TWO CASE STUDIES IN MODEL FITTING

by

Bradley Wilson Thomas

B.Sc., Simon Fraser University, 1984

PROJECT SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in the Department

of

Mathematics & Statistics

SIMON FRASER UNIVERSITY

September 1986

APPROVAL

Name: Bradley Wilson Thomas

Degree: Master of Science

Title of project:

Two case studies in model fitting

Examining Committee:

Chairman: Dr. A. Freedman

Dr. D. Eaves
Senior Supervisor

Dr. M. A. Stephens
Supervisor

Prof. C. Villegas
Supervisor

Dr. A. Harestad
External Examiner
Faculty of Biological Sciences
Simon Fraser University

Date Approved: September 18, 1986

PARTIAL COPYRIGHT LICENSE

I hereby grant to Simon Fraser University the right to lend my thesis, project or extended essay (the title of which is shown below) to users of the Simon Fraser University Library, and to make partial or single copies only for such users or in response to a request from the library of any other university, or other educational institution, on its own behalf or for one of its users. I further agree that permission for multiple copying of this work for scholarly purposes may be granted by me or the Dean of Graduate Studies. It is understood that copying or publication of this work for financial gain shall not be allowed without my written permission.

Title of Thesis/Project/Extended Essay

Two Case Studies in Model Fitting

Author:

(signature)

(name)

23 Sept. 1986

(date)

ABSTRACT

This project consists of analyses and model fittings for two different studies. Both studies arose out of M.Sc. research by graduate biology students.

In the first data set, the object was to determine whether or not an association exists between nesting colony size (or any other covariates) and rates at which food is carried to chicks. "Best k subsets" regression was used with a standard normal theory model. In addition a loglinear model was investigated for comparison.

In the second data set, the object was to determine how the proportion of deciduous trees bearing an active nest might be associated with characteristics of that tree, and how the explanatory variables in such a model may be ranked in importance. Linear logistic regression was pursued using BMDPLR and GLIM.

DEDICATION

To my parents, Connie & Len, for their unfailing love & support.

"*Charity* (love) *suffereth long, and is kind;...Charity never faileth: but whether there be prophecies, they shall fail;...whether there be knowledge, it shall vanish away.*"

1 Corinthians 13:4,8.

ACKNOWLEDGEMENTS

I wish to thank all the staff, faculty, and fellow graduate students in the Mathematics & Statistics department for making my M.Sc. work as pleasant as it was.

In particular, I wish to express my gratitude to my Senior Supervisor Dr. David Eaves for his availability, guidance, and willingness to help. I also wish to thank him for bringing the GLIM software package to SFU.

I wish also to acknowledge Professor Cesareo Villegas who, in addition to serving on my graduate committee, also recommended me into both the M.Sc. program and the NSERC Graduate Scholarship, which funded my entire M.Sc. program.

I would also like to thank the following computing centre staff for helping me to become more proficient in some much needed software: Mr. Reo Audette (BMDP & MINITAB), Mark the duty consultant (TEXTFORM & MTS in general), and Ms. Margaret Sharon (TEXTFORM).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PART A

MODELLING THE FEEDING RATES OF PIGEON GUILLEMOT CHICKS

CHAPTER 1

THE PROBLEM

Simon Emms, graduate student in the Dept. of Biological Sciences, presented data collected over two summers which dealt with the feeding of Pigeon Guillemot chicks by parent birds. The data was collected on Mitlenatch Island (near Campbell River, B.C.) and the Pigeon Guillemot is a sea bird species, so the diet of the chicks was one of fish. In a 2 hour period the number of fish delivered to chicks in a given nest by the parents was observed & the lengths of the fish estimated.

It was of interest to know if the feeding rate was influenced by the size of colony in which the observed nest was. This colony size is expressed in total number of nests in the colony. The interest in colony size arose out of the 'Information Centre Hypothesis', that is, the proposition that colony members can learn of the location of good feeding sites by following successful foragers, and can thereby increase their own foraging success. This effect would be greater in larger colonies.

Also of interest was whether or not any of the other measured variables exerted any influence on the feedrate. The measured variables contained both qualitative and quantitative effects.

The results of the analysis are to be applied to the population of Pigeon Guillemot birds in general. As it turned out, a final model (to be identified as the 'current best model' in Chapter 3) was obtained, but it did not contain colony size. Insofar as the sample data can be regarded as a random sample and representative of the population, this suggests that colony size is not associated with chick feeding rate.

Feeding rate does, however, seem to be positively associated with the number of chicks in the nest. For example, if the number of chicks is increased from 1 to 2, the model predicts that feeding rate should increase by 0.243 fish per hour. In addition feeding rate seems to be associated with both age of chicks and time of day in a changing pattern: in the case of chick age, feeding rate increases from some initial value at time of hatching until the chick is 27 days old, after which the feeding rate decreases. In contrast, the association with time of day is concave upward, i.e. feeding rate decreases from some initial value at dawn, reaches a minimum at 1330H, and then increases until dusk. Of course this does not suggest that Pigeon Guillemots continually feed their chicks while there is daylight, this is merely an observation in the trend in feeding rate averaged over all birds in the sample. The details of this aspect of the analysis can be found in Section 5.2.

Model building was the primary goal of the analysis, but a detailed inference was done in the case of difference in feeding rates between dawn and 1330H, the point of minimum feeding rate with respect to time of day as fitted by the model. Such inference is based on the validity of regarding the sample data as a representative and random sample from the target population of Pigeon Guillemot birds in general. It turned out that a 95% confidence interval for this difference in feeding rates between dawn at, say, 0530H and 1330H had a lower bound of 0.229 fish per hour (ignoring effects due to age) and an upper bound of 0.579 fish per hour. Details can be found in Section 5.4.

The client report is found in Chapters 1-5 and the technical supplements in Chapters 6-10 with Chapter 6 providing a technical supplement for Chapter 1 contents, and so on.

CHAPTER 2

THE DATA

Simon permitted copies of his 1984 and 1985 data files. A portion each of these files is shown in Tables 1 and 2 respectively, where there is one record for each observation. The record format is as follows.

	Variable	Column Range
(1)	Date	1-4
(2)	Colony Identification Number	6-7
(3)	Number of Nests Observed in the Colony	10-11
(4)	Colony Size	14-15
(5)	Tidal State Code	17-18
(6)	Time of Day	20-24
(7)	Accept/Reject Code	26-27
(8)	Nest Identification Number	29-31
(9)	Number of Chicks in Nest	34-36
(10)	Age of Chicks	38-40
(11)	Number of Fish Delivered--Species Type 1	41-43
(12)	Total Length of Fish Delivered--Species Type 1	45-49
(13)	Number of Fish Delivered--Species Type 2	50-52
(14)	Total Length of Fish Delivered--Species Type 2	54-58
(15)	Number of Fish Delivered--Species Type 3	59-61

Table 1: Portion of 1984 Raw Data

7.15	1.	9.	10.	2.	1000.	1.	18.	-1.	22.	2.	6.00	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.15	1.	9.	10.	2.	1000.	1.	24.	2.	9.	0.	0.0	0.	0.0	0.	0.0	0.	0.0	1.	2.00	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.15	1.	9.	10.	2.	1000.	1.	30.	-1.	10.	0.	0.0	0.	0.0	0.	0.0	0.	0.0	1.	2.50	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.15	1.	9.	10.	2.	1000.	1.	7.	1.	12.	0.	0.0	0.	0.0	0.	0.0	0.	0.0	1.	3.25	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.15	1.	9.	10.	2.	1000.	1.	28.	2.	12.	0.	0.0	1.	2.50	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.15	1.	9.	10.	2.	1000.	1.	-1.	-1.	-1.	1.	3.00	0.	0.0	0.	0.0	0.	0.0	1.	2.50	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.15	1.	9.	10.	2.	1000.	1.	9.	-1.	18.	2.	5.75	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.15	1.	9.	10.	2.	1000.	1.	5.	1.	6.	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.15	1.	9.	10.	2.	1000.	1.	8.	-1.	18.	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.15	1.	9.	10.	2.	1000.	1.	49.	-1.	3.	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.15	1.	9.	10.	1.	1300.	1.	30.	-1.	10.	1.	3.25	2.	4.25	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.15	1.	9.	10.	1.	1300.	1.	8.	-1.	18.	0.	0.0	0.	0.0	0.	0.0	0.	0.0	1.	2.00	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.15	1.	9.	10.	1.	1300.	1.	24.	2.	9.	1.	3.00	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.15	1.	9.	10.	1.	1300.	1.	7.	1.	12.	1.	2.00	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.15	1.	9.	10.	1.	1300.	1.	28.	2.	12.	1.	2.50	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.15	1.	9.	10.	1.	1300.	1.	5.	1.	6.	0.	0.0	0.	0.0	0.	0.0	0.	0.0	1.	2.00	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.15	1.	9.	10.	1.	1300.	1.	18.	-1.	22.	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.15	1.	9.	10.	1.	1300.	1.	49.	-1.	3.	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.15	1.	9.	10.	1.	1300.	1.	9.	-1.	18.	2.	6.25	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.15	1.	9.	10.	1.	1300.	1.	-1.	-1.	-1.	0.	0.0	2.	4.00	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.15	1.	9	10.	3.	1830.	1.	28.	2.	12.	2.	5.25	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	1.	2.50
7.15	1.	9.	10.	3.	1830.	1.	7.	1.	12.	1.	3.75	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.15	1.	9.	10.	3.	1830.	1.	30.	-1.	10.	0.	0.0	0.	0.0	0.	0.0	0.	0.0	1.	2.00	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.15	1.	9.	10.	3.	1830.	1.	49.	-1.	3.	1.	2.75	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.15	1.	9.	10.	3.	1830.	1.	24.	2.	9.	0.	0.0	1.	2.00	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.15	1.	9.	10.	3.	1830.	1.	8.	-1.	18.	0.	0.0	0.	0.0	0.	0.0	0.	0.0	1.	2.00	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.15	1.	9.	10.	3.	1830.	1.	-1.	-1.	-1.	2.	6.00	0.	0.0	0.	0.0	0.	0.0	1.	2.50	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.15	1.	9.	10.	3.	1830.	1.	9.	-1.	18.	1.	3.50	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.15	1.	9.	10.	3.	1830.	1.	18.	-1.	22.	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.15	1.	9.	10.	3.	1830.	1.	5.	1.	6.	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.16	1.	9.	10.	3.	700.	3.	28.	2.	12.	1.	2.75	2.	4.75	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	1.	2.50
7.16	1.	9.	10.	3.	700.	3.	7.	1.	12.	0.	0.0	1.	2.00	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.16	1.	9.	10.	3.	700.	3.	30.	-1.	10.	1.	2.50	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.16	1.	9.	10.	3.	700.	3.	8.	-1.	18.	1.	3.00	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.16	1.	9.	10.	3.	700.	3.	24.	2.	9.	0.	0.0	1.	2.00	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.16	1.	9.	10.	3.	700.	3.	9.	-1.	19.	0.	0.0	2.	4.75	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.16	1.	9.	10.	3.	700.	3.	5.	1.	7.	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.16	1.	9.	10.	3.	700.	3.	18.	-1.	23.	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.16	1.	9.	10.	3.	700.	3.	49.	-1.	4.	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.16	1.	9.	10.	2.	1030.	3.	49.	-1.	4.	0.	0.0	1.	3.00	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.16	1.	9.	10.	2.	1030.	3.	30.	-1.	11.	1.	2.50	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.16	1.	9.	10.	2.	1030.	3.	-1.	-1.	-1.	0.	0.0	1.	3.00	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.16	1.	9.	10.	2.	1030.	3.	9.	-1.	19.	0.	0.0	0.	0.0	0.	0.0	0.	0.0	1.	2.50	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.16	1.	9.	10.	2.	1030.	3.	5.	1.	7.	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.16	1.	9.	10.	2.	1030.	3.	7.	1.	13.	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.16	1.	9.	10.	2.	1030.	3.	8.	-1.	19.	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.16	1.	9.	10.	2.	1030.	3.	18.	-1.	23.	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.16	1.	9.	10	2.	1030.	3.	28.	2.	13.	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.16	1.	9.	10.	2.	1030.	3.	24.	2.	10.	2.	5.25	1.	2.50	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.16	1.	9.	10.	3.	1830.	1.	49.	-1.	4.	0.	0.0	0.	0.0	0.	0.0	0.	0.0	1.	3.00	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.16	1.	9.	10.	3.	1830.	1.	18.	-1.	23.	0.	0.0	0.	0.0	1.	3.50	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.16	1.	9.	10.	3	1830.	1.	28.	2.	13.	1.	3.25	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	1.	3.25

Table 2: Portion of 1985 Raw Data

7.20	1.	2.	2.	1.	1300.	1.	28.	2.	17.	1.	4.00	1.	3.75	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.20	1.	2.	2.	1.	1300.	1.	76.	-1.	13.	0.	0.0	1.	2.75	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.20	1.	2.	2.	1.	1830.	1.	28.	2.	17.	1.	3.75	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.20	1.	2.	2.	1.	1830.	1.	76.	-1.	13.	0.	0.0	0.	0.0	1.	3.00	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.21	1.	2.	2.	3.	600.	1.	28.	2.	18.	1.	3.50	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.21	1.	2.	2.	3.	600.	1.	76.	-1.	14.	1.	3.25	2.	2.50	0.	0.0	0.	0.0	0.	0.0	0.	0.0	1.	2.00	0.	0.0	0.	0.0
7.21	4.	1.	1.	2.	1000.	1.	102.	2.	15.	0.	0.0	1.	2.50	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.21	1.	2.	2.	3.	1800.	1.	28.	2.	18.	0.	0.0	1.	2.50	0.	0.0	0.	0.0	1.	4.00	0.	0.0	1.	2.50	0.	0.0	0.	0.0
7.21	1.	2.	2.	3.	1800.	1.	76.	-1.	14.	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.22	4.	1.	1.	3.	830.	1.	102.	2.	16.	1.	3.75	3.	7.50	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.22	4.	1.	1.	2.	1200.	1.	102.	2.	16.	0.	0.0	3.	8.00	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.22	4.	1.	1.	2.	1700.	1.	102.	2.	16.	0.	0.0	2.	6.00	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.23	4.	1.	1.	2.	1300.	1.	102.	2.	17.	0.	0.0	2.	6.00	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.23	2.	7.	2.	3.	1900.	1.	54.	1.	14.	2.	8.75	1.	2.75	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.23	2.	7.	2.	3.	1900.	1.	56.	2.	7.	1.	3.25	1.	2.75	1.	3.75	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.24	1.	2.	2.	2.	600.	1.	28.	2.	21.	2.	8.00	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.24	1.	2.	2.	2.	600.	1.	28.	2.	21.	2.	8.00	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	7.24
7.24	2.	7.	4.	3.	1900.	1.	54.	1.	15.	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.24	2.	7.	4.	3.	1900.	1.	56.	2.	8.	2.	7.50	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.24	2.	7.	4.	3.	1900.	1.	58.	2.	12.	1.	4.00	0.	0.0	0.	0.0	0.	0.0	1.	-1.00	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.24	2.	7.	4.	3.	1900.	1.	128.	2.	8.	2.	7.75	0.	0.0	1.	4.00	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.26	4.	1.	1.	2.	1000.	1.	102.	2.	20.	1.	4.00	2.	4.25	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.26	5.	4.	2.	2.	1000.	2.	26.	2.	13.	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.26	5.	4.	2.	2.	1000.	2.	129.	1.	8.	1.	3.50	1.	2.75	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.26	5.	4.	2.	3.	1400.	2.	26.	2.	13.	1.	4.75	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.26	5.	4.	2.	3.	1400.	2.	129.	1.	8.	0.	0.0	1.	2.75	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.27	5.	4.	2.	3.	1830.	1.	26.	2.	14.	1.	4.75	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.27	5.	4.	2.	3.	1830.	1.	129.	1.	9.	0.	0.0	1.	2.50	0.	0.0	0.	0.0	1.	3.75	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.28	1.	1.	1.	1.	800.	3.	28.	2.	25.	1.	4.00	0.	0.0	1.	4.25	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.28	4.	1.	1.	1.	800.	1.	102.	2.	22.	1.	4.00	1.	2.25	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.28	2.	7.	4.	2.	1200.	1.	54.	1.	19.	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.28	2.	7.	4.	2.	1200.	1.	56.	2.	12.	0.	0.0	0.	0.0	0.	0.0	1.	3.25	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.28	2.	7.	4.	2.	1200.	1.	58.	2.	16.	1.	4.00	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.28	2.	7.	4.	2.	1200.	1.	128.	2.	12.	2.	8.50	1.	2.50	1.	4.50	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.29	1.	1.	1.	1.	800.	1.	28.	2.	26.	0.	0.0	0.	0.0	0.	0.0	0.	0.0	1.	-1.00	0.	0.0	0.	0.0	1.	2.75	0.	0.0
7.29	5.	4.	2.	1.	1100.	1.	26.	2.	16.	1.	3.50	1.	3.00	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.29	5.	4.	2.	1.	1100.	1.	129.	1.	11.	1.	3.00	1.	2.25	2.	5.75	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.29	4.	1.	1.	1.	800.	1.	102.	2.	23.	0.	0.0	1.	2.50	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.29	4.	1.	1.	1.	1100.	1.	102.	2.	23.	0.	0.0	1.	2.50	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.29	2.	7.	4.	3.	1500.	1.	54.	1.	20.	0.	0.0	0.	0.0	1.	5.25	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.29	2.	7.	4.	3.	1500.	1.	56.	2.	13.	1.	4.25	0.	0.0	0.	0.0	1.	3.25	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.29	2.	7.	4.	3.	1500.	1.	58.	2.	17.	0.	0.0	0.	0.0	2.	9.00	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.29	2.	7.	4.	3.	1500.	1.	128.	2.	13.	0.	0.0	2.	5.75	1.	5.00	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.30	4.	1.	1.	1.	800.	1.	102.	2.	24.	2.	9.00	2.	4.50	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.30	1.	1.	1.	1.	800.	1.	28.	2.	27.	2.	8.00	1.	2.50	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.30	1.	1.	1.	1.	1200.	1.	28.	2.	27.	1.	4.25	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.30	1.	1.	1.	2.	1400.	1.	28.	2.	27.	1.	4.00	1.	2.75	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.30	5.	4.	2.	1.	1200.	1.	26.	2.	17.	1.	4.00	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.30	5.	4.	2.	1.	1200.	1.	129.	1.	12.	1.	7.00	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.30	5.	4.	2.	2.	1400.	1.	26.	2.	17.	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.30	5.	4.	2.	2.	1400.	1.	129.	1.	12.	1.	4.00	1.	2.50	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0
7.31	4.	1.	1.	1.	900.	1.	102.	2.	25.	1.	3.75	4.	10.00	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0	0.	0.0

(16)	Total Length of Fish Delivered--Species Type 3	63-67
(17)	Number of Fish Delivered--Species Type 4	68-70
(18)	Total Length of Fish Delivered--Species Type 4	72-76
(19)	Number of Fish Delivered--Species Type 5	77-79
(20)	Total Length of Fish Delivered--Species Type 5	81-85
(21)	Number of Fish Delivered--Species Type 6	86-88
(22)	Total Length of Fish Delivered--Species Type 6	90-94
(23)	Number of Fish Delivered--Species Type 7	95-97
(24)	Total Length of Fish Delivered--Species Type 7	99-103
(25)	Number of Fish Delivered--Species Type 8	104-106
(26)	Total Length of Fish Delivered--Species Type 8	108-112
(27)	Number of Fish Delivered--Species Type 9	113-115
(28)	Total Length of Fish Delivered--Species Type 9	117-121

The column ranges were not provided but were obtained by using the MTS file editor to add some column counting numbers to the end of the files. Once the column ranges were found, these numbers were removed.

Each of the variables selected for the data analysis is described below in more detail. These variable names will be

capitalized hereafter.

A) Response Variable, FEEDRATE

FEEDRATE itself does not appear in the data files, but was computed as a function of number of fish delivered for species types 1 through 9 (variables (11),(13),(15),(17),(19), (21),(23),(25),(27)) as follows:

$$\text{FEEDRATE} = \frac{\text{TOTFSH}}{(2.0 \text{ hours})}$$

where

$$\text{TOTFSH} = \sum_{i=1}^{9} \begin{pmatrix} \text{number of fish delivered} \\ \text{from species type } i \end{pmatrix}$$

Feedrate is thus expressed in units of total number of fish delivered per hour.

It should be noted that -1. is used as a missing value code in the data files. If such a code were encountered for the number of fish for any of the 9 species types, FEEDRATE would be not calculated, since it would be erroneous to include -1. in the TOTFSH sum. Furthermore that record was excluded from further analysis. This record selection was accomplished by a FORTRAN pre-processor program, to be discussed later.

Of course, FEEDRATE could have been defined in other ways. For example any measure of feeding rate should probably account for differing lengths of fish. This is important in considering the question of, say, if one chick receives 2 fish, each of which was 5.0cm in length, and

another chick received 1 fish which was 10.0cm in length, did the 2 chicks receive the same amount of food? Using only the number of fish in a FEEDRATE definition may not tell the whole story. Perhaps one should consider fish-mass per hour, where fish-mass could be measured by proxy as total fish volume, which in turn could be measured by proxy as the sum of the cubes of the lengths of all fish delivered.

Unfortunately the fish length measurements recorded contain more inaccuracies than do fish counts. First of all, it will be noticed from Tables 1 and 2 that for species where more than one fish was delivered, only one fish length is recorded, and this is the sum of the total lengths of the fish. Lengths of individual fish were not available from the files, but Simon stated that they were available elsewhere. Secondly, fish length was first estimated in units of bill length, that is, how may times larger did the fish appear to be than the parent bird's bill. Fish count, however, is easier to obtain accurately than fish length. Thirdly, not all birds have the same bill length, so this makes length measurement even more unreliable than fish count.

To use length in any definition of FEEDRATE would thus expand that variable's measurement error and variability. It was therefore decided not to use fish length in any definition of FEEDRATE.

Finally there are a number of cases where FEEDRATE is 0.0, that is, no fish were delivered during the observed 2 hour period. Such observations are to be retained for analysis since they contribute information (especially since time of day will be considered as an explanatory variable), but they may cause problems if transformations of the response variable are to be considered.

B) Description of Candidate Explanatory Variables

Each variable to be considered in the analysis is given below along with its position in the original files (Tables 1 and 2) as displayed earlier.

B.1)Date (columns 1-4)

This is expressed as

M.DD
(month) (day)

1984 observations have a starting date of 7.15 (15 July 1984), and those of 1985 started on 7.20 (20 July 1985). The difference in starting dates between years is intentional, since dates of first observation were chosen to be first day of actual feeding. This usually took place 3-4 days after hatching, and eggs certainly cannot be expected to hatch on the same day of each year.

For the purpose of analysis, this variable was recoded as follows:

Recoded Value	1984 Observation Date	1985 Observation Date
(Day) 1	7.15	7.20
(Day) 2	7.16	7.21

and so on. This variable will be referred to hereafter as DATE.

B.2)Colony Size (col. 14-15)

This is simply an integer showing the total number of nests in the colony, and will be referred to hereafter as COLSZE.

B.3)Tidal State (col. 17-18)

The three levels of tide were coded:

(Tide)= 1, if low tide
2, if midtide
3, if high tide

In addition, 2 design variables were created:

TIDEH= 1, if high tide
0, otherwise

TIDEM= 1, if midtide
0, otherwise

according to the recommendations on page 703 of Ref.(11). The reason for their creation is that many of the programs in the BMDP computer software package (whose regression routines will be utilized in the next chapter) do not generate design variables needed for qualitative or factor level type variables, such as tide, and design variables are necessary to replace qualitative variables in regression models, which will be used in the analysis. In general a

qualitative variable posessing k levels will require (k-1) design variables (See Sec. 10.1 of Ref.(11)).

TIDEH and TIDEM will be therefore used through most of the analysis, although some use will also be made of the original variable, to be referred to hereafter as TIDE.

B.4)Time of Day (col.20-24)

This is coded as a 24-hour military style clock time, and gives the starting time of the 2 hour observation period to the nearest half-hour. As there are 48 half-hours in a day, the new variable TIME will be the following recoding:

Original Time	Recoded as TIME
0000	0
0030	1
0100	2
0130	3
....	....
2330	47
2400	48

Although there is duplication in the first and last rows of the above table, this is not a point of concern since all observations were understandably done during daylight (0500-2100H approx.)

B.5)Number of Chicks in Nest (col. 34-36)

This is an integer showing total number of chicks in the nest. As it turns out, this number is either 1 or 2, and will be referred to hereafter as NUMCHK. A missing value code of -1 is also used, however. Any record containing such a code was rejected from further analysis.

B.6)Age of Chick (col. 38-40)

This is the age in days of the 1 or 2 chicks in the nest. Again -1 is used to record a missing value. For the case of 2 chicks in the nest, these chicks were taken to be born on the same day. This value will be referred to hereafter as AGECHK.

The variables of interest having been identified, each of the original data files was then subjected to its own FORTRAN pre-processor program, the purposes of which were to:

(a) calculate TOTFSH (save FEEDRATE for statistical software)

(b) perform re-coding previously indicated for DATE, TIME, and creation of extra TIDE variables: TIDEH and TIDEM

(c) flag records containing missing data codes for any fish counts or for either NUMCHK or AGECHK, and put the first reason encountered for such action into a 'reject message' file (records thus flagged are to be kept out of further analysis)

(d) put all acceptable records (those not in (c) above) into a new file containing variables selected for analysis and a coded tag to identify it

The tag referred to in (d) above has a 4-digit format: the first digit indicates the year of the data (4 for 1984 data file, 5 for 1985 data file) and the remaining 3 digits give

the actual line number for the record from its original data file. Thus a tag of 4097 means that the record is number 97 in the 1984 data file.

In addition the values to total length of fish delivered (summed over all species types) and average length of fish were calculated for possible later use. This use did not materialize.

Table 3 shows a portion of the file EMMSFDRT1, which contained the records taken from the 1984 data file along with their recoded values. The values are given in the order:

Identification Tag
Average Fish Length
TOTFSH
Total Fish Length
DATE
COLSZE
TIDE
TIDEH
TIDEM
TIME
NUMCHK
AGECHK

Table 4 shows a portion of the file EMMSREJECT1 which contained rejection messages for records kept out of EMMSFDRT1. Similarly, Table 5 shows a protion fo the file EMMSFDRT2, which contained the acceptable records from the 1985 data file, and Table 6 shows a portion of the file EMMSREJECT2 which contained the rejection messages for records kept out of EMMSFDRT2. The FORTRAN programs which performed these tasks are shown in the Technical Supplement

Table 3: Portion of EMMSFDRT1

4002.	2.000000	1.	2.	1.	10.	2.	0.	1.	20.	2.	9.
4004.	3.250000	1.	3.	1.	10.	2.	0.	1.	20.	1.	12.
4005.	2.500000	1.	3.	1.	10.	2.	0.	1.	20.	2.	12.
4008.	0.0	0.	0.	1.	10.	2.	0.	1.	20.	1.	6.
4013.	3.000000	1.	3.	1.	10.	1.	0.	0.	26.	2.	9.
4014.	2.000000	1.	2.	1.	10.	1.	0.	0.	26.	1.	12.
4015.	2.500000	1.	3.	1.	10.	1.	0.	0.	26.	2.	12.
4016.	2.000000	1.	2.	1.	10.	1.	0.	0.	26.	1.	6.
4021.	2.583333	3.	8.	1.	10.	3.	1.	0.	37.	2.	12.
4022.	3.750000	1.	4.	1.	10.	3.	1.	0.	37.	1.	12.
4025.	2.000000	1.	2.	1.	10.	3.	1.	0.	37.	2.	9.
4030.	0.0	0.	0.	1.	10.	3.	1.	0.	37.	1.	6.
4031.	2.500000	4.	10.	2.	10.	3.	1.	0.	14.	2.	12.
4032.	2.000000	1.	2.	2.	10.	3.	1.	0.	14.	1.	12.
4035.	2.000000	1.	2.	2.	10.	3.	1.	0.	14.	2.	9.
4037.	0.0	0.	0.	2.	10.	3.	1.	0.	14.	1.	7.
4044.	0.0	0.	0.	2.	10.	2.	0.	1.	21.	1.	7.
4045.	0.0	0.	0.	2.	10.	2.	0.	1.	21.	1.	13.
4048.	0.0	0.	0.	2.	10.	2.	0.	1.	21.	2.	13.
4049.	2.583333	3.	8.	2.	10.	2.	0.	1.	21.	2.	10.
4052.	3.250000	2.	7.	2.	10.	3.	1.	0.	37.	2.	13.
4053.	2.500000	1.	3.	2.	10.	3.	1.	0.	37.	2.	10.
4054.	2.500000	1.	3.	2.	10.	3.	1.	0.	37.	1.	7.
4056.	0.0	0.	0.	2.	10.	3.	1.	0.	37.	1.	13.
4060.	2.500000	1.	3.	3.	9.	2.	0.	1.	22.	1.	14.
4062.	3.000000	1.	3.	3.	9.	2.	0.	1.	22.	1.	8.
4064.	3.000000	1.	3.	3.	9.	2.	0.	1.	22.	2.	14.
4066.	0.0	0.	0.	3.	9.	2.	0.	1.	22.	2.	11.
4069.	3.000000	2.	6.	3.	9.	3.	1.	0.	38.	1.	14.
4070.	2.750000	1.	3.	3.	9.	3.	1.	0.	38.	1.	11.
4071.	3.375000	2.	7.	3.	9.	3.	1.	0.	38.	2.	14.
4072.	3.000000	1.	3.	3.	9.	3.	1.	0.	38.	1.	8.
4075.	3.000000	1.	3.	5.	12.	2.	0.	1.	28.	1.	24.
4076.	2.500000	1.	3.	5.	12.	2.	0.	1.	28.	1.	9.
4077.	2.500000	1.	3.	5.	12.	2.	0.	1.	28.	1.	19.
4078.	3.000000	1.	3.	5.	12.	2.	0.	1.	28.	2.	15.
4079.	2.750000	1.	3.	5.	12.	2.	0.	1.	28.	1.	14.
4080.	0.0	0.	0.	5.	12.	2.	0.	1.	28.	2.	21.
4081.	0.0	0.	0.	5.	12.	2.	0.	1.	28.	1.	8.
4082.	0.0	0.	0.	5.	12.	2.	0.	1.	28.	1.	9.
4083.	3.000000	1.	3.	5.	12.	2.	0.	1.	32.	1.	19.
4084.	4.000000	1.	4.	5.	12.	2.	0.	1.	32.	2.	21.
4085.	3.000000	1.	3.	5.	12.	2.	0.	1.	32.	1.	8.
4086.	3.500000	1.	4.	5.	12.	2.	0.	1.	32.	1.	9.
4087.	3.750000	3.	11.	5.	12.	2.	0.	1.	32.	2.	15.
4088.	3.500000	1.	4.	5.	12.	2.	0.	1.	32.	1.	24.
4089.	0.0	0.	0.	5.	12.	2.	0.	1.	32.	1.	14.
4090.	2.500000	1.	3.	5.	12.	2.	0.	1.	32.	1.	9.
4091.	3.000000	2.	6.	7.	5.	2.	0.	1.	16.	2.	17.
4092.	2.500000	1.	3.	7.	5.	2.	0.	1.	16.	1.	18.
4093.	3.000000	3.	9.	7.	5.	2.	0.	1.	16.	2.	19.
4097.	3.125000	2.	6.	8.	10.	2.	0.	1.	11.	2.	7.
4098.	2.000000	2.	4.	8.	10.	2.	0.	1.	11.	1.	19.
4099.	2.750000	2.	6.	8.	10.	2.	0.	1.	11.	1.	16.
4102.	3.375000	2.	7.	8.	10.	2.	0.	1.	11.	2.	19.
4107.	2.166666	3.	7.	8.	10.	2.	0.	1.	20.	1.	19.
4112.	0.0	0.	0.	8.	10.	2.	0.	1.	20.	1.	16
4113.	0.0	0.	0.	8.	10.	2.	0.	1.	20.	2.	19
4114.	0.0	0.	0.	8.	10.	2.	0.	1.	20.	2.	7
4115.	3.049999	5.	15.	9.	12.	2.	0.	1.	12.	1.	23.
4116.	3.750000	2.	8.	9.	12.	2.	0.	1.	12.	1.	13.
4117.	3.000000	1.	3.	9.	12.	2.	0.	1.	12.	1.	12.
4118.	2.875000	2.	6.	9.	12.	2.	0.	1.	12.	1.	18.
4119.	3.750000	1.	4.	9.	12.	2.	0.	1.	12.	1.	28.
4120.	3.250000	4.	13.	9	12.	2.	0.	1.	12.	2.	25.
4121.	2.750000	1.	3.	9.	12.	2.	0.	1.	12.	1.	13.
4122.	3.000000	1.	3.	9.	12.	2.	0.	1.	12.	2.	19.
4124.	3.000000	1.	3.	9.	12.	2.	0.	1.	22.	1.	23.
4125.	3.000000	1.	3.	9.	12.	2.	0.	1.	22.	1.	18.
4126.	2.500000	1.	3.	9.	12.	2.	0.	1.	22.	1.	12.
4127.	3.000000	1.	3.	9.	12.	2.	0.	1.	22.	1.	28.

Table 4: Portion of EMMSREJECT1

```
RECORD     1. REJECTED FOR MISSING VALUE CODE
              -1. FOR CHK

RECORD     3. REJECTED FOR MISSING VALUE CODE
              -1. FOR CHK

RECORD     6. REJECTED FOR MISSING VALUE CODE
              -1. FOR CHK

RECORD     7. REJECTED FOR MISSING VALUE CODE
              -1. FOR CHK

RECORD     9. REJECTED FOR MISSING VALUE CODE
              -1. FOR CHK

RECORD    10. REJECTED FOR MISSING VALUE CODE
              -1. FOR CHK

RECORD    11. REJECTED FOR MISSING VALUE CODE
              -1. FOR CHK

RECORD    12. REJECTED FOR MISSING VALUE CODE
              -1. FOR CHK

RECORD    17. REJECTED FOR MISSING VALUE CODE
              -1. FOR CHK

RECORD    18. REJECTED FOR MISSING VALUE CODE
              -1. FOR CHK

RECORD    19. REJECTED FOR MISSING VALUE CODE
              -1. FOR CHK

RECORD    20. REJECTED FOR MISSING VALUE CODE
              -1. FOR CHK

RECORD    23. REJECTED FOR MISSING VALUE CODE
              -1. FOR CHK

RECORD    24. REJECTED FOR MISSING VALUE CODE
              -1. FOR CHK

RECORD    26. REJECTED FOR MISSING VALUE CODE
              -1. FOR CHK

RECORD    27. REJECTED FOR MISSING VALUE CODE
              -1. FOR CHK

RECORD    28. REJECTED FOR MISSING VALUE CODE
              -1. FOR CHK

RECORD    29. REJECTED FOR MISSING VALUE CODE
              -1. FOR CHK

RECORD    33. REJECTED FOR MISSING VALUE CODE
              -1. FOR CHK

RECORD    34. REJECTED FOR MISSING VALUE CODE
              -1. FOR CHK

RECORD    36. REJECTED FOR MISSING VALUE CODE
              -1. FOR CHK

RECORD    38. REJECTED FOR MISSING VALUE CODE
              -1. FOR CHK

RECORD    39. REJECTED FOR MISSING VALUE CODE
              -1. FOR CHK

RECORD    40. REJECTED FOR MISSING VALUE CODE
              -1. FOR CHK
```

Table 5: Portion of EMMSFDRT2

5001.	3.875000	2.	8.	1.	2.	1.	0.	0.	26.	2.	17.
5003.	3.750000	1.	4.	1.	2.	1.	0.	0.	37.	2.	17.
5005.	2.750000	2.	6.	2.	2.	3.	1.	0.	12.	2.	18.
5007.	2.500000	1.	3.	2.	1.	2.	0.	1.	20.	2.	15.
5008.	3.000000	3.	9.	2.	2.	3.	1.	0.	36.	2.	18.
5010.	2.812500	4.	11.	3.	1.	3.	1.	0.	17.	2.	16.
5011.	2.666666	3.	8.	3.	1.	2.	0.	1.	24.	2.	16.
5012.	3.000000	2.	6.	3.	1.	2.	0.	1.	34.	2.	16.
5013.	3.000000	2.	6.	4.	1.	2.	0.	1.	26.	2.	17.
5014.	3.833333	3.	12.	4.	2.	3.	1.	0.	38.	1.	14.
5015.	3.250000	3.	10.	4.	2.	3.	1.	0.	38.	2.	7.
5016.	4.000000	2.	8.	5.	2.	2.	0.	1.	12.	2.	21.
5017.	7.620000	2.	15.	5.	2.	2.	0.	1.	12.	2.	21.
5018.	0.0	0.	0.	5.	4.	3.	1.	0.	38.	1.	15.
5019.	3.750000	2.	8.	5.	4.	3.	1.	0.	38.	2.	8.
5020.	-1.000000	2.	-1.	5.	4.	3.	1.	0.	38.	2.	12.
5021.	3.916666	3.	12.	5.	4.	3.	1.	0.	38.	2.	8.
5022.	2.750000	3.	8.	7.	1.	2.	0.	1.	20.	2.	20.
5023.	0.0	0.	0.	7.	2.	2.	0.	1.	20.	2.	13.
5024.	3.125000	2.	6.	7.	2.	2.	0.	1.	20.	1.	8.
5025.	4.750000	1.	5.	7.	2.	3.	1.	0.	28.	2.	13.
5026.	2.750000	1.	3.	7.	2.	3.	1.	0.	28.	1.	8.
5027.	4.750000	1.	5.	8.	2.	3.	1.	0.	37.	2.	14.
5028.	3.125000	2.	6.	8.	2.	3.	1.	0.	37.	1.	9.
5029.	4.125000	2.	8.	9.	1.	1.	0.	0.	16.	2.	25.
5030.	3.125000	2.	6.	9.	1.	1.	0.	0.	16.	2.	22.
5031.	0.0	0.	0.	9.	4.	2.	0.	1.	24.	1.	19.
5032.	3.250000	1.	3.	9.	4.	2.	0.	1.	24.	2.	12.
5033.	4.000000	1.	4.	9.	4.	2.	0.	1.	24.	2.	16.
5034.	3.875000	4.	16.	9.	4.	2.	0.	1.	24.	2.	12.
5035.	-1.000000	2.	-1.	10.	1.	1.	0.	0.	16.	2.	26.
5036.	3.250000	2.	7.	10.	2.	1.	0.	0.	22.	2.	16.
5037.	2.750000	4.	11.	10.	2.	1.	0.	0.	22.	1.	11.
5038.	2.500000	1.	3.	10.	1.	1.	0.	0.	16.	2.	23.
5039.	2.500000	1.	3.	10.	1.	1.	0.	0.	22.	2.	23.
5040.	5.250000	1.	5.	10.	4.	3.	1.	0.	30.	1.	20.
5041.	3.750000	2.	8.	10.	4.	3.	1.	0.	30.	2.	13.
5042.	4.500000	2.	9.	10.	4.	3.	1.	0.	30.	2.	17.
5043.	3.583333	3.	11.	10.	4.	3.	1.	0.	30.	2.	13.
5044.	3.375000	4.	14.	11.	1.	1.	0.	0.	16.	2.	24.
5045.	3.500000	3.	11.	11.	1.	1.	0.	0.	16.	2.	27.
5046.	4.250000	1.	4.	11.	1.	1.	0.	0.	24.	2.	27.
5047.	3.375000	2.	7.	11.	1.	2.	0.	1.	28.	2.	27.
5048.	4.000000	1.	4.	11.	2.	1.	0.	0.	24.	2.	17.
5049.	7.000000	1.	7.	11.	2.	1.	0.	0.	24.	1.	12.
5050.	0.0	0.	0.	11.	2.	2.	0.	1.	28.	2.	17.
5051.	3.250000	2.	7.	11.	2.	2.	0.	1.	28.	1.	12.
5052.	2.750000	5.	14.	12.	1.	1.	0.	0.	18.	2.	25
5053.	4.500000	2.	9.	12.	1.	1.	0.	0.	18.	2.	28.
5054.	3.250000	1.	3.	12.	4.	1.	0.	0.	24.	1.	22.
5055.	3.500000	1.	4.	12.	4.	1.	0.	0.	24.	2.	15.
5056.	5.000000	2.	10.	12.	4.	1.	0.	0.	24.	2.	19.
5057.	3.357142	7.	24.	12.	4.	1.	0.	0.	24.	2.	15.
5058.	3.500000	1.	4.	12.	4.	2.	0.	1.	28.	1	22.
5059.	3.250000	1.	3.	12.	4.	2.	0.	1.	28.	2.	15.
5060.	3.583333	3.	11.	12.	4.	2.	0.	1.	28.	2.	19.
5061.	3.875000	2.	8.	12.	4.	2.	0.	1.	28.	2.	15.
5062.	2.500000	1.	3.	13.	1.	1.	0.	0.	21.	2.	23
5063.	2.375000	4.	10.	13.	1.	1.	0.	0.	27.	2.	26
5064.	4.375000	2.	9.	13.	4.	1.	0.	0.	21.	1.	23.
5065.	4.000000	1.	4.	13.	4.	1.	0.	0.	21.	2.	16.
5066.	0.0	0.	0.	13.	4.	1.	0.	0.	21.	2.	20
5067.	3.562500	4.	14.	13.	4.	1.	0.	0.	21.	2.	16.
5068.	2.750000	1.	3.	13.	2.	1.	0.	0.	27.	2.	19.
5069.	4.500000	1.	5.	13.	2.	1.	0.	0.	27.	1.	14.
5070.	3.666666	3.	11.	13.	1.	3.	1.	0.	34.	2.	29.
5071.	2.250000	1.	2.	14.	1.	3.	1.	0.	14.	2.	27.
5072.	2.583333	3.	8.	14.	1.	1.	0.	0.	22.	2.	27.
5073.	2.750000	2.	6.	14.	1.	1.	0.	0.	26.	2.	27.
5074.	3.062500	4.	12.	14.	1.	3.	1.	0.	34.	2.	27.
5075.	4.250000	2.	9.	14.	4.	3.	1.	0.	14.	1.	24.

Table 6: Portion of EMMSREJECT2

```
RECORD     2  REJECTED FOR MISSING VALUE CODE
                -1. FOR CHK

RECORD     4  REJECTED FOR MISSING VALUE CODE
                -1. FOR CHK

RECORD     6  REJECTED FOR MISSING VALUE CODE
                -1. FOR CHK

RECORD     9  REJECTED FOR MISSING VALUE CODE
                -1. FOR CHK

RECORD    87  REJECTED FOR MISSING VALUE CODE
                -1. FOR CHK

RECORD    89  REJECTED FOR MISSING VALUE CODE
                -1. FOR CHK

RECORD    94  REJECTED FOR MISSING VALUE CODE
                -1. FOR CHK

RECORD    98  REJECTED FOR MISSING VALUE CODE
                -1. FOR CHK

RECORD   100  REJECTED FOR MISSING VALUE CODE
                -1. FOR CHK

RECORD   102  REJECTED FOR MISSING VALUE CODE
                -1. FOR CHK

RECORD   115  REJECTED FOR MISSING VALUE CODE
                -1. FOR CHK

RECORD   119  REJECTED FOR MISSING VALUE CODE
                -1. FOR CHK

RECORD   121  REJECTED FOR MISSING VALUE CODE
                -1. FOR CHK

RECORD   123  REJECTED FOR MISSING VALUE CODE
                -1. FOR CHK

RECORD   126  REJECTED FOR MISSING VALUE CODE
                -1. FOR CHK

RECORD   128  REJECTED FOR MISSING VALUE CODE
                -1. FOR CHK

RECORD   131  REJECTED FOR MISSING VALUE CODE
                -1. FOR CHK

RECORD   133  REJECTED FOR MISSING VALUE CODE
                -1. FOR CHK

RECORD   138  REJECTED FOR MISSING VALUE CODE
                -1. FOR CHK

RECORD   140. REJECTED FOR MISSING VALUE CODE
                -1. FOR CHK

RECORD   144. REJECTED FOR MISSING VALUE CODE
                -1. FOR CHK

RECORD   145. REJECTED FOR MISSING VALUE CODE
                -1. FOR CHK

RECORD   150. REJECTED FOR MISSING VALUE CODE
                -1. FOR CHK

RECORD   152. REJECTED FOR MISSING VALUE CODE
                -1. FOR CHK
```

for this chapter (found in Chapter 7).

Finally, using the MTS file editor, these two files were put together as one file, known as EMMSFDRT. With all the necessary data editing performed the data (as in the EMMSFDRT file) was now ready for analysis.

CHAPTER 3

FIRST ANALYSIS--FIND BEST MAIN EFFECTS MODEL

As a first attempt to explain the different values of Y=FEEDRATE, a multiple linear regression model is first attempted:

$$Y=\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p+\epsilon$$

where p is the number of explanatory variables to be included, and thus $p+1$ regression coefficients (including the intercept term β_0) must be estimated. ϵ is a random error component having a normal distribution with mean 0 and variance σ^2. An individual x_j could represent a single explanatory variable in first order (e.g. x_1=TIME), a higher order power of that variable (e.g. x_2=(TIME)^2), a cross-product of 2 or more variables (e.g. x_3=(TIME)(AGECHK)), or a more complex function of 1 or more variables. The qualifier 'linear' means only that the model is linear in the regression coefficients $\beta_0,\beta_1,\ldots,\beta_p$, not necessarily in the explanatory variables themselves.

If an x_j represents a function of a single explanatory variable, it is called a 'main effect' of that function. If an x_j represents a function of 2 or more explanatory variables in such a way that it can be written as the product of 2 or more main effects previously described (e.g. $x_4=x_2x_3$), it is called an 'interaction' between these main effects. Only main effects will be considered in this

section. Interactions are considered in chapter 4 for model fit improvement.

There is a total sample size of n=524 observations left after the data editing described in the previous chapter, so one can use a subscript to identify each observation & its corresponding explanatory variables:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$$

where i ranges from 1 to n, and the ϵ_i will have the additional assumption of no correlation amongst themselves. In general, the β_j represent unknown parameters, so suitable estimates, b_j will be sought to produce a 'fitted' model:

$$\hat{Y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_p x_{ip}$$

which give a fitting error of

$$e_i = Y_i - \hat{Y}_i$$

known as residual i which is defined to be the discrepancy between observation i and the corresponding outcome fitted by the model.

Next a set of x_j must be selected. To aid in this task, the BMDP program P6D (Ref. 4, Section 10.2) was used to produce scatter plots of TOTFSH against various variables and some higher order powers in order to get a visual appraisal of any trends or associations. TOTFSH was selected over FEEDRATE because TOTFSH takes on integer values. These plots are shown in Figures 1.a through 1.p

