three dimensional grid which covers the northern hemisphere and extends to the top of the atmosphere. The horizontal grid array is 129 x 129. The mesh length varies with latitude, being 153 kilometers at 30°N and 180 kilometers at 50°N. The vertical mesh length is initially defined by atmospheric pressure and the location of the tropopause. The boundary layer of the model is the first 50 millibars (mb) of pressure change. Between the top of the boundary layer and the tropopause there are 3 layers of initially equal pressure thickness. The layer pressure thickness is the pressure change from the top to the bottom of a layer. Between the tropopause and 50 mb of pressure there are another 3 layers of initially equal pressure thickness. An eighth layer extends from

50 to 0 mb, but is not included in the model name because it has no meteorological function. The layer pressure thicknesses change during the execution of the model. The initial conditions for the 7LPE model are assigned from observed data. Potential temperature, 2 horizontal components of wind, layer pressure thickness, and precipitable water are forecast directly by the model. Other variables are derived from these 5.

The LFM-II model was designed to provide increased forecast accuracy in the areas of greatest interest. The LFM-II model uses essentially the same equations as the 7LPE model, but differs from the 7LPE model in the horizontal mesh length, time step, and method of assigning boundary and initial conditions. The horizontal grid array is 79 by 67,with a mesh length of 116 km at 45°N. The grid covers North America and some of the surrounding ocean. The time step in the LFM-II model is reduced from that in the 7LPE model to retain numerical stability. Boundary and initial conditions for the LFM-II model are assigned from a combination of observed conditions and 7LPE model forecasts. The LFM-II model replaced the LFM model shortly before the 7LPE model replaced the 6LPE model. Rieck (1978) and Gerrity (1977) describe the LFM model and Brown (1977b) describes differences between the LFM and LFM-II models.

The trajectory model was designed to provide improved low level temperature and moisture forecasts, with special application to severe storm prediction (Rieck et al., 1976). Wind forecasts from the

7LPE model are used in the trajectory model to compute air parcel trajectories. Changes in temperature and moisture content along those trajectories are then calculated. The initial temperature and moisture conditions are assigned from observed data. 6LPE model wind forecasts were used in the trajectory model prior to the introduction of the 7LPE model.

The NMC runs each physical model twice each day. The runs are called the 0000 Greenwich Mean Time (GMT) and 1200 GMT forecast cycles. General information about forecast schedules and cutoff times for initialization can be found in Rieck et al.(1976).

Since physical models are always approximate descriptions of a real system, model predictions usually deviate from real system performance. Some of the deviations may be random and some may have patterns. The nature of the deviations between the real behavior of the atmosphere and the physical model predictions is an important part of the difference between the MOS and PP forecast methods.

2.2 Statistical Models

2.2.1 Perfect Prog (PP) Models

Perfect Prog models are linear combinations of predictor variables. The development and use of PP models is described in Klein and Glahn (1974), Klein et al. (1971), Klein and Lewis (1970), Klein et al. (1967), and Klein (1966). The predictors and associated weights for a particular model are usually chosen by applying a forward moving stepwise regression algorithm (see Draper and Smith, 1966) to a

list of potential predictors thought to be related to the predictand. Observed values of the predictors are used to develop the equations. However, when the equations are applied some of the predictors have not yet been observed and are replaced by forecasts of these predictors from physical models. Errors in the physical model forecasts are thus translated directly into errors in the PP model forecasts. The data records used to develop PP equations are usually 15 to 20 years long.

PP models were used by the NMC from 1964 through 1973 to issue max/min temperature forecasts out to 60 hours in advance for each of 143 cities in North America. The derivation and use of these models is described in Klein and Lewis (1970) and illustrates the general PP method. Data from 18 years were used to develop separate equations for each 2 month period. The potential predictors were the 700 mb heights and 700-1000 mb thicknesses observed approximately 12 hours before the valid time of the forecasts at 67 of the 6LPE model grid points, the observed maximum and minimum temperature from the preceeding day, and the day of the year. Observed values of the 700 mb heights and 700-1000 mb thicknesses were used to develop the PP equations and 6LPE model forecasts of the 700 mb heights and 700-1000 mb thicknesses were used to forecast with the PP equations. Observed values of the max/min temperatures from the preceeding day were used both develop and to forecast with the PP equations. The forward moving stepwise algorithm used to choose the predictors for each equation added pairs of variables

until no pair could increase the explained variance of the predictand by more than 2 percent.

PP models have been replaced by MOS models in most situations. PP and MOS models are compared in Section 2.2.2.

2.2.2 Model Output Statistics (MOS) Models

Guidance forecasts for air temperature, probability of precipitation, precipitation type, thunderstorm occurrence, cloud amount, wind speed, and wind direction are issued by the NMC using MOS models developed by the National Weather Service (NWS) Techniques Development Laboratory (TDL). Application of MOS models is described in Klein and Glahn (1974) and Glahn and Lowry (1972). Forecasting temperature with the MOS method is described in Carter et al. (1979), Hammons et al. (1976), and Klein and Hammons (1975).

There are 4 primary differences between the MOS and PP methods. First, in the MOS method physical model predictions are used both to develop and to forecast with the equations. Second, in the MOS method all the predictors are values for the forecast site. The values of the physical model forecasts at the forecast site are interpolated from the four grid points surrounding the site. Many of the physical model forecasts are smoothed by averaging each grid point value with 4, 8, or 24 surrounding grid point values prior to interpolation to the forecast site. Third, the data samples used to develop MOS equations are generally much shorter than the data samples used to develop the PP equations because there are longer records of observed

atmospheric conditions than of physical model forecasts. Also, periodic changes in the physical models further shorten the useful record length. Fourth, a greater variety of potential predictors is available to the MOS method than to the PP method because many of the variables forecast by the physical models are not observed directly.

The MOS method has 2 primary advantages over the PP method. First, some of the systematic errors in the physical model forecasts can be accounted for in the MOS method. If the physical model forecast errors were randomly distributed and unbiased, the MOS method would lose this advantage over the PP method and the short data records used to develop the MOS models might even introduce some errors not found in the PP models. Second, all of the predictors available to the PP method are also available to the MOS method, but some of the MOS predictors are not available to the PP method.

The PP method has 2 primary advantages over the MOS method. First, equations developed from long data samples tend to be relatively stable and thus do not need frequent redevelopment. Second, PP models improve directly with improved physical model forecasts. MOS equations are also likely to improve with improved physical models, but not until several years after the introduction of the new physical model, when there are enough archived forecasts to develop new equations.

2.3 MOS Temperature Forecast Schedule

MOS models are used to issue guidance forecasts of air temperature out to 60 hours in advance for approximately 240 cities.

Separate equations are used for each city. PP models are used for longer projection times but only the MOS models are discussed here.

The NMC issues 2 groups of MOS temperature forecasts in each of 2 daily forecast cycles. The 2 groups are called the early and final guidance packages and the forecast cycles are called the 0000 GMT and 1200 GMT cycles. The early guidance equations were developed from LFM model forecasts and are run with LFM-II model forecasts. The final guidance equations were developed from 6LPE and trajectory model forecasts and are run with 7LPE and trajectory model forecasts. These physical model changes were not considered sufficiently severe to warrant abandoning the old equations, but local forecasters are warned to watch for occasional unusual behavior in the MOS forecasts. This work is based on data used for the early guidance 0000 GMT forecast models.

The early guidance forecast schedule is shown in Figure 2.1 (from Carter et al., 1979). The upper time line is for general reference in the rest of the figure and the other lines show the times in each cycle for which temperature forecasts are issued. The 3 hourly forecasts are forecasts of temperature at the specified time and the max/min forecasts are forecasts of the maximum and minimum temperatures during calendar days. The max/min temperature forecasts are shown on separate lines from the 3 hourly forecasts and are marked at their approximate expected times of occurrence.

The equations within each set, for a given season, were

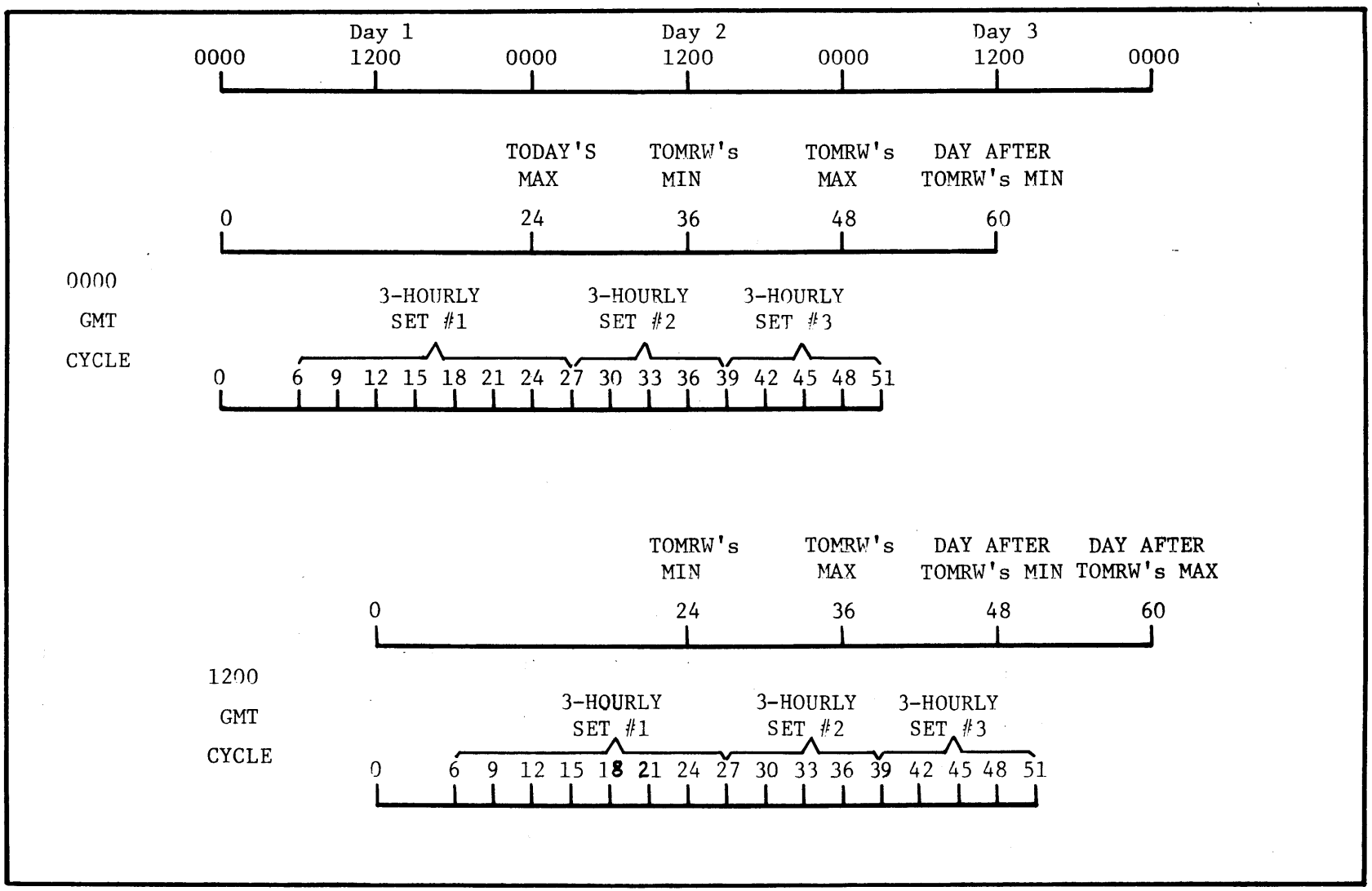

Figure 2.1 MOS Early Guidance Forecast Schedule
(from Carter et al., 1979)

constrained to use the same predictors. For example, the April through June (see Table 2.4) forecast equations for today's max and the temperatures at 6,9,12,15,18,21,24, and 27 hours in the 0000 GMT cycle all contain the same predictors, although the coefficients of the predictors vary between equations. The 24 hour max/min is associated with set 1, the 36 hour max/min is associated with set 2, and the 48 hour max/min is associated with set 3. The set numbers correspond to those shown in Tables 2.2 and 2.4. Primary and backup equations are available for each forecast. The primary equations were developed from the complete list of potential predictors (see Section 2.4). Observed predictors are not used in the backup equations.

The early guidance package is usually available to forecasters by 0004 GMT in the 0000 GMT forecast cycle and 1600 GMT in the 1200 GMT forecast cycle. Final guidance packages are usually available about 4 hours later.

2.4 MOS Temperature Forecast Equation Development Data

The potential predictors from which the MOS early guidance temperature prediction equations are developed include forecasts by physical models, observed conditions at the forecast site, and the first 2 harmonics of the day of the year. The physical model forecast variables and observed conditions are those thought to influence or be related to temperature. For example, the humidity variables

influence temperature through the effect of water on the heat budget and the layer temperature variables are related to surface temperature through the lapse rate for temperature. The first 2 harmonics of the day of the year are represented by the 4 functions

$$A_1 \sin(\frac{2\pi D}{365}) \qquad 2.1$$

$$A_2 \cos(\frac{2\pi D}{365}) \qquad 2.2$$

$$A_3 \sin(\frac{4\pi D}{365}) \qquad 2.3$$

$$A_4 \sin(\frac{4\pi D}{365}) \qquad 2.4$$

where D is the day of the year and A_1, A_2, A_3, and A_4 are weights chosen in the modeling process. The harmonic terms were originally introduced into the potential predictor list to correct a bias in the MOS forecasts noticed during the development of the MOS technique (Annet et al., 1972). The harmonic terms were later described as a method of modeling the seasonal trend of temperature (Carter et al., 1979). Modeling the seasonal trend with harmonic terms prior to developing the rest of a temperature prediction model is discussed in Section 5.2.

The observed predictors are listed in Table 2.1 (following Carter et al., 1979) and the physical model forecast variables are listed in Table 2.2. Table 2.2 was provided by the NWS TDL. The set numbers in Table 2.2 correspond to those listed in Figure 2.1. The observation times listed in Table 2.2 are for the physical model

forecasts and are relative to 0000 (1200) GMT of day 1 for the 0000 (1200) GMT forecast cycle. No observed predictors are used with set 3 or for the 60 hr. max/min. The abbreviations used in Tables 2.1 and 2.2 are listed in Table 2.3.

The years of data and seasonal stratification used to develop the currently operational early guidance forecast equations are listed in Table 2.4 (following Carter et al., 1979). Separate equations for each of the forecasts described in Section 2.3 were developed for each season. Sets 2 and 3 will also be stratified into 3 month seasons when sufficient data is available.

Set 1 data in the 0000 GMT forecast cycle for the Spring season at Huntsville, Alabama are used in this work. The predictand is today's maximum temperature. The NWS TDL supplied the data.

Element	0000 GMT cycle	1200 GMT cycle
	0300	1500
Sfc temperature	0000	1200
	2100 (yesterday)	
Sfc dew point temp	0300	1500
Cloud cover	0300	1500
Sfc U wind	0300	1500
Sfc V wind	0300	1500
Sfc wind speed	0300	1500
Ceiling height	0300	1500
Previous maximum temp		1200
Previous minimum temp	0000	
Snow cover	1200 (yesterday)	1200

Table 2.1 Potential Observed Predictors Used to Derive the MOS Early Guidance Temperature Prediction Equations

Variable	3-hr Set #1	3-hr Set #2	3-hr Set #3	60-hr max/min
1000-MB HEIGHT	12*,24*	24*,30*,36*	36**,42**,48**	48**,48***
850-MB HEIGHT	12,24	24,30,36	36,42,48	48*,48**
500-MB HEIGHT	12,24	24,30,36	36,42,48	36*,48*
500-1000 MB THICKNESS	0,6,12,18,24	24,30,36	36,42,48	48*
850-1000 MB THICKNESS	0,6,12,18,24	24,30,36	36,42,48*	48*,48**
500-850 MB THICKNESS	0,6,12,18,24	24,30,36	36,42,48*	48*
1000-MB TEMP	0,12*,24*	0,24*,36*	36**,48**	48**,48***
850-MB TEMP	0,6,12,18,24	0,24*,30*,36*	36*,42*,48*	48*,48**
700-MB TEMP	0,12,24	24,30,36	36*,42*,48*	48*,48**
BND LYR POTENTIAL TEMP	6,12,18,24	24*,30*,36*	36*,42*,48*,48**	48*,48**
BND LYR U	6,12,18*,24*	24*,30*,36*	36*,42*,48*	48*,48**
BND LYR V	6,12,18*,24*	24*,30*,36*	36*,42*,48*	48*,48**
BND LYR WIND SPEED	6,12,18*,24*	24*,30*,36*	36*,42*,48*	48*,48**
850-MB U	6,12,18*,24*	24*,30*,36*	36*,42*,48*	48**
850-MB V	6,12,18*,24*	24*,30*,36*	36*,42*,48*	48**
700-MB U	12,24*	24*,36*	36*,48*	48**
700-MB V	12,24*	24*,36*	36*,48*	48**
850-MB REL VORT	6*,12*,18*,24*	30**,36**	42**,48**	48**
500-MB REL VORT	12*,24*	30**,36**	42**,48**	48**
850-MB VERT VEL	12*,24*	36*	48**	48***
700-MB VERT VEL	12*,24*	30*,36*	42*,48*	48***
700-1000 MB TEMP DIF	12,24	36*	48*	48**
500-850 MB TEMP DIF	12,24	30*,36*	42*,48*	48**
BND LYR REL HUM	0*,6*,12*,18*,24*	24*,30*,36*	36**,42**,48**	48***
MEAN REL HUM	6*,12*,18*,24*	24*,30*,36*	36**,42**,48**	48***

Table 2.2 (cont'd on next page)

PRECIPITABLE WATER	6*,12*,18*,24*	30*,36*	42**,48**	48***
1000-MB DEW POINT	6*,12*,18*,24*	30*,36*	42*,48*	48**,48***
850-MB DEW POINT	12*,24*	30*,36*	42*,48*	48**
700-MB DEW POINT	12*,24*	30*,36*	42*,48*	48**
BND LYR WIND DIVERGENCE	6*,12*,18*,24*	30*,36*	42**,48**	48***
850-MB TEMP ADVECTION	12*,24*	30*,36*	42**,48**	48***
500-MB VORT ADVECTION	12*,24*	30*,36*	42**,48***	48***

Table 2.2 Projection Times of Potential Predictors from Physical Models Used to Derive the MOS Early Guidance (LFM based) Temperature Prediction Equations. The Stars Indicate the Field was Smoothed by 5 Points (*), 9 Points (**) or 25 Points (***).

Sfc	=	surface
temp	=	temperature
U	=	east-west wind component
V	=	north-south wind component
MB	=	millibar
BND LYR	=	boundary layer
REL VOR	=	relative vorticity
VERT VEL	=	vertical velocity
DIF	=	difference
REL HUM	=	relative humidity

Table 2.3 Abbreviations Used in Tables 2.1 and 2.2.

Season	24 h max/min 3-hourly set #1	36 h max/min 3-hourly set #2	48 & 60 h max/min 3-hourly set #3
Spring (April-June)	5(1973-77)	---	---
Summer (July-September)	5(1973-77)	---	---
Warm (April-September)	---	3(1975-77)	2(1976-77)
Fall (October-December)	6(1972-77)	---	---
Winter (January-March)	6(1973-78)	---	---
Cool (October-March)	---	3(1975-78)	2(1976-78)

Table 2.4 Number of Seasons of Archived Forecasts from the LFM Model Available for the Development of the Early Guidance Temperature Prediction Equations.

2.5 Forecast Evaluation

The quality of temperature forecasts can be measured in several different ways. The most rigorously justificable methods are

based on maximization of the utility of the forecast to the forecast consumers. Such measures of forecast quality require knowledge of both the forecast characteristics and the forecast consumer's utility functions. Forecast quality measures of this type are discussed by Thompson and Brier (1955), Gringorten (1959), Thompson (1962), Glahn (1964), Nelson and Winter (1964), and Murphy (1977), among others. These methods are generally used only to evaluate proposed models, though Glahn (1964) incorporated the forecast consumer's utility function into a model development scheme.

The most commonly used measures of forecast quality are statistical measures of forecast accuracy. The root mean squared error (rmse), mean absolute error (mae), correlation between forecasts and observations, number of large errors (nle), and forecast bias have been used to measure forecast accuracy. These statistics are usually calculated on independent data, but the rmse is also calculated on data used to develop the model. The reasons for preferring independent data for measuring forecast model accuracy are discussed in Chapter 3.

As it is rarely clear which measure of forecast accuracy is to be preferred, several writers have presented multiple measures of forecast accuracy, all based on independent data. Klein and Glahn (1974) presented the rmse, mae, correlation, and bias. Klein and Lewis (1970), Klein et al. (1967), and Klein (1966) presented the mae, rmse, and correlation. Hammons et al. (1976) and Klein and

Hammons (1975) presented the mae and correlation. Glahn and Lowry (1972) presented the mae, nle, and bias. Carter et al. (1979), Zurndorfer et al. (1979) and Klein et al. (1971) present only the mae. All of the preceeding writers also presented the rmse on estimation data when discussing model development characteristics. Sanders (1973) presented the percentage improvement of the mae over a control forecast as a measure of forecast accuracy. Sanders used climatology as the control forecast, but suggested that this was not the only valid choice.

Five statistical measures of forecast accuracy based on the data used to develop the models and two measures based on independent data are evaluated for the models developed in this work. These measures are described in section 3.2.3. Sanders' suggestion was not used because in the absence of a meaningful control forecast it would simply rescale all the numbers.

Chapter 3

EMPIRICAL MODELING

Empirical models can be used for many diverse purposes, including summarizing data, discovering cause and effect relations, and prediction. The appropriate techniques of modeling vary with the purpose for which a model will be used. The models developed in this work are used only for prediction. Thus the following discussion will concentrate on the aspects of empirical modeling relevant to problems of prediction.

3.1 Prediction

The basic problem is to predict the value of a variable we will call the dependent variable, given the values of a set of variables we will call the independent variables. The independent variables may themselves be arbitrary functions of other variables, but their values must be specified independently of the modeling process under consideration. Let y equal the dependent variable and let the vector $\underline{x}$ equal the set of independent variables. The expected value of y given x,

$$E[y/\underline{x}] = \int_{-\infty}^{\infty} y \; d(F_{y/\underline{x}}) \qquad 3.1$$

where $F_{y/\underline{x}}$ is the conditional probability mass function of y given $\underline{x}$, is probably the most commonly sought predictor of y. However, we usually neither know nor have enough data to estimate $F_{y/\underline{x}}$ for even one value of $\underline{x}$, and thus turn to the device of fitting functions to data to produce an approximate description of $E[y/\underline{x}]$ over a wide range of values for $\underline{x}$. All of the models in this work are developed

by fitting functions of the form

$$y = \underline{x}\,\underline{\beta} + \varepsilon \qquad 3.2$$

where y is the dependent variable

$\underline{x}$ is a 1 x k vector of independent variables

$\underline{\beta}$ is a k x 1 vector of coefficients

k is the number of independent variables in the model, including a constant term,

and ε is a zero mean random disturbance

The vector $\underline{x}$ contains an element with the constant value of 1. We assume that $E[\varepsilon] = 0$ and thus $E[y/x] = \underline{x}\,\underline{\beta}$. We also assume that $E[\varepsilon^2]$ is finite. Note that $V[y/\underline{x}] = E[\varepsilon^2]$ and when the independent variables are taken to have zero variance $V[y] = E[\varepsilon^2]$. We will always assume the variance of the independent variables is not a function of $\underline{\beta}$ or $E[\varepsilon^2]$. $E[z]$ and $V[z]$ are the expected value and variance of z.

Given the form of Equation 3.2 and the assumptions described above, the data samples used to develop a model may be described in the form

$$\underline{y} = \underline{X}\,\underline{\beta} + \underline{\varepsilon} \qquad 3.3$$

where $\underline{y}$ is a n x 1 vector of observations of the dependent variable

$\underline{X}$ is a n x k matrix of observations of the independent variables

$\underline{\beta}$ is a k x 1 vector of coefficients (same vector as in Equation 3.2)

$\underline{\varepsilon}$ is a n x 1 vector of sample disturbances

$$E[\underline{\varepsilon}] = \underline{0} \qquad 3.4$$

$$E[\underline{\varepsilon}\ \underline{\varepsilon}'] = \sigma^2\underline{\Omega} \qquad 3.5$$

σ^2 is a scale factor

$\underline{\Omega}$ is a symmetric positive definite matrix

and n is the number of observations in the sample

$\underline{\varepsilon}$ can account for many types of disturbances in the sample, including observation errors, system identification errors, and random behavior.

Three basic elements of fitting functions to data are identification, estimation, and validation. Identification is choosing the form for the model. Estimation is choosing the coefficients and other parameters for the model. For the class of models described by Equation 3.2, identification is choosing the variables in the vector $\underline{x}$ and estimation is choosing the vectors $\hat{\underline{\beta}}$ and $\hat{\underline{\varepsilon}}$, the matrix $\hat{\underline{\Omega}}$, and the scalar $\hat{\sigma}^2$. Validation is deciding if the proposed model is adequate.

Whenever $\underline{\beta}$ is estimated we will assume that the rank of $\underline{X}$ equals k and that k is less than n. When k is less than n and the rank of $\underline{X}$ does not equal k, at least one of the independent variables is a linear combination of the other independent variables, and is thus redundant. When k is greater than n the system of equations represented by Equation 3.3 has an infinite number of solutions and $\hat{\underline{\beta}}$ cannot be determined without assuming values for at least k-n values of $\hat{\underline{\beta}}$. When k equals n there is a unique solution for $\hat{\underline{\beta}}$. When k is

less than n, the system is overdetermined and no $\hat{\underline{\beta}}$ will satisfy the system exactly. Thus $\underline{\beta}$ must be estimated according to some criteria.

Identification is discussed in Section 3.2, estimation is discussed in Section 3.3, validation is discussed in Section 3.4, and prediction with linear models is discussed in Section 3.5.

3.2 Identification

The problem of identifying the best model for predicting the value of a given dependent variable has three parts. First, the independent variables to be measured or otherwise generated must be chosen. We will call these the original variables. Second, the complete set of independent variables to be considered for inclusion in the model must be developed. This set of variables can include both original variables and functions of the original variables. The dependent variable may also be transformed. For example, logarithms and reciprocals of the original variables, including the dependent variables, are often used in place of the original variables in econometric modeling (Johnston, 1972 and Durbin and Watson, 1951). Third, the subset of variables which produce the best model must be chosen.

The first part of the problem, choosing the original variables, is solved through prior knowledge of the system being modeled and is not amenable to general discussion. The original variables used in this work were chosen by the National Weather Service and are described in Chapters 2 and 5. The only variable transformations

discussed in this section are those generated by principal components analysis. Other transformations of the original variables are used in the Group Method of Data Handling (GMDH)(see Chapter 4). The rest of this discussion concentrates on the problem of choosing the subset of variables which should form the model.

Identification of the best subset of variables requires procedures to generate alternative models and measures of quality to rank the alternative models. Hocking (1976), Mosteller and Tukey (1977), Draper and Smith (1966), Chatterjee and Price (1977), and Cox and Snell (1974) present good discussion of the range of procedures which have been proposed both for generating alternative models and for measuring the quality of those models. Only the techniques used in this work are discussed here.

The best method of identifying a model is to know the correct form for the model prior to the beginning of the modeling process. An example of this type of situation is the experimental determination of the head-flow equation for a spillway. The form of the equation is frequently taken to be $Q = KH^{3/2} + e$, where Q is the flow over the spillway, H is the depth of the water over the crest of the spillway, K is the coefficient to be estimated, and e is an error term. If e were always equal to zero only one measurement would be needed to find K. However, boundary effects, properties of real water, and measurement error insure that e is rarely equal to zero. The analyst thus makes several measurements of Q and H, usually over a range of

values for H, and estimates K in whatever manner seems appropriate.

Identification is independent of estimation only when the form of the model is known in advance of the modeling process, as in the example given above. In most other situations identification procedures use values of the estimated sample residuals. Thus the estimation method will influence the model identification. Least squares estimation is used for all coefficient estimation in this work and is discussed in Section 3.3.

The independent variables may generally be divided into 2 groups, those which are known to belong in the model and those whose selection must be guided by the data sample. It is sometimes convenient to remove the effects of the variables in the first group from both the dependent variable and the independent variables in the second group prior to choosing variables from the second group. A typical method of removing the effects of some variables, call them x_k, from another variable, call it z, is to estimate a model of z using the x_k as the independent variables and replace z with the residuals from that model. z can be either a dependent or independent variable.

3.2.1 Principal Components

Principal components are mutually orthogonal linear transformations of the independent variables which have been constructed so that each succeeding component accounts for as much of the variation in the independent variables as possible. The variation of the variables in $\underline{X}$ can be completely represented with r components, where

r is the rank of $\underline{X}$, but a few components frequently account for a substantial portion of the variation. The results of a principal components analysis can be expressed in the form

$$\underline{Z} = \underline{X}^{*}\underline{A} \qquad 3.6$$

where $\underline{Z}$ is a n x c matrix of principal components

$\underline{X}^{*}$ is the n x k matrix of normalized independent variables

$\underline{A}$ is the k x c matrix which defines the transformation

and c is the number of components calculated, $c \leq r$

Each variable in $\underline{X}$ is usually normalized by subtracting the mean and dividing by the standard deviation before the principal components are generated. Normalization prevents variables represented by large numbers from dominating the components. Each column of $\underline{A}$ defines the transformation for one component. Coefficients for the variables in $\underline{X}^{*}$ may be retrieved from the coefficients of the variables in $\underline{Z}$ by using the transformation

$$\hat{\underline{\beta}} = \underline{A}\,\hat{\underline{\beta}}_{c} \qquad 3.7$$

where $\hat{\underline{\beta}}$ is the vector of coefficients for the variables in $\underline{X}^{*}$

and $\hat{\underline{\beta}}_{c}$ is the vector of coefficients of the variables in $\underline{Z}$.

The procedures for finding $\underline{A}$ are given in Kendall (1957), Theil (1971), and Johnston (1972). Example of the use of principle components can be found in Glahn (1962), Jeffers (1967), and Massy (1965).

Massy (1965) and Johnston (1972) suggest two general situations in which a transformation to principal components may be useful. The

first situation is when some of the independent variables which the analyst wishes to include in the model are highly collinear. Some of the problems caused by highly collinear variables are described in Section 3.3 A suggested procedure in this situation is to transform the variables into their principal components, delete the components which account for little or no variation, estimate coefficients for the remaining components, and transform those coefficients to get coefficients for the original variables (Equation 3.7). (Kendall, 1957 and Massy, 1965). The second situation is when it is desired to reduce the number of variables but no individual variables can be chosen for deletion. For this situation, as for the first, it has been suggested that components which account for the most variation be retained as independent variables, although the number of components retained is frequently smaller than for the first situation. (Mosteller and Tukey, 1977). The coefficients are not necessarily transformed back to coefficients on the original variables.

There is no assurance that the most variable principal components will also be good predictors of the dependent variable. Massy (1965) suggested that both the amount of variation explained and the correlation between the components and the dependent variable be examined when choosing which components to retain.

Some properties of the estimated coefficients help explain the selection criteria for principal components. When the least squares estimator of coefficients discussed in Section 3.3 is used, and

c,r, and k are equal, $\hat{\underline{\beta}}$ derived through the transformation given in Equation 3.7 is identical to $\hat{\underline{\beta}}$ calculated directly from $\underline{X}^*$ and is thus a minimum variance unbiased estimator of $\underline{\beta}$. As components are deleted, $\hat{\underline{\beta}}$ derived through Equation 3.7 becomes a biased estimator of $\underline{\beta}$. Greenberg (1975) shows that the coefficients of the components which account for the most variation are the linear combinations of coefficients of the original variables which can be estimated with the least variance, and suggests that the two criteria for retaining components proposed by Massy (1965) represent a tradeoff between increasing the variance and decreasing the bias of the coefficients of the original variables when derived through Equation 3.7. This tradeoff is most important in the first of the two situations described above because correct coefficients for particular variables are desired. In the second situation it is likely that only a few of the original variables are expected to be included in the model and the particular variables used are not significant. In either case, this tradeoff may affect the predictive power of the model.

The use of principal components in this work is closer in spirit to the second than the first of the situations described above. We would like to reduce the number of independent variables by replacing groups of similar variables with one or two representative variables, while retaining as much of the information contained in the original set as possible. A study by Kutzbach (1967) suggests that principal components may perform this function reasonably well for meteorological

variables. The few principal components for each variable subset which explain the most variation are then calculated and retained as independent variables. The correlations between the components and the dependent variable are not used to select components because the purpose of this exercise is not simply to produce as good a prediction model as possible, but rather to see if a good prediction model can be developed from summary variables which represent the various meteorological fields thought to influence temperature. A secondary purpose is to see if the variable selection with different identification techniques stabilizes when summary variables are used. The set of independent variables from which the model is developed then consists of some original variables and some linear transformations of subsets of the original variables. This application of principal components is described further in Section 5.2.2.

A frequent objection to using principal components in modeling is that they are hard to interpret. This objection is raised primarily when the model will be used to make inferences about the process being modeled. When a model is used only for prediction, as in this work, problems of interpretation are not as important. Principal components may even provide some clues for identifying the significant processes in the system which are not provided by other modeling techniques. However, no meteorological interpretations of equations are presented in this work. A potential advantage of using principal components is that model stability may be increased because the influence of errors

in individual variables is reduced

3.2.2. Generation of Alternative Models

Three methods of generating alternative models are used in this work. A variable selection algorithm called stepwise regression is used when the number of variable is large. A set of procedures we will call interactive stepwise regression is used when the number of variables is small. The division between large and small is not precise. Interactive stepwise regression requires more effort per variable and becomes unmanageable when the number of variables is to large. An experienced analyst will be able to choose the appropriate technique. Stepwise and interactive stepwise regression are discussed in this section. The third method of generating alternative models is the GMDH. The GMDH is discussed in Chapter 4.

The stepwise regression algorithm used in this work is that implemented in IMSL subroutine rlsep. (IMSL, 1977). Good descriptions of the stepwise variable selection procedure are given in Draper and Smith (1966) and Efroymson (1960). Briefly, in each step the independent variable having the highest partial correlation with the dependent variable is entered into the model. The hypothesis that there is no change in the sum of squared residuals (RSS)(see Equation 3.15) due to the addition of that variable is then tested using the F distribution. The F test is discussed in Section 3.4.2. If the null hypothesis is rejected at a specified probability level the variable is tentatively retained. The variables currently in the model are then removed one at

a time. The hypothesis that there is no change in RSS due to the removal of each of the variables is tested. Those variables for which that hypothesis is not rejected are removed from the model. The significance levels are specified as tail areas of the F distribution. For example, if the significance parameter for entering variables is 0.05, a variable will not be entered into the model unless the statistic for that variable is at least as large as the 95 percent point of the F distribution. The test for deleting variables can not be more severe than the test for entering variables. The procedure continues until no more variables can be entered or removed at the prespecified significance levels. The resulting equation has then already passed the F test validation procedure. The forward moving stepwise algorithm used by the NWS to choose variables for the Perfect Prog and Model Output Statistics models also selects variables in the order of their partial correlation with the dependent variable. However, a forward moving algorithm never removes a variable once entered into the model.

Stepwise regression is sometimes used to produce a single equation from a set of data. In this work stepwise regression is used mainly as a tool to sift through variables quickly and generate what are presumably reasonable alternative models. Different models can be developed by varying the significance levels for entering and deleting variables and by varying the portion of the data sample used to guide the variable selection. Those models are then examined using

the techniques discussed in Sections 3.2.3 and 3.4.

There are no standard techniques or references for what we call interactive stepwise regression. Rather, interactive stepwise regression is just a convenient name for a set of tools which may be used for model identification.

The tools used in this work are F ratios for entering and deleting variables, normal plots of residuals, the squared multiple correlation coefficient (R^2), the analysis of variance table (ANOVA), partial residual plots, and plots of residuals against time, the independent variables, and the predicted value of the dependent variable. Interactive stepwise regression crosses the boundaries within which we have chosen to discuss empirical modeling because validation procedures are used to guide the identification process rather than just applied to a proposed model. Validation procedures generally require relatively substantial amounts of effort and make interactive stepwise regression unsuitable for use on a large number of variables. F ratios and normal plots of residuals are discussed in Section 3.4.2. R^2 was examined, but was not emphasized for the reasons discussed in Section 3.2.3. ANOVA is discussed in Draper and Smith (1966) and includes the value of RMS_k, which is discussed in Section 3.2.3. Both the partial residual plots and the other residual plots are discussed in Section 3.4.1. All the procedures for interactive stepwise regression were implemented on the Consistent System (Laboratory of Architecture and Planning, 1978) and are described in Appendix C.

which in turn is implemented on the Honeywell Multics computer system at MIT.

3.2.3 Choosing Among Alternative Models

Choosing the best model from a specified set of alternative models is not a well defined procedure. The basic criterion for model quality is accurate prediction. When the correct form for the model is known in advance, as in the head-flow modeling example described in Section 3.2, previous experience with models of the same form inspires confidence in the future performance of the model. However, when the model identification is guided by primarily the data, confidence in future performance must also be guided primarily by the data. Several measures of model quality based on sample data are discussed in this section.

When estimating the future performance of a model we need to assume that both the relations between the dependent and independent variables and the relations between the independent variables will not change. When the number of variables in a model increases, the number of relations between independent variables increases and the probability that some of those relations will change also increases. Thus, as a second criterion for model quality, we prefer to keep the number of variables in the model as small as possible.

Some measures of model quality which are based on the data used to estimate the coefficients can be improved by simply adding more variables to the model. In the extreme case when n = k the data sample will be fit exactly and every sample residual will equal zero. Only

quality measures which include some adjustment for the number of variables in the model are used in this work. However, preference for models with fewer variables is retained as a separate evaluation criterion even though adjustments for the number of variables are included in the quality measures. The tradeoff between the value of the quality measure and the number of variables in the model is usually subjective.

The quality of forecasting models is best judged from data which were not used to estimate the model coefficients. Such data are called independent data. Independent data should not be confused with independent variables. Snee (1977) discusses several methods of choosing independent data and presents a general purpose data splitting algorithm. Mosteller and Tukey (1977) discuss the use of more than one independent data set. These methods are particularly useful when few data points are available. In this work the data are simply divided by years. For example, when 5 years of data are available, the first 3 years are used to estimate coefficients and the last 2 years are used as independent data. Quality measures based on independent data are not adjusted for the number of variables in the model. Preference for fewer variables is still used as separate choice criterion.

Numerous measures of model quality have been proposed in the literature. Hocking (1976), in an excellent summary article on model identification, listed eight commonly used quality measures, which he called criteria functions, and briefly discussed their use. Hocking

was careful to note that the properties of the various criteria functions have not been well established and firm rules which specify the best criteria function for a given situation do not exist. The choice of a quality measure is thus left to the judgement of the analyst.

The following seven quality measures were evaluated for the models examined in this work:

1) the mean squared residual, $RMS_k = RSS/(n-k)$ 3.8

2) the average prediction variance, $J_k = RMS_k(n+k)/n$ 3.9

3) the total squared error, $C_k = (RSS/\sigma^2) + 2k - n$ 3.10

4) the average prediction mean squared error, $S_k = RMS_k/(n-k)$ 3.11

5) the mean absolute residual, $RMA_k = RSA/(n-k)$ 3.12

6) the mean squared residual over independent data IRMS =

$$IRSS/(s-1) \qquad 3.13$$

7) the mean absolute residual over independent data, IRMA = IRSA/s 3.14

where

$$RSS = \sum_{i=1}^{n} (\hat{\varepsilon}_i)^2 \qquad 3.15$$

$$IRSS = \sum_{i=1}^{s} (\hat{\varepsilon}_i)^2 \qquad 3.16$$

$$RSA = \sum_{i=1}^{n} |\hat{\varepsilon}_i| \qquad 3.17$$

$$IRSA = \sum_{i=1}^{s} |\hat{\varepsilon}_i| \qquad 3.18$$

n is the number of observations in the data set used to estimate the coefficients

s is the number of observations in the independent data set

k is the number of coefficients estimated from the data, including the constant term,

and σ^2 is $Var(y/\underline{x})$ for the correct model

Hocking (1976) discusses the use of RMS_k, J_k, C_k, and S_k and provides references where further information about these functions can be found. Among these 4 functions, RMS_k and C_k are emphasized while J_k and S_k are considered supplementary. RMA_k is evaluated because temperature forecasts are frequently judged by their absolute errors (see Section 2.5). RMS_k, J_k, C_k, S_k, and RMA_k are calculated from the data used to estimate the coefficients. IRMS and IRMA are calculated from independent data. IRMS and IRMA are 2 of the statistical forecast accuracy measures mentioned in Section 2.5 The 3 others mentioned in Section 2.5, correlation between observed and predicted values of the dependent variable, number of large errors, and forecast bias, are not evaluated in this work. IRMS and IRMA are the most important of the 7 quality measures because they are the most direct measures of prediction accuracy and frequently reflect the problems caused by collinear variables. The problems caused by collinear variables are discussed in Section 3.3. Note that quality measures based on independent data are also used for model validation (Snee, 1977), but are used in this work primarily for guiding model identification.

A special problem associated with the use of C_k is that σ^2

must be estimated. Draper and Smith (1966) suggest that RMS approaches σ^2 as the number of variables in the model increases, provided all the important variables are in the model and there are an adequate number of observations. Hocking (1976) and Daniel and Wood (1971) similarly suggest that $\hat{\sigma}^2$ be taken as the RMS_k resulting when all the important independent variables are entered in the model. In this work the value of $\hat{\sigma}^2$ used in C_k is approximately the lowest value of RMS_k from the various models which were generated (see Chapters 5 and 6).

Perhaps the two most commonly used measures of model quality are the squared multiple correlation coefficient, R^2, and the adjusted squared multiple correlation coefficient, $\bar{R}^2$. R^2 is defined as

$$R^2 = [\sum_{i=1}^{n} (y_i-\bar{y})^2 - \sum_{i=1}^{n} (y_i-\hat{y}_i)^2]/\sum_{i=1}^{n} (y_i-\bar{y})^2 \qquad 3.19$$

and $\bar{R}^2$ is defined as

$$\bar{R}^2 = 1-(n-1)(1-R^2)/(n-k) \qquad 3.20$$

Neither R^2 nor $\bar{R}^2$ is used in this work because R^2 is not adjusted for the number of variables in the model and $\bar{R}^2$, for a given data set, provides no information not provided by RMS_k. However, the R^2 statistic is available in the interactive stepwise regression package used in this work. (see Section 3.2.2)

3.3 Estimation

Consider now the problem of finding $\hat{\underline{\beta}}$, given a data sample, without specifying the purpose of the model. $\hat{\underline{\varepsilon}}$ is determined along

with $\hat{\underline{\beta}}$ through equation 3.3 as

$$\hat{\underline{\varepsilon}} = \underline{y} - \underline{X}\,\hat{\underline{\beta}} \qquad 3.21$$

We will constrain $\hat{\underline{\beta}}$ to be linear, unbiased, and have the minimum variance among linear unbiased estimators of $\hat{\underline{\beta}}$. Linearity implies $\hat{\underline{\beta}} = \underline{c}'\underline{y}$, Unbiasedness implies $E[\hat{\underline{\beta}}] = \underline{\beta}$, and minimum variance implies $E[(\hat{\beta}_i - E[\hat{\beta}_i])^2] \leq E[(\theta_i - E[\theta_i])^2]$, $i = 1,\ldots,k$ where θ_i is any linear unbiased estimator of β_i.

The Gauss-Markov theorem states that the linear unbiased estimator of $\underline{\beta}$ which minimizes $\sum_{i=1}^{n} (\hat{\varepsilon}_i)^2$ has the smallest variance of any linear unbiased estimator of $\underline{\beta}$. This estimator is called the least squares estimator. A proof of the Gauss-Markov theorem is given in Meyer (1975).

The proof of the Gauss Markov theorem is usually given in two parts. First, the theorem is proved for the case when $\underline{\Omega} = \underline{I}$. $\underline{\Omega} = \underline{I}$ implies the system disturbances are independent and have equal variances. σ^2 need not be known. The least squares estimate for this situation is

$$\hat{\underline{\beta}} = (\underline{X}'\underline{X})^{-1}\,\underline{X}'\underline{y} \qquad 3.22$$

where $\underline{X}'$ is the transpose of $\underline{X}$

Second, it is shown that when $\underline{\Omega} \neq \underline{I}$ the data may be transformed by multiplying both sides of Equation 3.3 by $\underline{P}^{-1}$, where $\underline{PP}' = \underline{\Omega}$, to produce a model whose disturbances are independent and have equal variances, to which the first part of the proof may be applied.

The least squares estimator for this more general situation is

$$\hat{\underline{\beta}} = (\underline{X}'\underline{\Omega}^{-1}\underline{X})^{-1}\ \underline{X}'\underline{\Omega}^{-1}\underline{y} \qquad 3.23$$

The same estimator will be derived in Section 3.5 by minimizing the expected prediction variance.

$\underline{\beta}$ may be estimated either by using Equation 3.23 on the original data or by using Equation 3.22 on transformed data. The latter method is used in this work because most computer regression packages are based on the estimator given by Equation 3.22.

The covariance matrix for $\hat{\underline{\beta}}$ can be found from the rule for propagation of errors (Meyer, 1975).

$$E[\hat{\underline{\beta}}\ \hat{\underline{\beta}}'] = \sigma^2(\underline{X}'\underline{\Omega}^{-1}\underline{X})^{-1} \qquad 3.24$$

The independent variables must have zero variance for Equation 3.24 to apply.

The magnitude of the elements of $(\underline{X}'\underline{\Omega}^{-1}\underline{X})^{-1}$ increase as $(\underline{X}'\underline{\Omega}^{-1}\underline{X})$ approaches singularity. Thus, as the linear dependencies among the independent variables increase, the precision with which $\underline{\beta}$ may be estimated decreases. Chatterjee and Price (1977) and Farrar and Glauber (1967) discuss the problems which may be caused by such multicollinearity. Multicollinearity nearly always exists in real data sets. The problem for the analyst is thus not to discover if multicollinearity exists, but to determine if it causes problems in the application of a model. Multicollinearity is clearly a problem when the model coefficients will be used to make inferences about the system

being modeled. When the model is to be used only for prediction, the problems caused by multicollinearity are harder to define because the coefficients of particular variables are not an important product of the modeling process. Multicollinearity also increases the difficulty of identifying the correct model and makes the estimated coefficients very sensitive to particular data samples. These problems can sometimes be remedied by gathering more data or restricting the coefficients based on prior knowledge (Johnston, 1972). Prediction with models based on data always depends on the assumption that the relations represented by the data, and presumably captured in the model, will continue to apply in the future. Thus, even if some of the independent variables in a model are nearly perfectly collinear and the variances of the estimated coefficients of these variables are large, the model may be able to predict well if the same relations as exist in the data sample hold in the future. However, we prefer models without such highly collinear variables because their presence increases the effect on model performance of the stability of the relations between the independent variables. We see from Equation 3.24 that the value of σ^2, while not needed to calculate $\hat{\underline{\beta}}$, is needed to calculate the covariance matrix of $\hat{\underline{\beta}}$. Recall that σ^2 was also needed to calculate C_k in Section 3.2.3. Meyer (1975) shows that an unbiased estimate of σ^2 is given by

$$\hat{\sigma}^2 = \frac{\hat{\underline{\varepsilon}}' \underline{\Omega}^{-1} \hat{\underline{\varepsilon}}}{n-k} \qquad 3.25$$

When the data have been transformed prior to estimating $\underline{\beta}$, the unbiased

estimate of σ^2 is

$$\hat{\sigma}^2 = \frac{\hat{\underline{\varepsilon}}'\hat{\underline{\varepsilon}}}{n-k} \qquad 3.26$$

Note that this estimate of σ^2 is equal to RMS_k. The difference between the estimate of σ^2 in Equations 3.25 and 3.26 and the estimate of σ^2 desired for use in C_k is that the value being estimated for C_k is for a somewhat ficticious true model and the value being estimated by Equations 3.25 or 3.26 is for the particular model for which the coefficients have been estimated.

The matrix $\underline{\Omega}$ is usually not known. It may be estimated by using the sample residual estimates generated by assuming $\underline{\Omega} = \underline{I}$. (Draper and Smith 1966, Goldberger 1964, and Theil 1971). The sample residual estimates may be used to estimate $\underline{\Omega}$ in two ways. First, standard techniques for the estimation of a covariance matrix may be used if there are sufficient replicated observations. Replicated observations are multiple observations of the dependent variable for a given set of values of the independent variables. Suppose there are s sets each containing r replicated observations. Then

$$E[\varepsilon_i \varepsilon_j] = \frac{1}{r-1} \sum_{\ell=1}^{r} (y_{i\ell} - \underline{x}_i\hat{\beta})(y_{j\ell} - \underline{x}_j\hat{\beta}), \quad i,j = 1,\ldots,s \qquad 3.27$$

where $y_{i\ell}$ is the ℓ^{th} replicated observation in the i^{th} set

$\underline{x}_i$ is the i^{th} set of values of the independent values

and $\hat{\beta}$ is calculated assuming $\underline{\Omega} = \underline{I}$

It is rare to have enough replicated observations to use this technique.

In some cases the observations may be grouped to form sets of approximately replicated observations (Theil, 1971). When replicated observations are not available the residuals may be used to estimate the parameters of an assumed form for $\underline{\Omega}$. For example, if the system disturbances are assumed to be generated by the first order auto-regressive process

$$\varepsilon_t = \rho\, \varepsilon_{t-1} + \nu_t \qquad 3.28$$

where

$$E[\varepsilon_t^2] = \sigma^2 \qquad 3.29$$

$$E[\nu_t] = 0 \qquad 3.20$$

and

$$E[\nu_t^2] = (1-\rho^2)\sigma^2 \qquad 3.31$$

then $\underline{\Omega}$ takes the following form

$$\underline{\Omega} = \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \vdots & & & & \vdots \\ \rho^{n-1} & & \cdots & & 1 \end{bmatrix} \qquad 3.32$$

and ρ, the lag one correlation coefficient, could be estimated from the estimated sample residuals as

$$\hat{\rho} = \frac{1}{n} \sum_{i=1}^{n-1} (\hat{\varepsilon}_i \hat{\varepsilon}_{i+1}) \qquad 3.33$$

A more general iterative procedure for estimating $\underline{\Omega}$ while simultaneously identifying the equation is described in Mosteller and Tukey (1977) under the heading resistant stepwise fitting.

Note that the only assumption beyond those of Section 3.1 which

is used in this section is the assumption of zero variance independent variables required for Equation 3.24. However, the Gauss-Markov Theorem applies only when $\underline{\Omega}$ is correctly specified. Examination of the residuals for independence and constant variance is discussed in Section 3.4.

Sets of constraints on $\hat{\underline{\beta}}$ other than linearity, unbiasedness, and minimum variance may be specified. For example, in the procedure called ridge regression the estimator of $\underline{\beta}$ is allowed to be biased in expectation of reducing its variance (Hocking, 1976 and Hoerl and Kennard, 1970). Other methods of fitting a function to data may be found in Tukey (1977) and Mosteller and Tukey (1977). Only the least squares estimator is used in this work. The primary reasons for this restriction are that only the least squares estimator is commonly available on computer systems, and the constraints on and properties of $\hat{\underline{\beta}}$ calculated from the least squares estimator, while not ideal for every situation, are generally desirable.

3.4 Validation

Model validation is perhaps the fuzziest of the three steps in model building. Validation procedures never provide direct measures of model quality, but rather provide indications of possible problems. A model thus passes the validation procedures when no serious problems are indicated.

The three types of validation procedures used in this work are hypothesis tests, graphic analysis, and stability analysis. Hypothesis

tests depend on assumptions about the distribution of the residuals. Graphic analysis is used to examine the residuals directly. Stability analysis ean be used to examine many aspects of models and does not usually depend on assumptions. Graphic analysis is discussed in Section 3.4.1, hypothesis tests are discussed in Section 3.4.2, and stability analysis is discussed in Section 3.4.3.

3.4.1 Graphic Analysis

Graphic analysis is generally the most effective way to examine the sample residuals. Various statistics for examining residuals have been proposed in the literature. Draper and Smith (1966) list several such statistics and provide references where further information on these statistics may be found, but suggest that problems severe enough to require correction are nearly always revealed through the appropriate graphic examination. Good general discussions of the use of residual plots are given by Draper and Smith (1966), Chatterjee and Price (1977), Anscombe (1973), and Cox and Snell (1968).

The two basic requirements for residuals are that they lack structure and have equal variances. These two specifications are motivated by the conditions for validity of the Gauss-Markov theorem described in Section 3.3, but they also relate to correct model identification. When these two conditions are not met we suspect that either the model was not identified correctly or that $\underline{\Omega}$ was not estimated well.

Three of the most common types of structure in residuals are variation with time, variation with the magnitude of the dependent variable, and variation with the magnitudes of the independent variables. Residuals may be plotted in time sequence, against the estimated values of the dependent variables, and against the values of the independent variables to check for such variations and simultaneously check the constancy of the residual variance. Since only pattern, and not overall magnitude, is being examined in validation procedures, standardized residuals are often used in the plots. Some common residual patterns which indicate the presence of probably absence of problems and the interpretation of those patterns in the various plots are presented in Draper and Smith (1966).

In all the residual plots mentioned above the desirable pattern is usually considered to be a horizontal band of constant width. However, the variance of sample residuals varies with the values of the independent variables, being smallest toward the centroid of the variables and increasing towards the perimeters, even when the model assumptions are correct (Behnken and Draper, 1972). Thus, the expected pattern for a correct model is not precisely a horizontal band of constant width. Behnken and Draper (1972) suggest, however, that in many situations, particularly when k/n is small, failing to account for the expected variation in the sample residual variance does not have a large effect on the inferences drawn from graphic examination of residuals.

Another type of residual plot which is particularly useful in

detecting identification errors is the partial residual plot. Partial residuals and partial residual plots are discussed by Mosteller and Tukey (1977) and Larsen and McCleary (1972). The partial residuals of the dependent variable (pry_i) are the sample residuals of a model which does not contain the independent variable $\underline{x}_i$. The partial residuals of the independent variable $\underline{x}_i$(prx_i) are the residuals created by removing from $\underline{x}_i$ the effects of the independent variables already in the model. The relation between $\underline{x}_i$ and y, when the effects of the other variables have been removed, is revealed by a plot of pry_i against prx_i. If we assume the other variables are in the model correctly, this relation should be linear. When pry_i is modeled as a function of prx_i, the coefficient on prx_i is the same as the coefficient on $\underline{x}_i$ in the whole model (Mosteller and Tukey, 1977). Thus, drawing the line with slope $\underline{\beta}_i$ on the plot of pry_i against prx_i can reveal the possible influence on $\underline{\beta}_i$ of a few outlying data points. For example, if the cloud of points is oriented in one direction, but regression line does not follow that trend, one should suspect that $\underline{\beta}_i$ is not being estimated correctly, even if $\underline{x}_i$ does belong in the model. Two common causes of poor estimation are multicollinearity and outlying data points. Some partial residual plots are shown in Figures 5.7, 5.8, and 5.9.

Other residual plots may be useful. For example, when the data may be divided into a few categories, separate graphs for each category may reveal patterns not detected when the categories are aggregated.

Ideally, every model which seems promising should be subjected to graphic validation procedures. Even though validation has been separated from identification for the purpose of this discussion, graphic analysis of residuals can be used to help guide the identification process and the estimation of $\underline{\Omega}$. Recall that validation techniques are part of the procedure we call interactive stepwise regression in Section 3.2.2. Unfortunately, graphic examination of residuals requires more effort, both to produce and to interpret, than most of the identification techniques discussed in Section 3.2. Graphic examination also cannot be used to rank the equations. Thus, except for their use in interactive stepwise regression, graphic validation techniques are used primarily to examine a few of the best equations chosen in the identification process. The validation techniques may then be used to guide modifications of those equations, if necessary.

3.4.2 Hypothesis Tests

The two hypothesis tests used in this work are the F test and the Durbin-Watson test.

The F distribution is used to test hypotheses comparing two sums of squared residuals for different models. Meyer (1975) presents a good discussion of general hypothesis testing and of the F distribution. Chatterjee and Price (1977) describe the use of the F test on regression models. The typical null hypothesis is that the sums of squared residuals from two different models are equal. The statistic used to test this hypothesis is

$$F = \frac{(RSS_1 - RSS_2)/(k_1 - k_2)}{RSS_1/(n-k_1)} \qquad 3.34$$

where the subscripts 1 and 2 distinguish between the two different models and k_1 and k_2 are the numbers of variables in the 2 models.

The F distribution applies to the ratio of 2 variables, each of which has a chi-square distribution. Thus the F test applies only when $E[\underline{\varepsilon\varepsilon}'] \sim N(0, \sigma^2 \underline{I})$, that is, the residuals are independent multi-normally distributed variables with a common variance. This is a severe requirement on the distribution of the residuals and the F test should thus be used with caution. If desired, the normality of the distribution of the residuals may be examined by plotting the cumulative distribution of residuals against a scale which has been distorted according to a normal distribution. Such plots are discussed in Daniel and Wood (1971). Some normal plots are shown in Figures 6.5, 6.6, and 6.7. In this work the F test is used as part of both the automatic and interactive stepwise regression algorithms.

The Durbin-Watson statistic is used to test for lag one serial correlation in the residuals. The basic information about this test is in Durbin and Watson (1950, 1951) and further discussion of its use is given by Theil (1971). The statistics tested is

$$d = \sum_{i=1}^{n-1} (\hat{\varepsilon}_{i+1} - \hat{\varepsilon}_i)^2 / \sum_{i=1}^{n} (\hat{\varepsilon}_i)^2 \qquad 3.35$$

The null hypothesis is that the residuals are serially independent. The distribution of d is a function of $\underline{X}$, but upper and lower limits,

labeled d_u and d_ℓ, which are appropriate for any matrix $\underline{X}$ are normally used. To test for positive serial correlation, the null hypothesis is rejected if d is less than a specified point in the distribution of d_ℓ and not rejected if d is greater than a specified point in the distribution of d_u. No inference is drawn when d is between those values. When d is replaced by (4-d) the same procedure tests for negative serial correlation. The procedures developed by Durbin and Watson (1950, 1951) apply to all cases in which the independent variables may be considered to have zero variance, and thus do not apply when lagged values of the dependent variable are included in the model. Since most of the models developed in this work include a lagged value of the dependent variable, the Durbin-Watson test does not strictly apply. It will still be used as an approximate test.

The 5, 2 , and 1 percent significance levels for one tailed tests of d against d_u and $d_\ell$were tabulated by Durbin and Watson (1951) for models with from one to six variables whose coefficients were estimated with between 15 and 100 observations. Unfortunately, all of the models developed in this work are outside of those ranges. The distribution of d used to construct the tables of significance points for d_ℓ and d_u given in Durbin and Watson (1951) is fairly complex, but Durbin and Watson (1950) suggest that d is asymptotically normally distributed for large values of n-k. The equations for the means and variances of d_ℓ and d_u given on page 427 of Durbin and Watson (1950) and the normal distribution were used to construct

k'	$E[d_\ell]$	$E[d_u]$	$V[d_\ell]$	$V[d_u]$	5% d_ℓ	5% d_u	1% d_ℓ	1% d_u
1	1.99	2.01	0.0113	0.0114	1.82	1.83	1.75	1.76
2	1.99	2.01	0.0112	0.0115	1.81	1.84	1.74	1.76
3	1.98	2.02	0.0112	0.0116	1.81	1.84	1.74	1.77
4	1.98	2.02	0.0111	0.0116	1.80	1.85	1.73	1.77
5	1.97	2.03	0.0110	0.0117	1.80	1.85	1.73	1.78
6	1.97	2.03	0.0110	0.0118	1.79	1.86	1.72	1.78
7	1.96	2.04	0.0109	0.0118	1.79	1.86	1.72	1.79
8	1.95	2.05	0.0108	0.0119	1.78	1.87	1.71	1.79
9	1.95	2.05	0.0107	0.0120	1.78	1.87	1.71	1.80
10	1.94	2.06	0.0107	0.0120	1.77	1.88	1.70	1.80
11	1.94	2.06	0.0106	0.0121	1.77	1.88	1.70	1.81
12	1.93	2.07	0.0105	0.0122	1.76	1.89	1.69	1.81
13	1.92	2.08	0.0104	0.0123	1.75	1.90	1.69	1.82
14	1.92	2.08	0.0104	0.0123	1.75	1.90	1.68	1.83
15	1.91	2.09	0.0103	0.0124	1.74	1.91	1.67	1.83
16	1.90	2.10	0.0102	0.0125	1.74	1.91	1.67	1.84
17	1.90	2.10	0.0101	0.0126	1.73	1.92	1.66	1.84
18	1.89	2.11	0.0100	0.0126	1.73	1.92	1.66	1.85
19	1.89	2.11	0.0100	0.0127	1.72	1.93	1.65	1.85
20	1.88	2.12	0.0099	0.0128	1.72	1.93	1.65	1.86
21	1.87	2.13	0.0098	0.0129	1.71	1.94	1.64	1.86
22	1.87	2.13	0.0097	0.0129	1.70	1.95	1.64	1.87
23	1.86	2.14	0.0096	0.0130	1.70	1.95	1.63	1.87
24	1.85	2.15	0.0095	0.0131	1.69	1.96	1.63	1.88
25	1.85	2.15	0.0095	0.0132	1.69	1.96	1.62	1.89
26	1.84	2.16	0.0094	0.0133	1.68	1.97	1.62	1.89
27	1.83	2.17	0.0093	0.0133	1.68	1.98	1.61	1.90
28	1.83	2.17	0.0092	0.0134	1.67	1.98	1.60	1.90
29	1.82	2.18	0.0091	0.0135	1.66	1.99	1.60	1.91
30	1.81	2.19	0.0090	0.0136	1.66	1.99	1.59	1.91
31	1.81	2.19	0.0089	0.0137	1.65	2.00	1.59	1.92
32	1.80	2.20	0.0088	0.0137	1.65	2.01	1.58	1.93
33	1.79	2.21	0.0088	0.0138	1.64	2.01	1.58	1.93
34	1.79	2.21	0.0087	0.0139	1.63	2.02	1.57	1.94
35	1.78	2.22	0.0086	0.0140	1.63	2.02	1.57	1.94

Table 3.1 Mean, Variance, and Significance Points for d_ℓ and d_u, Durbin-Watson Statistic.
n = 350

Table 3.1. Table 3.1 is for n = 350.

3.4.3 Stability Tests

Testing model stability is perhaps the most subjective, but also the most robust, of the 3 types of validation procedures. Model stability can be tested in many ways, but the basic concept is to examine changes in model parameters with changes in the data used to estimate those parameters. Even model quality statistics based on independent data are a type of model stability test when compared with similar statistics based on the estimation data and thus, as mentioned in Section 3.2.3, are often used for model validation.

In this work the changes in the model coefficients with changes in the estimation data are examined. There are no definite acceptability criteria for this type of analysis, but in general we prefer that coefficients do not change sign or order of magnitude as the estimation data are changed.

3.5 Prediction with Linear Models (after Goldberger (1962))

The unbiased linear predictor of the value of a dependent variable, y_{n+s}, which has the smallest variance among all linear unbiased predictors of y_{n+s}, is sought. A linear predictor P has the form

$$P = \underline{c}'\underline{y} \qquad 3.36$$

where c is a n x 1 vector of constants and $\underline{y}$ is the n x 1 vector of

observations. An unbiased predictor has the property that

$$E[P_{n+s} - y_{n+s}] = 0 \qquad 3.37$$

where P_{n+s} is the predictor of y_{n+s}. Minimum variance implies that $E[(P_{n+s} - y_{n+s})^2]$ is minimized.

The subscript n+s generally indicates the time period for which the predictor is sought. Subscripts greater than n could also simply indicate variables not in the data sample, regardless of their time of occurence.

The basic model and data have the form described in Section 3.1. Thus,

$$y_{n+s} = \underline{x}_{n+s}\,\underline{\beta} + \varepsilon_{n+s} \qquad 3.38$$

Assume that

$$E[\varepsilon^2_{n+s}] = \sigma^2_{n+s} \qquad 3.39$$

and

$$E[\varepsilon_{n+s}\,\hat{\underline{\varepsilon}}] = \underline{\omega} \qquad 3.40$$

$\underline{\omega}$ is the vector of covariances between the system disturbance at time n+s and the sample disturbance estimates.

Combining Equations 3.3, 3.36, and 3.38 we have

$$P_{n+s} - y_{n+s} = (\underline{c}'\underline{X} - \underline{x}_{n+s})\,\underline{\beta} + \underline{c}'\underline{\varepsilon} - \varepsilon_{n+s} \qquad 3.41$$

Taking the expected value of both sides of Equation 3.41 and using Equation 3.37 shows that $c'\underline{X} = x_{n+s}$. Equation 3.41 can thus be simplified to

$$P_{n+s} - y_{n+s} = \underline{c}'\underline{\varepsilon} - \varepsilon_{n+s} \qquad 3.42$$

The vector $\underline{c}$ is then found by minimizing the expected value of the square of $P_{n+s} - y_{n+s}$ subject to $\underline{c}'\underline{X} = \underline{x}_{n+s}$. The technique of Lagrange multipliers may be used. Squaring and taking expected values of both sides of Equation 3.42 and using Equations 3.5, 3.39, and 3.40 yields

$$E[P_{n+s} - y_{n+s})^2] = E[(\underline{c}'\underline{\varepsilon} - \varepsilon_{n+s})(\underline{c}'\underline{\varepsilon} - \varepsilon_{n+s})'] = \underline{c}'\underline{\Omega}\,\underline{c} + \sigma^2_{n+s} - 2\underline{c}'\underline{\omega} \qquad 3.43$$

Note that $\underline{c}'\underline{\varepsilon} = \underline{\varepsilon}'\underline{c}$. Define

$$g = \underline{c}'\underline{\Omega}\,\underline{c} + \sigma^2_{n+s} - 2\underline{c}'\underline{\omega} - 2(\underline{c}'\underline{X} - \underline{x}_{n+s})\underline{\lambda} \qquad 3.44$$

where $\underline{\lambda}$ is the k x 1 vector of Lagrange multipliers and the term containing $\underline{\lambda}$ has been multiplied by 2 to facilitate later manipulations. Setting the derivatives of g with respect to $\underline{c}$ and $\underline{\lambda}$ equal to zero gives

$$\begin{bmatrix} \underline{\Omega} & \underline{X} \\ \underline{X}' & \underline{0} \end{bmatrix} \begin{bmatrix} \underline{c} \\ -\underline{\lambda} \end{bmatrix} = \begin{bmatrix} \underline{\omega} \\ \underline{x}'_{n+s} \end{bmatrix} \qquad 3.45$$

The solution to equation 3.45 is

$$\begin{bmatrix} \underline{c} \\ -\underline{\lambda} \end{bmatrix} = \begin{bmatrix} \underline{\Omega}^{-1}\,[\underline{I} - \underline{X}(\underline{X}'\underline{\Omega}^{-1}\underline{X})^{-1}\underline{X}'\underline{\Omega}^{-1}] & \underline{\Omega}^{-1}\underline{X}\,(\underline{X}'\underline{\Omega}^{-1}\underline{X})^{-1} \\ (\underline{X}'\underline{\Omega}^{-1}\underline{X})^{-1}\,\underline{X}'\underline{\Omega}^{-1} & -(\underline{X}'\underline{\Omega}^{-1}\underline{X})^{-1} \end{bmatrix} \begin{bmatrix} \underline{\omega} \\ \underline{x}'_{n+s} \end{bmatrix} \qquad 3.46$$

See Goldberger (1964) for the inverse of a partitioned matrix.

