1-1-1999

Analysis of increases in fishing power in the western rock lobster (Panulirus cygnus) fishery

John Fernandez
Edith Cowan University

Edith Cowan University

Copyright Warning

USE OF THESIS

The Use of Thesis statement is not included in this version of the thesis.

ANALYSIS OF INCREASES IN FISHING POWER IN THE WESTERN ROCK LOBSTER (PANULIRUS CYGNUS) FISHERY

BY

JOHN FERNANDEZ

A thesis submitted in partial fulfilment of the requirements for the degree of Master of Science (mathematics and planning) at the School of Engineering and Mathematics, Edith Cowan University, Western Australia.

Date of submission: 22nd March, 1999

ABSTRACT

The western rock lobster, *Panulirus cygnus*, fishery represents a significant commercial asset to Western Australia, and it is therefore important that appropriate strategies are developed to effectively manage it. Because the fishery has a very high level of exploitation, researchers and managers rely significantly on annual stock assessments which are based on catch and effort data. This study will identify and assess the effects that changes in fishing power factors (e.g. advances in fish-finding technology) have had on estimates of catch and effort. The fishing power increases can be used to adjust nominal fishing effort to produce a time series of standardised effort which can then be used to reassess stock abundance measures, particularly of the breeding stock. The study will utilise the theory and techniques of regression and generalised linear modelling. A comparison of the normal and gamma distributions as the specified probability distribution in the model will be made.

DECLARATION

I certify that this thesis does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any institution of higher education; and that to the best of my knowledge and belief it does not contain any material previously published or written by another person except where due reference is made in the text.

Signature 

Date 10 NOVEMBER, 1999

DECLARATION

I certify that this thesis does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any institution of higher education; and that to the best of my knowledge and belief it does not contain any material previously published or written by another person except where due reference is made in the text.

Signature 

Date 10 NOVEMBER, 1999

ACKNOWLEDGEMENTS

I express thanks to Dr James Cross who supervised my research with wisdom and patience and to the staff of Edith Cowan University who helped me in many ways. I thank Dr Nick Caputi who guided me continually with the statistical aspects of the research, and I acknowledge the help of the staff at the Western Australian Marine Research Laboratories, especially Wilf Lehre.

I am very grateful to my wife, Marie, who has supported me in countless ways throughout the course of this research, and to whom I dedicate this work.

TABLE OF CONTENTS

Chapter 1

INTRODUCTION

1.1 RATIONALE

The western rock lobster (WRL) fishery is situated along the west coast of Western Australia from Fremantle in the south, to just north of Kalbarri (see the map in Appendix A). As Australia's largest single-species fishery it constitutes a significant commercial asset for Western Australia (valued at over $300m annually). Hence, it is necessary that appropriate management strategies be developed and maintained. There are essentially two reasons for this: (i) Economic - there is a need to maintain sustainable catch rates to ensure acceptable commercial yields; (ii) Biological - there is a possibility of over-exploitation of the species, which would be both biologically and economically undesirable. There has been a decline in the level of the spawning stock of the WRL in the early 1990's, and the fishery was considered to be fully exploited with an exploitation rate exceeding 85% from recruitment to fishery and an annual exploitation rate exceeding 60% (Phillips & Brown, 1989; Bowen & Hancock, 1989). The maintenance of the economic and environmental balance of the fishery necessitates that the analyses and modelling of the available data collected from the fishery are as accurate and reliable as possible.

The main source of data with which analysis and modelling has been done in the WRL fishery is the catch and effort data collected from the commercial fishery. The catch is defined as the weight (in kilograms) of lobsters in the lobster traps (also called pots) that

are lifted from the ocean on a day, and the associated effort is taken to be the actual number of lobster traps lifted in order to yield that catch. These catch and effort data are particularly important for estimating the stock abundance, which plays a significant role in the management policy of the fishery. If the estimates of fishing effort are inaccurate then the estimates of stock abundance could be inaccurate, potentially leading to poor management decisions.

A potential source of inaccuracy in these estimates is the changing fishing power of the fleet. Fishing power is herein taken to be the level of ability and efficiency of a vessel (or the whole fleet) to catch lobsters. It can be measured by the catch per unit effort (CPUE), which, in this research, is calculated by dividing the weight of the daily catch by the number of pots lifted. CPUE can also be calculated using other denominators that are relevant and meaningful. Trawl-based fisheries, for example, generally use the number of hours trawled to obtain CPUE. Some factors that may affect the fishing power of a vessel, and hence of the fleet, are vessel length, the type of fishing traps used, and the presence of advanced, onboard technological equipment. The changes in fishing power factors have been telling in recent years, mainly owing, it is thought, to advances in onboard technology, and have lead to an obvious increase in the fishing efficiency of the fleet. It is anticipated that these sorts of factors will play the major role in this study, and it is the increase in efficiency associated with them, together with its implications, that primarily motivates this research. The effective fishing effort is influenced by such changes in efficiency and should be adjusted accordingly. This adjustment is based on the percentage increase in catch rates associated with each fishing power factor and the percentage of vessels in the fleet that make use of those factors for each season. This is important because not only has there been the emergence of more efficient equipment, but also more vessels are making use of that equipment. The incorporation of fishing power

into the catch and effort statistics will help to ensure that the stock abundance models are more accurate and reliable.

The changes in the use of various fishing power factors are identified and the catch rates are modelled in order to estimate the efficiency increases associated with these factors. A requirement of the modelling process is that there needs to be adequate numbers of vessels with contrasting fishing power characteristics for the parameter estimates to be reliable. For example, in order to estimate the increase in efficiency associated with the use of an onboard global positioning system (GPS), the catch rates of vessels with a GPS is compared to the catch rates of vessels without a GPS, and so there needs to be adequate numbers of vessels with and without a GPS. This may have the effect of rendering some seasons unsuitable in the modelling process. Hence, the first task is to obtain the time series of each fishing power factor under consideration and then determine which seasons are most appropriate for estimating its fishing power effect. Once the fishing power for each factor has been estimated it can then be applied to the estimates of fishing effort in the fishery for each year so that the stock assessment models incorporate the increases in fishing power.

The underlying probability distribution for catch and effort models has historically been assumed lognormal. The underlying distribution in the model affects the parameter estimates, and an incorrect specification of it may have serious consequences for the model's interpretation (Campbell, in prep.). Gulland (1956) gives empirical evidence that the variance increases with the mean in the fishing power of the North Sea trawlers, and that the corresponding logarithmically transformed data shows an approximately constant variance. It also asserts that the transformed data tend to be normally distributed. Kimura (1981) states that Gulland's findings are consistent with the assumption that the variance

is proportional to the square of the mean and is successfully stabilised by the logarithmic transformation. This seems to be supported by Caputi (in prep.) which states that the lognormal distribution is required when modelling catch rate data to reduce the extent of such a mean-to-variance relationship. In theory, the method of estimating the parameters associated with the fishing power factors in a generalised linear model, in which the identity link function is used and the lognormal distribution is assumed, is equivalent to estimating the parameters of an analogous multiplicative model, in which the log link function is used and the gamma distribution is assumed. There is a close connection between linear models with constant variance for log Y and multiplicative models with constant coefficient of variation for Y (See McCullagh and Nelder, 1983, pp 149-150, 156). Hence, this study will supplement its main purpose of identifying and analysing increases in fishing power by validating the assumptions of normality and constant variance using generalised linear modelling with the gamma distribution. The variance function will be analysed and the normality assumption will be tested by performing various residual analyses and by comparing the lognormal-based estimators with estimators based on the gamma distribution. The gamma distribution is additionally appropriate because it provides for data that is non-negative only, such as catch and effort data. In practice, the major thrust of the work will be based on the normal distribution with a logarithmic transformation, constructed as per the traditional linear multiple regression approaches to catch and effort models; whereas the validation analyses based on the gamma distribution will need to draw on the theory of generalised linear models.

1.2 RESEARCH OBJECTIVES

This objectives of this research are to:

1. Identify the changes in various fishing power factors in the WRL fishery.

2. Estimate the efficiency increases associated with the changes in fishing power.

3. Compare the specification of the normal and gamma distributions as the underlying probability distribution.

1.3 SIGNIFICANCE OF THE RESEARCH

If the fishing effort estimates are inaccurate or biased (e.g. owing to increases in fishing power) then the abundance estimates obtained from catch rates could be seriously flawed and possibly lead to poor management decisions. If there are increases in fishing power then a unit of fishing effort will catch a larger proportion of the stock, and if this is not accounted for the abundance estimates will be overestimated. Brown *et al.* (1995) showed that previous estimates of breeding stock abundance were overly optimistic, and clearly identified the need for further investigation of the effect of fishing power. This research is necessary to address the current need for a more rigorous analysis of the changes in fishing power, the assumption of normality, and their effect on fishing effort, catch rates and stock assessment.

1.4 DATA SOURCES

Data for the research was obtained from the Western Australian Marine Research Laboratories (WAMRL). In particular, the following sources at WAMRL were used:

Voluntary research log books. These have been completed by 20% to 30% of the fleet since 1964/65 and contain daily catch and fishing effort data, such as the number of traps lifted and soak-time (number of days the trap has been in the water), for legal-size, under-size and spawning lobsters, by 10' latitude transects and 5 depth zones.

Vessel, gear and equipment interviews. Interviews were held with fifty fishers throughout the fishery to obtain detailed information on changes that had occurred with their vessels, gear and technology between 1971/72 and 1989/90. Since these fishers had also been completing voluntary research log books, there are time series of catch and effort data that are synchronous with these changes.

Gear and equipment returns. Since 1989/90 fishers have been required to submit an annual form that shows any gear and technology changes made to their vessels. These give an indication of the presence or absence of various pieces of gear and equipment onboard each vessel. This database provided a larger sample of vessels than did the interviews database. Because this database was partially incomplete, information was interpolated when possible. If the forms indicated that a vessel had a piece of equipment, such as a GPS, in one season and again in a subsequent season, then it was assumed that the vessel had the equipment for all of the intervening seasons.

1.5 STRUCTURE OF THE THESIS

The dissertation will begin with some background of the WRL fishery and a review of the literature relevant to fishing power, standardisation of fishing effort and relevant aspects of generalised linear modelling in Chapter 2. The next two chapters set out the theoretical framework within which will lie the major part of the analyses in this thesis. Chapter 3 provides the background of the multiple linear regression techniques that are relevant to this study; and Chapter 4 outlines the theory of generalised linear models, its origins, assumptions, components, and inferential procedures. In Chapter 5 is presented the time series of changes in prevalence of the fishing power factors, in which we expect to see trends indicating, for example, increasing use of new technologies. Also in Chapter 5 is the application of the statistical methods of regression and generalised linear models to the catch and effort data, along with their parameter estimates (and thus the estimated fishing power effects) for the fishing power factors under consideration. Chapter 6 compares the two methods and their estimates and concludes the dissertation.

Chapter 2

LITERATURE REVIEW

2.1 FISHING POWER AND STANDARDISATION OF EFFORT

The motivation for standardising catch and effort data in stock assessments has mainly been to account for differences between vessels pertaining to fishing efficiency. Frameworks and techniques for standardisation were introduced by Gulland (1956) and Beverton and Holt (1957). These early papers used a process of standardisation that related vessel characteristics to a "standard vessel". It compared CPUE of the "standard vessel" with vessels that had all fished together a number of times. The relative fishing of the vessels were then obtained. Gulland's paper was significant for using an analysis of variance (ANOVA) model for log(CPUE) in the English demersal fisheries. (CPUE was defined as the total catch divided by the hours fished.) These methods were further developed to include least squares regression analysis in Robson (1966). It shows that commercial catch statistics may be converted to catch per standard unit of effort by estimating and adjusting for the effects of fishing power factors such as tonnage, skipper, age of vessel, and location of fishing grounds. Parrish and Keir (1959) discussed the relationship between measurable vessel characteristics and fishing power through the techniques of correlation and ANOVA. Pope (1975) gives several examples of using linear regression to estimate fishing power factors that are continuous in nature, such as vessel length. Parsons *et al.* (1976) used the simple and intuitive model,

$$U_1 = U_2 P, \tag{2.1}$$

where U_1 and U_2 are the catch rates of categories 1 and 2, respectively, and P is the fishing power of category 1 relative to category 2. The model was used for standardising effort based on relative fishing power for stock assessment, and is a simplification of Robson's (1966) methodology. Gavaris (1980) used Robson's work to develop a multiplicative model to estimate catch rate and effort. The model is

$$U = U_R \prod_{ij} (P_{ij}^{X_{ij}}) \tag{2.2}$$

where U is the catch rate; U_R is the catch rate for the particular combination of categories chosen as the reference; P_{ij} is the relative "power" of category j in category type i; and X_{ij} is 1 when category j occurs, and 0 otherwise.

A log-linear regression model for CPUE was used to standardise measures of relative abundance in two trawl-fished populations of Pacific ocean perch was developed by Kimura (1981). It modelled the effect that technological improvements, such as echo sounders, had on catch rates and efficiency, and has as a basic assumption that CPUE is proportional to abundance. It extended an accepted form for the catch equation, $c_{ij} = qf_{ij}N_i$, where q is the catchability coefficient and N_i is the average abundance during year i, so that it included relative efficiency. The equation for adjusted CPUE was given as

$$U_i^s = \sum_j c_{ij} / [\sum_j f_{ij} re(i, j \mid s)] \approx q_s N_i \tag{2.3}$$

where $re(i,j|s) = q_{ij} / q_s$ is the efficiency of vessel j in year i relative to some standard vessel s. The variables influencing efficiency were then modelled into the q_{ij} and multiple regression was used to estimate its respective coefficients. The study showed that technological improvements increased fishing efficiency particularly when abundance levels where relatively low. The model was expanded by Stocker and Fournier (1984), in

which forecast catch levels were improved by adjusting for vessel characteristics. Indices of abundance were also estimated by standardising catch and effort using a regression model by Allen and Punsly (1984). The model used for standardised catch rates was of the form

$$C = M + A_i + B_j + \ldots + F_k \qquad (2.4)$$

where M is the mean and A_i, B_j,..., F_k are factors which influence the catch rate. The data were transformed by log(catch rate + constant) and the regression model examined the main effects, first-order interactions and covariates. Similarly, Large (1992) used a multiplicative model to estimate abundance from CPUE data, and included interactions between the effects of year and other factors. It states that changes in fishing power can be included in the model and evaluated by examining the variation in CPUE explained by the interaction between vessel/vessel group and year. Caputi (in prep.) considers vessel characteristics, gear and equipment, and other factors that may affect catch rates should be examined when analysing catch and effort data. The impact of the global positioning system and associated plotter systems on the relative fishing power of the northern prawn fishery fleet in the tiger prawn fishery in Australia was investigated by Robins *et al.* (1996) and estimated to be about 12% when fishers had three years of experience with the equipment. This study highlighted the continuing gain in efficiency as a result of storing GPS information on a computer.

In the WRL fishery the spatial and temporal dynamics were modelled by Walters *et al.* (1993) in which the seasonal nature of moulting, recruitment, migration, fishing seasons and fishing effort were taken into account. Caputi *et al.* (1995) indicated that the effect of increases in fishing power still needed to be assessed in the prediction of catches based on indices of puerulus and juvenile abundance. Nominal fishing effort (number of traps

lifted) was used to adjust for increases in fishing power to analyse the relationship between spawning stock, environment, recruitment and fishing effort in the WRL fishery (Caputi *et al.*, 1993). A preliminary assessment of the increases in fishing power on stock assessment and fishing effort expended in the WRL fishery was undertaken by Brown *et al.* (1995). It examined the trends in the use of various components of fishing power and the effects on catch rates. To evaluate the impact of technology changes on catch rates, an ANOVA was used to account for some of the main factors thought to influence catch rates. The residuals of the ANOVA were then used in a regression model to test the effects of the fishing power factors. Results indicated that, in deeper waters during the sedentary period of the fishery (February to June), the increase in catch rates associated with using a colour echo sounder and GPS ranged from 13% to 17%. The fishing power increases were used to standardise nominal fishing effort and catch rates were adjusted for the standardised effort. The Walters *et al.* (1993) model was used with the adjusted catch rates and the results indicated that there had been a significant decline in the breeding stock in the previous two decades. The Brown *et al.* (1995) study is the natural precursor to this research and provides much of the necessary practical and theoretical background.

2.2 GENERALISED LINEAR MODELS

This section is included to review those aspects of generalised linear modelling techniques which will be relevant to this research and which have been used in previous fisheries research. Nelder and Wedderburn (1972) demonstrated the unity of many statistical methods involving linear combinations of parameters by developing a class of statistical models that is a natural generalisation of classical linear models. This concept of a generalised linear model was further outlined by McCullagh and Nelder (1983).

Campbell (in prep.) gives a thorough description of the use of generalised linear models in the analysis of catch rate data. Caputi (in prep) notes that the use of a generalised linear model is an extension of the regression/ANOVA modelling in the standardisation of catch rates which are discussed by Gulland (1956) and further developed and used by Robson (1966), Gavaris (1980), Kimura (1981) and others. The generalisation requires that a linear predictor, which is a systematic component, and the distribution of the random component of the model be specified. It also requires a link between these systematic and random components (McCullagh and Nelder, 1983). Good texts on the subject include McCullagh and Nelder (1983), Dobson (1990), Lindsey (1997) and Farhmeir and Tutz (1994).

For the *random* component of the model, the normal distribution has traditionally been used in the analysis of catch and effort data. Campbell (in prep.) states that such data is often skewed to the right and the variance tends to increase with the mean. Beverton and Holt (1957) gives evidence that the distribution of the errors in their analysis of fishing power statistics was lognormal. Campbell (in prep.) also stresses the importance of checking the validity of the assumption of normality because a mis-specification can lead to serious errors in the parameter estimates. The usefulness of a generalised linear modelling approach becomes apparent here, particularly because of the right-skewed, non-negative nature of the data and the mean-variance relationship. The gamma distribution, which allows only non-negative values, will be considered in this research as a means of validating the assumption of normality. Campbell's paper gives detailed examples of situations where the assumption of normality is clearly inappropriate. It shows the use of the gamma and Poisson distributions when the data allow only non-negative values, and suggests methods to employ when the variance is not constant.

Gulland (1956) gives some evidence that the logarithmic transformation normalises CPUE and stabilises the variance. Thus, the link component of a generalised linear model where the data are logarithmically transformed can be written as $\log(\mathbf{Y}) = \mathbf{X}'\beta$. Robins *et al.* (1996), however, which used a generalised linear model to examine fishing power, states that, although the logarithmic transformation linearises the model, it does not guarantee stabilisation of the variance of the error component. In consideration of this matter, an examination of the variance function will be undertaken in order to provide an appropriate link component for the generalised linear model.

Chapter 3

MULTIPLE LINEAR REGRESSION

3.1 INTRODUCTION

Among the variety of tools available to the statistical investigator, the methods of regression are the most widely used. Multiple regression is primarily concerned with the analysis of the relationships among a set of variables; in particular, between one or more *response* variables (also known as dependent or outcome variables) and several *explanatory* variables (also known as independent or predictor variables). The response variables are considered random and are free to vary in response to the explanatory variables, which are treated as though they are non-random measurements or observations (Dobson, 1990, p. 1). Regression is generally used to describe the system in which those relationships exist in order to make decisions concerning that system and, more particularly, about the response variables. It is also widely used for predictive purposes. (The term *multiple* in the nomenclature refers simply to the characteristic of multiple predictor variables; it is also often referred to as *general* linear regression.)

Regression was first developed by Sir Francis Galton in the latter part of the 19th century when he studied the relationship between the heights of fathers and sons. The term "regression" is a legacy of his observation that there existed a tendency for the heights of sons of both tall and short fathers to regress to the mean height of the group (Neter *et al.*, 1989, p.26).

We are herein concerned only with linear regression, in which each parameter in the function that describes the regression relationship is of the first-order. Consequently, unless otherwise specified, all references to regression pertain specifically to linear regression.

3.2 THE CLASSICAL MULTIPLE LINEAR REGRESSION MODEL

Consider a system in which a single response is thought to be related to a number of predictors. Define the response to be the random variable Y. Then, if we are willing to assume that the relationship between the predictor and response variables is linear, the system can be described as follows.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1} + \varepsilon \tag{3.1}$$

where:

$\beta_0, \beta_1, \ldots, \beta_{p-1}$ are the p model parameters

$X_1, X_2, \cdots, X_{p-1}$ are the known values of the p - 1 predictors

ε is a random error distributed as $N(0, \sigma^2)$.

Taken over n observations (or realisations), for observations Y_i, $i = 1, \ldots, n$, the system becomes

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{ip-1} + \varepsilon_i \tag{3.2}$$

where:

X_{ij} is the value of the j^{th} predictor ($j = 1, \ldots, p\text{-}1$) for observation i

ε_i are independent $N(0, \sigma^2)$.

Equivalently,

$$Y_i = \mu_i + \varepsilon_i \tag{3.3}$$

where $\mu_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{ip-1}$

Now, since the ε_i are $N(0, \sigma^2)$, $E[\varepsilon_i] = 0$ and, consequently, the mean response and thus the response function for this regression model is

$$E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1} \tag{3.4}$$

This means that the regression model with normal error terms implies that the observations Y_i are independent normal random variables with mean $E[Y]$ and constant variance σ^2. (Neter *et al.*, 1989, p.229-230). This model is also called the *general linear model* (which is not to be confused with the *generalised linear model*, which is discussed in chapter 4).

The variables in the model can be measured on a variety of scales. Variables that are numerically continuous are described as quantitative, and variables that are numerically or non-numerically discrete or categorical are described as qualitative variables. Quantitative explanatory variables are known as covariates, and qualitative explanatory variables are known as factors, which have categories called levels (Dobson, 1990, p. 2). For covariates, the respective parameter estimate represents the rate of change in the response variable corresponding to a unit change in the predictor. For qualitative

explanatory variables, there is a parameter estimate for each level of the factor. When a factor can take the values 0 or 1 it is called an indicator variable. (Dobson, 1990, p. 22.)

The multiple linear regression model can be used to describe a variety of situations, which may or may not have linear response-predictor relationships. The presence of non-linearity in these relationships do not necessarily imply non-linearity in the regression model because a linear model necessitates linearity in the parameters only. Consider, for example, a case where the regression function includes the term $\beta_j X_j^2$ for some predictor X_j. The parameter β_j is linear and so the response is linear, even though the predictor is quadratic. Similarly, when there is an interaction term for two predictor variables, that is when the relationship is different for different levels of the variables, then the parameter for the term $X_j X_k$ is still linear. This means that the regression model is not restricted to linear functions. When the regression model is not linear in the parameters, then there is need of a transformation within the regression function in order to make it linear.

To illustrate the point that the Y_i are independent normal random variables whose means vary in a (linearly) systematic way, and whose variances are all equal, consider Figure 3.1, which depicts the situation for a two-dimensional system (that is, with one response variable and one predictor variable). It can be seen that for every value of the predictor variable, X, there exists a distribution of probabilities for possible values of the response variable, Y. (Note that in the figure the probability distributions are shown in a third dimension.) A realisation of each Y_i for some value of the predictor may or may not lie on the regression line. However, the expected value of a realisation does lie on the line. This is because the distribution about the mean is a direct consequence of the error term in the model, which has zero mean. Further, since the errors are assumed to have constant variance, the distributions about each of the means also have equal variances. For the

$$\underset{p\times 1}{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \cdot \\ \cdot \\ \beta_{p-1} \end{bmatrix} \qquad \underset{n\times 1}{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdot \\ \cdot \\ \varepsilon_n \end{bmatrix}$$

The **X** matrix, also known as the model matrix or design matrix, contains the values of the predictor variables for each observation Y_i. Notice that there is a column of 1s in the **X** matrix, which corresponds to β_0 in the regression function. We can now define the regression model in matrix terms as

$$\underset{n\times 1}{\mathbf{Y}} = \underset{n\times p}{\mathbf{X}}\ \underset{p\times 1}{\beta} + \underset{n\times 1}{\varepsilon} \tag{3.5}$$

where:

Y is the random vector of responses

β is the vector of parameters

X is the matrix of known predictor values and represents the deterministic part of the model

ε is the vector of independent normal random variables with expectation $E[\varepsilon] = \mathbf{0}$ and whose variance-covariance matrix $\text{var}[\varepsilon] = \sigma^2\mathbf{I}$.

As a result of this last characteristic, **Y** has expectation $E[\mathbf{Y}] = \mathbf{X}\beta$ and its variance covariance matrix is $\text{var}[\mathbf{Y}] = \sigma^2\mathbf{I}$.

3.3 ASSUMPTIONS OF MULTIPLE LINEAR REGRESSION

The assumptions within the multiple regression model are now outlined.

1. Existence For every specific combination of values of predictor variables, Y is a univariate random variable with a certain probability distribution and having finite mean and variance.

2. Independence The observations for the response variable are independent of one another. This assumption is prone to violation when, for example, the observations are measured on the same subjects at different times.

3. Constant Variance For any fixed combination of the predictor variables, the variance of the response variable is constant. That is

$$\sigma^2{}_{Y|X_1,X_2,\ldots,X_{p-1}} = \mathrm{Var}\left[Y \mid X_1, X_2, \ldots, X_{p-1}\right] \equiv \sigma^2 \tag{3.6}$$

This property is also known as *homoscedasticity*.

4. Normality For any fixed combination of the predictor variables, the response variable is normally distributed. That is, $Y \sim N\left(\mu_{Y|X_1,X_2,\ldots,X_{p-1}}, \sigma^2\right)$, where $N(\mu, \sigma^2)$ denotes the normal probability distribution.

5. Linearity For every specific combination of values of the predictor variables, the expected value of the response variable is a linear combination of the predictor variables. That is,

$$\mu_{Y|X_1,X_2,\ldots,X_{p-1}} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1} \tag{3.7}$$

Comments: The assumptions imply that the random errors in the model for each observation are independent and normally distributed with mean 0 and constant variance. The assumption of normality is not necessary for estimation of the parameters by the method least squares. It is required, however, for inference-making purposes.

3.4 DATA COLLECTION AND EXPLORATION

Any information gained from regression analysis will be dependent on the data used in the analysis. It is vital to the validity of the results of the analysis that the collected data are representative of the population under study, free of errors and in an appropriate form. Any deficiencies in these regards may or may not be overcome, or at least compensated for, by appropriate statistical methods, but, at the vary least, cautionary caveats can be given in the inferences and conclusions. Yet regardless of this, it is obviously most desirable to have data which is as representative and accurate as possible. All affordable efforts should be directed towards achieving this goal.

Preliminary exploration of the data is a relatively easy task that could prevent the occurrence of costly subsequent errors. If data exploration is not effectively undertaken to eliminate or correct defective data, there may be undesirable consequences for the inferences and conclusions of the analysis. Data exploration is also useful for identifying basic patterns and outstanding features in the data. Some of the initial steps in the process of exploring the data are mentioned below.

- Visually scan the data. Identify obvious errors, extreme values, impossible values, and missing data.

- Plot the data. This will identify prima facie the errors and characteristics of the data.

- Obtain cross-tabulations of categorical data.

- Obtain summary statistics for each variable.

Other methods in data exploration involve checking the data for normality, linearity, homoscedasticity and independence. Graphical methods exist and are particularly useful for checking these assumptions prior to performing statistical analyses. Obtaining a histogram or a stem-and-leaf plot of the distribution of each of the variables is a good initial step for checking univariate normality. Histograms will also reveal the extent of skewness and kurtosis. (Descriptions of the above considerations are found in Jobson (1991) and Afifi and Clark (1984).)

Skewness pertains to the symmetry of a distribution, with a skewed distribution having its mean not in the centre. It is derived from the third moment about the mean and, for sample distributions, an index of skewness is given by

$$Skew(x) = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^3 . \tag{3.8}$$

In a large sample the mean of the index is 0 and the variance is approximately $6/n$, when the data are from a normal distribution. Kurtosis pertains to the proportion of the data lying near the centre of its distribution relative to that in the tails, and indicates the "peakedness" of a distribution, which can be too peaked or too flat when compared to the normal distribution. It is derived from the fourth moment about the mean and, for sample distributions, an index of kurtosis is given by

$$Kurt(x) = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}. \tag{3.9}$$

In a large sample the mean of the index is 0 and the variance is approximately $24/n$, when the data are from a normal distribution (Jobson, 1991, p.48). Hence, for both indexes we can use a significance test to test whether the mean is significantly different from 0. This can be done by defining the test statistic to be the index divided by its standard error (Tabachnick and Fidell, 1989, p.72).