Figure 1.a: TOTFSH vs. DATE

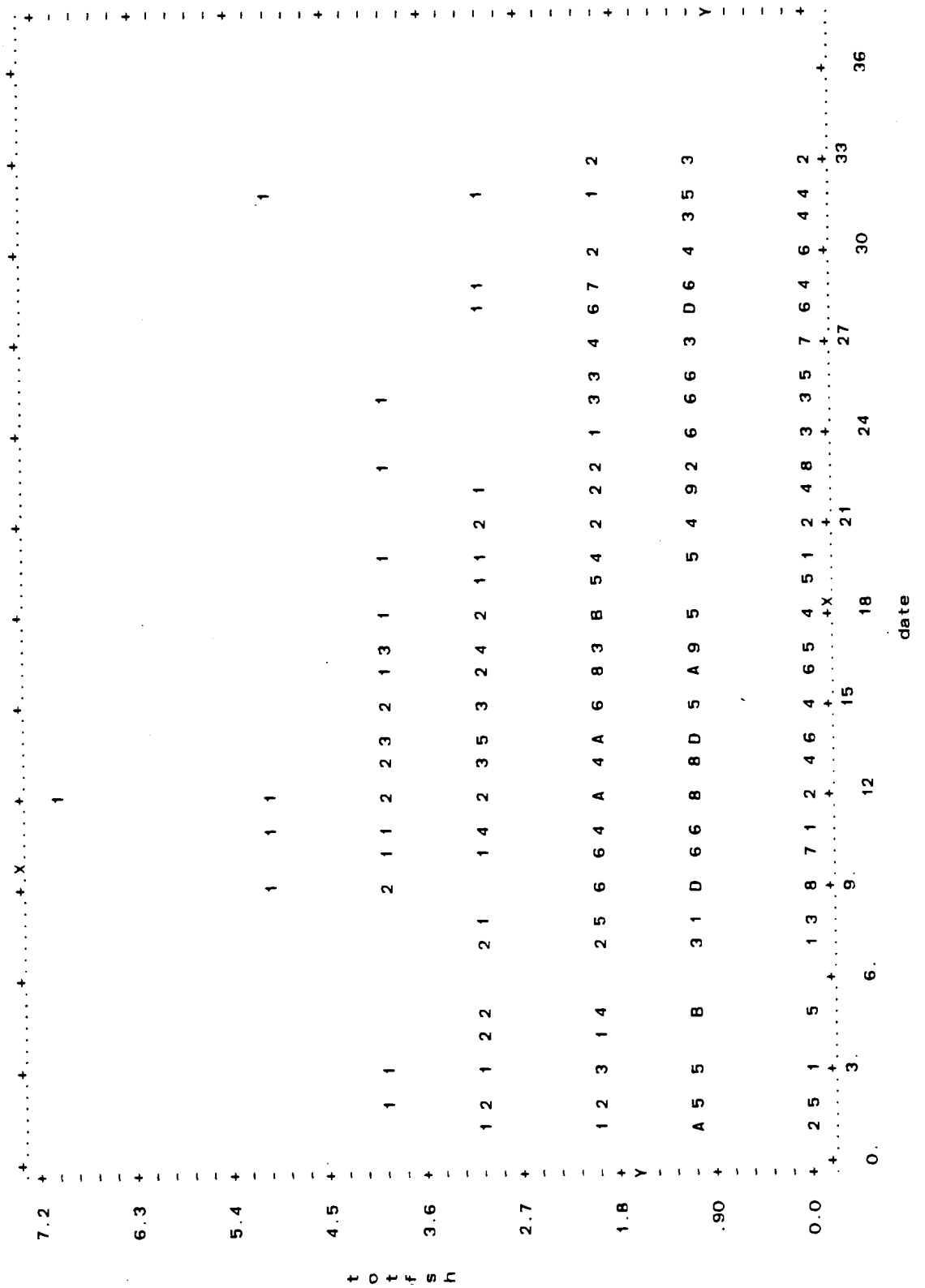


Figure 1.b: TOTFSH vs. COLSZE

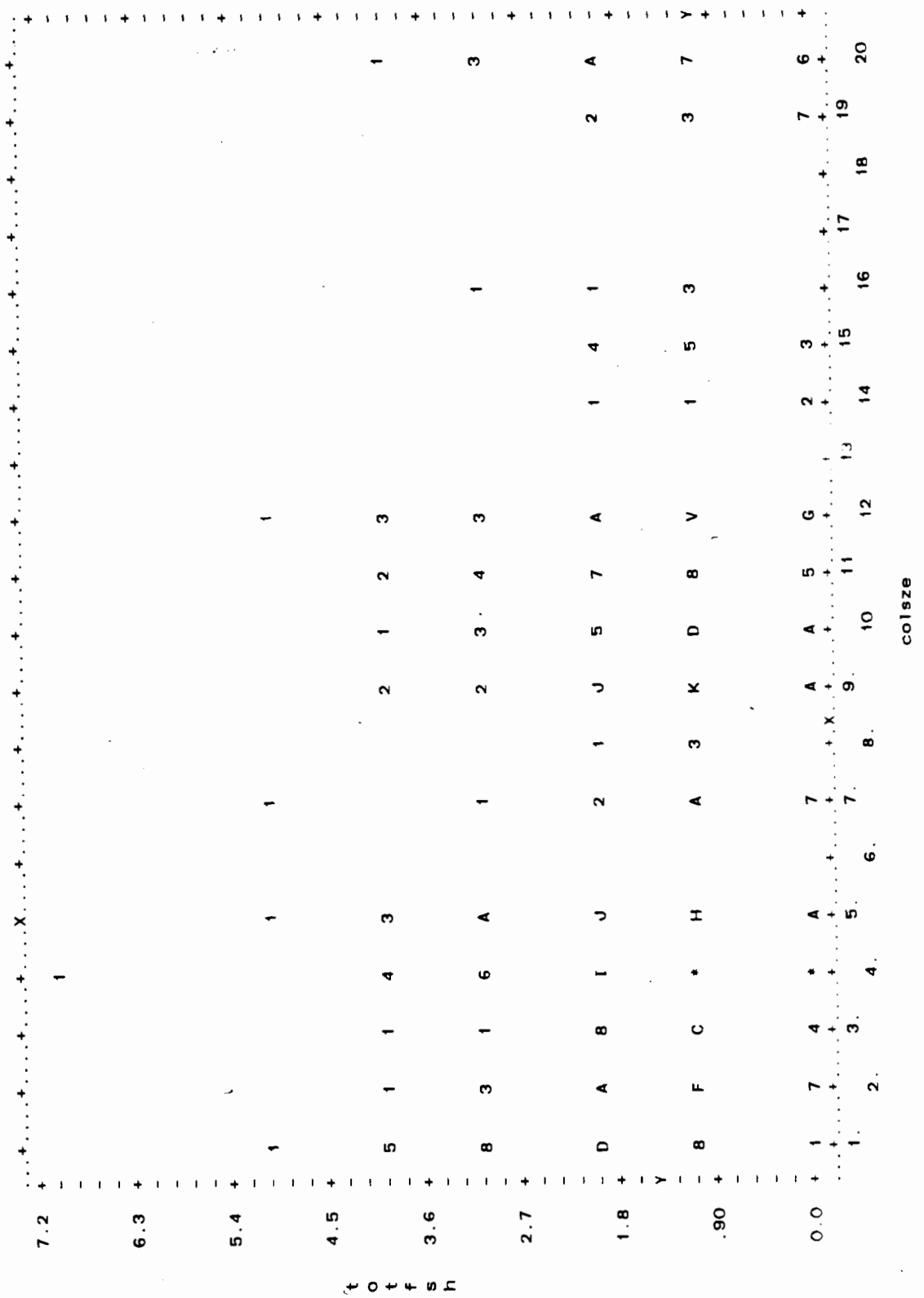


Figure 1.c: TOTFSH vs. TIDE

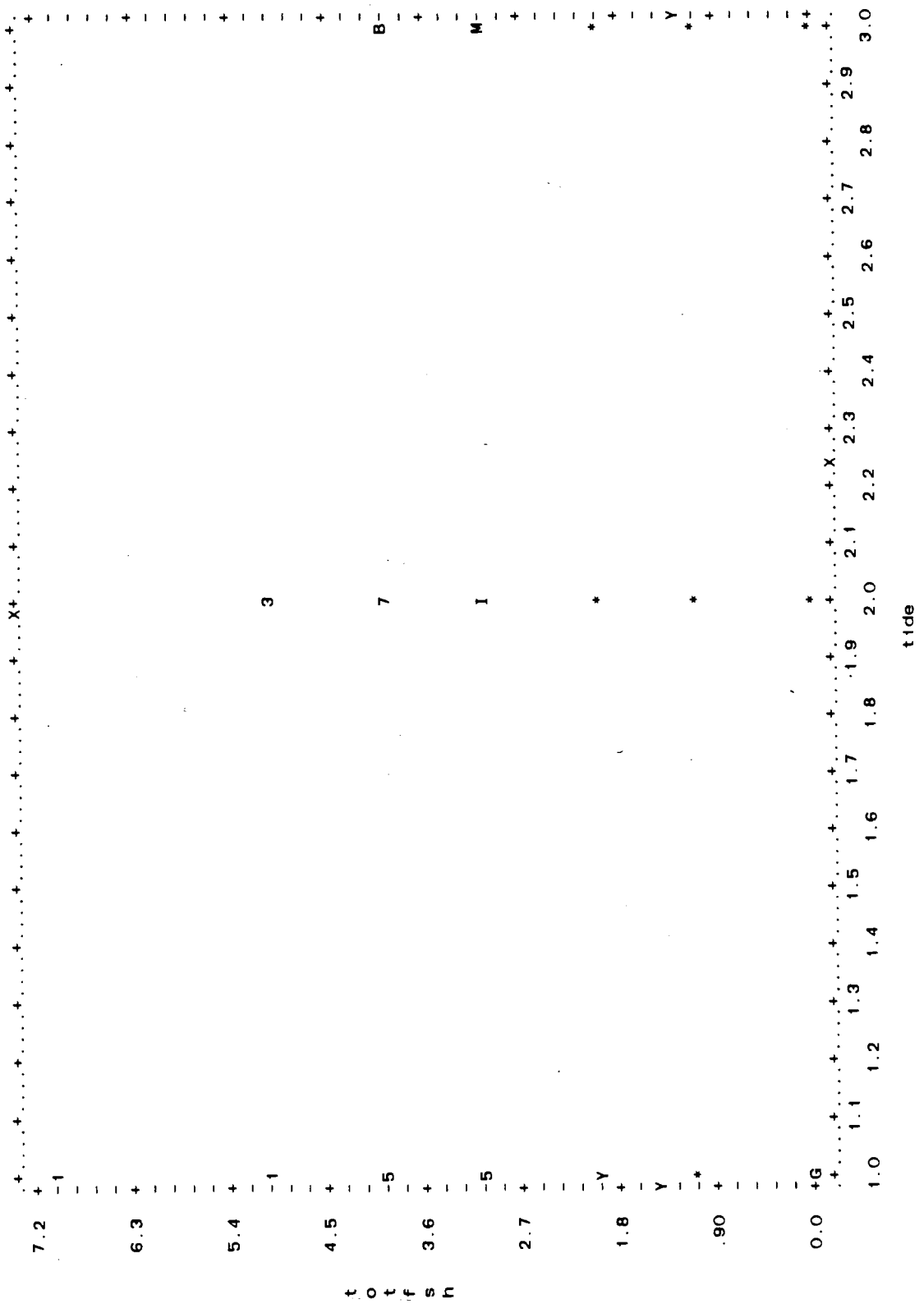


Figure 1.d: TOTFSH vs. TIDEH

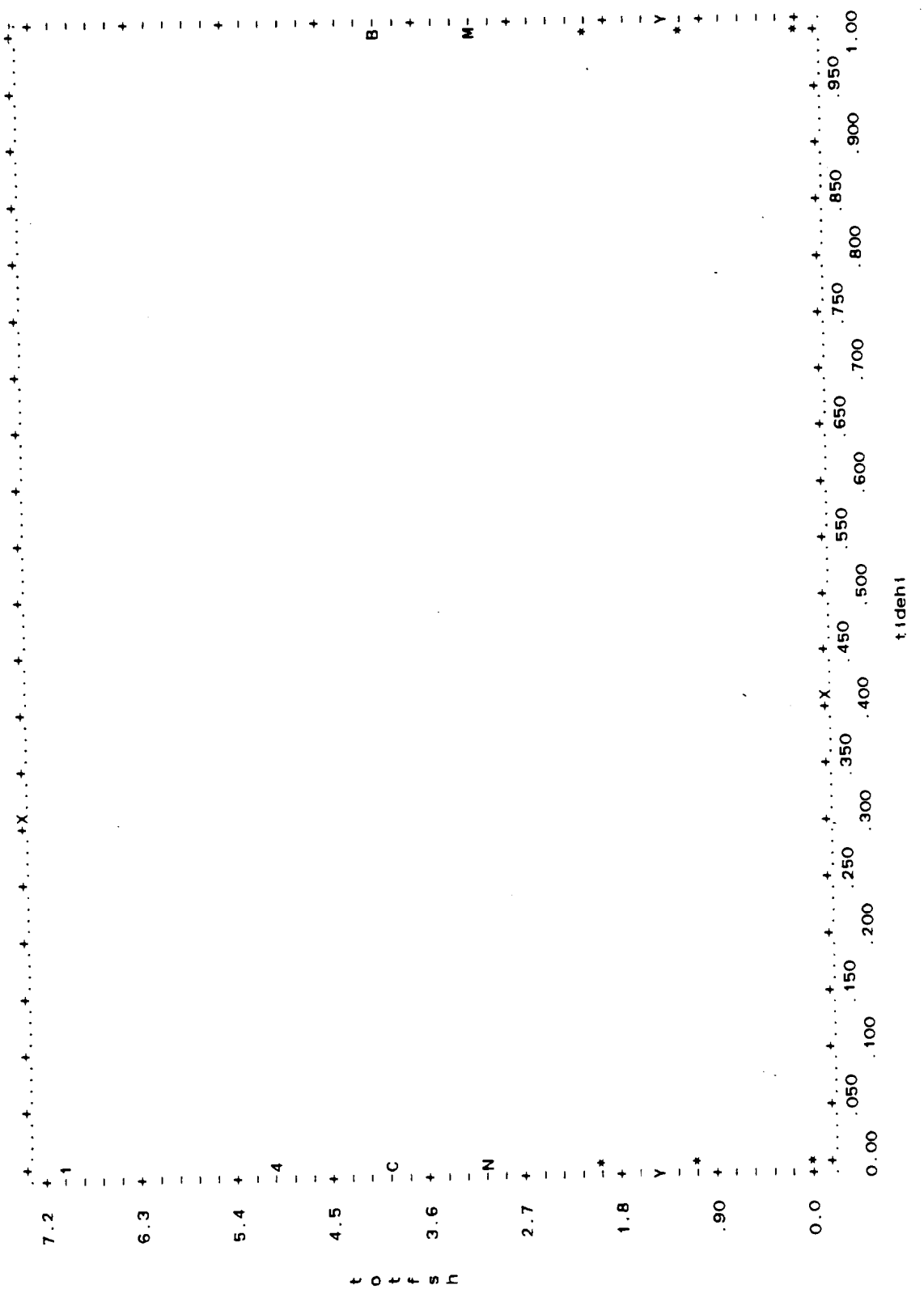


Figure 1.e: TOTFSH vs. TIDEM

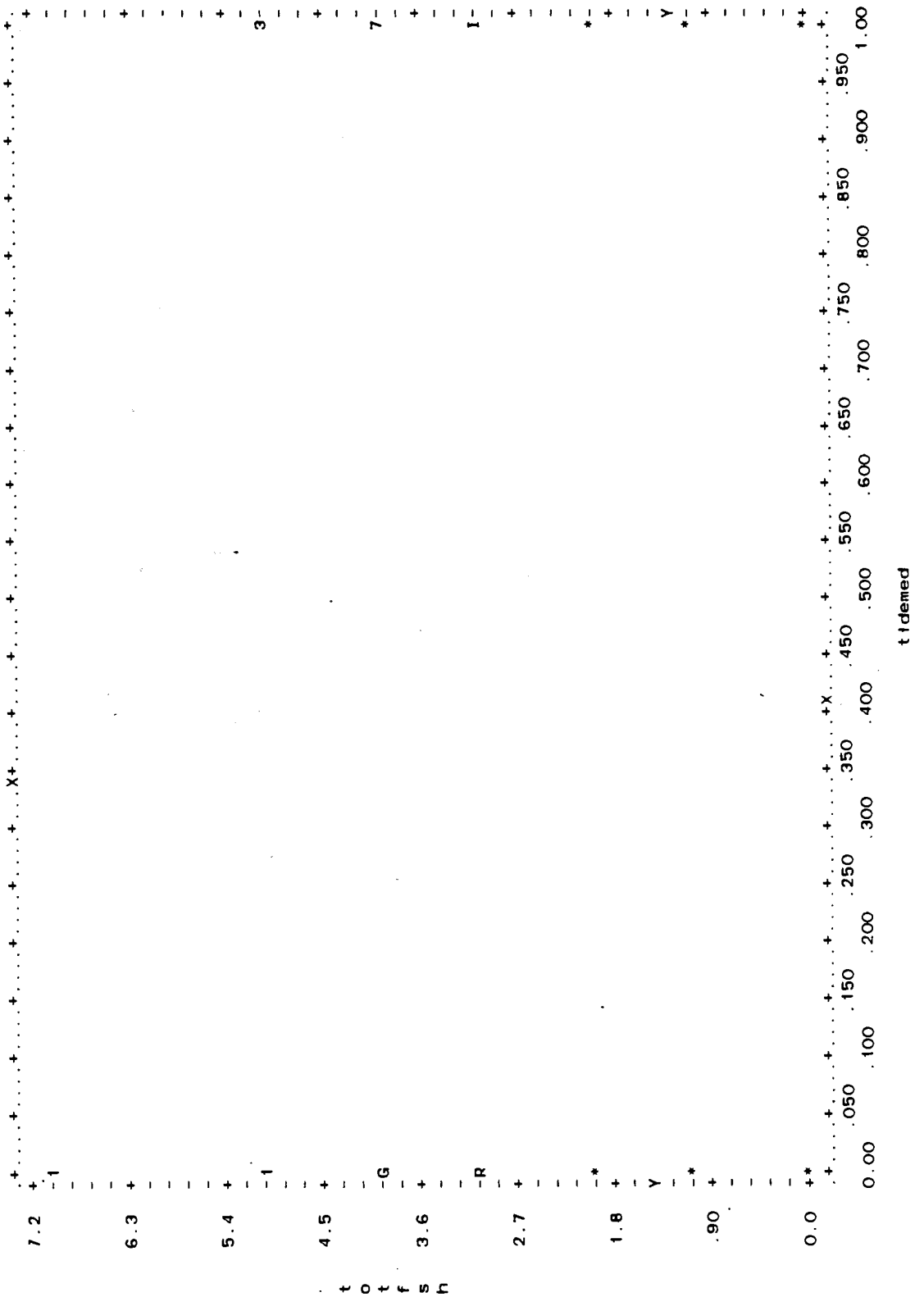


Figure 1.f: TOTFSH vs. NUMCHK

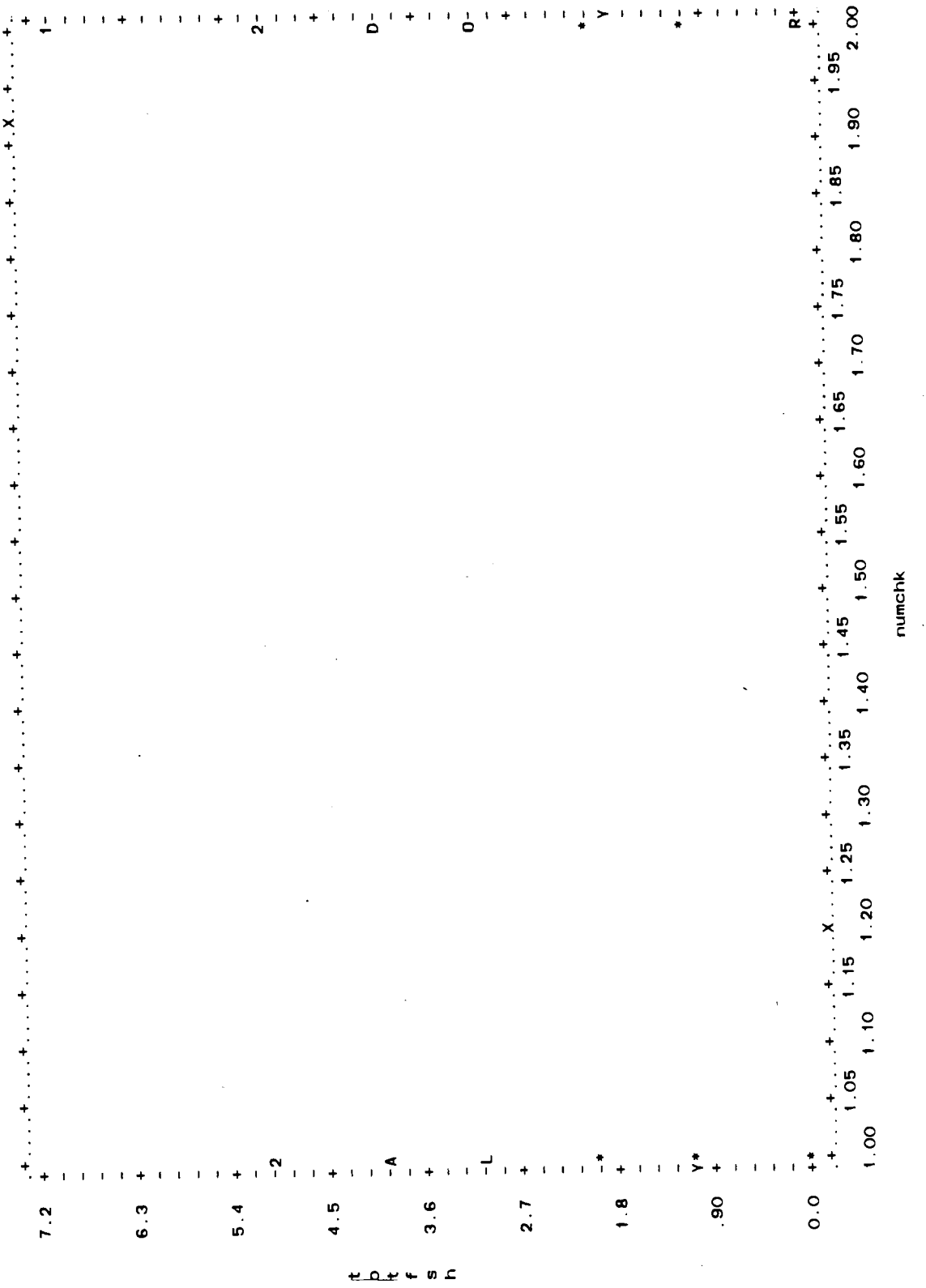


Figure 1.g: TOTFSH vs. AGECHK

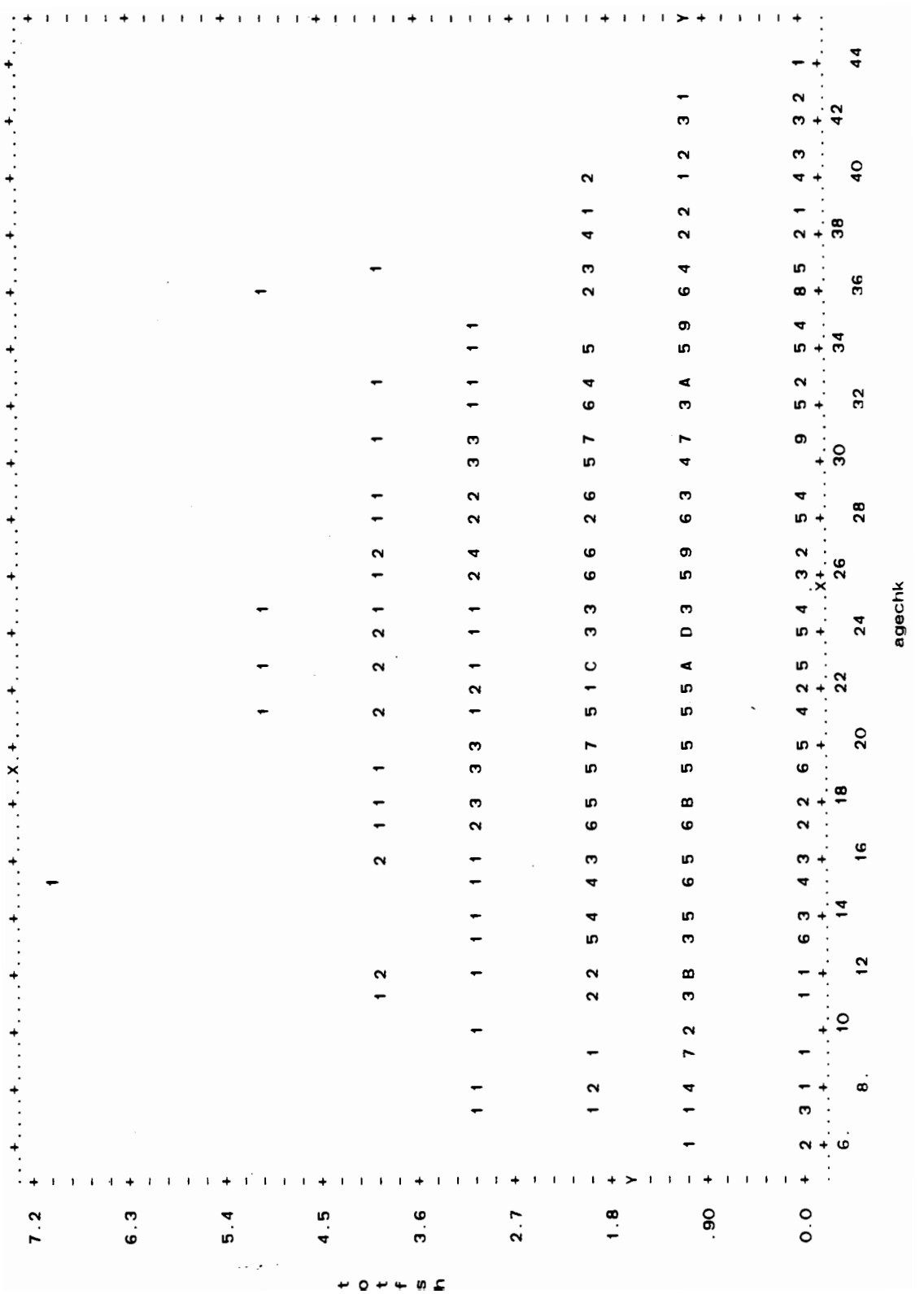


Figure 1.h: TOTFSH vs. TIME

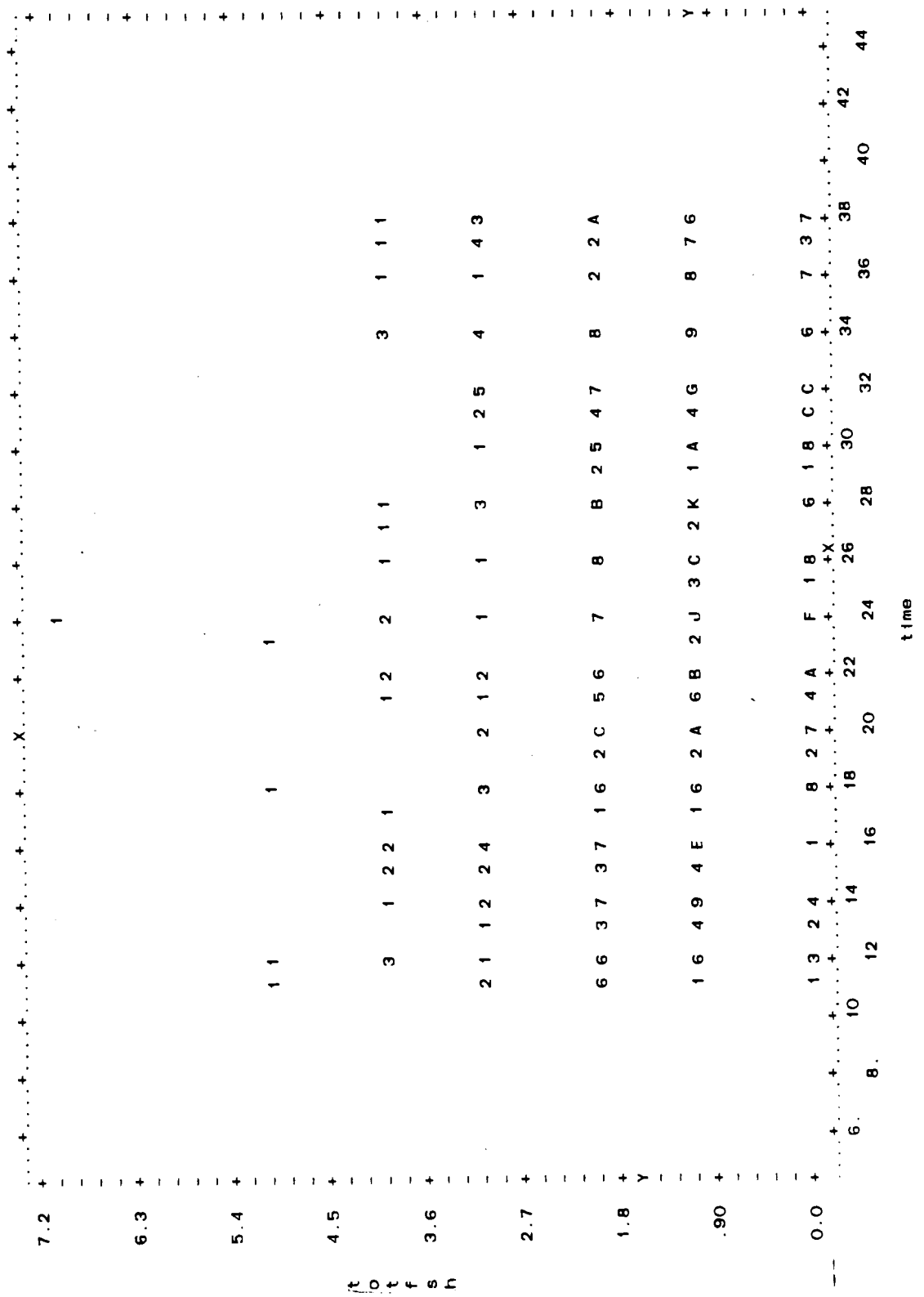


Figure 1.i: TOTFSH vs. SQDATE

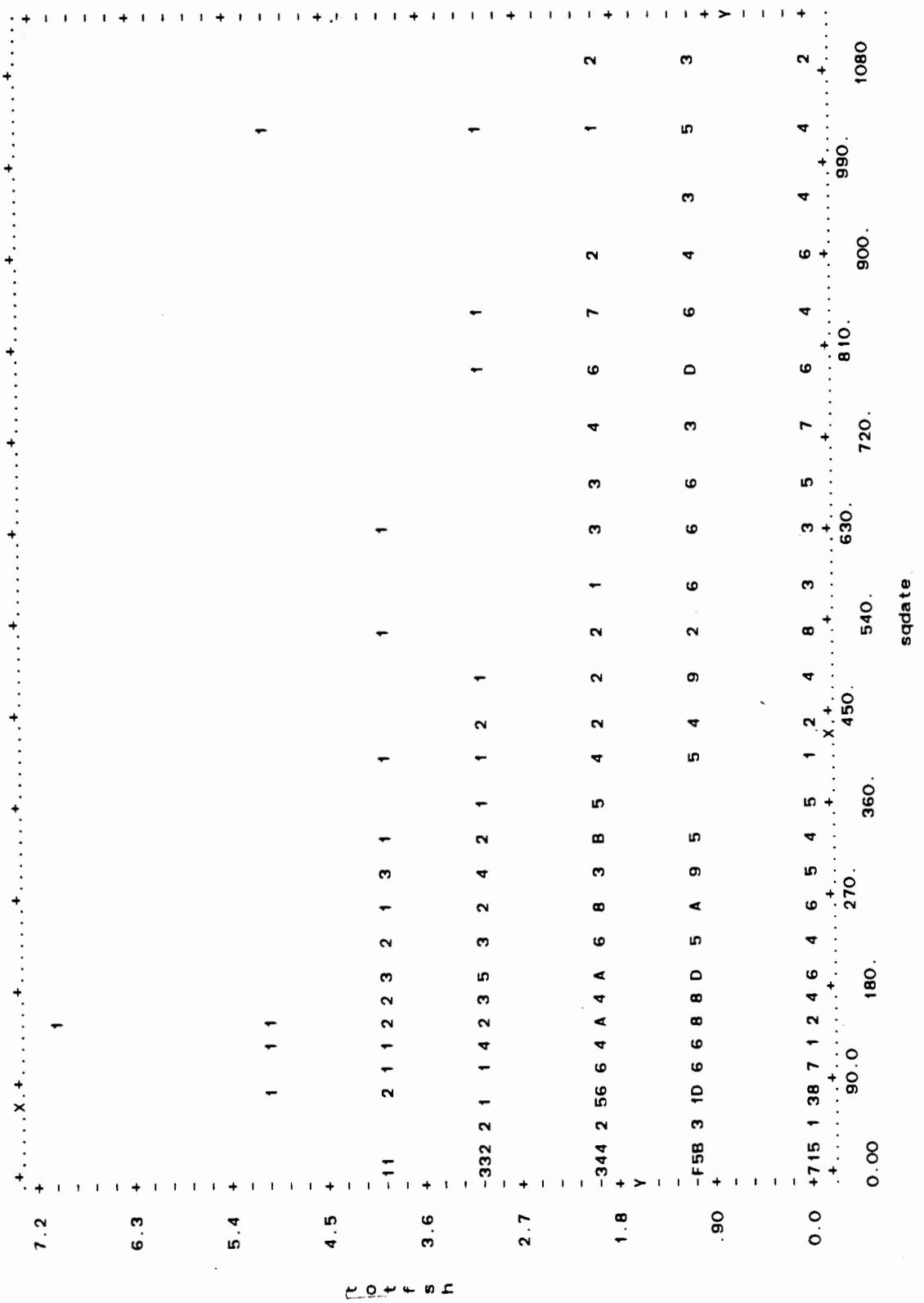


Figure 1.j: TOTFSH vs. SQTIME

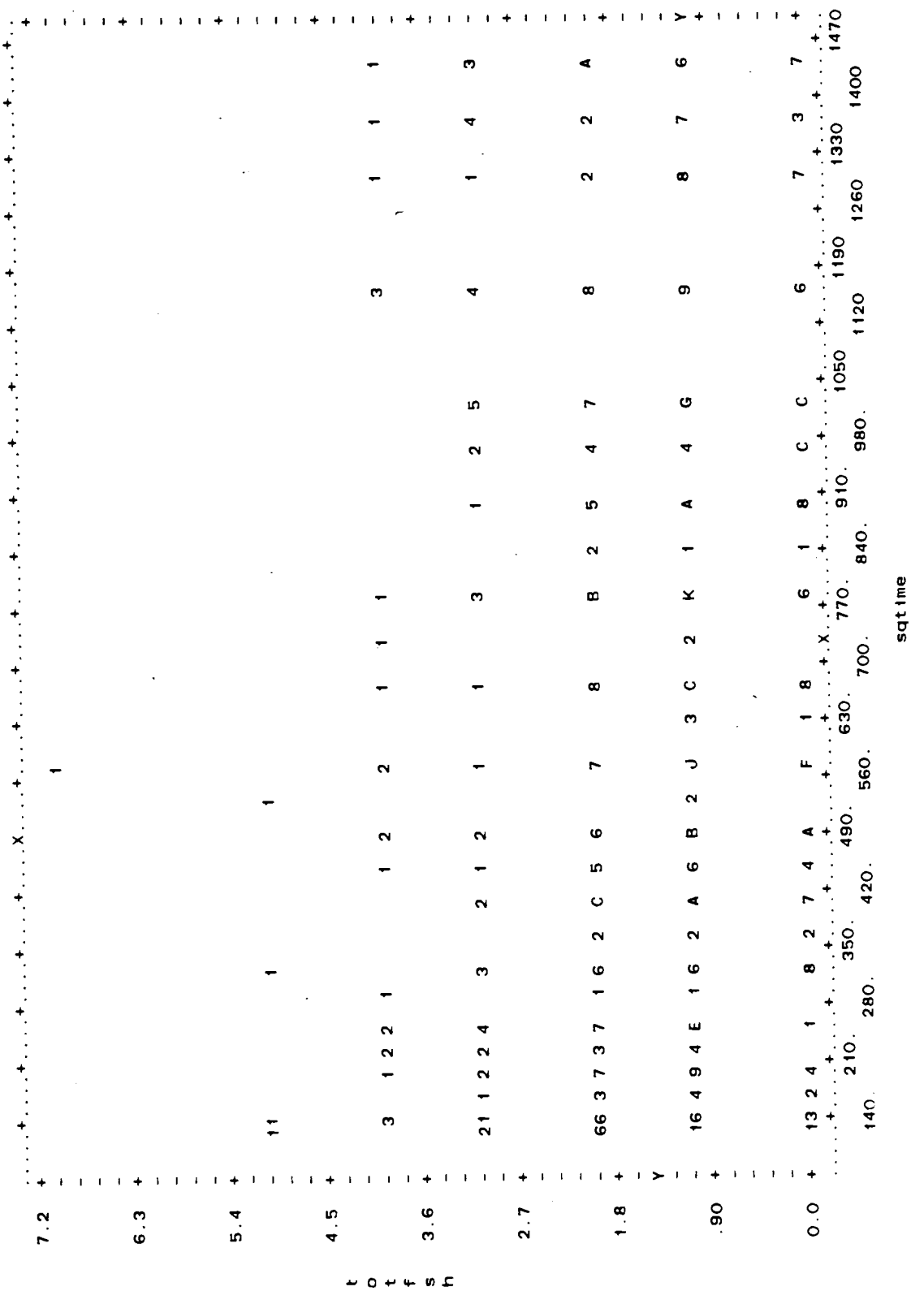


Figure 1.k: TOTFSH vs. SQAGE

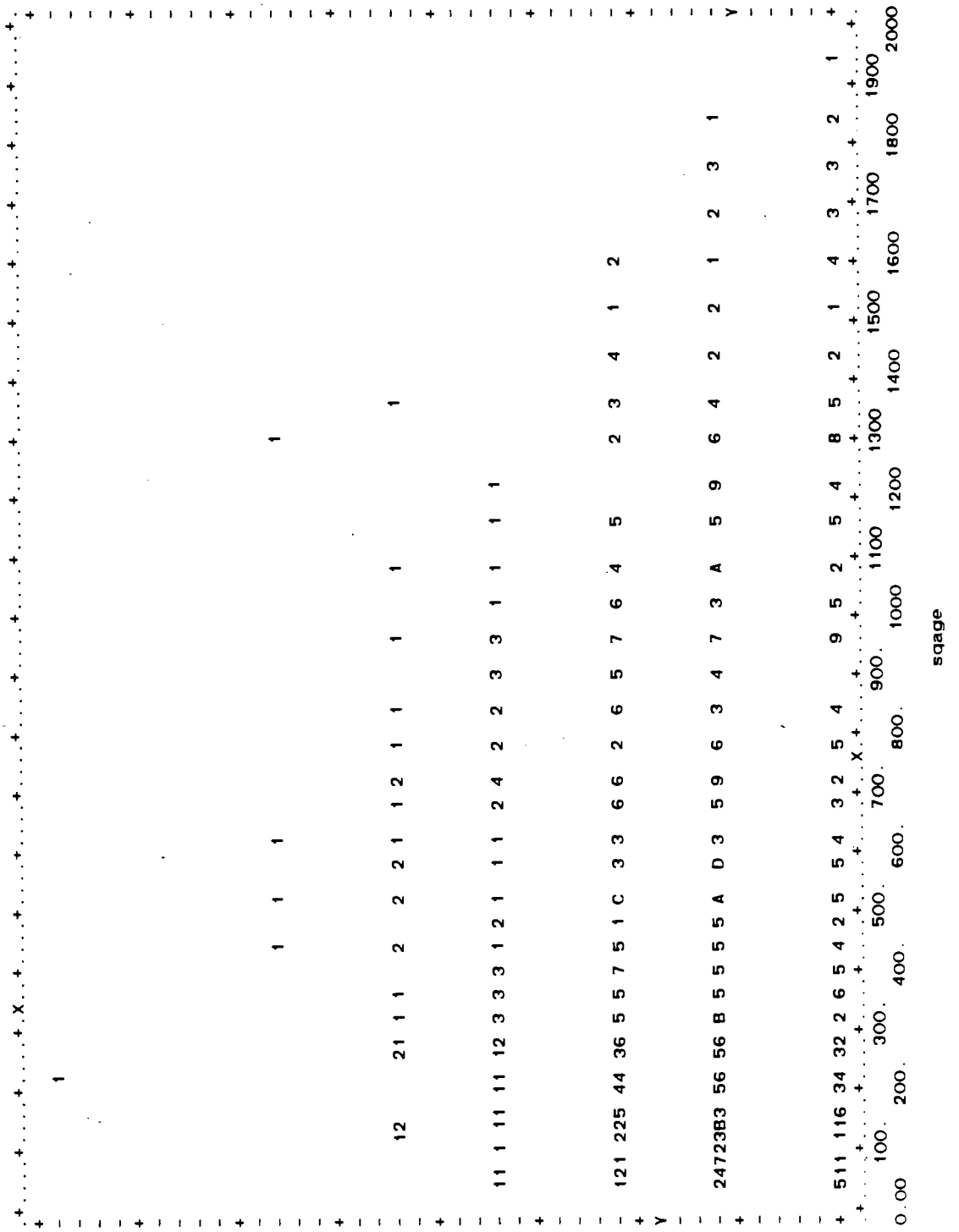


Figure 1.1: TOTFSH vs. SQCOL

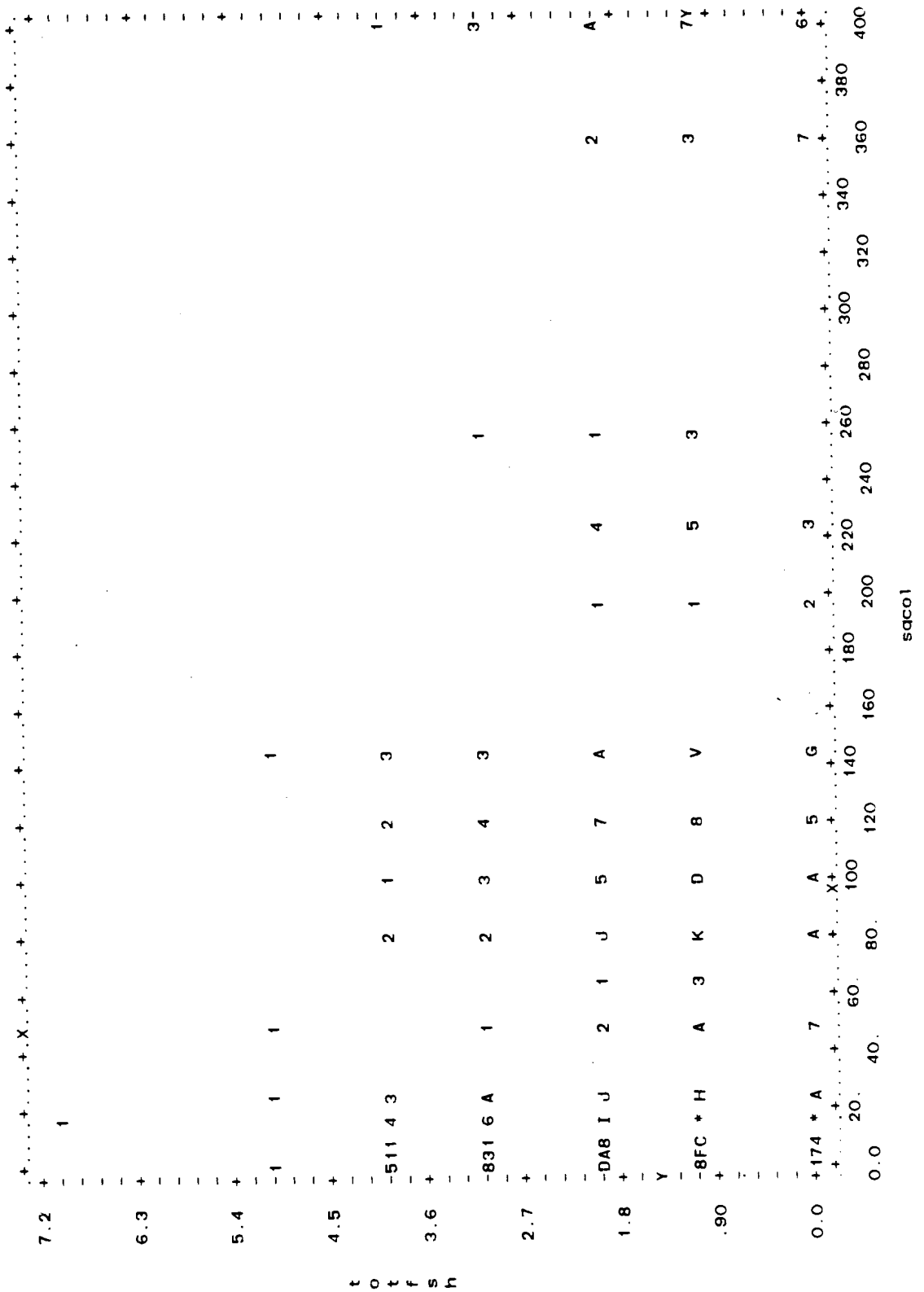


Figure 1.m: TOTFSH vs. SQNUM

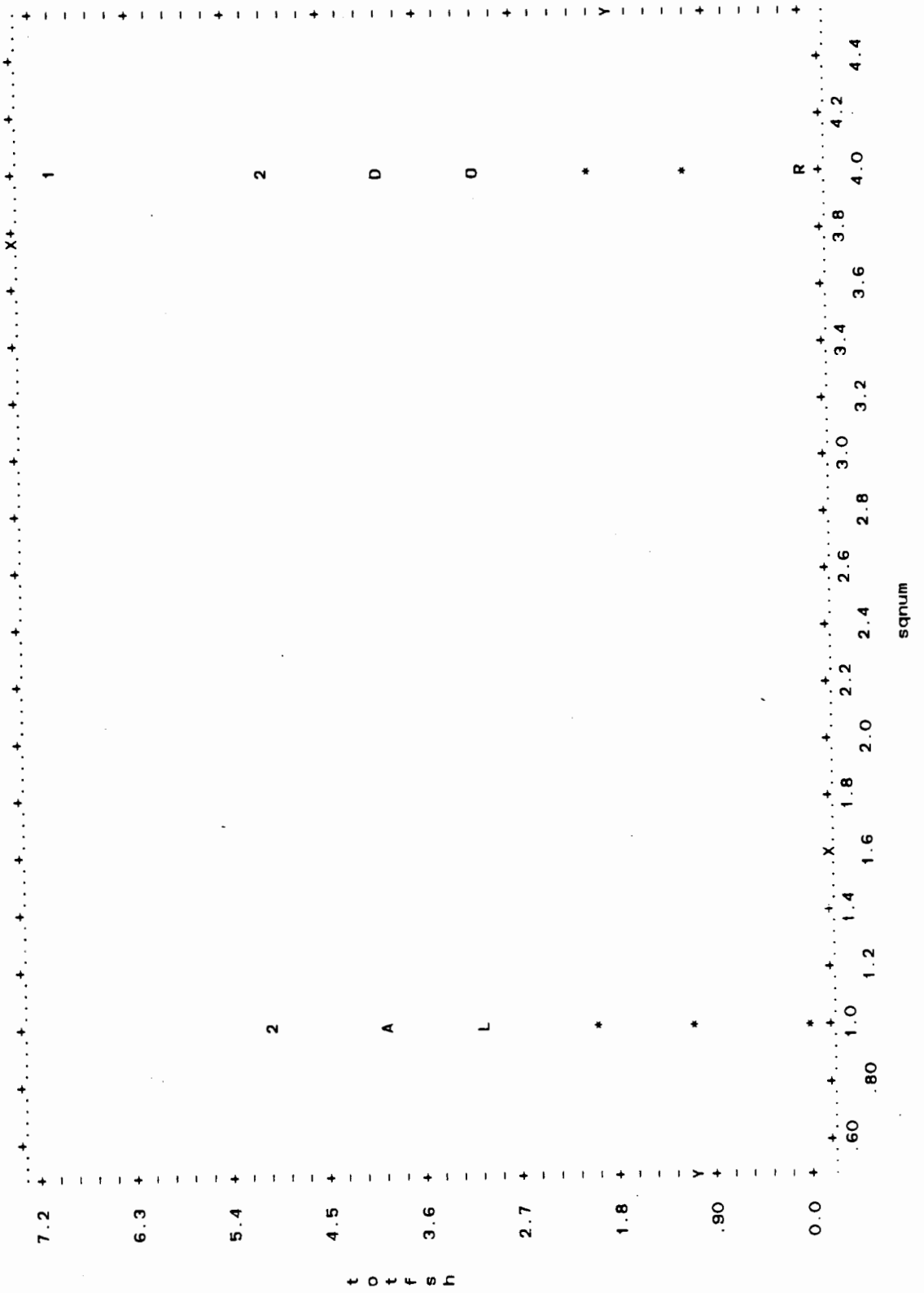


Figure 1.n: AGECHK vs. DATE

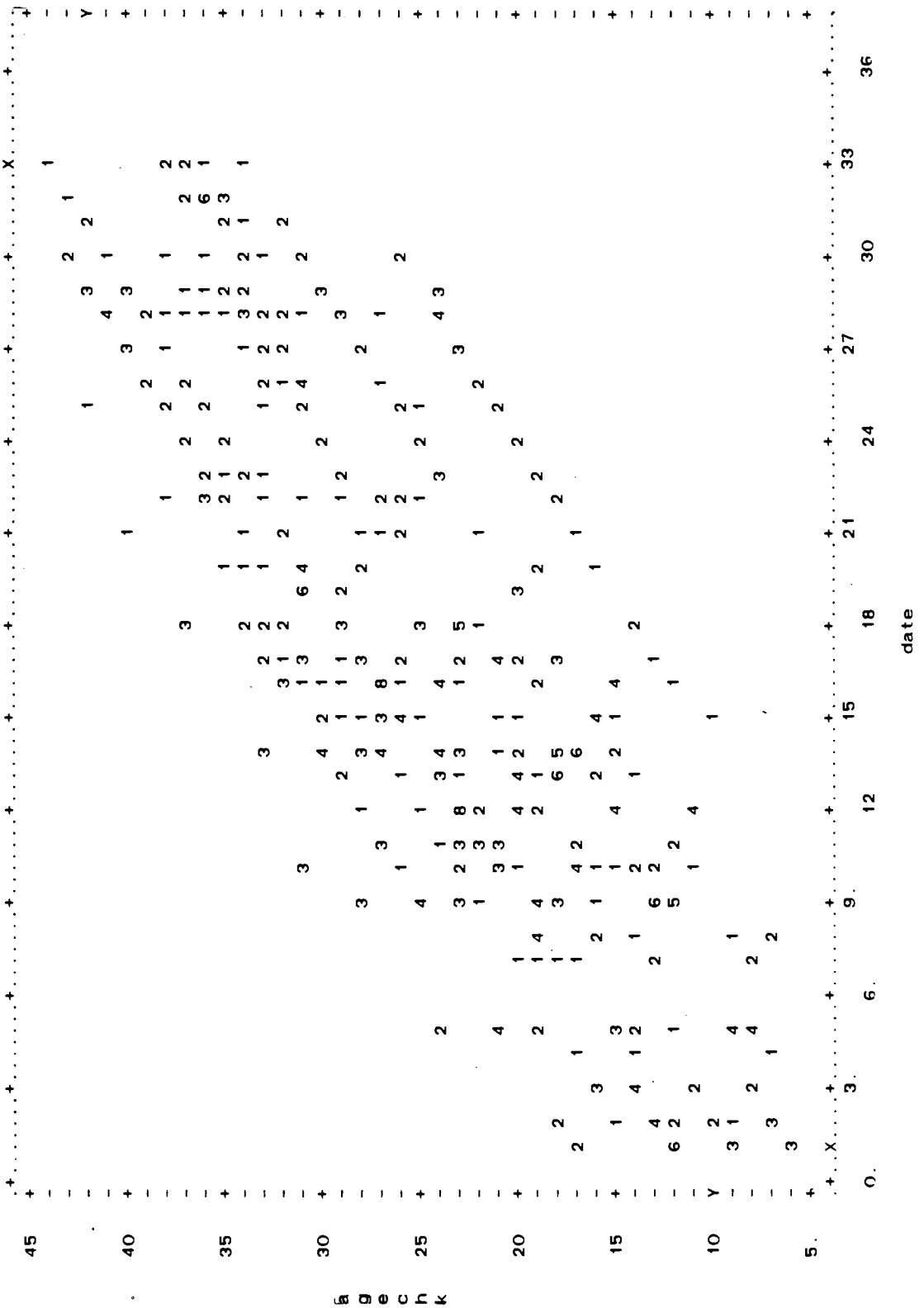


Figure 1.0: TIME vs. TIDE

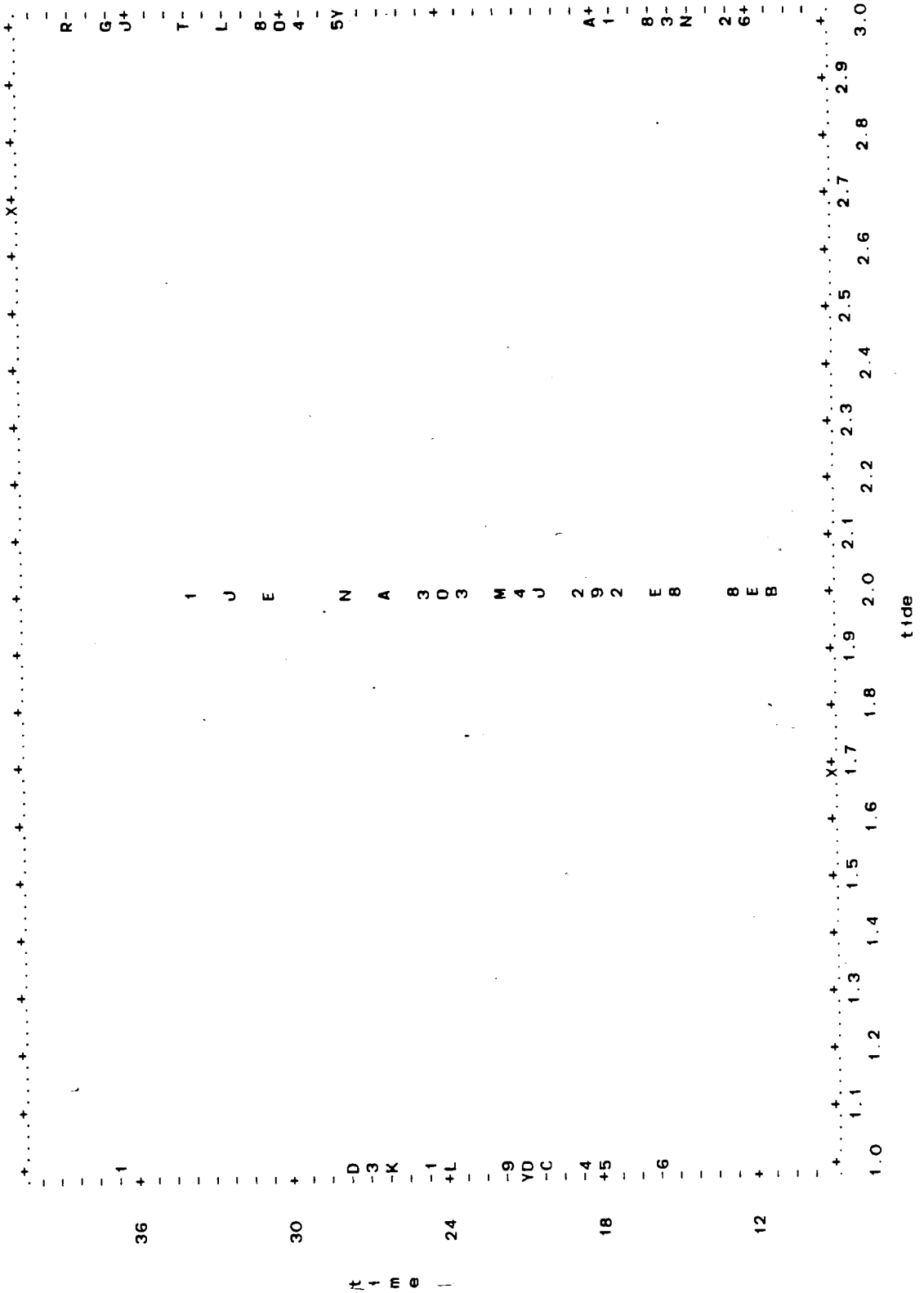
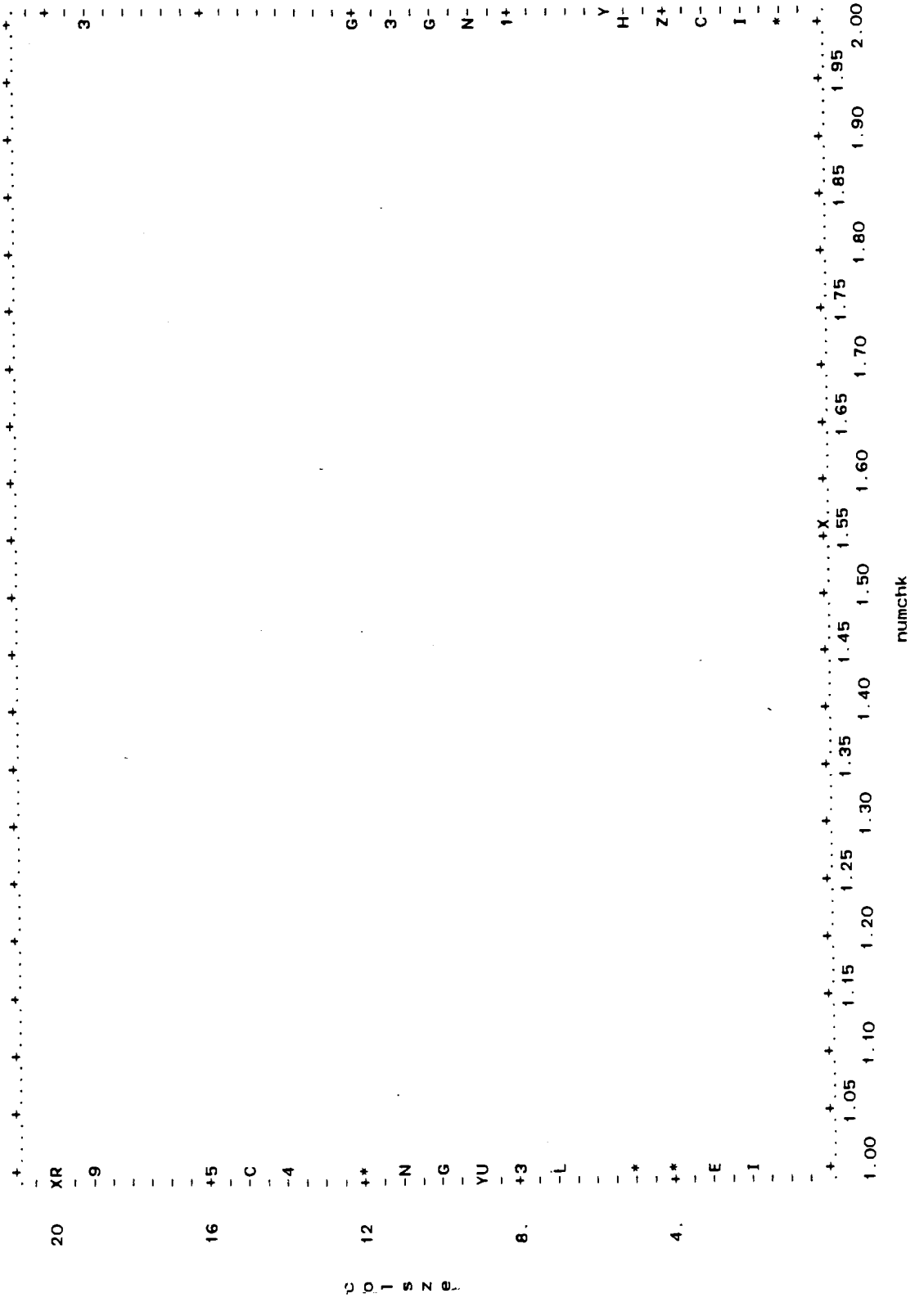


Figure 1.p: COLSZE vs. NUMCHK



One can spot the discrete nature of TOTFSH instantly by noticing in the plots that the points spread themselves out on horizontal parallel lines located at integer values of the TOTFSH axis. In these plots a '1' is used to indicate a single point, & higher numbers are used to show how many points are occupying the same location (or at least so nearly so as to be indistinguishable on the chosen scale). Furthermore an 'A' represents 10 points, 'B' represents 11 points, and so on up to 'Z' which represents 35 points. Finally, a '*' is used for 36 or more points.

The plots of Figures 1.a-1.l can be divided into 3 groups. The first group consists of Figures 1.a-1.h, the second of Figures 1.j-1.m, and the third of Figures 1.n-1.p.

In their first group, TOTFSH is plotted against candidate variables DATE, COLSZE, TIDE, TIDEH, TIDEM, NUMCHK, TIME, and AGECHK. The purpose of these plots is to inspect whether or not the explanatory variable is associated with the response (TOTFSH), and if so, then in what sort of way (linear, quadratic, logarithmic, and so on). For the quantitative variables, there seems to be no strong suggestion of any trend which could not be reasonably approximated by a curve which is quadratic in the explanatory variable (that is, by a half or whole parabola) over the given range of that variable. As for the solitary qualitative variable TIDE, the design variables should be sufficient to explain any effects on FEEDRATE. The purpose

of doing plots of TOTFSH against TIDEH and TIDEM as well as was to derive possible further information from the TOTFSH versus TIDE graph in the case that only one tide level is associated with the response and no others. In fact one of the advantages of converting a quantitative variable into a qualitative or factor one is that associations with the response variable can be investigated without specifying the nature of how the response variable may depend on the explanatory variable (e.g. linearly, quadratically and so on) (Ref. (11) page 518), although some information is lost in the conversion.

The second plot group consists of TOTFSH against squares of the quantitative variables. The purpose here is to see if any further information can be obtained on how these variables may be associated with TOTFSH. The variance (spread) in the data, however, is so wide that it seems unlikely that any further enlightenment can be obtained. Undoubtedly this wide variance will lead to problems with model fit later on.

The third group of plots do not involve TOTFSH at all. Their purpose is to explore possible correlations between pairs of explanatory variables. Such correlations are called multicollinearities and they are responsible for how the importance of a variable in predicting the response variable's outcome may change as further explanatory variables are added/dropped from the model. Selection of

pairs for these plots was done on the basis of prior suspicions about what variables may affect each other. One can see from these plots, for example, a high correlation between AGECHK and DATE, which is not surprising. Each variable is still to be considered as a separate candidate for the model since AGECHK may affect FEEDRATE through the demand for fish, and seasonal effects of fish migration (recorded through DATE) may affect FEEDRATE through the supply of fish.

Having now decided to use design variables for TIDE and to allow both linear and quadratic terms in the remaining quantitative variables in the model, attention could now be turned to model searching. Figure 2 shows some highlights of an attempt at finding a 'best' model using a best k subsets regression program, P9R, from BMDP (Section 13.3 of Ref.(4)). In the complete output (too voluminous for inclusion) the various subsets of explanatory variables are shown for models containing 1 variable up to the model which contains all of them. Here, 'best' is defined as maximal value of the quantity:

$$R^2 = 1 - \frac{SS(error)}{SS(total)}$$

$$= \frac{SS(regression)}{SS(total)}$$

familiar from standard regression texts, such as Ref.(11). $SS(total)$ depends only on observed feedrate values & will remain constant for all models, whereas $SS(error)$ will

Figure 2: Highlights of P9R run

SUBSETS WITH 5 VARIABLES

R-SQUARED	ADJUSTED R-SQUARED	CP	VARIABLE	COEFFICIENT	T-STATISTIC
0.123194	0.114731	10.55	time	-0.0844785	-4.16
			numchk	0.243053	4.84
			agechk	0.0544920	3.92
			sqtime	0.00156383	3.86
			sqage	-0.00120941	-4.33
			INTERCEPT	0.873322	
0.107497	0.098883	19.90	VARIABLE	COEFFICIENT	T-STATISTIC
			date	0.0286160	2.57
			time	-0.0818375	-3.98
			numchk	0.239917	4.56
			sqtime	0.00152468	3.72
			sqdate	-0.000977473	-3.09
			INTERCEPT	1.21852	

SUBSETS WITH 11 VARIABLES

R-SQUARED	ADJUSTED R-SQUARED	CP	VARIABLE	COEFFICIENT	T-STATISTIC
0.140891	0.122434	12.00	date	0.00902205	0.65
			colsze	-0.0292729	-1.73
			tidehi	-0.151389	-1.89
			tidemed	-0.114825	-1.74
			time	-0.108975	-4.55
			numchk	0.181468	3.12
			agechk	0.0451773	2.65
			sqtime	0.00210170	4.25
			sqdate	-0.000369618	-0.98
			sqage	-0.000990103	-2.98
			sqcol	0.00103880	1.27
			INTERCEPT	1.50021	

STATISTICS FOR 'BEST' SUBSET

MALLOWS' CP	12.00
SQUARED MULTIPLE CORRELATION	0.14089
MULTIPLE CORRELATION	0.37535
ADJUSTED SQUARED MULT. CORR.	0.12243
RESIDUAL MEAN SQUARE	0.285708
STANDARD ERROR OF EST.	0.534517
F-STATISTIC	7.63
NUMERATOR DEGREES OF FREEDOM	11
DENOMINATOR DEGREES OF FREEDOM	512
SIGNIFICANCE (TAIL PROB.)	0.0000

change with each model since it depends on which variables are in the model.

Note that the variable

$$\text{SQNUM} = (\text{NUMCHK})^2$$

has not been included in the list of explanatory variables. The reason is that since NUMCHK has a limited range of only 2 values (1 or 2 chicks in nest), one can write:

$$1 = \frac{3}{2}(\text{NUMCHK}) - \frac{1}{2}(\text{SQNUM})$$

Such linear dependencies are not acceptable. This would not occur if NUMCHK could vary over a wider range.

It should be pointed out that adding more variables simply because they deliver a higher R^2 is unwise since adding more explanatory variables to a model will never decrease R^2, and in fact usually increases it (Ref.(11) pg. 422). There must be a trade-off between maximizing R^2 and keeping the model as simple as possible, that is, by limiting the number of variables in the final model (Ref.(5) pg. 294). A helpful pictorial aid in this regard is an R^2-plot which plots R^2 against the number of regression coefficients, $(p+1)$, for a set of proposed models. Such a plot is shown in Figure 3, based on R^2-values for various models collected in Table 7. These models were obtained from various P9R and GLIM runs (the next chapter and the Technical Supplement will discuss the use of GLIM). Paths are drawn in Figure 3 to connect up successively nested

Figure 3: R^2-plot

Table 7: Collected Models

```
   1   0.0000   none                                %gm    +
   2     .0528  4                                   numchk+
   3     .0658  4,9                                 sqdate+
   4     .0769  4,9,3                               date   +
   5     .0840  4,9,3,6                             time   +
   6     .1075  4,9,3,6,10                          sqtime+
   8     .1169  4,9,3,6,10,5                        tide   +
   9     .1216  4,9,3,6,10,5,2                      colsze+
  10     .1257  4,9,3,6,10,5,2,7                    sqage  +
  11     .1386  4,9,3,6,10,5,2,7,1                  agechk+
  12     .1409  4,9,3,6,10,5,2,7,1,8                sqcol
   3     .0041  5                                                %gm+tide   +
   4     .0082  5,8                                                  sqcol  +
   5     .0170  5,8,10                                               sqtime+
   6     .0270  5,8,10,2                                             colsze+
   7     .0440  5,8,10,2,1                                           agechk+
   8     .0575  5,8,10,2,1,3                                         date   +
   9     .0722  5,8,10,2,1,3,9                                       sqdate+
  10     .0898  5,8,10,2,1,3,9,7                                     sqage  +
  11     .1063  5,8,10,2,1,3,9,7,4                                   numchk
   4     .0892  4,1,7
   5     .0980  6,4,1,7
   6     .1232  6,4,1,10,7
   7     .1278  2,6,4,1,10,7
   8     .1314  5,6,4,1,10,7
   9     .1350  2,5,6,4,1,10,7
  10     .1383  2,5,6,4,1,10,7,8
  11     .1402  2,5,6,4,1,10,9,7,8
C234567890123456789012345678901234567890123456789012345678901234567890123456789012
C         1         2         3         4         5         6         7
Number    R**2     Explanatory Variables Used (see label index below)
  of
Regression
Coefficients
Required,
  p
```

Explanatory Variable	Label
agechk	1
colsze	2
date	3
numchk	4
tide	5
time	6
sqage	7
sqcol	8
sqdate	9
sqtime	10

models (that is, any model in a path has all the explanatory variables of the model to the left of it, plus an extra one). One can look on the optimal model building process as finding the path of quickest ascent. Having found such a path, one can stop before the end when one finds a subset of explanatory variables which still deliver an R^2-value sufficiently close to the maximum available (obtained when all available variables are inside the model).

A certain amount of arbitrary choice on the part of the analyst is called for in this kind of data search analysis. It was decided that the following set of 5 variables gave a relatively high R^2 while still keeping the model simple: AGECHK,SQAGE,NUMCHK, TIME,SQTIME. From Figure 2, one can see that this is the "best" (in the R^2 sense) 5 variable model, and one can read off the calculated regression coefficients to propose the model:

$$\hat{Y}=0.873+0.243x_1+0.0545x_2-0.0845x_3-0.00121x_4+0.00156x_5$$

where:

$\hat{Y}$=fitted/predicted (not observed) FEEDRATE

x_1=NUMCHK

x_2=AGECHK

x_3=TIME

x_4=SQAGE=(AGECHK)$^2=x_2^2$

x_5=SQTIME=(TIME)$^2=x_3^2$

This model has

$$R^2=0.1232$$

which is not impressive. Consider

$$SS(total)=\sum_{i=1}^{n}(Y_i-\bar{Y})^2$$

which is a measure of the total spread/variation in the response variable Y, where

$$\bar{Y}=\frac{1}{n}\sum_{i=1}^{n}Y_i$$

is the sample mean. Consider also

$$SS(error)=\sum_{i=1}^{n}e_i^2$$

$$=\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2$$

which is a measure of total lack of fit (n=524 observations in each of the formulas) for a particualr model. Recall that for any model

$$0\leq SS(error)\leq SS(total)$$

so that R^2 is a measure of what proportion of the spread in FEEDRATE is explained by a particular model. Thus the model given previously explains 12.32% of this spread, so chance variation alone must cause the remaining 87.68%. Using all 11 explanatory variables does little to improve the situation for then

$$R^2=14.09\%$$

from Figure 2. The 1.77% improvement comes with a cost of adding 6 more explanatory variables to the model.

Nonetheless this 5-variable model shall be referred to hereafter as the 'current best model'. In the next chapter, means of improving this model (increasing R^2 without adding too many new variables) will be explored.

First some other aspects of this current best model need to be checked. These are the various plots which may give an indication of why the fit is so poor, and what could be done about it. Another P9R run was done, but on the current best model alone, in order to generate the plots. The results are shown in Figures 4.a-4.p. It will be noticed that the main emphasis in these plots is on the residuals, e_i.

The first plot, Figure 4.a, shows one of the most important plots for assessing a model's overall fit: residuals against fitted values, $\hat{Y}_i$. P9R labels this latter quantity 'PREDICTD'. Now if a model does fit the data well, such a plot should show a (nearly) horizontal band of constant width containing the points (Ref. (6), pg. 314). The plot shown here seems tricky to interpret, but one can gain some insight into what the plot is trying to convey by considering the second plot: residuals versus observed values, Y_i, that is, the FEEDRATE values. This plot is in Figure 4.b.

