MODEL SELECTION

by

ANNELIZE HILDEBRAND

submitted in part fulfilment of the
requirements for the degree of

MASTER OF SCIENCE

in the subject

STATISTICS

at the

UNIVERSITY OF SOUTH AFRICA

SUPERVISOR: PROF JL FRESEN

NOVEMBER 1995

SUMMARY

In developing an understanding of real-world problems, researchers develop mathematical and statistical models. Various model selection methods exist which can be used to obtain a mathematical model that best describes the real-world situation in some or other sense. These methods aim to assess the merits of competing models by concentrating on a particular ***criterion***. Each selection method is associated with its own criterion and is named accordingly. The better known ones include Akaike's Information Criterion, Mallows' C_p and cross-validation, to name a few. The value of the criterion is calculated for each model and the model corresponding to the minimum value of the criterion is then selected as the "best" model.

KEY TERMS:

Model selection; Discrepancy measures; Criteria; Akaike Information Criterion; Mallows' C_p; Cross-validation; R-square; Adjusted R-square; Mean Square Error.

---oOo---

ACKNOWLEDGEMENTS

My appreciation to Prof Fresen for his patient supervision and meaningful contributions.

My special thanks to my husband Marchand, to my mother and to my father for their continued moral support and encouragement.

My thanks to the three external examiners for their efforts and useful suggestions which were implemented in the final draft.

Above all, all the honour to the Lord for giving me the talent and for supporting me with my studies during the entire period.

Dedicated to my parents and my husband.

CONTENTS

CHAPTER 1

INTRODUCTION

1.1 EXAMPLES

1.1.1 The pork fat problem

Due to public health awareness a butcher wishes to indicate the fat percentage of the pork he sells. But measuring the percentage of fat in pork bellies directly is an expensive procedure. He thus requires a less expensive method of determining the fat content of a pork carcass.

After careful consideration, he realised that there are other, more easily measured properties of the pork carcass which could perhaps be used to predict the pork fat content. For example, an average of three measures of back fat thickness, live weight (*kg*) of the carcass, weight (*kg*) of the slaughtered carcass, the average of three determinations of the depth of the belly, to name a few. He identified nine possible ***predictors*** for pork fat content (see Table 1 of the appendix).

Should he use all nine predictors to predict pork fat content? It would be less expensive to use fewer than nine predictors! How will he use these predictors to forecast? He realises that he at least needs a mathematical equation to do so. This equation will contain some or all of the nine predictors he identified. It is clear that if fewer than nine predictors are used in the mathematical ***model***, there are numerous combinations of the predictors that could be included in the equation! How

will he then choose between competing ***models***?

1.1.2 The ethnicity problem

The Human Sciences Research Council (HSRC) undertakes research on issues that concern the public directly. One such project had as aim to investigate the social, economic and political, as well as psychological factors that influence a person's identification with her/his own ethnic group during the period of transition to a new political dispensation in South Africa. The initial step was a comprehensive literature study to determine what prior research had been done on the subject and what factors had been considered previously. A questionnaire containing questions from the literature as well as questions deduced from the researchers' own experience and initiative was then compiled. (Part of the questionnaire is included in the appendix.) This questionnaire was sent out to people from different racial groups over the period January/February 1994.

The questionnaire consisted of more than 100 items. The items were divided into sections with each section investigating a new variable in ethnic identification, for example section D08 investigates a person's ***social*** identity. A factor analysis and principal component analysis were performed on each section of the data to determine the dominant items per section influencing ethnic identification. The sum total of these dominant items then formed a variable corresponding to a section. From these results twenty variables were identified. These are shown in Table 2 of the appendix.

The objective of the HSRC project was to find a model that would best ***describe*** how a person identifies with her/his own ethnic group considering social, economic, political and psychological

factors. The objective is thus not to use these various factors to ***predict*** or ***forecast*** ethnic identification, but to ***explain*** and point out factors influencing a person's identification with his/her own ethnic group. It is thus not necessary to try and have as few factors as possible in the final model due to financial constraints or other reasons. The objective is to determine the relative influence of the factors.

These two examples illustrate some problems researchers encounter when trying to find models that describe the phenomena they are investigating. The butcher hopes to use his model to ***forecast*** and the HSRC researcher aims to ***explain*** and ***describe***. But what is a model exactly? How is a model constructed? And how do we select between competing models?

1.2 MODEL SELECTION

Although models can be physical objects, such as the models of molecules which are common in chemistry, in our context we are referring to mathematical and statistical models.

The use of mathematical/statistical models in solving real-world problems has become widespread in recent times. This is partly due to the increasing computational power of digital computers and the wide variety of statistical packages available. It is also due to our growing understanding and experience in applying these models.

The steps involved in using statistics and mathematics to solve real-world problems are shown in Figure 1.

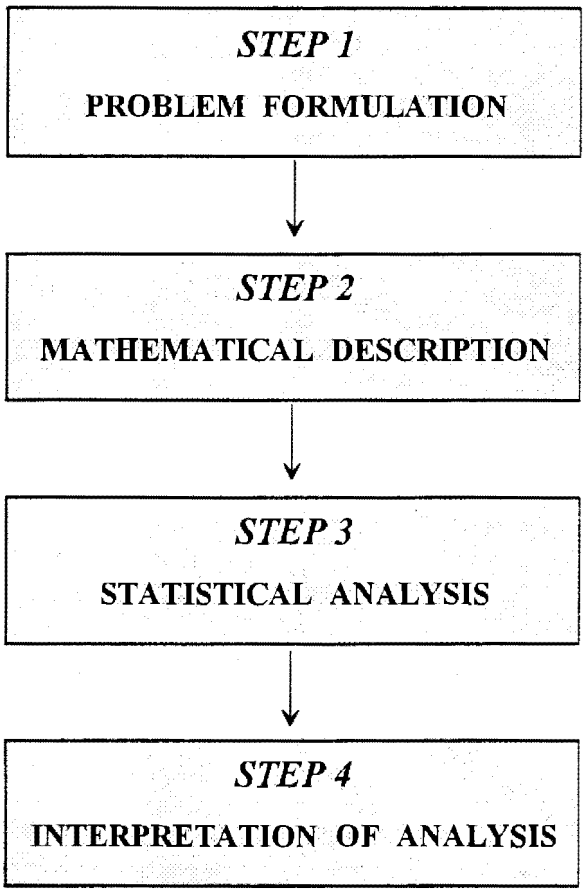

FIGURE 1: Problem solving using mathematics and statistics.

The real-world problem can be described as a collection of interacting objects. These objects as they appear in the real-world are represented in terms of abstract symbols, called parameters or variables. Parameters are attributes intrinsic to an object. Variables are attributes needed to describe interaction between objects. In this abstract form the original problem is divorced from the real-world and can be treated in mathematical terms only.

Once this is completed, model selection methods can be used to obtain a mathematical model that best describes the real-world situation. These methods aim to assess the merits of competing

models by concentrating on a particular ***criterion***. Each model selection method is associated with its own criterion and is named accordingly. Various selection methods exist and the better known ones include Akaike's Information Criterion (AIC), Mallows' C_p and Cross-validation, to name a few.

These methods are based on the following principle: a measure of the difference between the proposed model and the true situation, called a ***criterion*** (see Chapter 2 for details), is defined. The value of the criterion is calculated for each possible subset of the predictor variables. The subset corresponding to the minimum value of the criterion is then selected as the "best" subset. The coefficient of each predictor variable in the selected subset is then estimated. These coefficients are called the parameters of the model. The mathematical equation so obtained, together with possible assumptions and conditions, constitutes the final model.

If more than one selection method is used, it might happen that a variety of models are selected, each being the best model according to the different criteria. The model selection may then depend on other, perhaps non statistical criteria derived from real-world considerations. This is where experience and intuition play an important role.

It is thus clear that model selection is an art as well as a science. The scientific aspect deals with the statistical methods needed to execute the various steps in the model selection process. Because no two problems are the same in the real world, features such as creativity and intuition also play an important role.

1.3 LITERATURE REVIEW AND AIM OF THESIS

In doing research for the thesis, it became obvious that numerous model selection methods exist. But two specific methods seemed to form the basis of the development of various other methods, namely AIC and Mallows' C_p. Akaike (1969) was perhaps (Bozdogan, 1987) the first person who laid the foundation of the modern field of statistical data modelling. In Chapter 4 it will be shown that Mallows' C_p was derived from work done by Akaike. Therefore it seemed logical to investigate these two selection methods in detail (Chapters 3 and 4). Except for Akaike's own paper on the AIC (1969) and Mallows' paper on C_p (1973), various other statisticians published works on these topics. Bozdogan (1987) reviewed the general theory and analytical extensions of the AIC. The book by Sakamoto et al.(1986) gives a comprehensive discussion on the AIC. Thompson (1978a,b) discussed three model selection criteria in the multiple regression setup including Mallows' C_p and applied these to various examples. Stone (1977) discussed the asymptotic equivalence of choice of model by cross-validation and the AIC.

Other selection methods which were derived from or are equivalent to the AIC and C_p will be discussed in Chapter 5 and the relationship with the aforementioned will be shown. Research for this chapter was taken from Shao (1993) and Stone (1974). Shao considered linear model selection using cross-validation. Stone considered the choice and assessment of statistical predictors using cross-validation as a criterion. Chapter 2 gives a brief overview of the abstract theory behind model selection, including the important definitions of ***discrepancies*** and ***criteria***. Research for this chapter was taken from Linhart and Zucchini (1986) although their notation was adapted to be consistent with the notation of the thesis. Chapter 6 finally considers examples of model selection using the methods discussed in the previous

chapters. One example is taken from Thompson (1978b).

The aim of this thesis is thus twofold. Firstly to look at the theory of and relationships between model selection procedures and secondly to apply these methods to some examples.

1.4 NOTATION

The following notation will be used throughout:

x	:	The predictor variable
y	:	The response variable
$\mathcal{G}$	:	The family of all probability distributions
F	:	The true model
$\mathcal{F}$	:	The true family of models
Λ	:	Parameter space for true family of models
$\mathcal{E}$	:	The approximating family of models
Ω	:	Parameter space for approximating family of models
G	:	A member of the approximating family of models
β	:	Vector of parameters
G_{β_0}	:	The nearest approximating model
$G_{\hat{\beta}}$	:	The fitted model
$\Delta(\cdot,\cdot)$	:	A discrepancy
$\Delta_n(\beta, F)$	:	An empirical discrepancy
$\mathbb{R}$	:	Set of real numbers

---oOo---

CHAPTER 2

DISCREPANCY MEASURES

2.1 INTRODUCTION

In this chapter we give a review of the statistical theory associated with discrepancy measures which are used to measure the similarities between two probability distributions. Research for this chapter was strongly influenced by the book by Linhart and Zucchini (1986) although their notation was adapted to be consistent with the notation of the thesis.

2.2 THE TRUE MODEL AND THE TRUE FAMILY

When we do research on a set of data, the statistical perspective is that the data have been generated by a probability distribution, F say, which is usually unknown. The objective of the research is then to learn more about the nature and form of the model F, called the ***true model***. It often happens that it is only necessary to estimate some aspects of F, such as the mean or standard deviation, or perhaps the parameters which define F.

In the most general setting F is viewed as a member of the family of all probability distributions, denoted by $\mathcal{L}$. By the ***true model*** we mean that probability distribution, say F, $F \in \mathcal{L}$, which generates the data.

Information about the true model F is obtained from knowledge about the subject matter under investigation. However, it is only in exceptional cases that sufficient information is

available to fully specify the true model. It is often only possible to describe the group of models to which the true model belongs. This group of models will be called the ***true family of models*** and will be denoted by $\mathcal{F}$. In the parametric setting the true family is indexed by a parameter, say λ, in which case the true family is

$$\mathcal{F} = \{F_\lambda : \lambda \in \Lambda\}.$$

Here Λ is called the parameter space. Fitting a model then means finding an estimate of the parameters in the model, i.e. using the data to find an estimate of λ, denoted by $\hat{\lambda}$. In this case

$$\hat{F} = F_{\hat{\lambda}}$$

where $\hat{F}$ denotes our estimate of F.

The parameters λ are estimated from the data and the accuracy with which this can be done depends on the amount of data available, relative to the number of parameters to be estimated as well as the quality and precision of the data, and the method of estimation.

2.3 APPROXIMATING FAMILIES

In practice one may not have sufficient information about the random mechanism which generates the data to specify the true family of models. To proceed at all one is obliged to compromise by using a simple family of models which, in one or other sense, approximates the random generating mechanism. This family is called the ***approximating family***, which we will denote by

$$\mathscr{E} = \{G_\beta : \beta \in \Omega\}$$

In this instance, fitting a model means using the data to find an estimate, say $\hat{\beta}$, of β. The fitted model is

$$\hat{G} = G_{\hat{\beta}}$$

where $\hat{G}$ denotes our estimate of G.

2.4 DISCREPANCIES

The question which arises is how one should go about fitting a particular model from either the true family of models or, if this is unknown, the family of approximating models.

One way is to specify a function which increases in value as G (a member of the approximating family) and F (the true model) become less similar. This function is called a discrepancy and will be denoted by Δ. The manner in which we specify that F and G should be similar will influence our choice of discrepancy. The member of the approximating family $\mathscr{E}$ of models which minimizes an estimate of the expected discrepancy, is then selected.

2.4.1 Some definitions

Let $F, G \in \mathfrak{L}$. A ***discrepancy*** $\Delta : \mathfrak{L} \times \mathfrak{L} \to \mathbb{R}$ is a functional mapping $\mathfrak{L} \times \mathfrak{L}$ onto $\mathbb{R}$, the set of real numbers, with the following properties $\forall\, F, G \in \mathfrak{L}$:

DISCREPANCY	
1.	$\Delta(G,F) \geq 0$
2.	$\Delta(G,F) \geq \Delta(F,F)$

Let $\mathscr{E} = \{G_\beta ; \beta \in \Omega\}$ be the approximating family. Let G_{β_0} denote that member of $\mathscr{E}$ which has the smallest discrepancy between the true model F and any member of $\mathscr{E}$. Thus

$$G_{\beta_0} = arg\ \min\{\Delta(G_\beta, F) : \beta \in \Omega\}$$

and is called the nearest approximating model. The discrepancy between the nearest approximating model G_{β_0} and the true model F is called the ***discrepancy due to approximation*** and is denoted by

$$\Delta_{approx} = \Delta(G_{\beta_0}, F) .$$

The discrepancy due to approximation Δ_{approx} does not depend on the data, the sample size or method of estimation of parameters. However, Δ_{approx} does depend on the true model F, which is usually unknown, and the nearest approximating model G_{β_0} which is also unknown and thus Δ_{approx} cannot be calculated. In this respect, Δ_{approx} is a theoretical construct.

By contrast, the ***discrepancy due to estimation*** is the discrepancy between the fitted model $G_{\hat{\beta}}$ and the nearest approximating model G_{β_0}, i.e.

$$\Delta_{est} = \Delta(G_{\hat{\beta}}, G_{\beta_0}) .$$

The discrepancy due to estimation is a function of the data and the method of estimation and is thus a random variable. Because the nearest approximating model G_{β_0} is unknown, the distribution of Δ_{est} is not specified.

The ***overall discrepancy*** is the discrepancy between the fitted model $G_{\hat{\beta}}$ and the true model F, i.e.

$$\Delta_{overall} = \Delta(G_{\hat{\beta}}, F).$$

Like Δ_{est}, $\Delta_{overall}$ is a function of the data and the method of estimation, and is thus a random variable. Again it is a theoretical construct because the true model F is unknown. Hence the distribution of $\Delta_{overall}$ is not specified.