Bivariate scatter plots are useful for checking bivariate normality. If the observations came from a multivariate distribution then each bivariate distribution would be normal and each scatterplot should display a roughly elliptical contour. This result is true also if the relationship between the variables is linear and homoscedastic. Thus, a scatterplot is also a way of assessing the assumptions of linearity and homoscedasticity (Tabachnick and Fidell, 1989, pp. 80,82). The relationship between variables is linear and homoscedastic when the assumption of joint normality is met. A truly thorough investigation of normality would check the scatterplots of all possible joint distributions. This is of course impossible above two-dimensions and Johnson and Wichern (1992, p. 153) suggests that, for practical purposes, one-dimensional and two-dimensional investigations are ordinarily sufficient. An alternative technique for assessing the marginal distributions is the Q-Q plot, which plots the ordered sample quantiles versus the quantiles expected from a normal distribution. If the points lie in a straight line, the assumption of normality remains tenable. (Johnson and Wichern, 1992, p.14).

Data that contain observations on the same subjects taken over a course of time may not be independent. Checking the validity of the assumption of statistical independence between the observations, and hence the errors, will be discussed in a later chapter. (The assumption of normality as mentioned in the previous paragraph will also be further discussed there.) It will be sufficient to say here that a plot of the observations over a

naturally appropriate time-scale may reveal a departure of the independence assumption if a regularly occurring pattern or oscillation is apparent.

Statistical measures are available to quantify the extent of the above considerations. Indices of skewness and kurtosis are given in Jobson (1991) and significance tests for skewness and kurtosis are discussed in Tabachnick and Fidell (1989). A commonly used statistical measure for testing the hypothesis that a distribution is normal is the Kolmogorov-Smirnov (KS) D test statistic, which is obtained from a comparison of observed and hypothesised cumulative probability functions (See Jobson (1991), pp. 61-64). To quantify the extent of bivariate normality the correlation co-efficient of the points in a Q-Q plot may be used (See Johnson and Wichern, pp. 157-158). It should be emphasised that the methods mentioned in this section are given for checking the data prior to any analysis being done. Nevertheless, the analyst can and should employ similar and other techniques to check the data and the model's assumptions by using the residuals resulting from the actual analyses.

If a scatterplot suggests that the relationship between the response and a predictor variable is not linear, some form of transformation may be needed to induce linearity. In general, a transformation of the predictor variable is used in preference to a transformation of the response as the latter may induce non-linear relationships between the response and the other predictors. But sometimes a transformation of the response is the most appropriate action, and a range of possible transformations can be used to induce linearity in the model. Affifi and Clarke (1984) discusses a number of transformations for a range of situations where there is non-linearity. Among the more common transformations used are the logarithmic transformation, e.g., $log_e Y$, particularly where the distribution is positively skewed, and the power transformation, e.g., $Y^{1/2}$.

An example is now given which illustrates some of the above considerations. The example is not directly related to fisheries research and is chosen specifically to emphasise the statistical methodology used in later chapters. Table 3.1 gives measured values of paper density and strength. It is desirable to know how the density of a sheet of paper is related to its strength as measured in the direction of the manufacturing machine and as measured at right angles to the machine.

A visual scan of the data in Table 3.1 in the table does not give any indication of obvious outliers, impossible values, e.g., negative values, or missing data. Another visual aid in screening data is to plot the univariate and bivariate distributions of the variables. This can be done with histograms and scatterplots, which are given in Figures 3.2 and 3.3, respectively. Q-Q plots to compare the univariate distributions to the normal distribution are given in Figure 3.4.

Table 3.1 Paper strength data

Specimen	Density	Strength	
		Machine direction	Cross direction
1	0.801	121.41	70.42
2	0.824	127.70	72.47
3	0.841	129.20	78.20
4	0.816	131.80	74.89
5	0.840	135.10	71.21
6	0.842	131.50	78.39
7	0.820	126.70	69.02
8	0.802	115.10	73.10
9	0.828	130.80	79.28
10	0.819	124.60	76.48
11	0.826	118.31	70.25
12	0.802	114.20	72.88
13	0.810	120.30	68.23
14	0.802	115.70	68.12
15	0.832	117.51	71.62
16	0.796	109.81	53.10
17	0.759	109.10	50.85
18	0.770	115.10	51.68
19	0.759	118.31	50.60
20	0.772	112.60	53.51
21	0.806	116.20	56.53
22	0.803	118.00	70.70
23	0.845	131.00	76.35
24	0.822	125.70	68.29
25	0.971	126.10	72.10
26	0.816	125.80	70.64
27	0.836	125.50	76.33
28	0.815	127.80	76.75
29	0.822	130.50	80.33
30	0.822	127.90	75.68
31	0.843	123.90	78.54
32	0.824	124.10	71.91
33	0.788	120.80	68.22
34	0.782	107.40	54.42
35	0.795	120.70	70.41
36	0.805	121.91	73.68
37	0.836	122.31	74.93
38	0.788	110.60	53.52
39	0.772	103.51	48.93
40	0.776	110.71	53.67
41	0.758	113.80	52.42

Data taken from Johnson and Wichern (1992, p. 18)

Figure 3.2 Histograms of each variable in the paper strength example.

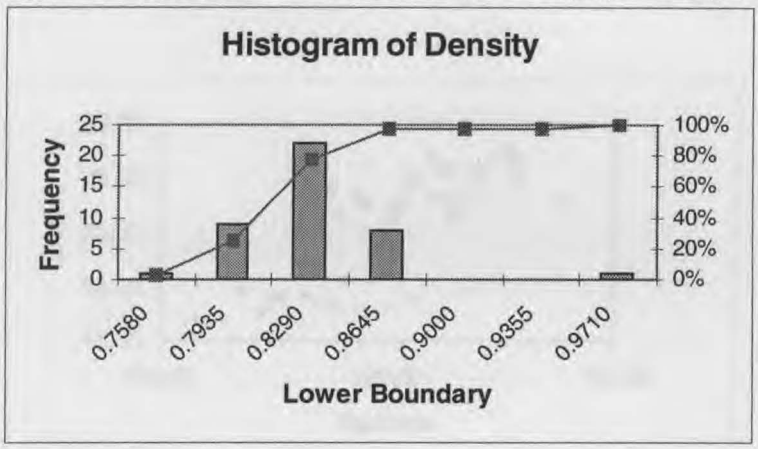


Figure 3.2a

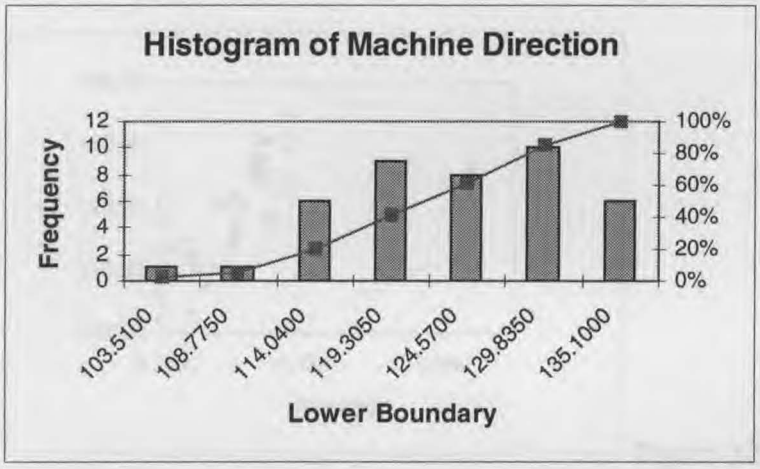


Figure 3.2b

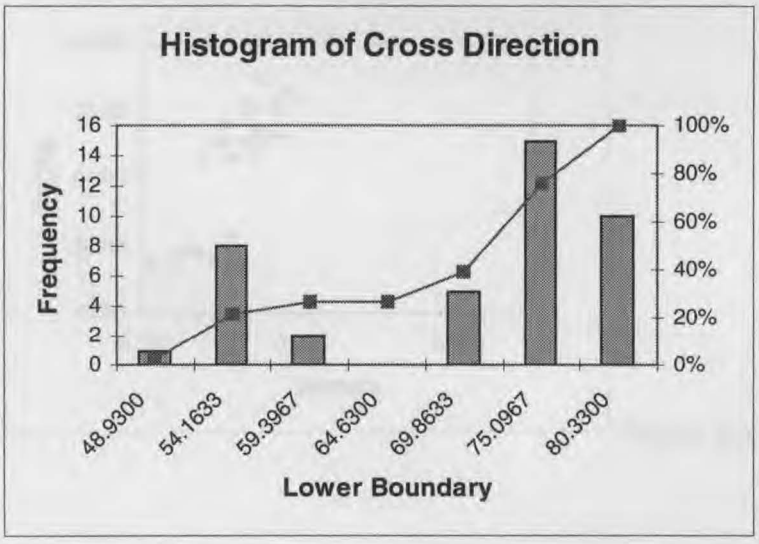


Figure 3.2c

Figure 3.3 Scatterplots of the variables in the paper strength example.

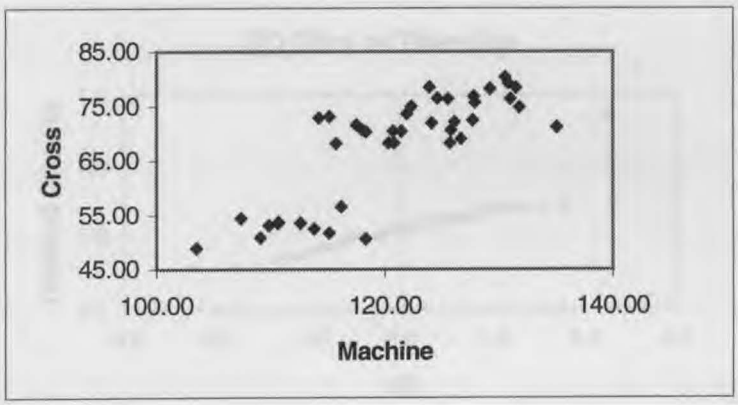


Figure 3.3a

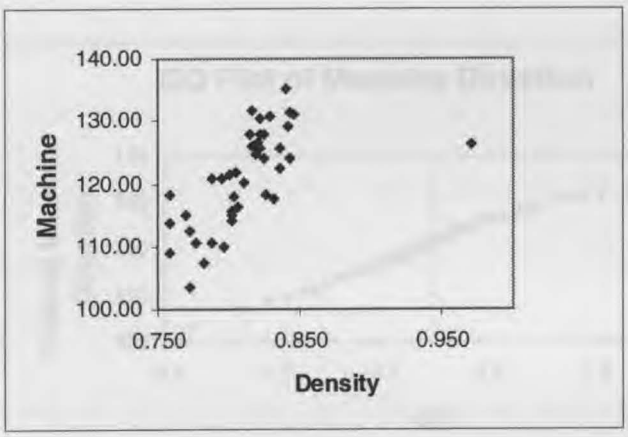


Figure 3.3b

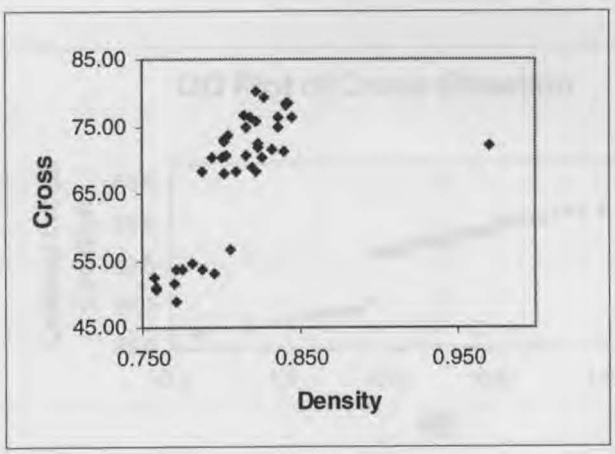


Figure 3.3c

Figure 3.4 Q-Q plots of the variables in the paper strength example.

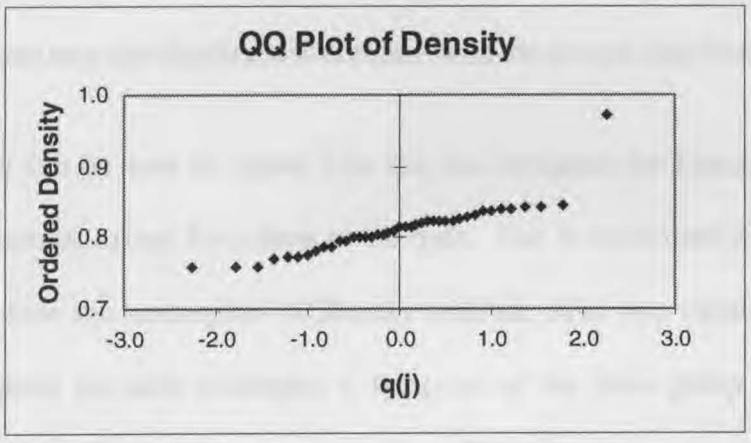


Figure 3.4a

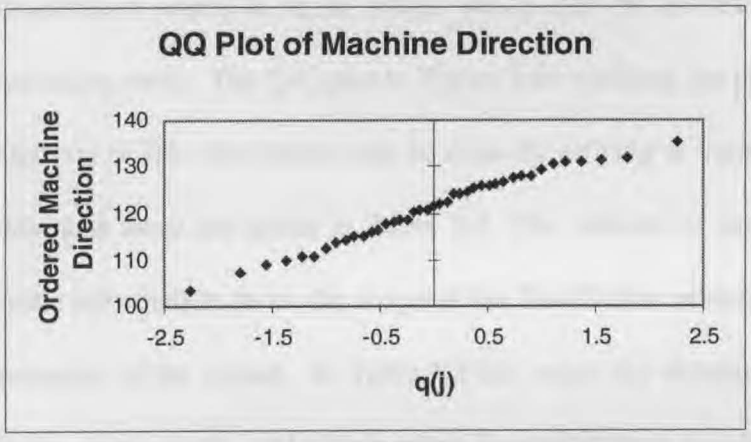


Figure 3.4b

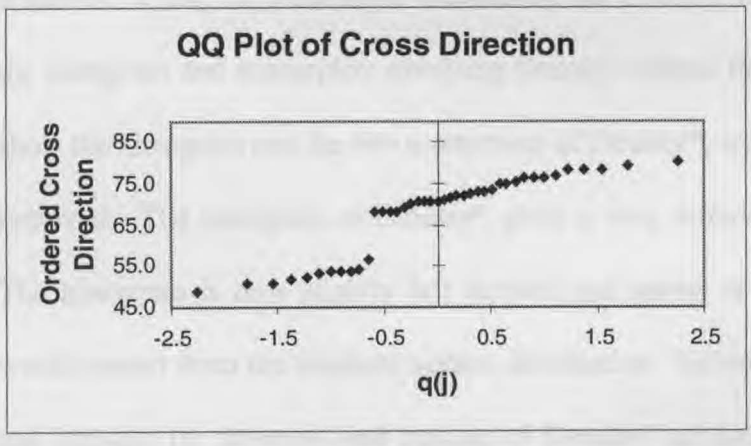


Figure 3.4c

We will interpret the information in these plots by looking at each variable separately; in particular, how each variable is distributed univariately, bivariately with other variables, and how the distributions compare with the normal distribution.

It can be seen in Figure 3.2a that the histogram for Density seems to be approximately normal except for a skew to the right. This is confirmed in Figures 3.3b and 3.3c, which show the scatterplots of Density with the other two variables. It appears that one data point for each scatterplot is lying out of the main group of data points. Upon closer inspection it is found that this is owing to the value of Density for observation 25. This observation seems to be an outlier which may be the result of, say, a measurement or recording error. The Q-Q plot in Figure 3.4a confirms the presence of the outlier. Further analysis of this distribution can be done by looking at various summary statistics for the data and these are given in Table 3.2. The indexes of skewness and kurtosis can give some information about the shape of the distribution, which in this case is affected by the presence of the outlier. In Table 3.2 the values for skewness and kurtosis are 2.021 and 9.154, respectively, and which, when divided by their standard errors, give significant test statistics. It may be worthwhile considering the deletion of the outlier, and then plotting the histogram and scatterplots involving Density without the outlier. Figures 3.5 and 3.6 show the histogram and the two scatterplots of Density*, which is Density with the outlier removed. The histogram of Density* gives a very different picture to that of Density. The histogram is now slightly left skewed and seems somewhat flatter than what we would expect from the standard normal distribution. However, Table 3.2 reveals that the test statistics for skewness and kurtosis of Density* are not significant. Figure 3.6a now shows the scatterplot of Density* and Machine direction in which the points lie in a group that does not seem to depart from bivariate normality. Figure 3.6b shows the scatterplot of Density* and Cross direction in which is seen two distinct clumps of points. As it turns

out, this was the result of an old roll of paper being used for some of the measurements. The Q-Q plot of Density* in Figure 3.7 confirms the notion that the removal of the outlier has normalised the distribution to an acceptable degree.

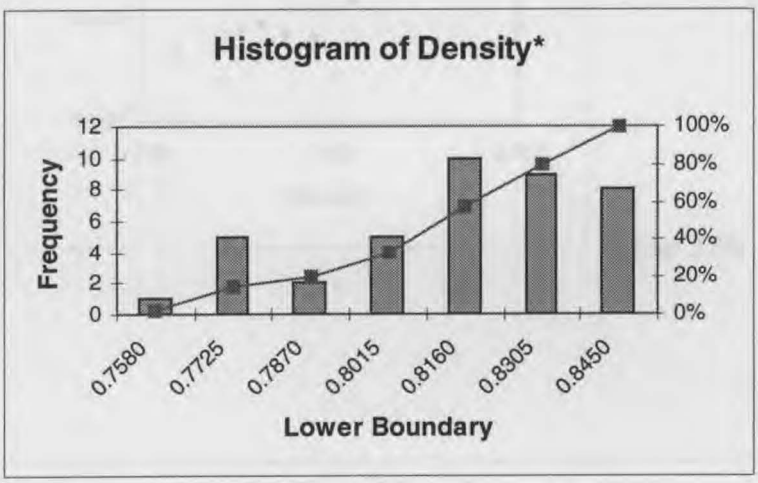


Figure 3.5

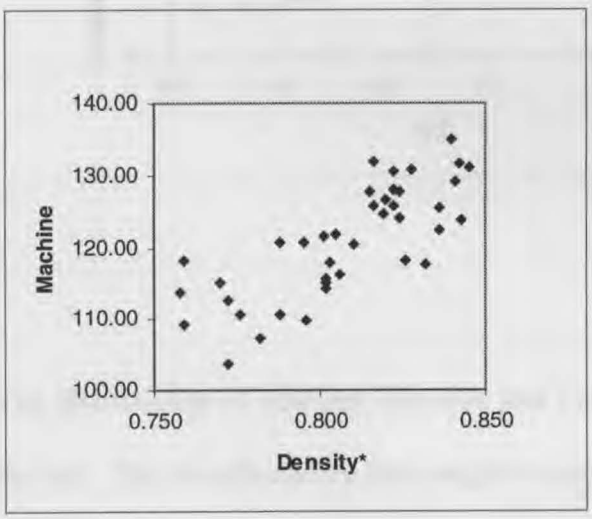


Figure 3.6a

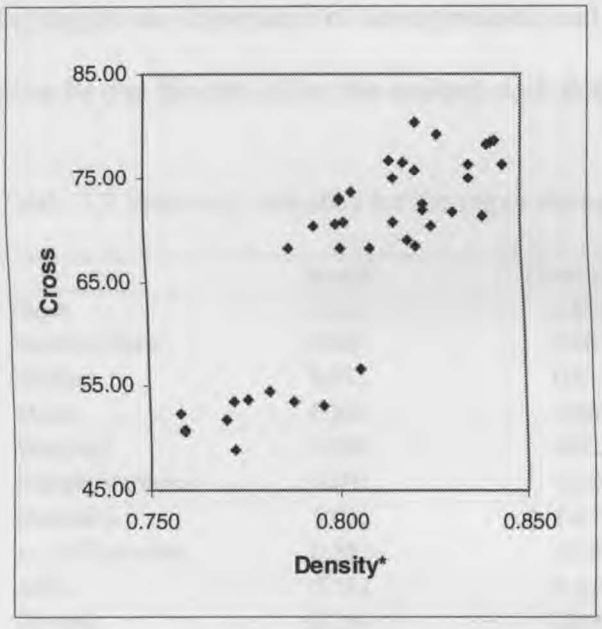


Figure 3.6b

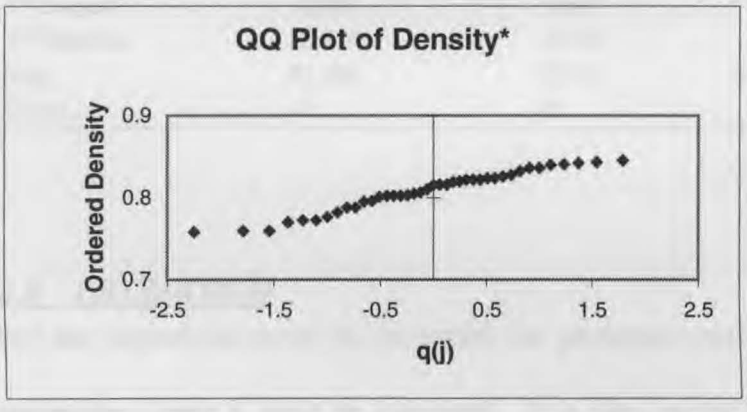


Figure 3.7

The distributions of Machine direction and Cross direction both seem to have a skew to the left. This is indicated by their negative value of the skewness index; however, the test statistics are not significant. Overall, it appears as though all of the variables are not grossly departing from normality, and transformations will not be needed in this regard. As mentioned above, a very noticeable characteristic of the distribution for Cross direction is the appearance of two separate groups of points. This explanation for this

highlights the importance of having reliable and consistent measurement practices. It may even be that the data affect the analysis such that the results are not reliable.

Table 3.2 Summary statistics for the paper strength data.

	Density	Density*	Machine direction	Cross direction
Mean	0.812	0.808	120.953	67.772
Standard Error	0.006	0.004	1.203	1.535
Median	0.815	0.813	121.410	70.700
Mode	0.802	0.802	115.100	
Standard	0.036	0.025	7.702	9.829
Sample Variance	0.001	0.001	59.321	96.617
Skewness	2.021	-0.471	-0.268	-0.776
s.e. of Skewness	0.383	0.387	0.383	0.383
z(Sk)	5.283	-1.216	-0.700	-2.029
Kurtosis	9.154	-0.692	-0.709	-0.902
s.e. of Kurtosis	0.765	0.775	0.765	0.765
z(Kur)	11.965	-0.893	-0.926	-1.179
Range	0.213	0.087	31.590	31.400
Minimum	0.758	0.758	103.510	48.930
Maximum	0.971	0.845	135.100	80.330
3rd Largest	0.843	0.842	131.500	78.540
3rd Smallest	0.759	0.759	109.100	50.850
Sum	33.286	32.315	4959.090	2778.650
Count	41	40	41	41

3.5 ESTIMATION

For the regression model to be useful for prediction and other purposes the unknown parameter vector β must be estimated. It is also useful to have some measure of the accuracy with which the parameters have been estimated. The regression model will thus have a deterministic component by providing the coefficients for each respective predictor variable. By associating a numerical factor with each of the observed or predicted values of the predictor variables a specific value of the response variable can be determined. The level of confidence associated with these predictions follow from the measures of accuracy of the parameter estimations.

The procedure for estimation requires some sort of acceptable measure of the goodness of fit between the data and the corresponding set of values that are fitted by the model. A criterion for determining the estimates based on the measure of goodness of fit must be chosen according to the method of estimation. There are two main criteria which are widely used for estimating the parameters in the general linear regression model:

- **Maximum likelihood** This method determines the likelihood of the parameters given the observed data. The criterion used for the best fit is that the likelihood must be maximised.

- **Least Squares** This method determines the differences between the observed data and the values fitted by the model. The criterion used for the best fit is that the sum of the squares of the differences must be minimised.

For the general linear regression models used in this research the method of least squares was chosen for estimation of the regression parameters. This method, as will be mentioned shortly, has some desirable properties under certain conditions. The procedure for estimation by the method of least squares will now be developed.

Consider the difference between an observed response value for given values of the predictors and the value fitted by the model for the same values of the predictors. The value fitted by the model is, of course, dependent on the parameter estimates. The question initially rises: Is there a set of parameter estimates for which the total of all such differences is a minimum? However, because some difference will be positive and some will be negative, there must be some other measure of goodness of fit. This measure is simply obtained by taking the sum of the squared differences.

Let y_i be the observed value of the i^{th} response variable Y_i, $\hat{y}_i$ be the value fitted by the model for the same values of the predictor variables corresponding to y_i, b_j be the estimate for the parameter β_j, and x_{ij} be the corresponding observed value of the j^{th} predictor variable X_j. Then the method of least squares chooses the parameter estimates b_j, ($j = 0, 1, 2, \ldots, p-1$) such that

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - b_0x_{i0} - b_1x_{i1} - b_2x_{i2} - \cdots - b_{p-1}x_{ip-1})^2 \qquad (3.10)$$

is minimised. Alternatively, in matrix notation, $(\mathbf{y} - \mathbf{Xb})'(\mathbf{y} - \mathbf{Xb})$ is minimised. Here $\mathbf{y}$ is an $n \times 1$ vector of observed responses, $\mathbf{X}$ is an $n \times p$ matrix of values of the predictors corresponding to each observed response (note that $x_{i0} = 1$, for all i), and $\mathbf{b}$ is a $p \times 1$ vector of least squares estimators chosen by the method of least squares. It is convenient here to define the difference between the observed and fitted values of the i^{th} response to be the *residual*, or error, denoted e_i. The vector, $\mathbf{e}$, of residuals is $\mathbf{y} - \mathbf{Xb}$ and, thus, the least squares estimate of $\mathbf{b}$ is chosen when $\mathbf{e}'\mathbf{e}$ is minimised. The vector of residuals also contains information about the error variance σ^2.

Now, for $\mathbf{e}'\mathbf{e}$ to minimised the first derivative must be set to zero. That is,

$$\frac{\partial}{\partial \mathbf{b}}\left[(\mathbf{y} - \mathbf{Xb})'(\mathbf{y} - \mathbf{Xb})\right] = \mathbf{0} \qquad (3.11)$$

and, performing the indicated differentiation, gives

$$-2[\mathbf{X}'\mathbf{y} - (\mathbf{X}'\mathbf{X})\mathbf{b}] = \mathbf{0}. \qquad (3.12)$$

This gives the system of least squares *normal equations*

$$\mathbf{(X'X)b = X'y}. \tag{3.13}$$

Therefore,

$$\mathbf{b = (X'X)^{-1}X'y} \tag{3.14}$$

which yields a unique set of parameter estimates **b** if the matrix **X** is of full column rank, and the square matrix **X'X** has an inverse. The matrix **X'X** has a very important role in the estimation of the parameters and can often be the major factor in the success or failure of the ordinary least squares method. Its diagonal elements are equal to the sums of the squares of the elements in columns of the matrix **X**, and its off-diagonal elements are equal to the sums of cross products of elements in the same columns. As a consequence it is a symmetric matrix. (Myers, 1990, p. 88)

As mentioned above, the vector of residuals provides information about the unknown variance σ^2. An estimate of σ^2 is obtained when the sum of squares of the residuals is divided by the appropriate number of degrees of freedom. This estimate, s^2, is used in assessing model quality and for screening variables via hypothesis testing. It expresses natural variation in the system being modelled, and is an unbiased estimator of σ^2, assuming that the model postulated is correct. (Myers, 1990, pp. 88-89)

The least squares estimators have the following properties.