In this plot vertical line segments will be noticed. These reflect the discrete nature of the horizontal-axis variable, Y. As was explained earlier, Y is one-half of

Figure 4.a: Residual vs. Fitted

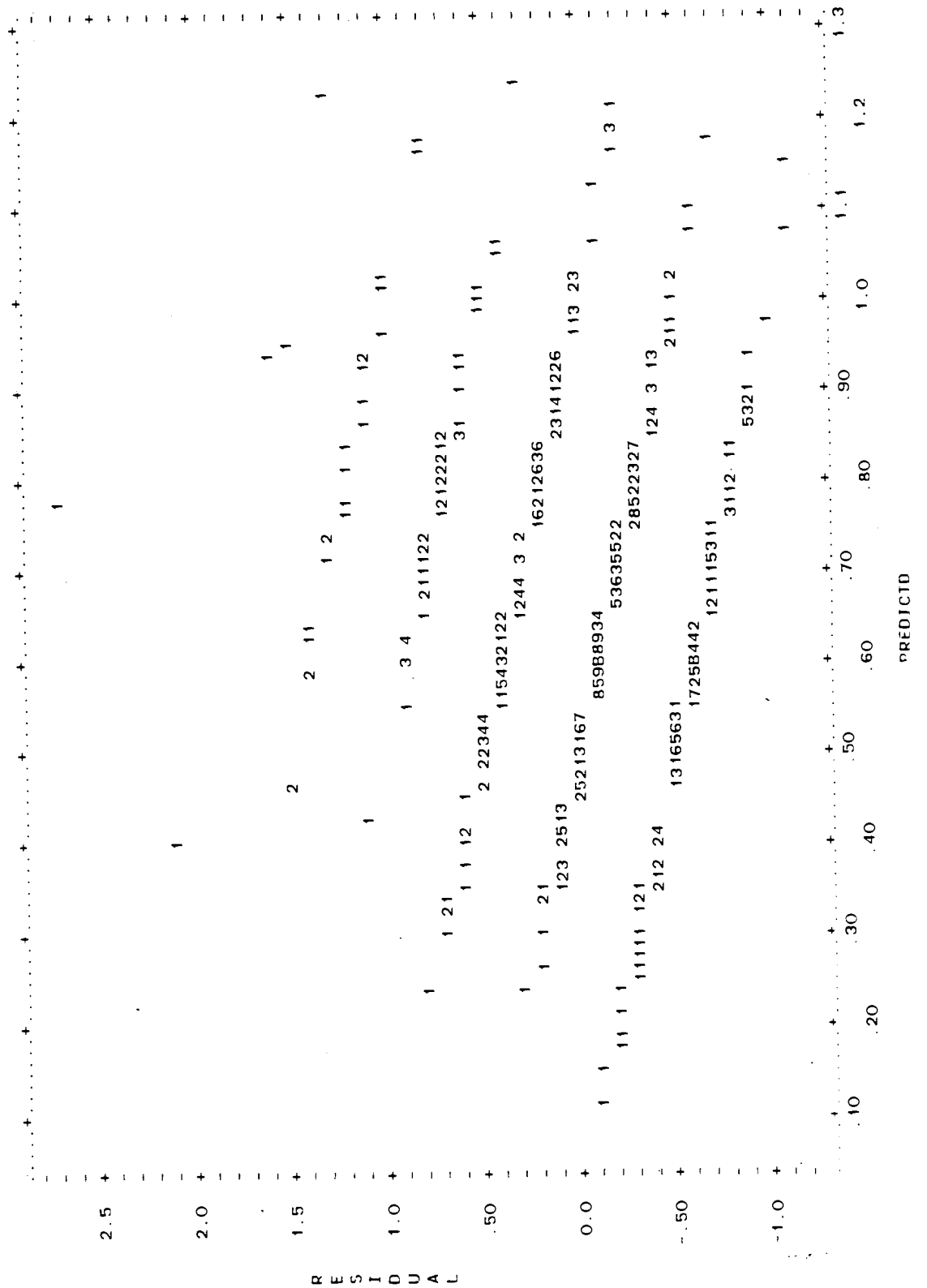


Figure 4.b: Residual vs. Observed

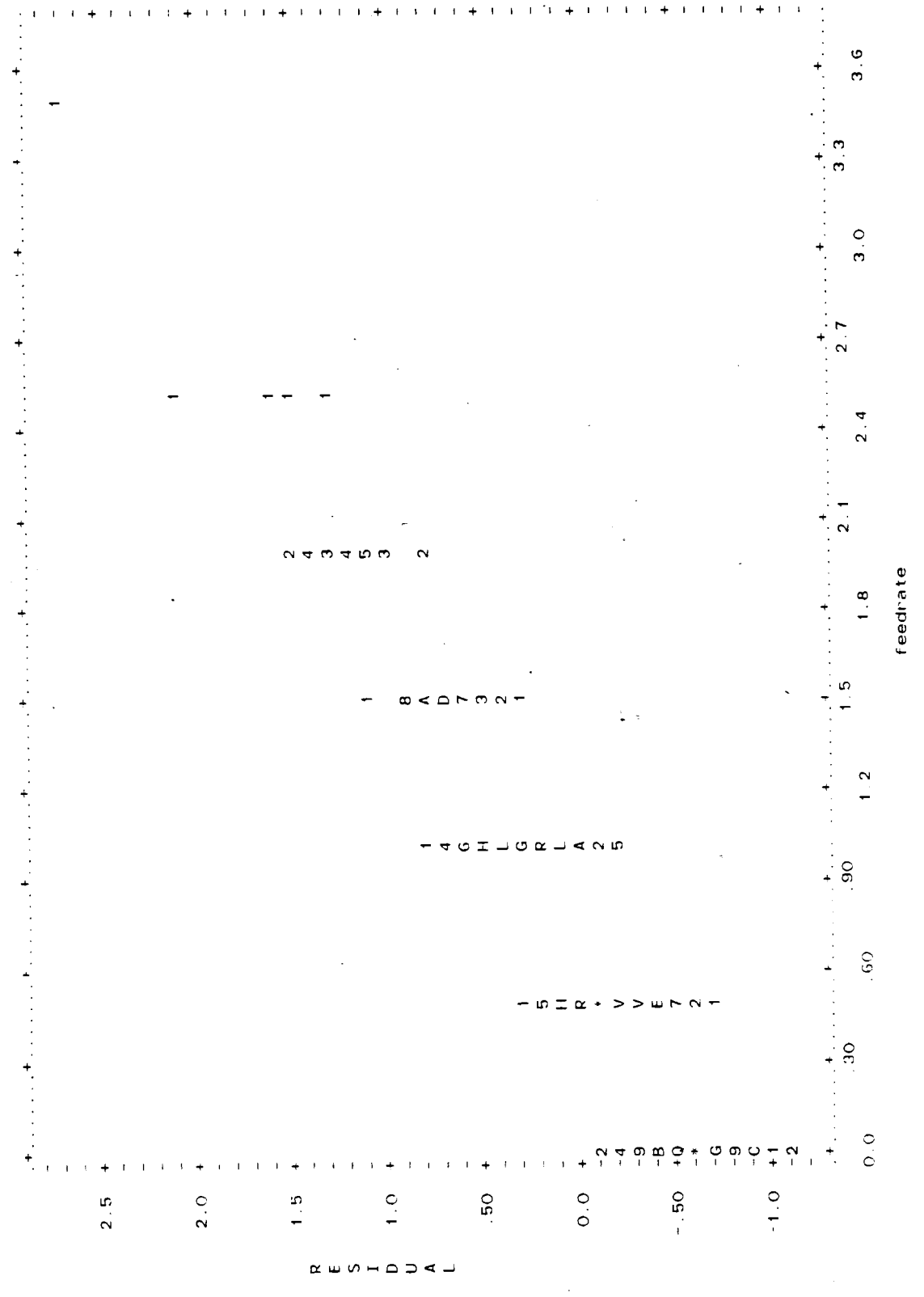


Figure 4.c: Residual vs. NUMCHK

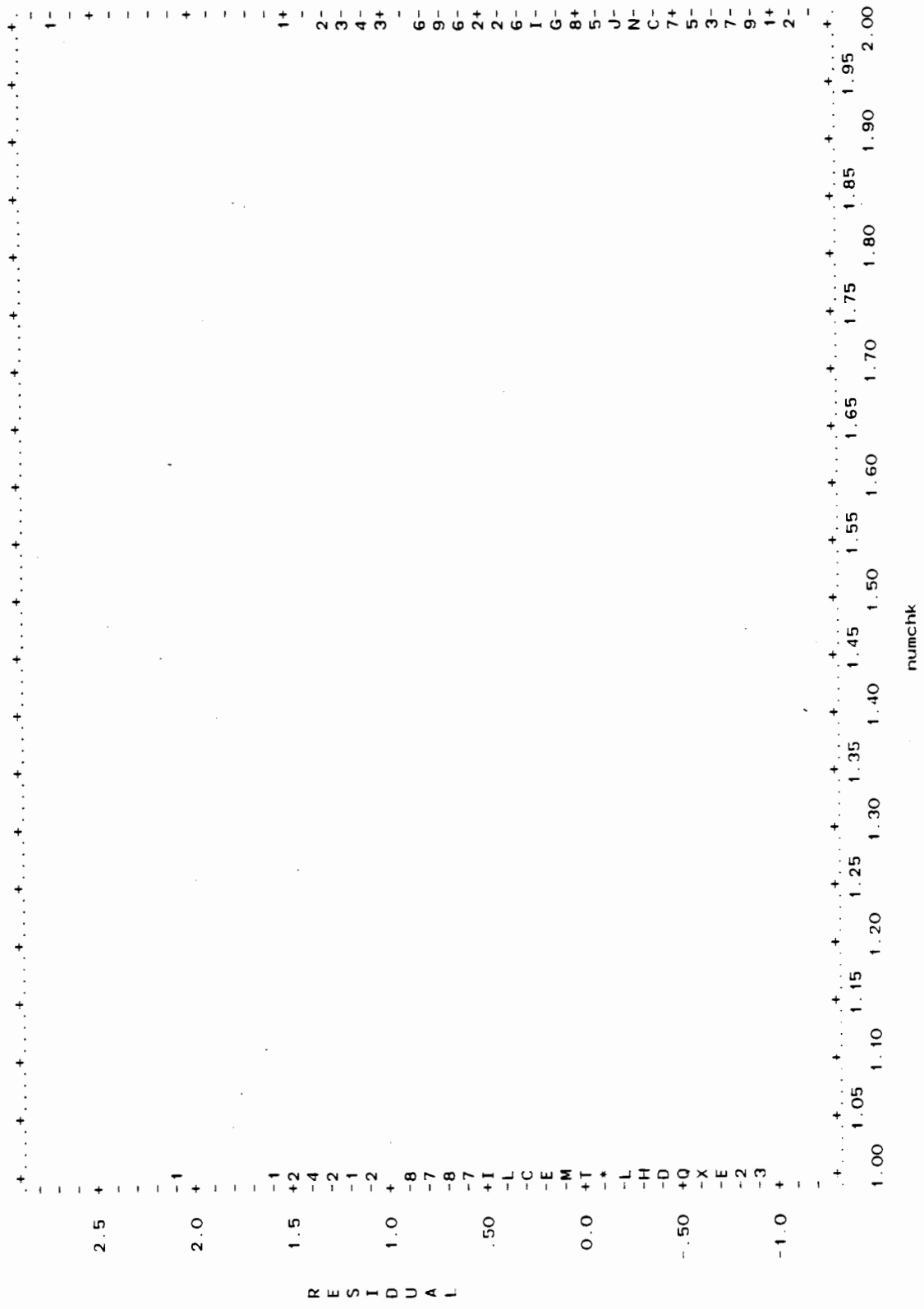


Figure 4.d: Residual vs. AGECHK

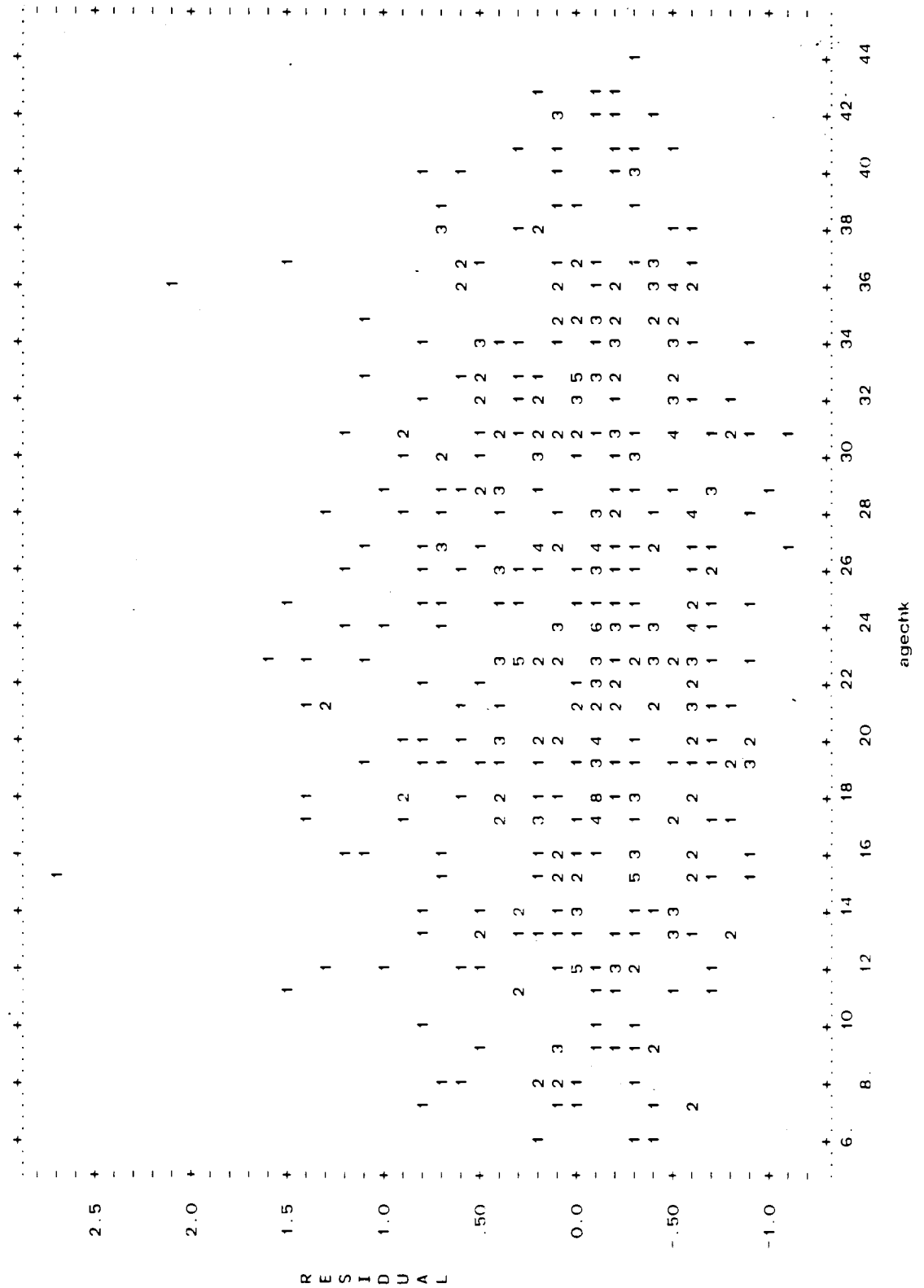


Figure 4.e: Residual vs. TIME

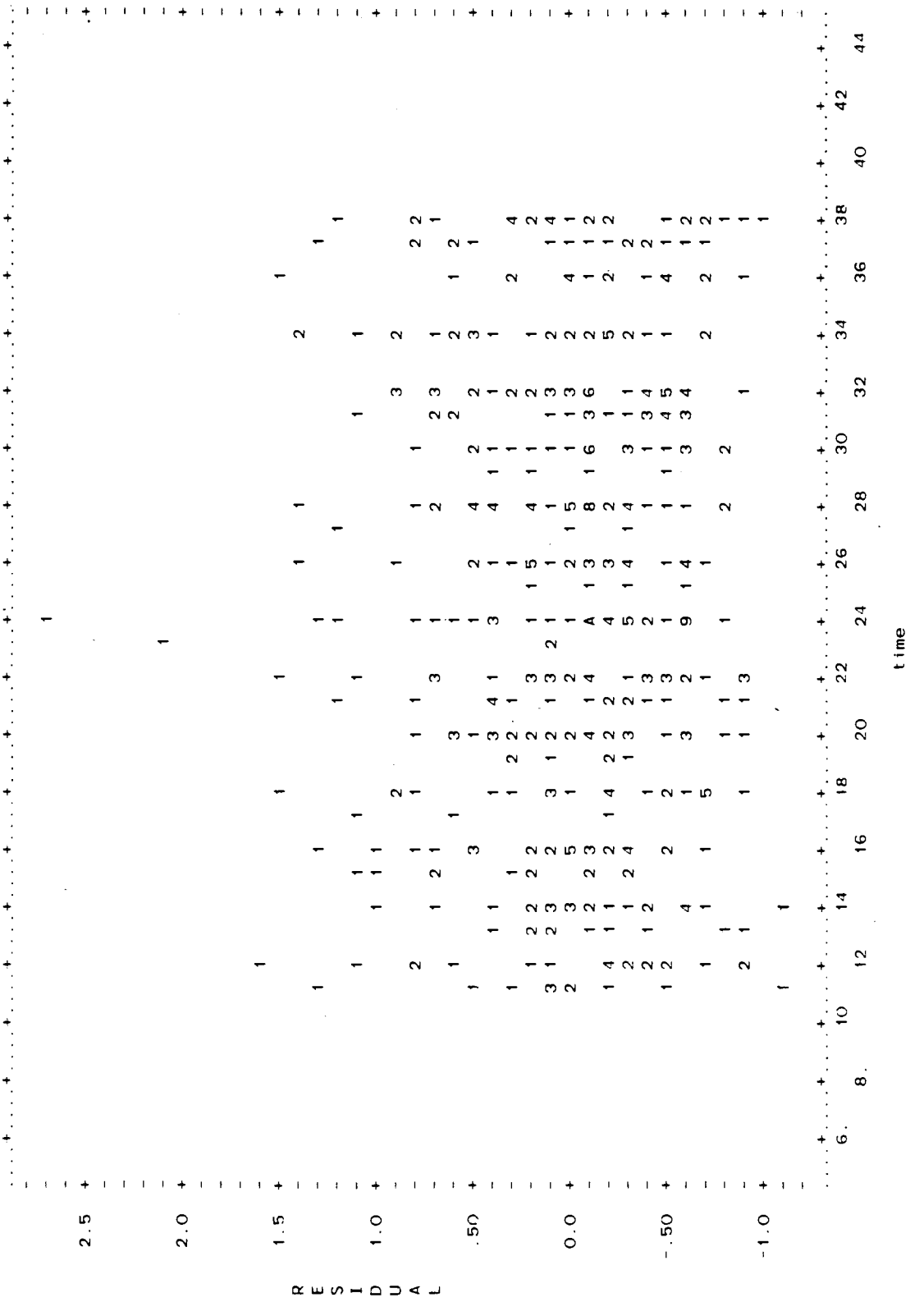


Figure 4.f: Residual vs. SQAGE

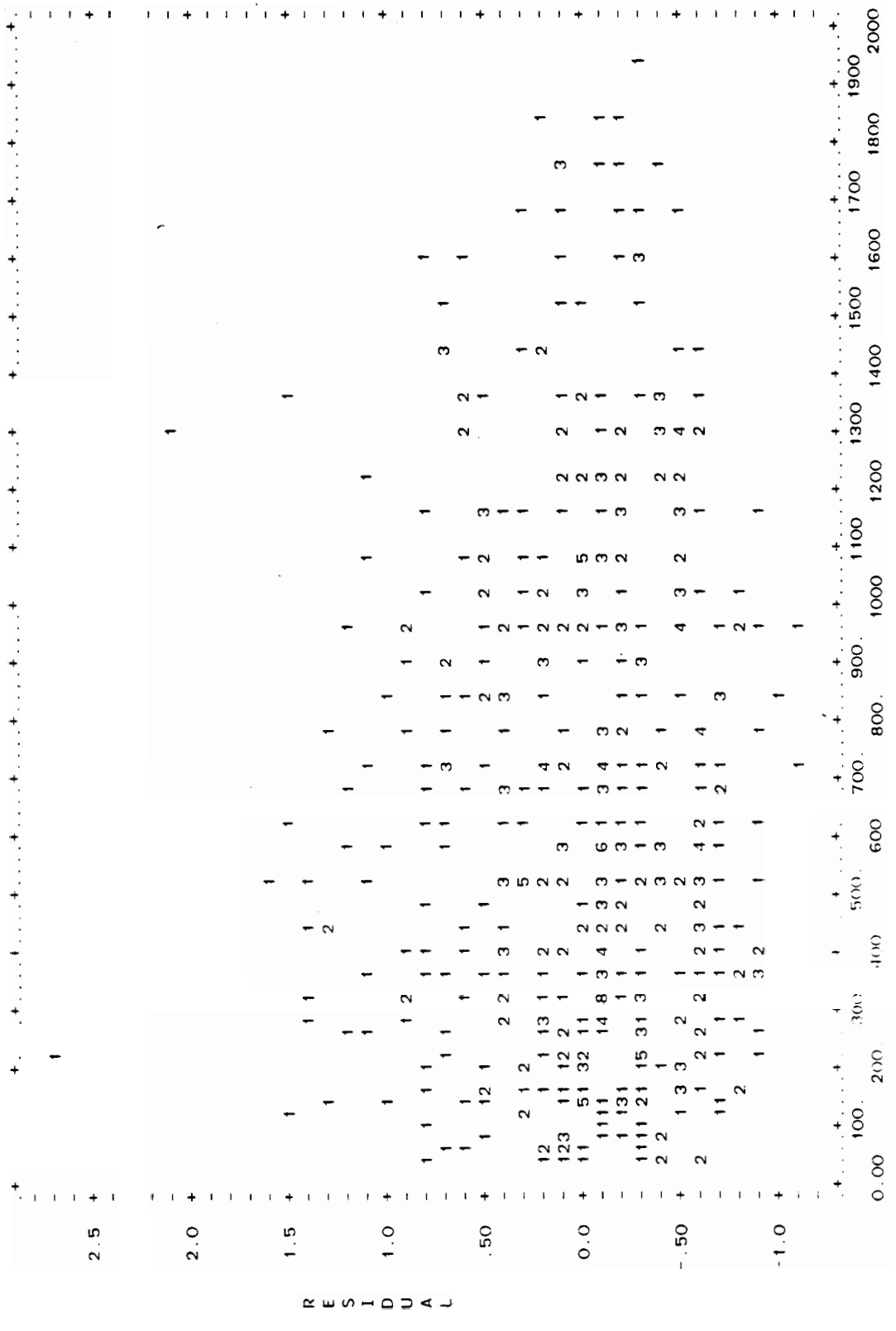


Figure 4.g: Residual vs. SQTIME

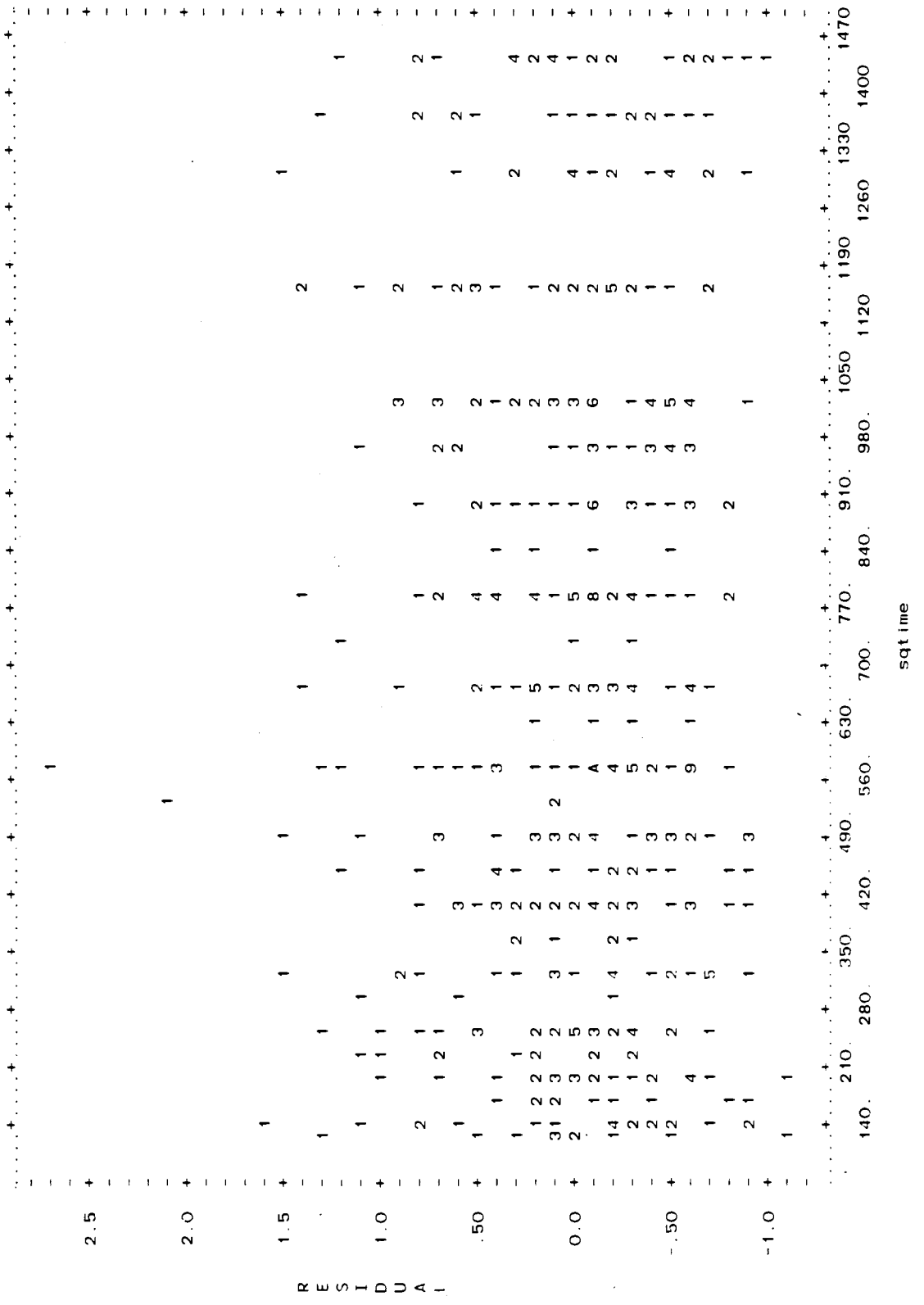


Figure 4.h: Residual vs. COLSZE

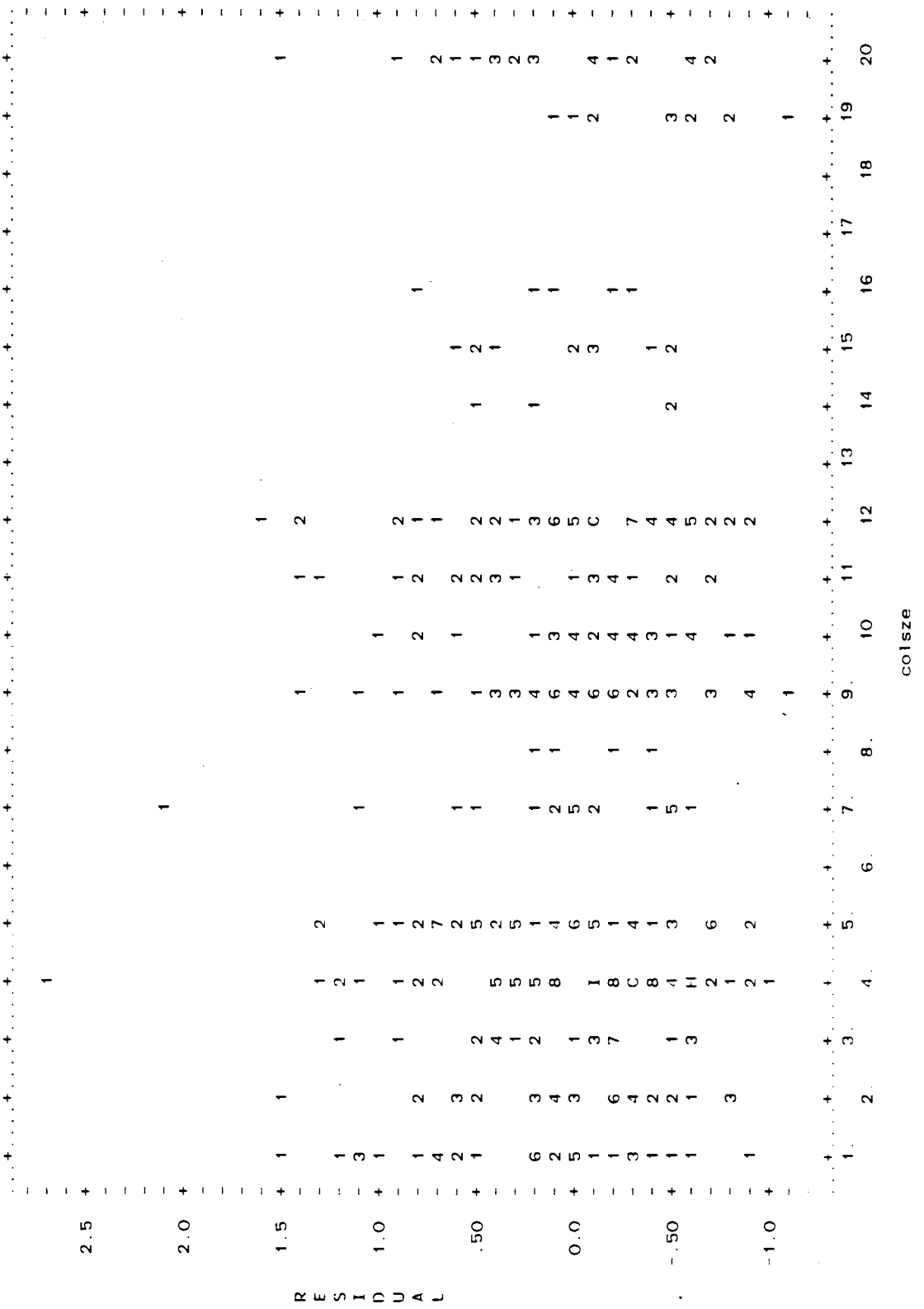


Figure 4.i: Residual vs. DATE

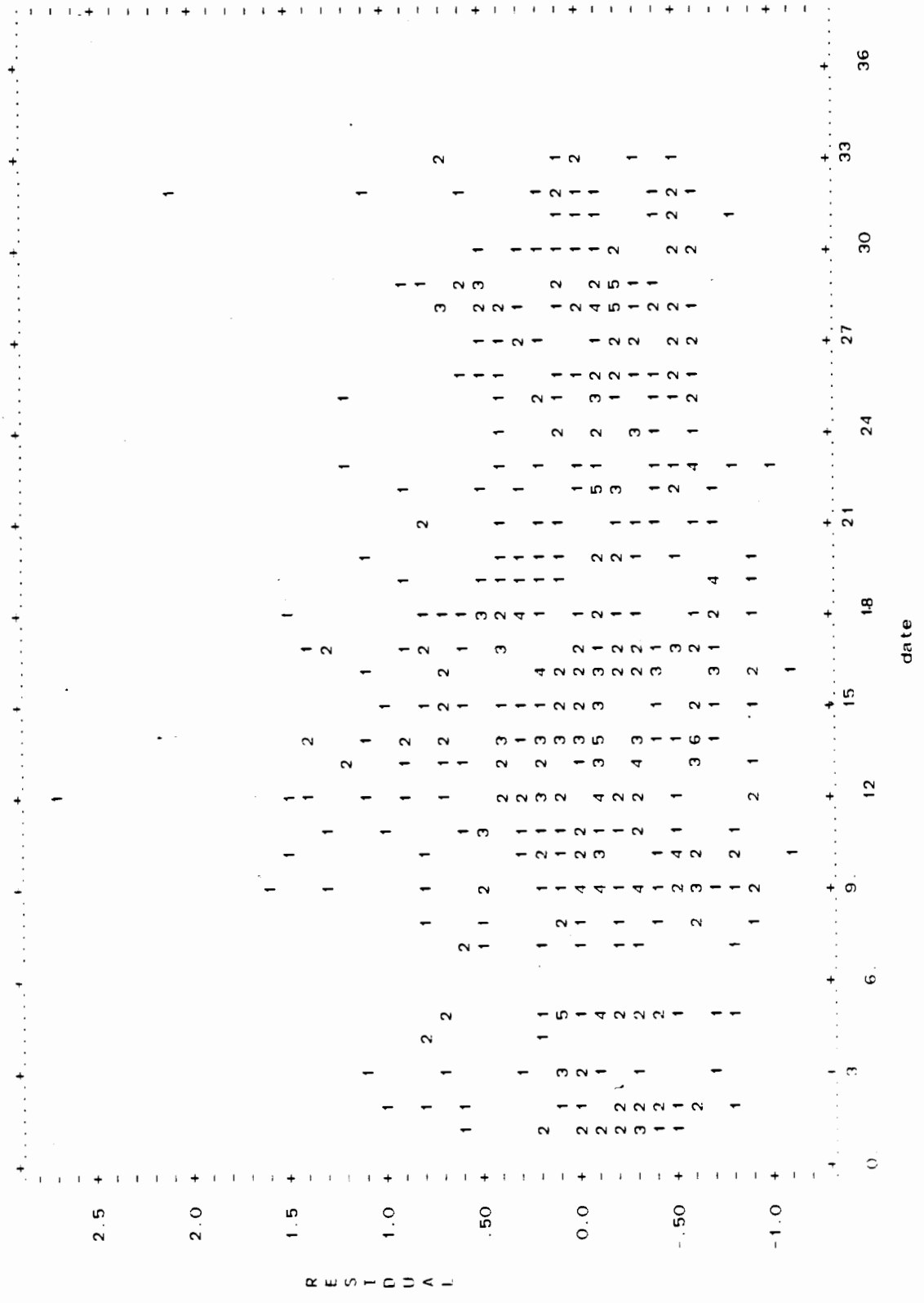


Figure 4.j: Residual vs. TIDE

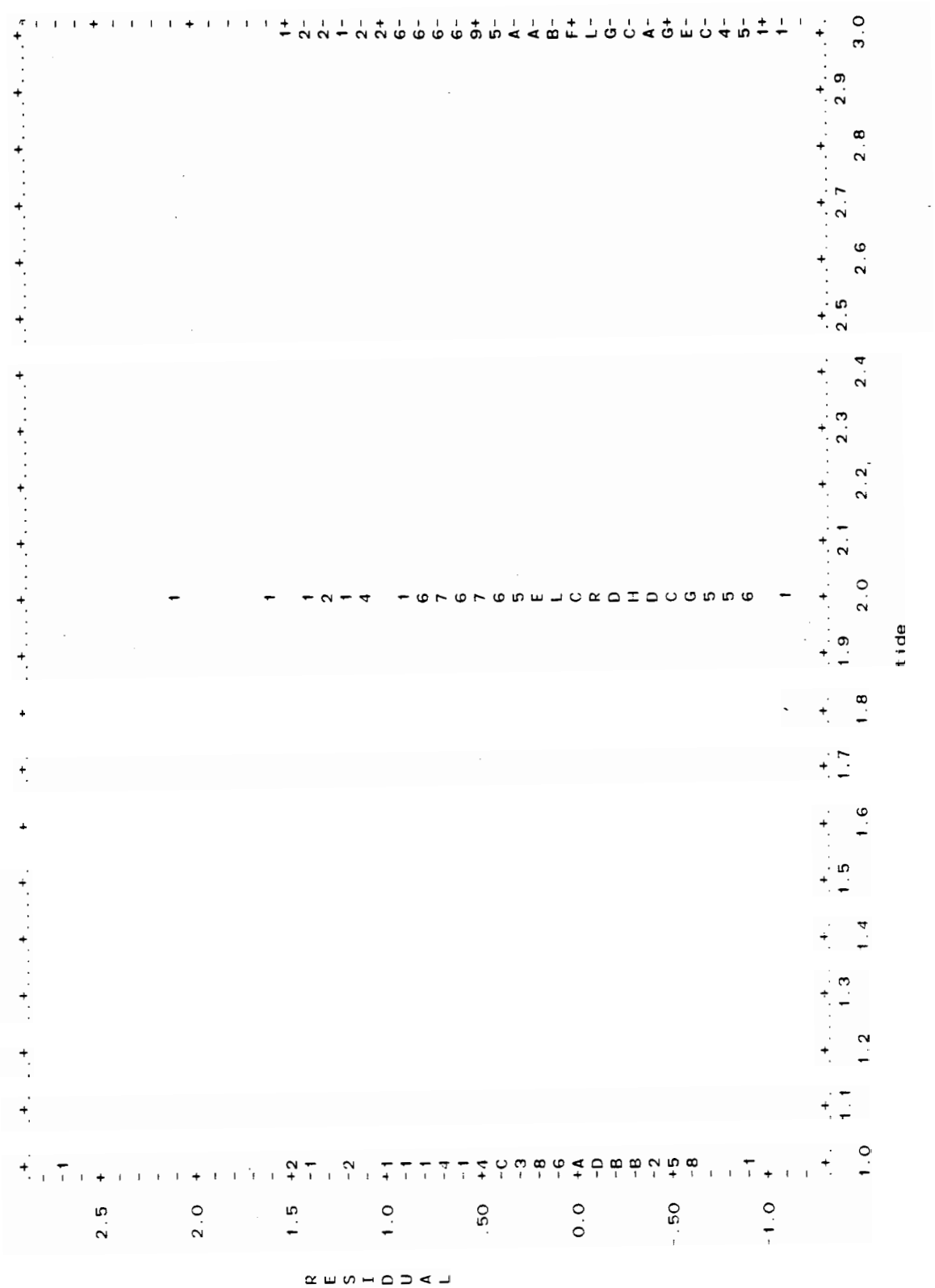


Figure 4.k: Reisudal vs. TIDEH

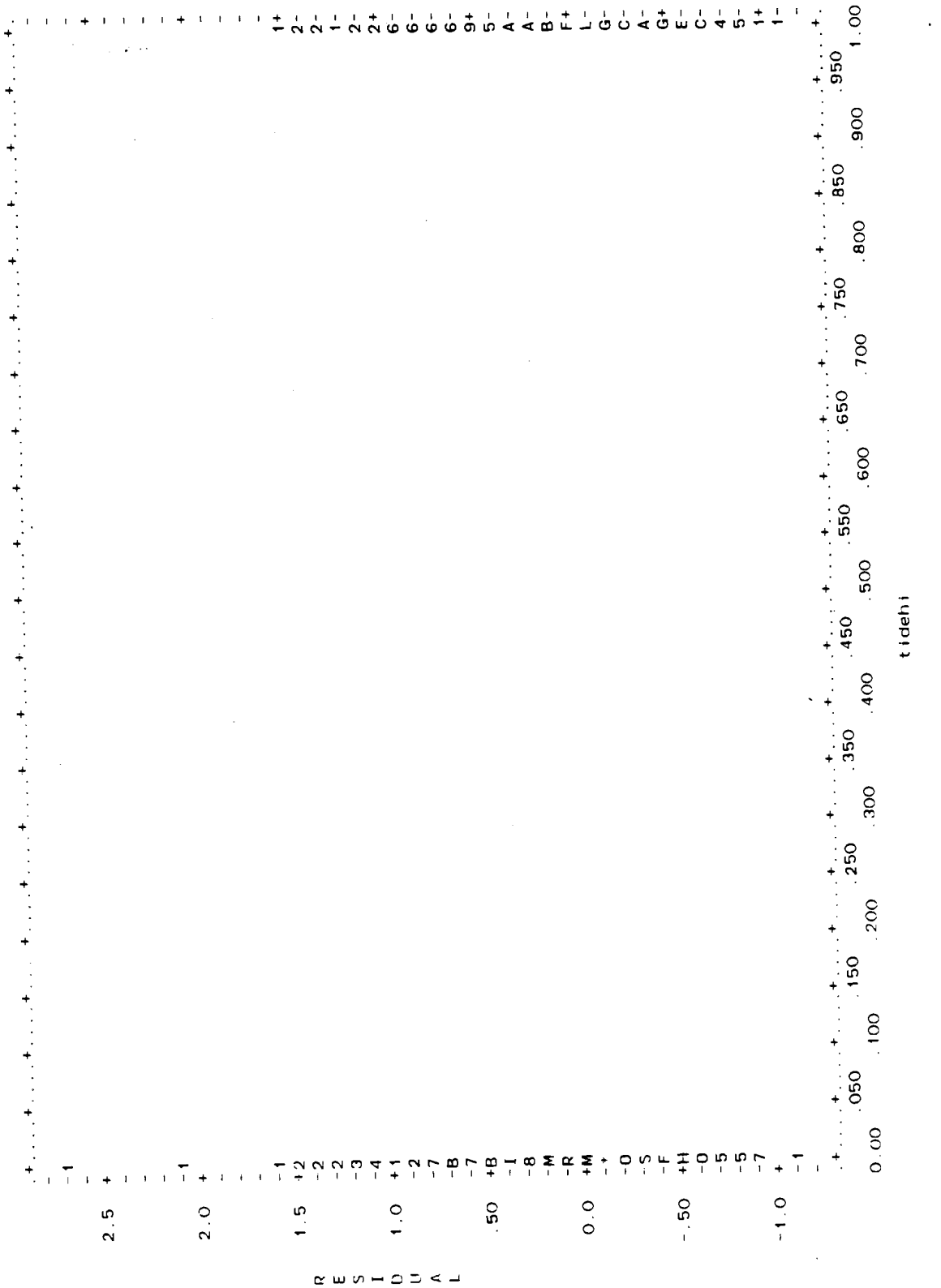


Figure 4.1: Residual vs. TIDEM

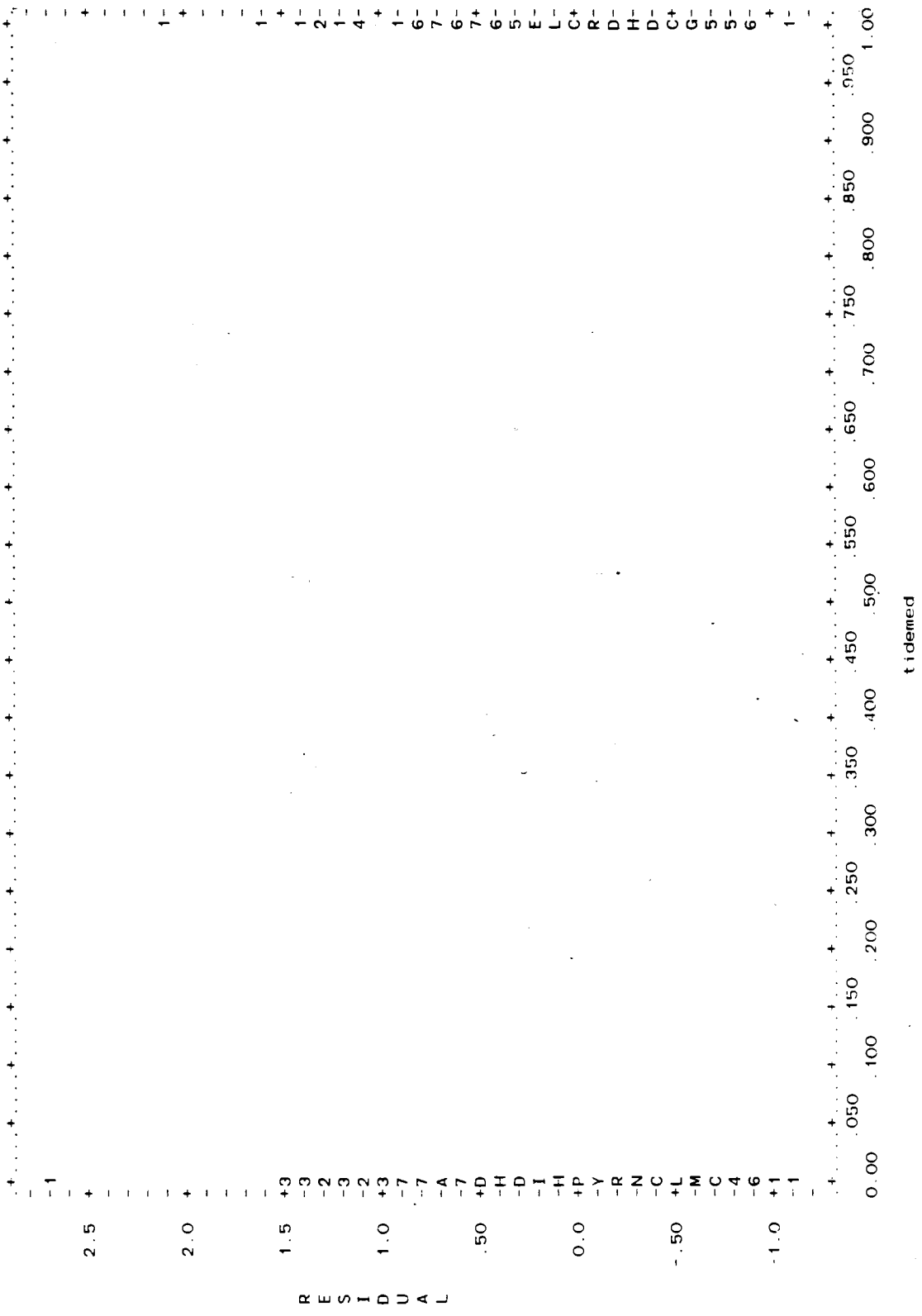


Figure 4.m: Residual vs. SQCOL

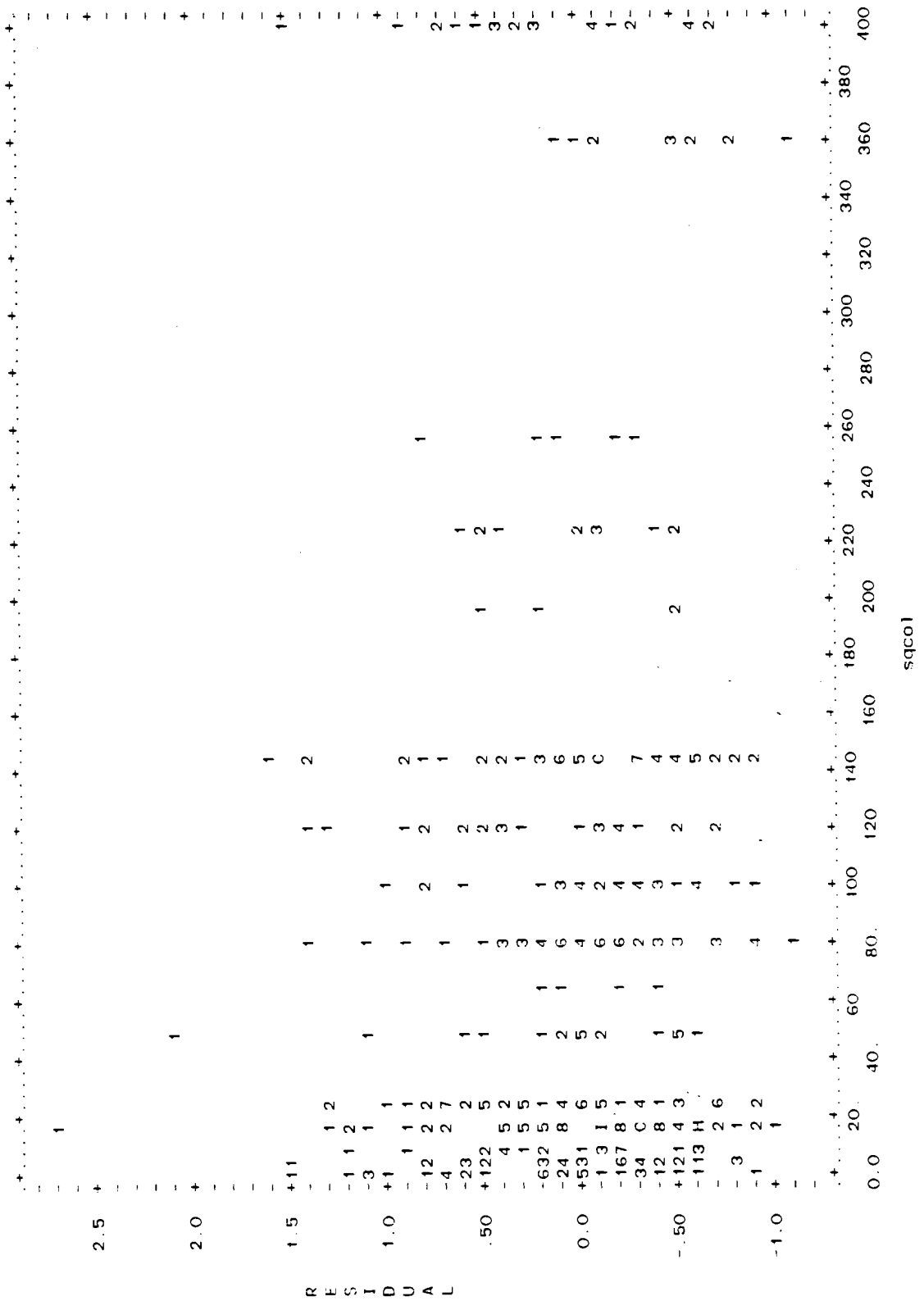


Figure 4.n: Residual vs. SQDATE

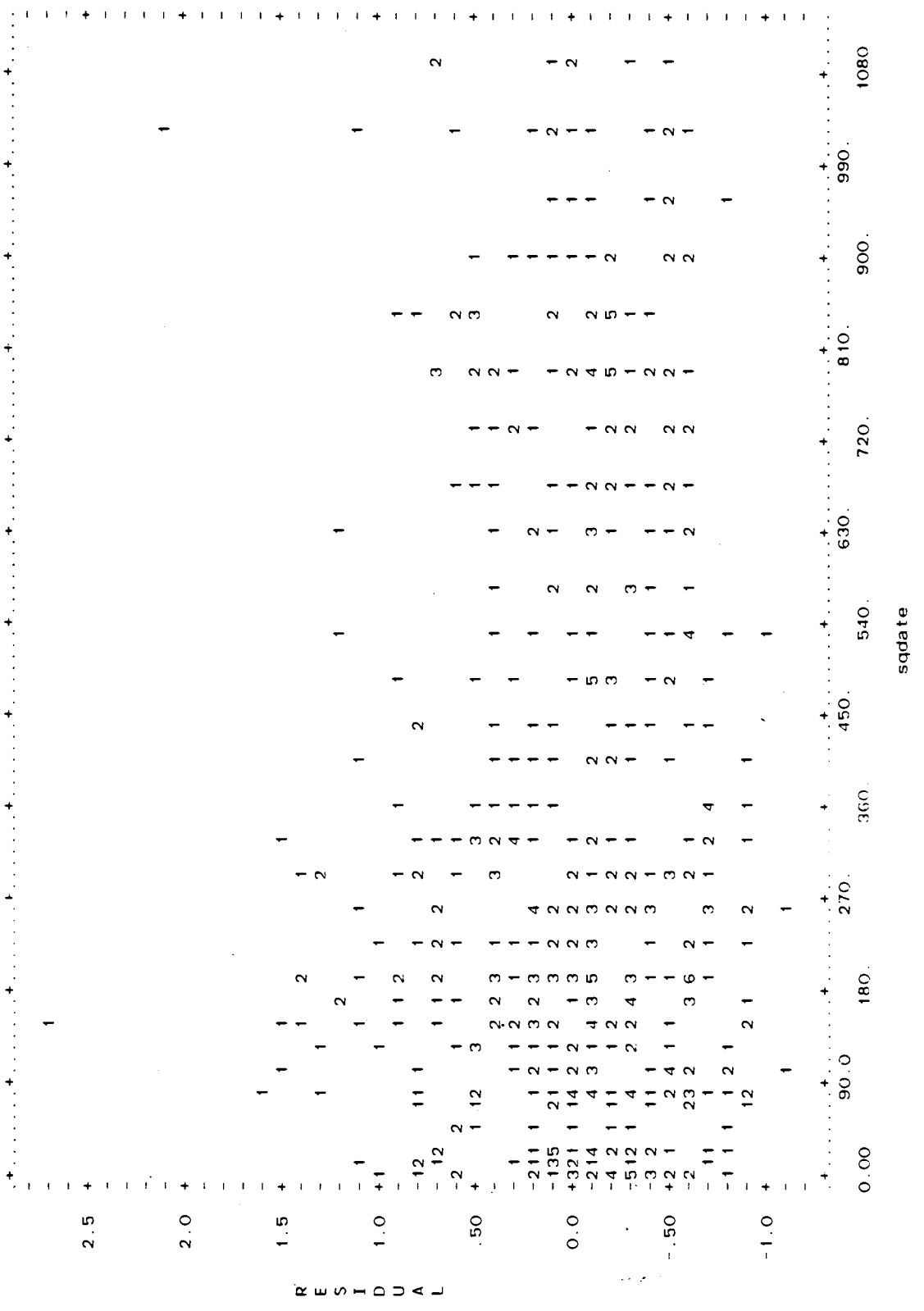


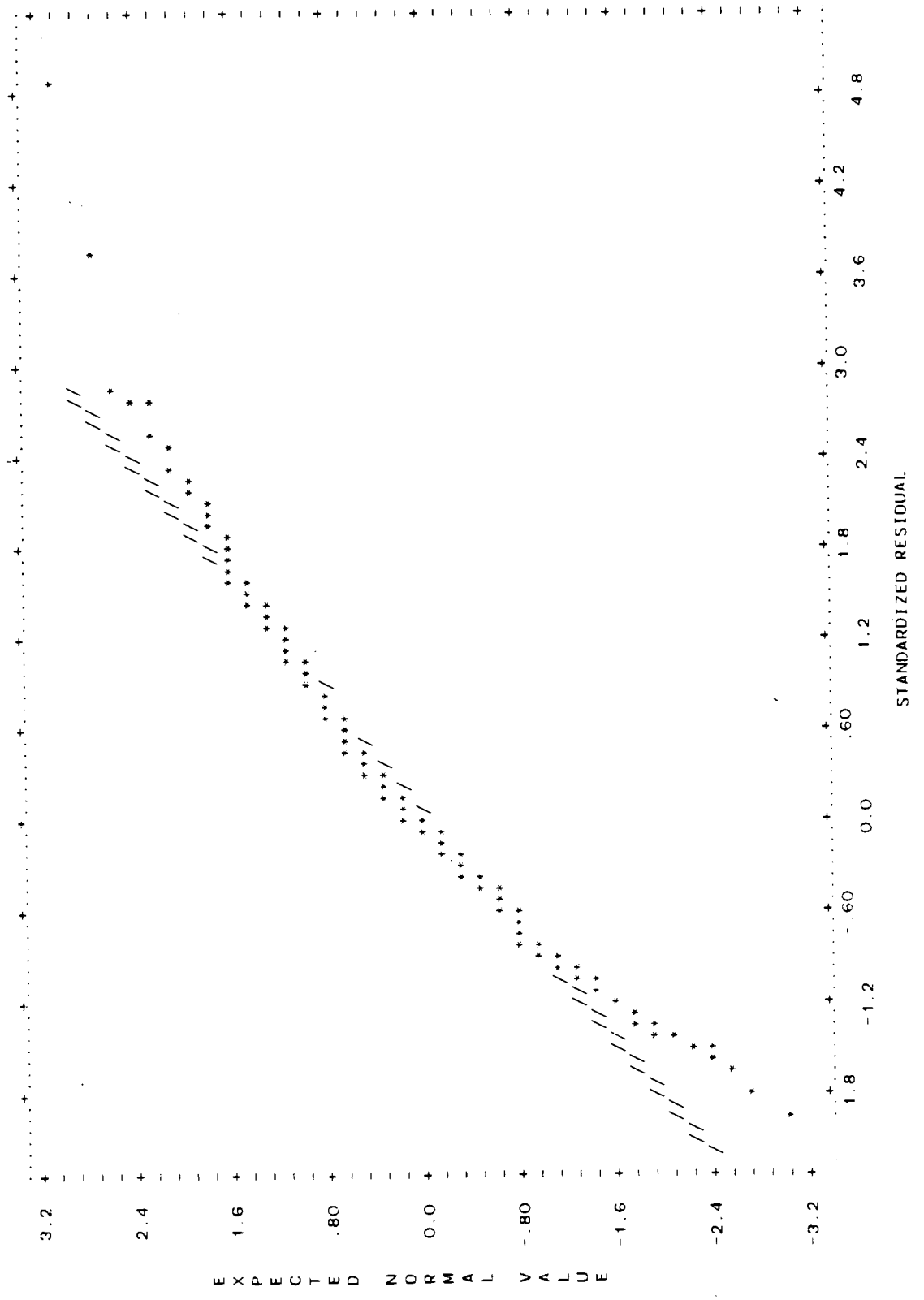
Figure 4.o: Histogram of Standardized Residuals

```
------------------------------------------------------------

HISTOGRAM OF STANDARDIZED (STUDENTIZED) RESIDUALS
EACH BIN OF THE HISTOGRAM IS LABELED WITH ITS LOWER LIMIT.
NOTE THAT IF THE COUNT FOR A BIN EXCEEDS 100, ONLY
100 ASTERISKS WILL BE PRINTED.

          -2.2     2 **
          -2.0     1 *
          -1.8    12 ************
          -1.6     9 *********
          -1.4    17 *****************
          -1.2    43 ****************************************
          -1.0    31 *******************************
          -0.8    35 ***********************************
          -0.6    40 ****************************************
          -0.4    54 ******************************************************
          -0.2    55 *******************************************************
           0.0    38 **************************************
           0.2    38 **************************************
           0.4    21 *********************
           0.6    26 **************************
           0.8    22 **********************
           1.0    10 **********
           1.2    22 **********************
           1.4    13 *************
           1.6     8 ********
           1.8     5 *****
           2.0     5 *****
           2.2     6 ******
           2.4     3 ***
           2.6     2 **
           2.8     4 ****
           3.0     0
           3.2     0
           3.4     0
           3.6     0
           3.8     1 *
GREATER THAN 4     1 *
```

Figure 4.p: Normal Prob. Plot for Stand. Residuals



TOTFSH, and TOTFSH must be a non-negative integer, so Y=FEEDRATE must be a non-negative integer multiple of one-half. These parallel line segments in turn suggest a nonhorizontal and positively sloped band which contains them, which also suggests a strong positive correlation between Y and e. In fact, according to Ref. (5), pg.147, this correlation should be:

$$(1-R^2)^{1/2}=0.9364$$

Furthermore, it can be shown that parallel vertical line segments pointing south to north in the e versus Y plot become parallel line segments pointing from southeast to northwest in the e versus $\hat{Y}$ plot (Ref.(8) pg. 216). For example, of the slanted parallel line segments suggested in Figure 4.a, the bottom most one is simply the line segment which corresponds to y=0 from Figure 4.b where it appears as the left most vertical line segment. Note also in Figure 4.a that no points appear below this bottom most slanted line segment. This is unavoidable because the data must satisfy $y \geq 0$ (that is, negative values of FEEDRATE are not possible here).

In addition, a non-horizontal positively sloped band will become an even more strongly sloped band. This time, however, the band has become so wide in Figure 4.a that the correlation between the axis variables almost vanishes. That is, while:

$$r(e,Y)=(1-R^2)^{1/2}=0.9364$$

where r stands for correlation coefficient calculated from sample data alone,

$$r(e,\hat{Y})\approx 0$$

But Ref.(5) states on pg. 148 that this is what should happen if the analysis proceeded correctly, which evidently it has. Exact equality with zero was prevented by accumulated round-off error.

These two residual plots, however, still reveal no new information on the quality of model fit. That the model does not fit the data well has already been indicated by the low R^2-value of 0.1232. The reasons for a poor fit may be important missing variables (including cross-products or interactions), need for transformations which involve the response variable, need for change in 'additivity scale' (this is subtly different from the need for transformations, as will be shown in the next chapter), or any combinations of these. While some of the above actions will be discussed in the next chapter, the next 3 sets of plots should hopefully suggest some corrective action in the meantime.

The next 5 plots (Figures 4.c-4.g) show residuals versus the various explanatory variables already in the current best model. The purpose of these plots is to look for suggestions of any systematic dependency of a non-linear nature between the residuals and the values of the other variable in the plot. The presence of that other variable in the model as is eliminates any further linear relation. If

any such dependency is found, then the appropriate function of that variable (e.g. higher-order power, reciprocal, root, logarithm, or whatever) should be added to the list of candidate explanatory variables and the analysis restarted. As it is, no such further dependency seems to be strongly suggested.

The plot of residual against TIME is important for another reason as well. Such a plot is used to check the assumption of zero correlation between the random error terms, ϵ_i. This plot appears as a band with no apparent upward or downward trend. Neither do any sort of 'cycle' effects seem present. On this basis, the assumption of uncorrelated ϵ_i appears justifiable.

The next 7 plots (Figures 4.h-4.n) show residuals versus the remaining candidate explanatory variables which did not make it into the current best model. Ordinarily these plots would be checked for both linear and non-linear trends, but the P9R program has already checked formally (that is, analytically) for the linear trends, and still a 5-variable model was selected as the current best one. Thus only non-linear trends need be investigated, as for the 5 previous plots. Again, however, no such trend appears to strongly suggest itself.