The three discrepancies Δ_{approx}, Δ_{est} and $\Delta_{overall}$ can be geometrically depicted as follows:

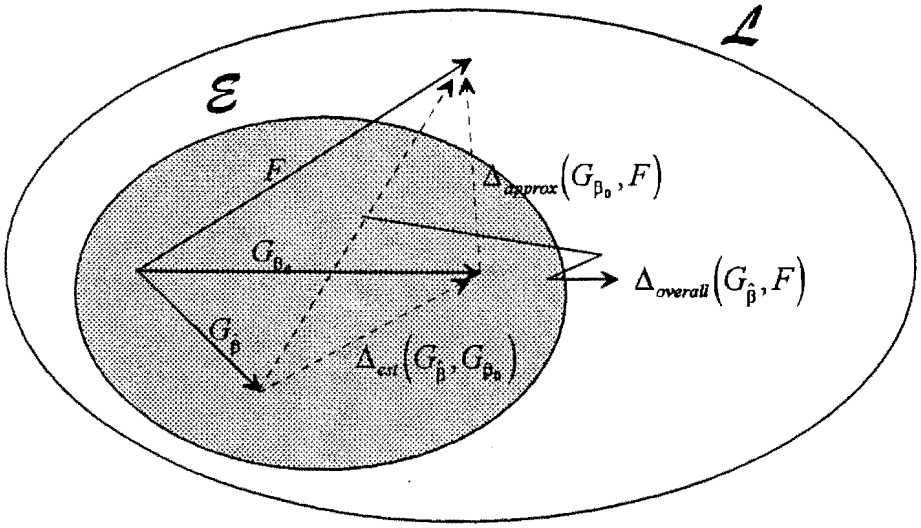

FIGURE 2: Geometric depiction of discrepancies due to approximation, estimation and the overall discrepancy.

2.4.2 Examples of discrepancies

Let F denote the true model and G_β an approximating model, and let f and g_β be the corresponding density or mass functions. The following are well known examples of discrepancy functions.

a) Kullback-Leibler discrepancy

$$\Delta_{K\text{-}L} = -E_F \ln g(\beta, x) = -\int \ln g(\beta, x)\, dF(x)$$

b) Kolmogorov discrepancy

$$\Delta_K = \sup_x |F(x) - G_\beta(x)|$$

c) Cramér-von Mises discrepancy

$$\Delta_{C\text{-}M} = E_{G_\beta}(F(x) - G_\beta(x))^2 = \int (F(x) - G_\beta(x))^2 dG_\beta(x)$$

d) Pearson chi-squared discrepancy

$$\Delta_P = \int \frac{(f(x) - g_\beta(x))^2}{g_\beta(x)} dx, \qquad g_\beta(x) \neq 0$$

The discrete version is

$$\Delta_P = \sum_x \frac{(f(x) - g_\beta(x))^2}{g_\beta(x)}, \qquad g_\beta(x) \neq 0$$

e) Neyman chi-squared discrepancy

$$\Delta_N = \int \frac{(f(x) - g_\beta(x))^2}{f(x)} dx, \qquad f(x) \neq 0$$

or

$$\Delta_N = \sum_x \frac{(f(x) - g_\beta(x))^2}{f(x)}, \qquad f(x) \neq 0$$

for the discrete case.

f) Gauss discrepancy

$$\Delta_G = \int (f(x) - g_\beta(x))^2 dx$$

or

$$\Delta_G = \sum_x (f(x) - g_\beta(x))^2$$

for discrete or grouped data.

2.4.3 Some general remarks on discrepancies

When two distributions F and G are "close" in some or other sense, the discrepancy between them $\Delta(F,G)$ is small. When they are not "close", the discrepancy $\Delta(F,G)$ is large. Thus, the discrepancy $\Delta(F,G)$ reminds us of a distance function or a metric. To determine whether a discrepancy is a metric or not, let us first list the properties of a metric $\forall F,G,H \in \mathfrak{L}$:

METRIC
1. $d(G,F) \geq 0$
2. $d(G,F) = d(F,G)$
3. $d(G,F) = 0 \Leftrightarrow G = F$
4. $d(G,F) \leq d(G,H) + d(H,F)$

Section 2.4.3: Some general remarks on discrepancies

Comparing these to the properties of a discrepancy as listed on page 11, it is clear that a metric is a discrepancy. But is a discrepancy a metric? Consider the examples of the previous section. The non-negativity property of a metric is also true for a discrepancy, (as stated in 2.4.1). The second property that the function is symmetric in its arguments, however, is clearly not true for the Pearson chi-squared or the Neyman chi-squared discrepancies, for:

$$\int \frac{(f(x) - g_\beta(x))^2}{g_\beta(x)} dx \neq \int \frac{(g_\beta(x) - f(x))^2}{f(x)} dx, \quad f(x) \neq g_\beta(x).$$

The third property of a metric holds for all the examples of a discrepancy in 2.4.2 except for the Kullback-Leibler discrepancy. For example, let us suppose that $F = G$ and $f(x) = e^{-x}$, $x > 0$. Then

$$\begin{aligned}
\Delta_{K-L} &= -E_F \ln g(\beta, x) \\
&= -\int \ln g(\beta, x)\, dF(x) \\
&= -\int \ln g(\beta, x) \,.\, f(x)\, dx \\
&= -\int \ln f(x) \,.\, f(x)\, dx \\
&= \int_0^\infty x e^{-x} dx \\
&= -x e^{-x} \Big]_0^\infty - \int_0^\infty -e^{-x} \\
&= -e^{-x} \Big]_0^\infty \\
&= 1 \neq 0
\end{aligned}$$

It is thus not true for Δ_{K-L} that $F = G \Rightarrow \Delta = 0$. However, Linhart and Zucchini points out that, and I quote:

Section 2.4.3: Some general remarks on discrepancies

"A discrepancy is simply a functional that one wishes to minimise - it is not important whether it achieves a minimum at zero..."

From the counter-examples above, it is clear that Δ is not a metric and that is why the term "discrepancy" is used to remind us that it is not a distance function.

Another aspect of discrepancies to be investigated is whether Δ is defined for ***all*** F and G. For example, is Δ defined for say a discrete distribution F and a continuous distribution G? On investigating the examples of discrepancies of the previous section, it is obvious that this is true for all the discrepancies except perhaps for the Kullback-Leibler. Consider an example:

Let $g(x) = e^{-x} \quad x > 0$ and $f(x) = \begin{cases} 1-\theta & x = 0 \\ \theta & x = 1 \end{cases} \quad \theta \in [0;1]$.

Using the Dirac delta function (Stremler, 1982, p59), $f(x)$ can be written as $f(x) = (1-\theta)\,\delta(x) + \theta\,\delta(x-1)$. For the Kullback-Leibler discrepancy it follows that:

$$\begin{aligned} \Delta_{K\text{-}L} &= -\int_0^\infty \ell n\, g(x).f(x)\, dx \\ &= -\int_0^\infty \ell n\, g(x).[(1-\theta)\,\delta(x) + \theta\,\delta(x-1)]\, dx \\ &= -[(1-\theta)\,\ell n\, g(0) + \theta\,\ell n\, g(1)] \\ &= -[0-\theta] \\ &= \theta \geq 0 \end{aligned}$$

It is thus mathematically possible to compare any two distributions using the Kullback-Leibler or any of the other discrepancies and we can therefore conclude that a discrepancy is defined for ***all*** $F, G \in \mathcal{G}$.

Section 2.4.3: Some general remarks on discrepancies

Because the true model F is usually unknown, the distribution of the discrepancies Δ_{est} and $\Delta_{overall}$ are unknown. One could estimate the distribution of the true model F and use this to find the distribution of the discrepancies but it is usually simpler to estimate some of the moments of the discrepancies, the simplest being the first moment, i.e. the expected value of Δ. Intuitively one wants to choose an estimation method which leads to ***small*** expected discrepancies, for the smaller the expected value of $\Delta(F,G)$, the "closer" F and G are to each other. An estimator which minimises the estimated expected discrepancy (the estimated first moment of a discrepancy), is called a minimum discrepancy estimator. Let $\Delta_n(G_\beta, F)$ be a consistent estimator of $\Delta(G_\beta, F)$. If

$$G_{\hat{\beta}} = arg\ \min\{\Delta_n(G_\beta, F) : \beta \in \Omega\}$$

exists almost surely, it is called a ***minimum discrepancy estimator*** of G_{β_0}.

A suitable empirical discrepancy for $\Delta(G_\beta, F)$ is $\Delta(G_\beta, \hat{F}_n)$, where F_n is the empirical distribution function, that is, the defining function of the distribution with mass $\frac{1}{n}$ at the observed points x_i, $i = 1,2,.....,n$, i.e.

$$\Delta_n(G_\beta, F) = \Delta(G_\beta, \hat{F}_n)$$

2.5 CRITERIA

The distribution of $\Delta_{overall} = \Delta(G_{\hat{\beta}}, F)$ under the true model F determines the quality of a model selection procedure, i.e. one would prefer model selection procedures which result in low overall discrepancies. The lower the overall discrepancy, the

"closer" the fitted model $G_{\hat{\beta}}$ and the true model F. Because $\Delta_{overall} = \Delta(G_{\hat{\beta}}, F)$ depends on the true model F which is unknown, its complete distribution is also unknown and must thus be estimated. It is, however, not always possible and/or easy to estimate the complete distribution of the overall discrepancy, but rather some characteristic of it such as the moments of the distribution. The first moment, the expectation, of the distribution is usually used for this purpose. And so it seems reasonable to judge a model selection procedure by the expected overall discrepancy, $E_F \Delta_{overall}(G_{\hat{\beta}}, F)$, simply called the expected discrepancy. This depends on the true model which is of course unknown, but it can be estimated. An estimator of the expected discrepancy $\widehat{E_F \Delta_{overall}(G_{\hat{\beta}}, F)} = E_{\hat{F}} \Delta_{overall}(G_{\hat{\beta}}, \hat{F})$ is called a ***criterion***. To select between competing approximating models, we choose the one which leads to the smallest value of the criterion.

---oOo---

CHAPTER 3
AKAIKE INFORMATION CRITERION (AIC)

3.1 INTRODUCTION

H. Akaike's interest in modelling came as an instinctive tendency because of his engineering orientation. His first opportunity to apply modelling occurred when a person from the ministry of Agriculture visited a colleague with whom he shared an office. He overheard that they were having trouble applying ordinary control chart techniques to the process of winding filaments of silk from bunches of cocoons into a thread and onto a reel. At that stage he had developed a model for the analysis of traffic flow and he suggested that this could be applied to this problem. The idea was accepted and the model was successfully implemented. This was essentially Akaike's first modelling experience.

In 1971 H. Akaike introduced an information criterion, called the AIC (Akaike Information Criterion), at the annual meeting of the Japan Statistical Society. His interest in factor analysis initially led to the discovery of the AIC, (Findley et al. 1995):

> *"One day I recognized that the factor analysis people were maximizing the likelihood, attempting to get a good distributional model for the purpose of prediction. However, in this case the prediction is not a value, but is*

the fitted distribution itself, which is applied to understand the next similar observation. For the next similar problem you use the present model, and the accuracy criterion for this prediction is given by the log-likelihood function. Then once I got this far, it was just one step to recognize that the expected log-likelihood is related to the Kullback information".

He developed the AIC for the identification of an optimal model in data analysis from a set of competing models. The AIC is a simple procedure which was designed to be an unbiased estimator of the expected Kullback-Leibler discrepancy.

Although the AIC was introduced in 1971, Akaike's paper on this was only published in 1973 and until recently was very inaccessible. Because his original paper is not readily available, our development follows Sakamoto et al. (1986). They introduce the ***mean expected log likelihood*** as a measure for the goodness of fit of a proposed model. The mean expected log likelihood and the necessary assumptions will be given in §3.2. It will be shown that the larger the mean expected log likelihood, the better the fit of the model. Since it is not always possible to find the exact value of the mean expected log likelihood, an estimator will be found in §3.3, namely

(*maximum log likelihood of a model*) - (*number of free parameters*)

which is an asymptotically unbiased estimator of the mean expected log likelihood. The criterion proposed in §3.3 for model selection is

$$AIC = -2\times(MLL \text{ of the model}) + 2\times(\text{number of free paramaters})$$

where *MLL* denotes the maximum log likelihood.

It will be shown that a model which minimizes the AIC is considered as the most appropriate model and that AIC prefers models with a small number of parameters.

3.2 THE MEAN EXPECTED LOG LIKELIHOOD - DEFINITIONS AND ASSUMPTIONS

The following definitions and assumptions will be used for the derivation of the AIC in the next paragraph.

Suppose that $X_1, X_2, \cdots, X_n$ are independently identically distributed with density function f. Let $X = (X_1, X_2, \cdots, X_n)$ and let $x = (x_1, x_2, \cdots, x_n)$ denote a realization of X. Denote the vector of true parameters by $\beta^* = (\beta_1^*, \beta_2^*, \cdots, \beta_P^*)$ and assume that all P parameters are unconstrained (free).

Define a parameter space Ω_k by restricting the original parameter space Ω as

$$\Omega_k = \{\beta \in \Omega \mid \beta_{k+1} = \beta_{k+2} = \cdots = \beta_P = 0\} \qquad 1 \leq k \leq P$$

For each Ω_k the number of restricted parameters is $P - k$ and the number of free parameters is k. Since $\Omega_1 \subset \Omega_2 \subset \cdots \subset \Omega_p$ there exists a Ω_k that satisfies $\beta^* \in \Omega_k$. For this parameter space Ω_k the number of free parameters is denoted by k^* and k^* is called the true number of free parameters.

Given the data x, the maximum likelihood estimate $\hat{\beta}_k$ of the k parameters is obtained by maximizing the log likelihood

Section 3.2: The mean expected log likelihood - definitions and assumptions

$$\ell(\beta|x) = \sum_{i=1}^{n} \log f(x_i, \beta)$$

over $\beta \in \Omega_k$. The maximum log likelihood is given by

$$\ell(\hat{\beta}_k|x) = \sup_{\beta \in \Omega_k} \ell(\beta|x)$$

The expected log likelihood of the distribution is defined by

$$E_z\{\log f(Z,\beta)\} = \int f(z,\beta^*) \log f(z,\beta)\, dz$$

where Z is a random variable with the same distribution as X_i but independent of X. Corresponding to the log likelihood $\ell(\beta|x)$, $\ell^*(\beta|x)$ is defined as n times the expected log likelihood, namely

$$\ell^*(\beta|x) = n\, E_z\{\log f(Z,\beta)\}$$

It should be noted here that $\ell^*(\beta|x)$ is equal to $-n\Delta_{K\text{-}L}$, where $\Delta_{K\text{-}L}$ is the Kullback-Leibler discrepancy of Chapter 2 (Hurvich and Tsai, 1989, p299).

The goodness of fit of the maximum likelihood model can be evaluated by $\ell^*(\hat{\beta}_k|x)$: The larger the value of $\ell^*(\hat{\beta}|x)$ the better the approximation of the distribution of $f(x,\beta_k)$ to the true one $f(x,\beta^*)$. However, since this quantity is dependent on the realization x of the random variable X, the model will be evaluated by its mean value called the mean expected log likelihood, denoted by $\ell^*_n(k|x)$. Thus

$$\ell^*_n(k|x,\beta^*) \equiv E\{\ell^*(\hat{\beta}_k|x)\} = \int \ell^*(\hat{\beta}_k|x) \prod_{i=1}^{n} g(x_i)\, dx_i$$

The mean expected log likelihood no longer depends on a particular realization, and depends only on the true model F, the

assumed model and the sample size n. The model with larger mean expected log likelihood is considered to be the better one.