- If $E[\varepsilon] = \mathbf{0}$ and $\text{Var}[\varepsilon] = \sigma^2\mathbf{I}$ then the estimates contained in **b** are unbiased, that is, $E[b_j)] = \beta_j$, or equivalently, $E[\mathbf{b}] = \beta$. They also have the minimum variance among all of the possible estimators that are both unbiased and linear functions of the elements of **y**. This property is the result of the *Gauss-Markoff Theorem* and does not

assume normality. The estimators are often described as BLUE (best linear unbiased estimators). (Neter *et al.*, 1989, p. 43)

- If $E[\varepsilon] = \mathbf{0}$, $\text{Var}[\varepsilon] = \sigma^2\mathbf{I}$ and the errors are normally distributed, then the estimators achieve uniformly minimum variance in the class of all unbiased estimators. (Myers, 1990, p. 92)

- In the case of normality, the least squares estimators are also the maximum likelihood estimators. (Neter *et al.*, 1989, p. 238)

- The variances of the estimators and the covariances among all possible pairs of estimators can be summarised in a matrix called the variance-covariance matrix. Each element of this matrix can be written as $\text{Cov}[b_j,b_k] = E[(b_j - E[b_j])(b_k - E[b_k])]$. The matrix is given by $\sigma_u^2(\mathbf{X}'\mathbf{X})^{-1} = \sigma_u^2\mathbf{C}$, where $\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1}$. The elements of $\mathbf{C}$ will be denoted by c_{ij}, where $i,j = 1, 2, 3, \ldots, p$ corresponding to the p parameters in $\mathbf{b}$. (Jobson, 1991, p.226)

- An unbiased estimator, s^2, of σ^2 is provided by

$$\frac{(\mathbf{y} - \mathbf{Xb})'(\mathbf{y} - \mathbf{Xb})}{(n - p)}.$$

  Thus, $E[s^2] = \sigma^2$ (Johnson and Wichern, 1992, p. 294).

- The parameter estimate b_j is independent of s^2 (Johnson and Wichern, 1992, p. 295).

3.6 THE ANALYSIS OF VARIANCE TABLE

The overall results of a regression analysis can be summarised in a table called an analysis of variance (ANOVA) table. ANOVA is a statistical procedure used to describe the

relationship between a continuous dependent variable and one or more not necessarily categorical independent variables. The ANOVA table contains several estimates of variance, as well as other information. Regression analysis and ANOVA are closely related and it is not surprising that the results of both procedures can be summarised similarly. The information contained in the ANOVA table can be used to answer the following key questions of linear regression analysis:

1. Is the model appropriate?

2. Does the predictor variable X_j have a significant effect in the model?

(Kleinbaum *et al.*, 1988, p. 96.) These questions will be addressed in the next section.

Before the ANOVA table is presented let us look at how the variability within the data can be partitioned according to that which is "explained" by the model and that which is not. The analyst, in a sense, seeks to explain the variability in the observations with a model. Obviously, if the fitted values, $\hat{y}_i$, are close to the observed values, y_i, then the model will explain much of the observed variability. This means that the variation of the $\hat{y}_i$ around the mean, $\bar{y}$, will be close to the variation of the y_i, around $\bar{y}$. (Myers, 1990, p. 22). The partitioning of the variability is achieved by considering these two sources of variation. The total variation in the observed data is the sum of the variation explained by the regression model and the remaining unexplained variation. This is the *fundamental equation of regression analysis*, which, for the general linear model, can be stated formally as

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y})^2 . \qquad (3.15)$$

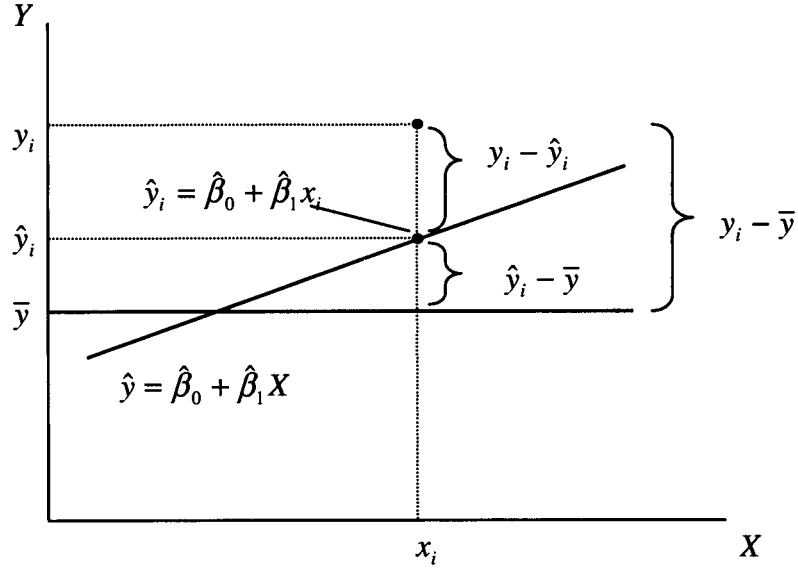
$\sum_{i=1}^{n}(y_i - \bar{y})^2$ is denoted by SST, the total sum of squares about, or corrected for, the mean.

$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ is denoted by SSR, the total sum of squares due to, or explained by, regression. SSR is sometimes called the regression sum of squares.

$\sum_{i=1}^{n}(y_i - \hat{y})^2$ is denoted by SSE, the total sum of squares not due to, or not explained by, regression. SSE is also sometimes called the error sum of squares or the residual sum of squares.

Thus, SST = SSR + SSE (see, for example, Neter *et al.*, 1989, pp. 87-91). SSR can be thought of as the variation due to the regression line and SSE can be thought of the variations around the regression line. Figure 3.2 depicts this relationship for a simple linear regression model (single predictor variable).

Figure 3.2 Partitioning of variation in a simple linear regression model.



Having described the partitioning of variability, we now need to justify the partitioning of the associated degrees of freedom, which are also of importance in the ANOVA table. SST has $n - 1$ degrees of freedom because there are n observations with one degree of freedom lost owing to the estimation of the population mean. (That is, the deviations y_i - $\bar{y}$ are not all independent because they must add to zero.) SSR has $p - 1$ degrees of freedom because there are p parameters with one degree of freedom lost owing, again, to the estimation of the population mean. (Here, again, the deviation $\hat{y}_i$ - $\bar{y}$ must sum to zero.) Finally, SSE has $n - p$ degrees of freedom because there are n observations with p degrees of freedom lost owing to the estimation of the p parameters.

The basic ANOVA table is now presented in Table 3.1. There are variations in the way an ANOVA table may be presented, with different formats being used and other statistics being included or omitted. Sometimes one may see the word "model" instead of the word "regression", and "error" instead of "residual".

Table 3.1 The basic ANOVA table. (Adapted from Neter *et al.*, 1989, p. 240)

Source of Variation	**Degrees of Freedom (df)**	**Sum of Squares (SS)**	**Mean Square (MS)**	***F* Statistic**
Regression	p - 1	SSR	MSR = SSR/(p – 1)	MSR/MSE
Residual	$n - p$	SSE	MSE = SSE/($n - p$)	
Total	n - 1	SST		

The mean squares are the sums of squares divided by their associated degrees of freedom. MSE represents s^2, mentioned in the previous section. It can be shown that $E[\text{MSE}] = \sigma^2$, and so s^2 is an estimate of σ^2 whether the regression model is appropriate or not. Also, $E[\text{MSR}]$ is σ^2 plus a non-negative quantity, and will only provide an estimate of σ^2 if the model is not appropriate for the data. (See Neter *et al.* (1989) for further details.)

However, if the model is appropriate, MSR will be inflated and will correspondingly overestimate σ^2. Thus, for testing whether a regression model is appropriate, we can compare MSR and MSE. If these two quantities are of similar magnitude, then that the model is not appropriate is suggested. But, if MSR is substantially larger than MSE, then this suggests that the model is appropriate. (Neter *et al.*, 1989, p. 94) The F statistic is the ratio MSR/MSE, and, in the light of this discussion, gives an indication of the appropriateness of the regression model. In general, the more closely the model fits the observed data the larger will be the value of the F statistic. This statistic will be discussed further in the next section.

Often one may see other sources of statistical information given with the ANOVA table. Among these are the coefficients of multiple determination and correlation. These two statistics are measures of the overall utility of the regression model and can be used as criteria for comparing several competing models. The coefficient of multiple determination is the proportion of the total variation in the observed values of the response variable which is explained by the regression model, and is computed by

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}} \tag{3.16}$$

R^2 can be thought of as the reduction in sums of squares achieved by using the model. Note that $0 \leq R^2 \leq 1$. It is important to recognise that a large value for R^2 does not necessarily imply that the model is appropriate. Also, including more independent variables in the model can only increase R^2 and never reduce it. This is so because adding more independent variables can never reduce SSE, while SST will always remain constant for a given set of observed values. To make this statistic more useful in this regard, it can be adjusted to cater for the number of variables in the model. By analogy with (3.16) the adjusted coefficient of multiple determination is given by

$$R_a^2 = \frac{\text{SSR}/p-1}{\text{SST}/n-1} = \left(\frac{n-1}{p-1}\right)\frac{\text{SSR}}{\text{SST}} \tag{3.17}$$

or,

$$R_a^2 = 1 - \frac{\text{SSE}/n-p}{\text{SST}/n-1} = 1 - \left(\frac{n-1}{n-p}\right)\frac{\text{SSE}}{\text{SST}} \tag{3.18}$$

(Neter *et al.*, 1989, p. 241)

The adjusted coefficient of multiple determination may actually become smaller when an extra independent variable is added to the model. Notice that, in the first of the two formulae given above, even though SSR may increase in the numerator, the denominator will increase because p is increased. (Neter *et al.*, 1989, p. 242)

Another statistic closely related to R^2 is the coefficient of multiple correlation, R. This is positive square root of R^2, and gives the correlation between the observed values and the fitted values of the response. As with R^2, $0 \leq R \leq 1$.

3.7 INFERENCES IN MULTIPLE LINEAR REGRESSION

The questions put forward in the previous section can be answered by making inferences based on the information in the ANOVA table. These are usually in the form of hypothesis tests. Other questions and hypotheses regarding the model, its parameter estimates, and other matters, can also be addressed with reference to other information. Let us take the previously stated questions one at a time.

The first question asks whether the model is appropriate or not. We set up the following alternative hypotheses:

$$H_0: \beta_0 = \beta_1 = \cdots = \beta_{p-1} = 0$$

$$H_A: \text{Not all } \beta_j \ (j = 0, 1, \ldots, p - 1) \text{ equals zero}$$

The null hypothesis, H_0, is stating that there is no overall significance in the regression model, and that the set of predictor variables does not collectively explain any variation in the observed responses. The test statistic is

$$F = \frac{\text{MSR}}{\text{MSE}}. \tag{3.19}$$

To construct a statistical decision rule for the F statistic, we need to know its sampling distribution. Consider firstly the case when the null hypothesis is true. Cochran's

theorem maintains that if all the observations Y_i are ~$N(\mu, \sigma^2)$ and SST is decomposed into k sums of squares SS_p, each with degrees of freedom df_p, then if the sum of the k df_p equals n-1, the k terms SS_p/σ^2 are independent χ^2 variables with df_p degrees of freedom (Neter *et al.*, 1989, p. 95). In the ANOVA table presented above, SST was decomposed into two distinct sums of squares, namely SSR and SSE, and their degrees of freedom were also additive. Therefore, if the null hypothesis is true, then both SSR/σ^2 and SSE/σ^2 are independent χ^2 variables. Further, consider that the ratio of these two variables is

$$\frac{\text{SSR}/\sigma^2}{p-1} + \frac{\text{SSE}/\sigma^2}{n-p} = \frac{\text{MSR}}{\text{MSE}}. \tag{3.20}$$

This is a ratio of two independent χ^2 variables each divided by its degrees of freedom, which means that it is an F random variable, and follows the $F_{p-1,\, n-p}$ distribution. (Neter *et al.*, 1989, p.96)

The decision rule follows by comparing the computed value of the F statistic with the critical point $F_{p-1,\, n-p;\, 1-\alpha}$ in the $F_{p-1,\, n-p}$ distribution, where α is the level at which the risk of a Type I error is set, typically 0.05 or 0.01. Note that it is an upper-tailed (one-tail) test. Conclude H_0 if the F statistic $\leq F_{p-1,\, n-p;\, 1-\alpha}$; otherwise, reject H_0 in favour of H_A. Alternatively, we can compare the probability of obtaining the computed F statistic with the desired level of significance, α. This probability is precisely the area under the curve of the $F_{p-1,\, n-p}$ distribution which is to the right of the computed F statistic. If this probability, or P-value, $\geq \alpha$, conclude H_0; otherwise reject H_0 in favour of H_A.

The F statistic for the model can also be expressed in terms of the coefficient of multiple determination by

$$F = \frac{R^2/p}{(1-R^2)/(n-p)} = \left(\frac{n-p}{p-1}\right)\frac{R^2}{1-R^2}. \tag{3.21}$$

To answer questions such as the second question stated in the previous section, we need to make inferences about the regression parameters. This requires knowledge of the sampling distribution of the parameters. Of considerable importance to this are the mean and variance of each parameter. The sampling distribution of **b** is normal with mean, $E[\mathbf{b}] = \beta$, and variance-covariance matrix, $\text{Var}[\mathbf{b}] = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. The mean of β is estimated by **b** since **b** is an unbiased estimator of β, and so $E[\beta_j] = b_j$. The variance-covariance matrix of β is estimated by $\text{MSE}(\mathbf{X}'\mathbf{X})^{-1}$, and so $\text{Var}[b_j] = s^2\sqrt{c_{jj}}$, where c_{jj} is the j^{th} diagonal element of $\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1}$.

If b_j and s^2 are independent, we can say that

$$\frac{b_j - \beta_j}{\sqrt{\text{Var}[b_j]}} \sim t_{n-p} \qquad j = 0, 1, \ldots, p - 1.$$

And so we can define the test statistic, t, as

$$t = \frac{b_j}{s\sqrt{c_{jj}}}. \tag{3.22}$$

The denominator is called the standard error of the estimate. This test statistic can be used to choose between the following two hypotheses:

$$H_0: \beta_j = 0$$

$$H_A: \beta_j \neq 0$$

The decision rule for this test follows by comparing the computed value of the t statistic with the critical point $t_{n-p;\,1-\alpha/2}$ of a t distribution with $n-p$ degrees of freedom. Here α is again the level of significance, but notice that the test is two-tailed. Conclude H_0 if $|t| \leq t_{n-p;\,1-\alpha/2}$; otherwise, reject H_0 in favour of H_A. Alternatively, we can compare the probability of obtaining the computed t statistic with the required significance level. If the probability $\geq \alpha$, conclude H_0; otherwise, reject H_0 in favour of H_A. As an interpretation of this test, we can say that if H_0 is rejected, then the predictor variable does not have a significant effect in the model. In effect, the inclusion of this variable did not significantly reduce SSE.

The t statistic can be used similarly to find confidence intervals for the parameter estimates. If the inference assumptions hold, then a $100(1-\alpha)\%$ confidence interval for β_j is

$$b_j \pm t_{n-p;\,\alpha/2}\, s\sqrt{c_{jj}}\,.$$

In deciding whether or not a predictor variable should be included in the model, we can use a different strategy to the t tests just described. This strategy is based on the analysis of the sums of squares attributed to the variable's inclusion in the model through what is called a partial F test. The partial F test can answer the question of whether a predictor variable, X_p, significantly contributes to the determination of the response after the other $p-1$ predictor variables have been accounted for. It is a procedure for testing a variable added last into the model. This method of inference for predictor variables is often shown in the output of statistical computer software.

The test proceeds firstly by computing the sums of squares, SSR and SSE, and mean squares, MSR and MSE, which result from adding X_p to the model. Then we obtain the ratio of sums of squares

$$\frac{\text{SSR}(X_p \mid X_1, X_2, \ldots, X_{p-1})}{\text{MSE}(X_1, X_2, \ldots, X_{p-1}, X_p)}$$

where $\text{SSR}(X_p \mid X_1, X_2, \ldots, X_{p-1})$ is the extra sum of squares obtained by adding X_p to the model, and $\text{MSE}(X_1, X_2, \ldots, X_{p-1}, X_p)$ is the mean square error for the model which contains all of the variables. This ratio has an $F_{1, n-p}$ distribution under H_0. Hence, we can consider it as an F statistic and use it in a decision rule for testing the following hypothesis:

H_0: $\beta_p = 0$ in the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1} + \beta_p X_p + \varepsilon$

H_A: $\beta_p \neq 0$ in the above model

Conclude H_0 if the computed F statistic $\leq F_{1, n-p; 1-\alpha}$, where α is the significance level; otherwise, reject H_0 in favour of H_A. Again, this test may be evaluated by comparing P-value of the test statistic and the significance level.

This partial F test can also be used to determine the extra sums of squares explained by the inclusion of successive predictors in the model. That is, the significance of X_j after accounting for the previous $j - 1$ predictors. These extra sums of squares are sometimes called Type I sums of squares, and are used for variables-added-in-order inferences. The sums of squares resulting from variables-added-last are sometimes called Type III sums of squares. In effect, the Type I sums of squares are a sequence of Type III sums of

squares - but Type I sums of squares will add to SSR, whereas Type III sums of squares will not, generally.

The paper strength data will now be used to illustrate the main ideas presented in this section. The method used to obtain parameters estimates will be the least squares method. The ANOVA table will be obtained and the parameter estimates and inferential information will be presented. For this example we will assume that the data has already been screened and validated, with the previously identified outlier deleted.

The output below was produced by SAS using the GLM (general linear model) procedure. The first section gives the ANOVA table resulting from the procedure. We see that the *F* value of 57.89 is significant at typical levels, which implies that the overall model is appropriate. The value of R-square means that about 75% of the variation in the data is explained by the model. The estimates for the model parameters are given in the last section of the output; specifically:

DENSITY = 0.6028228824 + 0.0007412022(**MACHINE**) + 0.0017069190(**CROSS**).

Even though the overall model seems to be reasonable, it is nevertheless important to test whether the individual parameter estimates are significant. The last section of the output gives the test statistics and P-values for the parameter estimates. We see that the estimates for the intercept and Cross are significant but that the estimate for Machine is marginally not significant at 0.1 and clearly not significant at 0.05. This is confirmed by the values given in the Type III sums of squares section, which gives the results of partial *F* tests wherein the significance of adding the variables last in the model is tested. We see that the P-values for each estimate are the same as in the last section of the output – this is

to be expected because the P-values in the last section are based on the inclusion of the parameter in the presence of the other parameters.

It is worth highlighting that the Type I sums of squares gave significant test statistics for both Machine and Cross, even though the Type III sums of squares gave Machine as not significant and Cross as significant at 0.05. This is because the Type I sums of squares are testing the significance of Machine only and then the significance of Cross in the presence of Machine, whereas the Type III sums of squares are testing the significance of Machine in the presence of Cross and then Cross in the presence of Machine.

Output of Paper Strength Data Analysis

General Linear Models Procedure

Dependent Variable: DENSITY

Source	DF	Sum of Squares	F Value	Pr > F
Model	2	0.01865911	57.89	0.0001
Error	37	0.00596327		
Corrected Total	39	0.02462238		

R-Square	C.V.	DENSITY Mean
0.757811	1.571437	0.80787500

Source	DF	Type I SS	F Value	Pr > F
MACHINE	1	0.01483156	92.02	0.0001
CROSS	1	0.00382755	23.75	0.0001

Source	DF	Type III SS	F Value	Pr > F
MACHINE	1	0.00044024	2.73	0.1068
CROSS	1	0.00382755	23.75	0.0001

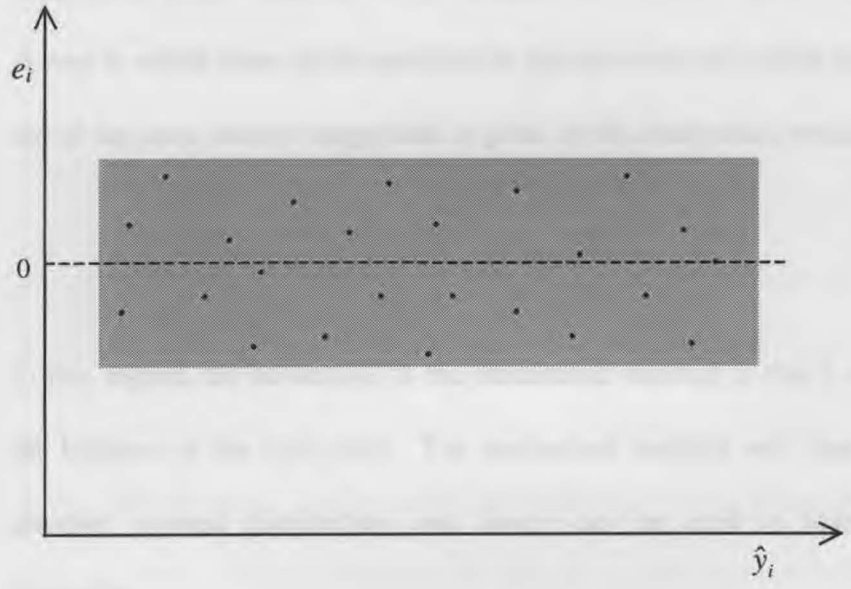
Parameter	Estimate	T for H0: Parameter=0	Pr > |T|	Std Error of Estimate
INTERCEPT	0.6028228824	16.01	0.0001	0.03765404
MACHINE	0.0007412022	1.65	0.1068	0.00044847
CROSS	0.0017069190	4.87	0.0001	0.00035026

3.8 RESIDUAL ANALYSIS

The validity of a fitted model depends greatly on the degree to which the assumptions upon which the model is based are satisfied. In this section, we will discuss methods for checking the satisfaction of these assumptions. To a large extent, this checking can be done by an examination of the residuals and the discussion here will give particular emphasis on graphical techniques for doing this. It is important to understand that, although a model may give a fairly good fit to the currently observed data set, there may still be violations of the underlying assumptions which may render the model less useful for a different set of observations and under different conditions. The method of least squares minimises the sum of the squared residuals for the observed data, but it tells us little about any patterns remaining in the residuals. When the residuals are plotted against a variety of values, for example against the predicted values or against a particular predictor's values, these patterns may reveal much about the validity of the model and problems associated it. The residuals can inform us about model mis-specification, violation of the constant variance assumption, existence of suspect data points, departure from normality and isolated high influence data points. (Myers, 1990, p.211)

When plotted against the predicted values, the residuals would ideally have the appearance of Figure 3.3, in which we see no particular pattern other than a uniform distribution of the residuals about the zero line.

Figure 3.3 Uniform residual distribution



If the fitted model is valid, each residual e_i is an estimate of the error ε_i, which is assumed to be a normal random variable with mean zero and variance σ^2. The expected value of the residuals $\boldsymbol{e}$ is $\mathbf{0}$, and the variance-covariance matrix is $\sigma^2[\mathbf{I} - \mathbf{H}]$,

where $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$;

and thus $\text{var}[e_i] = \sigma^2(1 - h_{ii})$,

where h_{ii} , the *leverage*, is the i^{th} diagonal of the hat matrix $\mathbf{H}$. It can be shown that the leverages range form 0 to 1 and sum to the number of parameters. A value of 0 for h_{ii} means the value of y_i must be 0, and a value of 1 means that the residual is 0, and thus the fitted and observed values are equal. The leverages are so named because they indicate the amount of influence an observation has in determining the fit. They are important for diagnosing possible problems with a fitted model, and are particularly useful for detecting the presence of outlying data points which may exert enough influence on the fitting process to warrant further investigation and perhaps consideration of deletion.

The variances of the residuals, in general, are not uncorrelated and there may be large differences in the variances of the residuals at different values of the predictor variables. A way in which these differences can be incorporated and which ensures that the residuals are of the same relative magnitude is given by the *studentised residual*, defined as

$$e_i^* = s\sqrt{1-h_{ii}}\,. \tag{3.23}$$

In this regard, the advantage of the studentised residual is that it eliminates the effect of the location of the data point. The studentised residual will also approximately have a standard normal distribution, and hence can be used to assess the assumption of normality.

The most commonly useful ways of plotting the residuals to check the regression assumptions are now discussed.

The Overall Plot

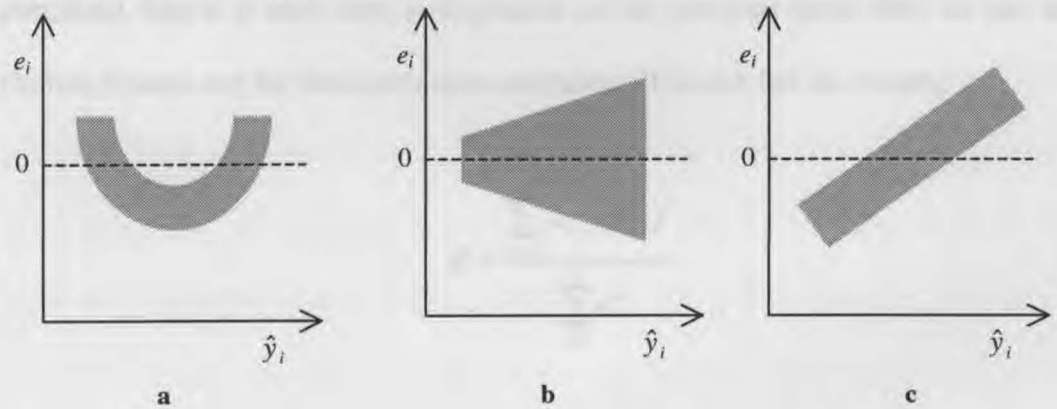
A histogram or dot-diagram will indicate the distribution of the residuals. If they are normally distributed we can expect the familiar bell-shaped curve. Coupled with this idea we can construct a normal probability plot of the ordered residuals against the normal quantiles. If the points in the plot lie nearly in a straight line we would not reject the idea that the errors are normally distributed. This concept was introduced in the section on data collection and exploration, and many of the other techniques discussed therein are also applicable, for example statistics relating to normality, and there is no need to repeat the process here. The main thrust of the residual plot is to assess whether all the systematic variation in the data has been accounted for and whether any points exert excessive influence in the analysis. Although this method can successfully be used to

check the assumption of normality of error terms, it should be noted that the shape of the distribution may also be influenced by violations of the other assumptions. Thus, it should be used in conjunction with other plots.

Plot Against Predicted Values

This plot will often highlight problems with model mis-specification or a deviation from the constant variance assumption. Ideally the plot should appear as in Figure 3.3, but typical problems may be indicated by error distributions shown in Figure 3.4. Figure 3.4a shows that the model may need to be modified to include another parameter, perhaps a quadratic term. Figure 3.4b suggests that the variance is increasing with the value of the response, and a transformation may be needed. This assumption violation may lead to biased standard error estimates. Figure 3.4c shows that the value of the residual is proportional to the value of the response, and suggests that there is an error in the calculations or wrongly omitting the intercept term.

Figure 3.4 Common patterns in residuals



Plot Against Predictor Values

If a systematic pattern is noticeable in the plot of the residuals against the values of a predictor variable then this is an indication that the model needs to be changed, perhaps to include another term. If patterns show up with the appearance similar to those shown in Figure 3.4, then this indicates similar problems to the plot of the residuals against the predicted values.

Plot Against Time

This plot tends to reveal a violation of the assumption of independence. The plot may show patterns similar to those shown in Figure 3.4. If the pattern shown in Figure 3.4a is noticed then this indicates that a linear or quadratic term in time should have been included in the model. A pattern as in Figure 3.4c indicates that a linear term in time should have been included. (Draper and Smith, 1981, p.145). If the error terms are auto correlated, that is if each error is dependent on the previous errors then we can use the Durbin-Watson test for first-order autocorrelation. This test has the statistic

$$d = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} \tag{3.24}$$

where e_t are the time-ordered errors, and compares d to critical values of $d_{L,\alpha}$ and $d_{U,\alpha}$ found in tables for the statistic, whereby we reject the hypothesis that the errors are not autocorrelated if $d < d_{L,\alpha}$.

Chapter 4

GENERALISED LINEAR MODELS

4.1 BACKGROUND

A generalised linear model is essentially an extension, or generalisation, of traditional linear models, such as classical linear models, logistic and probit models for binary data, and log-linear models for multinomial data. These models share some common properties, such as linearity, and they have similar methods of computing parameter estimates (McCullagh and Nelder, 1983). The class of generalised linear models includes as special cases such widely used and accepted techniques as linear regression and analysis of variance. The generalisation allows the mean of a population to depend on a linear predictor through a non-linear *link* function. It also allows the response variable (which in this research is the catch rate) to be distributed according to a wider class of probability distributions called the exponential family of distributions. This is advantageous by the fact that the second-order properties of the parameter estimates do not depend on the assumed underlying distribution, but on the variance-to-mean relationship (McCullagh and Nelder, 1983, p. 2). The generalisation of the form of the underlying distribution makes for the ability to model data that is continuous or discrete, or in the form of counts or proportions, or where the response variable is categorical and not continuous.