The remaining 2 plots (Figures 4.o and 4.p) deal with checking the assumption of a normal distribution with a mean

of zero for the random errors, ϵ_i. The residuals, e_i were first standardized

$$\frac{e_i}{\sqrt{\left(\frac{SS(error)}{n-p-1}\right)}}$$

before the plots were done. Figure 4.o shows a histogrm which should look approximately like a shaded-in normal distribution density curve if the normality assumption is correct. Similarly, Figure 4.p shows a normal probability plot of the standardized residuals (as indicated by the '*' characters), which if the normality assumption is correct should lie on the line indicated by the slash ('/') characters.

Having obtained the best possible model so far using no cross-products and modelling the response variable on a linear scale, the next chapter will investigate how to improve the poor fit in the current best model.

CHAPTER 4

SECOND ANALYSIS--IMPROVE FIT OF CURRENT BEST MODEL

There is basically one aspect of the current best model that needs to be improved: R^2 should be increased. This chapter sill investigate 3 methods of accomplishing this. Some methods will be used in combination. Other aspects of the current best model which were satisfactory will be re-checked to make sure they are not sacrificed.

4.1 Method 1. Investigate Interaction/Cross-Product Terms

It should be pointed out that an interaction or cross-product term must use at least 2 explanatory variables which measure different quantities. For example, COLSZE and AGECHK would form a cross-product of

$$(\text{COLSZE})(\text{AGECHK})$$

whereas TIME and SQTIME would form a product of

$$(\text{TIME})(\text{SQTIME}) = (\text{TIME})(\text{TIME})^2 = (\text{TIME})^3$$

which would not be considered a cross-product or interaction term.

Enumerating the possibilities then:

AGECHK
NUMCHK
TIME
TIDEH
TIDEM
COLSZE

DATE
SQAGE
SQTIME
SQCOL
SQDATE

AGECHK may form 9 cross-products with variables below it on the above list, since AGECHK and SQAGE do not form a true cross-product. NUMCHK may form 9 cross-products also with variables below it. The cross-product fo NUMCHK with AGECHK was already accounted for in the AGECHK count, so it must not be counted twice. Similarly TIME may form 7 cross-products with variables below it, TIDEH and TIDEM may form 6 each (a cross-product involving TIDEH and TIDEM would not make sense, especially since this product would always be zero), COLSZE may form 4, DATE and SQAGE may form 3 each, SQTIME may form 2, and SQCOL only 1. The total number of distinct cross-products available is thus 50. Furthermore this is only the possible number of 2-variable cross-products. 3 and higher variable products have not yet been considered (nor will they be).

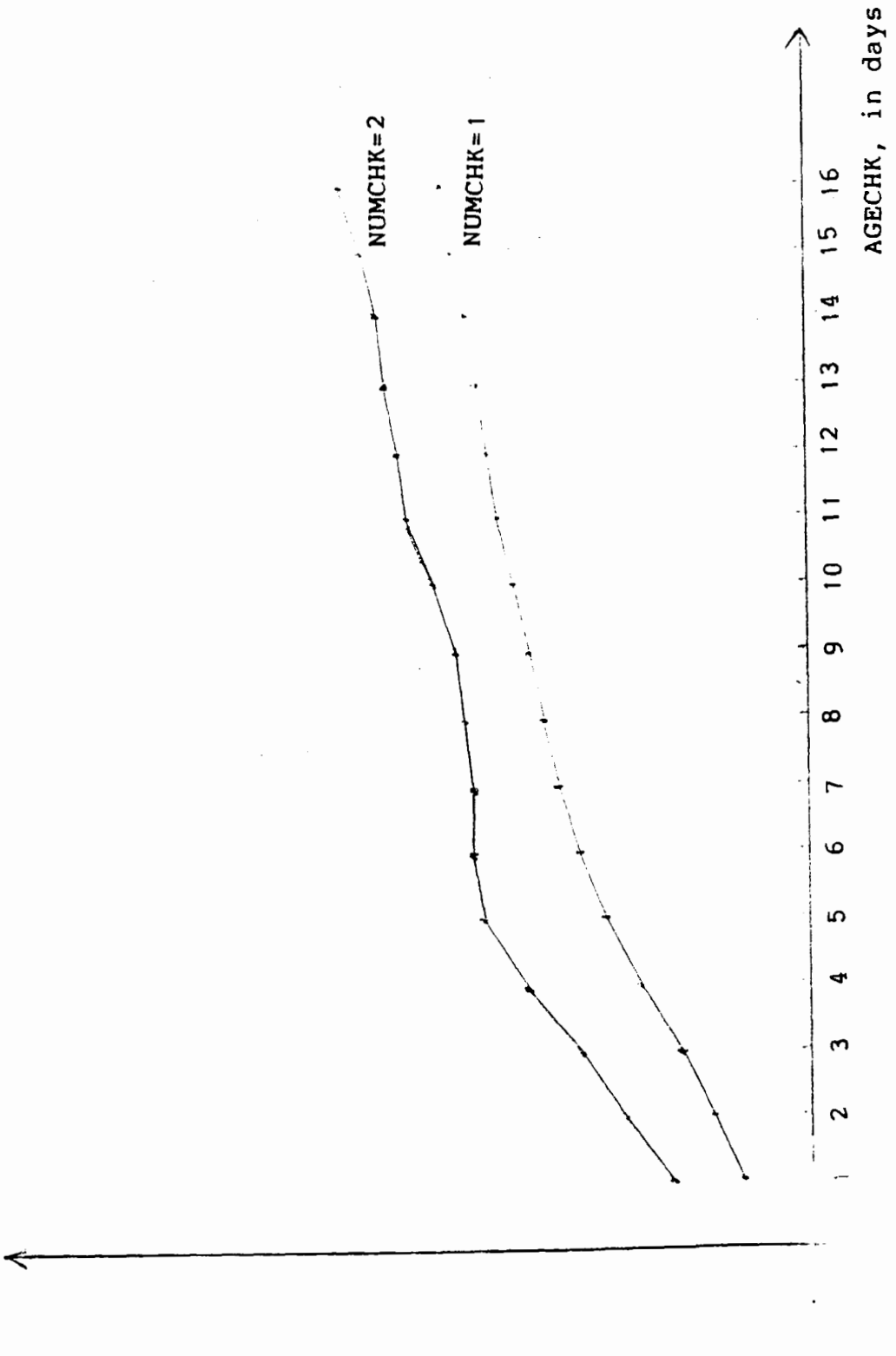
Not all of the 50 possible 2-variable interactions counted are worth considering, however. In fact, only those which make some sort of 'biological sense' will be investigated. Ref.(11) pg. 680 recommends selection of interactions be done by a subject area specialist. One could conceivably possess a data set of such a highly pathological nature that no significant improvement of model fit occurs until a certain 3rd order interaction is included. If the

choices of response variable and additivity scale are appropriate (refer to Methods 2 and 3 further on in this chapter), then this is more likely due to a quirk of the sample data itself (the 'luck of the draw', as it were) and not due to the phenomena being modelled. Such is one of the pitfalls of 'data snooping' (to be described in more detail in the next chapter). Certainly it is permissible however to consider cross-products where one of the variables does not already appear in the model as a main effect, although this is considered to be unusual since interaction effects are 'typically smaller' than the main effects (Ref. (11) pg. 681).

To select which interaction effects are worthy of investigation, the graphical techniques outlined in pages 675-681 of Ref. (11) are helpful. To illustrate, consider AGECHK and NUMCHK. The proposed graph in Figure 5 shows possibly negligible interaction between these two variables in fitting values of *Y*=FEEDRATE. Absolutely no interaction would occur if the 2 curves were perfectly parallel. The decision on whether or not to include an interaction was thus based on this prior expectation of parallelism in curves separated by levels of some factor.

In this way, interactions were anticipated between NUMCHK with AGECHK, TIME with AGECHK (reflecting possible differences in feeding schedules for older chicks), COLSZE with AGECHK, COLSZE with NUMCHK (COLSZE reflecting a sense

Figure 5: Proposed Graph to Visually Detect Interactions



of competition for finite fish supply), and TIME with both TIDEH and TIDEM. Furthermore, if an interaction between 2 variables was to be investigated, then interactions should also be attempted between any higher order terms of either variable as well. For example, not just TIME with AGECHK, but also SQTIME with AGECHK, TIME with SQAGE, and SQTIME with SQAGE should be considered.

A P9R run was done with these interactions attempted along with all of the original candidate explanatory variables, not just the 5 in the current best model. As was pointed out earlier, the multicollinearity present in the data means that although, say, COLSZE was not important without any interactions present in the model, it may become so after some are added. The naming of these interaction terms is shown in the source file for the run (Figure 22) found in the Technical Supplement (Chapter 9). A portion of the output is shown in Figure 6. The maximum attainable value available (R^2=0.1927) is still low however, indicating that the model still does not give a good fit to the data, as far as explaining variation in Y goes.

For a fitted value, $\hat{Y}$, produced by a model, its estimated variance, $s^2(\hat{Y})$, may also provide a useful criteria for choosing one model over another. If a model is to be valuable for estimating future mean outcomes (or predicting individual ones, although this requires a larger but related prediction variance, see pg. 312 of Ref.(6)) ,

Figure 6: P9R run with interactions

SUBSETS WITH 5 VARIABLES

R-SQUARED	ADJUSTED R-SQUARED	CP	VARIABLE	COEFFICIENT	T-STATISTIC
0.142086	0.133805	15.11	19 t1a1	-0.000318200	-2.95
			23 n1a1	0.0367827	7.44
			24 n1a2	-0.000844087	-6.23
			29 c2n1	-0.00248804	-5.29
			30 c2a1	0.000104966	4.92
			INTERCEPT	0.464412	
0.140845	0.132552	15.87	VARIABLE	COEFFICIENT	T-STATISTIC
			20 t1a2	-0.00000934075	-2.81
			23 n1a1	0.0333908	6.15
			24 n1a2	-0.000721877	-4.46
			29 c2n1	-0.00248880	-5.27
			30 c2a1	0.000104571	4.89
			INTERCEPT	0.431570	
0.139212	0.130904	16.87	VARIABLE	COEFFICIENT	T-STATISTIC
			9 numchk	0.525350	7.75
			16 c1n1	-0.0486801	-6.24
			17 c1a1	0.00223674	5.71
			20 t1a2	-0.0000326089	-4.22
			22 t2a2	0.4860595E-06	2.40
			INTERCEPT	0.352840	

SUBSETS WITH 27 VARIABLES

R-SQUARED	ADJUSTED R-SQUARED	CP	VARIABLE	COEFFICIENT	T-STATISTIC
0.192716	0.148771	28.00	3 date	0.00462269	0.31
			4 colsze	-0.105629	-0.74
			6 tideh	0.953796	0.65
			7 tidem	1.51885	1.04
			8 time	0.217280	1.21
			9 numchk	0.640454	1.57
			10 agechk	0.262125	1.63
			11 sqtime	-0.00432563	-1.19
			12 sqdate	-0.000191422	-0.48
			13 sqage	-0.00521267	-1.60
			14 sqcol	0.00869716	1.02
			16 c1n1	-0.0162931	-0.37
			17 c1a1	0.0108142	0.90
			18 c1a2	-0.000216398	-0.89
			19 t1a1	-0.0193626	-1.62
			20 t1a2	0.000406640	1.64
			21 t2a1	0.000363007	1.53
			22 t2a2	-0.00000790996	-1.59
			23 n1a1	-0.00827323	-0.25
			24 n1a2	-0.0000502875	-0.07
			25 tht1	-0.104348	-0.84
			26 tmt1	-0.153118	-1.22
			27 tht2	0.00244927	0.94
			28 tmt2	0.00347945	1.31
			29 c2n1	-0.00202752	-0.79
			30 c2a1	-0.000655491	-0.94
			31 c2a2	0.0000145394	1.06
			INTERCEPT	-2.68740	

then the best model would be the one which gives the narrowest confidence (or prediction) intervals. For a fixed confidence level, $100(1-\alpha)\%$, the width of these intervals increases with increasing $s^2(\hat{Y})$, so the model which gives the smallest $s^2(\hat{Y})$ values is preferable.

Figures 7.a through 7.c show portions of a GLIM run where the following 3 models were investigated:

Figure 7.a: Current Best Model
Figure 7.b: Model with All Explanatory Variables but No Interactions
Figure 7.c: Model with All Explanatory Variables and All Interactions from Figure 6

In each figure is shown fit results and a table of data for the first 50 observations. This table has numbered entries and each line displays the following data for a single observation:

1) Observation Tag (as was described in Chapter 2)
2) Fitted Outcome Value, $\hat{Y}$
3) Estimated Variance, $s^2(\hat{Y})$

From these figures, it seems that the current best model offers fitted outcomes with smaller estimated variances than the model with all explanatory variables. Furthermore, adding interactions to this latter model further increases the $s^2(\hat{Y})$ values. Thus the method of adding interactions was abandoned at this stage.

Before leaving this method altogether, it should be pointed out that 3rd order interactions were not attempted because it was suspected that the situation would not

Figure 7.a: Inspect $s^2(\hat{Y})$ for Current Best Model

CYCLE	DEVIANCE	DF
1	149.3	518

	ESTIMATE	S.E.	PARAMETER
1	0.8733	0.2903	%GM
2	0.5449E-01	0.1390E-01	AGEC
3	0.2431	0.5018E-01	NUMC
4	-0.8448E-01	0.2033E-01	TIME
5	-0.1209E-02	0.2796E-03	SQAG
6	0.1564E-02	0.4052E-03	SQTI

SCALE PARAMETER TAKEN AS 0.2882

1	4002.	0.6879	0.5651E-02
2	4004.	0.5321	0.3116E-02
3	4005.	0.7751	0.3426E-02
4	4008.	0.3358	0.9445E-02
5	4013.	0.6126	0.5783E-02
6	4014.	0.4568	0.3188E-02
7	4015.	0.6999	0.3590E-02
8	4016.	0.2605	0.9446E-02
9	4021.	0.8544	0.5443E-02
10	4022.	0.6113	0.4796E-02
11	4025.	0.7671	0.7366E-02
12	4030.	0.4150	0.1046E-01
13	4031.	0.9630	0.4462E-02
14	4032.	0.7199	0.4214E-02
15	4035.	0.8757	0.6637E-02
16	4037.	0.5624	0.8816E-02
17	4044.	0.3542	0.7841E-02
18	4045.	0.5360	0.2720E-02
19	4048.	0.7790	0.3042E-02
20	4049.	0.6990	0.4778E-02
21	4052.	0.8786	0.5098E-02
22	4053.	0.7986	0.6552E-02
23	4054.	0.4537	0.8949E-02
24	4056.	0.6356	0.4452E-02
25	4060.	0.5406	0.2441E-02
26	4062.	0.3733	0.6512E-02
27	4064.	0.7836	0.2780E-02
28	4066.	0.7109	0.4098E-02
29	4069.	0.6902	0.5248E-02
30	4070.	0.6174	0.6270E-02
31	4071.	0.9333	0.5924E-02
32	4072.	0.5229	0.8665E-02
33	4075.	0.5882	0.1849E-02
34	4076.	0.3695	0.5200E-02
35	4077.	0.5758	0.1842E-02
36	4078.	0.7653	0.2577E-02
37	4079.	0.5029	0.2342E-02
38	4080.	0.8311	0.2407E-02
39	4081.	0.3356	0.6300E-02
40	4082.	0.3695	0.5200E-02
41	4083.	0.6132	0.1810E-02
42	4084.	0.8685	0.2481E-02
43	4085.	0.3730	0.6019E-02
44	4086.	0.4069	0.4951E-02
45	4087.	0.8027	0.2565E-02
46	4088.	0.6256	0.1857E-02
47	4089.	0.5403	0.2225E-02
48	4090.	0.4069	0.4951E-02
49	4091.	0.9850	0.2515E-02
50	4092.	0.7541	0.2170E-02

Figure 7.b: Inspect $s^2(\hat{Y})$ for Model with All Explanatory Variables but No Interactions

CYCLE	DEVIANCE	DF
1	146.3	512

	ESTIMATE	S.E.	PARAMETER
1	1.500	0.3610	%GM
2	0.4518E-01	0.1707E-01	AGEC
3	0.1815	0.5809E-01	NUMC
4	-0.1090	0.2395E-01	TIME
5	-0.9901E-03	0.3328E-03	SQAG
6	0.2102E-02	0.4948E-03	SQTI
7	0.9022E-02	0.1399E-01	DATE
8	-0.2927E-01	0.1689E-01	COLS
0	ZERO	ALIASED	TIDE(1)
9	-0.1148	0.6600E-01	TIDE(2)
10	-0.1514	0.8024E-01	TIDE(3)
11	-0.3696E-03	0.3774E-03	SQDA
12	0.1039E-02	0.9190E-03	SQCO

SCALE PARAMETER TAKEN AS 0.2857

1	4002.	0.5557	0.8418E-02
2	4004.	0.4474	0.8395E-02
3	4005.	0.6288	0.8165E-02
4	4008.	0.2832	0.1041E-01
5	4013.	0.5967	0.1026E-01
6	4014.	0.4884	0.1027E-01
7	4015.	0.6699	0.9846E-02
8	4016.	0.3243	0.1265E-01
9	4021.	0.7762	0.8730E-02
10	4022.	0.5948	0.8854E-02
11	4025.	0.7031	0.9078E-02
12	4030.	0.4306	0.1107E-01
13	4031.	0.8253	0.9990E-02
14	4032.	0.6438	0.1009E-01
15	4035.	0.7521	0.1071E-01
16	4037.	0.5120	0.1181E-01
17	4044.	0.3007	0.8690E-02
18	4045.	0.4529	0.7278E-02
19	4048.	0.6344	0.7124E-02
20	4049.	0.5672	0.7205E-02
21	4052.	0.8046	0.7702E-02
22	4053.	0.7374	0.7885E-02
23	4054.	0.4709	0.9385E-02
24	4056.	0.6231	0.7755E-02
25	4060.	0.4695	0.6480E-02
26	4062.	0.3291	0.7340E-02
27	4064.	0.6509	0.6163E-02
28	4066.	0.5897	0.6088E-02
29	4069.	0.7069	0.7904E-02
30	4070.	0.6456	0.7506E-02
31	4071.	0.8884	0.7701E-02
32	4072.	0.5666	0.9076E-02
33	4075.	0.5114	0.8775E-02
34	4076.	0.3238	0.5981E-02
35	4077.	0.4984	0.6209E-02
36	4078.	0.6338	0.5285E-02
37	4079.	0.4359	0.4362E-02
38	4080.	0.6910	0.7195E-02
39	4081.	0.2955	0.7079E-02
40	4082.	0.3238	0.5981E-02

Figure 7.b, continued

41	4083.	0.5669	0.6257E-02
42	4084.	0.7595	0.7263E-02
43	4085.	0.3640	0.7371E-02
44	4086.	0.3924	0.6245E-02
45	4087.	0.7023	0.5455E-02
46	4088.	0.5799	0.8759E-02
47	4089.	0.5044	0.4504E-02
48	4090.	0.3924	0.6245E-02
49	4091.	0.9493	0.3353E-02
50	4092.	0.7783	0.4655E-02

Figure 7.c: Inspect $s^2(\hat{Y})$ for Model with All Explanatory Variables and All Interactions from Figure 6

CYCLE	DEVIANCE	DF	
1	137.5	496	
	ESTIMATE	S.E.	PARAMETER
1	-2.687	2.388	%GM
2	0.2621	0.1611	AGEC
3	0.6405	0.4074	NUMC
4	0.2173	0.1802	TIME
5	-0.5213E-02	0.3258E-02	SQAG
6	-0.4326E-02	0.3627E-02	SQTI
7	0.4623E-02	0.1507E-01	DATE
8	-0.1056	0.1436	COLS
0	ZERO	ALIASED	TIDE(1)
9	1.519	1.462	TIDE(2)
10	0.9538	1.465	TIDE(3)
11	-0.1914E-03	0.3997E-03	SQDA
12	0.8697E-02	0.8511E-02	SQCO
13	-0.8273E-02	0.3302E-01	A1N1
14	-0.5029E-04	0.6928E-03	A2N1
15	-0.1936E-01	0.1197E-01	A1T1
16	0.3630E-03	0.2373E-03	A1T2
17	0.4066E-03	0.2480E-03	A2T1
18	-0.7910E-05	0.4986E-05	A2T2
19	-0.1629E-01	0.4401E-01	C1N1
20	-0.2028E-02	0.2581E-02	C2N1
21	0.1081E-01	0.1197E-01	C1A1
22	-0.2164E-03	0.2437E-03	C1A2
23	-0.6555E-03	0.6941E-03	C2A1
24	0.1454E-04	0.1369E-04	C2A2
0	ZERO	ALIASED	TIME.TIDE(1)
25	-0.1531	0.1252	TIME.TIDE(2)
26	-0.1043	0.1245	TIME.TIDE(3)
0	ZERO	ALIASED	SQTI.TIDE(1)
27	0.3479E-02	0.2649E-02	SQTI.TIDE(2)
28	0.2449E-02	0.2601E-02	SQTI.TIDE(3)

SCALE PARAMETER TAKEN AS 0.2771

1	4002.	0.4733	0.2029E-01
2	4004.	0.3770	0.1278E-01
3	4005.	0.5452	0.1320E-01
4	4008.	0.1590	0.2516E-01
5	4013.	0.6117	0.2720E-01
6	4014.	0.4836	0.1586E-01
7	4015.	0.6519	0.1767E-01
8	4016.	0.3339	0.3099E-01
9	4021.	0.8080	0.1533E-01
10	4022.	0.6397	0.1421E-01
11	4025.	0.7156	0.2514E-01
12	4030.	0.3675	0.4031E-01
13	4031.	0.6392	0.1978E-01
14	4032.	0.4710	0.2012E-01
15	4035.	0.4930	0.3092E-01
16	4037.	0.1623	0.4088E-01
17	4044.	0.2066	0.2041E-01
18	4045.	0.3955	0.1076E-01
19	4048.	0.5542	0.1119E-01
20	4049.	0.4963	0.1618E-01
21	4052.	0.8356	0.1327E-01
22	4053.	0.7541	0.1977E-01

Figure 7.c, continued

23	4054.	0.4257	0.3186E-01
24	4056.	0.6768	0.1218E-01
25	4060.	0.4188	0.9568E-02
26	4062.	0.2528	0.1774E-01
27	4064.	0.6228	0.9814E-02
28	4066.	0.5766	0.1282E-01
29	4069.	0.7622	0.1319E-01
30	4070.	0.6529	0.1666E-01
31	4071.	0.9661	0.1387E-01
32	4072.	0.5076	0.3132E-01
33	4075.	0.5459	0.1191E-01
34	4076.	0.3587	0.1440E-01
35	4077.	0.5126	0.8583E-02
36	4078.	0.4825	0.1081E-01
37	4079.	0.4502	0.6649E-02
38	4080.	0.4865	0.1297E-01
39	4081.	0.3369	0.1826E-01
40	4082.	0.3587	0.1440E-01
41	4083.	0.6518	0.1333E-01
42	4084.	0.6233	0.1748E-01
43	4085.	0.4504	0.2263E-01
44	4086.	0.4773	0.1885E-01
45	4087.	0.6200	0.1581E-01
46	4088.	0.6750	0.1667E-01
47	4089.	0.5859	0.1131E-01
48	4090.	0.4773	0.1885E-01
49	4091.	1.051	0.6572E-02
50	4092.	0.7199	0.7622E-02

improve much. If it did, then this may be an indication that Methods 2 or 3 are a better approach. As for 4th and higher order interactions, they are usually found in practice to be small or non-existent (Ref. (11) pg. 809).

It is interesting to note, however, that the best 5 variable model in this run has only one main effect and 4 interactions. Again, this may suggest that a linear additive model on the original FEEDRATE scale may not be the best choice. Methods 2 and 3 will investigate some alternatives.

4.2 Method 2. Transform Response Variable

The multiple regression model presented at the start of Chapter 3 now has its response variable transformed before the model is fitted:

$$g(Y)=\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p+\epsilon$$

where g is some function chosen to show a scale where the linear additivity of explanatory varibles and random error are more suitable. The subscript i which tags individual observations is not shown for simplicity, but is implied.

This method is more commonly applied to data to make it look more like a random sample from a normal distribution. The problem with a regression model is that it is the error terms, ϵ_i, which should benefit from the transformation, since these are assumed to be normally distributed with a common mean 0 and common variance σ^2. The $g(Y_i)$, however are

then normally distributed with different means. since the x_{ij} vary from one observation to the next. It would thus be pointless to attempt to transform the Y_i so that they look like a sample from a single normal population (Ref. (15)), which can only have one mean.

Method 3 offers an even better alternative for the regression model. Nonetheless a logarithmic transformation was attempted since Ref. (5) pg. 221 suggests that it might have been the better way to proceed after the plots of Figure 1. The case of $Y_i=0$, however, causes a problem since the transformation

$$g(Y)=ln(Y)$$

is not defined there. Ref. (14) pg. 77 advises to replace such Y-values with a value less than one-half the available accuracy. The Y_i are derived from counts, however, so instead the advice of Ref. (6) pg. 161 was followed whereby a constant (chosen albeit arbitrarily) of 0.1 was added beforehand:

$$g(Y)=ln(Y+0.1)$$

A P9R run was done on these transformed values, the results of which are highlighted in Figure 8. No interactions were attempted.

As it turned out the best 5 variable model of Figure 8 uses the same 5 explanatory variables as in the current best model, but now R^2 has fallen to 0.1149. Furthermore, when all 11 explanatory variables are used in a model, the normal

Figure 8: P9R run on log-transformed observations

SUBSETS WITH 5 VARIABLES

R-SQUARED	ADJUSTED R-SQUARED	CP	VARIABLE	COEFFICIENT	T-STATISTIC
0.110800	0.102217	11.17	time	-0.130817	-3.59
			numchk	0.381655	4.24
			agechk	0.0906101	3.63
			sqtime	0.00237431	3.27
			sqage	-0.00207890	-4.14
			INTERCEPT	-0.324453	
0.096352	0.087629	19.67	VARIABLE	COEFFICIENT	T-STATISTIC
			tidemed	-0.136514	-1.49
			time	-0.0157781	-2.73
			numchk	0.376586	4.15
			agechk	0.0834191	3.31
			sqage	-0.00195473	-3.86
			INTERCEPT	-1.41973	
0.095153	0.086419	20.37	VARIABLE	COEFFICIENT	T-STATISTIC
			colsze	-0.0109627	-1.23
			time	-0.0132162	-2.41
			numchk	0.338401	3.51
			agechk	0.0877158	3.49
			sqage	-0.00204993	-4.04
			INTERCEPT	-1.44368	
0.094291	0.085549	20.88	VARIABLE	COEFFICIENT	T-STATISTIC
			time	-0.0132582	-2.41
			numchk	0.350399	3.69
			agechk	0.0886511	3.51
			sqage	-0.00206102	-4.04
			sqcol	-0.000436799	-1.01
			INTERCEPT	-1.52107	

SUBSETS WITH 11 VARIABLES

R-SQUARED	ADJUSTED R-SQUARED	CP	VARIABLE	COEFFICIENT	T-STATISTIC
0.129779	0.111082	12.00	VARIABLE	COEFFICIENT	T-STATISTIC
			date	0.00591546	0.24
			colsze	-0.0369024	-1.22
			tidehi	-0.387468	-2.69
			tidemed	-0.282428	-2.39
			time	-0.190823	-4.44
			numchk	0.304788	2.93
			agechk	0.0808545	2.64
			sqtime	0.00367653	4.14
			sqdate	-0.000199972	-0.30
			sqage	-0.00188969	-3.17
			sqcol	0.00130484	0.89
			INTERCEPT	0.906156	

probability plot of standardized residuals (not included here) showed a stronger deviation from normality than was the case back in Figure 2.

In Figure 9, P9R output is shown for another attempted transformation:

$$g(Y)=\sqrt{Y}$$

This transformation was selected on the basis of a remark on pg. 161 of Ref. (6) that square root transformations tend to make count-type data more 'normal-looking'. The situation with R^2 and the normal probability plot (not shown) had improved over the previous transformation, but overall the situation still seems to be better with untransformed Y_i.

Another problem caused by both of the above transformations could be seen in the histograms of standardized residuals (not shown here). Bimodality is very strongly suggested, and this situation should be avoided because of its unknown implications.

Although Method 2 has proved unfruitful, it will make and interesting comparison with the next attempt at R^2 improvement.

4.3 Method 3. Change Additivity Scale

For the original multiple regression model proposed:

$$Y=\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p+\epsilon$$

Figure 9: P9R run on root-transfromed observations

SUBSETS WITH 5 VARIABLES

R-SQUARED	ADJUSTED R-SQUARED	CP	VARIABLE	COEFFICIENT	T-STATISTIC
0.117785	0.109270	11.15	time	-0.0601809	-3.79
			numchk	0.175063	4.47
			agechk	0.0409328	3.78
			sqtime	0.00109775	3.47
			sqage	-0.000932062	-4.27
			INTERCEPT	0.837470	
0.101192	0.092516	20.99	VARIABLE	COEFFICIENT	T-STATISTIC
			tidemed	-0.0602573	-1.51
			time	-0.00693580	-2.76
			numchk	0.172764	4.37
			agechk	0.0376648	3.44
			sqage	-0.000875506	-3.98
			INTERCEPT	0.327617	
0.100209	0.091524	21.58	VARIABLE	COEFFICIENT	T-STATISTIC
			colsze	-0.00505068	-1.31
			time	-0.00580858	-2.43
			numchk	0.155130	3.69
			agechk	0.0395921	3.62
			sqage	-0.000918582	-4.16
			INTERCEPT	0.319768	
0.099174	0.090479	22.19	VARIABLE	COEFFICIENT	T-STATISTIC
			time	-0.00582593	-2.44
			numchk	0.160905	3.89
			agechk	0.0400009	3.64
			sqage	-0.000923118	-4.16
			sqcol	-0.000197421	-1.05
			INTERCEPT	0.283561	

SUBSETS WITH 11 VARIABLES

R-SQUARED	ADJUSTED R-SQUARED	CP	VARIABLE	COEFFICIENT	T-STATISTIC
0.136593	0.118043	12.00	date	0.00358991	0.33
			colsze	-0.0179204	-1.36
			tidehi	-0.159851	-2.56
			tidemed	-0.117512	-2.28
			time	-0.0851073	-4.56
			numchk	0.137315	3.03
			agechk	0.0359700	2.70
			sqtime	0.00163977	4.25
			sqdate	-0.000134536	-0.46
			sqage	-0.000828373	-3.19
			sqcol	0.000631767	0.99
			INTERCEPT	1.37207	

it will be noticed that the only random component is ϵ. The sum:

$$\mu=\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p$$

may be regarded as the systematic (nonrandom) component of the current model. Thus:

$$Y_i=\mu_i+\epsilon_i$$

for observation i. The additivity of the explanatory variable effects is on the same scale as μ and hence Y as well. Suppose instead the situation were altered so that additivity of explanatory variable effects no longer took place on the same scale as μ, but a function of it:

$$Y=\mu+\epsilon$$

but now

$$\eta=g(\mu)=\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p$$

For example, if

$$g(\mu)=ln(\mu)$$

then

$$\mu=e^{\eta}=exp\{\eta\}$$
$$=exp\{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p\}$$

Thus:

$$Y=exp\{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p\}+\epsilon$$

Back in Method 2, one of the attempted transformations was

$$ln(Y+0.1)=\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p+\epsilon$$

or equivalently

$$Y=exp\{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p+\epsilon\}-0.1$$

Ignoring the 0.1 subtraction, the difference between these two methods is that in Method 2, the 'linear predictor'

$$\eta=\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p$$

and the random error, ϵ, must be kept together as a sum before any transformations take place, whereas this is no longer necessary in Method 3. Also in Method 3, it is still possible to decompose the original observations into a sum of systematic and random components:

$$Y_i=\mu_i+\epsilon_i$$

which is more intuitively appealing. Method 3 achieves the flexibility of Method 2 by using a transformation on the systematic component alone, rather than on the observations, to obtain a linear predictor in the explanatory variables:

$$g(\mu)=\eta=\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p$$

This framework is known as a Generalized Linear Model, and the special case of

$$g(\mu)=ln(\mu)$$

known as the log-linear model will be pursued here.

Recall that FEEDRATE was calculated from total fish delivery counts, TOTFSH. According to pg. 127 of Ref. (8), the log-linear model is often suitable for count data. Thus TOTFSH will be used as a response variable rather than FEEDRATE, so now Y=TOTFSH. For the log-linear model the ideal distribution for the ϵ_i is no longer the normal one, but the Poisson, or quasi-Poisson in the case of over-dispersed data (that is, data with a wider spread than a Poisson distribution can accomodate).

The technique for fitting the log-linear model to the sample data in order to produce estimates of the β_j uses a different approach than the normal-theory multiple regression of Chapter 3. As a result there is no longer any R^2 or even $SS(total)$ or $SS(error)$ terms to work with. Instead one works with a more general goodness of fit measure known as the deviance, which will replace $SS(error)$. Also one considers a kind of generalized R^2:

$$R_g^2 = 1 - \frac{D(current\ model)}{D(null\ model)}$$

where $D(null\ model)$ is the deviance for a model which contains no explanatory variables (hence the term 'null'), but only a constant term, β_0, playing the role of an overall grand average or mean. This quantity will replace $SS(total)$. $D(current\ model)$ refers to the deviance of the current model being considered. It behaves like $SS(error)$ in the sense that if any other variable is added to the model, then it cannot increase. Thus R_g^2 cannot decrease. It turns out that for normal theory multiple regressions (such as in Chapter 3), R^2 and R_g^2 are the same. This is because under such conditions:

$$D(null\ model) = SS(total)$$

$$D(current\ model) = SS(error)$$

The Technical Supplement provides more details (see Chapter 9).

GLIM is the computer software package that does the work on Generalized Linear Models. Figure 10 displays a portions of the GLIM run used on the TOTFSH outcomes. It should be noted that GLIM generates its own design variables, so that TIDE was used in the input, rather than TIDEH and TIDEM. In fact one notices the similarity:

TIDE(2)=TIDEM

TIDE(3)=TIDEH

TIDE(1) is shown on the GLIM output as always having the pre-set value of zero, since although TIDE has 3 levels, only 2 design variables are needed.

One also notes from the output:

$$D(null\ model)=580.2$$

and for the fullest possible model without interactions:

$$D(current\ model)=509.9$$

so that:

$$R_g^2=1-\frac{509.9}{580.2}$$

$$=0.1212$$

But for the corresponding full model from Figure 2:

$$R^2=R_g^2=0.1409$$

so it seems as though this model will be a poor fit too.

Also as part of the GLIM run, though not shown here, the interactions from Method 1 were attempted as improvements to the fit of the full model. When all of them were inserted, the result was:

Figure 10: Log-Linear Modelling with GLIM

```
CYCLE  DEVIANCE      DF
   4     580.2       523

       ESTIMATE       S.E.        PARAMETER
   1    0.2995      0.3961E-01    %GM
  SCALE PARAMETER TAKEN AS         1.109

CYCLE  DEVIANCE      DF
   4     509.9       512

       ESTIMATE       S.E.        PARAMETER
   1     1.190       0.5719       %GM
   2    0.2542       0.9109E-01   NUMC
   3 -0.7175E-03     0.6227E-03   SQDA
   4  0.2046E-01     0.2216E-01   DATE
   5   -0.1462       0.3629E-01   TIME
   6  0.2834E-02     0.7544E-03   SQTI
   0      ZERO      ALIASED       TIDE(1)
   7   -0.1553       0.1051       TIDE(2)
   8   -0.1965       0.1267       TIDE(3)
   9 -0.3848E-01     0.2619E-01   COLS
  10 -0.1642E-02     0.5909E-03   SQAG
  11  0.7240E-01     0.2924E-01   AGEC
  12  0.1341E-02     0.1288E-02   SQCO
  SCALE PARAMETER TAKEN AS        0.9959
```

$$R_g^2 = 0.1649$$

which is no better than the analogous situation in the normal theory attempts in Method 1.

A comparison of residual sum of squares, that is, *SS(error)* from Chapter 3, is also worthwhile. In Figure 11.a is shown a portion of a GLIM run on the current best model with a separate calculation of *SS(error)* given below the fit results. This quantity was calculted as described in Chapter 3, and to 4 significant figures comes out to be 149.3, which is equal to the deviance of that model. But as has already been pointed out, this is what should happen for normal theory models. In Figure 11.b is shown another protion of the GLIM run where the same 5 explanatory variables are used, but now in a log-linear model. As is shown below the fit results, a *SS(error)* value of 596.8 is computed. In both cases the *SS(error)* is a measure of the spread in the discrepancy between fitted and observed outcome values (FEEDRATE for the normal theory model, TOTFSH for the log-linear model). Since a smaller *SS(error)* value is to be preferred, the log-linear model was therefore not pursued any further.

Overall it looks as though none of the methods of this chapter have produced a result which would make modification of the current best model worthwhile.

Figure 11.a: *SS(error)* for Normal Theory Model

```
CYCLE  DEVIANCE      DF
   1     170.3       523

       ESTIMATE      S.E.       PARAMETER
   1   0.6746      0.2493E-01   %GM
  SCALE PARAMETER TAKEN AS     0.3256

CYCLE  DEVIANCE      DF
   1     149.3       518

       ESTIMATE      S.E.       PARAMETER
   1   0.8733      0.2903       %GM
   2   0.2431      0.5018E-01   NUMC
   3   0.5449E-01  0.1390E-01   AGEC
   4  -0.8448E-01  0.2033E-01   TIME
   5  -0.1209E-02  0.2796E-03   SQAG
   6   0.1564E-02  0.4052E-03   SQTI
  SCALE PARAMETER TAKEN AS     0.2882

  1     149.292160
```

Figure 11.b: *SS(error)* for Log-Linear Model

```
CYCLE   DEVIANCE        DF
   4      580.2         523

       ESTIMATE         S.E.        PARAMETER
   1   0.2995        0.3961E-01     %GM
  SCALE PARAMETER TAKEN AS          1.109

CYCLE   DEVIANCE        DF
   4      517.9         518

       ESTIMATE         S.E.        PARAMETER
   1   0.3052        0.4573         %GM
   2   0.3385        0.7692E-01     NUMC
   3   0.9392E-01    0.2492E-01     AGEC
   4  -0.1132        0.3061E-01     TIME
   5  -0.2105E-02    0.5134E-03     SQAG
   6   0.2119E-02    0.6184E-03     SQTI
  SCALE PARAMETER TAKEN AS        0.9999

  1     596.764160
```

CHAPTER 5

CONCLUDING REMARKS AND OBSERVATIONS

The current best model of Chapter 3 is the final recommended model, suject to the following observations and remarks.

5.1 Remarks on Current Best Model

As has already been pointed out, this model suffers from lack of fit, as reflected by the low R^2-value of 0.1232. This lack of fit may be caused by omission of an unmeasured variable from the study or perhaps a still as yet undiscovered 'miracle' interaction/higher-order power of explanatory variables already in the study. The cause may even be in the model itself. The methods attempted in Chapter 4 did not seem to improve the situation sufficiently to justify their use over the current best model.

This insistent poor fit problem basically seems to be that FEEDRATE and TOTFSH values have much variation in themselves that seems to have little to do with any of the explanatory variables. That is, $r(Y,x_j)$, the sample correlation between FEEDRATE (Y) and a given candidate explanatory variable, is somewhat low for all of the x_j tried so far, which have been basically 1st order, 2nd order, and cross-product functions of the available

variables. This could be seen in the last row of the correlation matrices obtained in the various P9R and GLIM runs, such as the one displayed in Figure 12, which came from the first P9R run, whose results were previously highlighted in Figure 2. The methods of Chapter 4 did little to increase these low correlations.

Furthermore the residual plots for the current best model do not appear to suggest any strategy for improving the fit. It does appear, however, that negative residuals (overshooting of the Y_i by the $\hat{Y}_i$) tend to occur with the lower Y_i-values and positive residuals with the higher Y_i-values. Although this does not contradict the requirement that:

$$r(\hat{Y}, e) = 0$$

as was pointed out in Chapter 3, nonetheless Ref.(5) pg. 157 suggests that perhaps the major point of change from negative to positive residuals is caused by the change in factor levels of some as yet unconsidered (i.e. unmeasured or unobserved) qualitative variable, which in turn is also correlated with FEEDRATE or fish supply. Perhaps this missing variable is a weather or climate factor (wet/dry, clear/overcast, and so on).

Figure 12: Correlation Matrix from 1st P9R Run

CORRELATIONS

		date	colsze	tidehi	tidemed	time	numchk	agechk	sqtime	sqdate	sqage	sqcol	feedrate
		3	4	6	7	8	9	10	11	12	13	15	16
date	3	1.000											
colsze	4	-0.177	1.000										
tidehi	6	0.130	0.044	1.000									
tidemed	7	-0.125	0.047	-0.658	1.000								
time	8	0.063	-0.012	0.415	-0.316	1.000							
numchk	9	-0.318	-0.308	-0.089	0.015	-0.060	1.000						
agechk	10	0.791	-0.118	0.093	-0.095	0.006	-0.154	1.000					
sqtime	11	0.052	-0.004	0.480	-0.334	0.989	-0.061	0.003	1.000				
sqdate	12	0.968	-0.198	0.135	-0.110	0.096	-0.315	0.754	0.087	1.000			
sqage	13	0.778	-0.127	0.101	-0.088	0.020	-0.164	0.982	0.017	0.768	1.000		
sqcol	15	-0.161	0.959	0.030	0.041	-0.031	-0.265	-0.088	-0.027	-0.196	-0.106	1.000	
feedrate	16	-0.151	-0.096	-0.031	-0.022	-0.118	0.230	-0.107	-0.095	-0.181	-0.142	-0.067	1.000

5.2 Remarks on Best Model Search

Suppose now that this current best model is being considered for acceptance. For any model obtained by some empirical search procedure, mention should be made of the 'data snooping' phenomenon, which arises naturally in:

Stepwise Regression (Ref.(5) pg. 311-2, Ref.(12) pg. 389)
Discriminant Analysis (Ref.(6) pgs. 489, 518-9)
Abuse of Factor Effect Estimation in ANOVA (Ref.(11) pg. 574)

and hence in any regression model search procedure (Ref.(11) pg. 437). Basically what happens in data snooping is that one studies effects suggested by the data instead of first deciding on what specific effects are to be tested/studied before inspecting the sample data for these effects alone.

In the specific context of the current best model, the former procedure (data snooping) was followed. The latter procedure would correspond to deciding beforehand what model to try out before analyzing the sample data in order both to estimate the necessary coefficients and other unknown parameters, and to test whether or not the data do indeed support the *a priori* proposed model.

Both procedures do comprise valid statistical practice. But if one wished to perform the 'standard' F-test for regression:

$$H_0: \beta_1=\beta_2=\cdots=\beta_p=0 \quad (\text{i.e. } Y=\beta_0+\epsilon)$$

against

H_A: At least one of $\{\beta_1, \beta_2, \ldots, \beta_p\}$ is not zero.

this test would only be valid if the latter procedure were followed, which was not the case.

As will be seen later in Figure 13, the P9R run on the current best model provided an F-statistic of:

$$F=14.56$$

with

$$n-p-1=518$$

degrees of freedom for the denominator, and

$$p=5$$

degrees of freedom for the numerator. The associated significance level is:

$$\alpha=0.0000$$

to the available 4 decimal places. But when the former procedure (data snooping) is followed, the true distribution of the F-statistic under H_0 and its associated significance levels are difficult to obtain (Ref.(5) pg. 311-2). This is one reason why F-statistics have been thus far avoided in the analysis. Another reason will be given in the Technical Supplement (Section 8.4).

In a more general context of 'best subsets search' procedures for model building, an analyst will tend to prefer a model having a best fit according to some criterion, such as maxmizing R^2 (and perhaps allowing for other, sometimes opposing, considerations such as minimizing the number of explanatory variables in the final model). It

is thus possible for a model to fit a given set of sample data 'too well'. A fit may be best because it truly is, or simply because of sampling variability. It all depends on to what extent a given sample is truly representative of the population to which the model is to be applied. If the analyst selects a model by a data snooping procedure and further uses it to make predictions on outcomes for a new set of values for the explanatory variables already in the model, then a 'prediction bias' (Ref.(11) pg. 437) is further committed.

For any model obtained by data snooping, an advisable formal testing procedure might be to follow the example of discriminant analysis (Ref.(5) pg. 518-9) where a first sample (a 'training' sample) is taken and the final model is then subjected to formal hypothesis testing (such as with *F*-statistics) but using data from an independent second sample (a 'validation' sample) for all statistical calculations. This procedure is also recommended on pg. 437 of Ref.(11).

Incidentally, Simon's data set came naturally into 2 mutually exclusive data files (Tables 1 and 2 showed portions). It was thought at first to use one file as the training sample and the other as the validation sample. Unfortunately it turned out that the 1984 data file contained primarily colonies of size 10 nests or greater, whereas the 1985 data file contained mostly colonies of size

5 nests or less. This idea of 2 samples was thus abandoned in case COLSZE became an important explanatory variable.

5.3 Further Observations on the Current Best Model

Suppose that given the previous remarks, this current best model is accepted as final for the data at hand. Some further observations of interest can be made on it alone.

5.3.1 Significance of Regression Coefficients

Figure 13 shows some further details from the P9R run on the current best model. This is the P9R run that produced the plots shown in Figures 4.a-4.p. It can be seen that all of the regression coefficients, b_j, are highly statistically significant. This is due to the low sample standard errors, $s(b_j)$, which in turn lead to high t-statistics:

$$t_j = \frac{b_j}{s(b_j)}$$

with their associated significance levels (Ref.(11) pg. 243):

$$\alpha_j = Pr\{|T| > t_j \,|\, T \sim t(518)\}$$

that is, α_j is the probability that $T<-t_j$ or $T>t_j$ given that the random variable T has a t-distribution with $n-p-1=518$ degrees of freedom.

Consider for example the coefficient for NUMCHK:

$$b_1 = 0.243$$

Figure 13: Further Results of P9R Run on Current Best Model

STATISTICS FOR 'BEST' SUBSET

SQUARED MULTIPLE CORRELATION	0.12320
MULTIPLE CORRELATION	0.35099
ADJUSTED SQUARED MULT. CORR.	0.11473
RESIDUAL MEAN SQUARE	0.288215
STANDARD ERROR OF EST.	0.536857
F-STATISTIC	14.56
NUMERATOR DEGREES OF FREEDOM	5
DENOMINATOR DEGREES OF FREEDOM	518
SIGNIFICANCE (TAIL PROB.)	0.0000

VARIABLE NO.	NAME	REGRESSION COEFFICIENT	STANDARD ERROR	STAND. COEF.	T-STAT.	2TAIL SIG.	TOLERANCE	CONTRIBUTION TO R-SQ
	INTERCEPT	0.873322	0.290271	1.531	3.01	0.003		
9	numchk	0.243053	0.0501761	0.202	4.84	0.000	0.968776	0.03972
10	agechk	0.0544920	0.0138984	0.847	3.92	0.000	0.036309	0.02602
13	sqage	-0.00120941	0.000279562	-0.935	-4.33	0.000	0.036203	0.03168
8	time	-0.0844785	0.0203297	-1.152	-4.16	0.000	0.022026	0.02923
11	sqtime	0.00156383	0.000405224	1.070	3.86	0.000	0.022002	0.02521

THE CONTRIBUTION TO R-SQUARED FOR EACH VARIABLE IS THE AMOUNT BY WHICH R-SQUARED WOULD BE REDUCED IF THAT VARIABLE WERE REMOVED FROM THE REGRESSION EQUATION.

COVARIANCES OF THE ESTIMATES OF THE REGRESSION COEFFICIENTS

		numchk 9	agechk 10	sqage 13	time 8	sqtime 11
numchk	9	0.251764E-02				
agechk	10	-0.226674E-04	0.193165E-03			
sqage	13	0.888625E-06	-0.381248E-05	0.781552E-07		
time	8	-0.168059E-05	-0.126549E-04	0.231111E-06	0.413297E-03	
sqtime	11	0.203122E-06	0.310001E-06	-0.577500E-08	-0.814613E-05	0.164207E-06

It represents the rate of change (in this case, increase, since $b_1>0$) in predicted/fitted mean FEEDRATE per chick:

$$\frac{\partial \hat{Y}}{\partial(\text{NUMCHK})}=0.243$$

Thus in comparing 2 nests of different numbers of chicks, but where time of day and age of chicks is same for both nests, then according to the current best model the nest which has more chicks could expect on average to receive 0.243 more fish per hour for every chick that it has in excess of the other nest. Of course this is an estimate and as such is subject to random error as reflected in its standard error, $s(b_1)$, which is the square root of the estimated sample variance, $s^2(b_1)$. For this sample:

$$s(b_1)=0.0501761$$

One could still say, however, with 95% probability of being correct, that the true (but unknown) rate of mean FEEDRATE increase per chick, β_1, lies in the interval:

$$(b_1-t(518;0.975)s(b_1),b_1+t(518;0.975)s(b_1))$$

$$=(0.145,0.341)$$

where $t(\nu;\gamma)$ represents the 100γ% point of a t-distribution with ν degrees of freedom. So b_1 is statistically significantly nonzero (at a 95% confidence level), since the interval does not contain zero. Whether or not any β_1-value in this 95% confidence interval has a 'practical' significance over and above its 'statistical' significance is best left up to the subject matter expert to decide.

5.3.2 Quadratic Effects in TIME and AGECHK

If one takes the current best model and applies a 'completion of the square' procedure, one can re-express it as:

$$\hat{Y}=0.346+0.243(\text{NUMCHK})-0.00121(\text{AGECHK}-22.5)^2$$
$$+0.00156(\text{TIME}-27.0)^2$$

Thus for a group of nests at a given common time of day and all contining the same number of chicks, the average rate of number of fish delivered per hour, as a function of chick age, seems to follow a concave down parabola with vertex (and maximum value) located at 22.5 days. At least, this is the case over the observed range of chick ages. Thus it seems that FEEDRATE increases as the chick(s) gets older, reaches a maximum at an age of 22.5 days, and then decreases as the chick ages further.

A similar observation can be made in regards to time of day. For a group of nests all containing the same number of chicks, and all chicks having a common age, the average rate of number of fish delivered per hour as a function of time of day seems to follow a concave up parabola having a vertex (and minimum value) at half-hour 27.0, or 1330H (1:30 P.M.), at least over the range of observed times. Thus it seems that FEEDRATE decreases from some initial value as the day wears on until 1:30 P.M. after which it begins to increase again.

The fact that TIDE, COLSZE, and DATE are not in the model states that the given sample data suggest that the above observations do not seem to change with the state of the tide, vary with the number of nests in the colony, or even vary from one day to the next, at least over the available range of these variables.

Two points should be emphasized here. First, this apparent relation should not be imposed outside the limits of observed chick age (which according to the 'data summary statistics' of the P9R runs, ranged from 6 to 44 days) or time of day (0530H to 1900H), or for any other variable for that matter. Such a procedure is called extrapolation and is best avoided. Second, the parabolas indicated in this current best model are symmetric about a vertical axis when FEEDRATE is plotted against AGECHK or TIME, which implies equal rates of increase and decrease on either side of this axis. If this is considered undesirable for theoretical (biological) reasons, a cubic term could be incorporated for more flexibility, or perhaps the explanatory variable should first be transformed, although ideally such transformations should have some kind of *a priori* justifiability.

5.4 Further Investigation in Predicted Time of Day Differences

Consider the difference in estimated FEEDRATE between dawn and 1330H, the estimated vertex at which FEEDRATE, as a function of TIME, takes on a minimum value. If one accepts this 1330H value as exact and overlooks the fact that chicks in a given nest are 7 to 8 hours (at most 1/3 of a day) older at 1330H than at dawn, then one can estimate this difference and obtain a measurement of the accuracy of this estimate for a given nest on a single day.

Using the following notation:

t_D=time at dawn in half-hours, e.g. dawn at 0530H would give a t_D-value of 11 (cf. Chapter 2)

t_M=time at minimum FEEDRATE, taken to be 27 (i.e. 1330H)

$\hat{y}_D$=fitted/predicted FEEDRATE for a given nest at t_D

$\hat{y}_M$=fitted/predicted FEEDRATE for same nest as in $\hat{y}_D$ and on same day, but at t_M

then:

$$\hat{y}_D - \hat{y}_M$$

estimates

$$E(Y|t_D) - E(Y|t_M)$$

where $E(Y|t_D)$ refers to the true (but unknown) expected or mean FEEDRATE at time t_D. Furthermore, ignoring the increase in AGECHK mentioned earlier, one can approximate $\hat{y}_D - \hat{y}_M$ by:

$$d = -0.0845(t_D - t_M) + 0.00156(t_D^2 - t_M^2)$$

Taking

$$t_D = 11$$

that is, dawn at 0530H, gives:

$$d=0.404$$

So on average, the chicks in a given nest can expect to jointly receive approximately 0.404 fish per hour more at dawn than at 1330H.

It can be further shown (see Technical Supplement, Chapter 10 for details) that this estimate has an approximate standard error of:

$$s(d)\approx\sqrt{8.014*10^{-3}}=8.952*10^{-2}$$

so that a 95% confidence interval for d would be:

$$(0.404-1.96(8.952*10^{-2}),0.404+1.96(8.952*10^{-2}))$$

$$=(0.229,0.579)$$

If one did not overlook the AGECHK difference from t_D to t_M, then d would be a function of how old the chicks were at dawn.

CHAPTER 6

TECHNICAL SUPPLEMENT FOR CHAPTER 1

A multiple regression model with indicator variables for qualitative effects came to mind immediately. As noted on pages 6-7 of Ref.(13), such a model, especially with polynomial terms, can serve as a suitable approximation over the given range of the data.

It should be noted, however, that the data comes from an unplanned experiment (more specifically, an observation study) which means likely multicollinearity amongst explanatory variables in 1st order terms alone. Any regression done should be approached with some care since the random error component will represent some 'lurking' variables which were unmeasured and may be highly correlated with the variables which were measured. In an unplanned experiment, the regression analysis could more likely lead to some false results about which explanatory variables have a significant association and which ones do not than in a planned experiment. Ref.(3) gives more details on this important point.

CHAPTER 7

TECHNICAL SUPPLEMENT FOR CHAPTER 2

Figure 14 shows the FORTRAN program PROCEMMS1 which used the input file EMMSDATA1 to produce the output files EMMSFDRT1 and EMMSREJECT1. These 3 files have already been documented and partially displayed in Chapter 2. Figure 15 shows the command source file RUNPROCE1 which compiled and ran PROCEMMS1. This file is activated by submitting the MTS command:

```
$SO RUNPROCE1
```

Similarly Figure 16 shows the FORTRAN program PROCEMMS2 which used the input file EMMSDATA2 to produce the output files EMMSFDRT2 and EMMSREJECT2. The corresponding command source file to compile and run this program would be very similar to the previous one and so is not shown.