3.3 DERIVATION OF THE AIC

Using Young's form of Taylor's theorem (Serfling, 1980, p45) to expand the expected log likelihood $\ell^*(\beta|x) = n\, E_Z\{\log f(Z,\beta)\}$ around the true parameter β^* yields

$$\ell^*(\beta|x) - \ell^*(\beta^*|x) = n(\beta - \beta^*)^T E_Z\left[\frac{\partial \log f(Z,\beta)}{\partial \beta}\right]_{\beta^*} + \frac{1}{2} n(\beta - \beta^*)^T E_Z\left[\frac{\partial^2 \log f(Z,\beta)}{\partial \beta^2}\right]_{\beta^*} (\beta - \beta^*) + R_n$$

where $R_n = o(\|\beta - \beta^*\|^2)$.

The first term on the right hand side vanishes because

$$E\left[\frac{\partial \log f(Z,\beta)}{\partial \beta}\right]_{\beta^*} = 0.$$

Thus

$$\ell^*(\beta|x) - \ell^*(\beta^*|x) = \frac{1}{2}\sqrt{n}(\beta - \beta^*)^T E_Z\left[\frac{\partial^2 \log f(Z,\beta)}{\partial \beta^2}\right]_{\beta^*} \sqrt{n}(\beta - \beta^*) + R_n$$

Define J_* by

$$J_* = -E_Z\left[\frac{\partial^2}{\partial \beta^2} \log f(Z,\beta)\right]_{\beta^*}$$

It then follows that

$$\ell^*(\beta|x) - \ell^*(\beta^*|x) = -\frac{1}{2}\sqrt{n}(\beta - \beta^*)^T J_*\sqrt{n}(\beta - \beta^*) + R_n$$

The maximum likelihood estimator $\hat{\beta}_P$ is asymptotically normal with

$$\sqrt{n}(\hat{\beta}_P - \beta^*) \rightarrow N(0, J_*^{-1})$$

(Sakamoto et al. 1986, p55). Thus, $n(\hat{\beta}_P - \beta^*)^T J_*(\hat{\beta}_P - \beta^*)$ is asymptotically distributed as a chi-square distribution with P degrees of freedom (Sakamoto et al. 1986, p26). Consequently, since the mean of this chi-square distribution is P, it is true for large n that

$$E_X\{\ell^*(\hat{\beta}_P|x) - \ell^*(\beta^*|x)\} = E_X\{-\frac{1}{2}\sqrt{n}(\hat{\beta}_P - \beta^*)^T J_*\sqrt{n}(\hat{\beta}_P - \beta^*) + 0\}$$

$$\ell_n^*(P) - \ell^*(\beta^*|x) = -\frac{P}{2}$$

OR $$-2\{\ell_n^*(P) - \ell^*(\beta^*|x)\} = P$$

Let us now consider the Taylor expansion of the log likelihood

$$\ell(\beta|x) = \sum_{i=1}^{n} \log f(x_i;\beta)$$

The expansion around the maximum likelihood estimate $\hat{\beta}_P$ yields

$$\ell(\beta|x) - \ell(\hat{\beta}_P|x) = (\beta - \hat{\beta}_P)^T\left[\frac{\partial\ell}{\partial\beta}\right]_{\hat{\beta}_P} + \frac{1}{2}(\beta - \hat{\beta}_P)^T\left[\frac{\partial^2\ell}{\partial\beta^2}\right]_{\hat{\beta}_P}(\beta - \hat{\beta}_P) + R_n$$

where $R_n = o(\|\beta - \hat{\beta}_P\|^2)$.

Here the remainder tends to zero as n tends to infinity and the first term on the right hand side vanishes because

$$\left[\frac{\partial \ell}{\partial \beta}\right]_{\hat{\beta}_P} = 0.$$

Thus

$$\ell(\beta|x) - \ell(\hat{\beta}_P|x) = \frac{1}{2}(\beta - \hat{\beta}_P)^T\left[\frac{\partial^2 \ell}{\partial \beta^2}\right]_{\hat{\beta}_P}(\beta - \hat{\beta}_P) + R_n$$

$$= \frac{1}{2}\sqrt{n}(\beta - \hat{\beta}_P)^T\frac{1}{n}\left[\frac{\partial^2 \ell}{\partial \beta^2}\right]_{\hat{\beta}_P}\sqrt{n}(\beta - \hat{\beta}_P) + R_n$$

By the strong law of large numbers, as $n \to \infty$,

$$\frac{1}{n}\left[\frac{\partial^2 \ell}{\partial \beta^2}\right]_{\beta^*} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\partial^2}{\partial \beta^2}\log f(x_i;\beta)\right]_{\beta^*} \overset{a.s.}{\to} E_Z\left[\frac{\partial^2}{\partial \beta^2}\log f(Z;\beta)\right]_{\beta^*} = -J_*$$

where J_* is identical to the Fisher information matrix (Sakamoto et al. p64) and also equal to the second derivative of $\Delta_{K\text{-}L}$. Moreover, since $\hat{\beta}_P \overset{a.s.}{\to} \beta^*$ as $n \to \infty$, it follows that

$$\frac{1}{n}\left[\frac{\partial^2 \ell}{\partial \beta^2}\right]_{\hat{\beta}_P} \to -J_*.$$

Therefore, for large n

$$\ell(\beta|x) - \ell(\hat{\beta}_P|x) = -\frac{1}{2}\sqrt{n}(\beta - \hat{\beta}_P)^T J_* \sqrt{n}(\beta - \hat{\beta}_P) + R_n$$

Substituting β^* for β and taking expectations of both sides we have, similar to the way in which the expectation of $\ell(\hat{\beta}_P|x)$ was derived,

$$E_X\{\ell(\beta^*|x) - \ell(\hat{\beta}_P|x)\} = E_X\{-\frac{1}{2}\sqrt{n}(\beta^* - \hat{\beta}_P)^T J_*\sqrt{n}(\beta^* - \hat{\beta}_P) + R_n\}$$

$$E_X\{\ell(\beta^*|x)\} - E_X\{\ell(\hat{\beta}_P|x)\} = -\frac{P}{2} \qquad \ldots\ldots(A)$$

On the other hand, from the independence of the X_i's we get

$$\begin{aligned} \ell^*(\beta^*|x) &= nE_Z\{\log f(Z;\beta^*)\} \\ &= E_X\left\{\sum_{i=1}^{n} \log f(X_i;\beta^*)\right\} \qquad \ldots\ldots(B) \\ &= E_X\{\ell(\beta^*|x)\}. \end{aligned}$$

From the two equations (A) and (B) above, it follows that

$$\ell^*(\beta^*|x) - E_X\{\ell(\hat{\beta}_P|x)\} = -\frac{P}{2}$$

From this relation and $\ell^*_n(P) - \ell^*(\beta^*|x) = -\frac{P}{2}$, it follows that

$$\ell^*_n(P) - \ell^*(\beta^*|x) + \ell^*(\beta^*|x) - E_X\{\ell(\hat{\beta}_P|x)\} = -\frac{P}{2} - \frac{P}{2}$$

$$\ell^*(P) - E_X\{\ell(\hat{\beta}_P|x)\} = -P$$

and finally

$$\ell^*_n(P) = E_X\{\ell(\hat{\beta}_P|x)\} - P$$

This relation reveals that the maximum log likelihood is a biased estimator of the mean expected log likelihood and its bias is equal to the number of free parameters of the model, namely P. Thus, correcting for bias, the estimator

$$\ell(\hat{\beta}_P|x) - P$$

constitutes an unbiased estimator of the mean expected log likelihood. Define AIC as minus twice the above quantity, namely

$$AIC = -2\ell(\hat{\beta}_P|x) + 2P$$

$$= -2\,(MLL\ of\ the\ model) + 2\,(number\ of\ free\ parameters)$$

The reason why the AIC is multiplied by -2 is that in the literature the equation (A) is expressed in the form

$$-2E_X\{\ell(\beta^*|x) - \ell(\hat{\beta}_P|x)\} = P$$

(See Rao, 1965, p349).

Since

$$E_X\{-\frac{1}{2}AIC(P) - \ell^*(\hat{\beta}_P|x)\} = E_X\{\ell(\hat{\beta}_P|x) - P - \ell^*(\hat{\beta}_P|x)\}$$

$$= \ell^*(\hat{\beta}_P|x) - \ell^*(\hat{\beta}_P|x)$$

$$= 0$$

AIC can also be interpreted as an unbiased estimator of the -2 times expected log likelihood of the maximum likelihood model.

3.4 A FINAL REMARK

As mentioned shortly in the derivation of AIC, this criterion is based on the Kullback-Leibler discrepancy. AIC is an estimator of the expected value of $-n\Delta_{K-L}$, (Hurvich et al, 1989, p301) i.e:

$$AIC = \widehat{E_F\{-n\Delta_{K-L}\}} = E_{\hat{F}}\{-n\Delta_{K-L}\}$$

AIC is justifiably called a ***criterion*** since an estimator of an expected discrepancy was defined to be a criterion in Chapter 2.

---oOo---

CHAPTER 4

MALLOWS' C_p

4.1 INTRODUCTION

In this chapter we consider Mallows' C_p criterion. The criterion is defined to be

$$C_p = \frac{1}{\hat{\sigma}^2} RSS_p - n + 2p$$

where RSS_p is the residual sum of squares, i.e. the sum of squares of the deviations of the observed response variables from their estimated expected values, n is the sample size, p the number of parameters and $\hat{\sigma}^2$ an estimate of σ^2, the variance.

Mallows proposed this criterion based on what is called the final prediction error (FPE) criterion. The FPE criterion is an extension of a criterion proposed by H Akaike (1969) and is given in a general form as

$$FPE_\alpha(p) = RSS_p + \alpha k \frac{RSS_{(n-K)}}{(n-K)}$$

Almost equivalent to this is the general information criterion

$$C(\alpha, p) = \log RSS_p + \alpha k$$

proposed as an extension of the AIC (Atkinson, 1978). The AIC corresponds to $C(2,p)$ while the C_p criterion corresponds to the FPE with $\alpha = 2$ and $\frac{RSS_{(n-K)}}{(n-K)}$ used as $\hat{\sigma}^2$.

The criterion will be derived in §4.2 and it will also be shown

that using the C_p criterion for model selection, comes down to selecting that subset of the predictor variables that corresponds to the minimum C_p value and values which are very close to p, the number of parameters. Mallows also derived a graphical method for selecting a model, called the C_p plot, and this will be discussed in §4.3.

4.2 DERIVATION OF THE C_p-STATISTIC

Suppose the full approximating model is

$$y_i = \beta_0 + \sum_{j=1}^{k} x_{ij}\beta_j + \epsilon_i \qquad i = 1,2,\ldots n$$

with the residuals $\epsilon_1, \cdots, \epsilon_n$ independent random variables with mean zero and unknown common variance σ^2. We will assume that the $x_i's$ are ***non-stochastic*** and not random predictor variables since, although the C_p criterion is applicable to random variables, it performs much better when the $x_i's$ are non-stochastic.

Let k be the number of predictor variables. Consider a subset P of the set of indices $\{0,1,2,\cdots,k\}$. Let Q be the complementary subset of P. Suppose the number of elements in P and Q are p and q respectively, so that $p + q = k + 1$. Let X_P be the matrix obtained when the columns in X having subscripts in Q are removed. Denote by $\hat{\beta}_P$ the vector of estimates that is obtained when the coefficients with subscripts in P are taken into account. Then

$$\hat{\beta}_P = X_P^{-}y$$

where X_P^- is the Moore-Penrose generalized inverse of X_P, satisfying the following four conditions:

$$X_P X_P^- X_P = X_P$$

$$X_P^- X_P X_P^- = X_P^-$$

$$(X_P^- X_P)^T = X_P^- X_P$$

$$(X_P X_P^-)^T = X_P X_P^-$$

If X_P is of full rank, then $X_P^- = (X_P^T X_P)^{-1} X_P^T$. Thus

$$\hat{\beta}_P = X_P^- y$$

$$= (X_P^T X_P)^{-1} X_P^T y$$

Let RSS_P denote the corresponding residual sum of squares, i.e.

$$RSS_P = \sum (y_i - \hat{y}_{iP})^2$$

$$= \sum (y_i - x_{iP}^T \hat{\beta}_P)^2$$

It should be noted here that RSS_P is equivalent to the Gauss discrepancy $\Delta_G = \sum (f(x) - g_\beta(x))^2$ of Chapter 2.

For such an estimate $\hat{\beta}_P$, a measure of adequacy for prediction is the "scaled sum of squared errors"

$$J_P = \frac{1}{\sigma^2} RSS_P$$

$$= \frac{1}{\sigma^2} \sum_{i=1}^{n} (\hat{y}_{ip} - y_i)^2$$

$$= \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_{ip}^T \hat{\beta}_P - y_i)^2$$

$$= \frac{1}{\sigma^2} (X_P \hat{\beta}_P - Y)^T (X_P \hat{\beta}_P - Y).$$

Section 4.2: Derivation of the C_p-statistic

The expectation of J_p is

$$E(J_p) = E\left[\frac{1}{\sigma^2}\ (X_p\hat{\beta}_p - Y)^T(X_p\hat{\beta}_p - Y)\right]$$

$$= \frac{1}{\sigma^2}E\left[\hat{\beta}_p^T X_p^T X_p\hat{\beta}_p - Y^T X_p\hat{\beta}_p - \hat{\beta}_p^T X_p^T Y + Y^T Y\right]$$

$$= \frac{1}{\sigma^2}\left[E(\hat{\beta}_p^T X_p^T X_p\hat{\beta}_p) - E(Y^T X_p\hat{\beta}_p) - E(\hat{\beta}_p^T X_p^T Y) + E(Y^T Y)\right]$$

$$= \frac{1}{\sigma^2}\left[E(Y^T Y - Y^T X_p(X_p^T X_p)^{-1}X_p^T Y - Y^T X_p(X_p^T X_p)^{-1}X_p^T Y + Y^T X_p(X_p^T X_p)^{-1}X_p^T Y)\right]$$

$$= \frac{1}{\sigma^2}\left[E(Y^T Y) - E(Y^T X_p(X_p^T X_p)^{-1}X_p^T Y)\right]$$

$$= \frac{1}{\sigma^2}\left[n\sigma^2 + \beta^T X^T X\beta - E(Y^T AY)\right] \qquad \ldots\ldots (B)$$

where $A = X_p(X_p^T X_p)^{-1}X_p^T = AA$ and is thus idempotent. Using standard results from the theory of quadratic forms (see eg. Searle, 1971), it follows that

$$E(Y^T AY) = tr(A\sigma^2) + (X\beta)^T A(X\beta)$$

$$= rank(A)\,\sigma^2 + \beta^T X^T AX\beta$$

$$= p\sigma^2 + \beta^T X^T AX\beta$$

Substituting this into (B) gives

$$E(J_p) = \frac{1}{\sigma^2}\left[n\sigma^2 + \beta^T X^T X\beta - p\sigma^2 - \beta^T X^T AX\beta\right]$$

$$= (n - p) + \frac{1}{\sigma^2}\beta^T X^T [I - A]X\beta$$

Mallows (1973) recommends the use of the C_p statistic to

estimate $E(J_P)$ and defines it to be

$$C_P = \frac{1}{\hat{\sigma}^2} RSS_P - n + 2p$$

where $\hat{\sigma}^2$ is an estimate of σ^2. If the subset P of predictor variables adequately describes the data there should be negligible bias, i.e. RSS_P will estimate $(n - p)\sigma^2$ and

$$\begin{aligned} C_P &= \frac{RSS_P}{\hat{\sigma}^2} - n + 2p \\ &\approx \frac{(n - p)\sigma^2}{\hat{\sigma}^2} - n + 2p \\ &\approx p \end{aligned}$$

The search for the optimal set of values thus involves identifying those subsets of predictor variables which lead to the smallest C_P values and values which are very close to p.