4.2 THE CLASSICAL LINEAR MODEL

Let us consider first the form of the classical linear models given in equation 3.5, which is

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon$$

where $\mathbf{y}$ is an $n \times 1$ vector of observations assumed to be a realisation of an $n \times 1$ vector of random variables $\boldsymbol{Y}$ which are independently distributed with mean μ; $\mathbf{X}$ is an $n \times p$ model matrix of explanatory variables with covariates $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$, whose values are known; β is a $p \times 1$ vector of unknown parameters to be estimated; and ε is the vector of errors assumed to be independent normal random variables with mean zero and constant variance. The systematic part of the model is constituted by the vector of means $\mu = \mathbf{X}\beta$. The model assumes that the covariates which influence the systematic component are known and can be measured without error, and that there is independence in the error component, in which each error is normally distributed with constant variance. This form of model is used extensively in many statistical settings. It does have, however, a number of potential problems or limitations.

- The assumption of normality may not be true.
- The normal distribution may not be adequate for modelling certain types of data, for example, counts or proportions.
- The linear predictor in the model allows for any value, while some data may be necessarily restricted to a certain range of values.
- The assumption of constant variance may not be true.

4.3 THE GENERALISATION

To place the classical linear model in terms of a generalised linear model, the linear combination ($\mathbf{x}_i'\beta$) can be extended to include functions of linear combinations $g(\mathbf{x}_i'\beta)$, thus broadening the systematic component of the model. The random component, ε, also can be generalised by allowing the errors to be distributed not just specifically from the normal distribution but from any of the distributions from the exponential family. A generalised linear model can be thought of as comprising three components:

1. The *random* component, **Y**, consisting of independent response variables Y_i which all have the same distribution from the exponential family of distributions. This implies that the variance of the response depends on the mean by a variance function V such that $\text{var}(y_i) = V(\mu_i)\ \phi/\omega_i$, where ϕ is a constant and ω_i is a known weight for each observation. The dispersion parameter ϕ is either known, as in the case of the binomial and Poisson distributions, or must be estimated.

2. The *systematic* component, $\eta = \mathbf{x}_i'\beta$, which is a linear predictor.

3. The *link* function, g, such that $\eta = g(\mu)$, which describes the relation between the random and systematic components.

The main differences between the classical model and the generalised model are (1) the classical model for **Y** has $N(\mu,\sigma^2)$ while the generalised model allows any distribution from the exponential family, and (2) the classical model has g to be the identity function (hence $\eta = \mu$), whereas the generalised model has g to be any monotonic differentiable function. (McCullagh and Nelder, 1983, pp. 19-20). The assumption of independence in the observations is maintained in generalised linear models. Matters of scaling are greatly reduced in generalised linear models. Whereas in classical models a scale should be

chosen to combine constancy of variance and normality of errors, in generalised linear models these are not required, except that the way that the variance depends on the mean must be known. Fitted generalised linear models can be summarised through statistics such as parameter estimates, their standard errors, and goodness-of-fit statistics. Statistical inference can be made about the fitted model and its parameter estimates using hypothesis tests and confidence intervals.

4.4 EXPONENTIAL FAMILY OF DISTRIBUTIONS

Many of the commonly used distributions are members of the same family of distributions, which R. A. Fisher called the *exponential family*. As will be seen below, this family can be further broadened into what Jorgensen (for example, see Jorgensen, 1987) calls the *exponential dispersion family*, by transformation of the mean, the link function, being linearly related to the explanatory variables.

For a random variable Y whose probability function, if it is discrete, or probability density function, if it is continuous, depends on a single parameter, we can say that its distribution belongs to the (one parameter) exponential family of distributions if it can be written in the form

$$f(y;\theta) = s(y)t(\theta)e^{a(y)b(\theta)} \tag{4.1}$$

where a, b, s and t are known functions. Notice the symmetry of the observed value, y, and the parameter θ. θ acts as a location parameter indicating the position within the range of all possible observed values where the distribution lies. Alternatively, equation 4.1 can be rewritten in the form

$$f(y;\theta) = \exp[\omega(y)\xi(\theta) + \psi(\theta) + \zeta(y)] \tag{4.2}$$

where $s(y) = \exp[\zeta(y)]$ and $t(\theta) = \exp[\psi(\theta)]$. Equation 4.2 is said to be in the canonical form if $\alpha(y) = y$ and $\xi(\theta) = \theta$. The component $\xi(\theta)$ is referred to as the natural parameter and acts as a normalising constant in the distribution. As mentioned above, many of the more well known and much used distributions belong to the exponential family, for instance, the Normal, Poisson, Binomial and Gamma distributions (Dobson, 1990, p.27). (Barndorff-Nielsen (1978) has descriptions of these distributions).

As an example, consider the following formulation of the Normal distribution in terms of the exponential family. The Normal probability density function is

$$f(y;\mu) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left[-\frac{1}{2\sigma^2}(y-\mu)^2\right] \tag{4.3}$$

where μ is the parameter of interest and σ^2 is here regarded as a nuisance parameter and treated as part of the other functions. Equation 4.3 can be expressed in a similar form to equation 4.2, viz.,

$$f(y;\mu) = \exp\left[-\frac{y^2}{2\sigma^2} + \frac{y\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right] \tag{4.4}$$

which is in the canonical form. The natural parameter is $\xi(\mu) = \mu/\sigma^2$ and the other terms according to (4.2) are

$$\psi(\mu) = -\frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2) \quad \text{and} \quad \zeta(y) = -\frac{y^2}{2\sigma^2}\ .$$

The broadening of (4.2) includes a scale, or dispersion, parameter, ϕ, so that

$$f(y) = \exp\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\} \tag{4.5}$$

for some functions a, b and c, and known ϕ. Here, θ is still the canonical, or natural, location parameter, which is given in some form of function of the mean. Now, the distribution is an exponential dispersion family. For fixed ϕ, this reverts to a one parameter exponential family of distributions. It can be seen, then, that the exponential family is a special case of the exponential dispersion family, and so for convenience we shall use the former name to refer to both situations. The function $a(\phi)$ is often of the form $a(\phi) = \phi/w_i$, where ϕ is a dispersion parameter and w is a prior weight for each observation, which is assumed known (for grouped data it is usually n_i, the number of observations in group i, otherwise it is generally 1). The specific functions b and c correspond to the type of exponential family.

For comparison, let us again use as an example the Normal distribution density, given in (4.3), by expressing it in the above form of an exponential (dispersion) family:

$$f(y;\theta,\phi) = \exp\left\{\frac{yu - u^2/2}{\sigma^2} - \frac{1}{2}[y^2/\sigma^2 + \ln(2\pi\sigma^2)]\right\}.$$

Here, $\theta = \mu$, $b(\theta) = \mu^2/2$, $\phi = \sigma^2$, $a(\phi) = \phi^2$, and $c(y, \phi) = -[y^2/\phi^2 + \ln(2\pi\sigma^2)]/2$.

It is necessary to obtain expressions for the mean and variance of Y if the distribution is to be useful for most purposes. Firstly, we put $l(\theta,\phi;y) = \ln f_Y(y;\theta,\phi)$ as the log-likelihood function considered as a function of θ and ϕ, y being given. Then the mean and variance of Y can be derived easily from the relations

$$E\left(\frac{\partial l}{\partial \theta}\right) = 0 \qquad (4.6)$$

and

$$\mathrm{E}\left(\frac{\partial^2 l}{\partial\theta^2}\right)+\mathrm{E}\left(\frac{\partial l}{\partial\theta}\right)^2=0. \tag{4.7}$$

From (4.5) the log-likelihood function is $l=[y\theta-b(\theta)]/a(\phi)+c(y,\phi)$. Now

$$\frac{\partial l}{\partial\theta}=[y-b'(\theta)]/a(\phi) \tag{4.8}$$

and

$$\frac{\partial^2 l}{\partial\theta^2}=-b''(\theta)/a(\phi). \tag{4.9}$$

Thus from (4.6) and (4.8) we have

$$0=\mathrm{E}\left(\frac{\partial l}{\partial\theta}\right)=[\mu-b'(\theta)/a(\phi),$$

and so

$$\mathrm{E}(Y)=\mu=b'(\theta). \tag{4.10}$$

Similarly from (4.7), (4.8) and (4.9) we have

$$0=-\frac{b''(\theta)}{a(\phi)}+\frac{\mathrm{var}(Y)}{a^2(\phi)},$$

so that

$$\text{var}(Y) = b''(\theta)a(\phi), \tag{4.11}$$

where primes denote differentiation with respect to θ. For all members of the exponential family, the mean μ is a function of the canonical parameter θ, and is uniquely determined by the specific exponential family through the relation given by (4.10). There is also a special relationship for each member between the mean and the variance; specifically, the variance is a function of the mean, implied by (4.11). The variance is the product of two functions; one, $b''(\theta)$, which depends only on the canonical parameter (and hence on the mean) and will be called the *variance function*, while the other is independent of θ and depends only on ϕ (McCullagh and Nelder, 1983, pp. 20-21). The variance function is therefore also uniquely identified for different members of the exponential family. Notice that for the Normal distribution, $b''(\theta) = 1$, and so the variance function does not depend on the mean, as seen in the classical linear models.

4.5 LINK FUNCTIONS

The link function relates the linear predictor η to the mean of the observed value, μ. In classical linear models the link is the 'identity' function, that is, η and μ are the same. The power family of links is important and can be specified by

$$\eta = \mu^{\alpha}, \ \alpha \neq 0,$$

$$\eta = \log\mu, \ \alpha = 0.$$

The choice of link function is dependent on the particular exponential family. For each family there exists a natural, or canonical, link function which relates the natural parameter directly to the linear predictor: $\theta = \theta(\mu) = \eta = x'\beta$. Natural links lead to convenient and often desirable mathematical properties, while simplifying the numerical methods of estimation. However, their appropriateness to the particular application should be carefully considered in deciding their use, and non-natural links may sometimes be a better choice. (Fahrmeir and Tutz, 1994, p. 20)

Some of the above concepts are now described for the following members of the exponential family:

Normal: $$f(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2} \right], \quad -\infty < y < \infty$$

Poisson: $$f(y) = \frac{\mu^y e^{-\mu}}{y!}, \quad \text{for } y = 0, 1, 2, \ldots$$

Gamma: $$f(y) = \frac{1}{\Gamma(\nu) y} \left(\frac{y\nu}{\mu} \right)^{\nu} \exp\left(-\frac{y\nu}{\mu} \right), \quad \text{for } 0 < y < \infty$$

Notice that the canonical links for the described distributions are μ for the Normal distribution, $\ln(\mu)$ for the Poisson distribution and $-1/\mu$ for the Gamma distribution.

	Normal $N(\mu, \sigma^2)$	*Poisson* $P(\mu)$	*Gamma* $G(\mu, \nu)$
ϕ (scale)	σ	1	ν
$a(\phi)$	σ^2	1	$1/\nu$
θ (location)	μ	$\ln(\mu)$	$-1/\mu$
$b(\theta)$	$\theta^2/2 = \mu^2/2$	$\exp(\theta) = \mu$	$-\ln(-\theta) = \ln(\mu)$
$c(y, \phi)$	$-[y^2/\sigma^2 + \ln(2\pi\sigma^2)]/2$	$-\ln(y!)$	$\nu\ln(\nu) + (\nu-1)\ln(y) - \ln[\Gamma(\nu)]$
$E(Y) = b'(\theta)$	$\theta = \mu$	$\exp(\theta) = \mu$	$-1/\theta = \mu$
Var fn $= b''(\theta)$	1	$\exp(\theta) = \mu$	$(-1/\theta)^2 = \mu^2$
$\text{Var}(Y) = b''(\theta)\, a(\phi)$	σ^2	$\exp(\theta) = \mu$	$(-1/\theta)^2/\nu = \mu^2/\nu$

4.6 ESTIMATION

We need to obtain estimates for the unknown parameters. This is generally done by the method of least squares or by the method of maximum likelihood. The latter is used here to maximise the log-likelihood function $L(\mathbf{y},\mu,\phi)$. Theoretically, this is done by setting the first derivatives of the log-likelihood functions, called the score functions, equal to zero, and then solving the set of equations. The exponential family of distributions have a property such that they satisfy enough regularity conditions to ensure that the global maximum of the log-likelihood function is uniquely given by the solution of $\delta l/\delta\beta = \mathbf{0}$. In general, for the j^{th} parameter, the equations $\delta l/\delta\beta_j = 0$ are non-linear and need to be solved numerically. (Dobson, 1990, p. 40.) Practically, this will be done with an iterative Newton-Raphson method so that, on the r^{th} iteration, the parameter vector β_r is updated by

$$\beta_{r+1} = \beta_r - \mathbf{H}^{-1}\mathbf{s},$$

Here, **H** is the Hessian, or second derivative, matrix, such that $\mathbf{H} = [h_{ij}] = \left[\frac{\delta^2 L}{\delta\beta_i \delta\beta_j}\right]$. It is also the negative of what is called the observed information matrix. **s** is the gradient, or first derivative, vector of scores, such that $\mathbf{s} = [s_j] = \left[\frac{\delta L}{\delta\beta_j}\right]$. Both **H** and **s** are evaluated at the current value of the parameter vector. If $\mu_i = g^{-1}(\mathrm{x}_I'\beta)$ is an estimate of the mean of the i^{th} observation, obtained from the estimate of the parameter vector β, then we can write

$\mathbf{s} = \Sigma_i w_i(y_i - \mu_i)\mathbf{x}_i / V(\mu_i)g'(\mu_i)\phi$ and

$\mathbf{H} = \mathbf{X}'\mathbf{W}_o\mathbf{X}$

where **X** is the design matrix, $\mathbf{x}_i$ is the transpose of the i^{th} row of **X**, and V is the variance function. The matrix $\mathbf{W}_o$ is diagonal with i^{th} diagonal element

$$w_{oi} = w_{ei} + w_i(y_i - \mu_i)\frac{V(\mu_i)g''(\mu_i) + V'(\mu_i)g'(\mu_i)}{(V(\mu_i))^2(g'(\mu_i))^3\phi}$$

where

$w_{ei} = w_i / \phi V(\mu_i)(g'(\mu_i))^2$

in which primes denote derivatives of g and V with respect to μ.

The estimated variance-covariance matrix of the parameter estimator is given by $\Sigma = -\mathbf{H}^{-1}$ in which **H** is evaluated with the parameter estimates of the last iteration. The correlation

matrix has non-diagonal elements as $\sigma_{ij}/\sigma_i\sigma_j$, where σ_{ij} is an element of Σ, and diagonal elements of 1 (SAS Inst., 1993).

An example will now be given to illustrate the process of parameter estimation of a generalised linear model. Once the parameters have been estimated using the above methods we can check them against the least squares parameter estimates obtained from the classical linear regression of the last chapter. The paper strength data introduced in the last chapter was analysed using the SAS GENMOD procedure to assess the model with Density as the response variable and Machine direction and Cross direction as the two predictor variables.

The partial output below shows the parameter estimates after each iteration in the estimation process. We see that there was only one iteration before the process stopped. Shown also is the last evaluation of **s**, the first derivative of the log-likelihood function, which as expected is very close to zero. Likewise, we see the last evaluation of **H**, the second derivative. The section "Criteria for Assessing Goodness of Fit" provides information for assessing the adequacy of the model and will be dealt with in the next section of the dissertation. The maximum likelihood parameter estimates are shown in the following section. We see that MACHINE is highly significant and CROSS is significant at the 0.10 level but not significant at the 0.05 level. The estimates and standard errors match very closely those obtained by least squares in the previous chapter and these will also be discussed in the next section. The model given is

DENSITY = 0.6028 + 0.0007(MACHINE) + 0.0017(CROSS).

Output 4.1

The GENMOD Procedure

Model Information

Description	Value
Data Set	WORK.PAPER
Distribution	NORMAL
Link Function	IDENTITY
Dependent Variable	DENSITY
Observations Used	40

Parameter Information

Parameter	Effect
PRM1	INTERCEPT
PRM2	MACHINE
PRM3	CROSS

Iteration History For Parameter Estimates

Iter	Ridge	LogLikelihood	PRM1	PRM2	PRM3	Scale
0	0	119.462789	0.60282	0.0007412	0.001707	0.01221
1	0	119.462789	0.60282	0.0007412	0.001707	0.01221

Last Evaluation Of The Negative Of The Gradient

PRM1	PRM2	PRM3	Scale
-2.83E-11	-3.484E-9	-1.904E-9	7.068E-13

Last Evaluation Of The Negative Of The Hessian

Parameter	PRM1	PRM2	PRM3	Scale
PRM1	268309.4	32418414	18154819	4.6321E-9
PRM2	32418414	3.93268E9	2.2099E9	5.7065E-7
PRM3	18154819	2.2099E9	1.25422E9	3.1193E-7
Scale	4.6321E-9	5.7065E-7	3.1193E-7	536618.8

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	37	0.0060	0.0002
Scaled Deviance	37	40.0000	1.0811
Pearson Chi-Square	37	0.0060	0.0002
Scaled Pearson X2	37	40.0000	1.0811
Log Likelihood	.	119.4628	.

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCEPT	1	0.6028	0.0362	277.0860	0.0001
MACHINE	1	0.0007	0.0004	2.9530	0.0857
CROSS	1	0.0017	0.0003	25.6742	0.0001
SCALE	1	0.0122	0.0014	.	.

NOTE: The scale parameter was estimated by maximum likelihood.

Estimated Covariance Matrix

	PRM1	PRM2	PRM3	Scale
PRM1	0.001311	-0.000015	6.5663E-6	2.828E-19
PRM2	-0.000015	1.8604E-7	-1.179E-7	-4.14E-21
PRM3	6.5663E-6	-1.179E-7	1.1348E-7	2.728E-21
Scale	2.828E-19	-4.14E-21	2.728E-21	1.8635E-6

Estimated Correlation Matrix

	PRM1	PRM2	PRM3	Scale
PRM1	1.0000	-0.9283	0.5382	0.0000
PRM2	-0.9283	1.0000	-0.8114	-0.0000
PRM3	0.5382	-0.8114	1.0000	0.0000
Scale	0.0000	-0.0000	0.0000	1.0000

4.7 GOODNESS OF FIT AND INFERENCE

It is desirable to obtain an indication of the adequacy of the fitted model. This can be done by measuring of the size of the discrepancy between the values derived from the fitted model and the actual observed values. Alternatively, and in order to make use of various statistical properties, we can compare the likelihood functions of the fitted model and the *maximal* model at their maximum likelihood estimates, $\mathbf{b}$ and $\mathbf{b}_{max}$, respectively. (The maximal model is a model in which the number of parameters is equal to the number of observations.) This comparison can take the form $L(\mathbf{b}_{max}; \mathbf{y})/ L(\mathbf{b}; \mathbf{y}) = \lambda$, which is called the *likelihood ratio statistic.* After taking logs this has the form $l(\mathbf{b}_{max}; \mathbf{y}) - l(\mathbf{b}; \mathbf{y}) = \log \lambda$. If the fitted model describes the data well then $(\mathbf{b}; \mathbf{y})$ will be approximately equal to $(\mathbf{b}_{max}; \mathbf{y})$ and so $\log \lambda$ will be small; and conversely, if the fitted model is poor then $(\mathbf{b}; \mathbf{y})$ will be much smaller than $(\mathbf{b}_{max}; \mathbf{y})$ and so $\log \lambda$ will be relatively large.

4.7.1 The Deviance

Let us define the log- likelihood ratio statistic as

$$D = 2[l(\mathbf{b}_{max}; \mathbf{y}) - l(\mathbf{b}; \mathbf{y})]. \tag{4.12}$$

D (= $-2\log\lambda$) is often called the *scaled deviance*, but in the literature sometimes just the *deviance.* For the exponential family of distributions

$$D = \frac{2w_i \sum \left[y_i \left(\hat{\theta}_i - \tilde{\theta}_i \right) + b\left(\tilde{\theta}_i \right) - b\left(\hat{\theta}_i \right) \right]}{\phi} \tag{4.13}$$

where the hat (^) indicates the maximal model and the tilde (~) indicates the fitted model.

The numerator in (4.13) is also referred to in the literature as the deviance. To prevent any confusion between the two definitions, we will always call the former the scaled deviance, and the latter the *unscaled* deviance or simply the deviance. Denoting the unscaled deviance by D^*, it is simple to see that

$$D = D^*/a(\phi) \tag{4.14}$$

where $a(\phi) = \phi/w_i$ is the scale component of (4.5). The unscaled and scaled deviances are described below for some exponential families.

	(Unscaled) Deviance	Scaled Deviance
Normal	$\sum_i w_i(y_i - \mu_i)^2$	$\frac{1}{\sigma^2}\sum_i w_i(y_i - \mu_i)^2$
Poisson	$2\sum_i w_i[y_i \log(y_i/\mu_i) - (y_i - \mu_i)]$	$2\sum_i w_i[y_i \log(y_i/\mu_i) - (y_i - \mu_i)]$
Gamma	$2\sum_i w_i[-\log(y_i/\mu_i) + (y_i - \mu_i)/\mu_i]$	$2\nu\sum_i w_i[-\log(y_i/\mu_i) + (y_i - \mu_i)/\mu_i]$

For inferential purposes we need to know the distribution of the scaled deviance. It can be shown that

$$2[l(\mathbf{b}; \mathbf{y}) - l(\beta; \mathbf{y})] \sim \chi^2_p \tag{4.15}$$

where $\mathbf{b}$ is the maximum likelihood estimate vector of the p parameters in the parameter vector β. (Lindsey, 1997, p. 212) Now we can use (4.15) to advantage by expressing the scaled deviance with components as follows:

$$D = 2\{ [l(\mathbf{b}_{max}; \mathbf{y}) - l(\beta_{max}; \mathbf{y})] - [l(\mathbf{b}; \mathbf{y}) - l(\beta; \mathbf{y})] + [l(\beta_{max}; \mathbf{y}) - l(\beta; \mathbf{y})] \}. \quad (4.16)$$

The first component on the right hand side of (4.16) has the χ^2_n distribution because there are n parameters in the maximal model; likewise, the second term has the χ^2_p distribution because there are p parameters in the fitted model. If the fitted model adequately describes the data then the third component, which is a positive constant, will be close to zero. Further, if the random variables defined by the first two components are independent and the third component is close to zero, then

$$D \sim \chi^2_{n-p} \quad (4.17)$$

By comparing D with the value predicted by χ^2_{n-p} we have a measure of the adequacy of the fitted model. If D is larger then the model may not be adequate. Care must be taken here, because a large value for D does not necessarily provide proof for lack of fit, but merely adds support to that proposition (Fahrmeir and Tutz, 1994, p.48).

Another method, yet somewhat crude, can also provide an indication of the adequacy of the fitted model. We know that the expected value of a variable distributed as χ^2_m is m. It follows that if the model is adequate then (4.17) holds, and so $E(D) \cong (n - p)$. If D is larger than $(n - p)$, then this suggests that the model is not adequate. In general, however, the sampling distribution of the scaled deviance is not approximated well by (4.17). (Dobson, 1990, pp. 57-58.)

The above two methods are useful, but they necessitate that the value of D can be identified from the data. This means that the dispersion parameter ϕ must be known, and of course this is not readily the case for some distributions. For example, for the Normal distribution, the variance σ^2 is generally not known. But in the case of the Poisson distribution, the scale parameter is 1, and so D is D^*, which is readily computable from the data. Fortunately, though, the scale parameter can be estimated. One strategy is to compute the deviance from a maximal model, or at least a model with sufficient number of parameters to account for most of the systematic variation, and divide by the number of its degrees of freedom. The quotient can then be used as an estimate, $\hat{\phi}$, of the scale parameter. This follows from (4.14) and (4.17), because $\hat{\phi} = D^*/\mathrm{E}(D) = D^*/(n-p)$, if we are prepared to assume that the model is reasonable. Alternatively, the scale parameter can be regarded as an additional unknown parameter and estimated by maximum likelihood at each step of the fitting process.

An assessment of the adequacy of the fitted model will benefit from the above considerations, but other methods, especially examination of residuals, need to be employed for a truly thorough analysis.

4.7.2 Analysis of Deviance

In the process of finding suitable models to represent the data, we can compare two competing models by comparing their goodness of fit statistics; particularly, their likelihood ratio statistics. If the models have the same distribution, link function and dispersion parameter then we can compute the difference in the scaled deviances for the competing models, and use this difference in a test of hypothesis about the models.

Consider the two competing hypotheses:

$$H_0 : \beta = \beta_0 = [\beta_1, \beta_2, \ldots, \beta_q]'$$

$$H_1 : \beta = \beta_1 = [\beta_1, \beta_2, \ldots, \beta_p]'$$

where $q < p$. Define $\Delta D = D_0 - D_1$. Then

$$\Delta D = 2[l(\mathbf{b}_{max}; \mathbf{y}) - l(\mathbf{b}_0; \mathbf{y})] - 2[l(\mathbf{b}_{max}; \mathbf{y}) - l(\mathbf{b}_1; \mathbf{y})]$$

$$= 2[l(\mathbf{b}_1; \mathbf{y}) - l(\mathbf{b}_0; \mathbf{y})]$$

If the models fit the data well then $D_0 \sim \chi^2_{n-q}$ and $D_1 \sim \chi^2_{n-p}$. Therefore $\Delta D \sim \chi^2_{p-q}$, provided that some conditions of independence are satisfied. The decision rule is that we reject H_0 if $\Delta D > \chi^2_{p-q,\alpha}$, where α is the specified level of significance. If H_0 were rejected then this suggests that β_1 provides a significantly better description of the data than does β_0. (Dobson, 1990, pp. 61–62.) Note also that $E[\chi^2_{p-q}]$ is $p - q$, so if ΔD is much larger, say twice as large or more, than $p - q$ then H_0 should be viewed with suspicion.

If the deviances for the competing models are assumed to be distributed as chi-squared random variables, then an equivalent test would be to compare a ratio of them to the value of $F_{p-q, n-p, \alpha}$. This is particularly advantageous because it cancels out the need for the (usually unknown) dispersion parameter.

Using these methods for inferences about alternative models, we can develop a sequence of nested models, each successive model having only one parameter more than the previous, to build a table of differences in deviances analogous to sums of squares in an ANOVA table.

4.7.3 Inference About Parameters

The reliability of a parameter estimate of a fitted model can be judged by the magnitude of its standard error. Calculations of confidence intervals for the parameter estimate naturally follow. Firstly, though, we must identify the sampling distribution for the maximum likelihood estimate.

Many important results about generalised linear models relate to the first derivative $U = \mathrm{d}l/\mathrm{d}\theta$. This is called the *score*, and was referred to as **s** in section 4.6, concerning estimation. It can be shown that $\mathrm{E}[U] = 0$, and $\mathrm{E}[-U'] = \mathrm{E}[U^2]$. Thus, $\mathrm{Var}[U] = \mathrm{E}[U^2] = \mathrm{E}[-U']$, where primes denote the derivative of U. Hence, $\mathrm{Var}[U]$, called the *information matrix* **I**, is equal to the negative of the second derivative.

For **b**, an unbiased estimator of β, the variance-covariance matrix is

$\mathrm{E}[(\mathbf{b} - \beta)(\mathbf{b} - \beta)'] = \mathbf{I}^{-1}$. Thus, for large samples,

$$(\mathbf{b} - \beta)'\mathbf{I}(\mathbf{b} - \beta) \sim \chi^2_p$$

and so

$$\mathbf{b} \sim N(\beta, \mathbf{I}^{-1}) \tag{4.18}$$

The statistic $(\mathbf{b} - \beta)'\mathbf{I}(\mathbf{b} - \beta)$ is called the Wald statistic and is used to make inferences about β. (Dobson, 1990, p. 53.) The sampling distribution for **b**, given in (4.18), leads to the standard error, defined by,

$$\text{s.e.}(b_j) = \sqrt{v_{jj}} \tag{4.19}$$

where v_{jj} is the j^{th} term on the diagonal of the matrix $\mathbf{I}^{-1}$. Consequently, a 100(1 - α)% confidence interval for b_j is

$$b_j \pm t_{n-q;\,\alpha/2}\sqrt{v_{jj}}$$

Let us again consider the paper strength example to illustrate the concepts of goodness of fit and inference in a generalised linear model. The data set was analysed with the SAS GENMOD procedure and some of the output was presented in the estimation section. The parts concerning goodness of fit and parameter estimates are again presented here for discussion, together with Type 1 and Type 3 and Wald interval statistics. We see that the scaled deviance is 40.0000 and the when divided by the degrees of freedom gives a value close to unity. This suggests that the fitted model is adequate. The value of the scaled deviance is also approximately equal to the degrees of freedom, again suggesting that the model is adequate.