Both programs were written in standard FORTRAN-IV, even though FORTRAN-77 was available. Use was made, however, of an apparent MTS extension: standard FORTRAN supposedly imposes a maximum input field width of 80 columns (because of default treatment of input as 'data cards'), but the programs had no problem accessing all 121 columns in each file using the standard formatted READ and WRITE statements. Otherwise use could have been made of the file record splitting option of the 'correct' command of the MTS file editor in order to change each 121 column record into 2

Figure 14: FORTRAN Program PROCEMMS1

```
C PREPARE EMMSDATA1 FOR ANY STATS PACKAGE
C
C UNIT 10 IS EMMSDATA1
C UNIT 11 IS EMMSFORT1
C UNIT 12 IS EMMSREJECT1
C
C        1         2         3         4         5         6         7
C234567890123456789012345678901234567890123456789012345678901234567890123456789012
      DIMENSION FISH(9),FLEN(9)
      RECNUM=0.0
C
C BEGIN READING EMMSDATA1
C INCREMENT RECORD COUNTER FOR EACH RECORD PROCESSED
C PUT REJECTED RECORDS INTO REJECT FILE
C
   10 TOTFSH=0.0
      TOTLEN=0.0
      FLAG=0.0
      READ (10,101,ERR=998,END=999) DATE,COLSZE,TIDE,IHOUR,MIN,CHK,
     *AGECHK,(FISH(I1),FLEN(I1),I1=1,9)
      RECNUM=RECNUM+1.0
C
C TRANSFORM DATE FOR 1984 DATA
C
      IF (DATE .GE. 8.0) GO TO 20
      DATE=100*(DATE-7.0)-14.0
      GO TO 30
   20 DATE=100*(DATE-8.0)+17.0
C
C CREATE TIDE DESIGN VARIABLES
C
   30 TIDEH=0.0
      TIDEM=0.0
      IF (TIDE .EQ. 1.0) GO TO 50
      IF (TIDE .EQ. 2.0) GO TO 40
      TIDEH=1.0
      GO TO 50
   40 TIDEM=1.0
C
C TRANSFORM TIME
C
   50 TIME=FLOAT(IHOUR)*2.0+FLOAT(MIN)/30.0
C
C CHECK CHK
C
      IF (CHK .GE. 0.0) GO TO 60
      WRITE (12,202) RECNUM,CHK
      GO TO 10
C
C CHECK AGECHK
C
   60 CONTINUE
      IF (AGECHK .GE. 0.0) GO TO 70
      WRITE (12,203) RECNUM,AGECHK
      GO TO 10
C
C COMPUTE TOTAL FISH COUNT & LENGTH, PROVIDED NO MISSING DATA PRESENT
C IF MISSING LENGTHS, DON'T SKIP RECORD
C
   70 CONTINUE
      DO 55 I2=1,9
      IF (FISH(I2) .GE. 0.0) GO TO 80
      IBAD=I2
      GO TO 90
```

Figure 14, continued

```
   80 TOTFSH=TOTFSH+FISH(I2)
      IF (FLEN(I2) .LT. 0.0) FLAG=1.0
      IF (FLAG .EQ. 0.0) TOTLEN=TOTLEN+FLEN(I2)
   55 CONTINUE
      GO TO 100
   90 WRITE (12,204) RECNUM,FISH(IBAD),IBAD
      GO TO 10
C
C IF WE GOT THIS FAR WITHOUT LOOPING BACK, THEN RECORD CONTAINS NO
C       MISSING DATA CODES (EXCEPT FOR FISH LENGTHS)
C CALCULATE AVGLEN, PUT MODIFIED DATA IN NEW FILE & GO GET ANOTHER
C       RECORD
C ADD 4000 TO RECNUM IN ORDER TO CODE IT AS BEING FROM 1984 DATA FILE
C
  100 AVGLEN=0.0
      IF (FLAG .EQ. 0.0) GO TO 110
      AVGLEN=-1.0
      TOTLEN=-1.0
      GO TO 120
  110 CONTINUE
      IF (TOTFSH .GT. 0.0) AVGLEN=TOTLEN/TOTFSH
  120 RLABEL=RECNUM+4000.0
      WRITE(11,201) RLABEL,AVGLEN,TOTFSH,TOTLEN,DATE,COLSZE,TIDE,TIDEH,
     *TIDEM,TIME,CHK,AGECHK
      GO TO 10
  998 WRITE (11,206) RECNUM
  999 WRITE (11,207)
      WRITE (11,208)
      WRITE (11,209) RECNUM
      STOP
C
C FORMAT STATEMENTS
C
C        1         2         3         4         5         6         7
C234567890123456789012345678901234567890123456789012345678901234567890123456789012
  101 FORMAT(F4.2,T13,F4.1,F3.1,I2,I2,T34,F3.0,T38,F3.0,T41,9(F3.0,1X,
     *F5.2))
  201 FORMAT(T3,F6.0,1X,F10.6,2(1X,F8.0),7(1X,F3.0),1X,F4.0)
  202 FORMAT(' RECORD ',F5.0,' REJECTED FOR MISSING VALUE CODE'/T15,
     *F4.0,' FOR CHK'/)
  203 FORMAT(' RECORD ',F5.0,' REJECTED FOR MISSING VALUE CODE'/T15,
     *F4.0,' FOR AGECHK'/)
  204 FORMAT(' RECORD ',F5.0,' REJECTED FOR MISSING VALUE CODE'/T15,
     *F4.0,' FOR FISH ARRAY, POSITION ',I2/)
  206 FORMAT(' ***INPUT ERROR***'/' IN RECORD NUMBER ',F5.0)
  207 FORMAT('C234567890123456789012345678901234567890123456789012345678
     *90123456789012'/'C',8X,'1',9X,'2',9X,'3',9X,'4',9X,'5',9X,'6',9X,
     *'7')
  208 FORMAT(/T4,'LABEL',T11,'AVGLEN',T22,'TOTFSH',T31,'TOTLEN',T38,
     *'DATE',T43,'COL',3(1X,'TDE'),1X,'TIME',T64,'#',T67,'AGECHK'/T43,
     *'SIZE',T52,'HI',T55,'MED',T63,'CHK')
  209 FORMAT(/'LAST RECORD READ:   ',F6.0)
      END
```

Figure 15: Source File RUNPROCE1

```
$empty procemms1.ob ok
$run *ftn scards=procemms1 sprint=-news spunch=procemms1.ob
$empty emmsreject1 ok
$empty emmsfdrt1 ok
$run procemms1.ob 10=emmsdata1 11=emmsfdrt1 12=emmsreject1
```

Figure 16: FORTRAN Program PROCEMMS2

```
C PREPARE EMMSDATA2 FOR ANY STATS PACKAGE
C
C UNIT 10 IS EMMSDATA2
C UNIT 11 IS EMMSFDRT2
C UNIT 12 IS EMMSREJECT2
C
C        1         2         3         4         5         6         7
C234567890123456789012345678901234567890123456789012345678901234567890123456789012
      DIMENSION FISH(9),FLEN(9)
      RECNUM=0.0
C
C BEGIN READING EMMSDATA2
C INCREMENT RECORD COUNTER FOR EACH RECORD PROCESSED
C PUT REJECTED RECORDS INTO REJECT FILE
C
   10 TOTFSH=0.0
      TOTLEN=0.0
      FLAG=0.0
      READ (10,101,ERR=998,END=999) DATE,COLSZE,TIDE,IHOUR,MIN,CHK,
     *AGECHK,(FISH(I1),FLEN(I1),I1=1,9)
      RECNUM=RECNUM+1.0
C
C TRANSFORM DATE FOR 1985 DATA
C
      IF (DATE .GE. 8.0) GO TO 20
      DATE=100*(DATE-7.0)-19.0
      GO TO 30
   20 DATE=100*(DATE-8.0)+12.0
C
C CREATE TIDE DESIGN VARIABLES
C
   30 TIDEH=0.0
      TIDEM=0.0
      IF (TIDE .EQ. 1.0) GO TO 50
      IF (TIDE .EQ. 2.0) GO TO 40
      TIDEH=1.0
      GO TO 50
   40 TIDEM=1.0
C
C TRANSFORM TIME
C
   50 TIME=FLOAT(IHOUR)*2.0+FLOAT(MIN)/30.0
C
C CHECK CHK
C
      IF (CHK .GE. 0.0) GO TO 60
      WRITE (12,202) RECNUM,CHK
      GO TO 10
C
C CHECK AGECHK
C
   60 CONTINUE
      IF (AGECHK .GE. 0.0) GO TO 70
      WRITE (12,203) RECNUM,AGECHK
      GO TO 10
C
C COMPUTE TOTAL FISH COUNT & LENGTH, PROVIDED NO MISSING DATA PRESENT
C IF MISSING LENGTHS, DON'T SKIP RECORD
C
   70 CONTINUE
      DO 55 I2=1,9
      IF (FISH(I2) .GE. 0.0) GO TO 80
      IBAD=I2
      GO TO 90
```

Figure 16, continued

```
   80 TOTFSH=TOTFSH+FISH(I2)
      IF (FLEN(I2) .LT. 0.0) FLAG=1.0
      IF (FLAG .EQ. 0.0) TOTLEN=TOTLEN+FLEN(I2)
   55 CONTINUE
      GO TO 100
   90 WRITE (12,204) RECNUM,FISH(IBAD),IBAD
      GO TO 10
C
C IF WE GOT THIS FAR WITHOUT LOOPING BACK, THEN RECORD CONTAINS NO
C       MISSING DATA CODES (EXCEPT FOR FISH LENGTHS)
C CALCULATE AVGLEN, PUT MODIFIED DATA IN NEW FILE & GO GET ANOTHER
C       RECORD
C ADD 5000 TO RECNUM IN ORDER TO CODE IT AS BEING FROM 1985 DATA FILE
C
  100 AVGLEN=0.0
      IF (FLAG .EQ. 0.0) GO TO 110
      AVGLEN=-1.0
      TOTLEN=-1.0
      GO TO 120
  110 CONTINUE
      IF (TOTFSH .GT. 0.0) AVGLEN=TOTLEN/TOTFSH
  120 RLABEL=RECNUM+5000.0
      WRITE(11,201) RLABEL,AVGLEN,TOTFSH,TOTLEN,DATE,COLSZE,TIDE,TIDEH,
     *TIDEM,TIME,CHK,AGECHK
      GO TO 10
  998 WRITE (11,206) RECNUM
  999 WRITE (11,207)
      WRITE (11,208)
      WRITE (11,209) RECNUM
      STOP
C
C FORMAT STATEMENTS
C
C        1         2         3         4         5         6         7
C234567890123456789012345678901234567890123456789012345678901234567890123456789012
  101 FORMAT(F4.2,T13,F4.1,F3.1,I2,I2,T34,F3.0,T38,F3.0,T41,9(F3.0,1X,
     *F5.2))
  201 FORMAT(T3,F6.0,1X,F10.6,2(1X,F8.0),7(1X,F3.0),1X,F4.0)
  202 FORMAT(' RECORD ',F5.0,' REJECTED FOR MISSING VALUE CODE'/T15,
     *F4.0,' FOR CHK'/)
  203 FORMAT(' RECORD ',F5.0,' REJECTED FOR MISSING VALUE CODE'/T15,
     *F4.0,' FOR AGECHK'/)
  204 FORMAT(' RECORD ',F5.0,' REJECTED FOR MISSING VALUE CODE'/T15,
     *F4.0,' FOR FISH ARRAY, POSITION ',I2/)
  206 FORMAT(' ***INPUT ERROR***'/' IN RECORD NUMBER ',F5.0)
  207 FORMAT('C234567890123456789012345678901234567890123456789012345678
     *901234567890123'/'C',8X,'1',9X,'2',9X,'3',9X,'4',9X,'5',9X,'6',9X,
     *'7')
  208 FORMAT(/T4,'LABEL',T11,'AVGLEN',T22,'TOTFSH',T31,'TOTLEN',T38,
     *'DATE',T43,'COL',3(1X,'TDE'),1X,'TIME',T64,'#',T67,'AGECHK'/T43,
     *'SIZE',T52,'HI',T55,'MED',T63,'CHK')
  209 FORMAT(/'LAST RECORD READ:   ',F6.0)
      END
```

consecutive records of, say, 70 and 51 columns.

It should be noted that since all the well-known statistical software packages at SFU (GLIM, BMDP, MINITAB, SPSS, MIDAS) are programmed in FORTRAN and use FORTRAN in some of their options (such as user-specified I/O formats), a working knowledge of standard FORTRAN would therefore be helpful to any statistics graduate student in the non-thesis option. Such a knowledge will be assumed for the duration of this and future Techinical Supplements. A reference such as Ref.(7) can be consulted on this basis.

The necessary column specifications were obtained by using the MTS file editor to append column counter lines to the end of each EMMSDATA input file, as has already been described in Chapter 2. These lines were removed prior to the running of each program since the programs were designed to read to the end of each input file, and the column counter lines were not intended as input.

As for the EMMSREJECT files, it was noticed that no record (observation) was rejected for having a missing value code (-1.) for any one of the 9 fish species type counts. Of course one would hope that this was because no such missing data did indeed occur, otherwise some TOTFSH calculations would be incorrect. To double check this, as is good programming practice, a copy of EMMSDATA2 was made and an extra record inserted with a value of -1 for number of fish

delivered of species type 1. When PROCEMMS2 was run, the extra record was rejected and put into the EMMSREJECT2 file with the appropriate message. As the only differences between the 2 PROCEMMS programs are the I/O file specifications, date transformations, and program generated labels, it could be assumed that PROCEMMS1 would also treat such a record appropriately. Thus the original input files could be accepted as free from missing data for fish counts.

Missing data for fish lengths however, would by themselves not cause a record to be rejected, since these quantities were not used in the analysis. Instead they would serve to set the average length value to -1. A TOTFSH value of 0 caused this average to be set to 0. Otherwise the average length (AVGLEN) was computed as:

$$AVGLEN=\frac{TOTLEN}{TOTFSH}$$

where TOTLEN is the sum of the total lengths of fish summed over the 9 species types. Again this quantity was calculated for a possible future use, which did not materialize.

The design variables TIDEH and TIDEM created by the PROCEMMS programs for the qulitative variable TIDE are of course not the only ones possible. But some design variables may present additional difficulties. As an example some analysts would prefer the following design variables:

```
HITIDE =  1, if TIDE=3 (high tide)
          0, if TIDE=2 (midtide)
         -1, if TIDE=1 (low tide)
```

```
MEDTIDE=  0, if TIDE=3
          1, if TIDE=2
         -1, if TIDE=1
```

In an all possible or best k subsets search or stepwise routine, it is possible that one of the two design variables is used in a final accepted model, but the other one is not. In the case of TIDEH and TIDEM this presents no problem. If, say, TIDEH makes it into a final model but TIDEM does not, then it is because the data suggest that only a high tide (or some correlated 'lurking' variable which changes factor levels only at high tide--see Ref.(3)) has an association with FEEDRATE. If, however, HITIDE gets into a final model without MEDTIDE, then an equal interval scale of effects is implied:

```
high tide: HITIDE= 1 }
midtide  : HITIDE= 0 } increment of 1 } equalscaling
low tide : HITIDE=-1 } increment of 1 }
```

Now the gain/loss in having high tide over medium tide is equal to that of having medium tide over low tide, and hence one-half that of having high tide over low tide.

This is not a desirable situation, unless one has such prior information. A solution would be to force all design variables into the model for a particular factor when one of them is chosen. This is not necessary if the design variables are chosen as in the case of TIDEH and TIDEM.

CHAPTER 8

TECHNICAL SUPPLEMENT FOR CHAPTER 3

A number of observations and developments may be made on the contents of Chapter 3. In this chapter the usual matrix formulation:

$$\underline{Y}=X\underline{\beta}+\underline{\epsilon}$$

will be used, where:

(1) $\underline{Y}$ is a random vector in R^n containing the n response variables:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

Once these values are observed, $\underline{Y}$ is replaced by $\underline{y}$.

(2) X is an n by $(p+1)$ real matrix. The first column is all 1's, and each of the remaining columns contains the n observations for each of the p explanatory variables.

(3) $\underline{\beta}$ is the vector of regression parameters in $R^{(p+1)}$.

(4) $\underline{\epsilon}$ is the random vector of stochastically independent error terms and has the distribution

$$\underline{\epsilon}\sim N_n(\underline{0},\sigma^2 I_n)$$

That is, $\underline{\epsilon}$ has an n-variate normal distribution.

8.1 On the Stochastic Independence Assumption

One of the assumptions that must be made is that the components of $\underline{Y}$, the random response vector, must be

stochastically independent (or very nearly so). This assumption about FEEDRATE observations at first appears to be questionable since the fishers compete with one another over a finite supply of fish. In fact, depending on TIME and the season state (as reflected through DATE), the number of fishers could even exceed the number of fish. Such contemplation would imply covariance between feedrates for chicks belonging to the different fishers. Simon's assurance, however, was that the birds actually seem to fish from 'fishing territories' in which no other bird would invade or interfere. This territorial aspect of individual fishers and an assumption of random fish movement in the water suggests that to assume stochastic independence amongst FEEDRATE observations may be acceptable.

8.2 On the Model Selection Criterion Used

The criterion used for the best model in the P9R runs was the maximization of R^2, with attention being paid to the number of explanatory variables being inserted into the model. The Mallows' C_p-criterion, which is the default in P9R, could also have been used, although the conclusions might have been different.

With this criterion, the relation:

$$C_p \geq p$$

is supposed to be observable for the most part, the few

exceptions being due to random variation. The actual criterion is to minimize C_p and still keep C_p as close to p as possible (Ref.(11) pg. 426, Ref.(6) pg. 316, Ref.(5) pg. 300).

A P9R run was done using this criterion (details not shown). The result was that the program chose a set of 8 explanatory variables giving:

$$C_p=9.52$$

although a set of 9 explanatory variables with:

$$C_p=9.53$$

would also have been anexcellent choice. The model selected by this criterion might still however be too large (Ref.(5), pg. 305). The residual plots also showed little improvement over those in Figures 4.a-4.p. The current best model of Chapter 3 still seems preferable despite having:

$$C_p=10.53$$

although this is nonetheless the smallest C_p-value of all other 5-variable models.

8.3 On the (non) Use of *F*-Statistics

Throughout this report the use of *F*-statistics has been de-emphasized. One reason was given in Chapter 5, another one is the following argument.

Consider again the quantity:

$$R^2=1-\frac{SS(error)}{SS(total)}$$

$$=\frac{SS(regression)}{SS(total)}$$

In the usual formal regression test, mentioned in Chapter 5:

$$H_0: \beta_1=\beta_2=\cdots=\beta_p=0$$

against

$$H_A: \text{At least one of } \beta_j \text{ is not zero } (1\leq j\leq p)$$

the usual test statistic used is:

$$F^*=\frac{\left(\frac{SS(regression)}{p}\right)}{\left(\frac{SS(error)}{n-p-1}\right)}$$

Under H_0, $F^*\sim F(p,n-p-1)$. It follows that

$$F^*=\left(\frac{n-p-1}{p}\right)\left(\frac{1}{\frac{1}{R^2}-1}\right)$$

so that if R^2 is 'small', then:

$$\frac{1}{R^2}$$

will be 'large' and thus the quantity:

$$\frac{1}{\frac{1}{R^2}-1}$$

will again be 'small'. But F^* may still be significant for a sufficiently large factor of:

$$\frac{n-p-1}{p}$$

so that a poor fitting model (low R^2) may give significant F^*-values, thus triggering the decision to reject H_0.

The current best model shows this:

$$n=524$$
$$p=5$$
$$R^2=0.1232$$

so

$$F^*=14.56$$

results. This was the F-statistic value found in the P9R run on the current best model alone. Looking up a table of F-distribution percentage points (with 5 degreees of freedom for the numerator, 518 degreees of freedom for the denominator),

$$F=5.43$$

is significant at $\alpha=0.001$.

Of course it should be realized that a model can give a high F-value, but still be of little use for predictive purposes if one wishes to use the model in that capacity (Ref.(5), pg. 129-30). Any model search technique should thus be used with caution and judgement in this regard. Used mechanically the results could be misleading (Ref.(5), pg. 300). This is particularly true in unplanned experiments such as Simon's (Ref.(5) pg. 295, Ref.(3)).

8.4 An Extra Note on Outputs

Although the outputs are not shown in their entirety, one omitted table that should be investigated after running a BMDP program is the 'Summary Statistics for Each Variable'

in order to check for outlier or 'wild' values of any variable. As an example, a previous P9R run reported a value of 10 in that table under the column labelled 'Maximum Value', for a variable which was supposed to take on only values of 0 or 1. The reason for this particular misinterpretation was that in Simon's data files, all integer data values were supposed to have decimal points after them, but this particular '1' did not, and the FORTRAN format used when inputting the value interpreted the blank after it as a 0, thus changing 1 into 10.

8.5 On the Non-use of Centred Explanatory Variables

The current best model contains some second order terms in the quantitative variables, and is hence a polynomial multiple regression. For such variables, it is usually recommended (Ref.(11) pg. 300-1) to use a centre transformed variable:

$$x_{ij}^{*} = x_{ij} - \bar{x}_{\cdot j}$$

where:

$$\bar{x}_{\cdot j} = \frac{1}{n} \sum_{i=1}^{n} x_{ij}$$

The purpose in doing this is to reduce multicollinearity caused by typically large values for the correlation between x_{ij} and its higher order terms

$$x_{ik}=x_{ij}^2$$

and so on. Such multicollinearity makes the computation of $(X'X)^{-1}$ difficult to control for round-off error. This of course makes the regression coefficient least squares estimates:

$$\underline{b}=LSE(\underline{\beta})=(X'X)^{-1}X'\underline{y}$$

highly suspect with regards to accuracy. Some coefficients may even have the wrong sign (Ref.(10) pg. 287).

This approach was not pursued at length, however, since the BMDP programs typically provide a 'tolerance' control for matrix inversion. P9R, in particular, carries out all computations in double precision. Furthermore, if the multicollinearity did cause problems in the X'X inversion, then this would show up as large estimated variances of the regression coefficient estimates, so that statistically significant coefficients would be hard to find (Ref.(11) pg. 318). This is because the estimated variances are the diagonal elements of:

$$Cov(\underline{b})=(s^2)(X'X)^{-1}$$

where

$$s^2=\frac{SS(error)}{n-p-1}$$

and if the X'X matrix were difficult to invert, these diagonal elements would typically be large. In the case of the current best model, all the regression coefficients have sufficiently small estimated variances so that a hypothesis test of the form:

$$H_0: \beta_j = 0$$

for a single j in $\{1,2,\ldots,p\}$, would result in H_0 being rejected at a 1% level of significance.

8.6 On the Creation of the R^2-plot

The usefulness of an R^2-plot is already well documented on pages 422-4 of Ref.(11). The R^2-plot shown in Figure 3 was constructed using R^2-values for models found not only in Figure 2, but also from models using GLIM runs with two types of 'forward selection' procedures. Starting with an empty model, explanatory variables were added one-by-one on the basis of maximum R^2 improvement (procedure 1) or minimum R^2 improvement (procedure 2). Procedure 1 is analogous to maximum F improvement in the forward selection phase of stepwise regression (see pg. 430-6 of Ref. (11), for example). The purpose of procedure 2 was purely one of contrast in order to see how much worse a model fit could be if the 'wrong' explanatory variables were selected.

GLIM works with generalized linear models, the rudiments of which were indicated in Chapter 4, and are more thoroughly covered in Ref.(8). Briefly, instead of $SS(error)$ as a goodness of fit measure, GLIM uses deviance, D, defined as:

$$D=\sigma^2 S(\underline{\hat{\mu}},\underline{\tilde{\mu}})$$

where:

$$S(\hat{\underline{\mu}},\tilde{\underline{\mu}})=-2[l(\hat{\underline{\mu}},\sigma^2;\underline{y})-l(\tilde{\underline{\mu}},\sigma^2;\underline{y})]$$

$\hat{\underline{\mu}}$ is the n by 1 vector of fitted values (that is, $\hat{\underline{y}}=\hat{\underline{\mu}}$) for the current model being entertained

$\tilde{\underline{\mu}}$ is the n by 1 vector of fitted values for the fullest possible model, namely, when the number of regression coefficients, $p+1$, equals the number of observations, n. In such a case it can be shown that $\tilde{\underline{\mu}}=\underline{y}$, the observations themselves.

l is the log-likelihood function.

S is called the 'scaled deviance', because of the presence of the scale factor, σ^2, in the expression for D.

It can be further shown that for the special case of the normal-theory multiple regression, namely:

$$\underline{Y}=\underline{\mu}+\underline{\epsilon}$$

$$\underline{\eta}=\underline{\mu}=\mathbb{X}\underline{\beta}$$

$$\underline{\epsilon}\sim N_n(\underline{0},\sigma^2 I_n)$$

then for the null model ($(p+1)=1$; no regression coefficients except for constant term, β_0):

$$D(null\ model)=SS(total)$$

and for any current model with $1<p+1<n$:

$$D(current\ model)=SS(error)$$

Thus one can calculate R^2-values for a sequence of models of one's own choosing, which is easily implemented with GLIM's interactive atmosphere. One such program is shown in Figure 17.a, with the run results in Figure 17.b, where the selection is shown for the first explanatory variable. One can see for exmaple:

$$D(null\ model)=170.3=SS(total)$$

Figure 17.a: GLIM Command File which Generates Figure 17.b

```
$EMPTY EOUTGLIM18 OK
$RUN UNSP:GLIM 1=EMMSFORATE 2=EOUTGLIM18
$C
$C   GLIM RUN ON FILE EMMSFDRATE TESTING OUT NORMAL THEORY MODEL
$C
$C ************************
$C *  GET DATA & TRANSFORM  *
$C ************************
$C
$OUTPUT 2 132
$UNITS 524
$DATA LABEL TOTFSH DATE COLSZE TIDE TIME NUMCHK AGECHK
$FACTOR TIDE 3
$FORMAT
(3X,F5.0,17X,F3.0,9X,F4.0,2(1X,F3.0),8X,2(1X,F3.0),1X,F4.0)
$DINPUT 1
$LOOK 1 15 LABEL TOTFSH DATE COLSZE TIDE TIME NUMCHK AGECHK
$CALC FEEDRATE=TOTFSH/2.0
$C
$C   ADD QUADRATIC TERMS
$C
$CALC SQDATE=DATE*DATE: SQTIME=TIME*TIME: SQAGE=AGECHK*AGECHK
$CALC SQNUM=NUMCHK*NUMCHK: SQCOL=COLSZE*COLSZE
$C
$C ***********************************
$C *  SPECIFY MODEL TO BE ANALYZED   *
$C *  ALLOW FOR OVER-DISPERSION      *
$C ***********************************
$C
$YVAR FEEDRATE
$ERROR N
$LINK I
$SCALE 0
$C
$C ******************************
$C *  NOW FIT MODELS            *
$C *  FIRST NULL MODEL          *
$C *  THEN ALL POSSIBLE ONE     *
$C *     VARIABLE MODELS        *
$C ******************************
$C
$FIT %GM
$DISP A
$FIT DATE
$DISP A
$FIT COLSZE
$DISP A
$FIT TIDE
$DISP A
$FIT TIME
$DISP A
$FIT NUMCHK
$DISP A
$FIT AGECHK
$DISP A
$FIT SQDATE
$DISP A
$FIT SQTIME
$DISP A
$FIT SQAGE
$DISP A
$FIT SQNUM
$DISP A
$FIT SQCOL
$DISP A
$C
$C ***************************************************
$C *  NEXT PHASE OF FOWARD SELECTION TO BE FOUND IN   *
$C *       EMMSGLIM19 FILE                           *
$C ***************************************************
$C
$STOP
```

Figure 17.b: GLIM Run on Normal-Theory Model

```
CYCLE  DEVIANCE      DF
   1     170.3       523

      ESTIMATE       S.E.       PARAMETER
   1  0.6746      0.2493E-01  %GM
  SCALE PARAMETER TAKEN AS    0.3256

CYCLE  DEVIANCE      DF
   1     166.4       522

      ESTIMATE       S.E.       PARAMETER
   1  0.8426      0.5413E-01  %GM
   2 -0.1010E-01  0.2896E-02  DATE
  SCALE PARAMETER TAKEN AS    0.3188

CYCLE  DEVIANCE      DF
   1     168.7       522

      ESTIMATE       S.E.       PARAMETER
   1  0.7566      0.4476E-01  %GM
   2 -0.1067E-01  0.4849E-02  COLS
  SCALE PARAMETER TAKEN AS    0.3232

CYCLE  DEVIANCE      DF
   1     169.6       521

      ESTIMATE       S.E.       PARAMETER
   1  0.7454      0.5490E-01  %GM
   0     ZERO      ALIASED     TIDE(1)
   2 -0.8585E-01  0.6756E-01  TIDE(2)
   3 -0.9246E-01  0.6778E-01  TIDE(3)
  SCALE PARAMETER TAKEN AS    0.3255

CYCLE  DEVIANCE      DF
   1     167.9       522

      ESTIMATE       S.E.       PARAMETER
   1  0.8917      0.8342E-01  %GM
   2 -0.8685E-02  0.3187E-02  TIME
  SCALE PARAMETER TAKEN AS    0.3216

CYCLE  DEVIANCE      DF
   1     161.3       522

      ESTIMATE       S.E.       PARAMETER
   1  0.3043      0.7287E-01  %GM
   2  0.2756      0.5114E-01  NUMC
  SCALE PARAMETER TAKEN AS    0.3090

CYCLE  DEVIANCE      DF
   1     168.3       522

      ESTIMATE       S.E.       PARAMETER
   1  0.8429      0.7267E-01  %GM
   2 -0.6901E-02  0.2801E-02  AGEC
  SCALE PARAMETER TAKEN AS    0.3224

CYCLE  DEVIANCE      DF
   1     164.7       522

      ESTIMATE       S.E.       PARAMETER
   1  0.7944      0.3765E-01  %GM
   2 -0.3428E-03  0.8173E-04  SQDA
  SCALE PARAMETER TAKEN AS    0.3156

CYCLE  DEVIANCE      DF
   1     168.7       522

      ESTIMATE       S.E.       PARAMETER
   1  0.7696      0.5019E-01  %GM
   2 -0.1386E-03  0.6366E-04  SQTI
  SCALE PARAMETER TAKEN AS    0.3233

CYCLE  DEVIANCE      DF
   1     166.9       522

      ESTIMATE       S.E.       PARAMETER
   1  0.7980      0.4507E-01  %GM
   2 -0.1833E-03  0.5602E-04  SQAG
  SCALE PARAMETER TAKEN AS    0.3196

CYCLE  DEVIANCE      DF
   1     161.3       522

      ESTIMATE       S.E.       PARAMETER
   1  0.4881      0.4228E-01  %GM
   2  0.9187E-01  0.1705E-01  SQNU
  SCALE PARAMETER TAKEN AS    0.3090

CYCLE  DEVIANCE      DF
   1     169.5       522

      ESTIMATE       S.E.       PARAMETER
   1  0.7060      0.3217E-01  %GM
   2 -0.3686E-03  0.2391E-03  SQCO
  SCALE PARAMETER TAKEN AS    0.3247
```

so that amongst all models containing only one of the given candidate main effects, the one containing TIDE (both design variables moved as a unit) showed the lowest R^2 improvement:

$$D(current\ model)=169.6=SS(error)$$

$$R^2=1-\frac{169.6}{170.3}=0.0041$$

and the one containing only NUMCHK showed the highest R^2 improvement:

$$D(current\ model)=161.3$$

$$R^2=1-\frac{161.3}{170.3}=0.0528$$

Procedure 1 then takes the model with NUMCHK and tries out all the remaining effects (including SQNUM) in order to find the 2 effect model with the highest R^2 improvement when NUMCHK is already in. Similarly procedure 2 tries out all 2 effect models which contain TIDE. As it turned out, procedure 1 ended up with SQDATE being added to the NUMCHK model, and procedure 2 ended up with SQCOL being added to the TIDE model.

One then ends up with 2 sequences of nested models along with their R^2-values, which when plotted against p suggest 2 R^2 improvement paths, which were shown in Figure 3. Procedure 2 is shown in the bottom path, but procedure 1 did not result in the top path. This is because the forward selection did not find the best possible model at all values of p. This drawback is not unknown in stepwise regression (Ref.(11) pg. 435, Ref.(6) pg. 317) and is due to the

presence of multicollinearity in the data, since an explanatory variable's ability to decrease *SS(error)* when brought into the current model depends on what variables were already present in the model (Ref.(11) pg. 271-282). In particular the forward R^2 selection outlined in this section missed the current best model of Chapter 3, which can be found on the top path in Figure 3 above p=5, as can be confirmed from Table 7.

8.7 On the Generation of the Outputs in Chapter 3 of the Client Report

In Chapter 3 of the Client Report, Figures 1 through 4 are introduced along with Table 8. The command source files which generated each of Figures 1.a-p through 4.a-p are shown in Figures 18 through 21. All are examples of BMDP command files. The details of using any BMDP program can be looked up in Ref.(4). The layout of the '$run' command along with its input/ouput file specification is specific to the MTS system. Ref.(1) contains details on how to run the BMDP programs on the MTS system.

In Figure 18 is the P6D program which generated the plots of Figures 1.a-p. The fact that there are 524 observations was obtained by using the MTS '$list' command to output the file EMMSFDRT, and then observing the line number of the last record of observations. It also shows an

Figure 18: P6D Command File which Generated Figures 1.a-p

```
$empty eoutp6d1 ok
$run *bmdp sprint=eoutp6d1 7=emmsfdrate par=p6d
/ problem     title is 'EOUTP6D1:  plot TOTFSH against explanatory
                  variables & do some scatter plots without totfsh'.
/ input       unit is 7.
              cases are 524.
              variables are 10.
              format is '(3x,f5.0,17x,f3.0,9x,f4.0,6(1x,f3.0),1x,f4.0)'.
/ variable    names are label,totfsh,date,colsze,tide,tidehi,tidemed,
                        time,numchk,agechk,sqtime,sqdate,sqage,sqcol,
                        sqnum.
              add=5.
              label is label.
/ transform   sqtime=time*time.
              sqdate=date*date.
              sqage=agechk*agechk.
              sqcol=colsze*colsze.
              sqnum=numchk*numchk.
/ plot        yvar is totfsh.
              xvar are date,colsze,tide,numchk,agechk,time,
                       tidehi,tidemed,sqdate,
                       sqtime,sqage,sqcol,sqnum.
              cross.
              size=100,40.
/ plot        yvar is agechk.
              xvar is date.
              size=100,40.
/ plot        yvar is time.
              xvar is tide.
              size=100,40.
/ plot        yvar is colsze.
              xvar is numchk.
              size=100,40.
/ end
```

Figure 19: P9R Command File which Generated Figure 2

```
$empty eoutp9r6 ok
$run *bmdp sprint=eoutp9r6 7=emmsfdrate par=p9r
/ problem     title is 'EOUTP9R6:  Best (max R-square) 10 subsets
                       regression--feedrate response, no interactions'.
/ input       unit is 7.
              cases are 524.
              variables are 10.
              format is '(3x,a4,18x,f3.0,9x,f4.0,6(1x,f3.0),1x,f4.0)'.
/ variable    names are label,totfsh,date,colsze,tide,tidehi,tidemed,
                       time,numchk,agechk,sqtime,sqdate,sqage,sqnum,
                       sqcol,feedrate.
              add=6.
              label is label.
/ transform   sqtime=time*time.
              sqdate=date*date.
              sqage=agechk*agechk.
              sqnum=numchk*numchk.
              sqcol=colsze*colsze.
              feedrate=totfsh/2.0.
/ regress     dependent is feedrate.
              independent are date,colsze,tidehi,tidemed,time,numchk,
                              agechk,sqtime,sqdate,sqage,sqcol.
              method=rsq.
              number=10.
/ print       news.
              no shade.
/ plot        normal.
              yvar are predictd,residual,residual,residual,residual,
                       residual,residual,residual,residual,residual,
                       residual,residual,residual,residual,residual.
              xvar are feedrate,predictd,feedrate,date,colsze,tide,
                       tidehi,tidemed,time,numchk,agechk,sqtime,sqdate,
                       sqage,sqcol.
              size=115,50.
              hist.
/ end
```

Figure 20: P6D Command File which Generated Figure 3

```
$empty eoutp6d2 ok
$run *bmdp sprint=eoutp6d2 7=emmsrplt par=p6d
/ problem     title is 'EOUTP6D2:  R-sqaure plot for models without
                        interactions'.
/ input       unit is 7.
              cases are 29.
              variables are 2.
              format is '(2x,f2.0,3x,f6.4)'.
/ variable    names are p,Rsquare.
/ plot        yvar is Rsquare.
              xvar is p.
              size=100,40.
              no statistics.
              symbol='*'.
              minimum are 1,0.0.
              maximum are 12,0.20.
/end
```

Figure 21: P9R Command File which Generated Figures 4.a-p

```
$empty eoutp9r8 ok
$run *bmdp sprint=eoutp9r8 7=emmsfdrate par=p9r
/ problem     title is 'EOUTP9R8:  no search, but investigate current
                  "best" model'.
/ input       unit is 7.
              cases are 524.
              variables are 10.
              format is '(3x,a4,18x,f3.0,9x,f4.0,6(1x,f3.0),1x,f4.0)'.
/ variable    names are label,totfsh,date,colsze,tide,tidehi,tidemed,
                  time,numchk,agechk,sqtime,sqdate,sqage,sqnum,
                  sqcol,feedrate.
              add=6.
              label is label.
/ transform   sqtime=time*time.
              sqdate=date*date.
              sqage=agechk*agechk.
              sqnum=numchk*numchk.
              sqcol=colsze*colsze.
              feedrate=totfsh/2.0.
/ regress     dependent is feedrate.
              independent are numchk,agechk,sqage,time,sqtime.
              method=none.
/ print       news.
              no shade.
              matrices are corr,creg,rreg.
/ plot        normal.
              yvar are predictd,residual,residual,residual,residual,
                  residual,residual,residual,residual,residual,
                  residual,residual,residual,residual,residual
              xvar are feedrate,predictd,feedrate,date,colsze,tide
                  tidehi,tidemed,time,numchk,agechk,sqtime,sqdate,
                  sqage,sqcol.
              size=100,40.
              hist.
/ end
```

example of using the '/ plot' paragraph more than once, where the reason behind doing so was the choice of a new vertical axis variable.

Figure 19 shows the use of the P9R program with the specific request that the method of maximum R^2 be used to find the best 10 subsets as p varies from 1 to 11. Some of the results were given in Figure 2. Another interesting feature shown there is the 'news' sentence in the '/ plot' paragraph in order to get up-to-date information on the latest program modifications. This feautre is not mentioned in the BMDP manual (Ref.(4)).

Figure 20 shows the use of the P6D program to generate the R^2-plot, but using the file from Table 7 as input to produce the output in Figure 3. The table itself was written using the MTS file editor.

Figure 21 shows another P9R run but this time not for any model searching (hence the 'method=none.' sentence in the '/ regress' paragraph), but to take advantage of P9R's use of double precision arithmetic to obtain more accurate calcultions, and the program's plot facilities. The results of the latter were given in Figures 4.a-p.

CHAPTER 9

TECHNICAL SUPPLEMENT FOR CHAPTER 4

In Chapter 4, 3 methods were presented to try to modify the current best model in order to yield a higher R^2 without having to pay too high a penalty for doing so. There, computer run results of Figures 6 through 10 were introduced. The command source files which produced them will now be discussed along with other runs of interest.

9.1 Computer Runs for Method 1 of Chapter 4

Figure 22 shows a command source file similar to Figure 18, except that now some interaction terms are defined and included as potential explanatory variables. The parameter 'space=18000w' will be noticed in the '$run' command. The purpose of this is to increase the storage space required for the run. This will be explained in more detail in Chapter 7 of Part B where such a request plays a more predominant role. The results were partially shown in Figure 5.

The '/ plot' paragraph shows some interesting features. Firstly, 'residual' is a system vector for this particular program and contains residuals from the model of best fit, which the program has selected. Secondly, residual as a vertical axis for plots must be respecified for as many

Figure 22: P9R Command File which Generated Figure 6

```
$empty -eoutp9r7 ok
$run *bmdp sprint=-eoutp9r7 7=emmsfdrate par=p9r space=18000w
/ problem     title is 'EOUTP9R7:  Best 10 subsets regression with
                    feedrate response & important interactions'.
/ input       unit is 7.
              cases are 524.
              variables are 10.
              format is '(3x,a4,18x,f3.0,9x,f4.0,6(1x,f3.0),1x,f4.0)'.
/ variable    names are label,totfsh,date,colsze,tide,tideh,tidem,
                        time,numchk,agechk,sqtime,sqdate,sqage,sqcol,
                        feedrate,c1n1,c1a1,c1a2,t1a1,t1a2,t2a1,t2a2,
                        n1a1,n1a2,tht1,tmt1,tht2,tmt2,c2n1,c2a1,c2a2.
              add=21.
              label is label.
/ transform   sqtime=time*time.
              sqdate=date*date.
              sqage=agechk*agechk.
              sqcol=colsze*colsze.
              c1n1=colsze*numchk.
              c1a1=colsze*agechk.
              c1a2=colsze*sqage.
              c2a1=sqcol*agechk.
              c2a2=sqcol*sqage.
              c2n1=sqcol*numchk.
              t1a1=time*agechk.
              t1a2=time*sqage.
              t2a1=sqtime*agechk.
              t2a2=sqtime*sqage.
              n1a1=numchk*agechk.
              n1a2=numchk*sqage.
              tht1=tideh*time.
              tmt1=tidem*time.
              tht2=tideh*sqtime.
              tmt2=tidem*sqtime.
              feedrate=totfsh/2.0.
/ regress     dependent is feedrate.
              independent are date,colsze,tideh,tidem,time,numchk,
                              agechk,sqtime,sqdate,sqage,sqcol,c1n1,
                              c1a1,c1a2,t1a1,t1a2,t2a1,t2a2,n1a1,n1a2,
                              tht1,tmt1,tht2,tmt2,c2n1,c2a1,c2a2.
              method=rsq.
              tolerance=0.00001.
              number=10.
/ print       news.
              no shade.
/ plot        normal.
              yvar is residual,residual,residual,residual,residual,
                      residual,residual,residual,residual,residual,
                      residual,residual,residual,residual,residual,
                      residual,residual,residual,residual,residual,
                      residual,residual,residual,residual,residual,
                      residual,residual,residual,residual,residual.
              xvar are predictd,feedrate,date,colsze,tide,tideh,
                       tidem,time,numchk,agechk,sqtime,sqdate,sqage,
                       sqcol,c1n1,c1a1,c1a2,t1a1,t1a2,t2a1,t2a2,
                       n1a1,n1a2,tht1,tmt1,tht2,tmt2,c2n1,c2a1,c2a2.
              size=115,50.
              hist.
/ end
```

plots as are desired. This contrasts with the P6D program which has a 'cross' option so that a common vertical axis need be specified only once.

Figure 23 shows the GLIM command file which generated Figures 7.a through 7.c. It will be noted that use is made of the GLIM '$MACRO' command. This command is used to specify user-defined routines which require more than one line of typed instructions and are to be executed at least twice.

Another method of detecting important interactions was attempted along the lines of the graphical aids discussed in Chapter 4. First the necessary data had to be obtained.

Figure 24 shows a P9R command file which requests a fit on the current best model of Chapter 3 and a file for saving both the supplied input data and the model fit results, including the residuals. This data was saved for plots with P6D runs which could not be done with the more limited plot facilities in P9R (for example, P6D can do case selection for plots through the 'group' sentence in a '/ plot' paragraph whereas this is not possible in P9R). The '/ save' paragraph option was used to store the data in the file EP9RFILE and in the default unformatted binary layout. This default can be overridden by specifying a 'format' sentence in the '/ save' paragraph (Ref.(3) pg. 69). Nonetheless BMDP programs are able to read unformatted binary files, and so

Figure 23: GLIM Command File which Generated Figures 7.a-c

```
$EMPTY EOUTGLIM40 OK
$RUN UNSP:GLIM 1=EMMSFDRATE 2=EOUTGLIM40
$C
$C   GLIM RUN ON FILE EMMSFDRATE TESTING OUT NORMAL THEORY MODEL
$C
$C ***************************
$C *  GET DATA & TRANSFORM   *
$C ***************************
$C
$OUTPUT 2 132
$UNITS 524
$DATA LABEL TOTFSH DATE COLSZE TIDE TIME NUMCHK AGECHK
$FACTOR TIDE 3
$FORMAT
(3X,F5.0,17X,F3.0,9X,F4.0,2(1X,F3.0),8X,2(1X,F3.0),1X,F4.0)
$DINPUT 1
$LOOK 1 15 LABEL TOTFSH DATE COLSZE TIDE TIME NUMCHK AGECHK
$CALC FEEDRATE=TOTFSH/2.0
$C
$C   ADD QUADRATIC TERMS
$C
$CALC SQDATE=DATE*DATE: SQTIME=TIME*TIME: SQAGE=AGECHK*AGECHK
$CALC SQNUM=NUMCHK*NUMCHK: SQCOL=COLSZE*COLSZE
$C
$C ***********************************
$C *  SPECIFY MODEL TO BE ANALYZED   *
$C *  ALLOW FOR OVER-DISPERSION      *
$C ***********************************
$C
$YVAR FEEDRATE
$ERROR N
$LINK I
$SCALE 0
$C
$C ***************************
$C *  SET UP INPECT ROUTINE  *
$C ***************************
$C
$MACRO INSPECT $DISP A
               $EXTRACT %VL
               $LOOK 1 50 LABEL %FV %VL
               $ENDMAC
$C
$C ******************************************
$C *  NOW FIT MODELS                        *
$C *  FIRST NULL MODEL                      *
$C *  THEN CURRENT BEST MODEL               *
$C *  THEN FULL MODEL WITHOUT INTERACTIONS  *
$C *  THEN FULL MODEL WITH INTERACTIONS     *
$C ******************************************
$C
$FIT %GM
$DISP A
$FIT AGECHK+NUMCHK+TIME+SQAGE+SQTIME
$USE INSPECT
$FIT +DATE+COLSZE+TIDE+SQDATE+SQCOL
$USE INSPECT
$C
```

Figure 23, continued

```
$C  CALCULATE APPROPRIATE CROSS-PRODUCTS
$C
$CALC A1N1=AGEC*NUMC: A2N1=SQAG*NUMC: A1T1=AGEC*TIME: A1T2=AGEC*SQTI
$CALC A2T1=SQAG*TIME: A2T2=SQAG*SQTI: C1N1=COLS*NUMC: C2N1=SQCO*NUMC
$CALC C1A1=COLS*AGEC: C1A2=COLS*SQAG: C2A1=SQCO*AGEC: C2A2=SQCO*SQAG
$C
$C  NOW GO AHEAD WITH MODEL WITH INTERACTIONS
$C
$FIT +A1N1+A2N1+A1T1+A1T2+A2T1+A2T2+C1N1+C2N1+C1A1+C1A2+C2A1+C2A2+
      TIME.TIDE+SQTI.TIDE
$USE INSPECT
$STOP
```

Figure 24: P9R Run to Save Data for P6D Run

```
$empty eoutp9r9 ok
$run *bmdp sprint=eoutp9r9 7=emmsfdrate 8=ep9rfile par=p9r
/ problem     title is 'EOUTP9R9:  re-run of current best model, but
                       create data file for P6D runs'.
/ input       unit is 7.
              cases are 524.
              variables are 10.
              format is '(3x,a4,18x,f3.0,9x,f4.0,6(1x,f3.0),1x,f4.0)'.
/ variable    names are label,totfsh,date,colsze,tide,tidehi,tidemed,
                       time,numchk,agechk,sqtime,sqdate,sqage,sqnum,
                       sqcol,feedrate.
              add=6.
              label is label.
/ transform   sqtime=time*time.
              sqdate=date*date.
              sqage=agechk*agechk.
              sqnum=numchk*numchk.
              sqcol=colsze*colsze.
              feedrate=totfsh/2.0.
/ regress     dependent is feedrate.
              independent are numchk,agechk,sqage,time,sqtime.
              method=none.
/ print       news.
              no shade.
/ save        unit is 8.
              new.
              code is emms.
/ end
```

the default was used since the file's sole purpose was to provide input to a further P6D run.

Figure 25 shows the P6D run which was used to try out a graphical detection of an interaction between AGECHK with NUMCHK using the different values of NUMCHK:

single nest occupancy: NUMCHK=1
double nest occupancy: NUMCHK=2

The 3 graphs produced are shown in Figures 26.a through 26.c. The asterisk ('*') is reserved for cases where both 's' (single occupancy) and 'd' (double occupancy) are to occupy the same spot on the plot.

The question of whether the separate plots for single or double nest occupancies are sufficiently parallel to suggest no interaction, however, is not easy to answer from the graphs shown. Nevertheless it does seem that the separation effect due to the different levels of NUMCHK is not very strong. Note that for ease of comparison between the 3 plots P6D used the same horizontal and vertical scales in all of them. This is because the scales are determined from all cases before the subcases are selected for actual plotting. The user can override this by specifying different scales with each plot request. In any case, 'formal' methods such as the P9R run would still be used to quantify graphical intuition. This graphical approach was not pursued further.

Figure 25: P6D Command File which Generates Figures 26.a-c

```
$empty eoutp6d3 ok
$run *bmdp sprint=eoutp6d3 8=ep9rfile par=p6d
/ problem     title is 'EOUTP6D3:  plot of results from p9r8 run &
                       search for interactions with numchk'.
/ input       unit is 8.
              code is emms.
/ variable    grouping is numchk.
/ group       codes(9) are 1,2.
              names(9) are single,double.
/ plot        yvar is feedrate.
              xvar is agechk.
              group is single.
              group is double.
              groups are single,double.
              no statistics.
              size is 100,40.
/ end         .
```

Figure 26.a: FEEDRATE against AGECHK for NUMCHK=1

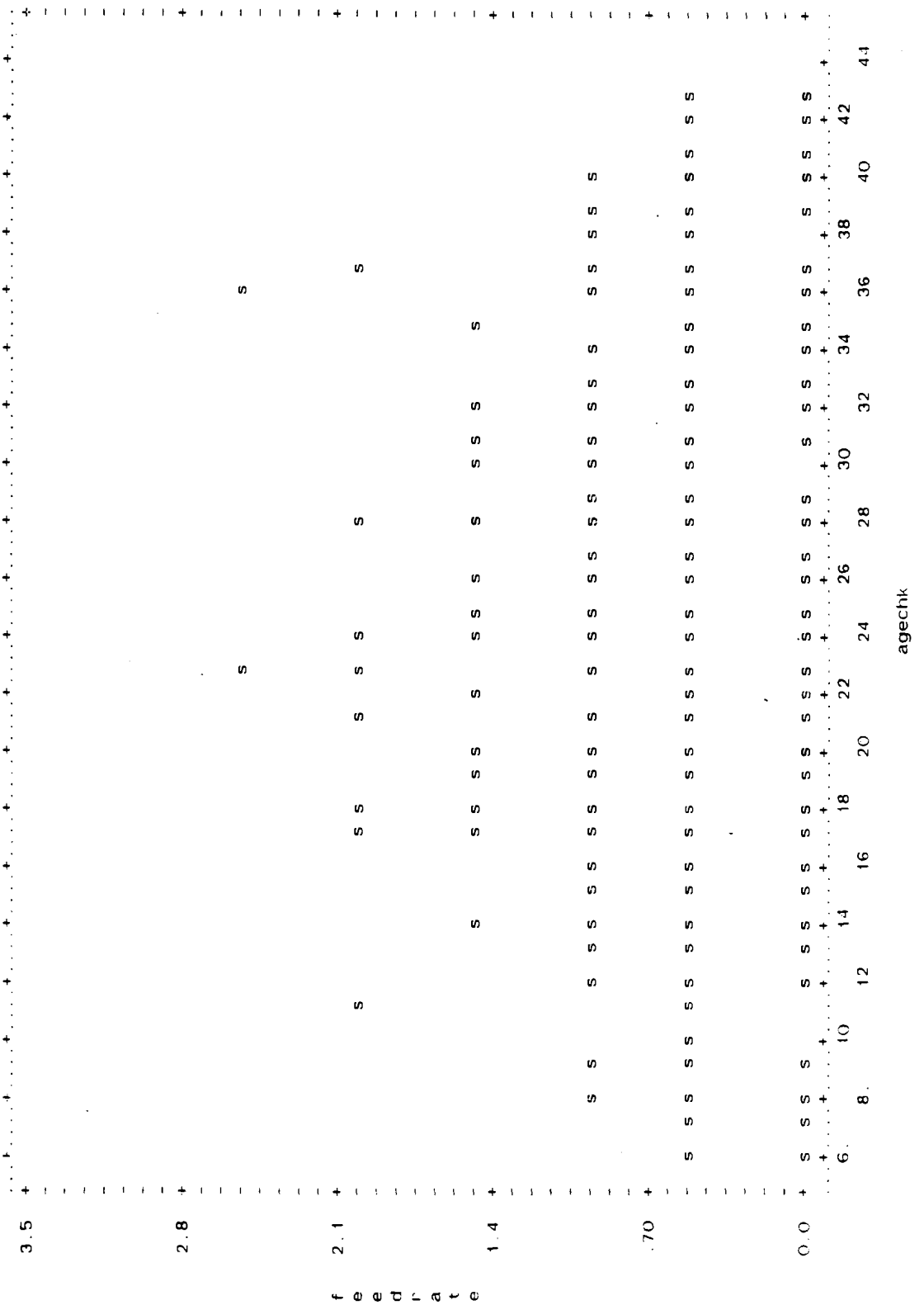


Figure 26.b: FEEDRATE against AGECHK for NUMCHK=2

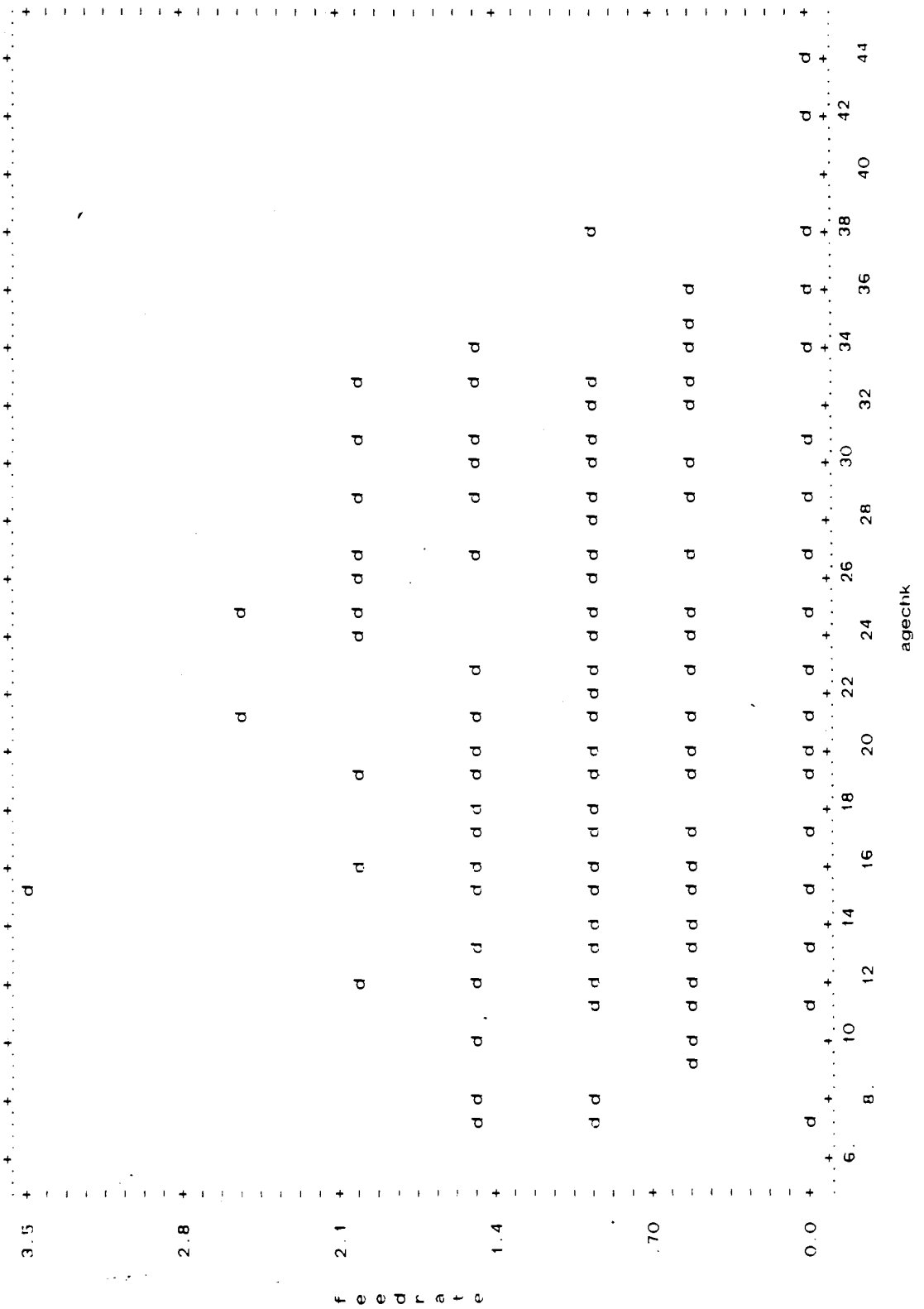
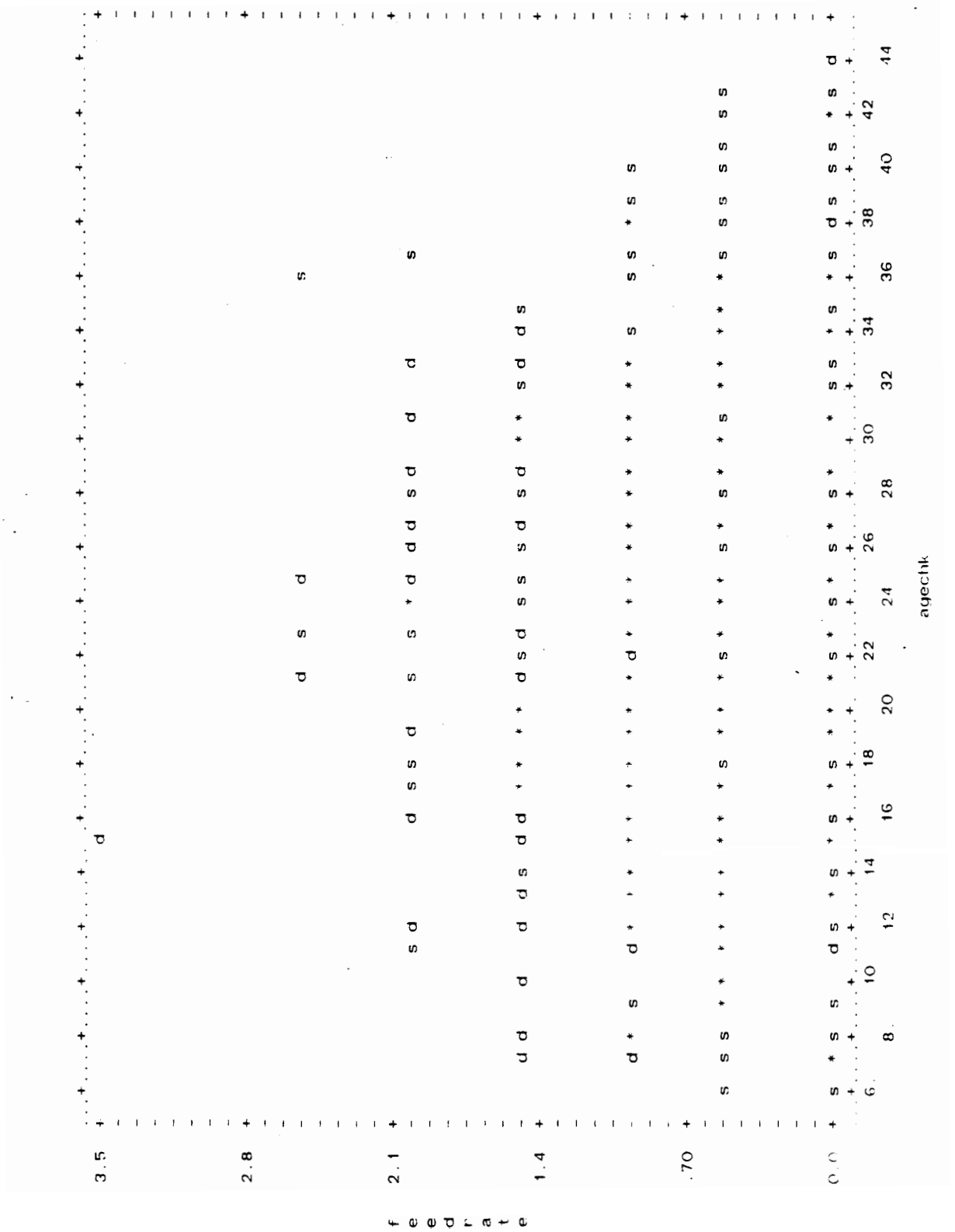


Figure 26.c: FEEDRATE against AGECHK for all NUMCHK values



9.2 Computer Runs for Method 2 of Chapter 4

Figure 27 shows the command source file of P9R commands which produced the model search for log-transformed FEEDRATE in Figure 8. Similarly Figure 28 shows the source file of P9R commands which did the search for square root transformed FEEDRATE in Figure 9.

9.3 Computer Runs for Method 3 of Chapter 4

Figure 29 shows the source file of commands from the GLIM statistical software package used to fit a null model, maximal model without interactions, and maximal model with interactions. The results were highlighted in Figure 10.