4.3 MALLOWS' GRAPHICAL METHOD OF SELECTING A MODEL

If the C_P values for different subsets of the predictor variables are plotted against p (the number of variables in the subset), those for subsets with small bias will tend to cluster about the line $C_P = p$ (Figure 3, point A), while those for subsets with substantial bias will fall above the line (Figure 3, point B).

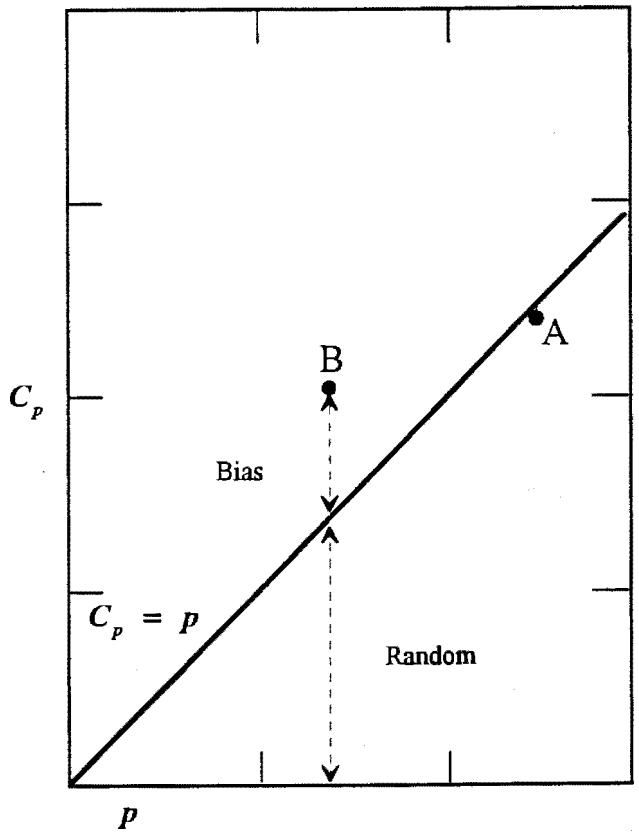

FIGURE 3

Point B in Figure 3 is above the line $C_p = p$ and thus has substantial bias. Point A, on the other hand, lies very close to the line $C_p = p$ and thus has very small bias.

When comparing points A and B it is seen that point B is lower than point A and consequently represents a model with slightly lower ***total error***. But point B corresponds to a subset of the predictor variables with less elements than A. Thus, adding predictor variables to the model equation may reduce bias but increase total error.

4.3.1 Some configurations

First, consider a set of only three predictor variables x_1, x_2, x_3 and suppose they are not highly correlated,

that $\beta = \beta_p$, and that every non-zero element of β is large (relative to the standard error of its least-squares estimate). Then the C_p - plot will look something like Figure 4, drawn for the subsets containing $k - 2$, $k - 1$, k and $k + 1$ predictor variables, where $k = 2$.

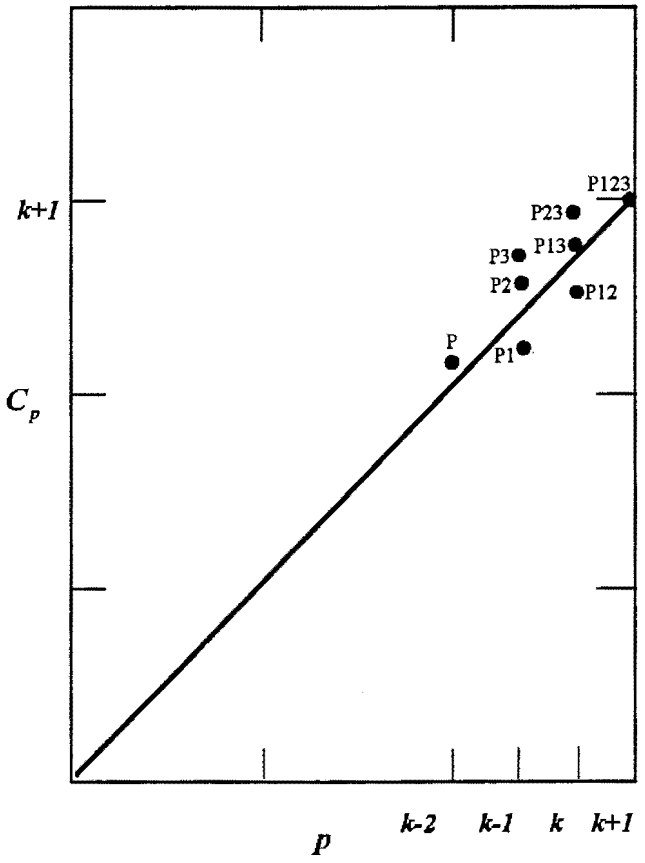

FIGURE 4

Notice that the points corresponding to the different subsets of variables form an approximate linear diagonal configuration.

Now suppose the predictor variables x_1, x_2, x_3 are highly correlated with each other, with each being equally correlated with y. Then any two of these variables, but not all three, can be deleted from the model without much effect. In this case the relevant point on the C_p - plot will look something like Figure 5a, if no other variables are of importance, or like Figure 5b

if another subset *P* is needed. Notice that the diagonal pattern is incomplete.

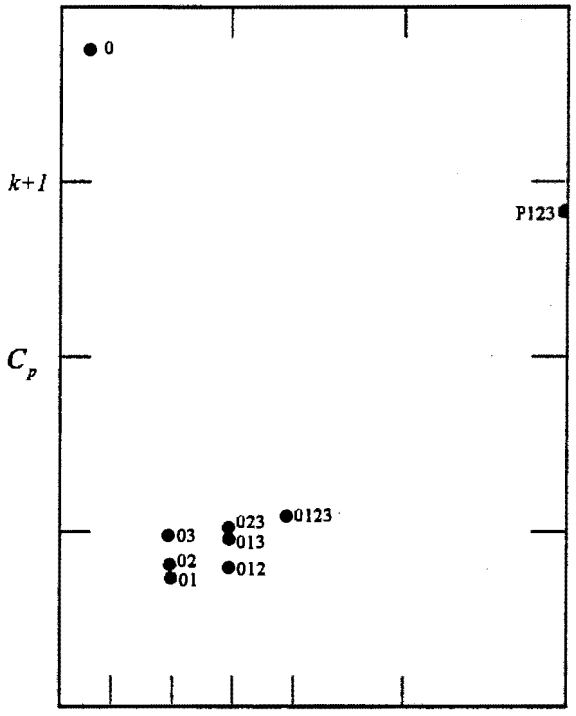

FIGURE 5a

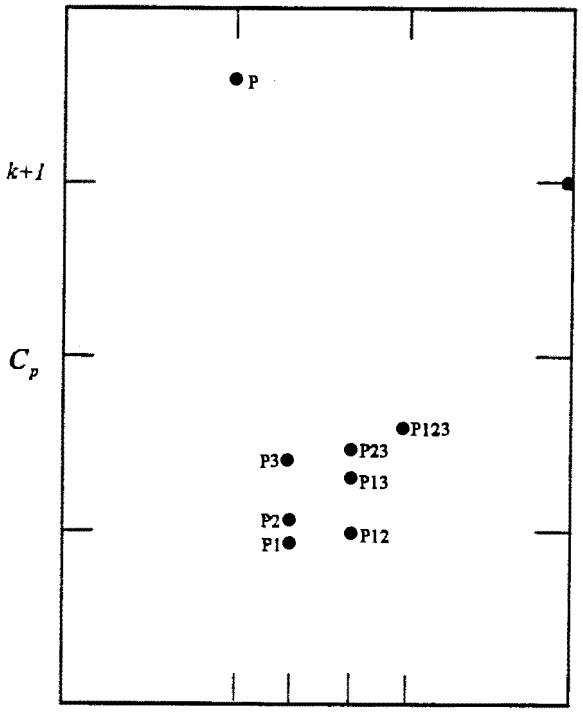

FIGURE 5b

Section 4.3.1: Some configurations

When the variables are jointly explanatory but separately not, the C_p - plot will look something like Figure 6.

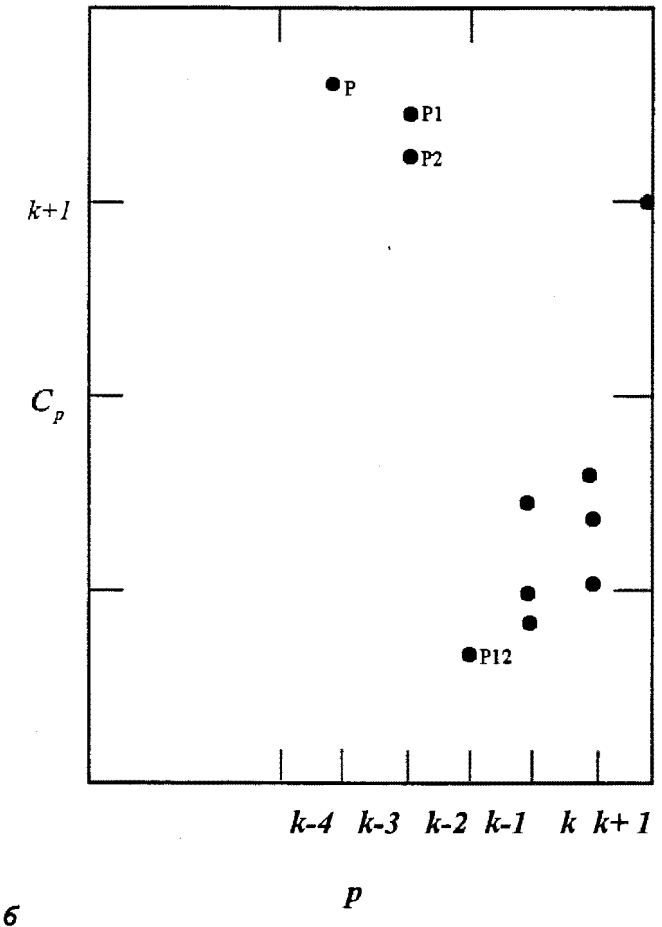

FIGURE 6

It seems then that if the C_p's for a number of subsets of the predictor variables are plotted against p, those for subsets with small bias will tend to cluster about the line $C_p = p$ (and $C_p \approx p$) while those for models with substantial bias will fall above the line. The search for the optimal set of parameters thus involves identifying those sets which lead to the smallest C_p value and very close to p.

Mallows (1973) concludes that this device helps the statistician to examine some aspects of the structure of her data.

4.4 A FINAL REMARK

It is clear from the derivation of Mallows' C_p, that it is based on the Gauss discrepancy. C_p is an estimator of the expected value of $\frac{1}{\sigma^2}\Delta_G$, i.e:

$$C_p = \widehat{E_F \frac{1}{\sigma^2}\Delta_G} = E_{\hat{F}} \frac{1}{\sigma^2}\Delta_G$$

Mallows' C_p is justifiably called a model selection ***criterion*** since an estimator of an expected discrepancy was defined to be a criterion in Chapter 2.

---oOo---

CHAPTER 5

OTHER MODEL SELECTION CRITERIA

5.1 INTRODUCTION

The AIC and C_p paved the way for the development of various other model selection criteria. These are either equivalent or asymptotically equivalent to either AIC or C_p. In this chapter a number of model selection criteria, most of which will be used in Chapter 6, will be listed and there relationship with AIC or C_p discussed.

5.2 CROSS-VALIDATION

Cross-validation originated in the 1930s in attempts to improve the estimation of true multiple correlation from the biased sample multiple correlation (Larson, 1931). In 1963, the cross-validation idea was further developed by F. Mosteller and D. Wallace and was refined in 1968 by Mosteller and J. W. Tukey.

Cross-validation is a method for determining how well a model describes the given data. The emphasis here is placed on the ***predictive ability*** of the model rather than the goodness of fit of the model.

Suppose that n data points are available. One approach is to split the data into two parts. The first part, containing say n_c data points, is used to fit the model. The second part, containing $n_v = n - n_c$ data points, is then used to validate the

model. This is repeated for all $\binom{n}{n_v}$ different subsets of size n_v. Another approach is to delete a single data point and then, using the estimated model parameters, obtain a prediction for the missing observation. This is repeated for each data point.

If two or more models are proposed for a data set, they may be assessed by comparing their cross-validatory scores C, where

$$C = \sum_{i=1}^{n} \frac{L(y_i, \hat{y}_i)}{n}$$

with L some appropriate loss function, for example

$$L(y, \hat{y}) = (y - \hat{y})^2.$$

It should be noted that this loss function is equivalent to the Gauss discrepancy of Chapter 2.

The smaller the value of C, the better the predictive ability of the model. Cross-validation selects the model with the best average predictive ability calculated on all different ways of splitting the data.

For the linear model

$$y_i = x_i^T \beta + \epsilon_i \qquad i = 1, 2, \ldots n$$

consider a subset P of the set of indices $\{0, 1, 2, \cdots, k\}$. Denote by β_P the vector of parameters that consists of the components of β corresponding to the integers in P. Then the model becomes

$$y_i = x_{iP}^T \beta_P + \epsilon_i.$$

For each possible subset P, a different model of this form

results, denoted by M_p.

Under model M_p, the least squares estimator of β_p is

$$\hat{\beta}_p = (X_p^T X_p)^{-1} X_p^T y$$

Suppose that z_i is the future value of the response variable to be predicted. Using the model M_p the average squared prediction error is

$$\frac{1}{n}\sum_{i=1}^{n}\left(z_i - x_{ip}^T\hat{\beta}_p\right)^2$$

Following a similar method of derivation as in Chapter 4, the overall expected squared prediction error is

$$\begin{aligned} E\left[\frac{1}{n}\sum_{i=1}^{n}\left(z_i - x_{ip}^T\hat{\beta}_p\right)^2\right] &= \frac{1}{n}E\left[(Z - X_p\hat{\beta}_p)^T(Z - X_p\hat{\beta}_p)\right] \\ &= \frac{1}{n}\left[n\sigma^2 + \beta^T X^T X\,\beta - P\sigma^2 - \beta^T X^T A X\,\beta\right] \\ &= \sigma^2 - \frac{P}{n}\sigma^2 + \frac{1}{n}\beta^T X^T [I - A] X\beta \\ &\overset{\text{say}}{=} \Gamma_{P;n} \end{aligned}$$

Note that $\Gamma_{P;n}$ consists of two components namely the variability of the future observations as well as $\frac{1}{n}\beta^T X^T [I - A] X\beta - \frac{P}{n}\sigma^2$, which reflects the error in model selection and estimation.

The cross-validation method selects a model by minimizing estimated $\Gamma_{P,n}$. For each model M_p the cross-validation estimate of $\Gamma_{P,n}$ is obtained. The model selected by cross-validation is the model that minimizes the cross-validation estimates over

all P. It is also shown by Shao (1993) and Stone (1977) that cross-validation is asymptotically equivalent to the AIC and Mallows' C_P.