The value of the unscaled deviance is 0.0060. The relationship between the scaled and unscaled deviances can be confirmed using the knowledge that the scaled deviance is just the deviance divided by the variance as estimated by the square of the scale parameter. Thus, the estimate of the variance should be approximately 0.00015, which is validated by the SCALE parameter value of 0.0122 given previously in the full output.

The output below shows the Type 1 and Type 3 likelihood ratio statistics for testing the significance of the variables in the model. The Type 1 table gives the estimates of the deviance and Chi-square statistics analogous to that obtained in the ANOVA table of multiple linear regression, except that the deviance replaces the sum of squares. Here, the statistics are based on the variables added-in-order method. In effect, it is comparing the model with only MACHINE to the model with both MACHINE and CROSS.

MACHINE is significant as the first variable in the model and CROSS is significant when MACHINE has already been accounted for. The Type 3 table shows the statistics for the variables when they are added last. In effect, it is comparing two models: one with MACHINE added last, and one with CROSS added last. The table shows that CROSS is significant (when MACHINE has already been accounted for), and MACHINE is significant only to about the 0.09 level (when CROSS has already been accounted for). It is interesting to note that CROSS is significant when it is included first or last, but MACHINE is only significant when added first.

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	37	0.0060	0.0002
Scaled Deviance	37	40.0000	1.0811
Pearson Chi-Square	37	0.0060	0.0002
Scaled Pearson X2	37	40.0000	1.0811
Log Likelihood	.	119.4628	.

LR Statistics For Type 1 Analysis

Source	Deviance	DF	ChiSquare	Pr>Chi
INTERCEPT	0.0246	0	.	.
MACHINE	0.0098	1	36.8884	0.0001
CROSS	0.0060	1	19.8331	0.0001

LR Statistics For Type 3 Analysis

Source	DF	ChiSquare	Pr>Chi
MACHINE	1	2.8491	0.0914
CROSS	1	19.8331	0.0001

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCEPT	1	0.6028	0.0362	277.0860	0.0001
MACHINE	1	0.0007	0.0004	2.9530	0.0857
CROSS	1	0.0017	0.0003	25.6742	0.0001
SCALE	1	0.0122	0.0014	.	.

NOTE: The scale parameter was estimated by maximum likelihood.

Normal Confidence Intervals For Parameters
Two-Sided Confidence Coefficient: 0.9500

Parameter	Confidence Limits	
PRM1	Lower	0.5318
PRM1	Upper	0.6738
PRM2	Lower	-0.000104
PRM2	Upper	0.001587
PRM3	Lower	0.001047
PRM3	Upper	0.002367
Scale	Lower	0.009534
Scale	Upper	0.0149

From the "Analysis of Parameter Estimates" section of the above output, we can make inferences about the parameter estimates based on the estimated standard errors. Confidence intervals based on the Wald statistics are computed using these standard errors and they are presented in the last section of the output. They show that the intercept and CROSS are significant different from zero, but that MACHINE is not.

4.8 RESIDUAL ANALYSIS

The analysis of residuals for generalised linear models is conceptually very similar to that under the classical linear models. The idea of plotting the residuals against predicted values of the response, against values of the predictor variables and over the time domain is central to the identification of possible violations of the assumptions underlying the model. Even though we may be satisfied that the goodness of fit criterion has provided us with a satisfactory measure for the overall adequacy of the proposed model, we still need to investigate the residuals in order to look at specific aspects of the model. However, we need a generalisation of the residuals in order to be applicable to all the distributions which may replace the Normal distribution, and which can be used for the same purposes as the residuals based on the Normal distribution. Apart from the raw residual, $r_i = y_i - \mu_i$, we will discuss three residuals for use in generalised linear models.

Pearson Residuals

The Pearson residual is the raw residual scaled by the estimated standard deviation. That is

$$r_P = \frac{y_i - \mu_i}{\sqrt{\frac{V(\mu_i)}{w_i}}} \quad . \tag{4.20}$$

Deviance Residuals

The deviance can be used as a measure of the discrepancy of a model, and when this is done, each observation contributes a quantity d_i to the value of the deviance, so that $\Sigma d_i = D$. The deviance residual is defined as

$$r_D = \text{sign}[y_i - \mu_i]\sqrt{d_i} \tag{4.21}$$

The form of d_i for some members of the exponential family of distributions can easily be taken from Table 4.2. For example, for the Gamma distribution

$$r_D = \text{sign}[y_i - \mu_i]\sqrt{\{\ 2\nu \sum_i w_i[-\log(y_i/\mu_i) + (y_i - \mu_i)/\mu_i]\}}. \tag{4.22}$$

The deviance residual is preferable to the Pearson residual for model checking purposes because its distributional properties are closer to those of the residuals from classical regression models.

Standardised Residuals

This residual facilitates the comparisons of individual residuals because it scales the raw residuals by their estimated standard deviations. That is

$$r_{\text{std}} = \frac{r_{D_i}}{s_i\sqrt{1-h_i}} \quad . \tag{4.23}$$

Many other forms of residuals have been proposed in the literature. For examples, see McCullagh and Nelder (1983) and Pierce and Schafer (1986), but these are not discussed

here and we will mainly use the deviance and standardised residuals for checking the fitted generalised linear model.

The residuals should be plotted against the fitted values, the explanatory variables and against the order in time that they were measured. These plots were discussed in chapter 3 and their usefulness is nonetheless applicable to the broader class of generalised linear models, but we need not illustrate them again here. For checking the variance function, the residuals would be plotted against the predicted responses to see if an increasing or decreasing pattern emerges and the model could be adjusted accordingly. For example, if the residuals show an increasing trend then this indicates that the variance function in the model is increasing too slowly and should be changed to a function which increases more rapidly. It is important to note that any pattern in the residuals indicates that the specified model is inadequate in some way. Essentially, the model is intended to account for all systematic variation in the data, leaving random errors.

Chapter 5

APPLICATION TO WESTERN ROCK LOBSTER DATA

5.1 FISHING POWER FACTORS

In the last decade or two the changes that have occurred in the Western Rock Lobster fishery with regard to fishing power have been substantive. In this section we will look at several fishing power factors and how their prevalence and usage has changed in that time. We will obtain a time series for each factor which will be used to determine which seasons are most appropriate for estimating fishing power increases. Of particular interest are the technologically advanced onboard fishing power factors and the ensuing analyses will focus on these. These are radar and global positioning systems (GPS), which are used as navigational tools, and black and white echo sounders and their successor, the colour echo sounders, which are fish-finding tools.

Figure 5.1, below, shows the time series of the percentages of vessels with various pieces of equipment in the fishery. It shows that in the early 1970's all, or nearly all, vessels had a black and echo sounder (BWES) and virtually none of the other equipment listed. It shows the introduction of radar during the same period. We see the emergence of the colour echo sounder (CES) in the mid 1980's and the coincident decline in BWES. Notable is the rapid increase in the use of GPS in the early 1990's, together with the gradual decline of radar. Functionally, CES replaced BWES as an underwater scanner, and GPS replaced radar as a navigational tool. By the 1995/96 season almost all vessels had a CES and a GPS onboard. However, as shown in the figure, many vessels retained

expected to have changed, however, by the introduction of radar, CES and GPS. For CES the suitable seasons chosen are 1982/83 to 1988/89; for GPS, the appropriate seasons are limited to 1989/90 to 1991/92. Radar appears to be suitable for analysis in all seasons from the mid 1970's and after but it is best to avoid the years when GPS was emerging because of the similar roles they perform on a boat. It was decided that radar could be examined concurrently with CES because they perform different roles on a boat. The analysis of fishing power will proceed in the next two sections by applying the statistical methods to the data using these subsets of seasons.

5.2 ESTIMATION OF FISHING POWER USING MULTIPLE LINEAR REGRESSION

This section will apply the methods discussed in chapter 3 to the catch and effort data. Before we proceed it is well worth noting that because of the environmental nature of the data we should appreciate that a "perfect" model is unlikely to be found, and indeed will not be sought. It is not within the purpose of this research to try to model the real world nature from which the data have come. That purpose would require a very thorough and possibly complex modelling process, involving many more variables. It does, however, have a specific purpose to work within the guidelines of previously established catch and effort models, using their traditionally tested predictors, to find how certain other predictors (fishing power factors) are influencing the catch rates. With this in mind we shall not seek to find the "best" regression or use forwards, backwards or other similar techniques which are common in regression analysis. It is well recognised that a least squares estimation of data such as this will leave a large proportion of the total variation unexplained. In fact, it is common to expect values for R^2 of less than 0.4. Also, because there is a very large number of observations in the dataset, we can expect the predictors in the models to be significant on most occasions. In this regard, if there is doubt as to a

predictor's significance in a model, it may be useful to look to the proportion of sums of squares explained by the predictor as a guide.

5.2.1 Preliminary Data Exploration

The variables in the dataset that will be considered in the modelling procedure are all categorical except for the response variable. The variables are listed below.

Region A geographical area of the fishery defined by latitude and longitude. The models will examine the coastal regions of the fishery.

Depth The depth of the ocean, in fathoms, where the catch was made. The models will examine depths of twenty or more fathoms, as it is thought that the effect of these fishing power factors is more easily seen (and measured) at these depths.

Pull The number of days the pot has been in the water before it is pulled up and emptied. This is typically one or two days. It is thought that the catch increases for the first couple of days, but not much after that, so the models will examine pulls of three days or less.

Month The fishing season runs from mid November to the end of June, but it is thought of as having two distinct phases: the whites period, from November to January, when newly moulted, highly catchable, immature lobsters leave the shallow reefs and move seaward to deeper waters; and the reds period, from February to June, when non-migratory lobsters are caught mainly in deeper waters. Because we are examining only the deeper waters, and also because the effect of echo sounders and navigational equipment may be more readily seen when

catchability levels are relatively low, the models will only examine the reds period.

Season The fishing season. For example, the 1982/83 season runs from November 1982 to June 1983. We have previously established which seasons are appropriate for the analyses.

Catch rate The total weight (in kilograms) of lobsters caught on one day divided by the total number of pots lifted to yield that catch. Because vessels have different numbers of pots, and each vessel may even pull different numbers of pots on different days, it may be appropriate to use the number of pots lifted to weight each observed catch rate. It is possible to have catch rates of zero, and this will need to be checked as a large proportion of zero rates may present difficulties for the model.

The other variables are those for the fishing power factors of radar, CES and GPS. The are binary variables, taking a value of 1 if the observation came from a vessel with the onboard equipment, and a value of 0 if the observation came from a vessel without the equipment.

Let us now look at the characteristics of the response variable. Output 5.1, below, shows the result of a SAS UNIVARIATE procedure for data from the seasons 1982/83 to 1991/92. The mean is 0.91 and the variance is 0.56. The data is heavily skewed to the right as is seen by the histogram and the quantile figures. This skewness and the apparent lack of normality in the distribution are not unacceptable; it is the errors from a regression analysis that should have a normal distribution, and not necessarily the response.

Output 5.1

The SAS System

Univariate Procedure

Variable=CATRATE

Moments				Quantiles(Def=5)			
N	50677	Sum Wgts	50677	100% Max	13.36458	99%	3.8125
Mean	0.913817	Sum	46309.5	75% Q3	1.133333	95%	2.282609
Std Dev	0.748417	Variance	0.560128	50% Med	0.71	90%	1.754717
Skewness	2.962079	Kurtosis	16.19141	25% Q1	0.447761	10%	0.293103
USS	70703.42	CSS	28385.02	0% Min	0	5%	0.214286
CV	81.90008	Std Mean	0.003325			1%	0.1
T:Mean=0	274.8661	Pr>|T|	0.0001	Range	13.36458		
Num ^= 0	50618	Num > 0	50618	Q3-Q1	0.685572		
M(Sign)	25309	Pr>=|M|	0.0001	Mode	1		
Sgn Rank	6.4056E8	Pr>=|S|	0.0001				

Extremes

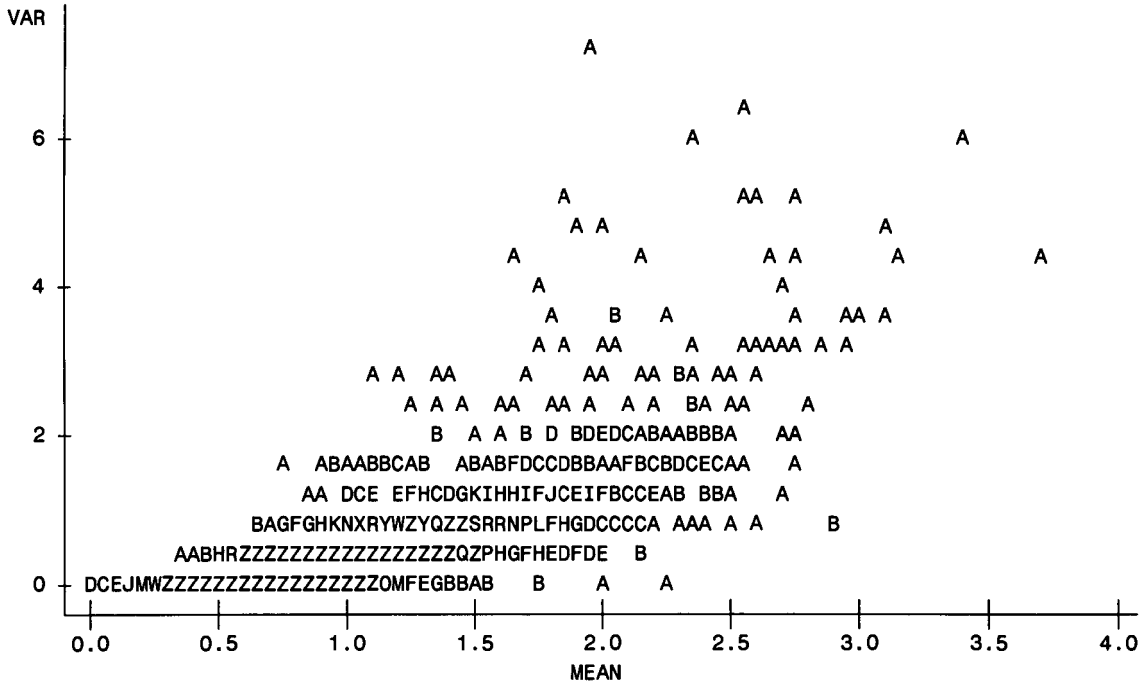
Lowest	Obs	Highest	Obs
0(	50216)	9.671875(	1636)
0(	48873)	9.973333(	13269)
0(	48841)	10(	35899)
0(	48298)	10.13605(	7631)
0(	45349)	13.36458(	8272)

```
                    Histogram                    #          Boxplot
 13.5+*                                          1             *
     .                                                         |
     .                                                         |
     .*                                          2             *
     .*                                          5             *
     .*                                          9             *
     .*                                         24             *
     .*                                         46             *
     .*                                         96             *
     .*                                        254             *
     .*                                        701             0
     .****                                    2599             0
     .*****************                      12225          +-----+
  0.5+*************************************** 34715          *--+--*
     ----+----+----+----+----+----+----+----+----+---
     * may represent up to 724 counts
```

The relationship between the mean and the variance is an important consideration in the analysis. Classical regression requires that the variance be constant over the distribution of the values of the variables. This can be checked visually by obtaining the mean and variance of the catch rate in each different combination of levels of the predictor variables and then plotting these pairs of data. Such a plot gives an indication of the nature of the probability distribution underlying the data. If the variance is constant then the plot will resemble a scatterplot of points with no apparent pattern or relation. Output 5.2 shows the plot of the variance against the mean for the catch rate. There is clearly a relationship between the two statistics. We see that the variance increases as the mean increases, and not just linearly but at an increasing rate. There is therefore a need for a transformation in the data. Because the plot seems to have an exponential characteristic about it, it seems likely that a logarithmic transformation of the catch rates may stabilise the variance.

Output 5.2

```
Plot of VAR*MEAN.  Legend: A = 1 obs, B = 2 obs, etc.
```

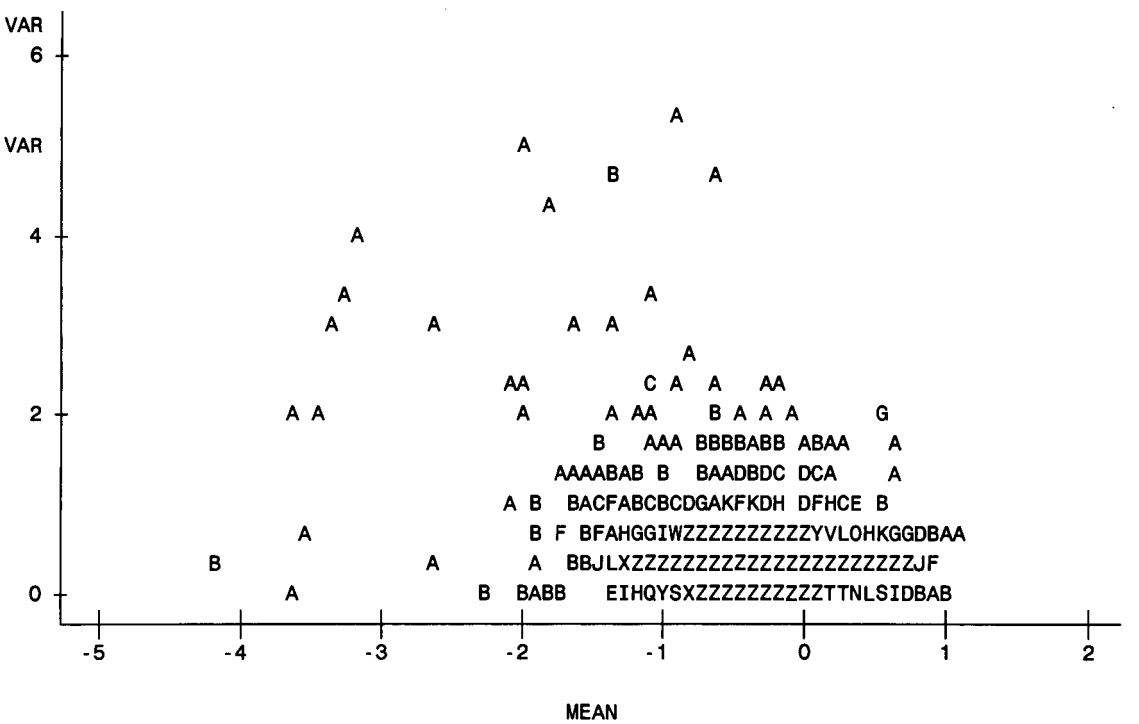



```
NOTE: 82 obs had missing values.  2350 obs hidden.
```

A logarithmic transformation was applied to the catch rates and the variance to mean relationship was re-examined. Because of the presence of zero catch rates for some observations, a constant was added to the catch rates to accommodate the log transformation. The size of the constant should be small enough not to affect the parameter estimates too much, and the decision should be made with consideration of the size of the mean of the catch rates. It is also helpful and desirable to have knowledge of the distribution of the smallest few observations because this will give an indication of the smallest values that are practically possible. The mean is 0.91 and there are only three observations with catch rates of less than 0.01, and they are all greater than 0.008. It was decided that a constant of 0.01 would be acceptable. The plot of the variance against the mean of the transformed catch rates are shown in Output 5.3. It shows that the previous variance to mean relationship is no longer apparent, and the variance has been stabilised considerably. This is similar to the findings of Gulland (1956) and Beverton and Holt (1957) who were instrumental in proposing the log transformation for catch and effort data. It does seem, though, that there are some points to the left of the central group which lessen the randomness of the scatterplot. Upon closer examination it was discovered that less than 1% of the observations of the dataset were contributing to the mean log catch rates of –2.3 or less, and the vast majority of observations were within the main group. Thus, we will accept the logarithmic transformation for the catch rates, and assume that the catch rates are lognormally distributed.

Output 5.3

```
Plot of VAR*MEAN.  Legend: A = 1 obs, B = 2 obs, etc.
```




```
NOTE: 82 obs had missing values.  2474 obs hidden.
```

5.2.2 Model Proposal

It is proposed that the estimation of fishing power be undertaken for the two subsets of seasons with the following models.

For the analysis of radar and CES, let model A be:

$$\text{Log}Y_i = \beta_0 + \beta_1\text{Region} + \beta_2\text{Month} + \beta_3\text{Season} + \beta_4\text{Pull} + \beta_5\text{Depth} + \beta_6\text{Region*Month} + \beta_7\text{CES} + \beta_8\text{Radar} + \varepsilon_i \quad (5.1)$$

for observations Y_i, $i = 1, \ldots, n$, where ε_i are independent $N(0, \sigma^2)$.

For the analysis of GPS, let model B be:

$$\text{Log}Y_i = \beta_0 + \beta_1\text{Region} + \beta_2\text{Month} + \beta_3\text{Season} + \beta_4\text{Pull} + \beta_5\text{Depth} + \beta_6\text{Region*Month} + \beta_7\text{GPS} + \varepsilon_i \quad (5.2)$$

for observations Y_i, $i = 1, \ldots, n$, where ε_i are independent $N(0, \sigma^2)$.

These model specifications imply that they are multiplicative in their parameters on the original scale, which is widely accepted for catch and effort data. Notice that the models include an interaction term between region and month. This is included because the catch rates vary over the regions for different months of the fishing season.

5.2.3 Estimation

The two models proposed were analysed with the SAS GLM procedure. The results are shown in Outputs 5.4 and 5.5 exactly as they were produced. We see from the values of R^2 that 25% and 29% of the variation in the data is explained by the respective models. This means that for each model about one quarter of the total sum of squared errors about the mean can be attributed to the regression model. One of the first considerations is to determine if the model is significant. In the ANOVA table in Output 5.4 the value for the F statistic is 34.98, which is significant at the 0.0001 level. Output 5.5 shows an F value for the model of 75.04, again significant at a very small level. One should not be too concerned about these statistics because of the large number of observations relative to the number of parameters in the model. This causes the error sums of squares to be divided by a large number of degrees of freedom, which results in a large F value. The predictor variables, as expected, are all significant (at the 0.0001 level). We can get an indication of the importance of these predictor variables by looking at the proportion of sums of squares attributed to them. For both models, the Type I sums of squares for the

fishing power factors contribute a significant amount. These sums of squares are calculated from the variables added-in-order method. Similarly, for both models, the factors contribute a significant proportion of the Type III sums of squares. This means that even when they are entered last in the model, they are still found to be significant.

Notice that the output contains a message saying that the unique estimates could not be found. The GLM procedure has given the parameter estimates for each (categorical) effect relative to the last level of each factor in the analysis. For each of the fishing power factors we can estimate the difference in the log-transformed catch rates between vessels with the factor and vessels without by comparing the parameter estimates for the two levels of the factor, the last of which is always zero.

Near the bottom of Output 5.4 we see that the parameter estimate for vessels without CES is -0.11, with a standard error of 0.024, and the parameter estimate for vessels without radar is -0.18, with a standard error of 0.026. From Output 5.5 we see that the parameter estimate for vessels without GPS is -0.17, with a standard error of 0.020. These three estimates are all significant and their confidence intervals do not include 0, even at the 0.0001 level. In order to obtain the increases in catch rates associated with these estimates, they need to be transformed back to their original scale. These are given by:

CES: $e^{0.11} = 1.12$

Radar: $e^{0.18} = 1.20$

GPS: $e^{0.17} = 1.18.$

Hence, we see that vessels with CES are catching 12% more than vessels without CES, vessels with radar are catching 20% more than vessels without radar, and vessels with GPS are catching 18% more than vessels without GPS.

Output 5.4

```
The SAS System

                    General Linear Models Procedure
                        Class Level Information

Class     Levels    Values

REGION         7    2 3 4 5 6 7 8

MM             5    2 3 4 5 6

SEASON         6    8283 8384 8485 8586 8687 8788

PULL           3    1 2 3

DEPTHCAT       3    2 3 4

COLECHO1       2    0 1

RADAR1         2    0 1

                Number of observations in data set = 4021

NOTE: Due to missing values, only 3700 observations can be used in this analysis.
```

The SAS System

General Linear Models Procedure

Dependent Variable: LCATRATE
Weight: POT

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	35	33765.67163460	964.73347527	34.98	0.0001
Error	3664	101041.11798123	27.57672434		
Corrected Total	3699	134806.78961583			

R-Square	C.V.	Root MSE	LCATRATE Mean
0.250475	-1433.901	5.25135452	-0.36622840

Source	DF	Type I SS	Mean Square	F Value	Pr > F
REGION	6	10557.35166180	1759.55861030	63.81	0.0001
MM	4	6718.58817340	1679.64704335	60.91	0.0001
SEASON	5	4725.06157633	945.01231527	34.27	0.0001
PULL	2	1059.54938869	529.77469434	19.21	0.0001
DEPTHCAT	2	611.17905626	305.58952813	11.08	0.0001
REGION*MM	14	5606.15311515	400.43950822	14.52	0.0001
COLECH01	1	3261.76258818	3261.76258818	118.28	0.0001
RADAR1	1	1226.02607480	1226.02607480	44.46	0.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
REGION	6	1572.13273351	262.02212225	9.50	0.0001
MM	4	2445.52559422	611.38139855	22.17	0.0001
SEASON	5	5745.03277827	1149.00655565	41.67	0.0001
PULL	2	1407.60020686	703.80010343	25.52	0.0001
DEPTHCAT	2	531.14282441	265.57141221	9.63	0.0001
REGION*MM	14	5749.78005250	410.69857518	14.89	0.0001
COLECH01	1	566.11345103	566.11345103	20.53	0.0001
RADAR1	1	1226.02607480	1226.02607480	44.46	0.0001

Parameter		Estimate	T for H0: Parameter=0	Pr > |T|	Std Error of Estimate
INTERCEPT		-0.074436081 B	-0.93	0.3501	0.07964927
REGION	2	0.015514820 B	0.03	0.9750	0.49486177
	3	-2.108111741 B	-5.37	0.0001	0.39275568
	4	-0.485289780 B	-2.71	0.0068	0.17930907
	5	-0.284493170 B	-0.62	0.5350	0.45857180
	6	-0.407859554 B	-3.33	0.0009	0.12246231
	7	0.076974772 B	0.85	0.3929	0.09008377
	8	0.000000000 B	.	.	.
MM	2	0.104157666 B	1.71	0.0880	0.06103436
	3	0.182810035 B	3.11	0.0019	0.05882655

The SAS System

General Linear Models Procedure

Dependent Variable: LCATRATE

Parameter		Estimate	T for H0: Parameter=0	Pr > |T|	Std Error of Estimate
MM	4	0.055263581 B	0.98	0.3267	0.05633329
	5	-0.214261875 B	-3.44	0.0006	0.06232127
	6	0.000000000 B	.	.	.
SEASON	8283	0.113326260 B	3.72	0.0002	0.03048087
	8384	-0.057085364 B	-1.81	0.0703	0.03153325
	8485	-0.251984468 B	-8.05	0.0001	0.03131107
	8586	-0.215366898 B	-7.69	0.0001	0.02799944
	8687	-0.140106530 B	-4.14	0.0001	0.03382545
	8788	0.000000000 B	.	.	.
PULL	1	-0.143822123 B	-5.33	0.0001	0.02698638
	2	-0.011632800 B	-0.45	0.6560	0.02610922
	3	0.000000000 B	.	.	.
DEPTHCAT	2	-0.160642313 B	-2.98	0.0029	0.05397116
	3	-0.030496211 B	-0.49	0.6222	0.06189334
	4	0.000000000 B	.	.	.
REGION*MM	2 2	-0.001459026 B	-0.00	0.9977	0.50515103
	2 3	0.000000000 B	.	.	.
	3 4	0.000000000 B	.	.	.
	4 2	0.133295850 B	0.68	0.4992	0.19723978
	4 3	0.000000000 B	.	.	.
	5 2	0.031610413 B	0.07	0.9455	0.46248270
	5 3	0.531841852 B	1.15	0.2487	0.46098361
	5 4	0.860778151 B	1.87	0.0616	0.46034261
	5 5	0.940003529 B	1.96	0.0496	0.47856709
	5 6	0.000000000 B	.	.	.
	6 2	0.113255450 B	0.85	0.3956	0.13331382
	6 3	0.566833490 B	4.42	0.0001	0.12814060
	6 4	0.648249318 B	5.09	0.0001	0.12732205
	6 5	0.604504747 B	4.12	0.0001	0.14655644
	6 6	0.000000000 B	.	.	.
	7 2	-0.145308334 B	-1.43	0.1526	0.10156784
	7 3	0.025202401 B	0.25	0.8034	0.10121243
	7 4	-0.034605430 B	-0.36	0.7214	0.09702936
	7 5	-0.139917143 B	-1.23	0.2172	0.11335367
	7 6	0.000000000 B	.	.	.
	8 2	0.000000000 B	.	.	.
	8 3	0.000000000 B	.	.	.
	8 4	0.000000000 B	.	.	.
	8 5	0.000000000 B	.	.	.
	8 6	0.000000000 B	.	.	.
COLECH01	0	-0.109851025 B	-4.53	0.0001	0.02424509
	1	0.000000000 B	.	.	.
RADAR1	0	-0.176607211 B	-6.67	0.0001	0.02648683
	1	0.000000000 B	.	.	.