As noted in Chapter 4, the generalized R^2 measure:

$$R_g^2 = 1 - \frac{D(current\ model)}{D(null\ model)}$$

seems to perform more poorly than the R^2 for the normal-theory multiple linear regression. Also as noted in Chapters 4 and 8, for normal-theory regression:

$$SS(error) = D(current\ model)$$

$$SS(total) = D(null\ model)$$

and thus $R^2 = R_g^2$ but in general this is not necessarily true. In particular, for the log-linear model pursued in Method 3 of Chapter 4, the scaled deviance, $S(\hat{\underline{\mu}}, \underline{y})$, is given by:

$$S(\hat{\underline{\mu}}, \underline{y}) = 2 \sum_{i=1}^{n} y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i)$$

Figure 27: P9R Command File which Generated Figure 8

```
$empty -eoutp9r3 ok
$run *bmdp sprint=-eoutp9r3 7=emmsfdrate par=p9r
/ problem     title is 'EOUTP9R3:  logfdrt1 with original explanatory
                       variables--best 5 subsets regression:
                       feedrate transformed'.
/ input       unit is 7.
              cases are 524.
              variables are 10.
              format is '(3x,a4,18x,f3.0,9x,f4.0,6(1x,f3.0),1x,f4.0)'.
/ variable    names are label,totfsh,date,colsze,tide,tidehi,tidemed,
                       time,numchk,agechk,sqtime,sqdate,sqage,sqcol,
                       feedrate,logfdrt1.
              add=6.
              label is label.
/ transform   sqtime=time*time.
              sqdate=date*date.
              sqage=agechk*agechk.
              sqcol=colsze*colsze.
              feedrate=totfsh/2.0.
              logfdrt1=ln(feedrate+0.1).
/ regress     dependent is logfdrt1.
              independent are date,colsze,tidehi,tidemed,time,numchk,
                              agechk,sqtime,sqdate,sqage,sqcol.
              method=rsq.
/ print       news.
              no shade.
/ plot        normal.
              yvar are residual,residual.
              xvar are predictd,logfdrt1.
              size=115,50.
              hist.
/ end
```

Figure 28: P9R Command File which Generated Figure 9

```
$empty -eoutp9r10 ok
$run *bmdp sprint=-eoutp9r10 7=emmsfdrate par=p9r
/ problem     title is 'EOUTP9R10:  sqrtfdrt with original explanatory
                    variables--best 5 subsets regression:
                    feedrate transformed'.
/ input       unit is 7.
              cases are 524.
              variables are 10.
              format is '(3x,a4,18x,f3.0,9x,f4.0,6(1x,f3.0),1x,f4.0)'
/ variable    names are label,totfsh,date,colsze,tide,tidehi,tidemed,
                    time,numchk,agechk,sqtime,sqdate,sqage,sqcol,
                    feedrate,sqrtfdrt.
              add=6.
              label is label.
/ transform   sqtime=time*time.
              sqdate=date*date.
              sqage=agechk*agechk.
              sqcol=colsze*colsze.
              feedrate=totfsh/2.0.
              sqrtfdrt=sqrt(feedrate).
/ regress     dependent is sqrtfdrt.
              independent are date,colsze,tidehi,tidemed,time,numchk,
                              agechk,sqtime,sqdate,sqage,sqcol.
              method=rsq.
/ print       news.
              no shade.
/ plot        normal.
              yvar are residual,residual.
              xvar are predictd,sqrtfdrt.
              size=115,50.
              hist.
/ end
```

Figure 29: GLIM Command File which Generated Figure 10

```
$EMPTY -EOUTGLIM37 OK
$RUN UNSP:GLIM 1=EMMSFDRATE 2=-EOUTGLIM37
$C
$C   GLIM RUN ON FILE EMMSFDRATE TESTING OUT LOG-LINEAR MODEL
$C       LOOK AT FULL MODEL FROM EOUTGLIM16, DO SOME PLOTS, & TRY TO
$C       IMPROVE FITS BY INTERACTIONS
$C
$C ***************************
$C *  GET DATA & TRANSFORM   *
$C ***************************
$C
$OUTPUT 2 132
$UNITS 524
$DATA LABEL TOTFSH DATE COLSZE TIDE TIME NUMCHK AGECHK
$FACTOR TIDE 3
$FORMAT
(3X,F5.0,17X,F3.0,9X,F4.0,2(1X,F3.0),8X,2(1X,F3.0),1X,F4.0)
$DINPUT 1
$LOOK 1 15 LABEL TOTFSH DATE COLSZE TIDE TIME NUMCHK AGECHK
$C
$C   ADD QUADRATIC TERMS
$C
$CALC SQDATE=DATE*DATE: SQTIME=TIME*TIME: SQAGE=AGECHK*AGECHK
$CALC SQNUM=NUMCHK*NUMCHK: SQCOL=COLSZE*COLSZE
$C
$C ***********************************
$C *  SPECIFY MODEL TO BE ANALYZED   *
$C *  ALLOW FOR OVER-DISPERSION      *
$C ***********************************
$C
$YVAR TOTFSH
$ERROR P
$LINK L
$SCALE 0
$C
$C ***************************
$C *  SET UP OUTPUT ROUTINE  *
$C ***************************
$C
$OUTPUT 2 115 50
$MACRO SEERESULTS $DISP A
                  $CALC RESID=TOTFSH-%FV
                  $PLOT %FV TOTFSH
                  $PLOT RESID %FV
                  $PLDT RESID TDTFSH
                  $ENDMAC
$C
$C ***************************************************************************
$C *  NOW FIT MODELS                                                         *
$C *  FIRST NULL MODEL & ITS PLOTS                                           *
$C *  THEN TRY SOME MODELS WITH INTERACTIONS & THEIR PLOTS                   *
$C ***************************************************************************
$C
$FIT %GM
$DISP A
$FIT NUMCHK+SQDATE+DATE+TIME+SQTIME+TIDE+COLSZE+SQAGE+AGECHK+SQCOL
$USE SEERESULTS
$C
```

Figure 29, continued

```
$C  CALCULATE APPROPRIATE CROSS-PRODUCTS
$C
$CALC A1N1=AGEC*NUMC: A2N1=SQAG*NUMC: A1T1=AGEC*TIME: A1T2=AGEC*SQTI
$CALC A2T1=SQAG*TIME: A2T2=SQAG*SQTI: C1N1=COLS*NUMC: C2N1=SQCO*NUMC
$CALC C1A1=COLS*AGEC: C1A2=COLS*SQAG: C2A1=SQCO*AGEC: C2A2=SQCO*SQAG
$C
$C  NOW TRY THEM OUT
$C
$FIT NUMC+SQDA+DATE+TIME+SQTI+TIDE+COLS+SQAG+AGEC+SQCO+A1N1+A2N1
$USE SEERESULTS
$FIT NUMC+SQDA+DATE+TIME+SQTI+TIDE+COLS+SQAG+AGEC+SQCO+A1T1+A1T2+
       A2T1+A2T2
$USE SEERESULTS
$FIT NUMC+SQDA+DATE+TIME+SQTI+TIDE+COLS+SQAG+AGEC+SQCO+C1A1+C1A2+
       C2A1+C2A2
$USE SEERESULTS
$FIT NUMC+SQDA+DATE+TIME+SQTI+TIDE+COLS+SQAG+AGEC+SQCO+C1N1+C2N1
$USE SEERESULTS
$FIT NUMC+SQDA+DATE+TIME*TIDE+SQTI*TIDE+COLS+SQAG+AGEC+SQCO
$USE SEERESULTS
$C
$C   NOW TRY FOR THE WHOLE THING
$C
$FIT NUMC+SQDA+DATE+TIME*TIDE+SQTI*TIDE+COLS+SQAG+AGEC+SQCO+A1N1+A2N1+
     A1T1+A1T2+A2T1+A2T2+C1N1+C2N1+C1A1+C1A2+C2A1+C2A2
$USE SEERESULTS
$STOP
```

(Ref.(8) pg. 25). Unscaled deviance, *D(current model)*, is defined by:

$$S(\hat{\underline{\mu}},\underline{y})=\frac{D(current\ model)}{\sigma^2}$$

where $\sigma^2=1$ is used for a Poisson likelihood without over-dispersion. Clearly, then:

$$D(current\ model)\neq SS(error)$$

since:

$$SS(error)=\underline{e}'\underline{e}=(\underline{y}-\hat{\underline{\mu}})'(\underline{y}-\hat{\underline{\mu}})$$

$$=\sum_{i=1}^{n}(y_i-\hat{\mu}_i)^2$$

It should be pointed out that this time:

$$\hat{\underline{\mu}}\neq\hat{\underline{Y}}$$

where

$$\hat{\underline{Y}}=X\underline{b}$$

Nonetheless, *D(current model)* is the goodness-of-fit measure which GLIM uses.

It will be recalled, however, that *SS(error)* for the log-linear model exceeded that for the normal theory model when the same 5 explanatory variables (NUMCHK, AGECHK, TIME, SQAGE, SQTIME) were used. This was shown in Figures 11.a and 11.b. The GLIM command file which generated these figures is shown in Figure 30. An important feature of this program is the calculation of *SS(error)* (stored in the scalar '%S') using the '%CU' function. This function is designed to give a result of the same size as its argument (here a vector of 524 entries), but the assignment of this result to the

Figure 30: GLIM Command File which Generated Figures 11.a-b

```
$EMPTY EOUTGLIM39 OK
$RUN UNSP:GLIM 1=EMMSFDRATE 2=EOUTGLIM39
$C
$C    GLIM RUN ON FILE EMMSFDRATE TESTING OUT
$C      CURRENT BEST MODEL
$C      USING BOTH NORMAL THEORY
$C      AND LOG-LINEAR REGRESSION MODELS
$C
$C ************************
$C *  GET DATA & TRANSFORM  *
$C ************************
$C
$OUTPUT 2 132
$UNITS 524
$DATA LABEL TOTFSH DATE COLSZE TIDE TIME NUMCHK AGECHK
$FACTOR TIDE 3
$FORMAT
(3X,F5.0,17X,F3.0,9X,F4.0,2(1X,F3.0),8X,2(1X,F3.0),1X,F4.0)
$DINPUT 1
$LOOK 1 15 LABEL TOTFSH DATE COLSZE TIDE TIME NUMCHK AGECHK
$CALC FEEDRATE=TOTFSH/2.0
$C
$C    ADD QUADRATIC TERMS
$C
$CALC SQDATE=DATE*DATE: SQTIME=TIME*TIME: SQAGE=AGECHK*AGECHK
$CALC SQNUM=NUMCHK*NUMCHK: SQCOL=COLSZE*COLSZE
$C
$C ********************************
$C *  SET UP MODEL ANALYSIS ROUTINE  *
$C *  FIT NULL MODEL                 *
$C *  FIT CURRENT BEST MODEL         *
$C *  AND CALCULATE RESIDUAL SUM OF  *
$C *  SQUARES                        *
$C ********************************
$C
$MACRO FIT $FIT %GM
           $DISP A
           $FIT NUMCHK+AGECHK+TIME+SQAGE+SQTIME
           $ACCURACY 9
           $DISP A
           $CALC RESID=%YV-%FV
           $CALC SQRESID=RESID*RESID
           $CALC %S=%CU(SQRESID)
           $LOOK %S
           $ENDMAC
$C
$C ********************************
$C *  NORMAL THEORY MODEL           *
$C *  SPECIFY MODEL TO BE ANALYZED  *
$C *  ALLOW FOR OVER-DISPERSION     *
$C ********************************
$C
$YVAR FEEDRATE
$ERROR N
$LINK I
$SCALE 0
$C
$C    NOW GET MODEL FIT RESULTS
```

Figure 30, continued

```
$C
$USE FIT
$C
$C **********************************
$C *  LOG-LINEAR MODEL                *
$C *  SPECIFY MODEL TO BE ANALYZED    *
$C *  ALLOW FOR OVER-DISPERSION       *
$C **********************************
$C
$YVAR TOTFSH
$ERROR P
$LINK L
$SCALE O
$C
$C   NOW GET MODEL FIT RESULTS
$C
$USE FIT
$STOP
```

scalar '%S' has the effect of '%S' taking on the last entry in that result vector. Consult Section 10.2 of Ref.(2) for details.

CHAPTER 10

TECHNICAL SUPPLEMENT FOR CHAPTER 5

This discussion is concerned with the estimates and associated standard errors mentioned in Section 5.4. First, some extended notation is needed:

$$\Delta t = t_D - t_M = t_D - 27$$

$$\Delta t^2 = t_D^2 - t_M^2 = t_D^2 - 729$$

a_D=age of chicks at dawn (t_D)

a_M=age of chicks at t_M

$$\Delta a = a_D - a_M$$

$$\Delta a^2 = a_D^2 - a_M^2$$

so:

$$\hat{y}_D = b_0 + b_1(\text{NUMCHK}) + b_2 a_D + b_3 t_D + b_4 a_D^2 + b_5 t_D^2$$

$$\hat{y}_M = b_0 + b_1(\text{NUMCHK}) + b_2 a_M + b_3 t_M + b_4 a_M^2 + b_5 t_M^2$$

$$\hat{y}_D - \hat{y}_M = d$$

$$= b_2 \Delta a + b_3 \Delta t + b_4 \Delta a^2 + b_5 \Delta t^2$$

$$= \underline{c}' \underline{b}$$

where:

$$\underline{c}' = (0, 0, \Delta a, \Delta t, \Delta a^2, \Delta t^2)$$

$$\underline{b}' = (b_0, b_1, b_2, b_3, b_4, b_5)$$

Furthermore, d is a 'best' (minimum variance) linear unbiased estimator of:

$$\delta = E(Y \mid t_D) - E(Y \mid t_M = 27)$$

by Gauss' Theorem (Ref.(6) pg. 301).

Now:

$$\Delta a^2 = a_D^2 - a_M^2$$
$$= (a_D - a_M)(a_D + a_M)$$
$$= \Delta a(2a_D + \Delta a)$$

Suppose dawn occurs at 0530H, that is,

$$t_D = 11$$

Then:

$$\Delta a = -\frac{1}{3}$$

$$\Delta t = -16$$

$$\Delta t^2 = -608$$

$$d = \hat{y}_D - \hat{y}_M$$
$$= -\frac{1}{3}b_2 - 16b_3 - \frac{1}{3}b_4\left(2a_D - \frac{1}{3}\right) - 608b_5$$

From the current best model:

$$\underline{b}' = (0.873, 0.243, 0.545, -0.0845, -0.00121, 0.00156)$$

so the only unknown left is a_D. Rather than estimate d separately for each a_D, suppose that Δa, being 0 to the nearest whole day, can be ignored. Then:

$$d \approx \underline{k}'\underline{b}$$

where

$$\underline{k}' = (0, 0, 0, \Delta t, 0, \Delta t^2)$$
$$= (0, 0, 0, -16, 0, -608)$$

so

$$d \approx -16(-0.0845) - 608(0.00156)$$
$$= 0.404$$

Now in general, if:

$$\underline{b} = (X'X)^{-1}X'\underline{Y}$$

then, according to Ref.(6) pg. 300:

$$E(\underline{b})=\underline{\beta}$$

$$Cov(\underline{b})=(X'X)^{-1}\sigma^2$$

so that:

$$\underline{Y}\sim N_n(\underline{0},\sigma^2 I_n)$$

gives:

$$\underline{b}\sim N_{p+1}(\underline{\beta},(X'X)^{-1}\sigma^2)$$

provided that the matrix X was full rank (Ref.(13) pg. 28), which was ensured in this analysis by removing all extrinsic and intrinsic aliasing (see Section 4.3 of Ref.(2)).

Furthermore:

$$\underline{k}'\underline{b}\sim N(\underline{k}'\underline{\beta},\underline{k}'(X'X)^{-1}\underline{k}\sigma^2)$$

$$\underline{c}'\underline{b}\sim N(\underline{c}'\underline{\beta},\underline{c}'(X'X)^{-1}\underline{c}\sigma^2)$$

(Ref.(13) pg. 28). Thus a 95% confidence interval for $\underline{k}'\underline{\beta}$ would be:

$$(\underline{k}'\underline{b}-z(0.975)\sigma\sqrt{\underline{k}'(X'X)^{-1}\underline{k}},\underline{k}'\underline{b}+z(0.975)\sigma\sqrt{\underline{k}'(X'X)^{-1}\underline{k}})$$

where $z(\gamma)$ represents the $100\gamma\%$ percentage point of a $N(0,1)$ distribution. As usual, however, σ^2 is unavailable and must be estimated by s^2 where:

$$s^2=\frac{SS(error)}{n-p-1}$$

which will require a t-distribution with $n-p-1$ degreees of freeedom in the consequent confidence interval:

$$(\underline{k}'\underline{b}-t(0.975;n-p-1)s\sqrt{\underline{k}'(X'X)^{-1}\underline{k}},$$

$$\underline{k}'\underline{b}+t(0.975;n-p-1)s\sqrt{\underline{k}'(X'X)^{-1}\underline{k}})$$

where now $t(\gamma;\nu)$ represents the $100\gamma\%$ percentage point of a t-distribution with ν degrees of freedom.

Now:

$$s^2\underline{k}'(X'X)^{-1}\underline{k}=\underline{k}'(\hat{Cov}(\underline{b}))\underline{k}$$

where $\hat{Cov}(\underline{b})$ is just the sample dispersion or variance-covariance matrix for $\underline{b}$ shown in Figure 10. Also, one can take advantage of the zeroes in $\underline{k}$ to obtain the simplification:

$$\underline{k}'(\hat{Cov}(\underline{b}))\underline{k}$$

$$=(\Delta t,\Delta t^2)\begin{pmatrix} s^2(b_3) & \hat{Cov}(b_3,b_5) \\ \hat{Cov}(b_3,b_5) & s^2(b_5)\end{pmatrix}\begin{pmatrix}\Delta t \\ \Delta t^2\end{pmatrix}$$

$$=(-16,-608)\begin{pmatrix} 0.413297*10^{-3} & -0.814613*10^{-5} \\ -0.814613*10^{-5} & 0.164207*10^{-6}\end{pmatrix}\begin{pmatrix}-16 \\ -608\end{pmatrix}$$

$$=8.014*10^{-3}$$

so the sample standard error of $\underline{k}'\underline{b}$ is:

$$s(\underline{k}'\underline{b})=\sqrt{s^2\underline{k}'(X'X)^{-1}\underline{k}}$$

$$=\sqrt{\underline{k}'(\hat{Cov}(\underline{b}))\underline{k}}$$

$$=\sqrt{8.014*10^{-3}}$$

$$=8.952*10^{-2}$$

But again, this is a standard error for $\underline{k}'\underline{b}$, not $d=\underline{c}'\underline{b}$, which would be a function of a_D.

Note that the $\hat{Cov}(\underline{b})$ matrix does not come with the P9R output by default but must be requested in the '/ print' paragraph (see the corresponding command file, Figure 21). Note also that this would be one way to get a scaled $(X'X)^{-1}$ matrix for possible future calculation requirements (such as Working-Hotelling confidence bands); the entries of $\hat{Cov}(\underline{b})$ are just those of $(X'X)^{-1}$ multiplied by:

$$s^2=\frac{SS(error)}{n-p-1}=0.288215$$

which is just the residual mean square from Figure 4.

Unforunately as can be seen from Figure 10 the entire $\hat{Cov}(\underline{b})$ matrix is not given. The row and column for the constant term are missing, which explains why the matrix given there is 5 by 5 rather than 6 by 6 as it should be. As will be seen in Part B (Figure 32), however, the full $\hat{Cov}(\underline{b})$ matrix may be obtained in a GLIM run, in this case on a normal theory regression model.

PART B

MODELLING ACTIVE NEST OCCURENCE IN DECIDUOUS TREES BY PRIMARY EXCAVATOR BIRD SPECIES

CHAPTER 1

THE PROBLEM

Dagmar Gook, graduate student in the Dept. of Biological Sciences, presented data collected in a study of trees in the B.C. interior, near Kamloops. The site at which she studied forms part of what is known as the Interior Douglas-fir Biogeoclimatic Zone. Such zones are used by biologists to classify areas in which trees grow. A tree was examined for presence or absence of an active nest, that is, a nest which was excavated in the tree itself and was currently being inhabited.

It was of interest to investigate whether the presence of such a nest in a tree (to be hereafter referred to as a 'success') was associated with any of the other characteristics measured for that tree. Once such a set of associated characteristics could be found, a ranking of their relative 'importance' would then be of interest. The results of such a study would be valuable to forestry and wildlife managers who may wish to encourage or discourage such activity.

The analysis was to be carried out specifically only for deciduous (non-evergreen) trees and for those species of birds who are 'primary excavators', that is, those birds who will dig their own nests and not use an already existing but vacant cavity. Those species will be identified in the next

chapter.

The results of this analysis are to be applied to the population of deciduous trees in all of the Interior Douglas-fir Biogeoclamatic Zones found in B.C., which are basically restricted to the low altitude regions of the southern interior. As it turned out a final model (to be identified in Chapter 3) was obtained using stepwise logistic regression. This model used up all the explanatory variables with which it was provided. It also contains no interaction terms. The 2 quantitative variables, length and height of tree, were entered into the analysis in natural log (base '*e*') scale. For the qualitative variables, strict 0,1-coding was used, where the first level of any such variable forced all applicable design variables to 0.

In logistic regression, the quantity being modelled is not directly probability of success, but is instead log odds of success. An increase in log odds, however, will lead to an increase in probability. In this sense, nesting was positively associated with height and diameter, so trees which were taller or thicker (or both) had a higher nesting rate than trees which were respectively shorter or thinner (or both). Similarly of the 3 tree species classes (aspen, birch, 'other'), aspen trees had higher nesting rates than birch, which in turn had higher rates than 'other' (all other quantities being equal). Nesting was further positively associated with fungal conks, scars, and broken

tops. Also, dead trees showed higher rates than live ones having similar characteristics (other than live/dead status). Details are given in Chapter 3.

As for the ranking of the variables in importance to the final model, it was proposed that the order of entry in the stepwise process would be the most useful to resource management personnel, but other rankings were also attempted. In all of them, presence or absence of fungal conks was clearly the most 'valuable' to the final model. Details are given in Chapter 4.

Regarding the sample data as a representative random sample from the target population mentioned above, an example inference was done for an aspen tree in the sample, where the fitted log odds value was found to be:

$$\hat{\eta}=1.5006$$

with an estimated sample variance of:

$$s^2(\hat{\eta})=\hat{var}(\hat{\eta})=0.0566$$

This value of $\hat{\eta}$ then leads to an estimated probability:

$$\hat{\mu}=0.8177$$

which estimates the probability of success for all trees in the target population which have similar characteristics to the example one selected here. By using $s^2(\hat{\eta})$ as well, one could further obtain a prediction probability:

$$\mu_p=0.8176$$

which predicts the probability of success for an individual tree within the population. This μ_p is of course subject to

a higher degree of imprecision than $\hat{\mu}$, which is why μ_p is closer to 0.5 than $\hat{\mu}$. A value of μ_p would reflect a toss of a fair coin as a guess of success or failure for a single tree. Details are found in Chapter 5.

The client report is found in Chapters 1-5 , and the technical supplements for each chapter in Chapters 6-9 of this part, with Chapter 6 providing the technical supplement for Chapter 2, and so on (no technical supplement was needed for Chapter 1).

CHAPTER 2

THE DATA

Dagmar permitted access to a data file which contained all of the characteristics measured on the trees along with a record of success (active nest present) or failure (no active nest). A portion of this file is shown in Table 8. For each observation (tree) in the file, there are 2 consecutive records, the formats of which are as follows.

	Variable	Column Range
RECORD 1		
(1)	Nest Tree Number/Plot Number	1-3
(2)	Blank or Tree Number	5-7
(3)	Tree Species	9-10
(4)	Diameter of Tree	12-15
(5)	Height of Tree	17-21
(6)	Tree Live or Dead Indicator	23
(7)	Decay Type 1 Indicator	25-26
(8)	Decay Type 2 Indicator	28-29
(9)	Decay Type 3 Indicator	31-32
(10)	Decay Type 4 Indicator	34-35
RECORD 2		
(11)	Presence/Absence of Active Nest (Success/Failure)	9-10
(12)	Birdspecies	12
(13)	Live or Dead Wood in Tree around Nest Indicator	30
(14)	Broken Top of Tree Indicator	32

Table 8: Portion of Raw Data File

```
  1    BI 24.0 12.95 L 37    43 44 U          1N            0
  1    NT N  0  S   9.36 48 D                               0
136    A  39.5 24.57 L 37 39 43    U  M   H?   1H            0
136    NT S  0  S  15.29 48                                  0
  2    BI 28.0  2.63 D 37 39    44 F  M   ?                 0
  2    NT B  F      2.12 48 D                                0
  4    A  38.3 18.43 L 37    43    U          1#S     1#S?   0
  4    NT S  F     12.21 48                                  0
  5    A  40.6 23.25 L 37 39 43    U  S   ?   1H?           0
  5    NT P  0  P? 12.62 16     A                            0
  6    A  38.3  7.68 D                                       0
  6    NT P  F      7.10 40 D                                0
  7    BI 24.4  4.71 D 37       44 F                         0
  7    NT F  0  F?  4.39 32 D 1                              0
  8    BI 20.5  8.45 D    39    44    DM  H?S                0
  8    NT Q  0  H?  8.21 40 D 1                              0
  9    BI 25.0 11.93 L 37 39 43 44 I          1S?           0
  9    NT N  F     11.21 24 D 1                              0
137    BI 30.3 13.73 D 37 39    44 IF S   ?   1S 1?   1#S?   0
137    NT ?F F     10.29 64 D 1                              0
 10    BI 16.3  6.70 D 37       44 F  D   S                  0
 10    NT B  F?     5.89 16 D 1                              0
 11    BI 26.8 12.53 L 37    43 44 F  DS  ?S? 1S 1?         0
 11    NT N  F     11.83 32 D 1                              0
 12    BI 26.0 11.56 L    39 43 44          1N?             0
 12    NT S  F      9.19 32 D                                0
 13    A  22.6 12.03 L       43 44    BS             3?     0
 13    NT N  ON N?  8.98 64 D                                0
 14    A  32.2 17.70 L    39 43                4S            0
 14    NT S  D  S   8.20 48                                  0
 15    BI 20.0  9.30 D 37    43 44 F  D   S                 0
 15    NT M  0  M?  5.54 48 D 1                              0
 16    A  29.2 16.35 L 37    43    I                         0
 16    NT H  F      9.11 24                                  0
 17    A  20.5  5.25 D 37 39    44 I  S   ?   1?            0
 17    NT M  F      3.50 40 D 1                              0
 18    A  60.0 23.42 L 37    43    I  BLS H?P 5S 1P         0
 18    NT S  F     13.44  8                                  0
 19    A  30.5 27.17 L 37    43    I  D   S   1S      1#?   0
 19    NT S  F      9.14  8                                  0
 24    A  35.5 18.81 L 37    43    I          2S 1H?        0
 24    NT P  F      9.61 40                                  0
 20    A  20.1 20.41 D 37          I  D   S   2S            0
 20    NT S  F      9.10 32 D                                0
 21    A  28.7  8.86 D    39    44    BSL ??P               0
 21    NT N  F  N?  5.39 16 D 1                              0
 22    A  20.9 12.32 L       43       S?                    0
 22    NT B  0? ?  11.39  8                                  0
 23    A  28.6 20.24 L 37    43    I          1S?           0
 23    NT S  F      8.61 32                                  0
 25    A  42.4 22.57 L 37    43 44 I  S   ?   3P 1S?        0
 25    NT P  F      7.79  8                                  0
 28    A  60.8 23.12 L 37    43    I                         0
 28    NT S  F      9.50 48                                  0
 26    A  28.1  5.35 D          44    SOL ?NP 2S?           0
 26    NT B  ?      4.39 40 D 1                              0
 27    A  23.1 13.62 L 37    43    I          1H?           0
 27    NT S  F      4.39 40                                  0
 30    BI 29.8 11.32 L 37 39 43 44 U  D   S                 0
 30    NT S  F      9.44 56 D 1                              0
 31    BI 35.0  8.55 L 37 39 43 44 UP         1B?           0
 31    NT B? F      7.83 64 D 1                              0
 34    BI 30.5  8.80 L 37    43 44 F          1S            0
 34    NT S  F      6.91 16 D                                0
 32    A  27.7 14.02 D          44                           0
 32    NT S  F     12.29 48 D 1                              0
 33    BI 16.2  3.15 D 37 39    44 F                         0
 33    NT B? 0      2.79 24 D 1                              0
 35    A  44.0 21.16 L 37    43    I  M   ?        2S       0
 35    NT F  0  P  11.59 48     A                            0
 35    A  44.0 21.16 L 37    43    I  M   ?        2S       0
 35    NT P  0     11.59 48     A                            0
```

Other data in the file was ignored for this analysis. The variables as given above are now described in more detail. A ∅ will be used to denote a blank character. The names of the variables to be retained in the analysis will be capitalized hereafter.

A) Response Variable, ACNEST (RECORD 2, column 9-10)

This is the outcome (success/failure) variable. As found in the raw data file, it takes on the following values:

NT: Nest Tree--active nest present
∅ : no active nest present

In order to make these values readable by all anticipated statistical software, they were recoded as follows

ACNEST= 1, if active nest present
0, if no active nest present

In the case of abandoned nests, which had been previously established by a primary excavator birdspecies, it was agreed to consider such a tree as a failure since the reason for the nest's abandoned state may be that the tree acquired a new characteristic which it did not have before the nest was established. This new characteristic may have led to the nest's abandonment.

B) Description of Candidate Explanatory Variables

B.1)Tree Species (RECORD 1, col. 9-10)

This qualitative variable may take on the following codes:

A -Aspen
BI-Birch
CT-Cottonwood
W -Willow

D -Alder

F -Douglas Fir
S -Spruce
PY-Ponderosa Pine
J -Juniper

Only the first 5 tree species were to be kept in the analysis since the remainig species were coniferous (evergreen). A blank line was used in the above list to separate visually the deciduous from the coniferous trees. For the species to be retained in the analysis, the following recoding was done:

SPTREE= 1, for tree code A (aspen)
2, for tree code BI (birch)
3, otherwise (codes CT, W, D)

B.2)Diameter of Tree (RECORD 1, col. 12-15)

This is measured in centimetres, and will be hereafter referred to as DIAM.

B.3)Height of Tree (RECORD 1, col.17-21)

This is measured in metres, and will be hereafter referred to as HEIGHT.

B.4)Decay Type 1 Indicator (RECORD 1, col.25-26)

This qualitative variable takes on the following codes:

37: fungal conk present
ø : no fungal conk

For the analysis, this was recoded as:

DI1= 1, if fungal conk is present
0, otherwise

B.5)Decay Type 2 Indicator (RECORD 1, col. 28-29)

This qualitative variable takes on the following codes:

39: scars present on tree
ø : no scars on tree

For the analysis, this was recoded as:

DI2= 1, if scars present on tree
0, otherwise

B.6)Decay Indicator Type 3 (RECORD 1, col. 31-32)

This qualitative variable takes on the following codes:

43: dead branches present
ø : no dead branches

For the analysis, this was recoded as:

DI3= 1, if dead branches present
0, otherwise

B.7)Deadwood

This variable did not appear on the data files, but at Dagmar's request was computed as a function of:

Tree Live or Dead Indicator (RECORD 1, col. 23)
Decay Type 4 Indicator (RECORD 1, col. 34-35)

The variable 'Tree Live or Dead Indicator' takes on the following values:

L: tree is alive
D: tree is dead

and 'Decay Type 4 Indicator' may take on the following values:

44: tree top either broken or dead
ø : tree top both intact and alive (full top)

The variable 'Deadwood', to be herafter referred to as DWOOD, was then created as follows:

Tree Live or Dead Indicator	Decay Type 4 Indicator	DWOOD
L	ø	1
L	44	2
D	ø	No Value
D	44	2

No value is assigned to DWOOD in the 3rd case above since 'D' and 'ø' is an illegal combination; a dead tree cannot have a live top. Thus:

DWOOD= 1, if tree is completely alive and has an intact top
2, otherwise

B.8)Broken Top (RECORD 2, col.32)

This qualitative variable is used to specifically detect cases of trees with broken tops, regardless of whether those trees are alive or not. It takes on the values:

1: top of tree broken
ø: top of tree intact (not broken)

For the analysis, this was recoded as:

BKTOP= 1, if top of tree is broken
0, otherwise

C) Further Case Selection Variable

The raw data file contains cases of both primary and secondary bird species. The variable 'Bird Species' (RECORD 2, col. 9-10) takes on the following values:

S-yellow bellied sapsucker
P-pileated woodpecker
N-red breasted nuthatch

F-northern flicker
H-hairy woodpecker
D-downy woodpecker
?-one of the above, possibly H but definitely not S or P; unconfirmed in any case
ø-necessary code for trees with no active nest, hence no bird species to classify

B-black capped chickadee
G-golden eye
K-American kestrel
M-mountain chickadee
Q-flying squirrel
R-red squirrel
T-tree swallow
W-white breasted nuthatch

The last 8 bird species in the above list are secondary excavators, and thus records containing such species were not to be included in the analysis. Again, a blank line was used to separate cases to be used in analysis from those which were to be rejected. For this analysis the remaining acceptable bird species were recoded:

SPBIRD= 1, for bird species code S
2, for bird species code P
3, for bird species code N
4, for bird species code F
5, for bird species code H
6, for bird species code D
7, for bird species code ø
8, for bird species code ?

Once the variables of interest were identified, the raw data file was edited by a FORTRAN program (shown in the Technical Supplement) in order to:

(a) remove observations which contained unwanted cases of SPTREE or SPBIRD

(b) remove duplicated records (since raw data file itself was a merger of 2 previous files), which were all cases of success after record 564 of the input file

(c) flag (but not remove) cases of '?' for SPBIRD so that such cases may be confirmed

(d) create and assign values to new variable DWOOD

(e) perform all recodings indicated thus far, since not all statistical software packages (e.g. BMDP) can accept alphabetic input for variables other than labels

(f) put all acceptable cases (including flagged ones in (c) above) into a new file containing variables selected for analysis and a coded tag to identify it

The tag referred to in (f) above is the line number from the raw data file in which a particular case began. Since every case required 2 consecutive records of data, all the tags are therefore odd numbers. This tag will be hereafter referred to as RECNUM.

Both the cases which were to be removed (excluded from the analysis) and those which were flagged in (c) above (but kept in the analysis) had their reasons for being singled out put into a 'reject' file. Table 9 shows a portion of the file GOOKNEST5, which contained the edited records which were to be kept in the analysis. The data values are given in the order:

RECNUM
ACNEST
SPTREE
HEIGHT
DIAM
DI1
DI2
DI3
DWOOD
BKTOP
SPBIRD

Table 9: Portion of GOOKNEST5

1.0	1.	2.	12.95	24.00	1.	0.	1.	2.	0.	3.
3.0	1.	1.	24.57	39.50	1.	1.	1.	1.	0.	1.
7.0	1.	1.	18.43	38.30	1.	0.	1.	1.	0.	1.
9.0	1.	1.	23.25	40.60	1.	1.	1.	1.	0.	2.
11.0	1.	1.	7.68	38.30	0.	0.	0.	2.	0.	2.
13.0	1.	2.	4.71	24.40	1.	0.	0.	2.	1.	4.
17.0	1.	2.	11.93	25.00	1.	1.	1.	2.	1.	3.
19.0	1.	2.	13.73	30.30	1.	1.	0.	2.	1.	8.
23.0	1.	2.	12.53	26.80	1.	0.	1.	2.	1.	3.
25.0	1.	2.	11.56	26.00	0.	1.	1.	2.	0.	1.
27.0	1.	1.	12.03	22.60	0.	0.	1.	2.	0.	3.
29.0	1.	1.	17.70	32.20	0.	1.	1.	1.	0.	1.
33.0	1.	1.	16.35	29.20	1.	0.	1.	1.	0.	5.
37.0	1.	1.	23.42	60.00	1.	0.	1.	1.	0.	1.
39.0	1.	1.	27.17	30.50	1.	0.	1.	1.	0.	1.
41.0	1.	1.	18.81	35.50	1.	0.	1.	1.	0.	2.
43.0	1.	1.	20.41	20.10	1.	0.	0.	2.	0.	1.
45.0	1.	1.	8.86	28.70	0.	1.	0.	2.	1.	3.
49.0	1.	1.	20.24	28.60	1.	0.	1.	1.	0.	1.
51.0	1.	1.	22.57	42.40	1.	0.	1.	2.	0.	2.
53.0	1.	1.	23.12	60.80	1.	0.	1.	1.	0.	1.
57.0	1.	1.	13.62	23.10	1.	0.	1.	1.	0.	1.
59.0	1.	2.	11.32	29.80	1.	1.	1.	2.	1.	1.
63.0	1.	2.	8.80	30.50	1.	0.	1.	2.	0.	1.
65.0	1.	1.	14.02	27.70	0.	0.	0.	2.	1.	1.
69.0	1.	1.	21.16	44.00	1.	0.	1.	1.	0.	4.
71.0	1.	1.	21.16	44.00	1.	0.	1.	1.	0.	2.
73.0	1.	1.	25.79	44.00	0.	1.	1.	1.	0.	2.
79.0	1.	1.	24.91	36.10	1.	1.	1.	1.	0.	1.
81.0	1.	1.	20.64	34.00	0.	1.	1.	1.	0.	1.
83.0	1.	1.	21.91	22.70	1.	1.	0.	2.	1.	6.
85.0	1.	1.	24.96	42.80	1.	1.	1.	1.	0.	1.
87.0	1.	1.	6.06	17.40	0.	1.	1.	2.	0.	5.
89.0	1.	1.	14.29	27.90	1.	0.	1.	1.	0.	1.
91.0	1.	1.	8.78	30.00	0.	0.	1.	2.	1.	1.
93.0	1.	1.	21.16	44.00	1.	0.	1.	1.	0.	1.
99.0	1.	1.	19.84	59.30	1.	0.	1.	2.	0.	1.
103.0	1.	1.	23.40	53.40	1.	0.	1.	1.	0.	2.
105.0	1.	1.	16.90	25.20	1.	1.	1.	2.	0.	1.
107.0	1.	1.	9.18	27.00	0.	1.	0.	2.	1.	3.
109.0	1.	2.	13.50	40.50	1.	0.	1.	2.	0.	1.
111.0	1.	1.	21.76	38.30	1.	1.	1.	1.	0.	1.
113.0	1.	1.	20.84	37.90	0.	1.	1.	1.	0.	8.
115.0	1.	1.	21.81	24.60	1.	0.	1.	1.	0.	1.
117.0	1.	1.	14.44	36.00	1.	0.	1.	1.	0.	1.
119.0	1.	1.	22.14	37.90	1.	0.	1.	1.	0.	2.
121.0	1.	2.	5.28	18.80	1.	0.	1.	2.	1.	3.
123.0	1.	1.	14.25	37.10	0.	0.	1.	1.	0.	1.
125.0	1.	3.	10.65	31.40	1.	0.	1.	2.	1.	6.
127.0	1.	3.	7.28	21.70	1.	1.	1.	2.	1.	3.
131.0	1.	1.	17.63	26.60	1.	1.	1.	1.	0.	1.
133.0	1.	1.	24.25	27.70	0.	1.	1.	1.	0.	1.
135.0	1.	1.	25.80	44.60	0.	1.	1.	1.	0.	2.
137.0	1.	1.	15.62	30.50	1.	1.	1.	1.	0.	1.
139.0	1.	1.	17.26	34.70	0.	1.	1.	1.	0.	1.
141.0	1.	1.	20.32	38.90	0.	1.	1.	1.	0.	1.
143.0	1.	2.	9.42	33.80	0.	1.	1.	2.	1.	3.
145.0	1.	1.	21.00	36.10	0.	0.	1.	1.	0.	1.
147.0	1.	1.	19.65	34.80	1.	0.	1.	1.	0.	1.
151.0	1.	1.	16.00	26.30	1.	0.	1.	1.	0.	8.
153.0	1.	1.	16.52	34.80	1.	0.	1.	1.	0.	1.
155.0	1.	2.	3.69	20.50	1.	0.	1.	2.	1.	4.
157.0	1.	2.	8.09	27.50	1.	0.	0.	2.	1.	1.
159.0	1.	1.	21.47	24.30	1.	0.	0.	1.	0.	1.
161.0	1.	1.	5.91	39.10	1.	1.	0.	2.	1.	1.
163.0	1.	1.	21.89	34.00	1.	0.	1.	1.	0.	1.
165.0	1.	1.	20.09	25.70	0.	1.	1.	1.	0.	8.
167.0	1.	1.	16.86	24.40	1.	1.	1.	1.	0.	1.
169.0	1.	1.	21.82	34.10	1.	1.	1.	1.	0.	1.
171.0	1.	1.	22.68	39.00	1.	0.	1.	1.	0.	1.
173.0	1.	1.	11.58	48.70	0.	0.	0.	2.	1.	4.

Table 10 shows a portion of the file GOOKREJECT5 which contains both the reasons why certain cases were kept out of GOOOKNEST5 (and hence further analysis) and the messages concerning the flagged observations mentioned earlier. Out of 1275 observations in the original raw data file, 1124 were put into GOOKNEST5 (including the 10 flagged for the '?' value for SPBIRD) and the remaining 151 rejected from further analysis.

Having performed the needed file editing, the remaining data (as in the GOOKNEST5 file) was now ready for analysis.

Table 10: Portion of GOOKREJECT5

```
RECORD      5.0 CONTAINS SECONDARY EXCAVATOR BIRD SPECIES B

RECORD     15.0 CONTAINS SECONDARY EXCAVATOR BIRD SPECIES Q

       ***CHECK RECORD     19.0 ***
       FOR BIRD SPECIES ?
       RECORD NOT SKIPPED

RECORD     21.0 CONTAINS SECONDARY EXCAVATOR BIRD SPECIES B

RECORD     31.0 CONTAINS SECONDARY EXCAVATOR BIRD SPECIES M

RECORD     35.0 CONTAINS SECONDARY EXCAVATOR BIRD SPECIES M

RECORD     47.0 CONTAINS SECONDARY EXCAVATOR BIRD SPECIES B

RECORD     55.0 CONTAINS SECONDARY EXCAVATOR BIRD SPECIES B

RECORD     61.0 CONTAINS SECONDARY EXCAVATOR BIRD SPECIES B

RECORD     67.0 CONTAINS SECONDARY EXCAVATOR BIRD SPECIES B

RECORD     75.0 CONTAINS SECONDARY EXCAVATOR BIRD SPECIES R

RECORD     77.0 CONTAINS SECONDARY EXCAVATOR BIRD SPECIES Q

RECORD     95.0 CONTAINS SECONDARY EXCAVATOR BIRD SPECIES B

RECORD     97.0 CONTAINS SECONDARY EXCAVATOR BIRD SPECIES Q

RECORD    101.0 CONTAINS SECONDARY EXCAVATOR BIRD SPECIES K

       ***CHECK RECORD    113.0 ***
       FOR BIRD SPECIES ?
       RECORD NOT SKIPPED

RECORD    129.0 CONTAINS SECONDARY EXCAVATOR BIRD SPECIES B

RECORD    149.0 CONTAINS SECONDARY EXCAVATOR BIRD SPECIES M

       ***CHECK RECORD    151.0 ***
       FOR BIRD SPECIES ?
       RECORD NOT SKIPPED

       ***CHECK RECORD    165.0 ***
       FOR BIRD SPECIES ?
       RECORD NOT SKIPPED

RECORD    175.0 CONTAINS SECONDARY EXCAVATOR BIRD SPECIES B

       ***CHECK RECORD    181.0 ***
       FOR BIRD SPECIES ?
       RECORD NOT SKIPPED

RECORD    185.0 CONTAINS SECONDARY EXCAVATOR BIRD SPECIES Q

RECORD    213.0 CONTAINS SECONDARY EXCAVATOR BIRD SPECIES T

       ***CHECK RECORD    217.0 ***
       FOR BIRD SPECIES ?
       RECORD NOT SKIPPED

RECORD    223.0 CONTAINS SECONDARY EXCAVATOR BIRD SPECIES K

RECORD    227.0 CONTAINS SECONDARY EXCAVATOR BIRD SPECIES W

RECORD    229.0 CONTAINS SECONDARY EXCAVATOR BIRD SPECIES M

RECORD    237.0 CONTAINS SECONDARY EXCAVATDR BIRD SPECIES T
```

CHAPTER 3

FIRST ANALYSIS--FIND INFLUENTIAL VARIABLES

This study gave binary responses (1-success, 0-failure) so logistic regression was considered appropriate. Letting:

$$\mu_i = Prob\left\{\begin{array}{l}\text{Deciduous tree } i \text{ has active nest with} \\ \text{primary bird species}\end{array}\right\}$$

that is, μ_i is the probability of a success for tree i, a linear model was used to explain the 'log odds':

$$\eta_i = g(\mu_i) = ln\left(\frac{\mu_i}{1-\mu_i}\right)$$

for $0<\mu_i<1$. g is known as a 'link' function because it will 'link' the linear regression model:

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

with the original quantity of interest, μ.

Also, in order to control data spread, HEIGHT and DIAM were logarithmically transformed using the natural logarithm (base e) befor being included in the linear model. This transformation was saved for the logistic regression program

The stepwise logistic regression program, PLR, from the BMDP software library was used (Ref. (4), Section 14.5). A protion of the final results is shown in Figure 31. It will be noticed that the variable DI3 was left out of the analysis. Dagmar requested this since she later believed the variable to be 'biologically unsound', that is, not worthy of inclusion. This request was granted since earlier analyses (not shown here) indicated that it was a variable

Figure 31: Results of PLR Run on GOOKNEST5 Data

TERM		COEFFICIENT	STANDARD ERROR	COEFF/S.E.
sptree	(1)	-2.9036	0.3729	-7.787
	(2)	-3.0541	0.9105	-3.354
di1		3.7217	0.2471	15.06
di2		1.1396	0.2226	5.121
dwood		0.93930	0.3280	2.864
bktop		1.3251	0.4205	3.151
lnheight		0.51375	0.3066	1.676
lndiam		1.7341	0.4396	3.945
CONSTANT		-10.039	1.481	-6.779

COVARIANCE MATRIX OF COEFFICIENTS

	sptre(1)	sptre(2)	di1	di2	dwood
sptre(1)	0.13905				
sptre(2)	0.04935	0.82897			
di1	-0.03761	-0.02440	0.06105		
di2	-0.01484	-0.01429	0.01254	0.04953	
dwood	-0.01827	-0.01082	0.00988	0.00271	-0.10760
bktop	-0.04355	-0.04328	0.00993	0.00669	-0.06058
lnheight	-0.00438	-0.00608	0.00477	0.00191	0.02017
lndiam	0.01098	-0.02309	0.01006	-0.00674	0.01372
CONSTANT	-0.01064	0.10471	-0.07489	-0.00725	-0.12929

	bktop	lnheight	lndiam	CONSTANT
sptre(1)				
sptre(2)				
di1				
di2				
dwood				
bktop	0.17681			
lnheight	0.05833	0.09400		
lndiam	-0.02414	-0.04951	0.19326	
CONSTANT	-0.09419	-0.11007	-0.52416	2.19295

of little importance anyway.

The final model from the PLR run is as follows:

$$\hat{\eta}=-10.039-2.9036[\text{SPTREE}(1)]-3.0541[\text{SPTREE}(2)]$$
$$+0.51375[ln(\text{HEIGHT})]+1.7341[ln(\text{DIAM})]$$
$$+3.7217[\text{DI1}]+1.1396[\text{DI2}]+0.9390[\text{DWOOD}]+1.3251[\text{BKTOP}]$$

where:

(1) $\hat{\eta}$ is the fitted log odds value. One could obtain a fitted probability by then using the inverse of the link function:

$$\hat{\mu}=g^{-1}(\hat{\eta})=(1+e^{-\hat{\eta}})^{-1}$$

but one may wish to use alternative methods to getting a μ value from η (see Chapter 5).

(2) The subscript i which tags individual observations or cases has been left off of the above fitted model (and the equation given in note (1) above) for simplicity, but is understood to be present on $\hat{\eta}$ and all explanatory variables.

(3) New design variables for SPTREE were created by PLR as follows:

SPTREE(1)= 1, if tree is birch (SPTREE=2)
0, otherwise

SPTREE(2)= 1, if tree is not aspen or birch (SPTREE=3)
0, otherwise

The effect of aspen trees (SPTREE=1) is already absorbed in the constant term, -10.039. These design variables were created because SPTREE is a qualitative variable with more than 2 possible factor levels. In general, a qualitative variable possessing k possible factor levels will give rise to k-1 design variables (Ref. (11) Section 10.1).

(4) The explanatory variable DWOOD was recoded as follows:

DWOOD= 1, if tree top dead or broken, or entire tree dead
0, if tree fully alive with intact top

This was done because PLR run was done with the 0,1-coding option.

(5) All other explanatory variables (HEIGHT, DIAM, DI1, DI2, BKTOP) are as previously discussed in Chapter 2.

Thus for a single deciduous tree one can observe values for all the explanatory variables shown in the model and then calculate a fitted value, $\hat{\eta}$, for the log odds of finding a success. One can then use this fitted value as an estimate of the mean of the distribution of the log odds value for all future trees that have the same values for the explanatory variables (at least as far as measurement accuracy will allow for HEIGHT and DIAM). This use of $\hat{\eta}$ will be hereafter referred to as '$\hat{\eta}$ as estimate'. Alternatively one could use the fitted log odds value as a prediction of the log odds value just for an individual tree, given the values of the explanatory variables for that tree. This is a different use of $\hat{\eta}$ and will be hereafter referred to as '$\hat{\eta}$ as prediction'. In this analysis, however, each use will produce a different 'fitted' μ (probability of success). This will be discussed later in Chapter 5, and in more detail in the Technical Supplement, Section 9.2.

As a further interpretation of the final fitted model, one could view it as a fitting for aspen trees:

$$\hat{\eta}=-10.039+0.51375[ln(\text{HEIGHT})]+1.7341[ln(\text{DIAM})]$$

with further penalties/awards as follows:

Deduct 2.9036 if tree is not aspen, but birch
3.0541 if tree is not aspen or birch,but some other kind of deciduous tree

Add 3.7217 if fungal conks are present

1.1396 if scars are present
0.9393 if tree is dead or has dead/broken top
1.3251 if tree has broken top but is still alive
(Note that the above 0.9393 will still be added too.)

The model obtained from the PLR run was confirmed by use of another software package, GLIM (Ref. (2)), the output of which is highlighted in Figure 32. Here different design variables for SPTREE:

SPTR(i)= 1, if SPTREE=i
0, otherwise

The GLIM run also provided data with which to construct some interesting plots with the P6D program from BMDP (Ref. (4) Section 10.2).

Figures 33.a through 33.d, for example, shows plots of fitted log odds against ln(DIAM) for all 3 values of SPTREE according to the following scheme:

'a' denotes 1 or more overlapping points for aspen trees (SPTREE=1)

'b' denotes 1 or more overlapping points for birch trees (SPTREE=2)

'o' denotes 1 or more overlapping points for 'other' deciduous trees, that is, trees not aspen or birch (SPTREE=3)

'*' denotes 1 or more overlapping points for different species of tree

The first three plots, Figures 33.a-c, are done for each SPTREE value separately. The last plot, Figure 33.d, is done for all SPTREE values. One can see definite clustering tendencies for aspen and birch trees in different areas of

Figure 32: Results of GLIM Run to Confirm those of PLR Run

```
      SCALED
CYCLE DEVIANCE      DF
   4    1174.      1123

      ESTIMATE      S.E.      PARAMETER
   1  -1.288      0.7228E-01  %GM
 SCALE PARAMETER TAKEN AS     1.000

      SCALED
CYCLE DEVIANCE      DF
   5    597.1      1115

      ESTIMATE      S.E.      PARAMETER
   1  -10.04        1.469     %GM
   0    ZERO      ALIASED     SPTR(1)
   2  -2.902        0.3676    SPTR(2)
   3  -3.053        0.8979    SPTR(3)
   4   1.734        0.4365    LNDI
   5  0.5136        0.3051    LNHE
   0    ZERO      ALIASED     DI1(1)
   6   3.721        0.2445    DI1(2)
   0    ZERO      ALIASED     DI2(1)
   7   1.139        0.2211    DI2(2)
   0    ZERO      ALIASED     DWOO(1)
   8  0.9389        0.3260    DWOO(2)
   0    ZERO      ALIASED     BKTO(1)
   9   1.325        0.4180    BKTO(2)
 SCALE PARAMETER TAKEN AS     1.000
```

```
(CO)VARIANCE MATRIX
  1   2.157
  2 -1.4686E-02  0.1352
  3  0.1024       4.6911E-02  0.8061
  4 -0.5156       1.1514E-02 -2.3175E-02  0.1906
  5 -0.1086      -3.9941E-03 -5.5639E-03 -4.9092E-02  9.3059E-02
  6 -7.1567E-02 -3.5880E-02 -2.2782E-02  9.4339E-03  4.6425E-03  5.9762E-02
  7 -6.4204E-03 -1.4214E-02 -1.3586E-02 -6.7614E-03  1.8297E-03  1.2079E-02  4.8877E-02
  8 -0.1267      -1.7474E-02 -1.0346E-02  1.3291E-02  1.9983E-02  9.2932E-03  2.6254E-03
  9 -9.1189E-02 -4.2195E-02 -4.1512E-02 -2.4216E-02  5.7547E-02  9.4029E-03  6.2937E-03
       1            2           3           4           5           6           7

  1
  2
  3
  4
  5
  6
  7
  8   0.1063
  9  -6.0148E-02  0.1748
        8           9
```

Figure 33.a: $\hat{\eta}$ against ln(DIAM) for Aspen Trees

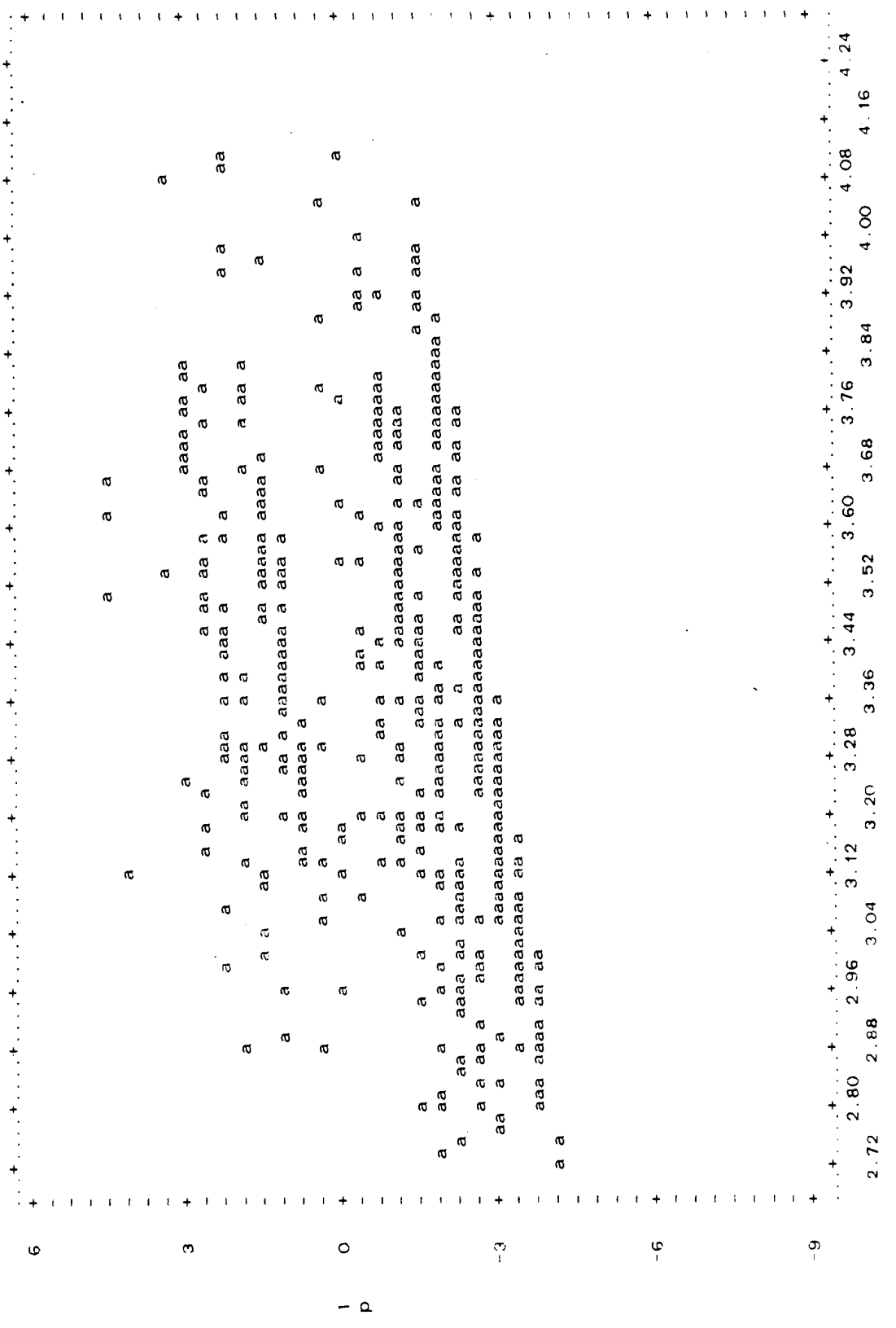


Figure 33.b: $\hat{\eta}$ against ln(DIAM) for Birch Trees

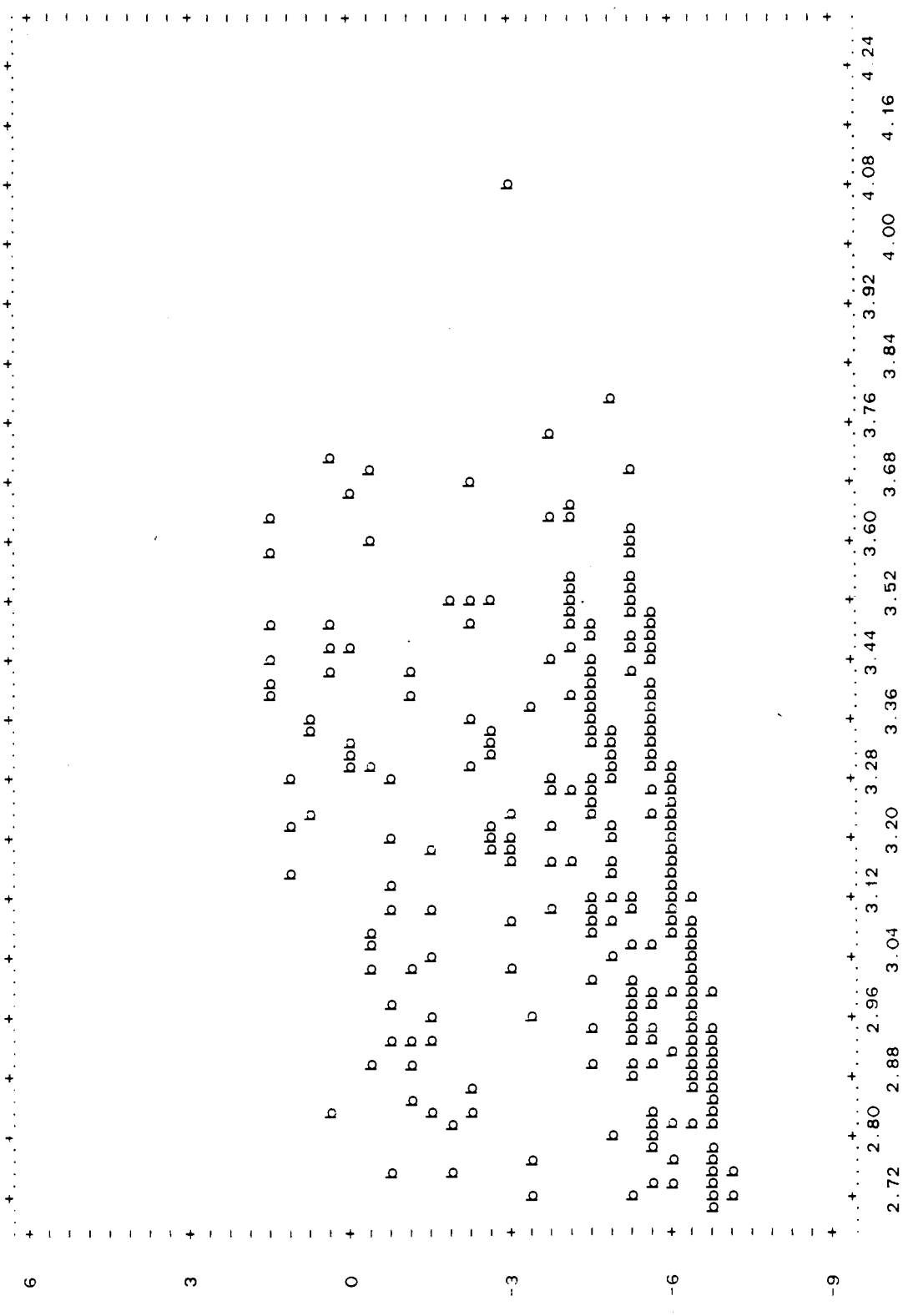


Figure 33.c: $\hat{\eta}$ against ln(DIAM) for Other Deciduous Trees

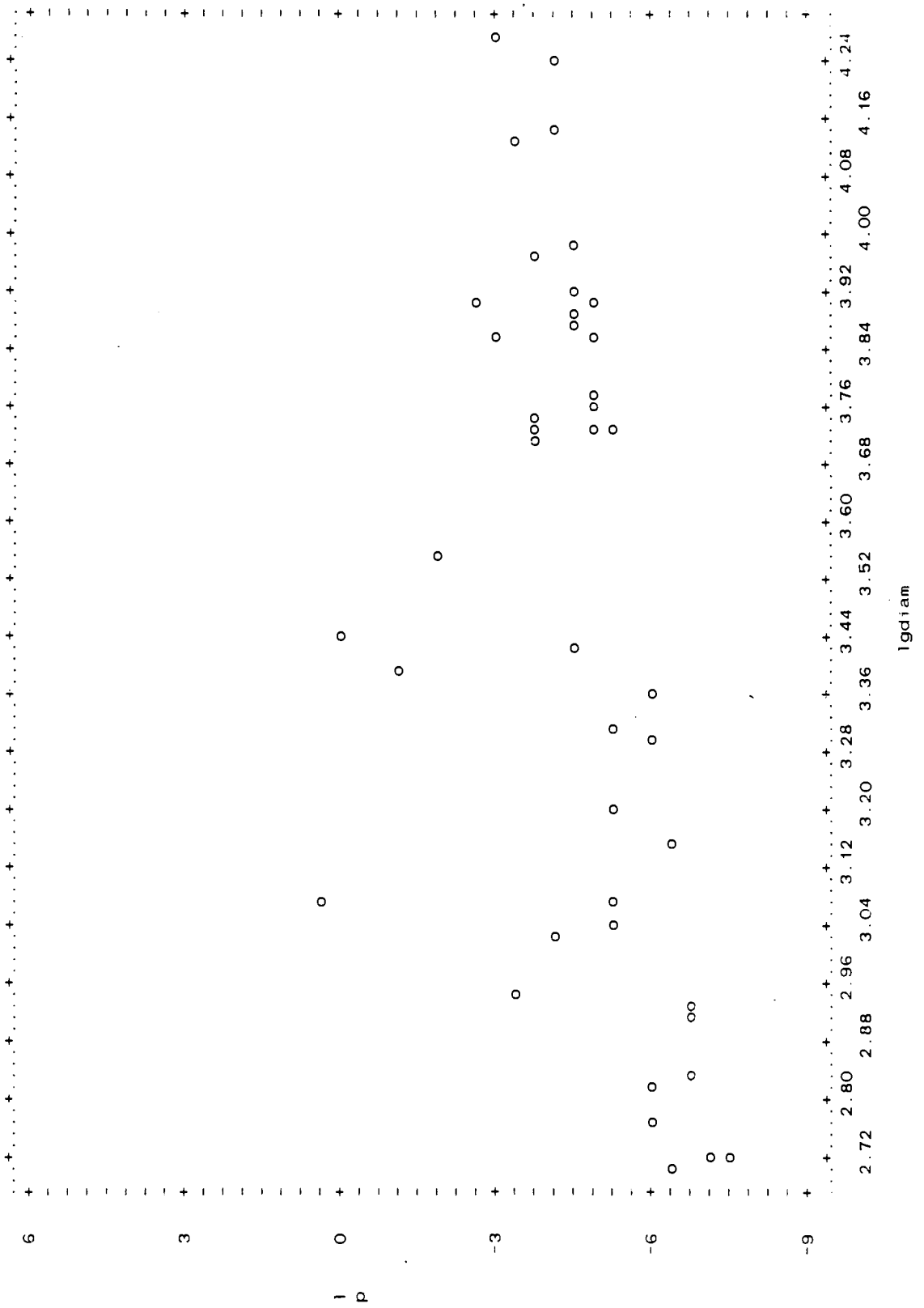


Figure 33.d: $\hat{p}$ against ln(DIAM) for All Deciduous Trees

the 4th plot. Furthermore these clusters seem to be taking place about parallel lines which go up as ln(DIAM) increases. This suggests a lack of interaction between SPTREE and ln(DIAM), so in the final model there would be little additional benefit for adding such an interaction.