Thus, using Equations 3.36 and 3.46,

$$P_{n+s} = \underline{x}_{n+s}\,\hat{\underline{\beta}} + \underline{\omega}'\underline{\Omega}^{-1}\hat{\underline{\varepsilon}} \qquad 3.47$$

where $$\hat{\underline{\beta}} = (\underline{X}'\underline{\Omega}^{-1}\underline{X})^{-1}\,\underline{X}'\underline{\Omega}^{-1}\underline{y} \qquad 3.48$$

and $$\hat{\underline{\varepsilon}} = \underline{y} - \underline{X}\,\hat{\underline{\beta}} \qquad 3.49$$

The vector $\hat{\underline{\beta}}$ can be considered an estimate of $\underline{\beta}$ and the scalar $\underline{\omega}'\underline{\Omega}^{-1}\hat{\underline{\varepsilon}}$ can be considered an estimate of ε_{n+s}. Note that Equation 3.48 is the least squares estimator of $\underline{\beta}$. When $\underline{\omega} = \underline{0}$ our predictor of y_{n+s} is $\underline{x}_{n+s}\,\hat{\underline{\beta}}$.

The loss in efficiency of prediction caused by assuming $\underline{\omega} = \underline{0}$ when $\underline{\omega} \neq \underline{0}$ is measured by the difference between the prediction variances when $\underline{c}$ is calculated with and without $\underline{\omega} = \underline{0}$. From Equation 3.43 we can write the difference in prediction variances as

$$\sigma^2_{p_o} - \sigma^2_{p_1} = (\underline{c}_o'\underline{\Omega}\underline{c}_o + \sigma^2_{n+s} - 2\underline{c}_o'\underline{\omega}) - (\underline{c}_1'\underline{\Omega}\underline{c}_1 + \sigma^2_{n+s} - 2\underline{c}_1'\underline{\omega}) =$$

$$\underline{\omega}'\underline{\Omega}^{-1}(\underline{I} - \underline{X}(\underline{X}'\underline{\Omega}^{-1}\underline{X})^{-1}\underline{X}'\underline{\Omega}^{-1})\underline{\omega} \qquad 3.50$$

where $\sigma^2_{p_o}$ is the prediction variance using $\underline{\omega} = \underline{0}$, and $\sigma^2_{p_1}$ is the prediction variance using $\omega \neq 0$. c_o is calculated from Equation 3.46 with $\underline{\omega} = \underline{0}$ and c_1 is calculated from equation 3.46 with $\underline{\omega} \neq \underline{0}$. The right hand term in Equation 3.50 is the inner product of $(\underline{H}\underline{\omega} - \underline{H}\underline{X}(\underline{X}'\underline{\Omega}^{-1}\underline{X})^{-1}\underline{X}'\underline{\Omega}^{-1}\underline{\omega})$ with itself, where $\underline{H}\underline{H}^T = \underline{\Omega}^{-1}$, and is thus a non-negative quantity. Since $\underline{\omega}$ is the only term in equation 3.50 which varies with s, the loss of efficiency from neglecting $\underline{\omega}$ decreases as the correlation between the sample residual estimates and the

prediction disturbance decreases.

$\underline{\omega}$, σ^2, and $\underline{\Omega}$ are not usually known and therefore must be estimated. $\underline{\omega}$ cannot usually be estimated unless a structure is assumed. Goldberger (1962) presents an example in which the sample disturbances are generated by the autoregressive process described by Equation 3.28. If the autoregressive process is assumed to continue after the sample data have been collected,

$$\underline{\omega} = \sigma^2 \begin{bmatrix} \rho^{n-1+s} \\ \rho^{n-2+s} \\ \cdot \\ \cdot \\ \cdot \\ \rho^{s} \end{bmatrix} \qquad 3.51$$

The estimation of ρ, σ^2, and $\underline{\Omega}$ were discussed in Section 3.3.

Chapter 4

THE GROUP METHOD OF DATA HANDLING (GMDH)

GMDH is the generic name of a method of empirical model identification developed by Ivakhnenko (1970, 1971 and 1976) which is characterized by a multilayer structure and self sampling. Partial models are constructed and evaluated in each layer. The GMDH is called self sampling because the output from the partial models in each layer is used as input to the next layer. The process continues until a stopping criterion is met, at which time the complete model is constructed from the partial models. The GMDH was developed primarily for generating complex models from short data records. In this work the GMDH is used on relatively long data records. Figure 4.1 shows the general GMDH algorithm.

The structure of the GMDH is motivated in part by the structure of perceptrons (Ivakhnenko, 1970). Perceptrons are pattern classifying systems used to model neuron networks. Pattern classifying systems are described in Nilson (1965) and perceptrons are described in Rosenblatt (1962), Block (1962), and Block et al. (1962). The GMDH has also been compared to breeding programs in which the specimens with desirable characteristics are cross bred until an optimal mix has been achieved (Ivakhnenko, 1970, 1971, and 1976). In the GMDH, partial models with desirable characteristics are combined to form new partial models until some specified criterion is met.

We will frequently speak of model complexity when discussing

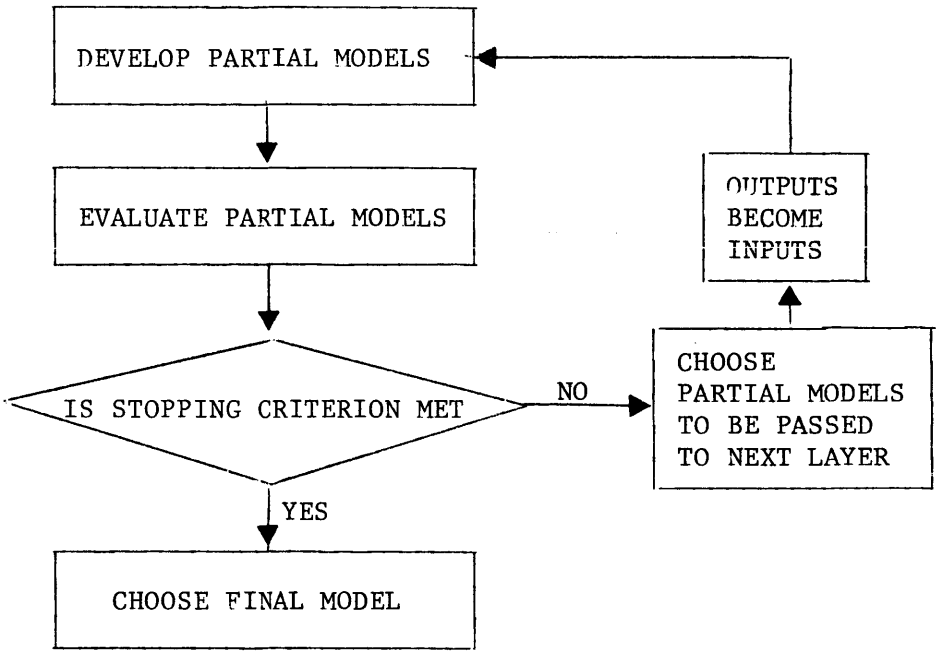

Figure 4.1 Structure of a GMDH Algorithm

the GMDH. Model complexity is not a well defined term, but larger numbers of model parameters or higher order terms generally indicate greater model complexity.

The basic principles of the GMDH are described in Section 4.1.1 and illustrated with an application in Section 4.1.2. Application of the GMDH to polynomial model identification is described in Section 4.2.1 and discussed in Section 4.2.2. The GMDH algorithm used in this work is described in Section 4.3.

4.1 The General GMDH Method

4.1.1 Elements of the GMDH

The three basic elements of a GMDH algorithm are partial models, a partial model quality criterion, and a stopping criterion.

The partial models are complete models in themselves but are called partial in the context of the GMDH because they are used as components of other models in the GMDH process. The partial model structure should be such that the complexity of the overall model structure increases in each layer. However, partial models should also be relatively simple because greater complexity in the partial models reduces the flexibility of the overall process. The partial model structure may vary both within and between layers, but is normally held constant.

The partial model quality criterion is used to rank the partial models. Typically, either all the partial models better than a specified level of quality or a specified number of the best

partial models are chosen to provide input to the next layer. The measure of quality used to rank the partial models should normally be based on independent data. The GMDH uses the same data, transformed through the partial models in the previous layers, to estimate the partial model parameters for each layer. Using the estimation data to also measure model quality might be appropriate when the models are intended only for interpolation, but would probably lead to problems when the models are used for prediction. Ivakhnenko (1969, 1970,and 1971) stresses the idea that measuring partial model quality on independent data filters out potential model components which do not have similar characteristics on the estimation and independent data sets.

The stopping criterion determines the number of layers which are developed. A common stopping criterion is the occurence of the first decrease, relative to the previous layer, of the quality of the best partial model in a layer. The final model is then constructed from the best partial model and the series of partial models which provide input to the chosen model. The number of layers developed may also be constrained.

There is no theory which explains the performance characteristics of the GMDH. Aspects of empirical modeling such as those discussed in Chapter 3 and modeling experience are the only guides to choosing the particular forms of the elements of a GMDH algorithm. However, the large number of partial models generated in most GMDH

algorithms precludes the use of any but the simplest procedures to develop the partial models. The usual strengths of the GMDH relate to reducing data requirements for constructing a model of given complexity and the beneficial aspects of using independent data to measure model quality. The particular advantages and disadvantages of the GMDH vary between applications.

4.1.2 Construction of Transition Probability Tables with the GMDH

A simplified version of an example presented by Ivakhnenko (1969) in which the GMDH is used to construct a transition probability table is presented in this section. Transition probability tables are tables of the probabilities of a system being in a specified condition, given some of the history of the system. The condition with the highest probability of occurence will be referred to as the prediction from a transition probability table. This example is not presented as a recommended method of constructing probability transition tables, but rather as an illustration of the general GMDH approach to problems other than polynomial model identification.

Ivakhnenko discretized the annual flows in the Volga River into three intervals. Flows less than 223 million m^3 are in interval 1, flows between 223 and 255 million m^3 are in interval 2, and flows greater than 255 million m^3 are in interval 3. Let I_t denote the flow interval in year t. The annual flow intervals for 74 years are listed in Table 4.1. Transition probability tables for the annual river flow show the estimated probabilities of being in intervals 1,

Year	I_t	$P1_t$	$P2_t$	$P3_t$	Year	I_t	$P1_t$	$P2_t$	$P3_t$	Year	I_t	$P1_t$	$P2_t$	$P3_t$
1	3	-	-	-	18	1	1	2	1	35	3	3	3	3
2	3	3	-	-	19	3	1	2	1	36	3	3	3	3
3	1	3	3	-	20	2	3	2	2	37	3	3	3	3
4	3	1	1	1	21	2	3	2	3	38	2	3	3	3
5	2	3	2	3	22	3	3	3	2	39	2	3	2	2
6	3	3	2	3	23	2	3	3	3	40	1	3	3	2
7	3	3	3	3	24	1	3	2	2	41	1	1	1	2
8	3	3	3	3	25	3	1	1	1	42	2	1	2	2
9	3	3	3	3	26	1	3	2	3	43	3	3	3	3
10	1	3	3	3	27	1	1	1	1	44	2	3	3	2
11	1	1	1	1	28	3	1	2	1	45	2	3	2	2
12	2	1	2	1	29	2	3	2	2	46	3	3	3	2
13	3	3	3	3	30	1	3	2	3	47	3	3	3	3
14	3	3	3	2	31	1	1	1	1	48	3	3	3	3
15	3	3	3	3	32	2	1	2	2	49	3	3	3	3
16	1	3	3	3	33	2	3	3	3	50	2	3	3	3
17	1	1	1	1	34	3	3	3	3	51	2	3	2	2

Table 4.1 Observed Annual Flow in the Volga River, I_t
Predicted Annual Flows in the Volga River, $P1_t$, $P2_t$, and $P3_t$
Interval 1 QA < 223 million m^3
Interval 2 223 $\leq$ QA $\leq$ 255 million m^3
Interval 3 255 million $m^3 \leq$ QA

Year	I_t	$P1_t$	$P2_t$	$P3_t$	Year	I_t	$P1_t$	$P2_t$	$P3_t$
52	3	3	3	2	69	1	3	3	2
53	1	3	3	3	70	2	1	1	2
54	1	1	1	2	71	1	3	3	2
55	1	1	2	1	72	2	1	1	1
56	1	1	2	1	73	3	3	3	2
57	1	1	2	1	74	1	3	3	2
58	1	1	2	1					
59	1	1	2	1					
60	1	1	2	1					
61	2	1	2	1					
62	2	3	3	3					
63	2	3	3	3					
64	2	3	3	2					
65	1	3	3	2					
66	2	1	1	2					
67	2	3	3	2					
68	2	3	3	3					

Table 4.1 cont'd

2, or 3, given some combination of previous flow intervals.

The first layer partial models are the transition probability tables defined by different combinations of previous flow intervals. The 3 partial models used in the first layer are shown in Tables 4.2, 4.3,and 4.4. The denominators in those tables are the total number of occurences of the previous flow interval pattern which defines the row. The numerators are the number of occurences of the intervals which define the columns. The predicted flow for each previous flow pattern is marked by a star. The first 50 years of data were used to construct the partial models and the quality of the partial models was measured by the percentage of correct predictions on the last 24 years of data. Model 1 predicted 39 percent correctly, model 2 predicted 14 percent correctly, and model 3 predicted 48 percent correctly.

Ivakhnenko (1969) suggests 2 ways of choosing the partial models to provide the input for the next layer. The first is to require W percent accuracy from partial models using one previous flow interval, 2W percent accuracy from those using 2 previous flow intervals, and 3W percent accuracy from those using 3 previous flow intervals. The second is to choose a specified number of the best models from the layer. The 2 best models, numbers 1 and 3, are chosen to provide input to layer 2 in this example.

The predictions of I_t from the first layer models, P1, P2, and P3, are listed in Table 4.1. The second layer partial model is the transition probability table using P1 and P3. This model is

I_t \ I_{t+1}	1	2	3
1	6/13*	3/13	4/13
2	3/14	5/14	6/14*
3	4/23	7/23	12/23*

Table 4.2 Model 1, Layer 1

$I_{t-1}I_t$ \ I_{t+1}	1	2	3
11	1/6	3/6*	2/6
12	0/3	1/3	2/3*
13	1/4	3/4*	0/4
21	2/3*	0/3	1/3
22	1/4	0/4	3/4*
23	0/6	2/6	4/6*
31	3/4*	0/4	1/4
32	2/7	4/7*	1/7
33	3/12	2/12	7/12*

Table 4.3 Model 2, Layer 1

I_{t-2}, I_t \ I_{t+1}	1	2	3
11	1/2*	0	1/2
12	1/7	2/7	4/7*
13	0/4	3/4*	1/4
21	1/3	2/3*	0/3
22	1/2	1/2*	0/2
23	1/8	1/8	6/8*
31	4/8*	1/8	3/8
32	1/5	2/5*	2/5
33	2/9	3/9	4/9*

Table 4.4 Model 3, Layer 1

$P1_t, P3_t$ \ I_t	1	2	3
11	5/10*	1/10	4/10
12	1/3	2/3*	0/3
13	-	-	-
21	-	-	-
22	-	-	-
23	-	-	-
31	-	-	-
32	2/10	5/10*	3/10
33	4/24	6/24	14/24*

Table 4.5 Model 1, Layer 2

shown in Table 4.5. As in the first layer, the first 50 years of data were used to construct the model and the last 24 years of data were used to measure the model accuracy. The layer 2 partial model predicted 46 percent correctly. Other partial models could have been used in both layers, but these illustrate the method.

The stopping criterion in this example is simply to construct 2 layers. The most accurate model, regardless of the layer in which it occurs, is the final model. Thus, model 3 in the first layer is the final model in this example.

4.2 Polynomial Model Identification with the GMDH

The GMDH has been used to develop polynomial models for economic (Ivakhnenko, 1971), environmental (Ikeda, et al., 1976 and Duffy and Franklin, 1975), and mechanical (Inooka and Inoue, 1978) systems. The basic elements of GMDH algorithms for polynomial model identification are described in Section 4.2.1 and their characteristics are discussed in Section 4.2.2.

4.2.1 Elements

A general polynomial of order p using q variables has the form

$$y = \sum_{i_1=1}^{q} \beta_{i_1} x_{i_1} + \sum_{i_1=1}^{q} \sum_{i_2=1}^{q} \beta_{i_1 i_2} x_{i_1} x_{i_2} + \ldots + \sum_{i_1=1}^{q} \sum_{i_2=1}^{q} \ldots \sum_{i_p=1}^{q} \beta_{i_1 i_2 \ldots i_p} x_{i_1} x_{i_2} \ldots x_{i_p} \qquad 4.1$$

The partial models in a GMDH algorithm for polynomial model identification are some portion of Equation 4.1. The most common choice is a second order 2 variable polynomial (p=q=2). This choice provides nonlinear and interaction terms at a relatively low cost in complexity. However, the nonlinear terms still cause the order and number of variables to double in each layer. The partial model structure is usually constant both within and between layers.

The large number of partial models typically developed in GMDH algorithms dictates the use of a least squares estimator for the partial model coefficients. Stepwise regression algorithms can be used to choose terms within the partial models. When all the terms in the partial models are retained the estimation technique is simply multiple regression.

Some partial model quality criteria and methods of choosing independent data are described in Section 3.2.3. Ivakhnenko (1976) discusses three partial model quality criteria and suggests that weighted combinations of the three are appropriate for many modeling situations. One of the 3 criteria is similar to IRMS (see Section 3.2.3), another measures the variation between models developed on different portions of a data set, and the last measures the departure of a model from prior knowledge of the system being modeled.

The partial model predictions of the dependent variables from one layer are used as independent variables in the next layer. A specified number of partial models are generally passed between

all layers to ease the programming of the algorithm on a computer. For example, if there are 10 independent variables in the first layer, 45 second order two variable partial models can be developed in the first layer. If five of those partial models are passed to the second layer, 10 second layer partial models can be developed from the 5 independent variables provided by the 5 first layer partial models. If 5 second layer models are then passed to the third layer, 10 third layer partial models can be developed.

There are generally no features in the stopping criterion which are not mentioned in Section 4.1.1.

A simple example of the process follows.

Consider a data set with 1 dependent variable, $\underline{y}$, and 3 independent variables, $\underline{x}_1$, $\underline{x}_2$, and $\underline{x}_3$. The 3 models given by Equations 4.2, 4.3, and 4.4 can be constructed in the first layer.

$$\hat{y}^1_1 = \hat{\beta}^1_{11}\underline{x}_1 + \hat{\beta}^1_{21}\underline{x}_2 + \hat{\beta}^1_{31}\underline{x}_1^2 + \hat{\beta}^1_{41}\underline{x}_2^2 + \hat{\beta}^1_{51}\underline{x}_1\underline{x}_2 + \hat{\beta}^1_{61} \qquad 4.2$$

$$\hat{y}^1_2 = \hat{\beta}^1_{12}\underline{x}_1 + \hat{\beta}^1_{22}\underline{x}_3 + \hat{\beta}^1_{32}\underline{x}_1^2 + \hat{\beta}^1_{42}\underline{x}_3^2 + \hat{\beta}^1_{52}\underline{x}_1\underline{x}_3 + \hat{\beta}^1_{62} \qquad 4.3$$

$$\hat{y}^1_3 = \hat{\beta}^1_{13}\underline{x}_2 + \hat{\beta}^1_{23}\underline{x}_3 + \hat{\beta}^1_{33}\underline{x}_2^2 + \hat{\beta}^1_{43}\underline{x}_3^2 + \hat{\beta}^1_{53}\underline{x}_2\underline{x}_3 + \hat{\beta}^1_{63} \qquad 4.4$$

where $\hat{\beta}^k_{ij}$ is the i^{th} coefficient in the j^{th} model in the k^{th} layer and $\hat{\underline{y}}^k_j$ is the prediction of $\underline{y}$ from the j^{th} model in the k^{th} layer.

Assume, for this example, that all 3 first layer partial models are passed to the second layer. The 3 models given by Equations

4.5, 4.6 and 4.7 can then be constructed in the second layer.

$$\hat{y}_1^2 = \hat{\beta}_{11}^2 \hat{y}_1^1 + \hat{\beta}_{21}^2 \hat{y}_2^1 + \hat{\beta}_{31}^2 \hat{y}_1^{1\,2} + \hat{\beta}_{41}^2 \hat{y}_2^{1\,2} + \hat{\beta}_{51}^2 \hat{y}_1^1 \hat{y}_2^1 + \hat{\beta}_{61}^2 \qquad 4.5$$

$$\hat{y}_2^2 = \hat{\beta}_{12}^2 \hat{y}_1^1 + \hat{\beta}_{22}^2 \hat{y}_3^1 + \hat{\beta}_{32}^2 \hat{y}_1^{1\,2} + \hat{\beta}_{42}^2 \hat{y}_3^{1\,2} + \hat{\beta}_{52}^2 \hat{y}_1^1 \hat{y}_3^1 + \hat{\beta}_{62}^2 \qquad 4.6$$

$$\hat{y}_3^2 = \hat{\beta}_{13}^2 \hat{y}_2^1 + \hat{\beta}_{23}^2 \hat{y}_3^1 + \hat{\beta}_{33}^2 \hat{y}_2^{1\,2} + \hat{\beta}_{43}^2 \hat{y}_3^{1\,2} + \hat{\beta}_{53}^2 \hat{y}_2^1 \hat{y}_3^1 + \hat{\beta}_{63}^2 \qquad 4.7$$

Subsequent layers could be constructed similarly. A model in terms of the original variables can be constructed from a partial model by substituting the partial models from previous layers into the selected partial model. For example, if Equation 4.6 is chosen, the model in terms of the original variables is constructed by substituting Equations 4.2 and 4.4 into Equation 4.6.

4.2.2 Characteristics

Two positive and 3 negative characteristics of GMDH algorithms for polynomial model identification are discussed in this section. The 2 positive characteristics relate to potential reductions in the computational burden of identifying polynomial models and the use of independent data in the identification process. The 3 negative characteristics relate to the sacrificed completeness through which the computational burden is reduced, induced multicollinearity among the independent variables, and misleading values for the variances of the coefficients.

In some situations the number of terms examined to develop a polynomial model of a given maximum order can be smaller in the GMDH than in other identification methods. Each term has a coefficient which must be estimated. In most models some of the coefficients will equal 0. The number of terms examined in each layer of a GMDH algorithm using 2 variable second order polynomials as the partial models is given by

$$T = 6\binom{q}{2} \qquad 4.8$$

where T is the number of terms examined and q is the number of independent variables in the layer. $\binom{q}{i} = q!/(q-i)!i!$ where ! is the factorial operator. The order of the overall model is 2 in the first layer and doubles in each succeeding layer. One example of an alternative identification method is using stepwise regression on a general polynomial (Equation 4.1) with the specified maximum order. The number of terms examined in this method is given by

$$T = \sum_{i=1}^{p} \binom{q}{i} \qquad 4.9$$

where p is the order of the model, q is the number of independent variables in the original data set. Table 4.6 shows the number of terms examined by the GMDH, when q independent variables are used in each layer, and by stepwise regression for some combinations of p and q. Only the approximate relative computational burdens of the 2 methods can be judged from the number of terms examined, because the number of operations associated with examining 1 term varies between

q \ p	2	4	8
10	270	540	810
15	630	1,260	1,890
20	1,140	2,280	3,420

GMDH

q \ p	2	4	8
10	55	385	1,012
15	120	1,940	22,818
20	210	6,195	263,949

Stepwise

Table 4.6 Number of Terms Examined in GMDH and Stepwise Algorithms for Identifying Polynomial Models.

p = Order of the Polynomial
q = Number of Variables

methods.

Any model can be tested with independent data, but the GMDH is one of the few, if not the only, identification methods which is guided in part by tests on independent data. It is probably impossible to generally establish which identification method or even guidance procedure within identification methods is best, but a test on independent data is at least as defensible as any other guidance statistic. However, the partial models must still be developed carefully because no selection method is capable of choosing a good model from a set of poor models.

The computational savings of a GMDH algorithm come at the expense of not considering all possible models of a given complexity. Thus, models which are better than any examined may be overlooked. Two methods which can help alleviate this problem without greatly increasing the computation burden of the overall modeling process have been proposed. The first is to create models from linear combinations of the input variables for each layer, perhaps using stepwise regression. The procedure was suggested by Duffy and Franklin (1975) and is used in this work. The second is to model and remove the effects of low order trends before applying the GMDH. This can be accomplished by developing, for example, a linear model of the process and using the residuals from that model in the GMDH. Ivakhnenko (1971) removed a third order trend in time before applying the GMDH. A harmonic trend in time is removed from some of the data used in this work (see Chapter

5). As with all identification methods,except an exhaustive search of all possible models, we simply hope that the algorithm is adequate to identify a model which is not too far from the best model.

The use of independent data to measure the partial model quality tends to eliminate the partial models in which multicollinearity is the greatest problem. The reasons for this are discussed in Section 3.2.3 and 3.3. However, the GMDH method also assures the strong multicollinearity of all the independent variables from the second layer on. When all the terms in every partial model are retained, the partial models from the second layer on will be strongly affected by multicollinearity. For example, the coefficients of such partial models frequently have the following structure. The magnitudes of the coefficients of the 2 linear terms are between 0 and 1 and their sum is approximately 1. The magnitudes of the coefficients of the 3 nonlinear terms are relatively large and their sum is approximately 0. This type of structure can be seen in the models developed by Ivakhnenko (1970a) and indicates the variables are strongly collinear. These models cannot be expected to have good predictive qualities over a wide range of independent variables because they will be very sensitive to small variations in the relations between the independent variables. The problems caused by multicollinearity in the GMDH can be alleviated in part by using a stepwise regression algorithm to develop the partial model structure.

The variance of the coefficients from the second layer on are

likely to be higher than indicated by Equation 3.24 because the independent variables are then functions of estimated coefficients which have non zero variances and Equation 3.24 uses the assumption of zero variance independent variables.

The value of k which should be used with the model quality statistics given by Equations 3.8 through 3.14 when evaluating equations developed with the GMDH is not clear. Equations 3.8 through 3.12 are functions of k and all of the statistics are evaluated with regard to k when comparing different equations. (see Figures 6.1 through 6.4)

If we are primarily concerned with producing unbiased estimates of the statistics which are functions of k, we might choose k to be the cumulative number of coefficients estimated in developing the final model. For example, if a second layer partial model has 3 terms and both of the first layer models which provide the input to the second layer model have 4 terms, k would equal 11 (3+4+4). However, the primary model quality evaluation statistics used in this work are calculated on independent data and are not functions of k. With these statistics we need k mostly to judge the number of independent variable interactions which affect the model and the general model complexity. For this purpose we might choose k to be somewhere in between the number of different original variables which appear in the model and the total number of terms in the model when expanded in the original variables. In this work we chose k to be the total number of terms in the model when expanded in the p original variables, including the constant term.

This value of k both suits the primary purpose and maintains something close to unbiased estimates of the statistics calculated from the estimation data because it is generally closer to the total number of coefficients estimated than is the number of different original variables which appear in the equation.

4.3 GMDH Algorithm Used in this Work

The partial models in each layer are second order 2 variable polynomials of the form given by Equation 4.1. The stepwise regression algorithm described in Section 3.2.2 is used to develop the partial models. The stepwise regression algorithm is also used to generate linear models from the complete set of input variables for each layer. A specified number of the partial models with the lowest values of IRMS (see Section 3.2.3) are passed between layers. The process stops when a specified number of layers have been calculated. Any of the partial models may be examined and the coefficients may be reestimated from all the data. A user's manual for the GMDH program is presented in Appendix B.

Chapter 5

DEVELOPMENT OF ALTERNATIVE MODELS

Forty-seven models for today's maximum temperature at Huntsville, Alabama were developed by applying stepwise regression to each of 5 data sets, the GMDH to 3 of the 5 data sets, and interactive stepwise regression to 1 of the 5 data sets. The 5 data sets are all variants of the set 1 data described in Section 2.4. Some general aspects of the data are described in Section 5.1 and the 5 data sets are described in Section 5.2. The use of stepwise regression, interactive stepwise regression, and the GMDH to generate alternative models is described in section 5.3.

5.1 General Characteristics of the Data

The set 1 variables (see Section 2.4) are listed in Table 5.1 along with the units and correlations with the dependent variable. The variable names, projection times, observation times, and smoothing information are the same as in Tables 2.1, 2.2, and 2.3. The sequential variable numbers listed in Table 5.1 will be used as the variable identifiers, even when some variables are deleted from the data set.

Set 1 data are available only on the days listed in Table 5.2. The missing days are caused by various malfunctions in the Limited Area Fine Mesh (LFM) forecasting system. In each 6 digit number

Table 5.1 Set 1 Variables

variable #	name	(hrs) projection	smoothing	correlation	units
Harmonic Terms					
1	sin(day of year)	-	-	-0.6598	-
2	sin(2*day of year)	-	-	-0.0756	-
3	cos(day of year)	-	-	-0.6792	-
4	cos(2*day of year)	-	-	0.6789	-
Layer Heights					
5	1000 mb	12	5	-0.0406	meters
6	1000 mb	24	5	-0.1838	"
7	850 mb	12	-	0.4345	"
8	850 mb	24	-	0.2787	"
9	500 mb	12	-	0.8021	"
10	500 mb	24	-	0.7452	"
Layer Thicknesses					
11	500-1000 mb	0	-	0.7608	meters
12	500-1000 mb	6	-	0.8054	"
13	500-1000 mb	12	-	0.8351	"
14	500-1000 mb	18	-	0.8490	"
15	500-1000 mb	24	-	0.8548	"
16	850-1000 mb	0	-	0.8123	"
17	850-1000 mb	6	-	0.8645	"
18	860-1000 mb	12	-	0.8954	"

Table 5.1 Set 1 Variables (Cont'd)

variable #	name	projection	smoothing	correlation	units
19	850-1000 mb	18	-	0.9139	meters
20	850-1000 mb	24	-	0.8997	"
21	500-850 mb	0	-	0.7190	"
22	500-850 mb	6	-	0.7620	"
23	500-850 mb	12	-	0.7827	"
24	500-850 mb	18	-	0.7891	"
25	500-850 mb	24	-	0.7943	"
Layer Temperatures					
26	surface	0	-	0.8468	°Kelvin
27	1000 mb	12	5	0.8971	"
28	1000 mb	24	5	0.8616	"
29	850 mb	0	-	0.7920	"
30	850 mb	6	-	0.8469	"
31	850 mb	12	-	0.8786	"
32	850 mb	18	-	0.8963	"
33	850 mb	24	-	0.8905	"
34	700 mb	0	-	0.7386	"
35	700 mb	12	-	0.8265	"
36	700 mb	24	-	0.8396	"
37	BND LYR POT	6	-	0.8579	"
38	BND LYR POT	12	-	0.8873	"

Table 5.1 Set 1 Variables (Cont'd)

variable #	name	projection	smoothing	correlation	units
39	BND LYR POT	18	-	0.8941	°Kelvin
40	BND LYR POT	24	-	0.8687	"
West Wind Component (U)					
41	BND LYR	6	-	-0.0545	(meter)(sec^{-1})
42	BND LYR	12	-	0.0579	"
43	BND LYR	18	5	0.0295	"
44	BND LYR	24	5	0.0826	"
53	850 mb	6	-	-0.2870	"
54	850 mb	12	-	-0.1381	"
55	850 mb	18	5	-0.1972	"
56	850 mb	24	5	-0.1727	"
61	700 mb	12	-	-0.4441	"
62	700 mb	24	5	-0.4126	"
North Wind Component (V)					
45	BND LYR	6	-	0.3139	(meter)(sec^{-1})
46	BND LYR	12	-	0.1673	"
47	BND LYR	18	5	0.1988	"
48	BND LYR	24	5	0.1543	"
57	850 mb	6	-	0.2209	"
58	850 mb	12	-	0.1329	"
59	850 mb	18	5	0.1190	"

Table 5.1 Set 1 Variables (Cont'd)

variable #	name	projection	smoothing	correlation	units
60	850 mb	24	5	0.1744	(meter)(sec^{-1})
63	700 mb	12	-	0.0321	"
64	700 mb	24	5	0.1062	"
Wind Speed					
49	BND LYR	6	-	-0.3642	(meter)(sec^{-1})
50	BND LYR	12	-	-0.3185	"
51	BND LYR	18	5	-0.3185	"
52	BND LYR	24	5	-0.2994	"
Relative Vorticity					
65	850 mb	6	5	-0.2645	(10^{-5})(sec^{-1})
66	850 mb	12	5	-0.2645	"
67	850 mb	18	5	-0.3324	"
68	850 mb	24	5	-0.2746	"
69	500 mb	12	5	-0.3648	"
70	500 mb	24	5	-0.3477	"
Vertical Velocity					
71	850 mb	12	5	-0.2522	(mb)(sec^{-1})
72	850 mb	24	5	-0.0993	"
73	700 mb	12	5	-0.2256	"
74	700 mb	24	5	-0.1418	"
Temperature Differences					
75	700-1000 mb	12	-	-0.5450	°Kelvin

Table 5.1 Set 1 Variables (Cont'd)

variable #	name	projection	smoothing	correlation	units
76	700-1000 mb	24	-	-0.4595	°Kelvin
77	500-850 mb	12	-	-0.6292	"
78	500-850 mb	24	-	-0.6002	"
Water Content					
79	BND LYR REL HUM	0	5	0.0884	Percent
80	"	6	5	0.0561	"
81	"	12	5	0.0778	"
82	"	18	5	0.0511	"
83	"	24	5	0.0650	"
84	MEAN REL HUM	6	5	0.0800	"
85	"	12	5	0.1445	"
86	"	18	5	0.1071	"
87	"	24	5	0.1159	"
88	PRECIP WAT	6	5	0.4984	$(kg)(meter^{-2})$
89	"	12	5	0.5247	"
90	"	18	5	0.5171	"
91	"	24	5	0.5206	"
92	1000 mb DEW PT	6	5	0.6202	°Kelvin
93	"	12	5	0.6642	"
94	"	18	5	0.6949	"
95	"	24	5	0.7148	"

Table 5.1 Set 1 Variables (Cont'd)

variable #	name	projection	smoothing	correlation	units
96	850 mb DEW PT	12	5	0.6234	°Kelvin
97	850 mb DEW PT	24	5	0.6123	"
98	700 mb DEW PT	12	5	0.5260	"
99	700 mb DEW PT	24	5	0.5176	"
Wind Divergence					
100	BND LYR	6	5	-0.0282	$(10^{-5})(sec^{-1})$
101	BND LYR	12	5	-0.1270	"
102	BND LYR	18	5	0.0470	"
103	BND LYR	24	5	-0.1100	"
Temperature Advection					
104	850 mb	12	5	0.3977	(10^{-5})(°Kelvin)(sec^{-1})
105	850 mb	24	5	0.2217	"
Vorticity Advection					
106	500 mb	12	5	-0.3780	$(10^{-10})(sec^{-2})$
107	500 mb	24	5	-0.1928	"
Observed Variables		(Observation Times)			
108	ceiling	03	-	-0.3049	feet
109	cloud cover	03	-	-0.1659	percent
110	dew point	03	-	0.6503	°Fahrenheit
111	sfc wind speed	03	-	-0.4076	knots
112	sfc wind U	03	-	0.2308	knots

Table 5.1 Set 1 Variables (Cont'd)

variable #	name	projection	smoothing	correlation	units
113	sfc wind V	03	-	-0.1928	knots
114	sfc temperature	03	-	0.7769	°Fahrenheit
115	sfc temperature	24	-	0.8162	"
116	sfc temperature	21(previous day)		0.8008	"
117	previous max temperature	-	-	0.7962	"
118	previous min temperature	-	-	0.6407	"
119	today's max temperature	-	-	1	"

730401	730402	730403	730404	730406	730407	730408	730410	730411
730412	730413	730415	730417	730418	730428	730429	730430	
730502	730503	730504	730505	730506	730507	730508	730509	730510
730511	730513	730514	730515	730517	730518	730522	730523	730524
730525	730526	730527	730528	730529	730530	730531		
730601	730603	730614	730615	730616	730617	730618	730619	730620
730621	730622	730623	730624	730625	730626	730629	730630	
740401	740402	740403	740404	740405	740406	740407	740408	740409
740410	740411	740412	740413	740414	740415	740416	740417	740418
740419	740420	740422	740423	740424	740425	740426	740427	740428
740429	740430							
740501	740502	740503	740504	740505	740506	740507	740508	740509
740510	740511	740512	740513	740514	740515	740516	740517	740518
740519	740520	740522	740524	740525	740526	740527	740528	740530
740531								
740602	740603	740604	740605	740606	740607	740611	740612	740613
740614	740615	740616	740618	740619	740620	740621	740622	740623
740624	740627	740628	740629	740630				
750403	750405	750406	750407	750409	750410	750411	750412	750414
750415	750416	750417	750418	750421	750422	750423	750425	750426
750427	750428	750429	750430					
750501	750502	750503	750504	750505	750507	750508	750509	750510
750511	750512	750513	750514	750515	750516	750517	750519	750520
750521	750522	750523	750524	750525	750526	750527	750528	750529
750530	750531							
750601	750602	750603	750604	750605	750606	750607	750608	750609
750610	750611	750612	750615	750616	750617	750618	750619	750620
750621	750622	750623	750624	750625	750626	750627	750628	750629
750630								

Table 5.2 Dates for which Equation Development Data is Available for 0000 GMT Forecast Cycle Early Guidance Set 1 Equations in the Spring Season.

760401	760402	760403	760404	760405	760406	760407	760408	760409
760410	760411	760412	760413	760414	760415	760416	760417	760418
760419	760420	760421	760422	760423	760424	760425	760426	760427
760428	760429	760430						
760501	760502	760503	760504	760505	760506	760507	760508	760509
760510	760511	760512	760513	760514	760515	760516	760517	760519
760520	760521	760522	760523	760524	760525	760526	760527	760528
760529	760530	760531						
760601	760603	760604	760605	760606	760607	760608	760609	760610
760611	760612	760613	760614	760615	760616	760617	760618	760619
760620	760621	760622	760623	760624	760625	760626	760627	760628
760629	760630							
770401	770402	770403	770404	770405	770406	770407	770408	770409
770410	770411	770412	770413	770414	770415	770416	770417	770418
770419	770420	770421	770422	770423	770424	770425	770426	770427
770428	770430							
770501	770502	770503	770504	770505	770506	770507	770508	770509
770510	770511	770512	770513	770514	770515	770516	770517	770518
770519	770520	770521	770522	770523	770524	770525	770526	770527
770528	770529	770530						
770601	770602	770603	770604	770605	770606	770607	770608	770609
770610	770611	770612	770613	770614	770616	770617	770618	770619
770620	770621	770622	770623	770624	770625	770626	770627	770628
770629	770630							

Table 5.2 (cont'd) Dates for which Equation Development Data is Available for 0000 FMT Forecast Cycle Early Guidance Set 1 Equations in the Spring Season.

Year	Number of Data Points
1973	59
1974	80
1975	79
1976	89
1977	88
Total	395

Table 5.3 Number of Days with Data During Each Year in Table 5.2.

the first pair of numbers is the year, the second pair is the month, and the third pair is the day. The number of days with data in each year is summarized in Table 5.3.

The daily maximum temperatures for 31 March through 6 October in 1968 through 1977 are shown in Figure 5.1. An "x" marks 31 March in each year. The abscissa is the consecutive point number within the complete set of plotted points. The days listed in Table 5.2 are a subset of the April-June, 1973-1977 portion of the days shown in Figure 5.1.

The natural variability of the maximum temperature changes through the year. The variances of the maximum temperatures for 31 March through 6 October, as calculated from the 10 years of data shown in Figure 5.1, are shown in Figure 5.2. A definite, though noisy, trend is apparent.

Large sets of meteorological variables, such as that used in this work, tend to be highly redundant. By redundant, we mean that essentially the same information is carried by several different variables or sets of variables. This redundancy is illustrated in Figure 5.3. Figure 5.3 is based on data set 2, which is used to predict tonight's minimum temperature (see Section 2.4), and on the fifth variant of data set 1, which includes principal components of the original set 1 variables and is used to predict today's maximum temperature (see Section 5.2). Each point in Figure 5.3 represents a model which uses variables not used in any of the other models. The models were created with stepwise regression. After each

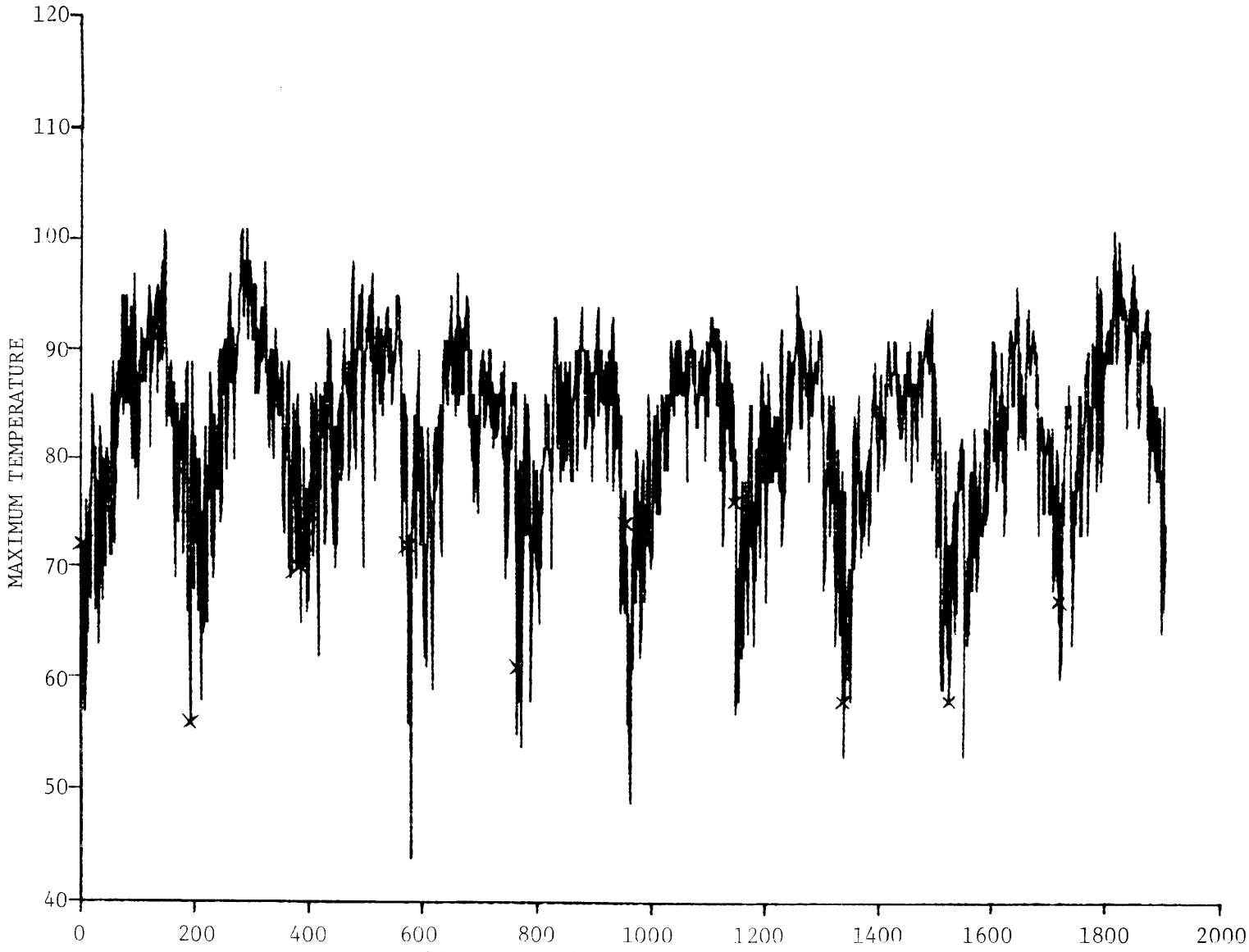

Figure 5.1 Daily Maximum Temperatures at Huntsville, Alabama

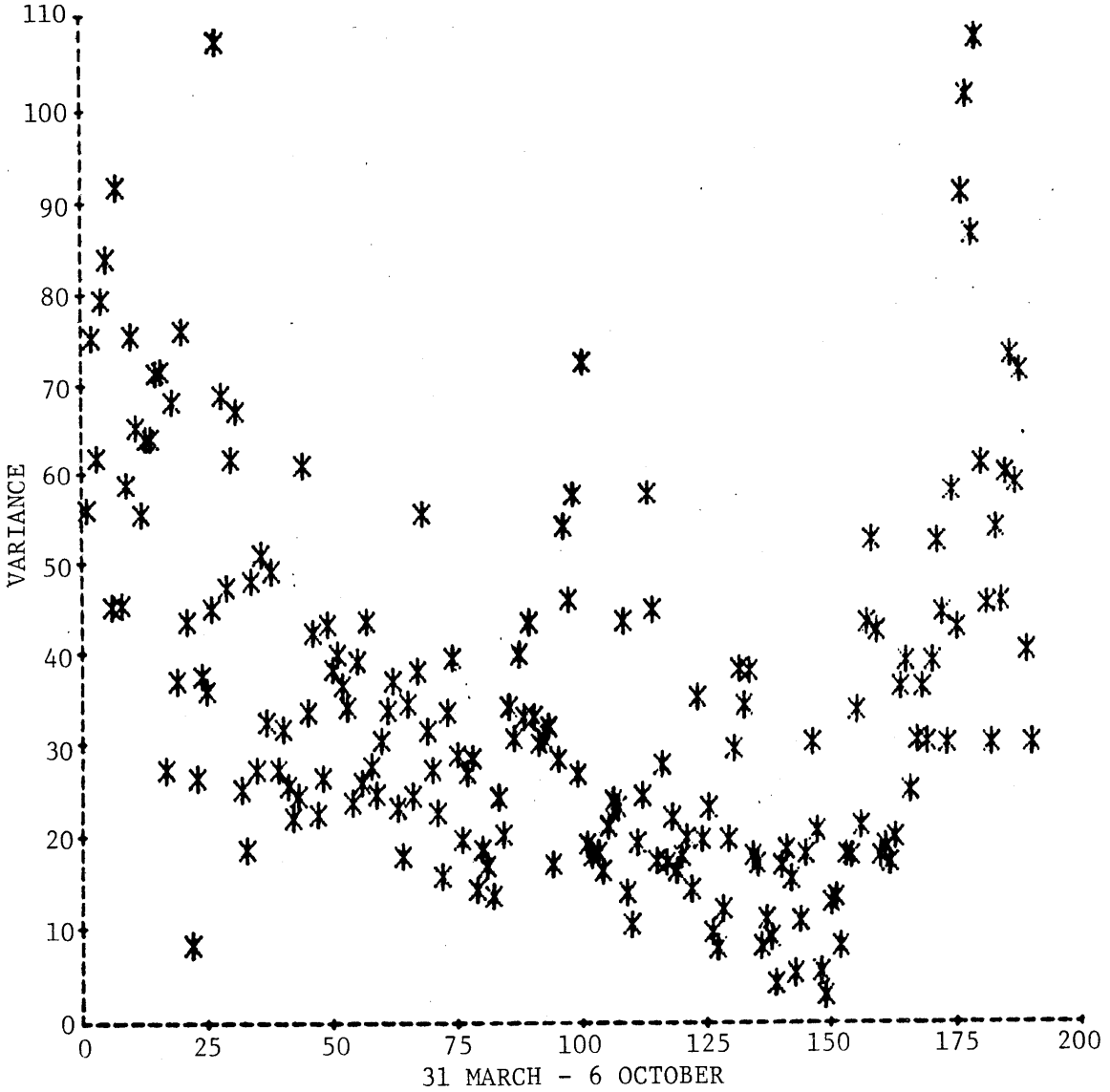

Figure 5.2 Variance of Daily Maximum Temperature at Huntsville, Alabama, 1968-1977

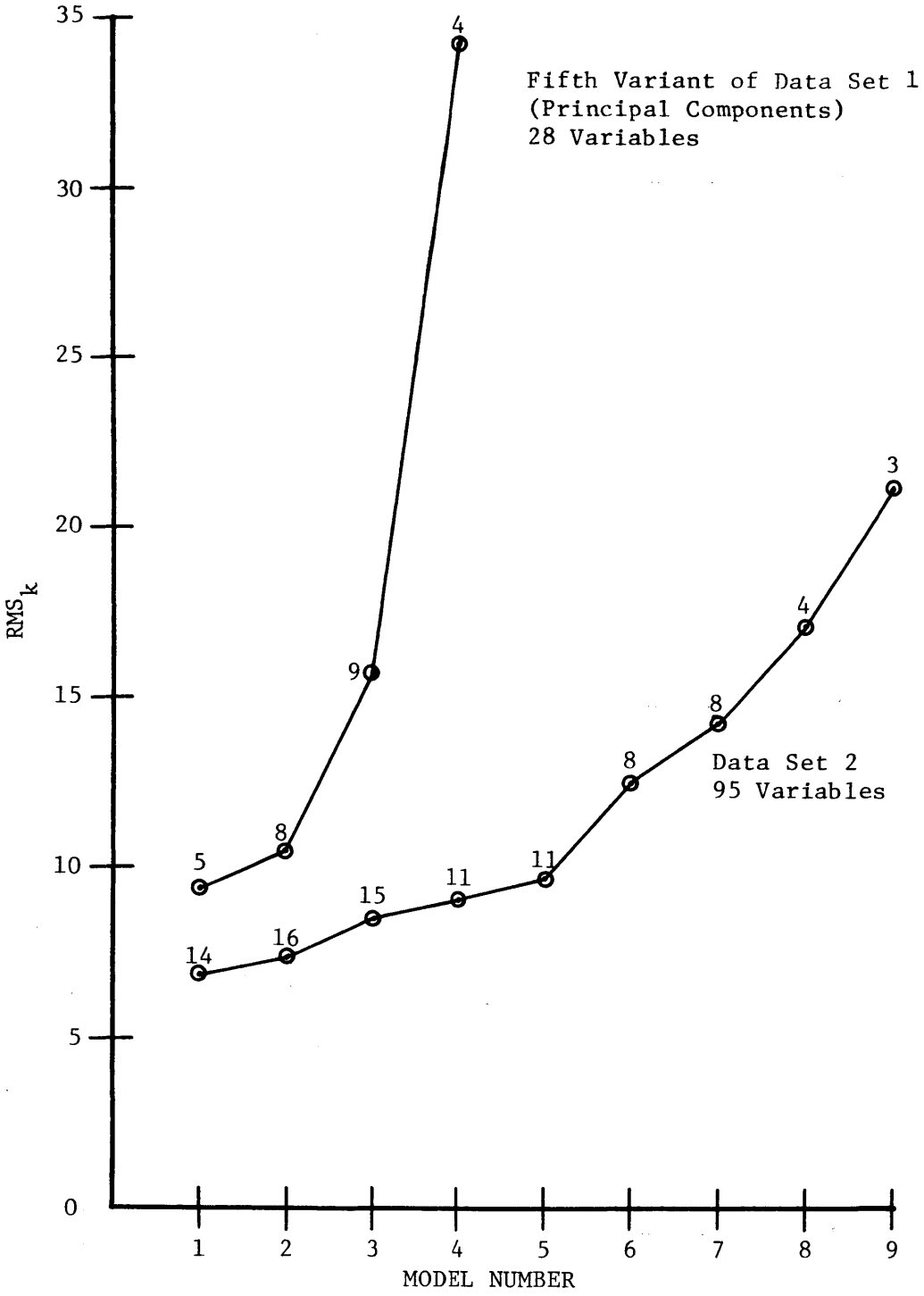

Figure 5.3 Successive Models Developed Using Stepwise Regression. The Numbers by the Points are the Numbers of Variables in Each Model. No Variable is Used in More than One Model.

model was created, the variables chosen for the model were removed from the data set. The significance level for entering and deleting variables was 0.05 for data set 2 and 0.01 for the fifth variant of data set 1. The quality of the successive models built from the principal components drops off more rapidly than those built from data set 2. This is expected because much of the redundancy is filtered out by the process of constructing principal components. This type of analysis was not performed on the original variables in data set 1, but they can be expected to have similar properties.

5.2 Data Sets Used in this Work

The first variant of data set 1 is simply the unmodified original variables. The second, third, and fourth variants were developed by removing harmonic components from the variables. The fifth variant includes principal components of the original variables. From here on, these 5 variants of data set 1 will be referred to simply as data sets 1 through 5. Since the original variables continue to be called data set 1 and no other data sets are used in this work, this renaming should not cause confusion. The development of data sets 2, 3, and 4 is described in Section 5.2.1 and development of data set 5 is described in Section 5.2.2.

5.2.1 Data Sets 2, 3, and 4

We may consider modeling temperature as the sum of two components, a mean and a departure from the mean. The existence of a

smooth mean trend of daily maximum temperatures produced by the regular pattern of the earth's orbit and relatively stationary geographical effects is physically plausible. Departures from the mean are caused by the more transient effects such as cloud cover and humidity. The potential advantages of explicit separation of the mean trend and deviations from the trend include the possibility of producing more robust models and facilitating interpretation of the model variables. The model robustness may be increased because a mean trend modeled separately with mathematical functions is perfectly stable and the meteorological variables are left only the task of modeling deviations from the trend. The interpretation of the models may be eased because meteorological variables which have little causal relation to temperatures, but whose trends match the trend of temperature, are not as likely to appear in the model.

Harmonic functions have been used to model temperature trends. For example, Craddock (1956) found that the first 2 terms of a Fourier series expansion adequately described the annual trend of 5 day mean temperatures at the 43 European cities he studied. Craddock had approximately 80 years of data for most of the cities. Taylor (1972), following the work of Craddock, used a 2 term harmonic model in a simulation of temperature in Britain.

The following two harmonic models of maximum temperature at Huntsville, Alabama, were developed with least squares regression. Each uses the first 4 variables in Table 5.1 (see Equations 2.1 through

2.4) and a constant term. Thus they are each equivalent to the first 2 terms in a Fourier series expansion.

$$T_{MAX} = 71.9 - 4.5 \sin(\frac{2\pi D}{365}) + 1.0 \sin(\frac{4\pi D}{365}) - 17.4 \cos(\frac{2\pi D}{365}) - 1.4 \cos(\frac{4\pi D}{365}) \quad (5.1)$$

$$T_{MAX} = -23.4+88.5 \sin(\frac{2\pi D}{365}) + 39.2 \sin(\frac{4\pi D}{365}) - 111.4 \cos(\frac{2\pi D}{365})-1.6 \cos(\frac{4\pi D}{365}) \quad (5.2)$$

Equation 5.1 was developed from the 10 years of data shown in Figure 5.1. Equation 5.2 was developed from only the data in set 1 (see Table 5.2). Thus, slightly more than 4 times as many data points were used to estimate the parameters for Equation 5.1 than were used for Equation 5.2. However, only half of the data used for Equation 5.1 are from the same season (April-June) as data set 1. The other half are from July-September. The 10 year average maximum temperatures and the harmonic model of these averages, Equation 5.1, are shown in Figure 5.4.

Data sets 2, 3, and 4 were created by 3 slightly different methods of modeling and removing harmonic components from data set 1. Data set 2 was created by replacing variables 5 through 118, and the dependent variable, with the residuals from separate harmonic models of each variable. Each harmonic model had the same form as Equations 5.1 and 5.2. The parameters of the models were estimated with least squares regression using the 5 years of data in set 1. Equation 5.2 was thus the harmonic model of the mean trend of the dependent variable. The residuals from this model, which are the dependent variable, are

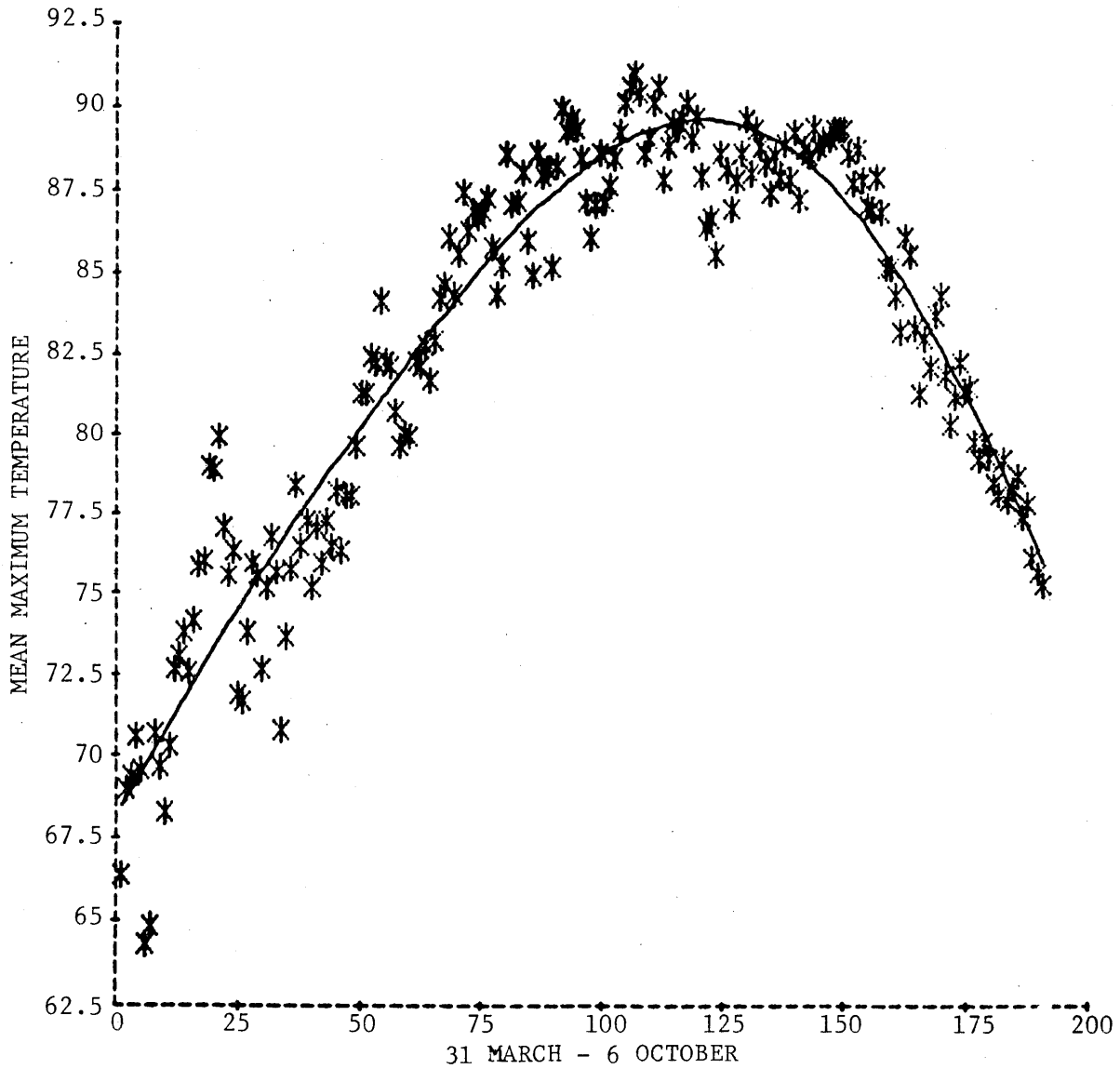

Figure 5.4 Mean Maximum Temperatures at Huntsville, Alabama 31 March - 6 October, 1968-1977

shown in Figure 5.5. Data set 3 uses the independent variables from data set 2, but the mean trend of the dependent variable is modeled with Equation 5.1. The residuals from Equation 5.1, which are the dependent variable, are shown in Figure 5.6. Note that Figure 5.6 is very similar to Figure 5.5. Data set 4 uses the original independent variables (from data set 1) and the detrended dependent variable from data set 3.

The method used to create data set 2 is perhaps the most conventional way of removing the effects of one set of variables from another set of variables. Although variables 1 through 4 were left in the data set unchanged, they were effectively removed from the modeling process in data set 2 because, following the removal of their effects, they were linearly uncorrelated with all of the other variables. Variables 1 through 4 were still available in data sets 3 and 4, because they had non-zero correlations with the dependent variables in data set 3 and with both the dependent and independent variables in data set 4. The methods used to create data sets 3 and 4 were attempts to incorporate information beyond that in the basic data set into the modeling process. Data set 4 was created to allow the harmonic terms which describe the net trends of the independent variables in the models to be chosen along with the independent variables, without the influence of the trend in the dependent variable.

The quality of models developed from data from which the effects of some variables have been removed can be estimated directly

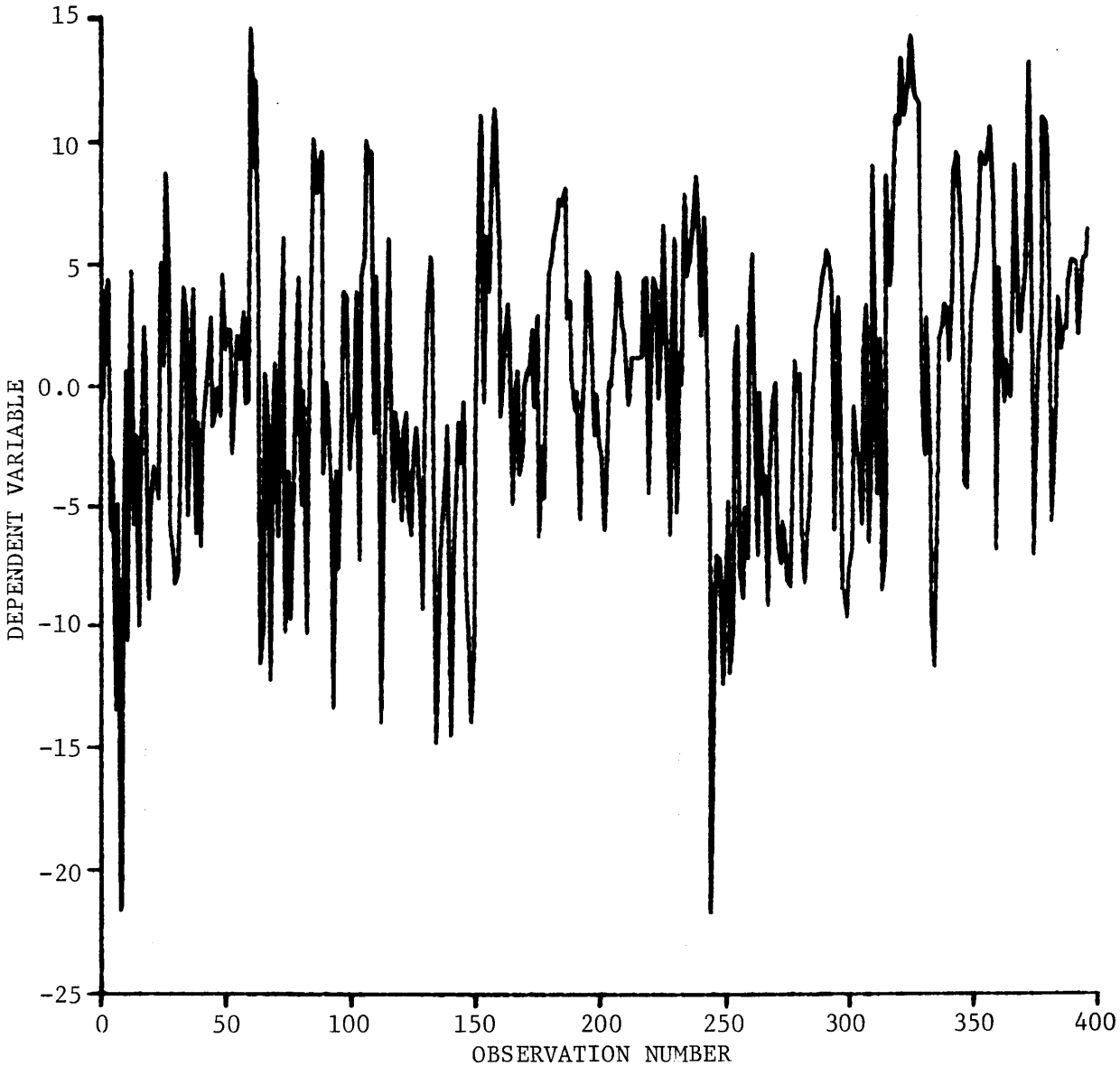

Figure 5.5 Dependent Variable in Data Set 2

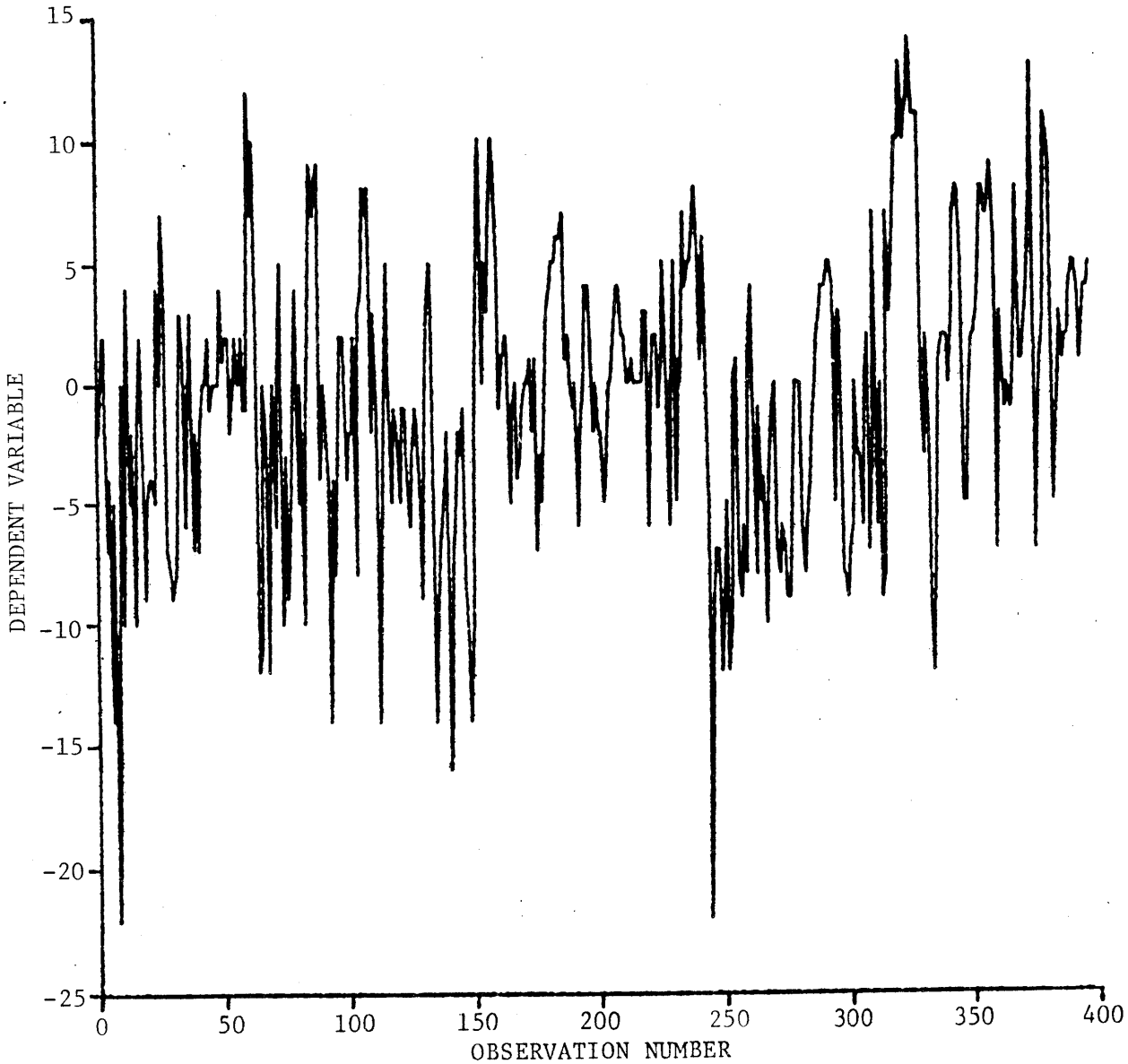

Figure 5.6 Dependent Variable in Data Sets 3 and 4

from the model residuals. This can be shown as follows. Let $\underline{X}$ and $\underline{y}$ be the original data and let $\underline{X}^*$ and $\underline{y}^*$ be the data from which the effects of some variables have been removed. We develop a model $\hat{\underline{y}}^* = \underline{X}^* \hat{\underline{\beta}}$ which has residuals $\underline{y}^* - \hat{\underline{y}}^*$. The residuals in terms of the original variables are $(\underline{y}^* + (\underline{y} - \underline{y}^*) - (\hat{\underline{y}}^* + (\underline{y} - \underline{y}^*)) = \underline{y}^* - \hat{\underline{y}}^*$, the same as the model residuals. However, when calculating a quality measure which is a function of the number of variables in a model (see Section 3.2.3), some account should be taken of effects already removed from the data. When the parameters of the removed components have been estimated using only the data from which they were removed, as in data set 2, it is fairly clear that k in Equations (3.7) through (3.11) should be increased by the number of parameters estimated. Thus, k was increased by 4 for equations developed from data set 2. However, when some of the data used to estimate parameters of the removed components are not part of the data set from which the components are removed, as in data sets 3 and 4, the situation is not as clear. If the removed component is estimated completely independently of the data set used for the rest of the modeling, no adjustment to k is needed. Since less than one fourth of the data used to estimate the parameters of Equation (5.1) is from data set 1, k was not agumented for models developed from data sets 3 and 4.