NOTE: The X'X matrix has been found to be singular and a generalized inverse was used to solve the normal equations. Estimates followed by the letter 'B' are biased, and are not unique estimators of the parameters.

Output 5.5

The SAS System

General Linear Models Procedure
Class Level Information

Class	Levels	Values
REGION	7	2 3 4 5 6 7 8
MM	5	2 3 4 5 6
SEASON	3	8990 9091 9192
PULL	3	1 2 3
DEPTHCAT	3	2 3 4
SNAVGPS1	2	0 1

Number of observations in data set = 8791

NOTE: Due to missing values, only 7653 observations can be used in this analysis.

The SAS System

General Linear Models Procedure

Dependent Variable: LCATRATE
Weight: POT

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	41	93764.74378123	2286.94497027	75.04	0.0001
Error	7611	231943.52712383	30.47477692		
Corrected Total	7652	325708.27090506			

R-Square	C.V.	Root MSE	LCATRATE Mean
0.287880	-2959.371	5.52039645	-0.18653954

Source	DF	Type I SS	Mean Square	F Value	Pr > F
REGION	6	40176.84141451	6696.14023575	219.73	0.0001
MM	4	27470.34944210	6867.58736052	225.35	0.0001
SEASON	2	8197.22723920	4098.61361960	134.49	0.0001
PULL	2	1419.53882236	709.76941118	23.29	0.0001
DEPTHCAT	2	3557.87681401	1778.93840701	58.37	0.0001
REGION*MM	24	10622.43642320	442.60151763	14.52	0.0001
SNAVGPS1	1	2320.47362586	2320.47362586	76.14	0.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
REGION	6	6908.19878283	1151.36646381	37.78	0.0001
MM	4	14663.73732162	3665.93433040	120.29	0.0001
SEASON	2	7945.93773358	3972.96886679	130.37	0.0001
PULL	2	2673.69309077	1336.84654538	43.87	0.0001
DEPTHCAT	2	2097.11574758	1048.55787379	34.41	0.0001
REGION*MM	24	11017.27251140	459.05302131	15.06	0.0001
SNAVGPS1	1	2320.47362586	2320.47362586	76.14	0.0001

Parameter		Estimate	T for H0: Parameter=0	Pr > |T|	Std Error of Estimate
INTERCEPT		0.031269917 B	0.43	0.6696	0.07327740
REGION	2	0.636060723 B	2.29	0.0219	0.27738702
	3	-1.054778041 B	-4.01	0.0001	0.26312559
	4	0.119568607 B	1.34	0.1812	0.08941895
	5	0.016068370 B	0.17	0.8646	0.09420456
	6	-0.284675490 B	-3.59	0.0003	0.07924358
	7	-0.186641204 B	-2.59	0.0096	0.07206402
	8	0.000000000 B	.	.	.
MM	2	0.207336093 B	2.97	0.0030	0.06975822
	3	0.236226126 B	3.52	0.0004	0.06717407
	4	0.181176843 B	2.87	0.0041	0.06315056
	5	-0.188196635 B	-2.73	0.0063	0.06885323

Dependent Variable: LCATRATE

Parameter		Estimate	T for H0: Parameter=0	Pr > |T|	Std Error of Estimate
MM	6	0.000000000 B	.	.	.
SEASON	8990	-0.066589411 B	-3.26	0.0011	0.02045194
	9091	-0.279558639 B	-15.42	0.0001	0.01812953
	9192	0.000000000 B	.	.	.
PULL	1	-0.181986272 B	-7.78	0.0001	0.02340174
	2	-0.060488236 B	-2.51	0.0120	0.02406278
	3	0.000000000 B	.	.	.
DEPTHCAT	2	-0.191800785 B	-4.24	0.0001	0.04527876
	3	-0.002518804 B	-0.05	0.9577	0.04750806
	4	0.000000000 B	.	.	.
REGION*MM	2 2	-0.164298891 B	-0.58	0.5637	0.28453991
	2 3	-0.162626241 B	-0.52	0.6003	0.31039788
	2 4	-0.106077474 B	-0.37	0.7108	0.28606405
	2 5	-0.091789953 B	-0.32	0.7481	0.28577038
	2 6	0.000000000 B	.	.	.
	3 2	0.090601931 B	0.33	0.7447	0.27819192
	3 3	0.775725987 B	2.86	0.0043	0.27126170
	3 4	1.483834921 B	5.30	0.0001	0.27978462
	3 5	0.574450228 B	1.76	0.0782	0.32612115
	3 6	0.000000000 B	.	.	.
	4 2	-0.462114459 B	-4.36	0.0001	0.10608492
	4 3	-0.040684290 B	-0.39	0.6955	0.10392638
	4 4	0.113728000 B	0.87	0.3828	0.13028898
	4 5	0.059463602 B	0.36	0.7189	0.16522196
	4 6	0.000000000 B	.	.	.
	5 2	-0.348972907 B	-3.13	0.0018	0.11164129
	5 3	0.315175240 B	3.05	0.0023	0.10349315
	5 4	0.432465961 B	4.23	0.0001	0.10219906
	5 5	0.474867830 B	4.15	0.0001	0.11440192
	5 6	0.000000000 B	.	.	.
	6 2	-0.326642637 B	-3.24	0.0012	0.10083953
	6 3	0.484046870 B	5.36	0.0001	0.09027303
	6 4	0.425132423 B	4.85	0.0001	0.08771217
	6 5	0.267717931 B	2.75	0.0060	0.09730871
	6 6	0.000000000 B	.	.	.
	7 2	-0.058511376 B	-0.64	0.5195	0.09083314
	7 3	0.308450976 B	3.60	0.0003	0.08563349
	7 4	0.231431400 B	2.84	0.0046	0.08154568
	7 5	0.208159534 B	2.35	0.0187	0.08847373
	7 6	0.000000000 B	.	.	.
	8 2	0.000000000 B	.	.	.
	8 3	0.000000000 B	.	.	.
	8 4	0.000000000 B	.	.	.
	8 5	0.000000000 B	.	.	.
	8 6	0.000000000 B	.	.	.
SNAVGPS1	0	-0.170921697 B	-8.73	0.0001	0.01958751
	1	0.000000000 B	.	.	.

NOTE: The X'X matrix has been found to be singular and a generalized inverse was used to solve the normal equations. Estimates followed by the letter 'B' are biased, and are not unique estimators of the parameters.

5.2.4 Residual Analysis

In this section we examine the residuals of the models in order to validate their assumptions. The reliability of the percentage increases in catch rates obtained from the parameter estimates in the previous section are to a large extent dependent on the assumptions of model specification, homoscedacity, normality, and independence. The accuracy of those estimates may also be greatly affected by influential observations. It is to these matters that we will now turn.

Consider first the residuals plotted against the predicted values of the model. Output 5.6a shows this plot for model A, involving CES and radar, and Output 5.6b show the plot for model B, involving GPS. Both plots show a broad scatter of points with no apparent trend or relationship between the residuals and the predicted values. Perhaps of concern are the few points in the first plot which lie to the left of the main group and which may be exerting too much influence in the model. These points are actually having the visual effect of compressing the main group laterally so that the horizontal axis can accommodate them, and thus misleadingly causing the plot to look stretched vertically. We will accept that both plots display enough randomness not to warrant further investigation.

The plots of the residuals against the predictor variables are considered next. Since the predictor variables are all categorical, we need to plot the residuals for the range of different levels they can take. By doing this we are actually checking if the variance of the underlying probability distribution is varying systematically with the predictors. All of the individual residuals are not plotted here, but an indication of their marginal distributions by plotting their mean, the first and third quantiles, and the lower and upper 95% limits of the residuals for each level of a predictor. We present, in Output 5.7a, the

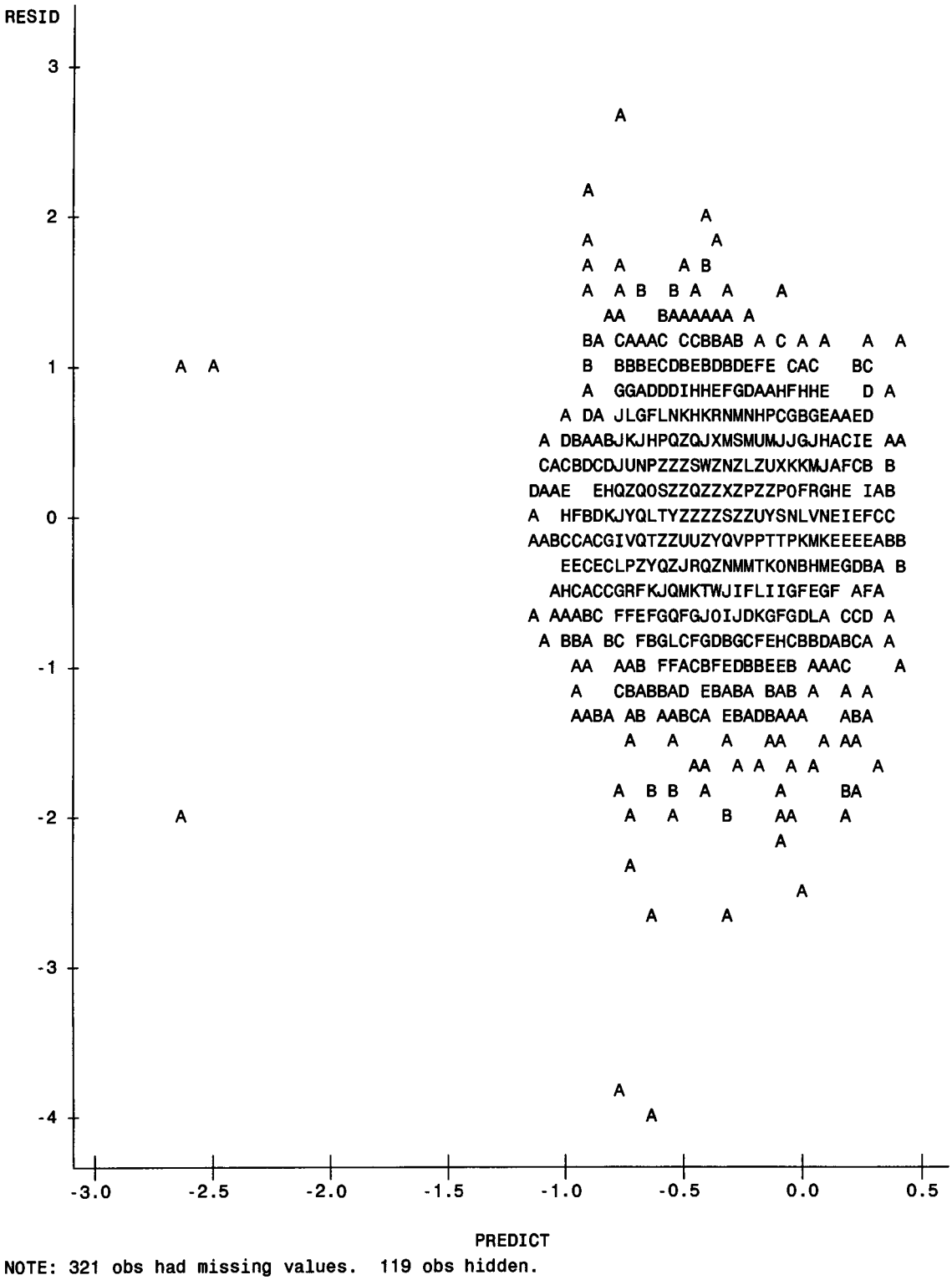
plot of the distribution of the residuals over the levels of the predictor, season, in model A. It is seen that the variability of the residuals is stable across the different seasons in the analysis. Output 5.7b shows the plot of the residuals over the levels of the predictor, depth, in model B. It is seen that the variability of the residuals is also stable for different ocean depths in the analysis.

The last check performed with the residuals is to see if they resemble a normal distribution. Since the models assumed that the errors are normally distributed, the standardised residuals should have a standard normal distribution if the probability distribution has been correctly specified. Outputs 5.8a and 5.8b, below, show the distributions of the standardised residuals resulting models A and B, respectively. It seems that both models have residuals that resemble a normal distribution. It should be noted, however, that both distributions appear slightly left skewed, with positive kurtosis.

The assessment of normality is now undertaken via a Q-Q plot, which compares how the distribution of the empirical errors compares with the theoretically assumed normal distribution. Outputs 5.9a and 5.9b show the Q-Q plots of the standardised residuals for models A and B, respectively. Both plots show that the residuals do not depart to any great extent from the line of normality. There is, however, a movement away from the line of normality near the extremes of the distribution; but not enough to warrant concern. It is likely that this feature is caused in part by the negative values of log catch rate resulting from the small constant added to zero catches. We will accept that the errors in model A and model B are normally distributed.

Output 5.6a

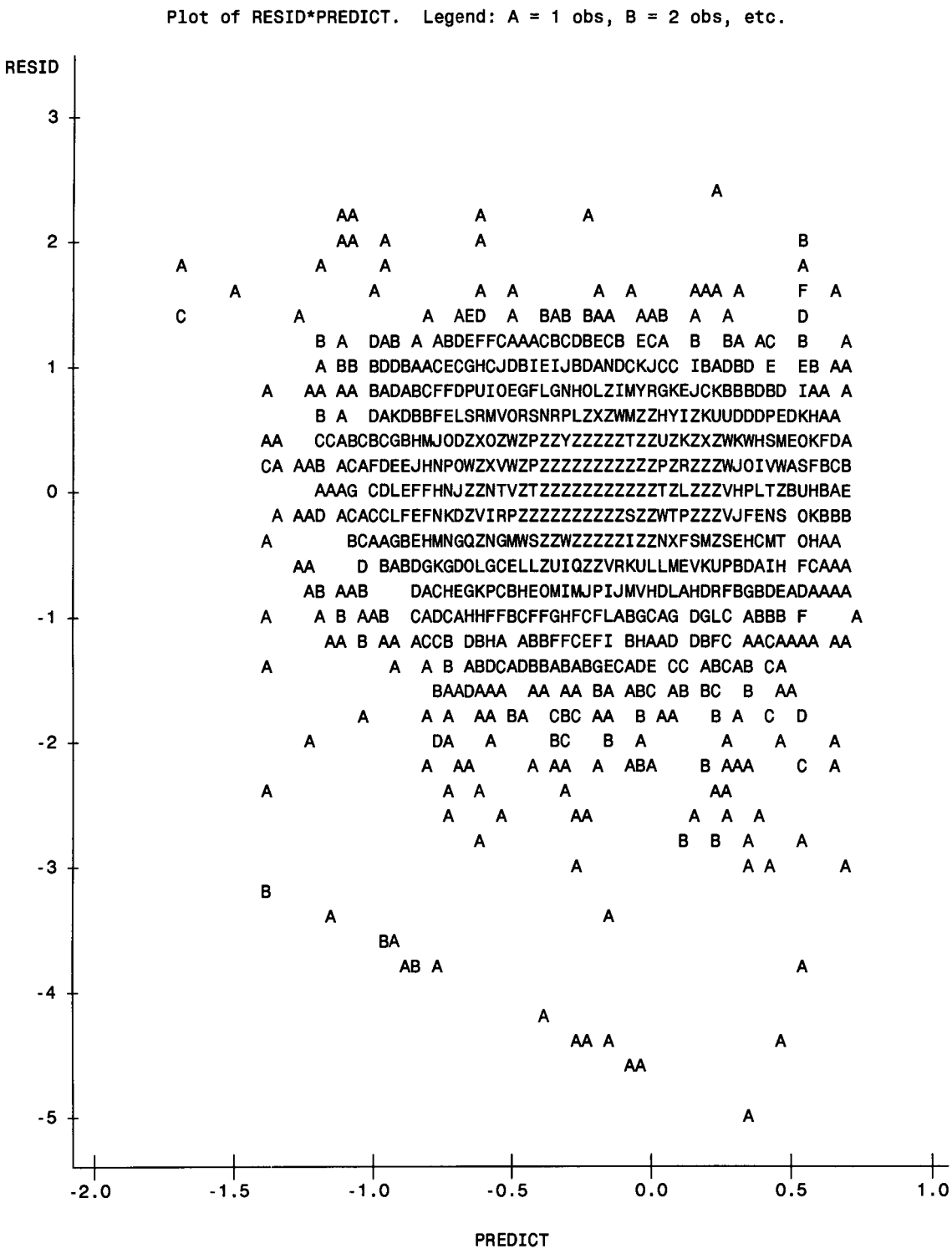
Residuals vs Predicted



NOTE: 321 obs had missing values. 119 obs hidden.

Output 5.6b

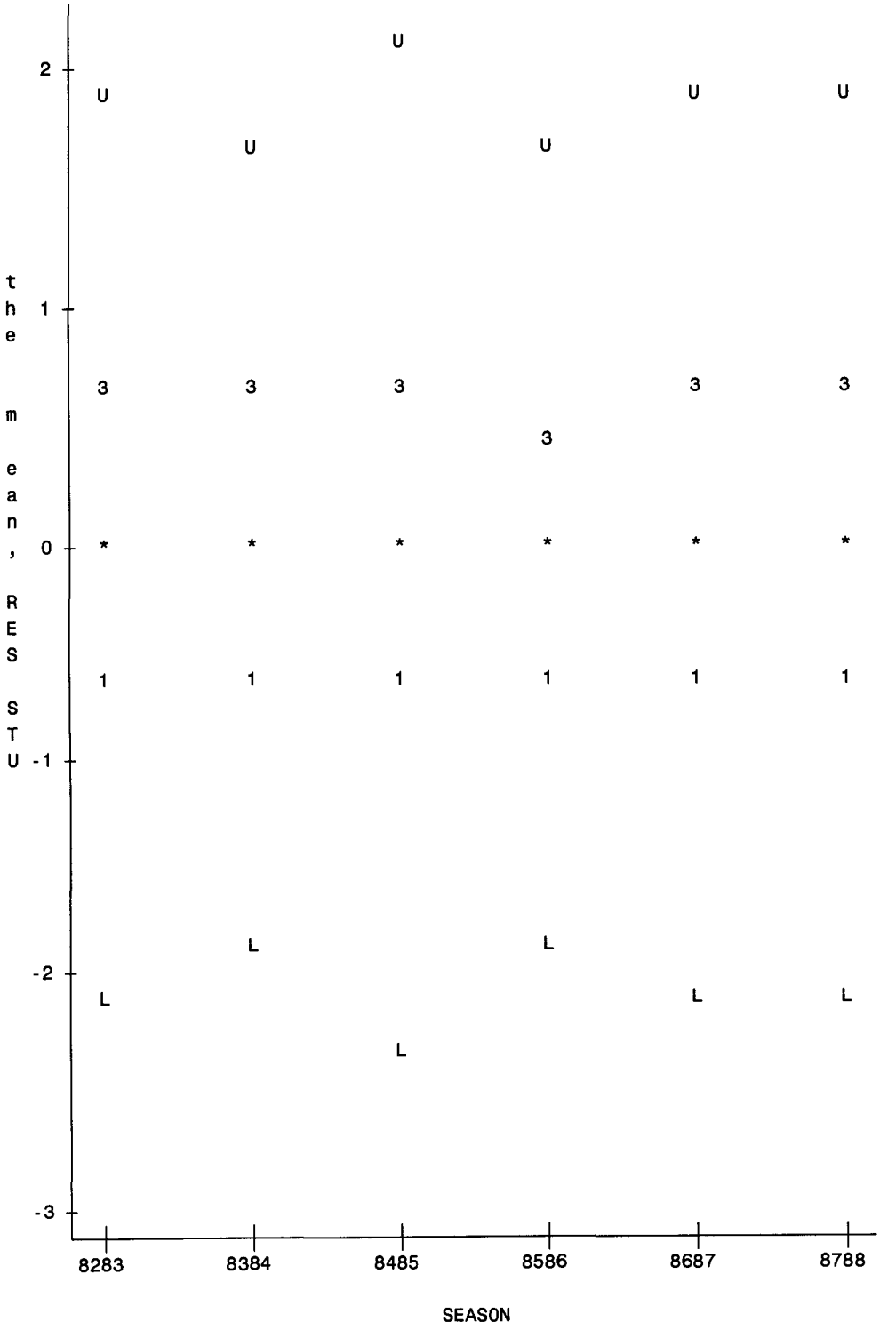
Residuals vs Predicted



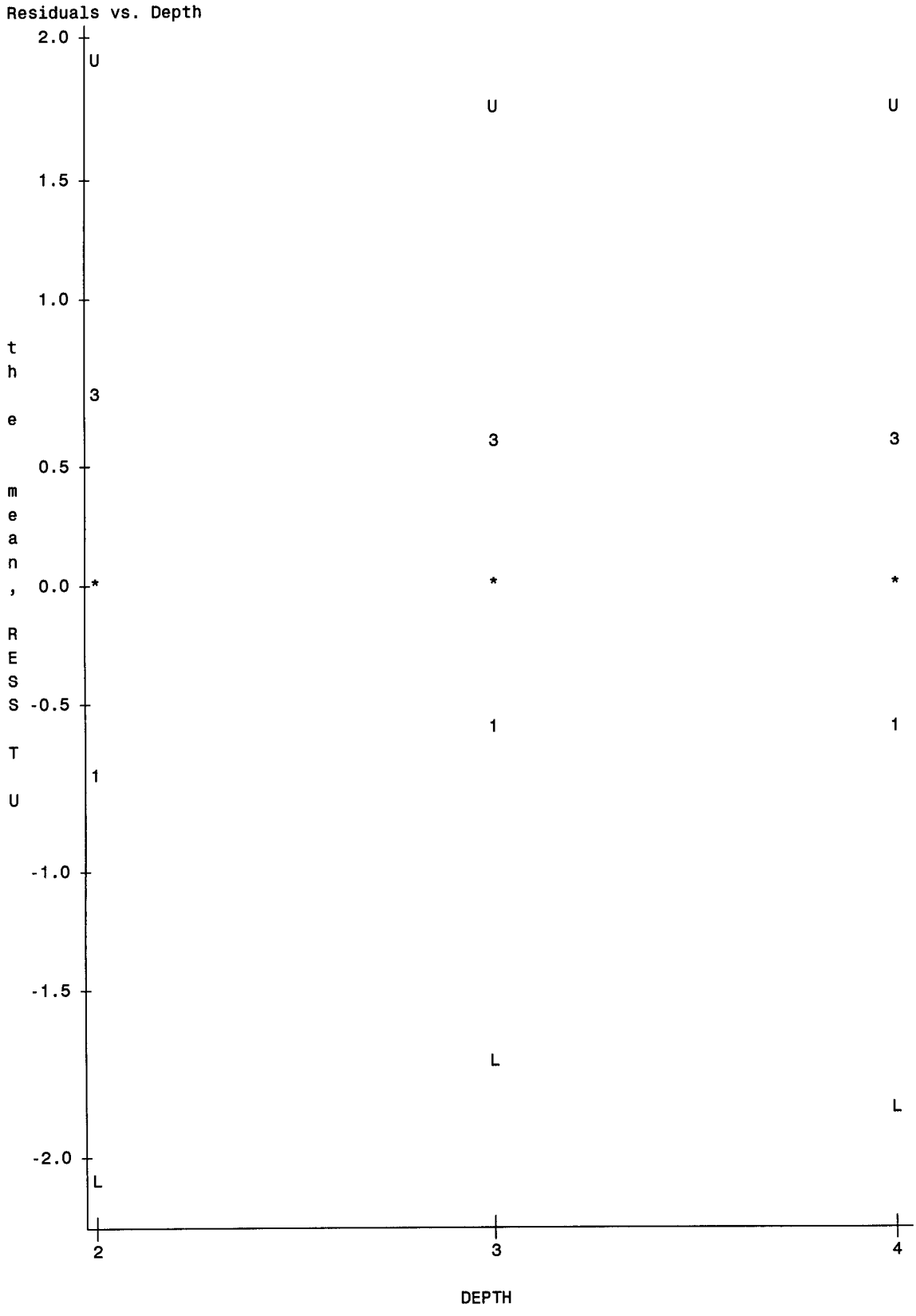
NOTE: 1138 obs had missing values. 1344 obs hidden.