As mentioned briefly in the derivation of the cross-validation method, $\Gamma_{P,n}$ is based on the Gauss discrepancy, i.e. $E\left(\frac{1}{n}\Delta_G\right)$. Cross-validation looks at the estimated value of $\Gamma_{P,n}$, thus $E\left(\frac{1}{n}\Delta_G\right)$, and is therefore, according to the definitions in Chapter 2, a model selection criterion. However, none of the major statistical computer programmes include this method. For this reason it was not used in the examples of Chapter 6.

5.3 R^2 AND ADJUSTED R^2

One way to measure the goodness of fit of a regression model is the multiple correlation coefficient R where

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

Remembering that the total sum of squares is $\sum(y_i - \bar{y})^2 = \sum(y_i - \hat{y}_i)^2 + \sum(\hat{y}_i - \bar{y})^2$, it is clear that R^2 represents the fraction of the total sum of squares which is accounted for by fitting the regression model. The larger the value of R^2, the closer the value of the regression sum of squares $\sum(\hat{y}_i - \bar{y})^2$ to the total sum of squares $\sum(y_i - \bar{y})^2$. And the closer these two values come to each other, the smaller $\sum(y_i - \hat{y}_i)^2$ becomes and thus the better the fit of the model.

Define a model for predicting y without any predictor variables as follows:

$$y_i = \beta_0 + \epsilon_i \qquad i = 1,2,\ldots n$$

Then the estimate of the standard deviation of this model is

$$s_y = \sqrt{\frac{\sum (y_i - \bar{y})^2}{(n-1)}}$$

When considering the full model

$$y_i = \beta_0 + \sum_{j=1}^{k} x_{ij}\beta_j + \epsilon_i \qquad i = 1,2,\ldots n$$

the estimate of the standard deviation is

$$s = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{(n-k-1)}}$$

It should be pointed out that the degrees of freedom is now equal to $n - k - 1$ due to using all the predictor variables x for predicting y.

The worth of x as the predictors of y is measured by comparing s with s_y using:

$$adj\ R^2 = \frac{s_y^2 - s^2}{s_y^2}$$

$$= 1 - \left(\frac{s}{s_y}\right)^2$$

The expression shows that unless s is appreciably smaller than s_y it is not worthwhile to use all the predictor variables x for predicting y. Thus, the larger the value of $adj\ R^2$ the greater the worth of the predictor variables included in the model.

Using the above measures for selecting a model, one would proceed as follows: the value of either R^2 or *adj* R^2 is calculated for each possible subset of the predictor variables. The subset corresponding to the largest value of the measure used, should then be included in the final model. However, R^2 is a non-decreasing function of the number of predictor variables. Therefore selection based on R^2 will always lead to the full model being selected. This is why R^2 is seldom used for model selection purposes.

These two measures are included in most of the major statistical computer programmes available. They form part of the analysis of variance programmes but could be calculated for each possible subset of the predictor variables.

To determine whether R^2 and *adj* R^2 are criteria as defined in Chapter 2, consider the relationship between R^2 or *adj* R^2 and C_p:

R^2 can also be written as

$$R^2 = 1 - \frac{RSS_p}{SST} \qquad \ldots(1)$$

where SST denotes the total sum of squares (Thompson, 1978, p15), and *adj* R^2 can be written as:

$$adj\ R^2 = 1 - \frac{n(1 - R^2)}{(n - p)} \qquad \ldots(2)$$

This notation for R^2 and *adj* R^2 is used for models containing only p of the k predictor variables. When the full model is used (all k predictor variables included), the following notation is used:

$$R_k^2 = 1 - \frac{RSS_k}{SST}$$

and the estimate of the variance is then given as

$$\hat{\sigma}^2 = \frac{RSS_k}{(n - k)}$$

Making RSS_p the subject of (1), it follows that:

$$RSS_p = (1 - R^2)\,SST$$

Substituting this into C_p gives:

$$C_p = \frac{RSS_p}{\hat{\sigma}^2} + 2p - n$$

$$= \frac{(1 - R^2)\,SST}{\frac{RSS_k}{(n - k)}} + 2p - n$$

$$= \frac{(1 - R^2)\,SST}{\frac{(1 - R_k^2)\,SST}{(n - k)}} + 2p - n$$

$$= (n - k)\,\frac{(1 - R^2)}{(1 - R_k^2)} + 2p - n$$

It is clear that C_p can be written as a function of the multiple correlation coefficient.

To find the relationship between C_p and $adj\ R^2$ make $(1 - R^2)$ the subject of (2):

$$(1 - R^2) = \frac{(n - p)}{n}\,(1 - adj\ R^2)$$

Then

$$C_p = (n - k)\,\frac{\left(\frac{n - p}{n}\right)(1 - adj\ R^2)}{\left(\frac{n - k}{n}\right)(1 - adj\ R_k^2)} + 2p - n$$

$$= (n - p)\,\frac{(1 - adj\ R^2)}{(1 - adj\ R_k^2)} + 2p - n$$

It is clear that C_p can also be written as a function of the adjusted multiple correlation coefficient.

It seems then that R^2 is not a criterion as defined in Chapter 2 but a function of a criterion:

$$R^2 = 1 - \frac{RSS_p}{SST}$$

$$= 1 - \frac{\hat{\sigma}^2 (C_p - 2p + n)}{SST}$$

That explains why, when selecting a model using R^2 and *adj* R^2, its ***maximum*** value is used in stead of its ***minimum*** as is the case with criteria.

5.4 MEAN SQUARE ERROR

The mean square error is also included in an analysis of variance output and could also be used for model selection. It is defined as

$$\text{MSE} = \frac{\text{SSE}}{(n - k)}$$

$$= \frac{\sum (y_i - \hat{y}_i)^2}{(n - k)}$$

(Searle, 1971, p92) where SSE will be referred to as the error sum of squares. The smaller the value of MSE, the closer the predicted values come to the real values of the response variables. When using MSE as a means of model selection, one would calculate the value of MSE for each possible subset of the predictor variables and then select the subset corresponding to the smallest value of MSE for inclusion in the final model.

Note that SSE is equal to the Gauss discrepancy of Chapter 2 and thus

$$\text{MSE} = \frac{1}{(n - k)} \Delta_G$$

It seems then that MSE can be described as a discrepancy. That is why one wants to find the ***minimum*** MSE, since a minimum discrepancy implies that the approximating model is very "close" to the true model.

One could even rewrite MSE as:

$$\Delta_{\text{MSE}} = \frac{\sum (y_i - \hat{y}_i)^2}{(n - k)}$$

---oOo---

CHAPTER 6

MODEL SELECTION EXAMPLES

6.1 INTRODUCTION

The previous chapters considered the theory of some model selection procedures. The objective of this chapter is to apply model selections methods to two different examples. The procedures of the previous chapters will be used in the model selection exercise and for convenience sake the following table is provided:

AIC	$-2MLL + 2p$
C_P	$\frac{1}{\hat{\sigma}^2} RSS_p - n + 2p$
R^2	$1 - \frac{RSS_p}{SST}$
$adj\ R^2$	$adj\ R^2 = 1 - \frac{n(1 - R^2)}{(n - p)}$
MSE	$\frac{SSE}{(n - k)}$

If a single method is to be used for model selection, the procedure is easy. For AIC the variable set corresponding to its minimum value will be selected for the final model. For either R-squared or Adjusted R-squared the variable set corresponding to the maximum value of either of the two criteria will be

selected for the final model. The choice for the number of predictors to be included in the model using MSE can be made at the value where MSE slowly decreases for further increase in the number of predictors or at its minimum value (if it exists). For Mallows' C_p, the number of variables to be included corresponds to its minimum value but it must also be true that $C_p \approx p$ or at least $C_p < p$. However, different criteria may result in different models being selected. This will be illustrated in the first model selection example of §6.3.

6.2 MULTICOLLINEARITY

Before applying model selection procedures to the data, it is advisable to test for multicollinearity amongst the predictor variables. The predictor variables are said to be multicollinear if they are linearly dependent or near linearly dependent. If this occurs then the design matrix X is not of full rank and $X'X$ is thus singular or near singular. One consequence of this is that the normal equations

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

do not have a unique solution. Another consequence is that the normal equations are ill conditioned, meaning that a small change in the predictor variables results in a large change in the estimation of β. This can result in large standard errors of the estimated regression coefficients. This can distort the relative importance of the predictors in the model. Another consequence of near linear relationships is that the smallest eigenvalue of $X'X$ may be zero or close to zero and be much smaller than the largest eigenvalue. If any linear or near linear relationships exist, some of the predictor variables involved in the relationship should be excluded from the model selection procedure in order to remove the linear dependence.

Belsley at al. (1980) propose the following diagnostic procedure for multicollinearity: The procedure is based on the eigenvalues of the matrix $X'X$ from which they derive condition numbers and variance proportions. The condition numbers are defined as the square roots of the ratios of the largest eigenvalue to the remaining eigenvalues of $X'X$. The variance proportions are obtained as follows: the predictor variables can be transformed to an equal number of pairwise uncorrelated variables called principal components. Each principal component corresponds to a different eigenvalue of the matrix $Z'Z$. The variance of each estimated regression coefficient can be decomposed into a sum of terms, each associated with a principle component. Each variance can thus be expressed as a sum of proportions. To examine near linear relationships, a useful guideline is to find the variance proportions greater than approximately 0,5 associated with a high condition number (Belsley, 1980). These predictor variables are then involved in near linear relationships. Various statistical computer programmes give the condition numbers, eigenvalues and variance proportions as standard output.

6.3 MODEL SELECTION EXAMPLES

6.3.1 The ethnicity problem of the HSRC researcher

Consider the ethnicity problem of the HSRC researcher described in Chapter 1. The data was examined for any evidence of multicollinearity using the collinearity diagnostics procedure from SAS. The results are shown in Table 3 of the appendix. However, no strong evidence of multicollinearity was found. If multicollinearity is present in these data, it is not evident from these diagnostics.

Model selection methods were then applied to the data using the PROC REG procedure from SAS, including the five best regression models for each number of independent variables. The following selection procedures were used in the analysis: R-squared, Adjusted R-squared, C_p, AIC and MSE. The results are shown in Table 4 of the appendix.

The minimum C_p value is equal to 5,587 with $p = 9$ and thus $C_p < p$. The plot of C_p versus p in Figure 7 shows that a minimum is reached at the subset containing the 9 variables A, B, G, H, J, K, L, R and S, and that $C_p < p$ is satisfied.

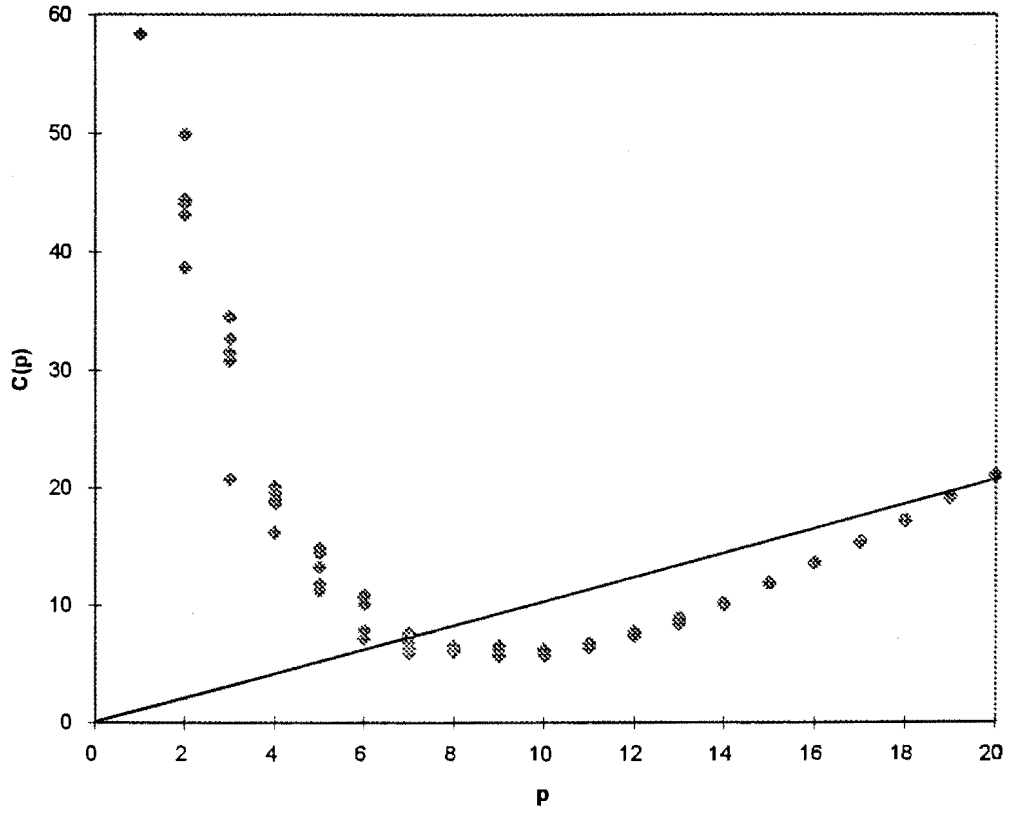

FIGURE 7: C_p-plot

The minimum AIC value is 1277, also obtained using the above mentioned variable subset. However, this minimum value is also obtained for another subset containing 9 variables as well as for two other subsets containing 10 variables each (see highlighted values in Table 4 in the appendix). It seems at this stage as if the subset containing the 9 variables A, B, G, H, J, K, L, R and S will be included in the model since both the criteria for AIC and C_p are satisfied.

The maximum values for R-squared and Adjusted R-squared were 0.578 and 0.564 respectively. The subsets corresponding to this maximum R-squared value, ranges between subsets containing 15 to all 20 variables (see Table 4 in appendix). A choice of any of these subsets would satisfy the R-squared criterion. The maximum Adjusted R-squared values are also highlighted in Table 4. Subsets containing between 11 and 13 variables satisfy this criterion.

Consider the subset containing the above mentioned 9 variables. The R-squared value for this subset is 0.572 which is very close to the maximum value of 0.578. The maximum value of Adjusted R-squared for this subset is even closer to its real maximum, namely 0.562. It thus seems as if this subset not only satisfies both the criteria for AIC and C_p, but also nearly satisfies the criteria for R-squared and Adjusted R-squared. The only criterion not satisfied using this subset is MSE. However, investigating the variables included when MSE reaches a minimum, it is seen that all 9 of the above mentioned variables are included in this subset.

The regression coefficients for the final model were estimated using the SAS PROC REG procedure. The results are shown in Table 5 of the appendix. The final model is as follows:

$$\hat{y} = 12.381 + 0.786A + 0.321B + 0.197G + 0.0187H$$
$$+ 0.0541J - 0.198K + 0.205L + 0.408R - 0.327S$$

Thus, it seems as if the psychological identification of a person with his/her own ethnic group is influenced by:

* Obtaining a well defined identity; involvement with ethnical group and exploration with ethnical identity.
* Feelings of ambivalence with respect to membership of ethnic group versus willingness to defend interests of the group.
* How far the symbols of a respondent's ethnical group will be recognised under a new dispensation.
* Attitudes towards negotiations with respect to the solutions of South Africa's problems.
* How far the identity of a respondent's ethnical group will be threatened under a new dispensation.
* Positive and acceptable social behaviour towards other racial groups (blacks versus whites and vice versa).
* Negative intergroup behaviour versus stereotyping of other racial groups.
* Positive feelings towards affirmative action.