These plots were later broken down by DI1 level. Figures 34.a through 34.d show the plots of Figures 33.a-d but now only for those trees for which DI1=0 (no fungal conks). Similarly, Figures 35.a through 35.d has the same plots but now only for those trees for whom DI1=1 (fungal conks present). It is interesting to note that for each value of SPTREE, the presence or absence of fungal conks separates the clusters in the plots of Figures 33.a-d into lower portions in Figures 34.a-d and upper portions in Figures 35.a-d. This suggest that the separation effects due to SPTREE and DI1 as estimated in the PLR run are strong indeed (more on this in the next chapter).

In regards to the earlier remark about no visual evidence of interaction between SPTREE and ln(DIAM), some other interactions were attempted in GLIM runs to see if their presence could significantly improve model fit. None of the 3 possible 2-way interactions between SPTREE, DI1, and ln(DIAM) could do so (the details are not shown). The search for significant interactions was restricted at first to these 3 variables on the basis that they seemed to offer the strongest associations with the log odds, η (see next

Figure 34.a: $\hat{\eta}$ against ln(DIAM) for Aspen Trees without Fungal Conks

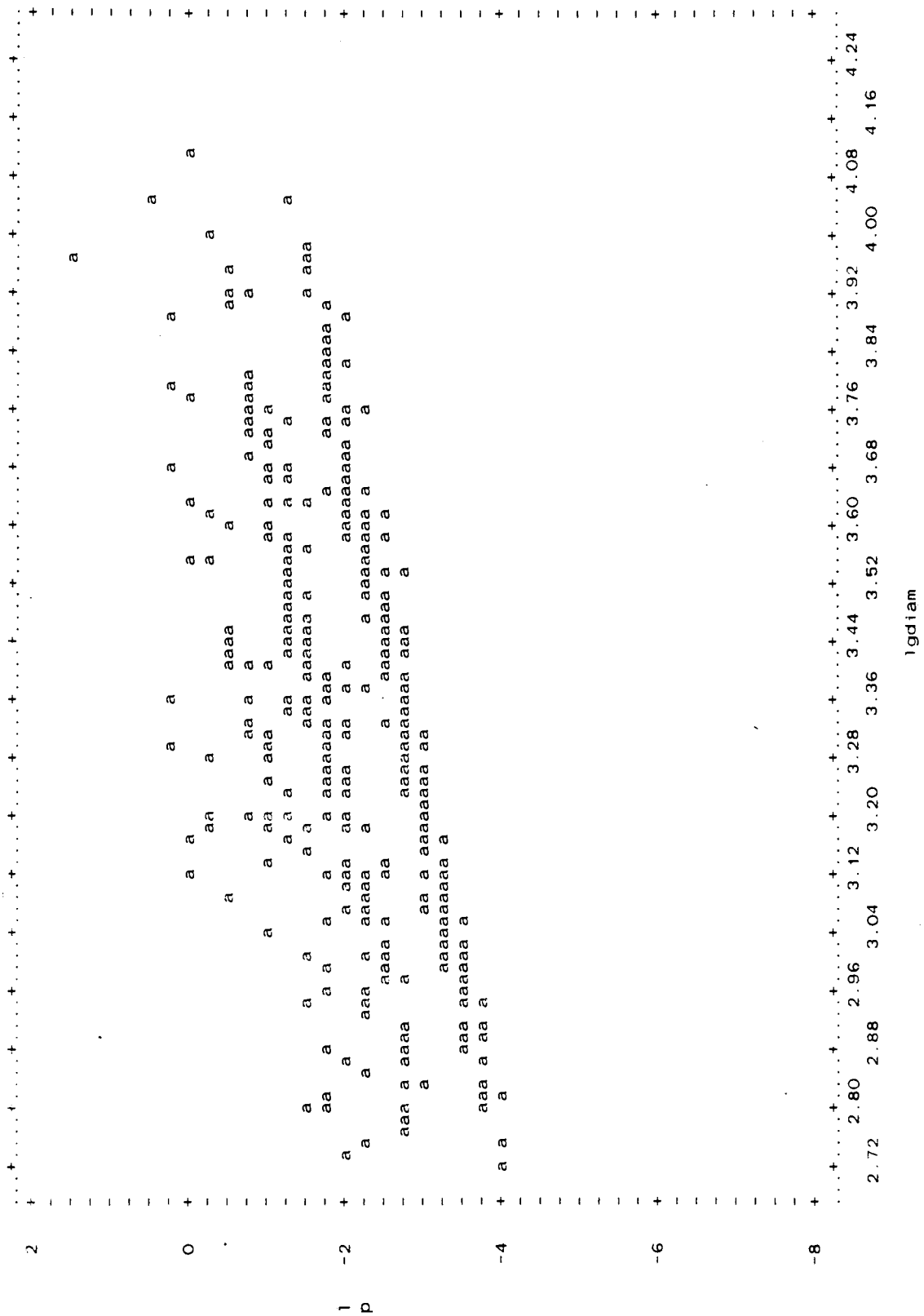


Figure 34.b: $\hat{\eta}$ against ln(DIAM) for Birch Trees without Fungal Conks

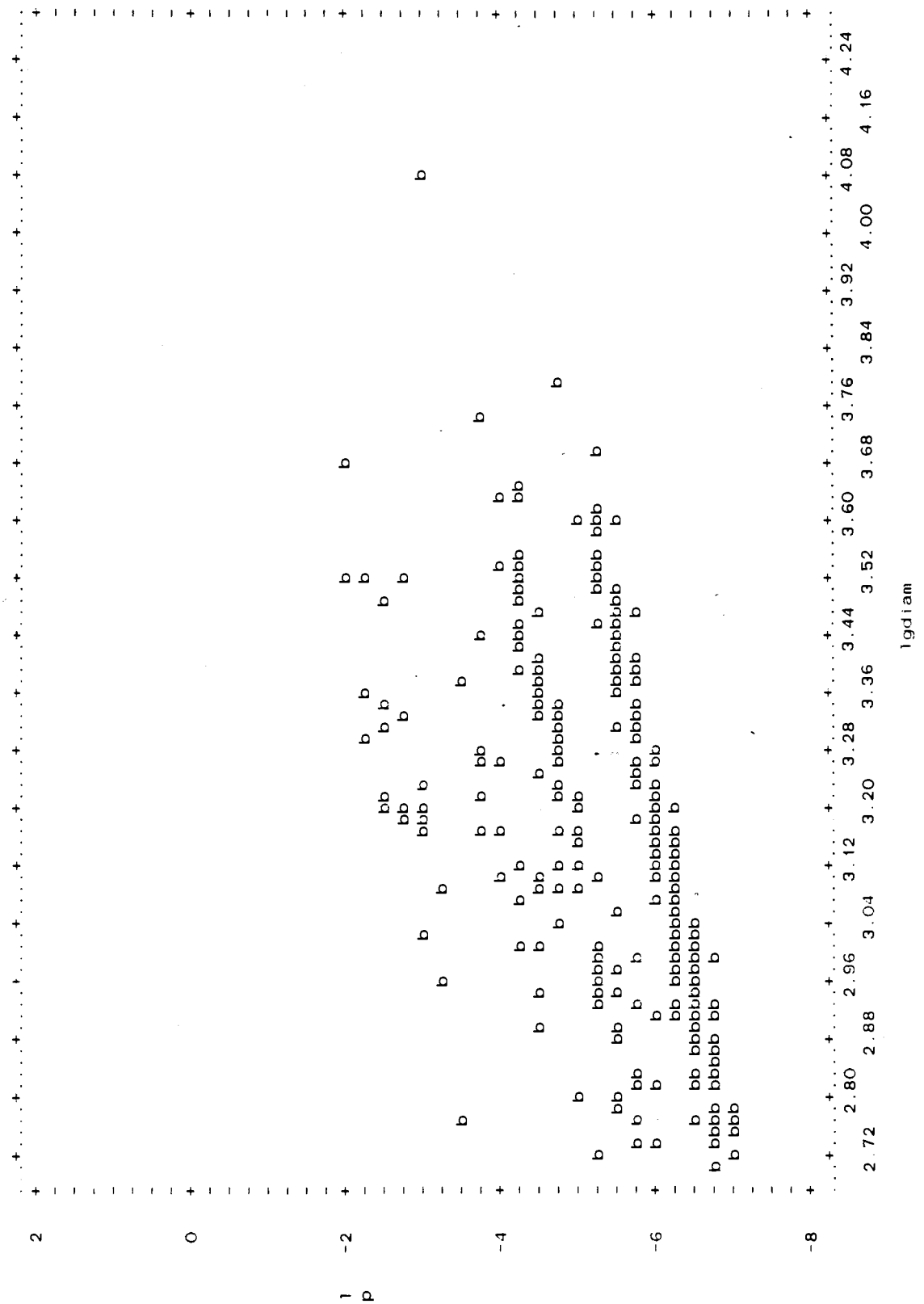


Figure 34.c: $\hat{\eta}$ against ln(DIAM) for Other Deciduous Trees without Fungal Conks

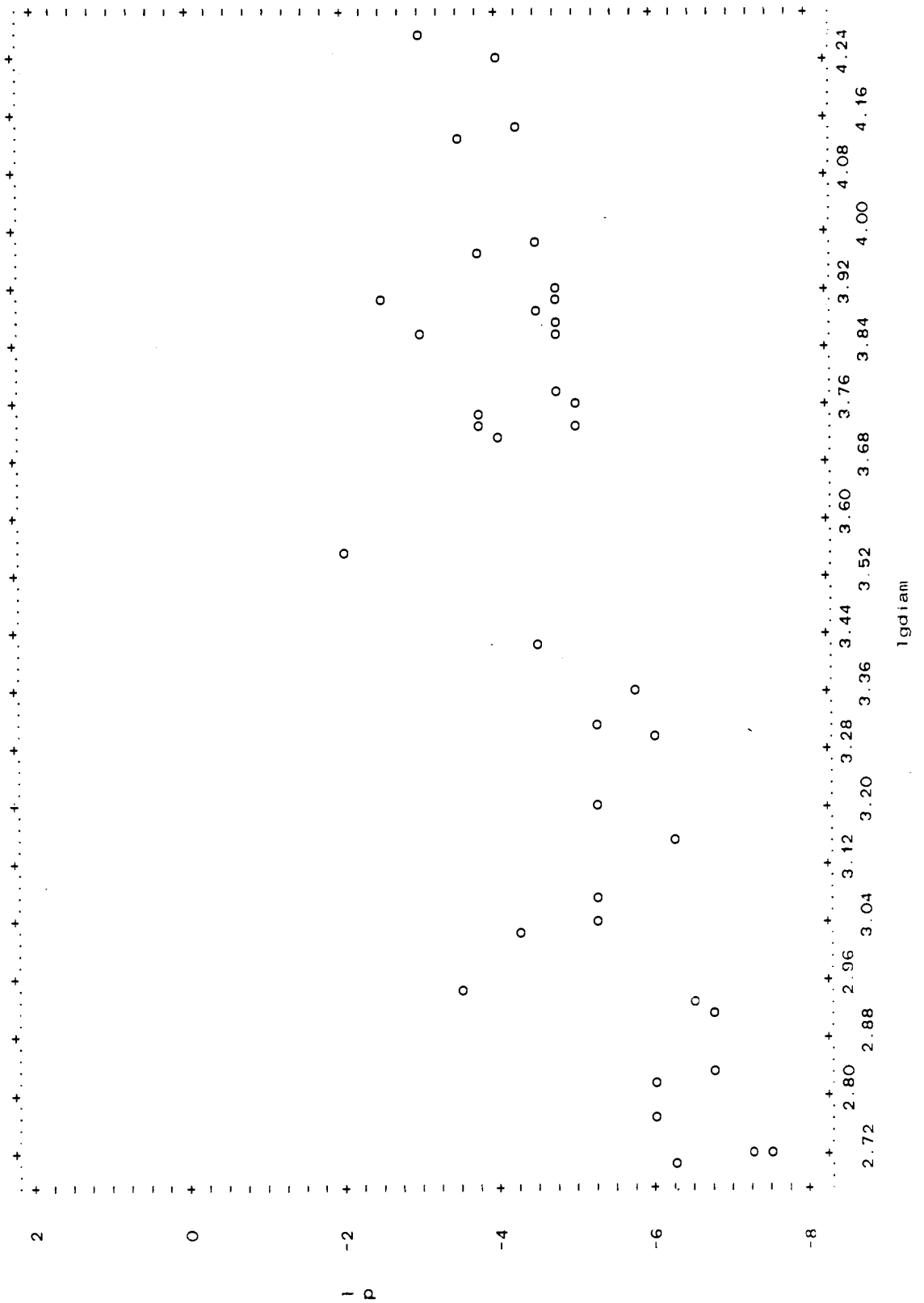


Figure 34.d: $\hat{\eta}$ against ln(DIAM) for All Deciduous Trees without Fungal Conks

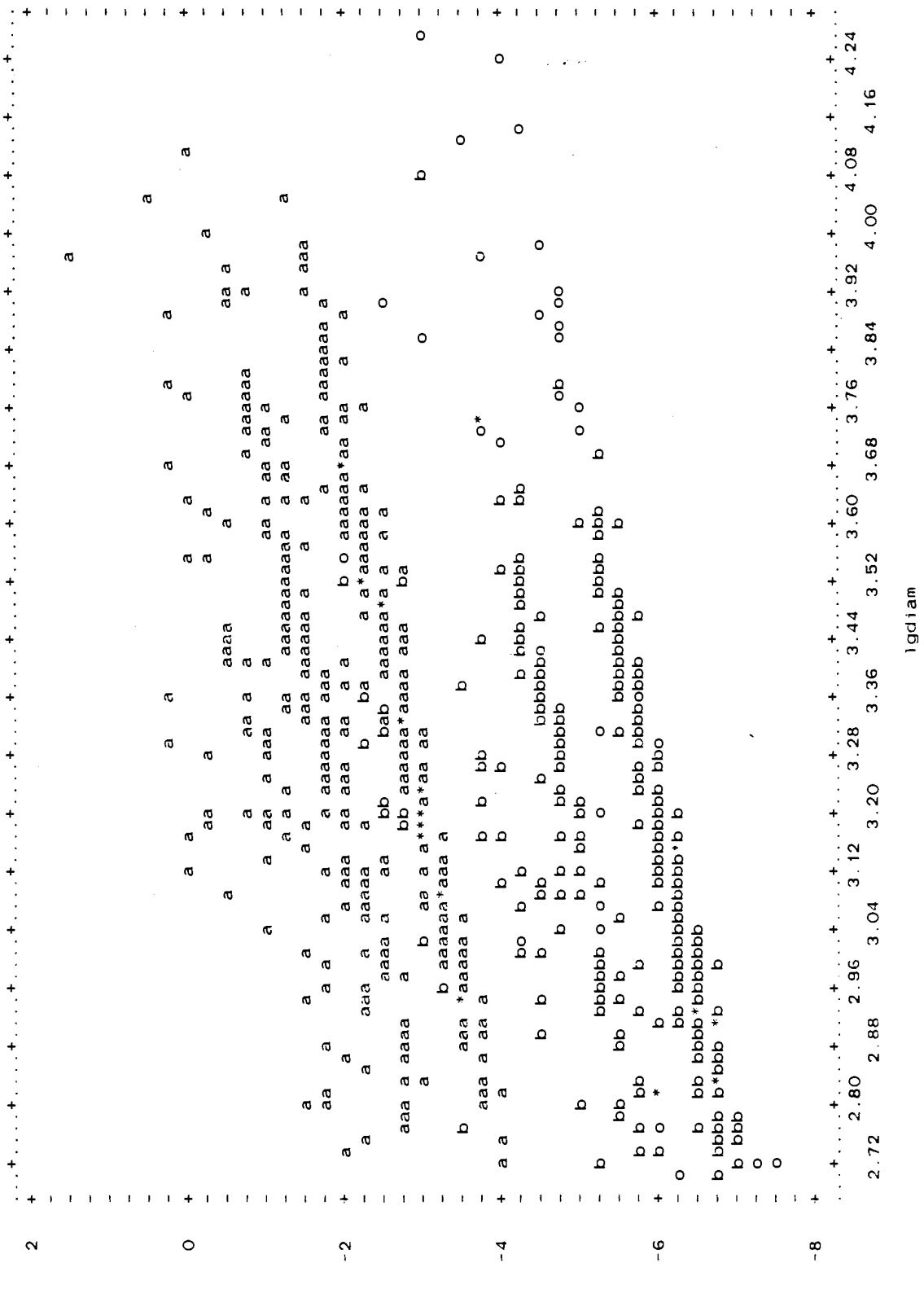


Figure 35.a: $\hat{\eta}$ against ln(DIAM) for Aspen Trees with Fungal Conks

Figure 35.b: $\hat{\eta}$ against ln(DIAM) for Birch Trees with Fungal Conks

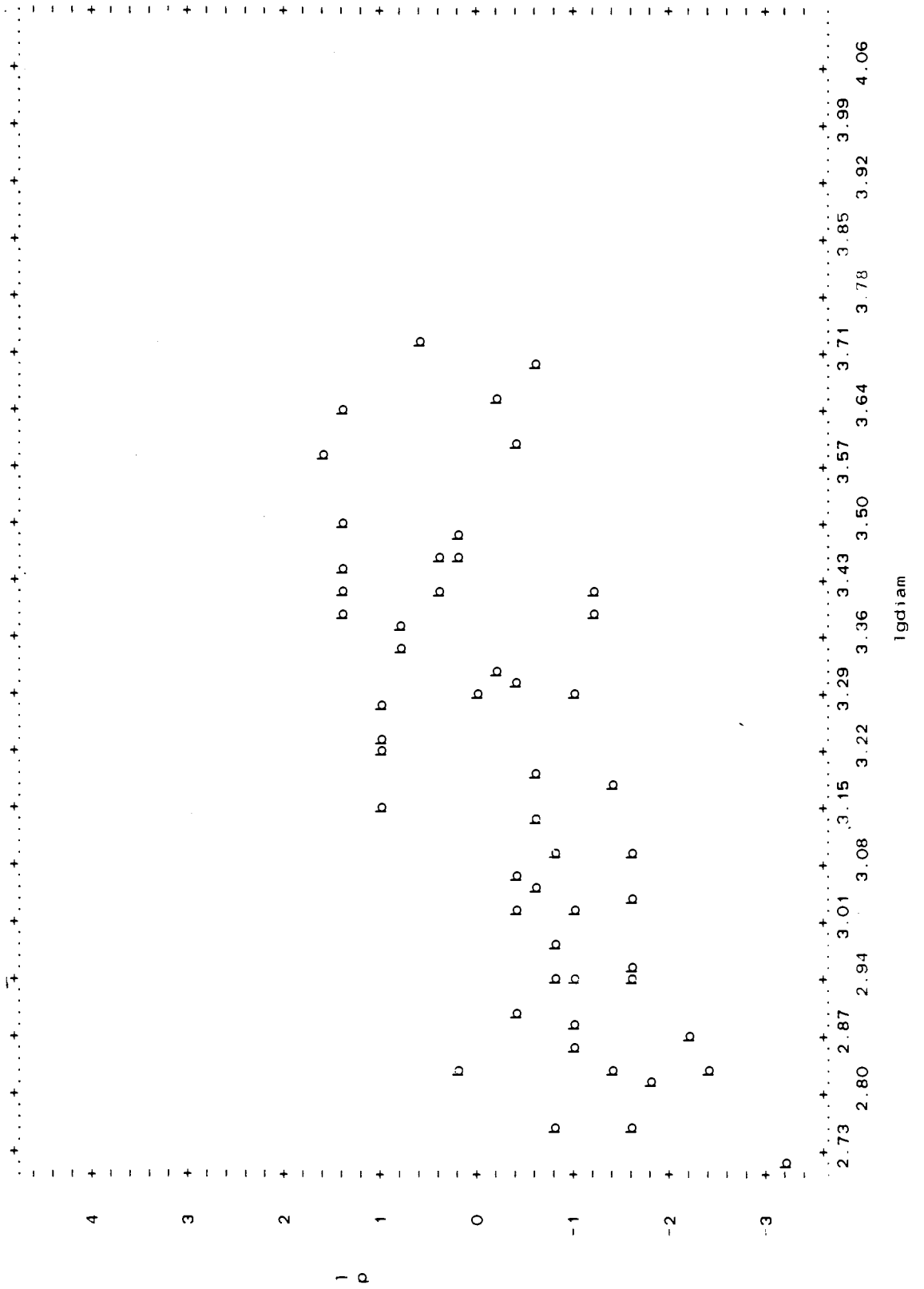


Figure 35.c: $\hat{\eta}$ against ln(DIAM) for Other Deciduous Trees with Fungal Conks

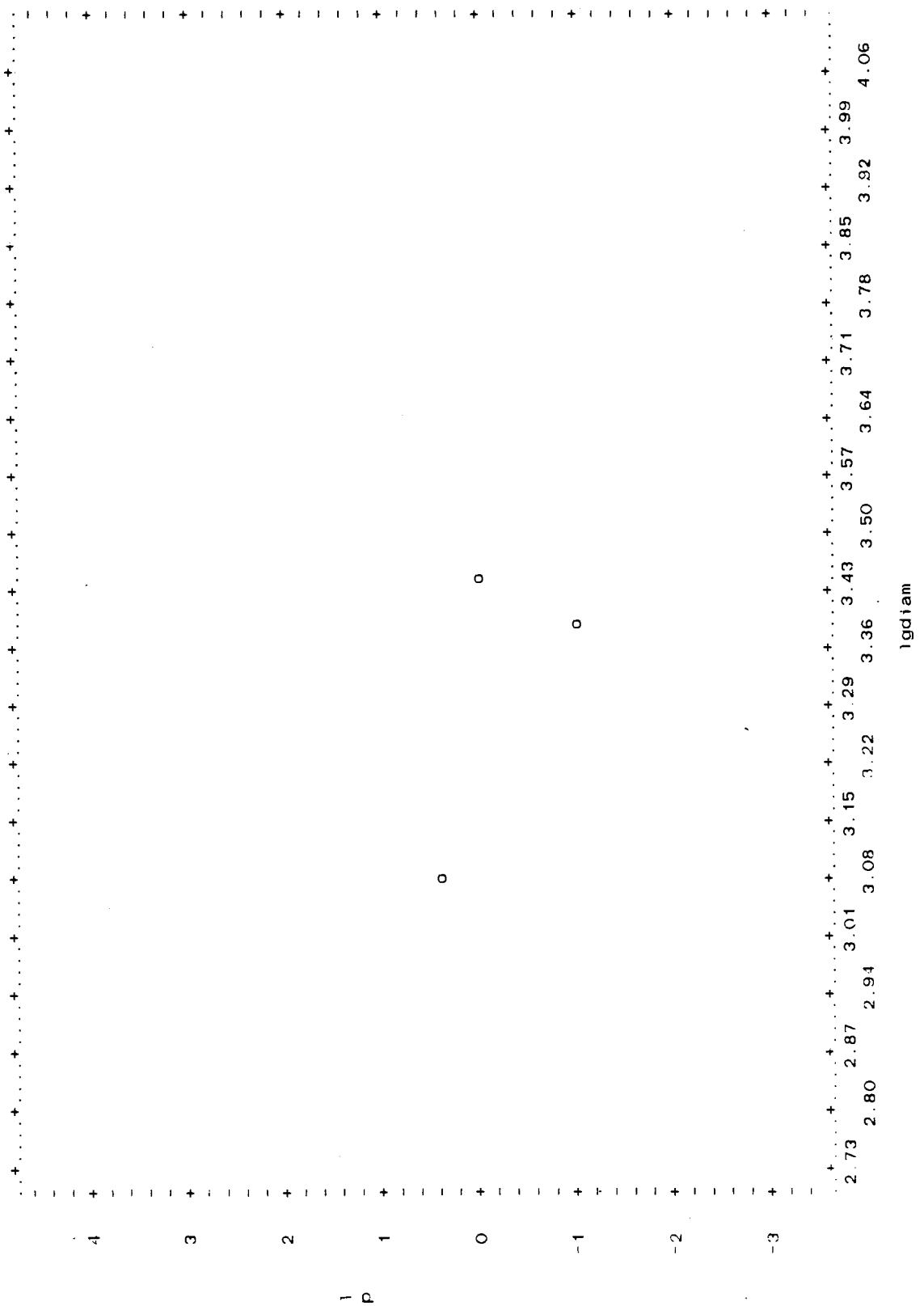
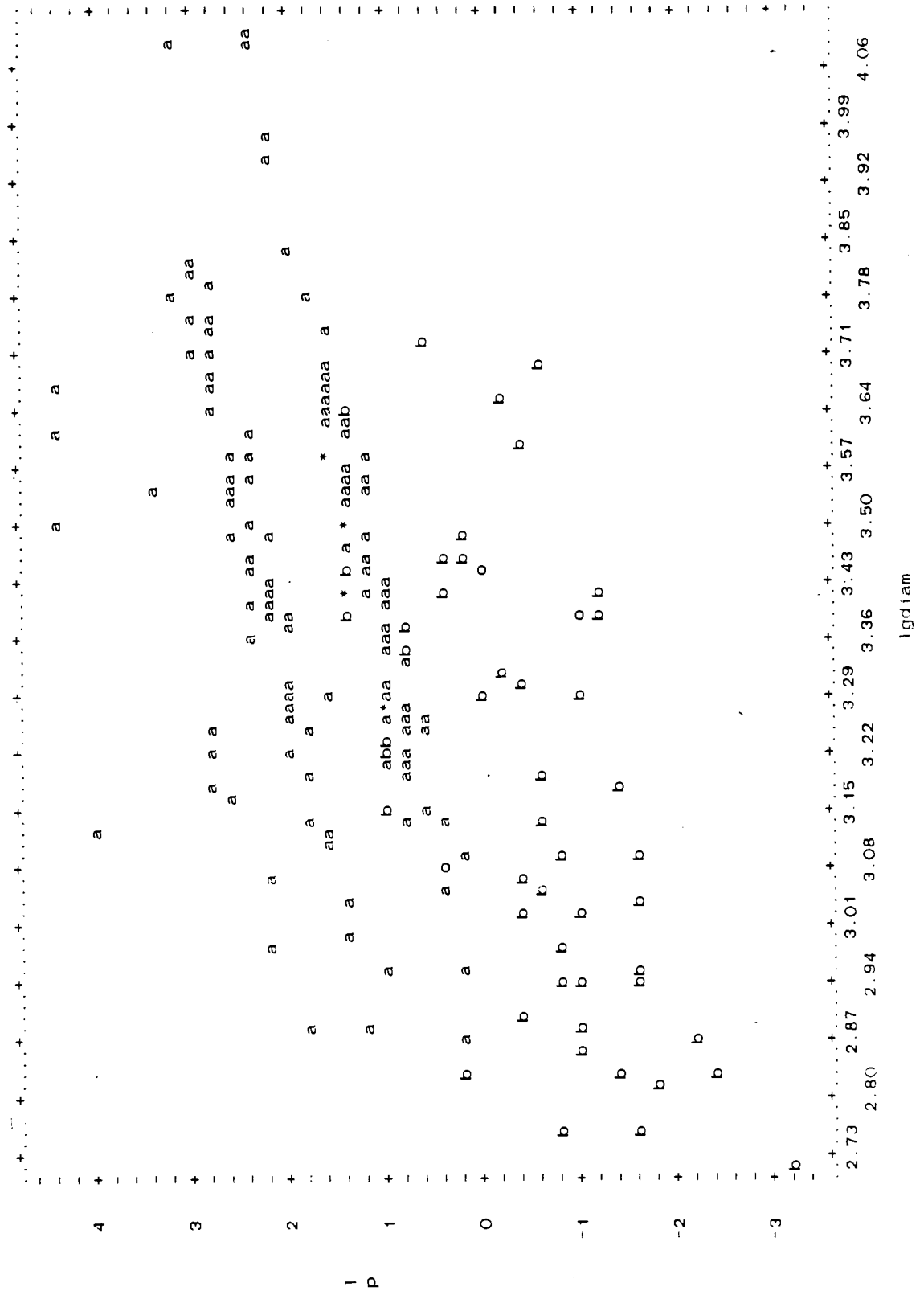


Figure 35.d: $\hat{\eta}$ against ln(DIAM) for All Deciduous Trees with Fungal Conks



chapter for a discussion). As no significant interactions were found at this stage, further searching was abandoned. In any case interaction effects are 'typically smaller' than main effects (Ref. (11) pg. 681), at least so long as the model chosen is the correct one.

CHAPTER 4

SECOND ANALYSIS--POSSIBLE RANKINGS FOR EXPLANATORY VARIABLES IN FINAL MODEL

The PLR run of Figure 31 used up all 7 expalanatory variables with which it was supplied. Note that SPTREE is counted here as one explanatory variable although both PLR and GLIM later split it up into 2 design variables. Dagmar next requested a ranking of these variables into which one was most 'important' to the final model, which came next in 'importance', and so on.

Unfortunately, no unique ranking scheme is possible because of nonzero correlations amongst the explanatory variables. For example if 2 variable, say X_1 and X_2, are highly correlated, then the order of their entry into a model becomes important. If X_1 enters the model first, then no significant improvement in model fit may result in adding X_2 when such correlation is present, since most of X_2's ability to explain variation in the response variable is already accounted for in X_1. Hence X_2 gets left out. But if X_2 enters the model ahead of X_1, then X_1 may end up getting left out. One is then faced with the problem of ranking the variables in 'importance' to the final model. This would not happen if X_1 and X_2 were uncorrelated, or very nearly so at least. This problem is known as multicollinearity (Ref.(11) Chapter 8).

Nonetheless of the distinct possible rankings which may present themselves, the following 3 are offered.

4.1 Order of Selection by PLR Run

The stepwise regression program starts off with an empty model (no explanatory variables, just a constant term to represent an overall average), and uses approximate *F*-to-enter values to search for the variable which offers the largest improvement over the empty model. Provided that the *p*-value associated with that *F*-to-enter value is sufficiently small, that variable then gets inserted into the model. DI1 was entered in the first step, since it offered the best improvement over an empty model.

Once DI1 was in the model, *ln*(DIAM) offered the best improvement over a model containing only DI1. Similarly in step 3, of the remaining candidate explanatory variables, SPTREE (through its associated 2 design variables) offered the best model fit improvement over a model which contained only DI1 and DIAM. Both of its design variables were entered at once, which is one of PLR's defaults.

This order of entry is summarized in Figure 36, and is probably most useful for management purposes. The reasoning is that if a deciduous tree is to be assessed for its probability of success, then if one intends to 'measure' only one explanatory variable, that variable should be DI1,

Figure 36: Explanatory Variables Order of Entry into Final PLR Model

SUMMARY OF STEPWISE RESULTS

STEP NO	TERM ENTERED	DF	TERM REMOVED	DF	LOG LIKELIHOOD	IMPROVEMENT CHI-SQUARE	IMPROVEMENT P-VALUE	GOODNESS OF FIT CHI-SQUARE	GOODNESS OF FIT P-VALUE
0					-586.609			1170.650	0.116
1	d11	1			-389.416	394.387	0.000	776.046	1.000
2	lndiam	1			-362.393	54.044	0.000	722.061	1.000
3	sptree	2			-330.935	62.917	0.000	659.093	1.000
4	bktop	1			-316.590	28.690	0.000	630.394	1.000
5	d12	1			-303.186	26.807	0.000	603.576	1.000
6	dwood	1			-299.951	6.471	0.011	597.096	1.000
7	lnheight	1			-298.504	2.893	0.089	594.202	1.000

since DI1 is the variable which gives the best possible 1-variable model. If one is willing to look at a second variable, then one should include *ln*(DIAM) as well, since *ln*(DIAM) offers the best improvement over a model which contains only DI1.

A couple of important points should be noted here. First, the discussion so far does <u>not</u> say that one may not obtain *ln*(DIAM) chronologically until after the value of DI1 is obtained. It does say that if one wishes to observe 2 variables (in whatever order) and one of those variables is to be DI1, then *ln*(DIAM) is the best possible choice for the other variable, since the best fitting 2-variable model in which DI1 is included contains *ln*(DIAM) as well. Second, although inclusion of *ln*(DIAM) offers the best improvement over a model which has only DI1, this does not mean that DI1 and *ln*(DIAM) form the best possible 2-variable model from all those possible with the given set of candidate explanatory variables. The determination of such a best possible 2-variable model is part of an 'all possible' or 'best *k*' subsets regression, which are distinct from stepwise procedures. Neither of these other procedures, however, was available in a package for logistic regression.

Continuing in the present discussion, if one similarly required a 3-variable model where DI1 and *ln*(DIAM) were to be used, then SPTREE offered the best improvement and so should be included as the 3rd 'variable'. Note that when a

tree's species is observed, both of the design variables SPTREE(1) and SPTREE(2) are assigned values, which is why they are counted as one 'variable'.

Figures 37 through 39 show what these 1, 2, and 3-variable models would be. From the appropriate table following 'Step Number 1' in Figure 37:

$$\hat{\eta}=-2.4036+3.400[DI1]$$

is the best possible 1-variable model. From the analogous table in 'Step Number 2' in Figure 38:

$$\hat{\eta}=-10.360+2.3688[ln(DIAM)]+3.3893[DI1]$$

is the best possible 2-variable model when DI1 is to be included. Again, from the analogous table in 'Step Number 3' in Figure 39:

$$\hat{\eta}=-7.8700+3.5907[ln(DIAM)]+1.7707[DI1]$$
$$-1.8599[SPTREE(1)]-2.2692[SPTREE(2)]$$

is the best possible 3-'variable' model of all those those which must include DI1 and ln(DIAM), and so on. Further steps are not shown.

One notes that as the number of variables increases, both the constant term and the coefficients for the common explanatory variables change from one model to the next. For example, in moving from the 2-variable to 3-'variable' model, the constant changed from -10.360 to -7.8700, and the coefficient for ln(DIAM) changed from 2.3688 to 3.5907. In general the constant term will change from one model to the next in any stepwise regression process, but the

Figure 37: Step 1 Selection of PLR Run

```
STEP NUMBER   1              di1                    IS ENTERED
---------------

                          LOG LIKELIHOOD =   -389.416
IMPROVEMENT CHI-SQUARE    (2*(LN(MLR) ) =    394.387  D.F.=    1  P-VALUE= 0.000
GOODNESS OF FIT CHI-SQ    (2*O*LN(O/E)) =    776.046  D.F.=1113  P-VALUE= 1.000
GOODNESS OF FIT CHI-SQ    ( C.C.BROWN ) =      0.0    D.F.=    0  P-VALUE= 1.000

                                          STANDARD
     TERM                 COEFFICIENT        ERROR    COEFF/S.E.

di1                          3.4000        0.1916        17.75
CONSTANT                    -2.4036        0.1214       -19.80

CORRELATION MATRIX OF COEFFICIENTS
----------------------------------

          di1         CONSTANT

di1         1.000
CONSTANT   -0.633       1.000

COVARIANCE MATRIX OF COEFFICIENTS
---------------------------------

             di1           CONSTANT

di1          0.03671
CONSTANT    -0.01473        0.01473

STATISTICS TO ENTER OR REMOVE TERMS
-----------------------------------
                    APPROX.                    APPROX.
     TERM            F TO    D.F.  D.F.        F TO    D.F.  D.F.
                     ENTER                     REMOVE                P-VALUE

sptree               46.21     2   1120                              0.0000
di1                                           314.62     1   1121    0.0000
di2                  30.45     1   1121                              0.0000
dwood                 0.87     1   1121                              0.3514
bktop                 0.66     1   1121                              0.4182
lnheight             11.38     1   1121                              0.0008
lndiam               57.61     1   1121                              0.0000
CONSTANT                                IS IN                                 MAY NOT BE
                                                                              REMOVED.
```

Figure 38: Step 2 Selection of PLR Run

```
STEP NUMBER   2              lndiam                IS ENTERED
--------------

                        LOG LIKELIHOOD =   -362.393
IMPROVEMENT CHI-SQUARE   (2*(LN(MLR) ) =     54.044  D.F.=    1  P-VALUE= 0.000
GOODNESS OF FIT CHI-SQ   (2*O*LN(O/E)) =    722.061  D.F.=1112  P-VALUE= 1.000
GOODNESS OF FIT CHI-SQ   ( D. HOSMER ) =     10.683  D.F.=   8  P-VALUE= 0.220
GOODNESS OF FIT CHI-SQ   ( C.C.BROWN ) =      0.961  D.F.=   2  P-VALUE= 0.618

                                        STANDARD
     TERM                COEFFICIENT      ERROR    COEFF/S.E.

di1                        3.4393        0.2033       16.92
lndiam                     2.3688        0.3323       7.128
CONSTANT                  -10.360         1.153      -8.988

CORRELATION MATRIX OF COEFFICIENTS
----------------------------------

          di1         lndiam     CONSTANT

di1         1.000
lndiam      0.222       1.000
CONSTANT   -0.287      -0.994       1.000

COVARIANCE MATRIX OF COEFFICIENTS
---------------------------------

              di1           lndiam         CONSTANT

di1          0.04134
lndiam       0.01501       0.11044
CONSTANT    -0.06726      -0.38079        1.32840

STATISTICS TO ENTER OR REMOVE TERMS
-----------------------------------
     TERM              APPROX.                      APPROX.
                       F TO    D.F.  D.F.    F TO   D.F.  D.F.
                       ENTER                 REMOVE              P-VALUE

sptree                 31.15     2  1119                        0.0000
di1                                                 293.65   1  1120  0.0000
di2                    22.62     1  1120                        0.0000
dwood                   7.63     1  1120                        0.0058
bktop                   4.57     1  1120                        0.0327
lnheight                0.00     1  1120                        0.9781
lndiam                                               52.14   1  1120  0.0000
CONSTANT                                          IS IN                    MAY NOT BE
                                                                           REMOVED.
```

Figure 39: Step 3 Selection of PLR Run

STEP NUMBER 3 sptree IS ENTERED

	LOG LIKELIHOOD =	-330.935			
IMPROVEMENT CHI-SQUARE	(2*(LN(MLR)) =	62.917	D.F.= 2	P-VALUE=	0.000
GOODNESS OF FIT CHI-SQ	(2*O*LN(O/E)) =	659.093	D.F.=1110	P-VALUE=	1.000
GOODNESS OF FIT CHI-SQ	(D. HOSMER) =	3.053	D.F.= 8	P-VALUE=	0.931
GOODNESS OF FIT CHI-SQ	(C.C.BROWN) =	2.933	D.F.= 2	P-VALUE=	0.231

TERM		COEFFICIENT	STANDARD ERROR	COEFF/S.E.
sptree	(1)	-1.8599	0.2821	-6.593
	(2)	-2.2692	0.7980	-2.844
di1		3.5907	0.2271	15.81
lndiam		1.7707	0.3720	4.760
CONSTANT		-7.8700	1.289	-6.107

CORRELATION MATRIX OF COEFFICIENTS

	sptre(1)	sptre(2)	di1	lndiam	CONSTANT
sptre(1)	1.000				
sptre(2)	0.051	1.000			
di1	-0.308	-0.048	1.000		
lndiam	0.224	-0.085	0.114	1.000	
CONSTANT	-0.242	0.074	-0.161	-0.995	1.000

COVARIANCE MATRIX OF COEFFICIENTS

	sptre(1)	sptre(2)	di1	lndiam	CONSTANT
sptre(1)	0.07959				
sptre(2)	0.01155	0.63682			
di1	-0.01971	-0.00864	0.05159		
lndiam	0.02346	-0.02531	0.00962	0.13836	
CONSTANT	-0.08803	0.07616	-0.04700	-0.47694	1.66072

STATISTICS TO ENTER OR REMOVE TERMS

TERM	APPROX. F TO ENTER	D.F.	D.F.	APPROX. F TO REMOVE	D.F.	D.F.	P-VALUE	
sptree				29.64	2	1117	0.0000	
di1				297.70	1	1118	0.0000	
di2	32.05	1	1118				0.0000	
dwood	35.69	1	1118				0.0000	
bktop	38.45	1	1118				0.0000	
lnheight	7.40	1	1118				0.0066	
lndiam				26.99	1	1118	0.0000	
CONSTANT				IS IN				MAY NOT BE REMOVED.

coefficients for a given explanatory variable should remain the same in all models when the variables are uncorrelated. As noted earlier, however, the given set of explanatory variables were correlated amongst themselves, so the coefficients will change from one model to the next larger one. Simple examples of this phenomenon are easily found, such as in Chapter 8 of Ref.(11). Most distressing is the case when the coefficients change sign.

4.2 All Possible 1-Variable Models

Referring back to Figure 36 again, one notes the column labelled 'Improvement Chi-Square'. The value given in that column for an individual variable is a measure of improvement in model fit once that variable is included. The PLR run does not show all the possible improvement chi-square values for 'Step Number 0' where all possible 1-variable models are considered, so another GLIM run was performed, the results of which are shown in Figure 40.

Along with each model fitted, a quantity called 'scaled deviance' is computed. It is the drop in scaled deviance in going from one model to another which gives the improvement chi-square values. One can see in Figure 40 that the empty model (no explanatory variables, just a constant term labelled '%GM' by GLIM) has a scaled deviance of 1174. The model with DI1 as the only variable gives a scaled deviance

Figure 40: GLIM Run of All Possible 1-Variable Models

```
        SCALED
CYCLE   DEVIANCE        DF
   4     1174.        1123

       ESTIMATE        S.E.        PARAMETER
   1   -1.288       0.7228E-01    %GM
  SCALE PARAMETER TAKEN AS        1.000

        SCALED
CYCLE   DEVIANCE        DF
   4     1057.        1121

       ESTIMATE        S.E.        PARAMETER
   1  -0.7590       0.8280E-01    %GM
   0     ZERO       ALIASED       SPTR(1)
   2   -1.859        0.2125       SPTR(2)
   3   -2.285        0.7240       SPTR(3)
  SCALE PARAMETER TAKEN AS        1.000

        SCALED
CYCLE   DEVIANCE        DF
   4     1095.        1122

       ESTIMATE        S.E.        PARAMETER
   1   -8.580        0.8707       %GM
   2    2.181        0.2559       LNDI
  SCALE PARAMETER TAKEN AS        1.000

        SCALED
CYCLE   DEVIANCE        DF
   4     1169.        1122

       ESTIMATE        S.E.        PARAMETER
   1   -2.303        0.4831       %GM
   2   0.3702        0.1727       LNHE
  SCALE PARAMETER TAKEN AS        1.000

        SCALED
CYCLE   DEVIANCE        DF
   4     779.0        1122

       ESTIMATE        S.E.        PARAMETER
   1   -2.404        0.1214       %GM
   0     ZERO       ALIASED       DI1(1)
   2    3.407        0.1915       DI1(2)
  SCALE PARAMETER TAKEN AS        1.000

        SCALED
CYCLE   DEVIANCE        DF
   4     1137.        1122

       ESTIMATE        S.E.        PARAMETER
   1   -1.622       0.9665E-01    %GM
   0     ZERO       ALIASED       DI2(1)
   2   0.9158        0.1492       DI2(2)
  SCALE PARAMETER TAKEN AS        1.000
```

Figure 40, continued

```
       SCALED
CYCLE  DEVIANCE       DF
   4     1161.       1122

       ESTIMATE       S.E.       PARAMETER
   1    -1.447      0.8746E-01   %GM
   0      ZERO      ALIASED      DWOD(1)
   2    0.5701       0.1575      DWOD(2)
 SCALE PARAMETER TAKEN AS        1.000

       SCALED
CYCLE  DEVIANCE       DF
   4     1161.       1122

       ESTIMATE       S.E.       PARAMETER
   1    -1.397      0.8025E-01   %GM
   0      ZERO      ALIASED      BKTO(1)
   2    0.6935       0.1903      BKTO(2)
 SCALE PARAMETER TAKEN AS        1.000
```

of 779.0 which means a drop of 395 in the scaled deviance. Indeed this is the chi-square improvment value given to DI1 back in 'Step Number 1' of Figure 31, given the available accuracy. A ranking of the explanatory variables on the basis of their chi-square improvement over the empty model is then possible:

DI1
SPTREE
ln(DIAM)
DI2
DWOOD
BKTOP
ln(HEIGHT)

The calculations are summarized in the Technical Supplement, Chapter 8.

Note that the order of SPTREE and *ln*(DIAM) has been changed over the previous ranking, and BKTOP has been displaced to a place of lesser importance. It is interesting to note that if a ranking is done on the basis of improvement chi-squares achieved in the stepwise model building shown back in Figure 31, one gets:

DI1
SPTREE
ln(DIAM)
BKTOP
DI2
DWOOD
ln(HEIGHT)

so that BKTOP becomes more important again. Thus one sees that while BKTOP is the most important variable to improve the 3-variable model in the stepwise process, it is almost the least important variable for improving an empty model.

Again this change in importance is due to multicollinearity.

Nonetheless, DI1 still manages to be the most important variable in these particular ranking schemes.

4.3 All Possible One-Less-Than-All Variable Models

If one now starts with the final model, which has a scaled deviance of 597.1 according to Figure 32, and tries out all possible 1-variable omissions, one can then rank the explanatory variables according to how much fit is lost when that variable is left out of the final model. The GLIM run to produce these fits is shown in Figure 41. One can see there that if DI1 is left out of the PLR final model, then the scaled deviance increases from 597.1 to 939.3. Thus if DI1 was the last variable to be added to the model, it would have given a chi-square improvement of 939.3-597.1=342.2. One could also look at this as a chi-square 'deprovement' if DI1 were singled out from omission from the PLR final model.

On this basis one can rank the explanatory variables as to how much loss in fit is encountered if that variable were singled out for omission from the PLR final model:

DI1
SPTREE
DI2
ln(DIAM)
BKTOP
DWOOD
ln(HEIGHT)

Figure 41: GLIM Run of All Possible 1-Variable Omissions from PLR Final Model

```
      SCALED
CYCLE DEVIANCE      DF
   4    692.7      1117

      ESTIMATE       S.E.      PARAMETER
   1    -11.88      1.284      %GM
   2     2.137     0.3903      LNDI
   0      ZERO    ALIASED      DI1(1)
   3     3.442     0.2099      DI1(2)
   0      ZERO    ALIASED      DI2(1)
   4    0.9274     0.2037      DI2(2)
   5    0.6108     0.3048      LNHE
   0      ZERO    ALIASED      DWOO(1)
   6    0.6871     0.3065      DWOO(2)
   0      ZERO    ALIASED      BKTO(1)
   7    0.3901     0.3943      BKTO(2)
 SCALE PARAMETER TAKEN AS      1.000

      SCALED
CYCLE DEVIANCE      DF
   5    613.3      1116

      ESTIMATE       S.E.      PARAMETER
   1    -5.537     0.8784      %GM
   0      ZERO    ALIASED      SPTR(1)
   2    -3.081     0.3635      SPTR(2)
   3    -2.957     0.9535      SPTR(3)
   0      ZERO    ALIASED      DI1(1)
   4     3.744     0.2423      DI1(2)
   0      ZERO    ALIASED      DI2(1)
   5     1.234     0.2194      DI2(2)
   6    0.9984     0.2873      LNHE
   0      ZERO    ALIASED      DWOO(1)
   7    0.8387     0.3202      DWOO(2)
   0      ZERO    ALIASED      BKTO(1)
   8     1.602     0.4123      BKTO(2)
 SCALE PARAMETER TAKEN AS      1.000

      SCALED
CYCLE DEVIANCE      DF
   4    939.3      1116

      ESTIMATE       S.E.      PARAMETER
   1    -8.508      1.108      %GM
   0      ZERO    ALIASED      SPTR(1)
   2    -1.901     0.2435      SPTR(2)
   3    -2.955     0.7331      SPTR(3)
   4     1.768     0.3313      LNDI
   0      ZERO    ALIASED      DI2(1)
   5    0.8832     0.1663      DI2(2)
   6    0.4077     0.2552      LNHE
   0      ZERO    ALIASED      OWOO(1)
   7    0.7302     0.2574      DWOO(2)
   0      ZERO    ALIASED      BKTO(1)
   8     1.129     0.3498      BKTO(2)
 SCALE PARAMETER TAKEN AS      1.000
```

Figure 41, continued

CYCLE	SCALED DEVIANCE	DF
5	624.3	1116

	ESTIMATE	S.E.	PARAMETER
1	-10.24	1.441	%GM
0	ZERO	ALIASED	SPTR(1)
2	-2.687	0.3530	SPTR(2)
3	-2.877	0.8691	SPTR(3)
4	1.941	0.4252	LNDI
0	ZERO	ALIASED	DI1(1)
5	3.625	0.2354	DI1(2)
6	0.4989	0.2995	LNHE
0	ZERO	ALIASED	DWOO(1)
7	0.9118	0.3197	DWOO(2)
0	ZERO	ALIASED	BKTO(1)
8	1.245	0.4029	BKTO(2)
SCALE PARAMETER TAKEN AS			1.000

CYCLE	SCALED DEVIANCE	DF
5	600.0	1116

	ESTIMATE	S.E.	PARAMETER
1	-9.481	1.422	%GM
0	ZERO	ALIASED	SPTR(1)
2	-2.888	0.3642	SPTR(2)
3	-3.036	0.8903	SPTR(3)
4	2.015	0.4047	LNDI
0	ZERO	ALIASED	DI1(1)
5	3.710	0.2438	DI1(2)
0	ZERO	ALIASED	DI2(1)
6	1.134	0.2203	DI2(2)
0	ZERO	ALIASED	DWOO(1)
7	0.8271	0.3180	DWOO(2)
0	ZERO	ALIASED	BKTO(1)
8	1.005	0.3696	BKTO(2)
SCALE PARAMETER TAKEN AS			1.000

CYCLE	SCALED DEVIANCE	DF
5	605.0	1116

	ESTIMATE	S.E.	PARAMETER
1	-9.083	1.408	%GM
0	ZERO	ALIASED	SPTR(1)
2	-2.792	0.3607	SPTR(2)
3	-3.002	0.8990	SPTR(3)
4	1.639	0.4315	LNDI
0	ZERO	ALIASED	DI1(1)
5	3.691	0.2416	DI1(2)
0	ZERO	ALIASED	DI2(1)
6	1.131	0.2202	DI2(2)
7	0.3546	0.2998	LNHE
0	ZERO	ALIASED	BKTO(1)
8	1.905	0.3797	BKTO(2)
SCALE PARAMETER TAKEN AS			1.000

Figure 41, continued

CYCLE	SCALED DEVIANCE	DF	
5	607.4	1116	
	ESTIMATE	S.E.	PARAMETER
1	-9.478	1.436	%GM
0	ZERO	ALIASED	SPTR(1)
2	-2.607	0.3449	SPTR(2)
3	-2.795	0.8529	SPTR(3)
4	1.948	0.4266	LNDI
0	ZERO	ALIASED	DI1(1)
5	3.694	0.2419	DI1(2)
0	ZERO	ALIASED	DI2(1)
6	1.116	0.2184	DI2(2)
7	0.7735E-01	0.2672	LNHE
0	ZERO	ALIASED	DWOO(1)
8	1.373	0.2857	DWOO(2)
SCALE PARAMETER TAKEN AS		1.000	

This differs from the ranking of the variables on the basis of improvement over the empty model in that the orders of DI2 and *ln*(DIAM) are reversed, as are the orders of BKTOP and DWOOD. Thus in omitting a variable from the PLR final model (perhaps because management is willing to observe 6 but not 7 varaibles for a given tree in order to cut back labour costs), the omission of *ln*(HEIGHT) would cause the least loss in fit, and the omission of DI1 would cause the greatest loss in fit (and hence predictive power). Note that the order reversal of DI2 and *ln*(DIAM) in this and the previous ranking means that while *ln*(DIAM) is more important than DI2 in improving the fit over an empty model, the omission of DI2 from the PLR final model would cause a larger loss of model fit than would the omission of *ln*(DIAM). Correlation between values of DI2 and DIAM (such as if, perhaps, thicker trees tend to have scars while thinner ones do not) and hence *ln*(DIAM) account for this order reversal. Once again, however, DI1 is still the most highly ranked explanatory variable.

4.4 Concluding Remarks on Ranking

The ranking schemes of the previous 3 sections confirm that there is no unique ranking possible among the explanatory variables, due to the presence of multicollinearity. Any ranking, however, should be selected on the basis of the usefulness of its interpretation to

those who would use it. Possibly the ranking scheme most useful to management would be that given in the first section, as has already been suggested there.

CHAPTER 5

USE OF PLR FINAL MODEL FOR ESTIMATION/PREDICTION OF PROBABILITIES

One of the intended uses of this model is for forest management or conservation workers who will want to know which factors they may control to encourage or discourage the population of primary excavator bird species through the availability of 'desirable' nest locations. Thus it will be useful to convert a fitted log odds value, $\hat{\eta}$, along with an estimated variance, $s^2(\hat{\eta})$, into a fitted value of the probability, μ, that a given tree or population of trees will have a success.

Rather than directly use the inverse of the link function, g, as was indicated in Chapter 3, a recent paper by Meester and Eaves (Ref.(9)) shows how to convert $\hat{\eta}$ and $s^2(\hat{\eta})$ into a predicted probability, μ_p for a single tree. This paper has been reproduced in its entirety in Appendix A. For example, for observation 3 in the GOOKNEST5 file, (records 7 and 8 in the original GOOKADD raw data file) which is an aspen tree:

SPTREE(1)=0
SPTREE(2)=0
ln(HEIGHT)=2.914
ln(DIAM)=3.645
DI1=1 (fungal conks)
DI2=0 (no scars)
DWOOD=0 (tree fully live and has intact top)
BKTOP=0 (intact top)

one can easily find from the PLR final model:

$$\hat{\eta}=1.5006$$

and after a more involved computation (Technical Supplement, Chapter 9), one obtains:

$$s^2(\hat{\eta})=0.0566$$

from which the table given in Ref.(9) gives:

$$\mu_p=0.8176$$

using the closest available values of 'm=1.5,v=0' from that table. So this particular aspen tree which has

$$\text{HEIGHT}=e^{2.914}=18.43 \text{ metres}$$

and

$$\text{DIAM}=e^{3.645}=38.28 \text{ centimetres}$$

has a predicted probability of 0.8176 of having a success. This is the '$\hat{\eta}$ as prediction' described in Chapter 3, which may prove more useful than '$\hat{\eta}$ as estimate'.

It is interesting to compare this predicted probability, μ_p, with the corresponding estimated mean probability:

$$\hat{\mu}=g^{-1}(1.5006)=0.8177$$

The difference is small, but in general one will find that μ_p is 'pulled' closer to 0.5 than $\hat{\mu}$. This would be more dramatically demonstrated if $s^2(\hat{\eta})$ were higher.

Consider instead observation 304 from the GOOKNEST5 file (records 709 and 710 from the original GOOKADD raw data file). For this observation it turns out that:

$$\hat{\eta}=-5.1703$$

$$s^2(\hat{\eta})=1.1477$$

Now whereas:

$$\hat{\mu}=g^{-1}(-5.1703)=0.0056508$$

(here $\hat{\mu}<0.5$ because $\hat{\eta}<0$), it turns out from the table in Ref.(9) (Appendix A), with '|m|=5.0,v=1':

$$\mu_p=1-0.9892=0.0108$$

If it turned out that $s^2(\hat{\eta})<0.5$, then 'v=0' would be used in the table look-up to produce:

$$\mu_p=1-0.9933=0.0067$$

which is closer to $\hat{\mu}$. In either case, the μ_p is pulled closer toward 0.5 than is $\hat{\mu}$, but this pull is more dramatic for a larger $s^2(\hat{\eta})$.

This pulling of a fitted μ closer to one-half in switching from $\hat{\mu}$ to μ_p reflects the greater caution taken in predicting a probability for a single tree than for estimating a mean probability for a population of trees which have similar values of SPTREE, HEIGHT, DIAM, DI1, and so on. This is described in more detail in the technical supplement, but is analogous to the situation in normal theory multiple linear regression where for a fitted response value, $\hat{Y}$, a $100(1-\alpha)\%$ prediction interval is wider than a $100(1-\alpha)\%$ confidence interval (Ref. (6) pg. 312). In that situation, the worst scenario would be one where the prediction interval becomes too wide to be useful. In the present situation of predicting probabilities, the worst scenario would be one where μ_p becomes 0.5, which states that for the given tree, one has no further information for

predicting success or failure other than by tossing a fair coin. One can further see from the table in Appendix A that as $s^2(\hat{\eta})$ increases, μ_p is pulled closer to 0.5. This too makes intuitive sense.

If one wishes to encourage the population of primary excavator birds, one would make a decision rule of cultivating trees such that, say, $\mu_p > 0.5$, or higher. μ_p can be increased above 0.5 by increasing $\hat{\eta}$ above 0, although any change to $\hat{\eta}$, done through the explanatory variables, will also change $s^2(\hat{\eta})$.

A useful item for field workers would be a series of tables showing μ_p values for all possible combinations of the qualitative explanatory variables (SPTREE, DI1, DI2, DWOOD, BKTOP) and for pre-selected values of HEIGHT and DIAM.

CHAPTER 6

TECHNICAL SUPPLEMENT FOR CHAPTER 2

For Chapter 2 there is little to add except to show the FORTRAN program (Figure 42) that was used to convert the raw data file into the file of data ready for analysis (GOOKNEST5), with rejected or flagged cases docmented in a separate reject file (GOOKREJECT5). Portions of these 3 files have already been shown in Tables 8-10 respectively. The commands to compile and run this program are not shown but would be similar to those shown in the Technical Supplement to the analysis of Simon Emms' data (PART A).

One feature of the program worth noting is the instruction to remove all trees having 'success' after line 564 in the original raw data file. The reason for doing so was that this data file itself was a concatenation of 2 other files, the first of which contained only successes, and was 564 lines long. The second file contained both successes and failures. Although this second file did not contain all the successes of the first file, Dagmar said that those which were contained were duplications of observations from the first file, and thus had to be omitted.