5.2.2 Data Set 5

The independent variables in data set 5 include principal components of groups of variables from data set 1 and some variables

unchanged from data set 1. The dependent variable was not modified. The groups of variables which were replaced by principal components are listed in Table 5.4. Each group was replaced by the principal component of that group which had the largest variance. The percentage of the total variance which was accounted for by the principal components used in this work are also listed in Table 5.4. The variables were normalized by subtracting the means and dividing by the standard deviations before the principal components were calculated. The variable transformations for the principal components which were retained are listed in Appendix A. Variables in Table 5.1 which are not listed in Table 5.4 were retained in their original form in data set 5, giving a total of 28 independent variables in data set 5. The variable numbering from data set 1 is again retained and the principal components are identified by the letters in Table 5.4.

5.3 Generating Alternative Models

Four models from each of the 5 data sets were generated using stepwise regression. The significance level for entering and deleting variables was 0.05 for 2 of the models from each data set and 0.01 for the other 2. At each significance level, the variables for one of the models were chosen using only the first three years of data. However, after the variables were chosen, the coefficients of those models were reestimated from all 5 years of data. The significance level for entering variables can be stricter than for deleting variables, but in this work the significance levels for entering and deleting were always

Group	Variables	Group Name	Maximum % of Total Variance in One Component
A	5-10	layer heights	69
B	11-25	layer thicknesses	90
C	26-40	layer temperatures	88
D	41-44;53-56; 61-62	wind U	71
E	45-48;57-60; 63-64	wind V	77
F	49-52	wind speed	70
G	65-70	relative vorticity	71
H	71-74	vertical velocity	72
I	75-78	temperature differences	75
J	79-99	water content	75
K	100-103	wind divergence	40
L	104-105	temperature advection	77
M	106-107	vorticity advection	62

Table 5.4 Groups of Variables Replaced by their Principal Components in Data Set 5

equal.

Five models from data set 5 and 10 models from each of data sets 1 and 2 were generated using the GMDH. The significance level for entering and deleting variables in the partial models was 0.05 for 5 of the models from data sets 1 and 2 and 0.01 for the other 5. Only 0.01 was used on data set 5. Fifty partial models were passed between layers. The 5 models for each significance level consisted of the best model in each of 3 layers and the models generated with stepwise regression from the input variables for layers 2 and 3. The first 3 years of data were used to estimate the coefficients and the last 2 years of data were used to measure the partial model quality. Following the variable selection, the coefficients of all the models were re-estimated using all 5 years of data. The linear models described in the previous paragraph whose variables were chosen using only the first 3 years of data are the same as the models developed in the GMDH algorithm from the input data for the first layer.

GMDH models were developed from only 3 data sets because of limitations on computer time. We expect that the relations between the quality of the linear models and of the GMDH models for data sets 3 and 4 is similar to that for data sets 1 and 2.

Two models from data set 5 were generated using interactive stepwise regression. The F statistics indicated that variables 1, A, B, and C should be included in any model and that variables K, M, and 109 were reasonable choices for a fifth variable. The partial residual plots of variables K, M, and 109, with variables 1, A, B, and C already

in the model, were examined to distinguish between the 3 variables. The partial residual plots are shown in Figures 5.7, 5.8, and 5.9. Two models, one using variable K and one using variable 109 for the fifth independent variable were then selected.

The 47 models generated are listed in Table 5.5 along with the number of terms in each model, the variables in those terms, the method of generation, the significance level for entering and deleting variables, and the data set from which the model was developed. The number of terms includes the constant term and is sometimes greater than the number of variables because the variables are used in different combinations and transformations in the models. The underlined groups of variables correspond to the groups from which principal components were calculated (see Table 5.4). The abbreviations used for the generation techniques are described in Table 5.6.

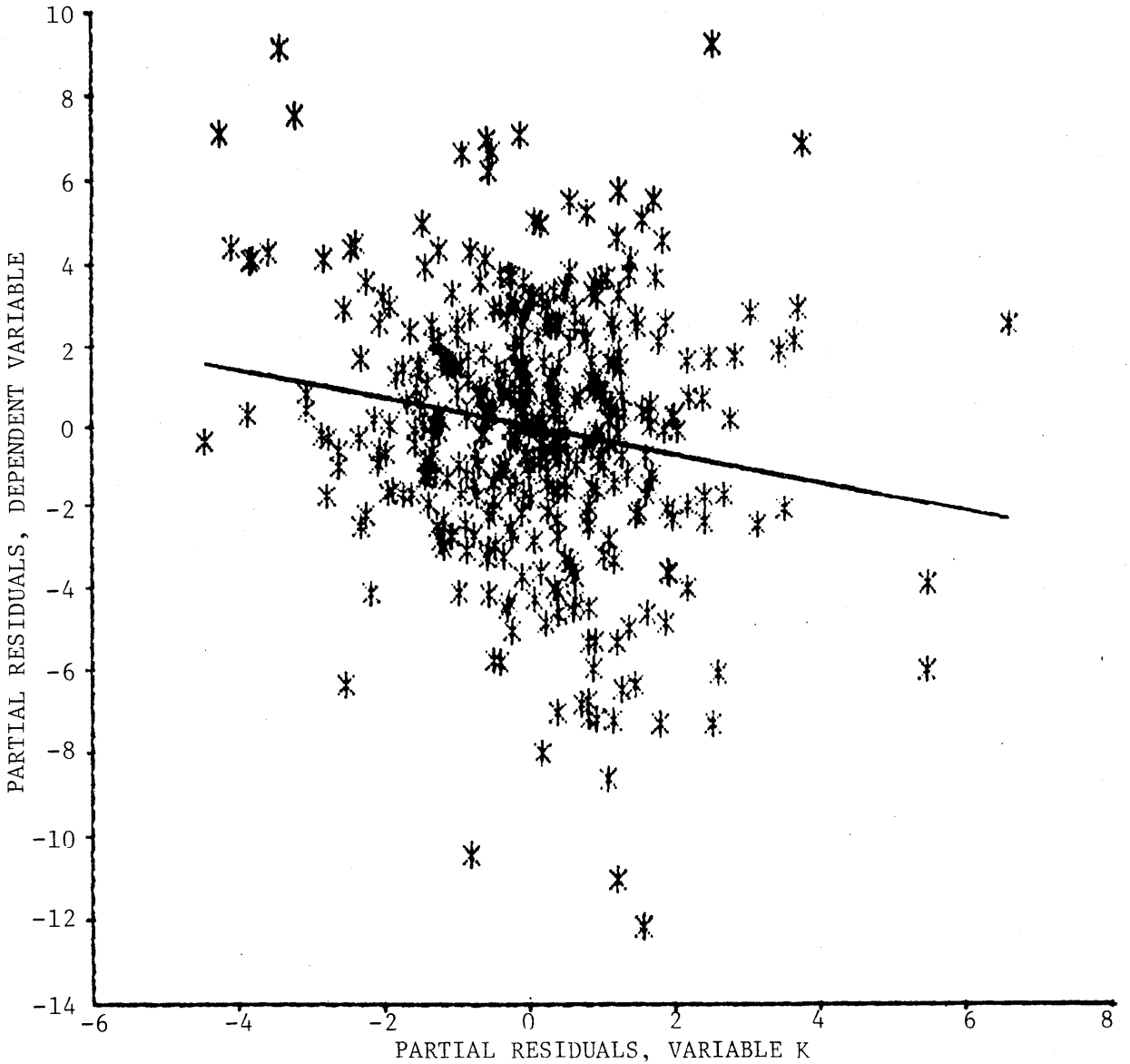

Figure 5.7 Partial Residuals of Variable K, Variables 1, A, B, and C in Model

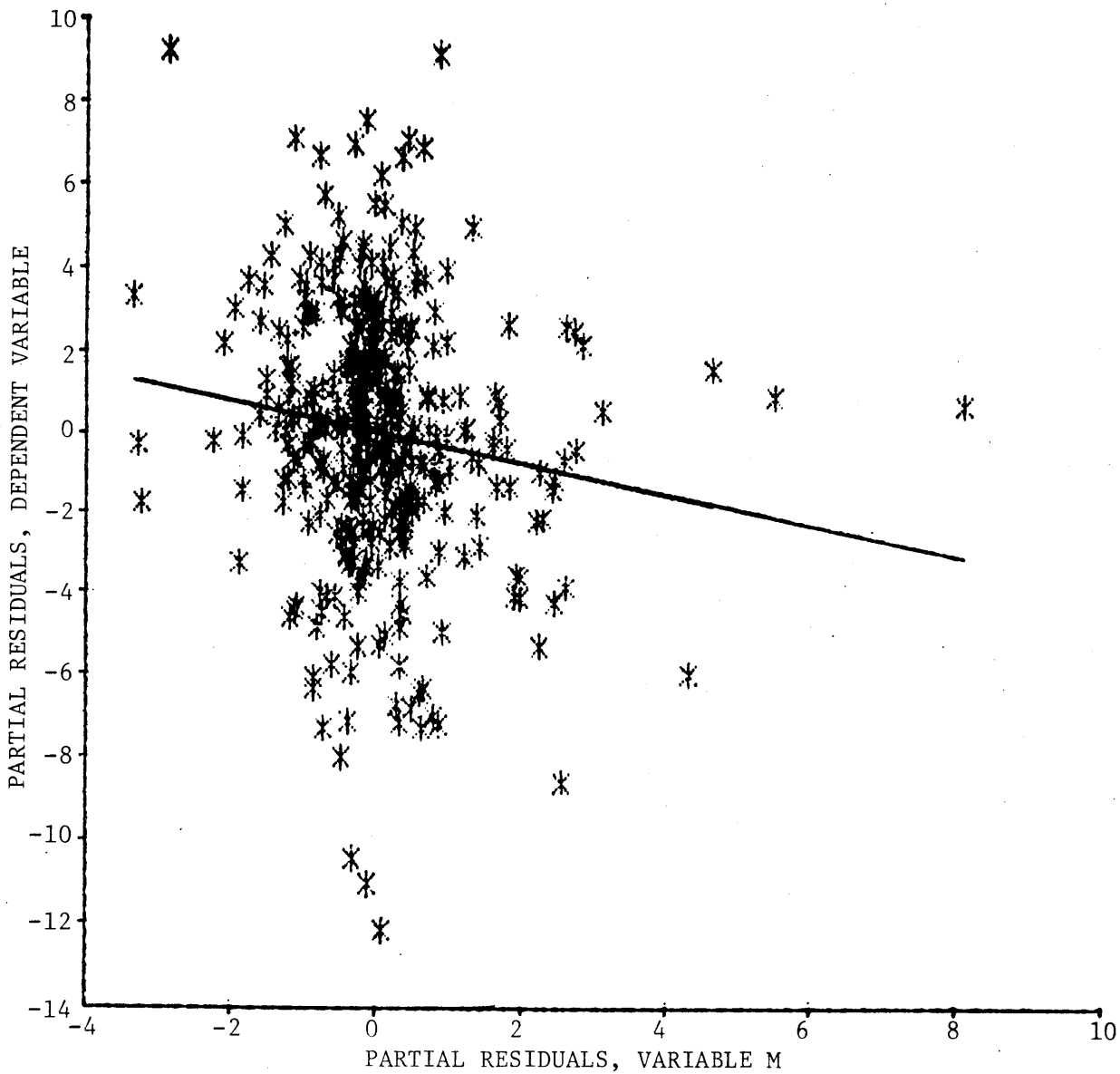

Figure 5.8 Partial Residuals of Variable M, Variables 1, A, B, and C in Model

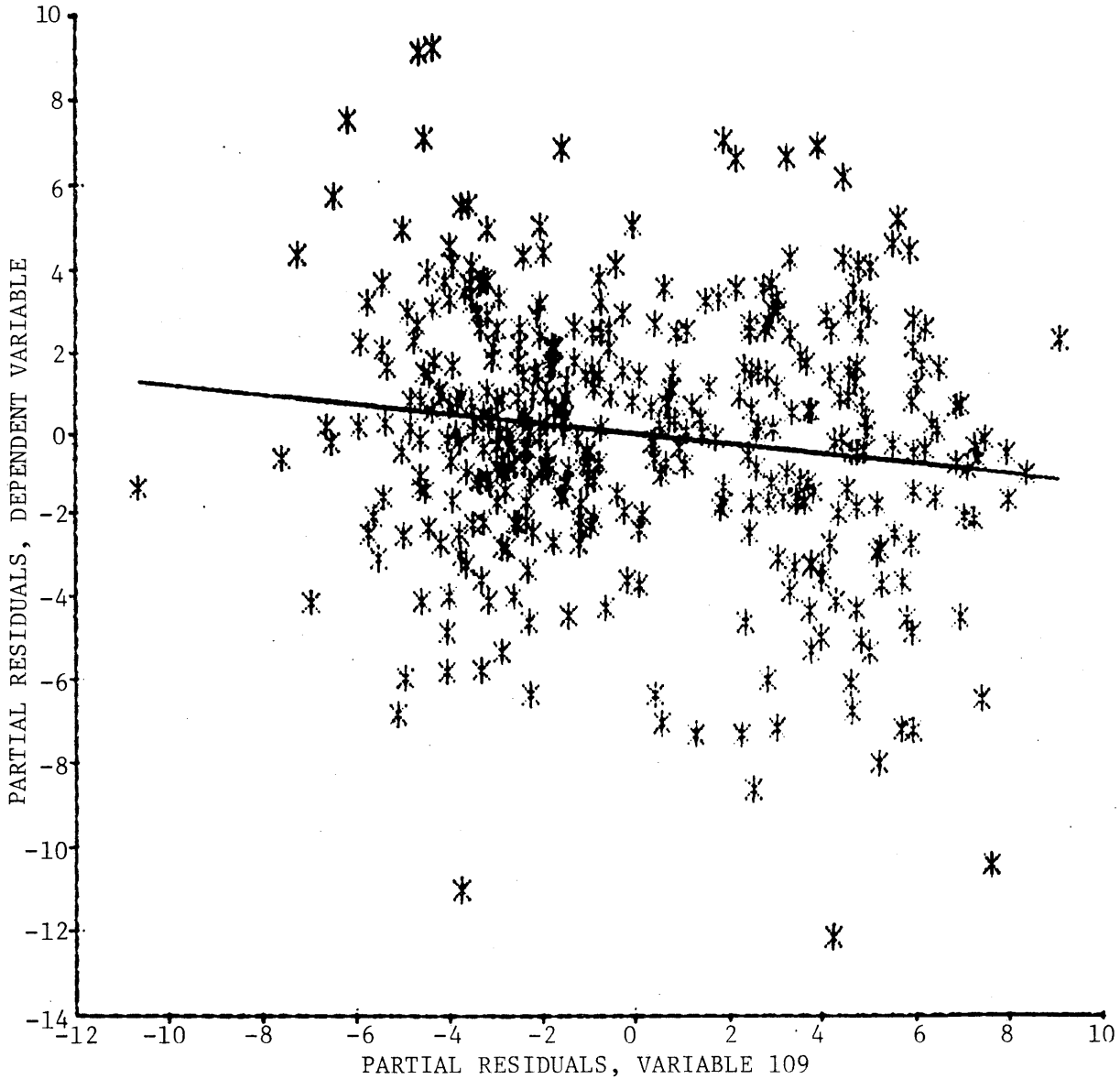

Figure 5.9 Partial Residuals of Variable 109, Variables 1, A, B, and C in Model

Model #	# Terms in Model	Variables in Model	Generation Technique	α_1 α_2	Data Set
1	12	1,7,8,19,68,85,89,96,107,115,117	STWS-3	0.05	1
2	12	2,10,19,43,61,72,74,102,104,109,117	STWS-5	0.05	1
3	6	1,19,68,107,117	STWS-3	0.01	1
4	7	7,19,74,102,104,117	STWS-5	0.01	1
5	2	20,26	GMDH mlℓ1	0.05	1
6	16	26,28,33,114	GMDH mlℓ2	0.05	1
7	14	20,26,40,114	GMDH mlℓ3	0.05	1
8	9	27,28,33,115,117	GMDH in ℓ2	0.05	1
9	36	16,26,28,32,40,114,116	GMDH in ℓ3	0.05	1
10	2	20,26	GMDH mlℓ1	0.01	1
11	8	20,26,28,114	GMDH mlℓ2	0.01	1
12	16	19,28,33,114,115,116	GMDH mlℓ3	0.01	1
13	7	26,28,32,116	GMDH in ℓ2	0.01	1
14	12	28,33,114,117	GMDH in ℓ3	0.01	1
15	9	5,19,68,85,96,107,115,117	STWS-3	0.05	2
16	11	10,19,43,61,67,72,74,102,104,117	STWS-5	0.05	2
17	5	19,68,107,117	STWS-3	0.01	2
18	6	19,43,61,68,117	STWS-5	0.01	2
19	3	20,26	GMDH mlℓ1	0.05	2
20	6	20,27,67,117	GMDH mlℓ2	0.05	2

Table 5.5 Alternative Models

Model #	# Terms in Model	Variables in Model	Generation Technique	α_1 α_2	Data Set
21	11	19,20,27,40,68,87,116,117	GMDH mlℓ3	0.05	2
22	21	17,19,26,27,28,30,39,40,61,68,86,115, 116,117	GMDH in ℓ2	0.05	2
23	17	19,20,26,27,39,40,62,68,86,115,116,117	GMDH in ℓ3	0.05	2
24	3	28,114	GMDH mlℓ1	0.01	2
25	5	20,27,68,115	GMDH mlℓ2	0.01	2
26	10	10,19,20,27,39,68,87,117	GMDH mlℓ3	0.01	2
27	13	10,19,26,27,39,40	GMDH in ℓ2	0.01	2
28	11	10,19,20,27,39,68,87,115,117	GMDH in ℓ3	0.01	2
29	8	7,19,68,85,96,107,117	STWS-3	0.05	3
30	11	7,19,43,61,46,63,72,74,102,117	STWS-5	0.05	3
31	5	19,68,107,117	STWS-3	0.01	3
32	6	19,43,61,68,117	STWS-5	0.01	3
33	8	3,27,32,43,61,68,107	STWS-3	0.05	4
34	13	3,10,32,39,43,61,72,74,77,102,109,116	STWS-5	0.05	4
35	6	3,27,32,68,107	STWS-3	0.01	4
36	12	3,10,21,32,39,43,61,72,74,102,116	STWS-5	0.01	4
37	7	1,A,C,I,K,M	STWS-3	0.05	5
38	11	1,A,B,C,D,K,L,M,109,114	STWS-5	0.05	5
39	6	1,A,C,I,M	STWS-3	0.01	5
40	6	1,A,B,C,K	STWS-5	0.01	5

Table 5.5 Alternative Models (cont'd)

Model #	# Terms in Model	Variables in Model	Generation Technique	α_2 α_1	Data Set
41	4	C,109	GMDH mlℓ1	0.01	5
42	15	A,C,109	GMDH mlℓ2	0.01	5
43	15	A,C,109	GMDH mlℓ3	0.01	5
44	7	A,B,C,K	GMDH in ℓ2	0.01	5
45	15	A,C,109	GMDH in ℓ3	0.01	5
46	6	1,A,B,C,K	instws	-	5
47	6	1,A,B,C,109	instws	-	5

Table 5.5 Alternative Models (cont'd)

Table 5.6 Abbreviations Used in Table 5.5

STWS-3	=	stepwise regression, 3 years of data used to choose the variables
STWS-5	=	stepwise regression, 5 years of data used to choose the variables
GMDH ml ℓ1	=	GMDH, the best model in layer 1
GMDH ml ℓ2	=	GMDH, the best model in layer 2
GMDH ml ℓ3	=	GMDH, the best model in layer 3
GMDH in ℓ2	=	GMDH, the linear model generated from all the input variables to layer 2
GMDH in ℓ3	=	GMDH, the linear model generated from all the input variables to layer 3
instws	=	interactive stepwise regression

Chapter 6

RESULTS

The 47 models described in Chapter 5, a Model Output Statistics (MOS) model used by the National Weather Service (NWS), and a model suggested by the validation procedures described in Section 6.2 are examined in this chapter. Model quality statistics are examined in Section 6.1, validation procedures are applied to some selected models in Section 6.2, and the relative performance of the 3 model generation techniques used in this work is considered in Section 6.3.

The NWS MOS equation for predicting today's maximum temperature at Huntsville, Alabama uses variables 3,10,20,27,51,87,96,107,109, and 114. The coefficients of these variables which were developed in this work are generally close to, but not the same as the coefficients developed by the NWS. The reason for the discrepancy is not known. The mean squared residual (RMS_k, see Equation 3.8) for this equation given by the NWS is 9.19 and the RMS_k calculated in this work is 9.26. Thus the quality of the fit to the estimation data is similar for both sets of coefficients. Because of the discrepancy and because the NWS equation is constrained to use the same predictors as the other equations in set 1 for the spring season (see Section 2.3) the equation labeled NWS in this chapter is not presented as the best single purpose equation which can be produced by the MOS system used by the NWS, but rather as an approximate representative of a model currently in use.

Any conclusions drawn from the information presented here must be tempered with the realization that only 1 dependent variable at 1 location was considered in this work.

6.1 Statistical Evaluation of Model Quality

The 7 statistics given by equations 3.8 through 3.14 are listed for each equation in Table 6.1. The value used for σ^2 in the equation for C_k was 7.0, approximately the lowest value of RMS_k from the 47 equations. The last 2 years of data were used to calculate IRMS and IRMA. IRMS and IRMA were calculated for the NWS model and for the models whose variables were chosen using all 5 years of data (models 2,4,16,18,30,32,34,36,38,40,46, and 47) by reestimating the coefficients using only the first 3 years of data and calculating the statistics from the last 2 years of data. The last 2 years of data are not totally independent in this procedure, since they were used to guide the variable selection, but statistics which can be compared with the other statistics on independent data are produced. k in Equations 3.8 through 3.14 is the number of terms in the model, including the constant term, plus any adjustments for modifications to the data (see Section 5.2.1). The value of k used for the models developed with the GMDH is discussed in Section 4.2.2.

IRMS, IRMA, RMS_k, and RMA_k are plotted against k in Figures 6.1 through 6.4. Plots of S_k, J_k, and C_k are not presented because they all show essentially the same result as the plot of RMS_k. Since we prefer low values on both axes in Figures 6.1 through 6.4, the models

Model #	Data Set	k	RMS_k	J_k	C_k	S_k	RMA_k	IRMS	IRMA
1	1	12	8.70	8.96	104.89	0.02271	2.27	11.58	2.61
2	1	12	7.89	8.13	60.60	0.02060	2.22	8.20	2.21
3	1	6	9.03	9.16	118.58	0.02320	2.29	9.73	2.38
4	1	7	8.73	8.89	103.04	0.02251	2.31	8.63	2.20
5	1	2	10.85	10.90	218.07	0.02760	2.46	9.90	2.37
6	1	16	10.48	10.91	204.60	0.02766	2.46	8.79	2.19
7	1	14	10.60	10.98	210.00	0.02782	2.49	8.60	2.14
8	1	9	9.83	10.05	165.07	0.02547	2.37	11.63	2.45
9	1	36	10.69	11.67	225.34	0.02978	2.57	9.28	2.26
10	1	2	10.85	10.90	218.07	0.02760	2.46	9.90	2.37
11	1	8	10.64	10.86	209.42	0.02750	2.47	8.69	2.17
12	1	16	10.28	10.69	193.44	0.02712	2.44	8.63	2.13
13	1	7	10.08	10.26	177.63	0.02598	2.38	9.69	2.30
14	1	12	10.33	10.65	194.35	0.02698	2.46	9.18	2.23
15	2	13	8.62	8.90	101.17	0.02255	2.24	10.32	2.44
16	2	15	7.84	8.14	60.64	0.02063	2.23	8.10	2.21
17	2	9	8.95	9.16	116.61	0.02319	2.29	9.34	2.30
18	2	10	8.45	8.66	89.59	0.02194	2.26	8.77	2.24
19	2	7	10.22	10.40	185.47	0.02634	2.46	9.74	2.42
20	2	10	9.20	9.43	131.08	0.02390	2.30	9.00	2.18
21	2	16	9.02	9.38	125.26	0.02379	2.31	8.57	2.14
22	2	25	8.85	9.41	122.89	0.02392	2.30	12.03	2.58
23	2	21	8.67	9.13	109.98	0.02317	2.28	10.39	2.45
24	2	7	11.48	11.68	255.18	0.02958	2.64	9.51	2.39
25	2	9	9.53	9.75	148.63	0.02469	2.35	9.06	2.21
26	2	14	9.09	9.41	127.70	0.02386	2.34	8.95	2.26
27	2	17	9.05	9.44	127.93	0.02395	2.33	10.24	2.45
28	2	15	9.05	9.39	126.12	0.02381	2.33	10.32	2.45
29	3	8	8.89	9.07	112.24	0.02296	2.27	9.29	2.04
30	3	11	7.12	7.32	17.43	0.01853	2.10	7.46	2.06
31	3	5	8.64	8.75	96.32	0.02215	2.25	8.88	2.11
32	3	6	7.66	7.78	42.90	0.01970	2.15	8.30	2.18

Table 6.1 Model Quality Statistics

Model #	Data Set	k	RMS_k	J_k	C_k	S_k	RMA_k	IRMS	IRMA
33	4	8	8.81	8.99	108.20	0.02277	2.27	9.13	2.12
34	4	13	7.16	7.40	21.79	0.01875	2.11	7.75	2.16
35	4	6	9.27	9.41	131.97	0.02382	2.32	9.61	2.14
36	4	12	7.20	7.42	23.01	0.01880	2.11	7.78	2.14
37	5	7	9.55	9.72	148.38	0.02461	2.39	10.77	2.51
38	5	11	9.27	9.53	135.38	0.02413	2.34	10.92	2.51
39	5	6	9.74	9.88	158.02	0.02503	2.37	10.87	2.51
40	5	6	9.68	9.83	155.01	0.02489	2.39	10.53	2.44
41	5	4	11.09	11.20	232.58	0.02837	2.59	11.63	2.67
42	5	15	10.54	10.94	207.32	0.02774	2.51	10.66	2.47
43	5	15	10.54	10.94	207.32	0.02774	2.51	10.66	2.47
44	5	7	9.92	10.10	168.84	0.02557	2.40	11.00	2.54
45	5	15	10.54	10.94	207.32	0.02774	2.51	10.66	2.47
46	5	6	9.68	9.83	155.01	0.02489	2.40	10.53	2.51
47	5	6	9.75	9.89	158.62	0.02505	2.40	10.29	2.43
NWS	1	11	9.26	9.51	134.74	0.02410	2.37	9.51	2.36

Table 6.1 Model Quality Statistics
(cont'd)

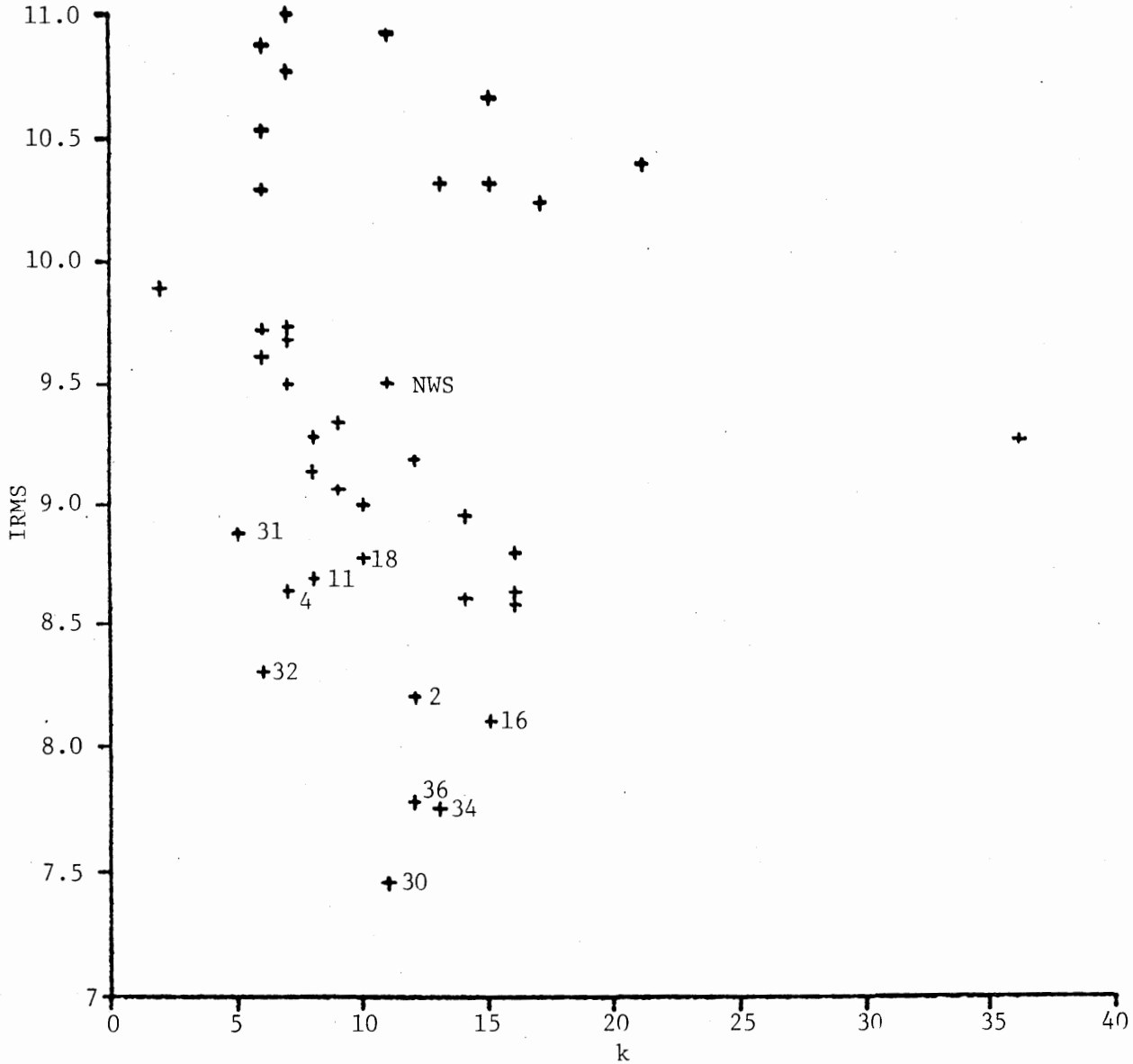

Figure 6.1 The Model Numbers are from Table 5.5

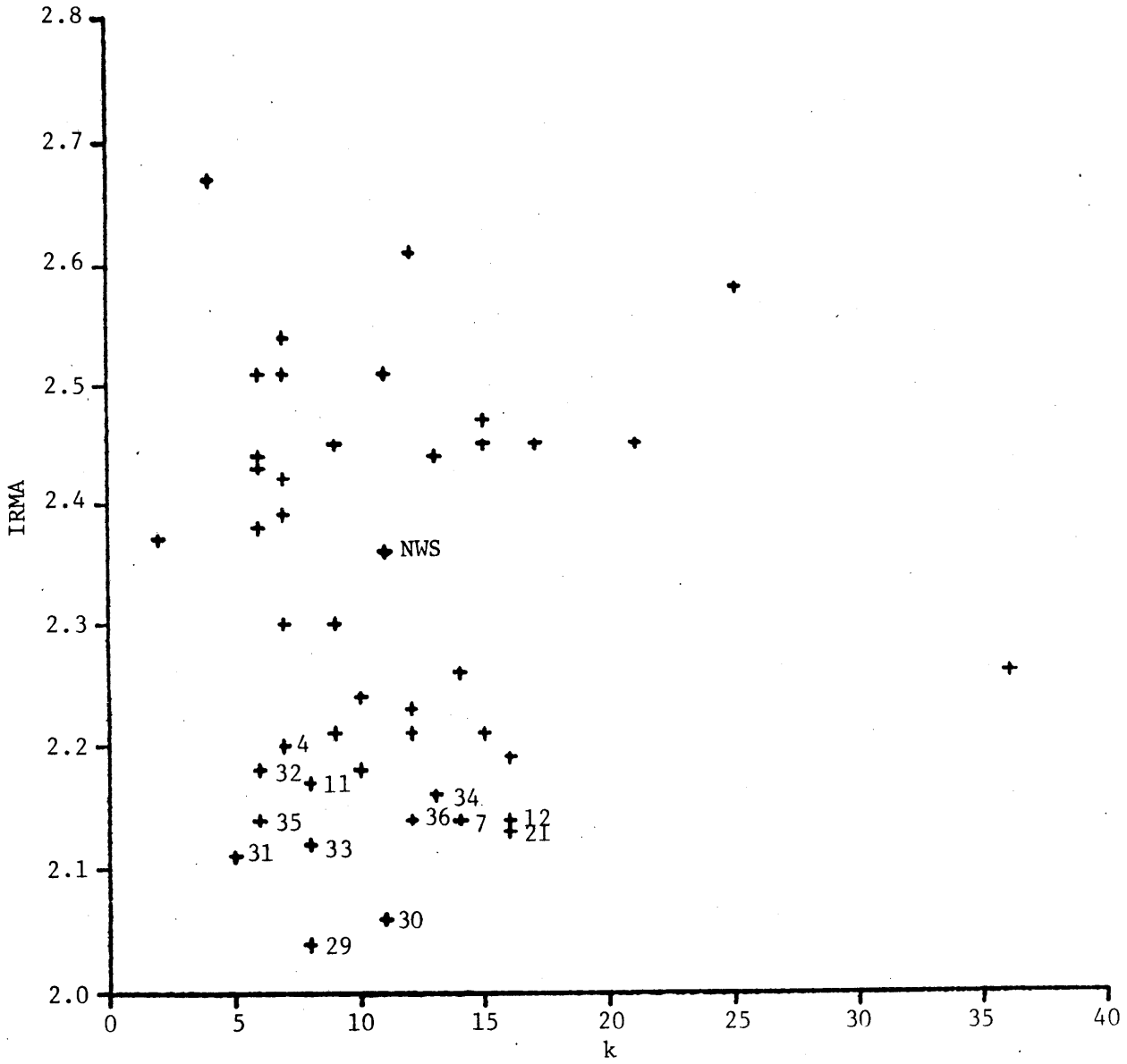

Figure 6.2 The Model Numbers are from Table 5.5

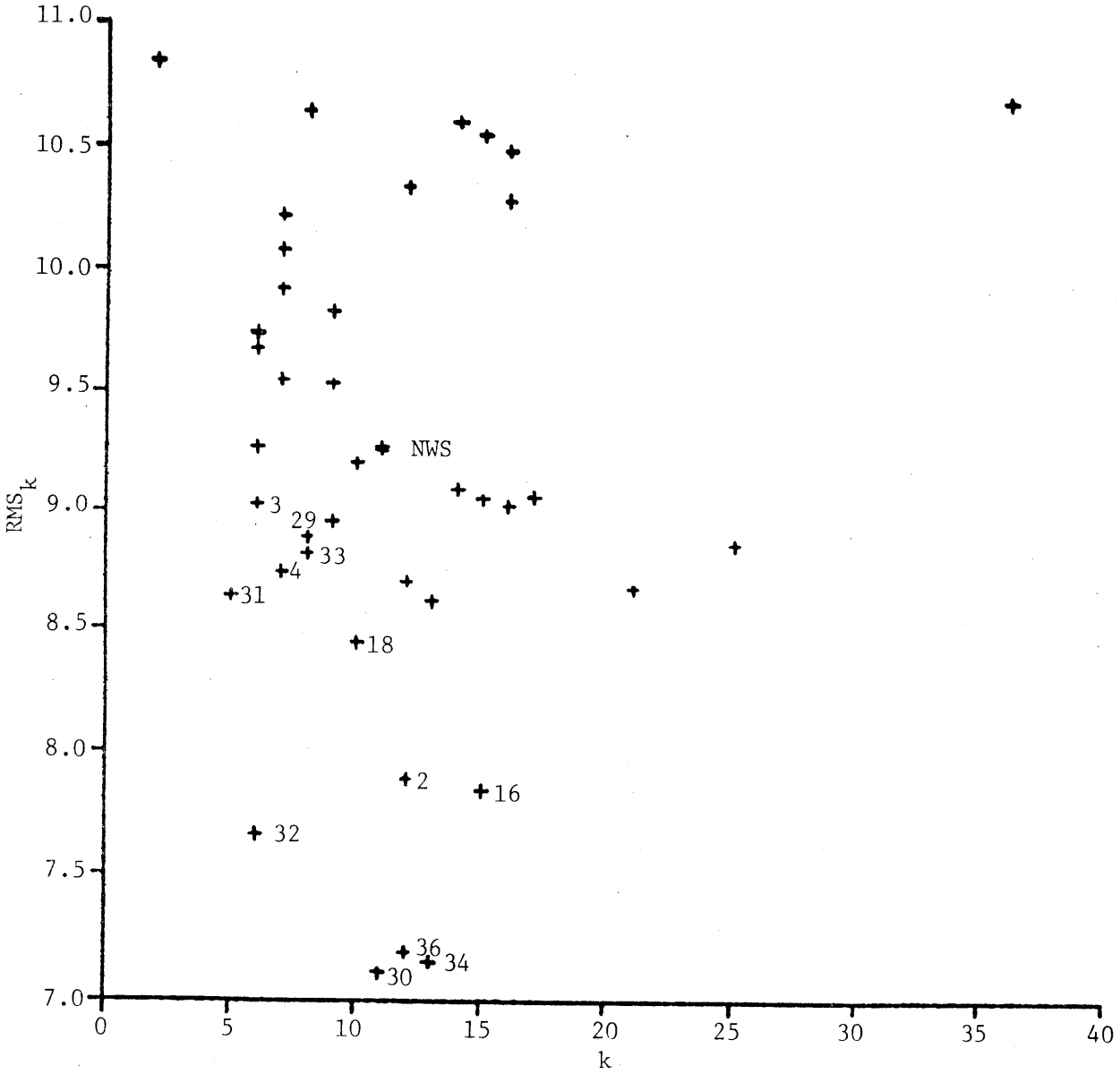

Figure 6.3 The Model Numbers are from Table 5.5

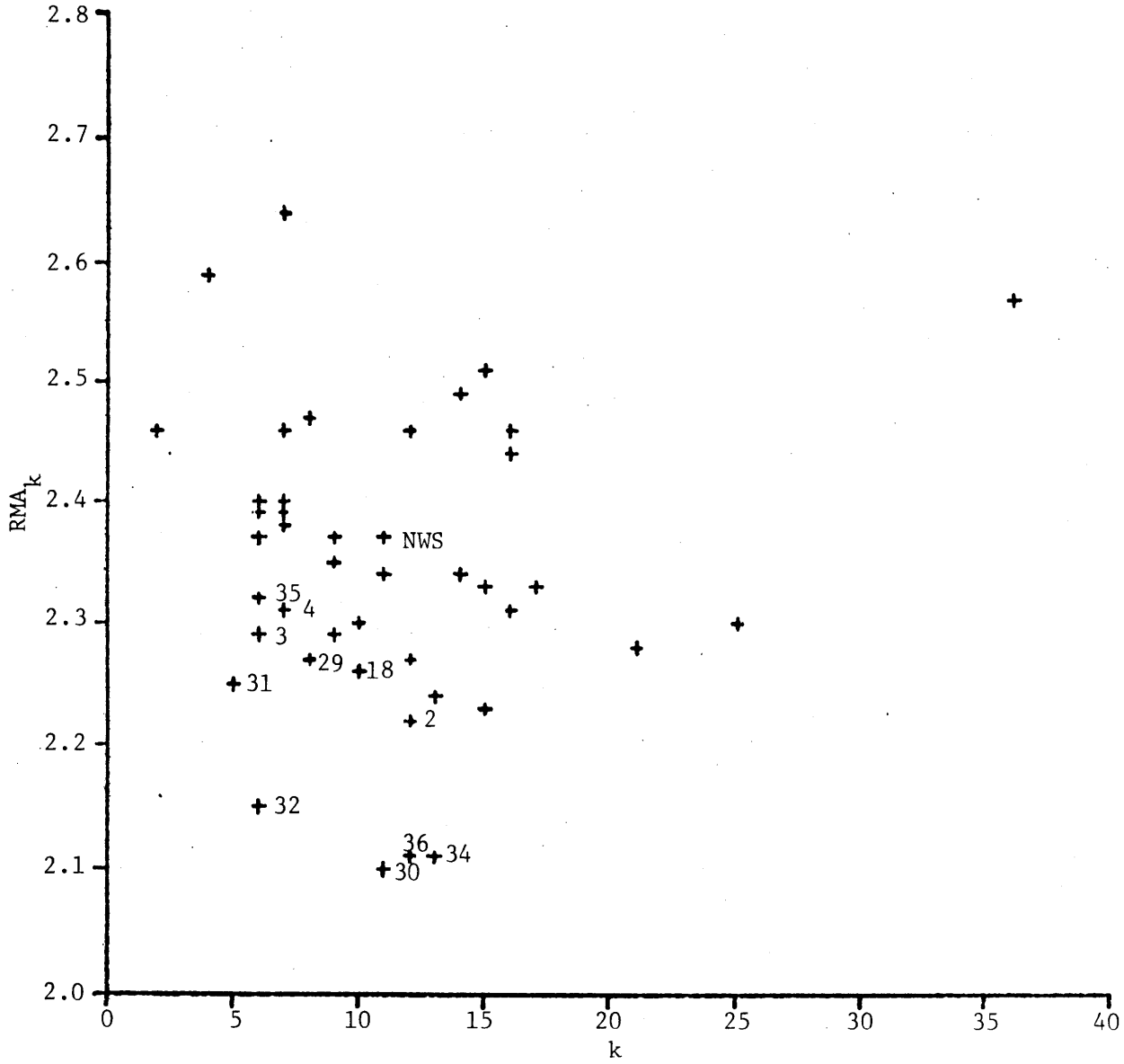

Figure 6.4 The Model Numbers are from Table 5.5

represented by points closer to the lower and left boundaries of the figures are preferred over other models. The identification numbers of the models on and close to the preferred boundaries are indicated in the figures. However, as discussed in Section 3.2.3, there are no definite rules to guide the trade-off between improved model statistics and increased model complexity. Also, subjective judgement concerning model qualities may sometimes induce the choice of a model not on or near the preferred boundaries. For example, if model stability were an overriding concern, the models developed from data set 5 (containing principal components) might be preferred in spite of their relatively poor statistics.

Models 30, 31, and 32, all linear models from data set 3, define the preferred boundaries for IRMS, RMS_k, and RMA_k. Model 4, from data set 1, and models 34 and 36 from data set 4, again linear models, appear close to the boundaries for each of these three statistics. The preferred boundary of IRMA is not clearly defined, but is generally dominated by linear models from data sets 3 and 4. Thus, while the GMDH may be successful in some situations, linear relations between the dependent and independent variables seem to produce the best prediction equations in this case. Using the GMDH on data sets 3 and 4 would probably have produced better nonlinear models than were produced from data sets 1, 2, and 5, but we expect that they still would have been dominated by the linear models.

The models from data set 3 clearly have the best model quality

statistics among the 5 data sets. However, data set 3 also requires much more effort to produce than data sets 1,2, and 4. In situations where this extra work is prohibitive we may wish to consider the other data sets. The creation of data set 2 requires nearly as much effort as data set 3, but the models from data set 2 are generally dominated by the models from data sets 1 and 4. Data set 4 requires only that the response be detrended with a model developed from a long data base,substantially less effort than required by data set 3, and data set 1 is the original data. Models 34 and 36, from data set 4, are generally close to the preferred boundaries for higher values of k. Some linear models developed from data set 4 by tightening the significance levels for entering and deleting variables in the stepwise regression algorithm (not presented in this work) were also close to the preferred boundaries at lower values of k. In summary, data set 3 yields the best models but requires the most effort to create; data set 4 yields reasonably good models and requires substantially less effort to create than data set 3, data set 1 yields reasonably good models only at lower values of k, and data set 2 requires more effort to create than data sets 1 or 4, but yields poorer models.

The Durbin Watson (DW) statistic d is normally used as a validation procedure for a few selected models and cannot easily be used to rank different models. However, it can, in this case, be used to help evaluate the different data sets. The DW statistics of the models are listed in Table 6.2. A "y" ("n") in the 5% and 1% columns

Model Number	k	Data Set	d	5%	1%
1	12	1	1.82	?	n
2	12	1	1.98	n	n
3	6	1	1.88	n	n
4	7	1	1.92	n	n
5	2	1	1.89	n	n
6	16	1	1.87	?	n
7	14	1	1.83	?	n
8	9	1	1.92	n	n
9	36	1	1.86	?	?
10	2	1	1.89	n	n
11	8	1	1.89	n	n
12	16	1	1.84	?	n
13	7	1	1.97	n	n
14	12	1	1.95	n	n
15	13	2	1.90	n	n
16	15	2	1.95	n	n
17	9	2	1.59	y	y
18	10	2	2.02	n	n
19	7	2	1.76	y	?
20	10	2	1.89	n	n
21	16	2	1.85	?	n
22	25	2	1.84	?	?
23	21	2	1.89	?	n
24	7	2	1.70	y	y
25	9	2	1.76	y	?
26	14	2	1.71	y	?
27	17	2	1.75	?	?
28	15	2	1.74	y	?
29	8	3	1.79	?	n
30	11	3	1.87	?	n

Table 6.2 Durbin-Watson Test Results

y indicates serial correlation
n indicates no serial correlation
? indicates the test was inconclusive

Note that Table 3.1 uses k' = k-1

Model Number	k	Data Set	d	5%	1%
31	5	3	1.85	n	n
32	6	3	1.81	?	n
33	8	4	1.59	y	y
34	13	4	1.84	?	n
35	6	4	1.59	y	y
36	12	4	1.87	?	n
37	7	5	1.64	y	y
38	11	5	1.89	n	n
39	6	5	1.64	y	y
40	6	5	1.71	y	y
41	4	5	1.62	y	y
42	15	5	1.69	y	?
43	15	5	1.69	y	?
44	7	5	1.63	y	y
45	15	5	1.69	y	?
46	6	5	1.71	y	y
47	6	5	1.77	y	?
NWS	11	1	1.73	y	?

Table 6.2 Durbin-Watson Test Results (cont'd)

y indicates serial correlation
n indicates no serial correlation
? indicates the test was inconclusive

Note that Table 3.1 uses k' = k-1

indicates the presence (absence) of serial correlation at the specified significance level. A question mark in either column indicates the test was inconclusive. The significance points of d for different values of k' (=k-1) are listed in Table 3.1. Table 3.1 is based on n = 350 because there are only 350 pairs of consecutive days from which to calculate d (see Table 5.2).

Serial correlation appears frequently in the models developed from data sets 2,4, and 5 and does not appear in the models developed from data sets 1 and 3. The serial correlation in models from data set 4 appears only when the coefficients have been selected using only just the first 3 years of data. Thus the quality of data set 3 is confirmed and models from data set 1 appear to be slightly less likely than models from data set 4 to exhibit the problem of serial correlation.

6.2 Model Validation

The application of validation procedures (see Section 3.4) to models 4, 32, and 36, from data sets 1, 3, and 4, are described in this section. Hypothesis tests, coefficient stability, and residual graphics are examined. Model 4 was examined in greater detail than models 32 and 36. Some revisions to model 4 suggested by the graphic validation procedures, including removing 1 variable, are also considered. Only hypothesis tests and residual graphics are examined for models 32 and 36.

6.2.1 Hypothesis Tests

The F and Durbin-Watson hypothesis tests were applied to models 4, 32 and 36. These tests are described in Section 3.4.2. The null hypothesis in the F test is that the change in the sum of squared residuals due to the presence of a variable in the model is equal to 0. This hypothesis is rejected at the 99 percent significance level for each variable in each of the 3 models because F test at that level was applied to each variable as part of the stepwise regression variable selection algorithm. The residuals are plotted against normal cumulative probability distribution functions in Figures 6.5, 6.6 and 6.7. There are no guidelines for accepting or rejecting the validity of the F test. It is simply subject to more or less suspicion as the residuals are less or more normally distributed.

The null hypothesis in the Durbin Watson test is that the residuals do not have positive first order serial correlation. This hypothesis is not rejected at the 5 percent level for model 4 and not rejected at the 1 percent level for models 32 and 36 (see Table 6.2).

6.2.2 Coefficient Stability

The examination of model stability was discussed in Section 3.4.3. Only coefficient stability of model 4 is considered here. Coefficient stability is examined by estimating the coefficients on different subsets of data and comparing the results. Estimates developed using each year of data separately are shown in Table 6.3. The

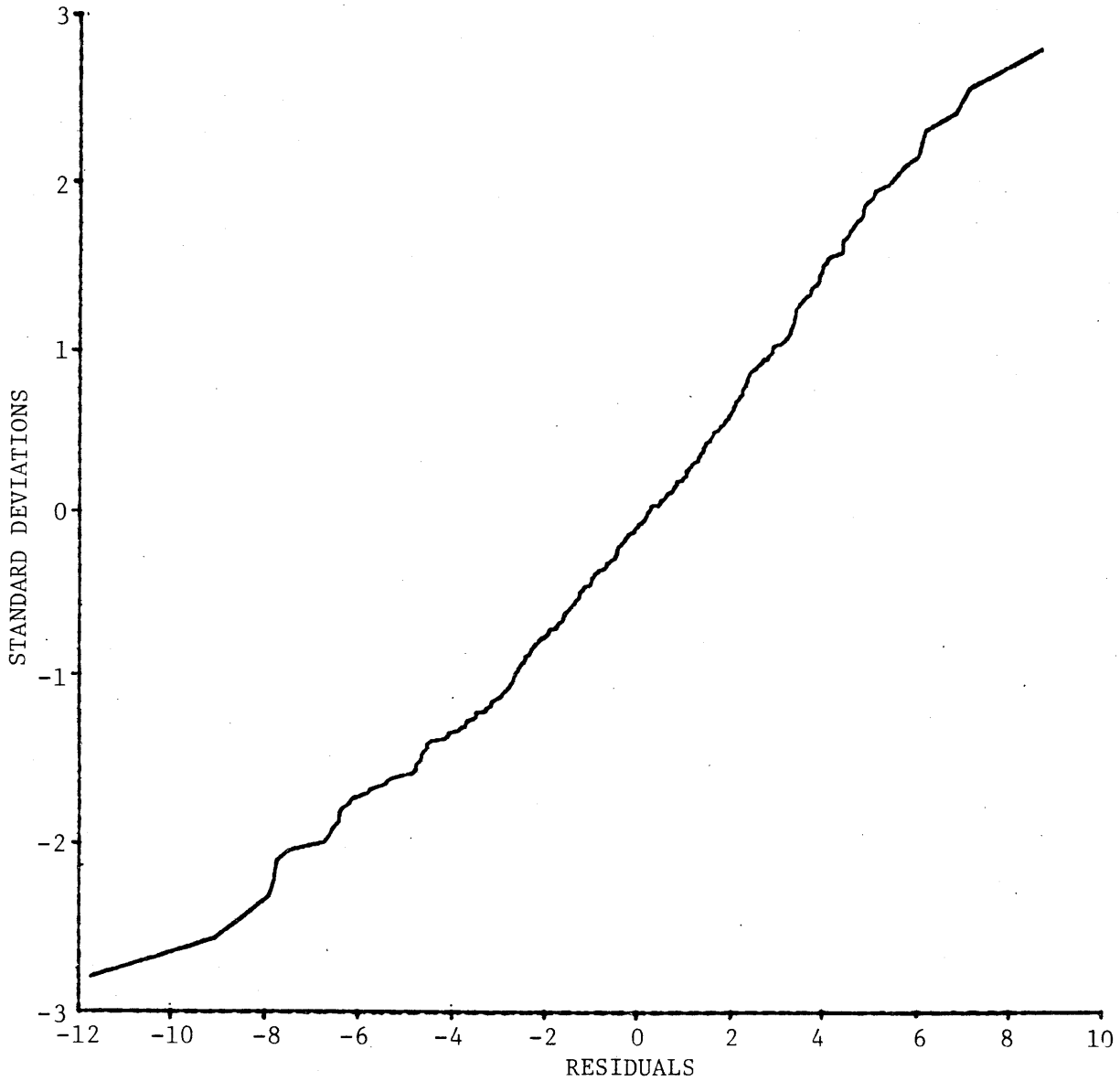

Figure 6.5 Normal Plot of Residuals, Model 4

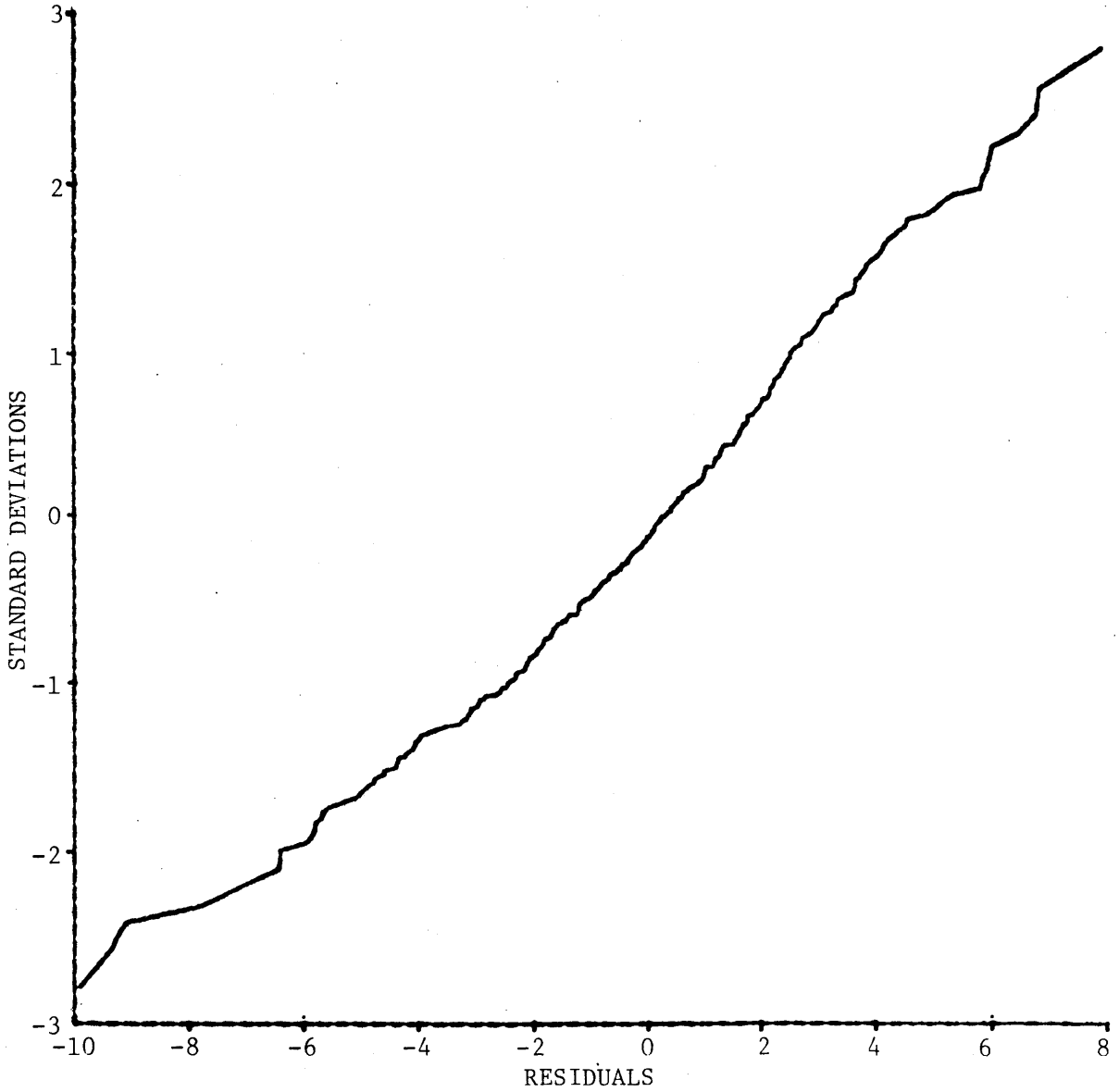

Figure 6.6 Normal Plot of Residuals, Model 32

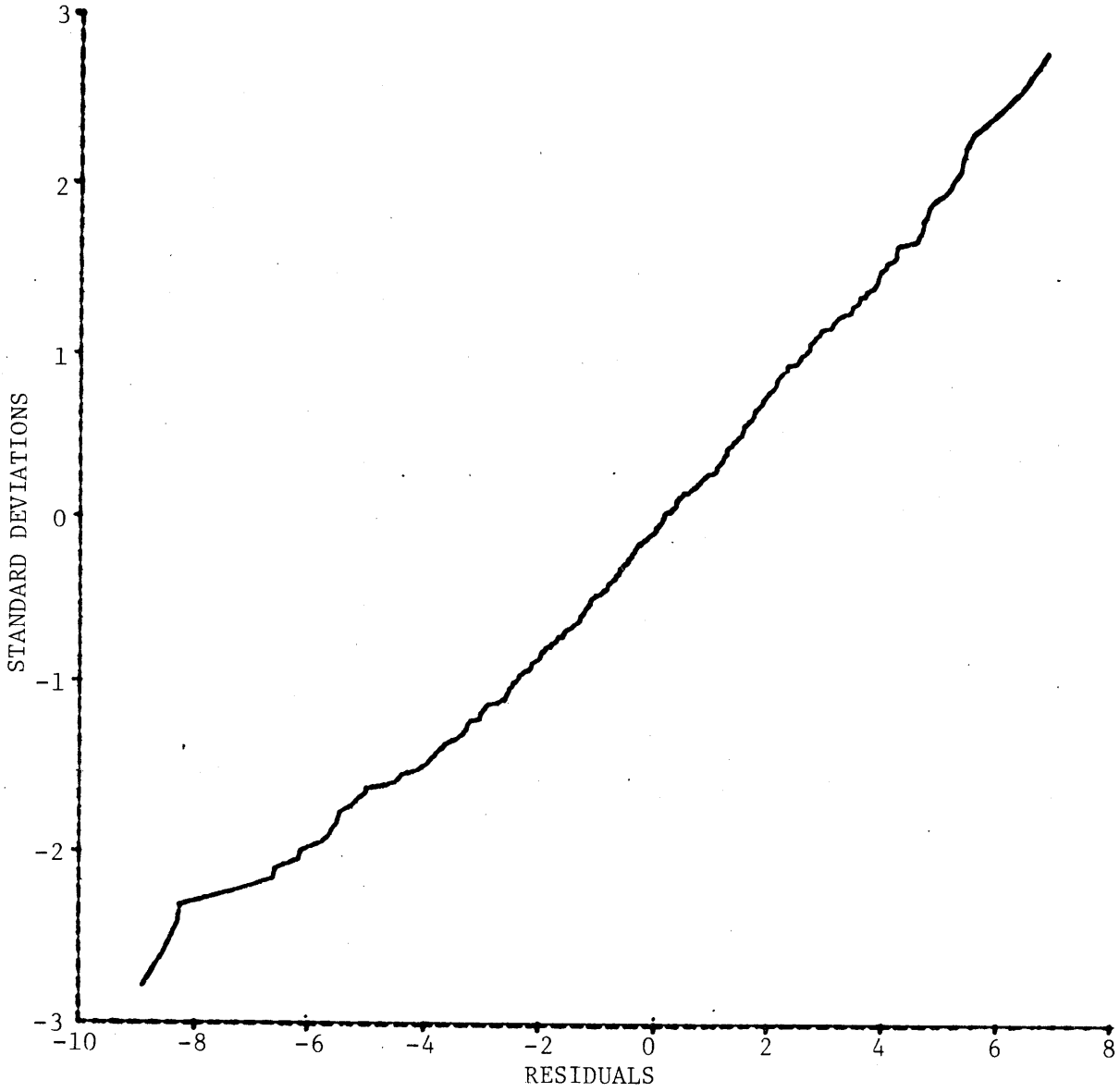

Figure 6.7 Normal Plot of Residuals, Model 36

	Year of Estimation Data						
Variable	1973	1974	1975	1976	1977	1973 through 1977	max. diff. ÷ 5 yr. coeff.
Constant	-320.852	-292.064	-260.383	-330.757	-384.470	-302.44	0.45
7	0.005	0.021	0.027	0.014	0.012	0.016	1.38
19	0.276	0.233	0.194	0.272	0.317	0.244	0.50
74	409.372	1,291.311	488.303	1,163.970	1,486.543	977.13	1.10
102	0.644	0.302	0.409	0.383	0.867	0.492	1.15
104	0.028	0.200	0.061	0.152	0.030	0.127	1.38
117	0.109	0.183	0.382	0.138	0.072	0.225	1.38

Table 6.3 Coefficients of Model 4, Estimated from Different Portions of the Data

maximum difference between the various estimates, divided by the estimate based on all 5 years of data, is also listed in Table 6.3. Since there are no reversals of sign or extreme changes in magnitude as the estimation data changes, no serious instability is indicated.

6.2.3 Graphic Analysis

In this section we examine residual and partial residual plots. These graphics were discussed in Section 3.4.1.

6.2.3.1 Residuals in time sequence

The standardized residuals of models 4, 32, and 36 are plotted in time sequence in Figures 6.8, 6.9, and 6.10. The residuals have been standardized by dividing by their standard deviation. The standard deviation varies between models, being 3.0 for model 4, 2.8 for model 32, and 2.7 for model 36. The mean of the residuals over the data used to estimate the coefficients is 0. It is fairly clear that there are no strong trends with periods greater than 1 season. However, there appears to be a bias towards large negative residuals for each model and there may be some trends within seasons.