6.3.2 A production process

This example was taken from Thompson (1978b) who used data from Draper and Smith (1966). They wanted to find a model to predict the pounds of steam used monthly in a certain production process, given ten independent predictor variables (see appendix Table 6).

The data was examined for any evidence of multicollinearity using

the collinearity diagnostics procedure from SAS. The results are shown in Table 7 of the appendix. No strong evidence of multicollinearity was found amongst the first nine variables. The unit variable was however excluded from the model selection procedure.

Only Mallows' C_p was used as model selection criterion. The reason for this being that the predictor variables are non-stochastic. In this instance Thompson (1978b) recommends the use of Mallows' C_p. The results of the SAS PROC REG procedure are shown in Table 8 of the appendix.

A minimum is achieved at $p = 6$ where $C_p = 5{,}368$. It must be noted, however, that this minimum value is not substantially smaller than the minimum associated with $p = 3$, namely $C_p = 5{,}586$. The C_p -plot of Figure 8 also illustrates this.

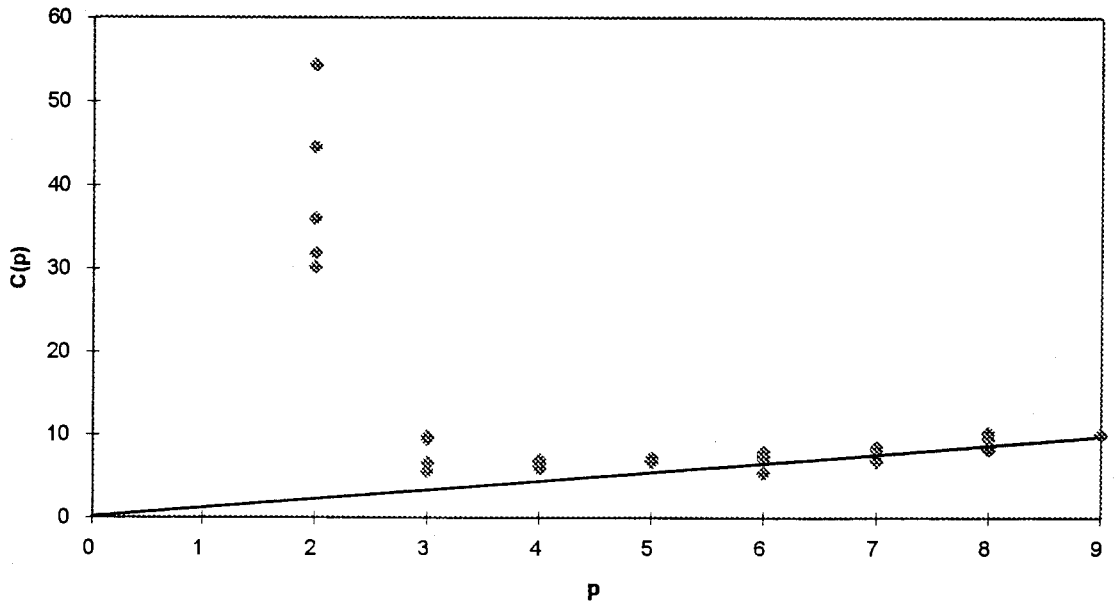

FIGURE 8: C_p-plot

However, for $p = 3$ it is not true that $C_p < p$. Although Thompson (1978b) suggests the use of this particular subset for inclusion in the final model, we suggest the use of the subset corresponding to $p = 6$ because $C_p < p$.

The estimated regression coefficients for both the abovementioned subsets were obtained using the SAS PROC REG procedure. The results are shown in Table 9 of the appendix. The final model for the subset corresponding to $p = 3$ is as follows:

$$y = 0,978 + 1,638X_4 - 5,106X_5 + 0,484X_7$$

and the pounds of steam used monthly in this production process can thus be predicted using:

* Calender days per month
* Operating days per month
* Average atmospheric temperature

The final model for the subset corresponding to $p = 6$ is as follows:

$$y = -9,168 + 1,249X_1 - 5,464X_3 + 0,135X_5 + 0,339X_7 + 0,175X_8 + 0,091X_9$$

and the pounds of steam used monthly in this production process can thus be predicted using:

* Pounds of real fatty acid in storage per month
* Average wind velocity in miles per hour
* Operating days per month
* Average atmospheric temperature
* (Average wind velocity)^2
* Number of startups

The above examples illustrate the use of model selection procedures under different circumstances. In some cases a model is required for prediction purposes (the production process example). In other cases a model is needed not for prediction but for explaining or understanding an existing phenomenon (the ethnicity example).

6.4 SUGGESTIONS FOR FURTHER RESEARCH

The problem of model selection occurs almost everywhere in statistics and is thus an important part of statistical theory. In this thesis we have considered a number of well known model selection criteria. However, it is clear that much work still has to be done to provide guidelines as to which criteria are appropriate under various circumstances. Although Thompson (1978) did make some recommendations about the choice of criteria, these criteria themselves have been extended since she wrote her paper and it is still not clear which is the best criterion in any given situation.

A further problem is that once a model has been selected then it is standard practise to proceed with inference about the selected model ignoring the model selection process. This results in very optimistic inferences. Faraway (1994) has done some pioneering work to address this problem in a linear regression setting and has shown that the strategy of ignoring the model selection procedure results in overly optimistic inference. Venter and Snyman (1995) also address this problem in the setting of estimating the mean of a normal population. Clogg (1994) addresses this problem within the context of causal inference using regression models. However, it is clear from these authors that the problem of including model selection in the inferential process is a difficult one and is not yet solved.

---oOo---

REFERENCES

AKAIKE, H. (1969), "Final Prediction Error (FPE) Criterion", *Annals of the Institute of Statistical Mathematics*, 21, 243 - 247.

ATKINSON, A.C. (1978), "Simple Bayesian Formula is misleading", *Biometrika*, 65, 39 - 48.

BELSLEY, D.A., KUH, E., WELSCH, R.E. (1980), *Regression diagnostics identifying influential data and sources of collinearity*, New York: Wiley.

BOZDOGAN, H. (1987), "Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions", *Psychometrika*, 52, 345 - 370.

CLOGG, C.C. (1994), "The regression method of causal inference and a dilemma with this method", Annual Conference of the South African Statistical Association, November 1994.

DRAPER, N.R., SMITH, H. (1966), *Applied Regression Analysis*, New York: John Wiley.

FARAWAY, J.J. (1994), *Choice of order in regression strategy. Selecting models from data*, Editors: CHEESEMAN, P., OLDFORD, R.W.

FINDLEY, D.F., PARZEN, E. (1995), "A conversation with Hirotugu Akaike", *Statistical Science*, Volume 10, No 1, 104 - 117.

HURVICH, C.M., TSAI, C. (1989), "Regression and time series model selection in small samples", *Biometrika*, 76, 297 - 307.

References

LARSON, S.C. (1931), *Journal of Educational Psychology*, 22, 45 - 55.

LINHART, H., ZUCCHINI, W. (1986), *Model Selection*, New York: John Wiley & Sons.

MALLOWS, C.L. (1973), "Some comments on C_p", *Technometrics*, 15, 661 - 675.

MOSTELLER, F., TUKEY, J.W. (1968), *Handbook of Social Psychology*, Addison-Wesley.

MOSTELLER, F., WALLACE, D. (1963), *Journal of the American Statistical Association*, 58, 275 - 309.

RAO, C.R. (1965), *Linear Statistical Inference and its Applications*, New York: John Wiley & Sons.

SAKAMOTO, Y., ISHIGURO, M., and KITAGAWA, G. (1986), *Akaike Information Criterion Statistics*, Tokyo: D. Reidel Publishing company.

SEARLE, S.R. (1971), *Linear Models*, New York: John Wiley & Sons.

SERFLING, R.J. (1980), *Approximation Theorems of Mathematical Statistics*, New York: John Wiley & Sons.

SHAO, J. (1993), "Linear model selection by Cross-validation", *Journal of the American Statistical Association*, 88, 486 - 494.

STONE, M. (1974), "Cross-validatory choice and assessment of Statistical Predictions", *Journal of the Royal Statistical Society*, Series B, 36, 111 - 147.

———— (1977), "An asymptotic equivalence of choice of model by Cross-validation and Akaike's Criterion", *Journal of the Royal Statistical Society*, Series B, 39, 44 - 47.

STREMLER, F.G. (1982), *Introduction to Communication Systems*, London: Addison-Wesley Publishing Company.

THOMPSON, M.L. (1978a), "Selection of Variables in Multiple Regression: Part I. A Review and Evaluation", *International Statistical Review*, 46, 1 - 19.

———— (1978b), "Selection of Variables in Multiple Regression: Part II. Chosen Procedures, Computations and Examples", *International Statistical Review*, 46, 129 - 146.

VENTER, J.H., SNYMAN, J.L.J. (1995), "A note on the generalised cross-validation in linear model selection", *Biometrika*, 82, 1 - 4.

The following references, although not explicitly referred to, were used in some or other sense in the development of the thesis:

DANIEL, C., WOOD, F.S. (1971), *Fitting Equations to Data*, New York: John Wiley & Sons.

MURTHY, D.N.P., PAGE, N.W., RODIN, E.Y. (1990), *Mathematical Modelling. A Tool for Problem Solving in Engineering, Physical, Biological and Social Sciences*, New York: Pergamon Press.

SAMSA, G., ODDONE, E.Z. (1994), "Integrating Scientific Writing Into a Statistics Curriculum: A Course in Statistically

Based Scientific Writing", *The American Statistician*, 48, 117 - 119.

TAMURA, H. (1994), "Model Comparison in Regression", *Teaching Statistics*, Volume 16, Number 2, 47 -49.

VAN DER MERWE, A.J., GROENEWALD, P.C.N., DE WAAL, D.J., VAN DER MERWE, C.A. (1983), "Model Selection for future data in the case of multivariate regression analysis", *South African Statistical Journal*, 17, 147 - 164.

ZHANG, P. (1992), "On the distributional properties of Model Selection Criteria", *Journal of the American Statistical Association*, 419, 732 - 737.

---oOo---

APPENDIX

TABLE 1

Nine possible predictors for pork fat content

1	An average of three measures of back fat thickness.
2	A muscling score for the carcass. The higher the number, the more the muscle and less fat.
3	An average of three measures of fat depth opposite the tenth rib.
4	Live weight (*kg*) of the carcass.
5	Weight (*kg*) of the slaughtered carcass.
6	A measure used to determine specific gravity. The higher the measure, the lower the percentage fat.
7	The average of three determinations of the depth of the belly.
8	The average measure of leanness of three cross sections of the belly.
9	Total weight (*kg*) of the belly.

	EXTREMELY IMPORTANT/ Besonder belangrik	REASONABLY IMPORTANT/ Redelik belangrik	CANNOT SAY/ NEUTRAL/ Kan nie sê nie/ Neutraal	REASONABLY UNIMPORTANT/ Redelik Onbelang rik	TOTALLY UNIMPORTANT/ Heeltemal onbelangrik	
A SPECIFIC STATUS OR POSITION IN SA SOCIETY/A SPECIFIC SOCIO-ECONOMIC CLASS/'n Spesifieke status of posisie in die SA gemeenskap/'n Spesifieke sosio-ekonomiese klas	1	2	3	4	5	58
SPECIFIC SONGS/FORMS OF MUSIC/FORMS OF DANCE/Spesifieke liedere/vorms van musiek/dansvorms	1	2	3	4	5	59
SPECIFIC UNIFORMS, ATTIRE OR ACCESSORIES (E.G. VOORTREKKER DRESS, KHAKI CLOTHES, CULTURAL WEAPONS, BEADS)/ Spesifieke uniforms, kleredrag of bykomstighede (bv. Voortrekkerdrag, Kakieklere, kulturele wapens, krale)	1	2	3	4	5	60
A SPECIFIC IDEOLOGY (E.G. COMMUNISM, SOCIALISM)/'n Spesifieke ideologie (bv. Kommunisme, sosialisme)	1	2	3	4	5	61
SPECIFIC IDEALS OR ASPIRATIONS (E.G. FREEDOM, INDEPENDENCE, POWER, SELF-DETERMINATION, ETC.)/Spesifieke ideale (bv. vryheid, onafhanklikheid, mag, selfbeskikking, ens.)	1	2	3	4	5	62

D08. THE FOLLOWING STATEMENTS MAY INDICATE HOW YOU FEEL ABOUT ..... **(FIELDWORKER: NAME THE GROUP THAT RESPONDENT MENTIONED IN QUESTION D01/QUESTION D06).** PLEASE INDICATE WHETHER YOU AGREE STRONGLY(1), AGREE(2), ARE NEUTRAL(3), DISAGREE(4) OR DISAGREE STRONGLY(5) WITH EACH STATEMENT./Die volgende stellings toon hoe u moontlik voel oor ..... **(Veldwerker: Noem die groep wat respondent in Vraag D01/Vraag D06 genoem het).** Dui asseblief aan of u van harte saamstem(1), Saamstem(2), Neutraal staan(3), Verskil(4) of sterk verskil(5) ten opsigte van elke stelling.

	STRONGLY AGREE/ Stem van harte saam	AGREE/ Stem saam	NEUTRAL/ Neutraal	DISAGREE/ Verskil	STRONGLY DISAGREE/ Verskil sterk	
LOYALTY TOWARDS MY OWN ETHNIC OR CULTURAL GROUP IS PARTICULARLY IMPORTANT TO ME/Trou aan my eie etniese of kulturele groep is vir my besonder belangrik	1	2	3	4	5	63
IT UPSETS ME WHEN OTHER PEOPLE SPEAK NEGATIVELY ABOUT MY OWN ETHNIC OR CULTURAL GROUP/Dit ontstel my wanneer ander mense iets afbrekends oor my eie etniese of kulturele groep sê	1	2	3	4	5	64

	STRONGLY AGREE/ Stem van harte saam	AGREE/ Stem saam	NEU-TRAL/ Neu-traal	DIS-AGREE/ Ver-skil	STRONGLY DISAGREE/ Verskil sterk	
PRESERVING THE IDENTITY OF MY OWN ETHNIC OR CULTURAL GROUP IS NOT VERY IMPORTANT TO ME/Die bewaring van die identiteit van my eie etniese of kulturele groep is nie vir my van groot belang nie	1	2	3	4	5	65
I DO NOT WANT TO BELONG TO ANY OTHER ETHNIC OR CULTURAL GROUP/ Ek wil niks anders as 'n lid van my eie etniese of kulturele groep wees nie	1	2	3	4	5	66
I WOULD BE WILLING TO TAKE ACTION IF THE IDENTITY OF MY OWN ETHNIC OR CULTURAL GROUP IS THREATENED/Ek sou bereid wees om aktief op te tree as die identiteit van my eie etniese of kulturele groep bedreig word	1	2	3	4	5	67
I RESPECT A PERSON WHO TAKES PRIDE IN THE SPECIAL QUALITIES OF HIS OR HER OWN ETHNIC OR CULTURAL GROUP/Ek het respek vir iemand wat trots is op die besondere eienskappe van sy eie etniese of kulturele groep	1	2	3	4	5	68
COMMITMENT TO THE CULTURE OF MY OWN ETHNIC OR CULTURAL GROUP IS A MAJOR SOURCE OF SECURITY IN MY LIFE/ Gebondenheid aan die kultuur van my eie etniese of kulturele groep is een van die belangrikste bronne van sekuriteit in my lewe	1	2	3	4	5	69
PROTECTING THE CUSTOMS OF MY OWN ETHNIC OR CULTURAL GROUP IS UNNECESSARY/Die instandhouding van die tradisies van my eie etniese of kulturele groep is onnodig	1	2	3	4	5	70
I HAVE SPENT TIME TRYING TO FIND OUT MORE ABOUT MY OWN ETHNIC OR CULTURAL GROUP, SUCH AS ITS HISTORY, TRADITIONS, AND CUSTOMS/Ek het tyd daaraan bestee om meer omtrent my eie etniese of kulturele groep uit te vind, bv. sy geskiedenis, tradisies en gebruike	1	2	3	4	5	71
I AM ACTIVE IN ORGANIZATIONS OR SOCIAL GROUPS THAT INCLUDE MOSTLY MEMBERS OF MY OWN ETHNIC OR CULTURAL GROUP/Ek is aktief betrokke by organisasies of sosiale groepe wat grootliks bestaan uit lede van my eie etniese of kulturele groep	1	2	3	4	5	72