Figure 42: FORTRAN Program Used to Edit Dagmar's Raw Data File

```
C PROGRAM TO PROCESS GOOKADD FILE INTO FILE READY FOR STAT PACKAGE INPUT
C
C    UNIT 10 = GOOKADD
C    UNIT 11 = GOOKNEST5
C    UNIT 12 = GOOKREJECT5
C
C        1         2         3         4         5         6         7
C234567890123456789012345678901234567890123456789012345678901234567890123456789012
      DIMENSION VTREES(5),VTLD(2),VPORA(2),VBIRDS(15)
      DATA VTREES /'A','BI','CT','W','D'/
      DATA VTLD /'L','D'/
      DATA VPORA /' ','NT'/
      DATA VBIRDS /'S','P','N','F','H','D',' ','B','G','K','M','Q','R',
     *'T','W'/
      RECNUM=-1.0
C
C  READ IN DATA
C
   10 READ (10,101,ERR=998,END=999) TREESP,DIAM,HEIGHT,TLD,DI1,DI2,DI3,
     *DI4,PORA,BIRDSP,BKTOP
      RECNUM=RECNUM+2.0
C
C REJECT NESTED TREES AFTER LINE 564 OF INPUT
C
      IF (RECNUM .LE. 563.0) GO TO 20
      IF (PORA .NE. VPORA(2)) GO TO 20
      WRITE (12,202) RECNUM,PORA
      GO TO 10
C
C  NOW REJECT CONIFEROUS TREES
C
   20 DO 55 I1=1,5
      IF (TREESP .EQ. VTREES(I1)) GO TO 30
   55 CONTINUE
      WRITE (12,203) RECNUM,TREESP
      GO TO 10
C
C  BEGIN TRANSFORMATIONS.....START WITH TREE SPECIES
C
   30 CONTINUE
      IF (TREESP .EQ. VTREES(1)) GO TO 40
      IF (TREESP .EQ. VTREES(2)) GO TO 50
      SPTREE=3.0
      GO TO 60
   40 SPTREE=1.0
      GO TO 60
   50 SPTREE=2.0
C
C  TRANSFORM DI1,DI2,DI3
C
   60 DI1=DI1/37.0
      DI2=DI2/39.0
      DI3=DI3/43.0
C
C  CREATE DWOOD VARIABLE
C
      DWOOD=2.0
      IF (TLD .EQ. VTLD(1)) GO TO 70
      IF (TLD .NE. VTLD(2)) GO TO 80
      GO TO 90
   70 CONTINUE
      IF (DI4 .EQ. 0.0) DWOOD=1.0
      GO TO 90
   80 WRITE (12,204) RECNUM,TLD
```

Figure 42, continued

```
      GO TO 10
   90 CONTINUE
C
C  TRANSFORM PORA
C
      IF (PORA .NE. VPORA(2)) GO TO 100
      ACNEST=1.0
      GO TO 110
  100 CONTINUE
      IF (PORA .EQ. VPORA(1)) GO TO 120
      WRITE (12,205) RECNUM,PORA
      GO TO 10
  120 ACNEST=0.0
C
C  TRANSFORM BIRDSPECIES
C
  110 SPBIRD=8.0
      DO 65 I2=1,7
      IF (BIRDSP .NE. VBIRDS(IS)) GO TO 65
      SPBIRD=I2
      GO TO 140
   65 CONTINUE
      DO 75 I3=8,15
      IF ( BIRDSP .EQ. VBIRDS(I3)) GO TO 130
   75 CONTINUE
      WRITE (12,206) RECNUM,BIRDSP
C
C  PROBABLY '?' SPECIES, BUT CARRY ON WITH REST OF RECORD ANYWAY & JUST
C    FLAG IT FOR CONFIRMATION
C
      GO TO 140
  130 WRITE (12,207) RECNUM,BIRDSP
      GO TO 10
C
C  TRANSFORMATIONS ALL DONE.....RECORD DATA & GO GET ANOTHER RECORD
C
  140 WRITE (11,201) RECNUM,ACNEST,SPTREE,HEIGHT,DIAM,DI1,DI2,DI3,
     *DWOOD,BKTOP,SPBIRD
      GO TO 10
  998 WRITE (12,208) RECNUM
      GO TO 10
  999 WRITE (12,209) RECNUM
      STOP
C
C  FORMAT STATEMENTS
C
C         1         2         3         4         5         6         7
C234567890123456789012345678901234567890123456789012345678901234567890123
  101 FORMAT (T9,A2,T12,F4.1,T17,F5.2,T23,A1,4(1X,F2.0)/T9,A2,T12,A1,
     *T32,F1.0)
  201 FORMAT (T3,F8.1,1X,F2.0,1X,F2.0,1X,F5.2,1X,F6.2,6(1X,F2.0))
  202 FORMAT (/T3,'IN RECORD ',F8.1,' (AFTER 564.0), PORA IS ',A2)
  203 FORMAT (/T3,'RECORD ',F8.1,' DELETED BECAUSE OF TREE SPECIES ',A2)
  204 FORMAT (/T3,'RECORD ',F8.1,' CONTAINS ILLEGAL TLD = ',A1)
  205 FORMAT (/T3,'RECORD ',F8.1,' CONTAINS ILLEGAL PORA = ',A2)
  206 FORMAT (/T10,'***CHECK RECORD ',F8.1,' ***'/T10,'FOR BIRD SPECIES
     *',A1/T10,'RECORD NOT SKIPPED')
  207 FORMAT (/T3,'RECORD ',F8.1,' CONTAINS SECONDARY EXCAVATOR BIRD SPE
     *CIES ',A1)
  208 FORMAT (/T10,'***ERROR IN DATA INPUT--RECORD ',F8.1,' ***')
  209 FORMAT (//T5,'ALL DONE'//'LAST 2 RECORDS READ IN BEGAN WITH RECORD
     * NUMBER ',F8.1)
      END
```

CHAPTER 7

TECHNICAL SUPPLEMENT FOR CHAPTER 3

A number of observations and developments may be made on the contents of Chapter 3. Basically they all have to do with the production of the outputs in Figures 31-35.

7.1 On the Production of the PLR Run in Figure 31

Figure 31, as well as Figures 36-39, contain a portion of the PLR run carried out by the command source file in Figure 43. There are 3 aspects particularly worthy of note.

Firstly, the 'space' modifier will be noted in the '$run' command. This requests that extra storage be made available for the PLR run. The default value is 15000 words, which was not enough for the run shown. This ability to request extra storage is a nice feature of BMDP. This particular method of doing so, however, is not that found in the BMDP manual (Ref.(4), Appendix B.1), but is that described on page 24 of Ref.(1) since it is an extension of the operating system, MTS.

Secondly, PLR offers 3 choices of design variables (Ref.(4) pg. 339). The strict 0,1-coding was chosen ('dvar=part.' sentence in the '/ regress' paragraph) where the first level of a factor (e.g. SPTREE=1) sets all applicable design variables to 0.

Figure 43: PLR Command File which Generated Figure 31

```
$empty goutplr4 ok
$run *bmdp 7=gooknest5 sprint=goutplr4 par=plr space=18000w
/ problem     title is 'GOUTPLR4:  logistic regression (no interactions)
                       on data in file GOOKNEST5, method is ACE'.
/ input       variables are 11.
              format is '(4x,a4,2x,2(1x,f2,0),1x,f5.2,1x,
                          f6.2,6(1x,f2.0))'.
              unit is 7.
              cases are 1124.
/ variable    names are recnum,acnest,sptree,height,diam,di1,di2,
                        di3,dwood,bktop,spbird,lnheight,lndiam.
              add=2.
              label is recnum.
/ transform   lnheight=ln(height).
              lndiam=ln(diam).
/ group       codes(2)=1,0.
              names(2)=nest,' '.
              codes(3)=1,2,3.
              names(3)=aspen,birch,'other deciduous'.
              codes(6)=1,0.
              names(6)=conks,' '.
              codes(7)=1,0.
              names(7)=scars,' '.
              codes(8)=1,0.
              names(8)='dead brn','full brn'.
              codes(9)=1,2.
              names(9)=full,'not full'.
              codes(10)=1,0.
              names(10)='bad top','gd top'.
              codes(11)=1,2,3,4,5,6,7,8.
              names(11)=s,p,n,f,h,d,'no nest',other.
/ regress     dependent=acnest.
              interval=lnheight,lndiam.
              categorical=sptree,di1,di2,dwood,bktop.
              dvar=part.
              cmove=2.
/ print       case=15.
              sort=none.
              hist.
              plot.
              cova.
              corr.
              news.
/ plot        size is 100,50.
/ end
```

Thirdly, the results are shown for the default ACE (Asymptotic Covariance Estimate) method rather than for the MLR (Maximum Likelihood Ratio) method which the manual states is the more 'reliable' of the two (Ref.(4) pg. 339). The MLR run results are not shown, but both methods did produce the same coefficient estimates and variance-covariance matrix for those estimates to the reported accuracy of 5 significant figures for the coefficients and 5 decimal places for the matrix entries. But the MLR method did not appear to perform the variable selection at each step of the stepwise process as the manual says that both it and the ACE method should (Ref.(4) pg. 339). Since the final model results are the same anyway, only the ACE method results are shown, in order to reduce confusion. The ACE method does work the way it should, namely by entry by smallest p-values, and not by largest F-statistics, since design variables for a given multi-level factor are to be entered as a group. Thus the degrees of freedom associated with the F-statistic may be different from those associated with other variables. One can see this in Figures 37-39.

7.2 On the GLIM Run which Produced Figure 37

In Figure 44 is shown the source file which produced the first GLIM run, part of which was shown in Figure 32. There are 4 features of special interest:

Figure 44: GLIM Command File which Generated Figure 32

```
$EMPTY GOUTGLIM1 OK
$EMPTY GOOKPLOT1 OK
$RUN UNSP:GLIM 1=GOOKNEST5 2=GOUTGLIM1 3=GOOKPLOT1
$C
$C GLIM RUN ON FILE GOOKNEST5 TESTING OUT LOGISTIC LINK ON BINARY
$C      RESPONSES
$C
$C ************************
$C *  GET DATA & TRANSFORM *
$C ************************
$C
$OUTPUT 2 132
$UNITS 1124
$DATA RECNUM ACNEST SPTREE LNHEIGHT LNDIAM DI1 DI2 DI3 DWOOD BKTOP
$FACTOR SPTREE 3 DI1 2 DI2 2 DI3 2 DWOOD 2 BKTOP 2
$FORMAT
(2X,F8.1,2(1X,F2.0),1X,F5.2,1X,F6.2,5(1X,F2.0))
$DINPUT 1
$LOOK 1 15 RECNUM ACNEST SPTREE LNHEIGHT LNDIAM DI1 DI2 DI3 DWOOD BKTOP
$C
$C          LOG TRANSFORM LNHEIGHT & LNDIAM
$C          SHIFT ALL QUALITATIVE VARIABLES UP BY 1 BECAUSE GLIM
$C              CAN'T HANDLE FACOTR LEVELS OF 0
$C
$CALC LNHEIGHT=%LOG(LNHEIGHT): LNDIAM=%LOG(LNDIAM): N=1: DI1=DI1+1
$CALC DI2=DI2+1: BKTOP=BKTOP+1
$C
$C ***************************************
$C *  SPECIFY MODEL TO BE ANALYZED       *
$C ***************************************
$C
$YVAR ACNEST
$ERROR B N
$LINK G
$C
$C ************************************************
$C *  FIT MODEL RECOMMENDED IN GOUTPLR4           *
$C ************************************************
$C
$FIT %GM
$DISP A
$FIT SPTREE+LNDIAM+LNHEIGHT+DI1+DI2+DWOOD+BKTOP
$DISP A
$DISP V
$C
$C ******************************************************************
$C *  GET FITTED VALUES, S.E.'S, & PUT INTO GOOKPLOT1 FILE          *
$C ******************************************************************
$C
$EXTRACT %VL
$OUTPUT 3 132
$ACCURACY 9
$LOOK RECNUM %LP %VL LNDIAM SPTREE DI1 BKTOP
$STOP
```

Firstly, each input record (as shown in Table 9) represents one observation or trial, and not the number of successes in a set of more than one trials. The data set is thus 'ungrouped' (Ref.(2) pg. 73) and hence 'N=1' is used in the '$CALC' command. Incidentally, the PLR run also gave the information that the 1124 observations could be assembled into 1115 different groups (that is, 1115 distinct patterns in the explanatory variables. Rather than assemble these groups for the GLIM run, the data was left in its ungrouped state.

Secondly, it will be noticed in Figure 44 that through '$CALC' commands, qualitative variables which had codings of 0 or 1 (DI1, DI2, BKTOP have these) are rescaled to have codings of 1 or 2 respectively, like DWOOD already has. This is necessary because GLIM will not process inputted factor levels (qualitative variable values) of 0. SPTREE, which has levels of 1, 2, and 3 for aspen, birch, and other deciduous trees respectively, could thus be inputted in as is. Once these variables were inputted, GLIM set up the necessary design variables in the same way that PLR set up its design variables through the 'dvar=part.' option mentioned earlier. These qualitative variables were identified to GLIM as factors along with their numbers of possible levels through the '$FACTOR' command, as can be seen in the figure.

Thirdly, for the vector of estimated coefficients, $\hat{\underline{\beta}}$, a dispersion or variance-covariance matrix, $C\hat{o}v(\hat{\underline{\beta}})$ was

estimated. The matrix of estimates differs in PLR (Figure 31) and GLIM (Figure 32) where it will be noticed that the corresponding entries agree to only 1 or 2 significant figures. One must further be careful when comparing these 2 matrices, since PLR and GLIM order the rows/columnsdifferently. This dilemma will be encountered again in Section 9.1.

Fourthly, $\hat{\eta}$ and $s^2(\hat{\eta})$, which are computed by GLIM for each observation in the given sample data and stored in the 'system vectors' %LP and %VL respectively, were requested to be put along with other data values of interest into another file (identified near the top of Figure 44 as GOOKPLOT1) according to the last 4 lines of Figure 44. This file, a portion of which is shown in Table 11, contains data which was used in the plots mentioned in Chapter 3, but not directly from GOOKPLOT1.

The data values, as they appear in Table 11, were not yet ready for inputting into the P6D program of BMDP. The reason is that GLIM uses exponential notation for all numbers less than 0.10 in absolute value. BMDP uses FORTRAN formats (unless free format is chosen) and only 'F' or 'I' formats are available for numeric input. 'G' or 'E' formats, which can handle exponential notation are not available in BMDP. Rather than take chances with free format, the file was subjected to a FORTRAN 'clean-up' program, as shown in Figure 45, to convert the exponential notation back into

Table 11: Portion of Data File GOOKPLOT1

1	1.00000000	-1.45338154	0.169185877	3.17805290	2.00000000	1.00000000
2	3.00000000	2.84211731	0.813454986E-01	3.67630005	1.00000000	2.00000000
3	7.00000000	1.50158501	0.561262630E-01	3.64544964	1.00000000	1.00000000
4	9.00000000	2.86137676	0.819050670E-01	3.70376778	1.00000000	1.00000000
5	11.0000000	-1.73006630	0.175053418	3.64544964	1.00000000	1.00000000
6	13.0000000	-0.619647980	0.107684910	3.19458199	2.00000000	2.00000000
7	17.0000000	1.03915882	0.122336328	3.21887493	2.00000000	2.00000000
8	19.0000000	1.44467640	0.132982910	3.41114712	2.00000000	2.00000000
9	23.0000000	0.455393717E-01	0.110201955	3.28840160	2.00000000	2.00000000
10	25.0000000	-3.95456982	0.203653574	3.25809574	2.00000000	1.00000000
11	27.0000000	-2.41409302	0.102817595	3.11794949	1.00000000	1.00000000
12	29.0000000	-1.40158081	0.373073705E-01	3.47196579	1.00000000	1.00000000
13	33.0000000	0.969761550	0.464834459E-01	3.37416840	1.00000000	1.00000000
14	37.0000000	2.40289879	0.124418974	4.09434414	2.00000000	1.00000000
15	39.0000000	1.30613136	0.588683560E-01	3.41772652	1.00000000	1.00000000
16	41.0000000	1.38045025	0.500395708E-01	3.56953239	1.00000000	1.00000000
17	43.0000000	1.37517166	0.153624356	3.00072002	1.00000000	1.00000000
18	45.0000000	0.306982517	0.984058976E-01	3.35689640	1.00000000	2.00000000
19	49.0000000	1.04338741	0.471495986E-01	3.35340691	1.00000000	1.00000000
20	51.0000000	2.72091770	0.160030961	3.74714756	1.00000000	1.00000000
21	53.0000000	2.41924095	0.127939343	4.10758972	1.00000000	1.00000000
22	57.0000000	0.469663918	0.595562868E-01	3.13983250	1.00000000	1.00000000
23	59.0000000	1.31669712	0.117482305	3.39450836	2.00000000	2.00000000
24	63.0000000	-1.23628902	0.206785917	3.41772652	2.00000000	1.00000000
25	65.0000000	-0.658148110	0.109468400	3.32143211	1.00000000	2.00000000
26	69.0000000	1.81306458	0.690288544E-01	3.78418922	1.00000000	2.00000000
27	71.0000000	1.81306458	0.690288544E-01	3.78418922	1.00000000	1.00000000
28	73.0000000	-0.666947126	0.489493981E-01	3.78418922	1.00000000	1.00000000
29	79.0000000	2.69313049	0.779743195E-01	3.58629227	1.00000000	2.00000000
30	81.0000000	-1.22835255	0.360173360E-01	3.52636051	1.00000000	1.00000000
31	83.0000000	4.08639431	0.288286626	3.12236404	1.00000000	2.00000000
32	85.0000000	2.98931408	0.878322124E-01	3.75653744	1.00000000	2.00000000
33	87.0000000	-2.08023739	0.181165636	2.85646915	1.00000000	1.00000000
34	89.0000000	0.821637630	0.505927429E-01	3.32862568	1.00000000	1.00000000
35	91.0000000	-0.760233760	0.827680826E-01	3.40119648	1.00000000	2.00000000
36	93.0000000	1.81306458	0.690288544E-01	3.78418922	1.00000000	1.00000000
37	99.0000000	3.23629093	0.227290392	4.08260918	1.00000000	2.00000000
38	103.000000	2.20042515	0.992392302E-01	3.97780991	1.00000000	1.00000000
39	105.000000	2.80963802	0.153126717	3.22684383	1.00000000	2.00000000
40	107.000000	0.219346106	0.100170255	3.29583645	1.00000000	1.00000000
41	109.000000	-0.524864674	0.217308462	3.70130157	2.00000000	2.00000000
42	111.000000	2.72625256	0.769442320E-01	3.64544964	1.00000000	1.00000000
43	113.000000	-1.03513718	0.384127386E-01	3.63495064	1.00000000	1.00000000
44	115.000000	0.820558786	0.592536591E-01	3.20274639	1.00000000	1.00000000
45	117.000000	1.26890659	0.610442534E-01	3.58351803	1.00000000	1.00000000
46	119.000000	1.57758236	0.541864745E-01	3.63495064	1.00000000	1.00000000
47	121.000000	-1.01299191	0.100823522	2.93385696	2.00000000	2.00000000
48	123.000000	-2.40671444	0.553725399E-01	3.61361694	1.00000000	1.00000000
49	125.000000	0.862799287E-01	0.805377007	3.44680691	3.00000000	2.00000000
50	127.000000	0.389656305	0.832920611	3.07731152	2.00000000	2.00000000
51	131.000000	1.98615360	0.743134618E-01	3.28091049	1.00000000	1.00000000
52	133.000000	-1.50084972	0.498603284E-01	3.32143211	1.00000000	1.00000000

Figure 45: FORTRAN 'Clean-Up' Program for File GOOKPLOT1

```
C PREPARE GOOKPLOT1 TO BMDP6D PLOTS
C
C   UNIT 10 = GOOKPLOT1
C   UNIT 11 = GPLOTDAT1
C
C        1         2         3         4         5         6         7
C234567890123456789012345678901234567890123456789012345678901234567890123
      N=0
   10 READ (10,101,ERR=998,END=999) RECNUM,YLP,VL,DIAMLG,SPTREE,DI1,
     *BKTOP
      N=N+1
      WRITE (11,201) RECNUM,YLP,VL,DIAMLG,SPTREE,DI1,BKTOP
      GO TO 10
  998 WRITE (11,202) N
      GO TO 10
  999 WRITE (11,203) N,RECNUM
      WRITE (11,204)
      STOP
  101 FORMAT (T10,F5.0,T24,2G17.9,T58,F13.8,7X,F2.0,2(15X,F2.0))
  201 FORMAT (T6,F5.0,1X,F14.9,1X,F14.9,1X,F14.8,3(1X,F2.0))
  202 FORMAT (' ***ERROR AFTER RECORD ',I4,' IN INPUT***')
  203 FORMAT (' >>>ALL DONE<<<'/' LAST RECORD WAS NUMBER ',I4,/' FOR GOO
     *KNEST5 FILE RECORD NUMBER ',F5.0)
  204 FORMAT ('C234567890123456789012345678901234567890123456789012345678
     *90123456789012'/'C',8X,'1',9X,'2',9X,'3',9X,'4',9X,'5',9X,'6',9X,
     *'7')
      END
```

fixed decimal, as shown in the 'cleaned up' file, GPLOTDAT1, a portion of which is shown in Table 12.

At the same time the program was used to remove superfluous 0's from the (unneeded) fractional parts of numbers which were intended to be integers (GLIM treats integers as it does real numbers in general) and to line up all the decimal points. This latter task was not necessary because when a number is inputted to a FORTRAN program (including the software packages used), a user-keyed decimal point will override where an input 'F' format indicates it should be found (Ref.(7) pg. 24). The alignment, however, comes naturally with 'F' format on output data. This is not true necessarily with 'G' format, which is evidently what GLIM used.

7.3 On the Production of the Plots in Figures 33-35

The plots in Figures 33.a-d, 34.a-d, and 35.a-d were produced by the command source files shown in Figures 46, 47, and 48 respectively. In Figures 47 and 48 note should be made of the usage of the 'use' variable for case selection in the '/ transform' paragraph. This feature is described on page 55 of Ref.(4) and is distinct from the 'use=' sentence in a '/ variable' paragraph as described on page 42 of the same reference. The 'use' variable is a BMDP supplied variable and as such does not need to be included in a

Table 12: Portion of 'Cleaned Up' Data File, GPLOTDAT1

1.	-1.453381538	0.169185877	3.17805290	2.	2.	1.
3.	2.842117310	0.081345499	3.67630005	1.	2.	1.
7.	1.501585007	0.056126263	3.64544964	1.	2.	1.
9.	2.861376762	0.081905067	3.70376778	1.	2.	1.
11.	-1.730066299	0.175053418	3.64544964	1.	1.	1.
13.	-0.619647980	0.107684910	3.19458199	2.	2.	2.
17.	1.039158821	0.122336328	3.21887493	2.	2.	2.
19.	1.444676399	0.132982910	3.41114712	2.	2.	2.
23.	0.045539372	0.110201955	3.28840160	2.	2.	2.
25.	-3.954569817	0.203653574	3.25809574	2.	1.	1.
27.	-2.414093018	0.102817595	3.11794949	1.	1.	1.
29.	-1.401580811	0.037307370	3.47196579	1.	1.	1.
33.	0.969761550	0.046483446	3.37416840	1.	2.	1.
37.	2.402898788	0.124418974	4.09434414	1.	2.	1.
39.	1.306131363	0.058868356	3.41772652	1.	2.	1.
41.	1.380450249	0.050039571	3.56953239	1.	2.	1.
43.	1.375171661	0.153624356	3.00072002	1.	2.	1.
45.	0.306982517	0.098405898	3.35689640	1.	1.	2.
49.	1.043387413	0.047149599	3.35340691	1.	2.	1.
51.	2.720917702	0.160030961	3.74714756	1.	2.	1.
53.	2.419240952	0.127939343	4.10758972	1.	2.	1.
57.	0.469663918	0.059556287	3.13983250	1.	2.	1.
59.	1.316697121	0.117482305	3.39450836	2.	2.	2.
63.	-1.236289024	0.206785917	3.41772652	2.	2.	1.
65.	-0.658148110	0.109468400	3.32143211	1.	1.	2.
69.	1.813064575	0.069028854	3.78418922	1.	2.	1.
71.	1.813064575	0.069028854	3.78418922	1.	2.	1.
73.	-0.666947126	0.048949398	3.78418922	1.	1.	1.
79.	2.693130493	0.077974319	3.58629227	1.	2.	1.
81.	-1.228352547	0.036017336	3.52636051	1.	1.	1.
83.	4.086394310	0.288286626	3.12236404	1.	2.	2.
85.	2.989314079	0.087832212	3.75653744	1.	2.	1.
87.	-2.080237389	0.181165636	2.85646915	1.	1.	1.
89.	0.821637630	0.050592743	3.32862568	1.	2.	1.
91.	-0.760233760	0.082768083	3.40119648	1.	1.	2.
93.	1.813064575	0.069028854	3.78418922	1.	2.	1.
99.	3.236290932	0.227290392	4.08260918	1.	2.	1.
103.	2.200425148	0.099239230	3.97780991	1.	2.	1.
105.	2.809638023	0.153126717	3.22684383	1.	2.	1.
107.	0.219346106	0.100170255	3.29583645	1.	1.	2.
109.	-0.524864674	0.217308462	3.70130157	2.	2.	1.
111.	2.726252556	0.076944232	3.64544964	1.	2.	1.
113.	-1.035137177	0.038412739	3.63495064	1.	1.	1.
115.	0.820558786	0.059253659	3.20274639	1.	2.	1.
117.	1.268906593	0.061044253	3.58351803	1.	2.	1.
119.	1.577582359	0.054186475	3.63495064	1.	2.	1.
121.	-1.012991905	0.100823522	2.93385696	2.	2.	2.
123.	-2.406714439	0.055372540	3.61361694	1.	1.	1.
125.	0.086279929	0.805377007	3.44680691	3.	2.	2.
127.	0.389656305	0.832920611	3.07731152	3.	2.	2.
131.	1.986153603	0.074313462	3.28091049	1.	2.	1.
133.	-1.500849724	0.049860328	3.32143211	1.	1.	1.
135.	-0.643266439	0.050052091	3.79773331	1.	1.	1.
137.	2.161178589	0.072113216	3.41772652	1.	2.	1.
139.	-1.284874916	0.038834143	3.54673958	1.	1.	1.
141.	-1.002964020	0.039903320	3.66099358	1.	1.	1.
143.	-2.280308723	0.124899089	3.52046013	2.	1.	2.
145.	-2.254927635	0.038123202	3.58629227	1.	1.	1.
147.	1.368361473	0.048646651	3.54961681	1.	2.	1.
151.	0.777302563	0.048891280	3.26956844	1.	2.	1.
153.	1.279248238	0.051838819	3.54961681	1.	2.	1.
155.	-1.046935081	0.126440823	3.02042484	2.	2.	2.
157.	-0.134461045	0.091147780	3.31418514	2.	2.	2.
159.	0.791216671	0.059640415	3.19047642	1.	2.	1.
161.	4.356127739	0.200334311	3.66612244	1.	2.	2.
163.	1.383487701	0.048721343	3.52636051	1.	2.	1.
165.	-1.727434158	0.049635485	3.24649048	1.	1.	1.
167.	1.813549042	0.080146015	3.19458199	1.	2.	1.
169.	2.526293755	0.072494030	3.52929688	1.	2.	1.
171.	1.639560699	0.056576263	3.66356087	1.	2.	1.
173.	0.221878707	0.138159871	3.88567829	1.	1.	2.

Figure 46: P6D Command File which Generated Figures 33.a-d

```
$empty goutp6d1 ok
$run *bmdp sprint=goutp6d1 7=gplotdat1 par=p6d
/ problem     title is 'GOOKP6D1: plot ACNEST vs. DIAM:  group by
                    sptree, all DI1'.
/ input       unit is 7.
              cases are 1124.
              variables are 7.
              format is '(5x,f5.0,2(1x,f14.9),1x,f14.8,3(1x,f2.0))'.
/ variable    names are recnum,lp,vl,lgdiam,sptree,di1,bktop.
              label is recnum.
              grouping is sptree.
/ group       codes(5)=1,2,3.
              names(5) are aspen,birch,other.
/ plot        yvar is lp.
              xvar is lgdiam.
              group is aspen.
              group is birch.
              group is other.
              groups are aspen,birch,other.
              size=100,40.
              no statistics.
/ end
```

Figure 47: P6D Command File which Generated Figures 34.a-d

```
$empty goutp6d2 ok
$run *bmdp sprint=goutp6d2 7=gplotdat1 par=p6d
/ problem     title is 'GOUTP6D2:  plot ACNEST vs. DIAM:  group by
                       SPTREE, no fungal conks'.
/ input       unit is 7.
              cases are 1124.
              variables are 7.
              format is '(5x,f5.0,2(1x,f14.9),1x,f14.8,3(1x,f2.0))'.
/ variable    names are recnum,lp,vl,lgdiam,sptree,di1,bktop.
              label is recnum.
              grouping is sptree.
/ transform   use=di1 eq 1.
/ group       codes(5)=1,2,3.
              names(5) are aspen,birch,other.
/ plot        yvar is lp.
              xvar is lgdiam.
              group is aspen.
              group is birch.
              group is other.
              groups are aspen,birch,other.
              size=100,40.
              no statistics.
/ end
```

Figure 48: P6D Command File which Generated Figures 35.a-d

```
$empty goutp6d3 ok
$run *bmdp sprint=goutp6d3 7=gplotdat1 par=p6d
/ problem     title is 'GOUTP6D3:  plot ACNEST vs. DIAM:  group by
                       SPTREE, fungal conks present'.
/ input       unit is 7.
              cases are 1124.
              variables are 7.
              format is '(5x,f5.0,2(1x,f14.9),1x,f14.8,3(1x,f2.0))'.
/ variable    names are recnum,lp,vl,lgdiam,sptree,di1,bktop.
              label is recnum.
              grouping is sptree.
/ transform   use=di1 eq 2.
/ group       codes(5)=1,2,3.
              names(5) are aspen,birch,other.
/ plot        yvar is lp.
              xvar is lgdiam.
              group is aspen.
              group is birch.
              group is other.
              groups are aspen,birch,other.
              size=100,40.
              no statistics.
/ end
```

'add=' sentence in the '/ variable' paragraph. The manual (Ref.(4)) neglects to point this out

CHAPTER 8

TECHNICAL SUPPLEMENT FOR CHAPTER 4

A number of observations and developments may be made on the contents of Chpater 4.

8.1 On the Production of Outputs in Figures 40 and 41

Figures 40 and 41 both showed outputs of GLIM runs, which were produced by the command source files in Figures 49 and 50. One can see that these source files are very similar to that of Figure 44 except for the actual model fit requests and the fact that no new data file gets created.

The calculations for drops in scaled deviance are straight forwrard. Using the notation of Section 5.4 in Ref.(2), let model 0 refer to the null model (no explanatory variables), model m refer to the maximal (PLR final) model presented in Chapter 3, and model f refer to the 'full' model, which would have 1124 coefficients (one for each observation in the analysis). Then for Figure 40:

i	Explanatory Variable Added to Model 0	$S(i,f)$	$S(0,i)$ $=S(0,f)-S(i,f)$
1	SPTREE	1057	117
2	ln(DIAM)	1095	79
3	ln(HEIGHT)	1169	5
4	DI1	779.0	395
5	DI2	1137	37
6	DWOOD	1161	13
7	BKTOP	1161	13

Figure 49: GLIM Command File which Generated Figure 40

```
$EMPTY GOUTGLIM3 OK
$RUN UNSP:GLIM 1=GOOKNEST5 2=GOUTGLIM3
$C
$C GLIM RUN ON FILE GOOKNEST5 TESTING OUT LOGISTIC LINK ON BINARY
$C      RESPONSES
$C
$C ************************
$C *  GET DATA & TRANSFORM *
$C ************************
$C
$OUTPUT 2 132
$UNITS 1124
$DATA RECNUM ACNEST SPTREE LNHEIGHT LNDIAM DI1 DI2 DI3 DWOOD BKTOP
$FACTOR SPTREE 3 DI1 2 DI2 2 DI3 2 DWOOD 2 BKTOP 2
$FORMAT
(2X,F8.1,2(1X,F2.0),1X,F5.2,1X,F6.2,5(1X,F2.0))
$DINPUT 1
$LOOK 1 15 RECNUM ACNEST SPTREE LNHEIGHT LNDIAM DI1 DI2 DI3 DWOOD BKTOP
$CALC LNHEIGHT=%LOG(LNHEIGHT): LNDIAM=%LOG(LNDIAM): N=1: DI1=DI1+1
$CALC DI2=DI2+1: BKTOP=BKTOP+1
$C
$C ***************************************
$C *  SPECIFY MODEL TO BE ANALYZED       *
$C ***************************************
$C
$YVAR ACNEST
$ERROR B N
$LINK G
$C
$C ***********************************************
$C *  FIT SINGLE-VARIABLE MODELS & NULL MODEL    *
$C ***********************************************
$C
$FIT %GM
$DISP A
$FIT SPTREE
$DISP A
$FIT LNDIAM
$DISP A
$FIT LNHEIGHT
$DISP A
$FIT DI1
$DISP A
$FIT DI2
$DISP A
$FIT DWOOD
$DISP A
$FIT BKTOP
$DISP A
$STOP
```

Figure 50: GLIM Command File which Generated Figure 41

```
$EMPTY GOUTGLIM4 OK
$RUN UNSP:GLIM 1=GOOKNEST5 2=GOUTGLIM4
$C
$C GLIM RUN ON FILE GOOKNEST5 TESTING OUT LOGISTIC LINK ON BINARY
$C      RESPONSES
$C
$C *************************
$C *  GET DATA & TRANSFORM *
$C *************************
$C
$OUTPUT 2 132
$UNITS 1124
$DATA RECNUM ACNEST SPTREE LNHEIGHT LNDIAM DI1 DI2 DI3 DWOOD BKTOP
$FACTOR SPTREE 3 DI1 2 DI2 2 DI3 2 DWOOD 2 BKTOP 2
$FORMAT
(2X,F8.1,2(1X,F2.0),1X,F5.2,1X,F6.2,5(1X,F2.0))
$DINPUT 1
$LOOK 1 15 RECNUM ACNEST SPTREE LNHEIGHT LNDIAM DI1 DI2 DI3 DWOOD BKTOP
$CALC LNHEIGHT=%LOG(LNHEIGHT): LNDIAM=%LOG(LNDIAM): N=1: DI1=DI1+1
$CALC DI2=DI2+1: BKTOP=BKTOP+1
$C
$C ***************************************
$C *  SPECIFY MODEL TO BE ANALYZED       *
$C ***************************************
$C
$YVAR ACNEST
$ERROR B N
$LINK G
$C
$C ***********************************************************
$C *  FROM GOUTPLR4 OUTPUT & GOUTGLIM1, TRY ALL POSSIBLE    *
$C *     MODELS WHICH HAVE A SINGLE VARIABLE MISSING.       *
$C ***********************************************************
$C
$FIT LNDIAM+DI1+DI2+LNHEIGHT+DWOOD+BKTOP
$DISP A
$FIT SPTREE+DI1+DI2+LNHEIGHT+DWOOD+BKTOP
$DISP A
$FIT SPTREE+LNDIAM+DI2+LNHEIGHT+DWOOD+BKTOP
$DISP A
$FIT SPTREE+LNDIAM+DI1+LNHEIGHT+DWOOD+BKTOP
$DISP A
$FIT SPTREE+LNDIAM+DI1+DI2+DWOOD+BKTOP
$DISP A
$FIT SPTREE+LNDIAM+DI1+DI2+LNHEIGHT+BKTOP
$DISP A
$FIT SPTREE+LNDIAM+DI1+DI2+LNHEIGHT+DWOOD
$DISP A
$STOP
```

where $S(0,f)=1174$.

Similarly, for Figure 41:

i	Explanatory Variable Removed from Model m	$S(i,f)$	$S(i,m)$ $=S(i,f)-S(m,f)$
1	SPTREE	692.7	95.6
2	ln(DIAM)	613.3	16.2
3	ln(HEIGHT)	600.0	2.9
4	DI1	939.3	342.2
5	DI2	624.3	27.2
6	DWOOD	605.0	7.9
7	BKTOP	607.4	10.3

where $S(m,f)=597.1$.

From these values the rankings for the last 2 schemes in Chapter 4 may be confirmed.

8.2 On Comparing Figure 36 Entries with GLIM Equivalent

Figure 51.a shows the command source file for a GLIM run which requests a sequence of model fits identical to that of the PLR run, which was summarized in Figure 36. The corresponding GLIM ouput is shown in Figure 51.b. A table of successive fit results for Figure 51.b similar to those for Figures 40 and 41 in the previous section is now given:

Step i	Explanatory Variable Added to Previous Step	$S(i,f)$	Improvement over Previous Model: $S(i-1,i)$ $=S(i-1,f)-S(i,f)$
0	%GM(Empty Model)	1174	---
1	DI1	779.0	395
2	ln(DIAM)	725.0	54.0
3	SPTREE	662.0	63.0
4	BKTOP	633.3	28.7
5	DI2	606.5	26.8

Figure 51.a: GLIM Command File which Generates Figure 51.b

```
$EMPTY GOUTGLIM5 OK
$RUN UNSP:GLIM 1=GOOKNEST5 2=GOUTGLIM5
$C
$C GLIM RUN ON FILE GOOKNEST5 TESTING OUT LOGISTIC LINK ON BINARY
$C      RESPONSES
$C
$C ************************
$C *  GET DATA & TRANSFORM *
$C ************************
$C
$OUTPUT 2 132
$UNITS 1124
$DATA RECNUM ACNEST SPTREE LNHEIGHT LNDIAM DI1 DI2 DI3 DWOOD BKTOP
$FACTOR SPTREE 3 DI1 2 DI2 2 DI3 2 DWOOD 2 BKTOP 2
$FORMAT
(2X,F8.1,2(1X,F2.0),1X,F5.2,1X,F6.2,5(1X,F2.0))
$DINPUT 1
$LOOK 1 15 RECNUM ACNEST SPTREE LNHEIGHT LNDIAM DI1 DI2 DI3 DWOOD BKTOP
$C
$C          LOG TRANSFORM LNHEIGHT & LNOIAM
$C          SHIFT ALL QUALITATIVE VARIABLES UP BY 1 BECAUSE GLIM
$C              CAN'T HANDLE FACOTR LEVELS OF O
$C
$CALC LNHEIGHT=%LOG(LNHEIGHT): LNDIAM=%LOG(LNDIAM): N=1: DI1=DI1+1
$CALC OI2=DI2+1: BKTOP=BKTOP+1
$C
$C ***************************************
$C *  SPECIFY MOOEL TO BE ANALYZED        *
$C ***************************************
$C
$YVAR ACNEST
$ERROR B N
$LINK G
$C
$C ********************************************************************
$C *  FIT NESTED MODELS IN ORDER SUGGESTED IN GOUTPLR4                 *
$C ********************************************************************
$C
$FIT %GM
$DISP A
$FIT DI1
$DISP A
$FIT DI1+LNDIAM
$DISP A
$FIT DI1+LNDIAM+SPTREE
$DISP A
$FIT DI1+LNDIAM+SPTREE+BKTOP
$DISP A
$FIT DI1+LNDIAM+SPTREE+BKTOP+DI2
$DISP A
$FIT DI1+LNDIAM+SPTREE+BKTOP+DI2+DWOOO
$DISP A
$FIT DI1+LNOIAM+SPTREE+BKTOP+OI2+DWOOO+LNHEIGHT
$OISP A
$STOP
```

Figure 51.b: GLIM Model Fits to Match PLR Sequence in Figure 36

```
      SCALED
CYCLE DEVIANCE      DF
   4    1174.      1123

      ESTIMATE      S.E.       PARAMETER
   1  -1.288      0.7228E-01   %GM
 SCALE PARAMETER TAKEN AS      1.000

      SCALED
CYCLE DEVIANCE      DF
   4    779.0      1122

      ESTIMATE      S.E.       PARAMETER
   1  -2.404      0.1214       %GM
   0    ZERO     ALIASED       DI1(1)
   2   3.407      0.1915       DI1(2)
 SCALE PARAMETER TAKEN AS      1.000

      SCALED
CYCLE DEVIANCE      DF
   4    725.0      1121

      ESTIMATE      S.E.       PARAMETER
   1  -10.36       1.147       %GM
   0    ZERO     ALIASED       DI1(1)
   2   3.439      0.2026       DI1(2)
   3   2.369      0.3310       LNDI
 SCALE PARAMETER TAKEN AS      1.000

      SCALED
CYCLE DEVIANCE      DF
   5    662.0      1119

      ESTIMATE      S.E.       PARAMETER
   1  -7.870       1.288       %GM
   0    ZERO     ALIASED       DI1(1)
   2   3.591      0.2270       DI1(2)
   3   1.771      0.3718       LNDI
   0    ZERO     ALIASED       SPTR(1)
   4  -1.860      0.2819       SPTR(2)
   5  -2.272      0.7976       SPTR(3)
 SCALE PARAMETER TAKEN AS      1.000

      SCALED
CYCLE DEVIANCE      DF
   5    633.3      1118

      ESTIMATE      S.E.       PARAMETER
   1  -8.973       1.346       %GM
   0    ZERO     ALIASED       DI1(1)
   2   3.594      0.2326       DI1(2)
   3   2.041      0.3859       LNDI
   0    ZERO     ALIASED       SPTR(1)
   4  -2.572      0.3417       SPTR(2)
   5  -2.784      0.8616       SPTR(3)
   0    ZERO     ALIASED       BKTO(1)
   6   1.525      0.2820       BKTO(2)
 SCALE PARAMETER TAKEN AS      1.000
```

Figure 51.b, continued

CYCLE	SCALED DEVIANCE	DF
5	606.5	1117

	ESTIMATE	S.E.	PARAMETER
1	-8.759	1.377	%GM
0	ZERO	ALIASED	DI1(1)
2	3.685	0.2412	DI1(2)
3	1.848	0.3953	LNDI
0	ZERO	ALIASED	SPTR(1)
4	-2.787	0.3580	SPTR(2)
5	-2.992	0.8947	SPTR(3)
0	ZERO	ALIASED	BKTO(1)
6	1.621	0.2931	BKTO(2)
0	ZERO	ALIASED	DI2(1)
7	1.125	0.2197	DI2(2)
SCALE PARAMETER TAKEN AS			1.000

CYCLE	SCALED DEVIANCE	DF
5	600.0	1116

	ESTIMATE	S.E.	PARAMETER
1	-9.481	1.422	%GM
0	ZERO	ALIASED	DI1(1)
2	3.710	0.2438	DI1(2)
3	2.015	0.4047	LNDI
0	ZERO	ALIASED	SPTR(1)
4	-2.888	0.3642	SPTR(2)
5	-3.036	0.8903	SPTR(3)
0	ZERO	ALIASED	BKTO(1)
6	1.005	0.3696	BKTO(2)
0	ZERO	ALIASED	DI2(1)
7	1.134	0.2203	DI2(2)
0	ZERO	ALIASED	DWOO(1)
8	0.8271	0.3180	DWOO(2)
SCALE PARAMETER TAKEN AS			1.000

CYCLE	SCALED DEVIANCE	DF
5	597.1	1115

	ESTIMATE	S.E.	PARAMETER
1	-10.04	1.469	%GM
0	ZERO	ALIASED	DI1(1)
2	3.721	0.2445	DI1(2)
3	1.734	0.4365	LNDI
0	ZERO	ALIASED	SPTR(1)
4	-2.902	0.3676	SPTR(2)
5	-3.053	0.8979	SPTR(3)
0	ZERO	ALIASED	BKTO(1)
6	1.325	0.4180	BKTO(2)
0	ZERO	ALIASED	DI2(1)
7	1.139	0.2211	DI2(2)
0	ZERO	ALIASED	DWOO(1)
8	0.9389	0.3260	DWOO(2)
9	0.5136	0.3051	LNHE
SCALE PARAMETER TAKEN AS			1.000

6	DWOOD	600.0	6.5
7	ln(HEIGHT)	597.1	2.9

Comparing these entries with those of Figure 36, it will be noticed that the improvement $S(i-1,i)$ is very nearly the same as the improvement χ^2 score for the same expalanatory variable, and also that the $S(i,f)$ differs from the goodness-of-fit χ^2 score by about 3 for the same explanatory variable. These are worthy of further consideration.

First of all, the 2 'improvement' scores should be exactly the same, with accumulated round-off errors and differences in the efficiencies of the respective numerical algorithms accounting for any discrepancies. The reason for the equality is that if, say, model i is nested in model j then the GLIM improvement measure is, according to Section 5.2 of Ref.(2):

$$S(i,j)=-2ln\left(\frac{L}{L}i\right)$$

where L_i is the likelihood function evaluated for the parameter estimates for model i, and L_j the likelihood function evaluated for parameter estimates for model j. According to page 683 of Ref.(4), this is also how PLR computes its improvement χ^2 score.

The goodness-of-fit measures, however, seem to be computed differently. GLIM uses a deviance formula (Ref.(8) pg. 25) but PLR does not reveal how its goodness-of-fit is computed, although the manual (Ref.(4)) hints on page 333

that the usual Pearson statistic for cell frequency counts is used. Certainly the observation that the two goodness-of-fit measures seem to differ always by 3 is of interest.

CHAPTER 9

TECHNICAL SUPPLEMENT FOR CHAPTER 5

On the contents of Chapter 5, the following observations and developments may be made.

9.1 On the Estimation of Future Log-Odds and their Variances

The linear predictor, in this analysis, occurs on the log-odds scale:

$$\eta=\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p=\underline{x}'\underline{\beta}$$

and the vector of estimated coefficients is $\hat{\underline{\beta}}$, estimated either by PLR or GLIM. For a future tree and its associated vector of explanatory variable values, $\underline{x}_0$, one can estimate the corresponding log odds:

$$\hat{\eta}=\underline{x}_0'\hat{\underline{\beta}}$$

and its associated variance:

$$s^2(\hat{\eta})=\underline{x}_0'\hat{Cov}(\hat{\underline{\beta}})\underline{x}_0$$

Now the exact variance would be:

$$Var(\hat{\eta})=\underline{x}_0'Cov(\hat{\underline{\beta}})\underline{x}_0$$

where $Cov(\hat{\underline{\beta}})$ is the true dispersion or variance-covariance matrix of $\hat{\underline{\beta}}$. But this itself must usually be estimated from the sample data by $\hat{Cov}(\hat{\underline{\beta}})$.

Computing the inner product $\hat{\eta}=\underline{x}_0'\hat{\underline{\beta}}$ and the quadratic form $\underline{x}_0'\hat{Cov}(\hat{\underline{\beta}})\underline{x}_0$ can easily be done with the MINITAB software package (Ref.(12)) as is shown in Figures 52.a and

52.b, where Figure 52.a shows the source file to produce the run, and Figure 52.b shows the input vector $\underline{x}_0$, and output results for the run. The actual input vector was originally stored in another file which is not reproduced here since its contents already appear in Figure 52.b. The following features of Figure 52.a will be noticed.

Firstly, the $\hat{Cov}(\underline{\hat{\beta}})$ matrix used is that from the PLR run, not from the GLIM run, since PLR does the computation in double precision, whereas the precision used in GLIM is not revealed in the manual. In addition the PLR program is specialized for logistic regression where GLIM is designed for greater generality, so it was felt that the PLR results were more reliable. It was noted earlier, in Section 7.2, that there are some differences between PLR and GLIM in the final reported results for both $\underline{\hat{\beta}}$ and $\hat{Cov}(\underline{\hat{\beta}})$.

Secondly, the $\underline{x}_0$ vector must be in the format:

$\underline{x}_0$ =
SPTREE(1)
SPTREE(2)
ln(HEIGHT)
ln(DIAM)
DI1
DI2
DWOOD
BKTOP
1.0

where the '1.0' is for the constant term, $\hat{\beta}_0$, in the $\underline{\hat{\beta}}$ vector. This order is imposed because of the row/column order of $\underline{\hat{\beta}}$ and $\hat{Cov}(\underline{\hat{\beta}})$ from the PLR output.

Figure 52.a: MINITAB Command File which Generates Figure 52.b

```
$empty gookmtbout ok
$run *minitab sprint=gookmtbout
noecho
#
##################################################
#  set up vector of regression coefficients  #
##################################################
#
set c1
-2.9036 -3.0541 0.51375 1.7341 3.7217 1.1396 0.93930 1.3251 -10.039
name c1 'coeff'
#
#######################################################
#  set up column vectors for dispersion matrix  #
#######################################################
#
set c2
0.13905 0.04935 -0.00438 0.01098 -0.03761 -0.01484 -0.01827 -0.04355
-0.01064
set c3
0.04935 0.82897 -0.00608 -0.02309 -0.02440 -0.01429 -0.01082 -0.04328
0.10470
set c4
-0.00438 -0.00608 0.09400 -0.04952 0.00477 0.00191 0.02017 0.05833
-0.11007
set c5
0.01098 -0.02309 -0.04952 0.19326 0.01006 -0.00674 0.01372 -0.02414
-0.52416
set c6
-0.03761 -0.02440 0.00477 0.01006 0.06105 0.01254 0.00988 0.00993
-0.07489
set c7
-0.01484 -0.01429 0.00191 -0.00674 0.01254 0.04953 0.00271 0.00669
-0.00725
set c8
-0.01827 -0.01082 0.02017 0.01372 0.00988 0.00271 0.10760 -0.06058
-0.12929
set c9
-0.04355 -0.04328 0.05833 -0.02414 0.00993 0.00669 -0.06058 0.17681
-0.09419
set c10
-0.01064 0.10470 -0.11007 -0.52416 -0.07489 -0.00725 -0.12929 -0.09419
2.19295
copy c2-c10 to m1
echo
#
#####################################################
#  here is vector of coefficients from GOUTPLR4  #
#####################################################
#
noecho
print 'coeff'
echo
#
###########################################################
#  here is sample dispersion matrix for above vector  #
###########################################################
#
noecho
print m1
#
#################################
#  get an estimation vector  #
#################################
```

Figure 52.a, continued

```
#
noecho
read 'gookmtbin' c11
echo
#
########################################
#  For the following input vector  #
########################################
#
noecho
print c11
#
#############################################################
#  find estimated future mean of linear predictor &  #
#     corresponding estimated variance                #
#############################################################
#
trans c11 put m2
echo
#
#########################################################
#  this is estimated future mean linear predictor  #
#########################################################
#
noecho
mult m2 'coeff' put k1
mult m2 m1 put m3
echo
#
#####################################
#  & here is estimated variance  #
#####################################
#
noecho
mult m3 c11 put k2
stop
```

Figure 52.b: MINITAB Run to Find Estimated Log-Odds and its Associated Variance

```
MTB > #
MTB > ##########################################################
MTB > #  here is vector of coefficients from GOUTPLR4  #
MTB > ##########################################################
MTB > #
coeff
  -2.9036    -3.0541     0.5137     1.7341     3.7217     1.1396     0.9393     1.3251
 -10.0390

MTB > #
MTB > ##############################################################
MTB > #  here is sample dispersion matrix for above vector  #
MTB > ##############################################################
MTB > #
MATRIX M1

 0.13905  0.04935 -0.00438  0.01098 -0.03761 -0.01484 -0.01827 -0.04355
 0.04935  0.82897 -0.00608 -0.02309 -0.02440 -0.01429 -0.01082 -0.04328
-0.00438 -0.00608  0.09400 -0.04952  0.00477  0.00191  0.02017  0.05833
 0.01098 -0.02309 -0.04952  0.19326  0.01006 -0.00674  0.01372 -0.02414
-0.03761 -0.02440  0.00477  0.01006  0.06105  0.01254  0.00988  0.00993
-0.01484 -0.01429  0.00191 -0.00674  0.01254  0.04953  0.00271  0.00669
-0.01827 -0.01082  0.02017  0.01372  0.00988  0.00271  0.10760 -0.06058
-0.04355 -0.04328  0.05833 -0.02414  0.00993  0.00669 -0.06058  0.17681
-0.01064  0.10470 -0.11007 -0.52416 -0.07489 -0.00725 -0.12929 -0.09419

-0.01064
 0.10470
-0.11007
-0.52416
-0.07489
-0.00725
-0.12929
-0.09419
 2.19295


     9 ROWS READ
MTB > #
MTB > ########################################
MTB > #  For the following input vector  #
MTB > ########################################
MTB > #
C11
  0.000    0.000    2.914    3.645    1.000    0.000    0.000    0.000    1.000

MTB > #
MTB > ##########################################################
MTB > #  this is estimated future mean linear predictor  #
MTB > ##########################################################
MTB > #
   ANSWER =          1.5006
MTB > #
MTB > ###################################
MTB > #  & here is estimated variance  #
MTB > ###################################
MTB > #
   ANSWER =          0.0566

*** MINITAB *** STATISTICS DEPT * PENN STATE UNIV. * RELEASE 82.1 *
STORAGE USED          624 STORAGE AVAILABLE       261744
```

9.2 On Estimating and Predicting μ

From the link function of Chapter 3:

$$\eta = g(\mu) = ln\left(\frac{\mu}{1-\mu}\right)$$

and its associated inverse:

$$\mu = Prob\{\text{Success}\} = g^{-1}(\eta)$$

$$= (1+e^{-\eta})^{-1}$$

one could then produce a fitted probability $\hat{\mu}$ from the fitted log-odds, $\hat{\eta}$, thus:

$$\hat{\mu} = (1+e^{-\hat{\eta}})^{-1}$$

and this would be the '$\hat{\eta}$ as estimate' case, mentioned in Chapter 3, that is, $\hat{\mu}$ estimates the mean probability μ of success for the population of all deciduous trees of same species, height, diameter, and values of DI1, DI2, DWOOD, and BKTOP. Whether this estimate is unbiased for μ or not is unsure, as is the question of unbiasedness of $\hat{\eta}$ for η.

The question of unbiasedness, however, is not important when using $\hat{\eta}$ to produce a prediction, μ_P for μ. The table in Ref.(9) mentioned earlier provides values of μ_P given $\hat{\eta}$ and $s^2(\hat{\eta})$. It is suggested that a prediction of μ for individual trees rather than a population estimate will be more useful to a field worker since prediction takes both $\hat{\eta}$ and $s^2(\hat{\eta})$ into account, whereas the estimate, $\hat{\mu}$, is a function only of $\hat{\eta}$.

Otherwise one could have the following scenario. Suppose for two vectors of explanatory variable values, $\underline{x}_1$ and $\underline{x}_2$:

$$\underline{x}_1 \neq \underline{x}_2$$

but

$$\hat{\eta}_1 = \hat{\eta}_2$$

where

$$\hat{\eta}_i = \underline{x}_i' \hat{\underline{\beta}}$$

for $i=1,2$. Then:

$$\hat{\mu}_1 = \hat{\mu}_2$$

But if:

$$s^2(\hat{\eta}_1) \neq s^2(\hat{\eta}_2)$$

then the corresponding μ_P values will also be different. In fact, the $\underline{x}_i$ vector having the higher $s^2(\hat{\eta}_i)$ will give a μ_P-value closer to 0.5 (so long as their $\hat{\eta}_i$ are equal). μ_P is thus in this sense, 'safer' than $\hat{\mu}$ as a fitted value for μ.

APPENDIX A

A TABLE OF PREDICTIVE SUCCESS PROBABILITIES FOR LOGISTIC REGRESSION

S. G. Meester and D. Eaves

Simon Fraser University
Burnaby, B.C.
Canada

Key Words and Phrases: dichotomous data; logistic regression; predictive probability.

ABSTRACT

A table of expected success rates under normally distributed success logit, used in conjunction with logistic regression analysis, enables easy calculation of expected win for betting on success of a future dichotomous trial.

INTRODUCTION

We wish to calculate a predictive probability of success under specified values $a_1, a_2, \ldots, a_k$ of k independent variables which might influence success probability. For example in the simple linear logistic regression case, $k=2$, $a_1 = 1$ and a_2 is the specified value of the independent variable.

The past information upon which we may base the calculation consists of an observed sequence of n successes and failures, each having occurred under a known set of values of the independent variables. Standard logistic regression (Cox, 1970) fits the model:

Success logit = $x_i'\beta$ $\quad i=1,2,\ldots,n$

where the row vector x_i' represents the independent variable values associated with observation i. This fitting produces a M.L.E. $\hat{\beta}$ and an estimated covariance matrix $\hat{\Sigma}$ for β.

It is conceptually convenient (but otherwise unnecessary) to interpret this in the Bayesian framework, as producing a normal approximation to the posterior Lebesgue density of β, as p-variate with posterior mean $\hat{\beta}$ and covariance $\hat{\Sigma}$. This idea is discussed in DeGroot (1970), Chapter 10. The posterior distribution of the success logit, θ, given $a = (a_1,\ldots,a_k)'$ may therefore be taken as univariate approximately normal with mean $m = a'\hat{\beta}$ and variance $s^2 = a'\hat{\Sigma}a$, which are easily calculated from the outputs of standard programs such as GLIM or BMD PLR. Since it is generally reasonable to assume that subsequent success/failure is conditionally independent of past data given θ, it follows readily that for a similar future Bernoulli trial,

$$\text{Prob[success}\,|\,\text{data]} = \int_{\Theta} [1 + \exp(-\theta)]^{-1} f(\theta\,|\,m,s)\,d\theta$$

where f is the normal density with mean m and variance s^2.

With the view of facilitating calculation of success probability on a future trial, values of this integral are given in table I. Note that as s increases, the integral shrinks from $[1+\exp(-m)]^{-1}$ toward 0.5. Also only values for $m > 0$ are given. If $m < 0$ then enter the table at $|m|$ and use 1-(value from the table for Prob[success|data]).

The production of this integral is analogous to the calculation of the predictive mean and standard error for a single future numerical observation associated with a classical regression model, since these two figures are the location and scale parameters of the predictive t-distribution. The amount of uncertainty of the logistic prediction is reflected in the amount of pulling toward 0.5 of the fitted success probability.

TABLE I

VARIANCE

MEAN	0	1	2	5	10	15	20	40
0.1	.5250	.5206	.5181	.5141	.5110	.5094	.5083	.5061
0.2	.5498	.5412	.5363	.5282	.5220	.5187	.5166	.5121
0.4	.5987	.5820	.5722	.5563	.5440	.5374	.5331	.5243
0.6	.6457	.6218	.6075	.5841	.5659	.5560	.5496	.5364
0.8	.6900	.6601	.6419	.6114	.5876	.5745	.5660	.5484
1.0	.7311	.6967	.6751	.6383	.6089	.5928	.5823	.5604
1.5	.8176	.7785	.7513	.7020	.6608	.6376	.6223	.5902
2.0	.8808	.8445	.8161	.7599	.7097	.6806	.6611	.6195
2.5	.9241	.8946	.8684	.8109	.7550	.7213	.6982	.6481
3.0	.9526	.9307	.9086	.8544	.7962	.7593	.7334	.6759
4.0	.9820	.9719	.9594	.9193	.8649	.8259	.7969	.7285
5.0	.9933	.9892	.9834	.9590	.9156	.8794	.8503	.7765
6.0	.9975	.9960	.9935	.9807	.9503	.9200	.8934	.8192
8.0	.9997	.9994	.9991	.9965	.9855	.9692	.9514	.8881
10.0	.9999	.9999	.9999	.9995	.9966	.9901	.9808	.9358

BIBLIOGRAPHY

Cox, D.R. (1970) Analysis of Binary Data. London: Methuen & Co. Ltd.

DeGroot, Morris, H. (1970) Optimal Statistical Decisions. New York: McGraw-Hill, Inc.

BIBLIOGRAPHY

1. Audette, Reo. BMDP User's Guide SFU Computing Centre, Revised Nov. 1982.

2. Baker, R.J. and Nelder, J.A. The GLIM System, Release 3: Generalized Linear Interactive Modelling. Oxford, England: Numerical Algorithms Group (NAG), 1978.

3. Box, George E.P. "Use and Abuse of Regression", Technometrics, Vol. 8 (1966), pp. 625-9.

4. Dixon, W.J. (Chief Editor) BMDP Statistical Software, 1985 Printing. Berkeley, California: University of California Press Ltd., 1983.

5. Draper, Norman R. and Smith, Harry. Appled Regression Analysis. 2nd ed. New York: John Wiley and Sons Inc., 1981.

6. Johnson, Richard A. and Wichern, Dean W. Appled Multivariate Statistical Analysis. Englewood Cliffs, N.J.: Prentice-Hall Inc., 1982.

7. Kreitzberg, Charles B. and Shneiderman, Ben. FORTRAN Programming: A Spiral Approach. New York: Harcourt Brace Jovanovich, 1975.

8. McCullagh, P. and Nelder, J.A. Generalized Linear Models. London, England: Chapman and Hall, 1983.

9. Meester, S. and Eaves, D. "A Table of Predictive Success Probabilities for Logistic Regression", in press.

10. Miller, Robert B. and Wichern, Dean W. Intermediate Business Statistics. New York: Holt, Rinehart and Winston, 1977.

11. Neter, John, Wasserman, William and Kutner, Michael H. Applied Linear Statistical Models. 2nd ed. Homewood, Illinois: Richard D. Irwin Inc., 1985.

12. Ryan, Thomas A. Jr., Joiner, Brian L. and Ryan, Barbara F. MINITAB Reference Manual. Uniersity Park, Pa.: Pennsylvannia State University, 1982.

13. Seber, G.A.F. Linear Regression Analysis. New York: John Wiley and Sons, 1977.

14. Tukey, John W. Exploratory Data Analysis. Reading, Mass: Addison-Wesley Pub. Co., 1977.

15. Weldon, K.L. Conversation 26 September 1985.