Each model has more residuals below -2 than above +2. Model 4 has 6 above and 15 below, model 32 has 9 above and 16 below, and model 36 has 4 above and 15 below. Thus each of the models is much more likely to severely overpredict than to severely underpredict. Also, no residual for any of the models is greater than +3, but each model has some residuals less than -3. The effect of removing the data point associated with the -4 residual in model 4 (day 358) is

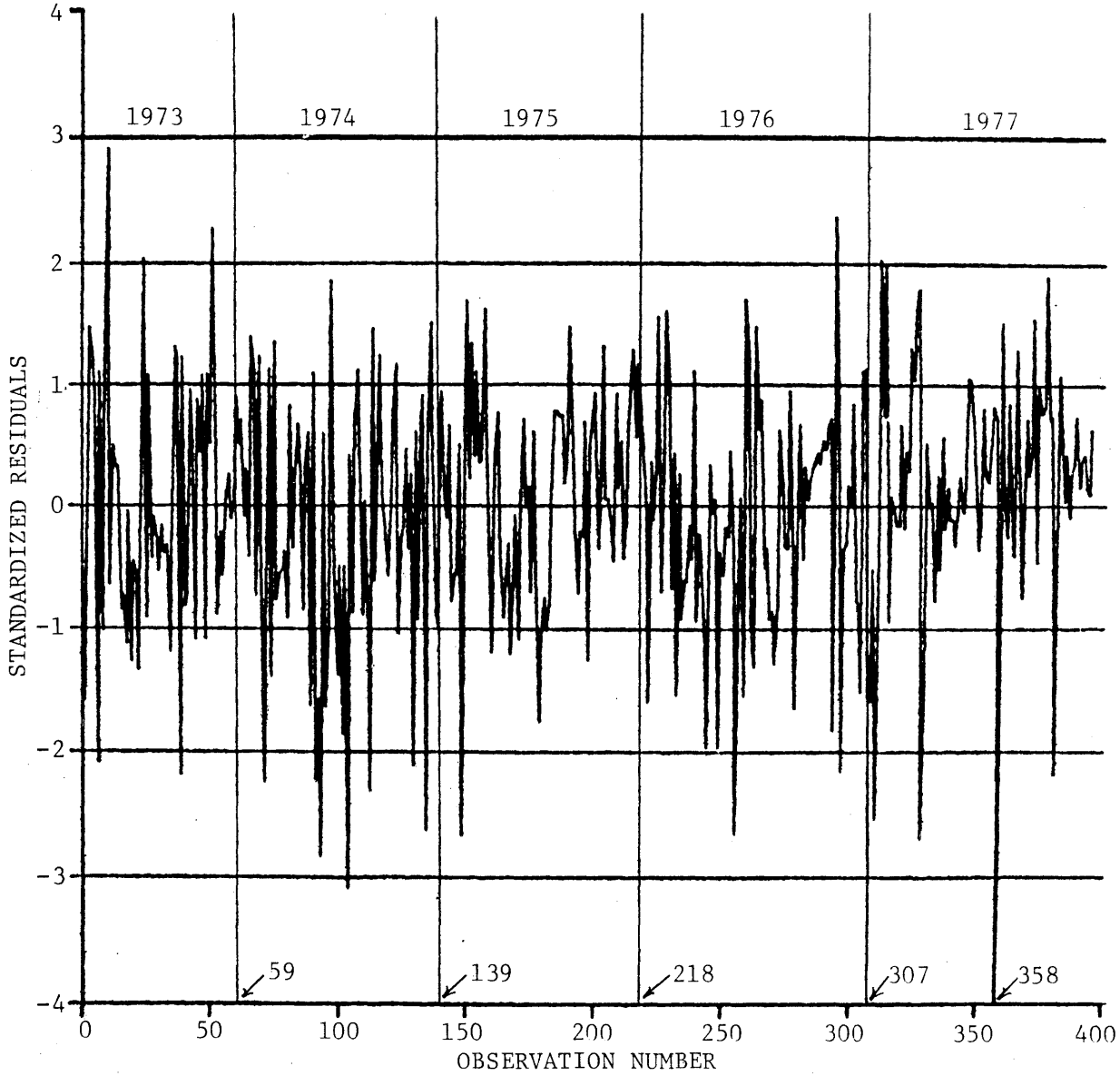

Figure 6.8 Standardized Residuals vs. Observation Number, Model 4

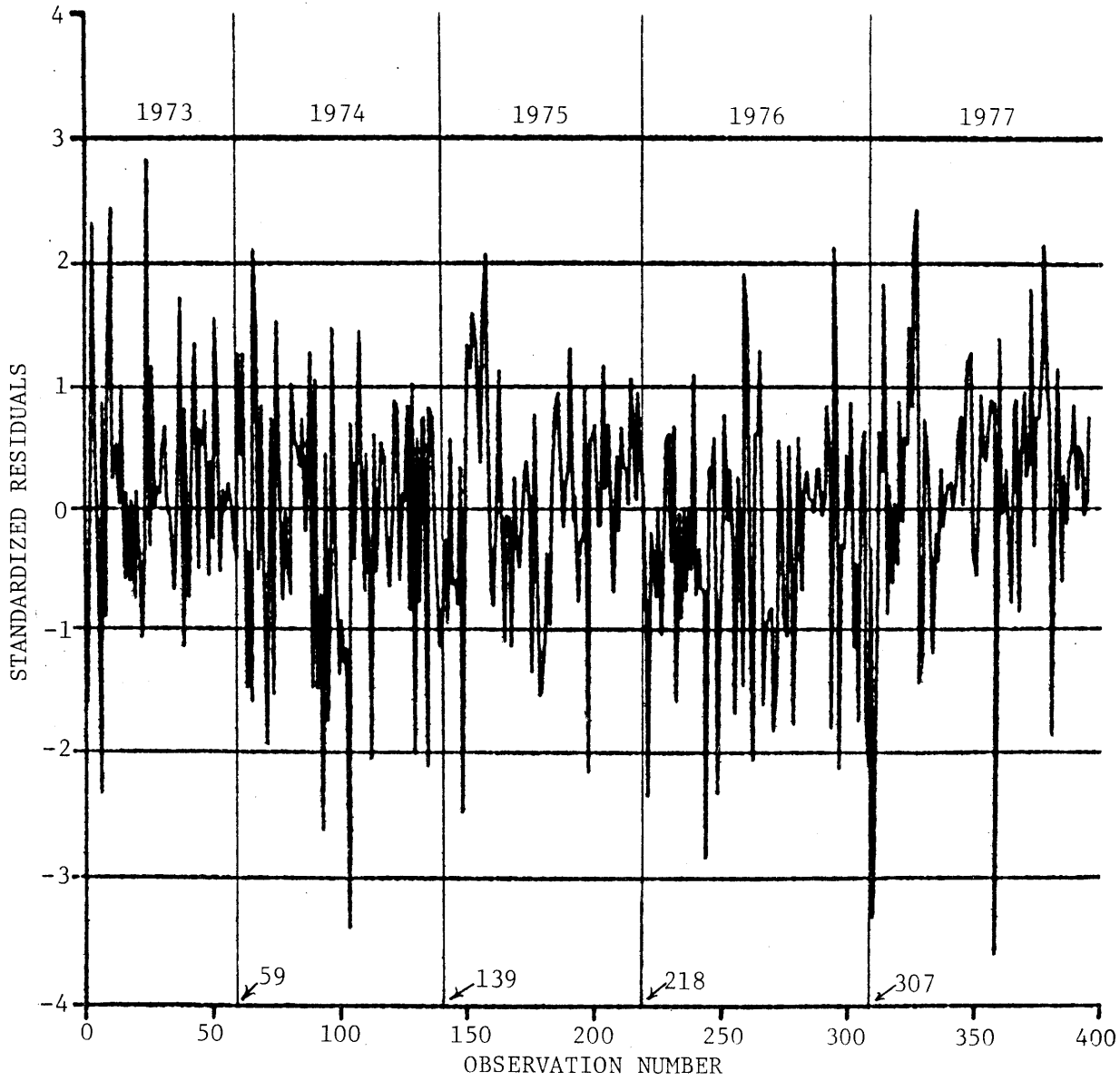

Figure 6.9 Standardized Residuals vs. Observation Number, Model 32

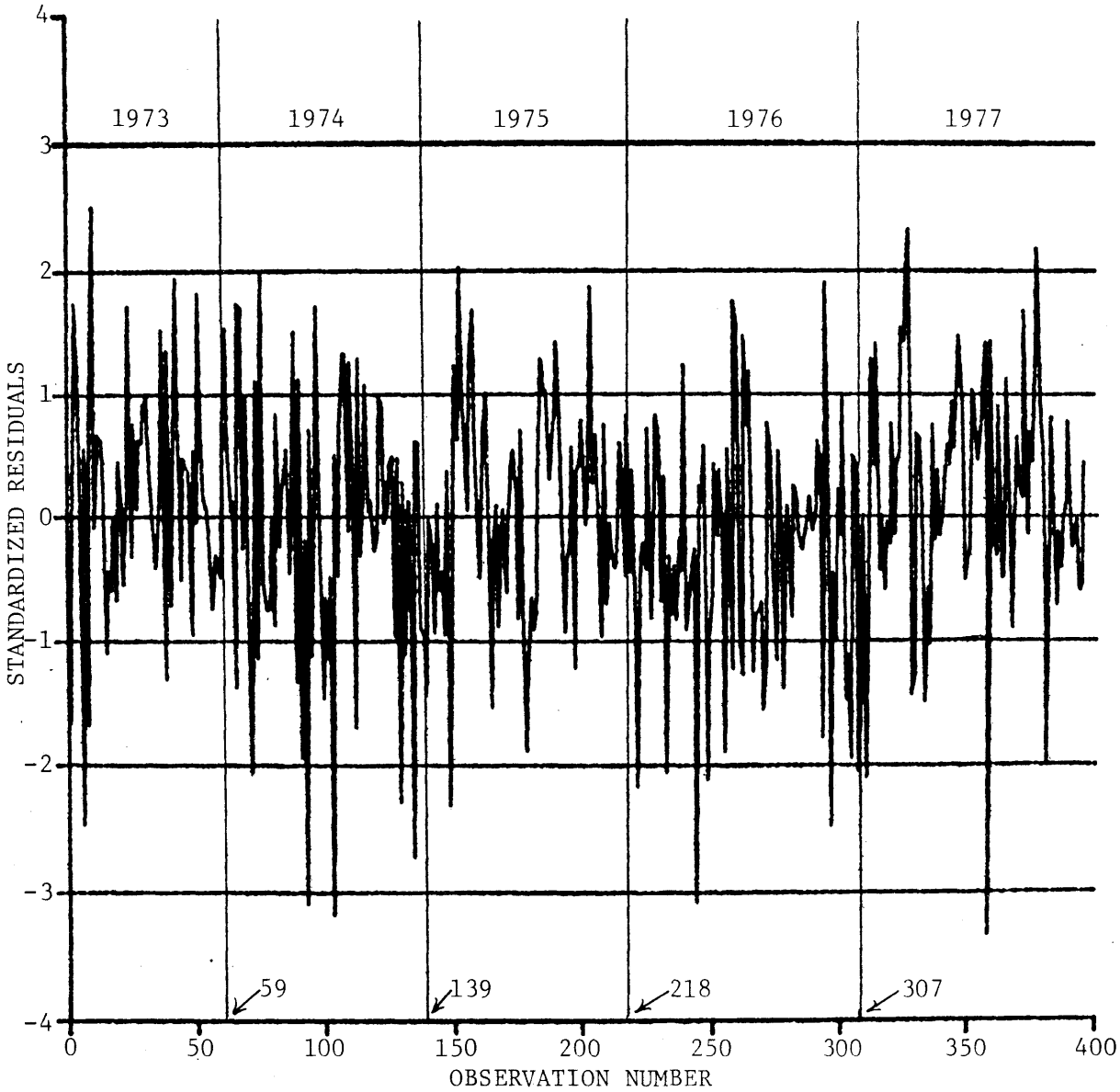

Figure 6.10 Standardized Residuals vs. Observation Number, Model 36

examined in Section 6.2.4.1.

High frequency variations in Figures 6.8, 6.9, and 6.10 make it difficult to visually detect patterns within seasons. Some of the high frequency variations were filtered out by calculating moving averages of 10 residuals. The averaging was performed only within seasons. The smoothed residuals are plotted in Figures 6.11, 6.12, and 6.13. Note the similarity between the plots for the different models. There also are some repeated within season patterns, but none which recur in every year. For example, in model 4 there is a pattern in years 1973, 1974, and 1976 which resembles the shape of the sine of twice the day of year. This pattern suggests that the inclusion of the sine of twice the day of year might improve the model by removing some of this pattern. However, the pattern is not as strong or clear in 1975 and 1977 and the new variable might radically increase errors in those years. That the sine of twice the day of the year was not chosen for inclusion in the model by the stepwise variable selection algorithm also indicates that the overall model quality is not improved, at the 99 percent significance level, by inclusion of that variable. No attempt was made to remove any of the patterns in the residuals from any of the models.

The absence of evidence that the residuals are either correlated or heteroscedastic indicates that the assumption $\underline{\Omega} = \underline{I}$ (see Section 3.3) is reasonable. Had the Durbin-Watson test indicated serial correlation or the residual plots indicated patterns of changing variance, a

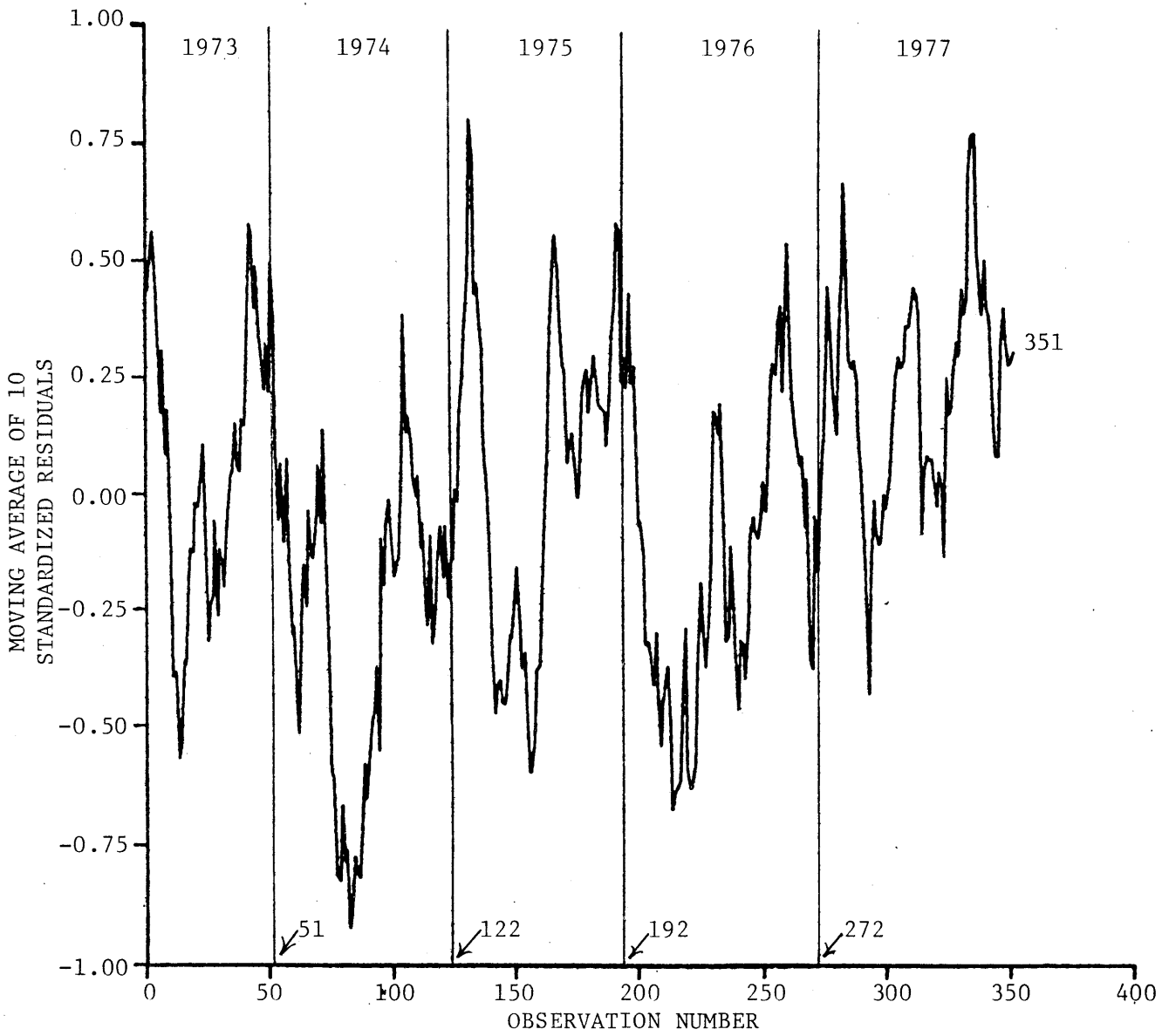

Figure 6.11 Moving Average of 10 Standardized Residuals vs. Observation Number, Model 4

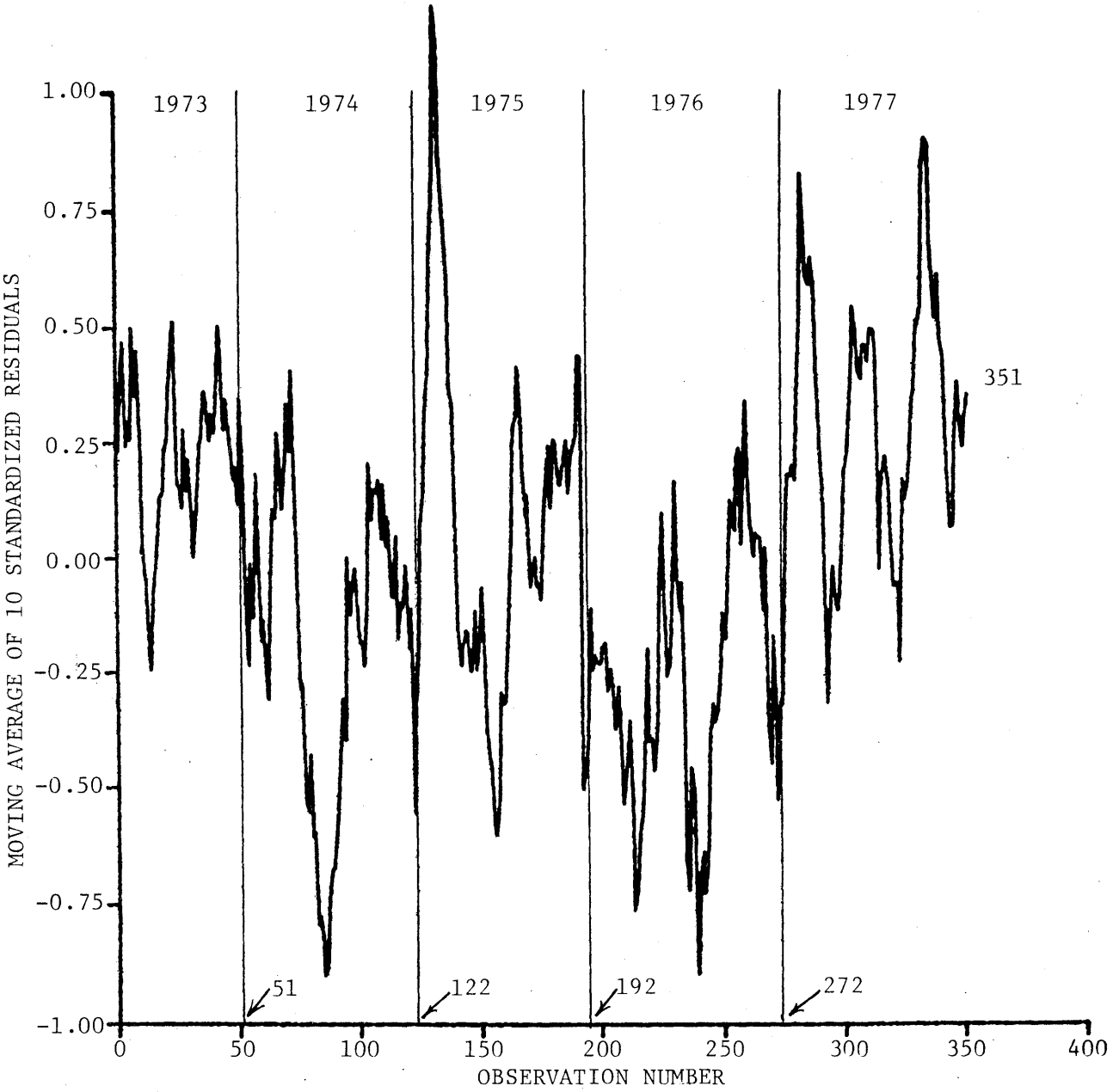

Figure 6.12 Moving Average of 10 Standardized Residuals vs. Observation Number, Model 32

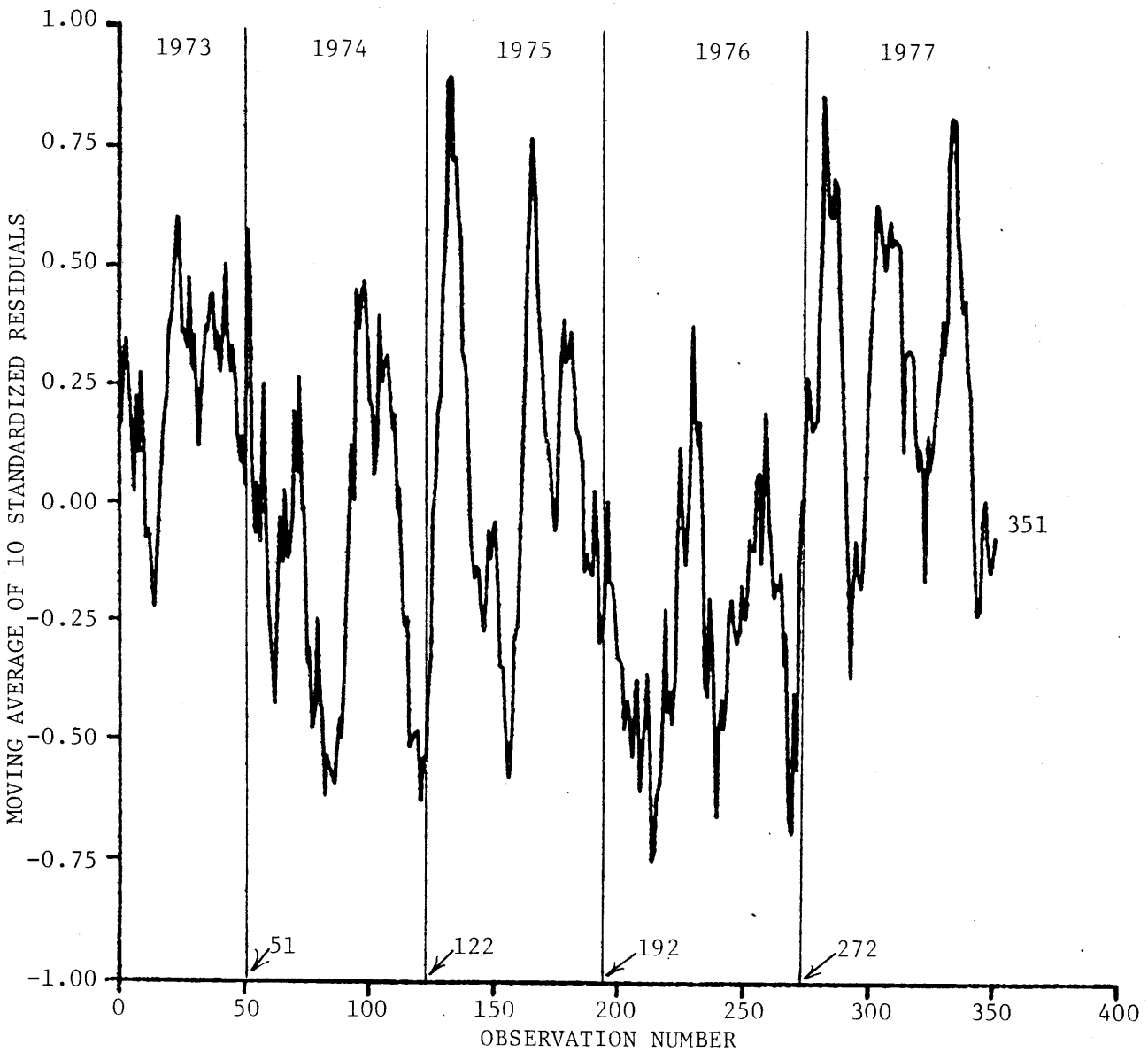

Figure 6.13 Moving Average of 10 Standardized Residuals vs. Observation Number, Model 36

reestimated $\underline{\Omega}$, as described in Section 3.3, would be expected to alleviate the problem.

6.2.3.2 Other residual plots

The standardized residuals are plotted against the predicted values of the independent variables in Figure 6.14, 6.15, and 6.16. Other than the bias towards large negative residuals noted in Section 6.2.3.1, no major patterns are evident in any of the plots. The reason for the diagonal bands in Figures 6.15 and 6.16 is not known.

The standardized residuals for model 4 are plotted against the independent variables in Figures 6.17 through 6.22. Some points well separated from the rest of the points were noted and are circled in Figures 6.19, 6.20, and 6.21. The effects of these outlying data points are examined in Section 6.2.4.2. These points could have been discovered before developing the model by plotting the independent variables in time sequence.

6.2.3.3 Partial residual plots, model 4

The partial residuals of the dependent variable are plotted against the partial residuals of each of the independent variables, given the presence in the model of the other 5 independent variables, in Figure 6.23 through 6.28. Recall that partial residual plots show the relation between an independent variable and the dependent variable when the effects of the other independent variables in the model have been removed.

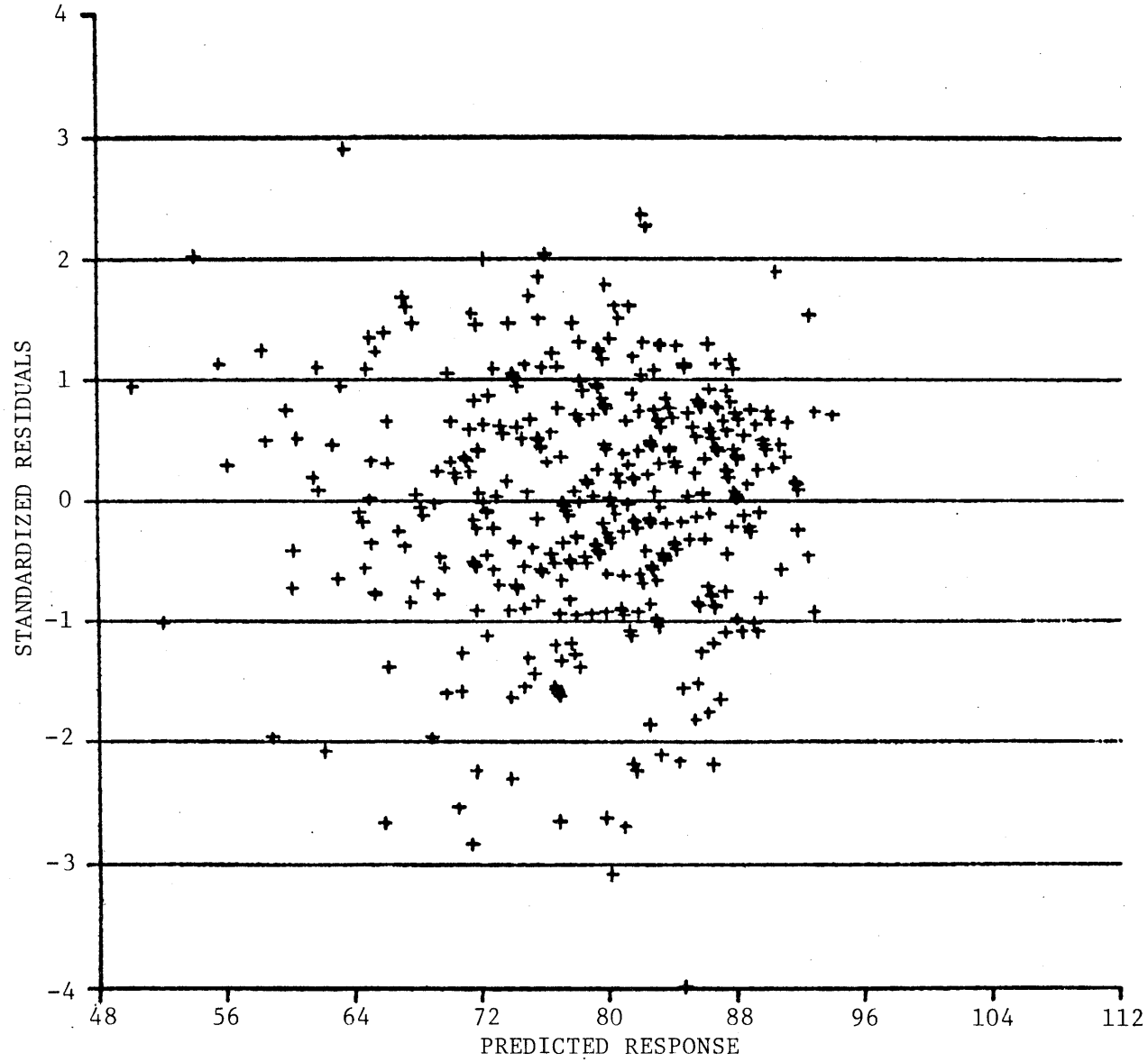

Figure 6.14 Standardized Residuals vs. Predicted Response, Model 4

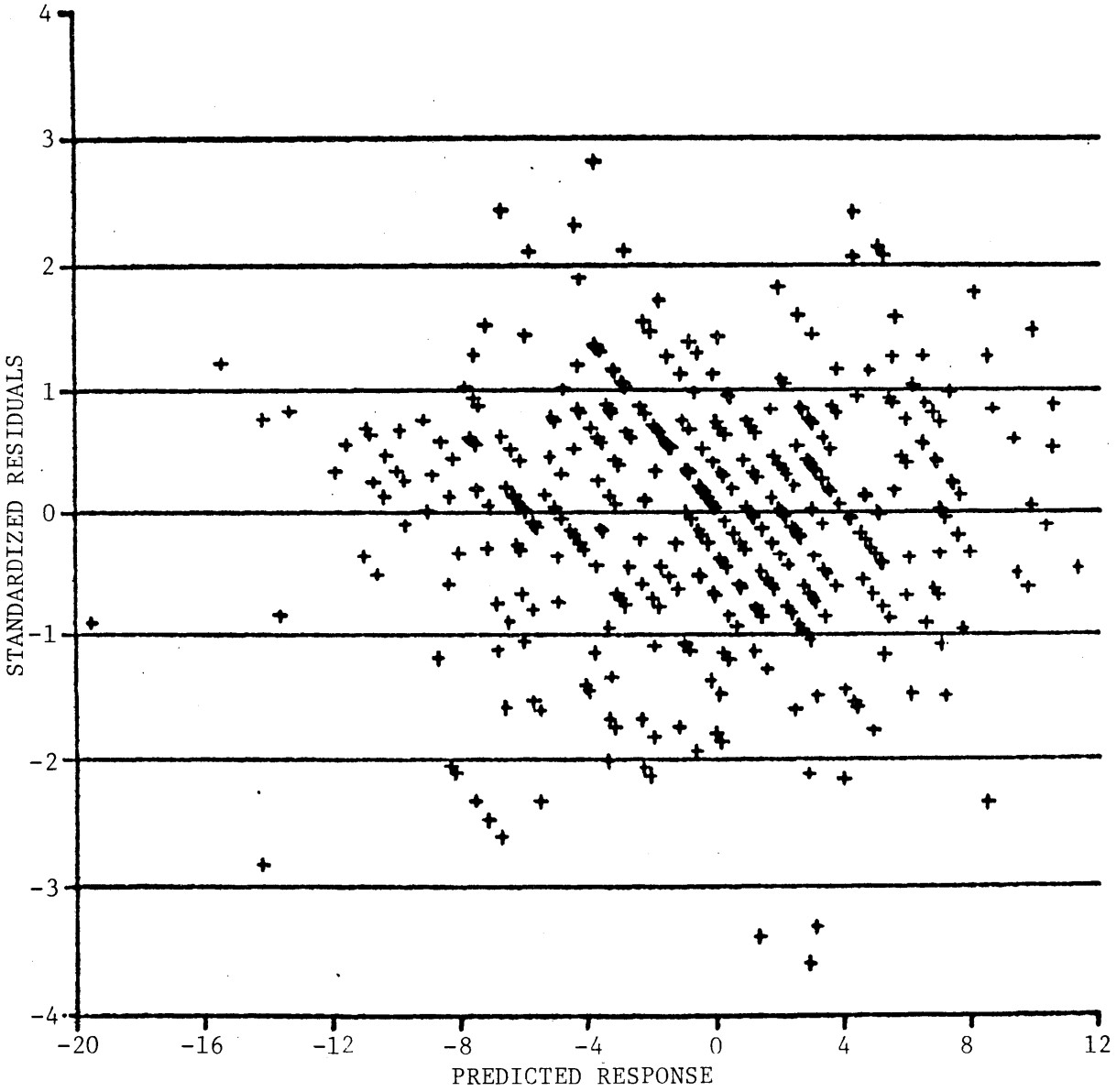

Figure 6.15 Standardized Residuals vs. Predicted Response, Model 32

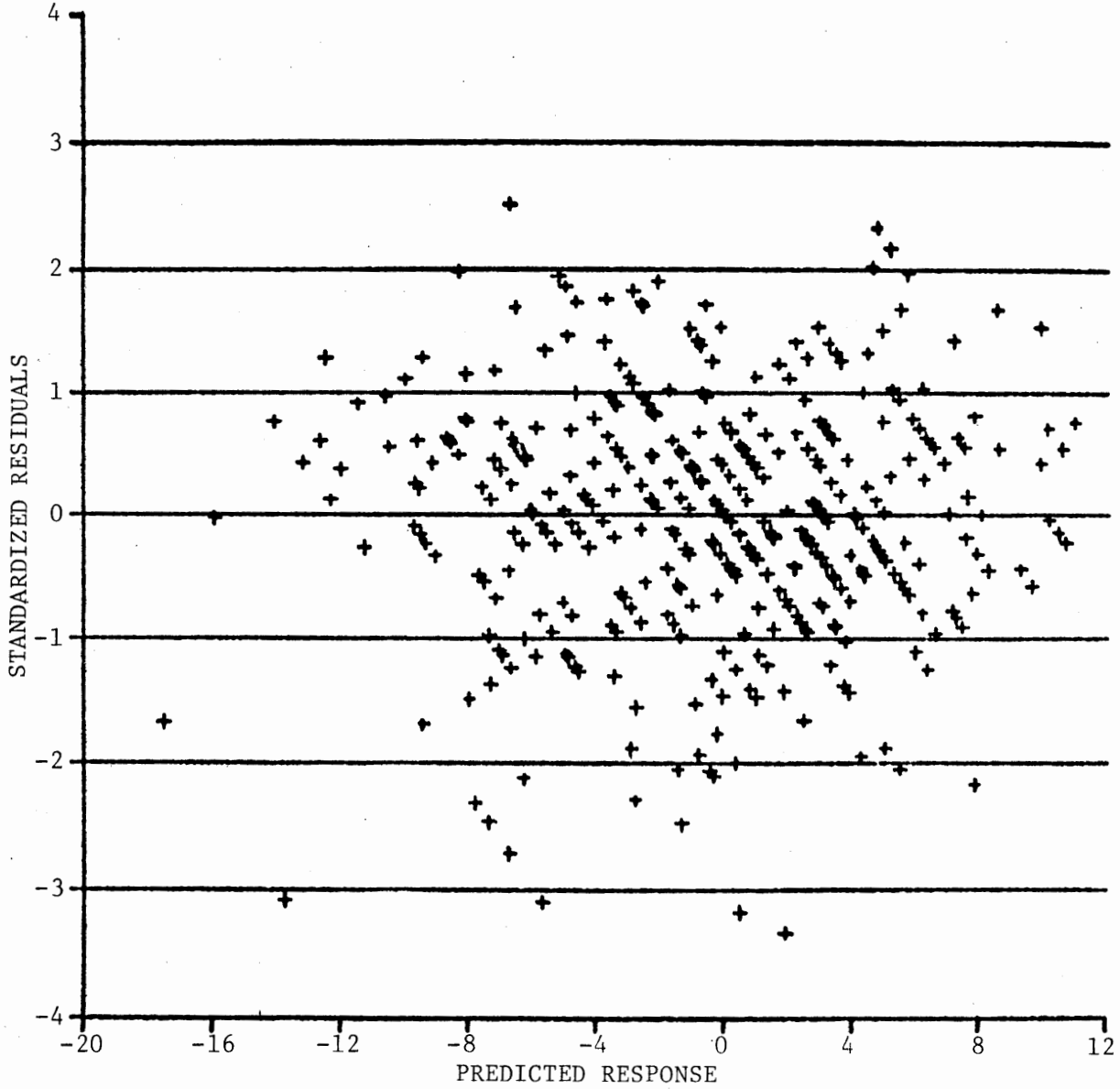

Figure 6.16 Standardized Residuals vs. Predicted Response, Model 36

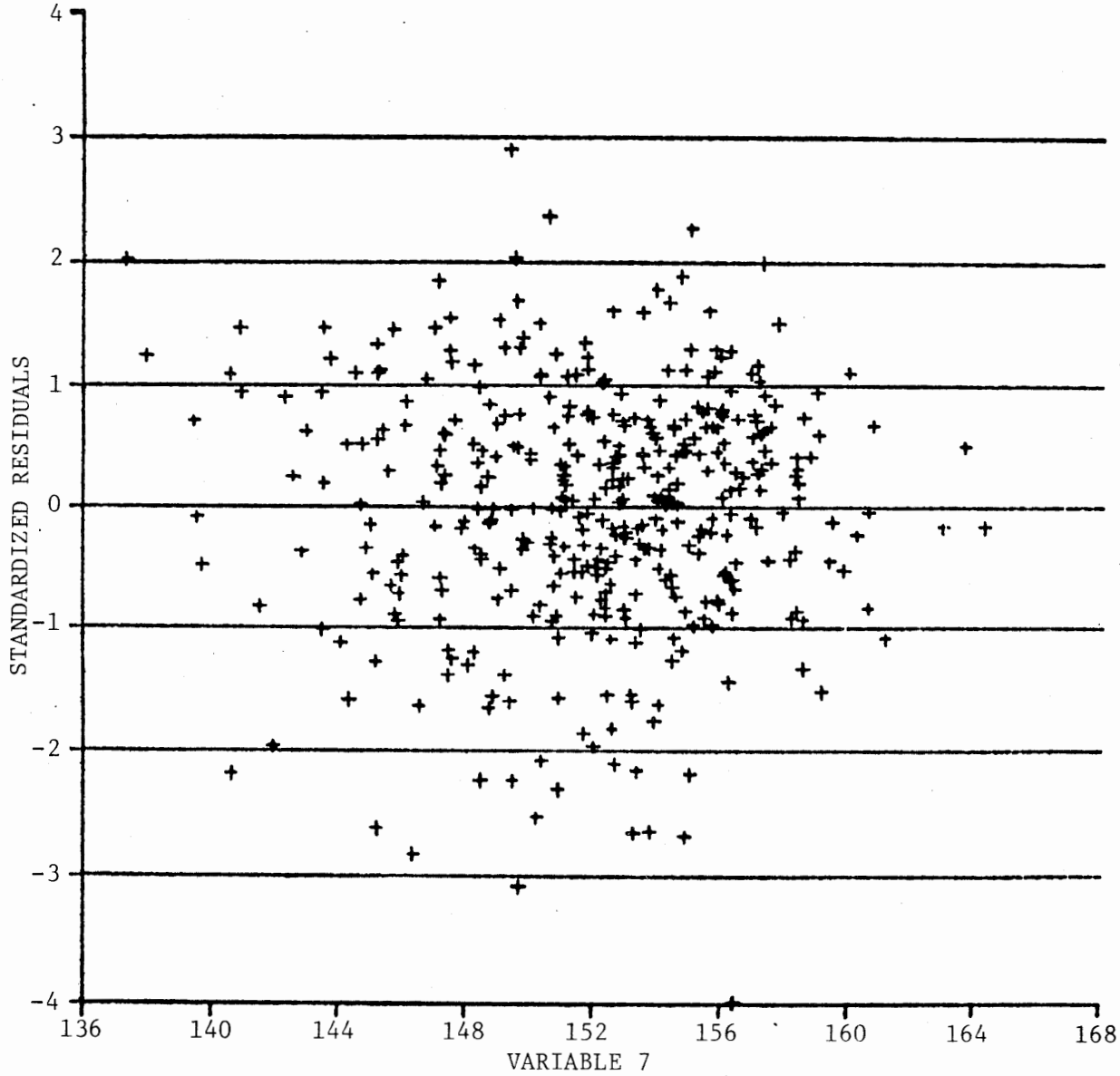

Figure 6.17 Standardized Residuals vs. Variable 7, Model 4

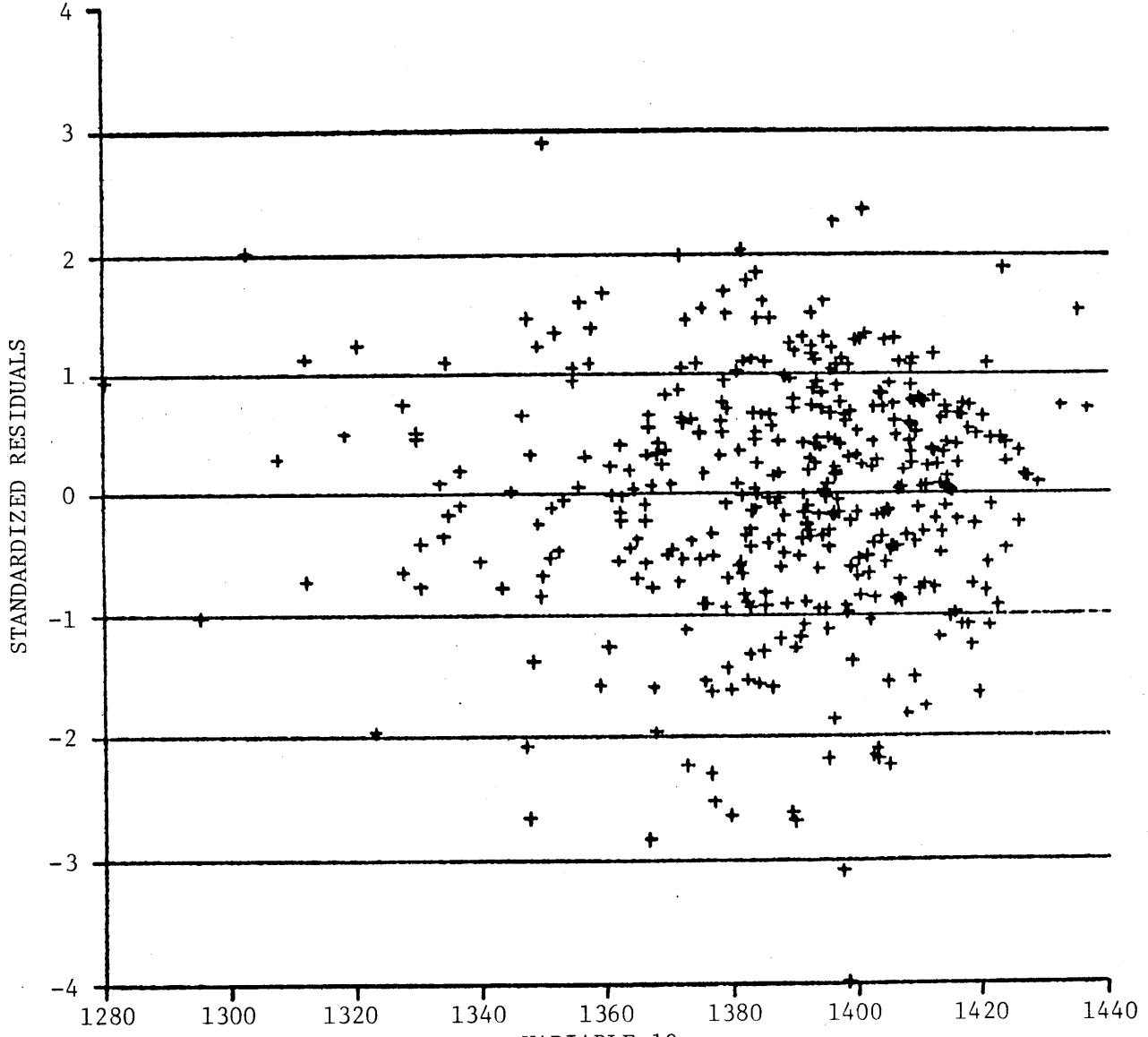

Figure 6.18 Standardized Residuals vs. Variable 19, Model 4

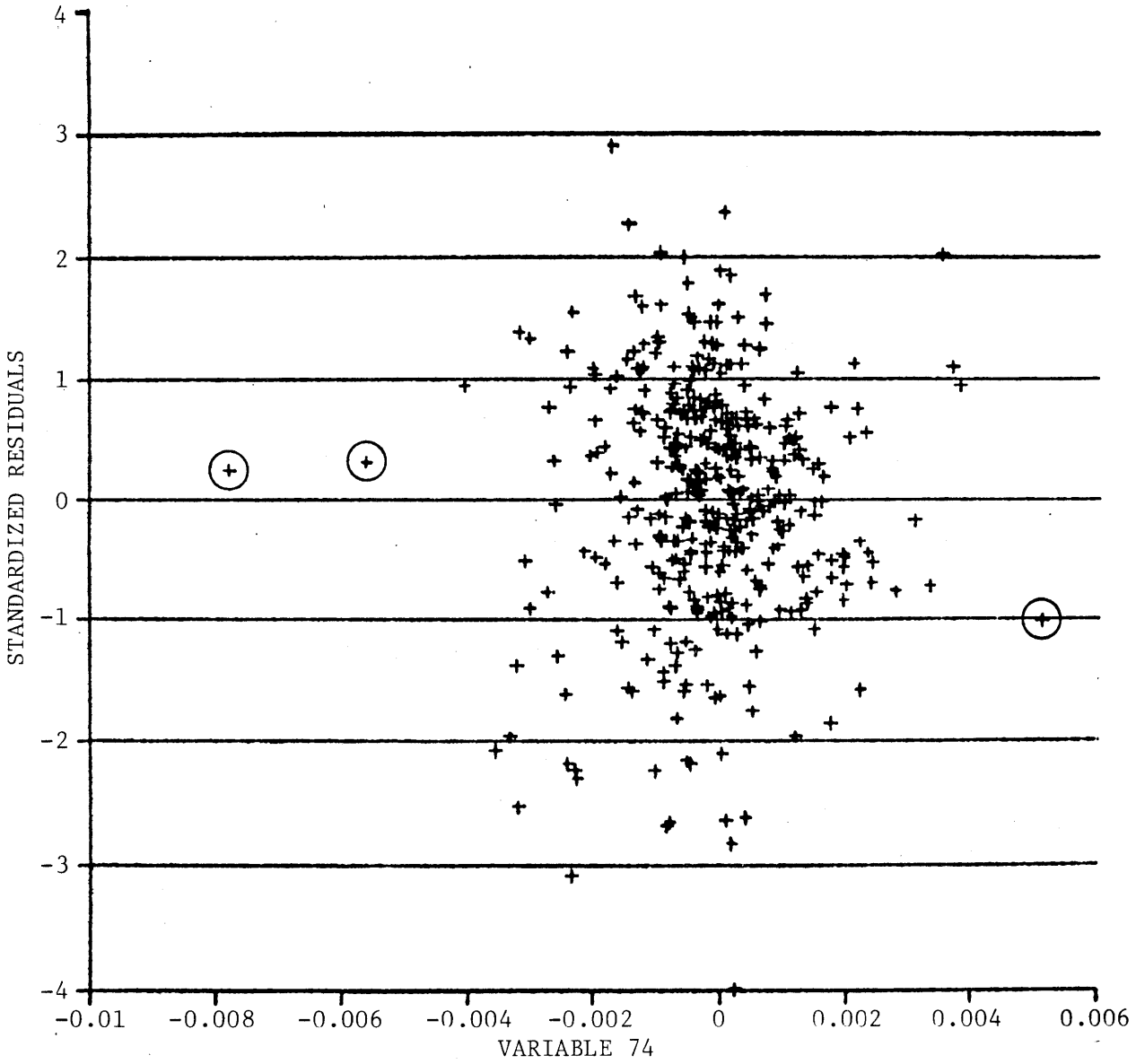

Figure 6.19 Standardized Residuals vs. Variable 74, Model 4

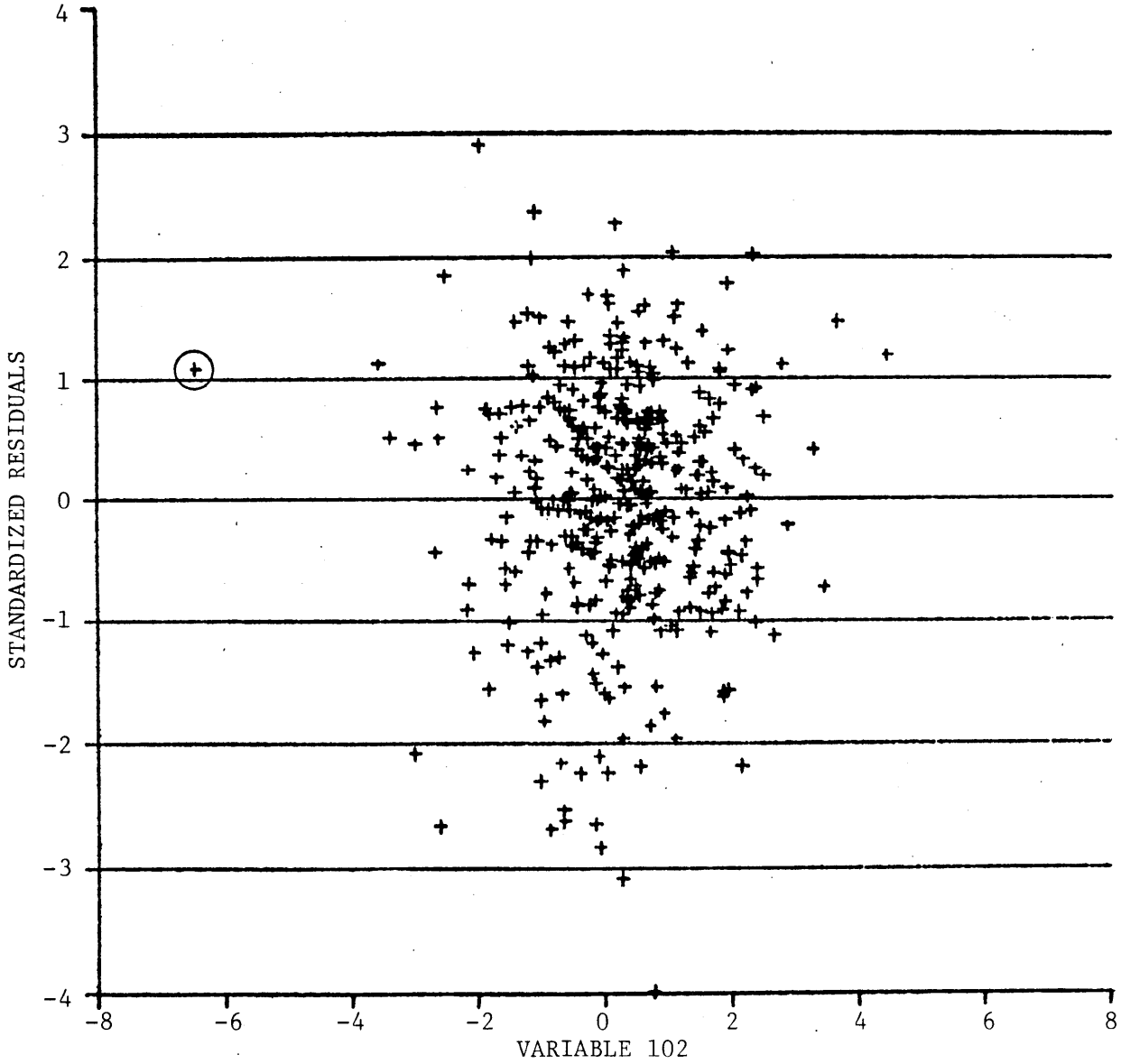

Figure 6.20 Standardized Residuals vs. Variable 102, Model 4

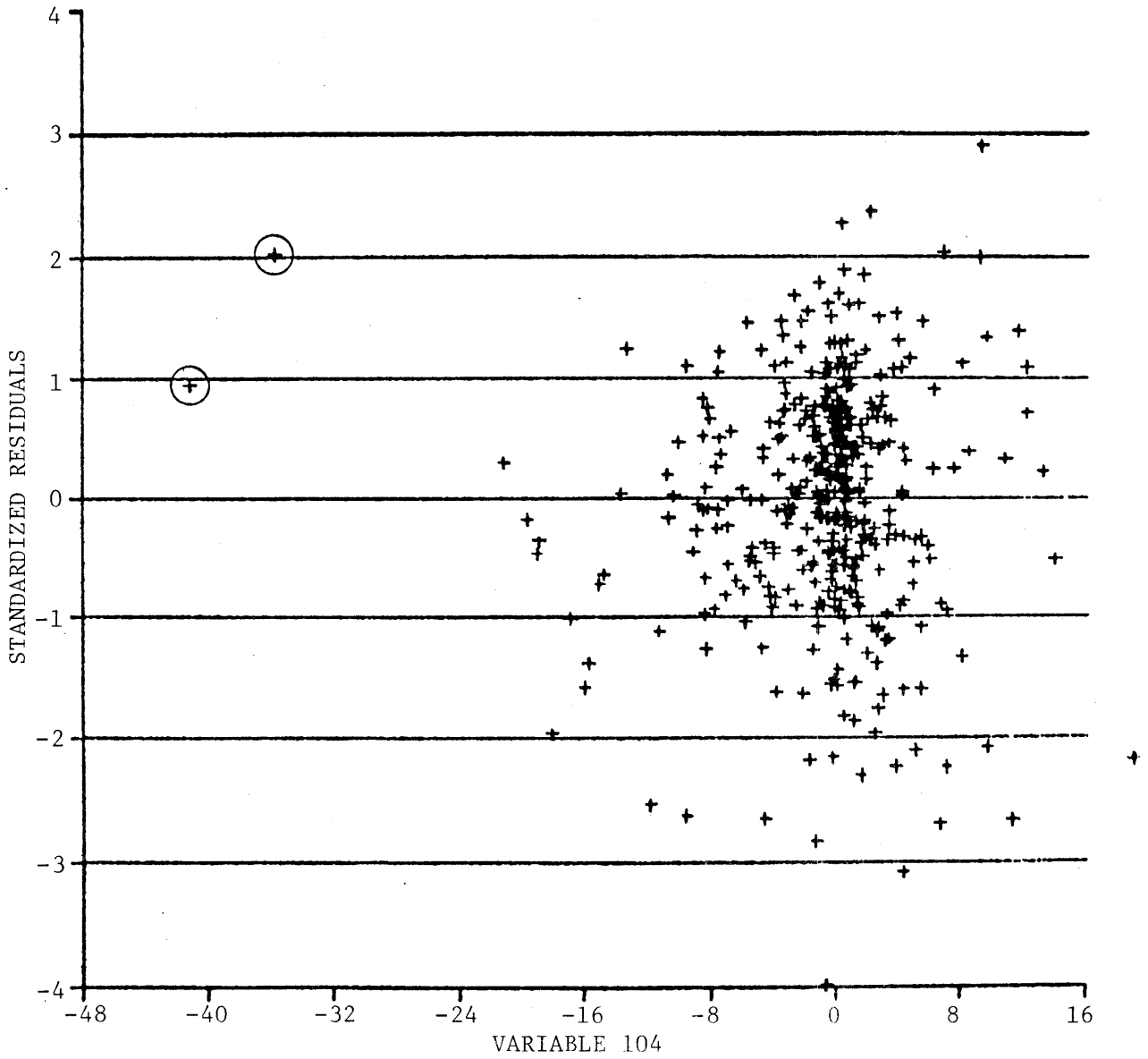

Figure 6.21 Standardized Residuals vs. Variable 104, Model 4

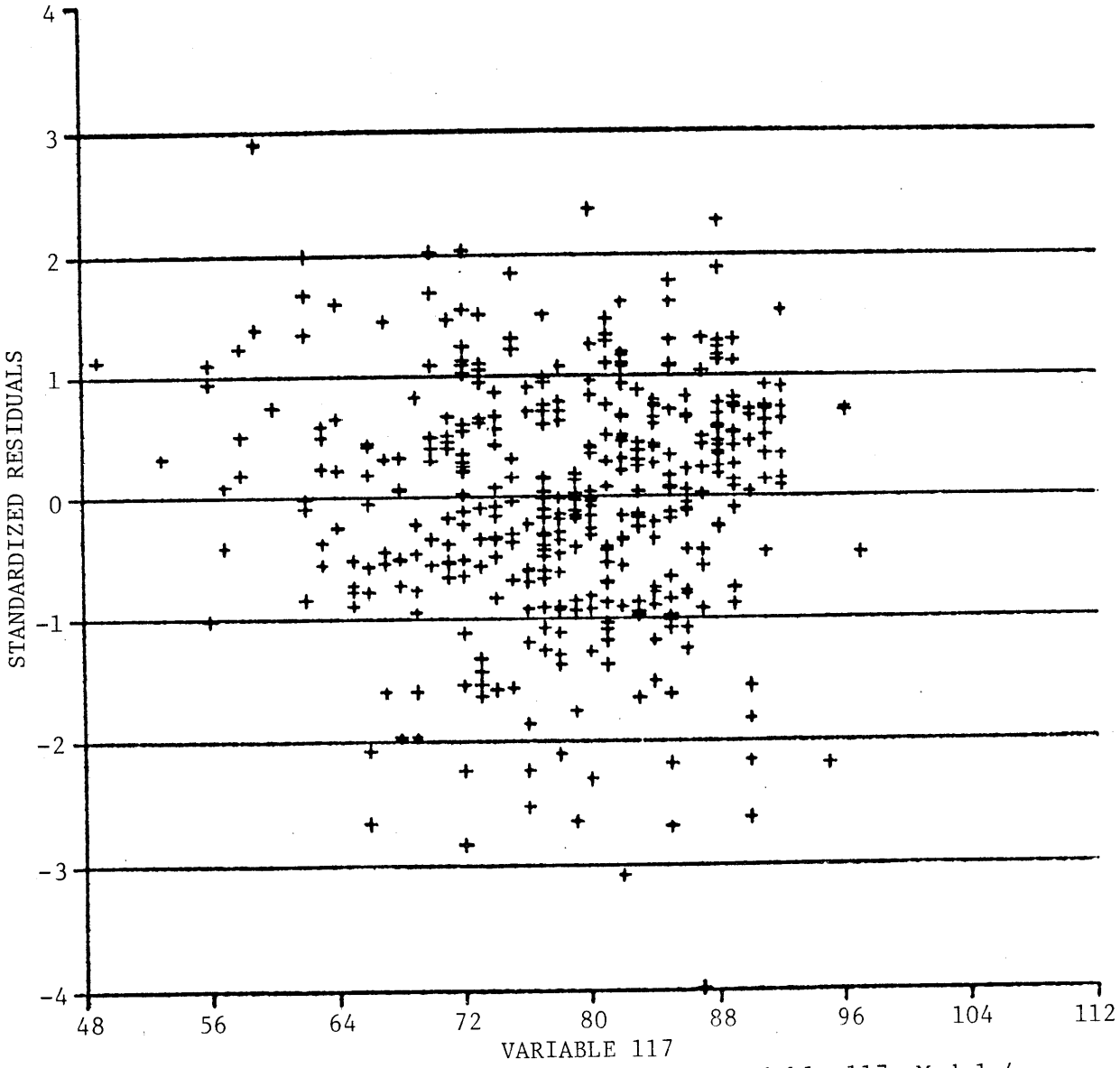

Figure 6.22 Standardized Residuals vs. Variable 117, Model 4

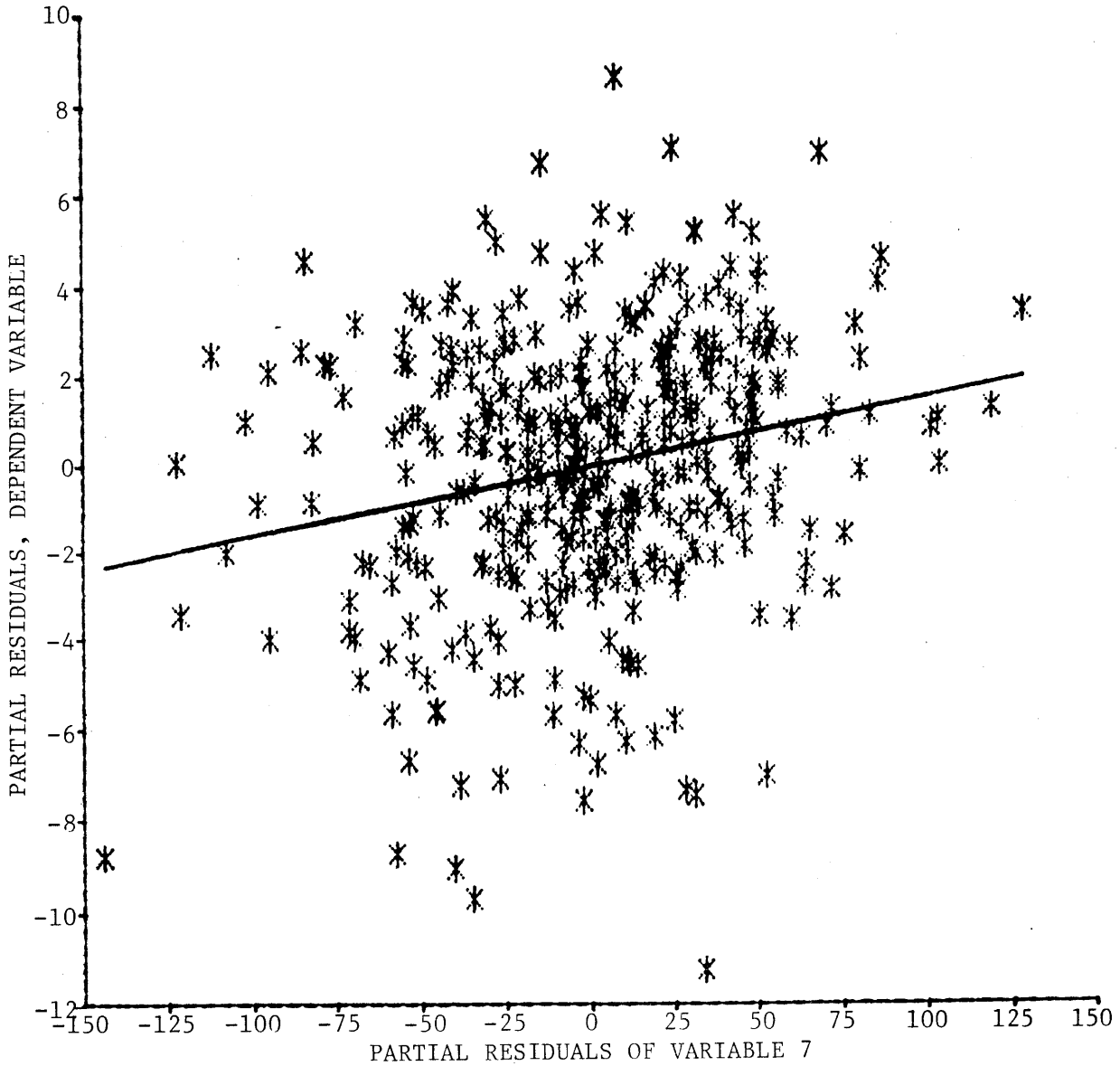

Figure 6.23 Partial Residuals of Variable 7, Variables 19, 74, 102, 104, and 117 in Model, Model 4

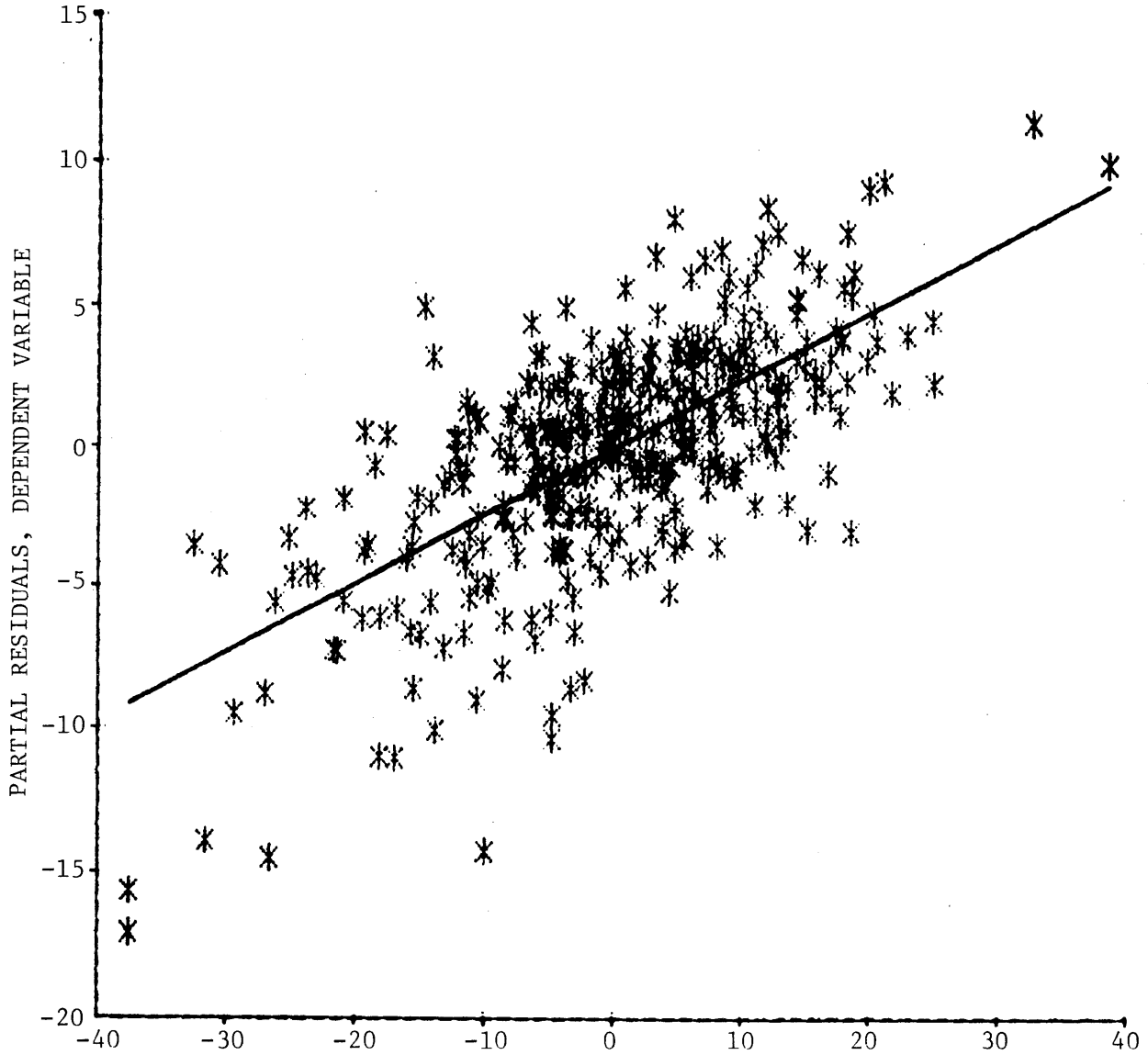

Figure 6.24 Partial Residuals of Variable 19, Variables 7, 74, 102, 104, and 117 in Model, Model 4