	STRONGLY AGREE/	AGREE/	NEU-TRAL/	DIS-AGREE/	STRONGLY DISAGREE/	
I HAVE A CLEAR SENSE OF MY ETHNIC OR CULTURAL BACKGROUND AND WHAT IS MEANS TO ME/Ek het 'n duidelike begrip van my etniese of kulturele agtergrond en wat dit vir my beteken	1	2	3	4	5	73
I THINK A LOT ABOUT HOW MY LIFE IS AFFECTED BY MY ETHNIC OR CULTURAL GROUP MEMBERSHIP/Ek dink baie oor hoe my lewe deur lidmaatskap van my etniese of kulturele groep beïnvloed word	1	2	3	4	5	74
						3 1 2-5
I AM HAPPY THAT I AM A MEMBER OF THE ETHNIC OR CULTURAL GROUP THAT I BELONG TO/Ek is gelukkig dat ek 'n lid is van die etniese of kulturele groep waaraan ek behoort	1	2	3	4	5	6
I AM NOT VERY CLEAR ABOUT THE ROLE OF ETHNICITY OR CULTURE IN MY LIFE/Ek het nie groot duidelikheid oor die rol van etnisiteit of kultuur in my lewe nie	1	2	3	4	5	7
I UNDERSTAND PRETTY WELL WHAT MY ETHNIC OR CULTURAL GROUP MEMBERSHIP MEANS TO ME, IN TERMS OF HOW TO RELATE TO MY OWN GROUP AND TO OTHER GROUPS/Ek begryp heeltemal goed wat lidmaatskap van my etniese of kulturele groep vir my beteken ten opsigte van die verhouding tot my eie groep en ander groepe	1	2	3	4	5	8
I HAVE A LOT OF PRIDE IN MY ETHNIC OR CULTURAL GROUP AND ITS ACCOMPLISHMENTS/Ek is baie trots op my etniese of kulturele groep en sy prestasies	1	2	3	4	5	9
I HAVE A STRONG SENSE OF BELONGING TO MY OWN ETHNIC OR CULTURAL GROUP/Ek voel baie sterk daaroor dat ek tot my eie etniese of kulturele groep behoort	1	2	3	4	5	10
I PARTICIPATE IN CULTURAL PRACTICES OF MY OWN GROUP, SUCH AS SPECIAL FOOD, MUSIC, OR CUSTOMS/Ek neem deel aan die kulturele gebruike van my eie groep, bv. spesiale kos, musiek of gewoontes	1	2	3	4	5	11

	STRONGLY AGREE/ Stem van harte saam	AGREE/ Stem saam	NEU-TRAL?/ Neu-traal?	DIS-AGREE/ Ver-skil	STRONGLY DISAGREE/ Verskil sterk	
I FEEL A STRONG ATTACHMENT TOWARDS MY OWN ETHNIC OR CULTURAL GROUP/Ek voel 'n sterk verbondenheid aan my eie etniese of kulturele groep	1	2	3	4	5	12
I SOMETIMES FEEL UNCOMFORTABLE ABOUT MY CULTURAL OR ETHNIC BACK-GROUND/Ek voel soms ongemaklik oor my kulturele of etniese agtergrond	1	2	3	4	5	13

FIELDWORKER: PLEASE MENTION THE SPECIFIC AREA IN WHICH THE RESPONDENT IS LIVING - THAT IS THE PRETORIA, WITWATERSRAND OF VAAL TRIANGLE AREA - IN QUESTION D09.1
Veldwerker: Noem asseblief die spesifieke area waarin die respondent woonagtig is - d.w.s. die Pretoria-, Witwatersrand- of Vaaldriehoekarea - in Vraag D09.

D09. IN THE QUESTIONS THAT FOLLOW WE ARE ESPECIALLY INTERESTED IN YOUR PERSONAL IMPRESSIONS OF WHAT THE POSITION OF ..... **(FIELDWORKER: NAME GROUP THAT RESPONDENT MENTION IN QUESTION D01/QUESTION D06)** WILL BE IN A NEW POLITICAL DISPENSATION OR IN THE SO-CALLED NEW SOUTH AFRICA, THAT IS AFTER THE ELECTIONS OF 27 APRIL 1994./In die volgende vrae stel ons in die besonder belang in u persoonlike indrukke van wat die posisie van ..... **(Veldwerker: Noem groep wat respondent in Vraag D01/Vraag D06 genoem het)** sal wees in 'n nuwe politieke bedeling of in die sogenaamde nuwe Suid-Afrika, d.w.s. na die verkiesing van 27 April 1994.

a) PLEASE ESTIMATE THE PERCENTAGE OF THE TOTAL POPULATION OF THE ..... AREA THAT WILL BE MADE UP BY MEMBERS OF YOUR GROUP:/Skat asseblief watter persentasie van die totale bevolking van die ..... area uit lede van u groep sal bestaan:

0 - 20 %	1
21 - 40 %	2
41 - 60 %	3
61 - 80 %	4
81 - 100 %	5

14

b) HOW HIGHLY WILL THE LANGUAGE OF YOUR GROUP BE REGARDED IN THE ..... AREA?/Hoe hoog sal die taal van u groep geag word in die ..... area?

NOT HIGHLY AT ALL / Glad nie hoog nie — 1 2 3 4 5 — EXTREMELY HIGHLY / Besonder hoog

15

c) HOW HIGHLY WILL THE LANGUAGE OF YOUR GROUP BE REGARDED INTERNATIONALLY?/Hoe hoog sal die taal van u groep internasionaal geag word?

NOT HIGHLY AT ALL / Glad nie hoog nie — 1 2 3 4 5 — EXTREMELY HIGHLY / Besonder hoog

16

d) HOW OFTEN WILL THE LANGUAGE OF YOUR GROUP BE USED IN GOVERNMENT SERVICES (E.G. REGISTRATION OFFICES, COURTS OF LAW, ETC.) IN THE ..... AREA?/Hoe dikwels sal die taal van u groep in regeringsinstellings (bv. registrasiekantore, howe, ens.) in die ..... area gebruik word?

NOT AT ALL / Glad nie — 1 2 3 4 5 — EXCLUSIVELY / Uitsluitlik [17]

e) HOW MUCH CONTROL WILL YOUR GROUP HAVE OVER ECONOMIC AND BUSINESS MATTERS IN THE ..... AREA?/Hoeveel beheer sal u groep oor ekonomiese en sake-aangeleenthede in die ..... area hê?

NONE AT ALL / Geen — 1 2 3 4 5 — EXCLUSIVE CONTROL / Uitsluitlike beheer [18]

f) TO WHAT EXTENT WILL THE LANGUAGE OF YOUR GROUP BE USED ON RADIO AND TELEVISION?/In watter mate sal die taal van u groep oor die radio en televisie gebruik word?

NOT AT ALL / Glad nie — 1 2 3 4 5 — EXCLUSIVELY / Uitsluitlik [19]

g) HOW HIGHLY WILL YOUR GROUP BE REGARDED IN THE ..... AREA?/Hoe hoog sal u groep in die ..... area geag word?

NOT HIGHLY AT ALL / Glad nie hoog nie — 1 2 3 4 5 — EXTREMELY HIGHLY / Besonder hoog [20]

h) TO WHAT EXTENT WILL YOUR GROUP REPRESENT A NUMERIC MINORITY OR MAJORITY IN THE ..... AREA?/In watter mate sal u groep 'n numeriese minderheid of meerderheid in die ..... area wees?

VERY SMALL MINORITY / Baie klein minderheid — 1 2 3 4 5 — VERY LARGE MAJORITY / Baie groot meerderheid [21]

i) TO WHAT EXTENT WILL THE LANGUAGE OF YOUR GROUP BE THE MEDIUM FOR TEACHING IN SCHOOLS IN THE ..... AREA?/In watter mate sal die taal van u groep die medium van onderrig in skole in die ..... area wees?

NOT AT ALL / Glad nie — 1 2 3 4 5 — EXCLUSIVELY / Uitsluitlik [22]

j) HOW MUCH POLITICAL POWER WILL YOUR GROUP HAVE IN THE ..... AREA?/Hoeveel politieke mag sal u groep in die ..... area hê?

COMPLETE POWER / Totale mag — 1 2 3 4 5 — NONE AT ALL / Geen [23]

k) TO WHAT EXTENT WILL THE LANGUAGE OF YOUR GROUP BE USED WITHIN BUSINESS INSTITUTIONS IN THE ..... AREA?/In watter mate sal die taal van u groep in die sakesektor van die ..... area gebruik word?

NONE AT ALL / Glad nie — 1 2 3 4 5 — EXCLUSIVELY / Uitsluitlik [24]

TABLE 2

Twenty variables for HSRC researcher

SYMBOL	VARIABLE	EXPLANATION
A	Involved	Composed of three elements: Obtaining a well defined identity; involvement with ethnic group and exploration of ethnic identity.
B	Ambivalent	Feelings of ambivalence with respect to membership to ethnic group versus willingness to defend interests of the group.
C	Perception	Future perceptions on circumstances in a new South Africa with respect to intergroup relationships, crime, safety and violence.
D	Language	Institutional support that the language of the respondent's ethnic group might enjoy under a new dispensation.
E	Power	The status and power that the respondent's group might enjoy under a new dispensation.
F	Assertiveness	How well an individual's ethnic group will succeed in maintaining their own identity under a new dispensation.
G	Symbols	How far the symbols of a respondent's ethnic group will be recognised under a new dispensation.
H	Negotiations	Attitudes towards negotiations with respect to the solutions of South Africa's problems.
I	Militant	Attitudes towards militant behaviour.
J	Threats	How far the identity of a respondent's ethnic group will be endangered under a new dispensation.
K	Acceptance	Positive and acceptable social behaviour towards other racial groups (blacks versus whites and vice versa).
L	Prejudice	Negative intergroup behaviour versus stereotyping of other racial group.
M	Prejudice2	Negative intergroup behaviour versus stereotyping of other racial group after items with low item total correlations were deleted.
N	Anxiety	Feelings of uncertainty because of social and political changes.
O	Liaison	Evaluation of the nature of contact with other race groups in the work situation.
P	Relationship	Evaluation of the relationships between racial groups in the respondent's work situation.
Q	AA_neg	Negative feelings towards affirmative action.
R	AA_pos	Positive feelings towards affirmative action.
S	AA_pos3	Positive feelings towards affirmative action after items with low item total correlations were deleted.
T	Merits	Evaluation of how far promotions and appointments in respondent's work situation are based on merit.

TABLE 3

Collinearity diagnostics

Number	Eigenvalue	Condition number	Var prop Intercept	Var prop A	Var prop B	Var prop C
1	18.57	1	0	0	0	0
2	0.725	5.063	0	0	0.001	0.005
3	0.501	6.086	0	0	0	0.001
4	0.198	9.673	0	0.002	0.003	0.001
5	0.173	10.368	0	0	0	0
6	0.134	11.79	0	0.002	0.007	0
7	0.124	12.22	0	0.003	0.001	0.006
8	0.098	13.756	0	0.001	0	0.012
9	0.083	14.916	0	0.031	0.019	0.02
10	0.079	15.329	0	0	0.006	0.071
11	0.065	16.964	0	0	0.107	0.389
12	0.051	19.061	0	0.086	0.285	0.093
13	0.043	20.861	0	0.145	0	0.016
14	0.037	22.394	0	0.041	0.006	0.005
15	0.029	24.932	0.001	0.304	0.179	0.034
16	0.025	27.28	0.002	0.001	0.075	0.099
17	0.022	29.223	0.001	0.209	0.115	0.193
18	0.014	35.847	0.007	0.123	0.15	0.002
19	0.014	36.985	0.002	0	0.018	0.002
20	0.004	71.898	0.984	0.028	0.021	0.014
21	0.011	41.868	0.003	0.026	0.009	0.036

Appendix: Table 3

Var prop D	Var prop E	Var prop F	Var prop G	Var prop H	Var prop I	Var prop J
0	0	0	0	0	0	0
0.193	0	0	0	0.001	0.001	0.009
0.119	0	0	0.002	0.001	0.015	0
0.001	0.002	0.003	0.027	0	0	0.006
0.114	0.014	0.006	0.141	0	0.006	0.306
0.146	0	0	0.009	0.004	0.189	0.042
0.003	0.006	0.002	0.112	0.001	0.035	0.448
0	0	0	0	0	0	0.035
0.036	0	0.001	0.006	0	0	0.043
0.008	0	0.001	0.008	0	0.061	0.005
0.196	0.004	0.004	0.001	0.007	0.006	0
0.014	0	0.003	0.061	0.001	0.004	0.001
0.052	0.121	0.026	0.243	0.019	0.001	0
0.005	0.07	0.052	0.137	0.063	0.006	0.006
0.018	0.113	0.003	0.112	0.204	0.051	0.032
0	0.029	0.03	0.09	0.13	0.019	0.005
0.044	0.051	0.015	0.019	0.304	0.115	0.005
0.001	0.126	0.144	0.011	0.005	0.002	0.002
0	0.437	0.652	0.013	0.007	0.001	0.001
0.03	0.018	0.01	0.008	0.237	0.16	0.036
0.001	0.007	0.049	0	0.004	0.003	0.018

Appendix: Table 3

Var prop K	Var prop L	Var prop M	Var prop N	Var prop O	Var prop P	Var prop Q
0	0	0	0	0	0	0
0	0.004	0.005	0	0	0	0.001
0.011	0.005	0.006	0.004	0.004	0.008	0.003
0.005	0.004	0.007	0.001	0	0	0.035
0.006	0	0.001	0.003	0.001	0.001	0.003
0.009	0.012	0.017	0	0.007	0.014	0.058
0.016	0	0.001	0.005	0.001	0	0.114
0.001	0	0.001	0	0.015	0.062	0.161
0.009	0.003	0.007	0.296	0.004	0.016	0
0.036	0.001	0.004	0.212	0	0.003	0.007
0	0	0.001	0.189	0.003	0.002	0.042
0.091	0.003	0.003	0.012	0.058	0.034	0.096
0.258	0	0.002	0.001	0.073	0.006	0.009
0.383	0.001	0	0.122	0.006	0.161	0.1
0.008	0.002	0.001	0.07	0.003	0.051	0.023
0.011	0	0.001	0	0.645	0.243	0.004
0.005	0.004	0.009	0.051	0.115	0.304	0.17
0.019	0.007	0	0.004	0.017	0.043	0.026
0.026	0.05	0.043	0.003	0.039	0	0
0.103	0.027	0.001	0.023	0.008	0.05	0.141
0.004	0.876	0.891	0.004	0.001	0.002	0.007