Output 5.7a

Residuals vs. Season



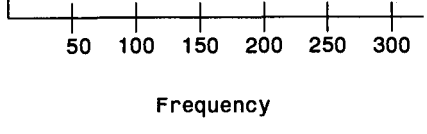
Output 5.7b


Residuals vs. Depth
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
the mean, RESSTU
U
3
1
L
*
2
3
4
DEPTH

Output 5.8a

RESSTU

Midpoint		Freq	Cum. Freq	Percent	Cum. Percent
-4.0000000000	*	7	7	0.19	0.19
-3.8000000000	*	6	13	0.16	0.35
-3.6000000000		2	15	0.05	0.41
-3.4000000000		1	16	0.03	0.43
-3.2000000000	*	5	21	0.14	0.57
-3.0000000000	*	8	29	0.22	0.78
-2.8000000000	*	9	38	0.24	1.03
-2.6000000000	*	9	47	0.24	1.27
-2.4000000000	**	18	65	0.49	1.76
-2.2000000000	***	32	97	0.87	2.62
-2.0000000000	****	41	138	1.11	3.73
-1.8000000000	*****	46	184	1.24	4.97
-1.6000000000	******	63	247	1.70	6.68
-1.4000000000	**********	102	349	2.76	9.43
-1.2000000000	**********	118	467	3.19	12.63
-1.0000000000	**********	164	631	4.43	17.06
-0.8000000000	**********	182	813	4.92	21.98
-0.6000000000	**********	235	1048	6.35	28.33
-0.4000000000	**********	266	1314	7.19	35.52
-0.2000000000	**********	283	1597	7.65	43.17
0.0000000000	**********	312	1909	8.43	51.61
0.2000000000	**********	320	2229	8.65	60.26
0.4000000000	**********	292	2521	7.89	68.15
0.6000000000	**********	270	2791	7.30	75.45
0.8000000000	**********	251	3042	6.79	82.24
1.0000000000	**********	193	3235	5.22	87.46
1.2000000000	**********	153	3388	4.14	91.59
1.4000000000	**********	112	3500	3.03	94.62
1.6000000000	*******	74	3574	2.00	96.62
1.8000000000	******	60	3634	1.62	98.24
2.0000000000	***	31	3665	0.84	99.08
2.2000000000	**	15	3680	0.41	99.49
2.4000000000	*	9	3689	0.24	99.73
2.6000000000	*	5	3694	0.14	99.86
2.8000000000		4	3698	0.11	99.97
3.0000000000		0	3698	0.00	99.97
3.2000000000		1	3699	0.03	100.00
3.4000000000		0	3699	0.00	100.00
3.6000000000		0	3699	0.00	100.00
3.8000000000		0	3699	0.00	100.00
4.0000000000		0	3699	0.00	100.00



Output 5.8b

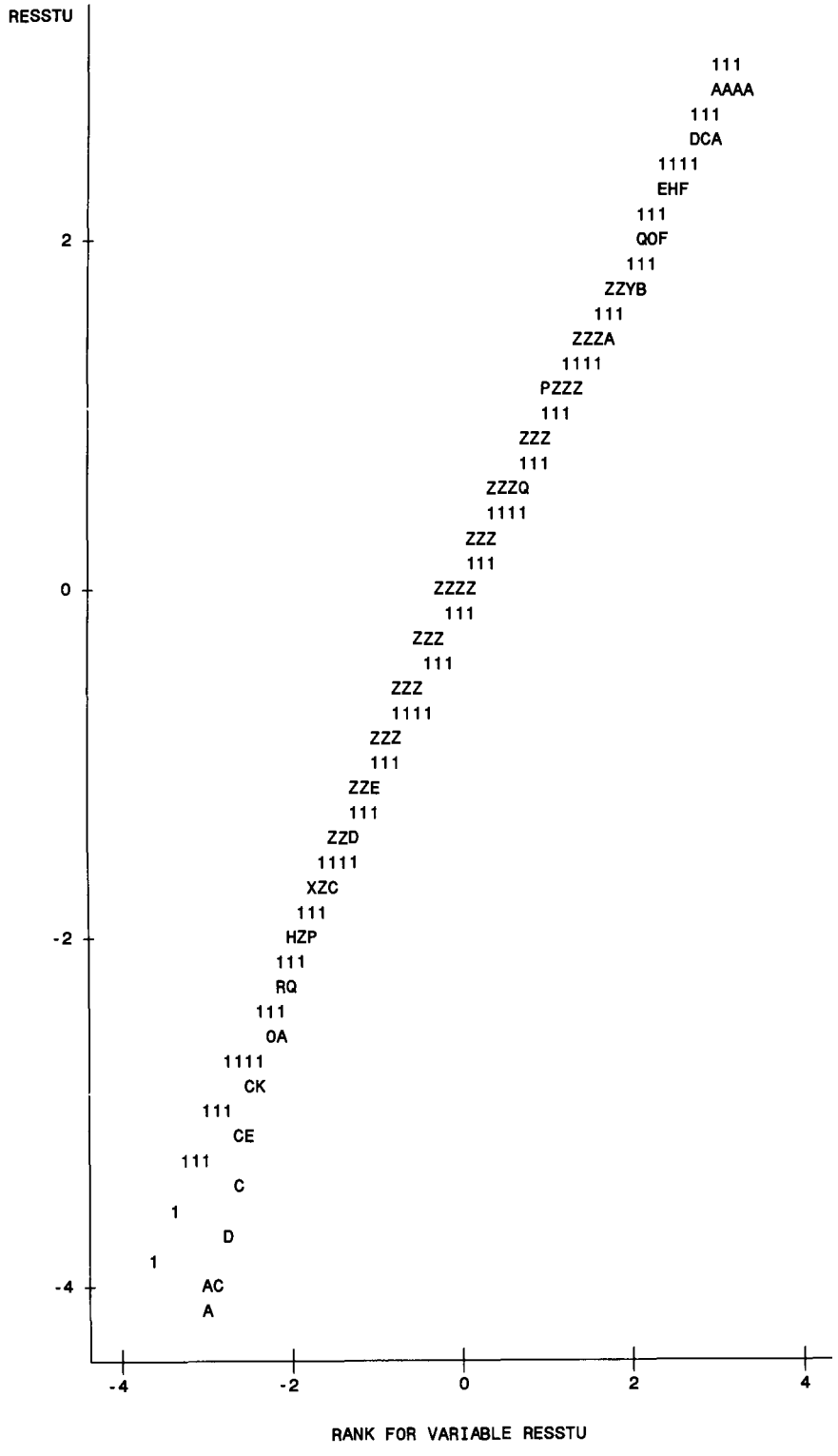
RESSTU

Midpoint		Freq	Cum. Freq	Percent	Cum. Percent
-4.0000000000	*	28	28	0.37	0.37
-3.8000000000	*	11	39	0.14	0.51
-3.6000000000		8	47	0.10	0.61
-3.4000000000	*	13	60	0.17	0.78
-3.2000000000	*	16	76	0.21	0.99
-3.0000000000	*	14	90	0.18	1.18
-2.8000000000	*	21	111	0.27	1.45
-2.6000000000	*	26	137	0.34	1.79
-2.4000000000	**	37	174	0.48	2.27
-2.2000000000	***	50	224	0.65	2.93
-2.0000000000	***	61	285	0.80	3.72
-1.8000000000	****	85	370	1.11	4.83
-1.6000000000	******	129	499	1.69	6.52
-1.4000000000	********	152	651	1.99	8.51
-1.2000000000	**********	227	878	2.97	11.47
-1.0000000000	**********	318	1196	4.16	15.63
-0.8000000000	**********	363	1559	4.74	20.37
-0.6000000000	**********	448	2007	5.85	26.23
-0.4000000000	**********	547	2554	7.15	33.37
-0.2000000000	**********	630	3184	8.23	41.60
0.0000000000	**********	687	3871	8.98	50.58
0.2000000000	**********	734	4605	9.59	60.17
0.4000000000	**********	672	5277	8.78	68.95
0.6000000000	**********	645	5922	8.43	77.38
0.8000000000	**********	534	6456	6.98	84.36
1.0000000000	**********	382	6838	4.99	89.35
1.2000000000	**********	266	7104	3.48	92.83
1.4000000000	**********	224	7328	2.93	95.75
1.6000000000	******	121	7449	1.58	97.33
1.8000000000	****	86	7535	1.12	98.46
2.0000000000	**	46	7581	0.60	99.06
2.2000000000	*	28	7609	0.37	99.43
2.4000000000	*	20	7629	0.26	99.69
2.6000000000		8	7637	0.10	99.79
2.8000000000		7	7644	0.09	99.88
3.0000000000		4	7648	0.05	99.93
3.2000000000		2	7650	0.03	99.96
3.4000000000		0	7650	0.00	99.96
3.6000000000		0	7650	0.00	99.96
3.8000000000		1	7651	0.01	99.97
4.0000000000		2	7653	0.03	100.00

100 200 300 400 500 600 700

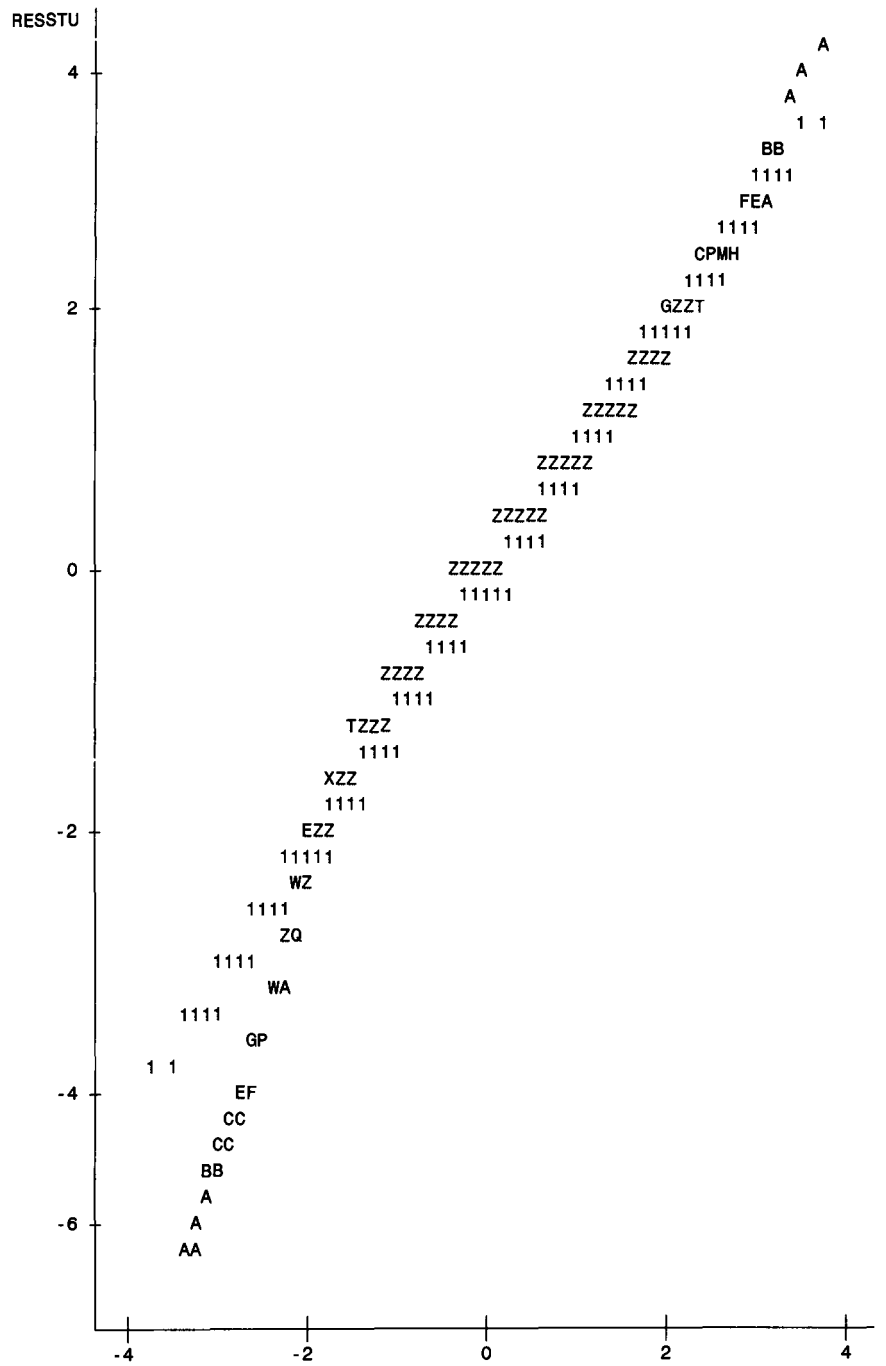
Frequency

Output 5.9a



NOTE: 644 obs had missing values. 6103 obs hidden.

Output 5.9b



NOTE: 2276 obs had missing values. 13858 obs hidden.

5.3 ESTIMATION OF FISHING POWER USING GENERALISED LINEAR MODELS

5.3.1 Motivation

This section will provide an alternative approach to the modelling of the catch and effort data. It will apply generalised linear models whose parameter estimates can be compared to the estimates obtained earlier from the classical regression models. The data exhibit the characteristic of an increasing variance as the mean increases. It was shown in the previous section that the log transformation sufficiently stabilised this relationship so that the assumption of constant variance in the classical linear regression model was not violated. Since the variance in the data varies in proportion to the square of the mean, and the catch rates take only non-negative values, the gamma distribution is appropriate. Further, instead of transforming the response variable, the original scale is retained and a log link function specified between the response and the systematic component of the model. The models will include the same predictor variables as in the previous sections, and the fishing power factors will be analysed over the same seasons as before.

5.3.2 Estimation

The catch and effort data were analysed with the SAS GENMOD procedure. Output 5.10 shows the results for the model with CES and radar and Output 5.11 shows the results for the model with GPS. A criterion for assessing if the models are adequate is the scaled deviance, which when close to unity, suggests that the model provides a good fit. We see that from the results below, the scaled deviances are 1.05 and 1.06 and we can tentatively assume that the models are adequate.

The analysis of residuals, however, will provide more surety about the models' appropriateness. The parameter estimates for the fishing power factors are 0.18 for CES, 0.14 for radar and 0.08 for GPS. In order to obtain the increases in catch rates associated

with these estimates, they need to be transformed back from the log link. These are given by:

CES: $e^{0.18} = 1.20$

Radar: $e^{0.14} = 1.15$

GPS: $e^{0.08} = 1.08$.

Hence, these results suggest that vessels with CES are catching 20% more than vessels without CES, vessels with radar are catching 15% more than vessels without radar, and vessels with GPS are catching 8% more than vessels without GPS.

Output 5.10

```
The SAS System

                    The GENMOD Procedure

                     Model Information

 Description                     Value

 Data Set                        RLDAT.LBGE7195
 Distribution                    GAMMA
 Link Function                   LOG
 Dependent Variable              CATRATE
 Scale Weight Variable           POT
 Observations Used               3697
 Invalid Response Values         3
 Missing Values                  321


                Class Level Information

 Class      Levels  Values

 REGION          7  2 3 4 5 6 7 8
 MM              5  2 3 4 5 6
 SEASON          6  8283 8384 8485 8586 8687 8788
 PULL            3  1 2 3
 DEPTHCAT        3  2 3 4
 COLECHO1        2  0 1
 RADAR1          2  0 1
```

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	3661	96291.1523	26.3019
Scaled Deviance	3661	3871.8940	1.0576
Pearson Chi-Square	3661	99844.5600	27.2725
Scaled Pearson X2	3661	4014.7775	1.0966
Log Likelihood	.	-1486.2476	.

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCEPT		1	0.1519	0.0747	4.1397	0.0419
REGION	2	1	-0.0837	0.4700	0.0317	0.8587
REGION	3	1	-1.2435	0.4562	7.4286	0.0064
REGION	4	1	-0.4089	0.1702	5.7679	0.0163
REGION	5	1	-0.0081	0.4357	0.0003	0.9851
REGION	6	1	-0.5477	0.1161	22.2589	0.0001
REGION	7	1	-0.0094	0.0853	0.0121	0.9123
REGION	8	0	0.0000	0.0000	.	.
MM	2	1	-0.0527	0.0581	0.8218	0.3646
MM	3	1	0.0322	0.0562	0.3291	0.5662

The SAS System

Analysis Of Parameter Estimates

Parameter			DF	Estimate	Std Err	ChiSquare	Pr>Chi
MM	4		1	-0.1027	0.0537	3.6617	0.0557
MM	5		1	-0.3257	0.0592	30.2375	0.0001
MM	6		0	0.0000	0.0000	.	.
SEASON	8283		1	0.1283	0.0293	19.1723	0.0001
SEASON	8384		1	-0.0652	0.0298	4.7745	0.0289
SEASON	8485		1	-0.2222	0.0297	55.9617	0.0001
SEASON	8586		1	-0.2274	0.0268	72.1368	0.0001
SEASON	8687		1	-0.1269	0.0322	15.5682	0.0001
SEASON	8788		0	0.0000	0.0000	.	.
PULL	1		1	-0.1743	0.0259	45.2135	0.0001
PULL	2		1	-0.0401	0.0249	2.5960	0.1071
PULL	3		0	0.0000	0.0000	.	.
DEPTHCAT	2		1	-0.1239	0.0500	6.1385	0.0132
DEPTHCAT	3		1	-0.0250	0.0579	0.1868	0.6656
DEPTHCAT	4		0	0.0000	0.0000	.	.
REGION*MM	2	2	1	0.1638	0.4800	0.1165	0.7329
REGION*MM	2	3	0	0.0000	0.0000	.	.
REGION*MM	3	4	0	0.0000	0.0000	.	.
REGION*MM	4	2	1	0.1355	0.1872	0.5233	0.4694
REGION*MM	4	3	0	0.0000	0.0000	.	.
REGION*MM	5	2	1	-0.2006	0.4393	0.2085	0.6480
REGION*MM	5	3	1	0.3158	0.4380	0.5198	0.4709
REGION*MM	5	4	1	0.6496	0.4375	2.2044	0.1376
REGION*MM	5	5	1	0.7764	0.4550	2.9117	0.0879
REGION*MM	5	6	0	0.0000	0.0000	.	.
REGION*MM	6	2	1	0.2945	0.1264	5.4320	0.0198
REGION*MM	6	3	1	0.7293	0.1216	35.9533	0.0001
REGION*MM	6	4	1	0.8360	0.1208	47.8849	0.0001
REGION*MM	6	5	1	0.7884	0.1391	32.1026	0.0001
REGION*MM	6	6	0	0.0000	0.0000	.	.
REGION*MM	7	2	1	-0.0486	0.0963	0.2550	0.6136
REGION*MM	7	3	1	0.1251	0.0962	1.6905	0.1935
REGION*MM	7	4	1	0.0672	0.0921	0.5327	0.4655
REGION*MM	7	5	1	-0.0471	0.1076	0.1920	0.6613
REGION*MM	7	6	0	0.0000	0.0000	.	.
REGION*MM	8	2	0	0.0000	0.0000	.	.
REGION*MM	8	3	0	0.0000	0.0000	.	.
REGION*MM	8	4	0	0.0000	0.0000	.	.
REGION*MM	8	5	0	0.0000	0.0000	.	.
REGION*MM	8	6	0	0.0000	0.0000	.	.
COLECH01	0		1	-0.1389	0.0235	35.0055	0.0001
COLECH01	1		0	0.0000	0.0000	.	.
RADAR1	0		1	-0.1648	0.0257	41.1395	0.0001
RADAR1	1		0	0.0000	0.0000	.	.
SCALE			1	0.0402	0.0009	.	.

NOTE: The scale parameter was estimated by maximum likelihood.

The SAS System

LR Statistics For Type 1 Analysis

Source	Deviance	DF	ChiSquare	Pr>Chi
INTERCEPT	130831.085	0	.	.
REGION	118685.443	6	381.7194	0.0001
MM	113404.203	4	177.7086	0.0001
SEASON	108780.145	5	162.1950	0.0001
PULL	107591.682	2	42.7495	0.0001
DEPTHCAT	107237.705	2	12.8199	0.0016
REGION*MM	101500.604	14	213.6193	0.0001
COLECH01	97308.3980	1	163.5309	0.0001
RADAR1	96291.1523	1	40.6980	0.0001

LR Statistics For Type 3 Analysis

Source	DF	ChiSquare	Pr>Chi
REGION	6	35.3109	0.0001
MM	4	81.3989	0.0001
SEASON	5	219.6821	0.0001
PULL	2	65.1093	0.0001
DEPTHCAT	2	12.5908	0.0018
REGION*MM	14	236.8838	0.0001
COLECH01	1	34.8870	0.0001
RADAR1	1	40.6980	0.0001

Output 5.11

The GENMOD Procedure

Model Information

Description	Value
Data Set	RLDAT.LBGE7195
Distribution	GAMMA
Link Function	LOG
Dependent Variable	CATRATE
Scale Weight Variable	POT
Observations Used	7636
Invalid Response Values	17
Missing Values	1138

Class Level Information

Class	Levels	Values
REGION	7	2 3 4 5 6 7 8
MM	5	2 3 4 5 6
SEASON	3	8990 9091 9192
PULL	3	1 2 3
DEPTHCAT	3	2 3 4
SNAVGPS1	2	0 1

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	7594	205579.3364	27.0713
Scaled Deviance	7594	8093.0381	1.0657
Pearson Chi-Square	7594	200634.2722	26.4201
Scaled Pearson X2	7594	7898.3658	1.0401
Log Likelihood	.	-5342.3651	.

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCEPT		1	0.1985	0.0682	8.4589	0.0036
REGION	2	1	0.6947	0.2533	7.5241	0.0061
REGION	3	1	-0.4553	0.2553	3.1810	0.0745
REGION	4	1	0.1381	0.0817	2.8552	0.0911
REGION	5	1	0.1267	0.0860	2.1671	0.1410
REGION	6	1	-0.2273	0.0728	9.7557	0.0018
REGION	7	1	-0.0899	0.0659	1.8636	0.1722
REGION	8	0	0.0000	0.0000	.	.
MM	2	1	0.1585	0.0638	6.1647	0.0130
MM	3	1	0.1827	0.0614	8.8458	0.0029
MM	4	1	0.1166	0.0577	4.0803	0.0434

The SAS System

Analysis Of Parameter Estimates

Parameter			DF	Estimate	Std Err	ChiSquare	Pr>Chi
MM	5		1	-0.1791	0.0629	8.1063	0.0044
MM	6		0	0.0000	0.0000	.	.
SEASON	8990		1	-0.0743	0.0191	15.1260	0.0001
SEASON	9091		1	-0.2733	0.0167	266.6342	0.0001
SEASON	9192		0	0.0000	0.0000	.	.
PULL	1		1	-0.2296	0.0214	114.6479	0.0001
PULL	2		1	-0.1029	0.0220	21.8341	0.0001
PULL	3		0	0.0000	0.0000	.	.
DEPTHCAT	2		1	-0.1976	0.0438	20.3393	0.0001
DEPTHCAT	3		1	-0.0173	0.0461	0.1402	0.7081
DEPTHCAT	4		0	0.0000	0.0000	.	.
REGION*MM	2	2	1	0.0716	0.2608	0.0753	0.7838
REGION*MM	2	3	1	-0.2010	0.2835	0.5026	0.4784
REGION*MM	2	4	1	-0.0746	0.2611	0.0817	0.7750
REGION*MM	2	5	1	-0.1443	0.2611	0.3056	0.5804
REGION*MM	2	6	0	0.0000	0.0000	.	.
REGION*MM	3	2	1	-0.1964	0.2686	0.5350	0.4645
REGION*MM	3	3	1	0.3346	0.2622	1.6289	0.2019
REGION*MM	3	4	1	0.9419	0.2695	12.2143	0.0005
REGION*MM	3	5	1	0.6516	0.3162	4.2461	0.0393
REGION*MM	3	6	0	0.0000	0.0000	.	.
REGION*MM	4	2	1	-0.3839	0.0970	15.6622	0.0001
REGION*MM	4	3	1	0.0280	0.0950	0.0869	0.7681
REGION*MM	4	4	1	0.2266	0.1191	3.6198	0.0571
REGION*MM	4	5	1	0.0046	0.1509	0.0009	0.9754
REGION*MM	4	6	0	0.0000	0.0000	.	.
REGION*MM	5	2	1	-0.3080	0.1019	9.1346	0.0025
REGION*MM	5	3	1	0.2550	0.0945	7.2778	0.0070
REGION*MM	5	4	1	0.3747	0.0934	16.1062	0.0001
REGION*MM	5	5	1	0.4721	0.1044	20.4354	0.0001
REGION*MM	5	6	0	0.0000	0.0000	.	.
REGION*MM	6	2	1	-0.3287	0.0924	12.6568	0.0004
REGION*MM	6	3	1	0.4682	0.0828	31.9394	0.0001
REGION*MM	6	4	1	0.4060	0.0804	25.4766	0.0001
REGION*MM	6	5	1	0.2075	0.0892	5.4062	0.0201
REGION*MM	6	6	0	0.0000	0.0000	.	.
REGION*MM	7	2	1	-0.1393	0.0830	2.8157	0.0933
REGION*MM	7	3	1	0.2129	0.0782	7.4035	0.0065
REGION*MM	7	4	1	0.1539	0.0745	4.2713	0.0388
REGION*MM	7	5	1	0.1180	0.0808	2.1320	0.1443
REGION*MM	7	6	0	0.0000	0.0000	.	.
REGION*MM	8	2	0	0.0000	0.0000	.	.
REGION*MM	8	3	0	0.0000	0.0000	.	.
REGION*MM	8	4	0	0.0000	0.0000	.	.
REGION*MM	8	5	0	0.0000	0.0000	.	.
REGION*MM	8	6	0	0.0000	0.0000	.	.
SNAVGPS1	0		1	-0.1572	0.0181	75.0934	0.0001
SNAVGPS1	1		0	0.0000	0.0000	.	.
SCALE			1	0.0394	0.0006	.	.

The SAS System

NOTE: The scale parameter was estimated by maximum likelihood.

LR Statistics For Type 1 Analysis

Source	Deviance	DF	ChiSquare	Pr>Chi
INTERCEPT	303742.323	0	.	.
REGION	247963.563	6	1668.4214	0.0001
MM	232078.109	4	540.1741	0.0001
SEASON	224774.419	2	260.1944	0.0001
PULL	222200.529	2	93.6063	0.0001
DEPTHCAT	217985.224	2	155.5435	0.0001
REGION*MM	207479.655	24	400.4296	0.0001
SNAVGPS1	205579.336	1	74.4832	0.0001

LR Statistics For Type 3 Analysis

Source	DF	ChiSquare	Pr>Chi
REGION	6	336.7056	0.0001
MM	4	351.7367	0.0001
SEASON	2	275.0356	0.0001
PULL	2	144.1641	0.0001
DEPTHCAT	2	79.4025	0.0001
REGION*MM	24	422.6742	0.0001
SNAVGPS1	1	74.4832	0.0001

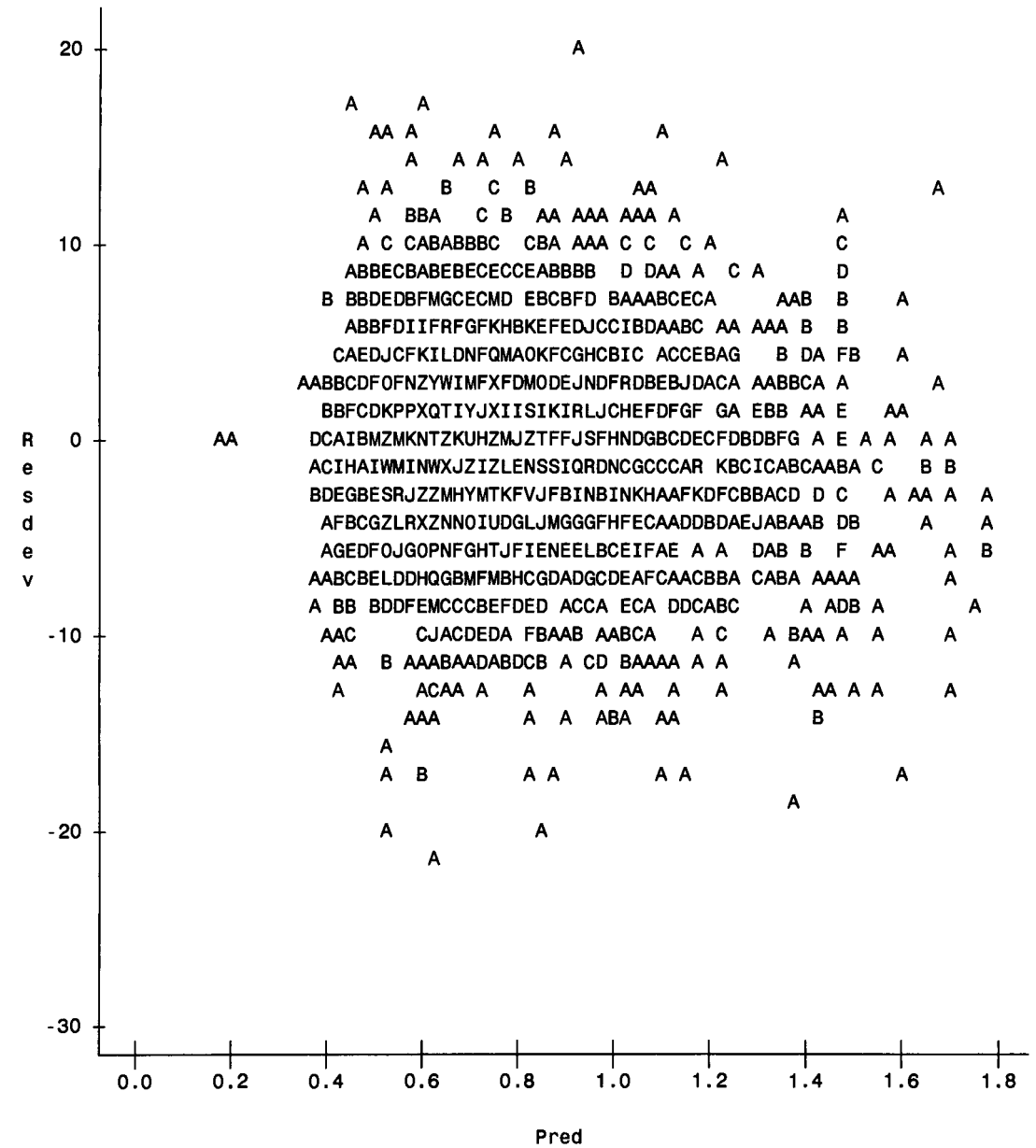
5.3.3 Residual Analysis

The generalised linear models that were estimated are further investigated by examining their residuals. Output 5.12 shows the residual analysis for the first model. Shown first is the output is the plot of the deviance residuals against the predicted values of the model. It shows no cause for concern about its randomness. Next is the output from the SAS UNIVARIATE procedure, which has given various statistics about the distribution of the standardised residuals, along with a histogram, boxplot and normal probability plot. The statistics show some departure from normality. For example the D statistic, which tests the hypothesis that the distribution is normal, has rejected such a hypothesis at the 0.03 level. The plots also show some non-normal features, such as mild skewness and kurtosis in the histogram, while the probability plot has problems at the extremes. However, these departures are not gross.