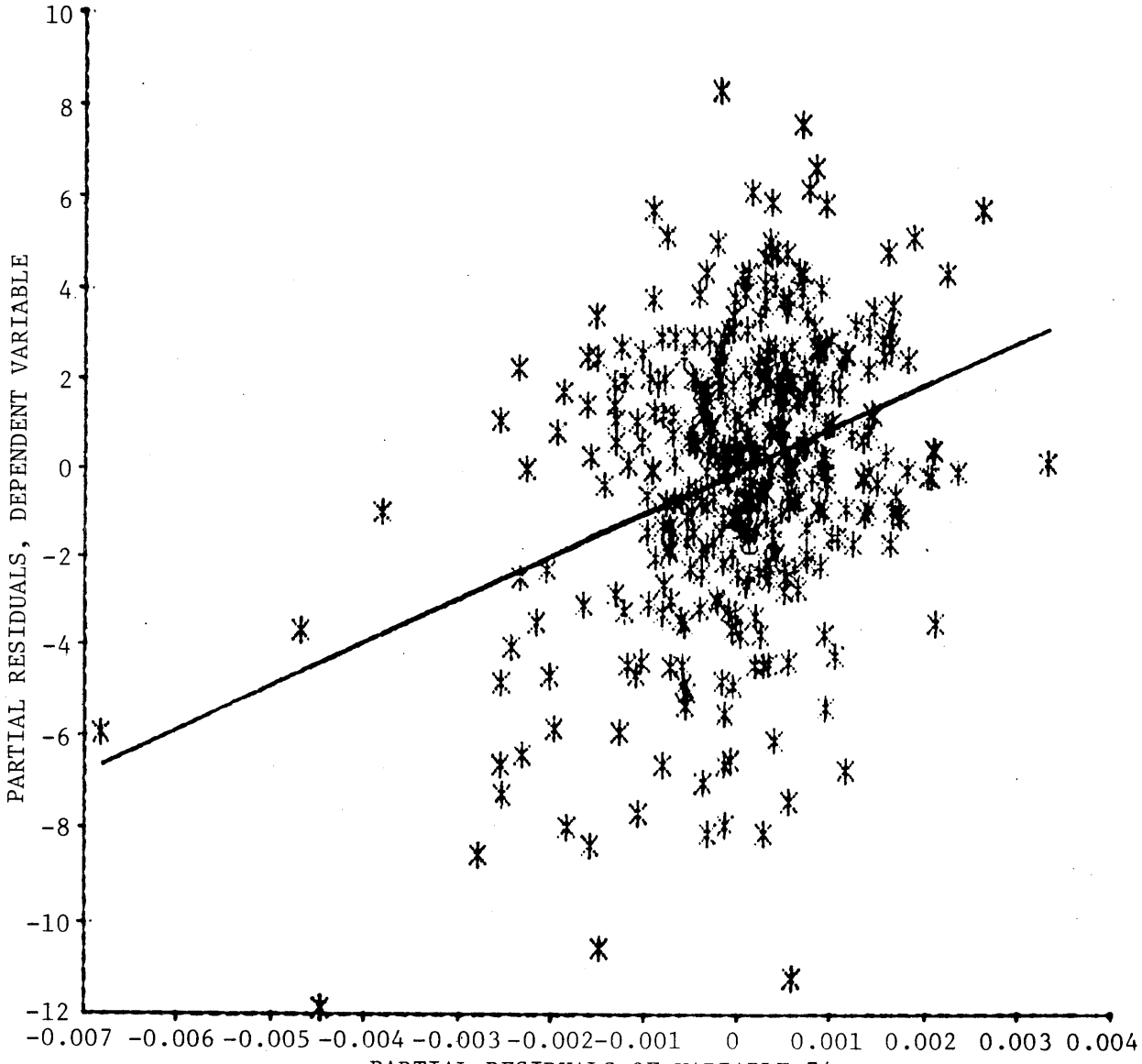

Figure 6.25 Partial Residuals of Variable 74, Variables 7, 19, 102, 104, and 117 in Model, Model 4

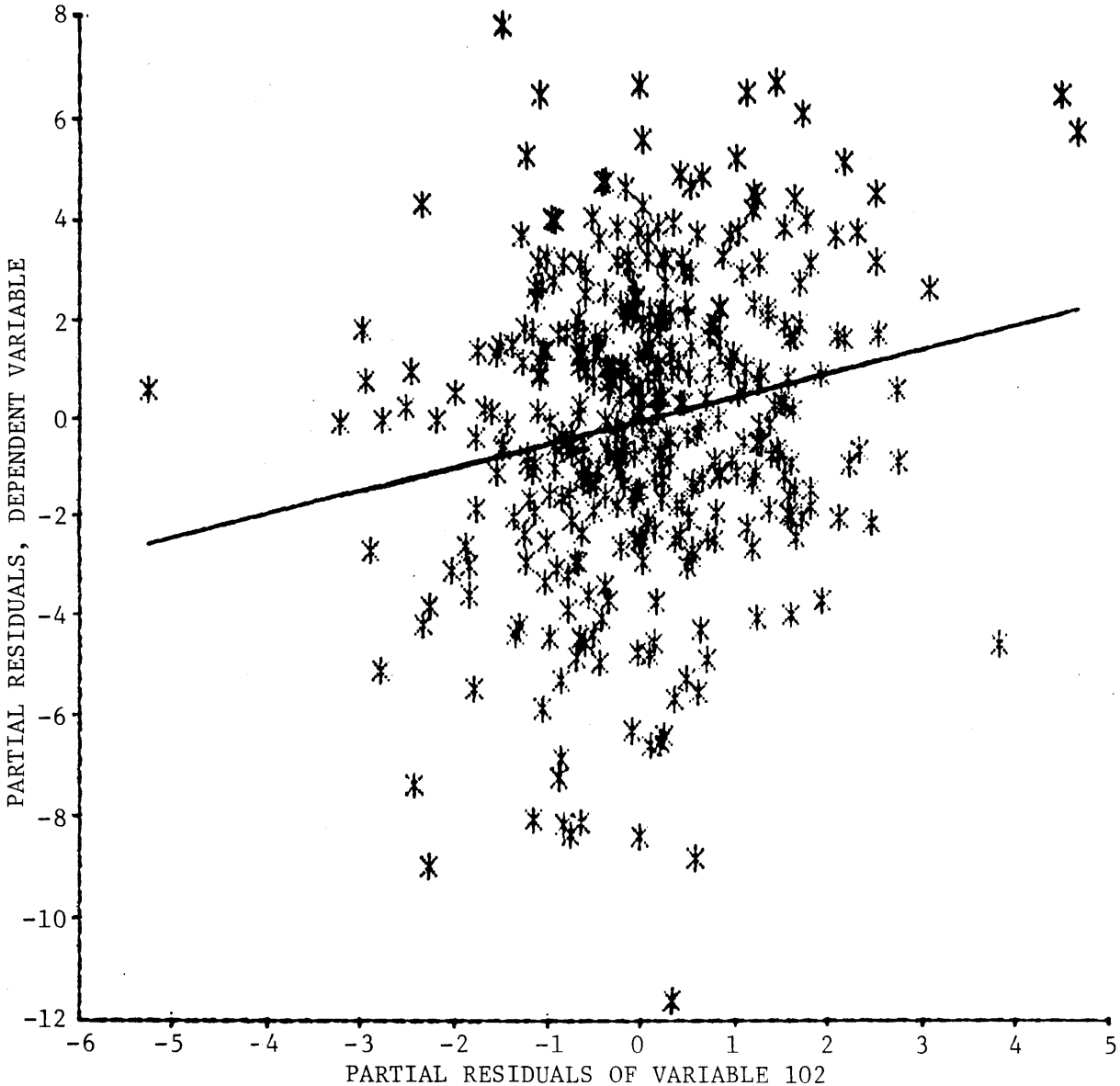

Figure 6.26 Partial Residuals of Variable 102, Variables 7, 19, 74, 104 and 117 in Model, Model 4

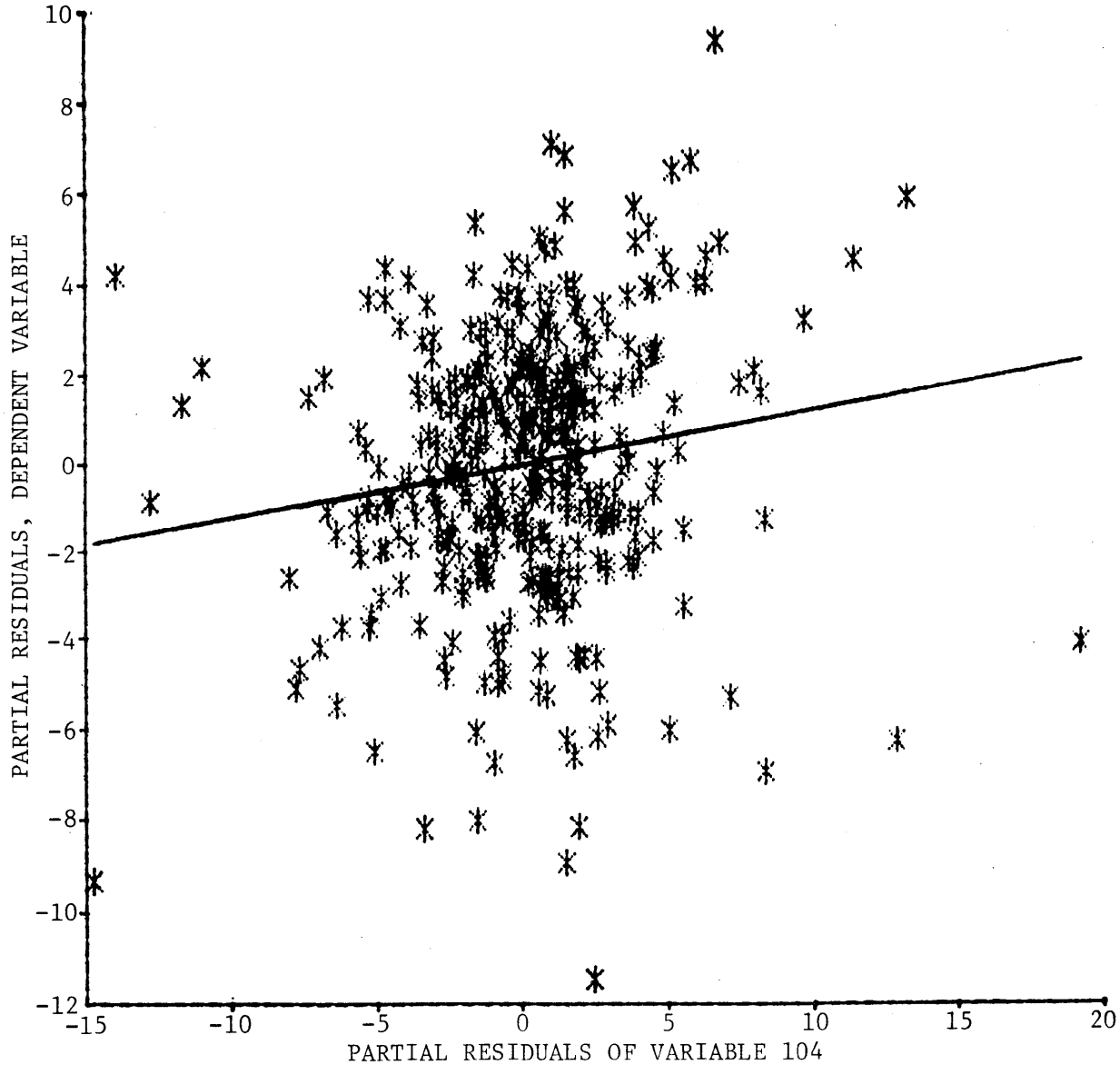

Figure 6.27 Partial Residuals of Variable 104, Variables 7, 19, 74, 102, and 117 in Model, Model 4

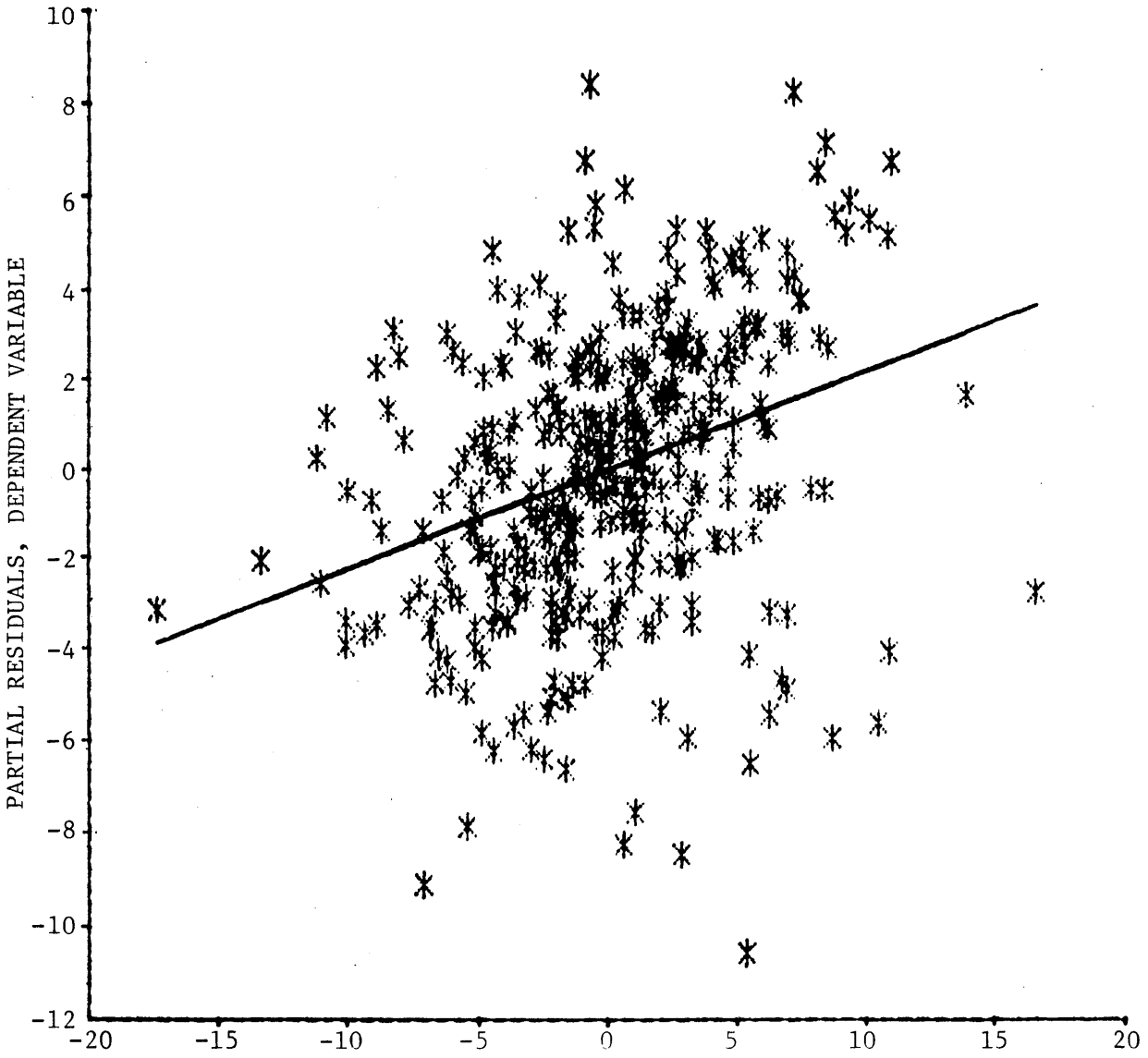

Figure 6.28 Partial Residuals of Variable 117, Variables 7, 19, 74, 102, and 104 in Model, Model 4

Variable 19 clearly has the strongest linear relation with the dependent variable, variables 7 and 117 show weaker relations, and variables 74, 102, and 104 show the weakest relations. Again, several outlying data points appear in the plots of variables 74, 102, and 104.

Variable 104 appears to have the weakest relation with the dependent variable. Thus we suspect that a model using only variables 7, 19, 74, 102, and 117 might perform nearly as well as model 4. A model using these variables was developed and labeled model 48. The quality statistics of model 48 are listed in Table 6.7 in Section 6.2.4. Based on these statistics, model 48 does perform nearly as well as model 4. The validation procedures applied to model 4 were also applied to model 48, but are not shown in this work. As for model 4, no serious problems were indicated. Thus choosing between models 4 and 48 is a subjective decision which would have to be made by the model user.

Model 48 could also have been generated by tightening the significance level for entering and deleting variables in the stepwise regression algorithm. As a test, after the work described in the above paragraph was performed, the significance level was tightened to 0.001. The variables in model 48 were then chosen. Model 4 was developed using a significance level of 0.01.

6.2.3.4 Partial residual plots, model 32

The partial residual plots for each of the independent variables in model 32 are shown in Figures 6.29 through 6.33. Variable

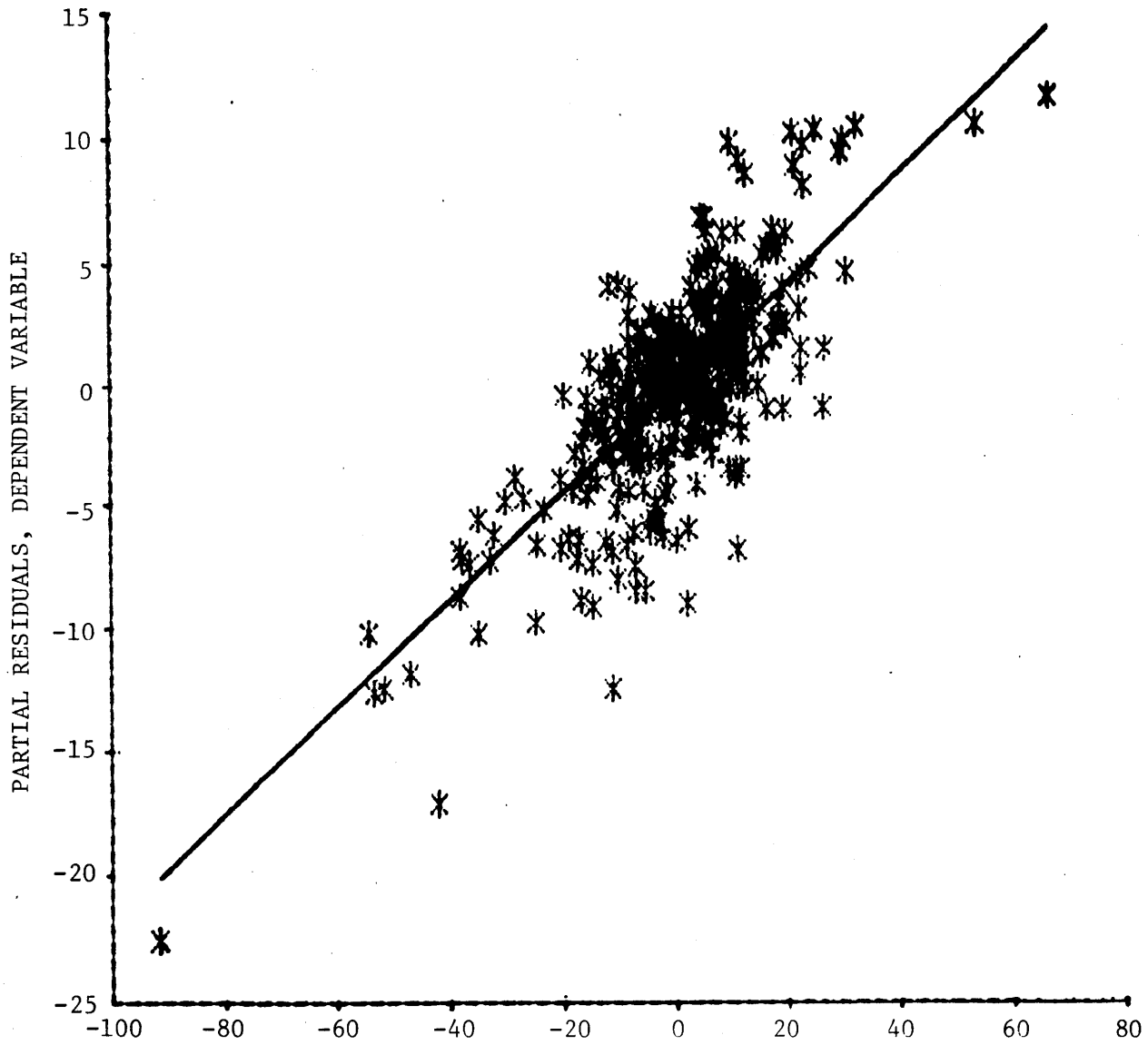

Figure 6.29 Partial Residuals of Variable 19, Variables 43, 61, 68, and 117 in Model, Model 32

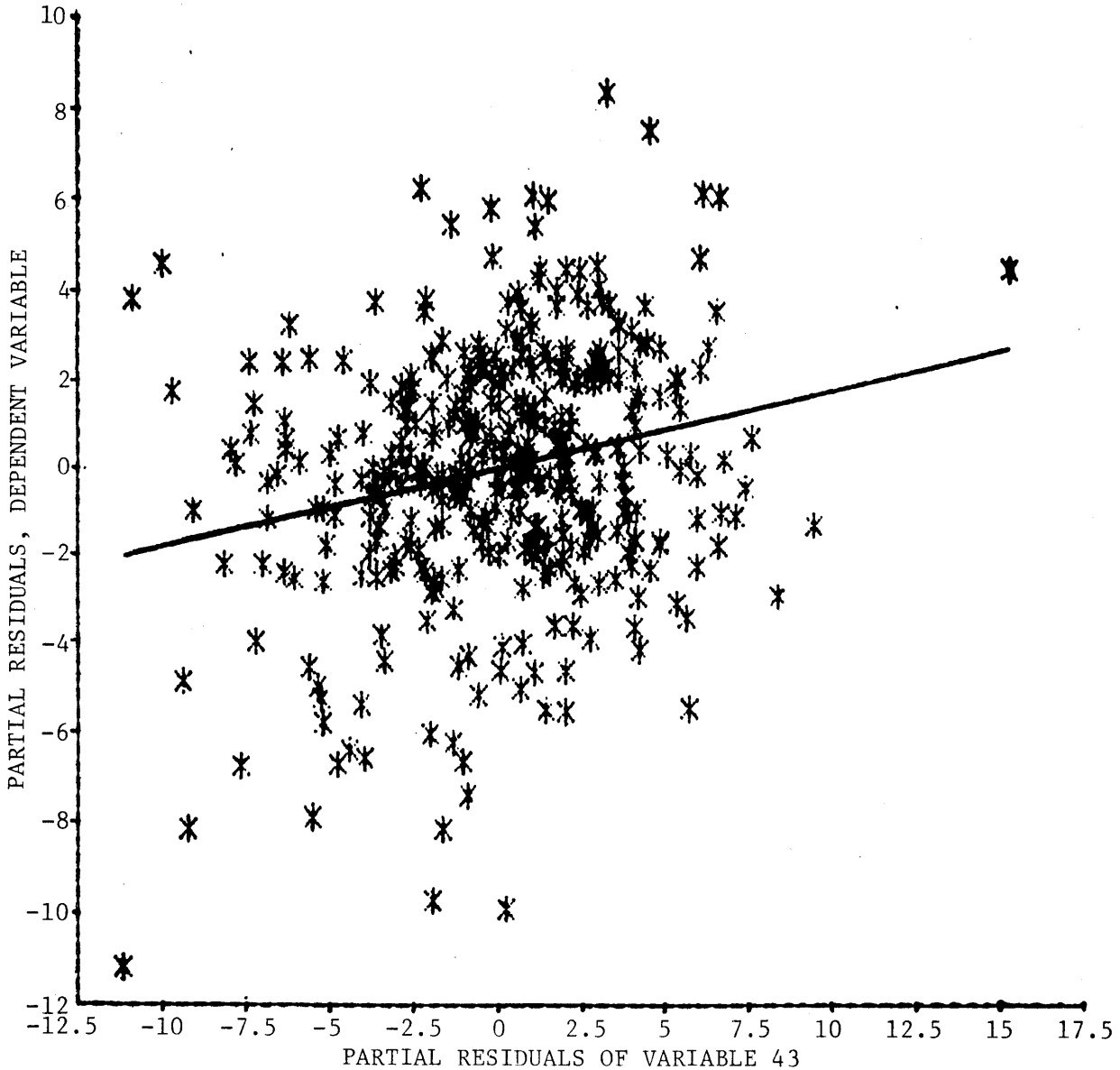

Figure 6.30 Partial Residuals of Variable 43, Variables 19, 61, 68, and 117 in Model, Model 32

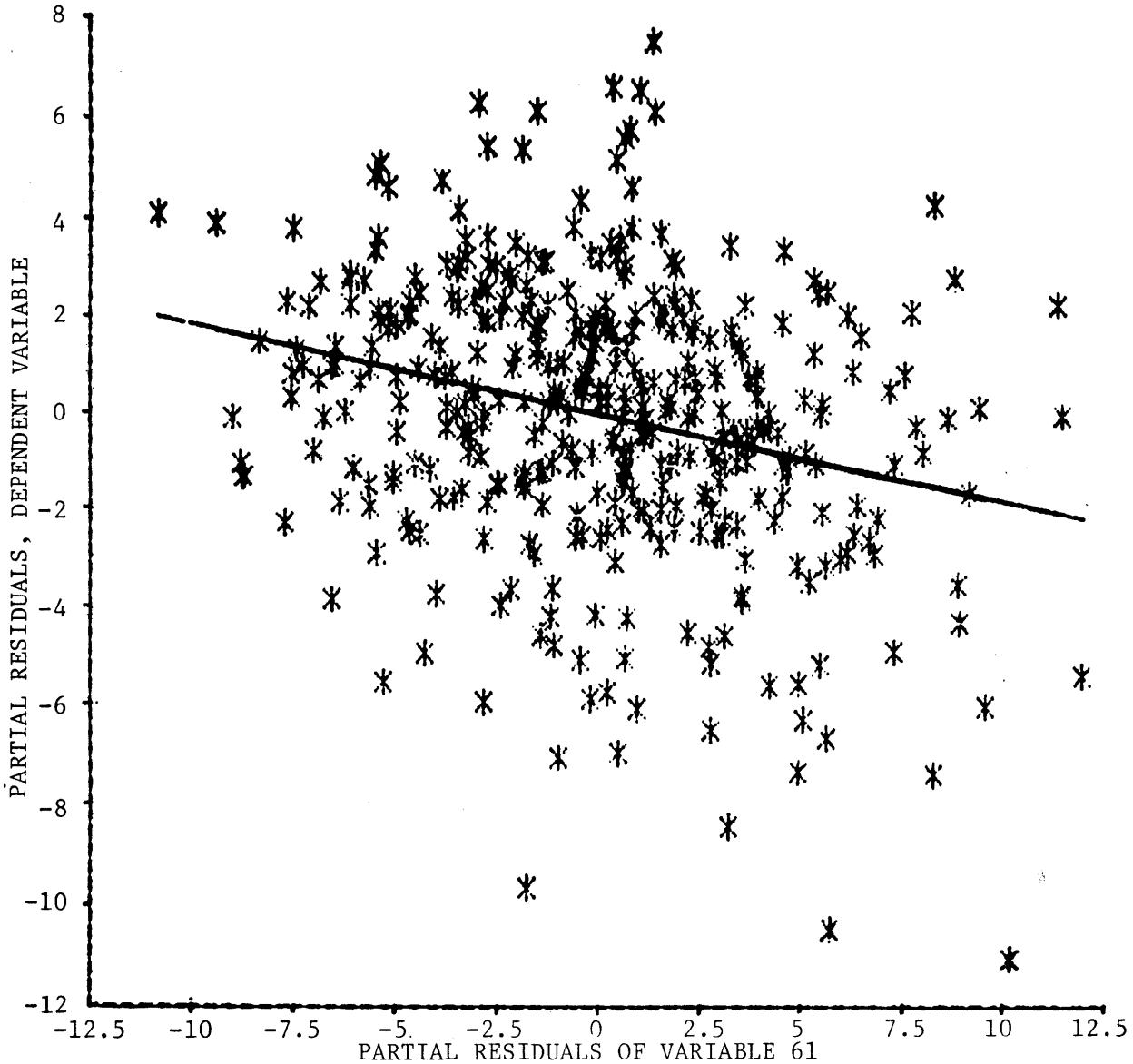

Figure 6.31 Partial Residuals of Variable 61, Variables 19, 43, 68, and 117 in Model, Model 32

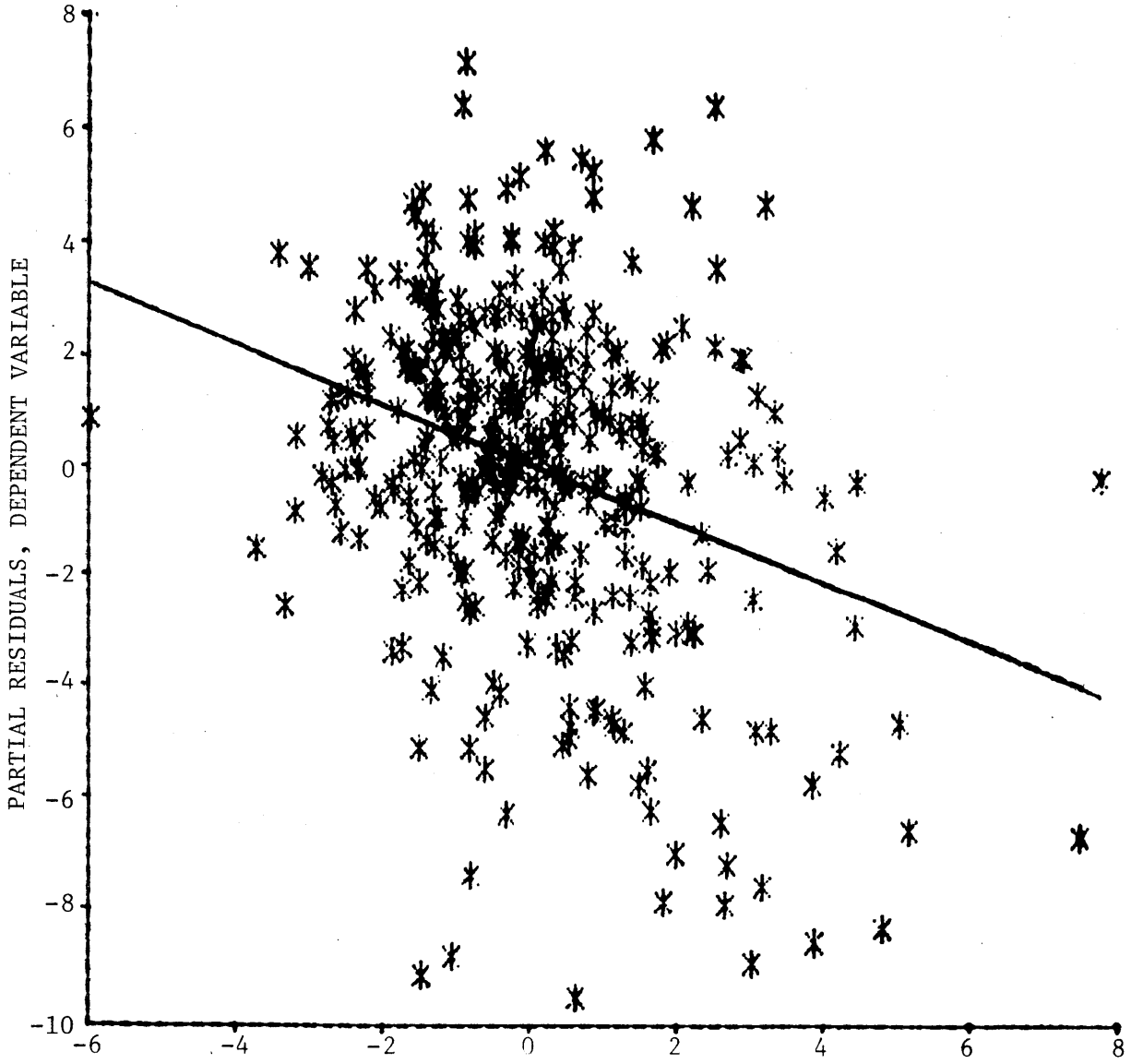

Figure 6.32 Partial Residuals of Variable 68, Variables 19, 43, 61, and 117 in Model, Model 32

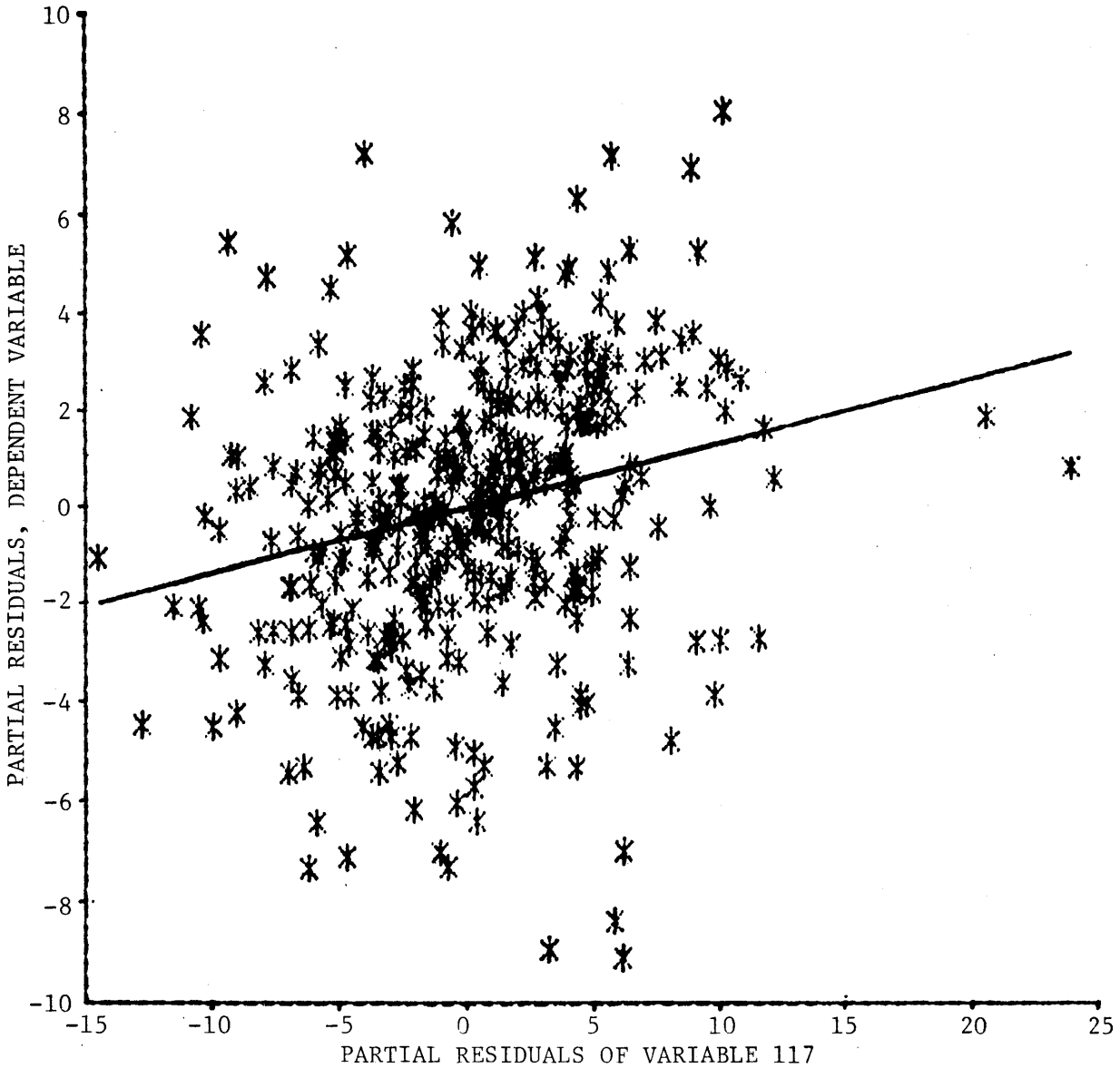

Figure 6.33 Partial Residuals of Variable 117, Variables 19, 43, 61, and 68 in Model, Model 32

19 again shows the strongest linear relation with the dependent variable. The variable coefficients appear to be generally better defined than for model 4. Variable 68 clearly has the most poorly defined coefficient in model 32, but appears relatively well defined when compared to variables 74, 102, or 104 of model 4. Thus we would expect model 32 to retain its prediction accuracy over a wider range of conditions than model 4. No further work with model 32 was performed. The partial residuals of model 36 were not plotted.

6.2.4 Analysis of Outliers

6.2.4.1 Residual outliers

The coefficients and model quality statistics of models 4 and 48 were reestimated after the data point associated with the standardized residual close to -4, day 358 noted in Figure 6.6, was deleted from the data set. The original and revised coefficients are listed in Table 6.4. The statistics IRMS and IRMA are not useful when comparing the models with and without day 358 because the coefficients estimated from the first 3 years of data are the same in both cases. Thus IRMS and IRMA change only due to the exclusion of the largest residual. However, RMS_k and RMA_k may be used to compare the models, after the day 358 residual, as calculated from the revised coefficients, has been included in the statistics of the revised models. RMS_k and RMA_k are listed in Table 6.5. The differences between the revised and original statistics are small. Since the original and revised day 358 residuals are approximately the same in both models,

the original and revised models are of approximately equal quality on the remaining points as well. Thus there is no apparent reason to exclude points with large residuals from the estimation data.

6.2.4.2 Data outliers

Outlying data points in variables 74, 102, and 104 were noticed in the plots of the residuals against the independent variables. The points which are circled in Figures 6.11, 6.12, and 6.13 were deleted from the data set and the coefficients of models 4 and 48 were reestimated from the reduced data set. Thus 6 points were deleted from the data used for model 4 and 4 points were deleted from the data used for model 48. The original and revised coefficients are listed in Table 6.6. The model quality statistics of the original and revised models are listed in Table 6.7. The statistics in Table 6.7 were calculated directly from the reduced data set. Had the statistics improved substantially, it would have been necessary to include in the calculation of the statistics the residuals from the deleted days, as was done in Section 6.4.2.1, to determine if the improvement was real. However, some of the coefficients, but none of the model quality statistics, changed substantially when the outlying data points were deleted. Thus, there appears to be little, if any, value in not allowing the models to reflect the full range of the available data. Model 48 appears slightly more resistant than model 4 to the removal of data points.

	Model 4			Model 48		
Variable	Original	Revised	% Change	Original	Revised	% Change
Constant	-302.44	-300.58	0.61	-334.25	-333.89	0.11
7	$(1.6153)10^{-2}$	$(1.6771)10^{-2}$	3.83	$(1.6617)10^{-2}$	$(1.7239)10^{-2}$	3.74
19	$(2.4439)10^{-1}$	$(2.4195)10^{-1}$	-1.00	$(2.6962)10^{-1}$	$(2.6839)10^{-1}$	-0.46
74	$(9.7713)10^{2}$	$(9.9314)10^{2}$	1.64	$(7.8700)10^{2}$	$(7.9404)10^{2}$	0.89
102	$(4.9231)10^{-1}$	$(4.9917)10^{-1}$	1.39	$(4.3351)10^{-1}$	$(4.3749)10^{-1}$	0.92
104	$(1.2748)10^{-1}$	$(1.3056)10^{-1}$	2.42	-	-	-
117	$(2.2542)10^{-1}$	$(2.3335)10^{-1}$	3.52	$(1.7388)10^{-1}$	$(1.7925)10^{-1}$	3.09
day 358 residual	11.75	11.86	0.94	11.45	11.53	0.70

Table 6.4 Effect on Model Coefficients of Deleting a Data Point Associated with an Outlying Residual (day 358)

	Model 4			Model 48		
Statistic	Original	Revised	% Change	Original	Revised	% Change
RMS_k	8.73	8.73	0.00	8.91	8.92	0.11
RMA_k	2.31	2.32	0.43	2.36	2.36	0

Table 6.5 Effect on Model Quality Statistics of Deleting a Data Point Associated with an Outlying Residual (day 358)

	Model 4			Model 48		
Variable	Original	Revised	% Change	Original	Revised	% Change
Constant	-302.44	-307.64	-1.72	-334.25	-341.85	-2.27
7	$(1.6153)10^{-2}$	$(1.7558)10^{-2}$	8.70	$(1.6617)10^{-2}$	$(1.7514)10^{-2}$	5.40
19	$(2.4439)10^{-1}$	$(2.4697)10^{-1}$	1.06	$(2.6962)10^{-1}$	$(2.7477)10^{-1}$	1.91
74	$(9.7713)10^{2}$	$(1.0387)10^{3}$	6.30	$(7.8700)10^{2}$	$(8.0032)10^{2}$	1.69
102	$(4.9231)10^{-1}$	$(4.9287)10^{-1}$	0.11	$(4.3351)10^{-1}$	$(4.5167)10^{-1}$	4.19
104	$(1.2748)10^{-1}$	$(1.4901)10^{-1}$	16.89	-	-	-
117	$(2.2542)10^{-1}$	$(2.1867)10^{-1}$	-2.99	$(1.7388)10^{-1}$	$(1.6185)10^{-1}$	-6.92

Table 6.6 Effect on Model Coefficients of Removing Outlying Data Points

	Model 4			Model 48		
Statistic	Original	Revised	% Change	Original	Revised	% Change
IRMS	8.63	8.49	-1.72	8.82	8.78	-0.47
IRMA	2.20	2.19	-0.45	2.23	2.22	-0.45
RMS_k	8.73	8.68	-0.63	8.91	8.91	0.00
RMA_k	2.31	2.32	0.43	2.36	2.36	0.00

Table 6.7 Effect on Model Quality Statistics of Removing Outlying Data Points

6.3 Comparison of Model Generation Methods

Data set 5, the principal components data set, was only data set to which all three model generation methods were applied. The relative effectiveness of the three methods is compared in this section by plotting IRMS, IRMA, RMS_k, and RMA_k against k for the models created from data set 5 and identifying the generation method of each model. These plots are shown in Figures 6.21 through 6.24. Interactive stepwise regression produced models with the smallest values of IRMS and IRMA. Automatic stepwise regression produced models with with the smallest values of RMS_k and RMA_k. Validation procedures were not applied to the models.

From this small sample, interactive stepwise regression appears to be the most effective model generation method of the three used in this work. However, automatic stepwise regression does nearly as well and requires substantially less effort. The GMDH did not perform as well as either stepwise regression method and requires more effort than automatic stepwise regression. The difference in effort between the GMDH and interactive stepwise regression for a data set with 28 variables is hard to judge. The GMDH probably uses more computer effort, but interactive stepwise regression requires more user effort.

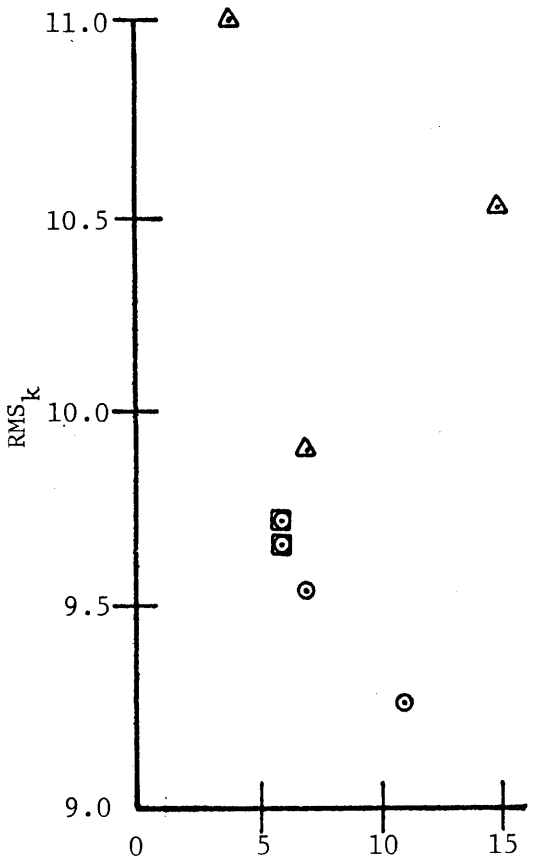

Figure 6.34

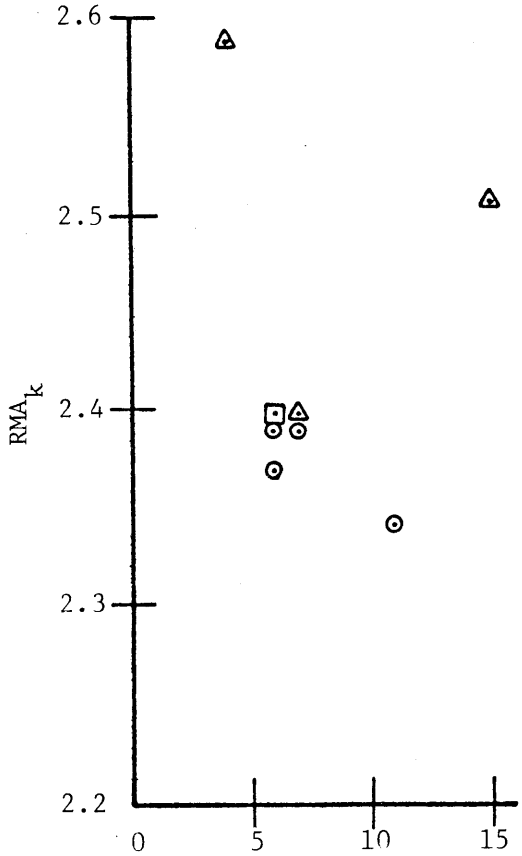

Figure 6.35

Comparison of Models Created from Data Set 5 with:

1) stepwise regression ⊙
2) GMDH △
3) interactive stepwise regression ⊡

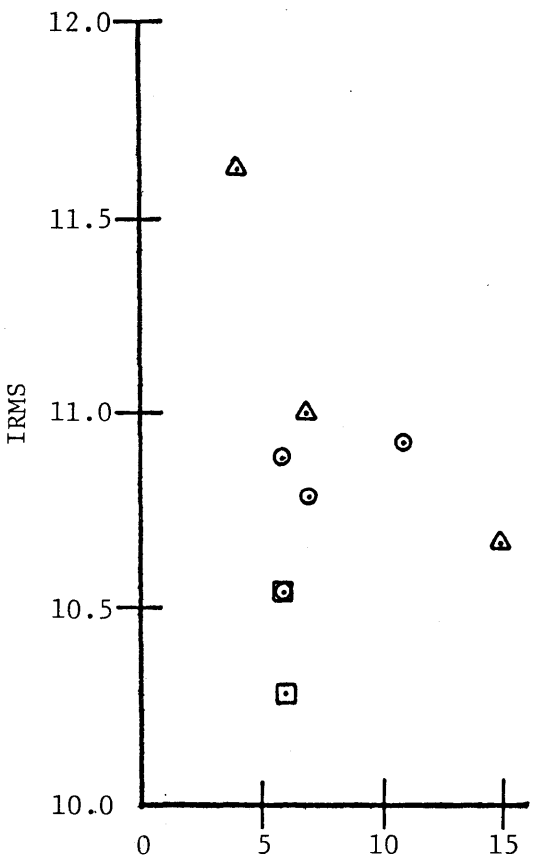

Figure 6.36

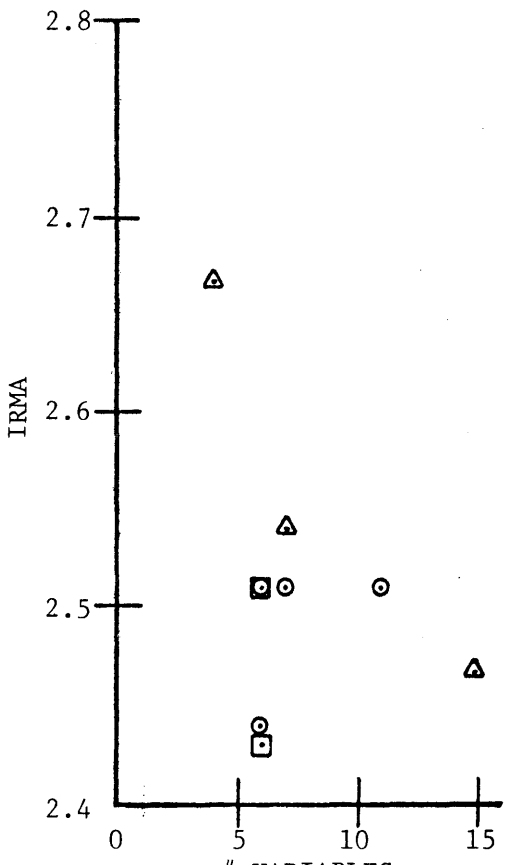

Figure 6.37

Comparison of Models Created from Data Set 5 with:

1) stepwise regression ⊙
2) GMDH △
3) interactive stepwise regression ⊡

Chapter 7

SUMMARY AND CONCLUSIONS

7.1 Summary

Models for predicting today's maximum temperature at Huntsville, Alabama were developed by applying different model generation methods to five variants of a data set provided by the National Weather Service. Temperature forecasting and the particular modeling method on which this work is based, Model Output Statistics, were discussed in Chapter 2. Some general aspects of empirical modeling and two of the model generation methods used in this work, automatic and interactive stepwise regression, were discussed in Chapter 3. The third model generation method, the Group Method of Data Handling, was discussed in Chapter 4. Details of the development of the 5 data sets and the use of the three model generation methods on those data sets were described in Chapter 5. The five data sets included the original data set, three data sets from which harmonic components had been removed, and one data set consisting primarily of principal components of groups of the original variables. The models were analysed in Chapter 6. First, the model quality statistics described in Chapter 3 were used to choose those models worth considering further. Linear models completely dominated this statistical analysis, and among the linear models those from which harmonic components had been removed were generally dominant. Some model validation procedures were then applied to three of the models, one from the original data

set and 2 from which some harmonic components had been removed. No serious problems were indicated.

Some potential changes to the model from the original data set suggested by the validation procedures were examined. These changes included removal of one of the independent variables and removal of some data points. The model with one independent variable removed was shown to perform nearly as well as the original model and removal of the selected data points had little effect on model quality.

7.2 Conclusions

Nonlinear transformations of the original variables chosen by the National Weather Service for the Model Output Statistics temperature forecast equations do not appear to be useful predictors. However, model quality can be improved by modeling mean trends separately from more transient effects. Also, the number of variables may be reduced from the 10, plus a constant term, currently used by the NWS without sacrificing much prediction accuracy or fit to the estimation data.

Carter (1979) has noted occasional irregular behavior in the temperature predictions from MOS models which appears to be caused by unstable relations between the independent variables. Reducing the number of variables in the equations should help reduce this source of instability.

The GMDH was not an effective modeling method in this situation. It is not clear whether this lack of effectiveness is due to some properties of the GMDH algorithm or simply to the insignificance of nonlinear relations between the original variables

in MOS temperature prediction equations. Even if the GMDH had produced the best models, the quality of those models would have had to have been substantially greater than the quality of models produced with simpler methods to justify the large computational burden imposed by the GMDH.

When the number of independent variables is sufficiently small to permit its use, interactive stepwise regression appears to be the most effective model generation method. However, the necessary human direction of the process makes interactive stepwise regression unsuitable for operations such as those of the NWS, in which thousands of equations must be developed. Thus, among the procedures examined in this work, the procedure closest to current NWS practice, linear stepwise regression is the best way to develop models for predicting today's maximum temperature at Huntsville, Alabama. Unfortunately, the coefficient discrepancies mentioned at the beginning of Chapter 6 prevented the direct comparison of models produced by the forward moving stepwise regression algorithm used by the NWS with the models produced by the stepwise regression algorithm used in this work.

This work was based on the prediction of one variable at one site. We suspect that similar variables at the same site have similar properties, but nothing can be said with certainty about the implications of these results to predicting other variables at the same site or predicting any variables at other sites.

References

1. Allen, R.A. and E.M. Vernon, "Objective Weather Forecasting," Compendium of Meteorology, American Meteorological Society, Boston, Mass., 1951, 796-801.

2. Annett, J.R., H.R. Glahn, and D.A. Lowry, "The Use of Model Output Statistics (MOS) to Estimate Daily Maximum Temperatures," NOAA Tech. Memo., NWS TDL-45, Mar. 1972.

3. Anscome, F.J.,"Graphs in Statistical Analysis," The American Statistician, 27(1), Feb. 1973, 17-21.

4. Behnken, D.W. and N.R. Draper, "Residuals and Their Variance Patterns," Technometrics, 14(1), Feb. 1972, 101-111.

5. Block, H.D., "The Perceptron: A Model for Brain Functioning. I," Reviews of Modern Physics, 34(1), Jan. 1962, 123-135.

6. Block, H.D., B.W. Knight, Jr., and F. Rosenblatt, "Analysis of a Four-Layer Series-Coupled Perceptron. II," Reviews of Modern Physics, 34(1), Jan. 1962, 135-142.

7. Bocchieri, J.R. and H.R. Glahn, "Use of Model Output Statistics for Predicting Ceiling Height," MWR, 100(2), Dec. 1972, 869-879.

8. Brown, J.A., "The Seven Layer Primitive Equation Model," NWS Tech. Proc. Bull. 218, 219, 30 Sept. 1977, 1977a.

9. Brown, J.A., "High Resolution LFM (LFM-II)," NWS Tech. Proc. Bull., 206, 28 July 1977b.

10. Carter, G.M., "Weather Forecasts, User's Economic Expenses, and Decision Strategies," San Jose State College, Dept. of Meteorology, San Jose, California, June 1972.

11. Carter, G.M., Personal communication, 1979.

12. Carter, G.M., J.P. Dallavalle, A.L. Forst, and W.H. Klein, "Improved Automated Surface Temperature Guidance," MWR, 107(10), Oct. 1979, pp. 1263-1274.

13. Chatterjee, S. and B. Price, Regression Analysis by Example, Wiley, N.Y., 1977.

14. Cox, D.R. and E.J. Snell, "A General Definition of Residuals," J. Royal Stat. Soc. (B), 30(2), 1968, 248-265.

15. Cox, D.R. and E.J. Snell, "The Choice of Variables in Observational Studies," Applied Statistics, 23(1), 1974, 51-59.

16. Craddock, J.M., "The Representation of the Annual Temperature over Central and Northern Europe by a Two-term Harmonic Form," J.R. Met. Soc., 82(353), July 1956, 275-288.

17. Daniel, C. and F.S. Wood, Fitting Equations to Data, Wiley, N.Y., 1971.

18. Dickey, W.W., "Forecasting Maximum and Minimum Temperatures," U.S. Weather Bureau, Forecasting Guide #4, 1960.

19. Draper, N.R. and H. Smith, Applied Regression Analysis, Wiley, N.Y., 1966.

20. Duffy, J.J. and M.A. Franklin, "A Learning Identification Algorithm and its Application to an Environmental System," IEEE Trans Syst., Man, and Cyb., SMC-5(2), March 1975, 226-240.

21. Durbin, J., and G.S. Watson, "Testing for Serial Correlation in Least Squares Regression. I," Biometrika, 37, 1950, 409-428.

22. Durbin, J. and G.S. Watson, "Testing for Serial Correlation in Least Squares Regression. II," Biometrika, 38, 1951, 159-178.

23. Efroymson, M.A., "Multiple Regression Analysis," Mathematical Methods for Digital Computers, A. Ralston and H.F. Wilf, (eds.) Wiley, N.Y., 1960.

24. Farrar, D.E. and R.R. Glauber, "Multicollinearity in Regression Analysis: The Problem Revisited," Rev. Econ. and Stat., 49(1), Feb. 1967, 92-107.

25. Gerrity, Jr., J.F., "The LFM Model -1976: A Documentation," NOAA Tech. Memo., NWS NMC-60, Dec. 1977.

26. Glahn, H.R., "An Experiment in Forecasting Rainfall Probabilities by Objective Methods," MWR, 90(2), Feb. 1962, 59-67.

27. Glahn, H.R., "The Use of Decision Theory in Meteorology: With an Application to Aviation Weather," MWR, 92(9), Sept. 1964, 383-388.

28. Glahn, H.R., "Objective Weather Forecasting by Statistical Methods," The Statistician, 15(2), 1965, 111-142.

29. Glahn, H.R. and D.A. Lowry, "The Use of Model Output Statistics (MOS) in Objective Weather Forecasting," JAM, 11(8), Dec. 1972, 1203-1211.

30. Gringorten, I.I., "Probability Estimates of the Weather in Relation to Operational Decisions," J. Meteorology, 16(6), Dec. 1959, 663-671.

31. Goldberger, A.S.,"Econometric Theory," Wiley, N.Y., 1964.

32. Goldberger, A.S.,"Best Linear Unbiased Prediction in the Generalized Linear Regression Model," J. Am. Stat. Assoc., 57(298), June 1962, 369-375.

33. Greenberg, E.,"Minimum Variance Properties of Principal Component Regression," J. Am. Stat. Assoc., 70(349), March 1975, 194-197.

34. Hammons, G.A., J.P. Dallavalle, and W.H. Klein, "Automated Temperature Guidance Based on Three Month Seasons," MWR, 104(12), Dec. 1976, 1557-1564.

35. Helbush, R.E., "Linear Programming Applied to Operational Decision Making in Weather Risk Situations," MWR, 96(12), Dec. 1968, 876-882.

36. Hocking, R.R., "The Analysis and Selection of Variables in Linear Regression," Biometrics, 32(1), March 1976, 1-49.

37. Hoerl, A.E., and R.W. Kennard, "Ridge Regression: Biased Estimation for Non-Orthogonal Problems," Technometrics, 12(1), 1970, 55-67.

38. Howe, C.W. and H.C. Cochrane, "A Decision Model for Adjusting to Natural Hazard Events with Application to Urban Snow Storms," Rev. Econ. and Stat., 58(1), Feb. 1976, 50-58.

39. Ikeda, S., M. Ochiai, and Y. Sawaragi, "Sequential GMDH Algorithm and its Application to River Flow Prediction," IEEE Trans. on Syst. Man, and Cyb., SMC-6(7), July 1976, 473-479.

40. Inooka, H. and A. Inoue, "Application of the GMDH Algorithm to a Manual Control System," IEEE Trans. on Syst., Man, and Cyb., SMC-8(11), Nov. 1978, 819-821.

41. International Mathematical and Statistical Libraries, Inc. (IMSL) Library 2, edition 6, 1977.

42. Ivakhnenko, A.G. and V.D. Dmitrov, "Stochastic Algorithms and the Group Method of Data Handling in Prediction of Random Events," Sov. Auto. Contr. 14(3), 1969, 20-33.

43. Ivakhnenko, A.G. "Heuristic Self-Organization in Problems of Engineering Cybernetics," Automatica, 6, 1970, 207-219.

44. Ivakhnenko, A.G., Y.V. Koppa, and W.S. Min, "Polynomial and Logical Theory of Dynamic Systems (Part II)., "Sov. Auto. Contr., 3(4), 1970a, 11-30.

45. Ivakhnenko, A.G.,"Polynomial Theory of Complex Systems," IEEE Trans. on Syst., Man, and Cyb., SMC-1(4), Oct. 1971, 364-378.

46. Ivakhnenko, A.G.,"The Group Method of Data Handling in Prediction Problems," Sov. Auto. Contr., 9(6), 1976, 21-30.

47. Jeffers, J.N.R., "Two Case Studies in the Application of Principal Components Analysis," Applied Statistics, 16(3), 1967, 225-236.

48. Johnston, J., Econometric Methods, 2nd ed., McGraw-Hill, New York, 1972.

49. Kendall, M.G., A Course in Multivariate Analysis, Hafner Publishing Co., N.Y., 1957.

50. Kernan, G.L., "The Cost-Loss Decision Model and Air Pollution Forecasting," JAM, 14(1), Feb. 1975, 8-16.

51. Klein, W.H., "Objective Forecasts of Surface Temperature from One to Three Days in Advance," JAM, 5(2), Apr. 1966, 137-147.

52. Klein, W.A., and H.R. Glahn, "Forecasting Local Weather by Means of Model Output Statistics," Bull. Am. Met. Soc., 55(10), Oct. 1974, 1217-1227.

53. Klein, W.H. and G.R. Hammons, "Maximum/Minimum Temperature Forecasts Based on Model Output Statistics," MWR, 103(9), Sept. 1975, 796-806.

54. Klein, W.H. and F. Lewis, "Computer Forecasts of Maximum and Minimum Temperatures," JAM, 9(3), June 1970, 350-359.

55. Klein, W.H., F. Lewis, and G.P. Casely, "Automated Nationwide Forecasts of Maximum and Minimum Temperature," JAM, 6(2), Apr. 1967, 216-228.

56. Klein, W.H., F. Lewis, and G.A. Hammons, "Recent Developments in Automated Maximum and Minimum Temperature Forecasting," JAM, 10(5), Oct. 1971, 916-920.

57. Kolb, L.L. and R.R. Rapp, "The Utility of Weather Forecasts to the Raisin Industry," JAM, 1(1), March 1962, 8-12.

58. Kutzbach, J.E., "Empirical Eigenvectors of Sea Level Pressure, Surface Temperature and Precipitation Complexes over North America," JAM, 6(5), Oct. 1967, 791-802.

59. Laboratory of Architecture and Planning, MIT, "Handbook of Programs and Data," revision 11, 1978.

60. Larsen, W.A. and S.J. McCleary, "The Use of Partial Residual Plots in Regression Analysis," Technometrics, 14(3), Aug. 1972, 781-790.

61. Lave, L.B.,"The Value of Better Weather Information to the Raisin Industry," Econometrica, 31(1-2), Jan.-Apr. 1963, 151-164.

62. Massy, W.F.,"Principal Components Regression in Exploratory Statistical Research," J. Am. Stat. Assoc., 60(309), Mar. 1965, 234-256.

63. Meyer, S.L., Data Analysis for Scientists and Engineers, Wiley, N.Y., 1975.

64. Miller, R.G.,"Statistical Prediction by Discriminant Analysis," Am. Met. Soc., Met. Monographs, 4(25), Oct. 1962.

65. Mosteller, F. and J.W. Tukey, Data Analysis and Regression: A Second Course in Statistics, Addison-Wesley, Reading, Mass., 1977.

66. Murphy, A.H., "The Value of Climatological, Categorical, and Probabilistic Forecasts in the Cost-Loss Ratio Situation," MWR, 105(7), July 1977, 803-816.

67. Nelson, R.R. and S.G. Winter, "A Case Study in the Economics of Information and Coordination: The Weather Forecasting System," Quarterly Journal of Economics, 78(3), Aug. 1964, 420-441.

68. Nilsson, N.J., Learning Machines, McGraw Hill, N.Y., 1965.

69. Petterssen, S.,"Observations, Analysis, and Forecasting," Am. Met. Soc., Met. Monographs, 3(15), July 1957.

70. Rieck, R.E., A.F. Sadowski, and J.J. Harrell, "National Weather Service Forecasting Handbook No. 1: Facsimile Products," NOAA NWS, July 1976.

71. Rieck, R.E.,"The Limited Area Fine Mesh Model (LFM)," NWS, Tech. Proc. Bull., 232, June 1978.

72. Rosenblatt, F.,Principles of Neurodynamics, Spartan Books, Wash. D.C. 1962.

73. Sanders, F.,"Skill in Forecasting Daily Temperature and Precipitation: Some Experimental Results," Bull. Am. Met. Soc., 54(11), Nov. 1973, 1171-1179.

74. Schuman, F.G. and J.B. Hovermale, "An Operational Six-Layer Primitive Equation Model," JAM, 7(4), Aug. 1968, 525-547.

75. Snee, R.D., "Validation of Regression Models: Methods and Examples," Technometrics, 19(4), Nov. 1977, 415-428.

76. Stackpole, J.D., "Operational Prediction Models at the National Meteorological Center," Development Division, National Meteorological Center, Dec. 1975, unpublished paper prepared for Advanced Prediction Techniques Course at National Weather Service headquarters, Silver Spring, Md.

77. Taylor, C.J., "A Stochastic Model of Temperature Variations at Weather Stations in Britain," Applied Statistics, 21(3), 1972, 248-260.

78. Theil, H., Principles of Econometrics, Wiley, N.Y., 1971.

79. Thompson, J.C.,"Economic Gains from Scientific Advances and Operational Improvements in Meteorological Prediction," JAM, 1(1), Mar. 1962, 13-17.

80. Thompson, J.C. and G.W. Brier, "The Economic Utility of Weather Forecasts," MWR, 83(11), Nov. 1955, 249-254.

81. Tukey, J.W., Exploratory Data Analysis, Addison-Wesley, 1977.

82. Zurndorfer, E.A. et al., "Trends in Comparative Verification Scores for Guidance and Local Aviation/Public Weather Forecasts," MWR, 107(7), July 1979, 799-811.