Var prop R	Var prop S	Var prop T
0	0	0
0.001	0	0.005
0.001	0.001	0.004
0.069	0.018	0
0.002	0	0.001
0.021	0.004	0
0	0	0.01
0.008	0.002	0.035
0.004	0.001	0.339
0.001	0	0.507
0.008	0.009	0.001
0.001	0.001	0.058
0.015	0	0.014
0	0	0.001
0.009	0.012	0
0.003	0	0
0.007	0.016	0.007
0.019	0.542	0.003
0.026	0.171	0.007
0.103	0.183	0.012
0.004	0.04	0

TABLE 4

Number in model	C(p)	AIC	R-square	Adjusted R-square	MSE	Variables in model
1	58.34	1327	0.499	0.499	22.757	A
1	409.98	1561	0.132	0.13	39.487	B
1	427.75	1570	0.113	0.111	40.332	L
1	443.88	1578	0.096	0.094	41.099	M
1	473.42	1592	0.066	0.063	42.505	N
2	38.641	1309	0.522	0.52	21.777	AK
2	43.173	1313	0.517	0.515	21.993	AL
2	44.139	1314	0.517	0.514	22.04	AM
2	44.402	1315	0.516	0.514	22.051	AB
2	49.952	1320	0.511	0.508	22.316	AJ
3	20.728	1292	0.543	0.54	20.877	ABK
3	30.807	1302	0.533	0.529	21.358	AJK
3	31.535	1303	0.532	0.529	21.393	ABL
3	32.671	1304	0.531	0.527	21.447	ABM
3	34.507	1306	0.529	0.525	21.535	AKL
4	16.219	1288	0.55	0.546	20.614	ABJK
4	18.681	1290	0.547	0.543	20.732	ABHK
4	18.982	1291	0.547	0.543	20.747	ABKL
4	19.53	1291	0.546	0.542	20.773	ABKM
4	20.105	1292	0.546	0.541	20.801	ABGK
5	11.211	1283	0.557	0.552	20.327	ABHJK
5	11.823	1284	0.557	0.551	20.357	ABHKL
5	13.303	1285	0.555	0.55	20.428	ABHKM
5	14.365	1286	0.554	0.549	20.479	ABGJK
5	14.803	1287	0.554	0.548	20.499	ABJKL
6	7.162	1279	0.564	0.557	20.085	ABHJKL
6	7.867	1280	0.563	0.557	20.119	ABHJKL
6	10.064	1282	0.561	0.554	20.225	ABGHJK
6	10.659	1282	0.56	0.554	20.253	ABHIJK
6	10.871	1283	0.56	0.553	20.263	ABHJKO

7	5.874	1278	0.567	0.559	19.974	ABGHJKL
7	6.287	1278	0.567	0.559	19.994	ABGHJKM
7	6.923	1279	0.566	0.559	20.025	ABHIJKL
7	7.4	1279	0.565	0.558	20.048	ABHJKLR
7	7.559	1279	0.565	0.558	20.056	ABCHJKL
8	6.017	1278	0.569	0.561	19.932	ABGHIJKL
8	6.182	1278	0.569	0.561	19.941	ABGHJKLR
8	6.446	1278	0.569	0.56	19.953	ABGHJKLO
8	6.5	1278	0.569	0.56	19.956	ABGHIJKM
8	6.519	1278	0.569	0.56	19.957	ABFGHJKL
9	**5.587**	**1277**	0.572	0.562	19.863	ABGHJKLRS
9	5.815	**1277**	0.571	0.562	19.874	ABGHJKMRS
9	6.314	1278	0.571	0.562	19.899	ABGHJKLOR
9	6.442	1278	0.571	0.561	19.904	ABEGHJKLR
9	6.602	1278	0.571	0.561	19.912	ABGHIJKLR
10	5.677	**1277**	0.574	0.563	19.818	ABEGHJKLRS
10	5.966	**1277**	0.573	0.563	19.832	ABEGHJKMRS
10	5.992	1278	0.573	0.563	19.834	ABGHJKLORS
10	6.073	1278	0.573	0.563	19.837	ABFGHJKLRS
10	6.204	1278	0.573	0.563	19.844	ABGHJKLORS
11	6.427	1278	0.575	**0.564**	19.805	ABEGHJKLORS
11	6.453	1278	0.575	**0.564**	19.807	ABFGHJKLORS
11	6.691	1278	0.575	0.563	19.818	ABEGHJKMORS
11	6.763	1278	0.575	0.563	19.822	ABFGHJKMORS
11	6.807	1278	0.575	0.563	19.824	ABEGHIJKLRS
12	7.356	1279	0.576	**0.564**	**19.801**	ABEGHJKLOQRS
12	7.456	1279	0.576	**0.564**	19.806	ABFGHJKLOQRS
12	7.494	1279	0.576	0.564	19.808	ABEGHJKMOQRS
12	7.647	1279	0.576	0.564	19.816	ABFGHJKMOQRS
12	7.715	1279	0.576	0.563	19.819	ABGHIJKLOQRS
13	8.372	1280	0.577	**0.564**	**19.801**	ABEGHIJKLOQRS
13	8.449	1280	0.577	**0.564**	19.805	ABFGHIJKLOQRS
13	8.574	1280	0.577	0.564	19.811	ABEGHIJKMOQRS
13	8.694	1280	0.577	0.563	19.817	ABFGHIJKMOQRS
13	8.948	1280	0.577	0.563	19.829	ABEFGHJKLOQRS

14	10.006	1281	0.577	0.563	19.832	ABEFGHIJKLOQRS
14	10.097	1281	0.577	0.563	19.836	ABEGHIJKLOQRST
14	10.162	1282	0.577	0.563	19.839	ABEGHIJKLNOQRS
14	10.19	1282	0.577	0.563	19.841	ABCEGHIJKLOQRS
14	10.232	1282	0.577	0.563	19.843	ABEGHIJKLOPQRS
15	11.779	1283	**0.578**	0.562	19.869	ABEFGHIJKLOQRST
15	11.782	1283	**0.578**	0.562	19.869	ABCEFGHIJKLOQRS
15	11.797	1283	**0.578**	0.562	19.87	ABEFGHIJKLNOQRS
15	11.873	1283	**0.578**	0.562	19.874	ABEFGHIJKLOPQRS
15	11.898	1283	**0.578**	0.562	19.875	ABEFGHIJKLMOQRS
16	13.568	1285	**0.578**	0.561	19.908	ABCEFGHIJKLNOQRS
16	13.582	1285	**0.578**	0.561	19.908	ABCEFGHIJKLNOQRS
16	13.585	1285	**0.578**	0.561	19.909	ABCEFGHIJKLOQRST
16	13.595	1285	**0.578**	0.561	19.909	ABEFGHIJKLNOQRST
16	13.614	1285	**0.578**	0.561	19.91	ABEFGHIJKLOPQRST
17	15.335	1287	**0.578**	0.561	19.945	ABCEFGHIJKLNOPQRS
17	15.351	1287	**0.578**	0.56	19.946	ABCEFGHIJKLOPQRST
17	15.394	1287	**0.578**	0.56	19.948	ABCEFGHIJKLNOQRST
17	15.405	1287	**0.578**	0.56	19.949	ABEFGHIJKLNOPQRST
17	15.46	1287	**0.578**	0.56	19.951	ABCDEFGHIJKLOPQRS
18	17.131	1288	**0.578**	0.56	19.984	ABCEFGHIJKLNOPQRST
18	17.245	1289	**0.578**	0.56	19.99	ABCEFGHIJKLMNOPQRS
18	17.268	1289	**0.578**	0.559	19.991	ABCDEFGHIJKLNOPQRS
18	17.276	1289	**0.578**	0.559	19.991	ABCEFGHIJKLMOPQRST
18	17.279	1289	**0.578**	0.559	19.992	ABCDEFGHIJKLOPQRST
19	19.042	1290	**0.578**	0.559	20.029	ABCDEFGHIKLMNOPQRST
19	19.096	1290	**0.578**	0.559	20.032	ABCDEFGHIJKLNOPQRST
19	19.167	1290	**0.578**	0.559	20.036	ABCDEFGHIJKLMNOPQRS
19	19.192	1290	**0.578**	0.558	20.037	ABCDEFGHIJKLMOPQRST
19	19.199	1290	**0.578**	0.558	20.037	ABDEFGHIJKLMNOPQRST
20	21	1292	**0.578**	0.558	20.077	ABCDEFGHIJKLMNOPQRST

TABLE 5

Analysis of Variance

Dependent variable: IDENTITY

Source	DF	Sum of squares	Mean square	F value	Prob > F
Model	9	10699.62	1188.846	57.747	0.0001
Error	425	8749.465	20.5869		
C Total	434	19449.08			

Root MSE	4.537	**R-square**	0.55
Dep mean	36.713	**Adj R-sq**	0.541
C.V.	12.359		

Parameter estimates

Variable	DF	Parameter estimate	Standard error	T for Ho: Parameter=0	Prob > /T/
Intercept	1	12.381	2.243	5.521	0.0001
A	1	0.786	0.049	15.853	0.0001
B	1	0.321	0.068	4.751	0.0001
G	1	0.197	0.094	2.085	0.0376
H	1	0.0187	0.074	2.531	0.0117
J	1	0.0541	0.173	3.133	0.0019
K	1	-0.198	0.041	-4.806	0.0001
L	1	0.205	0.099	2.061	0.0399
R	1	0.408	0.18	2.267	0.0239
S	1	-0.327	0.163	-2.012	0.0448

TABLE 6

Ten predictor variables for the prediction of pounds of steam used monthly in a production process

X1	Pounds of real fatty acid in storage per month
X2	Pounds of crude glycerine made
X3	Average wind velocity in miles per hour
X4	Calendar days per month
X5	Operating days per month
X6	Days below 32° F
X7	Average atmospheric temperature, degrees Fahrenheit
X8	(Average wind velocity)^2
X9	Number of start-ups
X10	Unit variables (intercept)

TABLE 7

Collinearity diagnostics

Number	Eigenvalue	Condition number	Var prop Intercept	Var prop X1	Var prop X2	Var prop X3
1	10.058	1	0	0	0	0
2	0.686	3.83	0	0	0	0
3	0.165	7.804	0	0	0	0
4	0.059	13.095	0	0.002	0.004	0
5	0.015	26.273	0	0.007	0.016	0
6	0.011	30.245	0	0.002	0.004	0.001
7	0.005	44.796	0	0.054	0.023	0.001
8	0.001	97.728	0	0.547	0.552	0.021
9	0.001	156.975	0	0.014	0.135	0.793
10	0	228.673	0	0.373	0.265	0.183
11	0	3171473	1	0	0	0

Var prop X4	Var prop X5	Var prop X6	Var prop X7	Var prop X8	Var prop X9	Var prop X10
0	0	0	0	0	0	0
0	0	0.113	0.004	0	0	0
0	0	0.135	0	0.007	0.003	0
0	0.005	0.089	0.093	0	0.042	0
0	0.018	0.008	0.008	0	0.721	0
0.005	0.044	0.414	0.485	0.004	0.007	0
0	0.591	0.199	0.108	0.007	0.007	0
0.001	0.152	0.016	0.105	0.022	0.014	0
0.14	0.14	0.001	0.074	0.813	0.136	0
0.853	0.05	0.024	0.123	0.146	0.069	0
0	0	0	0	0	0	0

TABLE 8

Number in model	C(p)	R-square	Variables in model
1	94.753	0.41	X6
1	118.9	0.287	X5
1	131.16	0.225	X3
1	144.78	0.156	X8
1	146.51	0.147	X1
2	30.214	0.749	X7 X9
2	31.877	0.741	X6 X7
2	36.021	0.719	X7 X8
2	44.658	0.676	X1 X6
2	54.405	0.626	X5 X6
3	**5.586**	0.885	X4 X5 X7
3	6.617	0.879	X1 X5 X7
3	9.492	0.865	X1 X7 X9
3	9.743	0.864	X1 X4 X7
3	9.765	0.864	X5 X7 X8
4	5.911	0.893	X1 X4 X5 X7
4	6.31	0.891	X1 X5 X7 X9
4	6.602	0.89	X4 X5 X6 X7
4	6.899	0.888	X4 X5 X7 X8
4	6.983	0.888	X1 X2 X5 X7
5	6.696	0.899	X1 X2 X5 X7 X9
5	6.857	0.898	X1 X4 X5 X7 X9
5	7.109	0.897	X1 X2 X4 X5 X7
5	7.118	0.897	X1 X5 X7 X8 X9
5	7.139	0.897	X3 X4 X5 X7 X8
6	**5.368**	**0.917**	X1 X3 X5 X7 X8 X9
6	7.135	0.907	X1 X2 X5 X7 X8 X9
6	7.336	0.907	X1 X3 X4 X5 X7 X8
6	7.58	0.905	X1 X2 X3 X5 X7 X9
6	7.882	0.904	X1 X2 X4 X5 X7 X9

7	6.721	0.92	X1 X3 X4 X5 X7 X8 X9
7	6.752	0.919	X1 X3 X5 X6 X7 X8 X9
7	6.968	0.919	X1 X2 X3 X5 X7 X8 X9
7	8.285	0.912	X2 X3 X4 X5 X7 X8 X9
7	8.471	0.911	X1 X2 X5 X6 X7 X8 X9
8	8.214	0.923	X1 X3 X4 X5 X6 X7 X8 X9
8	8.305	0.922	X1 X2 X3 X5 X6 X7 X8 X9
8	8.557	0.921	X1 X2 X3 X4 X5 X7 X8 X9
8	9.572	0.916	X2 X3 X4 X5 X6 X7 X8 X9
8	10.267	0.912	X1 X2 X4 X5 X6 X7 X8 X9
9	10	**0.924**	X1 X2 X3 X4 X5 X6 X7 X8 X9

TABLE 9

Analysis of Variance

Dependent variable: IDENTITY

Source	DF	Sum of squares	Mean square	F value	Prob > F
Model	3	280161	90387	5.529	0.0059
Error	21	350655	1.697		
C Total	24	630816			

Root MSE	1.303	**R-square**	0.441
Dep mean	9.424	**Adj R-sq**	0.362
C.V.	13.826		

Parameter estimates

Variable	DF	Parameter estimate	Standard error	T for Ho: Parameter=0	Prob > /T/
Intercept	1	0.978	2.217	0.441	0.664
X4	1	1.638	0.984	1.665	0.111
X5	1	-5.106	6.403	-0.797	0.434
X7	1	0.484	0.0153	3.161	0.005

Analysis of Variance

Dependent variable: IDENTITY

Source	DF	Sum of squares	Mean square	F value	Prob > F
Model	6	47.659	7.943	8.85	0.0001
Error	18	16.156	0.897		
C Total	24	63.815			

Root MSE	0.947	**R-square**	0.747
Dep mean	9.424	**Adj R-sq**	0.662
C.V.	10.053		

Parameter estimates

Variable	DF	Parameter estimate	Standard error	T for Ho: Parameter=0	Prob > /T/
Intercept	1	-9.168	9.186	-0.998	0.331
X1	1	1.249	0.84	1.487	0.154
X3	1	-5.464	6.245	-0.875	0.393
X5	1	0.135	0.148	0.908	0.376
X7	1	0.339	0.311	1.088	0.291
X8	1	0.175	0.125	1.406	0.177
X9	1	0.091	0.024	3.835	0.001

---oOo---