Output 5.13 shows the residual analysis for the second model. First is the plot of the deviance residuals against the predicted values of the model. It also shows no cause for concern about its randomness. Next is the output from the SAS UNIVARIATE procedure, which as in the first model, show some departure from normality. The probability plot again seems reasonable except at the extremes. These sorts of problems are virtually inevitable with this kind of data, and if we are willing to accept this then we can also accept the models with some degree of caution.

Output 5.12

Plot of RESDEV*PRED. Legend: A = 1 obs, B = 2 obs, etc.



NOTE: 324 obs had missing values. 41 obs hidden.

Univariate Procedure

Variable=STDDEV

Moments

				Quantiles(Def=5)			
N	3696	Sum Wgts	3696	100% Max	4.007469	99%	2.253874
Mean	-0.15302	Sum	-565.571	75% Q3	0.511196	95%	1.530675
Std Dev	1.018463	Variance	1.037267	50% Med	-0.16966	90%	1.173572
Skewness	0.058513	Kurtosis	0.400482	25% Q1	-0.81882	10%	-1.41333
USS	3919.245	CSS	3832.7	0% Min	-4.22759	5%	-1.76602
CV	-665.565	Std Mean	0.016752			1%	-2.54077
T:Mean=0	-9.13431	Pr>|T|	0.0001	Range	8.235058		
Num ^= 0	3696	Num > 0	1588	Q3-Q1	1.330016		
M(Sign)	-260	Pr>=|M|	0.0001	Mode	0.214464		
Sgn Rank	-612104	Pr>=|S|	0.0001				
D:Normal	0.015489	Pr>D	0.0305				

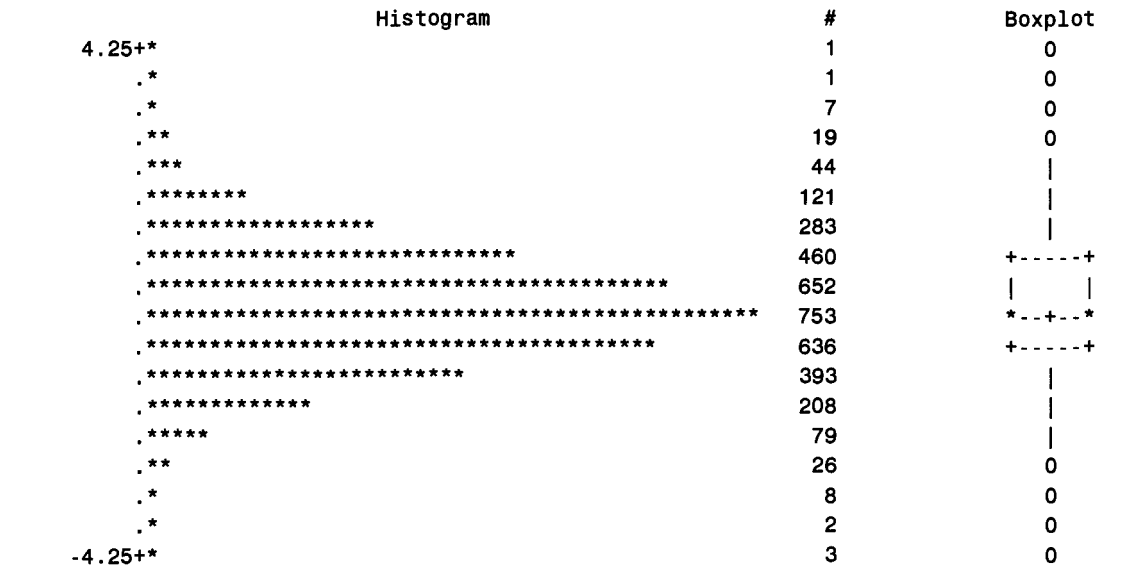
Extremes

Lowest	Obs	Highest	Obs
-4.22759(	1485)	3.275488(	3850)
-4.11667(	1216)	3.286267(	1970)
-4.09153(	1867)	3.448541(	1734)
-3.66978(	3378)	3.555491(	2696)
-3.52825(	906)	4.007469(	1234)

Missing Value	.
Count	325
% Count/Nobs	8.08

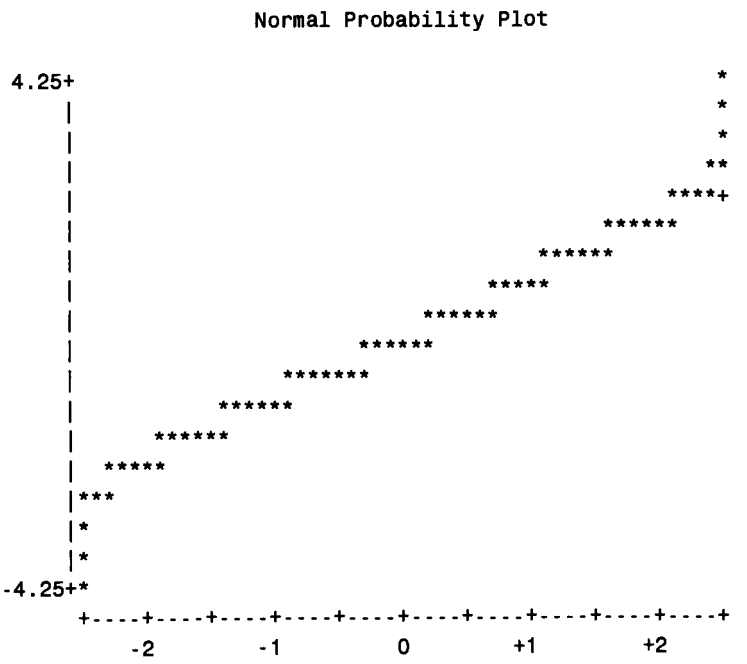
Univariate Procedure

Variable=STDDEV



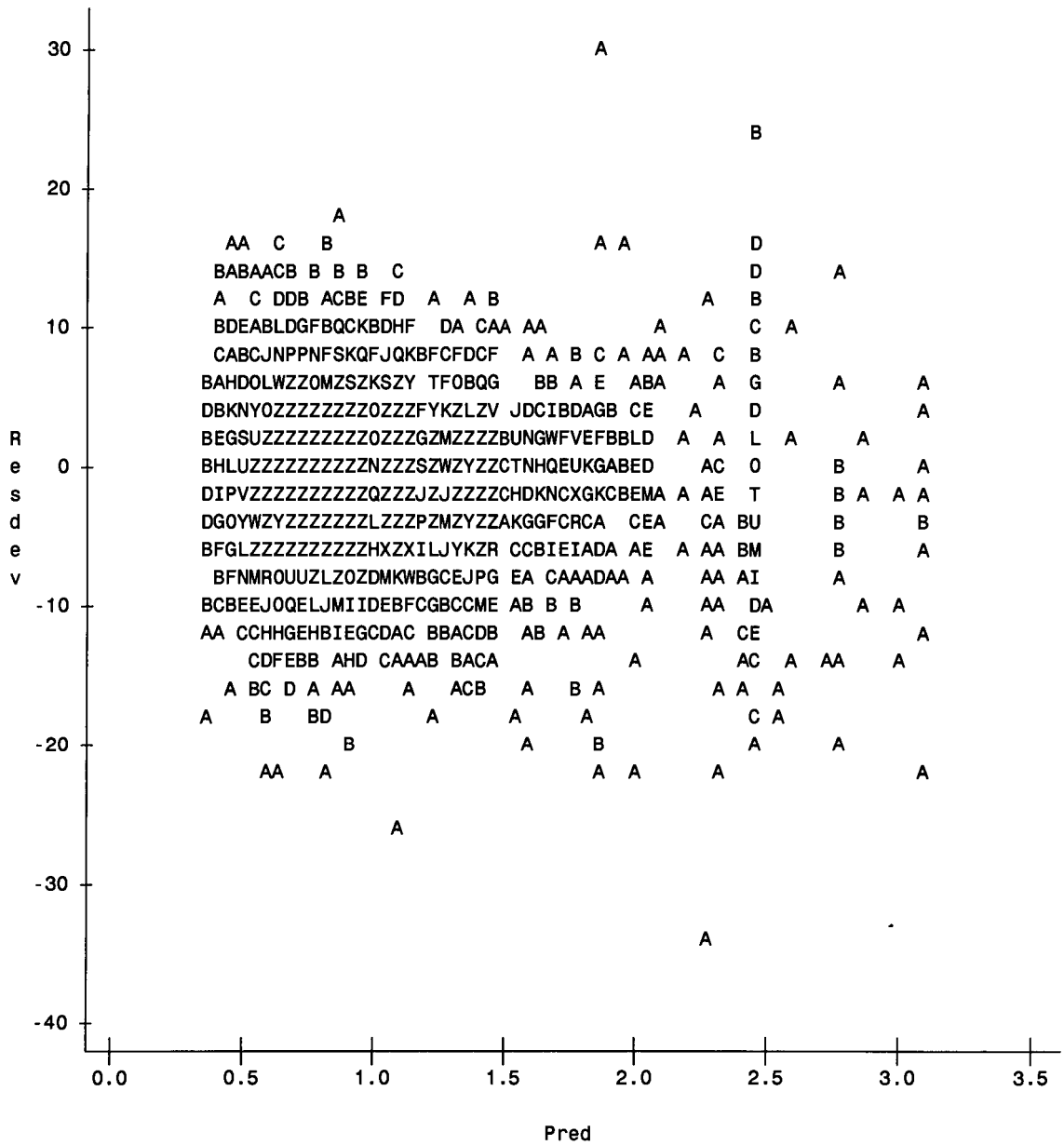
Univariate Procedure

Variable=STDDEV


Normal Probability Plot
4.25+
-4.25+
+----+----+----+----+----+----+----+----+----+----+----+
-2
-1
0
+1
+2

Output 5.13

Plot of RESDEV*PRED. Legend: A = 1 obs, B = 2 obs, etc.



NOTE: 1155 obs had missing values. 2239 obs hidden.

Univariate Procedure

Variable=STDDEV

Moments

N	7636	Sum Wgts	7636				
Mean	-0.16656	Sum	-1271.82				
Std Dev	1.019323	Variance	1.03902				
Skewness	-0.14579	Kurtosis	1.272961				
USS	8144.748	CSS	7932.918				
CV	-612	Std Mean	0.011665				
T:Mean=0	-14.2785	Pr>|T|	0.0001				
Num ^= 0	7636	Num > 0	3308				
M(Sign)	-510	Pr>=|M|	0.0001				
Sgn Rank	-2652307	Pr>=|S|	0.0001				
D:Normal	0.022428	Pr>D	<.01				

Quantiles(Def=5)

100% Max	6.060338	99%	2.266055
75% Q3	0.480152	95%	1.457047
50% Med	-0.14743	90%	1.060106
25% Q1	-0.79637	10%	-1.39695
0% Min	-6.7659	5%	-1.83767
		1%	-2.8108
Range	12.82623		
Q3-Q1	1.276521		
Mode	-0.77006		

Extremes

Lowest	Obs	Highest	Obs
-6.7659(	2523)	3.361867(	6470)
-5.10934(	380)	3.734845(	4206)
-4.44466(	3223)	4.715648(	8640)
-4.44122(	6464)	4.950618(	8025)
-4.4112(	5669)	6.060338(	5035)

Missing Value	.
Count	1155
% Count/Nobs	13.14

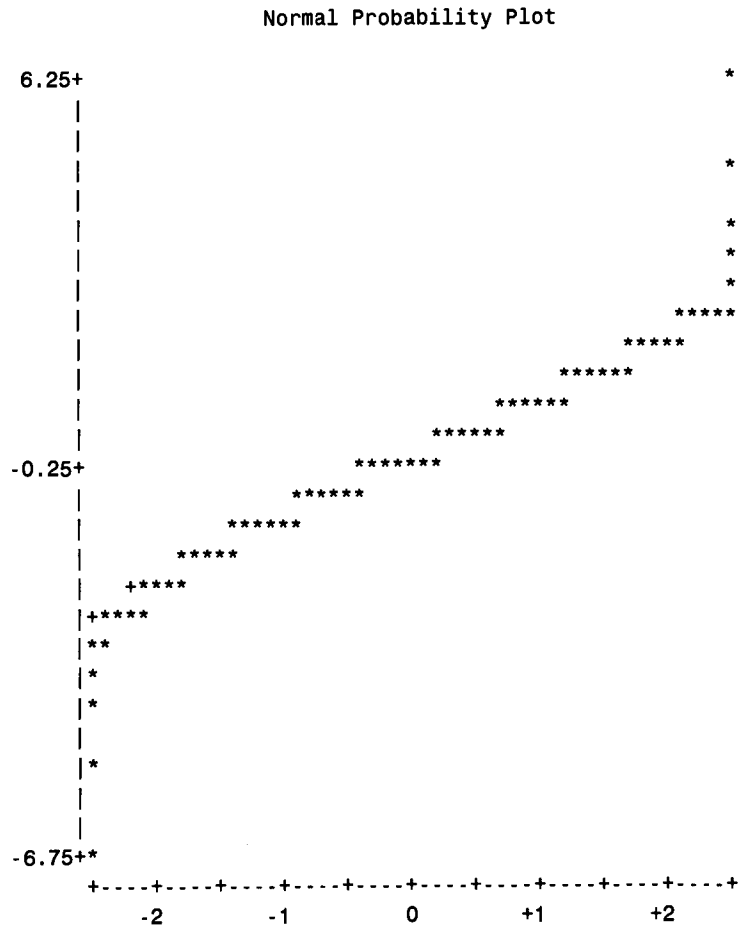
Univariate Procedure

Variable=STDDEV

```
                    Histogram                          #          Boxplot
      6.25+*                                           1             *
          .*                                           2             *
          .                                                          
          .*                                           1             0
          .*                                          13             0
          .**                                         35             0
          .***                                        71             0
          .*******                                   234             |
          .***************                           484             |
          .*****************************            1019             |
          .******************************************* 1448        +-----+
     -0.25+************************************************** 1626  *--+--*
          .*********************************          1248        +-----+
          .************************                    817             |
          .**********                                 335             |
          .******                                     172             |
          .***                                         72             0
          .*                                           32             0
          .*                                           16             0
          .*                                            8             0
          .                                                          
          .*                                           1             *
     -6.75+*                                           1             *
          ----+----+----+----+----+----+----+----+----+---
          * may represent up to 34 counts
```

Univariate Procedure

Variable=STDDEV


Normal Probability Plot
6.25+
-0.25+
-6.75+
-2
-1
0
+1
+2

Chapter 6

CONCLUSION

6.1 DISCUSSION OF ESTIMATES OF FISHING POWER FACTORS

Having fitted both general linear regression and generalised linear models to the catch and effort data, it is time to make a comparison of the two methods, their parameter estimates and their usefulness for modelling this kind of data. The regression followed standard procedures for modelling catch and effort data, using the well-established log transformation. It assumed normality in the error structure after the log transformation. The generalised linear model assumed a gamma structure in the model but also made use of the log link function between the response and the systematic part of the model. The variance function in the gamma model proved to have a similar effect to the log transformation. The results indicated that the method of linear regression with the log transformation yielded results and residual characteristics which were, in general, similar to those obtained by the generalised linear models method. The residual analyses for both methods yielded reasonably acceptable results. We saw that after the models were fit there were still some mild non-normal characteristics, but, in the main, these were acceptable.

It was likely that, with data of a biological and environmental nature as these, our models were not going to be very successful at accounting for much of the variation in the data. This does not, however preclude their usefulness in estimating the effects of certain variables of interest. Taken over many observations and seasons, the effect of

environmental factors on the model need not be of much concern because all vessels have to operate within the context of the same conditions. It would, however, be worthwhile to investigate the possibility that fishing power factors have differential effects under different environmental conditions. For example, a GPS may be more effective when the windspeed is greater. The models in this research, nevertheless, were intended not to explain all of the variation in the system but to obtain measures of how vessels are fairing relative to one another within the same environmental conditions of that system. With only a few variables and a large number of observations, we saw that the models and their parameter estimates were significant.

The parameter estimates obtained from both methods were all negative for vessels without the fishing power factors and showed that the fishing power factors were having a positive effect for vessels equipped with them. The regression models yielded estimates of fishing efficiency increases of 12%, 20% and 18% for CES, radar and GPS, respectively. The generalised linear models yielded corresponding efficiency increases of 20%, 15% and 8%. What can we say about the differences in these estimates? We know that they indicate positive fishing power effects, but we are not sure as to the extent of those effects. If we take the nature of the data and methodology into account, it may be prudent not to try to quantify these effects to a definite level. It would be acceptable to say that these particular pieces of equipment yield catch rates which are approximately 15% greater. On the other hand, in order to incorporate fishing power into the stock assessment models, the results need to be quantified at some level so that fishing effort can be standardised for each season.

The results in this study compare similarly to findings of previous research. The stabilisation of the variance using the log transformation confirmed the findings of

Gulland (1956) and Beverton and Holt (1957) that such a transformation was appropriate for catch and effort data. The use of least squares regression methods found in Allen and Punsley (1984) and Large (1992), which modelled the catch rate as a function of several variables, including fishing power factors, is supported by this research, which also validates their use for this kind of analyses. The estimated impact of GPS on the catch rates was similar to that estimated by Robins *et al.* (1996) for the northern prawn fishery fleet. The overall results for the impact of fishing power factors is also supportive of the preliminary analysis of the fishing power increases in the western rock lobster fishery (Brown *et al.*, 1995).

In this dissertation it was necessary to limit to scope of the fishing power analysis. These limitations included the number of fishing power factors and other variables in the models, and the number of observations were restricted to certain geographical regions, months of the year and seasons. However, in order to incorporate fishing power into the abundance models it may be necessary to perform a thorough analysis of all fishing power factors, over a comprehensive range of spatial and temporal dimensions, and it is acknowledged that more research into this area is necessary. However, from the results in this research, we can be satisfied that there are increases in efficiency associated with the factors used in this study, and that these can be applied to the current effort estimates of the fishery at least on an experimental basis.

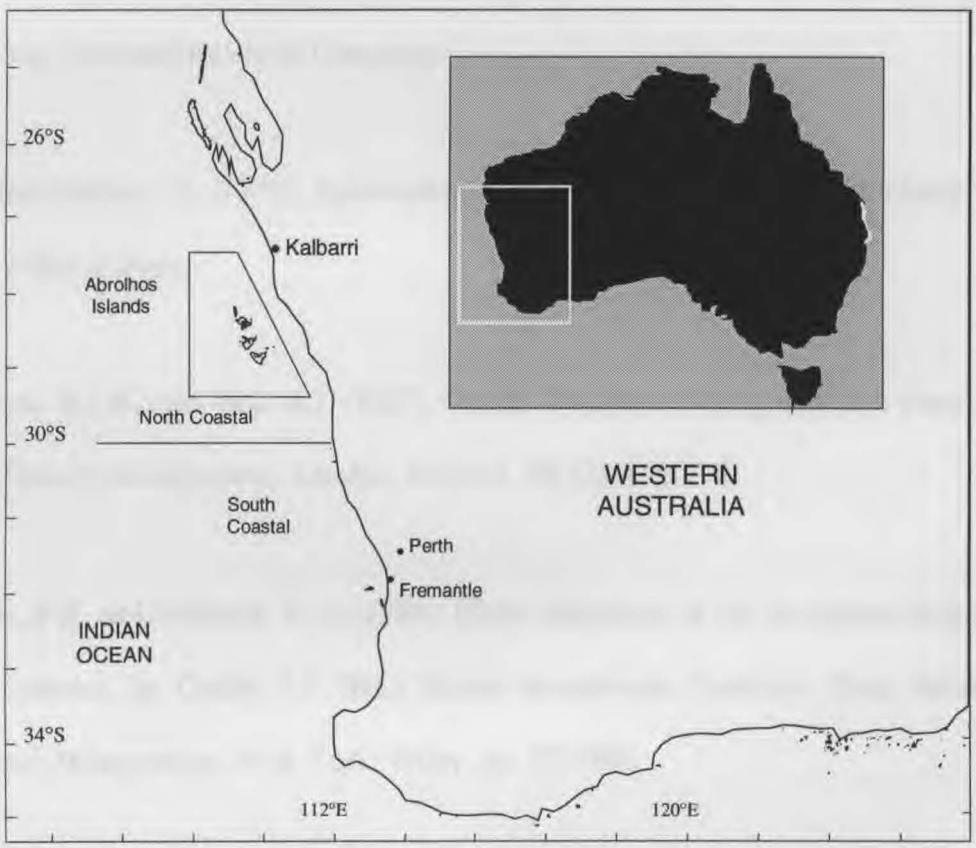
6.2 IMPLICATIONS FOR THE FISHERY

The findings of this research are useful for assessing the relative fishing power of the fleet in the western rock lobster fishery. The fishing efficiency increases can be applied to the models which estimate the overall effort being applied in the fishery. The model by Phillips and Brown (1989) which takes into account the spatial and temporal variation in

fishing effort could incorporate the estimated increases in fishing power in its effective effort calculations. Further, the abundance models which use those effort figures, for example, Walters *et al.* (1993), will be better able to describe and predict the fishery's course, and hence be better equipped to serve its management.

APPENDIX A

West coast of Western Australia showing western rock lobster fishing zones.



REFERENCES

Allen, R. and Punsly, R. (1984). Catch rates as indices of abundance of yellowfin tuna, *thunnus albacares*, in the eastern Pacific ocean. *Inter-American Tropical Tuna Commission Bulletin, LaJolla, Cal., 18(4)*: 303-379.

Afifi, A.A. and Clark, V. (1984). *Computer-Aided Multivariate Analysis.* New York: Van Nostrand Reinhold Company.

Barndorf-Nielsen, O. (1978). *Information and Exponential Families.* Chichester: John Wiley & Sons.

Beverton, R.J.H., and Holt, S.J. (1957). On the dynamics of exlpoited fish populations. *Fishery Invesigations, London, Series 2, 19(12)*: 172-178.

Bowen, B.K. and Hancock, D.A. (1989). Effort limitations in the Australian rock lobster fisheries. In: Caddy, J.F. (Ed.) Marine Invertebrate Fisheries: Their Assessment and Management. New York: Wiley, pp. 375-393.

Brown, R.S., Caputi, N. and Barker, E. (1995). A preliminary assessment of increases in fishing power on stock assessment and fishing effort expended in the western rock lobster (*Panulirus cygnus*) fishery. Proceedings 4th Internat. Workshop Lobster Biol. and Manage., 1993. *Crustaceana 68(2)*: 227-237.

Campbell, R.A. (in prep.). Notes on statistical modelling and the use of general linear models in the analysis of catch and effort data. CSIRO Marine Laboratories. Hobart.

Caputi, N. (in prep). Analysis of catch and effort data for redfish, blue grenadier and orange roughy using generalised linear models. Report for Bureau of resource sciences, department of primary industries and energy.

Caputi, N., Brown, R.S. and Phillips, B.F. (1995). Predicting catches of the western rock lobster (*Panulirus cygnus*) based on indices of puerulus and juvenile abundance. *ICES mar. Sci. Symp., 199*: 287-293.

Caputi, N., Chubb, C.F. and Brown, R.S. (1993). Relationship between spawning stock, environment, recruitment and fishing effort for the western rock lobster, Panulirus cygnus, fishery in Western Australia. Proceedings 4th Internat. Workshop Lobster Biol. and Manage., 1993.

Dobson, A.J. (1990). *An Introduction to Generalized Linear Models*. London: Chapman and Hall.

Draper, N.R. and Smith, H. (1981). *Applied Regression Analysis*. (2nd ed.). New York: John Wiley & Sons.

Fahrmeir, L. and Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models*. New York: Springer-Verlag.

Gavaris, S. (1980). Use of a multiplicative model to estimate catch rate and effort from commercial data. *Can. J. Fish. Aquat. Sci., 37* : 2272-2275.

Gulland, J.A. (1956). On the fishing effort in English demersal fisheries. *Fishery Invest., Lond., Ser. 2, 20 (5)* : 1-41.

Jobson, J.D. (1991). *Applied Multivariate Data Analysis.* New York: Springer-Verlag.

Johnson, R.A. and Wichern, D.W. (1992). *Applied Multivariate Statistical Analysis.* (3rd ed.). New Jersey: Prentice-Hall.

Jorgensen, B. (1987). Exponential dispersion models (with discussion). . *J. Roy. Statist. Soc. Ser., B 49*: 127-162.

Kimura, D.K., (1981). Standardized measures of relative abundance based on modelling log(c.p.u.e.), and their application to Pacific ocean perch (Sebastes alutus). *J.Con. int. Explor. Mer, 39* : 211-218.

Kleinbaum, D.G., Kupper, L.L. and Muller, K. E. (1988). *Applied Regression Analysis and Other Multivariate Methods.* (2nd ed.). Boston: PWS-KENT Publishing Company.

Large, P.A. (1992). Use of a multiplicative model to estimate relative abundance from commercial CPUE data. *ICES J. mar. Sci., 49*: 253-261.

Lindsey, James K. (1997). *Applying Generalized Linear Models*. New York: Springer-Verlag.

McCullagh, P. and Nelder, J.A. (1983). *Generalized Linear Models*. London: Chapman and Hall.

Myers, R.H. (1990). *Classical and Modern Regression with Applications*. (2nd ed.). California: PWS-KENT Publishing Company.

Nelder, J.A. and Wedderburn, R.W.M. (1972). Generalized linear models. *J. Roy. Statist. Soc., Ser. A 135* : 370-384.

Neter, J., Wasserman, W. and Kutner, M.A. (1989). *Applied Linear Regression Models*. (2nd ed.). Boston: Richard D. Irwin.

Parrish, B.B. and Keir, R.S. (1959). The measurement of fishing power and its relation to the characteristics of vessels. *Annual Proceedings, Vol. 9, International Commission for the Northwest Atlantic Fisheries*, pp. 106-112.

Parsons, L.S., Pinhorn, A.T. and Parsons, D.G. (1976). An evaluation of the Northern Newfoundland-Labrador and Flemish Cap Redfish fisheries. *ICNAF Res. Bull., 12* : 37-48.

Phillips, B.F. and Brown, R.S. (1989). The West Australian rock lobster fishery: research for management. In: Caddy, J.F. (Ed.) *Marine Invertebrate Fisheries: Their Assessment and Management*. New York: Wiley.

Pierce, D.A. and Schafer, D.W. (1986). Residuals in Gerneralized Linear Models. *J. Amer. Statist. Assoc., 81* : 977-986.

Pope, J.A. (Ed.) (1975). Measurement of fishing effort. (Meeting held at Charlottelund Slot, Charlottelund, 25 and 26 September, 1975.) *Rapp. P. -v. Reun. Cons. int. Explor. Mer, 168:* 102pp.

Robins, C.M., Wang, Y. and Die, D. (1996). The impact of new technology on fishing power in the Northern Prawn Fishery, Australia. CSIRO Division of Fisheries, Hobart.

Robson, D.S. (1966). Estimation of the relative fishing power of individual ships. *Res.. Bull. Int. Commn. N.W. Atl. Fish., 3*: 5-14.

SAS Institute Inc. (1993). *SAS Technical Report P-243*. Cary: SAS Institute Inc.

Stocker, M. and Fournier, D. (1984). Estimation of relative fishing power and allocation of effective fishing effort, with catch forecasts, in a multi-species fishery. *Bull. 42, Int. North Pac. Fish. Comm.*

Tabachnick, B.G. and Fidell, L.D. (1989). *Using Multivariate Statistics.* (2nd ed.). New York: Harper Collins.

Walters, C., Hall, N., Brown, R.S. and Chubb, C.F. (1993). A spatial model for the population dynamics and exploitation of the Western Australian rock lobster, Panulirus cygnus. *Can. J. Fish. Aquat. Sci., 5O* :1650-1662.