Abbreviations:

MWR = Monthly Weather Review
JAM = Journal of Applied Meteorology

Appendix A: Principal Component Transformations

Variable transformations defining the principal component with the largest variance for each of the groups listed in Table 5.4 These numbers are the elements of the matrix A in equation 3.6. For each group, A is a k x 1 matrix where k is the number of variables in the group. A has only 1 column because only one principal component was calculated for each group. The variable numbers refer to Table 5.1

Group	Variable	Weight	Group	Variable	Weight	Group	Variable	Weight
A	5	-0.834	C (cont'd)	35	-0.943	G	65	-0.819
	6	-0.749		36	-0.916		66	-0.919
	7	-0.948		37	-0.958		67	-0.928
	8	-0.942		38	-0.975		68	-0.803
	9	-0.699		39	-0.948		69	-0.778
	10	-0.786		40	-0.891		70	-0.806
B	11	-0.942	D	41	-0.683	H	71	-0.832
	12	-0.980		42	-0.782		72	-0.853
	13	-0.990		43	-0.833		73	-0.864
	14	-0.978		44	-0.743		74	-0.842
	15	-0.955		53	-0.869	I	75	-0.863
	16	-0.922		54	-0.942		76	-0.842
	17	-0.952		55	-0.940		77	-0.883
	18	-0.957		56	-0.878		78	-0.874
	19	-0.932		61	-0.879	J	79	-0.678
	20	-0.885		62	-0.820		80	-0.742
	21	-0.923	E	45	-0.785		81	-0.797
	22	-0.963		46	-0.852		82	-0.829
	23	-0.970		47	-0.902		83	-0.832
	24	-0.957		48	-0.827		84	-0.830
	25	-0.937		57	-0.838		85	-0.910
C	26	-0.921		58	-0.953		86	-0.887
	27	-0.969		59	-0.962		87	-0.831
	28	-0.870		60	-0.893		88	-0.923
	29	-0.909		63	-0.868		89	-0.957
	30	-0.968		64	-0.901		90	-0.947
	31	-0.982	F	49	-0.646		91	-0.894
	32	-0.976		50	-0.868		92	-0.882
	33	-0.948		51	-0.928		93	-0.907
	34	-0.879		52	-0.869		94	-0.882

Group	Variable	Weight
J (cont'd)	95	-0.813
	96	-0.946
	97	-0.907
	98	-0.948
	99	-0.854
K	100	-0.521
	101	-0.675
	102	-0.743
	103	-0.577
L	104	0.879
	105	0.879
M	106	0.787
	107	0.787

Appendix B

User's Manual, Group Method of Data Handling

B.1.1 Introduction

The program GMDH performs the Group Method of Data Handling described in Section 4.3 and forward moving, with a backwards glance, stepwise regression. Programs PLOT1 and PLWSV are used with GMDH to provide graphic output. GMDH, PLOT1, and PLWSV are written in Fortran IV as interactive programs and were implemented through the Multics Operating system on the Honeywell 6180 computer. Honeywell Multics documentation should be consulted for information about using this system. The program modifications which are known to be necessary for conversion to IBM Fortran are listed in Section B.1.2. All dimensioned variables and most frequently used scalars are declared in common blocks contained in file GMDH_COM. incl. fortran. This file is referenced in each of the subroutines in GMDH and PLWSV through an "% include" statement and must be present in the working directory when the programs are compiled. The parameters in GMDH_COM. incl. fortran are discussed in Section B.3.

IMSL (1977) subroutine RLSEP is used to perform all regression calculations in GMDH and a modification of IMSL (1977) subroutine RLRES is used to calculate residuals. IMSL (1977) documentation should be consulted for information about the parameters and performance of these subroutines.

The questions addressed to the user during program execution are described in Section B.2, the input and output formats and capabilities are described in Section B.3, a sample terminal session and associated output are presented in Section B.4, and the programs are listed in Section B.5.

B.1.2 Conversion to IBM FORTRAN IV

These programs have not yet been used on other systems. The following changes are believed to be required for conversion to IBM FORTRAN IV.

Character variables may need to be changed to real variables, though some compilers will accept character variables. The affected variables are SR, TITLE, CHAR, XAXIS, And YAXIS.

List directed input and output statements need to be changed from 'READ,' and 'PRINT,' to 'READ *,' and 'PRINT *,'.

B.2 INTERACTIVE INSTRUCTIONS

B.2.1 Introduction

GMDH prompts the user for information and choices of options. Some general aspects of interactive data entry are described in this section and the questions posed by GMDH are described in Section B.2.2.

All interactive numeric data entry is in list directed format. Thus entries are converted to the data type implied by the variable name. Different numbers on a line may be separated either by spaces or by comas. The format specifications given in the user prompts serve only to remind the user of the numbers and types of variables which should be entered. The entries may be on one or several lines. However, entries may not be changed after the line return key is pressed. Extra entries on a line are ignored, but extra entries on subsequent lines will be read by the next terminal READ statement. Thus it is good practice to both check the accuracy of all entries prior to pressing the line return key and to avoid extra entries.

Interactive character input is formatted and thus will accept blanks as characters. Each character variable must be entered on a single line and no other variables should be entered on the same line. This restriction is signaled by the slash (/) character in the format prompts.

One other general convention should be noted. When a particular entry is specified to perform an action, any other entry will cause that action to not be performed.

B.2.2 User Prompts

Question set 1; in subroutine UNIT

```
ENTER:
IREAD    =  INPUT FILE NUMBER
IWRIT    =  OUTPUT FILE NUMBER
FF       =  0 FOR A CHARACTER FILE
            1 FOR AN UNFORMATTED FILE
TITLE(1)=  DATA FILE NAME

3I, /, A50  FORMAT
```

IREAD and IWRIT are the unit numbers for the input and output data files. They should be 1 or 2 digit integers and must be specified in accordance with the operating system procedures. Numbers 5, 6, 7, 25, 26, 41, and 42 should generally not be used. FF is for the input file. When FF = 0 a list directed READ is used and when FF = 1 an unformatted READ is used. Unformatted files must have been written in a code compatible with that used by the system on which the program is being run. The organization of input data files is described in Section B.3.1. The first 50 characters on the line after that on which FF is entered will be read as TITLE (1).

Question set 2A; in subroutine RDTA

```
ENTER:
M    =  NUMBER OF INDEPENDENT VARIABLES
N    =  NUMBER OF DATA POINTS PER VARIABLE
IEX  =  1 TO USE A SUBSET OF THE VARIABLES

3I   FORMAT
```

M is one less than the total number of variables in the input data because one variable is designated as dependent. When IEX = 1 the user is asked, in question sets 2B and 2C, to specify which variables are to be used. When IEX ≠ 1, the first M columns in the data set are used as independent variables and the program proceeds to question set 3. The M + 1st column, as originally entered, is always used for the dependent variable.

Question set 2B; in subroutine RDATA

ENTER:
NV = NUMBER OF INDEPENDENT VARIABLES TO BE RETAINED
IEXIX = 1 TO SPECIFY THE VARIABLES TO BE RETAINED

2I FORMAT

Question set 2C; in subroutine RDATA

ENTER:
ISV = NUMBERS OF THE VARIABLES TO BE IN OR EXCLUDED

When IEXIX = 1 the response to question set 2C is a list of variables (by position in the data matrix) to be retained. When IEXIX ≠ 1 the response to question 2C is a list of variables to be excluded.

An operational point should be noted here. The data set reduction routine can be used several times during one program run. It always operates on the current data set and sufficient information to perform that reduction is automatically retained for possible later use. However, each time the reduction routine is used previous reduction information is overwritten. Thus, if the variable deletion option is exercised more than once some later portions of GMDH will not operate properly.

When the variable deletion option (IEX = 1) is used more than once on a data set, it is recommended that the program be restarted once the desired variable set is chosen. The desired variable selection can then be performed either outside the program or in a single step within the program.

Question set 2D; in subroutine RDATA

ENTER:
IEX = 1 TO USE A SUBSET OF THE VARIABLES

1I FORMAT

Question set 2D is asked after question set 3B to give the user a chance to delete variables after the statistics of the data have been examined. When IEX = 1 the program returns to question set 2B and when IEX ≠ 1 the program proceeds to question set 4.

Question set 3A; in subroutine RDATA

ENTER:
IMSD = 1 TO PRINT THE MEANS AND STANDARD DEVIATIONS
ICORR = 1 TO PRINT THE CORRELATION MATRIX

2I FORMAT

When either IMSD or ICORR = 1 the calculations for both are performed. However, only the requested data is printed. When IMSD = 1 the coefficient of variation is also printed. The program efficiency could be improved by calculating only the requested information. These statistics are currently calculated in subroutine MSDCORR which calls IMSL (1977) subroutine BECORI. IMSL (1977) documentation should be consulted for information about BECORI.

Question set 3B; in subroutine RDATA

ENTER:
IC1, IC2 TO PRINT VARIABLES IC1 THROUGH IC2

2I FORMAT

IC1 and IC2 are column numbers which identify variables. When IC1 = IC2 ≠ 0 one variable is printed and when IC1 or IC2 = 0 the program returns to question set 2D. The output is the data corresponding to variables IC1 through IC2.

Question set 4A; in subroutine INPAR:

ENTER:
LO = 0 FOR GMDH
1 FOR STEPWISE REGRESSION
NTR = NUMBER OF ESTIMATION DATA POINTS
JEM = 0 FOR THE MEAN SQUARED RESIDUAL ERROR MEASURE
1 FOR THE R SQUARED ERROR MEASURE
IADJ = 0 TO ADJUST THE ERROR MEASURE

4I FORMAT

When LO = 1 GMDH performs only a linear stepwise regression on the input data.

The first NTR data points are used to estimate the equation coefficients and the last N - NTR data points are used to calculate the error measure. When NTR = N the error measures are calculated from all N data points. The unadjusted R^2 on the first NTR data points is always calculated, regardless of the values of JEM or IADJ. The error measure controlled by JEM and IADJ is an additional calculation performed as follows:

$$\text{mean squared error (mse)} = \frac{1}{N-S} \sum_{i=S}^{N} \hat{e}_i^{\,2}$$

$$R^2 = 100\ (1 - {}^{mse}/V[y])$$

where e_i = the i^{th} residual

$$S = \text{NTR} + 1 \text{ when NTR} < N$$

$$1 \text{ when NTR} = N$$

$$V[y] = \frac{1}{N-S} \sum_{i=S}^{N} (y_i - \bar{y})^2$$

and

$$\bar{y} = \frac{1}{N-S} \sum_{i=S}^{N} y_i$$

When IADJ = 0 and NTR = N the error measures are adjusted as follows,

$$\text{mse adjusted} = (\text{mse})\ (N-k)$$

$$R^2 \text{ adjusted} = 100\ (1 - (1 - R^2)\ (N-1)/(N-k))$$

where k = the number of coefficients in the equation, including the constant.

The error measures are never adjusted when NTR < N. These calculations are performed in subroutine CEM.

When LO = 0 the program proceeds to question set 4B and when LO = 1 the program proceeds to question set 5.

Question set 4B; in subroutine INPAR

ENTER:
MS = NUMBER OF VARIABLES PASSED BETWEEN LAYERS
NLAY = NUMBER OF LAYERS

2I FORMAT

Question set 4B is asked only when LO = 0. MS must be between 3 and M(M-1)/2, where M is the number of independent variables. NLAY must be less than the value of ID3 set in the common block initialization. (See Section B.3). The limits for both MS and NLAY are printed with question set 4B during program execution.

Question set 5; in subroutine REGPAR

ENTER:
ALFA(1) = SIGNIFICANCE LEVEL FOR ENTERING VARIABLES
ALFA(2) = SIGNIFICANCE LEVEL FOR DELETING VARIABLES
IJOB(1) = 0 TO NOT PERFORM A LACK OF FIT TEST
IJOB(2) = 0 TO PERFORM ONLY AN OVERALL F TEST

2F, 2I FORMAT

These four values are parameters of IMSL (1977) subroutine RLSEP. The significance levels for entering and deleting variables are typically in the range 0.01 to 0.10, though other choices may be appropriate for

special situations. For example, when ALFA(1) = 0 and ALFA(2) = 1, only the variables forced into the equation will be chosen (see question set 6). ALFA(2) must always be greater than or equal to ALFA(1). Draper and Smith (1966) may be consulted for further information about these significance levels.

The user unfamiliar with lack of fit tests should always specify IJOB(1) = 0. The choice of IJOB(2) is subjective. A partial F test on every variable, performed when IJOB(2) $\neq$ 0, is more stringent than an overall F test.

Question set 6A; in subroutine VFORC

ENTER:
NVF = NUMBER OF FORCED VARIABLES

1I FORMAT

This question is asked only when LO = 1 (see question set 4A). When NVF $\neq$ 0 the program proceeds to question 6B. When NVF = 0 the program proceeds to quesiton set 7.

Question set 6B; in subroutine VFORC

ENTER:
NUMBERS OF THE FORCED VARIABLES

The column numbers of the variables to be forced into the equation should be entered here. The numbers may be entered in any order, but NVF entries are required.

Question set 7; in subroutine EXEQU

ENTER:
IPS = 0 TO EXAMINE ONLY INDIVIDUAL EQUATIONS
1 TO PRINT A SUMMARY OF THE EQUATION EVALUATION
2 TO PRINT A SUMMARY OF EQUATION COEFFICIENTS
IREV = 0 TO RETAIN COEFFICIENT ESTIMATES
1 TO REESTIMATE COEFFICIENTS

2I FORMAT

When IPS = 0 the program proceeds to question set 8. When IPS = 1 or 2 the requested information is printed in file IWRIT and when IWRIT $\neq$ 6 the summary of equation evaluation is also printed on the terminal. This information is printed to aid the user in choosing which equations to examine more thoroughly. (see question set 9). All information given when IPS = 1 is also given when IPS = 2. The equation evaluation summary is a list of the values of the chosen error measure (see question set 4A). Coefficients of the linear equations are listed in the order of occurence of the variables in the data file. The labeling system, equations numbers, and location indices used for the quadratic forms generated in the GMDH are explained in Section B.3.3.

When IREV = 0 the coefficients estimated from the first NTR data points are retained. When IREV = 1 the coefficients are reestimated using all of the data, but the model structure is not changed. When IREV = 1 and IPS $\neq$ 0, the model coefficients and error measures are replaced in program storage by the new values. When IREV = 1 and IPS = 0, the initial coefficient and error measure values are retained in program storage after the requested information is developed and printed.

Question set 8; in subroutine EXEQU

```
ENTER:
ICON =  0 TO STOP PROGRAM
        1 TO RESTART PROGRAM
        2 TO EXAMINE AN EQUATION
        3 TO RESTART SUBROUTINE EXEQU
IREV =  0 TO RETAIN COEFFICIENTS
        1 TO REESTIMATE COEFFICIENTS

2I   FORMAT
```

When ICON = 1 the program returns to question set 1, when ICON = 2 the program proceeds to question set 9, and when ICON = 3 the program returns to question set 7. The effect of IREV was explained with question set 7.

Question set 9; in subroutine EXEQU

ENTER:
IND2 = 0 FOR A QUADRATIC FORM
IEQU = 1 FOR A LINEAR FORM
IEQU = EQUATION NUMBER
LA = LAYER NUMBER
IPR = 1 TO PRINT THE EQUATION
IPA = 1 TO PRINT THE ANOVA TABLE
IPL = 1 TO VIEW GRAPHICS

6I FORMAT

The first three parameters identify the desired equation and the last three parameters identify the desired information. IND2 = 0 calls for an equation from within a layer of GMDH. IND2 = 1 calls for a linear equation, either an equation developed from all input variables to a layer or the single equation developed when LO = 1 (see question set 4). IEQU is the sorted position of an equation within a layer. LA is the layer in which the equation was developed. For example, to indicate the second best equation in the third layer, set IND2 = 0, IEQU = 2, and LA = 3. When IND2 = 1, IEQU should also equal 1. The equation and variable labeling systems used in the program output are explained in Section B.3.3. The entries in the ANOVA are described in Draper and Smith (1966). When IPL = 1 the program proceeds to question set 10A. When IPL $\neq$ 1 the program returns to question set 8.

Question set 10A; in subroutine EXEQU

```
ENTER:
ICTRL4    =  0 TO CONTINUE PROGRAM
             1 TO PRINT LIST OF PLOTS
           2-9 TO IDENTIFY A PLOT
ICTRL3    =  0 FOR AUTOMATIC PLOTTING
             1 TO CONTROL PLOT FORMAT
             2 TO STORE VECTORS
```

2I FORMAT

When ICTRL4 = 0 the program returns to question set 8. When ICTRL4 = 1 the program proceeds to question set 10B. If a number identifying a plot (see Table B.2.1 or question set 10B) is entered for ICTRL4 the program proceeds as though question set 10B had been asked. This option allows the user who is familiar with the available plots to avoid having them listed at the terminal.

When ICTRL3 = 0 or 1 a plotting routine named PLOT1 is called. PLOT1 is discussed in Section B.3.5.1. When ICTRL3 = 2 the program proceeds to question set 10D, after question set 10B or 10C, as controlled by ICTRL4, have been asked, and stores the information required to produce the selected plot in the file then designated. The stored information can be plotted later with a program called PLWSV or can be used to calculate model statistics. PLWSV is described in Section B.3.6.2.

Question set 10B; in subroutine EXEQU

```
ENTER:
ICTRL4 = 2, STAND. RES. VS. OBS. #
         3, ORIG. RES. VS. OBS. #
         4, STAND. RES. VS. PRED. RESP.
         5, ORIG. RES. VS. PRED. RESP.
         6, STAND. RES. VS. PREDICTOR
         7, ORIG. RES. VS. PREDICTOR
         8, OBS. RESP. VS. OBS. #
         9, PREDICTOR VS. OBS. #
```

1I FORMAT

The unabreviated plot names are listed in Table B.2.1. Standardized residuals are the original residuals divided by their standard deviation. The standard deviation is based on only the estimation data.

When ICTRL4 = 6, 7, or 9 the program proceeds to question set 10C. Otherwise the program returns to question set 10A or proceeds to question set 10D, as controlled by ICTRL3.

Question set 10C: in subroutine EXEQU

ENTER:
IP = PREDICTOR NUMBER

1I FORMAT

IP is the column number, in the data matrix PRD, of the desired predictor variable. Remember to make appropriate adjustments if some variables were deleted from the original data set through question set 2.

Question set 10D; in subroutine STORVEC

ENTER:
IFILE = THE FILE NUMBER FOR THE VECTORS

1I FORMAT

ICTRL4	PLOT		
2	Standardized Residuals	vs.	Observation Number
3	Original Residuals	vs.	Observation Number
4	Standardized Residuals	vs.	Predicted Response
5	Original Residuals	vs.	Predicted Response
6	Standardized Residuals	vs.	Predictor
7	Original Residuals	vs.	Predictor
8	Observed Response	vs.	Observation Number
9	Predictor	vs.	Observation Number

Table B.2.1

Plot Selection Controlled by ICTRL4

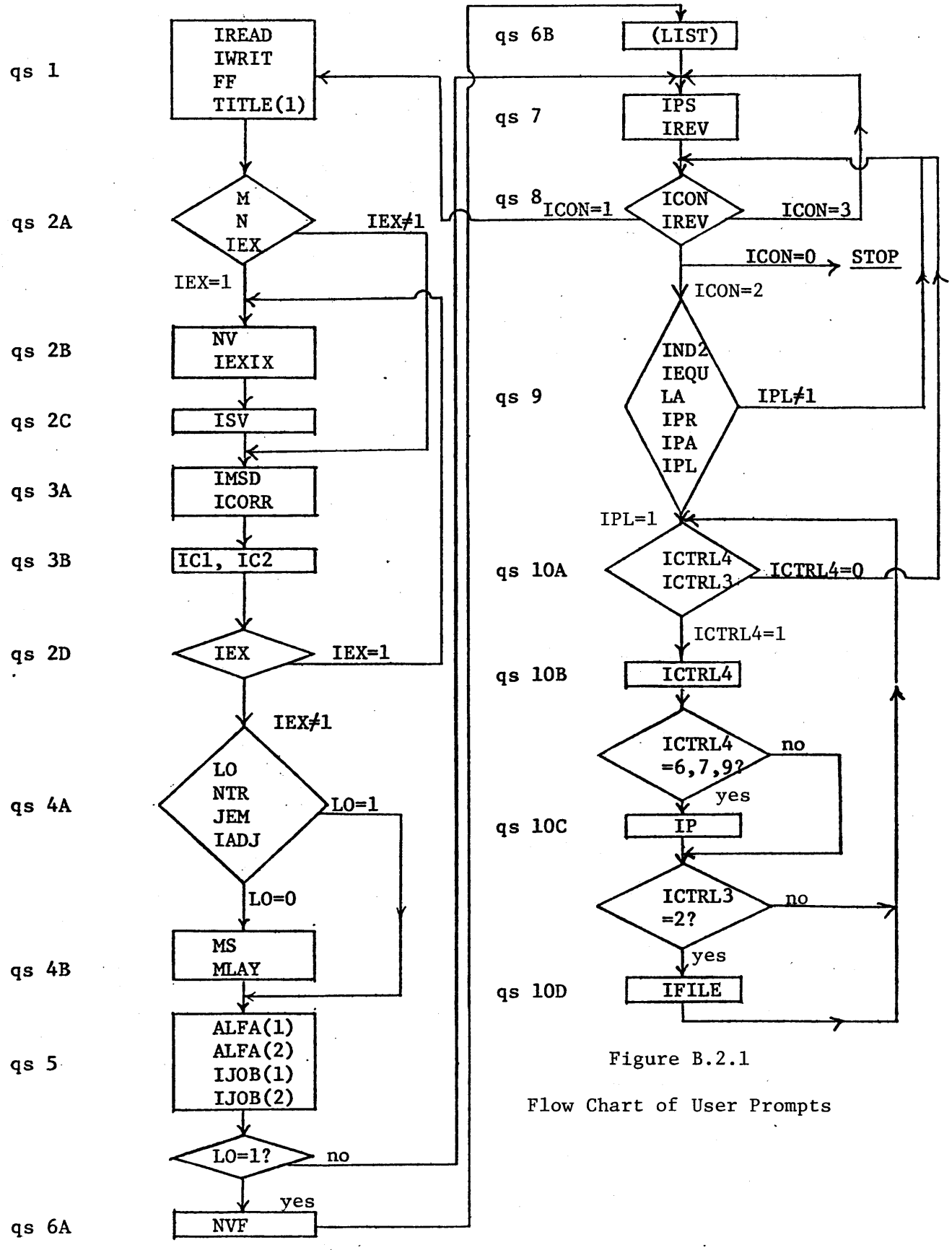

Figure B.2.1

Flow Chart of User Prompts

IFILE is a 2 digit integer designating the file in which information necessary to produce the requested plot is stored. A file named file NN, where NN is equal to IFILE, will be created and placed in the working directory. IFILE should not equal 5, 6, 7, 25, 26, 41, 42, or IWRIT (see question set 1).

A flow chart of the user prompts is shown in Figure B.2.1.

B.3 INPUT AND OUTPUT

B.3.1 Input Data Format

Data is read into GMDH through subroutine RDATA. RDATA is currently equipped to read either character or unformatted files. Both type of files must be arranged so all the values of one variable preceed all the values of the next variable and the dependent variable is listed last. The entries in character files must be separated by either commas or spaces and the first entry for each variable must begin a new record. Character files are read with a list directed format. Unformatted files must have all the values for one variable in one record and are read with an unformatted read statement.

Subroutine RDATA can easily be modified to read other file formats. The restriction is that each column of matrix PRD must contain all the values for one variable and the dependent variable must be in column M + 1. Thus each row of PRD contains all the variables in a given observation.

B.3.2 Common Block Parameter Assignments

The amount of space allocated for the dimensioned variables may be adjusted by changing parameter values in the file GMDH _ COM incl. fortran.

(see section B.5.1). Values for ID1, ID2, ID3, and IMS must be chosen according to the user's needs. Instructions are given in the file listing (see section B.5.1) for assigning values to ID1, ID2, ID3, IMS, IMX, ID7, ID8, ID9, and ID10. Parameters ID4, ID5, and ID6 are calculated from the other parameter values.

Parameter ID1 for common block P1 should be at least as large as ID2 and should be set to the same value in both GMDH_COM.incl. fortran and in PLOT1 (see Section B.3.5.1.3).

B.3.3 Equation Labels in Printed Output

Quadratic equations from all combinations of 2 different variables are developed in each layer. These equations are developed and numbered in an order determined by the rows in a lower triangular matrix, as shown in Figure B.3.1 where the pairs of entries are the variable numbers used in the

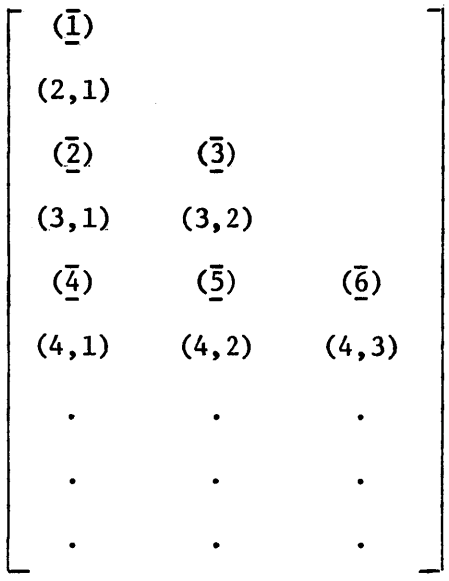

Figure B.3.1 Order of Quadratic Equation Development

equations and the circled number above each pair is called the location index in the program output.

Subroutine QMAP creates the map between the pairs of variables and the location indices illustrated in Figure B.3.1. This map can be printed by changing the main program to call subroutine QMAP with the argument IMAP = 1.

Each quadratic equation of variables x_1 and x_2 has the form

$$\hat{y} = \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_1^2 + \hat{\beta}_4 x_2^2 + \hat{\beta}_5 x_1 x_2 + \hat{\beta}_6$$

The subscript system in the above equation is used to label the coefficients in the program output. After all the equations in a layer have been developed they are reordered according to the values of the error measures, after which their sequential position is called the equation number. The equation number, location index, and layer are the 3 identifiers for equations. Each layer after the first is created from the reordered response variable of the previous layer and the location indices are based on the reordered variables of the previous layer. The development of a given equation is traced by subroutine DIAKA and may be printed by calling subroutine PREQU with IIACT and IKACT = 1. The vectors IACT and KACT are explained in the source listing for subroutine PREQU. (see Section B.5.2)

B.3.4 Additional Output Capabilities

Most matrices used in GMDH can be printed by changing the main program to call subroutines PRCOEFF, PRRSUM, PRDATA, or PRERRM with the appropriate arguments. The arguments are explained in the comments for each subroutine. (See Section B.5.2)

B.3.5 Graphic Output

The graphic output options controlled by question set 10 (see Section B.2) are implemented through programs named PLOT1 and PLWSV. PLOT1 is a general purpose plotting routine. PLWSV reads the files created when ICTRL3 = 2 (see question set 10, Section B.2) and calls PLOT1 as a subroutine. PLOT1 is described in Section B.3.5.1 and PLWSV is described in Section B.3.5.2.

B.3.5.1 PLOT1

PLOT1 can be used to produce either plots on a terminal with graphics capability, such as the tektronix 4015, or files which can be used by Cal Comp plotting equipment. The calcomp compatible subroutines, described in Honeywell Multics documentation, and SCLGPH, an MIT Information Processing Center (IPC) supplied subroutine described in IPC publication AP-59-3, are used in PLOT1.

Up to 5 different curves may be plotted on one set of axes. The points in each curve may be connected, marked with symbols at specified intervals, or both connected and marked. All the plotting parameters may be changed interactively.

The input arguments, user prompts, and common block parameter assignments are described in Sections B.3.5.1.1, B.3.5.1.2, and B.3.5.1.3. System documentation should be consulted for methods of producing plots at graphics terminals or peripheral plotting devices.

B.3.5.1.1 Input Arguments

The 2 input arguments are ICTRL3 and ICTRL4. The use of these variables in GMDH was discussed in Section B.2. When ICTRL3 = 0 a set of default variable values, assigned in the beginning of PLOT1, control the plot format. When ICTRL3 $\neq$ 0 these variables are assigned interactively.

The user prompts for this process are described in Section B.3.5.1.2. ICTRL4 is used to control choices in PLOT1 and is application specific. The current default variable assignments controlled by ICTRL4 are compatible with the plotting options described in question set 10 in Chapter 2. The plot titles, axis labels, and whether the points are connected or marked with symbols are currently determined by ICTRL4.

In addition to the 2 input arguments, NCURVE, ICTRL(k,1) for k = 1 to NCURVE, A2(I), B2(I), TITLE (1), and TITLE (2) are assigned in the calling program. NCURVE is the number of different curves on one plot. ICTRL(k,1) is the number of points in curve k. A2 and B2 are the complete sets of points for the abscissa and ordinate of the NCURVE curves. The point sets for the different curves are separated in PLOT1. TITLE (1) and TITLE (2) are titles for the plot. These and other variables are in a common block to allow the user to assign values in either the calling program or in PLOT1. However, the user will find it is generally convenient to assign at least the variables listed above in the calling program.

B.3.5.1.2 User Interaction

Every time PLOT1 is called the user is asked the following question.

```
ENTER
ILN = 1 TO DRAW HORIZONTAL LINES ON PLOTS

1I    FORMAT
```

when ILN = 1, horizontal lines at $\pm$ 1, $\pm$ 2, and $\pm$ 3 are drawn on the plots of standardized residuals (ICTRL4 = 2, 4, or 6). The user will normally find these lines desirable. The purpose of the question is primarily to simplify the use of PLOT1 with other programs.

When ICTRL3 $\neq$ 0 a series of questions is put to the user. Each set of questions has the same basic format. The user is given a list of plot format control variables and associated indicator variables. The values of the indicator variables must be set to 0 to retain the existing values of the plot format control variables. When a non zero value is entered the current value of the plot format control variable is printed at the terminal and the user is asked to enter a new value. Formats for entering both numeric and character variables are as described in Section B.2.1. The plot format control variables are described in the comments in the beginning of PLOT1. (See Section B.5.3) Further information may be found in Honeywell Multics documentation. Note that all coordinate positions are in a 1024 by 1024 device independent grid.

B.3.5.1.3 Common Block Parameter Assignment

Parameter IP1 must be set to the largest number of points in a single curve. Parameter IP2 is then automatically calculated to allow plotting of up to 5 curves, each of maximum length IP1.

B.3.5.2 PLWSV

PLWSV reads and plots files created by GMDH when ICTRL3 = 2 (see question 10, Section B.2). An option to smooth the curve is also available. The following 3 questions are put to the user. First,

```
ENTER:
IFILE = THE FILE NUMBER FOR THE VECTORS

1I   FORMAT
```

IFILE should be the same 2 digit integer specified in question set 10D in GMDH, unless the file has been renamed, in which case IFILE is the newly designated attachment number.

Second,

```
ENTER:
IND1  = 0 TO PLOT ALL POINTS
        1 TO PLOT MOVING AVERAGE OF NA
        2 TO PLOT SIMPLE AVERAGES OF NA
NA
```

2I FORMAT

The averages are calculated in subroutine SMOOTH. SMOOTH is currently set up specifically for set 1 data in the 0000 GMT cycle of the National Weather service Model Output Statistics equation development program. (see Chapter 2). The averages are calculated only within years of data. SMOOTH can be easily modified for other data sets.

Third,

```
ENTER:
ICTRL3 = 0 FOR AUTOMATIC PLOTTING
         1 TO CONTROL PLOT FORMAT
```

1I FORMAT

ICTRL3 was explained in Section B.3.5.1.1.

B.4 GMDH Sample Output

User responses are marked by arrows (←———). The question set numbers are indicated next to the questions to aid cross referencing with Section B.2. Everything else is typed by the program. When IWRIT does not equal 6 program output is stored in a file and does not appear at the terminal, except as noted in question set 7.

```
gmdh
 ENTER:
 IREAD     = INPUT FILE NUMBER                                    qs 1
 IWRIT     = OUTPUT FILE NUMBER
 FF        = 0 FOR A CHARACTER FILE
             1 FOR AN UNFORMATTED FILE
 TITLE(1) = DATA FILE NAME

 3I,/,A50 FORMAT
10 6 0 <----------
HALD DATA
 ENTER:                                                         qs 2A
 M    = NUMBER OF INDEPENDENT VARIABLES
 N    = NUMBER OF DATA POINTS PER VARIABLE
 IEX = 1 TO USE A SUBSET OF THE VARIABLES

 3I FORMAT
4 13 0 <----------
 ENTER:                                                         qs 3A
 IMSD  = 1 TO PRINT THE MEANS AND STANDARD DEVIATIONS
 ICORR = 1 TO PRINT THE CORRELATION MATRIX

 2I FORMAT
1 1 <----------
```


STATISTICS OF THE DATA SET HALD DATA

VARIABLE NUMBER	MEAN	STANDARD DEVIATION	COEFFICIENT OF VARIATION	VARIABLE NUMBER	MEAN	STANDARD DEVIATION	COEFFICIENT OF VARIATION
1	0.74615E+01	0.58824E+01	0.78836E+02	2	0.48154E+02	0.15561E+02	0.32315E+02
3	0.11769E+02	0.64051E+01	0.54423E+02	4	0.30000E+02	0.16738E+02	0.55794E+02
5	0.95423E+02	0.15044E+02	0.15765E+02				

CORRELATION MATRIX OF THE INPUT AND RESPONSE VARIABLES OF DATA SET HALD DATA

```
 1    1.0000
 2    0.2286  1.0000
 3   -0.8241 -0.1392  1.0000
 4   -0.2454 -0.9730  0.0295  1.0000
 5    0.7307  0.8163 -0.5347 -0.8213  1.0000
ENTER:                                                          qs 3B
IC1,IC2 TO PRINT VARIABLES IC1 THROUGH IC2

 2I FORMAT
1 5 <----------
```

THE DATA MATRIX HALD DATA
VARIABLE 5 IS THE RESPONSE

```
VARIABLE   1
 0.70000E+01  0.10000E+01  0.11000E+02  0.11000E+02  0.70000E+01  0.11000E+02  0.30000E+01  0.10000E+01  0.20000E+01  0.21000E+02
 0.10000E+01  0.11000E+02  0.10000E+02
```

```
VARIABLE   2
 0.26000E+02  0.29000E+02  0.56000E+02  0.31000E+02  0.52000E+02  0.55000E+02  0.71000E+02  0.31000E+02  0.54000E+02  0.47000E+02
 0.40000E+02  0.66000E+02  0.68000E+02
VARIABLE   3
 0.60000E+01  0.15000E+02  0.80000E+01  0.80000E+01  0.60000E+01  0.90000E+01  0.17000E+02  0.22000E+02  0.18000E+02  0.40000E+01
 0.23000E+02  0.90000E+01  0.80000E+01
VARIABLE   4
 0.60000E+02  0.52000E+02  0.20000E+02  0.47000E+02  0.33000E+02  0.22000E+02  0.60000E+01  0.44000E+02  0.22000E+02  0.26000E+02
 0.34000E+02  0.12000E+02  0.12000E+02
VARIABLE   5
 0.78500E+02  0.74300E+02  0.10430E+03  0.87600E+02  0.95900E+02  0.10920E+03  0.10270E+03  0.72500E+02  0.93100E+02  0.11590E+03
 0.83800E+02  0.11330E+03  0.10940E+03
 ENTER:
 IC1,IC2 TO PRINT VARIABLES IC1 THROUGH IC2                                    qs 3B

 2I FORMAT
0 0  <----
 ENTER:
 IEX = 1 TO USE A SUBSET OF THE VARIABLES                                      qs 2D

 1I FORMAT
1  <----
 ENTER:
 NV    = NUMBER OF INDEPENDENT VARIABLES TO BE RETAINED                        qs 2B
 IEXIX = 1 TO SPECIFY THE VARIABLES TO BE RETAINED

 2I FORMAT
3 0  <----
 ENTER:
 ISV = NUMBERS OF THE VARIABLES TO BE IN OR EX CLUDED                          qs 2C

              1 I FORMAT
3  <----
 ENTER:
 IMSD  = 1 TO PRINT THE MEANS AND STANDARD DEVIATIONS                          qs 3A
 ICORR = 1 TO PRINT THE CORRELATION MATRIX

 2I FORMAT
1 1  <----

                        STATISTICS OF THE DATA SET HALD DATA
      VARIABLE  MEAN           STANDARD       COEFFICIENT          VARIABLE  MEAN           STANDARD       COEFFICIENT
      NUMBER                   DEVIATION      OF VARIATION         NUMBER                   DEVIATION      OF VARIATION
         1     0.74615E+01    0.58824E+01    0.78836E+02             2     0.48154E+02    0.15561E+02    0.32315E+02
         3     0.30000E+02    0.16738E+02    0.55794E+02             4     0.95423E+02    0.15044E+02    0.15765E+02
                        CORRELATION MATRIX OF THE INPUT AND RESPONSE VARIABLES OF DATA SET HALD DATA

  1    1.0000
  2    0.2286  1.0000
```

```
  3   -0.2454 -0.9730  1.0000
  4    0.7307  0.8163 -0.8213  1.0000                                   qs 3B
 ENTER:
 IC1,IC2 TO PRINT VARIABLES IC1 THROUGH IC2

 2I FORMAT
1 4 <----

                 THE DATA MATRIX HALD DATA
                 VARIABLE    4 IS THE RESPONSE
VARIABLE   1
 0.70000E+01  0.10000E+01  0.11000E+02  0.11000E+02  0.70000E+01  0.11000E+02  0.30000E+01  0.10000E+01  0.20000E+01  0.21000E+02
 0.10000E+01  0.11000E+02  0.10000E+02
VARIABLE   2
 0.26000E+02  0.29000E+02  0.56000E+02  0.31000E+02  0.52000E+02  0.55000E+02  0.71000E+02  0.31000E+02  0.54000E+02  0.47000E+02
 0.40000E+02  0.66000E+02  0.68000E+02
VARIABLE   3
 0.60000E+02  0.52000E+02  0.20000E+02  0.47000E+02  0.33000E+02  0.22000E+02  0.60000E+01  0.44000E+02  0.22000E+02  0.26000E+02
 0.34000E+02  0.12000E+02  0.12000E+02
VARIABLE   4
 0.78500E+02  0.74300E+02  0.10430E+03  0.87600E+02  0.95900E+02  0.10920E+03  0.10270E+03  0.72500E+02  0.93100E+02  0.11590E+03
 0.83800E+02  0.11330E+03  0.10940E+03
 ENTER:                                                                qs 3B
 IC1,IC2 TO PRINT VARIABLES IC1 THROUGH IC2

 2I FORMAT
0 0 <----
 ENTER:
 IEX = 1 TO USE A SUBSET OF THE VARIABLES                              qs 2D

 1I FORMAT
0 <----
 ENTER:
 LO   = 0 FOR GMDH                                                     qs 4A
        1 FOR STEPWISE REGRESSION
 NTR  = NUMBER OF ESTIMATION DATA POINTS
 JEM  = 0 FOR THE MEAN SQUARED RESIDUAL ERROR MEASURE
        1 FOR THE R SQUARED ERROR MEASURE
 IADJ = 0 TO ADJUST THE ERROR MEASURE

 4I FORMAT                                                             qs 4B
0 7 0 1 <----
 ENTER:
 MS    = NUMBER OF VARIABLES PASSED BETWEEN LAYERS
         (3 ≤ MS ≤              3)
 NLAY = NUMBER OF LAYERS
         (NLAY ≤              3)

 2I FORMAT
```

```
3 3 <----------
 ENTER:                                                    qs 5
 ALFA(1) = SIGNIFICANCE LEVEL FOR ENTERING VARIABLES
 ALFA(2) = SIGNIFICANCE LEVEL FOR DELETING VARIABLES
 IJOB(1) = 0 TO NOT PERFORM A LACK OF FIT TEST
 IJOB(2) = 0 TO PERFORM ONLY AN OVERALL F TEST

 2F,2I FORMAT
0.05 0.05 0 1 <----------
     BEGINNING LAYER              1
     BEGINNING LAYER              2
     BEGINNING LAYER              3
 ENTER:                                                    qs 7
 IPS =  0 TO EXAMINE ONLY INDIVIDUAL EQUATIONS
        1 TO PRINT A SUMMARY OF THE EQUATION EVALUATION
        2 TO PRINT A SUMMARY OF EQUATION COEFFICIENTS
 IREV = 0 TO RETAIN COEFFICIENT ESTIMATES
        1 TO REESTIMATE COEFFICIENTS

 2I FORMAT
2 0 <----------
```


REGRESSION ON DATA FILE HALD DATA

REGRESSION PARAMETERS

THE SIGNIFICANCE LEVELS FOR ENTERING AND DELETING VARIABLES, ALFA(1) AND ALFA(2), = 0.50000E-01 AND 0.50000E-01
THE LACK OF FIT TEST PARAMETER, IJOB(1), = 0 THE PARTIAL OR OVERALL F TEST STATISTIC, IJOB(2), = 1
JEM = 0 IADJ = 1 IREV = 0 NTR = 7 NTE = 6

THE LOCATION INDICES AND COEFFICIENTS OF THE 3 BEST PREDICTORS IN LAYER 1

NUMBER	LOCATION INDEX	BETA 1	BETA 2	BETA 3	BETA 4	BETA 5	INTERCEPT
1	2	0.00000E+00	0.10168E+01	-0.86070E-02	0.00000E+00	0.00000E+00	0.98743E+02
2	1	0.67891E+00	0.00000E+00	0.00000E+00	0.11556E+00	0.00000E+00	0.54402E+02
3	3	0.00000E+00	0.00000E+00	-0.92592E-02	0.00000E+00	0.00000E+00	0.10713E+03

THE LOCATION INDICES AND COEFFICIENTS OF THE 3 BEST PREDICTORS IN LAYER 2

NUMBER	LOCATION INDEX	BETA 1	BETA 2	BETA 3	BETA 4	BETA 5	INTERCEPT
1	2	0.00000E+00	0.00000E+00	0.00000E+00	0.55483E-02	0.00000E+00	0.44186E+02
2	1	0.00000E+00	0.00000E+00	0.55073E-02	0.00000E+00	0.00000E+00	0.44529E+02
3	3	0.00000E+00	0.00000E+00	0.00000E+00	0.55073E-02	0.00000E+00	0.44529E+02

THE LOCATION INDICES AND COEFFICIENTS OF THE 3 BEST PREDICTORS IN LAYER 3

NUMBER	LOCATION INDEX	BETA 1	BETA 2	BETA 3	BETA 4	BETA 5	INTERCEPT
1	1	0.10000E+01	0.00000E+00	0.00000E+00	0.00000E+00	0.00000E+00	0.27780E-05
2	2	0.10000E+01	0.00000E+00	0.00000E+00	0.00000E+00	0.00000E+00	0.27780E-05
3	3	0.00000E+00	0.10000E+01	0.00000E+00	0.00000E+00	0.00000E+00	0.27780E-05

```
                THE COEFFICIENTS OF THE REGRESSIONS IN EACH LAYER ON ALL THE PREDICTORS IN THAT LAYER
                (THE LAST ENTRY IS THE INTERCEPT)
                LAYER 1
0.13362E+01  0.00000E+00 -0.58708E+00  0.10361E+03
                LAYER 2
0.00000E+00  0.10000E+01  0.00000E+00  0.27780E-05
                LAYER 3
0.00000E+00  0.00000E+00  0.10000E+01  0.27780E-05

                THE LOCATION INDICES AND ERROR MEASURES OF THE  3 BEST PREDICTORS IN EACH LAYER
                                                LAYER  1
  NUMBER     LOCATION          MS RES          NUMBER    LOCATION          MS RES          NUMBER    LOCATION          MS RES
              INDEX                                         INDEX                                     INDEX
    1           2          0.37526E+02          2           1          0.96145E+02          3           3          0.16497E+03
                                                LAYER  2
  NUMBER     LOCATION          MS RES          NUMBER    LOCATION          MS RES          NUMBER    LOCATION          MS RES
              INDEX                                         INDEX                                     INDEX
    1           2          0.29740E+02          2           1          0.21942E+03          3           3          0.21942E+03
                                                LAYER  3
  NUMBER     LOCATION          MS RES          NUMBER    LOCATION          MS RES          NUMBER    LOCATION          MS RES
              INDEX                                         INDEX                                     INDEX
    1           1          0.21942E+03          2           2          0.21942E+03          3           3          0.21942E+03

                THE MS RES     OF THE REGRESSIONS ON ALL THE PREDICTORS IN EACH LAYER
  LAYER          MS RES
     1       0.99758E+01
     2       0.96145E+02
     3       0.21942E+03
ENTER:
ICON = 0 TO STOP PROGRAM                                                    qs 8
       1 TO RESTART PROGRAM
       2 TO EXAMINE AN EQUATION
       3 TO RESTART SUBROUTINE EXEQU
IREV = 0 TO RETAIN COEFFICIENTS
       1 TO REESTIMATE COEFFICIENTS

2I FORMAT
2 0 <----
ENTER:                                                                      qs 9
IND2 = 0 FOR A QUADRATIC FORM
       1 FOR A LINEAR FORM
IEQU = EQUATION NUMBER
LA   = LAYER NUMBER
IPR  = 1 TO PRINT THE EQUATION
IPA  = 1 TO PRINT THE ANOVA TABLE
IPL  = 1 TO VIEW GRAPHICS

6I FORMAT
```

0 2 3 1 1 1 ←

VARIABLE 2 LAYER 3

REGRESSION PARAMETERS

THE SIGNIFICANCE LEVELS FOR ENTERING AND DELETING VARIABLES, ALFA(1) AND ALFA(2), = 0.50000E-01 AND 0.50000E-01
THE LACK OF FIT TEST PARAMETER, IJOB(1), = 0 THE PARTIAL OR OVERALL F TEST STATISTIC, IJOB(2), = 1
JEM = 0 IADJ = 1 IREV = 0 NTR = 7 NTE = 6

THE COMPLETE EQUATION, BY LAYERS
RESPONSE (Y) =

```
VR   2 LY3 =
    0.27780E-05
 +  0.10000E+01*VR   3 LY2
VR   3 LY2 =
    0.44529E+02
 +  0.55073E-02*VR   2 LY1**2
VR   1 LY2 =
    0.44186E+02
 +  0.55483E-02*VR   1 LY1**2
VR   3 LY1 =
    0.10713E+03
 + -0.92592E-02*VR   3 LY0**2
VR   2 LY1 =
    0.54402E+02
 +  0.67891E+00*VR   2 LY0
 +  0.11556E+00*VR   1 LY0**2
VR   1 LY1 =
    0.98743E+02
 +  0.10168E+01*VR   1 LY0
 + -0.86070E-02*VR   3 LY0**2
```

THE COMPLETE EQUATION
RESPONSE (Y) =

```
VR   2 LY3 =
    0.60828E+02
 +  0.40682E+00 *VR   2
 +  0.25384E-02 *VR   2**2
 +  0.69248E-01 *VR   1**2
 +  0.73550E-04 *VR   1**4
 +  0.86418E-03 *VR   2    *VR   1**2
 +  0.11142E+01 *VR   1
 + -0.94309E-02 *VR   3**2
 +  0.57368E-02 *VR   1**2
 +  0.41102E-06 *VR   3**4
 + -0.97117E-04 *VR   1    *VR   3**2
```

ANALYSIS OF VARIANCE TABLE

SOURCE	D.F.	SS	MS	F RATIO	F TAIL AREA

REGRESSION	0.10000E+01	0.10607E+04	0.10607E+04	0.25582E+03	0.17397E-04
RESIDUALS	0.50000E+01	0.20731E+02	0.41462E+01		
CORRECTED TOTAL	0.60000E+01	0.10814E+04			

LACK OF FIT TEST	0.00000E+00	0.00000E+00 0.00000E+00
THE PERCENTAGE OF THE RESPONSE VARIATION EXPLAINED BY THE REGRESSION =		0.98083E+02
THE STANDARD DEVIATION OF THE RESIDUALS =		0.20362E+01
THE STANDARD DEVIATION OF THE RESIDUALS AS A PERCENTAGE OF THE RESPONSE MEAN =		0.21845E+01
THE ERROR MEASURE USED TO ORDER THE EQUATIONS =		0.21942E+03

REGRESSION MODEL SUMMARY, TXYB

VARIABLE	MEAN	COEFFICIENT	ADJ. SS	F RATIO	F TAIL AREA	VARIANCE
1	0.93214E+02	0.10000E+01	0.10607E+04	0.25582E+03	0.17397E-04	0.94279E-03
		INTERCEPT				
RESPONSE	0.93214E+02	0.27780E-05				

THE INVERSE OF THE INFORMATION MATRIX

1 0.9428E-03

```
ENTER:
ICTRL4 = 0 TO CONTINUE PROGRAM                                  qs 10A
         1 TO PRINT LIST OF PLOTS
         2-9 TO IDENTIFY A PLOT
ICTRL3 = 0 FOR AUTOMATIC PLOTTING
         1 TO CONTROL PLOT FORMAT
         2 TO STORE VECTORS

2I FORMAT
1 2 <----
ENTER:                                                          qs 10B
ICTRL4 = 2, STAND. RES. VS. OBS. NUM.
         3,  ORIG. RES. VS. OBS. NUM.
         4, STAND. RES. VS. PRED. RESP.
         5,  ORIG. RES. VS. PRED. RESP.
         6, STAND. RES. VS. PREDICTOR
         7,  ORIG. RES. VS. PREDICTOR
         8,  OBS. RESP. VS. OBS. NUM.
         9, PREDICTOR   VS. OBS. NUM.

1I FORMAT
2 <----
ENTER IFILE = THE FILE NUMBER FOR THE VECTORS                   qs 10D

1I FORMAT
50 <----
ENTER:
ICTRL4 = 0 TO CONTINUE PROGRAM                                  qs 10A
         1 TO PRINT LIST OF PLOTS
         2-9 TO IDENTIFY A PLOT
ICTRL3 = 0 FOR AUTOMATIC PLOTTING
```

```
          1 TO CONTROL PLOT FORMAT
          2 TO STORE VECTORS

 2I FORMAT
0 0 <----
 ENTER:                                                          qs 8
 ICON = 0 TO STOP PROGRAM
        1 TO RESTART PROGRAM
        2 TO EXAMINE AN EQUATION
        3 TO RESTART SUBROUTINE EXEQU
 IREV = 0 TO RETAIN COEFFICIENTS
        1 TO REESTIMATE COEFFICIENTS

 2I FORMAT
3 1 <----
 ENTER:                                                          qs 7
 IPS =  0 TO EXAMINE ONLY INDIVIDUAL EQUATIONS
        1 TO PRINT A SUMMARY OF THE EQUATION EVALUATION
        2 TO PRINT A SUMMARY OF EQUATION COEFFICIENTS
 IREV = 0 TO RETAIN COEFFICIENT ESTIMATES
        1 TO REESTIMATE COEFFICIENTS

 2I FORMAT
2 1 <----
     BEGINNING LAYER                1
     BEGINNING LAYER                2
     BEGINNING LAYER                3

              REGRESSION ON DATA FILE HALD DATA

                                  REGRESSION PARAMETERS
     THE SIGNIFICANCE LEVELS FOR ENTERING AND DELETING VARIABLES, ALFA(1) AND ALFA(2), = 0.50000E-01   AND   0.50000E-01
     THE LACK OF FIT TEST PARAMETER, IJOB(1), = 0     THE PARTIAL OR OVERALL F TEST STATISTIC, IJOB(2), = 1
     JEM =  0     IADJ =  1     IREV =  1     NTR =  7     NTE =  6

              THE LOCATION INDICES AND COEFFICIENTS OF THE   3 BEST PREDICTORS IN LAYER 1
     NUMBER   LOCATION      BETA 1          BETA 2          BETA 3          BETA 4          BETA 5       INTERCEPT
               INDEX
          1       2     0.00000E+00     0.14082E+01    -0.89165E-02     0.00000E+00     0.00000E+00     0.95247E+02
          2       1     0.72641E+00     0.00000E+00     0.00000E+00     0.69423E-01     0.00000E+00     0.54361E+02
          3       3     0.00000E+00     0.00000E+00    -0.10917E-01     0.00000E+00     0.00000E+00     0.10807E+03

              THE LOCATION INDICES AND COEFFICIENTS OF THE   3 BEST PREDICTORS IN LAYER 2
     NUMBER   LOCATION      BETA 1          BETA 2          BETA 3          BETA 4          BETA 5       INTERCEPT
               INDEX
          1       2     0.00000E+00     0.00000E+00     0.00000E+00     0.53070E-02     0.00000E+00     0.46054E+02
          2       1     0.00000E+00     0.00000E+00     0.52016E-02     0.00000E+00     0.00000E+00     0.47016E+02
          3       3     0.00000E+00     0.00000E+00     0.00000E+00     0.52016E-02     0.00000E+00     0.47016E+02
```

THE LOCATION INDICES AND COEFFICIENTS OF THE 3 BEST PREDICTORS IN LAYER 3

NUMBER	LOCATION INDEX	BETA 1	BETA 2	BETA 3	BETA 4	BETA 5	INTERCEPT
1	1	0.10000E+01	0.00000E+00	0.00000E+00	0.00000E+00	0.00000E+00	0.21329E-05
2	2	0.10000E+01	0.00000E+00	0.00000E+00	0.00000E+00	0.00000E+00	0.21329E-05
3	3	0.00000E+00	0.10000E+01	0.00000E+00	0.00000E+00	0.00000E+00	0.21329E-05

THE COEFFICIENTS OF THE REGRESSIONS IN EACH LAYER ON ALL THE PREDICTORS IN THAT LAYER
(THE LAST ENTRY IS THE INTERCEPT)

LAYER 1

0.14400E+01 0.00000E+00 -0.61395E+00 0.10310E+03

LAYER 2

0.00000E+00 0.10000E+01 0.00000E+00 0.42657E-05

LAYER 3

0.00000E+00 0.00000E+00 0.10000E+01 0.21329E-05

THE LOCATION INDICES AND ERROR MEASURES OF THE 3 BEST PREDICTORS IN EACH LAYER

LAYER 1

NUMBER	LOCATION INDEX	MS RES	NUMBER	LOCATION INDEX	MS RES	NUMBER	LOCATION INDEX	MS RES
1	2	0.15334E+02	2	1	0.10735E+02	3	3	0.83303E+02

LAYER 2

NUMBER	LOCATION INDEX	MS RES	NUMBER	LOCATION INDEX	MS RES	NUMBER	LOCATION INDEX	MS RES
1	2	0.14161E+02	2	1	0.13143E+02	3	3	0.13143E+02

LAYER 3

NUMBER	LOCATION INDEX	MS RES	NUMBER	LOCATION INDEX	MS RES	NUMBER	LOCATION INDEX	MS RES
1	1	0.13143E+02	2	2	0.13143E+02	3	3	0.13143E+02

THE MS RES OF THE REGRESSIONS ON ALL THE PREDICTORS IN EACH LAYER

LAYER	MS RES
1	0.74762E+01
2	0.97588E+01
3	0.13143E+02

ENTER:
ICON = 0 TO STOP PROGRAM
1 TO RESTART PROGRAM
2 TO EXAMINE AN EQUATION
3 TO RESTART SUBROUTINE EXEQU
IREV = 0 TO RETAIN COEFFICIENTS
1 TO REESTIMATE COEFFICIENTS

qs 8

2I FORMAT
2 0 ←

ENTER:
IND2 = 0 FOR A QUADRATIC FORM
1 FOR A LINEAR FORM
IEQU = EQUATION NUMBER

qs 9

```
 LA    = LAYER NUMBER
 IPR   = 1 TO PRINT THE EQUATION
 IPA   = 1 TO PRINT THE ANOVA TABLE
 IPL   = 1 TO VIEW GRAPHICS

 6I FORMAT
0 2 3 1 1 0 <----------
```

VARIABLE 2 LAYER 3

REGRESSION PARAMETERS

THE SIGNIFICANCE LEVELS FOR ENTERING AND DELETING VARIABLES, ALFA(1) AND ALFA(2), = 0.50000E-01 AND 0.50000E-01
THE LACK OF FIT TEST PARAMETER, IJOB(1), = 0 THE PARTIAL OR OVERALL F TEST STATISTIC, IJOB(2), = 1
JEM = 0 IADJ = 1 IREV = 0 NTR = 7 NTE = 6

```
   THE COMPLETE EQUATION, BY LAYERS
   RESPONSE (Y) =
VR   2 LY3 =
   -0.37177E+01
 +  0.10532E+01*VR   3 LY2
VR   3 LY2 =
    0.47016E+02
 +  0.52016E-02*VR   2 LY1**2
VR   1 LY2 =
    0.46054E+02
 +  0.53070E-02*VR   1 LY1**2
VR   3 LY1 =
    0.10807E+03
 + -0.10917E-01*VR   3 LY0**2
VR   2 LY1 =
    0.54361E+02
 +  0.72641E+00*VR   2 LY0
 +  0.69423E-01*VR   1 LY0**2
VR   1 LY1 =
    0.95247E+02
 +  0.14082E+01*VR   1 LY0
 + -0.89165E-02*VR   3 LY0**2

   THE COMPLETE EQUATION
   RESPONSE (Y) =
VR   2 LY3 =
    0.61991E+02
 +  0.43268E+00 *VR   2
 +  0.28909E-02 *VR   2**2
 +  0.41351E-01 *VR   1**2
 +  0.26404E-04 *VR   1**4
 +  0.55257E-03 *VR   2    *VR   1**2
 +  0.14236E+01 *VR   1
 + -0.90141E-02 *VR   3**2
```

```
+  0.10524E-01 *VR   1**2
+  0.42193E-06 *VR   3**4
+ -0.13327E-03 *VR   1    *VR   3**2
```

ANALYSIS OF VARIANCE TABLE

SOURCE	D.F.	SS	MS	F RATIO	F TAIL AREA
REGRESSION	0.10000E+01	0.10103E+04	0.10103E+04	0.71035E+02	0.38570E-03
RESIDUALS	0.50000E+01	0.71112E+02	0.14222E+02		
CORRECTED TOTAL	0.60000E+01	0.10814E+04			
LACK OF FIT TEST		0.00000E+00		0.00000E+00	0.00000E+00

THE PERCENTAGE OF THE RESPONSE VARIATION EXPLAINED BY THE REGRESSION = 0.93424E+02
THE STANDARD DEVIATION OF THE RESIDUALS = 0.37713E+01
THE STANDARD DEVIATION OF THE RESIDUALS AS A PERCENTAGE OF THE RESPONSE MEAN = 0.40458E+01
THE ERROR MEASURE USED TO ORDER THE EQUATIONS = 0.13143E+02

REGRESSION MODEL SUMMARY, TXYB

VARIABLE	MEAN	COEFFICIENT	ADJ. SS	F RATIO	F TAIL AREA	VARIANCE
1	0.92032E+02	0.10532E+01	0.10103E+04	0.71035E+02	0.38570E-03	0.10980E-02
		INTERCEPT				
RESPONSE	0.93214E+02	-0.37177E+01				

THE INVERSE OF THE INFORMATION MATRIX

```
 1    0.1098E-02
```



```
 ENTER:
 ICON = 0 TO STOP PROGRAM                                    qs 8
        1 TO RESTART PROGRAM
        2 TO EXAMINE AN EQUATION
        3 TO RESTART SUBROUTINE EXEQU
 IREV = 0 TO RETAIN COEFFICIENTS
        1 TO REESTIMATE COEFFICIENTS

 2I FORMAT
2 0 <----
 ENTER:                                                      qs 9
 IND2 = 0 FOR A QUADRATIC FORM
        1 FOR A LINEAR FORM
 IEQU = EQUATION NUMBER
 LA   = LAYER NUMBER
 IPR  = 1 TO PRINT THE EQUATION
 IPA  = 1 TO PRINT THE ANOVA TABLE
 IPL  = 1 TO VIEW GRAPHICS

 6I FORMAT
1 1 2 1 0 0 <----
```

REGRESSION ON INPUT VARIABLES FOR LAYER 2

REGRESSION PARAMETERS

```
        THE SIGNIFICANCE LEVELS FOR ENTERING AND DELETING VARIABLES, ALFA(1) AND ALFA(2), = 0.50000E-01   AND   0.50000E-01
        THE LACK OF FIT TEST PARAMETER, IJOB(1), = 0     THE PARTIAL OR OVERALL F TEST STATISTIC, IJOB(2), = 1
        JEM =  0     IADJ =  1     IREV =  0     NTR =  7     NTE =  6

   THE COMPLETE EQUATION, BY LAYERS
   RESPONSE (Y) =
 0.42657E-05
 +  0.10000E+01*VR   2LY1
VR   2 LY1 =
    0.54361E+02
 +  0.72641E+00*VR   2 LY0
 +  0.69423E-01*VR   1 LY0**2

   THE COMPLETE EQUATION
   RESPONSE (Y) =
 0.54361E+02
 +  0.72641E+00*VR   2 LY0
 +  0.69423E-01*VR   1 LY0**2
 ENTER:
 ICON = 0 TO STOP PROGRAM                                        qs 8
        1 TO RESTART PROGRAM
        2 TO EXAMINE AN EQUATION
        3 TO RESTART SUBROUTINE EXEQU
 IREV = 0 TO RETAIN COEFFICIENTS
        1 TO REESTIMATE COEFFICIENTS

 2I FORMAT
0 0 <----

STOP
```

```
C*****************************************************************
C***                                                           ***
C***                    GMDH_COM.INCL.FORTRAN                  ***
C***                                                           ***
C*****************************************************************
C THE INTEGER CONSTANTS IN THE FOLLOWING PARAMETER STATEMENTS MUST
C BE ASSIGNED ACCORDING TO THE NEEDS OF THE LARGEST
C DATA SET ON WHICH THE PROGRAM IS TO BE USED.
C ID1 = THE MAXIMUM NUMBER OF INDEPENDENT VARIABLES
C ID2 = THE MAXIMUM NUMBER OF DATA POINTS PER OBSERVATION
C ID3 = THE MAXIMUM NUMBER OF LAYERS TO BE DEVELOPED
C       ID3 MUST BE _ 5
C IMS = THE MAXIMUM NUMBER OF INDEPENDENT VARIABLES TO BE PASSED
C       BETWEEN LAYERS, 3 _ IMS _ (ID1)*(ID1-1)/2
C
      PARAMETER( ID1=118,ID2=518,ID3=3,IMS=50)
      PARAMETER( ID4=ID1+1)
C IMX = MAX(ID1,IMS)
      PARAMETER( IMX=118)
      PARAMETER( ID5=IMX+1,ID6=IMX*(IMX-1)/2)
      PARAMETER( IMXL2=2*ID5,IMM=ID5*(ID5+1)/2)
C ID7 = MAX(ID5,16)
C ID8 = MAX(ID5,6)
C ID9 = MAX(IMXL2,12)
C ID10 = MAX(IMM,21)
      PARAMETER( ID7=119,ID8=119,ID9=238,ID10=7140)
C
C*****************************************************************
      COMMON /C1/ PRD(ID2,ID5),IC,ICL,FF,IADJ.MGL,ICON,IEXDD.MDCL,IEXIX
      COMMON /C2/ WSM(ID2,ID8)
      COMMON /C3/ QCO(6,ID6,ID3)
      COMMON /C4/ IQCO(6,ID6,ID3)
      COMMON /C5/ IREAD,IWRIT,QP(ID2,6),CO(ID5,ID3).EM(ID6,ID3).
     &      IEMLOC(ID6,ID3),EMAT(ID3),ICO(ID5.ID3),
     &             IXD(ID9),ANOVA(ID7),XYB(ID8.5),TXYB(ID8.5).VARB(ID10).
     &      IH(ID9),BETA(ID8),RES(ID2,4),ISV(ID1),
     &             MAP(ID6,2),IACT(63),KACT(31),C(54.2),
     &             ALFA(2),IJOB(2),XMEAN(ID4).SD(ID4),CV(ID4).CORR(ID10),
     &             M,MS,N,NTR,NLAY,MX,NTE,ML,M2,M2L,MC2,MC2L,MSL.LA,
     &      MS2,MS2L,MSC2,MSC2L,MXL,MX2,MX2L,MXC2,MM.LM.NTRL.MXLAY2,
     &      MXLAYL,LO,JEM
      EXTERNAL RLSEP(DESCRIPTORS)
      EXTERNAL BECORI(DESCRIPTORS)
      CHARACTER*10 SR
C THE FOLLOWING SECTION IS FOR SUBROUTINE PLOT1.FORTRAN
      PARAMETER( IP1=2000)
```

```
      PARAMETER( IP2=5*IP1+2)
      CHARACTER*50 TITLE,CHAR,XAXIS,YAXIS
      COMMON /P1/ TITLE(11),XTITLE(11),YTITLE(11),ATITLE(11),
     &HTITLE(11),HSMBL(11),ASMBL(11),YSMBL(11),XSMBL(11),ISMBL(11),
     &NSMBL,NCHAR,NTITLE,A1(4),B1(4),A(IP1),B(IP1),NCURVE,
     &ICTRL(5,5),A2(IP2),B2(IP2),XAXIS,YAXIS
      EXTERNAL CCS_$AXIS(DESCRIPTORS)
      EXTERNAL CCS_$DFACT(DESCRIPTORS)
      EXTERNAL CCS_$LINE(DESCRIPTORS)
      EXTERNAL CCS_$PLOT(DESCRIPTORS)
      EXTERNAL CCS_$PLOTS
      EXTERNAL CCS_$SCALE(DESCRIPTORS)
      EXTERNAL CCS_$SYMBOL(DESCRIPTORS)
```

```
C*********************************************************
C*********************************************************
C**                                                     **
C**              GROUP METHOD OF DATA HANDLING          **
C**                                                     **
C*********************************************************
C*********************************************************
C
C THE MATRICES SHOULD TO BE DIMENSIONED
C ACCORDING TO THE REQUIREMENTS OF THE LARGEST DATA SET
C FOR WHICH THE PROGRAM WILL BE USED. ENTER
C THE APPROPRIATE VALUES IN FILE GMDH_COM.INCL.FORTRAN.
C
C DO NOT USE  FILES 25 OR 26 FOR ANY INPUT-OUTPUT OPERATIONS.
C FILE 25 IS USED IN THE PLOTTING SUBROUTINE TO DISPLAY
C MESSAGES ON THE TERMINAL AT THE COMPLETION OF A PLOT.
C FILE 26 IS USED TO LABEL PLOTS AND MAY BE DELETED
C AT THE END OF A RUN.
C
C THE DATA MUST BE ENTERED INTO MATRIX PRD SO THAT EACH COLUMN
C OF PRD CONTAINS ALL THE OBSERVATIONS OF ONE VARIABLE AND THE LAST
C COLUMN CONTAINS THE RESPONSE VARIABLE.  DATA IS ENTERED THROUGH
C SUBROUTINE RDATA.  THE READ STATEMENT IN RDATA MAY BE
C MODIFIED AS NEEDED TO READ THE ORIGINAL DATA FILE.
C
C MAJOR VARIABLES
C
C PRD = THE DATA MATRIX.
C WSM = A WORK SPACE.
C QP = A QUADRATIC SETTING IN STANDARD ORDER OF TWO VARIABLES
C      FROM PRD.  THE RESPONSE VARIABLE IS IN COLUMN 6.
C EM = THE ERROR MEASURE MATRIX FOR THE
C                 REGRESSIONS ON QP.
C IEMLOC = A LOCATION INDEX FOR THE ENTRIES IN EM.
C      SEE SUBROUTINE QMAP FOR DETAILS.
C MAP = A MAPPING OF QUADRATIC FORMS OF TWO VARIABLES
C      INTO ORDERED PAIRS OF INTEGERS.
C QCO = THE COEFFICIENTS OF THE REGRESSIONS ON EACH
C      QP. THE INTERCEPT IS IN ROW 6.  EACH COLUMN CONTAINS ONE EQUATION.
C IQCO = 1  IF QCO(I,J,K) = 0.
C                0 IF QCO(I,J,K) = 0.
C CO = THE COEFFICIENTS OF THE REGRESSIONS ON PRD
C      IN EACH LAYER.  EACH COLUMN CONTAINS ONE EQUATION.
C ICO = 1 IF CO(I,J) = 0
C           0 IF CO(I,J) = 0
C EMAT = THE ERROR MEASURES FROM THE REGRESSIONS ON PRD
```

```
C       IN EACH LAYER.
C XMEAN = VECTOR OF MEANS OF THE VARIABLES
C SD = VECTOR OF STANDARD DEVIATIONS OF THE VARIABLES
C CV = VECTOR OF COEFFICIENTS OF VARIATION OF THE VARIABLES
C CORR = VECTOR OF CORRELATIONS AMONG THE VARIABLES
C
C M = THE NUMBER OF INDEPENDENT VARIABLES IN THE ORIGINAL INPUT
C       FILE.
C MS = THE NUMBER OF VARIABLES TO BE TRANSFERRED BETWEEN LAYERS.
C MX = THE MAXIMUM OF M AMD MS.
C N = THE NUMBER OF DATA POINTS PER VARIABLE.
C NTR = THE NUMBER OF TRAINING DATA POINTS.
C NTE = THE NUMBER OF TESTING DATA POINTS.  NTR + NTE = N.
C
C
      %INCLUDE GMDH_COM
  100 CALL UNIT
      CALL RDATA(0)
      CALL INPAR
      CALL QMAP(0)
      CALL REGPAR(IREV,1,0)
      CALL CORE(0)
      CALL EXEQU
      IF(ICON.EQ.1) GO TO 100
      STOP
      END


      SUBROUTINE INPAR
C
C   THE BASIC DATA SIZE PARMETERS ARE ENTERED THROUGH INPAR
C AND OTHER PARAMETERS ARE CALULATED.
C
C NOTATION:
C
C   *C2 = THE NUMBER OF COMBINATIONS OF * THINGS TAKEN 2 AT A TIME
C   *L = * PLUS 1
C   *2 = * MULTIPLIED BY 2
C
      %INCLUDE GMDH_COM
      PRINT,' ENTER:'
      PRINT,' LO    = 0 FOR GMDH '
      PRINT,'         1 FOR STEPWISE REGRESSION'
      PRINT,' NTR   = NUMBER OF ESTIMATION DATA POINTS'
      PRINT,' JEM   = 0 FOR THE MEAN SQUARED RESIDUAL ERROR MEASURE'
      PRINT,'         1 FOR THE R SQUARED ERROR MEASURE'
```

```
      PRINT,' IADJ = 0 TO ADJUST THE ERROR MEASURE'
      PRINT
      PRINT,' 4I FORMAT'
      READ,LO,NTR,JEM,IADJ
      IF(LO.NE.0) GO TO 100
      PRINT,' ENTER:'
      PRINT,' MS     = NUMBER OF VARIABLES PASSED BETWEEN LAYERS'
      PRINT,'          (3 ≤ MS ≤ ',MC2,')'
      PRINT,' NLAY = NUMBER OF LAYERS'
      PRINT,'          (NLAY ≤ ',ID3,')'
      PRINT
      PRINT,' 2I FORMAT'
      READ,MS,NLAY
      GO TO 110
  100 MS = 0
      NLAY = 1
  110 MX = MAX(M,MS)
      NTE = N-NTR
      ML = M+1
      M2 = M*2
      M2L = M2+1
C     MC2 = M*(M-1)/2 WAS CALCULATED IN RDATA
      MC2L = MC2+1
      MSL = MS+1
      MS2 = MS*2
      MS2L = MS2+1
      MSC2 = MS*(MS-1)/2
      MSC2L = MSC2+1
      MXL = MX+1
      MX2 = MX*2
      MX2L = MX2+1
      MXC2 = MX*(MX-1)/2
      MM = MXL*(MXL+1)/2
      LM = MAX(MXL,16)
      NTRL = NTR+1
      NLAY2 = NLAY*2
      NLAYL = NLAY + 1
      MXL2 = MXL*2
      RETURN
      END


      SUBROUTINE UNIT
      %INCLUDE GMDH_COM
      PRINT,' ENTER:'
      PRINT,' IREAD    = INPUT FILE NUMBER'
```

```
      PRINT,' IWRIT     = OUTPUT FILE NUMBER'
      PRINT,' FF        = 0 FOR A CHARACTER FILE'
      PRINT,'             1 FOR AN UNFORMATTED FILE'
      PRINT,' TITLE(1) = DATA FILE NAME'
      PRINT
      PRINT,' 3I,/,A50 FORMAT'
      READ,IREAD,IWRIT,FF
      READ 900,TITLE(1)
  900 FORMAT(A50)
      LA = 0
      RETURN
      END


      SUBROUTINE RDATA(IND)
C
C INPUT:
C
C IND = 0 WHEN INTIALLY ENTERING DATA
C       1 TO REPEAT THE PROCESS, RETAINING THE INITIAL PARAMETERS
C
      %INCLUDE GMDH_COM
      IF(IND.EQ.1) GO TO 215
      PRINT,' ENTER:'
      PRINT,' M    = NUMBER OF INDEPENDENT VARIABLES'
      PRINT,' N    = NUMBER OF DATA POINTS PER VARIABLE'
      PRINT,' IEX = 1 TO USE A SUBSET OF THE VARIABLES'
      PRINT
      PRINT,' 3I FORMAT'
      READ,M,N,IEX
      IEXDD = IEX
      ML = M + 1
      MDCL = ML
  215 REWIND IREAD
      DO 310 J=1,ID5
      DO 310 I=1,ID2
  310 PRD(I,J) = 0.
      DO 201 J=1,MDCL
      IF(FF.EQ.0) READ(IREAD,225) (PRD(I,J),I=1,N)
      IF(FF.EQ.1) READ(IREAD) (PRD(I,J),I=1,N)
  201 CONTINUE
  225 FORMAT(V)
      IF(IEXDD.NE.1) GO TO 440
      IF(IND.EQ.1) GO TO 230
  207 PRINT,' ENTER:'
      PRINT,' NV     = NUMBER OF INDEPENDENT VARIABLES TO BE RETAINED'
```

```
      PRINT,' IEXIX = 1 TO SPECIFY THE VARIABLES TO BE RETAINED'
      PRINT
      PRINT,' 2I FORMAT'
      READ,NV,IEXIX
      IF(NV.LE.M) GO TO 205
      PRINT,'      YOU ASKED FOR MORE VARIABLES THAN YOU ENTERED INTO'
      PRINT,'       THE DATA FILE.  PLEASE TRY AGAIN'
      GO TO 207
  205 DO 220 J=1,M
  220 ISV(J) = 0
      MISV = NV
      IF(IEXIX.NE.1) MISV = M - NV
      M = NV
      ML = M + 1
      PRINT,' ENTER:'
      PRINT,' ISV = NUMBERS OF THE VARIABLES TO BE IN OR EX CLUDED'
      PRINT
      PRINT,MISV,' I FORMAT'
      READ,(ISV(J),J=1,MISV)
  230 IF(IEXIX.NE.1) GO TO 835
      DO 810 J=1,M
      DO 810 I=1,N
  810 WSM(I,J) = PRD(I,ISV(J))
      GO TO 850
  835 ICNT4 = 1
      ICNT3 = 0
      DO 840 J=1,MDCL-1
      IF(J.EQ.ISV(ICNT4)) GO TO 805
      ICNT3 = ICNT3 + 1
      DO 860 I=1,N
  860 WSM(I,ICNT3) = PRD(I,J)
      GO TO 840
  805 ICNT4 = ICNT4 + 1
  840 CONTINUE
  850 DO 820 I=1,N
  820 WSM(I,ML) = PRD(I,MDCL)
      DO 830 J=1,ML
      DO 830 I=1,N
  830 PRD(I,J) = WSM(I,J)
  440 MC2 = M*(M-1)/2
      IF(LA.NE.0) RETURN
      MGL = ML
      PRINT,' ENTER:'
      PRINT,' IMSD  = 1 TO PRINT THE MEANS AND STANDARD DEVIATIONS'
      PRINT,' ICORR = 1 TO PRINT THE CORRELATION MATRIX'
      PRINT
```

```
      PRINT,' 2I FORMAT'
      READ,IMSD,ICORR
      IF(IMSD.EQ.1.OR.ICORR.EQ.1) CALL MSDCORR(IMSD.ICORR.1.MGL.0)
  450 PRINT,' ENTER:'
      PRINT,' IC1,IC2 TO PRINT VARIABLES IC1 THROUGH IC2'
      PRINT
      PRINT,' 2I FORMAT'
      READ,IC1,IC2
      IF(IC1.EQ.0.OR.IC2.EQ.0) GO TO 460
      CALL PRDATA(0,1,0,0,0,MQ1,MQ2,IC1,IC2,IR1,IR2)
      GO TO 450
  460 PRINT,' ENTER:'
      PRINT,' IEX = 1 TO USE A SUBSET OF THE VARIABLES'
      PRINT
      PRINT,' 1I FORMAT'
      READ,IEX
      IF(IEX.EQ.1) IEXDD = IEX
      IF(IEX.EQ.1) GO TO 207
      RETURN
      END


      SUBROUTINE QMAP(IMAP)
C
C    A MAP OF ALL COMBINATIONS OF MX THINGS TAKEN 2 AT A TIME.
C     MX*(MX-1)/2 ENTRIES, IS CREATED.  MAP(I,1) AND MAP(I,2)
C     ARE THE FIRST AND SECOND ELEMENTS OF THE ITH COMBINATION.
C     THE COMBINATIONS ARE ORDERED BY ROWS OF A LOWER TRIANGULAR
C     MATRIX, EXCLUDING THE DIAGONAL.
C
C MATRIX MAP IS PRINTED WHEN IMAP = 1
C
      %INCLUDE GMDH_COM
      IF(IMAP.EQ.1) GO TO 220
      DO 710 I=2,MX
      J = (I-1)*(I-2)/2
      DO 710 K=J+1,J+I-1
      MAP(K,1) = I
  710 MAP(K,2) = K-J
  220 IF(IMAP.NE.1) RETURN
      WRITE(IWRIT,200)
  200 FORMAT(/,T40,'MAP OF QUADRATIC FORMS',/,T5,'NUMBER',T15,
     &'ELEMENT 1',T25,'ELEMENT 2',T40,'NUMBER',T50,'ELEMENT 1',
     &T60,'ELEMENT 2',T75,'NUMBER',T85,'ELEMENT 1',T95,
     &'ELEMENT 2')
      WRITE(IWRIT,210)(I,MAP(I,1),MAP(I,2),I=1,MXC2)
```

```
  210 FORMAT((6X,3(I4,7X,I4,6X,I4,10X)))
      RETURN
      END


      SUBROUTINE REGPAR(IREV,IENTER,IPRINT)
C
C THE ALFA AND IJOB PARAMETERS FOR THE IMSL STEPWISE REGRESSION
C SUBROUTINE ARE ENTERED WITH THIS SUBROUTINE
C
C INPUT:
C
C IREV   = 0 IF THE ORIGINAL COEFFICIENTS HAVE BEEN RETAINED
C          1 IF THE COEFFICIENTS HAVE BEEN REESTIMATED
C IENTER = 1 TO ENTER THE REGRESSION PARAMETERS
C IPRINT = 1 TO PRINT THE REGRESSION PARAMETERS
C
      %INCLUDE GMDH_COM
      IF(IENTER.NE.1) GO TO 200
      PRINT,' ENTER:'
      PRINT,' ALFA(1) = SIGNIFICANCE LEVEL FOR ENTERING VARIABLES'
      PRINT,' ALFA(2) = SIGNIFICANCE LEVEL FOR DELETING VARIABLES'
      PRINT,' IJOB(1) = 0 TO NOT PERFORM A LACK OF FIT TEST'
      PRINT,' IJOB(2) = 0 TO PERFORM ONLY AN OVERALL F TEST'
      PRINT
      PRINT,' 2F,2I FORMAT'
      READ,ALFA(1),ALFA(2),IJOB(1),IJOB(2)
  200 IF(IPRINT.NE.1) RETURN
      WRITE(IWRIT,310) ALFA(1),ALFA(2),IJOB(1),IJOB(2)
     &,JEM,IADJ,IREV,NTR,NTE
  310 FORMAT(/,T40,'REGRESSION PARAMETERS'/,T10,'THE SIGNIFICANCE'.
     & ' LEVELS FOR ENTERING AND DELETING VARIABLES. ALFA(1) AND '.
     &'ALFA(2), =', E12.5,3X,'AND',2XE12.5,
     &/,T10,'THE LACK OF FIT TEST PARAMETER, IJOB(1), = '.
     &I1,5X,'THE PARTIAL OR OVERALL F TEST STATISTIC, IJOB(2), = '.
     &I1,/,T10,'JEM =',I3,5X,'IADJ =',I3,5X,'IREV =',I3,5X,'NTR ='.
     &I4,5X,'NTE =',I4)
      RETURN
      END


      SUBROUTINE VFORC
C
C VARIABLES MAY BE FORCED INTO THE MODEL WITH VFORC
C
      %INCLUDE GMDH_COM
```

```
      DO 310 I=1,ID9
  310 IXD(I) = 0
      PRINT,' ENTER:'
      PRINT,' NVF = NUMBER OF FORCED VARIABLES'
      PRINT
      PRINT,' 1I FORMAT'
      READ, NVF
      IF(NVF.EQ.0) RETURN
      PRINT,' ENTER:'
      PRINT,' NUMBERS OF THE FORCED VARIABLES'
      PRINT
      PRINT,NVF,' I FORMAT'
      READ,(VARB(I),I=1,NVF)
      DO 300 I=1,NVF
  300 IXD(IFIX(VARB(I))) = 1
      RETURN
      END


      SUBROUTINE CORE(IREV)
C
C CORE PERFORMS A LAYERED REGRESSION USING QUADRATIC POLYNOMIALS
C OF ALL POSSIBLE COMBINATIONS OF TWO VARIABLES.
C THE BEST MS POLYNOMIALS FROM EACH LAYER ARE TRANSFERRED
C TO THE SUCCEEDING LAYER.
C
C INPUT:
C IREV = 0 TO PERFORM THE INITIAL REGRESSION
C        1 TO REEVALUATE THE COEFFICIENTS USING ALL OF THE DATA
C          WHILE RETAINING THE MODEL STRUCTURE
C
      %INCLUDE GMDH_COM
      NTRT = NTR
      ATEMP = ALFA(1)
      IF(IREV.EQ.0) GO TO 1992
      NTR = N
      ALFA(1) = 0.0000001
 1992 DO 1990 LAY=1,NLAY
      IER36 = 0
      PRINT,'      BEGINNING LAYER',LAY
      LA = LAY
      IF(LAY.GT.1) GO TO 1995
      MG = M
      MGL = ML
      MGC2L = MC2L
      GO TO 2005
```

```
 1995 MG = MS
      MGL = MSL
      MGC2L = MSC2L
 2005 IF(LO.EQ.1) GO TO 2300
      DO 2000 I=2,MG
      II = (I-1)*(I-2)/2
      DO 2000 J=1,I-1
      IJ = II + J
      IF((IJ/50)*50.EQ.IJ) PRINT,'      IJ = '.IJ
C
C I = THE COLUMN NUMBER OF THE FIRST ELEMENT IN QP.
C J = THE COLUMN NUMBER OF THE SECOND ELEMENT IN QP.
C IJ = THE SEQUENCE NUMBER OF THE QUADRATIC FORM AS
C       FOUND IN THE MATRIX MAP.
C
      CALL QUAD( I,J)
      DO 2025 K=1,6
      IXD(K) = 0
      DO 2025 KI=1,N
 2025 WSM(KI,K) = QP(KI,K)
      IF(IREV.EQ.0) GO TO 2320
      DO 2330 K=1,5
 2330 IXD(K) = IQCO(K,IJ,LAY)
 2320 CALL RLSEP(WSM,NTR,5,ID2,ALFA,IJOB,IXD,ANOVA,XYB,ID8,VARB,IER)
      IF(IER.EQ.36) IER36 = IER36 + 1
      IER = 0
      IF(IER.NE.0) CALL PRERRM(IER,IJ,0,'RLSEP      ')
      IF(JEM.EQ.1.AND.NTR.EQ.N) GO TO 2035
      CALL COMPCT(0)
      IF(ANOVA(12).EQ.0) ANOVA(12) = SQRT(ANOVA(5)/ANOVA(2))
      CALL RLRESA(QP,ID2,6,N,IH,ICL,BETA,ANOVA(12),BES,ID2,IER,ID9)
      IF(IER.NE.0) CALL PRERRM(IER,IJ,0,'RLRESA    ')
 2035 DO 2030 K=1,6
      IF(K.EQ.6.OR.IREV.EQ.1) GO TO 2030
      IQCO(K,IJ,LAY) = IXD(K+5)
 2030 QCO(K,IJ,LAY) = XYB(K,2)
      IF(IREV.EQ.1) GO TO 2010
      IQCO(6,IJ,LAY) = 0
      IF(QCO(6,IJ,LAY).NE.0.) IQCO(6,IJ,LAY) = 1
 2010 CALL CEM(IJ,0)
 2000 CONTINUE
C
C    A LINEAR REGRESSION ON ALL THE TERMS IN EACH LAYER
C     IS ALSO PERFORMED.
C
 2300 DO 2047 K=1,MGL
```

```
      IXD(K) = 0
      DO 2047 KI=1,N
 2047 WSM(KI,K) = PRD(KI,K)
      IF(IREV.EQ.0) GO TO 2350
      DO 2360 K=1,MG
 2360 IXD(K) = ICO(K,LAY)
      GO TO 2351
 2350 IF(LO.EQ.1) CALL VFORC
 2351 CALL RLSEP(WSM,NTR,MG,ID2,ALFA,IJOB,IXD,ANOVA,XYB,ID8,VARB,IER)
      IF(IER.NE.0) CALL PRERRM(IER,IJ,1,'RLSEP      ')
      CALL COMPCT(1)
      IF(ANOVA(12).EQ.0) ANOVA(12) = SQRT(ANOVA(5)/ANOVA(2))
      CALL RLRESA(PRD,ID2,MGL,N,IH,ICL,BETA,ANOVA(12),RES,ID2,IER,ID9)
      IF(IER.NE.0) CALL PRERRM(IER,IJ,1,'RLRESA     ')
 2045 DO 2050 K=1,MGL
      IF(K.EQ.MGL.OR.IREV.EQ.1) GO TO 2050
      ICO(K,LAY) = IXD(K+MG)
 2050 CO(K,LAY) = XYB(K,2)
      ICO(MGL,LAY) = 0
      IF(CO(MGL,LAY).NE.0.) ICO(MGL,LAY) = 1
      CALL CEM(IJ,1)
      IF(MG.EQ.MX) GO TO 2070
      IF(LO.EQ.1) GO TO 2310
C
C   WHEN MS DOES NOT EQUAL M THE EXTRA SPACES IN THE
C    COEFFICIENT AND SUMS OF SQUARES MATRICES ARE SET TO
C    DEFAULT VALUES.
C
      DO 2080 I=MGC2L,MXC2
      DO 2065 K=1,6
 2065 QCO(K,I,LAY) = 0.
      EM(I,LAY) = 0.
      IF(JEM.EQ.0) EM(I,LAY) = 9999999.
 2080 CONTINUE
 2310 DO 2085 I=MGL+1,MX
 2085 CO(I,LAY) = 0.
 2070 IF(LO.NE.1) GO TO 2072
      NTR = NTRT
      ALFA(1) = ATEMP
      RETURN
 2072 CALL CSORT(LAY,LAY,MS,IREV,JEM)
      IF(IER36.NE.0) PRINT,'     IER36 =',IER36
      IF(LAY.EQ.NLAY) GO TO 1990
      CALL CLAY(0,0)
 1990 CONTINUE
      NTR = NTRT
```

```
      ALFA(1) = ATEMP
      RETURN
      END


      SUBROUTINE QUAD(MQ1,MQ2)
C
C   A QUADRATIC FORM FROM COLUMNS MQ1 AND MQ2 IN THE ORDER
C    MQ1, MQ2, MQ1*MQ1, MQ2*MQ2, MQ1*MQ2, IS CREATED AND RETURNED
C    IN QP.
C
C INPUT:
C MQ1 = FIRST VARIABLE IN THE QUADRATIC FORM.
C MQ2 = SECOND VARIABLE IN THE QUADRATIC FORM.
C
      %INCLUDE GMDH_COM
      DO 720 I=1,N
      QP(I,1) = PRD(I,MQ1)
      QP(I,2) = PRD(I,MQ2)
      QP(I,3) = PRD(I,MQ1)*PRD(I,MQ1)
      QP(I,4) = PRD(I,MQ2)*PRD(I,MQ2)
      QP(I,5) = PRD(I,MQ1)*PRD(I,MQ2)
  720 QP(I,6) = PRD(I,MGL)
      RETURN
      END


      SUBROUTINE COMPCT(IND)
C
C   A SUMMARY MATRIX TXYB WHICH INCLUDES ONLY THE VARIABLES
C    CHOSEN BY THE STEPWISE REGRESSION IS EXTRACTED FROM THE
C    MATRIX XYB.  BETA AND IH ARE CREATED FOR INPUT TO
C    SUBROUTINE RLRESA.
C
C INPUT:
C IND = 0 FOR QUADRATIC FORMS (QP).
C       1 FOR THE REGRESSIONS ON ALL THE VARIABLES IN A LAYER.
C
C OUTPUT:
C IC = THE NUMBER OF VARIABLES CHOSEN BY THE STEPWISE
C     REGRESSION.
C ICL = IC+1.
C
      %INCLUDE GMDH_COM
      IC = 0
      IF(IND.EQ.0) GO TO 1330
```

```
      IF(LA.LE.1) GO TO 1335
      MA = MSL
      MB = MS2
      MC = MS
      GO TO 1340
 1335 MA = ML
      MB = M2
      MC = M
      GO TO 1340
 1330 MA = 6
      MB = 10
      MC = 5
 1340 DO 1300 I=MA,MB
      IF(IXD(I).EQ.0) GO TO 1300
      IC = IC+1
      IH(IC) = I-MC
      DO 1310 J=1,5
 1310 TXYB(IC,J) = XYB(I-MC,J)
 1300 CONTINUE
      ICL = IC+1
      IH(ICL) = MA
      TXYB(ICL,1) = XYB(MA,1)
      TXYB(ICL,2) = XYB(MA,2)
      DO 1320 I=1,ICL
 1320 BETA(I) = TXYB(I,2)
      IF(ICL.EQ.MA) GO TO 1370
      IF(ICL.GT.MA) PRINT,'AN ERROR HAS OCCURRED IN SUBROUTINE COMPCT'
      DO 1360 I=(ICL+1),MA
      BETA(I) = 0.
 1360 IH(I) = 0
 1370 RETURN
      END


      SUBROUTINE CSORT(MA,MB,NO,IXL,IXO)
C
C THE NO SMALLEST (IF IXO = 0) OR LARGEST (IF IXO = 1)
C ELEMENTS IN EACH OF THE COLUMNS MA
C THROUGH MB OF EM ARE CHOSEN AND ARRANGED IN ASCENDING
C (IF IXO = 0) OR DESCENDING (IF IXO = 1) ORDER IN
C THE FIRST NO ROWS OF THOSE COLUMNS, FOR IXL = 0.
C IF IXL = 1, EM IS SORTED ACCORDING TO THE INPUT IEMLOC.
C
C INPUT:
C MA = THE FIRST COLUMN TO BE SORTED.
C MB = THE LAST COLUMN TO BE SORTED.
```

```
C NO = THE NUMBER OF ELEMENTS SORTED FROM EACH COLUMN.
C IXL = 0 TO SORT IEMLOC ACCORDING TO EM
C       1 TO SORT EM ACCORDING TO IEMLOC
C IXO = 0 TO SORT IN ASCENDING ORDER
C       1 TO SORT IN DESCENDING ORDER
C
C OUTPUT:
C EM IS RETURNED SORTED.
C IEMLOC.  A I IN THE JTH (J<NO) ROW OF COLUMNS MA
C          THROUGH MB INDICATES THAT ELEMENT I IN THE
C          INPUT MATRIX EM IS THE JTH SMALLEST
C          ELEMENT IN THAT COLUMN OF EM.  THAT IS, THE
C          ELEMENT WHICH WAS IN THE ITH ROW OF EM WHEN
C          INPUT SHOULD BE IN THE JTH ROW OF EM
C          WHEN OUTPUT.
C
      %INCLUDE GMDH_COM
      IF(IXL.EQ.1) GO TO 665
      DO 660 J=MA,MB
      DO 660 I=1,MXC2
  660 IEMLOC(I,J) = I
      DO 670 J=MA,MB
      DO 680 I=1,NO
      ICOUNT = 0
      NI = MXC2-I
      DO 690 K=1,NI
      KM = MXC2+1-K
      IF(IXO.EQ.1) GO TO 730
      IF(EM(KM,J).GE.EM(KM-1,J)) GO TO 690
      GO TO 740
  730 IF(EM(KM,J).LE.EM((KM-1),J)) GO TO 690
  740 ICOUNT = ICOUNT+1
      TEMP = EM(KM,J)
      EM(KM,J) = EM(KM-1,J)
      EM(KM-1,J) = TEMP
      TEMP = IEMLOC(KM,J)
      IEMLOC(KM,J) = IEMLOC(KM-1,J)
      IEMLOC(KM-1,J) = TEMP
  690 CONTINUE
      IF(ICOUNT.EQ.0) GO TO 670
  680 CONTINUE
  670 CONTINUE
      RETURN
  665 DO 710 J=MA,MB
      DO 700 I=1,NO
```

```
      K = IEMLOC(I,J)
  700 WSM(I,1) = EM(K,J)
      DO 710 I=1,NO
  710 EM(I,J) = WSM(I,1)
      RETURN
      END


      SUBROUTINE CLAY(IND1,IND2)
C
C A NEW PRD MATRIX IS CREATED FROM THE MS BEST PREDICTORS
C IN THE INPUT LAYER.  CLAY SHOULD BE CALLED ONLY AFTER
C THE LAYER HAS BEEN SORTED (SEE SUBROUTINE CSORT).
C
C IND1 = 0 TO REPLACE THE ENTIRE MATRIX PRD
C      = 1 TO REPLACE ONLY SELECTED COLUMNS
C IND2 = THE NUMBER OF COLUMNS TO BE REPLACED
C
C
C NOTE:WHEN IND1 = 1 THE MATRIX RES MUST CONTAIN THE LIST OF
C      COLUMNS TO BE REPLACED (SEE SUBROUTINE RDE)
C
      %INCLUDE GMDH_COM
      MT = MS
      IF(IND1.EQ.1) MT = IND2
      DO 900 J=1,MT
      J1 = J
      K = IEMLOC(J,LA)
      IF(IND1.EQ.0) GO TO 1000
      K = IFIX(RES(J,LA))
      IF(J.EQ.1) GO TO 995
      DO 990 I=1,J-1
      IF(IFIX(RES(I,LA)).EQ.K) GO TO 900
  990 CONTINUE
  995 DO 980 I=1,MS
      IF(IEMLOC(I,LA).EQ.K) J1 = I
  980 CONTINUE
 1000 IA = MAP(K,1)
      IB = MAP(K,2)
      DO 900 I=1,N
      WSM(I,J1) = QCO(1,K,LA)*PRD(I,IA) + QCO(2,K,LA)*PRD(I,IB) +
     & QCO(3,K,LA)*PRD(I,IA)*PRD(I,IA) + QCO(4,K,LA)*PRD(I,IB)
     &*PRD(I,IB) + QCO(5,K,LA)*PRD(I,IA)*PRD(I,IB) + QCO(6,K,LA)
  900 CONTINUE
      DO 920 I=1,N
  920 WSM(I,MSL) = PRD(I,MGL)
```

```
      DO 930 J=1,MSL
      DO 930 I=1,N
  930 PRD(I,J) = WSM(I,J)
C
C WHEN THE INPUT PRD MATRIX IS LARGER THAN THE OUTPUT PRD
C MATRIX THE EXTRA POSITIONS ARE SET EQUAL TO 0.
C
      IF(MSL.GE.MGL) GO TO 970
      DO 960 J=MSL+1,MGL
      DO 960 I=1,N
  960 PRD(I,J) = 0.
  970 RETURN
      END


      SUBROUTINE CEM(IJ,IND)
C
C INPUT:
C IJ = QUADRATIC FORM NUMBER WITHIN LAYER
C IND = 0 FOR QUADRATIC FORMS
C       1 FOR REGRESSIONS ON ALL THE VARIABLES IN A LAYER
C
C OUTPUT:
C RETURNS THE ERROR MEASURE FOR A REGRESSION.
C
C NOTE THAT WHEN A STATISTIC IS CALCULATED ON
C INDEPENDENT DATA IT IS NOT ADJUSTED FOR THE DEGREES
C OF FREEDOM IN THE COEFFICIENT ESTIMATION.
C
      %INCLUDE GMDH_COM
      VEQ = 0.
      XNTE = NTE
      XN = N
C
C KDC = DEGREES OF CONSTRAINT
C KDF = DEGREES OF FREEDOM
C
      KDC = 0
      IF(IND.EQ.1) GO TO 820
      DO 800 KL=1,6
  800 KDC = KDC + IQCO(KL,IJ,LA)
      GO TO 830
  820 DO 810 KL=1,MGL
  810 KDC = KDC + ICO(KL,LA)
  830 KDF = N - KDC
      IF(NTR.NE.N) KDF = NTE - 1
```

```
      DF = KDF
      DC = KDC
      SSRA = 0.
C
C CALCULATE THE SUM OF SQUARED RESIDUALS
C
      DO 710 K=NTRL,N
  710 SSRA = SSRA + RES(K,3)*RES(K,3)
      IF(JEM.NE.0) GO TO 700
C
C CALCULATE THE MEAN SQUARED RESIDUAL
C
      IF(NTR.EQ.N) GO TO 760
      VEQ = SSRA/DF
      GO TO 748
  760 VEQ = ANOVA(12)*ANOVA(12)
  715 IF(IADJ.EQ.0) VEQ = VEQ*DC
      GO TO 748
  700 IF(JEM.NE.1) GO TO 750
C
C CALULATE THE REDUCTION OF VARIANCE
C
      IF(NTR.LT.N) GO TO 745
      VEQ = 100.-((XN-1.)/DF)*(100.-ANOVA(11))
      IF(IADJ.NE.0) VEQ = ANOVA(11)
      GO TO 748
  745 SRA = 0.
      TV = 0.
      SUM = 0.
      DO 740 K=NTRL,N
      SRA = SRA + RES(K,3)
      SUM = SUM + RES(K,1)
  740 TV = TV + RES(K,1)*RES(K,1)
      YRS = SSRA - SRA*SRA/XNTE
      YBS = TV - SUM*SUM/XNTE
      VEQ = 100.*(1.-(YRS/YBS))
  748 IF(IND.EQ.0) EM(IJ,LA) = VEQ
      IF(IND.EQ.1) EMAT(LA) = VEQ
  750 RETURN
      END


      SUBROUTINE EXEQU
C
C EQUATIONS CAN BE PRINTED
C OR EXAMINED GRAPHICALLY.
```

```
C
      %INCLUDE GMDH_COM
 2484 PRINT,' ENTER:'
      PRINT,' IPS =  0 TO EXAMINE ONLY INDIVIDUAL EQUATIONS'
      PRINT,'        1 TO PRINT A SUMMARY OF THE EQUATION EVALUATION'
      PRINT,'        2 TO PRINT A SUMMARY OF EQUATION COEFFICIENTS'
      PRINT,' IREV = 0 TO RETAIN COEFFICIENT ESTIMATES'
      PRINT,'        1 TO REESTIMATE COEFFICIENTS'
      PRINT
      PRINT,' 2I FORMAT'
      READ,IPS,IREV
      IF(IREV.EQ.0.OR.IPS.EQ.0) GO TO 2485
      CALL RDATA(1)
      CALL CORE(IREV)
 2485 WRITE(IWRIT,2480) TITLE(1)
 2480 FORMAT(/,T20,'REGRESSION ON DATA FILE '.A50./)
      LA = NLAY
      CALL REGPAR(IREV,0,1)
      IF(IPS.EQ.0) GO TO 2490
      II1 = 1
      IF(IPS.EQ.1) II1 = 0
      IF(LO.EQ.0) CALL PRCOEFF(II1,0,II1,0,1,0,1)
      IF(LO.EQ.1) CALL PRCOEFF(0,0,II1,0,0,0,1)
      IF(IWRIT.EQ.6) GO TO 2490
      IWTEMP = IWRIT
      IWRIT = 6
      IF(LO.EQ.0) CALL PRCOEFF(0,0,0,0,1,0,1)
      IF(LO.EQ.1) CALL PRCOEFF(0,0,0,0,0,0,1)
      IWRIT = IWTEMP
 2490 PRINT,' ENTER:'
      PRINT,' ICON = 0 TO STOP PROGRAM'
      PRINT,'        1 TO RESTART PROGRAM'
      PRINT,'        2 TO EXAMINE AN EQUATION'
      PRINT,'        3 TO RESTART SUBROUTINE EXEQU'
      PRINT,' IREV = 0 TO RETAIN COEFFICIENTS'
      PRINT,'        1 TO REESTIMATE COEFFICIENTS'
      PRINT
      PRINT,' 2I FORMAT'
      READ,ICON,IREV
      IF(ICON.EQ.3) GO TO 2484
      IF(ICON.EQ.0.OR.ICON.EQ.1) RETURN
      PRINT,' ENTER:'
      PRINT,' IND2 = 0 FOR A QUADRATIC FORM'
      PRINT,'        1 FOR A LINEAR FORM'
      PRINT,' IEQU = EQUATION NUMBER'
      PRINT,' LA   = LAYER NUMBER'
```

```
      PRINT,' IPR  = 1 TO PRINT THE EQUATION'
      PRINT,' IPA  = 1 TO PRINT THE ANOVA TABLE'
      PRINT,' IPL  = 1 TO VIEW GRAPHICS'
      PRINT
      PRINT,' 6I FORMAT'
      READ,IND2,IEQU,LA,IPR,IPA,IPL
      IF(IPR.NE.1.AND.IPL.NE.1.AND.IPA.NE.1) GO TO 2490
C
C THESE LABELS ARE USED IN SUBROUTINE PLOT1
C
      REWIND 26
      IF(IND2.EQ.0) WRITE(26,2724) IEQU,LA
      IF(IND2.EQ.1) WRITE(26,2725) LA
      REWIND 26
      READ(26,261) TITLE(2)
  261 FORMAT(A50)
      WRITE(IWRIT,2720) TITLE(2)
 2724 FORMAT(1X,'VARIABLE ',I4,'      LAYER'.I2.
     &'                    ')
 2725 FORMAT(1X,'REGRESSION ON INPUT VARIABLES FOR LAYER'.I2.
     &'          ')
 2720 FORMAT(/,T40,A50)
      IMRK = 1
      IF(IPA.NE.1.AND.IREV.EQ.0) GO TO 2410
      IF(LO.EQ.1.AND.IREV.EQ.0) GO TO 2410
      IMRK = 0
      CALL RDATA(1)
      ATEMP = ALFA(1)
      ALFA(1) = 0.0000001
      CALL RDE(IEQU,IREV,IND2,0)
      ALFA(1) = ATEMP
 2410 CALL REGPAR(IREV,0,1)
      IF(IPR.EQ.1) CALL PREQU(0,0,1,1,IEQU,1,IND2)
      IF(IPA.EQ.1) CALL PRRSUM(0,0,1,0,1,1,0.0,0,IER,IND2,IEQU)
      IF(IMRK.EQ.0.AND.IREV.EQ.1) CALL RDE(IEQU,IREV,IND2,1)
      IF(IPL.NE.1) GO TO 2490
 2675 PRINT,' ENTER:'
      PRINT,' ICTRL4 = 0 TO CONTINUE PROGRAM'
      PRINT,'          1 TO PRINT LIST OF PLOTS'
      PRINT,'          2-9 TO IDENTIFY A PLOT'
      PRINT,' ICTRL3 = 0 FOR AUTOMATIC PLOTTING'
      PRINT,'          1 TO CONTROL PLOT FORMAT'
      PRINT,'          2 TO STORE VECTORS'
      PRINT
      PRINT,' 2I FORMAT'
      READ,ICTRL4,ICTRL3
```

```
      IF(ICTRL4.NE.1) GO TO 100
      PRINT,' ENTER:'
      PRINT,' ICTRL4 = 2, STAND. RES. VS. OBS. NUM.'
      PRINT,'          3,  ORIG. RES. VS. OBS. NUM.'
      PRINT,'          4, STAND. RES. VS. PRED. RESP.'
      PRINT,'          5,  ORIG. RES. VS. PRED. RESP.'
      PRINT,'          6, STAND. RES. VS. PREDICTOR'
      PRINT,'          7,  ORIG. RES. VS. PREDICTOR'
      PRINT,'          8,  OBS. RESP. VS. OBS. NUM.'
      PRINT,'          9, PREDICTOR   VS. OBS. NUM.'
      PRINT
      PRINT,' 1I FORMAT'
      READ,ICTRL4
  100 IF(ICTRL4.EQ.0) GO TO 2490
      IF(ICTRL4.NE.2) GO TO 2560
      DO 2540 J=1,N
      A2(J) = J
 2540 B2(J) = RES(J,4)
 2560 IF(ICTRL4.NE.3) GO TO 2580
      DO 2570 J=1,N
      A2(J) = J
 2570 B2(J) = RES(J,3)
 2580 IF(ICTRL4.NE.4) GO TO 2600
      DO 2590 J=1,N
      A2(J) = RES(J,2)
 2590 B2(J) = RES(J,4)
 2600 IF(ICTRL4.NE.5) GO TO 2620
      DO 2610 J=1,N
      A2(J) = RES(J,2)
 2610 B2(J) = RES(J,3)
 2620 IF(ICTRL4.NE.6.AND.ICTRL4.NE.7) GO TO 2640
      PRINT,' ENTER: IP = PREDICTOR NUMBER'
      PRINT
      PRINT,' 1I FORMAT'
      READ,IP
      DO 2710 J=1,N
      A2(J) = PRD(J,IP)
      IF(ICTRL4.EQ.6) B2(J) = RES(J,4)
 2710 IF(ICTRL4.EQ.7) B2(J) = RES(J,3)
 2640 IF(ICTRL4.NE.8) GO TO 2655
      DO 2650 J=1,N
      A2(J) = J
 2650 B2(J) = RES(J,1)
 2655 IF(ICTRL4.NE.9) GO TO 2660
      PRINT,' ENTER: IP = PREDICTOR NUMBER'
      PRINT
```

```
      PRINT,' 1I FORMAT'
      READ,IP
      DO 2690 J=1,N
      A2(J) = J
 2690 B2(J) = PRD(J,IP)
 2660 NCURVE = 1
      IF(ICTRL3.EQ.2) CALL STORVEC(ICTRL4,0)
      IF(ICTRL3.EQ.2) GO TO 2663
      ICTRL(1,1) = N
      CALL PLOT1(ICTRL3,ICTRL4)
 2663 GO TO 2675
      END


      SUBROUTINE MSDCORR(IMSD,ICORR,L1,L2,IWAL)
C
C INPUT:
C IMSD  = 1 TO PRINT THE MEAN, SD, AND C OF V FOR EACH VARIABLE
C ICORR = 1 TO PRINT THE CORRELATION MATRIX
C L1,L2 = THE FIRST AND LAST VARIABLES FROM PRD TO BE USED
C IWAL  = 1 IF THE DATA IS IN PRD
C         0 IF THE DATA IS IN WSM
C
      %INCLUDE GMDH_COM
      NV = L2 - L1 + 1
      IF(IWAL.NE.0) GO TO 305
      ICNT = 0
      DO 300 K=L1,L2
      ICNT = ICNT + 1
      DO 300 KI=1,N
  300 WSM(KI,ICNT) = PRD(KI,K)
  305 CALL BECORI(WSM,N,NV,ID2,XMEAN,SD,CORR,IER)
      DO 310 J=1,NV
  310 CV(J) = (SD(J)/XMEAN(J))*100.
      IF(IMSD.NE.1) GO TO 320
      WRITE(IWRIT,330) TITLE(1),(J+L1-1,XMEAN(J),SD(J),CV(J),J=1,NV)
  330 FORMAT(/,T30,'STATISTICS OF THE DATA SET ',A50,/,
     &T10,'VARIABLE',T20,'MEAN',T35,'STANDARD',T50,
     &'COEFFICIENT',T70,'VARIABLE',T80,'MEAN',T95,'STANDARD',
     &T110,'COEFFICIENT',/,T10,'NUMBER',T35,'DEVIATION',T50,
     &'OF VARIATION',T70,'NUMBER',T95,'DEVIATION',T110,
     &'OF VARIATION',/,(T10,I4,4X,3(E12.5,3X),T70,I4,4X,3(E12.5,3X)))
  320 IF(ICORR.NE.1) RETURN
      WRITE(IWRIT,360) TITLE(1)
  360 FORMAT(1X,/,T30,'CORRELATION MATRIX OF THE INPUT AND RESPONSE',
     & ' VARIABLES OF DATA SET ',A50)
```

```
      K = 0
      DO 350 I=1,NV
      K = K + I
      L= K - I + 1
  350 WRITE(IWRIT,340) I+L1-1,(CORR(J),J=L,K)
  340 FORMAT(1X,I3,2X,(15(1X,F7.4)))
      RETURN
      END


      SUBROUTINE RLRESA  (XY,IX,MM,N,IH,M,BETA,SDR,RES,IR,IER,ID9)
C
C THIS SUBROUTINE IS BASED ON IMSL SUBROUTINE RLRES
C
C
      DIMENSION            XY(IX,MM),IH(ID9),BETA(M),RES(IR,4)
      REAL               BETA,SDR,XY,RES
      DOUBLE PRECISION   STAT
      IER = 0
      IF(M.LE.MM.AND.M.GE.1) GO TO 5
      IER = 130
      RETURN
C  TERMINAL ERROR 2, MISSPECIFIED PARAMETERS
    5 DO 10 I=1,MM
   10 IH(M+I) = I
      DO 25 I = 1,M
         DO 15 J=I,MM
            JJ = J
            IF(IH(M+J).EQ.IH(I)) GO TO 20
   15    CONTINUE
         IER = 129
      RETURN
C  TERMINAL ERROR 1, NO TERMS IN EQUATION
   20    ITEMP = IH(M+I)
         IH(M+I) = IH(M+JJ)
         IH(M+JJ) = ITEMP
   25 CONTINUE
      L = IH(M)
      DO 35 I=1,N
         STAT = BETA(M)
         RES(I,1) = XY(I,L)
         M1 = M-1
      IF(M1.EQ.0) GO TO 40
         DO 30 J=1,M1
            K = IH(J)
            STAT = STAT + DBLE(BETA(J))*DBLE(XY(I,K))
```

```
   30    CONTINUE
   40    RES(I,2) = STAT
         RES(I,3) = RES(I,1) - STAT
         RES(I,4) = RES(I,3)/SDR
   35 CONTINUE
      RETURN
      END


      SUBROUTINE RDE(IEQU,IREV,IND2,IND3)
C
C THE REQUESTED EQUATION IS RECREATED
C
C INPUT:
C
C IEQU = THE SORTED POSITION OF AN EQUATION WITHIN A LAYER
C IREV = 0 TO RETAIN THE OLD COEFFICIENTS
C        1 TO REESTIMATE THE COEFFICIENTS BASED ON ALL THE DATA
C IND2 = 0 FOR QUADRATIC FORMS
C        1 FOR REGRESSION ON ALL THE VARIABLES IN A LAYER
C IND3 = 0 TO RECREATE THE REQUESTED EQUATION
C        1 TO RESTORE THE ORIGINAL COEFFICIENTS
C
      %INCLUDE GMDH_COM
      NTRT = NTR
      LAT = LA
      ICNT = 0
      MGL = ML
      MG = M
      IF(LA.LE.1) GO TO 230
      IF(IREV.EQ.0.AND.IND3.EQ.1) GO TO 230
      IF(IC.LE.IFIX(ID2/16.)) GO TO 231
      IF(IWRIT.NE.6) WRITE(6,232)
      WRITE(IWRIT,232)
  232 FORMAT(T5,'THE REQUESTED EQUATION COULD NOT BE RECOVERED'.
     & ' BECAUSE IT HAD TOO MANY TERMS',/,T5.'AND THE FOLLOWING OUTPUT'.
     & ' IS PROBABLY INCORRECT. THE MAXIMUM NUMBER OF TERMS'.
     & ' IS ID2/16.')
C
C THIS ROUTINE WILL HANDLE UP TO ID2/16 TERMS
C
  231 IF(IND2.EQ.1) LA = LAT - 1
      ICNT3 = 0
      KJ = IEQU
      IF(IND2.EQ.0) GO TO 2082
      DO 2090 KJ = 1,MS
```

```
      IF(ICO(KJ,LAT).EQ.0) GO TO 2090
 2082 ICNT3 = ICNT3 + 1
      CALL DIAKA(KJ,IV)
      IVT = IV
      DO 2080 MJ1=1,LAT-1
      DO 2085 MJ2=1,IV
 2085 RES(MJ2+(ICNT3-1)*IV,MJ1) = KACT(MJ2+IV-1)
 2080 IV = IV/2
      IF(IND2.EQ.0) GO TO 2070
 2090 CONTINUE
 2070 IV = IVT * ICNT3
      DO 105 LA=1,LAT-1
      IF(LA.GE.2) MGL = MSL
      MG = MGL - 1
      IF(IREV.EQ.0) GO TO 175
      DO 110 J=1,IV
      IF(J.EQ.1) GO TO 180
      DO 190 JJ=1,J-1
      IF(IFIX(RES(J,LA)).EQ.IFIX(RES(JJ,LA))) GO TO 110
  190 CONTINUE
  180 IF(IND3.EQ.1) GO TO 185
      CALL QUAD(MAP(IFIX(RES(J,LA)),1),MAP(IFIX(RES(J,LA)),2))
C
C     THE ORIGINAL COEFFICIENTS ARE STORED IN CORR
C
  185 DO 120 K=1,6
      IF(IND3.EQ.1) QCO(K,IFIX(RES(J,LA)),LA) = CORR(ICNT+K)
      IF(IND3.EQ.1) GO TO 120
      CORR(ICNT+K) = QCO(K,IFIX(RES(J,LA)),LA)
      IXD(K) = IQCO(K,IFIX(RES(J,LA)),LA)
      DO 120 KI=1,N
      WSM(KI,K) = QP(KI,K)
  120 CONTINUE
      ICNT = ICNT + 6
      IF(IND3.EQ.1) GO TO 110
      CALL RLSEP(WSM,N,5,ID2,ALFA,IJOB,IXD,ANOVA,XYB,ID8,VARB,IER)
      IF(IER.NE.0) CALL PRERRM(IER,IFIX(RES(J,LA)),0,'RLSEP      ')
      DO 110 K=1,6
      QCO(K,IFIX(RES(J,LA)),LA) = XYB(K,2)
  110 CONTINUE
  175 CALL CLAY(0,0)
  105 IV = IV/2
  230 LA = LAT
      IF(LA.GE.2) MGL = MSL
      MG = MGL - 1
      IF(IREV.EQ.0) GO TO 240
```

```
      IF(IND2.EQ.1) GO TO 220
      DO 150 K=1,6
      IF(IND3.EQ.1) QCO(K,IEMLOC(IEQU,LA),LA) = CORR(ICNT+K)
      IF(IND3.NE.1) CORR(ICNT+K) = QCO(K,IEMLOC(IEQU,LA),LA)
  150 CONTINUE
      IF(IND2.EQ.0) GO TO 240
  220 DO 160 K=1,MGL
      IF(IND3.EQ.1) CO(K,LA) = CORR(ICNT+K)
      IF(IND3.NE.1) CORR(ICNT+K) = CO(K,LA)
  160 CONTINUE
  240 NTR = NTRT
      IF(IND3.EQ.1) RETURN
C
C THE FINAL LAYER MUST BE CALCULATED BY REGRESSION TO
C PRODUCE THE EVALUATION STATISTICS AND RESIDUALS
C
      IF(IREV.EQ.1) NTR = N
      IF(IND2.EQ.1) GO TO 2400
      CALL QUAD(MAP(IEMLOC(IEQU,LA),1),MAP(IEMLOC(IEQU,LA),2))
      DO 2025 K=1,6
      IXD(K) = IQCO(K,IEMLOC(IEQU,LA),LA)
      DO 2025 KI=1,N
 2025 WSM(KI,K) = QP(KI,K)
      IXD(6) = 0
      CALL RLSEP(WSM,NTR,5,ID2,ALFA,IJOB,IXD,ANOVA,XYB,ID8,VARB,IER)
      IF(IER.NE.0) CALL PRERRM(IER,IEMLOC(IEQU,LA),0,'RLSEP      ')
      DO 2027 K=1,6
 2027 QCO(K,IEMLOC(IEQU,LA),LA) = XYB(K,2)
      CALL COMPCT(0)
      IF(ANOVA(12).EQ.0) ANOVA(12) = SQRT(ANOVA(5)/ANOVA(2))
      CALL RLRESA(QP,ID2,6,N,IH,ICL,BETA,ANOVA(12),RES,ID2,IER,ID9)
      NTR = NTRT
      IF(IER.NE.0) CALL PRERRM(IER,KACT(1),0,'RLRESA     ')
      RETURN
 2400 DO 2047 K=1,MGL
      IXD(K) = ICO(K,LA)
      DO 2047 KI=1,N
 2047 WSM(KI,K) = PRD(KI,K)
      IXD(MGL) = 0
      CALL RLSEP(WSM,NTR,MG,ID2,ALFA,IJOB,IXD,ANOVA,XYB,ID8,VARB,IER)
      IF(IER.NE.0) CALL PRERRM(IER,IJ,1,'RLSEP      ')
      DO 2049 K=1,MGL
 2049 CO(K,LA) = XYB(K,2)
      CALL COMPCT(1)
      IF(ANOVA(12).EQ.0) ANOVA(12) = SQRT(ANOVA(5)/ANOVA(2))
      CALL RLRESA(PRD,ID2,MGL,N,IH,ICL,BETA,ANOVA(12),RES,ID2,IER,ID9)
```

```
      IF(IER.NE.0) CALL PRERRM(IER,IU,1,'RLRESA    ')
      NTR = NTRT
      RETURN
      END


      SUBROUTINE DIAKA(IEQU,IV)
C
C INPUT:
C IEQU = THE SORTED LOCATION WITHIN LAYER LA
C        OF THE DESIRED EQUATION
C
C OUTPUT:
C IV = THE NUMBER OF PARTIAL MODELS CREATED FROM THE
C      ORIGINAL DATA MATRIX
C IACT,KACT = MAPS OF THE VARIABLES IN THE FINAL EQUATION.
C
C KACT(I) = THE LOCATION OF A VARIABLE IN A LAYER BEFORE
C           SORTING.
C IACT(2*I) AND IACT(2*I + 1) = THE SORTED LOCATIONS OF THE
C           VARIABLES IN THE PREVIOUS LAYER USED TO
C           CREATE VARIABLE KACT(I) IN THE CURRENT LAYER.
C (IACT CAN EASILY BE RECOVERED FROM THE MATRICES MAP
C AND KACT, BUT IS RETAINED FOR CONVENIENCE)
C
      %INCLUDE GMDH_COM
C
C DO LAST LAYER SEPARATELY
C
      IV = 1
      IACT(1) = IEQU
      KACT(1) = IEMLOC(IEQU,LA)
      IACT(2) = MAP(KACT(1),1)
      IACT(3) = MAP(KACT(1),2)
      IF(LA.EQ.1) RETURN
      DO 2500 IL=(LA-1),1,-1
      IV = 2*IV
      DO 2500 J=IV,2*IV-1
C
C UNSORTED POSITION IN CURRENT LAYER
C
      KACT(J) = IEMLOC(IACT(J),IL)
C
C SORTED POSITION IN PREVIOUS LAYER
C
      IACT(2*J) = MAP(KACT(J),1)
```

```
      IACT(2*J+1) = MAP(KACT(J),2)
 2500 CONTINUE
      RETURN
      END


      SUBROUTINE CALCOF(JCL,I1)
C
C THIS SUBROUTINE CALCULATES THE COEFFICIENTS FOR THE
C TERMS IN A SECOND LAYER EQUATION.
C
C INPUT:
C JCL = THE COLUMN NUMBER OF C INTO WHICH THE COEFFICIENTS ARE WRITTEN
C I1  = THE SORTED LAYER TWO EQUATION NUMBER
C
C OUTPUT:
C C(54,2) = THE VECTOR OF COEFFICIENTS OF ALL THE TERMS WHICH MAY
C           APPEAR IN A SECOND LAYER EQUATION
C
      %INCLUDE GMDH_COM
      I2 = MAP(IEMLOC(I1,2),1)
      I3 = MAP(IEMLOC(I1,2),2)
      DO 150 I=1,6
      SD(I) = QCO(I,IEMLOC(I1,2),2)
      SD(I+6) = QCO(I,IEMLOC(I2,1),1)
  150 SD(I+12) = QCO(I,IEMLOC(I3,1),1)
      C(1,JCL) = SD(6) + SD(1)*SD(12) + SD(2)*SD(18) + SD(3)
     &*SD(12)*SD(12) + SD(4)*SD(18)*SD(18) + SD(5)*SD(12)*SD(18)
      C(2,JCL) = SD(1)*SD(7) + 2.*SD(3)*SD(12)*SD(7)
     &+ SD(5)*SD(7)*SD(18)
      C(3,JCL) = SD(1)*SD(8) + 2.*SD(3)*SD(12)*SD(8)
     &+ SD(5)*SD(8)*SD(18)
      C(4,JCL) = SD(1)*SD(9) + SD(3)*SD(7)
     &*SD(7) + 2.*SD(3)*SD(12)*SD(9) + SD(5)
     &*SD(9)*SD(18)
      C(5,JCL) = SD(1)*SD(10) + SD(3)*SD(8)
     &*SD(8) + 2.*SD(3)*SD(12)*SD(10)
     &+ SD(5)*SD(10)*SD(18)
      C(6,JCL) = SD(1)*SD(11) + 2.*SD(3)*SD(12)
     &*SD(11) + 2.*SD(3)*SD(7)*SD(8)
     &+ SD(5)*SD(11)*SD(18)
      C(7,JCL) = SD(2)*SD(13) + 2.*SD(4)*SD(18)
     &*SD(13) + SD(5)*SD(12)*SD(13)
      C(8,JCL) = SD(2)*SD(14) + 2.*SD(4)*SD(18)
     &*SD(14) + SD(5)*SD(12)*SD(14)
      C(9,JCL) = SD(2)*SD(15) + SD(4)*SD(13)
```

```
     &*SD(13) + 2.*SD(4)*SD(18)*SD(15)
     &+ SD(5)*SD(12)*SD(15)
      C(10,JCL) = SD(2)*SD(16) + SD(4)*SD(14)
     &*SD(14) + 2.*SD(4)*SD(18)*SD(16)
     &+ SD(5)*SD(12)*SD(16)
      C(11,JCL) = SD(2)*SD(17) + 2.*SD(4)*SD(18)
     &*SD(17) + 2.*SD(4)*SD(13)*SD(14)
     &+ SD(5)*SD(12)*SD(17)
      IF(SD(3).EQ.0.) GO TO 510
      C(12,JCL) = SD(3)*SD(9)*SD(9)
      C(13,JCL) = SD(3)*SD(10)*SD(10)
      C(14,JCL) = SD(3)*(SD(11)*SD(11) + 2.*SD(9)*SD(10))
      C(15,JCL) = 2.*SD(3)*SD(7)*SD(9)
      C(16,JCL) = 2.*SD(3)*(SD(7)*SD(10) + SD(8)*SD(11))
      C(17,JCL) = 2.*SD(3)*(SD(7)*SD(11) + SD(8)*SD(9))
      C(18,JCL) = 2.*SD(3)*SD(8)*SD(10)
      C(19,JCL) = 2.*SD(3)*SD(9)*SD(11)
      C(20,JCL) = 2.*SD(3)*SD(10)*SD(11)
      GO TO 530
510   DO 520 J=12,20
520   C(J,JCL) = 0.
530   IF(SD(4).EQ.0.) GO TO 540
      C(21,JCL) = SD(4)*SD(15)*SD(15)
      C(22,JCL) = SD(4)*SD(16)*SD(16)
      C(23,JCL) = SD(4)*(SD(17)*SD(17) + 2.*SD(15)*SD(16))
      C(24,JCL) = 2.*SD(4)*SD(13)*SD(15)
      C(25,JCL) = 2.*SD(4)*(SD(13)*SD(16) + SD(14)*SD(17))
      C(26,JCL) = 2.*SD(4)*(SD(13)*SD(17) + SD(14)*SD(15))
      C(27,JCL) = 2.*SD(4)*SD(14)*SD(16)
      C(28,JCL) = 2.*SD(4)*SD(15)*SD(17)
      C(29,JCL) = 2.*SD(4)*SD(16)*SD(17)
      GO TO 560
540   DO 550 J=21,29
550   C(J,JCL) = 0.
560   IF(SD(5).EQ.0.) GO TO 570
      ICOUNT = 29
      DO 200 J=7,11
      DO 200 I=13,17
      ICOUNT = ICOUNT + 1
200   C(ICOUNT,JCL) = SD(5)*SD(J)*SD(I)
      RETURN
570   DO 580 J=30,54
580   C(J,JCL) = 0.
      RETURN
      END
```

```
      SUBROUTINE OUTFOR(JCL,IA,IB,IC1,ID)
C
C THIS SUBROUTINE CONTAINS THE OUTPUT FORMATS FOR LAYER 2 EQUATIONS
C
C JCL = COLUMN NUMBER IN C
C IA,IB,IC,ID = THE LAYER ZERO TERM NUMBERS
C
      %INCLUDE GMDH_COM
C
      IF(C(2,JCL).NE.0.) WRITE(IWRIT,600) C(2,JCL),IA
      IF(C(3,JCL).NE.0.) WRITE(IWRIT,600) C(3,JCL),IB
      IF(C(4,JCL).NE.0.) WRITE(IWRIT,601) C(4,JCL),IA
      IF(C(5,JCL).NE.0.) WRITE(IWRIT,601) C(5,JCL),IB
      IF(C(6,JCL).NE.0.) WRITE(IWRIT,602) C(6,JCL),IA,IB
      IF(C(7,JCL).NE.0.) WRITE(IWRIT,600) C(7,JCL),IC1
      IF(C(8,JCL).NE.0.) WRITE(IWRIT,600) C(8,JCL),ID
      IF(C(9,JCL).NE.0.) WRITE(IWRIT,601) C(9,JCL),IC1
      IF(C(10,JCL).NE.0.) WRITE(IWRIT,601) C(10,JCL),ID
      IF(C(11,JCL).NE.0.) WRITE(IWRIT,602) C(11,JCL),IC1,ID
      IF(C(12,JCL).NE.0.) WRITE(IWRIT,603) C(12,JCL),IA
      IF(C(13,JCL).NE.0.) WRITE(IWRIT,603) C(13,JCL),IB
      IF(C(14,JCL).NE.0.) WRITE(IWRIT,604) C(14,JCL),IA,IB
      IF(C(15,JCL).NE.0.) WRITE(IWRIT,605) C(15,JCL),IA
      IF(C(16,JCL).NE.0.) WRITE(IWRIT,606) C(16,JCL),IA,IB
      IF(C(17,JCL).NE.0.) WRITE(IWRIT,606) C(17,JCL),IB,IA
      IF(C(18,JCL).NE.0.) WRITE(IWRIT,605) C(18,JCL),IB
      IF(C(19,JCL).NE.0.) WRITE(IWRIT,607) C(19,JCL),IB,IA
      IF(C(20,JCL).NE.0.) WRITE(IWRIT,607) C(20,JCL),IA,IB
      IF(C(21,JCL).NE.0.) WRITE(IWRIT,603) C(21,JCL),IC1
      IF(C(22,JCL).NE.0.) WRITE(IWRIT,603) C(22,JCL),ID
      IF(C(23,JCL).NE.0.) WRITE(IWRIT,604) C(23,JCL),IC1,ID
      IF(C(24,JCL).NE.0.) WRITE(IWRIT,605) C(24,JCL),IC1
      IF(C(25,JCL).NE.0.) WRITE(IWRIT,606) C(25,JCL),IC1,ID
      IF(C(26,JCL).NE.0.) WRITE(IWRIT,606) C(26,JCL),ID,IC1
      IF(C(27,JCL).NE.0.) WRITE(IWRIT,605) C(27,JCL),ID
      IF(C(28,JCL).NE.0.) WRITE(IWRIT,607) C(28,JCL),ID,IC1
      IF(C(29,JCL).NE.0.) WRITE(IWRIT,607) C(29,JCL),IC1,ID
      IF(C(30,JCL).NE.0.) WRITE(IWRIT,602) C(30,JCL),IA,IC1
      IF(C(31,JCL).NE.0.) WRITE(IWRIT,602) C(31,JCL),IA,ID
      IF(C(32,JCL).NE.0.) WRITE(IWRIT,606) C(32,JCL),IA,IC1
      IF(C(33,JCL).NE.0.) WRITE(IWRIT,606) C(33,JCL),IA,ID
      IF(C(34,JCL).NE.0.) WRITE(IWRIT,608) C(34,JCL),IA,IC1,ID
      IF(C(35,JCL).NE.0.) WRITE(IWRIT,602) C(35,JCL),IB,IC1
      IF(C(36,JCL).NE.0.) WRITE(IWRIT,602) C(36,JCL),IB,ID
      IF(C(37,JCL).NE.0.) WRITE(IWRIT,606) C(37,JCL),IB,IC1
```

```
      IF(C(38,JCL).NE.0.) WRITE(IWRIT,606) C(38,JCL),IB,ID
      IF(C(39,JCL).NE.0.) WRITE(IWRIT,608) C(39,JCL),IB,IC1,ID
      IF(C(40,JCL).NE.0.) WRITE(IWRIT,606) C(40,JCL),IC1,IA
      IF(C(41,JCL).NE.0.) WRITE(IWRIT,606) C(41,JCL),ID,IA
      IF(C(42,JCL).NE.0.) WRITE(IWRIT,604) C(42,JCL),IA,IC1
      IF(C(43,JCL).NE.0.) WRITE(IWRIT,604) C(43,JCL),IA,ID
      IF(C(44,JCL).NE.0.) WRITE(IWRIT,609) C(44,JCL),IC1,ID,IA
      IF(C(45,JCL).NE.0.) WRITE(IWRIT,606) C(45,JCL),IC1,IB
      IF(C(46,JCL).NE.0.) WRITE(IWRIT,606) C(46,JCL),ID,IB
      IF(C(47,JCL).NE.0.) WRITE(IWRIT,604) C(47,JCL),IB,IC1
      IF(C(48,JCL).NE.0.) WRITE(IWRIT,604) C(48,JCL),IB,ID
      IF(C(49,JCL).NE.0.) WRITE(IWRIT,609) C(49,JCL),IC1,ID,IB
      IF(C(50,JCL).NE.0.) WRITE(IWRIT,608) C(50,JCL),IA,IB,IC1
      IF(C(51,JCL).NE.0.) WRITE(IWRIT,608) C(51,JCL),IA,IB,ID
      IF(C(52,JCL).NE.0.) WRITE(IWRIT,609) C(52,JCL),IA,IB,IC1
      IF(C(53,JCL).NE.0.) WRITE(IWRIT,609) C(53,JCL),IA,IB,ID
      IF(C(54,JCL).NE.0.) WRITE(IWRIT,610) C(54,JCL),IA,IB,IC1,ID
      RETURN
  600 FORMAT(1X,' + ',E12.5,' *VR',I4)
  601 FORMAT(1X,' + ',E12.5,' *VR',I4,'**2')
  602 FORMAT(1X,' + ',E12.5,' *VR',I4,4X,'*VR',I4)
  603 FORMAT(1X,' + ',E12.5,' *VR',I4,'**4')
  604 FORMAT(1X,' + ',E12.5,' *VR',I4,'**2 *VR',I4,'**2')
  605 FORMAT(1X,' + ',E12.5,' *VR',I4,'**3')
  606 FORMAT(1X,' + ',E12.5,' *VR',I4,4X,'*VR',I4,'**2')
  607 FORMAT(1X,' + ',E12.5,' *VR',I4,4X,'*VR',I4,'**3')
  608 FORMAT(1X,' + ',E12.5,' *VR',I4,4X,'*VR',I4,4X,'*VR',I4)
  609 FORMAT(1X,' + ',E12.5,' *VR',I4,4X,'*VR',I4,4X,'*VR',I4
     &,'**2')
  610 FORMAT(1X,' + ',E12.5,' *VR',I4,4X,'*VR',I4,4X,'*VR',I4
     &,4X,'*VR',I4)
      END


      SUBROUTINE PRCOEFF(IPQCO,IPIQCO,IPCO,IPICO,IEM,IIEMLOC,IEMAT)
C
C   MATRIX **** IS PRINTED WHEN I**** , OR IP****. = 1.
C
C INPUT:
C ALL ARGUMENTS ARE INPUTS.
C
      %INCLUDE GMDH_COM
      IF(JEM.EQ.0.AND.IADJ.EQ.0) SR = 'MS RES ADJ'
      IF(JEM.EQ.1.AND.IADJ.EQ.0) SR = 'R SQRD ADJ'
      IF(JEM.EQ.0.AND.IADJ.EQ.1) SR = 'MS RES'
      IF(JEM.EQ.1.AND.IADJ.EQ.1) SR = 'R SQRD'
```

```
      IF(IPQCO.NE.1) GO TO 650
      DO 500 J=1,LA
      WRITE(IWRIT,510)MS,J
  510 FORMAT(/,T20,'THE LOCATION INDICES AND COEFFICIENTS',
     & ' OF THE ',I3,' BEST PREDICTORS IN LAYER',I2./,T10,'NUMBER'
     &,T20,'LOCATION',T35,'BETA 1',T50,'BETA 2',T65,'BETA 3',T80
     &,'BETA 4',T95,'BETA 5',T107,'INTERCEPT'./,T21,'INDEX')
C
C EM MUST ALREADY HAVE BEEN SORTED FOR THE OUTPUT
C TO BE MEANINGFUL.
C
      DO 520 I=1,MS
      K = IEMLOC(I,J)
  520 WRITE(IWRIT,530)I,K,(QCO(L,K,J),L=1,6)
  530 FORMAT((T12,2(I4,5X),T32,6(E12.5,3X)))
  500 CONTINUE
  650  IF(IPCO.NE.1) GO TO 700
      WRITE(IWRIT,660)
  660 FORMAT(/,T20,'THE COEFFICIENTS OF THE REGRESSIONS IN EACH',
     & ' LAYER ON ALL THE PREDICTORS IN THAT LAYER'./,T20,
     & '(THE LAST ENTRY IS THE INTERCEPT)')
      J = 1
      WRITE(IWRIT,720) J,(CO(I,J),I=1,ML)
      IF(LA.LE.1.OR.LO.EQ.1) GO TO 700
      DO 715 J=2,LA
  715 WRITE(IWRIT,720)J,(CO(I,J),I=1,MSL)
  720 FORMAT(T20,'LAYER ',I1,/,(10(1X,E12.5)))
  700 IF(IEM.NE.1) GO TO 750
      WRITE(IWRIT,560)MS
  560 FORMAT(/,T20,'THE LOCATION INDICES AND ERROR MEASURES OF',
     &' THE',I3,' BEST PREDICTORS IN EACH LAYER')
      DO 570 J=1,LA
      WRITE(IWRIT,590)J,SR,SR,SR
  590 FORMAT(T50,'LAYER',I3,/,T5,'NUMBER',T15,'LOCATION',T33,A10
     &,T49,'NUMBER',T59,'LOCATION',T77,A10,T93,
     &'NUMBER',T103,'LOCATION',T121,A10,/,T16,'INDEX',T60,
     &'INDEX',T104,'INDEX')
      WRITE(IWRIT,605)(I,IEMLOC(I,J),EM(I,J),I=1,MS)
  605 FORMAT((T5,3(I3,7X,I4,11X,E12.5,7X)))
  570 CONTINUE
  750 IF(IIEMLOC.NE.1) GO TO 800
      WRITE(IWRIT,760)MS
  760 FORMAT(/,T20,'LOCATION MAP OF THE BEST',I3,' PREDICTORS IN EACH',
     &' LAYER',/,T5,'NUMBER',T15,'LAYER 1',T30,'LAYER 2',T45,'LAYER 3',
     & T60,'LAYER 4',T75,'LAYER 5')
  770 WRITE(IWRIT,780)(I,(IEMLOC(I,J),J=1,5),I=1,MS)
```

```
  780 FORMAT((T7,I4,5X,5(I4,11X)))
  800 IF(IEMAT.NE.1) GO TO 850
      WRITE(IWRIT,810)SR
  810 FORMAT(/,T20,'THE ',A10,' OF THE REGRESSIONS ON ALL',
     & ' THE PREDICTORS IN EACH LAYER')
      WRITE(IWRIT,830)SR,(J,EMAT(J),J=1,LA)
  830 FORMAT(T5,'LAYER',T20,A10,/,(T7,I3,T15,E12.5))
  850 IF(IPIQCO.NE.1) GO TO 900
      DO 860 J=1,LA
      WRITE(IWRIT,870)MS,J
  870 FORMAT(/,T20,'THE LOCATION INDICES AND ',
     & ' NON-ZERO COEFFICIENT INDICATORS OF THE ',
     & I3,' BEST PREDICTORS IN LAYER',I2,/,T10,'NUMBER',
     &T20,'LOCATION',T35,'BETA 1',T50,'BETA 2',T65,'BETA 3',T80,
     &'BETA 4',T95'BETA 5',T107,'INTERCEPT',/,T21,'INDEX')
C
C    EM MUST ALREADY HAVE BEEN SORTED FOR THE OUTPUT
C TO BE MEANINGFUL.
C
      DO 880 I=1,MS
      K = IEMLOC(I,J)
  880 WRITE(IWRIT,890)I,K,(IQCO(L,K,J),L=1,6)
  890 FORMAT((T12,2(I4,5X),T38,6(I1,14X)))
  860 CONTINUE
  900 IF(IPICO.NE.1) RETURN
      WRITE(IWRIT,960)
  960 FORMAT(/,T20,'THE NON-ZERO COEFFICIENT INDICATORS ',
     & ' OF THE REGRESSIONS IN EACH',
     & ' LAYER ON ALL THE PREDICTORS IN THAT LAYER',/,T20,
     & '(THE LAST ENTRY IS FOR THE INTERCEPT)')
      J = 1
      WRITE(IWRIT,920) J,(ICO(I,J),I=1,ML)
      IF(LA.EQ.1.OR.LO.EQ.1) RETURN
      DO 915 J=2,LA
  915 WRITE(IWRIT,920)J,(ICO(I,J),I=1,MSL)
  920 FORMAT(T20,'LAYER ',I1,/,(5X,10(8X,I1,4X)))
      END


       SUBROUTINE PRRSUM(IIER,IIND,IANOVA,IXYB,ITXYB,IVARB,IIH
     &,IBETA,IRES,IER,IND,IEQU)
C
C MATRIX **** IS PRINTED WHEN I**** = 1.
C
C INPUT:
C ALL ARGUMENTS ARE INPUTS.
```

```
C IER = THE ERROR PARAMETER FROM AN IMSL SUBROUTINE.
C IND = 0 FOR QUADRATIC FORMS.
C       1 FOR REGRESSIONS ON ALL THE VARIABLES IN A LAYER.
C IJ  =  THE LOCATION NUMBER OF A QUADRATIC FORM
C
      %INCLUDE GMDH_COM
      MAL = MGL
      IF(IND.EQ.0) MAL = 6
      MA = MAL - 1
      MMA = MAL*(MAL+1)/2
      IF(IIER.NE.1) GO TO 120
      WRITE(IWRIT,110)IER
  110 FORMAT(/,T10,'THE ERROR PARAMETER, IER(SEE IMSL DOCUMENTATION)'.
     &', = ',I3,/)
  120 IF(IIND.NE.1) GO TO 150
      WRITE(IWRIT,130)MA,MA
  130 FORMAT(/,T20,'IND: A 1 IN THE FIRST',I4,' LOCATIONS INDICATES'.
     & ' THE VARIABLE WAS FORCED INTO THE MODEL'./.T25,'A 1 IN THE'.
     & ' SECOND',I4,' LOCATIONS INDICATES THE VARIABLE IS IN THE'.
     & ' DEVELOPED MODEL.',//,T5,'VARIABLE',T15,'IND',T30,'VARIABLE'
     &,T40,'IND',T55,'VARIABLE',T65,'IND',T80,'VARIABLE',T90
     &,'IND',T105,'VARIABLE',T115,'IND')
      WRITE(IWRIT,140)(I,IXD(I),I=1,MA)
  140 FORMAT((T9,5(I4,4X,I1,16X)))
      WRITE(IWRIT,139)
  139 FORMAT(/)
      WRITE(IWRIT,140)(I,IXD(I+MA),I=1,MA)
  150 IF(IANOVA.NE.1) GO TO 200
      EMJ = EM(IEQU,LA)
      IF(IND.EQ.1) EMJ = EMAT(LA)
      WRITE(IWRIT,160)ANOVA(1),ANOVA(4),ANOVA(7),ANOVA(9),ANOVA(10)
     &,ANOVA(2),ANOVA(5),ANOVA(8),ANOVA(3),ANOVA(6),ANOVA(14)
     &,ANOVA(15),ANOVA(16)
  160 FORMAT(/,T50,'ANALYSIS OF VARIANCE TABLE',/,T10,'SOURCE',T35,
     &'D.F.',T57,'SS',T77,'MS',T92,'F RATIO',T108,'F TAIL AREA'./,T10.
     &'REGRESSION',T30,5(E12.5,8X),/,
     &T10,'RESIDUALS',T30,3(E12.5,8X),/,T10,'CORRECTED TOTAL',
     & T30,2(E12.5,8X),//,T10,'LACK OF FIT TEST',T50,E12.5,
     &29X,2(E12.5,8X) /)
      WRITE(IWRIT,170)ANOVA(11),ANOVA(12),ANOVA(13),EMJ
  170 FORMAT(T10,'THE PERCENTAGE OF THE RESPONSE VARIATION EXPLAINED'.
     & ' BY THE REGRESSION = ',T90,E12.5,/,T10,'THE STANDARD DEVIATION'.
     & ' OF THE RESIDUALS = ',T90,E12.5,/,T10,'THE STANDARD DEVIATION',
     &' OF THE RESIDUALS AS A PERCENTAGE OF THE RESPONSE MEAN = ',
     &T90,E12.5,/,T10,'THE ERROR MEASURE USED TO ORDER THE EQUATIONS =',
     &T90,E12.5,/)
```

```
200 IF(IXYB.NE.1) GO TO 250
    WRITE(IWRIT,210)
210 FORMAT(/,T50,'REGRESSION MODEL SUMMARY. XYB',/,T10,'VARIABLE'
   &,T30,'MEAN',T41,'COEFFICIENT',T57,'ADJ. SS',T72,'F RATIO',T86
   &,'F TAIL AREA',T100,'FORCED',T115,'CHOSEN',/,T99,
   &'VARIABLES',T114,'VARIABLES')
    WRITE(IWRIT,230)(I,(XYB(I,J),J=1,5),IXD(I),IXD(I+MA),I=1,MA)
230 FORMAT((T12,I3,T25,5(E12.5,3X),T102,2(I1,14X)))
    WRITE(IWRIT,240)XYB(MAL,1),XYB(MAL,2)
240 FORMAT(T40,'INTERCEPT',/,T10,'RESPONSE',T25,2(E12.5,3X))
250 IF(ITXYB.NE.1) GO TO 300
    WRITE(IWRIT,260)
260 FORMAT(/,T50,'REGRESSION MODEL SUMMARY. TXYB',/,T10,'VARIABLE'
   &,T30,'MEAN',T41,'COEFFICIENT',T57,'ADJ. SS',T72,'F RATIO',T86,
   &'F TAIL AREA',T101,'VARIANCE')
    IF(IC.EQ.0) GO TO 283
    K = 0
    DO 285 I=1,IC
    K = K + I
285 WRITE(IWRIT,280) IH(I),(TXYB(I,J),J=1,5),VARB(K)
280 FORMAT((T12,I3,T25,6(E12.5,3X)))
283 WRITE(IWRIT,290)TXYB(ICL,1),TXYB(ICL,2)
290 FORMAT(T40,'INTERCEPT',/,T10,'RESPONSE',T25,2(E12.5,3X))
300 IF(IVARB.NE.1) GO TO 350
    WRITE(IWRIT,310)
310 FORMAT(/,T50,'THE INVERSE OF THE INFORMATION MATRIX')
    K = 0
    DO 315 I=1,IC
    K = K + I
    L = K - I + 1
315 WRITE(IWRIT,320) I,(VARB(J),J=L,K)
320 FORMAT(1X,I3,2X,(10(1X,E11.4)))
350 IF(IIH.NE.1) GO TO 400
    WRITE(IWRIT,360)(IH(I),I=1,ICL)
360 FORMAT(/,T30,'THE FOLLOWING VARIABLES(COLUMN NUMBERS)',
   &' WERE CHOSEN IN THE STEPWISE REGRESSION'/,T30,'(THE LAST',
   & ' NUMBER IS THE COLUMN CONTAINING THE RESPONSE VARIABLE)',/,
   & (T12,10I10))
400 IF (IBETA.NE.1) GO TO 450
    WRITE(IWRIT,410)
410 FORMAT(/,T30,'THE VARIABLE NUMBERS(COLUMN NUMBERS)AND ',
   &'COEFFICIENTS CHOSEN IN THE STEPWISE REGRESSION',/,T30,'(THE ',
   &'LAST LISTING IS THE RESPONSE VARIABLE AND INTERCEPT',/,T5,
   &'VARIABLE',T15,'COEFFICIENT',T35,'VARIABLE',T45,
   &'COEFFICIENT',T65,'VARIABLE',T75,'COEFFICIENT',
   &T95,'VARIABLE',T105,'COEFFICIENT')
```

```
      WRITE(IWRIT,420)(IH(I),BETA(I),I=1,ICL)
  420 FORMAT((T6,4(I4,7X,E12.5,7X)))
  450 IF(IRES.NE.1) RETURN
      WRITE(IWRIT,460)
  460 FORMAT(/,T50,'COMPLETE RESIDUAL LIST'./.T10,'OBSERVATION',T26
     &,'OBSERVED',T40,'PREDICTED',T55,'RESIDUAL',T66,'STANDARDIZED'./.
     &T10,'NUMBER',T25,'RESPONSE',T40,'RESPONSE',T66,'RESIDUAL')
      WRITE(IWRIT,480)(I,(RES(I,J),J=1,4),I=1,N)
  480 FORMAT((T14,I5,T25,4(E12.5,3X)))
      RETURN
      END


      SUBROUTINE PRDATA(IPRDBR,IPRDBC,IWSMBR,IWSMBC,IQP,MQ1,MQ2,IC1,IC2,
     &                  IR1,IR2)
C
C MATRIX **** IS PRINTED WHEN I**** = 1 .
C THE SUFFIX BR MEANS BY ROWS
C THE SUFFIX BC MEANS BY COLUMNS
C
C INPUT:
C ALL ARGUMENTS ARE INPUTS.
C MQ1 = THE NUMBER OF THE FIRST VARIABLE USED FOR QP.
C MQ1 = THE NUMBER OF THE SECOND VARIABLE USED FOR QP.
C IC1,IC2 = FIRST AND LAST COLUMNS TO BE PRINTED
C IR1,IR2 = FIRST AND LAST ROWS TO BE PRINTED
C
C NOTE: IC1,IC2,IR1,IR2 = 0 TO PRINT ALL ROWS OR COLUMNS
C
C
      %INCLUDE GMDH_COM
      IF(IPRDBR.NE.1.AND.IPRDBC.NE.1) GO TO 500
      WRITE(IWRIT,460)TITLE(1),ML
  460 FORMAT(/,T20,'THE DATA MATRIX ',A50./.T20,'VARIABLE ',I4.
     & ' IS THE RESPONSE')
      IF(IPRDBR.NE.1) GO TO 485
      IF(IR1.EQ.0) IR1 = 1
      IF(IR2.EQ.0) IR2 = N
      DO 470 I=IR1,IR2
  470 WRITE(IWRIT,480)I,(PRD(I,J),J=1,MGL)
  480 FORMAT(1X,'OBSERVATION NUMBER',I4,/,(10(1X,E12.5)))
      IF(IPRDBC.NE.1) GO TO 500
      IF(IC1.EQ.0) IC1 = 1
      IF(IC2.EQ.0) IC2 = MGL
  485 DO 486 J=IC1,IC2
  486 WRITE(IWRIT,481) J,(PRD(I,J),I=1,N)
```

```
  481 FORMAT(1X,'VARIABLE',I4,/,(10(1X,E12.5)))
  500 IF(IWSMBR.NE.1.AND.IWSMBC.NE.1) GO TO 550
      WRITE(IWRIT,510)
  510 FORMAT(/,T20,'THE TEMPORARY DATA MATRIX,WSM: THE LAST COLUMN
     & CONTAINS THE RESPONSE VARIABLE')
      IF(IWSMBR.NE.1) GO TO 585
      DO 530 I=1,N
  530 WRITE(IWRIT,480)I,(WSM(I,J),J=1,MGL)
      IF(IWSMBC.NE.1) GO TO 550
  585 DO 586 J=1,MGL
  586 WRITE(IWRIT,481) J,(WSM(I,J),I=1,N)
  550 IF(IQP.NE.1) RETURN
      WRITE(IWRIT,560)MQ1,MQ2
  560 FORMAT(/,T10,'THE QUADRATIC SETTING OF COLUMNS MQ1 =',I3,' AND',
     &' MQ2 =',I3,' OF MATRIX PRD',/,1X,'OBS. NUM.',T21,'MQ1',T36,'MQ2'
     &,T47,'MQ1*MQ1',T62,'MQ2*MQ2',T77,'MQ1*MQ2',T91,'RESPONSE')
      WRITE(IWRIT,580)(I,(QP(I,J),J=1,6),I=1,N)
  580 FORMAT((T4,I3,T15,6(E12.5,3X)))
      RETURN
      END


      SUBROUTINE PRERRM(IER,IJ,IND1,SR)
C
C INPUT:
C IER = THE IMSL ERROR CODE(SEE IMSL DOCUMENTATION).
C IJ = THE STEP WITHIN THE LAYER FROM WHICH THE ERROR
C      CODE WAS RETURNED.
C IND = 0 FOR QUADRATIC FORMS
C       1 FOR REGRESSIONS ON ALL VARIABLES IN A LAYER.
C SR = THE NAME OF THE IMSL SUBROUTINE FOR WHICH THIS
C      SUBROUTINE IS BEING USED.  THE MAXIMUM LENGTH
C      OF SR IS 10 CHARACTERS.
C
C OUTPUT:
C A MESSAGE LISTING THE IMSL ERROR CODE AND LOCATION OF OCCURRENCE.
C
      %INCLUDE GMDH_COM
      IF(IND1.EQ.1) GO TO 1200
      WRITE(IWRIT,1110)IER,SR,IJ,LA
 1110 FORMAT(T20,'IMSL ERROR CODE ',I3,' WAS RETURNED BY SUBROUTINE '
     &,A10,' AT THE ',I4,' CALL IN LAYER ',I1)
      IF(IWRIT.EQ.6) GO TO 1195
      WRITE(6,1110)IER,SR,IJ,LA
 1195 RETURN
 1200 WRITE(IWRIT,1210)IER,SR,LA
```

```
 1210 FORMAT(T20,'IMSL ERROR CODE ',I3,' WAS RETURNED BY SUBROUTINE '
     &,A10,' AT THE CALL TO ALL THE VARIABLES IN LAYER ',I1)
      IF(IWRIT.EQ.6) RETURN
      WRITE(6,1210)IER,SR,LA
      RETURN
      END


      SUBROUTINE PREQU(IIACT,IKACT,ISE,ICE,IEQU,IND1,IND2)
C
C ITEM **** IS PRINTED WHEN I**** = 1.
C
C INPUT:
C IEQU = THE SORTED LOCATION WITHIN LAYER LA
C        OF THE DESIRED EQUATION.
C IND1 = 0 TO PRINT ALL COEFFICIENTS, INCLUDING ZEROS
C        1 TO PRINT ONLY NON ZERO COEFFICIENTS.
C IND1 APPLIES ONLY TO SE OUTPUT(SEE BELOW)
C IND2 = 0 FOR QUADRATIC FORMS
C        1 FOR REGRESSIONS ON ALL THE VARIABLES IN A LAYER
C
C
C OUTPUT:
C IACT,KACT = MAPS OF THE VARIABLES IN THE FINAL EQUATION.
C
C KACT(I) = THE LOCATION OF A VARIABLE IN A LAYER BEFORE
C           SORTING.
C IACT(2*I) AND IACT(2*I + 1) = THE SORTED LOCATIONS OF THE
C           VARIABLES IN THE PREVIOUS LAYER USED TO
C           CREATE VARIABLE KACT(I) IN THE CURRENT LAYER.
C (IACT CAN EASILY BE RECOVERED FROM THE MATRICES MAP
C AND KACT, BUT IS RETAINED FOR CONVENIENCE)
C
C SE = A LIST OF THE DESIRED EQUATION BY LAYERS.
C CE = A LIST OF THE DESIRED EQUATION.
C
C THIS SUBROUTINE SHOULD GENERALLY NOT BE USED TO PRINT
C IACT OR KACT AND PRINT AN EQUATION IN THE SAME CALL.
C ALL ITEMS REQUESTED WILL ALWAYS BE PRINTED. BUT IN SOME
C CASES THE ORDER OF OUTPUT MAY BE MIXED UP.
C
      %INCLUDE GMDH_COM
 2590 LAT = LA
      ICJ = 1
      IND3 = 0
      IND4 = 0
```

```
      IF(IND2.EQ.1) GO TO 2690
 2910 CALL DIAKA(IEQU,IV)
      IF(IIACT.NE.1) GO TO 2670
      WRITE(IWRIT,2660)IACT
 2660 FORMAT(/,T20,'THE MAPPING VECTOR IACT'./,(10I10))
 2670 IF(IKACT.NE.1) GO TO 2690
      WRITE(IWRIT,2680)KACT
 2680 FORMAT(/,T20,'THE MAPPING VECTOR KACT'./,(10I10))
 2690 IF(ICE.EQ.1.AND.LA.EQ.1) GO TO 2600
      IF(IND3.EQ.1) GO TO 2600
      IF(ISE.NE.1) GO TO 2700
 2600 IF(IND2.EQ.ICU) GO TO 2760
      ICNT1 = 1
      IF(IND2.EQ.1) GO TO 2940
      WRITE(IWRIT,2629)
 2629 FORMAT(/,T5,'THE COMPLETE EQUATION, BY LAYERS'./.T5.
     &'RESPONSE (Y) =',/)
C
C OUTPUT FORMATS
C
 2940 DO 2610 IL=LA,1,-1
      L2 = IL - 1
      DO 2650 I=ICNT1,(2*ICNT1-1)
      IF(I.EQ.ICNT1) GO TO 2652
      DO 2655 II=ICNT1,(I-1)
 2655 IF(IACT(I).EQ.IACT(II)) GO TO 2650
 2652 IU = I+I
      IUL = IU+1
      IF(IND3.EQ.1) GO TO 2712
      IF(IND1.EQ.1) GO TO 2657
      WRITE(IWRIT,2630)IACT(I),IL,QCO(6,KACT(I),IL),QCO(1,KACT(I),IL),
     &IACT(IU),L2,QCO(2,KACT(I),IL),IACT(IUL),L2,QCO(3,KACT(I),IL),
     &IACT(IU),L2,QCO(4,KACT(I),IL),IACT(IUL),L2,QCO(5,KACT(I),IL),
     &IACT(IU),L2,IACT(IUL),L2
      GO TO 2650
 2657 WRITE(IWRIT,2710) IACT(I),IL,QCO(6,KACT(I),IL)
 2710 FORMAT(1X,/,1X,'VR',I4,' LY',I1,' = ',/,4X,E12.5)
 2712 DO 2650 LI=1,5
      IF(IQCO(LI,KACT(I),IL).EQ.0) GO TO 2650
      IF(LI.EQ.1) WRITE(IWRIT,2720) QCO(LI,KACT(I),IL),IACT(IU),L2
      IF(LI.EQ.2) WRITE(IWRIT,2720) QCO(LI,KACT(I),IL),IACT(IUL),L2
      IF(LI.EQ.3) WRITE(IWRIT,2730) QCO(LI,KACT(I),IL),IACT(IU),L2
      IF(LI.EQ.4) WRITE(IWRIT,2730) QCO(LI,KACT(I),IL),IACT(IUL),L2
      IF(LI.EQ.5) WRITE(IWRIT,2740) QCO(LI,KACT(I),IL),IACT(IU),L2
     &,IACT(IUL),L2
 2650 CONTINUE
```

```
      ICNT1 = ICNT1 + ICNT1
 2630 FORMAT(1X,/,1X,'VR',I4,' LY',I1,' = ',/,4X,E12.5,/,1X,' + ',
     &E12.5,'*VR',I4,' LY',I1,' +',E12.5,'*VR',I4,' LY',
     &I1,' + ',E12.5,'*VR',I4,' LY',I1,'**2 + ',E12.5,'*VR',I4
     &,' LY',I1,'**2',/,1X,' + ',E12.5,'*VR',I4,' LY',I1,'*VR'
     &,I4,' LY',I1)
 2720 FORMAT(1X,' + ',E12.5,'*VR',I4,' LY',I1)
 2730 FORMAT(1X,' + ',E12.5,'*VR',I4,' LY',I1,'**2')
 2740 FORMAT(1X,' + ',E12.5,'*VR',I4,' LY',I1,'*VR',I4,' LY',I1)
 2610 CONTINUE
      IF(IND4.EQ.1) RETURN
      IF(IND2.EQ.0) GO TO 2700
      IF(IND3.EQ.0) GO TO 2900
      GO TO 3330
C
C BEGIN SECTION FOR REGRESSION ON ALL INPUT VARIABLES FOR A LAYER
C
 2760 MG = M
      IF(LA.GE.2) MG = MS
      MGL = MG + 1
      LA = LA - 1
      WRITE(IWRIT,2629)
      WRITE(IWRIT,2810) CO(MGL,LAT)
 2810 FORMAT(1X,E12.5)
      IF(IND1.EQ.1) GO TO 2840
      WRITE(IWRIT,2820)(CO(I,LAT),I,LA,I=1,MG)
 2820 FORMAT(4(1X,' + ',E12.5,'*VR',I4,'LY',I1))
      GO TO 2865
 2840 DO 2850 J=1,MG
      IF(ICO(J,LAT).NE.0) WRITE(IWRIT,2860) CO(J,LAT),J,LA
 2850 CONTINUE
 2860 FORMAT(1X,' + ',E12.5,'*VR',I4,'LY',I1)
 2865 IF(LAT.LE.1) GO TO 2920
      ICJ = 0
      DO 2900 J=1,MG
      IF(ICO(J,LAT).EQ.0) GO TO 2900
      IEQU = J
      GO TO 2910
 2900 CONTINUE
 2920 LA = LAT
C
C BEGIN SECTION FOR FULLY EXPANDED EQUATION
C
 2700 IF(LA.EQ.1) RETURN
      IF(ICE.NE.1) RETURN
      WRITE(IWRIT,3010)
```

```
 3010 FORMAT(/,T5,'THE COMPLETE EQUATION',/,T5,'RESPONSE (Y) =')
      IF(IND2.EQ.1) GO TO 3300
      IF(LA.EQ.3) GO TO 3060
C
C FOR LAYER 2
C
      CALL CALCOF(1,IEQU)
      WRITE(IWRIT,2710) IEQU,LA,C(1,1)
      CALL OUTFOR(1,IACT(4),IACT(5),IACT(6),IACT(7))
      RETURN
C
C FOR LAYER 3
C
 3060 IF(IQCO(3,IEQU,3).EQ.0.AND.IQCO(4,IEQU,3).EQ.0.AND.IQCO(5,IEQU,3)
     &.EQ.0) GO TO 3090
      WRITE(IWRIT,3070)
      IF(IWRIT.NE.6) WRITE(6,3070)
 3070 FORMAT(/,T5,'THIS PROGRAM IS NOT CURRENTLY EQUIPPED TO FULLY',
     & ' EXPAND GENERAL THIRD LAYER EQUATIONS')
      IF(ISE.EQ.1) RETURN
      LAT = LA
      ICJ = 1
      IND3 = 0
      IND4 = 1
      CALL DIAKA(IEQU,IV)
      GO TO 2600
 3090 CALL CALCOF(1,IACT(2))
      CALL CALCOF(2,IACT(3))
      CONST= QCO(6,IEQU,3) + QCO(1,IEQU,3)*C(1,1) + QCO(2,IEQU,3)*C(1,2)
      WRITE(IWRIT,2710) IEQU,LA,CONST
      DO 3100 J=1,2
      IF(IQCO(J,IEQU,3).EQ.0) GO TO 3100
      DO 3100 I=2,54
      C(I,J) = C(I,J)*QCO(J,IEQU,3)
 3100 CONTINUE
      CALL OUTFOR(1,IACT(8),IACT(9),IACT(10),IACT(11))
      CALL OUTFOR(2,IACT(12),IACT(13),IACT(14),IACT(15))
      RETURN
C
C SECTION FOR REGRESSION ON ALL INPUTS FOR A LAYER
C
C LAYER 2
C
 3300 IF(LA.EQ.3) GO TO 3400
      CONST = CO(MSL,2)
      DO 3310 I=1,MS
```

```
      IF(ICO(I,2).EQ.0) GO TO 3310
      DO 3320 J=1,6
 3320 QCO(J,IEMLOC(I,1),1) = QCO(J,IEMLOC(I,1),1)*CO(I,2)
      CONST = CONST + QCO(6,IEMLOC(I,1),1)
 3310 CONTINUE
      WRITE(IWRIT,2810) CONST
      ICJ = 0
      IND3 = 1
      LA = 1
      DO 3330 J=1,MS
      IF(ICO(J,2).EQ.0) GO TO 3330
      IEQU = J
      GO TO 2910
 3330 CONTINUE
      LA = 2
      DO 3340 I=1,MS
      IF(ICO(I,2).EQ.0) GO TO 3340
      DO 3350 J=1,6
 3350 QCO(J,IEMLOC(I,1),1) = QCO(J,IEMLOC(I,1),1)/CO(I,2)
 3340 CONTINUE
      RETURN
 3400 CONST = CO(MSL,3)
      LA = 2
      DO 3410 I=1,MS
      IF(ICO(I,3).EQ.0) GO TO 3410
      CALL DIAKA(I,IV)
      CALL CALCOF(1,I)
      DO 3420 J=1,54
 3420 C(J,1) = C(J,1)*CO(I,3)
      CONST = CONST + C(1,1)
      CALL OUTFOR(1,IACT(4),IACT(5),IACT(6),IACT(7))
 3410 CONTINUE
      WRITE(IWRIT,3430) CONST
 3430 FORMAT(1X,' + ',E12.5)
      LA = 3
      RETURN
      END


      SUBROUTINE STORVEC(ICTRL4,IND)
C
C THIS SUBROUTINE STORES VECTORS FOR LATER PLOTTING
C
C INPUT:
C TITLE(1) = THE FILE NAME
C TITLE(2) = A LABEL FOR THE GRAPH
```

```
C ICTRL4 = THE PLOT IDENTIFIER (SEE SUBROUTINE EXEQU IN TEST7.FORTRAN)
C IND = 0 TO STORE VECTORS
C       1 TO PLOT VECTORS PREVIOUSLY STORED
C
      %INCLUDE GMDH_COM
      PRINT,' ENTER IFILE = THE FILE NUMBER FOR THE VECTORS'
      PRINT
      PRINT,' 1I FORMAT'
      READ,IFILE
      IF(IND.NE.0) GO TO 200
      WRITE(IFILE) N,NCURVE,ICTRL4,TITLE(1),TITLE(2),(A2(I),I=1,N),
     &(B2(I),I=1,N)
      RETURN
  200 REWIND IFILE
      READ(IFILE) N,NCURVE,ICTRL4,TITLE(1),TITLE(2),(A2(I),I=1,N),
     &(B2(I),I=1,N)
      PRINT,' ENTER ICTRL3 = 0 FOR AUTOMATIC PLOTTING'
      PRINT,'                1 TO CONTROL PLOT FORMAT'
      PRINT
      PRINT,' 1I FORMAT'
      READ,ICTRL3
      ICTRL(1,1) = N
      CALL PLOT1(ICTRL3,ICTRL4)
      RETURN
      END
```

```
      %INCLUDE GMDH_COM
      IND = 1
      DO 100 J=1,20
  100 CALL STORVEC(ICTRL4,IND)
      STOP
      END


      SUBROUTINE STORVEC(ICTRL4,IND)
C
C THIS SUBROUTINE STORES VECTORS FOR LATER PLOTTING
C OR PLOTS VECTORS WHICH HAVE BEEN STORED
C
C INPUT:
C TITLE(1) = THE FILE NAME
C TITLE(2) = A LABEL FOR THE GRAPH
C ICTRL4 = THE PLOT IDENTIFIER ( SEE SUBROUTINE
C          EXEQU IN GMDH.FORTRAN)
C IND = 0 TO STORE VECTORS
C       1 TO PLOT VECTORS PREVIOUSLY STORED
C
      %INCLUDE GMDH_COM
C
      PRINT,' ENTER IFILE = THE FILE NUMBER FOR THE VECTORS'
      PRINT
      PRINT,' 1I FORMAT'
      READ,IFILE
      IF(IFILE.EQ.0) STOP
      IF(IND.NE.0) GO TO 200
      WRITE(IFILE) N,NCURVE,ICTRL4,TITLE(1),TITLE(2),
     &             (A2(I),I=1,N),(B2(I),I=1,N)
      RETURN
  200 REWIND IFILE
      READ(IFILE) N,NCURVE,ICTRL4,TITLE(1),TITLE(2),
     &            (A2(I),I=1,N),(B2(I),I=1,N)
      PRINT,' ENTER IND1 = 0 TO PLOT ALL POINTS'
      PRINT,'              1 TO PLOT MOVING AVERAGE OF NA'
      PRINT,'              2 TO PLOT AVERAGES OF NA'
      PRINT,'       NA'
      PRINT
      PRINT,' 2I FORMAT'
      READ,IND1,NA
      IF(IND1.NE.0) CALL SMOOTH(IND1,NA)
      PRINT,' ENTER ICTRL3 = 0 FOR AUTOMATIC PLOTTING'
      PRINT,'                1 TO CONTROL PLOT FORMAT'
```

```
      PRINT
      PRINT,' 1I FORMAT'
      READ,ICTRL3
      ICTRL(1,1) = N
      CALL PLOT1(ICTRL3,ICTRL4)
      RETURN
      END
      SUBROUTINE SMOOTH(IND1,NA)
      %INCLUDE GMDH_COM
      IH(1) = 59
      IH(2) = 80
      IH(3) = 79
      IH(4) = 89
      IH(5) = 88
      IF(IND1.EQ.2) GO TO 400
      ICNT = 1
      IBP = 0
      DO 100 J=1,5
      ND = IH(J)
      RS = 0.
      DO 110 I=1+IBP,NA+IBP
  110 RS = RS + B2(I)
      VARB(ICNT) = RS/NA
      ICNT = ICNT + 1
      DO 120 I=NA+IBP+1,ND+IBP
      RS = RS - B2(I-NA) + B2(I)
      VARB(ICNT) = RS/NA
      ICNT = ICNT+1
  120 CONTINUE
      PRINT,' YEAR, IH, ICNT',J,IH(J),ICNT
      IBP = IBP + IH(J)
  100 CONTINUE
      N = ICNT - 1
      DO 130 J=1,N
  130 B2(J) = VARB(J)
      RETURN
  400 ICNT = 1
      IBP = 0
      DO 200 J=1,5
      ND = IH(J)/NA
      DO 210 I=1,ND
      RS = 0.
      DO 220 IL =1+IBP,NA+IBP
  220 RS = RS + B2(IL+(I-1)*NA)
      VARB(ICNT) = RS/NA
      ICNT = ICNT + 1
```

```
210 CONTINUE
    PRINT,' YEAR, IH, ICNT',J,IH(J),ICNT
    IBP = IBP + IH(J)
200 CONTINUE
    N = ICNT - 1
    DO 230 J=1,N
230 B2(J) = VARB(J)
    RETURN
    END
```

```
      SUBROUTINE PLOT1(ICTRL3,ICTRL4)
C
C INPUT ARGUMENTS
C
C ICTRL3 = 0 TO USE DEFAULT PLOTTING FORMAT
C          1 TO CONTROL PLOTTING FORMAT
C ICTRL4 CONTROLS INPUT DATA SET
C
C VARIABLES WHICH MUST BE ASSIGNED IN THE CALLING PROGRAM
C
C NCURVE
C ICTRL(K,1), K=1,NCURVE
C A2(I),I=1,NPTS
C B2(I),I=1,NPTS
C TITLE(I),I=1,2
C
C COMMON BLOCK P1 AND THE CHARACTER VARIABLES MUST BE
C DECLARED IN THE CALLING PROGRAM.
C
C OTHER VARIABLES
C
C ICTRL5 IS THE CONTINUATION PARAMETER
C TITLE = CHARACTER STRINGS OF LABELS FOR THE GRAPH
C NTITLE = THE NUMBER OF LABELS TO BE WRITTEN ON THE GRAPH
C XTITLE = X COORDINATE OF TITLE
C YTITLE = Y COORDINATE OF TITLE
C ATITLE = ANGLE OF TITLE
C HTITLE = HEIGHT OF TITLE
C XAXIS = LABEL FOR X AXIS
C YAXIS = LABEL FOR Y AXIS
C XXIN,XYIN = COORDINATES OF BEGINNING OF X AXIS
C YXIN,YYIN = COORDINATES OF BEGINNING OF Y AXIS
C XDELTA = QUANTITY BETWEEN TIC MARKS ON X AXIS
C YDELTA = QUANTITY BETWEEN TIC MARKS ON Y AXIS
C XFIRST = X VALUE OF THE ORIGIN
C YFIRST = Y VALUE OF THE ORIGIN
C CHAR = COMMENTS WRITTEN AFTER GRAPH IS PRODUCED
C        (CHAR IS READ FROM FILE 25)
C NCHAR = NUMBER OF LINES FROM CHAR TO BE WRITTEN
C
C THE NEXT 6 VARIABLES DEALING WITH SYMBOLS ARE FOR THE
C LABELS ON THE GRAPH.  THE SYMBOLS USED TO PLOT POINTS ON THE GRAPH
C ARE CONTROLLED BY THE ICTRL MATRIX.
C
C NSMBL = NUMBER OF SYMBOLS
C HSMBL = HEIGHT OF SYMBOLS
```

```
C ASMBL = ANGLE OF SYMBOLS
C XSMBL = X COORDINATE OF SYMBOL
C YSMBL = Y COORDINATE OF SYMBOL
C ISMBL = SYMBOL CODE NUMBER
C A = VECTOR OF DATA FOR THE ORDINATE (FOR ONE CURVE)
C B = VECTOR OF DATA FOR THE ABSCISSA (FOR ONE CURVE)
C NCURVE = NUMBER OF CURVES TO BE PLOTTED ON ONE GRAPH
C A2 = COMBINED DATA FOR ORDINATE (NCURVE DATASETS)
C B2 = COMBINED DATA FOR ABSCISSA (NCURVE DATASETS)
C NPTS = NUMBER OF DATA POINTS IN EACH OF A2 AND B2
C XLENGT = LENGTH OF X AXIS
C YLENGT = LENGTH OF Y AXIS
C ICTRL = MATRIX WITH 1 ROW PER CURVE (NCURVE ROWS)
C   COL 1 = NUMBER OF POINTS IN THE CURVE
C   COL 2 = 0 TO NOT CONNECT POINTS
C   COL 3 = 0 TO NOT PLOT WITH SYMBOLS
C           OTHERWISE, THE INTERVAL BETWEEN POINTS
C           TO BE MARKED WITH SYMBOLS
C   COL 4 = SYMBOL CODE, FOR COL 3 = 0
C   COL 5 = SYMBOL SIZE, FOR COL 3 = 0
C
      PARAMETER(IP1=2000)
      PARAMETER(IP2=5*IP1+2)
      CHARACTER*50 TITLE,CHAR,XAXIS,YAXIS
      COMMON/P1/TITLE(11),XTITLE(11),YTITLE(11),ATITLE(11),
     &HTITLE(11),HSMBL(11),ASMBL(11),YSMBL(11),XSMBL(11),ISMBL(11),
     &NSMBL,NCHAR,NTITLE,A1(4),B1(4),A(IP1),B(IP1),NCURVE
     &,ICTRL(5,5),A2(IP2),B2(IP2),XAXIS,YAXIS
      EXTERNAL CCS_$AXIS (DESCRIPTORS)
      EXTERNAL CCS_$DFACT (DESCRIPTORS)
      EXTERNAL CCS_$LINE (DESCRIPTORS)
      EXTERNAL CCS_$PLOT (DESCRIPTORS)
      EXTERNAL CCS_$PLOTS
      EXTERNAL CCS_$SCALE (DESCRIPTORS)
      EXTERNAL CCS_$SYMBOL (DESCRIPTORS)
C**********************************************************
C**********************************************************
C
C DEFAULT VARIABLE ASSIGNMENTS
C
C
      IF(ICTRL3.NE.0) GO TO 200
C
      TITLE(3) = 'STANDARDIZED RESIDUALS VS. OBSERVATION NUMBER'
      TITLE(4) = 'ORIGINAL RESIDUALS VS. OBSERVATION NUMBER'
      TITLE(5) = 'STANDARDIZED RESIDUALS VS. PREDICTED RESPONSE'
```

```
      TITLE(6) = 'ORIGINAL RESIDUALS VS. PREDICTED RESPONSE'
      TITLE(7) = 'STANDARDIZED RESIDUALS VS. PREDICTOR'
      TITLE(8) = 'ORIGINAL RESIDUALS VS. PREDICTOR'
      TITLE(9) = 'OBSERVED RESPONSE VS. OBSERVATION NUMBER'
      TITLE(10) = 'PREDICTOR VS. OBSERVATION NUMBER'
      TITLE(11) = 'EXTRA TITLE'
      XTITLE(1) = 660.
      YTITLE(1) = 980.
      ATITLE(1) = 0.
      HTITLE(1) = 10.
      XTITLE(2) = 650.
      YTITLE(2) = 950.
      ATITLE(2) = 0.
      HTITLE(2) = 10.
      DO 100 I=3,11
      XTITLE(I) = 660.
      YTITLE(I) = 930.
      ATITLE(I) = 0.
  100 HTITLE(I) = 10.
      DO 110 I=1,11
      HSMBL(I) = 5.
      ASMBL(I) = 0.
      XSMBL(I) = 0.
      YSMBL(I) = 0.
  110 ISMBL(I) = I + 2
      NSMBL = 0
      NCHAR = 0
      NTITLE = 3
      XLENGT = 800.
      YLENGT = 800.
      XXIN = 50.
      XYIN = 100.
      YXIN = 50.
      YYIN = 100.
C
C NOTE THAT THERE IS NO DEFAULT VALUE FOR ICTRL(K,1).  IT MUST BE
C ASSIGNED IN THE CALLING PROGRAM
C
      DO 120 K=1,NCURVE
      ICTRL(K,2) = 0
      IF(ICTRL4.EQ.8.OR.ICTRL4.EQ.9.OR.ICTRL4.EQ.2.OR.ICTRL4.EQ.3)
     &ICTRL(K,2) = 1
      ICTRL(K,3) = 1
      IF(ICTRL(K,2).EQ.1) ICTRL(K,3) = 0
      ICTRL(K,4) = K + 2
  120 ICTRL(K,5) = 10
```

```
      IF(NCURVE.EQ.5) GO TO 140
      DO 130 J=1,5
      DO 130 I=(NCURVE+1),5
  130 ICTRL(I,J) = 0
  140 XAXIS = 'OBSERVED RESPONSE'
      YAXIS = 'STANDARDIZED RESIDUALS'
      IF(ICTRL4.EQ.4.OR.ICTRL4.EQ.5) XAXIS = 'PREDICTED RESPONSE'
      IF(ICTRL4.EQ.2.OR.ICTRL4.EQ.3.OR.ICTRL4.EQ.8.OR.ICTRL4.EQ.9)
     & XAXIS = 'OBSERVATION NUMBER'
      IF(ICTRL4.EQ.6.OR.ICTRL4.EQ.7) XAXIS = 'PREDICTOR'
      IF(ICTRL4.EQ.3.OR.ICTRL4.EQ.5.OR.ICTRL4.EQ.7)
     &YAXIS = 'RAW RESIDUALS'
      IF(ICTRL4.EQ.8) YAXIS = 'OBSERVED RESPONSE'
      IF(ICTRL4.EQ.9) YAXIS = 'PREDICTOR'
C
C*************************************************************
C*************************************************************
  200 IF(ICTRL3.EQ.0) GO TO 210
      PRINT
      PRINT,' ENTER IBP = 0 TO RETAIN NCURVE, NTITLE, NSMBL, AND NCHAR'
      PRINT
      PRINT,'      1I FORMAT'
      READ,IBP
      IF(IBP.EQ.0) GO TO 210
      PRINT,'      ENTER NCURVE, NTITLE, NSMBL, NCHAR'
      PRINT,'      CURRENT VALUES: NCURVE =',NCURVE
      PRINT,'                      NTITLE =',NTITLE
      PRINT,'                      NSMBL  =',NSMBL
      PRINT,'                      NCHAR  =',NCHAR
      PRINT
      PRINT,'      4I FORMAT'
      READ,NCURVE,NTITLE,NSMBL,NCHAR
C CALCULATE NPTS
  210 NPTS = 0
      DO 220 I=1,NCURVE
  220 NPTS = ICTRL(I,1) + NPTS
C INITIALIZE THE SCREEN
      CALL CCS_$PLOTS
      IF(ICTRL3.NE.0) GO TO 230
C THE AXES ARE SCALED USING ALL THE POINTS
      CALL CCS_$SCALE(A2,XLENGT,NPTS,1)
      CALL CCS_$SCALE(B2,YLENGT,NPTS,1)
C THE SCALING VALUES ARE RETURNED IN A2 AND B2
      XDELTA=A2(NPTS+2)
      YDELTA=B2(NPTS+2)
      XFIRST=A2(NPTS+1)
```

```
      YFIRST=B2(NPTS+1)
  230 IF(ICTRL3.EQ.0)GO TO 270
      PRINT
      PRINT,'     ENTER, IS = 0 TO RETAIN AXIS SCALES'
      PRINT,'            IO = 0 TO RETAIN AXIS LOCATIONS'
      PRINT,'            IL = 0 TO RETAIN AXIS LABELS'
      PRINT,'            IG = 0 TO RETAIN AXIS LENGTHS'
      PRINT
      PRINT,'     4I FORMAT'
      READ,IS,IO,IL,IG
      IF(IS.EQ.0) GO TO 240
      PRINT,'     ENTER XDELTA, YDELTA, XFIRST, YFIRST'
      PRINT,'     CURRENT VALUES: XDELTA =',XDELTA
      PRINT,'                     YDELTA =',YDELTA
      PRINT,'                     XFIRST =',XFIRST
      PRINT,'                     YFIRST =',YFIRST
      PRINT
      PRINT,'     4F FORMAT'
      READ,XDELTA,YDELTA,XFIRST,YFIRST
  240 IF(IO.EQ.0) GO TO 250
      PRINT,'     ENTER XXIN, XYIN = COORDINATES OF BEGINNING OF X AXIS'
      PRINT,'           YXIN, YYIN = COORDINATES OF BEGINNING OF Y AXIS'
      PRINT,'     CURRENT VALUES: XXIN =',XXIN
      PRINT,'                     XYIN =',XYIN
      PRINT,'                     YXIN =',YXIN
      PRINT,'                     YYIN =',YYIN
      PRINT
      PRINT,'     4F FORMAT'
      READ,XXIN,XYIN,YXIN,YYIN
  250 IF(IL.EQ.0) GO TO 260
      PRINT,'     ENTER XAXIS = X AXIS LABEL'
      PRINT,'           YAXIS = Y AXIS LABEL'
      PRINT,'     CURRENT VALUES: XAXIS =',XAXIS
      PRINT,'                     YAXIS =',YAXIS
      PRINT
      PRINT,'     A50,/,A50 FORMAT, ONE LABEL PER LINE'
      READ 900,XAXIS,YAXIS
  900 FORMAT(A50,/,A50)
  260 IF(IG.EQ.0) GO TO 270
      PRINT,'     ENTER XLENGT = LENGTH OF XAXIS'
      PRINT,'           YLENGT = LENTH OF Y AXIS'
      PRINT,'     CURRENT VALUES: XLENGT =',XLENGT
      PRINT,'                     YLENGT =',YLENGT
      PRINT
      PRINT,'     2F FORMAT'
      READ,XLENGT,YLENGT
```

```
 270  CALL CCS_$AXIS(XXIN,XYIN,XAXIS,50,XLENGT,0.,XFIRST,XDELTA)
      CALL CCS_$AXIS(YXIN,YYIN,YAXIS,-50,YLENGT,90.,YFIRST,YDELTA)
      IF(NSMBL.EQ.0)GO TO 310
      DO 300 I=1,NSMBL
      IF(ICTRL3.EQ.0)GO TO 300
      PRINT
      PRINT,'     ENTER IHC = 0 TO RETAIN COORDINATES OF SYMBOL',I
      PRINT,'           IHA = 0 TO RETAIN HEIGHT AND ANGLE OF SYMBOL',I
      PRINT,'           IHS = 0 TO RETAIN THE SAME SYMBOL CHARACTER'
      PRINT
      PRINT,'     3I FORMAT'
      READ,IHC,IHA,IHS
      IF(IHC.EQ.0) GO TO 280
      PRINT,'     ENTER XSMBL(I),YSMBL(I) = COORDINATES OF SYMBOL',I
      PRINT,'     CURRENT VALUES: XSMBL(I) =',XSMBL(I)
      PRINT,'                     YSMBL(I) =',YSMBL(I)
      PRINT
      PRINT,'     2I FORMAT'
      READ,XSMBL(I),YSMBL(I)
 280  IF(IHA.EQ.0) GO TO 290
      PRINT,'     ENTER, HSMBL(I) = HEIGHT OF SYMBOL I'
      PRINT,'            ASMBL(I) = ANGLE OF SYMBOL I'
      PRINT,'     CURRENT VALUES: HSMBL(I) =',HSMBL(I)
      PRINT,'                     ASMBL(I) =',ASMBL(I)
      PRINT
      PRINT,'     2I FORMAT'
      READ,HSMBL(I),ASMBL(I)
 290  IF(IHS.EQ.0) GO TO 300
      PRINT,'     ENTER ISMBL(I) = CODE FOR SYMBOL',I
      PRINT,'     CURRENT VALUE: ISMBL(I) =',ISMBL(I)
      PRINT
      PRINT,'     1I FORMAT'
      READ,ISMBL(I)
 300  CALL CCS_$SYMBOL(XSMBL(I),YSMBL(I),HSMBL(I),ISMBL(I),ASMBL(I),0)
 310  IF(NTITLE.EQ.0)GO TO 370
      DO 340 I=1,NTITLE
C*********************************************************************
C THE USE OF J AND ICTRL4 IS APPLICATION SPECIFIC.
C 1 IS ADDED TO ICTRL4 BEAUSE OF THE NUMBERING SYSTEM
C USED IN SUBROUTINE EXEQU OF PROGRAM TEST7 FOR THE
C VARIOUS PLOTS WHICH CAN BE CALLED FROM TEST7.
C
      J = I
      IF(I.GE.3) J = ICTRL4 + 1
C*********************************************************************
      IF(ICTRL3.EQ.0)GO TO 340
```

```
      PRINT
      PRINT,'     ENTER ITC = 0 TO RETAIN TITLE(I), I=',J
      PRINT,'           ITH = 0 TO RETAIN HEIGHT AND ANGLE OF TITLE(I)'
      PRINT,'           ITL = 0 TO RETAIN LOCATION OF TITLE(I)'
      PRINT
      PRINT,'     3I FORMAT'
      READ,ITC,ITH,ITL
      IF(ITC.EQ.0) GO TO 320
      PRINT,'     ENTER TITLE(I), I =',J
      PRINT,'     CURRENT VALUE: TITLE(I) =',TITLE(J)
      PRINT
      PRINT,'      (A50) FORMAT'
      READ 910,TITLE(J)
  910 FORMAT(A50)
  320 IF(ITH.EQ.0) GO TO 330
      PRINT,'     ENTER, HTITLE(I) = HEIGHT OF TITLE',J
      PRINT,'            ATITLE(I) = ANGLE OF TITLE',J
      PRINT,'     CURRENT VALUES: HTITLE(I) =',HTITLE(J)
      PRINT,'                     ATITLE(I) =',ATITLE(J)
      PRINT
      PRINT,'     2I FORMAT'
      READ,HTITLE(J),ATITLE(J)
  330 IF(ITL.EQ.0) GO TO 340
      PRINT,'     ENTER XTITLE(I),YTITLE(I) = COORDINATES OF TITLE',J
      PRINT,'     CURRENT VALUES: XTITLE(I) =',XTITLE(J)
      PRINT,'                     YTITLE(I) =',YTITLE(J)
      PRINT
      PRINT,'     2I FORMAT'
      READ,XTITLE(J),YTITLE(J)
  340 CALL CCS_$SYMBOL(XTITLE(J),YTITLE(J),HTITLE(J),TITLE(J),
     &                 ATITLE(J),50)
      IF(ICTRL3.EQ.0) GO TO 370
      DO 360 I=1,NCURVE
      PRINT
      PRINT,' ENTER INPTS = 0 TO RETAIN NUMBER OF POINTS IN CURVE',I
      PRINT,'       ICV   = 0 TO RETAIN ICTRL PARAMETERS FOR CURVE',I
      PRINT
      PRINT,'     2I FORMAT'
      READ,INPTS,ICV
      IF(INPTS.EQ.0) GO TO 350
      PRINT,'     ENTER ICTRL(I,1) = NUMBER OF POINTS IN CURVE',I
      PRINT,'     CURRENT VALUE: ICTRL(I,1) =',ICTRL(I,1)
      PRINT
      PRINT,'     1I FORMAT'
      READ,ICTRL(I,1)
  350 IF(ICV.EQ.0) GO TO 360
```

```
      PRINT,'     ENTER COL 2 = 0 TO NOT CONNECT POINTS'
      PRINT,'           COL 3 = 0 TO NOT PLOT SYMBOLS'
      PRINT,'                   OTHERWISE, THE INTERVAL BETWEEN POINTS'
      PRINT,'                   TO BE MARKED WITH SYMBOLS'
      PRINT,'           COL 4 = SYMBOL CODE (FOR COL 3 = 0)'
      PRINT,'           COL 5 = SYMBOL SIZE (FOR COL 3 = 0)'
      PRINT,'     CURRENT VALUES: ICTRL(I,2) =',ICTRL(I,2)
      PRINT,'                     ICTRL(I,3) =',ICTRL(I,3)
      PRINT,'                     ICTRL(I,4) =',ICTRL(I,4)
      PRINT,'                     ICTRL(I,5) =',ICTRL(I,5)
      PRINT
      PRINT,'     4I FORMAT'
      READ,ICTRL(I,2),ICTRL(I,3),ICTRL(I,4),ICTRL(I,5)
  360 CONTINUE
  370 CALL CCS_$PLOT(YXIN,YYIN,-3)
      XDELTA=XDELTA/100.
      YDELTA=YDELTA/100.
C*****************************************************************
      PRINT
      PRINT,' ENTER ILN = 1 TO DRAW HORIZONTAL LINES ON PLOTS'
      PRINT
      PRINT,'     1I FORMAT'
      READ,ILN
      IF(ILN.NE.1) GO TO 390
C DRAW SOME LINES ON THE RESIDUAL PLOT GRAPHS BEFORE PLOTTING
C THE CURVES
C
C DRAW A LINE AT Y = 0 FOR ALL RESIDUAL PLOTS
C
      IF(ICTRL4.GE.8) GO TO 390
      A(1) = XFIRST
      A(2) = XFIRST + XDELTA*800.
      B(1) = 0.
      B(2) = 0.
      NPTS = 2
      CALL SCLGPH(A,B,NPTS,0.,INTEQV,XFIRST,XDELTA,YFIRST,YDELTA)
C
C DRAW LINES AT Y = _1,_2,_3 FOR STANDARDIZED RESIDUAL PLOTS
C
      IF(ICTRL4.NE.2.AND.ICTRL4.NE.4.AND.ICTRL4.NE.6) GO TO 390
      DO 380 I1=1,3
      DO 380 I2=1,2
      B(1) = I1
      B(2) = I1
      IF(I2.EQ.2) B(1) = -I1
      IF(I2.EQ.2) B(2) = -I1
```

```
  380 CALL SCLGPH(A,B,NPTS,0.,INTEQV,XFIRST.XDELTA.YFIRST.YDELTA)
C
C*****************************************************************
C
C NOW PLOT THE CURVES
C
  390 LPTS = 0
      DO 410 I=1,NCURVE
      J1=LPTS+1
      J2=LPTS+ICTRL(I,1)
      LPTS = LPTS + ICTRL(I,1)
      IF(ICTRL(I,2).EQ.0)GO TO 410
      ICOUNT = 0
      DO 400 J=J1,J2
      ICOUNT=ICOUNT+1
      A(ICOUNT)=A2(J)
  400 B(ICOUNT)=B2(J)
      CALL SCLGPH(A,B,ICTRL(I,1),0.,INTEQV,XFIRST.XDELTA.YFIRST.YDELTA)
  410 CONTINUE
C
C THE SYMBOLS ARE PLOTTED
C
      LPTS=0
      DO 430 I=1,NCURVE
      J1=LPTS+1
      J2=LPTS+ICTRL(I,1)
      LPTS = LPTS + ICTRL(I,1)
      IF(ICTRL(I,3).EQ.0)GO TO 430
      ICOUNT=0
      DO 420 J=J1,J2,ICTRL(I,3)
      ICOUNT=ICOUNT+1
      A(ICOUNT)=A2(J)
  420 B(ICOUNT)=B2(J)
      NPTS1=ICOUNT*(-1)
      CALL SCLGPH(A,B,NPTS1,FLOAT(ICTRL(I,5)),ICTRL(I,4),XFIRST.XDELTA,
     &            YFIRST,YDELTA)
  430 CONTINUE
      CALL CCS_$PLOT(0.,0.,33)
C
C ILMJK IS A DUMMY VARIABLE WHICH MUST BE ENTERED AT THE TERMINAL
C BEFORE PROCEEDING.  THE PURPOSE OF THE FOLLOWING LINE
C IS TO ALLOW PRODUCTION OF A CLEAN COPY OF A GRAPH WHEN AT A
C SCOPE TERMINAL SUCH AS THE TEK 4015.
C
      READ,ILMJK
      REWIND 25
```

```
      IF(NCHAR.EQ.0) GO TO 450
      DO 440 I=1,NCHAR
      READ(25,920)CHAR
  920 FORMAT(A50)
  440 WRITE(6,930)CHAR
  930 FORMAT(10X,A50)
  450 PRINT
      PRINT,'     ENTER ICTRL5 = 0 IF FINISHED WITH PLOT'
      PRINT
      PRINT,'     1I FORMAT'
      READ,ICTRL5
      IF(ICTRL5.EQ.0) RETURN
      ICTRL3 = 1
      XDELTA = XDELTA*100.
      YDELTA = YDELTA*100.
      GO TO 200
      END
```

Appendix C

Consistent System Macros for Interactive Stepwise Regression

C.1 Introduction

The 5 programs described in this Appendix were written for use on the Consistent System (CS) as implemented on the Multics operating system on the Honeywell 6800 computer facility at the Massachusetts Institute of Technology. The programs comprise CS commands and are called CS macros.

BSR, ENT_VAR, and DEL_VAR are used to perform interactive stepwise multiple least squares regression. CALC_RES, PQS_RES, and NM_PLOT are used for analysis of a developed model. The input arguments for and restrictions on the use of these macros are explained in Section C.2. The macros are listed in Section C.3. CS documentation should be consulted for further information.

C.2 Using the Macros

C.2.1 General Information

Runnable macros must be created from the files mentioned above before they can be used. The runnable macros should be given names which are different from, but similar to, the original file names. An example of the necessary command is

```
make_macro file filem
```

where make_macro is the CS command, file is the name of the file which contains the text of the macro, and filem is the name of the runnable macro. The macro is then invoked by typing filem followed by the argument list. The arguments do not need to have the same names as in the macro text. They are recognized by position in the argument string.

Files named input, cpm, and sw will appear in the working directory when these macros are used. They may be deleted when the work is completed.

C.2.2 Adding and Removing Variables

BSR, ENT_VAR, and DEL_VAR are used to add and remove variables from a model.

BSR prepares a means and cross-products matrix for use in the macros ENT_VAR and DEL_VAR. BSR must be called before any of the other macros are called and then is not called again until a new data set is used. The input arguments are PR and DTMX. When PR = 0 the coefficient matrix is printed at the terminal. When PR is positive nothing is printed. The coefficient matrix printed when PR = 0 includes coefficients, degrees of freedom for the coefficients, F statistics, and significance levels for the F statistics. DTMX is a labeled genarray file in which the dependent variable is in the last column. Some of the macros will not work unless the variables are labeled. DTMX is also an input argument for the model analysis macros.

ENT_VAR enters specified variables into the model using the CS routine QSWEEP. The input arguments are PR followed by a list of variables to be entered. PR has the same function as in BSR. The variables are identified by column number in DTMX.

DEL_VAR removes variables from the model using the CS routine QRSWEEP. The input arguments are the same as for ENT_VAR.

C.2.3. Model Analysis

CALC_RES, PQS_RES, and NM_PLOT calculate model residuals and produce information from those residuals.

CALC_RES produces 2 files containing residuals. 'fit' contains the residuals and associated predicted responses. 'seq' contains residuals in the sequence in which the data was given in DTMX. The various CS plotting routines may be used to produce graphs from these files. They may also be printed. The Durbin-Watson statistic is calculated and printed at the terminal. DTMX is the input argument.

PQS_RES plots the partial residuals of the dependent variable against the partial residuals of a variable which is not yet entered into the model. A line with the slope of the coefficient of that variable, were it entered in the model along with the variables already in the model, is also plotted on the graph. PQS_RES is set up for use at graphics terminals but may easily be changed to work at line printers by changing the plotting calls. The input arguments are DTMX and VARNBR, the number of the independent variable.

NM_PLOT plots the residuals against a normal cumulative probability distribution. NM_PLOT is also set up for use at a graphics terminal and may also be changed to work at a line printer by changing the plotting call. The input argument for NM_PLOT is DTMX.

C.3 Macro Listings

The 6 macros described in this section are listed on the following pages.

```
&c BSR
&c
&c      This macro prepares a means and crossproducts
&c      matrix for use in an interactive stepwise regression
&c      using qsweep and qrsweep.  It also prints the regression with
&c      only the constant term entered.
&c      pr = 0 to print the coefficient matrix
&c      pr = any positive number to not print the coefficient matrix
&c      dtmx = the data matrix with the response in the last column.
&p pr dtmx
eval:a cpm:=crossp:x(dtmx)
copy_file:a cpm sw
&if pr end
eval:a rgqsig:x(cpm,cpm) print coefs,rsq
&label end


&c ENT_VAR
&c
&c      This macro enters variables into the regression model
&c      using qsweep.  cpm must be prepared by crossp, sw is overwritten
&c      Input arguments: pr followed by a list of variables to be entered, where:
&c      pr = 0 to print the coefficient matrix
&c      pr = any positive number to not print the coefficient matrix
&a directions input
&t pr vnbr
concatenate:a "make_attribute:a " input " vnbr" input
run:a input
eval:a pr:=subset:a(vnbr,"(1=1)")
eval:a vnbr:=subset:a(vnbr,"(1^=1)")
cm:a vnbr:=vnbr+1
eval:a sw:=qsweep:a(sw with swa(vnbr$))
&if pr end
eval:a rgqsig:x(sw,cpm) print coefs,rsq
&label end
```

```
&c NM_PLOT
&c
&c     This macro produces a normal plot of the residuals from
&c     a regression model developed with the CS sweep operators.
&c     dtmx = complete data matrix with the response in the last column
&c     cpm = means and cross products matrix from crossp
&c     sw = current swept matrix from qwseep or qrsweep
&p dtmx
&t res coefs jres
eval:a coefs:=rgqsig:x(sw,cpm)
eval:a res:=residuals:x(dtmx,coefs)
eval:a jres:=extract_attr:x(res with attr(3))
eval:a norm_plot:x(jres) print using plot1:a


&c PQS_RES
&c
&c     This macro creates a partial residual plot for a variable not
&c     already entered in the model.  This is a companion
&c     macro for bsr, ent_var, and del_var.
&c
&c     dtmx = the complete data matrix with the response in the last column
&c     varnbr = the variable number for which the partial residual
&c              plot is to be created.
&c     cpm = means and cross products matrix from crossp
&c     sw = current swept matrix from qsweep or qrsweep
&p dtmx varnbr
&t res1 coefs1 res2 coefs2 b rall presx presy prmtx res3 fit
cm:a varnbr:=varnbr+1
eval:a coefs1:=rgqsig:x(sw,cpm)
eval:a res1:=residuals:x(dtmx,coefs1)
eval:a sw:=qsweep:a(sw with swa(varnbr$))
eval:a coefs2:=rgqsig:x(sw,cpm)
eval:a res2:=residuals:x(dtmx,coefs2)
eval:a b:=subset:a(coefs2,"(1=varnbr$)","(2=1)")
eval:a rall:=extract_attr:x(res2 with attr(3))
eval:a presy:=extract_attr:x(res1 with attr(3))
cm:a presx:=(presy-rall)/b
eval:a res3:=rgattr:x(presy on presx ret(residuals))
eval:a fit:=extract_attr:x(res3 with attr(1))
plot2:a presx presy
plot1:a presx fit -noerase
eval:a sw:=qrsweep:a(sw with swa(varnbr$))
```

```
&c DEL_VAR
&c
&c     This macro deletes variables from a regression model.
&c     It is a companion macro for ent_varm.
&a directions input
&t pr vnbr
concatenate:a "make_attribute:a " input " vnbr" input
run:a input
eval:a pr:=subset:a(vnbr,"(1=1)")
eval:a vnbr:=subset:a(vnbr,"(1^=1)")
cm:a vnbr:=vnbr+1
eval:a sw:=qrsweep:a(sw with swa(vnbr$))
&if pr end
eval:a rgqsig:x(sw,cpm) print coefs,rsq
&label end


&c CALC_RES
&c
&c     This macro is used to calculate the residuals of a
&c     model created with the sweep operators.  The residuals.
&c     coefficient matrix, R squared, vcvcf, ANOVA, seq, fit.
&c     and the Durbin-Watson statistic are calculated.
&c     seq and fit are matrices for plotting the residuals in
&c     sequence and against the fitted response.
&c     A file named days with consecutive numbers for
&c     each observation should be created before
&c     this macro is used.
&p dtmx
&t r1
eval:a coefs:=rgqstats:x(sw,cpm ret(coefs) save(rsq:=rsq) save(vcvcf:=vcvcf) save(anov:=anov))
eval:a res:=residuals:x(dtmx,coefs)
eval:a fit:=extract_attr:x(res with attr(2))
eval:a r1:=extract_attr:x(res with attr(3))
eval:a fit:=dtmx_join:x(fit,r1)
eval:a seq:=dtmx_join:x(days,r1)
eval:a durbin_watson:x(r1) print
```